Universitat de Girona

Estadística

Departament d'Informàtica, Matemàtica Aplicada i Estadística (IMAE)

Aquest material Estadística és el resultat d'una beca de col·laboració per a l'elaboració de material d'Estadística sota el Programa de Millora de la Docència impulsat pel rectorat de la UdG amb el Departament d'Informàtica, Matemàtica Aplicada i Estadística

Aquest treball l'han fet la Maria Simon Font i l'Èric Pairet Artau, alumnes d'enginyeries de la Escola Politècnica Superior, basant-se en el material ja existent de teoria, problemes i pràctiques que ha elaborat el professorat de l'àrea d'Estadística i Investigació Operativa.

El resultat final ha estat supervisat pels propis professors de l'àrea i una primera versió d'aquests Apunts d'Estadística és la que teniu entre mans.

Girona, setembre del 2014

ÍNDEX

TEMA 1: Tècniques d'anàlisi exploratòria de dades univariants

TEORIA

1. Dades i variables

Una dada és aquella informació que descriu una peculiaritat d'un individu. Així doncs, totes les possibles informacions que puguin definir una mateixa característica formaran una variable. Depenent de la característica a definir, obtindrem variables de diferents tipus.

1.1. Variable categòrica

Una variable categòrica, qualitativa o nominal o factor, informa sobre el grup o categoria al que pertany un individu. Aquesta variable no és interpretable com a un número, tot i que a vegades es pot codificar amb números

Segons la relació entre les categories d'una mateixa variable podem diferenciar tres tipus de variables categòriques:

- **Nominal:** categories entre les que no es poden establir relacions d'ordre, com per exemple, la tipologia de producte, l'olor...
- **Ordinal:** categories que poden ser ordenades segons la seva magnitud, com per exemple, les escales d'abundància, el tipus de magnitud d'un defecte...
- **Binàries:** categories que són dicotòmiques, com per exemple, el gènere (home-dona), la presència-absència d'un determinat defecte de fabricació o d'un contaminant...

1.2. Variable numèrica

Una variable numèrica o quantitativa pren valors numèrics, amb els quals té sentit fer operacions aritmètiques i establir relacions d'ordre.

Segons els valors numèrics que puguin agafar les variables podem diferenciar tres tipus de variables numèriques:

- **Contínues:** poden agafar infinits valors, com per exemple, longituds, pesos, temperatures… No ho hem de confondre amb els valors que les mesurem i la precisió de la mesura.
- **Discretes:** només poden agafar un número fix de valors i mai valors intermedis entre aquests, com per exemple, el nombre de defectes, les unitats reprocessades...

- **Mixtes:** és contínua excepte en un conjunt petit de punts on és discreta, com per exemple, la quantitat de pluja recollida en un observatori per dia (el rendiment que se'n treu d'un vedell).

1.3. Variable multivariant

Una variable multivariant és un conjunt de dues o més variables observades sobre un mateix individu. Tot depenent del tipus de variables agrupades, podem diferenciar tres tipus de variables multivariants:

- **Homogènies:** agrupació de variables del mateix tipus, com per exemple, les dimensions en mm d'una peça, els nivells de contaminació per substàncies químiques...
- **Heterogènies:** agrupació de variables de diferents tipus, com per exemple, la caracterització d'un estany (pH, conductivitat, profunditat, nivell de contaminació...) o d'un vehicle (velocitat màxima, potencia, dimensions, consum...).
- **Composicionals:** variable que defineix la participació relativa que cada component té dins un total, com per exemple, la composició d'un contaminant, la composició d'un sòl, la composició dels tipus de residus...

Exemple: *La següent taula conté informació sobre els treballadors d'una empresa:*

Nom	***Triennis***	***Sexe***	***Estudis***	***Salari***	***Escala salarial***
Costa, Ester	*3*	*Dona*	*Universitaris*	*52100*	*1*
Fernández, Joan	*2*	*Home*	*Universitaris*	*27350*	*2*
Kaur, Irina	*2*	*Dona*	*Secundaris*	*18250*	*3*
Martí, Josep	*4*	*Home*	*Secundaris*	*47600*	*1*

Cada fila és un individu descrit; en l'exemple són els treballadors. Cada columna pertany a una variable, que conté el valor característic de la variable per cada treballador.

La variable triennis és una variable numèrica discreta, la variable sexe és una variable categòrica binària, la variable estudis és una variable categòrica ordinal, la variable salari és una variable numèrica contínua i la variable escala salarial és una variable categòrica, possiblement ordinal.

2. Taules, gràfics i diagrames

Les taules, gràfiques i diagrames són útils per a l'agrupament i visualització d'un conjunt de dades d'una forma gràfica i esquematitzada, que ens proporciona una idea ràpida del repartiment que segueix aquest conjunt.

2.1. Taula de freqüències

La taula de freqüències ens mostra per una dada categòrica, per una dada numèrica discreta o per un interval de dades numèriques contínues, les següents freqüències:

- **Freqüència absoluta:** nombre de vegades que observem un valor.
- **Freqüència relativa**: nombre de vegades que observem un valor en relació a les observacions totals. És la freqüència absoluta dividida pel nombre d'observacions fetes. Sempre serà un valor entre 0 i 1.
- **Freqüència percentual:** expressió de la freqüència relativa en tant per cent. Sempre serà un valor entre 0 i 100.
- **Freqüència acumulada:** nombre de vegades que observem un valor o els valors més petits que ell. Pot ser en absolut, relatiu o percentual. Només té sentit per a dades amb ordre (numèriques o categòriques ordinals).

En les taules de freqüències formades per dades categòriques, cada categoria formarà una fila, on es poden calcular les tres freqüències corresponents. En el cas de treballar amb dades categòriques ordinals es poden calcular les quatre freqüències.

Exemple: *Es realitza una taula de freqüències a partir d'un conjunt de dades categòriques. Aquest conté el sistema operatiu utilitzat per 50 ordinadors.*

Conjunt de dades:

NT LINUX UNIX XP XP LINUX XP XP NT NT NT XP UNIX XP NT XP XP LINUX XP NT UNIX XP XP NT LINUX XP NT XP LINUX XP XP XP LINUX XP XP NT UNIX NT NT XP UNIX NT NT XP UNIX XP NT XP XP NT

Taula de freqüències:

Categoria	Freqüència Absoluta	Freqüència Relativa	Freqüència Percentual
NT	15	0.30	30 %
LINUX	6	0.12	12 %
XP	23	0.46	46 %
UNIX	6	0.12	12 %
	50	1.00	100%

Per les taules formades de dades numèriques discretes, el procediment a seguir i la constitució de la taula serà igual a una taula de freqüències de dades categòriques ordinals.

Per les taules formades de dades numèriques contínues, les dades s'hauran d'agrupar en intervals. Definint n el nombre de dades totals i k com el nombre d'intervals, es decidirà aquest últim amb un dels següents criteris:

Criteri 1	Criteri 2		Criteri 3
$k = \sqrt{n}$	$n < 50$ $50 \leq n < 100$ $100 \leq n < 250$ $n \geq 250$	$5 \leq k \leq 7$ $6 \leq k \leq 10$ $7 \leq k \leq 12$ $10 \leq k \leq 20$	Regla Sturges: $k = 1 + \frac{\ln(n)}{\ln(2)}$

Un cop sapiguem el nombre d'intervals que ha de tenir la taula de freqüències, haurem de calcular l'amplada d'aquests. Definint h com l'amplada de l'interval, aquest es calcularà com:

$$h = \frac{(\text{dada màxima} - \text{dada mínima})}{k}$$

Exemple: *Volem realitzar una taula de freqüències a partir d'un conjunt de dades numèriques. Aquest conté la longitud en centímetres de 90 peces fabricades per una màquina.*

Conjunt de dades sense ordenar:

49.2	53.9	50.0	44.5	42.2	42.3	32.3	31.3	60.9	47.5	43.5	37.9	41.1	57.6	40.2
45.3	51.7	52.3	45.7	53.4	51.0	45.7	45.9	50.0	32.5	67.2	55.1	59.6	48.6	50.3
45.1	46.8	47.4	38.3	41.5	44.0	62.2	62.9	56.3	35.8	38.3	33.5	48.5	47.4	49.6
41.3	55.2	52.1	34.3	31.6	38.2	46.0	47.0	41.2	39.8	48.4	49.2	32.8	47.9	43.3
49.3	54.5	54.1	44.5	46.2	44.4	45.1	41.5	43.3	39.1	39.1	41.6	43.1	43.7	48.8
37.2	33.6	28.7	33.8	37.4	43.5	44.2	53.0	45.1	51.9	50.6	48.5	39.0	47.3	48.8

Conjunt de dades ordenat:

28.7	31.3	31.6	32.3	32.5	32.8	33.5	33.6	33.8	34.3	35.8	37.2	37.4	37.9	38.2
38.3	38.3	39.0	39.1	39.1	39.8	40.2	41.1	41.2	41.3	41.5	41.5	41.6	42.2	42.3
43.1	43.3	43.4	43.5	43.5	43.7	44.0	44.2	44.4	44.5	44.5	45.1	45.1	45.1	45.3
45.7	45.7	45.9	46.0	46.2	46.8	47.0	47.3	47.4	47.4	47.5	47.9	48.4	48.5	48.5
48.6	48.8	48.8	49.2	49.2	49.3	49.6	50.0	50.0	50.3	50.6	51.0	51.7	51.9	52.1
52.3	53.0	53.4	53.9	54.1	54.5	55.1	55.2	56.3	57.6	59.6	60.9	62.2	62.9	67.2

Primer de tot decidim el nombre d'intervals en que dividirem el conjunt de dades. Utilitzem el criteri 1:

$$k = \sqrt{90} = 9.486 \quad \rightarrow \quad \text{Arrodonirem a l'alça tot agafant 10 intervals.}$$

Seguidament es decideix l'amplada de l'interval:

$$h = \frac{(67.2 - 28.7)}{10} = 3.85 \quad \rightarrow \quad \text{Arrodonirem a 4.}$$

L'amplada no pot ser amb una precisió més gran que la precisió amb la que estem treballant. En el nostre cas h no el podrem prendre com 3.85, amb dos decimals, ja que les dades només les tenim mesurades amb un decimal.

Comencem el primer interval amb un valor fàcil d'operar i que contingui almenys el valor mínim 28.7, per exemple 28. Construïm la resta d'intervals i afegim tota la informació necessària.

Interval de classe	Marca de Classe C_i	Freq. Absoluta fa_i	Freq. Relativa f_i	Freq. Abs. Ac. Fa_i	Freq. Rel. Ac. F_i
[28.0, 32.0)	30.0	3	0.033	3	0.033
[32.0, 36.0)	34.0	8	0.089	11	0.122
[36.0, 40.0)	38.0	10	0.111	21	0.233
[40.0, 44.0)	42.0	15	0.167	36	0.400
[44.0, 48.0)	46.0	21	0.233	57	0.633
[48.0, 52.0)	50.0	17	0.189	74	0.822
[52.0, 56.0)	54.0	9	0.100	83	0.922
[56.0, 60.0)	58.0	3	0.033	86	0.955
[60.0, 64.0)	62.0	3	0.033	89	0.988
[64.0, 68.0)	66.0	1	0.012	90	1.000
		90	1.000		

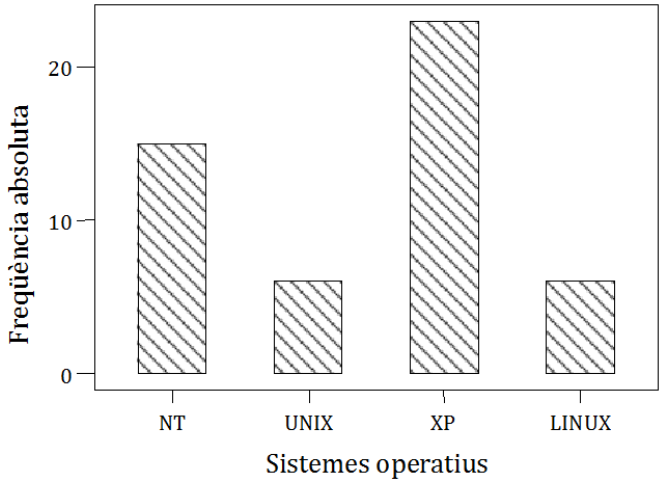
Quan fem intervals per a dades contínues, perdem els valors de les dades d'un interval. Per a tenir un representant definim la **marca de classe**, com al valor mig dels extrems de l'interval. Podem afegir una columna nova amb les marques de classe.

2.2. Gràfic de barres

El gràfic de barres mostra de forma visual per una variable categòrica o una quantativa discreta amb pocs valors diferents la quantitat de mostres de cada categoria. Cal tenir en compte que les barres entre elles se solen representar separades ja que l'eix de les x no ens indica cap seqüència de valors.

En aquests gràfics l'escala de l'eix vertical pot estar en freqüències absolutes, relatives o percentuals. No importa ja que són equivalents.

Exemple: *Continuant amb l'exemple de les dades del sistema operatiu de 50 ordinadors, realitzem un gràfic de barres.*



2.3. Histogrames i polígons de freqüències

L'**histograma** és un gràfic que representa de forma visual freqüències la distribució d'una variable numèrica contínua. En aquest gràfic es dibuixa per a cada interval de classe un rectangle que té per base l'interval i per alçada la freqüència de la classe. L'àrea de cada rectangle és proporcional a la freqüència que es tracti en l'histograma.

Una representació equivalent a l'histograma és el **polígon de freqüències,** línia que uneix els punts determinats per la freqüència i un element de l'interval i ens mostra la tendència que segueix la distribució de la variable.

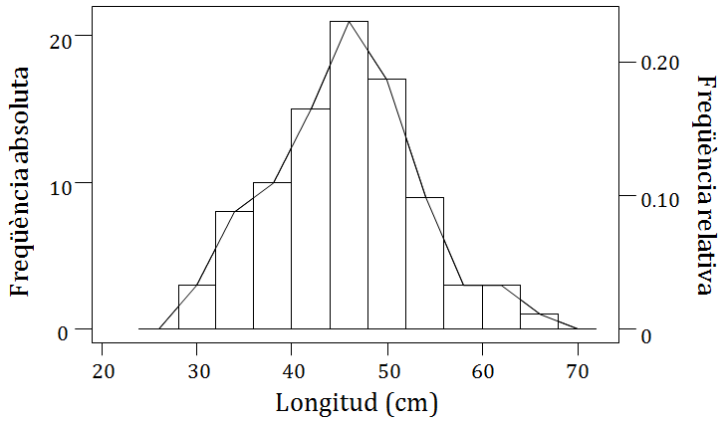
- **Histograma i polígon de freqüències:** aquest histograma tracta la freqüència absoluta i la relativa en relació la quantitat de mostres per cada marca de classe d'un interval.

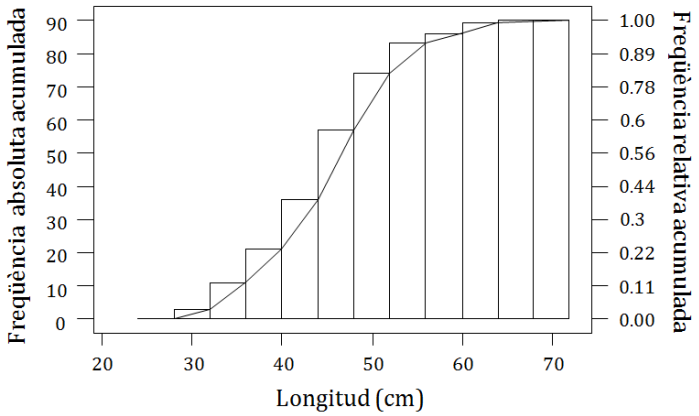
  El polígon de freqüències uneix els punts determinats pels la marca de classe de l'interval i la freqüència.

- **Histograma i polígon de freqüències acumulades:** aquest histograma tracta la freqüència absoluta acumulada i la relativa acumulada en relació la quantitat de mostres per cada marca de classe d'un interval.

  El polígon de freqüències uneix l'extrem dret superior de tots els rectangles, és a dir, el punt determinat per l'extrem de l'interval i la freqüència en aquell interval. El perfil resultant del qual rep el nom d'**ogiva**.

Exemple: *Continuant amb l'exemple de les dades de la longitud de 90 peces fabricades per una màquina, realitzem els dos histogrames comentats anteriorment amb el seu corresponent polígon de freqüències. A més aprofitem que les freqüències són proporcionals per posar dos eixos verticals al gràfic un etiquetat amb freqüències absolutes i l'altre amb freqüències relatives.*



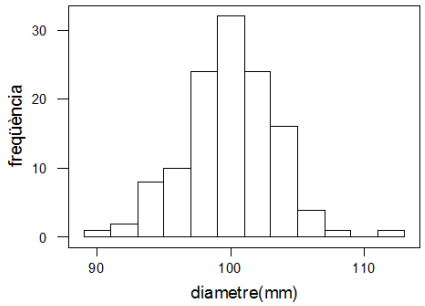


Un cop realitzat l'histograma podem observar quina distribució té les dades representades. Tot depenent de la forma obtinguda, parlarem que les dades segueixen un tipus de distribució o una altra.

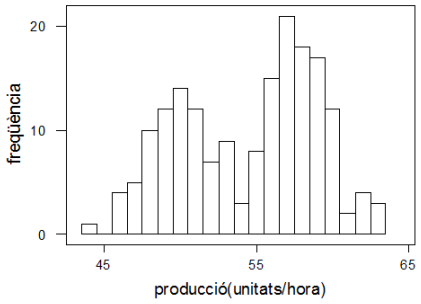
Parlarem de **distribució simètrica** si les dades es trobem distribuïdes simètricament en relació al seu centre, tenint a dreta i esquerra el mateix nombre de dades repartides de forma semblant. Si no hi ha aquesta simetria, parlarem d'una **distribució esbiaixada**. El **biaix** serà cap a la dreta o cap a l'esquerra si la cua de la distribució és a la dreta o a l'esquerra respectivament.

Si tenim una única dada més repetit direm que la distribució és **unimodal**. Si tenim dues dades molt repetides i prou separades, dins un mateix conjunt de dades, parlarem d'una **distribució bimodal**.

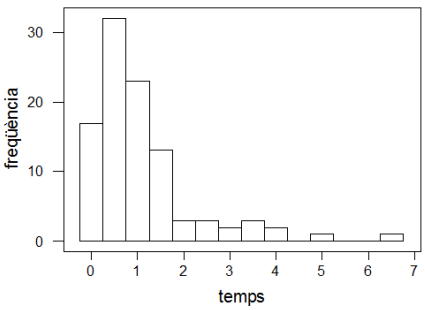
Distribució simètrica



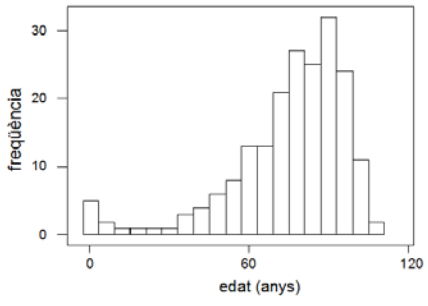
Distribució bimodal



Distribució esbiaixada a la dreta



Distribució esbiaixada a l'esquerra

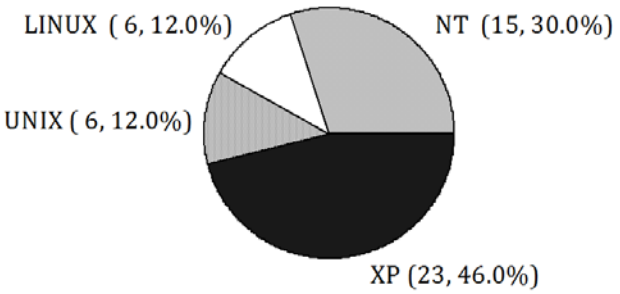


2.4. Diagrama de sectors

El diagrama de sectors mostra de forma genèrica i visual per una variable categòrica la quantitat de mostres de cada categoria.

L'àrea de tot el diagrama representa el 100% de les dades, i es fracciona de tal manera que cada categoria li pertoca un sector circular que té una àrea equivalent al percentatge de dades que té respecte el total.

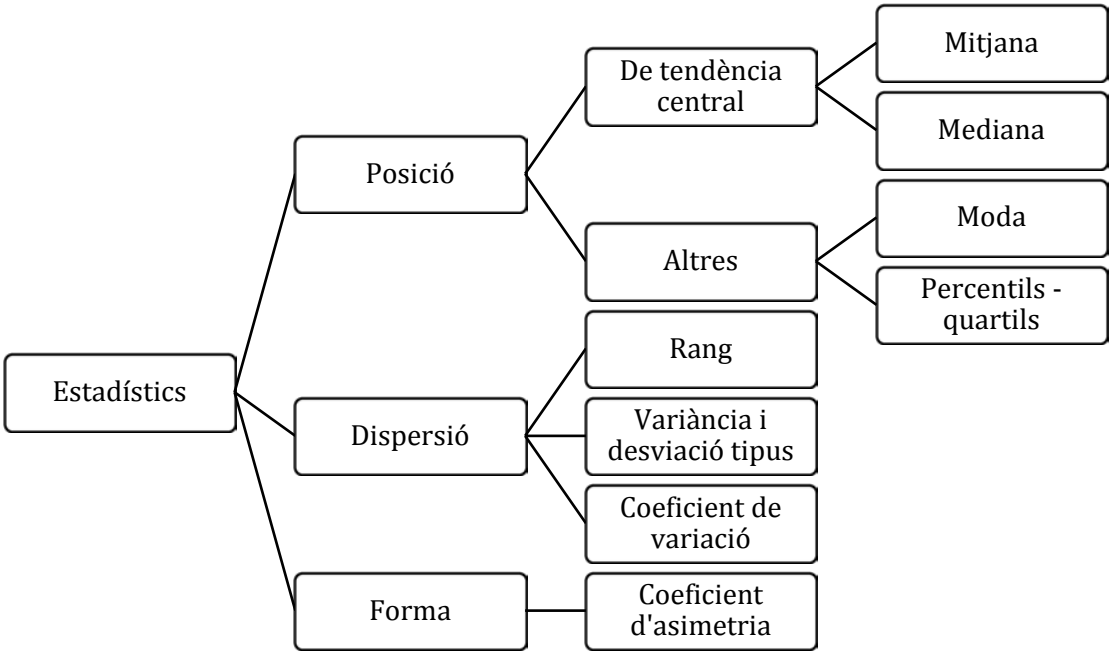
Exemple: *Continuant amb l'exemple de les dades del sistema operatiu de 50 ordinadors, realitzem un gràfic de barres.*



3. Estadístics

Entenem per estadístic aquell valor numèric calculat a partir de les dades d'una mostra que proporciona informació característica de la variable.

Classificarem els estadístics segons el tipus d'informació que proporcionin de la següent manera:



3.1. Mitjana, $\bar{x}$

La mitjana d'una mostra amb dades x_1, x_2, …, x_n de mida n és la suma de totes les observacions dividida per n:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Per dades agrupades, si definim c_i com la marca de classe, k el nombre de classes, n_i la freqüència absoluta de l'interval representat per la marca de classe i n el nombre total de dades:

$$\bar{x} = \frac{\sum_{i=1}^{k} c_i n_i}{n}$$

Característiques de la mitjana:

- Indica el centre de gravetat de la distribució.
- Té en compte totes les dades de la distribució.
- És molt sensible a la presència de dades atípiques o extremes.

D'acord amb aquest últim punt calculem la **mitjana truncada**, que s'obté calculant la mitjana de la variable després de suprimir el 5% dels valors superiors i el 5% dels inferiors (o un altre percentatge). Així, la mitjana que s'obté és més robusta davant la possible presència de dades atípiques. Si els valors obtinguts de la mitjana i la mitjana truncada són bastant diferents, és senyal que hi ha dades atípiques. Si són iguals, pot o no pot haver-hi dades atípiques.

Exemple: *Continuant amb l'exemple anterior, calculem les tres mitjanes esmentades.* Mitjana: $\bar{x} = 45.507$ cm, mitjana amb dades agrupades: $\bar{x} = 45.56$ cm, mitjana truncada al 5%: $\bar{x}\,\text{trunc}5\% = 45.33$ cm.

3.2. Mediana, Md

La mediana és el punt mig de les dades ordenades, és a dir, és el valor que té per sobre i per sota el 50% de les dades ordenades. Dades ordenades es refereix a tenir-les organitzades en ordre numèricament ascendent.

El càlcul de la mediana es realitzarà de la següent manera:

$$\text{Md} = x_{\left(\frac{n}{2}+0.5\right)}$$

Cal tenir en compte que el resultat de la fórmula serà la posició en què es troba el valor de la mediana. El resultat no serà sempre un valor enter. Si no és així, s'haurà de fer la mitjana o mitjana ponderada entre els valors immediatament inferior i superior per trobar el valor de la mediana.
Característiques de la mediana:

- Indica el centre de la distribució.
- No té en compte el valor de totes les observacions.
- És un estadístic robust a la presència de dades atípiques.

Exemple: *Continuant amb l'exemple anterior, calculem la mediana.* Md = 45.5 cm.

3.3. Moda

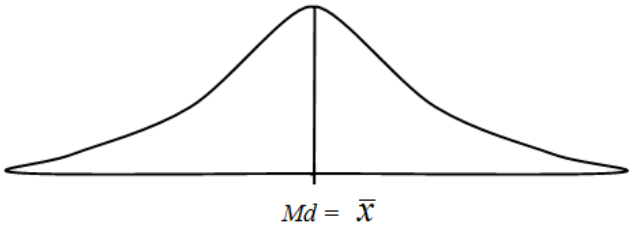
La moda és el valor de l'observació més freqüent. Cal tenir en compte que podem tenir més d'una moda, i que en el cas que totes les observacions d'una mostra tinguessin la mateixa freqüència, diríem que la mostra no té moda.

Exemple: *Continuant amb l'exemple anterior, calculem la moda.* Moda = 45.1 cm.

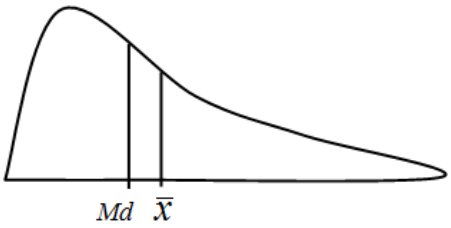
3.4. Centre i simetria

Comparant la mitjana amb la mediana podem definir el tipus de distribució que segueix una variable. Observem que el resultat d'aquesta comparació és el mateix que el deduït a partir de l'histograma i del diagrama de caixa:

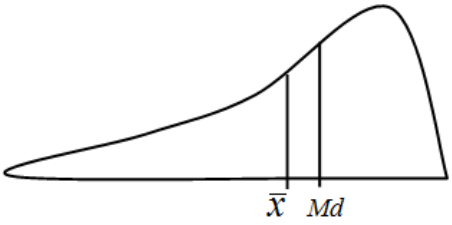
Distribució simètrica ($\bar{x} = Md$)



Distribució esbiaixada a la dreta ($\bar{x} > Md$)



Distribució esbiaixada a l'esquerra ($\bar{x} < Md$)



Hem d'anar amb compte si tenim una distribució bimodal resultant de la barreja de dues de les distribucions anteriors, la mitjana i la moda de les dades no ens servirien per identificar la distribució.

3.5. Percentils, P_q

Un percentil P_q és aquell valor que té per sota seu un q% de les dades ordenades d'una variable. El càlcul d'aquest estadístic es pot fer en dos sentits:

- **Volem saber el percentil corresponent a una certa dada:** tenint les dades ordenades, contem la posició que correspon a la dada i calculem un tant per cent. Depenent de si n és parell o senar, es calcularà diferent:

  Si n és senar:

  $$P_q(x_i) \text{ significa que } x_i \text{ acumula } P_{\left(\frac{100 \cdot i}{n+1}\right)}$$

  Si n és parell:

  $$P_q(x_i) \text{ significa que } x_i \text{ acumula } P_{\left(\frac{100 \cdot (i-0.5)}{n}\right)}$$

- **Volem saber la dada que li pertoca un percentil en concret:** calculem la posició que li pertoca un percentil a partir de les següents fórmules, que seran diferents depenent de si n és parell o senar:

  Si n és senar:

  $$P_q \text{ és el percentil corresponent a } x_{\left(\frac{q \cdot n+q}{100}\right)}$$

Si n és parell:

$$P_q \text{ és el percentil corresponent a } x_{\left(\frac{q \cdot n}{100}+0.5\right)}$$

Cal dir que la fórmula de càlcul dels percentils no és única, sinó que cada paquet estadístic o calculadora té la seva pròpia. Això comporta que els resultats poden diferir segons l'eina que s'utilitzi per realitzar el càlcul.

Quan es treballi amb dades agrupades, s'haurà de buscar l'interval que conté el percentil P_q a través de la columna de freqüències relatives acumulades. Tot seguit s'haurà de fer una interpolació lineal entre els valors límits de l'interval per obtenir la dada x_i que acumula el tant per cent q.

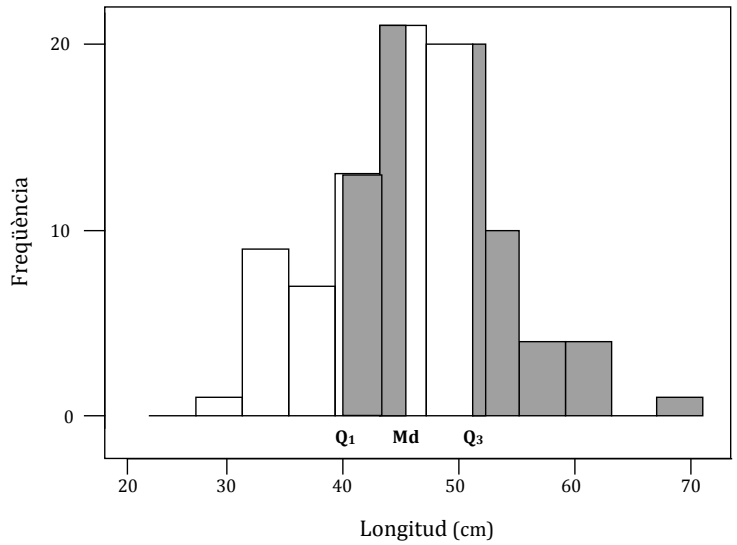
Exemple: *Continuant amb l'exemple anterior, calculem el percentil que acumula el 45% de les dades.* P_{45} = 44.86 cm.

3.6. Quartils, Q

Rep el nom de quartil aquell percentil que acumula el 25%, el 50% o el 75% del conjunt de dades. El percentil que acumula el 25% de les dades s'anomena 1r quartil (Q_1). El percentil que acumula el 50% de les dades és la mediana (Md). El percentil que acumula el 75% de les dades s'anomena 3r quartil (Q_3).

Els quartils Q_1, Md i Q_3, divideixen l'histograma en quatre parts d'igual àrea.

Exemple: *Continuant amb l'exemple anterior, calculem els tres quartils.* Q_1 = 41.1 cm, Md = 45.5 cm, Q_3 = 50 cm.



3.7. Variància, s^2

La variància o variància mostral és un estadístic que mesura el grau de dispersió de les dades al voltant de la mitjana $\bar{x}$. Diferenciem dues formes per el seu càlcul:

- **Variància:**

Conjunt de dades:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Dades agrupades:

$$s^2 = \frac{\sum_{i=1}^{k} n_i(c_i - \bar{x})^2}{n-1}$$

- **Variància no corregida:**

Conjunt de dades:

$$s^{*2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Dades agrupades:

$$s^{*2} = \frac{\sum_{i=1}^{k} n_i(c_i - \bar{x})^2}{n}$$

Exemple: *Continuant amb l'exemple anterior, calculem totes les variàncies explicades.* s^2 = 58.74 cm, s^2(dades agrupades) = 58.77 cm. s^{*2} = 58.09 cm, s^{*2}(dades agrupades) = 58.12 cm,

3.8. Desviació estàndard, s

La desviació estàndard, o desviació tipus, és l'arrel quadrada positiva de la variància. Així doncs, qualsevol tipus de desviació estàndard que s'hagi de calcular es farà seguint la següent fórmula:

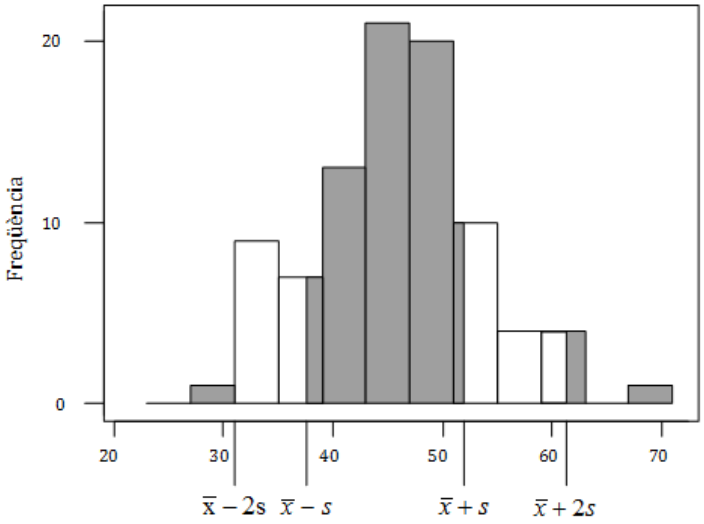
$$s = +\sqrt{s^2}$$

Observacions sobre s:

- s = 0 només quan no hi ha dispersió. Això passa quan totes les observacions són iguals. En cas contrari, sempre s > 0.
- s té les mateixes unitats que la variable d'estudi i per tant és més adequada que s^2, amb unitats al quadrat.
- s no és robusta en front de dades extremes, és a dir, es veu molt influenciada per aquestes.

Per un conjunt de dades que segueixi una distribució en forma de campana i n sigui gran, la desviació estàndard es pot interpretar de la següent manera:

- Entre $\bar{x}$ – s i $\bar{x}$ + s hi haurà aproximadament un 68% de les dades.
- Entre $\bar{x}$ – 2s i $\bar{x}$ + 2s hi haurà aproximadament un 95% de les dades.
- Entre $\bar{x}$ – 3s i $\bar{x}$ + 3s hi haurà aproximadament un 100% de les dades.



Exemple: *Continuant amb l'exemple anterior, calculem totes les desviacions estàndard possibles*. s = 7.66 cm, s(dades agrupades) = 7.67 cm s^* = 7.62 cm, s^*(dades agrupades) = 7.62 cm,.

També calculem el tant per cent de dades incloses en l'interval format per una, dos o tres desviacions d'allunyament per cada costat de la mitjana. Considerarem la mitjana de 45.507 cm i la desviació estàndard de 7.66 cm:

- Interval $[\bar{x} - s, \bar{x} + s]$ = [37.847, 53.23]. Inclou 64 dades (71.1%).
- Interval $[\bar{x} - 2s, \bar{x} + 2s]$ = [30.187, 60.827]. Inclou 85 dades (94.4%).
- Interval $[\bar{x} - 3s, \bar{x} + 3s]$ = [22.527, 68.487]. Inclou 90 dades (100%).

3.9. Coeficient de variació, CV

El coeficient de variació d'un grup de dades expressa la desviació estàndard com a tant per cent de la mitjana. Aquest es calcularà de la següent forma:

$$CV(\%) = \frac{s}{\bar{x}} \cdot 100$$

Exemple: *Continuant amb l'exemple anterior, calculem el coeficient de variació.* CV = 16.84%.

3.10. Coeficient d'asimetria, CA

El coeficient d'asimetria valora la simetria o no d'una distribució d'un conjunt de dades, representant-la amb un valor numèric. Aquest es calcularà de la següent forma:

$$CA = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Depenent del resultat d'aquest coeficient, podrem tenir una idea de la distribució del grup de dades:

- Si $CA \approx 0 \Rightarrow$ distribució simètrica.
- Si $CA \gg 0 \Rightarrow$ distribució esbiaixada a la dreta.
- Si $CA \ll 0 \Rightarrow$ distribució esbiaixada a l'esquerra.

Exemple: *Continuant amb l'exemple anterior, calculem el coeficient d'asimetria.* CA = 0.19, pel qual podem dir que la distribució de les dades és bastant simètrica.

3.11. Diagrama de caixa

El diagrama de caixa és un gràfic que representa de forma visual on es troben situats dins les diferents dades 5 diferents estadístics resum (mínim, Q1, mediana, Q3 i màxim). Deixarem definits i representats els estadístics relacionats amb aquest diagrama, els quals s'entendran a mida que s'avanci en el següent apartat.

- **L'amplitud interquartílica (AIQ o IQR):** espai entre dos punts del diagrama que inclou el 75% de les dades ordenades (Q_3) menys el primer 25% (Q_1), és a dir, un total del 50% de valors. Matemàticament es definirà com:

$$AIQ = Q_3 - Q_1$$

- **Dada atípica:** una dada atípica serà aquella que quedi molt desplaçada cap un dels dos extrems de la majoria de dades. Les representarem amb un asterisc (*). Sent *x* una dada, es considerarà atípica si compleix alguna de les dues següents condicions:

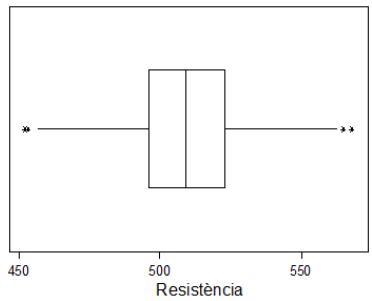
$$x < Q_1 - 1.5 \cdot AIQ \qquad x > Q_3 + 1.5 \cdot AIQ$$

- **Límits efectius:** són les dues primeres dades que incompleixen el criteri de dada atípica. Anomenant l_i a la dada més petita en incomplir-ho i l_s a la ultima dada més gran en fer-ho, podem escriure el següent:

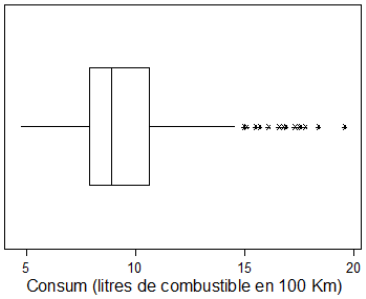
$$l_i \geq Q_1 - 1.5 \cdot AIQ \qquad l_s \leq Q_3 + 1.5 \cdot AIQ$$

Un cop realitzat el diagrama de caixa, on marquem una caixa central amb els Quartils i la Mediana, uns braços o bigotis que arriben fins als límits efectius i les dades atípiques, podem observar quina tendència té la distribució de dades representada. Depenent de la forma obtinguda, parlarem de diferents distribucions:

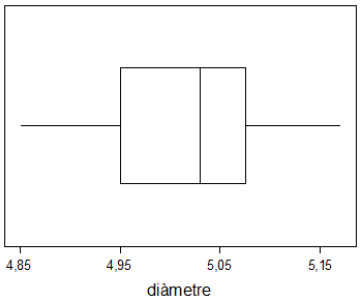
Distribució simètrica



Distribució esbiaixada a la dreta

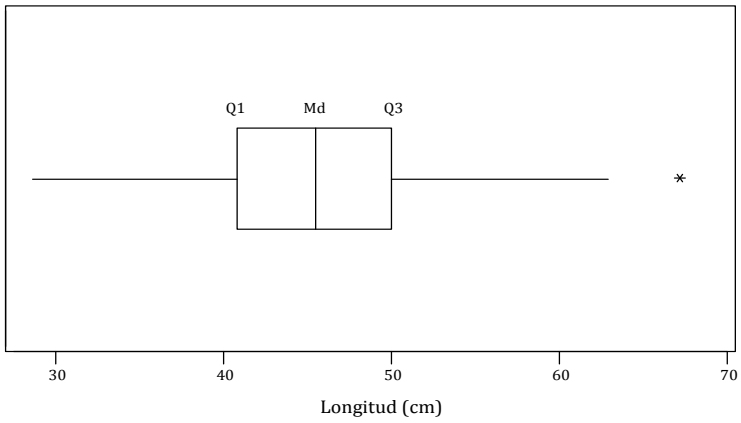


Distribució esbiaixada a l'esquerra



Podem observar que tant a partir dels diagrames de caixa com dels histogrames es podrà identificar quin tipus de distribució segueixen les dades.

Exemple: *Continuant amb l'exemple de les dades de la longitud de 90 peces fabricades per una màquina, realitzem el diagrama de caixa corresponent.*



A partir del diagrama de caixa podem observar que les dades segueixen una distribució bastant simètrica i que únicament hi ha una dada atípica, amb un valor aproximat de 67 cm.

4. Transformacions

L'objectiu de la transformació d'una variable és millorar la seva anàlisi estadística en aspectes com són:

- Que les mesures de tendència central (mitjana i mediana) estiguin realment situades en el centre de la distribució de les dades, és a dir, augmentar la simetria.
- Que les mesures de dispersió mostrin la variabilitat respecte al centre.
- Que la forma de la distribució de les dades s'assemblin més a una distribució en forma de campana, com té la llei Normal, que veurem més endavant.
- Que millori la relació amb altres variables.

4.1. Transformacions lineals. Estandardització

Tenim un conjunt de dades $x_1, ..., x_n$ que tenen com a mitjana $\bar{x}$ i desviació estàndard s_x. Si s'aplica la transformació lineal **t = k·x + a** per a cada dada:

$$x_1 \rightarrow t_1 = k \cdot x_1 + a$$
$$...$$
$$x_n \rightarrow t_n = k \cdot x_n + a$$

Es compleix que:

$$\bar{t} = k \cdot \bar{x} + a$$
$$s_t = |k| \cdot s_x$$

Geomètricament, **t = x + a** equival a traslladar l'histograma **a** unitats en la direcció de l'eix d'abscisses, tenint en compte el signe de **a**.

Per altra banda, **t = k·x** equival a encongir (si |k| < 1) o a allargar (si |k| > 1) horitzontalment l'histograma respecte de l'origen de l'eix d'abscisses. Si k < 0, tots els valors de **x** passaran a pertànyer a l'eix d'abscisses negatiu, si x > 0, o al positiu, si x < 0.

Estandarditzar significa aplicar la següent transformació lineal per a cada dada:

$$z = \frac{x - \bar{x}}{s}$$

Si tenim en compte que es tracta d'una transformació lineal **z = k·x + a**:

$$k = \frac{1}{s} \qquad a = -\frac{\bar{x}}{s}$$

Per tant, les dades quedaran transformades de la següent manera:

$$x_1 \rightarrow z_1 = \frac{x_1 - \bar{x}}{s}$$
$$...$$
$$x_n \rightarrow z_n = \frac{x_n - \bar{x}}{s}$$

El valor $|z_i|$ ens diu a quantes desviacions estàndard (s) està el valor x_i de la mitjana $\bar{x}$. El signe de z_i ens informa de si x_i està a la dreta ($z_i > 0$) o a l'esquerra ($z_i < 0$) de la mitjana $\bar{x}$.

Les dades estandarditzades $z_1, ..., z_n$ compleixen que $\bar{z} = 0$ i $s_z = 1$.

L'estandardització elimina les unitats de les dades.

4.2. Transformacions no lineals. Transformació logit

Si un conjunt de dades $x_1, ..., x_n$ presenta un **fort biaix**, és preferible aplicar una transformació $y = f(x)$ de manera que les dades transformades $t_1, ..., t_n$ no tinguin tant de biaix.

En aquest cas, a diferència de les transformacions lineals, si s'aplica una transformació qualsevol $y = f(x)$ a unes dades $x_1, ..., x_n$ cal tenir en compte que, en general, $\bar{t} \neq f(\bar{x})$ i $s_t \neq f(s_x)$.
Segons el biaix de les dades sigui a la dreta o a l'esquerra tenim diferents tipus de transformacions:

- **Transformacions per disminuir el biaix de la dreta:**

  Transformació potencial amb $\alpha < 1$: $t = x^{\alpha}$

  Transformació logarítmica: $t = \ln(x)$

- **Transformacions per disminuir el biaix de l'esquerra:**

  Transformació potencial amb $\alpha > 1$: $t = x^{\alpha}$

La **transformació logit** s'aplica a variables que prenen valors en intervals (a, b) per a tenir valors a tots els reals. Si $x \in (a, b)$ aleshores:

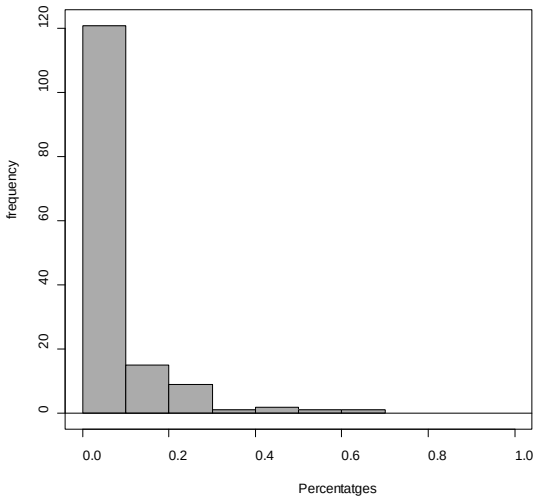
La transformació inversa serà:

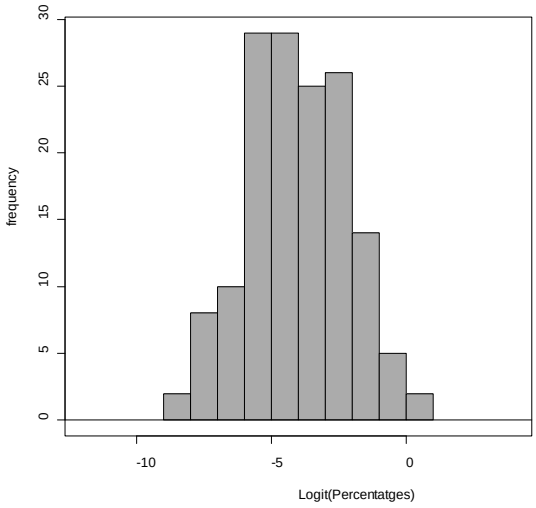
$$x = \frac{a + b \cdot e^t}{1 + e^t}$$

Exemple*:* Si $x \in (0, 1)$, aleshores podem realitzar la transformació logit del següent grup de dades seguint la fórmula següent:

$$\text{logit}(x) = \ln\left(\frac{x - 0}{1 - x}\right)$$

Observem la distribució del conjunt de dades originals a l'esquerra i, a la dreta, la distribució d'aquest grup de dades un cop aplicada la transformació logit:





PROBLEMES

1. Exercicis resolts

1.1. *Des d'un pont de l'autopista AP7 del municipi de Salt, s'ha controlat 70 vegades el nombre de vehicles que circulen de Barcelona en direcció Figueres. Cada control realitzat ha durat 30 segons. Els resultats obtinguts queden resumits a continuació:*

x_i	0	1	2	3	4	5	6	7	8	9
F_i	4	8	17	20	9	6	2	2	1	1

x_i: Nombre de vehicles que han circulat durant 30 segons.
F_i: Freqüència absoluta (nombre d'intervals de temps).

Es demana:

a) Caracteritzeu numèricament aquesta distribució.
b) Caracteritzeu gràficament aquesta distribució.
c) Representeu i expliqueu el diagrama de caixa i detecteu dades atípiques.

a) Caracteritzar numèricament o fer un estudi numèric significa calcular la mitjana, la desviació, la mediana, el mínim, màxim, el primer i tercer quartil i el coeficient de variació.

$$\bar{x} = \frac{\sum_{i=1}^{k} c_i n_i}{n} = \frac{0 \cdot 4 + 1 \cdot 8 + \ldots + 8 \cdot 1 + 9 \cdot 1}{70} = \mathbf{3.014\ cotxes}$$

$$s^2 = \frac{\sum_{i=1}^{k} n_i (c_i - \bar{x})^2}{n - 1} = \frac{4 \cdot (0 - 3.014)^2 + \ldots + 1 \cdot (9 - 3.014)^2}{70 - 1} = 3.32$$

$$\mathbf{s} = \sqrt{s^2} = \sqrt{3.32} = \mathbf{1.82\ cotxes}$$

$$\mathbf{Md} = x_{\left(\frac{n}{2}+0.5\right)} = x_{\left(\frac{70}{2}+0.5\right)} = x_{35.5} = \frac{x_{35} + x_{36}}{2} = \frac{3+3}{2} = \mathbf{3\ cotxes}$$

$$x_{Q_1} = x_{P_{25\%}} = \frac{25 \cdot (70 - 1)}{100} + 1 = 18.25$$

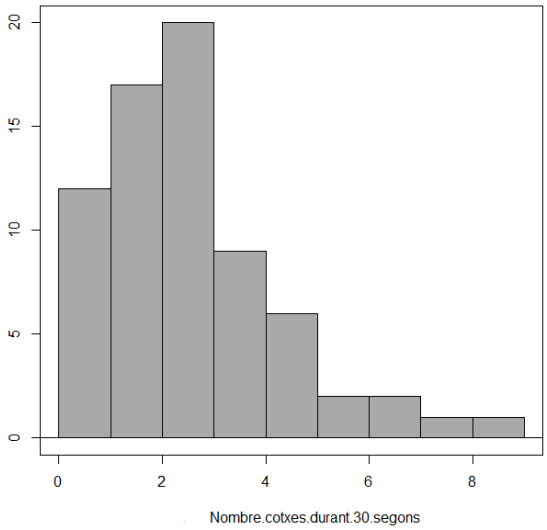
Com que la posició 18 i la 19 tenen el mateix valor, $\mathbf{x_{Q_1} = 2\ cotxes}$

$$x_{Q_3} = x_{P_{75\%}} = \frac{75 \cdot (70 - 1)}{100} + 1 = 52.75$$

Com que la posició 52 i la 53 tenen el mateix valor, $\mathbf{x_{Q_3} = 4\ cotxes}$

$$\mathbf{CV}(\%) = \frac{s}{\bar{x}} \cdot 100 = \frac{1.82}{3.014} \cdot 100 = \mathbf{0.604\ cotxes}$$

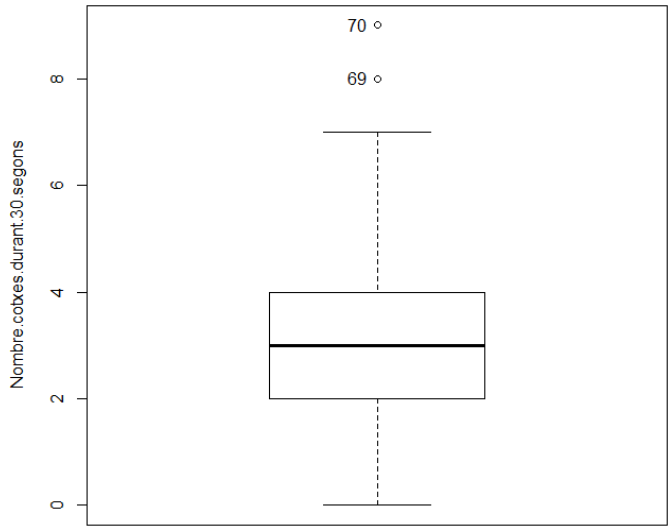
b) Caracteritzar gràficament o fer un estudi gràfic significa representar les dades amb algun mètode gràfic que ens doni una idea de la tendència que pren el conjunt de mostres. Aquest mètode gràfic ha d'estar d'acord amb la naturalesa de les dades. Per a dades numèriques contínues, podem utilitzar un histograma.



En l'histograma podem observar que la observació més freqüent és la de 3 cotxes observats durant 30 segons, i que al tenir una freqüència baixa de les dades extremes, la mitjana s'aproxima al mateix valor que la mediana.

També veiem que l'amplitud de nombre de cotxes vistos en 30 segons va des dels 0 als 9, i que la distribució d'aquest conjunt de dades és unimodal, esbiaixat a la dreta. Aquest comportament pot ser degut a que no es poden obtenir observacions de valor negatiu..

c) Amb aquest gràfic de caixa veiem que la meitat d'observacions prenen un valor entre 2 i 4 cotxes observats durant 30 segons. També veiem que les observacions 69 i 70 són dades atípiques.

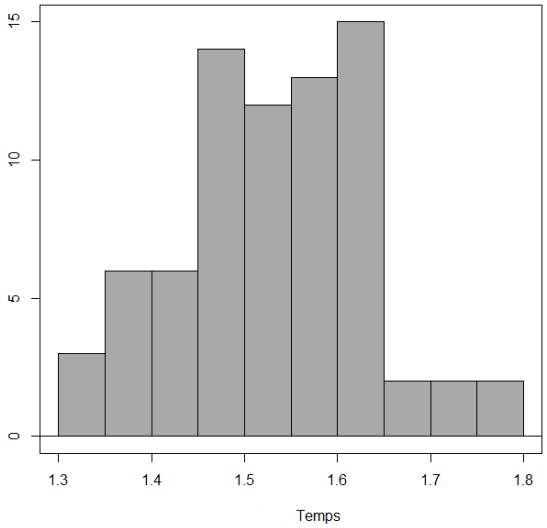


1.2. *La taula mostrada a continuació conté els temps (en segons) que ha costat establir una connexió a Internet en 75 ocasions diferents. Les dades s'han agrupat mitjançant intervals de classe de 0.05 s.*

Es demana:

a) Dibuixeu el corresponent histograma i comenteu-ne els trets més característics.
b) Indiqueu l'interval modal.
c) Calculeu la mitjana, la mediana, i la variància no corregida.
d) Quin percentatge de vegades es tarda menys de 1.625 s? I més de 1.40 s? I entre 1.40 i 1.625 s?

Temps (s)	***Connexions***
1.30 – 1.35	*3*
1.35 – 1.40	*6*
1.40 – 1.45	*6*
1.45 – 1.50	*14*
1.50 – 1.55	*12*
1.55 – 1.60	*13*
1.60 – 1.65	*15*
1.65 – 1.70	*2*
1.70 – 1.75	*2*
1.75 – 1.80	*2*



a) Observem que l'histograma presenta un perfil unimodal amb un lleuger biaix a l'esquerra. També veiem que la dispersió de les dades és baixa, ja que els valors centrals són entre 1.45 i 1.65 segons i l'amplitud total de les dades és poc més ampli, de 1.3 a 1.8 segons.

b) L'interval modal conté l'interval de classe amb més dades i és igual a [1.6, 1.65).

c) Per calcular la mitjana i la variància no corregida d'un conjunt de dades agrupat en intervals, utilitzarem la mitjana de cada interval.

Temps	1.325	1.375	1.425	1.475	1.525	1.575	1.625	1.675	1.725	1.775
Conn.	3	6	6	14	12	13	15	2	2	2

$$\bar{x} = \frac{\sum_{i=1}^{k} c_i n_i}{n} = \frac{1.325 \cdot 3 + 1.375 \cdot 6 + \ldots + 1.725 \cdot 2 + 1.775 \cdot 2}{75} = \mathbf{1.5324\ segons}$$

$$s^{*2} = \frac{\sum_{i=1}^{k} n_i (c_i - \bar{x})^2}{n} = \frac{3 \cdot (1.325 - 1.5324)^2 + \ldots + 2 \cdot (1.775 - 1.5324)^2}{75}$$

$$\mathbf{s^{*2} = 0.01038\ segons}$$

Pel càlcul de la mediana utilitzarem directament la taula que se'ns ha proporcionat a l'enunciat.

$$\mathrm{Md} = x_{\left(\frac{n}{2}+0.5\right)} = x_{\left(\frac{75}{2}+0.5\right)} = x_{38}$$

La posició 38 es troba dins l'interval [1.50, 1.55), el qual va des de la posició 29 a la 41. Realitzem una interpolació lineal per saber el valor que pren la posició on es troba la mediana.

$$\frac{1.55 - 1.50}{41 - 29} = \frac{1.55 - \mathrm{Md}}{41 - 38} \rightarrow \mathbf{Md = 1.5375\ segons}$$

d) Calculem el percentatge de vegades que es tarda menys de 1.625 segons en realitzar una connexió. El valor de 1.625 segons correspon a la posició 61.5. Per això, calcularem el percentil d'aquesta dada.

$$P_q(61.5) \text{ significa que } 61.5 \text{ acumula } P_{\left(\frac{100\cdot(61.5-1)}{75-1}\right)} = P_{81.76\%}$$

Per tant, el percentatges de vegades que es realitza una connexió en menys de 1.625 segons és del **81.76%.**

Ara calculem el percentatge de vegades que es tarda més de 1.4 segons en realitzar una connexió. El valor de 1.4 segons correspon a la posició 9. Per això, del 100% restarem el percentil d'aquesta dada.

$$P_q(9) \text{ significa que } 9 \text{ acumula } P_{\left(\frac{100\cdot(9-1)}{75-1}\right)} = P_{10.81\%}$$

$$100\% - 10.81\% = 89.91\%$$

Per tant, el percentatges de vegades que es realitza una connexió en més de 1.4 segons és del **89.19%.**

Finalment calculem el percentatge de vegades que es tarda entre 1.4 i 1.625 segons en realitzar una connexió. Aquest percentatge el calculem de la següent manera:

$$100 - (89.19\% - 81.76\%) = 92.57\%$$

Per tant, el percentatges de vegades que es realitza una connexió entre 1.4 i 1.625 segons és del **92.57%.**

2. Exercicis proposats

2.1. *Per tal de determinar el nivell de coneixements matemàtics que té un alumne, aquest ha de realitzar 5 tests que es puntuen de 0 a 100 punts. Després de fer els 5 tests, el nivell de coneixements d'un alumne s'estableix d'acord amb el següent barem:*

Nivell A: 90 – 100 Nivell B: 80 – 89 Nivell C: 70 – 79
Nivell D: 60 – 69 Nivell E: 0 – 59

Un alumne ha obtingut 80, 96, 84, 95 i 90 punts en els 5 tests.

Es demana:

a) Determineu la mediana i la mitjana de les puntuacions de l'alumne.
b) A quin nivell de coneixements seria assignat aquest alumne segons fos la mitjana o la mediana el paràmetre que s'utilitzés per a resumir el conjunt de puntuacions?

c) Quin dels dos paràmetres creieu que resumeix millor les puntuacions obtingudes per l'alumne en els 5 tests?

Solució: a) $\bar{x} = 90$ i $Md = 89$; b) B i A; c) La mediana, té més notes d'A que de B

2.2. *Una empresa de maquinària forestal ha recollit els equip de desbrossament que ha venut en els darrers 200 dies laborables. En la taula mostrada el nombre d'equips venuts és simbolitzat per x_i, i la freqüència (dies) de cada valor d'aquesta variable per F_i.*

x_i	0	1	2	3	4	5	6	7	8	9	10	11	12
F_i	38	35	28	21	18	19	17	15	5	1	2	0	1

Caracteritzeu numèricament i gràficament aquesta distribució.

Solució: $\bar{x} = 3.045; Md = 2; Moda = 0; s = 2.585$

2.3. *Les comandes rebudes durant els darrers 20 dies laborables en un taller de fusteria d'alumini són: 7, 5, 5, 0, 3, 9, 5, 3, 7, 9, 7, 5, 5, 3, 3, 5, 3, 5, 5, 3. Calculeu la mitjana, la mediana, la moda, l'amplitud, la desviació i variància corregides, el coeficient de variació, el primer i el tercer quartil.*

Solució: $\bar{x} = 4.85; Md = 5; Moda = 5; ampl = 9;\ s = 2.207;\ s^2 = 4.87;$
$CV = 44.36\%;\ Q_1 = 3;\ Q_2 = 6$

2.4. *Una empresa està interessada en analitzar el temps que dura una peça d'una certa màquina (temps de vida) abans no s'espatlla. Per fer aquest estudi, s'han recollit les hores de durada de les darreres 30 peces que aquesta màquina ha utilitzat. En la següent taula s'adjunten els temps de vida de les diferents peces en hores.*

47	63	66	58	32	61	57	44	44	56	38	35	76	58	48
59	67	33	69	53	51	28	25	36	49	78	48	42	72	52

Caracteritzeu numèricament i gràficament aquesta distribució.

Solució: $\bar{x} = 51.5; Md = 51.6;\ s = 13.8;\ Q_1 = 40.83;\ Q_2 = 63.33$

PRÀCTIQUES

1. Introducció

L'anàlisi exploratòria de dades consisteix en un conjunt de tècniques estadístiques per a descriure gràficament i numèrica les dades d'una mostra. Els gràfics que es representen i els estadístics que es calculen per a descriure una variable s'escullen segons la seva tipologia: qualitativa, quantitativa discreta, i quantitativa contínua.

En aquesta pràctica treballarem les tècniques bàsiques de l'anàlisi exploratòria univariant tot aplicant-les a un cas pràctic. Concretament estudiarem el conjunt de dades `cargols`. Les dades es troben al directori:

```
s:\practica\estadist\R\Introd_R\
```

Primer de tot, feu una còpia d'aquesta carpeta al vostre llapis de memòria (o en el directori `c:\temp`) per poder treballar més còmodament. Una vegada feta la còpia, executeu el programa R. A continuació obriu R-Commander i anar a:

```
Paquetes, Cargar Paquetes, Rcmdr ...
```

I tot seguit escriure l'ordre `library(Rcmdr)` i polsar Enter.

Una vegada en R-Commander, declareu el directori del vostre llapis de memòria (o el directori `c:\temp`) com a directori treball anant a:

```
Fitxer, Canvia el directori de treball ...
```

Aquesta operació s'ha de repetir cada vegada al inici d'una sessió.

El fitxer cargols es troba en el directori de treball. Per carregar-lo feu:

```
Dades, Carrega taula de dades ...
```

I escolliu `cargols.rda` que trobareu al vostre directori de treball. Una vegada carregat, cargols passa a ser el conjunt de dades actiu i el podeu visualitzar polsant a sobre del botó `Visualiza la taula de dades`.

El fitxer cargols consta de 237 registres organitzat en 6 variables. Els registres corresponen a una mostra de 237 cargols fabricats per una empresa en dues plantes de producció amb tres línies de producció cadascuna.

- *Planta_Prod*: planta de producció (P1, P2).
- *Línia_Prod*: línia de producció (L1, L2, L3).
- *Recobriment*: tipus de recobriment (Zn o Cr).
- *Sup_Rovell*: superfície lateral rovellada (en cm^2).
- *Longitud*: longitud del cargol (en cm).

- *Diàmetre*: diàmetre del cargol (en cm).
- *Nom_Def*: nombre de defectes trobats.

Els individus descrits són els cargols. Cada fila descriu un individu o cas. Cada columna conté els valors d'una variable per a cada individu. Hi ha 7 variables; *Planta*, *Línia* i *Recobriment* són variables categòriques. La resta són variables numèriques. Noteu que *Nom_Def* és una variable numèrica discreta la qual rebrà un tractament específic.

2. Estudi de la variable categòrica *Línia_Prod*

2.1. Tabulació de les dades

Els valors d'una variable categòrica són noms o categories, per exemple home o dona. La distribució d'una variable categòrica descriu les categories i la seva freqüència o percentatge. Això es presenta en forma de taula. Per exemple, la distribució dels 237 cargols de la mostra en les diferents línies de producció és:

Tipus de família	Freqüència	Percentatge (%)
L1	73	30.80
L2	86	36.29
L3	78	32.91

Per a obtenir la informació de les freqüències de les diferents línies de producció anem a:

`Estadístics, Resums, Distribució de freqüències ...`

Li indicarem que ens ho faci sobre la variable *Línia_Prod* i polsem `D'acord`. Tot observant la finestra de resultats responeu les següents preguntes:

- Quantes mostres s'han produït en la línia 1? 73.
  Quin percentatge representen? 30.8%.

- Quantes mostres s'han produït en la línia 2? 86.
  Quin percentatge representen? 36.29%.

- Quantes mostres s'han produït en la línia 3? 78.
  Quin percentatge representen? 32.91%.

2.2. Representacions gràfiques

Els gràfics més adequats per les variables de tipus qualitatiu són el diagrama de barres i el diagrama de sectors. El diagrama de barres compara de forma ràpida la freqüència de cada categoria, mentre que el diagrama de sectors o de pastís visualitza la importància relativa de cada categoria respecte el total de dades. En

conseqüència, si es tracta d'una variable categòrica ordinal s'aconsella el diagrama de barres. Pel cas de variables qualitatives nominals amb poques categories s'aconsella el diagrama de pastís.

Per representar el diagrama de barres de la variable *Línia_Prod* cal anar a:

`Gràfics, Gràfic de barres ...`

I escollir *Línia_Prod*.

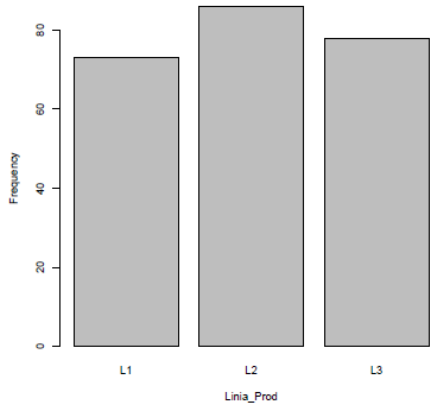


Figura 1: Diagrama de barres de la variable *Linia_Prod*.

Es pot obrir una altra finestra gràfica independent anant a `Finestra d'instruccions` i executar l'ordre `windows()`.

Feu ara el diagrama de sectors anant a:

`Gràfics, Gràfic de sectors ...`

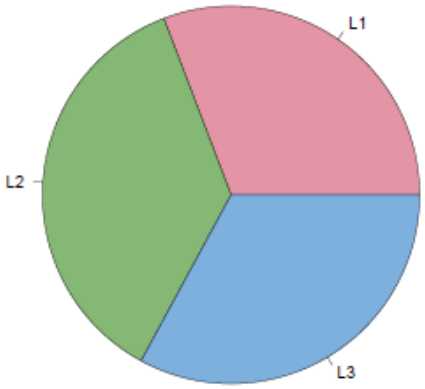


Figura 2: Diagrama de sectors de la variable *Linia_Prod*.

- A partir d'aquestes gràfiques, descriviu els trets més importants de la variable *Línia_Prod*. Quin dels dos gràfics és el més adient? Variable qualitativa no ordinal amb 3 categories. La moda de la distribució és Línia 2 amb un 36.26% de la freqüència. La moda no és molt acusada, totes les categories tenen una freqüència similar. La distribució s'assembla a la

distribució d'un model uniforme, on cada categoria té la mateixa freqüència, 1/3 en aquest cas. El gràfic més adient és el de sectors atès que la variable té poques categories (3) i no és ordinal.

3. Estudi de la variable contínua *Longitud*

Estudiarem a nivell descriptiu la variable *Longitud* (longitud (mm) dels cargols). Tabularem les dades i realitzarem una anàlisi descriptiva gràfica i numèrica.

3.1. Tabulació de les dades

R-Commander no permet fer resums de freqüències de variables numèriques, en canvi la funció `table()` que hem de fer servir a la finestra d'instruccions, sí que ho permet. Ara bé, si hi ha molts valors diferents, els resultats no són gaire clars. Escriviu a la finestra d'instruccions o copieu i enganxeu la comanda:

```
table(cargols$Longitud)
```

I polseu `Executar`. Observareu que el resultat és una mica difícil de resumir. Es a dir, la taula de freqüències no és informativa i per tant calen altres tècniques estadístiques per l'estudi de variables numèriques contínues.

3.2. Anàlisi descriptiva gràfica

Com ja hem vist, els gràfics més adequats per representar variables numèriques contínues són l'histograma i la caixa de dispersió.

- **Histograma**

Quan una variable numèrica pren molts valors diferents, ni els diagrames de barres ni els de sectors són informatius: cal agrupar els valors més pròxims en classes o interval i construir un histograma.

Farem un histograma de la variable *Longitud*. R-Commander ens calcularà automàticament el nombre d'intervals que cal fer servir, encara que nosaltres podrem canviar-ho.

```
Gràfics, histograma ...
```

Escollim la variable *Longitud*. Observeu que en la casella `Nombre de segments`, la qual determina el nombre de barres que tindrà l'histograma, conté la informació `auto`. Amb aquesta opció R-Commander calcularà el nombre d'intervals segons la Regla de Sturges.

Polseu `D'acord`.

- En quants intervals de classe han quedat agrupades les 237 dades? 9.
- Observeu i descriviu el perfil de l'histograma. Perfil en forma simètrica i amb decreixement ràpid en els extrems (campana de Gauss). Molt simètric i centrat en el 5. L'amplitud del gràfic va des de 4.8 fins 5.25.
- Creus que es pot considerar simètric? Cap a on té el biaix? Sí. Lleuger biaix al costat dret.
- Creus que es pot considerar unimodal? Sí, hi ha dos intervals amb freqüència molt alta en la zona central. La tendència de la distribució és d'augment de la freqüència des de 4.8 fins a la zona central (5) per a disminuir la freqüència fins a 5.25.

Si es vol canviar el nombre d'intervals hem de posar-ho a l'opció `Nombre de segments` del menú `Gràfics, Histograma ...` Representeu diversos histogrames provant amb diferents nombres d'intervals. Observareu que, a vegades, R no fa el que nosaltres li demanen.

Es pot forçar un histograma amb 8 intervals de classe escrivint a la finestra d'instruccions:

```
hist(cargols$Longitud, seq(4.8,5.21,length=9))
```

I polsant `Executar`. Feu-ho. Representeu diferents histogrames canviant el valor del paràmetre `length`.

- Observeu el perfil dels histogrames i decidiu quin és el nombre de classes que proporciona una millor representació de la distribució. 8 o 9 intervals de classe dóna una representació adient de la distribució; No hi ha forats de freqüència (massa intervals).

- **Diagrama de caixa o caixa de dispersió**

Amb la informació dels quartils i la mediana es pot construir un gràfic de gran importància: el diagrama de caixa. Aquest gràfic dóna una descripció molt clara de la forma de la distribució i de l'existència de valors atípics (**Figura 3**).

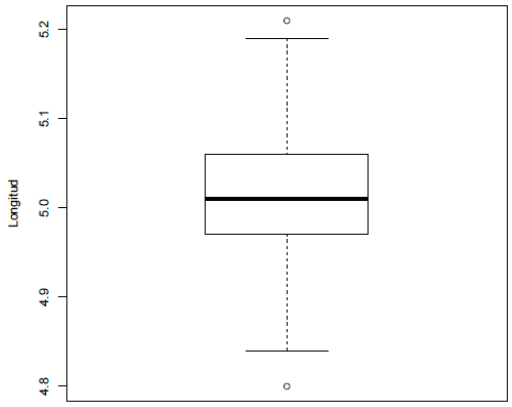


Figura 3: Diagrama de caixa de la longitud dels cargols

- Els costats inferior i superior de la caixa corresponen als quartils.
- El segment interior de la caixa correspon a la mediana.
- Per calcular els extrems de les línies es calcula l'amplitud interquartilíca. Es consideren atípiques les dades per sota de $Q_1 - 1.5 \cdot AIQ$ o per sobre de $Q_3 + 1.5 \cdot AIQ$. En cas que hi hagi dades atípiques aquestes apareixen en la figura marcades amb símbols específics (asteriscs, cercles...). Les línies o braços de la caixa central arriben fins el valor de les dades no atípiques. Així, la línia de inferior marca la primera dada no atípica, i el braç superior acaba en el valor de la darrera dada no atípica.

Per a representar la caixa de dispersió de la variable *Longitud* cal anar a:

`Gràfics, Caixa de dispersió ...`

I escollir la variable *Longitud* i deixar la resta d'opcions sense modificar. Si preferiu representar horitzontalment el diagrama de caixa cal afegir l'opció `horizontal=TRUE` en el codi de la comanda. Així, cal anar a la finestra d'instruccions i executar el següent ordre:

```
boxplot(cargols$Longitud, ylab="Longitud",horizontal=TRUE)
```

A la vista dels resultats, contesteu les següents qüestions:

- Quant valen aproximadament el primer quartil, la mediana i el tercer quartil? 4.97, 5.02 i 5.08.
- Cap a on està esbiaxada la distribució? Una mica a la dreta.
- Hi observeu alguna dada atípica? Un parell de dades atípiques.

Com veieu, han aparegut dades atípiques. R-Commander permet identificar aquestes dades marcant l'opció `Identify atypical points` amb el ratolí en el menú. Amb el botó de l'esquerra del ratolí podem identificar la dada que vulguem. Per acabar fareu un clic en el botó de la dreta del ratolí. Heu de tenir en compte que els números que apareixen en el gràfic no són els valors atípics sinó que són el nom o identificador del registre de la dada original. En canvi, en acabar, a la finestra de resultats sí que hi apareixeran els números de fila que correspon a cada dada atípica. Per saber el valor atípic concret cal fer clic en `Visualitza la taula de dades` i buscar la fila corresponent.

- Indiqueu els valors de les dades atípiques. Files 39 i 85 amb els valors: cargols$Longitud[39] = 4.8; cargols$Longitud[85] = 5.21.

3.3. Anàlisi descriptiva numèrica

- **La mitjana**

La descripció d'una variable numèrica ha d'incloure mesures de centre i de dispersió. La mesura més comuna de centre és la mitjana.

Mitjançant comandes escrites a la finestra d'instruccions de R-Commander es pot calcular la mitjana:

```
mean(cargols$Longitud)
```

I polsant Executar.

- Quin valor pren la mitjana de la mostra? 5.012658.

- **La mediana**

La mediana Med és el punt mitjà de la distribució quan les dades estan ordenades de més petites a més grans. Per calcular-la escriurem la següent comanda a la finestra d'instruccions de R-Commander.

```
median(cargols$Longitud)
```

I polsant Executar.

- Quin valor pren la mediana de la mostra? 5.01.

- **Comparació de la mitjana i la mediana**

La mitjana és molt sensible a les dades extremes o atípiques, en canvi, la mediana no ho és. Si canviem la dada de valor 5.21 per 20.21, es pot comprovar que la mitjana s'incrementa i en canvi la mediana resta constant.

Cliqueu en la pestanya Edita taula de dades de la barra d'icones. Localitzeu la dada 5.21 en la fila 85, canvieu-la pel valor 500.21, tanqueu la finestra. Tot seguit aneu a:

```
Dades, Taula de dades activa, Refresca la taula de dades activa ...
```

Aneu a la finestra d'instruccions de R-Commander i feu un clic en la línia on heu calculat la mitjana. Polseu Executar. Feu el mateix amb la línia de la mediana.

- Quins valors prenen la mitjana i la mediana de la mostra? 7.101266 i 5.01.

La mitjana i la mediana d'una distribució simètrica prenen valors semblants. En canvi, en una distribució asimètrica, la mitjana queda desplaçada cap a la cua més llarga de la distribució, és a dir, la cua on es situen les dades extremes.

Cliqueu en la pestanya Edita taula de dades de la barra d'icones i desfeu el canvi en la dada de la fila 85.

- **Els percentils**

El percentil p_q on $q \in (0, 1)$ d'un conjunt de dades es defineix com aquell valor que limita de forma aproximada el $q \cdot 100$ % de les dades, on $0 \leq q \leq 1$. Recordem que cada calculadora o programa pot tenir fórmules diferents pel càlcul dels percentils. Per al càlcul dels percentils s'utilitza la funció `quantile()` tot especificant quins percentils es volen calcular.

Per exemple, per calcular els percentils 40, 60 i 85% escriurem el següent:

```
quantile(cargols$Longitud, c(0.4, 0.6, 0.85))
```

Amb la funció `c()` s'indica el vector de percentils a calcular.

Si en canvi es vol obtenir els percentils 0, 10, 20, ..., 90, 100% s'escriurà:

```
quantile(cargols$Longitud, seq(0, 1, 0.1))
```

La funció `seq()`, fa una seqüència de números, des de 0 fins a 1, en salts de 0.1.

- Quins valors agafen els percentils mostrals de 1, 10, 30, 70 i 95? Escriurem l'instrucció `quantile(cargols$Longitud, c(0.01, 0.1, 0.3, 0.7, 0.95))`. Resultats: 4.8436, 4.9260, 4.9800, 5.0500, 5.1220.

- **Els quartils**

La mitjana i la mediana són dues mesures de posició. Però una distribució no sols queda definida per mesures de centre. Dues províncies amb una mateixa mediana d'ingressos per família poden ser molt diferents si en una d'elles hi ha extrems de pobresa o riquesa, mentre que l'altra té poca variació entre els extrems. Calen per tant altres mesures, anomenades de dispersió o de variabilitat, que ajudin a definir d'una manera més concreta una distribució.

Una primera forma és donar les observacions mínima i màxima de la distribució. La diferència entre les dues s'anomena amplitud. L'amplitud mesura la dispersió de la distribució. Però pot resultar poc efectiva si hi ha dades extremes.

Els quartils determinen entre quins valors es troben (de forma aproximada) el 25% i el 75% de les dades. El primer quartil, Q_1 limita o separa el primer 25% de les observacions, mentre que el tercer quartil, Q_3, limita el 75%. El segon quartil és la mediana Med.

Mitjançant comandes, es poden calcular els quartils i la mediana escrivint la següent comanda a la finestra d'instruccions de R-Commander:

```
quantile(cargols$Longitud)
```

I polsant Executar.

- Quin valor agafen els quartils de la mostra? 4.97 i 5.06.

- **La variància i la desviació típica**

Una mesura més comuna i adequada per mesurar la dispersió és la desviació tipus o desviació típica, que mesura la dispersió de les dades en relació a la seva mitjana.

Es defineix la variància d'una mostra de dades com la suma dels quadrats de la distància de les observacions respecte a la mitjana dividit per n – 1. La desviació típica és l'arrel quadrada de la variància

Mitjançant comandes, es pot calcular la variància i la desviació escrivint les comandes:

```
var(cargols$Longitud)
sd(cargols$Longitud)
```

A la finestra d'instruccions de R-Commander i polsant Executar, respectivament sobre cada línia.

- Quin valor té la variància i la desviació de la mostra? 0.004686548 i 0.06845837.

- **Coeficient de variació**

Per poder catalogar de gran o petita la desviació de les dades d'una mostra s'utilitza el coeficient de variació (CV). Alguns autors indiquen que un CV superior al 30% és indicador d'una dispersió molt elevada.

- Aneu a la finestra d'instruccions i calculeu el CV de la variable *Longitud*. Com catalogaries la desviació de la mostra?
  CV=sd(cargols$Longitud)/mean(cargols$Longitud)*100 = 1.365710. La distribució de les dades de la mostra té poquíssima dispersió.

- **Resum numèric descriptiu**

Per calcular els estadístics descriptius més habituals de la variable *Longitud* ho farem a partir del menú:

```
Estadístics, Resums, Resums numèrics ...
```

A la finestra resultant es marca la variable *Longitud* i la resta d'opcions es deixen tal com estan i es polsa Acceptar.

- Indiqueu quin és el nombre de dades (n), la mitjana , la desviació estàndard, la dada mínima (Min), la mediana i la dada màxima (Max) i els quartils. 237, 5.012658, 0.06845837, 4.8, 5.01, 5.21, 4.97 i 5.06.

Com haureu observat a la finestra del resum numèric, es poden calcular els percentils que es vulguin indicant-ho al costat de quantils.

- Quins valors agafen els percentils mostrals de 1, 10, 30, 70 i 95? Posem literalment 0.01, 0.10, 0.30, 0.70, 0.95 i polsem `Acceptar`. Resultats: 4.8436, 4.9260, 4.9800, 5.0500, 5.1220.

4. Transformacions

L'objectiu de la transformació d'una variable és millorar la seva anàlisi estadística en varis aspectes importants. Els tipus de transformació més usuals són la lineal, l'estandardització, la logarítmica, la de potència i la lògit.

Anem a estandarditzar la variable *Longitud*:

```
Dades, Modifica dades de la taula, Calcula nova variable ...
```

A `Nom de la variable` posem el nom de la nova variable. Per seguir la nomenclatura estàndard posarem *Z.Longitud*. A `Expressió a calcular` escriurem la fórmula d'estandardització:

```
(Longitud-mean(Longitud))/sd(Longitud)
```

Comproveu que s'obté el mateix resultat realitzant el següent:

```
Dades, Modifica dades de la taula, Estandarditza les variables ...
```

- Compareu els gràfics i els estadístics de la variable *Longitud* i la seva estandardització. A quantes desviacions es troben les dades atípiques? La mitjana val aproximadament zero i la desviació val 1. Els gràfics (Ex: diagrama de caixa) mostren que la forma de la distribució no ha canviat. Les dades atípiques es troben al voltant de 3 desviacions.
- Assumiu que els cargols tenen forma cilíndrica. Mitjançant les variables *Sup_Rovell*, *Longitud* i *Diàmetre* calculeu el percentatge de superfície lateral rovellada de cada cargol. Analitzeu i compareu la distribució d'aquesta nova variable i de la seva logit transformada. Anem a:

  ```
  Dades, Modifica, Calcula ...
  ```

  Escrivim *Percrov* com a `Nom de la nova variable` i com a expressió a calcular `SupRovell/(2*pi*(Diametre/2)*Longitud)`. La nova columna (Percrov) recull la proporció de superfície rovellada. Anem a:

`Estadístics, Resums, Resums numèrics ...`

I ens dóna: mean = 0.3317927, sd = 0.05625684, 0% = 0.1932590, 25% = 0.2971957, 50% = 0.3310905, 75% = 0.365256, 100% = 0.5459769 i n = 237.

Calculem el CV:

```
sd(cargols$Percrov)/mean(cargols$Percrov)*100
```

Veiem que la desviació de la distribució és moderada (aproximadament 17%). Comparant la mitjana (0.3317) amb la mediana (0.3311) veiem que suggereixen simetria, la qual cosa queda recollida en l'histograma i el diagrama de caixa. El diagrama de caixa mostra dues dades atípiques. Són els valors de les files 47 i 134, que corresponen als cargols núm. 418 i 245. Els valors atípics són 0.1932590 i 0.5459769, els mínim i màxim de la distribució. Les dades estan en l'interval (0, 1).

Fem la transformació logit amb a = 0 i b = 1. Anem a:

`Dades, Modifica, Calcula ...`

Escrivim *Logitrov* com a `Nom de la nova variable` i com a expressió a calcular `log(Percrov/(1-Percrov))`. La nova columna (Logitrov) recull els valors logit-transformats de Percrov. Calculem els estadístics bàsics de nou:

`Estadístics, Resums, Resums numèrics ...`

I ens donen: mean = –0.7111689, sd = 0.2582874, 0% = –1.428971, 25% = –0.8606878, 50% = –0.703257, 75% = –0.5526232, 100% = 0.1844286 i n = 237.

Veiem que mitjana i mediana són similars (–0.71 i –0.70) però no tant com abans. Fent els gràfics histograma i diagrama de caixa observem que la simetria es conserva però apareixen més dades atípiques: cargols núm. 329 i 839.

5. Anàlisi descriptiva numèrica segons una variable categòrica

També podem fer un diagrama de caixa d'una variable numèrica segons una altra variable. En el menú anterior cal polsar a sobre de `Gràfic segons grup` i escollir la variable categòrica adequada.

- Feu la caixa de dispersió de la variable *Longitud* segons la variable *Planta_Prod* i indiqueu si hi ha diferències en les longituds segons la planta de producció. No s'aprecien diferències en les distribucions de les dades de les longituds segons la planta de producció atès que les dues caixes estan

situades en els mateixos valors (quartils) i tenen amplades similars (desviació).

- Feu el mateix amb *Diàmetre* i *Planta_Prod*. A la vista del diagrama de caixa múltiple sí que s'observen diferències entre les distribucions dels diàmetres segons la planta. Si bé les amplades de les caixes (dispersió) són similars, en canvi, la posició (quartils) de les caixes són diferents. Per exemple, el Q_3 de la P1 val aproximadament igual que el Q_1 de la P2. Això suggereix que en la P2 estan fabricant peces de diàmetres majors que no pas en la P1.

Ara valorarem numèricament les diferències entre la longitud segons la planta de producció, calculant per a cada grup els estadístics habituals (mitjana, desviació estàndard...) de la variable *Longitud*. Per fer-ho, anem com abans a:

`Estadístics, Resums, Resums numèrics ...`

Escollim *Longitud* i polsem `Resums per grups`, i escollim de nou la variable *Planta Prod*.

- Valoreu numèricament les diferències de longitud entre plantes de producció. Comenteu les diferències més destacades.

	n	Mín.	Màx.	Mitjana	Med.	Desv. Típ.
P1	118	4.8	5.19	5.007542	5.01	0.06612519
P2	119	4.87	5.21	5.017731	5.02	0.07060798

Quasi tots els estadístics de posició (mitjana i percentils) prenen valors majors en la P2. Les dues distribucions tenen molt poca dispersió, ja que CV = 1.2%.

6. Estudi de la variable *Nom_Def*

Aquesta es una variable numèrica de tipus discret ja que només pot prendre uns certs valors: 1 defecte, 2 defectes... És evident que no té sentit de parlar de 1.25 defectes. A més té la característica que hi ha un reduït nombre de valors que es repeteixen moltes vegades (a diferència d'una variable numèrica contínua que té molts valors que es repeteixen poques vegades).
Com a variable numèrica, es poden calcular els mateixos estadístics que una variable numèrica contínua, fent servir els mateixos menús. Moltes vegades es fan servir aquestes variables com a referències per a una variable contínua. En aquests casos, i per poder treballar millor amb R, cal convertir-les en variables categòriques. Aleshores, els càlculs que es poden fer són els mateixos que amb una variable categòrica.

- Abans de fer la conversió, feu un resum numèric de la variable. mean = 0.9662447, sd = 1.517874, 0% = 0, 25% = 0, 50% = 0, 75% = 2, 100% = 6 i n = 237.

Es poden fer servir operadors lògics per a obtenir determinats resultats, per exemple, la proporció de productes amb menys de 3 defectes. Aquestes operacions només es poden fer sobre variables numèriques, per tant, si volem aquests resultats ho hem de fer abans de la conversió. Escriviu a la finestra d'instruccions:

```
table(cargols$Nom_Def<3)
```

I polseu el botó Executar. Per obtenir les freqüències relatives caldrà dividir pel total de dades:

```
table(cargols$Nom_Def<3)/length(cargols$Nom_Def)
```

- Quin tant per cent de productes de la mostra tenen menys de 3 defectes? 0.8523207 = 85.2%.
- Escriviu una funció similar per calcular el percentatge de productes de la mostra que tenen més de 2 defectes. Quant val? table(cargols$Nom Def¿2)/length(cargols$Nom_Def) = 14.77%.
- Es poden combinar els operadors per a obtenir resultats una mica més complexos com ara el percentatge de productes de la mostra que tenen entre 2 i 4 defectes (ambdós valors inclosos). Com s'ha de fer: table(cargols$Nom_Def>=2&cargols$Nom_Def<=4)/length(cargols$Nom_Def)) = 20.25%.

6.1. Tabulació de les dades

Primer de tot cal tenir en compte que R considera la variable com a numèrica i per tant no ens deixarà fer les taules de freqüències. Haurem de convertir-la en una variables categòrica (factor) fent:

```
Dades, Modifica variables, Converteix variables numèriques en factors ...
```

Seleccionem la variable *Nom_Def*, indiquem Emprar els números i polsem D'acord. Ens demanarà si volem sobreescriure la variable i li diem que sí. Ara R ja ho considera com a factor i els passos a seguir són els mateixos que per la variable *Línia_Prod*.
El que farem primerament és una taula senzilla de freqüències de les :

```
Estadístics, Resums, Distribució de freqüències ...
```

- Quantes dades hi ha? 237.
- Quina és la dada màxima? 6.
- I quina la mínima? 0.
- Quina dada és la moda? 0.

Si volguéssim fer servir operadors lògics hem de fer una petita operació per a poder obtenir resultats. Per obtenir els resultats de la secció anterior, escriurem:

```
table(as.numeric(cargols$Nom_Def)<3)
```

I polseu el botó Executar.

Per obtenir les freqüències relatives caldrà dividir pel total de dades:

```
table(as.numeric(cargols$Nom_Def)<3)/length(cargols$Nom_Def)
```

I de forma similar per a càlculs semblants.

6.2. Representacions gràfiques

Les representacions gràfiques d'aquests tipus de variables són les mateixes que les variables categòriques: diagrama de barres i de sectors.

- Feu els gràfics i comenteu els resultats. El gràfic més adient és el de barres perquè la variable és ordinal. La distribució de dades té una tendència a disminuir a mesura que augmenta el nombre de defectes.

TEMA 2: La probabilitat i els seus elements

TEORIA

1. Teoria de la probabilitat

La teoria de la probabilitat és el model matemàtic que es construeix per explicar i analitzar la regularitat estadística que caracteritza els fenòmens aleatoris.

1.1. Fenòmens aleatoris i fenòmens deterministes

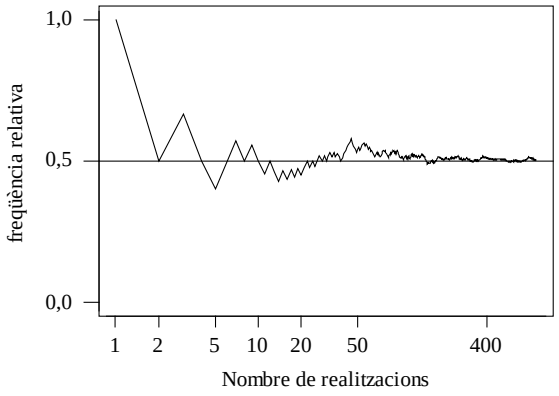
Molts fenòmens naturals o artificials no són deterministes, és a dir, els seus resultats no són predictibles. Malgrat això, molts d'ells presenten patrons de comportament regular a llarg termini. Aquests fenòmens s'anomenen **aleatoris**.

Exemple: El llançament d'una moneda és un cas de fenomen aleatori. Si llancem una moneda una vegada no sabem si el resultat serà CARA o CREU, però si la llancem moltes vegades sabem que aproximadament el 50% d'elles ens sortirà CARA i l'altre 50% CREU.

1.2. Llei de la regularitat estadística

En augmentar el nombre de realitzacions d'un fenomen aleatori, la freqüència relativa d'un dels possibles resultats tendeix a aproximar-se cap a un valor fix, el qual anomenarem **probabilitat** d'aquest resultat.

Exemple: En el cas del llançament d'una moneda, en augmentar el nombre de llançaments, la freqüència relativa de l'esdeveniment CARA tendeix a aproximar-se a 0.5. Per tant, direm que la probabilitat d'obtenir CARA és igual a 0.5 o 50%.



2. El fenomen aleatori

2.1. Espai mostral

L'espai mostral és el conjunt dels possibles resultats o esdeveniments simples del fenomen aleatori. Es simbolitza amb la lletra Ω.

Exemple: Tornant amb l'exemple del llançament d'un dau: Ω = {1, 2, 3, 4, 5, 6}.

2.2. Esdeveniment

Un esdeveniment és qualsevol possible resultat d'un fenomen aleatori o bé qualsevol col·lecció de possibles resultats d'un fenomen aleatori. Per tant, un esdeveniment és qualsevol subconjunt de l'espai mostral. Es simbolitzen amb lletres majúscules.

Exemple: Continuant amb l'exemple anterior del llançament d'un dau podem considerar diferents esdeveniments com:

- Esdeveniment A = obtenir un 2.
- Esdeveniment B = obtenir un número parell.
- Esdeveniment C = obtenir un número més gran de 4.

2.3. Probabilitat d'un esdeveniment

La probabilitat d'un esdeveniment A és el valor al qual s'aproxima la seva freqüència relativa quan el nombre de repeticions del fenomen aleatori, que anomenarem N, es fa infinitament gran.

Per assignar les probabilitats ho podem fer de dues maneres:

- **Per condicions de simetria (Regla de Laplace):** si a priori, per raons de simetria, es pot suposar que tots els resultats individuals d'un fenomen aleatori tenen la mateixa probabilitat de sortir, aleshores, la probabilitat d'un esdeveniment A qualsevol associat al fenomen aleatori és igual a:

$$P(A) = \frac{\text{nombre elements A}}{\text{nombre elements } \Omega}$$

Exemple: Podem fer servir la regla de Laplace per assignar les probabilitats dels esdeveniments de l'exemple anterior, ja que al llançar un dau qualsevol número te les mateixes probabilitats de sortir que un altre:

$P(A) = P(\{2\}) = 1/6$
$P(B) = P(\{2, 4, 6\}) = 3/6$
$P(C) = P(\{5, 6\}) = 2/6$

- **Empíricament:** si no podem suposar simetria entre els possibles resultats individuals d'un fenomen aleatori, per calcular la probabilitat d'un esdeveniment hem de realitzar un nombre suficientment gran de repeticions i observar quants elements pertanyen a aquest. Per tant, per calcular la probabilitat d'un esdeveniment A farem:

$$P(A) \cong \frac{\text{nombre elements A després d'N repeticions}}{\text{N repeticions}}$$

Exemple: Per calcular la probabilitat que una màquina produeixi una peça defectuosa calcularem, experimentalment, el nombre de peces defectuoses després de N peces produïdes i aplicarem:

$$P(A) \cong \frac{\text{nombre peces defectuoses}}{\text{N peces produïdes}}$$

2.4. Tipus d'esdeveniments i les seves probabilitats

Podem considerar diferents tipus d'esdeveniments:

- **Esdeveniment segur (Ω):** aquell que està format per l'espai mostral i que, per tant, es verifica sempre. La probabilitat que succeeixi és 1. Si llancem un dau, la probabilitat que ens surti un número entre 1 i 6 és $P(\Omega) = 1$.

- **Esdeveniment impossible ($\varnothing$):** aquell que és impossible que passi. La probabilitat que succeeixi és 0. Si llancem un dau, la probabilitat que ens surti un número > 6 és $P(\varnothing) = 0$.

- **Esdeveniment unió d'esdeveniments A i B ($A \cup B$):** succeeix sempre que obtenim resultats de A **o** B. Al llençar un dau, la probabilitat que el número sigui > 4 (esdeveniment C) o parell (esdeveniment B) és la probabilitat que ens surtin els números 2, 4, 6 i 5. Per tant, $P(C \cup B) = 4/6$.

- **Esdeveniment intersecció d'esdeveniments A i B ($A \cap B$):** succeeix quan obtenim resultats de A **i** de B. **Exemple:** Si llancem un dau, la probabilitat que ens surti un número > 4 (esdeveniment C) i parell (esdeveniment B) és la probabilitat que ens surti el número 6. Per tant, $P(C \cap B) = 1/6$.

- **Complementari d'un esdeveniment A (A^c o $\bar{A}$):** A^c és complementari de A si $A \cap A^c = \varnothing$ i $A \cup A^c = \Omega$. Si llancem un dau i definim com a esdeveniment

B que ens surti un número parell, B^c serà que ens surti un número senar. Així: $P(B \cup B^c) = 1$ i $P(B \cap B^c) = 0$.

- **Esdeveniments A i B incompatibles:** A i B són incompatibles quan no és possible que A i B es realitzin simultàniament, és a dir $A \cap B = \varnothing$. Si llancem un dau, que ens surti un número < 2 (esdeveniment D) i parell (esdeveniment B) és impossible. Per tant, $P(D \cap B) = 0$.

3. Propietats de la probabilitat

- Qualsevol probabilitat és un nombre entre 0 i 1: $\mathbf{0 \leq P(A) \leq 1}$
- Tots els resultats possibles junts han de sumar probabilitat 1: $\mathbf{P(\Omega) = 1}$
- La probabilitat que un esdeveniment no passi és igual a 1 menys la probabilitat que l'esdeveniment passi: $\mathbf{P(A^c) = 1 - P(A)}$
- Si dos esdeveniments són **incompatibles**, és a dir, $A \cap B = \varnothing$, la probabilitat que passi qualsevol d'ells és igual a la suma de les seves probabilitats individuals: $\mathbf{P(A \cup B) = P(A) + P(B)}$

Exemple: Si tirem un dau i definim els següents esdeveniments A i B, la probabilitat que passi A o B serà:

A = puntuació ≥ 4 → P(A) = P({4, 5, 6}) = 3/6
B = puntuació < 3 → P(B) = P({1, 2}) = 2/6

Veiem que A i B són incompatibles, ja que $A \cap B = \varnothing$. Per tant:

$P(A \cup B) = P(A) + P(B) = 3/6 + 2/6 = 5/6$

- Si dos esdeveniments A i B tenen **resultats en comú**, la probabilitat que passi qualsevol d'ells és: $\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$

Exemple: Si sabem quina és la probabilitat que una peça sigui defectuosa i quina és la probabilitat que una peça hagi estat produïda per la màquina B, quan calculem la probabilitat que passi A o B farem:

A = peça defectuosa → P(A) = 0.03
B = peça produïda per la màquina B → P(B) = 0.6

Com que A i B tenen resultats en comú, és a dir, una peça pot ser defectuosa i produïda per la màquina B:

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.03 + 0.6 - 0.03 \cdot 0.6 = 0.612$

4. Probabilitat condicionada

La probabilitat de l'esdeveniment **A condicionat a** l'esdeveniment **B**, és a dir, la probabilitat que passi A sabent que ha passat B es defineix com:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Utilitzant la definició de la probabilitat condicionada, podem expressar la intersecció com:

$$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

Exemple: Si considerem com a fenomen aleatori el llançament d'un dau i com a esdeveniments A i B els següents, la probabilitat que passi A sabent que ha passat B serà:

A = puntuació parell → P(A) = P({2, 4, 6}) = 3/6
B = puntuació ≥ 4 → P(B) = P({4, 5, 6}) = 3/6

P(A ∩ B) = P({4, 6}) = 2/6

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/6}{3/6} = 2/3$$

També podem calcular P(A | B) calculant P(B | A) = 2/3 i aplicant les propietats de la probabilitat condicionada:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} = \frac{3/6 \cdot 2/3}{3/6} = 2/3$$

4.1. Esdeveniments independents

Si dos esdeveniments A i B són **independents** aleshores es compleix que:

$$\mathbf{P(A \cap B) = P(A) \cdot P(B)}$$

Per tant, la fórmula de probabilitat condicionada quedarà de la següent manera:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

$$\mathbf{P(A \mid B) = P(A)}$$

És a dir , el condicionament no afecta.

Exemple: Tornem a considerar com a fenomen aleatori el llançament d'un dau i establim els següents esdeveniments A, B i C:

A = puntuació parell → $P(A) = P(\{2, 4, 6\}) = 3/6 = 1/2$
B = puntuació ≥ 4 → $P(B) = P(\{4, 5, 6\}) = 3/6$
C = puntuació > 4 → $P(C) = P(\{5, 6\}) = 2/6$

$P(A \cap B) = P(\{4, 6\}) = 2/6$
$P(A \cap C) = P(\{6\}) = 1/6$

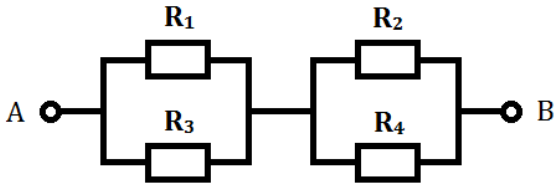
Calculem P(A | B) i observem que, com que P(A | B) ≠ P(A), A i B són dependents:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/6}{3/6} = 2/3$$

Calculem P(A | C) i observem que, com que P(A | C) = P(A), A i C són independents:

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{1/6}{2/6} = 1/2$$

Exemple: *Suposem que tenim 4 components idèntics i independents. La probabilitat que el corrent passi per R_i és $p_i = 0.95$. Aleshores, quant val la probabilitat que arribi el corrent des de A fins a B (P(AB))?*



Esdeveniment R_i = corrent passa per R_i

Tenint en compte que $(A \cup B)^C = A^C \cap B^C$ per a esdeveniments qualssevol A i B:

$P(AB) = P((R_1 \cup R_2) \cap (R_3 \cup R_4)) = P(R_1 \cup R_2) \cdot P(R_3 \cup R_4) =$
$= [1 - P((R_1 \cup R_2)^c)] \cdot [1 - P((R_3 \cup R_4)^c)] = [1 - P(R_1^c \cap R_2^c)] \cdot [1 - P(R_3^c \cap R_4^c)] =$
$= [1 - P(R_1^c) \cdot P(R_2^c)] \cdot [1 - P(R_3^c) \cdot P(R_4^c)] =$
$= [1 - (1 - p_1) \cdot (1 - p_2)] \cdot [1 - (1 - p_3) \cdot (1 - p_4)] =$
$= [1 - (1 - 0.95) \cdot (1 - 0.95)] \cdot [1 - (1 - 0.95) \cdot (1 - 0.95)]$

Per tant, P(AB) = 0.995

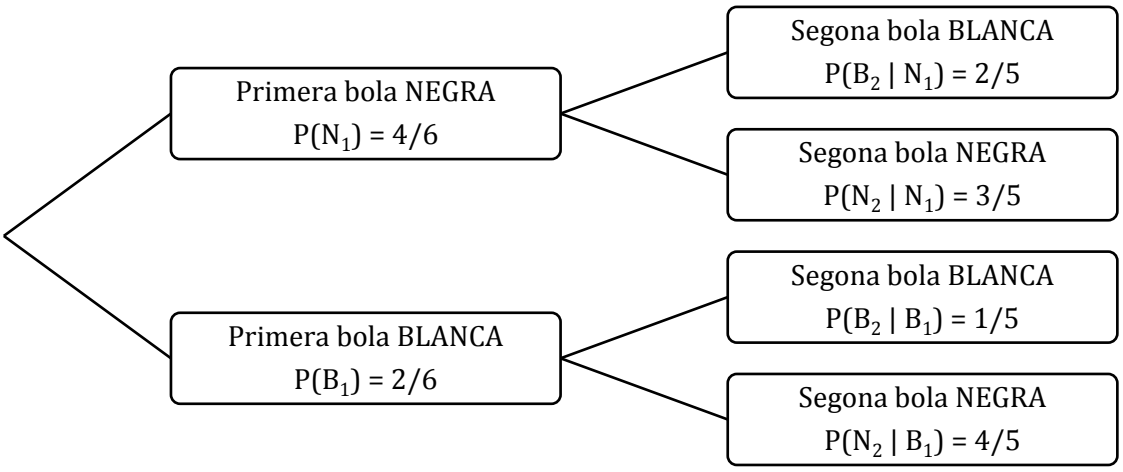
5. Arbres de probabilitat

Els arbres de probabilitat són **diagrames** que ens permeten calcular, més fàcilment, les probabilitats de diferents esdeveniments. Per realitzar-los hem de tenir en compte tots els possibles resultats d'un fenomen aleatori i les respectives probabilitats.

Exemple: *Tenim una caixa amb sis boles, 4 de negres i 2 de blanques. Prenem com a fenomen aleatori l'extracció successiva de dues boles sense devolució. Quina és la probabilitat que la segona bola sigui BLANCA? I la probabilitat que la primera bola sigui NEGRA sabent que la segona ha estat BLANCA?*

Esdeveniment B_1 = primera bola blanca
Esdeveniment B_2 = segona bola blanca
Esdeveniment N_1 = primera bola negra
Esdeveniment N_2 = segona bola negra

ARBRE DE PROBABILITAT:



Per tant, podem calcular les següents probabilitats:

$P(N_1 \cap B_2) = P(N_1) \cdot P(B_2 \mid N_1) = 4/6 \cdot 2/5 = 4/15$
$P(N_1 \cap N_2) = P(N_1) \cdot P(N_2 \mid N_1) = 4/6 \cdot 3/5 = 6/15$
$P(B_1 \cap B_2) = P(B_1) \cdot P(B_2 \mid B_1) = 2/6 \cdot 1/5 = 1/15$
$P(B_1 \cap N_2) = P(B_1) \cdot P(N_2 \mid B_1) = 2/6 \cdot 4/5 = 4/15$

- Probabilitat que la segona bola sigui BLANCA:

$$P(B_2) = P(N_1 \cap B_2) + P(B_1 \cap B_2) = 4/15 + 1/15 = 1/3$$

- Probabilitat que la primera bola sigui NEGRA sabent que la segona ha estat BLANCA:

$$P(N_1|B_2) = \frac{P(N_1 \cap B_2)}{P(B_2)} = \frac{4/15}{1/3} = 4/5$$

PROBLEMES

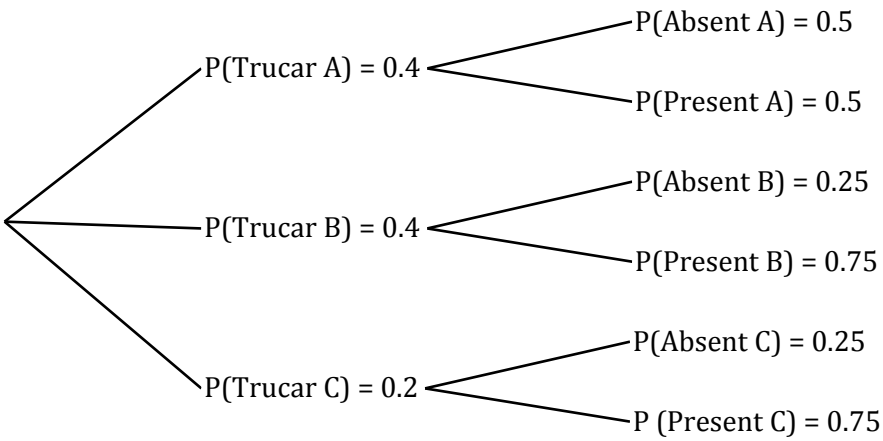
1. Exercicis resolts

1.1. *En una oficina amb un únic telèfon hi ha tres persones A, B i C. Les trucades que es reben a l'oficina es produeixen aleatòriament durant el dia en les proporcions següents: 2/5 dirigides a la persona A, 2/5 dirigides a la persona B i 1/5 dirigides a la persona C. Per motius de la seva feina A, B i C han d'absentar-se de l'oficina aleatòriament i independentment, de manera que A és absent la meitat del seu horari laboral, mentre que tant B com C són absents la quarta part del seu horari laboral.*

Per a les trucades que es reben a l'oficina en horari laboral, trobeu la probabilitat que:

a) No hi hagi ningú per atendre la trucada.
b) Que una trucada pugui ser atesa per la persona a la qual va dirigida.
c) Que tres trucades consecutives vagin dirigides a la mateixa persona.
d) Que tres trucades consecutives vagin dirigides a tres persones diferents.
e) Que una persona que vol contactar amb B hagi de trucar més de 3 vegades per localitzar-la.
f) S'ha rebut una trucada que se sap que ha pogut ser atesa per la persona a la qual anava dirigida. Calculeu la probabilitat que anés dirigida a A.

De l'enunciat podem extreure les següents probabilitats:



a) La probabilitat que no hi hagi ningú per atendre la trucada serà la mateixa que la probabilitat que siguin absents A, B i C:

P(No hi ha ningú) =

= P(Absent A ∩ Absent B ∩ Absent C) =

= P(Absent A) · P(Absent B) · P(Absent C) = 0.5 · 0.25 · 0.25 = **1/32**

b) La probabilitat que una trucada pugui ser atesa per la persona a la qual va dirigida serà la probabilitat que, al trucar a A, B o C, aquest es trobi present a l'oficina:

Tenint present que no hi ha intersecció en trucar a A , B o C:

P(Trucada atesa) =

= P[(Trucar A ∩ Present A) ∪ (Trucar B ∩ Present B) ∪ (Trucar C ∩ Present C)] =

= P(Trucar A) · P(Present A) + P(Trucar B) · P(Present B) +

P(Trucar C) · P(Present C) = 0.4 · 0.5 + 0.4 · 0.75 + 0.2 · 0.75 = **13/20**

c) La probabilitat que tres trucades seguides vagin dirigides a la mateixa persona es pot escriure de la següent manera:

P(Tres trucades seguides al mateix) =

= P([Trucar A ∩ Trucar A ∩ Trucar A] ∪ [Trucar B ∩ Trucar B ∩ Trucar B] ∪

[Trucar C ∩ Trucar C ∩ Trucar C]) =

= P(Trucar A) · P(Trucar A) · P(Trucar A) + P(Trucar B) · P(Trucar B) · P(Trucar B)

+ P(Trucar C) · P(Trucar C) · P(Trucar C) = $0.4^3 + 0.4^3 + 0.2^3$ = **17/125**

d) La probabilitat que tres trucades seguides vagin dirigides a persones diferents es pot escriure de la següent manera:

P(Tres trucades seguides vagin dirigides a tres persones diferents) =

= P([Trucar A ∩ Trucar B ∩ Trucar C] ∪ [Trucar A ∩ Trucar C ∩ Trucar B] ∪

[Trucar B ∩ Trucar A ∩ Trucar C] ∪ [Trucar B ∩ Trucar C ∩ Trucar A] ∪

[Trucar C ∩ Trucar A ∩ Trucar B] ∪ [Trucar C ∩ Trucar B ∩ Trucar A]) =

= P(Trucar A) · P(Trucar B) · P(Trucar C) +

P(Trucar A) · P(Trucar C) · P(Trucar B) +

P(Trucar B) · P(Trucar A) · P(Trucar C) +

P(Trucar B) · P(Trucar C) · P(Trucar A) +

P(Trucar C) · P(Trucar A) · P(Trucar B) +

P(Trucar C) · P(Trucar B) · P(Trucar A) = 6 · (0.4 · 0.4 · 0.2) = **24/125**

e) La probabilitat que una persona hagi de trucar més de 3 vegades a B per localitzar-la és la probabilitat que B sigui absent els 3 cops, per tant:

P(Mínim 3 trucs seguits a B absent) =

= P(Absent B ∩ Absent B ∩ Absent B) =

= P(Absent B) · P(Absent B) · P(Absent B) = 0.25^3 = **1/64**

f) La probabilitat que una trucada vagi dirigida a A sabent que aquesta ha estat atesa és:

P(Trucar A | Trucada atesa) =

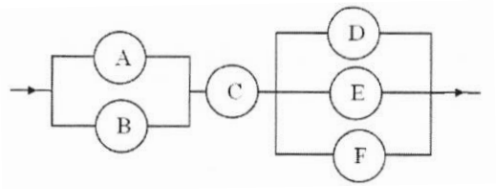
= P(Trucar A ∩ Trucada atesa) / P(Trucada atesa) =

= P(Trucar A) · P(Trucada atesa | Trucar A) / P(Trucada atesa) =

= P(Trucar A) · P(Present A | Trucar A) / P(Trucada atesa) =

= 0.4 · 0.5 / 0.65 = **4/13**

1.2. *Considerem el circuit de la figura:*



Cada component pot funcionar correctament amb les següents probabilitats:

P(A) = 0.9; P(B) = 0.8; P(C) = 0.95; P(D) = 0.9; P(E) = 0.9; P(F) = 0.5

Quina és la probabilitat que el circuit funcioni?

Per començar, dividim el circuit en tres parts: AB, C i DEF.

El corrent passarà per AB quan funcionin A i B, quan funcioni A (B no) o bé quan funcioni B (A no):

P(AB) =

= $P(A \cap B) \cup P(A \cap B^c) \cup P(A^c \cap B)$ =

= $P(A) \cdot P(B) + P(A) \cdot P(B^c) + P(A^c) \cdot P(B) = 0.9 \cdot 0.8 + 0.9 \cdot 0.2 + 0.1 \cdot 0.8 = 0.98$

El corrent passarà per C quan aquest component funcioni:

P(C) = 0.95

El corrent passarà per DEF quan funcionin D, E i F, quan funcionin D i E (F no), quan funcionin D i F (E no), quan funcionin E i F (D no), quan funcioni D (E i F no), quan funcioni E (D i F no) o bé quan funcioni F (D i E no):

P(DEF) =

= $P(D \cap E \cap F) \cup P(D \cap E \cap F^c) \cup P(D \cap E^c \cap F) \cup P(D^c \cap E \cap F) \cup P(D \cap E^c \cap F^c) \cup$

$P(D^c \cap E \cap F^c) \cup P(D^c \cap E^c \cap F) =$

$= P(D) \cdot P(E) \cdot P(F) + P(D) \cdot P(E) \cdot P(F^c) + P(D) \cdot P(E^c) \cdot P(F) + P(D^c) \cdot P(E) \cdot P(F) +$
$P(D) \cdot P(E^c) \cdot P(F^c) + P(D^c) \cdot P(E) \cdot P(F^c) + P(D^c) \cdot P(E^c) \cdot P(F) =$
$= 0.9 \cdot 0.9 \cdot 0.5 + 0.9 \cdot 0.9 \cdot 0.5 + 0.9 \cdot 0.1 \cdot 0.5 + 0.1 \cdot 0.9 \cdot 0.5 + 0.9 \cdot 0.1 \cdot 0.5 +$
$0.1 \cdot 0.9 \cdot 0.5 + 0.1 \cdot 0.1 \cdot 0.5 = 0.995$

La probabilitat que el circuit funcioni, és a dir, la probabilitat que el corrent passi pel circuit serà la mateixa que la probabilitat que el corrent passi per AB, per C i per DEF:

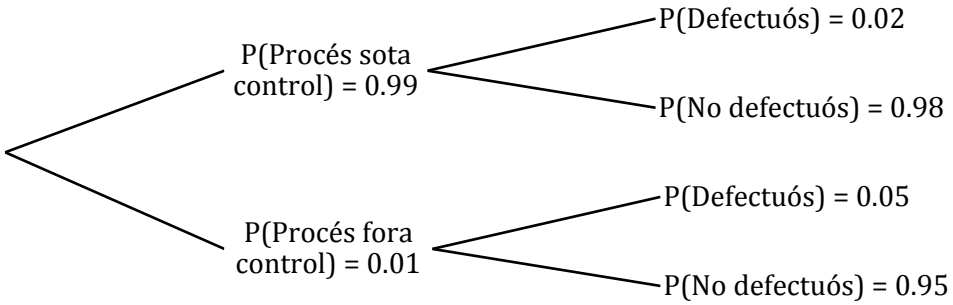
P(AB) ∩ P(C) ∩ P(DEF) = $0.98 \cdot 0.95 \cdot 0.955$ = **0.9263**

1.3. *Una fàbrica de components electrònics quan el procés de fabricació està sota control produeix un 2% d'unitats defectuoses, mentre que aquest percentatge és del 5% si està fora de control. Se sap que el 99% de les vegades el procés es troba sota control.*

Es demana:

Quina és la probabilitat que escollit a l'atzar un component, aquest sigui defectuós? S'agafa a l'atzar un component i resulta que no és defectuós, quina és la probabilitat que el procés estigui fora de control?

De l'enunciat podem extreure les següents probabilitats:



La probabilitat que un producte sigui defectuós ve donada per:

P(Defectuós) =
= P(Procés sota control ∩ Defectuós) ∪ P(Procés fora control ∩ Defectuós) =
= P(Procés sota control) · P(Defectuós) + P(Procés fora control) · P(Defectuós) =
= $0.99 \cdot 0.02 + 0.01 \cdot 0.05$ = **0.0203**

La probabilitat que el procés estigui fora de control sabent que el producte no és defectuós ve donada per:

P(Procés fora control | No defectuós) =

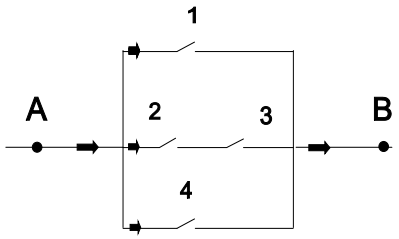
= P(Procés fora control ∩ No defectuós) / P(No defectuós) =

= P(Procés fora control) · P(No defectuós | Procés fora control) /

[1 – P(No defectuós)] = 0.01 · 0.95 / (1 – 0.0203) = **0.009697**

2. Exercicis proposats

2.1. *Es vol enviar un senyal des del punt A al B a través de la xarxa de comunicació que es mostra a la figura. El senyal es bifurca per les 3 branques intentant arribar fins al punt B. Els nodes 1, 2, 3 i 4 poden estar oberts (no passa el senyal) o tancats (passa el senyal) de forma independent. La probabilitat que un node estigui tancat és igual a p. Es demana:*



a) Calculeu, en funció de p, la probabilitat que un senyal enviat des de A arribi fins a B.
b) S'ha enviat un senyal des del punt A que finalment no ha arribat a B. Calculeu, en funció de p, la probabilitat que el node 3 estigui obert.
c) S'ha enviat un senyal des del punt A que sí ha arribat a B. Calculeu, en funció de p, la probabilitat que dos nodes estiguin oberts i dos tancats.

Solució: a) $p(p^3 - 2p^2 + 2)$; b) $\frac{1}{1+p}$; c) $\frac{6(1-p)^2 p}{p^3-2p^2+2}$

2.2. *Una urna conté 1 bola blanca i 3 boles negres. Quatre jugadors A, B, C i D treuen –per aquest ordre i sense reemplaçament– una bola de la urna. Guanya el primer jugador que treu la bola blanca. Calculeu la probabilitat que té cada jugador de guanyar la partida. Creieu que l'ordre en què els jugadors treuen les boles té influència en el resultat final?*

Solució: 0.25. No

2.3. *Es tenen dues monedes: una d'elles és normal i simètrica mentre que l'altra té dues cares. S'escull a l'atzar una de les monedes i es fa un llançament. Suposant que la probabilitat d'escollir la moneda normal és 3/4 i que s'ha obtingut CARA en el llançament, quina és la probabilitat que s'hagi escollit la moneda trucada?*

Solució: 0.4

2.4. *En una bossa A hi ha 5 boles negres i 3 blanques, mentre que en una altra bossa B hi ha 1 bola negra i 2 blanques. Es tira un dau i, si surt un 1 o un 2, es treu a l'atzar una bola de la bossa B i, sense mirar el seu color, s'introdueix a la bossa A. A continuació s'extreu a l'atzar una bola de la bossa A. Si la puntuació del dau és major que 2, s'extreu a l'atzar una bola de la bossa A i, sense mirar el seu color, s'introdueix a la bossa B. A continuació s'extreu a l'atzar una bola de la bossa B.*

Es demana:

a) Calculeu la probabilitat que la bola extreta la segona vegada sigui de color negra.
b) Si la bola extreta la segona vegada resulta que és de color negre, quina probabilitat hi ha que la bola extreta en primer lloc també ho sigui?

Solució: a) 607/1296; b) 366/607

2.5. *Un canal de telecomunicació transmet missatges codificats en un sistema binari. La probabilitat que sigui emès el senyal 0 és p i la probabilitat que sigui emès el senyal 1 és 1 – p. Certes pertorbacions en la transmissió, anomenades soroll de fons, poden alterar el senyal emès –canviant 0's per 1's i 1's per 0's– essent p_0 la probabilitat d'alteració quan el senyal emès és 0, i p_1 la probabilitat d'alteració quan el senyal emès és 1.*

Es demana:

a) Si s'ha rebut el senyal 0, quina és la probabilitat que el senyal emès hagi estat efectivament el 0?
b) Si s'ha rebut el senyal 1, quina és la probabilitat que el senyal emès hagi estat el 0?

Solució: a) $\frac{p(1-p_0)}{p(1-p_0)+p_1(1-p)}$; b) $\frac{p\cdot p_0}{p\cdot p_0+(1-p)(1-p_1)}$

2.6. *En un sistema d'alarma, la probabilitat que es produeixi un perill és p = 0.1. Si es produeix el perill, la probabilitat que l'alarma funcioni és p_1 = 0.95, i la probabilitat que l'alarma funcioni sense que s'hagi produït el perill és p_2 = 0.03.*

Calculeu:

a) La probabilitat que, havent funcionat l'alarma, el perill no s'hagi presentat.
b) La probabilitat que hi hagi un perill i l'alarma funcioni.
c) La probabilitat que, no havent funcionat l'alarma, hi hagi un perill.

Solució: a) 27/122; b) 19/200; c) 5/878

2.7. *En una certa instal·lació industrial, dues màquines M1 i M2 ocupen respectivament el 10% i el 90% de la producció total d'un determinat article. La probabilitat que la primera màquina fabriqui una peça defectuosa és $p_1 = 0.01$, i la probabilitat que fabriqui una peça defectuosa la segona màquina és $p_2 = 0.05$. Agafant a l'atzar una peça de la producció d'un dia, s'observa que és defectuosa. Quina és la probabilitat que aquesta peça procedeixi de la primera màquina?*

Solució: 1/46

2.8. *Si un ordinador personal està contaminat per un determinat virus V, un programa PR_1 detecta la seva presència amb probabilitat $p_1 = 0.92$. Si l'ordinador no té el virus V, el programa detecta efectivament la seva absència amb probabilitat $p_2 = 0.87$. S'estima que la probabilitat que un ordinador contingui el virus V és igual a 0.32.*

Es demana:

a) Probabilitat que l'ordinador contingui realment el virus V quan el programa PR1 detecta la seva presència.
b) Probabilitat que l'ordinador no contingui efectivament el virus V quan el programa PR1 no detecta la seva presència.
c) Probabilitat que el programa PR1 realitzi una diagnosi correcta de la presència o no del virus V en un ordinador personal escollit a l'atzar.
d) Un segon programa PR2, preparat per detectar el mateix tipus de virus V, té probabilitats $p_1 = 0.99$ i $p_2 = 0.82$, respectivament. Quin dels dos programes, PR1 o PR2, és més eficient? Raoneu la vostra resposta.

Solució: a) 0.791; b) 0.9585; c) 0.886; d) PR_1

2.9. *Dues màquines A i B estan funcionant correctament. La probabilitat que la màquina A continuï funcionant correctament durant 10 dies més és igual a 1/4, mentre que la probabilitat que ho faci la màquina B és igual a 1/3. El funcionament d'una de les màquines no influeix en el de l'altra.*

Calculeu la probabilitat que:

a) Passats 10 dies, les dues màquines continuïn funcionant correctament.
b) Passats 10 dies, una de les màquines –com a mínim– funcioni correctament.
c) Passats 10 dies, cap de les dues màquines funcioni correctament.
d) Passats 10 dies, només funcioni correctament la màquina B.
e) Passats 10 dies, només funcioni correctament una de les dues màquines.

Solució: a) 1/12; b) 1/2; c) 1/2; d) 1/4; e) 5/12

2.10. *Siguin A i B dos esdeveniments associats a un fenomen aleatori de manera que:*

$$P(A) = 1/2,\ P(B) = 1/3 \text{ i } P(A \cap B) = 1/4$$

Es demana:
a) P(A | B)
b) P(B | A)
c) P(A ∪B)
d) P(A^c | B^c)
e) P(B^c | A^c)
f) P(A^c | B)
g) P(A | B^c)

Solució: a) 3/4; b) 2/4; c) 7/12; d) 5/8; e) 5/6; f) 1/4; g) 3/8

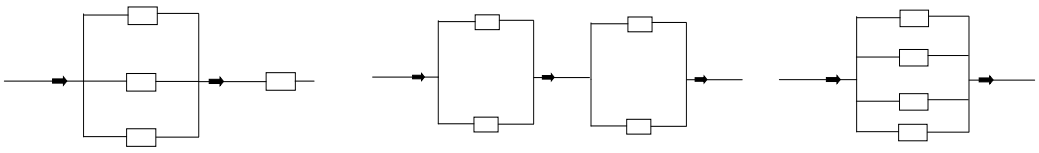
2.11. *Una empresa dedicada a la fabricació de motocicletes vol llançar al mercat un nou model per l'any 2010. No obstant això, la possible situació econòmica en aquest any presenta tres alternatives: inflació, estabilitat o depressió. Suposant les tres situacions equiprobables i que les probabilitats de llançar el nou model al mercat segons les tres alternatives són: 0.7 si existeix inflació, 0.4 si existeix estabilitat i 0.2 si existeix depressió.*

Es demana:

a) La probabilitat que el nou model estigui en el mercat l'any 2010.
b) La probabilitat que, estant ja el nou model en el mercat, existeixi depressió.

Solució: a) 0.43; b) 0.153

2.12. *S'ha utilitzat quatre components electrònics, idèntics i independents, del model CE-123 per a construir tres circuits diferents (Figures A, B, i C). Segons les especificacions del fabricant del component CE-123, la probabilitat que un d'aquests elements funcioni correctament és igual a 0.95. Calculeu en cada cas, quant val la probabilitat que el circuit funcioni correctament.*



A **B** **C**

Solució: P(A) = 0.94988; P(B) = 0.995; P(C) = 0.99999

TEMA 3: Lleis de probabilitat contínua

TEORIA

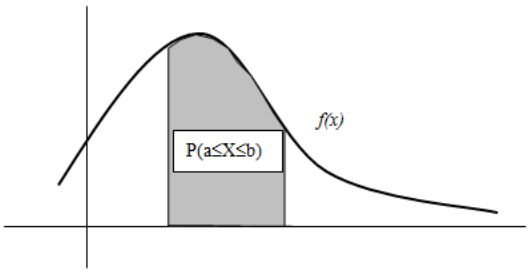
1. Variable aleatòria

Una variable X es diu que és aleatòria (v.a.) quan, a priori, no se sap exactament quin valor pren però si que es coneix quins valors numèrics pot arribar a prendre i la probabilitat que els adquireixi.

Direm que una v.a. X és **contínua** quan aquesta pot prendre valors de forma contínua al llarg de la recta real. Per saber amb quina probabilitat la v.a. X pot prendre aquests valors, es proporciona una funció real positiva f(x) anomenada **funció de densitat**. Aquesta funció es pot considerar com el límit d'un histograma quan el nombre d'intervals es fa molt gran i l'amplada molt petita.

1.1. Funció de densitat f(x)

La funció de densitat d'una v.a. proporciona la probabilitat que aquesta variable prengui algun valor comprés entre dos números a i b qualsevols. Aquesta probabilitat s'interpreta com l'**àrea** que queda sota la gràfica de la funció f(x) entre els punts a i b. Per això aquesta funció només té significat quan es busca la probabilitat entre dos punts.



$$\mathbf{P(a \leq X \leq b) = \int_a^b f(x)\,dx}$$

La funció de densitat f(x) es caracteritza per les següents **propietats**:

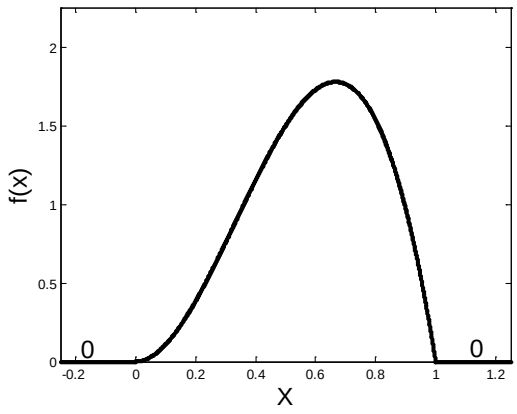
- La funció de densitat és sempre no negativa: $\mathbf{f(x) \geq 0}$
- L'àrea total és sempre igual a 1: $\mathbf{\int_{-\infty}^{+\infty} f(x)dx = 1}$
- La probabilitat que una v.a. agafi un valor puntual és zero: $\mathbf{P(X = x) = 0}$
- $\mathbf{P\left(x - \frac{\Delta x}{2} \leq X \leq x + \frac{\Delta x}{2}\right) = f(x) \cdot \Delta x}$

Exemple: *Volem que la funció $f(x) = kx^2(1 - x)$ sigui una funció de densitat per $0 \leq x \leq 1$. Quin valor de k permetrà complir el requisit?*

Es sap que l'àrea total de la funció de densitat ha de ser 1, per això es planteja que:

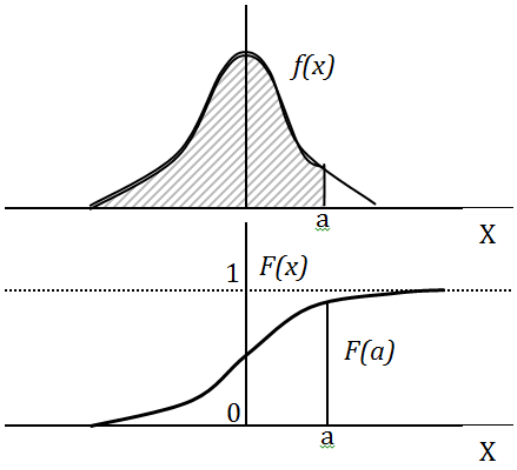
$$\int_0^1 f(x)\,dx = \int_0^1 k \cdot x^2 \cdot (1-x)\,dx = k \cdot \left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1 = \frac{k}{12}$$

Per tant, si k/12 ha de ser igual a 1 per tal que la funció f(x) sigui una funció de densitat, el valor de k ha de ser de 12. Finalment, representem f(x):



1.2. Funció de distribució F(x)

La funció de distribució d'una v.a. representa la **probabilitat acumulada** fins a un valor a. Aquesta probabilitat s'interpreta com l'àrea que queda sota la gràfica de la funció f(x) entre el –∞ i el punt a . Aquesta probabilitat és el valor que assoleix F(x) per x = a.



$$\mathbf{F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)\,dx}$$

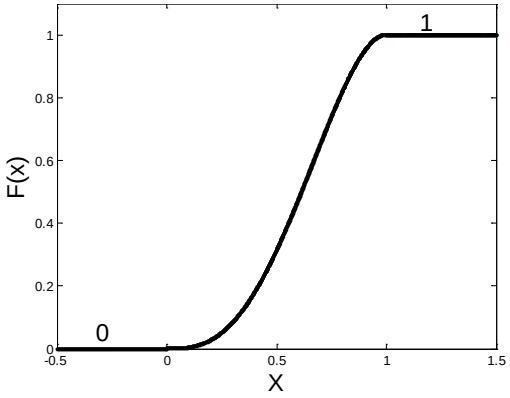
La funció de distribució F(x) es caracteritza per les següents **propietats**:

- La funció de distribució únicament pot agafar valors entre 0 i 1: $0 \leq F(x) \leq 1$. Per això, $\mathbf{F(-\infty) = 0}$ i $\mathbf{F(\infty) = 1}$.
- La probabilitat no acumulada d'un punt a és tota la probabilitat de la v.a. menys la probabilitat acumulada: $\mathbf{P(X > x) = 1 - F(x)}$

- La probabilitat compresa entre dos punts és la resta de probabilitats acumulades d'aquests dos punts: $\mathbf{P(a \leq X \leq b) = F(b) - F(a)}$

Exemple: *A partir de la funció de densitat $f(x) = 12x^2(1 - x)$ trobem la funció de distribució i la representem.*

$$F(x) = \int f(x)\,dx = \int 12 \cdot x^2 \cdot (1 - x)\,dx = 12 \cdot \left[\frac{x^3}{3} - \frac{x^4}{4}\right]$$



2. Estadístics per variables aleatòries contínues

Un grup de dades pot tenir associat diferents estadístics representatius de la seves propietats, per tant podem donar estadística associats a variables aleatòries contínues. Per les v.a. contínues ens centrarem amb els dos estadístics més representatius, l'**esperança** i la **variància**.

2.1. Esperança E{X}

La mitjana d'una v.a. contínua rep el nom d'esperança. Si considerem una v.a. de n mostres amb les dades agrupades en K classes i ajustem la freqüència absoluta n_i de cada marca de classe c_i a una funció de densitat $f(c_i)$, podem escriure el següent:

$$\bar{x} = \frac{\sum_{i=1}^{K} c_i n_i}{n} \rightarrow \bar{x} = \sum_{i=1}^{K} c_i\, f(c_i) \text{ on } f(c_i) = \frac{n_i}{n}$$

A partir de l'expressió anterior s'escriu l'expressió d'esperança:

$$E\{X\} = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

L'esperança es caracteritza per les següents **propietats**:

- L'esperança de la suma de dues v.a. és la suma de les esperances de dues v.a.: $E\{X + Y\} = E\{X\} + E\{Y\}$
- L'esperança és lineal: $E\{a + bX\} = a + bE\{X\}$

- Si X i Y són dues v.a. independents, l'esperança del producte de les dues v.a. és el producte de les esperances de les dues v.a: $E\{X \cdot Y\} = E\{X\} \cdot E\{Y\}$

Exemple: *Continuant amb la funció de densitat anterior, calculem la seva esperança per 0 ≤ x ≤ 1.*

$$E\{X\} = \int_0^1 x \cdot f(x)\, dx = \int_0^1 x \cdot 12 \cdot x^2 \cdot (1-x)\, dx = 12\left[\frac{x^4}{4} - \frac{x^5}{5}\right]_0^1 = 0.6$$

2.2. Variància var{X}

La variància d'una v.a. contínua és l'esperança del quadrat de les distàncies de tots els valors de X respecte l'esperança E{X}.

$$var\{X\} = E\{(X - E\{X\})^2\} = \int_{-\infty}^{+\infty} (x - E\{X\})^2 f(x)\, dx$$

La variància es caracteritza per les següents **propietats**:

- La variància no és lineal: $var\{aX\} = a^2 var\{X\}$ i $var\{X + a\} = var\{X\}$
- Si X i Y són dues v.a. independents, la variància de la suma o resta de les dues v.a. és la suma de les variàncies de les dues v.a.: $var\{X + Y\} = var\{X - Y\} = var\{X\} + var\{Y\}$
- $var\{X\} = E\{X^2\} - E^2\{X\}$
- $var\{X\} = \int_{-\infty}^{+\infty} x^2 f(x)\, dx - E^2\{X\}$
- La desviació desv{X} serà l'arrel positiva de la variància

Exemple: *Continuant amb la funció de densitat anterior, calculem la seva variància per 0 ≤ x ≤ 1.*

$$var\{X\} = \int_0^1 x^2 f(x)\, dx - E^2\{X\} = \int_0^1 x^2 \cdot 12 \cdot x^2 \cdot (1-x) dx - 0.6^2 =$$

$$= 12 \cdot \left[\frac{x^5}{5} - \frac{x^6}{6}\right]_0^1 - 0.6^2 = 0.04$$

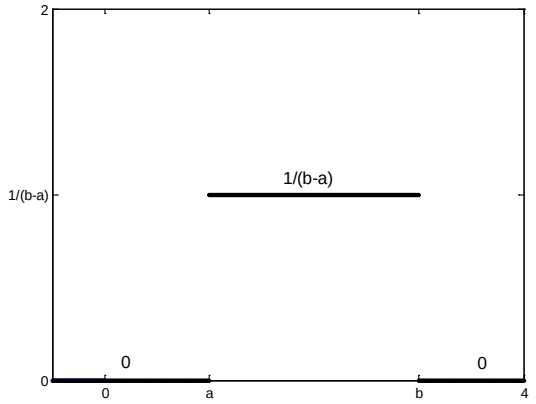
També podem dir que la desviació d'aquesta funció de densitat és de 0.2.

3. Distribucions per variables aleatòries contínues

3.1. Distribució uniforme contínua en un interval

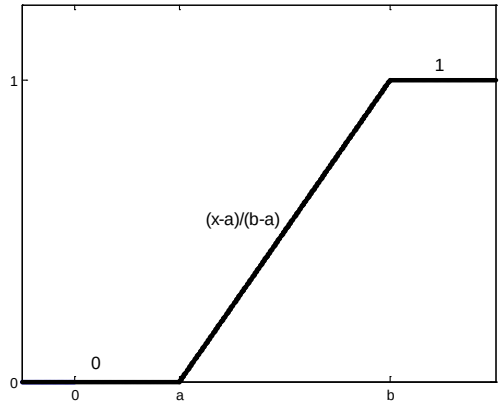
La distribució uniforme contínua en un interval [a, b] modela una situació d'**equiprobabilitat**, és a dir, la probabilitat d'obtenir un valor a l'atzar dins d'aquest interval és igual. S'anota com **X~U[a, b]**, el que significa que la v.a. X segueix una distribució uniforme U dins l'interval [a, b].

La **funció de densitat** d'aquesta distribució es caracteritza per una funció a trossos, on fora de l'interval val 0 i dins val tota la probabilitat (1) repartida entre tots els valors que conté l'interval, obtenint així una situació d'equiprobabilitat.



$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } x \in [a,b] \\ 0 & \text{si } x \notin [a,b] \end{cases}$$

La **funció de distribució** es caracteritza també per una funció a trossos, resultat d'integrar la funció de densitat anterior.



$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \dfrac{1}{b-a} \cdot (x-a) & \text{si } x \in [a,b] \\ 1 & \text{si } x > b \end{cases}$$

Per tant, els **estadístics** d'una distribució uniforme contínua en un interval seran els següents:

$$E\{X\} = \frac{a+b}{2}$$

$$var\{X\} = \frac{(b-a)^2}{12}$$

Exemple: *Considerem X una variable que és el temps transcorregut per un producte des de la seva venda fins que entra en el servei tècnic per ser reparat. El temps de*

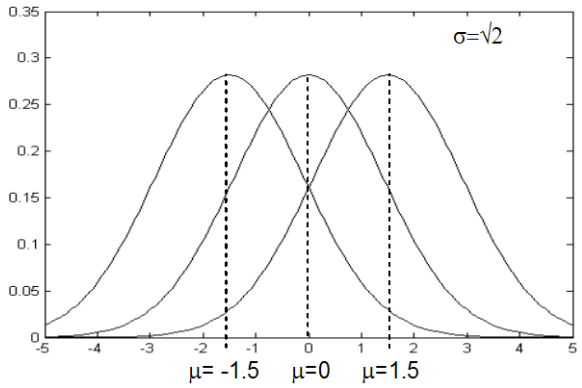
garantia és de 2 anys, i se'ns diu que la probabilitat de fallada és uniforme al llarg del temps de garantia. Calculem l'esperança i la variància de la distribució.

Primer de tot i a partir de les dades de l'enunciat identifiquem que es tracta d'una distribució uniforme contínua en un interval, i per tant podem escriure X~U[0,2].

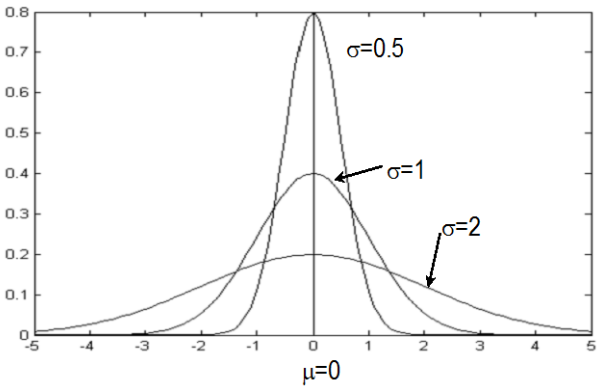
Per tant podem calcular que E{X} = 1 i var{X} = 1/3.

3.2. Distribució normal de Gauss-Laplace

La distribució normal o llei Normal és aquella que a partir d'un valor central μ i una variabilitat σ es modela la probabilitat d'obtenir un valor a l'atzar d'una v.a. amb aquests paràmetres. S'anota com **X~N(μ,σ) o X~N(μ,σ²)**, el que significa que la v.a. X segueix una distribució normal amb paràmetres **μ** i **σ**.

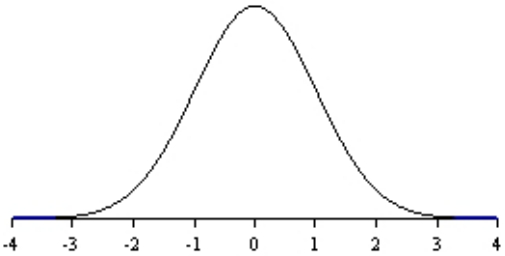


El paràmetre de valor central, de posició o μ indica on es troba situat el centre de la distribució normal. Aquest coincideix amb l'esperança E{X}, la mediana i la moda.



El paràmetre de forma o σ regula el grau de dispersió de les dades d'una v.a. al voltant de μ. Aquest paràmetre és sempre positiu i correspon a la desviació estàndard desv{X}.

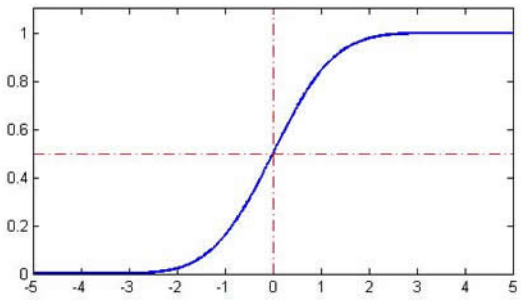
La **funció de densitat** d'aquesta distribució té una forma que rep el nom de **Campana de Gauss**. Es caracteritza perquè els valors més propers al valor central μ són més probables d'obtenir que els que estan més allunyats.



$$\frac{1}{\cdot\sqrt{2\cdot\pi}}\cdot e^{\frac{-1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{per } -\infty < x < +\infty$$

Cal dir que la funció de densitat és simètrica respecte el valor central μ, i que per tant, $f(\mu + a) = f(\mu - a)$.

La **funció de distribució** s'obté d'integrar la funció de densitat anterior.



$$F(x) = \int_{-\infty}^{+\infty} f(x)\,dx \qquad \text{per } -\infty < x < +\infty$$

La funció de distribució també es pot representar a partir de gràfica de la funció de densitat, marcant l'àrea de $-\infty$ al valor desitjat, el que representarà la probabilitat acumulada fins aquest valor.

Per no haver de realitzar la integral numèricament, ja que no té primitiva, cada cop que es necessita conèixer una probabilitat acumulada d'una v.a., hi ha tabulat el resultat d'aquest càlcul per tots els valors compresos entre –3.5 i 3.5 d'una normal amb paràmetres $\mu = 0$ i $\sigma = 1$, la qual rep el nom de **llei Normal estàndard**. Aquesta s'anotarà com **Z~N(0;1)**.

Per tant, qualsevol sigui la combinació de paràmetres que defineixi la distribució normal que segueix una v.a. X, interessarà tractar-la com una distribució normal estàndard Z per poder extreure de les taules estadístiques les probabilitats d'interès. D'aquest procés se'n diu **estandardització**, i es basa en el següent:

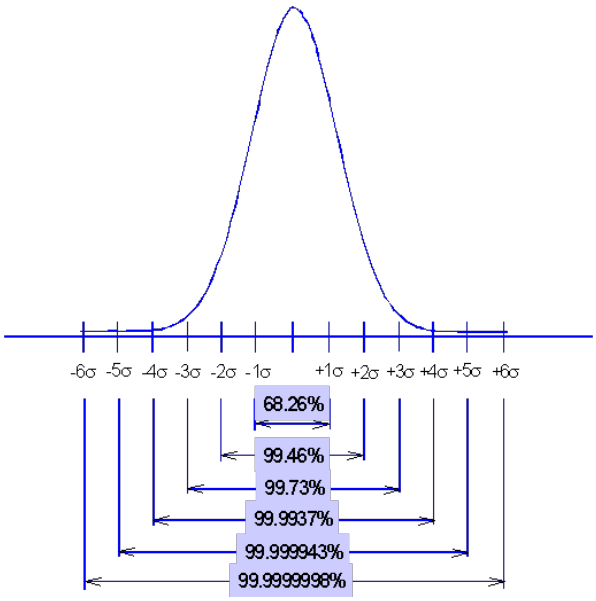
$$Z = \frac{X - \mu}{\sigma}$$

$$z_i = \frac{x_i - \mu}{\sigma}$$

Amb aquest càlcul s'aconsegueix passar un valor x_i d'una v.a. $X\sim N(\mu;\sigma)$ de la que no sabem la probabilitat acumulada, a un valor z_i d'una v.a. $Z\sim N(0;1)$ de la qual podem calcular la probabilitat acumulada a partir de les **taules estadístiques**. Aquesta probabilitat acumulada de z_i serà igual a la de x_i en les respectives distribucions normals.

Veiem tot un seguit de **probabilitats** importants:

- $P[X \leq \mu] = P[Z \leq 0] = 0.5$
- $P[\mu - \sigma \leq X \leq \mu + \sigma] = P[-1 \leq Z \leq +1] = 0.6826$
- $P[\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma] = P[-1.96 \leq Z \leq +1.96] = 0.95$
- $P[\mu - 2\sigma \leq X \leq \mu + 2\sigma] = P[-2 \leq Z \leq +2] = 0.9544$
- $P[\mu - 2.576\sigma \leq X \leq \mu + 2.576\sigma] = P[-2.576 \leq Z \leq +2.576] = 0.995$
- $P[\mu - 3\sigma \leq X \leq \mu + 3\sigma] = P[-3 \leq Z \leq +3] = 0.9975$



Exemple: *Considerem una v.a. X~N(100,2). Se'ns demana calcular el següent:*

- *Probabilitat que la v.a. agafi valors inferiors o iguals a 97.*
- *Probabilitat que la v.a. agafi valors superiors a 97.*
- *Probabilitat que la v.a. agafi valors entre 97 i 104.*

A priori observem que la normal amb què hem de treballar no és l'estàndard. Per això, el primer que fem és estandarditzar els valors amb què hem de treballar i la variable X en una variable Z~N(0;1):

$$\frac{X - 100}{2} = Z$$

$$\frac{97 - 100}{2} = -1.5 \qquad \frac{104 - 100}{2} = 2$$

Aquests dos valors obtinguts de –1.5 i 2 corresponents a una normal N(0,1) seran els que buscarem a les taules estadístiques de la distribució normal:

- Probabilitat acumulada de –1.5 $= P[Z \leq -1.5] = 0.0668$

- Probabilitat acumulada de 2 = $P[Z \leq 2] = 0.9772$

Podem reescriure l'enunciat amb el següent i donar el seu resultat:

- $P[X \leq 97] = P[Z \leq -1.5] = 0.0668 = 6.68\%$
- $P[X > 97] = 1 - P[X \leq 97] = 1 - P[Z \leq -1.5] = 1 - 0.0668 = 0.9332 = 93.32\%$
- $P[97 \leq X \leq 104] = P[-1.5 \leq Z \leq 2] = P[Z \leq 2] - P[Z \leq -1.5] = 0.9772 - 0.0668 = 0.9104 = 91.04\%$

La suma de la probabilitat que la v.a. X agafi un valor menor a un cert valor amb la probabilitat que la v.a. X agafi un valor major al mateix número és 1:

$$P[X \leq 97] + P[X > 97] = 0.0668 + (1 - 0.0668) = 0.0668 + 0.9332 = 1$$

En comptes de treballar amb una sola v.a. normal ens pot interessar fer **combinacions** entre dues o més **distribucions normals**. Siguin $X_1 \sim (\mu_1;\sigma_1)$ i $X_2 \sim (\mu_2;\sigma_2)$ **independents**, llavors podem tractar-les de la següent manera:

$$X_1 + X_2 \sim N\left(\mu_1 + \mu_2; \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

$$X_1 - X_2 \sim N\left(\mu_1 - \mu_2; \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

De forma genèrica i sabent que a és un nombre real, podem escriure el següent:

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n \sim N\left(a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n; \sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2}\right)$$

Així doncs, fent qualsevol combinació lineal de v.a. normals independents s'obté una normal.

Exemple: *X_k és una v.a. normal que defineix la producció diària de paper en metres d'una màquina k i a_k el cost en euros/metre de la mateixa. Volem calcular la v.a. T que descriurà el cost total de la producció diària per les n màquines d'una empresa*.

De forma genèrica podem escriure el següent:

$$\sum_{k=1}^{n} a_k \cdot X_k \sim N\left(a_1\mu_1 + \cdots + a_n\mu_n; \sqrt{a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2}\right) \sim T$$

Si vàries v.a. independents segueixen la mateixa distribució normal, llavors la mitjana d'aquestes:

$$\frac{X_1 + X_2 + \cdots + X_n}{n} = \overline{X}_n \sim \left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

Conseqüentment, alhora d'estandarditzar la mitjana haurem de tenir en compte el número de mostres, ja que aquestes reduiran la variabilitat de les dades:

$$Z = \frac{X - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z_i = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Exemple: *Considerem una v.a. X que segueix una distribució N(10;2). Volem veure com varia la distribució de la mitjana d'aquesta normal quan el nombre de mostres n és una, quatre o nou.*

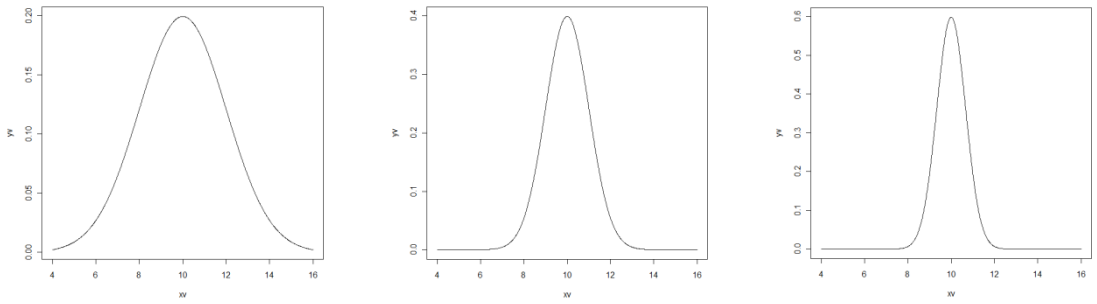
Quan $n = 1$, la normal que obtenim és $N(10; \frac{2}{\sqrt{1}}) = N(10;2)$

Quan $n = 4$, la normal que obtenim és $N(10; \frac{2}{\sqrt{4}}) = N(10;1)$

Quan $n = 9$, la normal que obtenim és $N(10; \frac{2}{\sqrt{9}}) = N(10;0.66)$

Aquest canvi en la variabilitat de les dades segons el nombre de mostres que es tingui, justifica la forma que agafen les diferents campanes de Gauss:

N(10;2) N(10;1) N(10;0.66)



Diverses v.a. independents amb **qualsevol distribució** es poden tractar conjuntament a partir del **teorema del límit central**. Aquest diu que quan el nombre de mostres tendeix a infinit, el total de la mitjana de les distribucions s'ajusta a una normal estàndard:

$$\lim_{n \to \infty} \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0; 1)$$

Això implica que la v.a. $\overline{X}_n$ s'ajusta a una normal sigui quina sigui la seva distribució:

$$\overline{X}_n \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

Al terme $\sigma/\sqrt{n}$ se'l coneix com **l'error estàndard de la mitjana mostral** (SE Mean). Com major sigui el nombre de mostres, la variabilitat del conjunt serà menor.

Exemple: *Tenim una v.a. X que segueix una distribució uniforme contínua en l'interval [9.5, 10.5]. Volem veure com es compleix el teorema del límit central, és a dir, que a mesura que augmentem el nombre de mostres amb aquesta distribució, el total s'aproxima millor a una normal.*

Primer de tot hem de calcular els estadístics de la distribució uniforme.

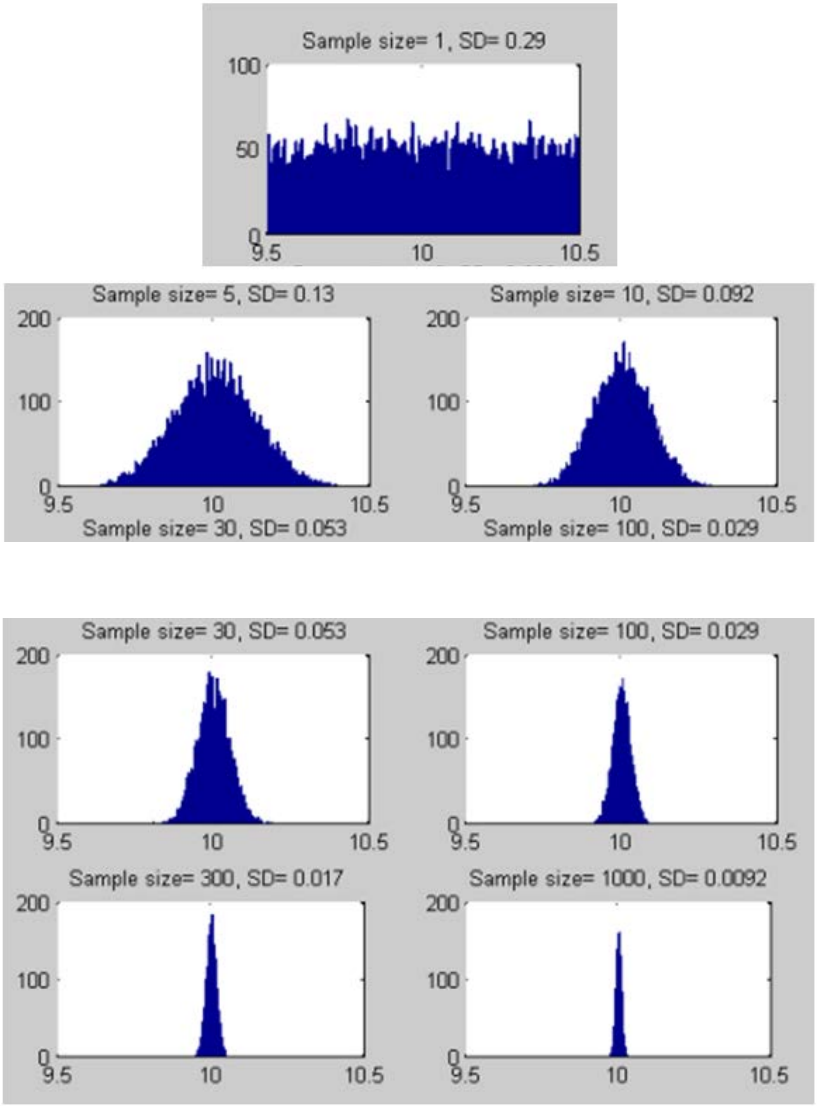
$$\mu = E\{X\} = \frac{a+b}{2} = \frac{9.5+10.5}{2} = 10$$

$$\sigma^2 = var\{X\} = \frac{(b-a)^2}{12} = \frac{(10.5-9.5)^2}{12} = 0.0833 \;\rightarrow\; \sigma = 0.2887$$

A partir d'aquests estadístics podem veure com disminueix la variància total a mesura que augmentem el nombre de mostres:

- Si n = 1, llavors $\overline{X}_1$ té $E(X) = 10$; $Desv(X) = 0.2887/\sqrt{1} = 0.2887$
- Si n = 5, llavors $\overline{X}_5$ té $E(X) = 10$; $Desv(X) = 0.2887/\sqrt{5} = 0.13$
- Si n = 10, llavors $\overline{X}_{10}$ té $E(X) = 10$; $Desv(X) = 0.2887/\sqrt{10} = 0.092$
- Si n = 30, llavors $\overline{X}_{30}$ té $E(X) = 10$; $Desv(X) = 0.2887/\sqrt{30} = 0.053$
- Si n = 100, llavors $\overline{X}_{100}$ té $E(X) = 10$; $Desv(X) = 0.2887/\sqrt{100} = 0.029$
- Si n = 300, llavors $\overline{X}_{300}$té $E(X) = 10$; $Desv(X) = 0.2887/\sqrt{300} = 0.017$
- Si n = 1000, llavors $\overline{X}_{1000}$té $E(X) = 10$; $Desv(X) = 0.2887/\sqrt{1000} = 0.0092$

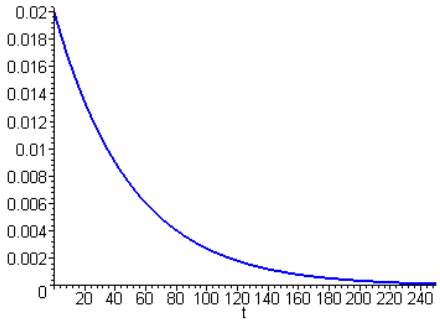
I gràficament observem la distribució uniforme que segueix realment la v.a. X i la distribució que segueix $\overline{X}_n$ que en augmentar n s'acosta més a la Normal:



3.3. Distribució exponencial

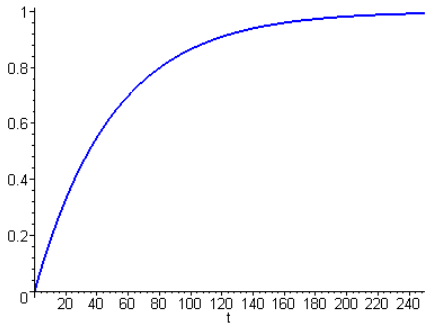
La distribució exponencial o llei del temps de vida és aquella que a partir d'un valor positiu **λ** es modelen diverses situacions com poden ser alguns temps de vida. S'anota com **X~Exp(λ)**, el que significa que la v.a. X segueix una distribució exponencial Exp amb el paràmetre λ.

La **funció de densitat** d'aquesta distribució és:



$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ \lambda \cdot e^{-\lambda \cdot x} & \text{si } x \geq 0 \end{cases}$$

La **funció de distribució** s'obté d'integrar la funció de densitat anterior. Aquesta representa la probabilitat de mort d'un component quan aquest ja té un cert temps de vida.



$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - e^{-\lambda \cdot x} & \text{si } x \geq 0 \end{cases}$$

Per tant, els **estadístics** d'una distribució exponencial seran els següents:

$$E\{X\} = \frac{1}{\lambda}$$

$$var\{X\} = \frac{1}{\lambda^2}$$

Anomenem T a la v.a. del **temps de vida d'un component**, és a dir, el temps que transcorre des que el component inicia el seu funcionament fins que s'espatlla. Així, enlloc de parlar d'una variable x parlarem d'una t, i obtenim les següents funcions:

- **Funció de densitat f(t):** no té cap ús directe alhora de fer càlculs. Aquesta funció defineix la distribució del temps de vida que segueix un component.
- **Funció de distribució o de mortalitat F(t):** d'aquesta es pot extreure la probabilitat que la v.a. T adquireixi valors menors a un cert temps t. $F(t) = P(T \leq t)$.
- **Funció de supervivència o fiabilitat R(t):** d'aquesta es pot extreure la probabilitat que la v.a. T adquireixi valors majors a un cert temps t. $R(t) = P(T > t) = 1 - F(t)$.

El paràmetre λ es coneix com **taxa instantània de fallada**. S'interpreta com la probabilitat que un component que en un instant concret està funcionant, falli en l'instant immediatament següent. Aquest paràmetre no sempre és constant, sinó que pot venir determinat per una funció $\lambda(t)$, la qual es calcula de la següent manera:

$$\lambda(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)}$$

En el cas que el paràmetre λ no sigui constant, pel càlcul de la funció de densitat i de distribució s'ha de tenir en compte:

$$f(t) = \lambda(t) \cdot e^{-\int \lambda(t)dt}$$
$$F(t) = 1 - e^{-\int \lambda(t)dt}$$

Si el paràmetre λ és constant, significa que el component no envelleix mai, és a dir, la v.a. T segueix una llei Exponencial que compleix una propietat anomenada **falta de memòria**. Aquesta propietat diu que si en el període [0, a] d'una distribució exponencial no ha mort un component, la probabilitat que ho faci en el període [a, a + t] és la mateixa que ho faci en el [0, t]. Dit d'una altra manera:

$$P(T < t + a \mid T > a) = P(T < t)$$

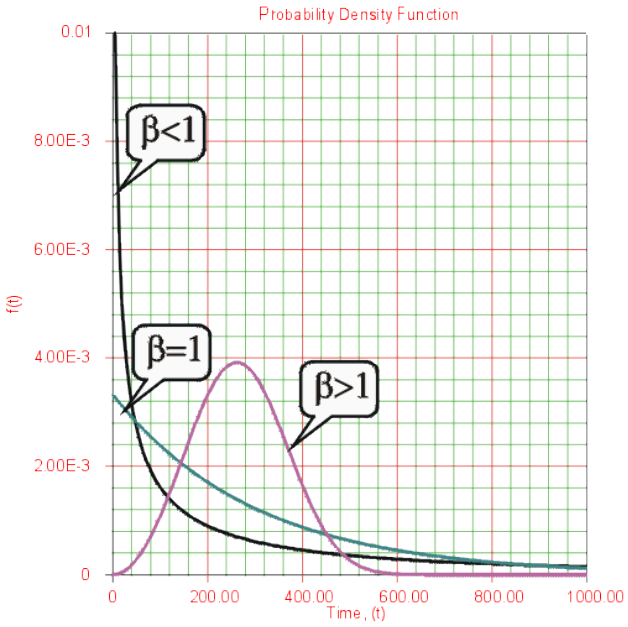
A més per a la distribució exponencial, la funció de densitat representa la probabilitat de supervivència d'una mostra multiplicada per λ per qualsevol instant de la vida d'un component.

Si el paràmetre λ no és constant i la funció λ(t) pren la forma següent:

$$\frac{\beta}{\alpha} \cdot \left(\frac{t}{\alpha}\right)^{\beta-1}$$

Si β és major a 1, la taxa de fallada és creixent, i per tant, el component envelleix. Si β és menor a 1, la taxa de falla és decreixent. Si β és igual a 1, la taxa de falla és constant, i pren el valor de $1/\alpha$.

Així doncs, la v.a. T segueix una **llei de Weibull**.



Funció de densitat:

$$f(t) = \frac{\beta}{\alpha^{\beta}} t^{\beta-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}}$$

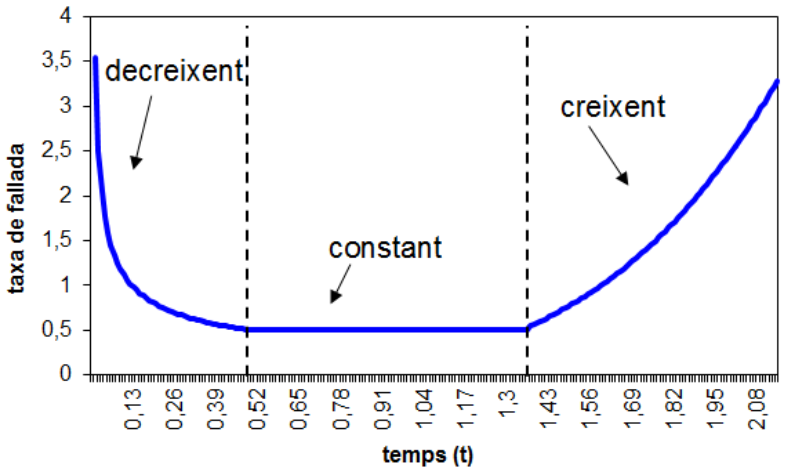
Funció de distribució:

$$F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\beta}}$$

A partir del paràmetre β podem descriure les etapes per les quals passa qualsevol component. La corba de la taxa de fallada λ(t) o també anomenada de banyera, descriu com evoluciona la taxa de fallada al llarg de la vida d'un component. Podem distingir tres etapes:

- **Etapa 1 o de mortalitat inicial**: la taxa de fallada és decreixent (β < 1).

- **Etapa 2 o de no envelliment**: la taxa de fallada és constant (β = 1).
- **Etapa 3 o de desgast**: la taxa de fallada és creixent (β > 1).



Normalment però a la vida quotidiana no tractarem amb un sol component, sinó que ens trobarem amb varis components independents relacionats entre ells. Ho poden estar en **sèrie** o en **paral·lel**.
Si tenim dos components independents **connectats en sèrie**, la probabilitat de supervivència del conjunt serà menor a les probabilitats individuals de cada component. Això és degut a que si en falla un o l'altre, el conjunt deixarà de funcionar. Veiem una combinació de components en sèrie:

Si considerem T_s la v.a. que descriu el temps de vida del sistema i T_1 i T_2 les v.a. que descriuen el temps de vida de cada component, podem escriure el següent:

$$R_s(t) = P(T_s > t) = P((T_1 > t) \cap (T_2 > t)) = P(T_1 > t) \cdot P(T_2 > t) = R_1(t) \cdot R_2(t)$$

$$F_s(t) = 1 - R_s(t) = 1 - R_1(t) \cdot R_2(t) = \big(1 - F_1(t)\big) \cdot \big(1 - F_2(t)\big)$$

Exemple: *Tenim dos components connectats en sèrie, els quals segueixen una distribució T_1 i T_2. A més sabem que $T_1 = T_2 = T_i \sim Exp(0.03)$. Interessa saber la distribució T_s que seguirà tot el sistema.*

Per a poder resoldre aquest problema interessa treballar amb la funció de supervivència $R_i(t)$. Per això cal tenir molt clar com passar d'una funció a una altra.

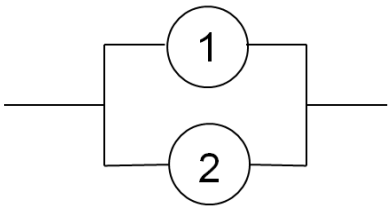
$$f_i(t) = 0.03 \cdot e^{-0.03t} \rightarrow \quad F_i(t) = 1 - e^{-0.03t} \rightarrow \quad R_i(t) = e^{-0.03t}$$

$$R_s(t) = R_1(t) \cdot R_2(t) = e^{-0.03t} \cdot e^{-0.03t} = e^{-0.06t}$$

$$R_s(t) = e^{-0.06t} \rightarrow \quad F_s(t) = 1 - e^{-0.06t} \rightarrow \quad f_s(t) = 0.06 \cdot e^{-0.06t}$$

Per tant podem dir que el sistema segueix una distribució $T_s \sim Exp(0.06)$.

Si tenim dos components independents **connectats en paral·lel**, la probabilitat de supervivència del conjunt serà major a les probabilitats individuals de cada component. Això és degut a que si en falla un o l'altre, el conjunt continua funcionant; únicament deixa de funcionar si els dos components fallen. Veiem una combinació de components en paral·lel:



Si considerem T_s la v.a. que descriu el temps de vida del sistema i T_1 i T_2 les v.a. que descriuen el temps de vida de cada component, podem escriure el següent:

$$F_s(t) = P(T_s \leq t) = P((T_1 \leq t) \cap (T_2 \leq t)) = P(T_1 \leq t) \cdot P(T_2 \leq t) = F_1(t) \cdot F_2(t)$$

$$R_s(t) = 1 - F_s(t) = 1 - F_1(t) \cdot F_2(t)$$

Exemple: *Tenim dos components connectats en paral·lel, els quals segueixen una distribució T_1 i T_2. A més sabem que $T_1 = T_2 = T_i \sim Exp(0.03)$. Interessa saber la distribució del temps de vida que seguirà tot el sistema.*

Per a poder resoldre aquest problema interessa treballar amb la funció de distribució $F_i(t)$. Per això cal tenir molt clar com passar d'una funció a una altra.

$$f_i(t) = 0.03 \cdot e^{-0.03t} \rightarrow \quad F_i(t) = 1 - e^{-0.03t}$$

$$F_s(t) = F_1(t) \cdot F_2(t) = (1 - e^{-0.03t}) \cdot (1 - e^{-0.03t}) = (1 - e^{-0.03t})^2$$

Per tant podem dir que el sistema segueix una distribució del temps de vida igual a $F_s(t) = (1 - e^{-0.03t})^2$.

PROBLEMES

1. Exercicis resolts

1.1. *El nivell de qualitat X d'un producte manufacturat és una v.a. que es distribueix segons una funció de densitat:*

$$f(x) = \begin{cases} 0 & si\ x \leq 0 \\ 2\lambda^{-2}x & si\ 0 < x < \lambda \\ 0 & si\ x \geq \lambda \end{cases}$$

essent λ una constant positiva que pot ser controlada en el procés de manufactura del producte. Es demana:

a) Calculeu la mitjana i la variància de X en funció de λ.
b) Cada ítem fabricat és inspeccionat abans de posar-lo a la venda. Els ítems que tenen un nivell de qualitat X major o igual que 8 són enviats a la venda mentre que la resta d'ítems són destruïts. Per cada ítem posat a la venda, el benefici és de 27 – λ milers de euros, mentre que cada ítem rebutjat suposa una pèrdua de 5 + λ milers de euros. Calculeu el valor de λ que fa màxim el benefici per ítem manufacturat, i calculeu el valor d'aquest benefici màxim.

a)

$$\mathbf{E\{X\}} = \int_0^{\lambda} x \cdot f(x)\, dx = \int_0^{\lambda} x^2 \cdot 2\lambda^{-2}\, dx = \left[\frac{x^3}{3} \cdot \frac{2}{\lambda^2}\right]_0^{\lambda} = \mathbf{\frac{2\lambda}{3}}$$

$$\text{var}\{X\} = \int_0^{\lambda} x^2 f(x)\, dx - E^2\{X\} = \int_0^{\lambda} x^3 \cdot 2\lambda^{-2}\, dx - \left(\frac{2\lambda}{3}\right)^2$$

$$\mathbf{var\{X\}} = \left[\frac{x^4}{4} \cdot \frac{2}{\lambda^2}\right]_0^{\lambda} - \frac{4\lambda^2}{9} = \mathbf{\frac{\lambda^2}{18}}$$

b) Calculem la funció de distribució per poder calcular el nombre d'ítems que es venen i es destrueixen:

$$F(x) = \int f(x)\, dx = \int 2\lambda^{-2}x\, dx = \frac{x^2}{\lambda^2}$$

$$\text{ítems destruïts} = P(\text{Qualitat} < 8) = F(8) = \frac{8^2}{\lambda^2} = \frac{64}{\lambda^2}$$

$$\text{ítems per vendre} = 1 - P(\text{Qualitat} < 8) = 1 - \frac{64}{\lambda^2}$$

Calculem el benefici de l'empresa.

$$\text{Guanys}(\lambda) \;=\; \text{ítems per vendre} \;\cdot\; \text{benefici per ítem venut} = \left(1 - \frac{64}{\lambda^2}\right) \cdot (27 - \lambda)$$

$$\text{Pèrdues}(\lambda) \;=\; \text{ítems destruïts} \;\cdot\; \text{pèrdua per ítem destruït} = \frac{64}{\lambda^2} \cdot (5 + \lambda)$$

$$\text{Benefici}(\lambda) = \text{Guanys}(\lambda) - \text{Pèrdues}(\lambda) = \frac{-2048}{\lambda^2} - \lambda + 27$$

Busquem els punts crítics de l'equació del benefici per trobar el valor de lambda que la fa màxim.

$$\text{Benefici}(\lambda)' = 0$$

$$\frac{4096}{\lambda^3} - 1 = 0$$

$$\boldsymbol{\lambda = 16}$$

Fem la segona derivada per determinar si és un màxim o un mínim.

$$\text{Benefici}(\lambda)'' = \frac{12288}{\lambda^4}$$

$$\frac{12288}{16^4} = -0.1875$$

Trobem que $\lambda = 16$ correspon a un màxim relatiu en l'interval $[0, \infty]$. Calculem el benefici màxim que pot obtenir l'empresa.

$$\textbf{Benefici}(\mathbf{16}) = \frac{-2048}{16^2} - 16 + 27 = \textbf{3 milers d'euros}$$

1.2. *La llargada de les peces fabricades per una determinada màquina s'ajusta a una distribució normal de mitjana 150 cm i desviació tipus 0.4 cm. Les peces es consideren acceptables si la seva llargada pertany a l'interval obert (149.2, 150.4).*

Es demana:

a) La proporció de peces defectuoses que contindrà la fabricació.
b) Trobeu un interval [150 – δ, 150 + δ] que contingui el 95% de la producció.
c) Se sap que una peça mesura més de 149 cm. Calculeu la probabilitat que mesuri menys de 150.1.
d) Quin ha de ser el valor de σ si es vol que la probabilitat de ser defectuosa sigui del 1% a la part baixa?

Definim la v.a. que segueix la longitud de les peces fabricades com L~N(150;0.4).

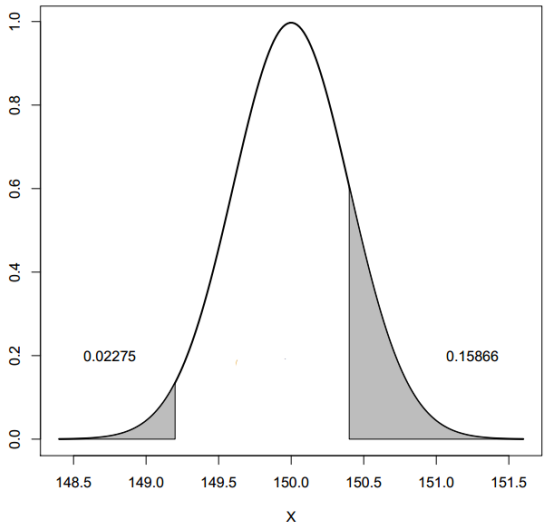
a)

$$\mathbf{P(Defectuosa)} = P(L \leq 149.2) + P(L \geq 150.4) =$$

$$= P\left(Z \leq \frac{149.2 - \mu}{\sigma}\right) + P\left(Z \geq \frac{150.4 - \mu}{\sigma}\right) =$$

$$= P\left(Z \leq \frac{149.2 - 150}{0.4}\right) + P\left(Z \geq \frac{150.4 - 150}{0.4}\right) = P(Z \leq -2) + P(Z \geq 1) =$$

$$= P(Z \leq -2) + \big(1 - P(Z < 1)\big) = 0.0228 + (1 - 0.8413) = \mathbf{0.1815}$$



b)

$$1 - 0.95 = 0.05 = P(L \leq 150 - \delta) + P(L \geq 150 + \delta)$$

$$0.025 = P(L \leq 150 - \delta)$$

Com que l'interval proposat és simètric, busquem a les taules estadístiques el valor que acumula una probabilitat de 0.025. Trobem el següent:

$$0.025 = P(Z \leq -1.96)$$

Busquem el valor no estandarditzat que equival al valor trobat a les taules.

$$-1.96 = \frac{(150 - \delta) - \mu}{\sigma} = \frac{(150 - \delta) - 150}{0.4}$$

$$\delta = 1.69 \cdot 0.4 = 0.784$$

Per tant, l'interval que inclourà el 95% de la producció serà [150 –0.784, 150 + 0.784] = **[149.216, 150.784]**.

c)

$$\mathbf{P(L < 150.1 \mid L > 149)} = \frac{P(149 < L < 150.1)}{P(L > 149)} =$$

$$= \frac{P(L < 150.1) - P(L < 149)}{1 - P(L < 149)} = \frac{P\left(Z < \frac{150.1 - \mu}{\sigma}\right) - P\left(Z < \frac{149 - \mu}{\sigma}\right)}{1 - P\left(Z < \frac{149 - \mu}{\sigma}\right)} =$$

$$= \frac{P\left(Z < \frac{150.1 - 150}{0.4}\right) - P\left(Z < \frac{149 - 150}{0.4}\right)}{1 - P\left(Z < \frac{149 - 150}{0.4}\right)} = \frac{P(Z < 0.25) - P(Z < -2.5)}{1 - P(Z < -2.5)} =$$

$$= \frac{0.5987 - 0.0062}{1 - 0.0062} = \mathbf{0.59619}$$

d)

$$P(L \leq 149.2) = 0.01$$

$$P\left(Z \leq \frac{149.2 - \mu}{\sigma}\right) = 0.01$$

Busquem a les taules estadístiques el valor que acumula un 0.01. Com que no és exacte hem de fer una interpolació lineal.

$$\frac{-2.32 - (-2.33)}{0.0102 - 0.0099} = \frac{-2.32 - z}{0.0102 - 0.01}$$

$$z = -2.3267$$

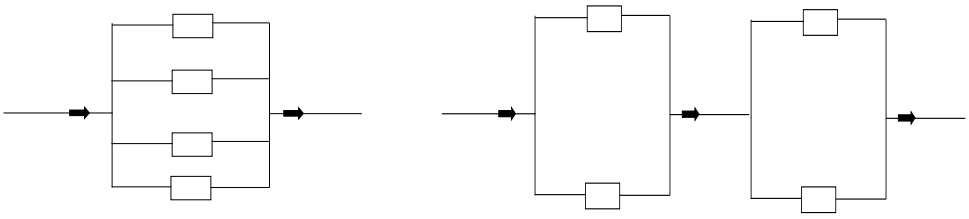
Busquem el valor no estandarditzat que equival al valor trobat a les taules.

$$-2.3267 = \frac{149.2 - \mu}{\sigma} = \frac{149.2 - 150}{\sigma}$$

$$\boldsymbol{\sigma = 0.3438}$$

1.3. *La vida T (en setmanes) d'un determinat tipus de component electrònic, ve donada per la funció de densitat: $f(t) = 3t^2 / 400^3$, $0 < t < 400$. Es demana:*

a) Calculeu la F(t), R(t), la mitjana i la desviació estàndard de la vida d'un d'aquests components.
b) Si per augmentar la vida mitjana connectem en paral·lel 4 d'aquests components electrònics (figura A), trobeu la f(t), R(t) i F(t) de la vida d'aquesta instal·lació o kit (aquesta només deixarà de funcionar si fallen els 4 components).
d) Calculeu el mateix que en l'apartat anterior pel cas de la figura B.



A **B**

a)

$$\mathbf{f(t)} = \frac{\mathbf{3 \cdot t^2}}{\mathbf{400^3}}$$

$$\mathbf{F(t)} = \int f(t)\,dx = \int \frac{3 \cdot t^2}{400^3}\,dt = \frac{\mathbf{t^3}}{\mathbf{400^3}}$$

$$\mathbf{R(t)} = 1 - F(t) = \mathbf{1} - \frac{\mathbf{t^3}}{\mathbf{400^3}}$$

$$\mathbf{E\{T\}} = \int_0^{400} t \cdot f(t)\,dt = \int_0^{400} \frac{3 \cdot t^3}{400^3}\,dt = \left[\frac{3 \cdot t^4}{4 \cdot 400^3}\right]_0^{400} = \mathbf{300}$$

$$var\{T\} = \int_0^{400} t^2 f(t)\,dx - E^2\{T\} = \int_0^{400} \frac{3 \cdot t^4}{400^3}\,dt - 300^2$$

$$\mathbf{var\{T\}} = \left[\frac{3 \cdot t^5}{5 \cdot 400^3}\right]_0^{400} - 300^2 = \mathbf{6000}$$

b)

$\mathbf{F_{KIT}(t)} = P(T_{KIT} \leq t) = P((T_1 \leq t) \cap (T_2 \leq t) \cap (T_3 \leq t) \cap (T_4 \leq t)) =$
$= P(T_1 \leq t) \cdot P(T_2 \leq t) \cdot P(T_3 \leq t) \cdot P(T_4 \leq t) = F_1(t) \cdot F_2(t) \cdot F_3(t) \cdot F_4(t) =$
$= F(t)^4 = \left(\frac{t^3}{400^3}\right)^4 = \frac{\mathbf{t^{12}}}{\mathbf{400^{12}}}$

$$\mathbf{f_{KIT}(t)} = F_{KIT}(t)' = \frac{\mathbf{12 \cdot t^{11}}}{\mathbf{400^{12}}}$$

$$\mathbf{R_{KIT}(t)} = 1 - F(t) = \mathbf{1} - \frac{\mathbf{t^{12}}}{\mathbf{400^{12}}}$$

d)

$$F_{PARAL\cdot LEL}(t) = P(T_{PARAL\cdot LEL} \leq t) = P((T_1 \leq t) \cap (T_2 \leq t)) =$$
$$= P(T_1 \leq t) \cdot P(T_2 \leq t) = F_1(t) \cdot F_2(t) =$$
$$= F(t)^2 = \left(\frac{t^3}{400^3}\right)^2 = \frac{t^6}{400^6}$$

$$R_{PARAL\cdot LEL}(t) = 1 - F_{PARAL\cdot LEL}(t) = 1 - \frac{t^6}{400^6}$$

$$\mathbf{R_{KIT}(t)} = R_{PARAL\cdot LEL1}(t) \cdot R_{PARAL\cdot LEL2}(t) = \left(1 - \frac{t^6}{400^6}\right)^2 = \mathbf{1 - \frac{2 \cdot t^6}{400^6} + \frac{2 \cdot t^{12}}{400^{12}}}$$

$$\mathbf{F_{KIT}(t)} = 1 - R(t) = \mathbf{\frac{2 \cdot t^6}{400^6} - \frac{2 \cdot t^{12}}{400^{12}}}$$

$$\mathbf{f_{KIT}(t)} = F_{KIT}(t)' = \mathbf{\frac{12 \cdot t^5}{400^6} - \frac{24 \cdot t^{11}}{400^{12}}}$$

1.4. *Un avió amb els dipòsits plens de carburant i sense passatgers ni tripulació pesa 120 Tm. L'avió té una capacitat de 100 places (inclosa la tripulació). El pes d'una persona segueix una v.a. N(70;10) kg. El pes de l'equipatge de cada persona és una v.a. N(μ_0;5) kg. Si el pes total de l'avió supera els 129890 kg, hi ha el perill de patir problemes per sobrecàrrega.*

Calculeu el valor màxim que pot tenir μ_0 si la probabilitat que l'avió sobrepassi el pes crític anterior (quan és ple) ha de ser menor o igual que 0.0002.

$$\text{Pes Total} = \text{Pes Avió} + \text{Pes Passatgers} + \text{Pes Equipatges} =$$

$$= 120000 + 100 \cdot N(70; 10) + 100 \cdot N(\mu_0; 5) =$$

El pes de l'avió el podem escriure com una v.a. N(120000;0). D'aquesta manera podem simplificar el càlcul.

$$\text{Pes Total} = N(120000; 0) + 100 \cdot N(70; 10) + 100 \cdot N(\mu_0; 5) \sim$$

$$\sim N(1 \cdot 120000 + 100 \cdot 70 + 100 \cdot \mu_0; \sqrt{1 \cdot 0^2 + 100 \cdot 10^2 + 100 \cdot 5^2} =$$

$$= N(127000 + 100 \cdot \mu_0; 111.8034)$$

Busquem a les taules estadístiques quin és el valor estandarditzat que deixa a la seva dreta el 0.0002. Trobem que z val 3.49, i a partir d'aquí trobem el valor de μ_0.

$$P(\text{Pes Total} > 129890) = P(Z > 3.49) = 0.0002$$

$$3.49 = \frac{129890 - (127000 + 100 \cdot \mu_0)}{111.8034}$$

$$\boldsymbol{\mu_0 = 25\ kg}$$

1.5. *La duració X de les bombones de butà de 40 kg es distribueix segons una llei N(200;20) hores. Es demana:*

a) Calculeu la probabilitat que una bombona duri més de 220 hores.
b) Quin és el temps de vida mínima que es pot garantir amb un risc d'equivocar-nos del 20%?
c) Si una bombona porta 160 h funcionant, quina és la probabilitat que duri més de 220 h?
d) Calculeu la probabilitat que el temps de vida total de 25 bombones sigui com a mínim de 5200 hores.

En el mercat existeix un altre tipus de bombones de 10 kg, el temps de vida Y de les quals es distribueix segons una llei Y~N(50;8) hores.

e) Quina duració total es pot garantir, amb un risc del 5%, si disposem de 4 bombones de 40 kg?
f) I si tenim 16 bombones de 10 kg?
g) Quina conclusió traieu d'aquests resultats?

a)

$$\mathbf{P(X > 220)} = 1 - P(X < 220) = 1 - P\left(Z < \frac{220 - 200}{20}\right) = 1 - P(Z < 1) =$$

$$= 1 - 0.8413 = \mathbf{0.1587}$$

b) Se'ns demana quin valor de la v.a. X no assoliran el 20% de les bombones de butà.

$$P(X \leq x) = 0.2$$

Busquem a les taules estadístiques el valor que acumula un 0.2. Com que no és exacte hem de fer una interpolació lineal.

$$\frac{-0.85 - (-0.84)}{0.1977 - 0.2005} = \frac{-0.85 - z}{0.1977 - 0.2}$$

$$z = -0.8418$$

Busquem el valor no estandarditzat que equival al valor trobat a les taules.

$$-0.8418 = \frac{x - \mu}{\sigma} = \frac{x - 200}{20}$$

$$\mathbf{x = 183.164\ hores}$$

c)

$$\mathbf{P(X > 220 \mid X > 160)} = \frac{P(X > 220 \cap X > 160)}{P(X > 160)} = \frac{P(X > 220)}{P(X > 160)} =$$

$$= \frac{1 - P(X < 220)}{1 - P(X < 160)} = \frac{1 - P(X < 220)}{1 - P(X < 160)} = \frac{1 - P(Z < 220)}{1 - P(Z < 160)} =$$

$$= \frac{1 - P\left(Z < \frac{220 - 200}{20}\right)}{1 - P\left(Z < \frac{160 - 200}{20}\right)} = \frac{1 - P(Z < 1)}{1 - P(Z < -2)} = \frac{1 - 0.8413}{1 - 0.0228} = \mathbf{0.1624}$$

d) Primer de tot calculem la normal que seguiran el grup de les 25 bombones.

$$X_{25} = 25 \cdot X \sim N(25 \cdot 200; \sqrt{25 \cdot 20^2} = N(5000; 100)$$

$$\mathbf{P(X_{25} \geq 5200)} = 1 - P(X_{25} < 5200) = 1 - P\left(Z < \frac{5200 - 5000}{100}\right) =$$

$$= 1 - P(Z < 2) = 1 - 0.9972 = \mathbf{0.0228}$$

e) Primer de tot calculem la normal que seguiran el grup de les 4 bombones.

$$X_4 = 4 \cdot X \sim N(4 \cdot 200; \sqrt{4 \cdot 20^2} = N(800; 40)$$

Busquem a les taules estadístiques el valor que acumula un 0.05. Com que no és exacte hem de fer una interpolació lineal.

$$\frac{-1.65 - (-1.64)}{0.0495 - 0.0505} = \frac{-1.65 - z}{0.0495 - 0.05}$$

$$z = -1.645$$

Busquem el valor no estandarditzat que equival al valor trobat a les taules.

$$-1.645 = \frac{x - \mu}{\sigma} = \frac{x - 800}{40}$$

$$\mathbf{x = 734.2\ hores}$$

f) Primer de tot calculem la normal que seguiran el grup de les 16 bombones.

$$Y_{16} = 16 \cdot Y \sim N(16 \cdot 50; \sqrt{16 \cdot 8^2} = N(800; 32)$$

Sabem que el valor estandarditzat z que acumula un 0.05 és –1.645. Busquem el valor no estandarditzat que equival al valor trobat a les taules.

$$-1.645 = \frac{x - \mu}{\sigma} = \frac{x - 800}{32}$$

$$\mathbf{x = 747.36 \text{ hores}}$$

g) Els resultats ens fan veure que disposem de més hores d'oxigen amb les 16 bombones de 10 kg. Tot i això, cal que l'usuari valori el fet de si manipular 16 bombones enlloc de 4 li compensa les aproximadament 13 hores de guany de funcionament, ja que també implica 12 canvis d'ampolla més.

1.6. *El temps de supervivència t (en hores) d'un cert component electrònic segueix una variable aleatòria amb funció de densitat mostrada. Es demana:*

$$f(t) = \begin{cases} 0.01 \cdot e^{-0.01 \cdot t} & si\ t > 0 \\ 0 & si\ t \leq 0 \end{cases}$$

a) Trobeu F(t) i R(t).
b) Quina és la fiabilitat del component a l'instant t = 25 hores?
c) Trobeu λ(t) i interpreteu els resultats.

a)

$$\mathbf{F(t)} = \int_0^t f(t)\, dt = \int_0^t 0.01 \cdot e^{-0.01 \cdot t}\, dt = \mathbf{1 - e^{-0.01 \cdot t}}$$

$$\mathbf{R(t)} = 1 - F(t) = \mathbf{e^{-0.01 \cdot t}}$$

b) Calculem la funció de fiabilitat per t = 25.

$$\mathbf{R(25)} = e^{-0.01 \cdot 25} = \mathbf{0.779}$$

c)

$$\boldsymbol{\lambda(t)} = \frac{f(t)}{R(t)} = \frac{0.01 \cdot e^{-0.01 \cdot t}}{e^{-0.01 \cdot t}} = \mathbf{0.01}$$

Com que λ(t) és constant, amb aquesta taxa, un de cada 100 components serà defectuós.

1.7. *Es considera que un disc té una "fallada inicial" si esdevé abans del temps t=α i una "fallada d'utilització" si esdevé després del temps t=β. Suposeu que la distribució del temps de fallada durant la vida d'un disc ve donada per la següent funció de densitat:*

$$f(t) = \frac{1}{\beta - \alpha} \quad \alpha \leq t \leq \beta$$

Es demana:

a) Trobeu F(t), R(t) i λ(t).
b) Dibuixeu la taxa instantània de falla del disquet per a α = 100 hores i β = 1500 hores.
c) Per a α = 100, β = 1500 quina és la fiabilitat d'un disc a l'instant t = 500 hores? Quina és la taxa instantània de falla?

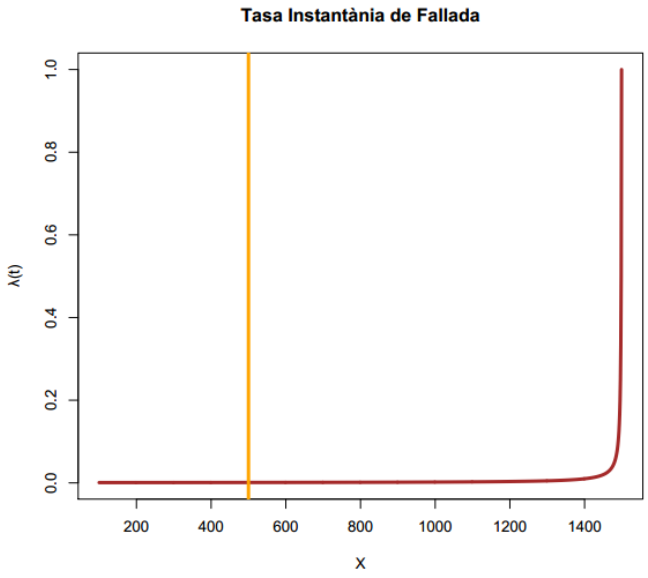
a)

$$\mathbf{F(t)} = \int_{\alpha}^{t} f(t)\, dt = \int_{\alpha}^{t} \frac{1}{\beta - \alpha}\, dt = \frac{t - \alpha}{\beta - \alpha}$$

$$\mathbf{R(t)} = 1 - F(t) = \frac{\boldsymbol{\beta - t}}{\boldsymbol{\beta - \alpha}}$$

$$\boldsymbol{\lambda(t)} = \frac{f(t)}{R(t)} = \frac{\frac{1}{\beta - \alpha}}{\frac{\beta - t}{\beta - \alpha}} = \frac{\mathbf{1}}{\boldsymbol{\beta - t}}$$

b) Veiem de color groc la taxa instantània de fallada per α = 100 hores i de color vermellós per β = 1500.



c) La fiabilitat d'un disc amb els paràmetres α i β donats a l'instant de 500 hores és:

$$\mathbf{R(500)} = \frac{1500 - 500}{1500 - 100} = \mathbf{0.714}$$

La taxa instantània de fallada d'un disc amb els paràmetres α i β donats a l'instant de 500 hores és:

$$\boldsymbol{\lambda(500)} = \frac{1}{1500 - 500} = \mathbf{0.001}$$

2. Exercicis proposats

2.1. *Sigui X una variable aleatòria contínua amb funció de densitat:*

$$f(x) = \begin{cases} \dfrac{1 + x^2}{12} & si\ x \in [0,3] \\ 0 & si\ x \notin [0,3] \end{cases}$$

Es demana:

a) Comproveu que f(x) és efectivament una funció de densitat.
b) Determineu la funció de distribució F.
c) Representeu gràficament sobre uns mateixos eixos les funcions f(x) i F(x).
d) Calculeu la probabilitat que X prengui un valor entre 1 i 2.
e) Calculeu la probabilitat que X prengui un valor menor que 1.
f) Si se sap que X > 1, trobeu la probabilitat que X sigui major que 2.
g) Calculeu l'esperança i la variància de X.

Solució: a) f(x) sempre és major a 0, l'àrea sota la corba és 1 i P(X=x) = 0;
b) $F(x) = \frac{1}{12}(x + \frac{x^3}{3})$; d) $P(1 \leq X \leq 2) = \frac{5}{8}$; e) $P(X \leq 1) = \frac{1}{9}$;
f) $P(X > 2 \mid X > 1) = \frac{11}{16}$; g) $E\{X\} = \frac{33}{16}$; $var\{X\} = \frac{699}{1280}$

2.2. *El temps de vida T d'un tub fluorescent ve donat per una llei exponencial. Es demana:*

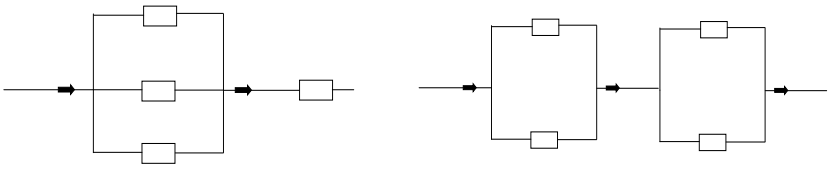
a) Si la vida mitjana dels tubs fos de 1500 hores, quina seria la probabilitat que la vida d'un tub fos major de 3000 hores?
b) Si el fabricant volgués assegurar que com a màxim l'1 per 1000 dels tubs tinguessin una vida inferior a les 5 hores, quina hauria de ser la vida mitjana mínima que haurien de tenir els tubs fluorescents?

Solució: a) $P(X > 3000) = e^{-2}$; b) E{X} = 49997.5

2.3. *La vida T (en setmanes) d'un determinat tipus de component electrònic, ve donada per la funció de densitat $f(t) = 3t^2 / a^3$, per $0 \le t \le a$.*

Es demana:

a) Un dispositiu A està format per un bloc de 3 components connectats en paral·lel, seguit d'un quart component connectat en sèrie amb el bloc anterior (figura esquerra). Perquè el dispositiu funcioni correctament, n'hi ha prou que funcioni un dels components del bloc juntament amb el quart component. Calculeu la probabilitat que el dispositiu continuï funcionant després de a/2 setmanes.
b) La mateixa pregunta en relació al dispositiu B (figura dreta). En aquest cas, el dispositiu funcionarà correctament sempre que funcionin cada un dels dos blocs connectats en sèrie.



A **B**

Solució: a) $P\left(\text{funciona després de } t = \frac{a}{2}\right) = 0.873$;
b) $P\left(\text{funciona després de } t = \frac{a}{2}\right) = 0.969$

2.4. *El temps de vida T d'un determinat component electrònic es pot considerar que s'ajusta a una llei exponencial. La vida mitjana d'aquests components és de 2000 dies.*

a) L'empresa que comercialitza aquests components vol donar als seus compradors una garantia d'un temps mínim de funcionament. Si com a màxim vol haver de reparar en període de garantia un 15% dels components venuts, quin és, com a màxim, el temps mínim de funcionament que l'empresa pot garantir?

Un determinat kit consta de la connexió en sèrie de dos d'aquests components electrònics. Per tal que el kit funcioni correctament, els dos components electrònics han de funcionar correctament. Es demana:

b) Caracteritzeu la v.a. T que proporciona el temps de vida del kit. Busqueu la seva funció de densitat i comproveu que es tracta d'una llei exponencial.*
c) Calculeu la probabilitat que el kit funcioni correctament més de 1000 dies.
d) Repetiu els dos apartats anteriors b) i c) en el supòsit que el kit té els dos components electrònics connectats en paral·lel i que, per tant, funciona correctament sempre que hi hagi com a mínim un component que funcioni correctament. (NOTA: En aquests cas, la v.a. T no segueix una llei exponencial).*

Solució: a) 325 dies; b) $F_{T^*}(t) = 1 - e^{-\frac{t}{1000}}$; $f_{T^*}(t) = \frac{1}{1000} \cdot e^{-\frac{t}{1000}}$;
c) 0.368; d) $F_{T^*}(t) = 1 - 2e^{-\frac{t}{2000}} + e^{\frac{t}{1000}}$; $f_{T^*}(t) = \frac{1}{1000} \cdot (e^{-\frac{t}{2000}} - e^{-\frac{t}{1000}})$; 0. 845

2.5. *En una cadena de producció, una màquina d'envasament automàtic omple els envasos amb un determinat producte. La quantitat introduïda en l'envàs es una v.a. X de mitjana 81.5 g i desviació tipus 8 g. El pes dels envasos buits es distribueix segons una v.a. de mitjana 14.5 g i desviació tipus 6 g. Ambdues distribucions són normals i independents. Es demana:*

a) La funció de densitat del pes dels envasos plens.
b) Si els paquets es distribueixen en caixes de cartró de 40 unitats el pes de les quals es distribueix segons una v.a. normal de mitjana 520 g i desviació tipus 50 g, calculeu la funció de densitat del pes de les caixes plenes.

Solució: a) N(96;100); b) N(4360;6500)

2.6. *El procés de fabricació d'un lot de resistències de tipus A és tal que la resistència R_A de les peces es distribueix segons una llei normal $N(\mu_A = 2000\Omega;\sigma_A)$. Un altre lot de resistències de tipus B és tal que la resistència R_B de les peces es distribueix segons una llei normal $N(\mu_B = 1000\Omega;\sigma_B)$. Se sap que $\sigma_A = 2\sigma_B$. En connectar en sèrie una resistència de tipus A amb una de tipus B escollides a l'atzar de cadascun dels lots, la probabilitat que la resistència R_{A+B} del conjunt estigui compresa entre 2999.5 Ω i 3000.5 Ω és igual a 0.9973. Es demana:*

a) Quina distribució segueix R_{A+B}?
b) Calculeu σ_A i σ_B.

Solució: a) $N(3000; 5\sigma_B)$; b) $\sigma_A = 0.1491$; $\sigma_B = 0.0745$

2.7. *Suposeu 3 components C1, C2 i C3 independents amb distribució del temps de vida idèntica. Amb aquests components construïm dos circuits diferents: A i B. El sistema A és el resultat de connectar C1 i C2 en sèrie i, a continuació, C3 en paral·lel. El sistema B és el resultat de connectar C1 i C2 en paral·lel i, a continuació, C3 en sèrie.*

a) Trobeu la fiabilitat del sistema per a cada un d'aquests esquemes.
b) Quin sistema té una fiabilitat major?

Solució: a) $1 - F(t) \cdot [1 - [1 - F(t)]^2]$; $[1 - F^2(t)] \cdot [1 - F(t)]$; b) Esquema A

2.8. *Suposeu un circuit electrònic té quatre transistors. El temps de vida de cada un dels transistors és independent i segueix una distribució exponencial de paràmetres $\lambda_1 = 0.00006$, $\lambda_2 = 0.00003$, $\lambda_3 = 0.00012$ i $\lambda_4 = 0.000018$ (fallades per minut). Es demana:*

a) Quina serà l'esperança de vida del circuit electrònic si els transistors estan organitzats en sèrie?
b) Quina és la fiabilitat del circuit electrònic si els transistors estan organitzats en paral·lel?

c) En el supòsit anterior, dibuixeu la gràfica de R(t) al llarg del temps.

Solució: a) 4386 minuts;
b) $R(t) = 1 - (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t})(1 - e^{-\lambda_3 t})(1 - e^{-\lambda_4 t})$

2.9. *Tres components estan organitzats en sèrie. Els seus respectius temps de vida són mútuament independents i segueixen una distribució exponencial de paràmetres $\lambda_1 = 0.003$, $\lambda_2 = 0.01$ i $\lambda_3 = 0.008$. Es demana:*

a) Trobeu la fiabilitat i la taxa instantània de falla del sistema.
b) Quina és l'esperança μ de vida del sistema?
c) Avalueu $R(\mu)$.

Solució: a) $R(t) = 1 - (1 - e^{-0.021 \cdot t})$; $\lambda = 0.021$; b) 47.6; c) 0.368

2.10. *Se suposa que el diàmetre extern d'un cert tipus de coixinets es troba, de manera aproximada, distribuït normalment amb $\mu = 3.5$ cm i $\sigma = 0.02$ cm. Si el diàmetre d'aquests coixinets no ha de ser menor de 3.47 cm ni major de 3.53 cm, quin és el percentatge de coixinets que, durant el procés de manufactura, s'han de descartar?*

Solució: 13.36%

2.11. *El temps de vida dels components elèctrics d'un circuit segueix una llei exponencial Exp(0.01), mesurat en hores. Es demana:*

a) Calculeu el valor esperat de funcionament, així com la desviació típica.
b) Calculeu la probabilitat que les components funcionin entre 2 i 3 hores.
c) Comproveu que $P[X \geq 5 \mid X \geq 2] = P[X \geq 3]$ (propietat de no memòria).

Solució: a) 100 hores; 100 hores; b) 0.0097;
c) $P(X \geq 5 \mid X \geq 2) = P(X \geq 3) = e^{-0.03}$

PRÀCTIQUES

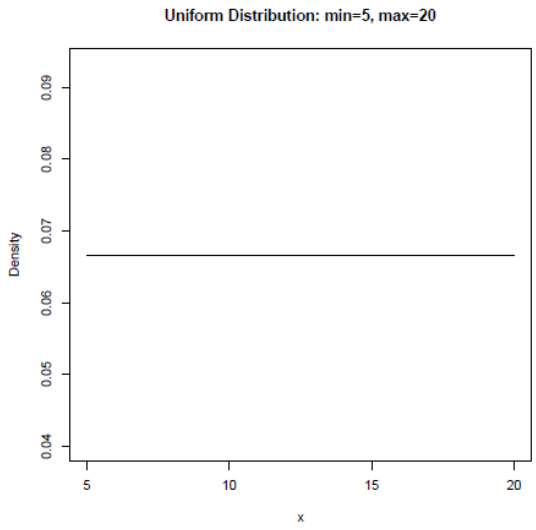
1. El model uniforme [a, b]

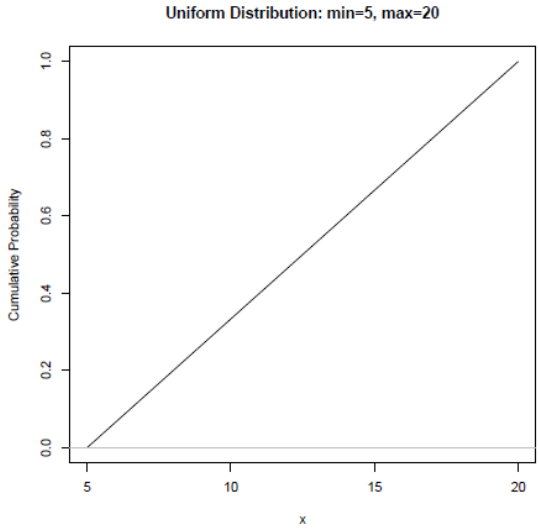
Un satèl·lit que ha finalitzat el seu cicle de vida en òrbita al voltat de la Terra està a punt de caure sobre la superfície del planeta. Els enginyers calculen que el mal funcionament d'unes plaques metàl·liques de l'aparell de posicionament provocarà que el satèl·lit aterri a la Terra en qualsevol punt situat a una distància entre 0 i 30 km del punt de retorn inicialment programat.

Per representar la funció de densitat f(x) i la de distribució F(x) d'un model uniforme, anem a:

`Distribucions, Distribucions contínues, Distribució uniforme, Traça una distribució uniforme ...`

- Representeu la funció de densitat f(x) i la de distribució F(x) d'un model uniforme U~[5,20].





Si es vol obtenir dues figures en una mateixa finestra gràfica, prèviament s'ha d'executar la instrucció `par(mfrow=c(1, 2))`. Una vegada representades les figures, per tornar al modus habitual s'ha d'executar l'ordre `par(mfrow=c(1, 1))`.

Utilitzeu les instruccions anteriors per a representar les funcions de la v.a. *Distància del punt d'aterratge al punt programat*.

- Quant val el valor constant que pren la funció de densitat? $1/30 = 0.033$.
- A la vista del gràfic de la funció de distribució, quant diríeu que val aproximadament la probabilitat que la distància sigui inferior a 15 km? $F(15) \leq 0.48$.

Per calcular probabilitats acumulades d'un valor d'una distribució uniforme anirem a:

`Distribucions, Distribucions contínues, Distribució uniforme, Probabilitats uniformes ...`

- Quina és la probabilitat que la distància sigui inferior a 10km?
  F(10) = 0.333.
- Quina és la probabilitat que la distància sigui superior a 10km?
  P(X ≥ 10) = = 0.667.
- Quina és la probabilitat que la distància sigui entre 10 i 20km?
  P(10 ≤ X ≤ 20) = F(20) − F(10) = 0.667 − 0.333 = 0.334.

Els enginyers estan interessats en tenir una idea de les distàncies d'aterratge que corresponen a les probabilitats principals. Per calcular-ho s'ha d'anar a:

`Distribucions, Distribucions contínues, Distribució uniforme, Quantils uniformes ...`

- Quins són els percentils de les probabilitats 0.05, 0.1, 0.25, 0.5, 0.75, 0.9 i 0.95 de la v.a. *distància del punt d'aterratge del satèl·lit al punt programat*?
  1.5, 3.0, 7.5, 15.0, 22.5, 27.0 i 28.5.

Si es vol obtenir la simulació d'una mostra de valors aleatoris d'una v.a. uniforme cal anar a:

`Distribucions, Distribucions contínues, Distribució uniforme, Mostra d'una distribució uniforme ...`

- Representeu i descriviu l'histograma i el diagrama de caixa d'un conjunt de dades format per 500 mostres distribuïdes uniformement des de 0 fins a 30.
  Primer de tot generem una distribució uniforme amb la instrucció anterior. Posem un nom qualsevol a la v.a., `Valor mínim` = 0 i `Valor màxim` = 30. A `Nombre de mostres`, escriurem 500 i a `Nombre d'observacions` escriurem 1, de forma que tindrem una columna amb 500 observacions. Finalment desactivarem l'opció `Mitjanes de la mostra`.

  Si es vol, es pot obtenir una representació conjunta dels dos gràfics; per això instal·lem el paquet StatDA de la següent manera, el carregarem i finalment l'executem:

```
install.packages("StatDA")
library(StatDA)
edaplot(SimUnif500$obs, S.cex=2).
```

  El paràmetre `S.cex` = 2 fa augmentar la mida dels punts representats en el gràfic.

Histograma amb amplitud de 0 a 30 i un perfil força horitzontal, és a dir, tots els intervals de classe tenen més o menys la mateixa freqüència. El diagrama de caixa és molt simètric i mostra que els quartils i la mediana divideixen la'mplitud en quatre parts molt similars, suggerint que les dades s'han repartit uniformement en l'interval [0, 30].

- Calculeu els estadístics mostrals més importants (mitjana, desviació i quartils) i compareu-los amb els valors teòrics. Valors mostrals: mitjana = 15.71944, desviació = 8.66375, Q_1 = 7.632644, mediana = 16.61679 i Q_3 = 23.40402. Valors teòrics: mitjana = 15, desviació = 8.66, Q_1 = 7.5, mediana = 15 i Q_3 = 22.5.

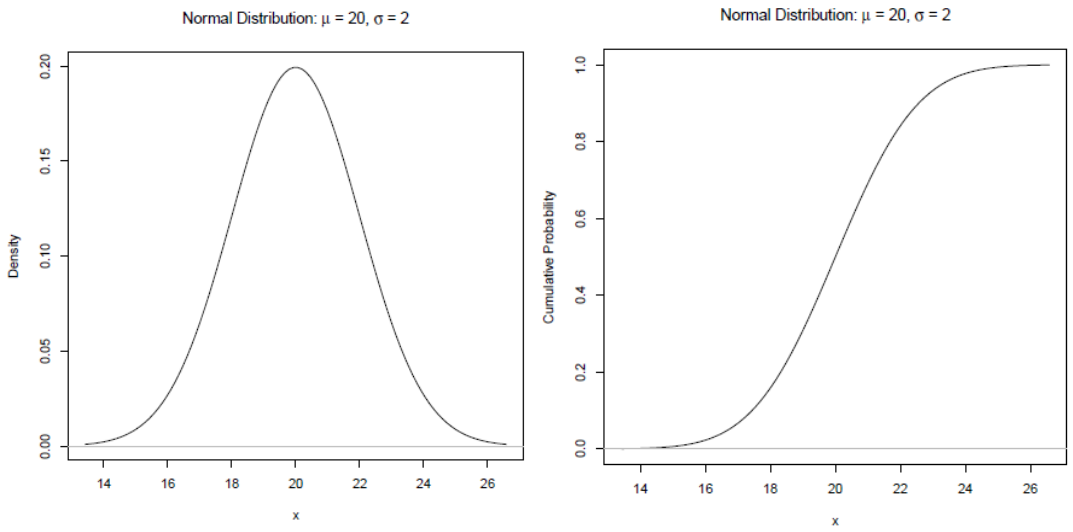
2. El model normal N(μ;σ)

El model normal N(μ;σ) també se'l coneix com el model dels errors aleatoris. Quan μ = 0, el model N(0;σ) s'anomena model del soroll blanc. Un model normal està caracteritzat per dos paràmetres, μ i σ.

Per representar un model normal en R anirem a:

`Distribucions, Distribucions contínues, Distribució normal, Traça d'una distribució normal ...`

- Representeu una la funció de densitat f(x) i la funció de distribució F(x) d'una v.a. X~N(20;2).



- Considereu una v.a. Z~N(0;1). A partir dels gràfics de les funcions de densitat i distribució, trobeu quin és l'amplitud de valors de la v.a. de Z? Aproximadament és de (–3.25, 3.25).
- A la vista del gràfic de la funció de distribució anterior, quant diríeu que val aproximadament la probabilitat P(Z ≤ –2)? Aproximadament F(–2) = 0.02.

Si es volen calcular probabilitats acumulades d'una distribució normal anirem a:

`Distribucions, Distribucions contínues, Distribució normal, Probabilitats normals ...`

- Quan val la probabilitat que P(Z ≤ 2)? Introduirem a `Valor de la Variable` un 2, a `Mitjana` un 0 i a `Desviació estàndard` un 1. Aquesta probabilitat val 0.9772499.

Observeu que a `Valor de la variable` podem posar diferents valors separats per espais.

Queda clar, a la vista de la finestra anterior, que per calcular probabilitats de la forma P (X > 3) caldrà marcar `Cua a la dreta`. Amb una mica d'observació podem calcular fàcilment probabilitats de la forma P(1 ≤ X ≤ 5) i similars.

- Quina és la probabilitat P(−1 ≤ Z ≤ 1)? 0.68.
- Quina és la probabilitat P(−1.96 ≤ Z ≤ 1.96)? 0.95.
- Quina és la probabilitat P(−2 ≤ Z ≤ 2)? 0.955.
- Quina és la probabilitat P(−3 ≤ Z ≤ 3)? 0.9975.

Per calcular quantils, anirem a:

`Distribucions, Distribucions contínues, Distribució normal, Quantils normals ...`

- Considereu una v.a. X~N(30;4). Calculeu la mediana i els quartils de la v.a. X. Q_1 = 27.30204, mediana = 30 i Q_3 = 32.69796.
- Calculeu els percentils 10, 44, 57 i 98. P_{10}% = 24.87379, P_{44}% = 29.39612, P_{57}% = 30.70550 i P_{98}% = 38.215.

Si es vol obtenir la simulació d'una mostra de valors aleatoris d'una v.a. normal cal anar a:

`Distribucions, Distribucions contínues, Distribució normal, Mostra d' una distribució normal ...`

- Representeu i descriviu l'histograma i el diagrama de caixa d'una distribució normal N(30;4) formada per 1000 mostres. El procediment emprat per crear la v.a. normal és molt similar al realitzat per crear la v.a. uniforme. L'histograma té una amplitud aproximada del 18 al 44, la qual cosa equival aproximadament a $\mu \pm 3 \cdot \sigma = 30 \pm 3 \cdot 4$. Té un perfil en forma de campana, centrat en el 30 i molt simètric. El diagrama de caixa està situat sobre el valor de mediana 30 i és molt simètric. Es detecten dades atípiques per sota de 20 i per sobre de 40.
- Quina és l'esperança de la N(30;4)? I la seva desviació estàndard? Compareu aquests dos estadístics teòrics amb els mostrals. Valors mostrals: mitjana = 29.75969 i desviació = 3.995514. Valors teòrics: mitjana = 30 i desviació = 4. Són molt similars.

3. El model Exponencial Exp(λ)

Amb el model exponencial podem realitzar els mateixos càlculs que en el cas del model normal (observeu que els menús són similars).

Considereu que el temps de vida d'un component electrònic segueix un model T∼Exp(λ = 0.1) i responeu el següent:

- Quin valor pren la funció de densitat per t = 0? f(0) = λ = 0.1.
- Calculeu la mediana i els quartils. Q_1 = 2.876821, mediana = 6.934172 i Q_3 = 13.862944.
- Calculeu els percentils 10, 44, 57 i 98. P_{10}% = 1.053605, P_{44}% = 5.798185, P_{57}% = 30.70550 i P_{98}% = 38.215.
- Quina és la mitjana mostra i la desviació estàndard d'una mostra de mida 500 generada per vosaltres? Mitjana mostral = 9.56126 i desviació estàndard mostral = 10.22465.
- Quina és l'esperança de la llei exponencial en estudi? I la seva desviació estàndard? Compareu els valors teòrics amb els mostrals. Esperança = 10 i desviació estàndard = 10. Els valors entre ells són semblants.
- Realitzeu l'histograma i el diagrama de caixa de les dades de la mostra i comenteu els resultats. L'amplitud de l'histograma és entre 0 i 60. El perfil és típic de la distribució exponencial. La màxima freqüència es dóna en el primer interval de classe i després sempre va decreixent de manera ràpida. La conseqüència és una asimetria molt forta a la dreta. Aquesta asimetria també es detecta en el diagrama de caixa, en el qual es veu un munt de dades atípiques en els valors més grans (cua a la dreta).

4. El model Weibull(α(scale), β(shape))

Considereu el model de distribució Weibull(α = 4(scale), β = 2(shape)) que segueix el temps de vida T (anys) d'un electrodomèstic. Responeu les següents qüestions:

- Comenteu el gràfic de la funció de densitat de T. Forma acampanada en una amplitud de 0 a 11, centrat aproximadament en el 3.5, amb asimetria a la dreta.
- Calculeu P(T ≤ 2), la P(T > 2.5) i la P(2 < T < 6). 0.2211992, 0.6766338 i 0.6734016.
- Calculeu la mediana i els quartils. Q_1 = 2.145440, mediana = 3.330218 i Q_3 = 4.709640.
- Calculeu els percentils 10, 44, 57 i 98. P_{10}% = 1.298371, P_{44}% = 3.045833, P_{57}% = 3.674714 i P_{98}% = 7.911534.
- Simuleu una distribució de Weibull amb els paràmetres donats i de mida 400. Calculeu els estadístics mostrals. Mitjana = 3.732112, desviació estàndard = 1.988478 i mediana = 3.623.
- Considereu la v.a. T_1∼Weibull(4, 0.5). Compareu el gràfic de la seva funció de densitat amb el de les variables T_2∼Weibull(4, 1) i T_3∼Weibull(4, 2). Per fer-ho executem `par(mfrow=c(1,3))` per a tenir les tres figures en una mateixa finestra. A mesura que el paràmetre forma augmenta, l'amplitud de

valors va disminuint i els valors de la variable temps són més petits. Per un valor de forma igual a 1 la corba és una exponencial de paràmetre $\lambda=1/a=0.25$.

5. Fiabilitat

5.1. Ajust per un model exponencial

Obriu l'arxiu de dades `tempsfallada.rda`. Aquest arxiu conté el temps de vida de 20 components electrònics del mateix tipus, és a dir, el temps que ha passat des que s'han posat en funcionament fins que han fallat.

- Feu un anàlisi numèric de la variable temps. mean = 104.582, sd = 118.6589, 0% = 8.33, 25% = 36.2275, 50% = 78.475, 75% = 98.3275, 100% = 519.26 i n = 20.
- Feu un histograma de la variable temps i comenteu-ne la forma. Quin model penseu que es pot ajustar a les dades? La freqüència s'acumula en els valors petits i després decreix ràpidament. Entre la forma de l'histograma i que la mitjana i la desv. est. són similars, sembla que un model exponencial podria ser apropiat.

Sembla que el model més adequat per aproximar el temps de vida dels components és l'exponencial. Aquest model té un paràmetre λ, i sabem que $\mu = 1/\lambda$. Per tant, l'estimació més adequada del paràmetre λ és $1/\bar{x}$.

- Quant val l'estimació de λ? $1/\bar{x} = 1/104.582 = 0.009561875$.

Una vegada estimat el valor de λ, podem aprofitar-ho per conèixer més sobre el temps de vida del producte: percentils, quantils, etc. Només cal revisar el que hem fet en els apartats anteriors d'aquesta pràctica.

- Calculeu la proporció de components que viuran més de 200 unitats de temps. Per fer-ho anem a:

  `Distribucions, Distribucions contínues, Distribució exponencial, Probabilitats exponencials ...`

  I escrivim 200, amb una raó 0.009561875 i activem la cua a la dreta.

  El resultat és de 0.1477291.

- Suposeu que el fabricant dona una garantia de 100 unitats de temps. Calculeu la proporció de components que poden gaudir de la garantia. Per calcular-ho anem a:

  `Distribucions, Distribucions contínues, Distribució exponencial, Probabilitats ...`

I obtenim 0.6156445, el que indica que gaudiran de la garantia un 61.56% dels components.

- Queda clar que donar una garantia de 100 unitats no és rendible per al fabricant. Calculeu quina hauria de ser la garantia perquè només el 15 % dels components en poguessin gaudir. Per calcular-ho anem :

  `Distribucions, Distribucions contínues, Distribució exponencial, Quantils ...`

  I obtenim t = 16.99655, el que podem aproximar a 17 unitats de temps.

- Calculeu la fiabilitat d'un component quant t=150. Calculem la probabilitat de la cua a la dreta de 150; P (T > 150) = 0.2382866.
- Calcula la taxa de fallada d'un component per a t=150. El model exponencial té una taxa de fallada constant λ = 0.009561875.

5.2. Ajust per un model Weibull

Obriu l'arxiu de dades `tempsvida.rda`. Aquest arxiu conté el temps de vida de 20 components electrònics del mateix tipus, és a dir, el temps que ha passat des que s'han posat en funcionament fins que han fallat.

- Feu un anàlisi numèric de la variable temps. mean = 163.9733, sd = 102.3691, 0% = 19.320, 25% = 74.1820, 50% = 149.834, 75% = 222.083, 100% = 397.636 i n = 20.
- Feu un histograma de la variable temps i comenteu-ne la forma. Quin model penseu que es pot ajustar a les dades? La freqüència s'acumula en els valors petits i després decreix ràpidament. Perfil en forma de campana en l'amplitud (0, 400), centrat en el 100 i amb asimetria a la dreta. No sembla un model exponencial. Si és un temps de vida aleshores podria tractar-se d'un model Weibull.

Sembla que ara no segueix un model exponencial. El model més adequat potser serà el Weibull. Aquest model té dos paràmetres: paràmetre d'escala α, i paràmetre de forma β. El paràmetre d'escala indica el percentil 0.632. A diferència del model exponencial, l'estimació dels paràmetres no és gaire senzilla i hem de recórrer a la funció `fitdistr()`.

Escriviu a la finestra d'instruccions:

```
fitdistr(temps.vida$temps, ''weibull'')
```

- Quant val l'estimació del paràmetre d'escala α? I l'estimació del paràmetre de forma β? shape = 1.6625858 i scale = 183.8511742.

Una vegada estimat els valors de α i β, podem aprofitar-ho per conèixer més sobre el temps de vida del producte: percentils, quantils, etc. Només cal revisar el que hem fet en els apartats anteriors d'aquesta pràctica.

- Calculeu la proporció d'elements que viuran més de 250 unitats de temps. 0.1888303.
- Suposeu que el fabricant dóna una garantia de 125 unitats de temps. Calculeu la proporció de components que poden gaudir de la garantia. 0.409349, el que significa el 40.93% dels components.
- Queda clar que donar una garantia igual a t=125 no és rendible per al fabricant. Calculeu quina hauria de ser la garantia perquè només el 10 % de les unitats en poguessin gaudir. 47.49346 unitats de temps.
- Calculeu la fiabilitat d'una unitat quan t=1751. 0.3980246.
- Calcula la taxa de fallada d'una unitat per a t=1752. $(1.6625858/183.8511742) \cdot (175/183.8511742)(1.6625858 - 1) = 0.008752246$.

TEMA 4: La qualitat en un procés de producció

TEORIA

1. Paràmetres poblacionals. Inferència estadística

La **inferència estadística** s'ocupa d'extreure conclusions sobre els paràmetres d'una v.a. X d'una població a partir de la informació que ens dóna una mostra representativa d'unitats d'aquesta població.

Si la distribució de probabilitat d'una v.a. X d'una població s'ajusta a una llei Normal, necessitem conèixer els paràmetres μ i σ de la població en relació a la variable. μ és la mitjana de la població i σ la desviació estàndard d'aquesta, però no podem conèixer exactament aquests dos valors, ja que la població està formada per una gran quantitat d'unitats (infinites o inabastables). Per aquesta raó, el que fem és prendre una mostra de n unitats d'aquesta població i estimar els valors d'aquests dos paràmetres.

S'ha de tenir en compte que la mostra escollida ha de ser representativa perquè les conclusions o inferències en relació a la població siguin encertades. D'això se'n encarrega la **teoria del mostreig**. Una manera d'obtenir una mostra representativa és fer un sorteig entre totes les unitats de la població i escollir un nombre n d'unitats a l'atzar, el que anomenem mostra aleatòria simple.

2. Estimadors dels paràmetres poblacionals

L'**estimador $\hat{\theta}$** ($\bar{x}$, s, ...) és l'estadístic calculat sobre la mostra de n unitats que ens serveix per tenir bona informació sobre el paràmetre θ (μ, σ, ...).

En la taula següent veiem diferents paràmetres i els seus respectius estimadors:

Paràmetre θ	**Estimador $\hat{\theta}$**
Esperança μ	Mitjana mostral $\bar{x}$
Variància σ^2	Variància mostral s^2
Proporció p	Proporció mostral p_n

La llei de probabilitat que segueixen els valors de $\hat{\theta}$ calculats sobre infinites mostres de mida n és el que s'anomena **distribució mostral de $\hat{\theta}$**.

Per altra banda, sabem que el valor de $\hat{\theta}$ varia segons la mostra escollida i el nombre d'unitats, el que s'anomena **variabilitat mostral**. Per aquesta raó és important escollir una mostra amb un nombre d'unitats adequat.

2.1. Mitjana mostral, $\bar{x}$

La mitjana mostral $\bar{x}$ serveix per estimar l'esperança d'una v.a. X de la població d'esperança μ.

Per calcular $\bar{x}$ el que fem és prendre una mostra aleatòria de n unitats, mesurar el valor de la v.a. X de cada una d'elles i aplicar la fórmula del càlcul de la mitjana per aquesta mostra. Com que realitzem el càlcul per n dades, expressem $\bar{x}$ com $\bar{x}_n$.

Si prenem la mitjana mostral com una v.a. $\overline{X}_n$, és a dir, si calculem la mitjana mostral per infinites mostres, totes de mida n, obtenim la **distribució de la v.a. $\overline{X}_n$**.

D'aquesta manera, l'esperança i la desviació estàndard de $\overline{X}_n$ prenen els següents valors:

$$E\{\overline{X}_n\} = \mu$$

$$var\{\overline{X}_n\} = \frac{\sigma^2}{n} \quad \rightarrow \quad desv\{\overline{X}_n\} = \frac{\sigma}{\sqrt{n}}$$

La $desv\{\overline{X}_n\}$ és el que s'anomena **error estàndard de la v.a. $\overline{X}_n$**, per tant, com més gran sigui n, menys variabilitat o error tindrà la v.a. $\overline{X}_n$. Això significa que les mitjanes calculades amb un nombre de mostres gran tindran més probabilitat d'acostar-se al valor real de μ que no pas les calculades amb un nombre de mostres petit.

Si la v.a. $X \sim N(\mu;\sigma)$, aleshores:

$$\overline{X}_n \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

En conseqüència, la v.a. $\overline{X}_n$ estandarditzada Z segueix la següent llei Normal estàndard:

$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$$

- **Teorema del límit central**

Encara que la v.a. X no s'ajusti a una llei Normal $N(\mu;\sigma)$, la v.a. $\overline{X}_n$ tendeix a aproximar-se a una llei Normal $N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$ a mesura que augmenta el nombre d'unitats n de les mostres. Aquesta aproximació és tant millor com més gran és la mida n de les mostres i com més s'aproxima la variable X a una llei Normal.

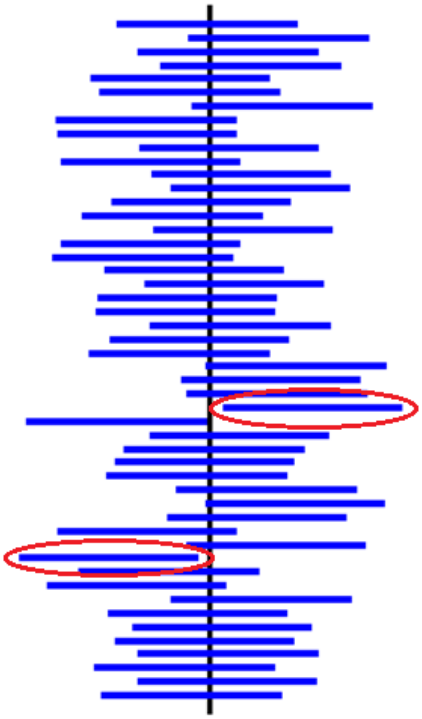
3. Interval de confiança de l'esperança μ

Si d'una v.a. X en desconeixem μ, podem estimar el valor de μ calculant $\bar{x}_n$, és a dir, la mitjana mostral per una mostra aleatòria de n unitats de la població. Tenint en compte que probablement $\bar{x}_n$ pren un valor semblant a μ, podem construir el següent interval al voltant de $\bar{x}_n$:

$$(\bar{x}_n - \varepsilon,\ \bar{x}_n + \varepsilon)$$

La probabilitat que μ es trobi dins d'aquest interval ha de ser alta, gairebé 1, diguem-n'hi 1 – α, de manera que α pot valdre 0.01, 0.05 o 0.10. Aquest interval s'anomena **interval de confiança de nivell 1 – α de l'esperança μ**, i el podem escriure com **IC(μ, 1 –α)**. Ens diu que μ es troba dins l'interval només amb un risc α d'equivocar-nos o el que és el mateix, que μ es troba dins l'interval amb una probabilitat 1 – α.

Si agafem 50 mostres diferents d'una mateixa població i en calculem l'IC amb una confiança del 95%, és a dir, α = 5%, podem afirmar que dels 50 intervals n'hi haurà 2 (≈ 5%) que no contindran el valor real de μ. Podem observar-ho en el següent dibuix, on la línia vertical de color negre representa el valor real de μ i cada una de les línies horitzontals un IC(95%):



La **precisió** d'un IC ve determinada per ε: com menor sigui aquest valor, més petit serà l'IC i per tant, més precisa serà l'estimació. Aquesta precisió és el radi de l'IC.

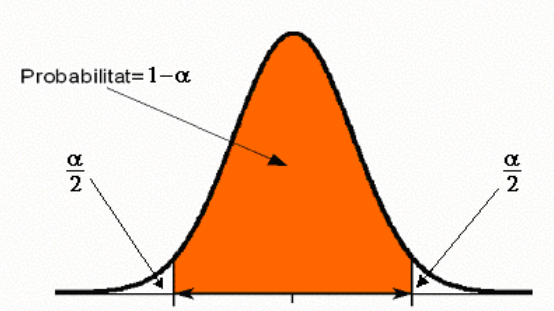
3.1. IC(μ, 1 – α) coneixent σ

Si el número n de mostres és prou gran:

$$\bar{X}_n \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right) \quad \rightarrow \quad Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$$

Per construir l'IC(μ, 1 – α) triem dos valors $-z_{\alpha/2}$ i $+z_{\alpha/2}$ de la llei Z~N(0;1), de manera que deixin entre ells una probabilitat (àrea) igual a 1 – α:

$$P\left(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}\right) = 1 - \alpha$$



D'aquesta manera, a partir de $\bar{x}_n$:

$$\mathbf{IC(\mu, 1-\alpha)} = \left(\bar{x}_n - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x}_n + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

on $-z_{\alpha/2}$ i $+z_{\alpha/2}$ s'han de buscar a les taules.

Com es pot veure, per a un nivell de confiança 1 – α prefixat, la precisió o radi de l'IC depèn de σ i de la mida n de la mostra. Si volem saber quina ha de ser la mida n que ha de tenir la mostra per tal que l'IC tingui una determinada ε prefixada:

$$\varepsilon = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{\varepsilon}\right)^2$$

Exemple: *A partir de la següent mostra aleatòria de 25 valors de la variable X, calculeu l'IC(μ, 95%) de μ sabent que μ és desconeguda i σ = 2.*

100.9	*100.0*	*98.1*	*97.8*	*96.4*
102.8	*103.3*	*100.2*	*98.7*	*101.9*
102.1	*99.5*	*99.5*	*101.0*	*102.4*

98.4	*99.2*	*99.6*	*102.7*	*99.6*
99.4	*102.7*	*98.8*	*99.2*	*102.2*

Calculem la mitjana de la mostra: $\bar{x}_{25} = 100.256$.

Un IC(μ,95%) significa que α = 5%, per tant, a partir de les taules de la llei Normal busquem els valors següents:

$$-z_{0.05/2} = -1.960$$
$$+z_{0.05/2} = +1.960$$

Aquests dos valors deixen entre ells una probabilitat o àrea del 95%.

D'aquesta manera l'IC(95%) de la μ és igual a:

$$IC(\mu, 95\%) = \left(100.256 - 1.960 \cdot \frac{2}{\sqrt{25}}, 100.256 + 1.960 \cdot \frac{2}{\sqrt{25}}\right)$$

$$IC(\mu, 95\%) = (99.472, 101.04\ )$$

3.2. IC(μ, 1 – α) desconeixent σ

En aquest cas, com que σ és desconeguda, s'ha d'estimar el seu valor a partir de s, és a dir, la desviació estàndard mostral. Per fer-ho hem de prendre una mostra aleatòria de n unitats i aplicar la fórmula del càlcul de la desviació estàndard mostral.

Ara, la v.a. $\overline{X}_n$ estandarditzada segueix una llei t-Student amb ν = n – 1 graus de llibertat:

$$t = \frac{\overline{X}_n - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Per construir l'IC(μ, 1 – α) es parteix del mateix concepte que en el cas anterior, en què σ és coneguda. Així, el nou IC(μ, 1 – α) es pot escriure com:

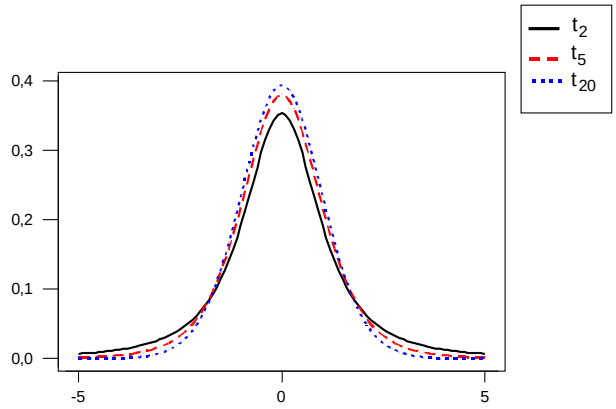
$$\mathbf{IC(\mu, 1-\alpha) = \left(\bar{x}_n - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}},\ \bar{x}_n + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)}$$

on $-t_{n-1,\alpha/2}$ i $+t_{n-1,\alpha/2}$ s'han de buscar a les taules.

3.3. Distribució t d'Student

La distribució t d'Student és una distribució de probabilitat contínua que partir d'un paràmetres anomenat els graus de llibertat, n, modela la probabilitat d'obtenir un valor a l'atzar d'una v.a. Es nota com $\mathbf{X \sim t_n}$, el que significa que la v.a. X segueix una distribució t d'Student amb n graus de llibertat.

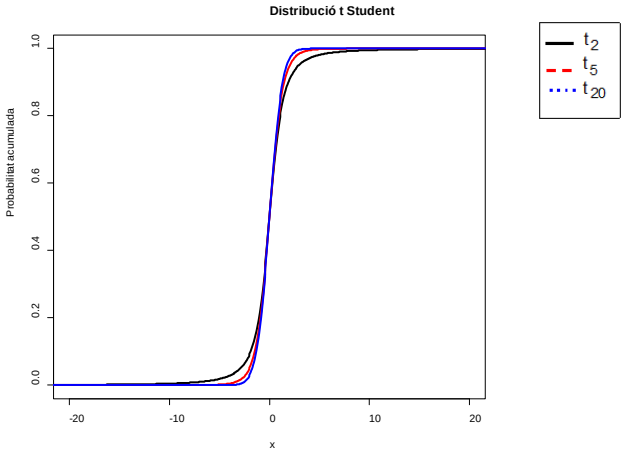
La **funció de densitat** d'aquesta distribució té la següent forma i es defineix per la fórmula que l'acompanya.



$$f(x) = k\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Aquesta equació és vàlida per a valors de menys infinit a més infinit i per graus de llibertat nombres naturals. És una funció simètrica respecte al 0 i unimodal.

La **funció de distribució** s'obté d'integrar la funció de densitat anterior. Degut a la dificultat de calcular aquesta funció, tenim unes taules estadístiques on tenim tabulat el valor que assoleix t per a diferents graus de llibertat.



$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

La funció de distribució també es pot representar a partir de gràfica de la funció de densitat, marcant l'àrea de -∞ al valor desitjat, el que representarà la probabilitat acumulada fins aquest valor.

Per tant, els **estadístics** d'una distribució t seran els següents:

$$E(X) = 0$$
$$var(X) = \frac{n}{n-2} \quad \text{on} \quad n > 2$$

Exemple: *A partir de la següent mostra aleatòria de 25 valors de la variable X, calcular l'IC(μ, 95%) de μ sabent que μ i σ és desconeguda.*

100.9	*100.0*	*98.1*	*97.8*	*96.4*
102.8	*103.3*	*100.2*	*98.7*	*101.9*
102.1	*99.5*	*99.5*	*101.0*	*102.4*
98.4	*99.2*	*99.6*	*102.7*	*99.6*
99.4	*102.7*	*98.8*	*99.2*	*102.2*

Calculem la mitjana de la mostra: $\bar{x}_{25} = 100.256$

Calculem la desviació estàndard de la mostra: $s = 1.845$

Com que n = 25, ν = n – 1 = 24 graus de llibertat.

Sabent que α = 5%, a partir de les taules de la llei t-Student busquem els següents valors, que deixen entre ells una probabilitat del 95%:

$$-t_{24,0.05/2} = -2.064$$
$$+t_{24,0.05/2} = +2.064$$

D'aquesta manera l'IC(μ, 95%) de la μ és igual a:

$$IC(\mu, 95\%) = \left(100.256 - 2.064 \cdot \frac{1.845}{\sqrt{25}}, 100.256 + 2.064 \cdot \frac{1.845}{\sqrt{25}}\right)$$

$$IC(\mu, 95\%) = (99.484, 101.02\ )$$

4. Control Estadístic de Processos, SPC

El **Control Estadístic de Processos** (en anglès, *Statistical Process Control*) és la minimització, a partir d'estudis estadístics de control, de la producció d'unitats defectuoses. Amb ell es volen aconseguir els següents objectius:

- Minimitzar la producció defectuosa.
- Mantenir una actitud de millora continua del procés.
- Comparar la producció respecte les especificacions.

Segons Shewhart, l'estratègia per millorar la qualitat d'un procés de producció es basa en la identificació de les causes que produeixen variabilitat i en una correcta assignació a una de les dues categories següents:

Causes comunes	Causes assignables o específiques
Són moltes i cadascuna produeix petites variabilitats.	Són poques i cadascuna produeix efectes importants.
Són la part permanent del procés. La seva superposició genera la variabilitat total del procés.	Són esporàdiques. Fàcils d'identificar mitjançant els gràfics de control.

Són difícils d'eliminar, però reduir-les millora la qualitat del producte.	Són fàcils d'eliminar pels operaris i/o tècnics.
Afecten al conjunt de màquines, operaris, etc.	Afecten específicament a una màquina, operari, etc.
La variabilitat que produeixen admet modelització estadística.	La variabilitat que produeixen no admet modelització estadística.

4.1. Definició dels gràfics de control

Un procés en estat de control és aquell que només està afectat per causes comunes de variabilitat. Per tant, podem dir que l'objectiu dels gràfics de control és monitoritzar l'evolució d'un procés de producció per tal de controlar gràficament que la seva variabilitat (en relació a una determinada característica de qualitat) sigui deguda només a causes comunes i no a causes assignables.

- **Elements d'un gràfic de control**

Els elements fonamentals d'un gràfic de control són:

- La **v.a.** de qualitat **X** a controlar, la qual ha de seguir una distribució de probabilitat, ja sigui una distribució normal, exponencial, etc.
- L'**estadístic mostral $\hat{\theta}$** ($\bar{X}$, R, S) calculat sobre mostres de mida n procedents de la v.a. X. Així, $\hat{\theta}$ és una v.a. que té una $E(\hat{\theta})$ i una $var(\hat{\theta})$.

Realitzar un gràfic de control significa representar, en forma de diagrama temporal, els valors $\hat{\theta}_1$, $\hat{\theta}_2$, ..., $\hat{\theta}_m$ que adquireix l'estadístic $\hat{\theta}$ sobre les successives mostres M_1, M_2, ..., M_m procedents del procés de producció:

	$\mathbf{M_1}$	$\mathbf{M_2}$	⋯	$\mathbf{M_m}$
Dades procedents de la v.a. X	x_{11}	x_{21}	⋯	x_{m1}
	⋮	⋮	⋮	⋮
	x_{1n}	x_{2n}	⋯	x_{mn}
Estadístic mostral $\hat{\theta}$	$\bar{x}_1$ R_1 s_1	$\bar{x}_2$ R_2 s_2	⋯	$\bar{x}_m$ R_m s_m

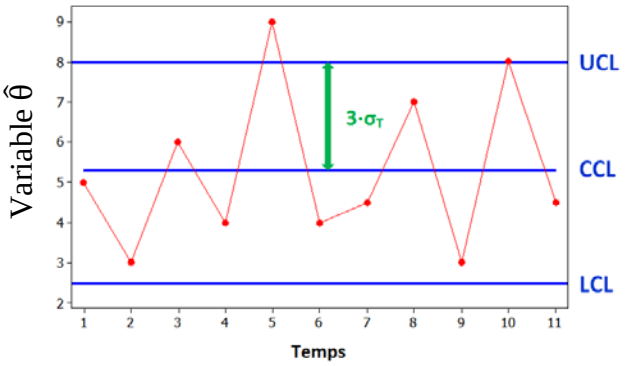
El nombre de mostres es designa amb la lletra m i la mida comuna de les mostres es designa amb la lletra n. Habitualment, les mostres M_i es prenen de la mateixa mida n, tot i que també es poden prendre mostres de mides diferents.

- **Línies de control**

Els gràfics de control estan formats per les tres línies de control següents:

	Històric	No històric
Línia central de control (CCL)	$\mu_{\hat{\theta}}$	$\overline{\mu_{\hat{\theta}}}$
Línia superior de control (UCL)	$\mu_{\hat{\theta}} + k_u$	$\overline{\mu_{\hat{\theta}}} + \overline{k_u}$
Línia inferior de control (LCL)	$\mu_{\hat{\theta}} - k_l$	$\overline{\mu_{\hat{\theta}}} - \overline{k_l}$

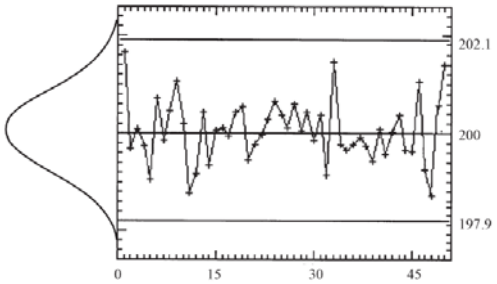
Si la informació de què disposem **no és històrica**, és a dir no es coneix la mitjana i la desviació del procés, es fan servir $\overline{\mu_{\hat{\theta}}}$, $\overline{k_u}$, i $\overline{k_l}$, que són estimacions calculades en base a les dades de les mostres.
Les línies UCL i LCL es calculen de manera que, quan el procés funciona correctament en relació a la v.a. X, la probabilitat que l'estadístic $\hat{\theta}$ estigui comprès entre elles sigui igual a un determinat valor 1 – α prefixat. Habitualment, 1 – α = 0.9975, que equival a 6σ en una distribució N(0;1). Per tant, tal com ens mostra el gràfic, a 3 desviacions per sobre de la línia CCL trobaríem la línia UCL i a 3 desviacions per sota de la línia CCL trobaríem la línia LCL:



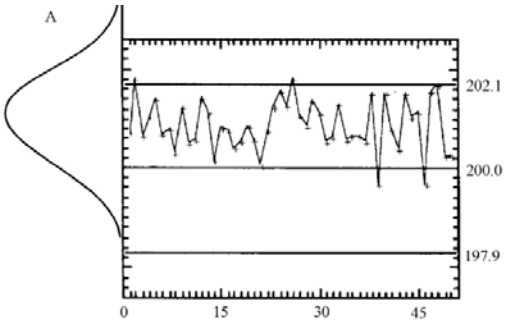
4.2. Gràfics de control per a variables

Un procés en estat de control pot convertir-se en un procés fora de control per canvis en la seva mitjana, en la seva variabilitat o per ambdós estadístics de cop:

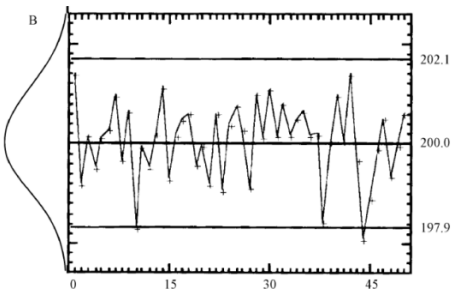
Procés en estat de control



Procés fora de control (mitjana fora de control)



Procés fora de control (variabilitat fora de control)



Per tant, per poder conèixer l'estat en què es troba una v.a. X calen gràfics de control per els dos paràmetres, és a dir, per a la mitjana (**gràfic $\overline{X}$**) i per a la variabilitat (**gràfic R** o **gràfic S**).

Els gràfics de control que s'estudiaran a continuació són vàlids per a **característiques contínues X** del producte que presenten una **distribució normal** (amb esperança μ i desviació estàndard σ). Per posar un exemple, les v.a. X que es poden tractar són el contingut en cm^3 d'un líquid, el pes d'un sac de pinso, la viscositat d'una resina, etc.

- **Gràfic $\overline{X}$**

Per controlar la **mitjana** del procés s'utilitza l'estadístic mostral $\overline{X}$. Es tracta doncs de representar, en forma de diagrama temporal, els valors $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_m$ que adquireix l'estadístic $\overline{X}$ sobre les successives mostres.

La v.a. X~N(μ;σ) per tant, en el cas que partim d'informació **històrica**:

$$E\{\overline{X}\} = \mu$$

$$\text{desv}\{\overline{X}\} = \frac{\sigma}{\sqrt{n}}$$

Per altra banda, en el cas que partim d'informació **no històrica**, s'ha de tenir en compte que els valors μ i σ s'han d'estimar.

Per estimar μ ho fem a partir de la mitjana de la mitjana de les mostres:

$$\bar{\bar{x}} = \frac{\sum \bar{x}_i}{m}$$

Per estimar σ tenim dues vies:

- A partir de l'amplitud $\mathbf{R_i}$ de les mostres: tenint en compte que la v.a. $W = R/\sigma$ compleix que $E(W) = d_2$, aleshores $E(W) = d_2 = E(R/\sigma)$ i

$$\hat{\sigma} = \frac{\bar{R}}{d_2}, \text{ on } \bar{R} = \frac{\sum R_i}{m}$$

- A partir de $\mathbf{s_i}$ de les mostres: Tenint en compte que la v.a. s compleix que $E\{s\} = c_4\sigma$.

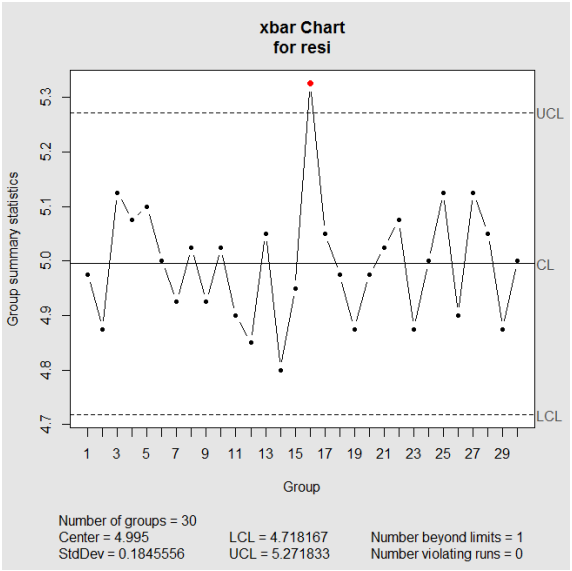
$$\hat{\sigma} = \frac{\bar{s}}{c_4}, \text{ on } \bar{s} = \frac{\sum s_i}{m}$$

D'aquesta manera, les línies de control CCL, LCL i UCL valdran:

	Històric	**No històric**	
CCL	μ	$\bar{\bar{x}}$	
LCL	$\mu - A\sigma$ $\left(\text{on } A = \frac{3}{\sqrt{n}}\right)$	**A partir de R_i** $\bar{\bar{x}} - A_2\bar{R}$ $\left(\text{on } A_2 = \frac{3}{d_2\sqrt{n}}\right)$	**A partir de s_i** $\bar{\bar{x}} - A_3\bar{s}$ $\left(\text{on } A_3 = \frac{3}{c_4\sqrt{n}}\right)$
UCL	$\mu + A\sigma$ $\left(\text{on } A = \frac{3}{\sqrt{n}}\right)$	**A partir de R_i** $\bar{\bar{x}} + A_2\bar{R}$ $\left(\text{on } A_2 = \frac{3}{d_2\sqrt{n}}\right)$	**A partir de s_i** $\bar{\bar{x}} + A_3\bar{s}$ $\left(\text{on } A_3 = \frac{3}{c_4\sqrt{n}}\right)$

on A, A_2, A_3, d_2 i c_4 es poden trobar a les taules estadístiques dels gràfics de control.

Exemple: Gràfic $\bar{X}$.



Representació d'un gràfic de control de la $\bar{X}$. Tenim n=30 mostres amb $\bar{\bar{x}} = 4.995$. Una mostra (núm. 16) es troba fora de la zona de control. El procés està fora de control de manera puntual, segurament esporàdica.

- **Gràfic R**

Per controlar la **variabilitat** del procés una de les opcions és utilitzar l'estadístic mostral R que ens calcula l'amplitud. Es tracta doncs de representar, en forma de diagrama temporal, els valors R_1, R_2, ..., R_m que adquireix l'estadístic R sobre les successives mostres.

En el cas que partim d'informació **històrica**, sabent que la v.a. $W = R/\sigma$ compleix que $E\{W\} = d_2$. Aleshores $R = \sigma W$ i $E(R) = \sigma$ i , per tant, $E(W) = d_2\sigma$

Com que $desv\{W\} = d_3$, $desv\{R\} = desv\{W\} * \sigma = d_3\sigma$

En el cas que partim d'informació **no històrica**, s'ha d'estimar σ com hem fet anteriorment en el cas del gràfic $\overline{X}$. Per tant:

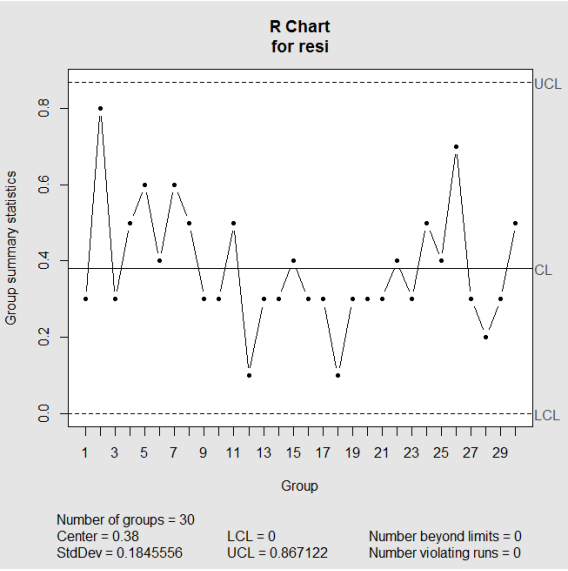
$$\hat{\sigma} = \frac{\overline{R}}{d_2}, \quad \text{on} \quad \overline{R} = \frac{\sum R_i}{m}$$

D'aquesta manera, les línies de control CCL, LCL i UCL valdran:

	Històric	**No històric**
CCL	$d_2\sigma$	$\overline{R}$
LCL	$D_1\sigma$ $(\text{on } D_1 = d_2 - 3d_3)$	$D_3\overline{R}$ $\left(\text{on } D_3 = 1 - \frac{3d_3}{d_2}\right)$
UCL	$D_2\sigma$ $(\text{on } D_2 = d_2 + 3d_3)$	$D_4\overline{R}$ $\left(\text{on } D_4 = 1 + \frac{3d_3}{d_2}\right)$

on D_1, D_2, D_3, D_4, d_2 i d_3 es poden trobar a les taules dels gràfics de control.

Exemple: Gràfic R.



Procés sota control. Observem que els límits no són simètrics respecte la línia central CL=0.38, ja que no tenen sentit les amplituds R negatives.

- **Gràfic S**

Per controlar la **variabilitat** del procés l'altra opció és utilitzar l'estadístic mostral S. Es tracta doncs de representar, en forma de diagrama temporal, els valors s_1, s_2,..., s_m que adquireix l'estadístic S sobre les successives mostres.

En el cas que partim d'informació **històrica**, s'ha de saber que la v.a. s compleix que E{s} = $c_4\sigma$ (com hem dit anteriorment) i que desv{s} = $\sigma\sqrt{1-c_4^2}$.

En el cas que partim d'informació **no històrica**, s'ha d'estimar σ com hem fet anteriorment en el cas del gràfic $\overline{X}$. Per tant:

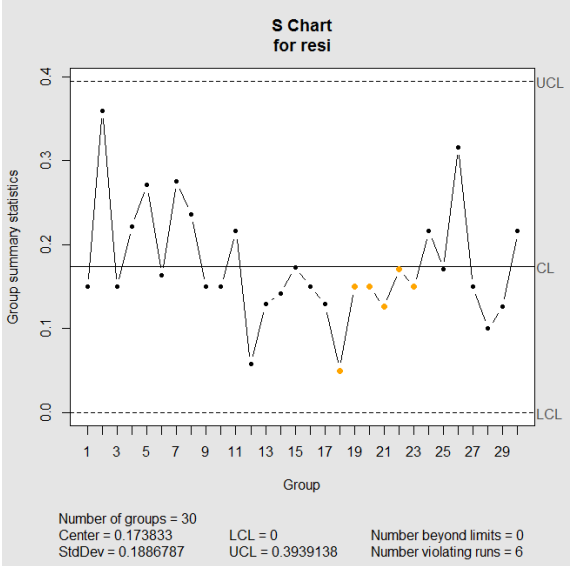
$$\hat{\sigma} = \frac{\bar{s}}{c_4}$$

D'aquesta manera, les línies de control CCL, LCL i UCL valdran:

	Històric	**No històric**
CCL	$c_4\sigma$	$\bar{s}$
LCL	$B_5\sigma$ $\left(\text{on } B_5 = c_4 - 3\sqrt{1-c_4^2}\right)$	$B_3\bar{s}$ $\left(\text{on } B_3 = 1 - \frac{3}{c_4}\sqrt{1-c_4^2}\right)$
UCL	$B_6\sigma$ $\left(\text{on } B_6 = c_4 + 3\sqrt{1-c_4^2}\right)$	$B_4\bar{s}$ $\left(\text{on } B_4 = 1 + \frac{3}{c_4}\sqrt{1-c_4^2}\right)$

on B_3, B_4, B_5, B_6 i c_4 es poden trobar a les taules dels gràfics de control.

Exemple: Gràfic S.

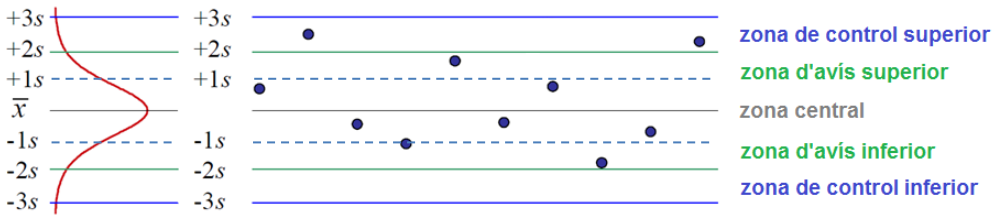


Cap observació es troba fora de la zona de control.
Tanmateix hi ha massa observacions consecutives per sota la línia de control.
Ens informa que hi ha hagut una zona temporal on hi ha hagut reducció de la variabilitat.

4.3. Interpretacions dels gràfics de control

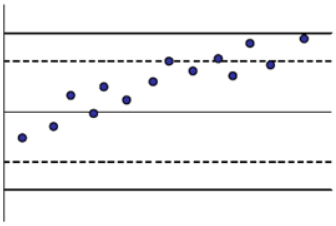
Podem dividir els gràfics de control en tres zones:

- **Zona central**: de –1s a +1s.
- **Zona d'avís**: de –1s a –2s i de +1s a +2s.
- **Zona de control**: de –2s a –3s i de +2s a +3s.

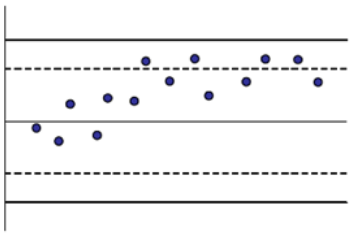


Així, podem dir que un procés està fora de control si en el gràfic existeixen:

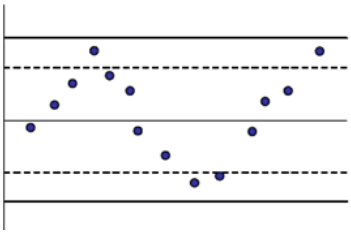
- 1 punt més enllà de la zona de control (la probabilitat que sigui degut a l'atzar és < 0.3%).
- 2 de 3 punts consecutius en la zona de control (la probabilitat que sigui degut a l'atzar és < 0.0625%).
- 6 punts consecutius en línia ascendent o descendent: deriva.



- 9 punts consecutius a un costat de la línia central, ja sigui per sobre o per sota: desplaçament del valor central.



- 14 punts consecutius alternant a dalt o a baix: fenomen cíclic o sèries temporals.



- 15 punts consecutius en la zona central: millora de la precisió i reducció de la variabilitat associada → tornar a calcular els límits UCL i LCL.
- 4 de 5 punts consecutius en la zona d'avís o més enllà.
- 8 punts consecutius per sobre i per sota de la zona central: 2 poblacions diferents.

4.4. Capacitat d'un procés

La capacitat d'un procés és l'aptitud per a generar un producte que verifiqui certes especificacions respecte una certa característica de qualitat X, la qual considerem que s'ha de distribuir segons una llei Normal (μ,σ).

Per calcular la capacitat d'un procés parlarem dels següents elements:

- **Límits de tolerància o especificacions:**
  Límit inferior (LSL, *Lower Specific Limit*)
  Límit superior (USL, *Upper Specific Limit*)
- **Valor nominal** (VN) o **objectiu** (*target*)**:** valor que es troba a la mateixa distància de LSL que de USL, centre de l'interval de tolerància
- **Tolerància:** requeriments establerts; criteri habitual 6σ.

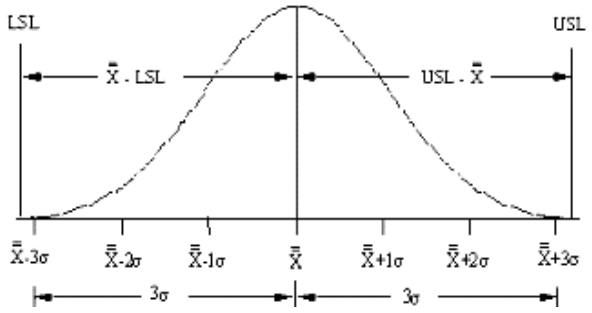
Per mesurar quina és la capacitat d'un procés s'utilitzen els següents **índexs de capacitat**:

Índex de capacitat	Valor
Índex Potencial	$C_p = \frac{USL - LSL}{6\sigma}$
Índex Lateral (superior)	$C_{pu} = \frac{USL - \mu}{3\sigma}$
Índex Lateral (inferior)	$C_{pl} = \frac{\mu - LSL}{3\sigma}$
Índex Real	$C_{pk} = \min\{C_{pu}; C_{pl}\}$

Si no es disposa d'informació històrica de μ es reemplaça per $\bar{\bar{x}}$. El valor de σ, es reemplaça el seu valor per una estimació, com pot ser, per exemple, l'estimació:

$$\hat{\sigma} = \frac{s_p}{c_4} \text{ on } s_p = \sqrt{\frac{s_1^2 + s_2^2 + \ldots + s_m^2}{m}}.$$

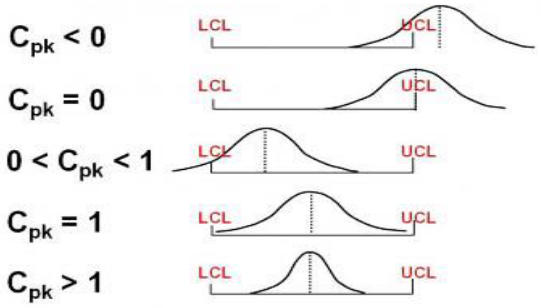
Quan els límits de tolerància LSL i USL es troben separats 6σ, com marca la tolerància, i el VN coincideix amb $\bar{\bar{x}}$, podem dir que $C_p = C_{pk} \approx 1$:



- **Interpretació de l'índex C_{pk}**

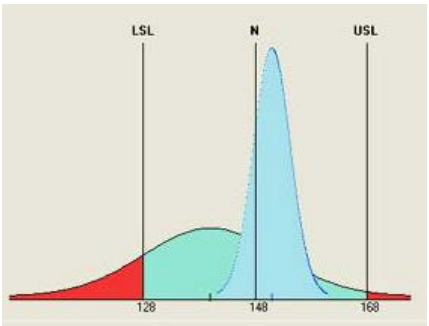
Es considera que un índex de capacitat adequat és aquell que es troba entre 1 i 1.33, tot i que es recomana que sigui superior a 1.33.

Segons el valor de l'índex de capacitat trobem els següents tipus de distribucions:



Les actuacions recomanades per a millorar la capacitat són:

- Centrar la distribució.
- Reduir la variabilitat de la distribució.



En el cas de la distribució normal pintada de color blau de la figura, una millora de la capacitat seria centrar la distribució al VN. En canvi, en el cas de la distribució normal pintada de color verd (amb cues vermelles), per millorar la capacitat, a més de centrar la distribució al VN s'ha de reduir la seva variabilitat.

PROBLEMES

1. Exercicis resolts

1.1. *Els errors aleatoris que es produeixen en les pesades que es realitzen amb una determinada balança es distribueixen segons una llei Normal de mitjana 0 i desviació tipus 0.5 decigrams.*

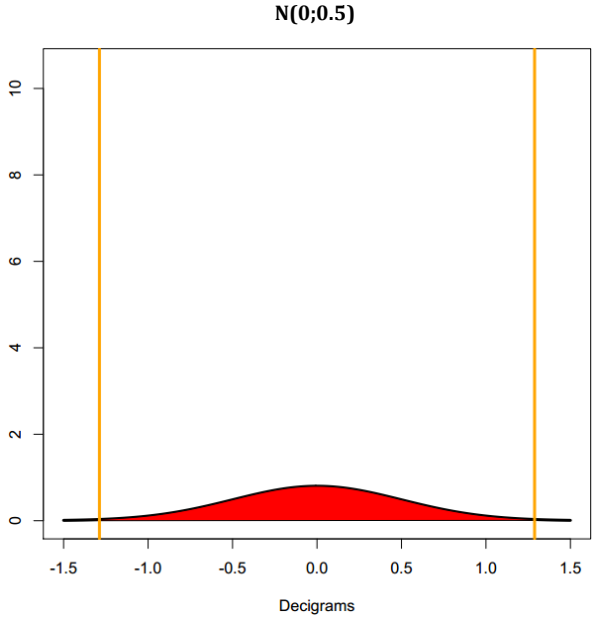
Es demana:

a) Calculeu l'error màxim, per defecte i per excés, que es pot produir en una pesada, amb probabilitat 0.99.
b) Si es fan 10 pesades d'un mateix objecte i es pren com a pes la mitjana de les 10 pesades, calculeu l'error màxim d'aquest pes final, amb probabilitat 0.99.
c) Calculeu el nombre mínim n de pesades que cal realitzar d'un mateix objecte per tal que, si es pren com a pes la mitjana de les n pesades, l'error màxim d'aquest pes final sigui inferior a 0.1 decigram amb probabilitat 0.99.

a) Ens demanen que busquem l'error màxim (per defecte i per excés) que es pot arribar a cometre en una pesada amb una confiança del 99%. Podem dir que aquesta probabilitat es tracta d'un IC(1 – α), per tant, que α = 1%.

Diem que la v.a. X = "Pes que marca la balança" segueix la llei Normal X~N(0;0.5).

Gràficament obtenim:



Numèricament hem de buscar a les taules de la llei Normal els valors $\pm z_{\alpha/2}$:

$$\pm z_{\alpha/2} = \pm z_{0.01/2} = \pm 2.575$$

Si desestandarditzem aquests valors trobarem l'error màxim que es pot cometre:

$$Z = \frac{X - \mu}{\sigma}$$

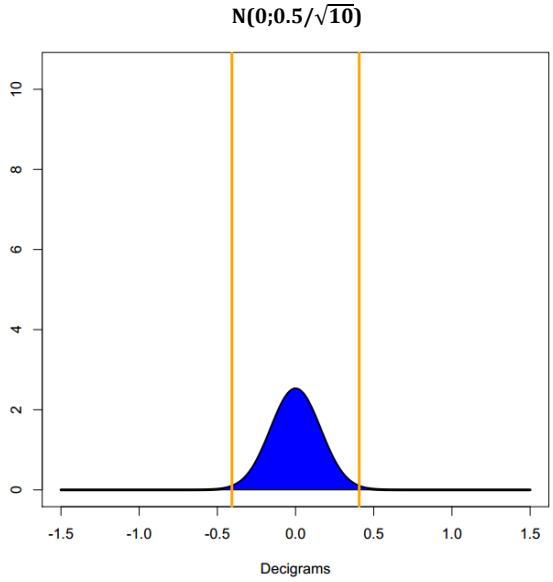
$$\pm 2.575 = \frac{\pm x - 0}{0.5}$$

$$\boldsymbol{\pm x = 1.288 \text{ decigrams}}$$

b) Igual que en l'apartat anterior, ens demanen quin és l'error màxim que es pot arribar a cometre amb una confiança del 99%. El procediment a seguir és el mateix, però ara hem de tenir en compte que n = 10.

Diem que la v.a. $\overline{X}_{10}$ = "Mitjana de 10 pesades" segueix la llei Normal $X \sim N(0; 0.5/\sqrt{10})$.

Gràficament obtenim:



Numèricament, alhora de desestandarditzar:

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

$$\pm 2.575 = \frac{\pm \bar{x}_{10} - 0}{0.5/\sqrt{10}}$$

$$\boldsymbol{\pm \bar{x}_{10} = 0.4071 \text{ decigrams}}$$

c) Perquè l'error màxim sigui inferior a 0.1 decigrams (amb una confiança del 99%) hem d'aplicar la següent fórmula, on ε es tracta de l'error màxim:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{\varepsilon}\right)^2$$

$$\mathbf{n} = \left(\frac{2.575 \cdot 0.5}{0.1}\right)^2 = 165.8 \cong \mathbf{166}$$

1.2. *Una màquina fabrica peces la longitud de les quals és una v.a. de llei Normal N(μ;σ). En un moment donat el procés es descentra (desplaçament de la mitjana) passant a fabricar peces segons una llei Normal N(μ+2σ;σ).*

Es demana:

a) Si s'escullen 4 peces consecutives de la cadena de producció. Quina probabilitat hi ha que la mostra contingui com a mínim una peça fora de l'interval [μ – 3σ, μ + 3σ]?
b) Calculeu la probabilitat que la mitjana de les 4 observacions estigui fora de l'interval [μ – 3σ$_{\bar{x}}$, μ + 3σ$_{\bar{x}}$] (sigui σ$_{\bar{x}}$ la desviació tipus de $\bar{x}_4$).
c) A la vista dels resultats anteriors, si es vol assegurar que el procés és centrat què es preferible, controlar les peces individuals, o bé controlar la mitjana de n peces?

a) Per calcular la probabilitat que una mostra de 4 peces contingui, com a mínim, una peça fora de l'interval [μ – 3σ, μ + 3σ] comencem calculant la probabilitat que una mostra d'una sola peça es trobi fora de l'interval que ens demanen.

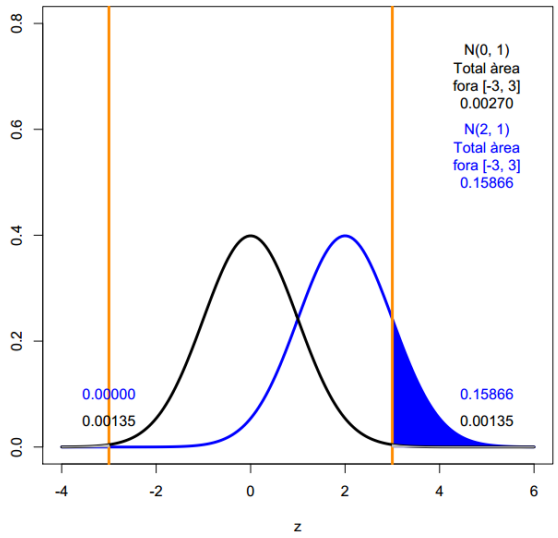
P(Longitud fora interval) =

= 1 – P(Longitud dins interval) =

= 1 – P(μ – 3σ ≤ Longitud ≤ μ + 3σ) =

$$= 1 - P\left(\frac{\mu - 3\sigma - (\mu + 2\sigma)}{\sigma} \le Z \le \frac{\mu + 3\sigma - (\mu + 2\sigma)}{\sigma}\right) = 1 - P(-5 \le Z \le 1) =$$

= 1 – [P(Z ≤ 1) – P(Z ≤ –5)] = 1 – [0.8413 – 0] = 0.1587



Continuem fent els càlculs tenint en compte que la probabilitat d'obtenir una o més peces fora de l'interval és contrària a la de no obtenir cap peça fora d'aquest:

X = "Nombre de peces fora de l'interval"

P(X ≥ 1) =

= 1 – P(X = 0) =

= 1 – P(1a peça dins interval ∩ 2a peça dins interval ∩ 3a peça dins interval ∩ 4a peça dins interval) =

= 1 – P(1a peça dins interval) · P(2a peça dins interval) · P(3a peça dins interval) · P(4a peça dins interval) =

= 1 – [P(Peça dins interval)]^4 =

= 1 – [1 – P(Peça fora interval)]^4 = 1 – [1 – 0.1587]^4 = **0.4990**

b) La llei Normal que segueix la v.a. Mitjana és la següent:

$$\text{Mitjana} \sim N\left(\mu + 2\sigma; \sigma/\sqrt{4}\right) = N(\mu + 2\sigma; \sigma/2)$$

Per calcular la probabilitat que aquesta mitjana estigui fora de l'interval $[\mu - 3\sigma_{\bar{x}}, \mu + 3\sigma_{\bar{x}}] = [\mu - 3\sigma/2, \mu + 3\sigma/2]$ ho farem igual que en l'apartat anterior:

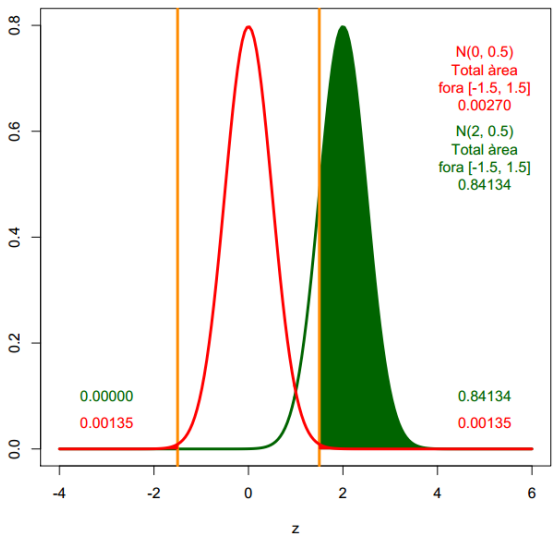
P(Mitjana fora interval) =

= 1 – P(Mitjana dins interval) =

= 1 – P(μ – 3σ/2 ≤ Mitjana ≤ μ + 3σ/2) =

$$= 1 - P\left(\frac{\mu - 3\sigma/2 - (\mu + 2\sigma)}{\sigma/2} \leq Z \leq \frac{\mu + 3\sigma/2 - (\mu + 2\sigma)}{\sigma/2}\right) = 1 - P(-7 \leq Z \leq -1) =$$

= 1 – [P(Z ≤ –1) – P(Z ≤ –7)] = 1 – [0.1587 – 0] = **0.8413**



c) Si es vol assegurar que el procés està centrat és preferible controlar la mitjana de les 4 peces, ja que aquest cas, la probabilitat de trobar una longitud fora de l'interval que ens diu l'enunciat és 0.8413, mentre que si es realitza un control individual la probabilitat passa a ser de 0.4990.

1.3. *Es vol estimar la mitjana μ del pes dels fulls de paper que es produeixen en una cadena de producció. S'escullen a l'atzar 22 d'aquests fulls i s'obté una mitjana mostral de $\bar{x}$ = 2.4 dg.*

Es demana:

a) Si la desviació estàndard del pes d'una fulla de paper és de 0.2 dg, trobeu un interval de confiança del 95% del valor de μ.
b) Si la desviació estàndard és desconeguda però la desviació estàndard s de la nostra mostra és igual a 0.2 dg, determineu l'interval de confiança del 95% del valor de μ.

Definim que la v.a. Pes segueix una llei Normal: Pes ~ N(μ;σ)

a) Com que σ és coneguda i σ = 0.2, l'IC(μ,95%) de μ vindrà donat per:

$$\mathrm{IC}(\mu, 1-\alpha) = \left(\bar{x}_n - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x}_n + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Busquem a les taules de la llei Normal els valors $\pm z_{\alpha/2}$:

$$\pm z_{\alpha/2} = \pm z_{0.05/2} = \pm z_{0.025} = \pm 1.960$$

Sabent que n = 22 i $\bar{x}_n = 2.4$, l'IC(95%) serà:

$$\mathrm{IC}(\mu, 95\%) = \left(2.4 - 1.960 \cdot \frac{0.2}{\sqrt{22}}, 2.4 + 1.960 \cdot \frac{0.2}{\sqrt{22}}\right)$$

$$\mathbf{IC}(\mu, \mathbf{95}\%) = (\mathbf{2.316}, \mathbf{2.484})$$

b) Com que ara σ és desconeguda però sabem que s = 0.2, l'IC(95%) de μ valdrà:

$$\mathrm{IC}(\mu, 1-\alpha) = \left(\bar{x}_n - t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}},\ \bar{x}_n + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}\right)$$

Sabent que n = 22, busquem a les taules de la distribució t-Student els valors $\pm t_{n-1,\alpha/2}$:

$$\pm t_{n-1,\alpha/2} = \pm t_{22-1,0.05/2} = \pm t_{21,0.025} = 2.080$$

Per tant, l'IC(μ, 95%) serà:

$$\text{IC}(\mu, 95\%) = \left(2.4 - 2.080 \cdot \frac{0.2}{\sqrt{22}}, 2.4 + 2.080 \cdot \frac{0.2}{\sqrt{22}}\right)$$

$$\mathbf{IC}(\mu, \mathbf{95}\%) = (\mathbf{2.311}, \mathbf{2.489})$$

1.4. *En una empresa envasadora de cerveses s'han realitzat gràfics de control, basats en mostres de 4 ampolles, per a la mitjana i l'amplitud del contingut (cm³) en cervesa de les ampolles que comercialitza. Aquests diagrames han proporcionat la següent informació:*

Diagrama de la mitjana: UCL = 330.99, CCL = 328, LCL = 325.01
Diagrama d'R: UCL = 9.4, CCL = 4.1, LCL = 0

S'ha confirmat que el procés està sota control.

Es demana:

a) Calculeu una estimació de la desviació típica σ del procés.
b) Si a partir de les dades del diagrama de rang, construïm un diagrama de control per a la desviació típica, quins límits tindria?
c) Si se suposa que aquestes ampolles volen vendre's com ampolles d'un terç de litre i que les especificacions que desitja el fabricant per al contingut de les ampolles són 333 ± 8 cm³, quin és l'índex de capacitat real? Quin percentatge d'ampolles no compleixen les especificacions? Quina correcció caldria fer en el procés d'envasatge per a augmentar el percentatge d'ampolles que compleixen les especificacions?

a) L'estimació de la desviació típica del procés $\hat{\sigma}$ la calcularem a partir dels rangs. Tenint en compte que la v.a. W compleix que $\sigma = R/W$ i que $E\{W\} = d_2$ trobem que:

$$\hat{\sigma} = \frac{\overline{R}}{d_2}$$

Com que ens basem en grups de 4 ampolles, buscant a les taules dels gràfics de control trobem que $d_2 = 2.059$. Per altra banda, sabem que $\overline{R} = CCL = 4.1$. Calculem el valor de $\hat{\sigma}$:

$$\hat{\boldsymbol{\sigma}} = \frac{4.1}{2.059} = \mathbf{1.991}$$

b) Per mostres de mida 4, busquem a les taules dels gràfics de control els valors $c_4 = 0.9213$, $B_3 = 0$ i $B_4 = 2.266$. Així, les línies de control del gràfic s (no històric) són les següents:

$$\mathbf{CCL} = \bar{s} = \hat{\sigma} \cdot c_4 = 1.991 \cdot 0.9213 = \mathbf{1.834}$$

$$\mathbf{LCL} = B_3 \cdot \bar{s} = 0 \cdot 1.834 = \mathbf{0}$$

$$\mathbf{UCL} = B_4 \cdot \bar{s} = 2.266 \cdot 1.834 = \mathbf{4.156}$$

c) L'índex de capacitat real C_{pk} es calcula de la següent manera:

$$C_{pk} = \min\{C_{pu}; C_{pl}\} = \left\{\frac{USL - \bar{\bar{x}}}{3\sigma}; \frac{\bar{\bar{x}} - LSL}{3\sigma}\right\}$$

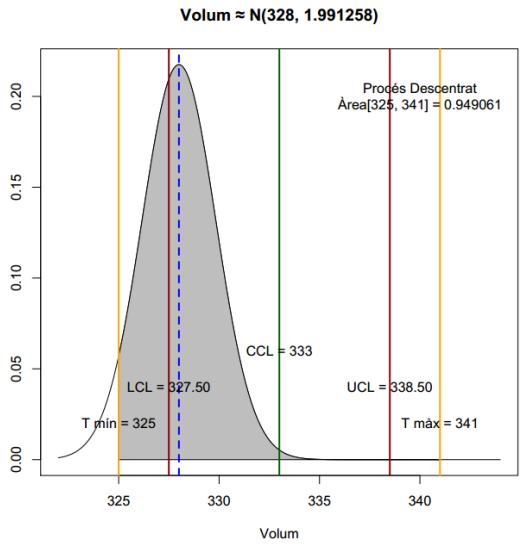
Atès que $\bar{\bar{x}} = CCL = 328$ i VN=333 l'índex C_{pk} és més informatiu que C_p.
Com que no disposem del valor σ, el substituïm per la seva estimació $\hat{\sigma} = 1.991$.
Per altra banda, sabem que $\bar{\bar{x}} = CCL = 328$. USL i LSL són la tolerància superior i inferior, respectivament, que ens dóna l'enunciat:

$$C_{pk} = \min\left\{\frac{(333 + 8) - 328}{3 \cdot 1.991}; \frac{328 - (333 - 8)}{3 \cdot 1.991}\right\}$$

$$C_{pk} = \min\{2.176; 0.5023\}$$

$$\mathbf{C_{pk} = 0.5023}$$

Com que $C_{pk} < 1$, es tracta d'un índex de capacitat real inadequat.



Per calcular el percentatge d'ampolles que no compleixen les especificacions, calculem primer el percentatge d'ampolles que sí que les compleixen:

P(Ampolles compleixen especificacions) =

$= P(333 - 8 \leq X \leq 333 + 8) = P(325 \leq X \leq 341) =$

$= P\left(\frac{325-328}{1.991} \leq Z \leq \frac{341-328}{1.991}\right) = P(-1.507 \leq Z \leq 6.529) =$

$= P(Z \leq 6.529) - P(Z \leq -1.507) = 1 - 0.06589 = 0.93411$

Per tant, el percentatge d'ampolles que no compleixen les especificacions és:

P(Ampolla no compleixen especificacions) =

= 1 – P(Ampolles compleixen especificacions) = 1 – 0.93411 = 0.06589 = **6.589%**

Per a augmentar el percentatge d'ampolles que compleixen les especificacions caldria ajustar la maquinària per a què la mitjana $\bar{\bar{x}} = 328$ s'acostés a 333. D'aquesta manera aconseguiríem el següent:

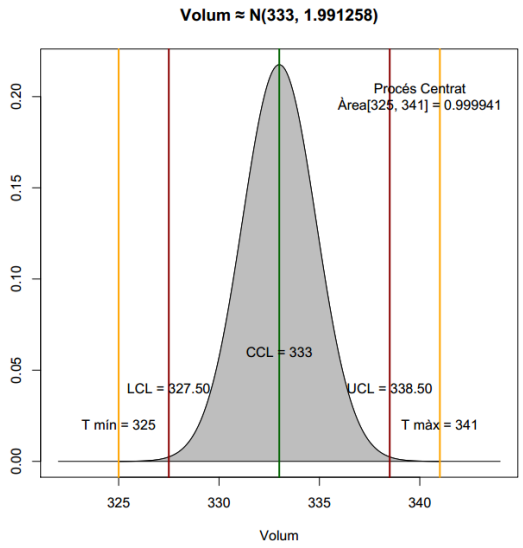
P(Ampolles compleixen especificacions) =

= P(333 – 8 ≤ X ≤ 333 + 8) = P(325 ≤ X ≤ 341) =

$= P\left(\frac{325-333}{1.991} \leq Z \leq \frac{341-333}{1.991}\right) = P(-4.018 \leq Z \leq 4.018) =$

= P(Z ≤ 4.018) – P(Z ≤ –4.018) = 1 – 0 = 1 = 100%

Podem observar que, en aquest cas, la distribució sí que es trobaria centrada:



Si calculem C_p podrem saber si el procés és capaç o no:

$$C_p = \frac{USL - LSL}{6\hat{\sigma}}$$

$$C_p = \frac{(333 + 8) - (333 - 8)}{6 \cdot 1.991}$$

$$C_p = 1.339$$

Com que $C_p > 1$ el procés és capaç, però com que $C_p > C_{pk}$ podem afirmar que la distribució no es troba centrada.

2. Exercicis proposats

2.1. *X és una v.a. d'esperança μ i variància σ^2 i X_1, ..., X_n v.a. independents que es distribueixen igual que la variable X.*

Es demana:

a) Quant valen l'esperança i la variància de la v.a. $X_1 + ... + X_n$?
b) Quant valen l'esperança i la variància de la v.a nX ?
c) Quina diferència conceptual hi ha entre la v.a. $X_1 + ... + X_n$ i la v.a. nX ?

Solució: a) $n\mu$; $n\sigma^2$; b) $n\mu$; $n^2\sigma^2$;
c) Sumar n valors aleatoris i multiplicar per un valor aleatori

2.2. *Per quant s'ha de multiplicar la mida n d'una mostra si volem reduir a la quarta part l'error estàndard de l'estimació puntual $\bar{X}_n$ de la mitjana μ d'una població?*

Solució: 16

2.3. *En un procés d'envasatge automàtic d'ampolles d'aigua mineral, el contingut net X de les ampolles segueix una llei Normal X~N(μ;σ=2) (en cl). Per tal de controlar el procés, s'ha seleccionat una mostra de 5 ampolles i s'ha mesurat el seu contingut net d'aigua mineral (en ml). Els resultats obtinguts són: 1002, 1000, 1002, 999 i 1001.*

Determineu un interval del 99% de confiança pel valor esperat μ del contingut net de les ampolles. Repetiu el càlcul però en el cas de σ desconeguda. Compareu els intervals obtinguts.

Solució: IC (σ coneguda) = (998.50, 1003.10);
IC (σ desconeguda) = (998.12, 1003.48) (menys precís)

2.4. *Un investigador estima sempre el valor de la mitjana μ d'una població fent servir intervals de confiança del 90%. Després de 400 estimacions, quin és el nombre aproximat d'intervals de confiança que contindran el verdader valor de μ?*

Solució: 360 intervals

2.5. *Una població té una mitjana μ desconeguda i una desviació estàndard igual a 5. Trobeu la mida n que ha de tenir una mostra per tal de tenir un 95% de confiança que l'estimador $\bar{x}$ calculat sobre aquesta mostra està dins l'interval (μ – 1.5, μ + 1.5).*

Solució: n = 43

2.6. *Es vol estimar –a un nivell de confiança del 0.99– l'esperança d'una distribució normal $N(\mu,\sigma^2)$ de variància coneguda. Calculeu quina ha de ser la mida n de la mostra perquè l'interval d'estimació tingui una longitud igual a 2δ.*

Solució: $n = (2.576\sigma/\delta)^2$

2.7. *A partir d'una mostra de mida n = 12 s'ha calculat un interval de confiança del 95% del valor de μ que val: IC(95%) = (18.6, 26.2). Quins són els valors de $\bar{x}$ i de s?*

Solució: $\bar{x} = 22.4$; $s = 5.98$

2.8. *Hom vol estimar la mitjana μ del temps d'espera dels clients en un caixer d'un gran supermercat a una determinada hora punta. Se sap d'altres vegades que la variància del temps d'espera és aproximadament de 8.0 min^2. Quants clients caldrà controlar si hom vol, amb un 90% de confiança, que el verdader valor μ difereixi com a màxim en 1 min de la mitjana de la nostra mostra?*

Solució: n = 21

2.9. *Les mesures dels diàmetres d'una mostra aleatòria de 10 boles de rodaments produïdes per una màquina al llarg d'un determinat període de temps van donar un valor de la mitjana mostral igual a 0.84 cm i una desviació típica de 0.06 cm.*

Es demana:

a) Trobeu l'interval de confiança del 95% per al diàmetre mig.
b) Trobeu l'interval de confiança del 99% per al diàmetre mig.
c) Compareu la longitud dels intervals obtinguts. Per què un té longitud superior a l'altra?

Solució: a) IC(95%) = (0.7971, 0.8829); b) IC(95%) = (0.7783, 0.9017);
c) L'interval augmenta a mesura que augmenta el nivell de confiança

2.10. *En un procés industrial es controla la resistència a la tensió de certes peces metàl·liques. S'ha mesurat la resistència de 30 mostres, de 6 elements cada una, i s'ha obtingut una mitjana igual a $\bar{\bar{x}}=200$ i una desviació igual a $\sigma=5$.*

Es demana:

a) Calculeu, a partir d'aquestes mostres, els límits de control per a la mitjana i per a la desviació.
b) S'ha conclòs que el procés està sota control. Determineu l'índex de capacitat si els límits de tolerància són 200 ± 5.
c) Quantes peces metàl·liques defectuoses produeix aquest procés? (NOTA: s'entén que una peça és defectuosa si sobrepassa els límits de tolerància)

d) En un moment concret es desajusta el procés i es fabriquen peces metàl·liques amb mitjana 199, però sense modificar-se'n la variància. Quina és la probabilitat de detectar el desajust en la següent mostra que es prengui de 6 elements?

Solució: a) Gràfic $\overline{X}$ (no històric): LCL = 193.56, UCL = 206.44;
Gràfic S (no històric): LCL = 0.15, UCL = 9.85;
b) C_p = 0.317; c) 34.13%; d) 0.27%

2.11. *Si la mitjana del pes d'unes llaunes de conserva és 41.5 g i la desviació típica és 0.5 g.*

Es demana:

a) Trobeu els límits de control teòrics per a la mitjana mostral si el nombre d'elements de cada mostra és n = 5.
b) Trobeu els límits de control teòrics per a la desviació típica.
c) La taula ens mostra els valors obtinguts per a la mitjana i la desviació típica de 20 mostres de mida n = 5 del mateix procés:

$\overline{x}$	41.9	41.3	42.1	41.6	41.8	42.3	41.4	41.6	41.8	42.0
s	0.8	0.2	0.3	0.7	0.9	0.1	0.4	0.5	0.6	0.2

$\overline{x}$	42.0	41.8	41.3	42.0	42.0	41.7	41.5	41.49	41.6	41.4
s	0.3	0.3	0.2	0.5	0.6	0.2	0.4	0.4	0.3	0.6

Indiquen aquests valors que el procés està sota control en mitjana? I en desviació?

Solució: a) Gràfic $\overline{X}$ (històric): LCL = 40.829, UCL = 42.171;
b) Gràfic S (històric): LCL = 0, UCL = 0.982;
c) En mitjana, la sisena mostra està fora de control;
En desviació, el procés està sota control

2.12. *El diàmetre de les varetes metàl·liques VM032 és una característica important en la seva qualitat i és important realitzar-ne un control en el seu procés de producció. D'acord amb el disseny original, les especificacions per al diàmetre de les varetes són 0.50350 ± 0.0010 polzades. La taula següent mostra els valors de les mitjanes ($\bar{x}$) i dels rangs (R) per a 20 mostres de 5 varetes cada una fabricades en un màquina-torn específic. Noteu que els valors que apareixen per a la mitjana en la taula són només les tres darreres xifres de la mesura. És a dir, 342 significa 0.50342. Els valors de R cal multiplicar-los per 10^{-4}, és a dir, 8 representa 0.0008.*

Mostra	$\overline{x}$	**R**
1	342	3
2	316	4
3	318	4
4	334	5
5	350	4
6	321	2

Mostra	$\overline{x}$	**R**
11	354	8
12	340	6
13	360	4
14	372	7
15	352	3
16	334	10

7	*326*	*7*	***17***	*350*	*4*
8	*338*	*9*	***18***	*344*	*7*
9	*348*	*10*	***19***	*339*	*8*
10	*386*	*4*	***20***	*340*	*4*

Es demana:

a) Trobeu els límits de control del gràfic $\bar{x}$ i, si cal, reviseu-los, eliminant les mostres fora de control.
b) Calculeu l'índex de capacitat del procés.
c) Quin percentatge de varetes metàl·liques defectuoses està produint el procés?

Solució: a) Gràfic $\overline{X}$ (no històric): CCL = 0.50343, LCL = 0.5031, UCL = 0.50373;
Eliminem la mostra 10 (es troba fora de control) i tornem a fer els càlculs;
Gràfic $\overline{X}$ (no històric): CCL = 0.503409, LCL = 0.50308, UCL = 0.50374;
b) C_p = 1.3515; c) 0.009%

2.13. *En el seu procés de producció, la longitud de l'encenedor de cigarretes d'automòbil és controlada mitjançant gràfics de control per a la mitjana i el rang. La taula següent mostra les mesures de la longitud de 20 mostres de mida 4.*

	Observacions					***Observacions***			
Mostra	***1***	***2***	***3***	***4***	***Mostra***	***1***	***2***	***3***	***4***
1	*5.15*	*5.10*	*5.08*	*5.09*	***11***	*5.13*	*5.08*	*5.09*	*5.05*
2	*5.14*	*5.14*	*5.10*	*5.06*	***12***	*5.10*	*5.15*	*5.08*	*5.10*
3	*5.09*	*5.10*	*5.09*	*5.11*	***13***	*5.08*	*5.12*	*5.14*	*5.09*
4	*5.08*	*5.06*	*5.09*	*5.13*	***14***	*5.15*	*5.12*	*5.14*	*5.05*
5	*5.14*	*5.08*	*5.09*	*5.12*	***15***	*5.13*	*5.16*	*5.09*	*5.05*
6	*5.09*	*5.10*	*5.07*	*5.13*	***16***	*5.14*	*5.08*	*5.08*	*5.12*
7	*5.15*	*5.10*	*5.12*	*5.12*	***17***	*5.08*	*5.10*	*5.16*	*5.09*
8	*5.14*	*5.16*	*5.11*	*5.10*	***18***	*5.08*	*5.14*	*5.10*	*5.09*
9	*5.11*	*5.07*	*5.16*	*5.10*	***19***	*5.13*	*5.15*	*5.10*	*5.08*
10	*5.11*	*5.14*	*5.11*	*5.12*	***20***	*5.09*	*5.07*	*5.15*	*5.08*

Es demana:

a) Trobeu els límits de control superiors i inferiors per a la mitjana i el rang, eliminant, si fos necessari, les mostres fora de control.
b) Si l'interval de tolerància és (5.05, 5.15), calculeu l'índex de capacitat i estimeu la proporció d'encenedors que quedarien fora de l'interval.

Solució: a) Gràfic $\overline{X}$ (no històric): LCL = 5.0575, UCL = 5.1565
(no hi ha mostres fora de control);
Gràfic R (no històric): LCL = 0, UCL = 0.15518
(no hi ha mostres fora de control);
b) C_p = 0.50465; 10.238%

PRÀCTIQUES

1. Estimació de la mitjana a partir d'una mostra

Treballem les dades de la mostra que conté l'arxiu `vel.llum`. Les dades d'aquest fitxer corresponen a les mesures repetides de la velocitat de la llum (en km/s) que el físic Albert Michelson va realitzar a finals del segle XIX.

Si Michelson hagués estimat a partir de les seves mesures, a un nivell de confiança del 95%, la velocitat de la llum, quin interval de confiança hauria obtingut? Vegem com fer-ho ràpidament.

R-Commander fa servir la mateixa funció per calcular intervals i per fer contrastos d'una variable, en la qual assumeix el cas de variància desconeguda i utilitza la distribució t-Student per a l'estadístic mostral $\frac{\bar{X}_n-\mu}{s/\sqrt{n}}$:

`Estadístics, Mitjanes, t-test per una variable ...`

Escolliu la variable (en aquest cas l'única al panell) i marqueu el nivell de confiança 0.95. Premeu `D'acord`.

Observant la resposta donada per R-Commander, contesteu:

- Quina és l'estimació puntual de la mitjana? $\bar{x}$ = 299896.7.
- Quina és el valor de l'estadístic $t = \frac{\bar{X}_n-\mu}{s/\sqrt{n}}$? t = 10378.85.
- Quants graus de llibertat té la distribució t-Student utilitzada? n – 1 = 14.
- Quin és l'interval d'estimació del 95% de confiança de la mitjana de la velocitat de la llum? IC(95%): (299834.7, 299958.6).
- Repetiu l'estimació a un nivell de confiança del 99%. Quin és ara el nou interval de confiança? IC(99%): (299810.7, 299982.7).
- Perquè aquest interval és més llarg que l'anterior? A les Taules t-Student, per 14 graus de llibertat trobem $t_{0.025}$ = 2.145 i $t_{0.005}$ = 2.977. Aleshores, l'amplada de l'interval augmenta quan el nivell de confiança augmenta. També es pot calcular anat a: `Distribucions, Distribucions contínues, Distribució t, Quantils ...`
- Feu una estimació al 90% de confiança: IC(90%): (299845.8, 299947.6). Ara ha disminuït l'amplada de l'IC.

2. Gràfics de control

R-Commander té el complement (plug-in) `qcc` per a representar els gràfics de control i els diagrames de capacitat d'un procés. Aquest plug-in permet fer els anàlisis per a processos no històrics. Pels processos històrics utilitzarem comandes que executarem des de la finestra d'instruccions. Les comandes que utilitzarem es

basen en funcions del mateix paquet del plug-in qcc. En primer lloc hem d'instal·lar (en cas que no ho estigui ja) i carregar el plug-in qcc. Per instal·lar-lo executem la següent instrucció:

```
install.packages( 'RcmdrPlugin.qcc' )
```

A continuació anem al menú:

```
Eines, Carrega plugin del Rcmdr ...
```

Escollim el plug-in qcc. El programa ens demanarà reiniciar el R-commander. Una vegada reiniciat el programa, carregueu l'arxiu de dades resist.rda mitjançant el menú:

```
Dades, Carrega taula de dades ...
```

Aquest arxiu conté dades sobre una mostra de 120 resistors escollits en el decurs d'un procés de control de fabricació automàtica de resistors de l'empresa ResGirona S.A.. El procés de control consistia en escollir cada 6 minuts una mostra de 4 resistors i mesurar-ne la resistència. Les variables de què disposem són:

- *Temps_Hora*: Moment en què es va recollir la mostra, des de 6 min. fins a 180 min.
- *Num_Mos*: Número de la mostra, des de la 1 fins a la 30.
- *Num_Uni*: Número de la unitat (1, 2, 3 o 4) dins de cada mostra.
- *Res*: Resistències (en ohms) dels 120 resistors.

Visualitzeu la taula de dades. La manera com estan organitzades les dades no és l'adequada per a poder executar les facilitats del plug-in qcc. Aleshores, abans de representar els gràfics de control és necessari agrupar les dades convenientment. Copieu i executeu la comanda:

```
resi=as.data.frame(qcc.groups(resist$Res,resist$Num_Mos))
```

Canvieu la taula de dades activa i visualitzeu les dades resi. Observareu que resi és una matriu on cada fila és una mostra i les quatre columnes són les unitats de cada mostra.

2.1. Control de la mitjana

Per a construir el gràfic de control de la mitjana Xbar anem a:

```
Control Charts, Xbar, Graph ...
```

En la finestra que s'obre hi seleccionem les quatre columnes i deixem el valor de k igual a 3. Aquest paràmetre controla la quantitat de sigmes amb les que calculem

els límits de control. El gràfic que s'obté en fer un clic en `D'acord` ha de ser com el de la **figura 1**.

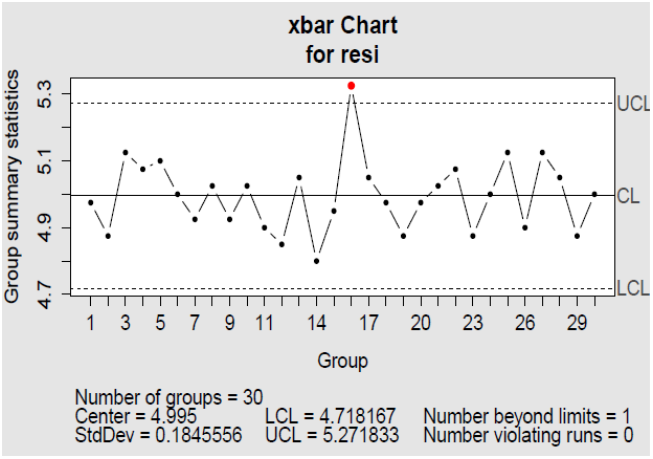


Figura 1: Gràfic de control `Xbar` per als resistors de l'empresa ResGirona S.A..

La funció `qcc()` és la responsable de crear el gràfic. Aquesta funció, per defecte, estima la desviació tipus σ mitjançant els rangs de les mostres. Per tant, utilitza el valor de $\overline{R}$ per a calcular els límits del gràfic. La funció `stats.R(resi)` calcula l'amplitud de cada mostra i la corresponent mitjana de les amplituds $\overline{R}$. Per altra banda, el valor `StdDev = 0.1845556`, que apareix en la **figura 1** del gràfic de control de la mitjana és l'estimació de σ a partir dels rangs (recordeu que aquesta estimació s'obté mitjançant l'expressió $\overline{R}/d_2$). En R-Commander es pot calcular directament executant `sd.xbar(resi, std.dev=c('UWAVE-R'))`, on el paràmetre `UWAVE-R` especifica que la desviació s'estima mitjançant una mitjana (*AVErage*) no-ponderada (*UnWeighted*) dels rangs (R) de les mostres.

A la vista dels resultats obtinguts i de la figura del gràfic de control de la mitjana responeu a les següents qüestions:

- Quant val $\overline{R}$ (`stats.R(resi)`)? $\overline{R} = 0.38$.
- Quant val el valor de la línia central del gràfic? $\overline{\overline{x}} = 4.995$. Com es calcula? És la mitjana de les mitjanes de les mostres.
- Quant valen els valors dels dos límits de control del gràfic? Com es calculen? [LCL, UCL] = [4.718, 5.272]. Es calculen fent: $\overline{\overline{x}} \pm A_2\overline{R} = 4.995 \pm 0.729 \cdot 0.38$.
- Està el procés fora de control? Per què? Sí, perquè la mostra número 16 té la mitjana per sobre del UCL.
- A la vista de l'evolució global del gràfic, creieu que el descontrol del procés va ser degut a una causa fortuïta o a una causa sistemàtica? Raoneu la vostra resposta. El descontrol és degut a una causa fortuïta perquè no s'observa cap mena de patró (ratxa, tendència) en les mostres anteriors ni posteriors.

Queda clar que la mostra 16 es troba fora de control. Podem filtrar aquesta dada i estudiar el nou gràfic si executem l'ordre:

```
resi2=resi[-16,]
```

Canvieu el conjunt de dades actiu i visualitzeu-lo.

- Feu el gràfic Xbar del conjunt de dades resi2 i comenteu-lo. Els valors del centre, desviació i límits canvien lleugerament si eliminem la mostra número 16. Ara, les 29 mostres estan sota control.

2.2. Control de la variabilitat

- **Gràfic S**

Anem a estudiar si la variabilitat del procés està sota control a partir d'un gràfic S (gràfic de desviacions estàndard). El procediment és similar a l'anterior. Treballarem amb el conjunt de dades resi2, el qual té extreta la mostra 16, perquè ja hem vist que no forma part de la variabilitat natural del procés.

Per a construir el gràfic de control S anem a:

```
Control Charts, S, Graph ...
```

Seleccionem les quatre columnes i deixem el valor de 3 sigmes que apareix per defecte.

A la vista del gràfic responeu les següents qüestions:

- Quant val el valor de la línia central (stast.S(resi2))? $\bar{s}$ = 0.17465. Com es calcula? Es calcula fent la mitjana de les desviacions de les mostres.
- Quant valen els valors dels dos límits de control del gràfic? Com es calculen?
  [LCL, UCL] = [0, 0.39578].
  Es calculen fent: $[B_3\bar{s}, B_4\bar{s}] = [0 \cdot 0.17465, 2.266 \cdot 0.17465]$.
- Està el procés fora de control? Per què? El procés està controlat perquè cap observació surt fora dels límits. No s'observa cap ratxa ni tendència. Sí que s'observen massa punts consecutius (del 12 al 24) per sota de la línia central, fet que indica un desplaçament (reducció) del valor central.
- Interpreteu la informació que dóna el gràfic referent a Number violating runs. A partir del sisè punt consecutiu per sota de la línia, el gràfic marca amb color les observacions que segueixen estant en el mateix costat de la línia.

- **Gràfic R**

Utilitzant el mateix plug-in sobre les dades resi2 estudiem si la variabilitat d'un procés està sota control, però ara a partir d'un gràfic R.

A la vista del gràfic responeu a les següents qüestions:

- Quant val el valor de la línia central? $\overline{R}$ = 0.3827586. Com es calcula? És la mitjana dels rangs de les mostres.
- Quant valen els valors dels dos límits de control del gràfic? Com es calculen?
  [LCL, UCL] = [0, 0.8734169].
  Es calculen fent: $[D_3\overline{R}, D_4\overline{R}] = [0 \cdot 0.3828, 2.282 \cdot 0.3828]$
- Està el procés fora de control? Per què? El procés està controlat, ja que no s'observa cap ratxa ni tendència.

2.3. Dades amb informació històrica

Suposem que se sap que quan el procés funciona correctament la resistència dels resistors es distribueix segons una llei normal amb $\mu = 5$ i $\sigma = 0.2$.

Ens interessa saber, a partir de les dades `resi`, si el procés de fabricació està sota control. Començarem representant un gràfic de variabilitat per veure si està sota control. Farem els gràfics `S` i `R`. Farem servir la funció `qcc` del plug-in afegint les informacions històriques com a paràmetres, en aquest cas, `center` i `std.dev`.

Anem a estudiar la variabilitat d'un procés a partir d'un gràfic `S` (gràfic de desviacions estàndard). Com que es tracta d'un procés històric, ho hem de indicar a la funció `qcc`. Recordeu que el centre del gràfic es calcula fent $c_4\sigma = 0.9213 \cdot 0.2 = 0.18426$, per tant hem d'executar l'ordre:

```
qcc(resi, type='S', center=0.18426, std.dev=0.2)
```

El gràfic que obtenim ens mostra l'evolució de la desviació estàndard de les 30 mostres.

Responeu a les següents preguntes:

- Quant val el valor S que figura a la línia central? Valor central = 0.18426. Com es calcula el seu valor a partir de les taules? És l'estimació de la mitjana de les desviacions mostrals. Es calcula fent $c_4\sigma$.
- Quant val el valor corresponent a la línia superior del gràfic S? Com es calcula? Valor UCL = 0.4175. Es calcula fent: $B_6\sigma = 2.085 \cdot 0.2$.
- Per què el límit inferior és igual a 0?
  Es calcula mitjançant $B_5\sigma = 0 \cdot 0.2 = 0$. LCL no serà mai negatiu, ja que una desviació estàndard mai serà negativa.
- Sembla que tot està sota control? Què observeu? El procés està controlat, ja que cap observació surt fora del límits. No s'observa cap ratxa ni tendència, però sí massa punts consecutius (del 12 al 24) en el mateix costat de la línia central, suggerint un desplaçament (reducció) de la variabilitat.

Si volem fer el gràfic `R`, hem d'indicar el centre, $\sigma d_2 = 0.2 \cdot 2.059 = 0.4118$ i executar l'ordre:

```
qcc(resi, type= 'R' , center=0.4118, std.dev=0.2)
```

- Comenteu i comproveu els resultats d'aquest gràfic. La línia central (0.4118) es calcula fent σd_2. Els límits [LCL, UCL] = [0, 0.9396865] es calculen fent $[D_1\sigma, D_2\sigma]$. El procés està controlat i es detecten massa punts consecutius per sota de la línia central.

Ara estudiarem el gràfic de control de `Xbar`. En aquest cas l'ordre que s'ha d'executar és molt senzilla:

```
qcc(resi, type= 'xbar' , center=5,  std.dev=0.2)
```

Responeu a les següents preguntes:

- Què hi figura a la línia central? La mitjana = 5.
- Quina és la fórmula per calcular el valor del límit superior a partir dels valors de μ, σ i n (mida dels subgrups)? $\mu \pm A\sigma = \mu \pm \frac{3}{\sqrt{n}}\sigma$.
- Quina és la mostra que cau fora de control? La número 16.
- A la vista de l'evolució global del gràfic, creieu que el descontrol del procés va ser degut a una causa fortuïta o a una causa sistemàtica? Raoneu la vostra resposta. Degut a una causa fortuïta perquè ni abans ni després de l'observació descontrolada s'observa cap mena de tendència, patró o ratxa.
- Com hem fet abans, podem excloure la mostra 16 i repetir el gràfic de control. Feu-ho i compareu els gràfics. Executant l'ordre `qcc(resi[-16,], type= 'xbar' , center=5,  std.dev=0.2)` obtenim un gràfic molt semblant al obtingut sense informació històrica. Ara el procés està controlat.

3. Capacitat d'un procés

L'empresa multinacional ResWorld Corporation necessita que una empresa subministradora li proporcioni resistors d'unes determinades característiques per incorporar-los als aparells electrònics que fabrica. La multinacional necessita concretament resistors amb una resistència de 5 Ω, tot i que pot arribar a admetre resistència dins del marge de tolerància 4,25 Ω – 5,75 Ω. Si per defecte de fabricació la resistència d'un resistor està fora d'aquest límits, l'aparell electrònic que els conté pot tenir greus problemes de funcionament. Per això, la ResWorld Corporation, abans d'acceptar a la ResGirona S.A. com a proveïdora de resistors, vol assegurar-se que aquesta petita empresa és capaç de complir les especificacions que la multinacional necessita per als seus resistors.

Amb aquest objectiu, la ResWorld Corporation realitza una auditoria a l'empresa ResGirona S.A. per controlar el seu procés de producció de resistors. Imagineu-vos que les dades de la variable *Res* són les que s'han obtingut durant l'auditoria del procés de producció. Les mostres s'han obtingut seguint el procediment que s'ha explicat al començament de la pràctica. Recordem que la ResWorld Corporation necessita resistors de resistència compreses dins l'interval 4,25 Ω – 5,75 Ω.

Les dades són les mateixes que es troben al fitxer `resist` i l'objecte que analitzarem és `resi2`, el qual hem definit tot eliminant l'observació número 16.

Per estudiar la capacitat del procés anem al menú:

```
Control Charts, Xbar, Capability ...
```

On seleccionarem les quatre columnes, deixarem el nombre de sigmes igual a 3 i escriurem els límits de tolerància [LSL, USL] = [4.25, 5.75].

El gràfic ens ajusta una distribució normal a l'histograma de les dades i ens marca el valor central (*target*) de l'interval de tolerància i els seus extrems (LSL i USL).

Responeu a les següents qüestions:

- Quantes dades estan fora dels límits LSL i USL especificats? Totes les dades estan dins dels límits, no n'hi ha cap que estigui fora.
- A la vista del gràfic, diríeu intuïtivament que el procés de producció de ResGirona S.A. és capaç de complir amb les especificacions demanades per la ResWorld Corporation? Per què? Sí, totes les dades de la distribució mostral estan dins de l'interval i encara queda un marge entre l'amplitud de l'histograma i els extrems de l'interval.

A més del gràfic, R ens informa sobre diferents característiques del procés:

- LSL = Límit inferior de l'interval de tolerància. LSL = 4.25.
- USL = Límit superior de l'interval de tolerància. USL = 5.75.
- *Center* $\bar{\bar{x}}$. $\bar{\bar{x}}$ = 4.983621.
- Desviació tipus ponderada. Desv. Tip. Pond. = 0.1858954.
- Nombre d'observacions. Nomb. Obs. = 116.

Índexs de capacitat (anomenada *Within*):

- Índex de capacitat global. CP = 1.344842.
- Índex de capacitat superior. CPU = 1.374212.
- Índex de capacitat inferior. CPL = 1.315473.
- Índex real. CPK = 1.315473.

R també ens informa dels valors observats i esperats que queden fora de les especificacions:

- Percentatge de valors observats sota LSL. 0%.
  Percentatge de valors esperats sota LSL. 0%.
- Percentatge de valors observats sobre USL. 0%.
  Percentatge de valors esperats sobre USL. 0%.
- Si fóssiu els responsables de la ResWorld Corporation, agafaríeu a la ResGirona S.A. com a proveïdora de resistors? Sí, perquè la probabilitat de trobar-nos un resistor defectuós (fora de la tolerància) és pràcticament nul·la.

Repetiu l'estudi de capacitat en el supòsit que els límits de tolerància imposats per la ResWorld Corporation fossin resistències compreses entre 4.5 Ω i 5.5 Ω.

- CP. CP = 0.897.
- CPU. CPU = 0.926.
- CPL. CPL = 0.867.
- CPK. CPK=0.867 (inferor a 1, procés no capaç).

- Percentatge de valors observats sota LSL. 0%.
  Percentatge de valors esperats sota LSL. 0.46%.
- Percentatge de valors observats sobre USL. 0%.
  Percentatge de valors esperats sobre USL. 0.27%.
- Diríeu que el procés de producció de ResGirona S.A. és capaç de complir les últimes especiacions marcades per la ResWorld Corporation? Malgrat en la mostra no s'observen valors fora de l'interval de tolerància, no podem dir que el procés és capaç de complir les especificacions ja que CPK=0.867<1.
- Si fóssiu els responsables de la ResWorld Corporation, agafaríeu a la ResGirona S.A. com a proveïdora de resistors? No, perquè ens subministraria resistors defectuosos.
- Si fóssiu els responsables de la ResGirona S.A, quines accions emprendríeu per tal de millorar la capacitat del procés i poder servir a l'empresa multinacional? Prendre mesures per ajustar millor la producció i reduir la variància.

TEMA 5: Contrast d'hipòtesi

TEORIA

1. Contrast d'hipòtesi

L'objectiu del contrast d'hipòtesi és contrastar, a partir d'unes dades mostrals, la versemblança entre dues hipòtesis complementàries que ens plantejarem. Una d'elles l'anomenarem **hipòtesi nul·la (H_0)**, la qual és la més conservadora ja que a priori es considera que serà la certa. L'altra l'anomenarem **hipòtesi alternativa (H_1)**, la qual es considera d'entrada que no serà certa i, si de cas, seran les dades que ens la faran escollir.

Així doncs, d'un contrast de la H_0 envers la H_1 extraurem una les dues següents conclusions:

- "No tenim motius suficients per rebutjar la H_0".
- "Tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

La majoria de contrasts plantegen hipòtesis que fan referència a un determinat paràmetre θ (**contrast paramètric**). Tot i això, a final del tema estudiarem el contrast de bondat d'ajust que no és paramètric, és a dir, les hipòtesis no fan referència a un paràmetre.

1.1. Contrasts paramètrics i p-valor

Partim d'un paràmetre poblacional qualsevol θ que no podem saber-ne el valor però sabem que seguirà una distribució determinada (per exemple $\overline{X}_n \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$ segons el teorema del límit central). Per a estimar el valor θ, s'extreu una mostra i se n'obté un paràmetre estimat $\hat{\theta}$, el qual anomenarem $\hat{\theta}_{0observat}$, en endavant $\hat{\theta}_{0obs}$.

Per altra banda a través de la H_0 plantejarem si el paràmetre estudiat θ és igual, menor o major que un valor en concret θ_0. Per tant podrem plantejar tres tipus de contrasts paramètrics:

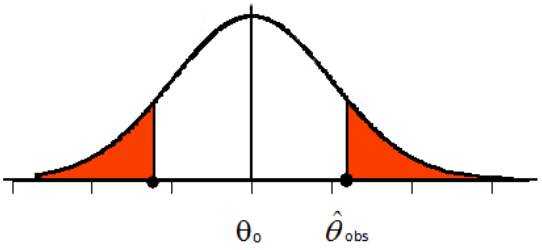
Bilateral:	**Unilateral dreta:**	**Unilateral esquerra:**
$H_0 : \theta = \theta_0$ $H_1 : \theta \neq \theta_0$	$H_0 : \theta \leq \theta_0$ $H_1 : \theta > \theta_0$	$H_0 : \theta \geq \theta_0$ $H_1 : \theta < \theta_0$

Encara que no s'observi una igualtat idèntica entre $\hat{\theta}_{0obs}$ i θ_0, moltes vegades, quan $\hat{\theta}_{0obs}$ es troba a prop de θ_0, podem afirmar estadísticament que la H_0 és certa. Per

això calculem la probabilitat p d'obtenir un valor mostral $\hat{\theta}$ tant o més allunyat de θ_0 que el valor $\hat{\theta}_{0bs}$. Aquesta probabilitat l'anomenarem **p-valor**.

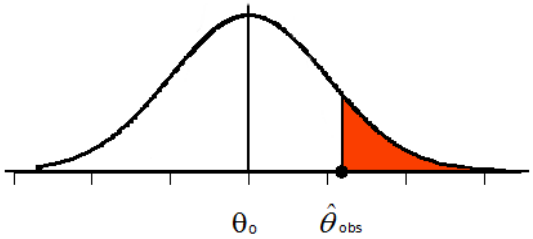
Sabent això podem entendre els tres tipus de contrasts de la següent manera:

- **Bilateral:**



$$p = P[|\hat{\theta} - \theta_0| \geq |\hat{\theta}_{0bs} - \theta_0|]$$

- **Unilateral dreta:**



$$p = P[(\hat{\theta} - \theta_0) \geq (\hat{\theta}_{0bs} - \theta_0)]$$

- **Unilateral esquerra:**

$$p = P[(\theta_0 - \hat{\theta}) \geq (\theta_0 - \hat{\theta}_{0bs})]$$

Observem que la zona de color taronja s'ajusta a la definició del p-valor. Depenent de si el valor d'aquesta variable és gran o petit, extraurem una conclusió o una altra.

Per jutjar si el p-valor és gran o petit necessitem una referència. Depenent del criteri agafat i de la referència marcada, un mateix p-valor podria veure's gran o petit. Per no tenir aquest problema utilitzarem l'anomenat **nivell de significació**.

1.2. Nivell de significació, α

Entenem per nivell de significació (α) com el valor mínim que ha de tenir el p-valor per tal de continuar acceptant la H_0 com a certa. D'aquesta manera la regla de decisió serà la següent:

- Si **p-valor ≥ α** considerarem que la diferència entre θ_0 i $\hat{\theta}_{0bs}$ és deguda a l'atzar, i per tant direm que "No tenim motius suficients per rebutjar la H_0".

- Si **p-valor < α** considerarem que la diferència entre θ_0 i $\hat{\theta}_{0bs}$ no és deguda a l'atzar, sinó que la H_0 és falsa. Per això direm que "Tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

Com que d'entrada la H_0 sempre la considerarem com la hipòtesi certa, assignarem al nivell de significació un valor molt petit. Els valors més comuns que sol prendre α són del **10%**, **5%** i **1%**.

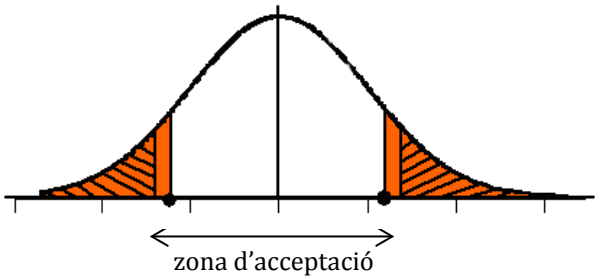
A partir del nivell de significació podem definir la regió anomenada **zona d'acceptació**. Si l'estadístic cau dins d'aquesta zona es complirà que el p-valor serà major al nivell de significació α.

Exemple: *Se'ns proporciona la gràfica d'una distribució normal amb les probabilitats que representen el nivell de significació α i el p-valor marcades. Per cada un dels casos següents dir quina de les hipòtesis plantejades acceptaríem. La llegenda pels tres casos és la següent:*

p-valor

α

CAS 1: el primer que deduïm és que es tracta d'un contrast bilateral amb les següents hipòtesis:

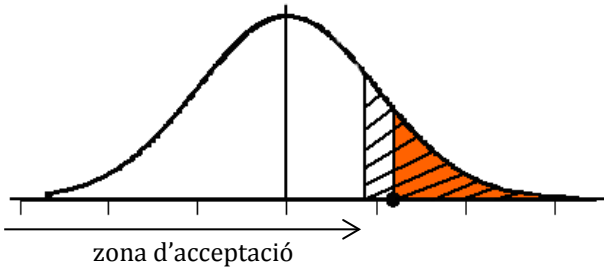


$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

Observem que el p-valor és major al nivell de significació, per tant diem que "no tenim motius suficients per rebutjar la H_0".

CAS 2: el primer que deduïm és que es tracta d'un contrast unilateral amb les següents hipòtesis:

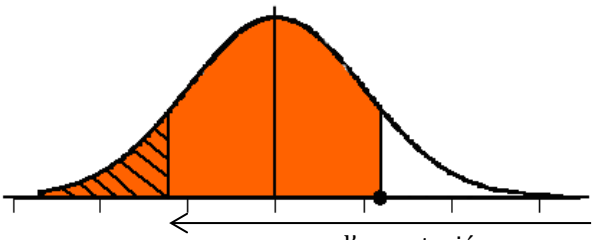


$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta > \theta_0$$

Observem que el p-valor és menor al nivell de significació, per tant diem que "tenim motius suficients per rebutjar la H_0 i per això acceptem la H_1".

CAS 3: el primer que deduïm és que es tracta d'un contrast unilateral amb les següents hipòtesis:



$$H_0 : \theta \geq \theta_0$$

$$H_1 : \theta < \theta_0$$

Observem que el p-valor és major al nivell de significació, per tant diem que "no tenim motius suficients per rebutjar la H_0".

1.3. Errors en la decisió

Malgrat acceptar amb fonaments estadístics una de les dues hipòtesis no és segur que haguem escollit correctament. En cas d'equivocar-nos podem cometre dos tipus d'error:

- **Error tipus I**: originat quan es rebutja la H_0 quan és certa.
- **Error tipus II**: originat quan s'accepta la H_0 quan la H_1 és la certa.

En el següent quadre podem veure els quatre casos amb què ens podem trobar:

	Decisió que es pren	
Hipòtesi certa	**Acceptar H_0**	**Rebutjar H_0**
H_0	Decisió correcta	Error de tipus I
H_1	Error de tipus II	Decisió correcta

La probabilitat de cometre un error de tipus I és igual al p-valor associat a la mostra obtinguda per realitzar el contrast:

$$P(\text{error tipus I}) = p - \text{valor}$$

Atès que només decidirem rebutjar H_0 quan p-valor < α, el nivell de significació és l'error màxim de tipus I que es pot cometre.

La probabilitat de cometre un error de tipus II dependrà del valor que realment tingui el paràmetre θ sobre el que es realitza el contrast. Com que el valor real de θ és desconegut, aquesta probabilitat es calcula en funció dels possibles valors del paràmetre a través de les anomenades **corbes de potència o característiques**:

$$\mathrm{P}(\text{error tipus II}) = \beta$$

Altrament d'un contrast podem determinar altres característiques com la potència d'un contrast, la corba de potència i la corba característica.

Fixat el nivell de significació α d'un contrast, es defineix la **potència d'un contrast** d'hipòtesis com la probabilitat de rebutjar la hipòtesi nul·la H_0 quan el valor real del paràmetre és igual a θ, o el que és el mateix:

$$\mathrm{Pot}(\theta) = \mathrm{P}[\text{rebutjar } H_0 \mid \theta]$$

Es compleix que la $\mathrm{Pot}(\theta_0) = 1 - \beta$.

La **corba de potència** d'un contrast que es realitza a un nivell de significació α es defineix com la funció $f(\mathrm{Pot}(\theta))$ que per cada valor de θ proporciona la potència del contrast, és a dir, la probabilitat de rebutjar la H_0.

Complementàriament a la corba de potència existeix la **corba característica**. La corba característica d'un contrast que realitza a un nivell de significació α es defineix com la funció $1 - \mathrm{Pot}(\theta)$, o el que és el mateix:

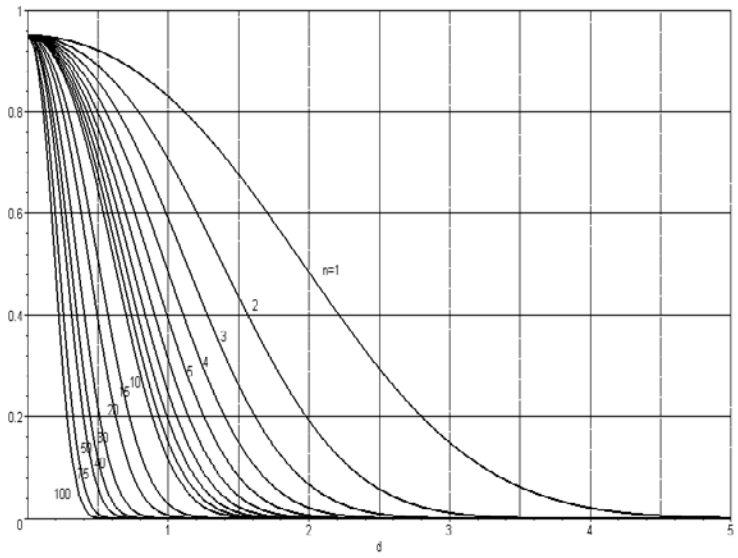
$$1 - \mathrm{Pot}(\theta) = \mathrm{P}[\text{acceptar } H_0 \mid \theta]$$

Quan el valor de θ faci certa la H_1, $1 - \mathrm{Pot}(\theta)$ esdevindrà la probabilitat de cometre un error de tipus II, és a dir, β.

Alhora de treballar amb les corbes característiques, primer de tot haurem d'escollir quin dels dos grups de corbes utilitzem:

- Si la variància és coneguda utilitzarem les corbes basades en una **distribució normal (σ^2 coneguda)**.
- Si la variància és desconeguda utilitzarem les corbes basades en una **distribució de t-Student (σ^2 desconeguda)**.

Escollit un dels dos grups, ens centrarem en una sola corba, la qual triarem a partir del tipus de contrast realitzat i del nivell de significació fixat. A més, tenint el valor de θ_0 i la mida n de la mostra, podrem obtenir de les corbes la probabilitat de cometre un error de tipus II, és a dir, β. Tanmateix podem calcular alguna dada que ens falti a partir del valor β:



Observem un exemple de corba característica. En aquest cas escollim la corresponent a un conjunt de dades mostral amb variància coneguda i pel contrast ens marquem a priori un nivell de significació $\alpha = 5\%$.

L'eix de les ordenades pertany a la probabilitat d'acceptar H_0. Les diferents corbes que trobem al mig del gràfic corresponen a les diferents mides que pot agafar un conjunt de dades mostrals.

L'eix de les abscisses pertany a un paràmetre d. El seu càlcul depèn de si la variància és coneguda o no. Veiem com es calcula:

Variància coneguda:

$$d = \frac{|\mu - \mu_0|}{\sigma}$$

Variància desconeguda:

$$d = \frac{|\mu - \mu_0|}{s}$$

Recordem que quan la variància és coneguda s'utilitza el valor poblacional i per això les corbes es basen en la distribució normal. En canvi, quan la variància és desconeguda, es calcula el seu valor a partir de les dades de la mostra, i per això les corbes es basen en la distribució de t-Student.

2. Contrast d'una esperança

A partir del contrast d'una esperança es vol comprovar si una v.a. numèrica X compleix una certa H_0 relativa a la seva $E\{X\}=\mu$. Com hem dit, les hipòtesis que podem plantejar són tres, i en aquest cas els contrasts agafen les següents formes:

Bilateral:	**Unilateral dreta:**	**Unilateral esquerra:**
$H_0 : E\{X\} = \mu_0$	$H_0 : E\{X\} \leq \mu_0$	$H_0 : E\{X\} \geq \mu_0$
$H_1 : E\{X\} \neq \mu_0$	$H_1 : E\{X\} > \mu_0$	$H_1 : E\{X\} < \mu_0$

El valor de μ_o serà aquell que voldrem contrastar amb el valor de l'esperança E{X} de la v.a. Com que no sabem el valor real de l'esperança s'extraurà de la v.a. una mostra aleatòria de n valors, a partir dels quals calcularem la seva mitjana $\bar{x}_n$ i que ens servirà d'estimador de l'esperança E{X}. En cas que no coneguem la variància var{X}, calcularem també la variància mostral s^2.

Si estandarditzem $\bar{x}_n$, segons si coneixem el valor de la variància o no, obtindrem un valor z o t que pertanyerà a una v.a. Z o t respectivament. Aquest valor representarà la diferència entre la $\bar{x}_n$ observada i la μ_0.

Si la **variància** és **coneguda**, farem servir el següent estadístic de contrast:

$$z = \frac{\bar{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad \text{pertany a} \quad Z = \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$$

Si la **variància** és **desconeguda**, farem servir el següent estadístic de contrast:

$$t = \frac{\bar{x}_n - \mu_0}{\frac{s}{\sqrt{n}}} \quad \text{pertany a} \quad t = \frac{\bar{X}_n - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Comparant el p-valor obtingut amb un nivell de significació α fixat a priori, extraurem una de les dues conclusions següents:

- Si **p-valor ≥ α** considerarem que la diferència entre μ_0 i μ, estimada a partir de la mitjana d'una mostra $\bar{x}_n$, és deguda a l'atzar, i per tant direm que "no tenim motius suficients per rebutjar la H_0".

- Si **p-valor < α** considerarem que la diferència entre μ_0 i μ, estimada a partir de la mitjana d'una mostra $\bar{x}_n$, no és deguda a l'atzar, sinó que la H_0 és falsa. Per això direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

Exemple: *En un laboratori les mostres de bacteris es guarden en un frigorífic, la temperatura T del qual està regulada per un termòstat graduat a 5°C. D'aquesta manera s'assegura que la temperatura del frigorífic segueix una normal N(5;0.5).*

Per tal de controlar el correcte funcionament del frigorífic es mesura sempre en el mateix punt la temperatura cada 6 hores. Les mesures obtingudes en les darreres 24 hores són 5.4°C, 4.9°C, 5.3°C i 5.8°C.

Es calcula la mitjana de la mostra i resulta $\bar{T}_4$ = 5.35°C. És clar que la mitjana no és igual al valor esperat de 5°C, però podem considerar la diferència entre aquestes dues dades és suficientment diferent com per dir que el termòstat no funciona correctament?

Per resoldre aquest problema ens plantegem un contrast paramètric bilateral d'una esperança amb un nivell de significació $\alpha = 5\%$ i amb les següents hipòtesis:

$$H_0 : \mu = 5°C = \mu_0$$
$$H_1 : \mu \neq 5°C$$

Com que sabem la desviació que segueix la v.a. de les temperatures, realitzem el següent càlcul:

$$z = \frac{\bar{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{5.35 - 5}{\frac{0.5}{\sqrt{4}}} = 1.4$$

Com que estem realitzant un contrast bilateral, busquem a les taules la probabilitat acumulada que deixa a l'esquerra el valor de –1.4 i la probabilitat que deixa a la dreta el valor de +1.4. Com que sabem que el valor de la probabilitat serà el mateix, busquem el que està directament tabulat i el multipliquem per dos.

Veiem a la taula estadística per una normal N(0,1) que el valor de –1.4 li correspon una probabilitat acumulada de 0.0808. Si multipliquem per dos aquest valor obtenim que el p-valor = 0.1616 = 16.16%.

Com que el p-valor és major al nivell de significació α, diem no tenim motius suficients per rebutjar la H_0 i per tant, acceptem la H_0.

Aquest resultat ens dóna a conèixer que:

- La probabilitat que, funcionant bé el termòstat, la mitjana de 4 temperatures surti 0.35°C o més per sobre o per sota de la temperatura esperada de 5°C és del 16.16%. Per tant, el càlcul realitzat també es podria haver fet de la següent manera:

$$p - valor = P[(\overline{T}_4 \leq 4.65) \text{ o } (\overline{T}_4 \geq 5.35)] =$$

$$p - valor = P\left[\left(Z \leq \frac{4.65 - 5}{0.5/\sqrt{4}}\right) \text{ o } \left(Z \geq \frac{5.35 - 5}{0.5/\sqrt{4}}\right)\right]$$

$$p - valor = P\{Z \leq -1.4\} + P\{Z \geq 1.4\} = 0.1616$$

- Vist el resultat, si decidíssim rebutjar la H_0 i per tant acceptéssim la H_1, tindríem una probabilitat d'un 16.16% d'equivocar-nos en la nostra decisió. Aquesta tant per cent representa la probabilitat de cometre un error de tipus I.
- Si decidim continuar considerant com a certa la H_0, també és possible que ens estem equivocant amb la nostra decisió si realment la $\mu \neq 5°C$. En aquest cas podríem cometre un error de tipus II. La seva probabilitat dependria de l'esperança real, que desconeixem.

3. Contrast d'igualtat de dues esperances

A partir del contrast de dues esperances es vol comprovar si dues v.a. numèriques X i Y compleixen una certa H_0. Com hem dit, les hipòtesis que podem plantejar són tres, i en aquest cas els contrasts agafen les següents formes:

Bilateral:	**Unilateral dreta:**	**Unilateral esquerra:**
$H_0 : \mu_X = \mu_Y$	$H_0 : \mu_X \leq \mu_Y$	$H_0 : \mu_X \geq \mu_Y$
$H_1 : \mu_X \neq \mu_Y$	$H_1 : \mu_X > \mu_Y$	$H_1 : \mu_X < \mu_Y$

D'altra manera podem escriure les mateixes hipòtesis com:

$H_0 : \mu_X - \mu_Y = 0$	$H_0 : \mu_X - \mu_Y \leq 0$	$H_0 : \mu_X - \mu_Y \geq 0$
$H_1 : \mu_X - \mu_Y \neq 0$	$H_1 : \mu_X - \mu_Y > 0$	$H_1 : \mu_X - \mu_Y < 0$

Per realitzar aquest contrast serà necessari extreure una mostra aleatòria de n unitats de la població vinculada a la v.a. X i una mostra m unitats a la població vinculada a la v.a. Y. De les dues mostres caldrà calcular el següent:

- La mitjana mostral. De la v.a. X obtindrem $\bar{x}_n$ mentre que de la v.a. Y obtindrem $\bar{y}_m$.
- En cas de no conèixer la variància de les v.a., caldrà calcular la variància mostral. En cas que sigui necessari, de la v.a. X obtindrem s_X^2 mentre que de la v.a. Y obtindrem s_Y^2.

Quan realitzem el contrast de dues esperances ens podem trobar en dues situacions:

- La v.a. X i la v.a. Y no tenen res en comú, és a dir, les mostres no són iguals per les dues v.a., no es realitzen les mesures en el mateix instant de temps ni en el mateix lloc. Quan totes aquestes condicions es compleixin, parlarem de **v.a. independents** que ens proporcionaran **mostres independents**.
- La v.a. X i la v.a. Y tenen com a mínim un fet dels esmentats anteriorment en comú. En aquest cas parlarem **d'una mostra de dades aparellades**.

3.1. Contrast d'igualtat de dues esperances a partir de dues mostres independents

Per realitzar el contrast en qüestió necessitem saber el valor de les mitjanes poblacionals μ_x i μ_y. Com que no les podem saber utilitzarem l'estimador de les mitjanes mostrals $\bar{x}_n$ i $\bar{y}_m$. Així podem escriure el següent:

$$E\{\overline{X}_n - \overline{Y}_m\} = \mu_X - \mu_Y$$

Si les **variàncies** són **conegudes**, utilitzarem el següent raonament:

Sabem que si les mides n i m són prou grans, aleshores l'estimador s'aproximarà a una distribució normal de les següents característiques:

$$\overline{X}_n - \overline{Y}_m \sim N\left(\mu_X - \mu_Y; \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)$$

I per tant, podrem estandarditzar-ho de la següent manera:

$$z_{obs} = \frac{(\bar{x}_n - \bar{y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \quad \text{pertany a} \quad Z = \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0;1)$$

El valor de z_{obs} representarà la diferència entre les dues mitjanes mostrals.

Si les **variàncies** són **desconegudes**, utilitzarem el següent raonament:

Calcularem el valor de s_X^2 i de s_Y^2. Encara que no sapiguem el valor de les variàncies poblacionals, en alguns casos podem considerar que seran aproximadament iguals, aleshores podem fer una sola estimació calculant la variància ponderada de les dues variàncies mostrals de la següent manera:

$$s_p^2 = \frac{(n-1)\cdot s_X^2 + (m-1)\cdot s_Y^2}{n+m-2}$$

A partir de la variància ponderada podem escriure el següent:

$$t_{obs} = \frac{(\bar{x}_n - \bar{y}_m) - (\mu_X - \mu_Y)}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \text{pertany a} \quad t = \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

El valor de t_{obs} representarà la diferència entre les dues mitjanes.

Si no podem suposar que les variàncies desconegudes són iguals, podem obtenir el següent

$$t_{obs} = \frac{(\bar{x}_n - \bar{y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \quad \text{pertany a} \quad t = \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \sim t_{n+m-2-\Delta}$$

on els graus de llibertat de la t-Student són menys que l'anterior cas.

Indiferentment amb quin dels casos ens trobem, sabent quin contrast estem realitzant i a partir de les taules estadístiques, podrem calcular el p-valor que li correspon a l'estadístic de contrast $\bar{x}_n - \bar{y}_m$ sobre la corresponent distribució.

Comparant la probabilitat obtinguda amb un nivell de significació α fixat a priori, extraurem una de les dues conclusions següents:

- Si **p-valor ≥ α** considerarem que la diferència entre μ_X i μ_Y, estimada a partir de l'estimador de contrast $\bar{x}_n - \bar{y}_m$, és deguda a l'atzar, i per tant direm que "no tenim motius suficients per rebutjar la H_0".

- Si **p-valor < α** considerarem que la diferència entre μ_X i μ_Y, estimada a partir de l'estimador de contrast $\bar{x}_n - \bar{y}_m$, no és deguda a l'atzar, sinó que la H_0 és falsa. Per això direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

Podem, també, construir un **interval de confiança** per a quantificar la diferència entre les mitjanes.

Aquest interval anirà des del valor $-z_{\alpha/2}$ de la normal N(0;1) que acumuli la probabilitat de $\alpha/2$ fins a un altre valor $z_{\alpha/2}$ de la mateixa normal que deixi a la seva dreta la mateixa probabilitat, anàlogament per a altres distribucions. Així, a l'interior d'aquest interval ens quedarà una probabilitat de 1 – α. Per aquesta raó anotarem **IC(**$\mu_X - \mu_Y$**,1 – α)**.

L'interval de confiança quan la **variància** de les dades sigui **coneguda** el calcularem de la següent manera:

$$IC(\mu_X - \mu_Y, 1-\alpha) = (\bar{x}_n - \bar{y}_m) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

L'interval de confiança quan la **variància** de les dades sigui **desconeguda** el calcularem de la següent manera:

$$IC(\mu_X - \mu_Y, 1-\alpha) = (\bar{x}_n - \bar{y}_m) \pm t_{n+m-2,\alpha/2} \cdot s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}$$

Si el valor plantejat a la H_0 per la diferència entre μ_X i μ_Y està inclòs a l'interval de confiança, direm que "no tenim motius suficients per rebutjar la H_0". Altrament direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1". L'interval ens informarà dels valors mínim i màxim d'aquesta diferència.

Exemple: *Es vol estudiar si hi ha o no diferència significativa entre la durada en dies de dos marques X i Y de bateries. Volem comparar l'esperança d'una variable (durada bateria) segons el factor binari (marca comercial).*

Amb aquest objectiu s'ha pres una mostra de 20 bateries de la marxa X i s'ha estudiat la seva durada. El mateix s'ha fet amb 25 bateries de la marca Y. Per tal de poder comparar els resultats, totes les bateries han estat sotmeses a les mateixes condicions de treball.

Les durades en dies són:

Marca X	*58*	*71*	*41*	*76*	*68*	*76*	*64*	*48*	*72*	*60*
	59	*68*	*64*	*96*	*84*	*62*	*65*	*81*	*70*	*77*
Marca Y	*48*	*63*	*57*	*50*	*63*	*76*	*62*	*58*	*67*	*73*
	74	*69*	*57*	*42*	*51*	*58*	*63*	*56*	*74*	*70*
	64	*54*	*55*	*67*	*55*					

Per saber si podem considerar la durada de les dues marques de bateries iguals, realitzem un contrast d'igualtat de dues esperances amb un nivell de significació del 5% i les següents hipòtesis:

$$H_0 : \mu_X - \mu_Y = 0$$
$$H_1 : \mu_X - \mu_Y \neq 0$$

Per realitzar el contrast el primer que fem és calcular els estadístics de la durada de les bateries de la marca X i Y.

Marca X:	n = 20	$\bar{x} = 68$ dies	$s_X = 12.35$ dies
Marca Y:	m = 25	$\bar{y} = 61.04$ dies	$s_Y = 8.88$ dies

Si estimem la diferència de $\mu_X - \mu_Y$ a partir de $\bar{x} - \bar{y}$, veiem que aquesta és de 6.96 dies. Tot i que esperem que la diferència entre mitjanes doni 0, donat que estem treballant amb dades mostrals i no amb tota la població, hem de valorar si aquesta diferència és prou significativament diferent de 0.

Com que no coneixem les variàncies de X i Y, hem calculat les variàncies de cada una de les mostres, i suposant que les variàncies poblacionals són iguals, estimem la variància comuna o variància ponderada.

$$s_p^2 = \frac{(n-1)\cdot s_X^2 + (m-1)\cdot s_Y^2}{n+m-2} = \frac{(20-1)\cdot 12.35^2 + (25-1)\cdot 8.88^2}{20+25-2} = 111.41$$

Per tant, la desviació estàndard ponderada l'estimem com:

$$s_p = \sqrt{111.41} = 10.56 \text{ dies}$$

Tenint aquesta dada, podem calcular l'estadístic de contrast corresponent:

$$t_{obs} = \frac{(\bar{x}_n - \bar{y}_m) - (\mu_X - \mu_Y)}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{(68 - 61.04) - (0)}{10.56 \cdot \sqrt{\frac{1}{20} + \frac{1}{25}}} = 2.198$$

Busquem aquest valor a les taules t-Student per 20 + 25 – 2 = 43 graus de llibertat. Tot i no estar tabulat podem veure com el seu p-valor és inferior al nivell de significació $\alpha = 5\%$ i superior a $\alpha = 1\%$ (amb un paquet estadístic podem trobar que val 0.0334). Per això diem que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1: les mitjanes de la durada de les dues marques X i Y són diferents".

A aquesta mateixa conclusió hi podem arribar a partir del càlcul de l'interval de confiança. Com que a priori ens hem marcat un nivell de significació del 5%, buscarem l'interval de confiança IC(1 – α) = IC(95%).

Busquem a les taules t-Student el valor que correspon a $t_{n+m+2,\alpha/2} = t_{43,0.025}$. Interpolant els dos valors tabulats més pròxims obtenim que val 2.01785.

$$IC(\mu_X - \mu_Y, 1-\alpha) = (\bar{x}_n - \bar{y}_m) \pm t_{\alpha/2} \cdot s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}$$

$$IC(\mu_X - \mu_Y, 95\%) = (68 - 61.04) \pm 2.01785 \cdot 10.56 \cdot \sqrt{\frac{1}{20} + \frac{1}{25}}$$

$$IC(\mu_X - \mu_Y, 95\%) = (0.567\ , 13.35)$$

Veiem que l'interval no inclou el valor zero, per tant arribem a la mateixa conclusió que per l'altre mètode.

D'aquest resultat també podem dir que la durada mitjana de les bateries de la marca X és major que la de la Y. A més a més, s'estima que la diferència entre ambdues mitjanes està situada a un nivell de confiança del 95% entre 0.567 i 13.35 dies.

3.2. Contrast d'igualtat de dues esperances a partir d'una mostra de dades aparellades

Quan tenim dades que es relacionen entre elles (moment en realitzar la mesura, lloc de la mesura...), dissenyarem el contrast d'igualtat de dues esperances a partir de mostres aparellades. Això significa que tindrem dos conjunts de dades amb el mateix número de valors, que un valor de la mostra X estarà relacionat amb un valor de la mostra Y, quedant així tots els valors relacionats entre ells.

Tot aquest doble conjunt de dades l'estudiarem a partir d'una **variable de diferència D**. Aquesta l'obtindrem restant un valor de la mostra X amb el valor que tingui relació de la mostra Y:

$$d_i = x_i - y_i$$

Així aconseguirem un conjunt de dades D del mateix número de mostres que la mostra X i la mostra Y.

D'aquesta variable de diferència en podrem calcular la mitjana $\bar{d}$, que serà l'estimador de la mitjana poblacional μ_D. Aquest estimador serà el que ens permetrà realitzar els següents contrasts:

Bilateral:	**Unilateral dreta:**	**Unilateral esquerra:**
$H_0 : \mu_X - \mu_Y = 0$	$H_0 : \mu_X - \mu_Y \leq 0$	$H_0 : \mu_X - \mu_Y \geq 0$
$H_1 : \mu_X - \mu_Y \neq 0$	$H_1 : \mu_X - \mu_Y > 0$	$H_1 : \mu_X - \mu_Y < 0$

D'altra manera podem escriure les mateixes hipòtesis com:

$H_0 : \mu_D = 0$	$H_0 : \mu_D \leq 0$	$H_0 : \mu_D \geq 0$
$H_1 : \mu_D \neq 0$	$H_1 : \mu_D > 0$	$H_1 : \mu_D < 0$

Com és de suposar, mai sabrem la variància de la variable de diferència D. Per això sigui quin sigui el contrast, aquest el realitzarem a partir de la distribució de t-Student. El càlcul a fer serà el següent:

$$t_{obs} = \frac{\bar{d} - \mu_D}{\frac{s_D}{\sqrt{n}}} \quad \text{pertany a} \quad t = \frac{\bar{d} - \mu_D}{\frac{s_D}{\sqrt{n}}} \sim t_{n-1}$$

Buscant la probabilitat que acumula el valor t_{n-1} a les taules de la distribució t-Student i comparant el p-valor obtingut amb el nivell de significació α marcat a priori, traurem de l'estudi una de les següents conclusions:

- Si **p-valor ≥ α** considerarem que la diferència entre μ_X i μ_Y, estimada a partir de l'estimador de contrast $\bar{d}$, és deguda a l'atzar, i per tant direm que "no tenim motius suficients per rebutjar la H_0".

- Si **p-valor < α** considerarem que la diferència entre μ_X i μ_Y, estimada a partir de l'estimador de contrast $\bar{d}$, no és deguda a l'atzar, sinó que la H_0 és falsa. Per això direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

Tanmateix, enlloc de calcular el p-valor per determinar la certesa de la H_0, podem construir un **interval de confiança**.

Aquest interval anirà des del valor $-t_{\alpha/2}$ de la distribució de t-Student t_{n-1} que acumuli la probabilitat de $\alpha/2$ fins a un altre valor $t_{\alpha/2}$ de la mateixa distribució que deixi a la seva dreta la mateixa probabilitat. Així, a l'interior d'aquest interval ens quedarà una probabilitat de 1 – α. Per aquesta raó anotarem **$IC(\mu_X - \mu_Y, 1 - \alpha)$** i el calcularem de la següent manera:

$$IC(\mu_X - \mu_Y, 1 - \alpha) = \bar{d} \pm t_{n-1,\alpha/2} \cdot \frac{s_D}{\sqrt{n}}$$

Si el valor plantejat a la H_0 per la diferència entre μ_X i μ_Y està inclòs a l'interval de confiança, direm que "no tenim motius suficients per rebutjar la H_0". Altrament direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

Exemple: *Es vol estudiar l'efectivitat d'unes mesures mediambientals per disminuir la contaminació per nitrats d'un determinat col·lector d'una indústria. Per això es va agafar una mostra d'aigua en cada un dels 20 punts del col·lector just abans d'aplicar les mesures mediambientals. A dia d'avui, 5 mesos després de l'aplicació de les mesures que es volen estudiar, s'agafa una mostra d'aigua en els mateixos 20 punts del col·lector. De cada una de les mostres s'estudia la concentració de nitrats en mg/l. Les dades obtingudes són les següents:*

Punt de mesura del col·lector (i)	***Concentració de nitrats (mg/l) abans d'aplicar les mesures mediambientals***	***Concentració de nitrats (mg/l) després d'aplicar les mesures mediambientals***
1	*50.4*	*51.9*
2	*50.2*	*46.9*
3	*53*	*53.6*
4	*49.8*	*49.1*
5	*52.1*	*50*
6	*50.8*	*46.3*
7	*46.1*	*46.3*
8	*48*	*47.2*
9	*48.2*	*45.6*
10	*50.6*	*50*
11	*52.1*	*50.8*
12	*53*	*49.7*
13	*51.3*	*48.9*
14	*50.4*	*48.5*
15	*49.7*	*47.7*
16	*50.3*	*48.6*
17	*50.2*	*47.1*
18	*51.8*	*51.9*
19	*52.3*	*53.4*
20	*53.2*	*50.1*

A partir d'aquestes dades es vol estudiar si efectivament les mesures protectores han fet disminuir la concentració de nitrats en el col·lector.

Primer de tot observem que les dues mesures de nitrats s'han realitzat en els mateixos punts del col·lector, és a dir, cada valor de la primera mostra té relació amb un valor de la segona. Per això veiem que haurem de fer un contrast d'igualtat de dues esperances a partir d'una mostra de dades aparellades.

Com que volem estudiar si les mesures mediambientals han fet reduir la concentració de nitrats, plantejarem el següent contrast d'hipòtesis amb un nivell de significació $\alpha = 0.05$:

$$H_0 : \mu_D \geq 0$$
$$H_1 : \mu_D < 0$$

Com que la mitjana de la variable de diferència D no la coneixem, l'haurem de calcular, igual que la variància d'aquesta distribució. Però abans d'això, haurem de

calcular els valors que pren la variable D depenent del punt de mesura del col·lector (i):

$$d_i = \text{CONCENTRACIÓ DESPRÉS}_i - \text{CONCENTRACIÓ ABANS}_i$$

Obtenim el següent conjunt de dades:

Punt de mesura del col·lector (i)	**Diferència concentració de nitrats (mg/l) (concentració després – concentració abans)**
1	1.5
2	–3.3
3	0.6
4	–0.7
5	–2.1
6	–4.5
7	0.2
8	–0.8
9	–2.6
10	–0.6
11	–1.3
12	–3.3
13	–2.4
14	–1.9
15	–2
16	–1.7
17	–3.1
18	0.1
19	1.1
20	–3.1

Un cop tenim les dades de la v.a. D, calculem els dos estadístics que necessitarem.

$$\bar{d} = -1.495 \qquad s_D = 1.639 \qquad n = 20$$

Veiem que la mitjana de la v.a. D surt negativa, i per això sembla que les mesures mediambientals han fet disminuir la concentració de nitrats. Tot i això des d'un punt estadístic cal comprovar si aquesta diferència és o no significativament menor que 0. Per això realitzem el següent càlcul:

$$t_{obs} = \frac{\bar{d} - \mu_D}{\frac{s_D}{\sqrt{n}}} = \frac{-1.495 - 0}{\frac{1.639}{\sqrt{20}}} = -4.08$$

Busquem aquest valor a les taules t-Student per 20 – 1 = 19 graus de llibertat. Tot i no estar tabulat podem veure com el seu p-valor és inferior al nivell de significació $\alpha = 0.5\%$. Per això diem que "tenim motius suficients per rebutjar la H_0", és a dir, les mesures mediambientals han fet disminuir en mitjana la concentració de nitrats en el col·lector.

A aquesta mateixa conclusió hi podem arribar a partir del càlcul de l'interval de confiança. Com que a priori ens hem marcat un nivell de significació del 5%, buscarem l'interval de confiança $IC(\mu_X - \mu_Y, 1 - \alpha) = IC(\mu_X - \mu_Y, 95\%)$.

Busquem a les taules t-Student el valor que correspon a $t_{n-1,\alpha/2} = t_{19,0.025} = 2.093$.

$$IC(\mu_X - \mu_Y, 1 - \alpha) = \bar{d} \pm t_{n-1,\alpha/2} \cdot \frac{s_D}{\sqrt{n}}$$

$$IC(\mu_X - \mu_Y, 95\%) = (-1.495) \pm 2.093 \cdot \frac{1.639}{\sqrt{20}}$$

$$IC(\mu_X - \mu_Y, 95\%) = [-2.262\ , -0.728]$$

Veiem que l'interval no inclou cap valor positiu, per tant arribem a la mateixa conclusió que per l'altre mètode.

Per tant, amb una confiança del 95% podem afirmar que, en mitjana, les mesures protectores han fet disminuir la concentració de nitrats del col·lector en un valor comprès entre 0.728 i 2.262 mg/l.

4. ANOVA

L'ANOVA (abreviació de *Analysis of Variance*) o anàlisi de la variància és una tècnica d'inferència estadística per comparar les mitjanes de més de dues poblacions a partir de les seves dades mostrals. Aquesta tècnica és molt útil perquè permet comparar més de dues v.a. alhora, cosa que amb els contrasts explicats fins ara únicament podíem comparar les v.a. de dues en dues.

La terminologia que utilitzarem d'ara endavant quan utilitzem aquest mètode és la següent:

- A la variable aleatòria numèrica que segueixi cada població li direm **variable de resposta** o $\mathbf{Y_i}$.

- A la variable categòrica que descrigui cada població li direm **variable explicativa**, **factor** o **X**.

- Als diferents valors que agafi la variable categòrica li direm **tractaments** o **i.** Ens referirem a cada una de les poblacions a contrastar com $\mathbf{X_i}$.

- Al nombre total de tractaments li assignarem la lletra **k.**

Havent definit la terminologia que utilitza l'ANOVA, podem explicar de forma genèrica que les hipòtesis que sempre ens plantejarem per aquest mètode seran, definint un nivell de significació α a priori, les següents:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$
$$H_1 : \text{les mitjanes } \mu_1, \mu_2, \mu_3, \ldots, \mu_k \text{ no són totes iguals}$$

Abans d'explicar el procediment a seguir, cal tenir en compte que per poder aplicar aquest mètode cal que es compleixin les següents condicions:

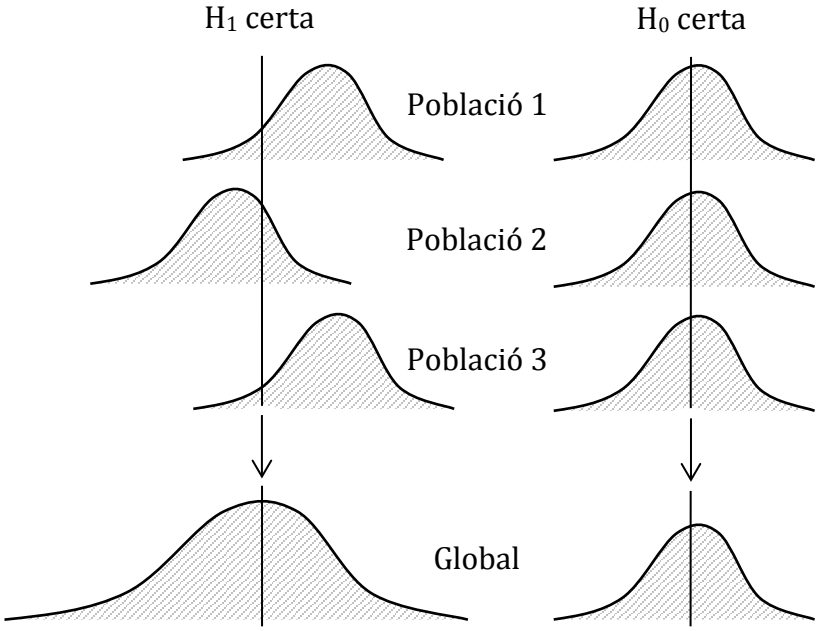
- La distribució de les dades té una distribució normal per a cada tractament del factor X_i.

- La variància ha de ser la mateixa per a cada un dels tractaments, és a dir, $Y_i \sim N(\mu_i; \sigma)$ per $i = 1,2,3, \ldots, k$. Equivalentment podem dir que $Y_i = \mu_i + \varepsilon$ amb $\varepsilon \sim N(0; \sigma)$ per $i = 1,2,3, \ldots, k$. D'aquesta equivalència s'extreu que la variable ε ha de ser també normal. Aquesta igualtat de variàncies s'anomena **homoscedasticitat**.

- Les observacions per a cada tractament han de ser aleatòries i independents. Conseqüentment a això, els residus de les diferents Y_i hauran de ser independents entre ells.

La importància d'aquestes condicions és deguda a que l'ANOVA, tal i com diu el seu nom, analitza les variàncies dels diferents tractaments per determinar quina de les dues hipòtesis del contrast esmentat anteriorment és la certa.

Si el mètode ens indica que la hipòtesi certa és la H_0, vol dir que les mitjanes de tots els tractaments són iguals i que la variància global de totes les dades serà la mateixa que la variància de les dades de cada tractament. Pel contrari, si ens indica que la hipòtesi certa és la H_1, significarà que no totes les mitjanes són iguals i que la variància global de les dades serà més gran que la variància de les dades dels tractaments. Això es pot reflectir de la següent manera:



Veiem que quan la H_0 és la hipòtesis certa, la variància poblacional global és la mateixa que en les altres poblacions. En canvi, quan es rebutja la H_0 i es considera la H_1 com a certa, podem observar que la variància poblacional global és més gran que la variància poblacional dels tractaments.

Quan apliquem ANOVA normalment partirem com a mínim de les dades de tres tractaments. De forma genèrica ho podem simbolitzar com:

	Variable explicativa o factor (X)			
Tractaments	**1**	**2**	**…**	**k**
Variable de resposta (Y)	y_{11}	y_{21}	…	y_{k1}
	y_{12}	y_{22}	…	y_{k2}
	…	…	…	…
	y_{1n_1}	y_{2n_2}	…	y_{kn_k}

Cada observació agafarà el format y_{ij}, el que significa observació número j de la variable Y_i. D'aquesta informació en podem calcular les mitjanes poblacions i la global de la següent manera:

- Mitjana poblacional: $$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

- Mitjana global: $\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$ on $n = \sum_{i=1}^{k} n_i$

Cada una de les dades mostrals o observacions es trobarà a una certa distància de la mitjana de la població a la que pertany. Això ho podem escriure com:

$$y_{ij} = \mu_i + e_{ij}$$

El terme **e** que anomenarem **residu** és un valor observat de la variable ε. Recordem que $\varepsilon \sim N(0; \sigma)$.

L'objectiu d'aquest mètode és obtenir un estadístic de contrast que ens ajudi a decantar-nos per una hipòtesi o altra. Aquest estadístic s'obté a partir d'un seguit d'operacions relacionades entre elles. Per fer-ho de forma més entenedora, quan realitzem aquest mètode sempre recollirem els resultats de cada pas en la **taula d'anàlisi de la variància** que mostrem a continuació:

	Graus de llibertat (DF)	**Suma dels quadrats (SS)**	**Mitjana dels quadrats (MS)**	**Estadístic de contrast (F)**
Entre tractaments	k –1	SSTract	MSTract	F_{obs}
Dins tractaments	n – k	SSErr	MSErr	
Total	n – 1	SSTot = SSTract + SSErr		

Com hem comentat anteriorment, l'anàlisi de la variància ANOVA es basa en l'anàlisi de la variabilitat de les dades. Per això, després d'haver calculat els **graus de llibertat (DF),** calculem la **suma dels quadrats (SS).**

Definim **SSTract** com la suma de les diferències al quadrat entre les mitjanes dels diferents tractaments amb la mitjana global. Si totes les mitjanes fossin similars, llavors SSTract ≈ 0.

$$SSTract = \sum_{i=1}^{k} n_i \cdot (\bar{y}_i - \bar{y})^2$$

Definim **SSErr** com la suma de les diferències al quadrat entre cada dada i la mitjana del tractament a que correspon:

$$SSErr = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{k} (n_i - 1) \cdot s_i^2$$

Finalment, definim **SSTot** com la suma de les diferències al quadrat entre cada dada i la mitjana global:

$$\text{SSTot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2 = (n-1)\cdot s_Y^2$$

Tenint en compte que SSTot = SSTract + SSErr, realitzant els dos càlculs més fàcils (SSErr i SSTot) i restant-los entre ells de forma adequada, obtindrem fàcilment SSTract.

Un cop omplertes les primeres dues columnes de la taula, procedim a calcular la tercera. Aquesta l'emplenarem amb les **mitjanes dels quadrats (MS)**, que ho calcularem a partir de les dades de les dues primeres columnes.

Definim **MSTract** com la variabilitat entre tractaments o, dit d'una altra manera, la variabilitat poblacional global:

$$\text{MSTract} = \frac{\sum_{i=1}^{k} n_i \cdot (\bar{y}_i - \bar{y})^2}{k-1} = \frac{\text{SSTract}}{k-1}$$

Definim **MSErr** com la variabilitat dins dels tractaments. Com que per aplicar ANOVA hem suposat que les variàncies dels diferents tractaments eren aproximadament iguals, MSErr ens donarà la mitjana de les variàncies de cada tractament:

$$\text{MSErr} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2}{n-k} = \frac{\text{SSError}}{n-k}$$

Recordem que anteriorment hem dit que quan la H_0 sigui certa, la variància dels diferents tractaments amb la variància poblacional global serà aproximadament igual. En canvi, si la variància poblacional global és major a la dels tractaments, es rebutjarà la H_0 i acceptarem la H_1.

La variància poblacional global ens vindrà representada per MSTract, mentre que la variància dels tractaments ens l'indicarà MSErr. Aquests dos els compararem mitjançant el següent **estadístic de contrast**:

$$F = \frac{\text{MSTract}}{\text{MSErr}}$$

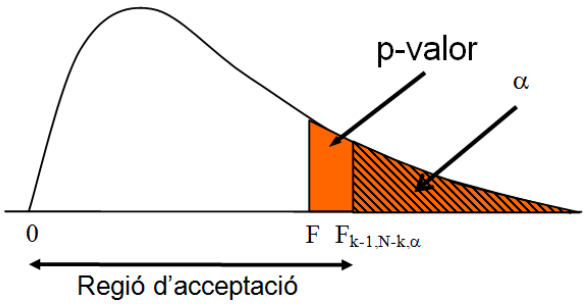
Aquest estadístic de contrast segueix una distribució F de Fisher amb $\nu_1 = k - 1$ i $\nu_2 = n - k$ graus de llibertat. Buscant la probabilitat que acumula el valor $F_{\nu_1,\nu_2} = F_{k-1,n-k}$ a les taules de la distribució de Fisher i comparant el p-valor obtingut amb el nivell de significació α marcat a priori, finalitzarem l'estudi amb una de les següents conclusions:

- Si **p-valor ≥ α** considerarem que la diferència entre els estadístics de les diferents poblacions és deguda a l'atzar, i per tant direm que "no tenim motius suficients per rebutjar la H_0".

- Si **p-valor < α** considerarem que la diferència entre els estadístics de les poblacions no és deguda a l'atzar, sinó que la H_0 és falsa. Per això direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

Tanmateix, enlloc de calcular el p-valor per determinar la certesa de la H_0, ens podem regir per una **regió d'acceptació**. Aquesta regió no serà més que un interval que anirà des de 0 fins al valor F que deixi a la dreta una probabilitat igual al nivell de significació α.

$$\text{Regió d'acceptació} = \left[0, F_{k-1,n-k,\alpha}\right]$$

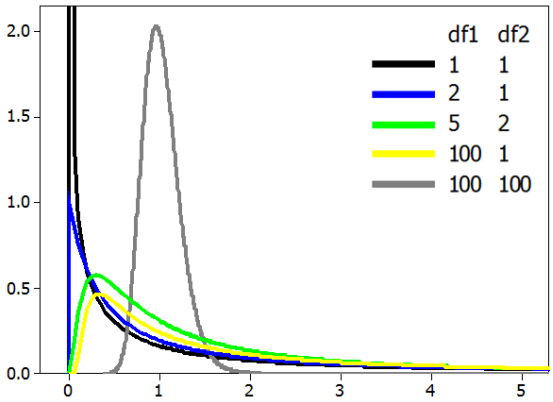


Així doncs, veiem que si el valor de l'estadístic de contrast F es troba inclòs dins la regió d'acceptació, llavors el p-valor serà major al nivell d'acceptació i direm que "no tenim motius suficients per rebutjar la H_0". Altrament direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

4.1. Distribució de Fisher-Snedecor

La distribució de Fisher és una distribució de probabilitat contínua que partir de dos paràmetres anomenats ν_1 i ν_2 (graus de llibertat del numerador i del denominador respectivament). Es modela la probabilitat d'obtenir un valor a l'atzar d'una v.a. S'anota com **$X \sim F(\nu_1; \nu_2)$**, el que significa que la v.a. X segueix una distribució de Fisher F amb paràmetres ν_1 i ν_2.

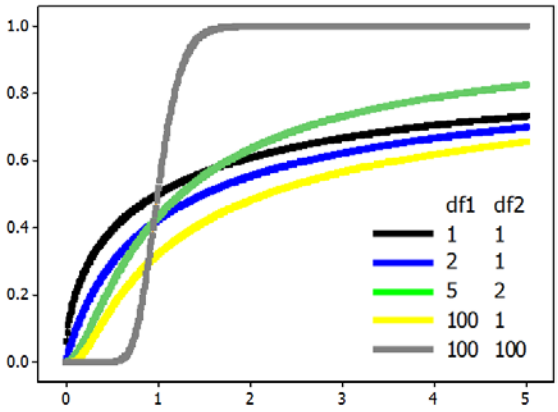
La **funció de densitat** d'aquesta distribució té la següent forma i es defineix per la fórmula que l'acompanya.



$$f(x) = \frac{\sqrt{\frac{(\nu_1 \cdot x) \cdot \nu_1 \cdot \nu_2^{\nu_2}}{(\nu_1 \cdot x + \nu_2) \cdot \nu_1 + \nu_2}}}{x \cdot \beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}$$

Aquesta equació és vàlida per valors de x majors a 0 i per nombres naturals dels paràmetres ν_1 i ν_2.

La **funció de distribució** s'obté d'integrar la funció de densitat anterior. Degut a la dificultat de calcular aquesta funció, tenim unes taules estadístiques on tenim tabulat el valor que assoleix F per varis conjunts de paràmetres i probabilitats no acumulades (1-probabilitat acumulada fins el punt F de la distribució).



$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

La funció de distribució també es pot representar a partir de gràfica de la funció de densitat, marcant l'àrea de -∞ al valor desitjat, el que representarà la probabilitat acumulada fins aquest valor.

Per tant, els **estadístics** d'una distribució F seran els següents:

$$E(X) = \frac{\nu_2}{\nu_2 - 2} \qquad \text{on} \quad \nu_2 > 2$$

$$\text{var}(X) = \frac{2 \cdot \nu_2^2 \cdot (\nu_1 + \nu_2 - 2)}{\nu_1 \cdot (\nu_2 - 2)^2 \cdot (\nu_2 - 4)} \qquad \text{on} \quad \nu_2 > 4$$

Exemple: *En una empresa hi ha tres línies de producció, suposadament iguals, d'un mateix producte. Interessa saber si hi ha diferències de productivitat entre les línies. Per això, durant cada hora al llarg de 15 hores s'ha mesurat la productivitat (unitats produïdes/hora) en les tres línies de producció. La mesura en cada línia s'ha fet en hores diferents. Les dades obtingudes són les següents:*

LÍNIA		
1	**2**	**3**
75.9973	75.9446	78.6007
75.2505	76.0937	78.7304
75.8129	75.474	79.1946
75.3044	74.9464	76.9601
76.5921	74.8067	76.0524
74.3179	75.4593	78.7872
76.0692	79.0766	78.6391

75.6494	77.679	78.7956
75.5623	77.3934	78.5298
75.2429	78.7216	77.9155
75.6873	75.5584	80.7565
75.3064	74.4243	76.6777
73.9138	75.3141	77.4621
72.189	73.9436	76.7714
74.8154	75.3661	77.6008

Primer de tot observem en quin tipus de problema ens trobem. Observem que hem de comparar més de dues poblacions, el que ens dóna una pista que haurem d'aplicar ANOVA. Abans però, hem de corroborar si el tipus de problema compleix les condicions necessàries per utilitzar aquest mètode.

A l'enunciat se'ns deixa clar que es creu que la producció és teòricament igual, pel que considerarem que la variància poblacional també ho és. Per altra banda també se'ns diu que les dades han sigut obtingudes a partir d'observacions independents al no realitzar-se a la mateixa hora. Així doncs, podem aplicar ANOVA.

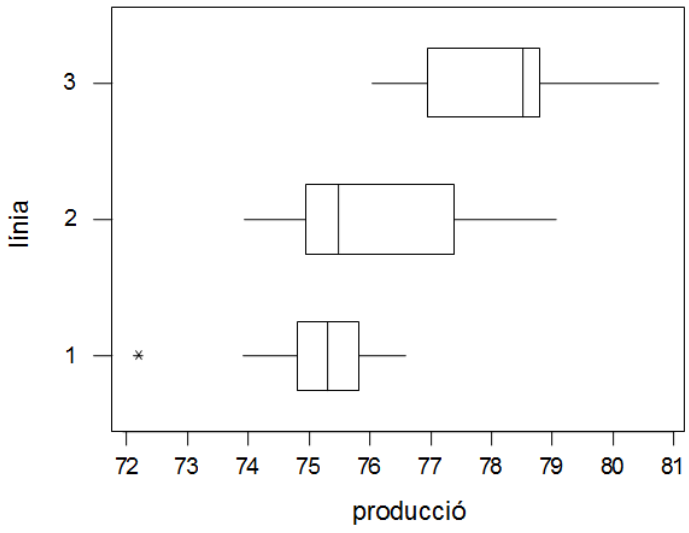
Llegint l'enunciat veiem que la variable explicativa és la línia de producció mentre la variable de resposta és la productivitat del procés. Veiem que el nombre de tractaments a realitzar el contrast és de tres.

Abans d'aplicar el mètode ANOVA, calculem la mitjana de cada línia i realitzem un estudi gràfic per veure si cal procedir. Si fos molt evident que les tres línies de producció no són iguales, no caldria fet tot el càlcul de l'anàlisi de variància.

Mitjana línia 1: $\bar{y}_1 = 75.181$

Mitjana línia 2: $\bar{y}_2 = 76.013$

Mitjana línia 3: $\bar{y}_3 = 78.098$



Observem que la diferència de mitjanes no és molt gran, tot i que no podem afirmar si aquestes tres distribucions són significativament iguals o diferents. Així doncs plantejarem el contrast d'hipòtesis següent amb un nivell de significació α del 5%.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \text{les mitjanes } \mu_1, \mu_2, \mu_3 \text{ no són totes iguals}$$

Mitjançant mètodes que explicarem més endavant al tema 6, es calcula la normalitat dels residus e i la seva homoscedasticitat. Recordem que $e_{ij} = y_{ij} - \bar{y}_i$, i que les dues condicions esmentades s'han de complir per cada tractament. En aquest cas, els dos contrastos de normalitat i d'homoscedasticitat ens porten a acceptar que es compleixen aquests supòsits, continuem desenvolupant l'anàlisi ANOVA.

Realitzant els càlculs explicats a l'apartat de teoria, s'omple la taula de la variància i se n'estudia l'estadístic de contrast.

	Graus de llibertat (DF)	**Suma dels quadrats (SS)**	**Mitjana dels quadrats (MS)**	**Estadístic de contrast (F)**
Entre tractaments	2	67.76	33.88	20.69
Dins tractaments	42	68.77	1.64	
Total	44	136.53		

Buscant a les taules el valor que deixa una probabilitat del 5% a la dreta en una distribució de Fisher $F_{k-1,n-k} = F_{2,42}$, trobem que correspon aproximadament a F = 3.23. Com que l'estadístic de contrast obtingut és major a 3.23, la probabilitat que deixarà el valor de F = 20.69 a la seva dreta serà menor del 5%. Així doncs direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

5. Anàlisi de la bondat d'ajust

Realitzar un contrast de bondat d'ajust significa comparar la distribució que segueixen les dades d'una mostra amb una llei de probabilitat. D'aquesta manera podem conèixer si una v.a. X segueix una determinada distribució o no.

Analitzar la bondat d'ajust consta de dues etapes que són l'**etapa descriptiva** i l'**etapa confirmatòria**. Finalment, a la vista dels resultats l'investigador de l'anàlisi ha de decidir si l'ajust a la llei teòrica de les dades és bo o no.

5.1. Etapa descriptiva. Diagrames Q-Q

Per fer una anàlisi descriptiva hem de dibuixar l'**histograma** corresponent a les dades, calcular els **estadístics** principals i realitzar els **diagrames Quantil-Quantil (Q-Q)** que explicarem més endavant.

Per una banda, la forma de l'**histograma** ens pot suggerir quina llei de distribució es podria ajustar millor a les dades. Per exemple, que l'histograma presenti forma de campana simètrica pot significar que les dades realment segueixin una distribució normal. Per altra banda, els valors dels **estadístics** calculats ajuden a seleccionar els paràmetres de la distribució seleccionada anteriorment. Per exemple, si fent l'histograma hem vist que les dades poden seguir una distribució normal, calculant els estadístics mitjana $\bar{x}$ i desviació estàndard s seleccionem els paràmetres μ i σ de la llei Normal.

Els **diagrames Q-Q** són un mètode gràfic per jutjar la bondat d'ajust a un model de distribució teòrica d'una distribució de dades experimentals. Consisteixen en un diagrama de punts on a l'eix d'abscisses s'hi col·loquen les dades experimentals, mentre que a l'eix d'ordenades s'hi col·loquen els valors teòrics. Aquests últims és troben a partir de les taules o fórmules de la distribució corresponent, tenint en compte que cada valor acumula la mateixa probabilitat que la dada experimental que li pertoca.

Per construir un diagrama Q-Q fa falta omplir una taula amb les cinc columnes que es descriuen a continuació:

- **Dades empíriques (C1):** col·loquem totes les observacions de la v.a. X ordenades de menor a major.
- **Rang o número d'ordre (C2):** anotem la posició que li pertoca a cada observació. En el cas que ens trobem amb dos o més valors iguals, la posició que anotarem serà la mateixa per cada observació i correspondrà a la mitjana de les seves posicions.
- **Percentatges acumulats (C3):** calculem la probabilitat que acumula cada una de les observacions mitjançant la fórmula $C3 = (C2 - 0.5)/N$.
- **Dades teòriques estandarditzades (C4):** per cada probabilitat acumulada, busquem a les taules o mitjançant fórmules el valor teòric estandarditzat que acumula la mateixa probabilitat.
- **Dades teòriques (C5):** calculem quin valor teòric correspon a cada valor teòric estandarditzat, és a dir, realitzem el procés contrari a l'estandardització.

Exemple: *Comprova si les següents dades s'ajusten a una llei Normal mitjançant un diagrama Q-Q.*

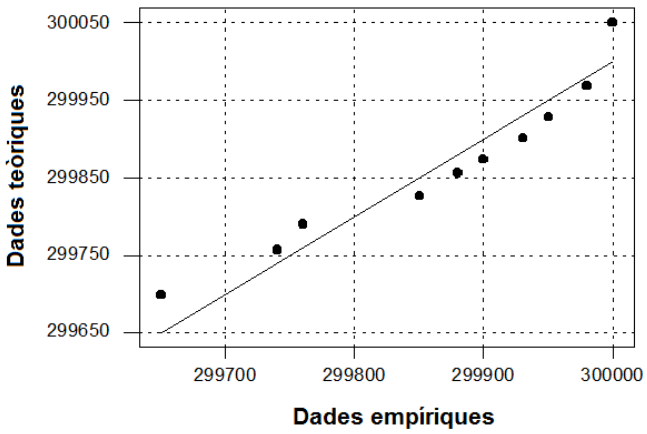
299650, 299740, 299760, 299850, 299850, 299880, 299900, 299930, 299930, 299950, 299980, 299980, 299980, 300000, 300000

Com que partim que les dades segueixen una distribució normal, calculant els estadístics mitjana $\bar{x}$ i desviació estàndard s estimem els paràmetres μ i σ d'aquesta distribució:

$$\hat{\mu} = \bar{x} = 299892$$
$$\hat{\sigma} = s = 105.438$$

Per tant, com que les dades teòriques han de seguir una N(299892;105.438), podem omplir la següent taula i dibuixar el diagrama Q-Q corresponent:

Dades empíriques	Rang	Percentatges acumulats	Dades teòriques N(0;1)	Dades teòriques N(299892;105.438)
299650	1.0	0.033333	−1.83391	299699
299740	2.0	0.100000	−1.28155	299757
299760	3.0	0.166667	−0.96742	299790
299850	4.5	0.266667	−0.62293	299826
299850	4.5	0.266667	−0.62293	299826
299880	6.0	0.366667	−0.34069	299856
299900	7.0	0.433333	−0.16789	299874
299930	8.5	0.533333	0.08365	299901
299930	8.5	0.533333	0.08365	299901
299950	10.0	0.633333	0.34069	299928
299980	12.0	0.766667	0.72791	299969
299980	12.0	0.766667	0.72791	299969
299980	12.0	0.766667	0.72791	299969
300000	14.5	0.933333	1.50109	300050
300000	14.5	0.933333	1.50109	300050



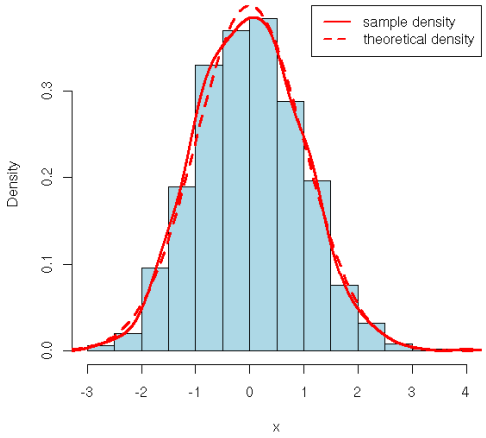
Tot i que per confirmar que les dades s'ajusten a un model un gràfic no és suficient, sí que el gràfic ens pot donar indicacions que el model pot ser aproximadament apropiat. En aquest cas, gràficament les dades s'ajusten prou bé al model Normal.

5.2. Etapa confirmatòria. Contrast χ^2

L'anàlisi confirmatòria es pot realitzar a través de diferents contrastos com són el **contrast de bondat d'ajust X_i-Quadrat (χ^2)**, el **contrast de Kolmogorov-**

Smirnov (K-S) i/o el **contrast d'Anderson-Darling (A-D)**. No obstant, en aquest apartat només explicarem el contrast χ^2.

El **contrast χ^2** ens permet comparar la distribució de freqüències absolutes d'una mostra amb la funció de densitat d'una distribució teòrica. D'aquesta manera podem valorar, a partir d'unes dades experimentals, si una determinada llei de probabilitat s'ajusta bé a la distribució de les dades.



Tenint en compte que partim d'una v.a. contínua X, hem de seguir els següents passos per realitzar un contrast χ^2:

1) Prenem una mostra de mida n de la població i anotem quin valor pren la v.a. X per cadascuna de les observacions.
2) Com que es tracta d'una v.a. contínua, dividim els valors que pot prendre la v.a. X en k intervals, com si féssim un histograma: x_1, x_2, ..., x_k. Anomenem O_i al nombre observat d'individus de la mostra que pertanyen a l'interval de x_i. D'aquesta manera podem dir que: $\sum_{i=1}^{k} O_i = n$.
3) A través de la funció de distribució F(x) de la llei de probabilitat donada, calculem les probabilitat que la v.a. X pertanyi a cada interval: p_1, p_2, ..., p_k. Si no ens donen els paràmetres necessaris d'aquesta llei de probabilitat amb la qual hem de fer la comparació, estimem els valors a partir de la mostra. Diem que r és el nombre de paràmetres que ha calgut estimar. Com a conseqüència, si aquests paràmetres són coneguts r valdrà 0.
4) Calculem E_1, E_2, ..., E_k, essent $E_i = np_i$ el nombre teòric o esperat d'unitats que pertanyen a l'interval de x_i.

Les hipòtesis H_0 i H_1 que ens hem de plantejar en aquest contrast són les següents:

$$H_0 : \text{la llei ajusta bé les dades}$$
$$H_1 : \text{la llei no ajusta bé les dades}$$

Si acceptem H_0, és a dir, si les dades segueixen una llei de probabilitat donada, l'estadístic de contrast ha de seguir una distribució X_i-Quadrat amb $\nu = k - r - 1$ graus de llibertat:

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-r-1}$$

On *r* és el número de paràmetres estimats i, per tant, graus de llibertat perduts.
Donat un nivell de significació α, la regió d'acceptació d'aquest contrast és $[0, \chi^2_{k-r-1,\alpha}]$.

A les taules de la distribució χ^2 podem trobar-hi les diverses probabilitats acumulades per diferents valors de α i ν.

Exemple: *Després d'observar el temps de vida de 70 bateries de cotxe d'unes determinades especificacions, s'ha obtingut la següent distribució:*

Temps de vida (anys)	*[0, 1)*	*[1, 2)*	*[2, 3)*	*[3, 4)*	*≥ 4*
Freqüència observada O_i	*30*	*23*	*6*	*5*	*6*

Segueixen aquestes dades una llei Exponencial? (α = 5%)

Els passos 1 i 2 que hem explicat anteriorment ja ens els donen fets, per tant, passem directament a fer el passos 3 i 4.

Com que ens demanen si la mostra segueix una llei Exponencial, necessitem calcular $p_1, p_2, ..., p_k$ a través de la funció de distribució de la llei Exponencial:

$$F(x) = \begin{cases} 0, & \text{si } x < 0 \\ 1 - e^{-\lambda x}, & \text{si } x \geq 0 \end{cases}$$

Veiem que no ens donen el paràmetre λ necessari per poder calcular aquestes probabilitats, per tant, l'estimem a partir de la mostra:

Marca de classe Ci	0.5	1.5	2.5	3.5	5
Freqüència observada O_i	30	23	6	5	6

$$\hat{\lambda} = (\bar{x})^{-1} = \left(\frac{30 \cdot 0.5 + 23 \cdot 1.5 + \ldots + 5 \cdot 6}{70}\right)^{-1} = 0.625$$

Ara ja podem calcular $p_1, p_2, ..., p_k$ i, posteriorment, $E_1, E_2, ..., E_k$:

$$F(0) = 0,\ F(1) = 0.465,\ F(2) = 0.713,\ F(3) = 0.847,\ F(4) = 0.919$$

x_i	[0, 1)	[1, 2)	[2, 3)	[3, 4)	≥ 4
p_i	F(1) – F(0) = = 0.465	F(2) – F(1) = = 0.248	F(3) – F(2) = = 0.134	F(4) – F(3) = = 0.072	1 – F(4) = = 0.081
E_i	32.55	17.36	9.38	5.04	5.67

Fem el càlcul de l'estadístic del contrast:

$$\chi^2_{obs} = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} = \frac{(30 - 32.55)^2}{32.55} + \ldots + \frac{(6 - 5.67)^2}{5.67} = 3.27$$

Busquem a les taules el valor de $\chi^2_{k-r-1,\alpha}$:

$$\chi^2_{k-r-1,\alpha} = \chi^2_{5-1-1,0.05} = \chi^2_{3,0.05} = 7.815$$

Com que $\chi^2_{obs} \leq \chi^2_{3,0.05}$, l'estadístic de contrast entra dins la regió d'acceptació, és a dir, no tenim motius suficients per rebutjar H_0. Per tant, podem afirmar amb un 95% de confiança que el temps de vida de les bateries de cotxe d'aquest problema segueix una distribució de probabilitat exponencial amb $\lambda = 0.625$.

PROBLEMES

1. Exercicis resolts

1.1. *Un procés industrial fabrica peces les longituds de les quals es distribueixen segons una llei normal L~N(190;10) mm. Es vol realitzar un contrast de la hipòtesi H_0: $\mu = 190$ versus la hipòtesi contrària a partir d'una mostra de mida $n = 5$, a un nivell de significació $\alpha = 0.05$. Es demana:*

a) Indiqueu l'estadístic a utilitzar per a realitzar el contrast.
b) Calculeu, amb l'ajuda de les corbes característiques, la probabilitat β de cometre un error de tipus II quan $\mu = 180$ mm.
c) Si la mostra obtinguda ha resultat ser igual a 187, 212, 195, 208 i 200, es demana quina decisió cal prendre.
d) Suposeu ara que la variància poblacional es desconeguda. Realitzeu novament el contrast sobre la mostra anterior i indiqueu la decisió adoptada.

El contrast que se'ns planteja a l'enunciat és bilateral d'una esperança.

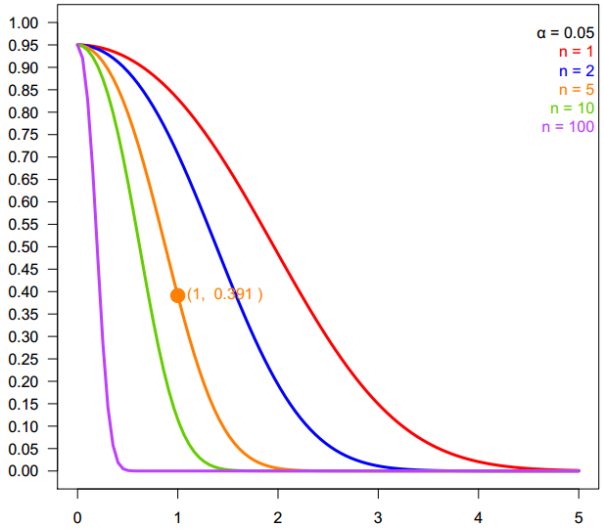
$$H_0 : E\{L\} = 190 \text{ mm}$$
$$H_1 : E\{L\} \neq 190 \text{ mm}$$

a) Com que la variància es coneguda, utilitzarem el següent estadístic de contrast:

$$\mathbf{z_{obs}} = \frac{\mathbf{\bar{l}_5 - 190}}{\frac{\mathbf{10}}{\mathbf{\sqrt{5}}}} \quad \text{pertany a} \quad Z = \frac{\bar{L}_5 - 190}{\frac{10}{\sqrt{5}}} \sim N(0;1)$$

b) Ens centrem en la corba característica per variància coneguda, nivell de significació de 0.05 i número de mostres igual a 5.

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|180 - 190|}{10} = 1$$



Busquem la intersecció de d = 1 amb la corba n = 5 i ens dóna **β =0.4**. Per tant, la probabilitat de cometre un error de tipus II és de 0.4.

c) Calculem la mitjana de la mostra donada a l'enunciat, $\bar{l}_5 = 200.4$ mm. Tot seguit calculem l'estadístic de contrast sobre la mostra:

$$z_{obs} = \frac{\bar{l}_5 - 190}{\frac{10}{\sqrt{5}}} = \frac{200.4 - 190}{\frac{10}{\sqrt{5}}} = 2.3255$$

Busquem a la distribució normal de les taules estadístiques el p-valor d'aquest estadístic de contrast. Com que no és un valor exacte hem de fer una interpolació lineal:

$$\frac{2.33 - 2.32}{0.9901 - 0.9898} = \frac{2.33 - 2.3255}{0.9901 - p}$$

$$p = 0.9899$$

$$p - \text{valor} = 2 \cdot (1 - p) = 0.202$$

Com que el p-valor < α, **rebutgem H_0 i acceptem la H_1**; la mitjana de les peces és diferent a 190 mm.

d) Com que la variància es desconeguda, utilitzarem el següent estadístic de contrast:

$$t_{obs} = \frac{\bar{l}_5 - 190}{\frac{\sigma}{\sqrt{5}}} \quad \text{pertany a} \quad t = \frac{\bar{L}_5 - 190}{\frac{\sigma}{\sqrt{5}}} \sim t_{n-1}$$

Calculem la desviació sobre la mostra $\sigma = 10.015$ mm. Tot seguit calculem l'estadístic de contrast sobre la mostra. Recordem que $\bar{l}_5 = 200.4$ mm.

$$t_{obs} = \frac{\bar{l}_5 - 190}{\frac{\sigma}{\sqrt{5}}} = \frac{200.4 - 190}{\frac{10.015}{\sqrt{5}}} = 2.322$$

Busquem a la distribució t-Student de les taules estadístiques el p-valor d'aquest estadístic de contrast per v = n – 1 = 4. Com que no és un valor exacte hem de fer una interpolació lineal:

$$\frac{2.776 - 2.132}{0.025 - 0.05} = \frac{2.776 - 2.322}{0.025 - p}$$

$$p = 0.0426$$

$$p - \text{valor} = 2 \cdot p = 0.0852$$

Com que el p-valor ≥ α, **acceptem la H_0** i per tant diem que la mitjana poblacional de les peces és de 190 mm.

1.2. *Els registres de producció de l'any passat en un procés d'enllaunat de pastanagues va mostrar que la longitud mitjana de les pastanagues era de 12.46 cm i la desviació estàndard de 1.80 cm. En la collita d'enguany una mostra de 100 pastanagues ha donat una longitud mitjana de 12.82 cm. Es demana:*

a) Podem dir que les pastanagues d'ara són més llargues que les de l'any passat?
b) Suposem que un exportador ens vol comprar un gran carregament però vol una longitud mitjana de 12.50 cm. Agafa una mostra de 31 pastanagues i troba una longitud mitjana de 12.02 cm i una desviació estàndard de 1.74 cm. Basat en la inspecció d'aquesta mostra, ha d'acceptar l'exportador el carregament, si s'estableix un nivell de significació del 5%, per a una prova a dues cues? (suposem ara desconeguda la desviació tipus).
c) Si realment μ = 10.76 cm, determineu, mitjançant les corbes característiques, quina hauria de ser la mida de la mostra per a que l'exportador ho detectés amb probabilitat 0.8 (considereu el contrast de l'apartat b amb la mateixa desviació estàndard mostral).

a) Definim la v.a. N(12.46;1.8) que defineix la longitud de les pastanagues de l'any passat com a P. En canvi, la v.a. que defineix la longitud de les d'aquest any com a A. El contrast que hem de realitzar és un contrast d'un esperança unilateral dreta.

$$H_0 : E\{A\} \leq 12.46$$
$$H_1 : E\{A\} > 12.46$$

Calculem l'estadístic de contrast sobre la mostra:

$$z = \frac{\overline{P} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{12.82 - 12.46}{\frac{1.8}{\sqrt{100}}} = 2$$

Busquem a la distribució Normal de les taules estadístiques el p-valor d'aquest estadístic de contrast.

$$P(Z < 2) = 0.9772$$

$$P(Z > 2) = p - valor = 1 - p = 0.0228$$

Com que el p-valor < α, **rebutgem H_0 i acceptem la H_1**; les pastanagues d'aquest any són més llargues que les de l'any passat.

b) Definim la v.a. N(12.02;1.74) que defineix la longitud de les 31 pastanagues agafades com a mostra com a M. El contrast que hem de realitzar és un contrast d'un esperança bilateral.

$$H_0 : E\{M\} = 12.5$$
$$H_1 : E\{M\} \neq 12.5$$

Calculem l'estadístic de contrast sobre la mostra.

$$t_{obs} = \frac{\bar{M} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{12.02 - 12.5}{\frac{1.74}{\sqrt{31}}} = 1.5359$$

Busquem a la distribució t-Student de les taules estadístiques el p-valor d'aquest estadístic de contrast per v = n – 1 = 30. Com que no és un valor exacte hem de fer una interpolació lineal:

$$\frac{1.697 - 1.310}{0.05 - 0.1} = \frac{1.697 - 1.5359}{0.05 - p}$$

$$p = 0.0708$$

$$p - valor = 2 \cdot p = 0.1415$$

Com que el p-valor ≥ α, **acceptem la H_0** i per tant l'exportador hauria d'acceptar el carregament.

c) Ens centrem en la corba característica per variància desconeguda, nivell de significació de 0.05 i potència del contrast igual a 0.8.

$$P(\text{error tipus II}) = \beta = 1 - Pot(\theta) = 1 - 0.8 = 0.2$$

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|10.76 - 12.5|}{1.74} = 1$$

Busquem la intersecció de d = 1 amb $\beta = 0.2$ i ens dóna **n = 10**. Per tant, la mida de la mostra hauria de ser de 10 pastanagues.

1.3. *MOLTNET S.A. produeix un mateix tipus de detergent en dues plantes, una a Granollers i l'altra a Getafe. A Granollers fan servir matèria prima del proveïdor A, i a Getafe fan servir matèria prima del proveïdor B. Un investigador, per comparar la influència dels dos proveïdors en la producció, recull les quantitats de detergent produïdes en les dues plantes durant 25 dies. Els resultats obtinguts són els següents:*

	Proveïdor A	***Proveïdor B***
Producció mitjana	*130.0 Tm*	*127.2 Tm*
Desviació tipus s	*4.5 Tm*	*3.1 Tm*

Es demana:

a) En base a aquest estudi, quin dels dos proveïdors és preferible?
b) Creieu que el disseny d'aquesta investigació és correcte? Perquè?

a) Per saber quin dels dos proveïdors és preferible, haurem de realitzar un contrast bilateral d'igualtat d'esperances a partir de dues mostres independents amb un nivell de significació de 0.05.

$$H_0 : \mu_A - \mu_B = 0$$
$$H_1 : \mu_A - \mu_B \neq 0$$

Calculem l'estadístic de contrast sobre les mostres, les variàncies poblacionals de les quals són desconegudes però les suposem iguals.

$$s_p^2 = \frac{(n-1)\cdot s_A^2 + (m-1)\cdot s_B^2}{n+m-2} = \frac{(25-1)\cdot 4.5^2 + (25-1)\cdot 3.1^2}{25+25-2} = 14.93$$

$$t_{obs} = \frac{(\overline{A}_n - \overline{B}_m) - (\mu_A - \mu_B)}{s_p \cdot \sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{(130-127.2)-0}{\sqrt{14.93}\cdot\sqrt{\frac{1}{25}+\frac{1}{25}}} = 2.562$$

Busquem a la distribució t-Student de les taules estadístiques el p-valor d'aquest estadístic de contrast per v = n + m + 2 = 48.

Tot i no estar tabulat, si ens fixem que tant per n = 40 com n = 60 l'estadístic de contrast t = 2.562 es troba entre la probabilitat de 0.01 i 0.005, podem afirmar que per n = 48 també estarà en aquest interval. Així doncs, és segur que serà menor a 0.01.

Com que el p-valor < α, **rebutgem H_0 i acceptem la H_1**; el rendiment de les cadenes de producció són diferents.

Fent el IC(1 – α), podrem saber quina de les dues cadenes produeix més.

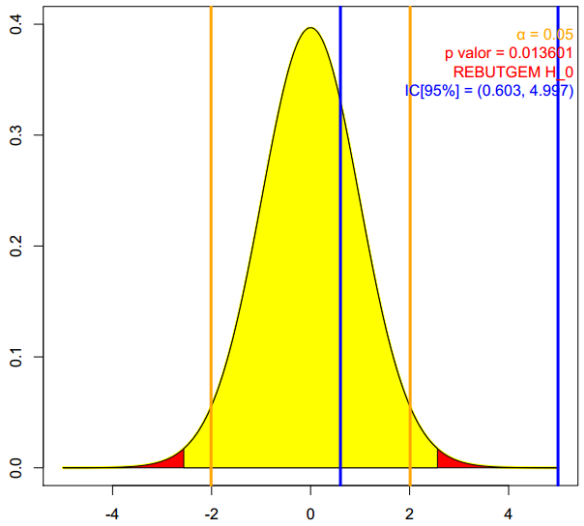
$$IC(1-\alpha) = (\overline{A}_n - \overline{B}_m) \pm t_{n+m-2,\alpha/2}\cdot s_p \cdot \sqrt{\frac{1}{n}+\frac{1}{m}}$$

$$IC(1-0.05) = (\overline{A}_n - \overline{B}_m) \pm t_{48,0.025}\cdot s_p \cdot \sqrt{\frac{1}{n}+\frac{1}{m}}$$

$$IC(95\%) = (130-127.2) \pm 2.010635\cdot\sqrt{14.93}\cdot\sqrt{\frac{1}{25}+\frac{1}{25}}$$

$$IC(95\%) = (0.603, 4.997)$$

Com que l'interval de confiança trobat no inclou el zero i tot ell és positiu, podem dir que **la matèria prima del proveïdor A és més rentable**.



b) No és correcte. Per ser-ho totalment s'hauria d'haver fet l'estudi a la mateixa cadena productiva.

1.4. *Es vol controlar la contaminació difusa per nitrats en un determinat riu. Per aquesta raó s'ha mesurat la quantitat de nitrats (NO_3 en mg/l) en 3 estacions de control. La primera estació està situada en el riu objecte del nostre estudi, les altres dues estan situades en els seus dos principals afluents. En cada estació s'han realitzat 10 mesures de la quantitat de nitrats. Les observacions en cada estació són independents. Els resultats obtinguts són els següents:*

Estació 1	***Estació 2***	***Estació 3***
$n_1 = 10$	$n_2 = 10$	$n_3 = 10$
$\bar{x}_1 = 10.5$	$\bar{x}_2 = 9.4$	$\bar{x}_3 = 12.4$
$s_1 = 2.4$	$s_2 = 1.6$	$s_3 = 2$

Apliqueu un contrast a partir d'un model d'anàlisi de la variància per comprovar si la mitjana del contingut de nitrats és la mateixa en cada estació de control.

Com que tenim més de dues distribucions a comparar aplicarem ANOVA. El contrast que plantegem amb un nivell de significació de 0.05 és el següent:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \text{les mitjanes } \mu_1, \mu_2, \mu_3 \text{ no són totes iguals}$$

Realitzem tots els càlculs necessàries per emplenar la taula d'anàlisis de la variància.

$$k = 3$$

$$n = 30$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{30} \cdot (10 \cdot 10.5 + 10 \cdot 9.4 + 10 \cdot 12.4) = 10.77$$

$$SSTract = \sum_{i=1}^{k} n_i \cdot (\bar{y}_i - \bar{y})^2 =$$

$$= 10 \cdot (10.5 - 10.77)^2 + 10 \cdot (9.4 - 10.77)^2 + 10 \cdot (12.4 - 10.77)^2$$

$$SSTract = 46.067$$

$$SSErr = \sum_{i=1}^{k} (n_i - 1) \cdot s_i^2 = (10 - 1) \cdot 2.4^2 + (10 - 1) \cdot 1.6^2 + (10 - 1) \cdot 2^2$$

$$SSErr = 110.88$$

$$SSTot = SSTract + SSErr = 46.067 + 110.88 = 156.947$$

$$MSTract = \frac{SSTract}{k - 1} = \frac{46.067}{3 - 1} = 23.0335$$

$$MSErr = \frac{SSError}{n - k} = \frac{110.88}{30 - 3} = 4.1067$$

$$F = \frac{MSTract}{MSErr} = \frac{23.0335}{4.1067} = 5.609$$

Calculats tots els valors, la taula quedaria de la següent manera:

	Graus de llibertat (DF)	**Suma dels quadrats (SS)**	**Mitjana dels quadrats (MS)**	**Estadístic de contrast (F)**
Entre tractaments	2	46.067	23.0335	5.609
Dins tractaments	27	110.88	4.1067	
Total	29	156.947		

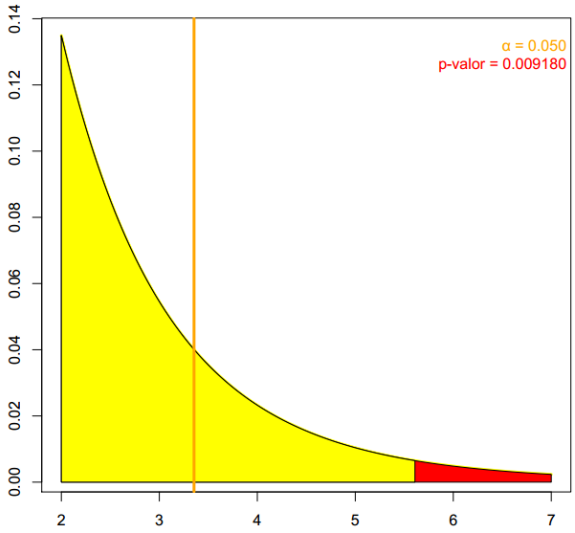
Busquem a la distribució de Fisher de les taules estadístiques quin és la probabilitat que li correspon a l'estadístic de contrast trobat amb els graus de llibertat corresponents.

$$v_1 = k - 1 = 2$$

$$v_2 = n - k = 27$$

Veiem que 5.609 és major que el valor de 5.49 en la taula de la distribució de Fisher amb $\alpha = 0.01$. Per això podem afirmar que la probabilitat que acumularà serà menor a 0.01.

Com que el p-valor < α, **rebutgem H_0 i acceptem la H_1**; no totes tres estacions de de control mesuren el mateix contingut de nitrats.



2. Exercicis proposats

2.1. *En els laboratoris d'una empresa petroquímica es vol realitzar un assaig per contrastar la hipòtesi H_0 que la mitjana del contingut de plom d'unes barreges de petroli és major o igual que 300 µg, versus la hipòtesi contrària. Es fixa l'error de tipus I en $\alpha = 0.01$. Per una experiència anterior se sap que la desviació tipus en el contingut de plom és igual a $\sigma = 30$µg. Es demana:*

a) Si la mida de la mostra és n = 25, determineu la regió d'acceptació o no rebuig de la H_0. Determineu també quina és la probabilitat β de cometre un error de tipus II quan $\mu = 280$ µg.

Suposeu ara que el control es vol realitzar de forma que si la mitjana real del contingut de plom és de 280 µg, la probabilitat que la hipòtesi H_0 sigui acceptada com a bona sigui només de $\beta = 0.05$.

b) Consulteu les corbes característiques corresponents per tal de determinar la mida n aproximada de la mostra per tal de dur a terme aquest contrast.
c) Suposeu que s'ha extret una mostra de mida n = 36, la mitjana de la qual ha resultat ser igual a $\bar{x} = 290$ µg. Quina seria la decisió que caldria adoptar?

Solució: a) Acceptació = [286.05, +∞]; $\beta \approx 0.15$; b) n = 45; c) No rebutgem H_0

2.2. *Es fa una remesa de tubs de grassa industrial, dels quals s'afirma que el pes mig és de 1000 g. Examinada una mostra de 5 d'aquests tubs, s'han obtingut els següents pesos: 995, 992, 1005, 998, 1000 g. Hi ha motius per rebutjar la H_0: μ = 1000 g versus la hipòtesis H_1: μ < 1000 g, amb un nivell de significació del 5%?*

Solució: $\bar{x} = 998$, s=4.95, t=-0.903, p-valor=0.209, no tenim motius suficients per rebutjar H_0.

2.3. *Es desitja contrastar la hipòtesi H_0: $\mu \geq 70$ versus la hipòtesi contrària. El contrast d'hipòtesi es vol realitzar de manera que α = 0.05 i $\beta(\mu = 60)$ = 0.1. S'ha extret una petita mostra per conèixer una primera aproximació de la desviació tipus de la variable objecte de contrast, i s'ha trobat que s = 6.2. Busqueu quina ha de ser la mida aproximada de la mostra per dur a terme aquest contrast.*

Solució: d=1.61 mirant les corbes característiques $n \approx 5$

2.4. *Un cap de planta afirma que el temps per a produir un determinat lot de producte és, en mitjana, igual a 15 h amb σ = 2 h. Per a contrastar la hipòtesi H_0: μ = 15 h versus la hipòtesi H_1: μ < 15 h, es controla 25 vegades el temps de durada de la producció d'aquest tipus de lot. Es demana:*

a) Indiqueu l'estadístic a utilitzar per realitzar el contrast. Calculeu la probabilitat β de cometre un error de tipus II quan μ = 14h.
b) Si després de produir 25 lots del tipus esmentat, la mitjana del temps resulta ser igual a 13.8 h, indiqueu la decisió a prendre i el valor crític p d'aquest contrast.

Solució: a) d=0.5; si α=0.05 llavors $\beta \approx 0.2$, si α=0.01 llavors $\beta \approx 0.4$
b) z=-3, p-valor=0.0013, es conclou rebutjar H_0.

2.5. *La duració mitjana d'unes certes peces en les que intervenen diferents materials és igual a 1800 h. Variant un d'aquests materials, la vida mitjana d'una mostra de 10 peces ha resultat ser igual a 2000 h amb una desviació tipus mostral s = 150 h.*

Creieu que el canvi del material ha incrementat significativament la vida mitjana de les peces?

Solució: t=4.216, p-valor=0.00112, es conclou rebutjar H_0 i per tant es conclou que el canvi ha incrementat significativament la vida mitjana

2.6. *El rendiment d'una màquina A en diferents dies ha resultat ser de: 137.5, 140.7, 106.9, 175.1, 177.3, 120.4, 77.9 i 104.2. El rendiment d'una altra màquina B de característiques semblants ha estat igual a: 103.3, 121.7, 98.4, 161.5, 167.8 i 67.3. El control de la màquina B s'ha realitzat en dies diferents als de la màquina A. Contrasteu, a un nivell de significació α = 0.05, la hipòtesi H_0: "les mitjanes dels*

rendiments de les dues màquines són iguals" versus la hipòtesi contrària (suposeu igualtat de variàncies).

Solució: t=0.508, p-valor=0.621, no tenim motius suficients per rebutjar H_0.

2.7. *En una empresa es vol estudiar si la implantació d'un nou pla de seguretat en el treball és realment efectiu. Amb aquesta intenció s'han calculat les hores de treball setmanals perdudes per accidents laborals –abans i després de la implantació del pla– en les sis plantes de producció de les que es composa l'empresa.*

Aquests són els valors obtinguts:

Hores setmanals perdudes	***Planta***					
	1a	***2a***	***3a***	***4a***	***5a***	***6a***
Abans del pla	*12*	*29*	*16*	*37*	*28*	*15*
Després del pla	*10*	*28*	*17*	*35*	*25*	*16*

Confirmen aquestes dades l'eficàcia del pla de seguretat? Especifiqueu les hipòtesis teòriques necessàries per aplicar el contrast.

Solució: t=1.463, p-valor=0.102, no tenim motius suficients per rebutjar H_0.

2.8. *Sis màquines són sotmeses a unes millores tècniques. Es controla la productivitat per hora (pes producte en kg) de les màquines immediatament abans d'aplicar les millores i al cap de tres mesos després. A partir dels resultats obtinguts es demana:*

Prod. després (y)	*73.5*	*75.0*	*74.5*	*75.0*	*75.5*	*82.0*
Prod. abans (x)	*72.5*	*74.5*	*60.0*	*70.0*	*76.5*	*80.0*

a) Contrasteu la hipòtesi que no hi ha augment significatiu de la producció, versus la hipòtesi contrària ($\alpha = 0.05$).
b) Calculeu la mida que hauria de tenir la mostra de màquines per tal que una augment real de 2 kg en la mitjana del pes per aplicació de les millores fos detectada amb probabilitat 0.90 ($\alpha = 0.05$). Suposeu que la desviació tipus de la variable D = Y – X és igual a 2.5 kg.

Solució: a) t=1.583, p-valor=0.087, no tenim motius suficients per rebutjar H_0.
b) n=15

2.9. *Contrasteu a partir de la mostra següent de la durada en dies d'un determinat coixinet d'un engranatge, la hipòtesi segons la qual la població de procedència es distribueix segons una llei de distribució normal.*

54	*54*	*55*	*55*	*57*	*58*	*58*	*58*	*59*	*59*	*59*	*59*	*60*	*60*	*60*	*60*
60	*60*	*61*	*61*	*62*	*62*	*62*	*63*	*63*	*64*	*64*	*64*	*65*	*65*	*65*	*65*
66	*66*	*66*	*66*	*67*	*68*	*70*	*70*	*70*	*71*	*71*	*71*	*71*	*72*	*72*	*72*
73	*73*	*74*	*74*	*74*	*74*	*75*	*75*	*75*	*75*	*76*	*76*	*77*	*77*	*80*	*80*

80 84 84 85 85 85 90 94 97

Solució: Rebutgem la normalitat

2.10. *Contrasteu la hipòtesi de normalitat de les dades recollides per Michelson (1879) sobre la velocitat de la llum en km/s:*

299850, 299740, 299900, 300070, 299930, 299850, 299950, 299980, 299980, 299880, 300000, 299980, 299930, 299650, 299760

Solució: No es pot rebutjar la normalitat

2.11. *S'han mesurat 12 valors d'una variable física que se sospita que no és normal. Els valors obtinguts han estat els següents. Ajudeu-vos d'un diagrama Q-Q per contrastar la hipòtesi de normalitat de la variable.*

30.2, 30.8, 29.3, 29.0, 30.9, 30.8, 29.7, 28.9, 30.5, 31.2, 31.3, 28.5.

Solució: No es pot rebutjar la normalitat

2.12. *S'estudia el temps de vida (en hores) de 10 bateries de 9 volts seleccionades a l'atzar de la producció diària, i s'obtenen els següents resultats. Ajudeu-vos d'un diagrama Q-Q per contrastar si la mostra anterior s'ajusta a una llei exponencial.*

28.9, 15.2, 28.7, 72.5, 48.6, 52.4, 37.6, 49.5, 62.1, 54.5

Solució: Rebutgem el model exponencial

2.13. *Estudieu si la següent mostra procedeix o no d'una distribució uniforme sobre l'interval [0, 2].*

0.7017, 1.3871, 1.4120, 0.0236, 1.8542, 1.7697, 0.3041, 0.2344, 0.0504, 0.6095 1.1827, 1.2472, 0.9881, 0.4483, 1.7020, 1.9235

Solució: No es pot rebutjar el model U[0,2]

2.14. *Hom vol estudiar si la resistència d'una determinada peça depèn o no del tipus de material amb que és fabricada. Per a cada tipus de material s'han construït tres exemplars del mateix tipus de peça i, tot seguit, s'ha sotmès a una prova d'esforços, mesurant el nombre de quilos que podien suportar fins a trencar-se.*

Els resultats foren:

Tipus material

		A	B	C	D
Número d'observació	1	15	25	17	10
	2	9	21	23	13
	3	14	19	20	16

Es demana:

a) Feu un estudi exploratori preliminar (gràfic i numèric) i comenteu els resultats.
b) Apliqueu un contrast a partir d'un model d'anàlisi de la variància. A quina conclusió s'arriba?

Solució: b) SSTract=196.333 MSTract=65.44 Fobs=6.950 p-valor=0.013
SSError=75.333 MSError=9.417
SSTotal=271.667 Es conclou rebutjar H_0.

2.15. *En una planta de producció de piles alcalines, hi ha quatre línies que fabriquen el mateix tipus de pila. S'han escollit de cada línia una mostra de piles i s'ha mesurat la seva durada (en hores). Els resultats obtinguts són els següents:*

Línia de producció			
1	2	3	4
380	350	354	376
376	356	360	344
360	358	362	342
368	376	352	372
372	338	366	374
366	342	372	360
374	366	362	
382	350	344	
	344	342	
	364	358	
		351	
		348	
		348	

a) Estudieu gràfica i numèricament les dades.
b) Comproveu a partir d'un model d'anàlisi de la variància si la durada de les piles depèn de la línia de producció de procedència.

Solució: b) SSTract=12.189 MSTract=0.642 Fobs=0.436 p-valor=0.958
SSError=25.000 MSError=1.471
SSTotal=37.189
Es conclou que no tenim motius suficients per rebutjar H_0.

2.16. *En una fàbrica metal·lúrgica de l'Alt Empordà especialitzada en la fabricació de cables d'acer, s'estan fent servir tres tipus de material (A, B i C). L'empresa vol realitzar un control per tal d'analitzar si el tipus de material influeix en la temperatura màxima que pot suportar el cable (en °C). Per aquesta raó s'ha*

seleccionat una mostra aleatòria de 11 cables fabricats amb el material A, 10 cables fabricats amb el material B i 9 cables fabricats amb el material C i s'ha mesurat la temperatura màxima que poden suportar. Els resultats obtinguts són els següents:

Material A	$n_A = 11$	$\bar{y}_A = 805.36$	$s_A = 17.22$
Material B	$n_B = 10$	$\bar{y}_B = 785.9$	$s_B = 16.58$
Material C	$n_C = 9$	$\bar{y}_C = 800.22$	$s_C = 26.64$

Aplica un model ANOVA per contrastar la hipòtesi segons la qual la temperatura mitjana que poden suportar els cables és la mateixa pels 3 materials de fabricació (la mitjana global de totes les dades és $\bar{y}$ = 797.33 i la desviació estàndard corregida global és s = 20.29).

Solució: SSTract=2089.6 MSTract=1044.8 Fobs=2.53 p-valor=0.098
SSError=11116.8 MSError=417.7
SSTotal=13206.4
Es conclou que no tenim motius suficients per rebutjar H_0.

2.17. *Una indústria utilitza una màquina per polir peces metàl·liques. El temps mig que triga la màquina en polir una peça és de 10 segons. L'encarregat sospita que la màquina no funciona gaire bé i que triga més temps. Amb la intenció d'estudiar l'estat de la màquina s'han mesurat els temps de polit de 12 peces obtenint els següents valors (en segons):*

10.2 9.8 10.3 10.5 10.1 9.9 9.7 10.2 10.4 10.2 9.7 10.6

Fes un contrast d'hipòtesi amb un nivell de significació del 0.10 per determinar si l'encarregat té raó amb la seva sospita.

Solució: $\bar{x} = 10.133$, s=0.302, t=1.527, p-valor=0.0775<0.10, rebutgem H_0.

2.18. *La vida d'uns components electrònics es distribueix segons una llei normal. S'ha extret una mostra de 10 components d'una certa marca A la vida dels quals resulta ser: 1614, 1294, 1293, 1643, 1466, 1270, 1340, 1380, 1028, 1497 h. Es fa el mateix amb els components d'una segona marca B, resultant: 1383, 1138, 1092, 1143, 1017, 1061, 1627, 1021, 1711, 1965 h. Calculeu, per un nivell de confiança 1 – α = 0.95, una estimació de la diferència d'esperances, suposant la igualtat de variàncies de les dues poblacions.*

Solució: IC(95%) = (–188.7, 322.7)

2.19. *En la fabricació de bigues d'acer poden utilitzar-se dos aliatges A i B diferents. En una siderúrgia es vol comparar aquests dos aliatges pel que fa referència a la capacitat de carga de les bigues. Com a tal, cal entendre el pes màxim que es pot*

aplicar sobre la biga abans no es trenqui. S'han seleccionat dues mostres a l'atzar de bigues. En les bigues de la primera mostra s'ha utilitzat l'aliatge A mentre que en les de la segona s'ha utilitzat l'aliatge B. Amb els resultats obtinguts es demana:

$$\begin{array}{llll} \textit{Aliatge A:} & n_A = 11 & \bar{y}_A = 43.7\,Tm & s_A^2 = 24.4\,Tm^2 \\ \textit{Aliatge B:} & n_B = 17 & \bar{y}_B = 48.5\,Tm & s_A^2 = 19.9\,Tm^2 \end{array}$$

a) Feu una estimació, al nivell de confiança del 99%, de la diferència entre les mitjanes de les capacitats de carga dels dos aliatges. Especifiqueu els supòsits teòrics a tenir en compte per poder fer aquesta estimació.
b) A la vista del resultat anterior, es pot concloure que les mitjanes de les capacitats de carga dels dos aliatges són significativament diferents?

Solució: a) IC(99%) = (–9.8, 0.2); b) No, el valor zero està dins l'IC

2.20. *Una mostra de mida 10 d'una N(μ_1;σ_1 = 225) té una mitjana mostral igual a $\bar{x}_1$ = 170.2. Una altra mostra de mida 12 d'una N(μ_2;σ_2 = 256) té una mitjana mostral $\bar{x}_2$ = 176.7. Calculeu un interval de confiança del 95% de la diferència $\mu_1 - \mu_2$.*

Solució: IC(95%) = (–19.48, 6.48)

2.21. *En una ciutat s'han fet anàlisis d'aigua per a comprovar la potabilitat de la mateixa. Es van prendre 10 mostres a l'atzar del sistema d'aigües de la ciutat i es van dividir a fi i a efecte que fossin analitzades per dos laboratoris diferents. En quan a la quantitat de sulfats (mg/l), els resultats obtinguts foren:*

Lab. A	*50*	*58*	*49*	*53*	*55*	*54*	*59*	*49*	*53*	*56*
Lab. B	*53*	*59*	*48*	*55*	*54*	*56*	*63*	*55*	*52*	*63*

Estudieu si hi ha diferències en els mètodes dels laboratoris en quan a aquesta variable en particular.

Solució: t=-2.4328, p-valor=0.03781, es conclou rebutjar H_0
$\bar{d} = 2$, IC(95%) per la mitjana de la diferència és (-4.2457, -0.1543)

PRÀCTIQUES

1. Contrast de la mitjana a partir d'una mostra

Treballarem en primer lloc amb l'arxiu `vel.llum` que ja coneixem. Les dades d'aquest vector corresponen a les mesures repetides de la velocitat de la llum (en km/s) que el físic Albert Michelson (1852–1931) va realitzar a finals del segle XIX.

Actualment, la comunitat científica accepta com a velocitat de la llum el valor de 299792.5 km/s.

- Calculeu la mitjana de la nostra mostra i observeu que no coincideix amb aquest valor teòric. Mitjana = 299896.7.

Tot seguit utilitzarem les dades de Michelson per contrastar les següents dues hipòtesis:

$$H_0: \mu = 299792.5$$
$$H_1: \mu \neq 299792.5$$

El nostre nivell de significació serà igual a $\alpha = 0.05$, que equival a un nivell de confiança del 95%. Recordem que R-Commander fa servir la mateixa funció per calcular intervals i fer contrastos d'una variable:

`Estadístics, Mitjanes, t-test per una variable ...`

Al requadre `Hipòtesi nul·la` escrivim `299792.5`. Com que es tracta d'un contrast bilateral deixem marcat `Population mean !=mu0`.

Fixeu-vos en la informació que apareix a la finestra de resultats. R-Commander ens informa del valor de l'estadístic de contrast t_{obs}, del p-valor i d'altres informacions.

Responeu:

- Quants graus de llibertat té la llei t-Student? Graus llibertat = Df = 14. Quina és la mida de la mostra utilitzada per fer l'estimació? Mida mostra = 15.
- Quina és l'estimació puntual de la mitjana? Mitjana = 299896.7.
- Quant val el valor de l'estadístic de contrast (t_{obs}) d'aquest test? t_{obs} = 3.605.
- Quant val el p-valor d'aquest contrast? p-valor = 0.00287.
- A la vista d'aquesta informació, quina és la decisió que preneu? Amb un nivell de significació α del 5 % (0.05) decidim rebutjar la hipòtesi nul·la perquè el p-valor (0.00287 = 0.287%) és inferior a α.
- Amb aquesta decisió, quin tipus d'error –I o II– podeu cometre? Estem rebutjant la H_0 i podria passar que realment fos certa, aleshores podem cometre un error de tipus I.
- Quina probabilitat teniu de cometre'l? 0.00287.

Si volguéssim fer el contrast unilateral per la dreta, és a dir:

$$H_0: \ \mu \leq 299792.5$$
$$H_1: \ \mu > 299792.5$$

Només caldria marcar Mitjana poblacional `mu > mu0`.

- Feu-ho i calculeu el p-valor. p-valor = 0.001435.
- Quina decisió hauríeu de prendre? Rebutjar la hipòtesi nul·la.
- Com estan relacionats els p-valors dels contrastos unilateral i bilateral? El contrast unilateral és la meitat del bilateral. La distribució t-Student és perfectament simètrica.

2. Contrast d'igualtat de mitjanes a partir de dues mostres independents

Si tenim totes les dades posades en una columna i tenim una altra columna que contingui un factor que distingeixi entre les dues mostres, R-Commander ens permetrà fer un contrast de mostres independents a través del següent menú:

`Estadístics, Mitjanes, t-test per mostres independents ...`

Mitjançant el menú també podrem fer contrastos amb dades aparellades si tenim dues columnes: una per la primera mostra i l'altra per la segona.

Per un procediment estàndard s'ha determinat la demanda d'oxigen en aigües residuals (en mg/l). S'han fet 10 determinacions d'aigües residuals amb els següents resultats:

74.4, 67.2, 66.1, 71.2, 68.7, 69.9, 71.0, 77.8, 72.4, 70.1

Al mateix temps, i per un altre mètode, s'han fet 10 determinacions diferents d'aigües residuals amb els següents resultats:

71.6, 71.4, 71.3, 74.5, 71.9, 72.6, 69.1, 73.4, 69.5, 70.2

Carreguem aquestes dades des del fitxer `oxigen.rda`. Visualitzeu les dades.

Volem comparar, a un nivell de significació del $\alpha = 0.05 = 5\ \%$, si hi ha diferències entre els dos mètodes de determinació. A la vista de les dades, es tracta de fer un contrast de comparació de dues mitjanes provinents de mostres independents:

$$H_0: \ \mu_1 = \mu_2$$
$$H_1: \ \mu_1 \neq \mu_2$$

Abans de realitzar el contrast farem una anàlisi descriptiva, la qual ens ajudarà en les interpretacions i a decidir si a l'hora de fer el contrast assumim o no que les variàncies poblacionals són iguals.

Per a obtenir una anàlisi descriptiva numèrica de les dades, anem al menú:

`Estadístics, Resums, Resums numèrics ...`

Fem un clic en la pestanya `Resum per grups` per confirmar que volem els estadístics de la variable segons els grups determinats pel factor *metode*.

Ho completem amb una anàlisi descriptiva gràfica fent un diagrama de caixa múltiple. Anem al menú:

`Gràfics, Caixa de dispersió ...`

Fem un clic en la pestanya `Gràfic segons grup` per a dir que volem un diagrama de caixa per cada grup determinat per la variable *metode*. El gràfic obtingut ha de ser el de la **figura 1**.

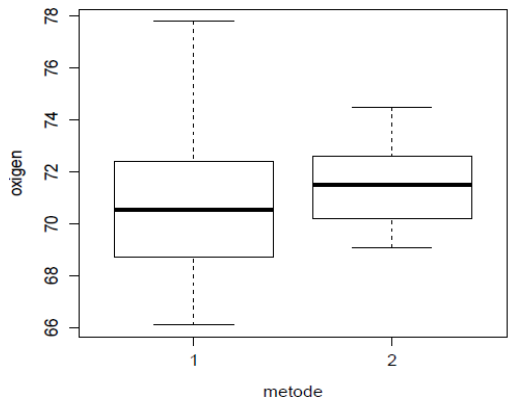


Figura 1: Diagrama de caixa múltiple per les dades del fitxer `oxigen.rda`.

- A la vista dels resultats numèrics i del diagrama de caixa múltiple, creus que hi ha indicis que les mitjanes són diferents?

mean	sd	0%	25%	50%	75%	100%	n
70.88	3.422410	66.1	69.000	70.55	72.100	77.8	10
71.55	1.682095	69.1	70.475	71.50	72.425	74.5	10

  Els diagrames de caixa suggereixen que no hi ha gaire diferència entre els valors centrals de les dues distribucions. Aquest indici queda corroborat pels resultats numèrics: la mitjana 70.88 té associat un error estàndard de $3.42/\sqrt{10} = 1.082261$ i la mitjana 71.55 un error estàndard de $1.68/\sqrt{10} = 0.5319251$. Si sumem i restem 2 errors estàndards a cada mitjana veiem que l'altra mitjana està dins els límits. Això ens suggereix que les mitjanes no estan molt allunyades entre si.

- Quina és la teva decisió respecte l'assumpció que les variàncies de les dues poblacions són iguals? Els diagrames de caixa suggereixen que les variabilitats (amplituds) són diferents. La variància mostral de la primera mostra és $3.42^2 = 11.71289$ i la de la segona $1.682095^2 = 2.829444$. Malgrat que les mides de les mostres són molt petites (n = 10), sembla que podem

dir que les variàncies són diferents. Si anem al menú `Estadístics, Variàncies, Test F per dues variàncies ...` i fem el contrast d'igualtat de variàncies, el p-valor = 0.04588 és una mica inferior a un 5% de significació. Per tant, decidim rebutjar la igualtat de variàncies (H_0).

Suposarem que les variàncies són desconegudes i diferents. Anem al menú:

`Estadístics, Mitjanes, t-test per mostres independents ...`

Deixem seleccionades totes les opcions que apareixen per defecte. Fixeu-vos en la informació numèrica de la finestra de resultats i contesteu les següents preguntes:

- Quant val l'estadístic de contrast? $t_{obs} = -0.5556$.
- Quants graus de llibertat té la t-Student associada a aquest estadístic? Graus llibertat = 13.108. Fixem-nos que és menor que 10+10-2=18.
- Quant val el p-valor d'aquest contrast? p-valor = 0.5878.
- A la vista d'aquest p-valor, quina decisió prendríeu en relació al contrast de les dues hipòtesis? No tenim motius suficients per rebutjar H_0, ja que el p-valor del contrast (0.5878 = 58.78%) és major que el nivell de significació α = 0.05 = 5%
- Quin tipus d'error (I o II) podeu cometre amb aquesta decisió? Atès que estem acceptant com a certa H_0 quan realment podria ser certa la H_1, podem estar cometent un error de tipus II.
- Quina és la probabilitat que teniu d'equivocar-vos amb la vostra decisió? La probabilitat de cometre aquest error s'anomena β i és un valor que depèn de la diferència real entre les dues mitjanes (valor real de la H_1). No es pot calcular sense saber aquesta dada.
- També tenim informació de l'interval de confiança per a la diferència de les mitjanes. Quant val l'estimació, a un nivell de confiança del 95%, de la diferència $\mu_1 - \mu_2$ de mitjanes poblacionals?
  IC(95%) = (–3.273034, 1.933034).
- Us sembla lògic que aquest interval de confiança tingui els extrems de diferent signe, és a dir, que contingui el valor 0? L'interval inclou el zero i, per tant, la diferència entre les mitjanes poblacionals podria ser zero, és a dir, podrien ser iguals.

- Torneu a repetir l'anàlisi però ara assumint que les variàncies poblacionals són iguals. Comenteu els resultats. Quan estem en el menú `t-test per mostres independents` seleccionem l'opció `Sí` en la pregunta `Assumir variàncies iguals?`. Els resultats (p-valor, IC per la diferència de mitjanes...) han variat una mica (Ex: Graus de llibertat = 18), però la conclusió de l'anàlisi és la mateixa: no hi ha motius suficients per rebutjar la igualtat de mitjanes.

3. Contrast d'igualtat de mitjanes a partir de d'un disseny de dades aparellades

Carreguem les dades del fitxer `oxigenpaired.rda`. A la vista de l'organització de les dades podem dir que estem suposant que les mesures de la variable *metod2* s'han fet sobre les mateixes mostres que la variable *metod1*. Si, com abans, volem comparar les mitjanes poblacionals dels dos mètodes, ara haurem de fer un contrast amb dades aparellades. En aquest cas, les anàlisis descriptives preliminars s'han de realitzar sobre la variable diferència.
Anem al menú:

`Dades, Modifica variables ... Calcula una nova variable ...`

Creem una variable diferència amb el nom *dif*, que sigui igual a l'expressió *metod2* – *metod1*.

- Calculeu els estadístics bàsics i representeu el diagrama de caixa de la variable *dif*. Què suggereixen aquests resultats? Mitjançant el menú `Estadístics, Resums, Resums numèrics ...` i el menú `Gràfics, Caixa de dispersió ...` obtenim els resultats mean = 0.67, sd = 3.461551, n = 10 i el diagrama de caixa. El diagrama de caixa mostra una distribució asimètrica al voltant del zero, amb una amplitud de −5 a +5, aproximadament. Aleshores, l'error estàndard associat a la mitjana és, més o menys, $3.461551/\sqrt{10} = 1.094639$. Clarament, si sumem i restem dos errors estàndards a la mitjana 0.67 ens queden uns límits que inclouen el zero. En resum, els resultats descriptius suggereixen que no hi ha diferència entre els dos mètodes.

Per a realitzar el contrast d'hipòtesis ($\alpha = 0.05$) següent:

$$H_0: \mu_{dif} = 0$$
$$H_1: \mu_{dif} \neq 0$$

Anem al menú:

`Estadístics, mitjanes, t-test per dades aparellades ...`

Sseleccionem com a `Primera variable` la *metod2* i com a `Segona variable`, *metod1*. Deixem les altres opcions amb els seus valors per defecte. Polsem `D'acord`.

- Quant val la mitjana i la desviació estàndard de la variable diferència: *dif* = *metod2* – *metod1*? (Escriviu a la finestra d'instruccions `sd(oxigenpaired$metod2-oxigenpaired$metod1)`).
  $\bar{x}_{dif} = 0.67$ i $s_{dif} = 3.461551$.
- Quant val el valor de l'estadístic de contrast (t_{obs}) d'aquest test?
  $t_{obs} = 0.6121$.
- Quant val el p-valor d'aquest contrast? p-valor = 0.5556.

- A la vista d'aquesta informació, quina és la decisió que preneu? El p-valor és major que $\alpha = 0.05$, per tant, no tenim motius suficients per rebutjar la hipòtesi H_0. La conclusió és que els dos mètodes, en mitjana, treballen igual.
- Amb aquesta decisió, quin tipus d'error –I o II– podeu cometre? Error de tipus II. Quina probabilitat teniu de cometre'l? No la podem conèixer.
- Quant val l'interval de confiança de la diferència?
  IC(95%) = (–1.806244, 3.146244).
- Com s'interpreta el fet que l'interval de confiança tingui els extrems de diferent signe? Implica que el zero pertany a interval i, per tant, que és un valor possible per a μ_{dif}, essent possible que $\mu_{metod2} = \mu_{metod1}$.

4. Contrast d'igualtat de mitjanes a partir d'un model d'anàlisi de la variància (ANOVA)

L'anàlisi de la variància és una tècnica d'inferència estadística per comparar més de dues mitjanes poblacionals. Tot seguit veurem el procés a seguir per a realitzar una anàlisi de la variància amb una sola via de classificació o un sol factor.

Un enginyer electrònic està interessat en l'efecte de diferents tipus de recobriment sobre la conductivitat d'un determinat component d'un tub de raigs catòdics. En un experiment realitzat amb aquest objectiu, s'ha mesurat 4 vegades la conductivitat per a cada tipus de recobriment. S'han obtingut els següents resultats, que heu de carregar del fitxer `recobriment.rda`:

	Tipus de recobriment				
	1	**2**	**3**	**4**	**5**
Conductivitat	143	152	134	129	147
	141	149	133	127	148
	150	144	132	132	144
	146	143	127	129	142

- D'acord amb l'objectiu que persegueix l'investigador, quina és la variable resposta i quina la variable explicativa? Var. resposta = *conduct* i var. explicativa = *TipReco*.
- Quina és la tipologia de cadascuna de les dues variables? *conduct* és una variable contínua, mentre que *TipReco* és categòrica.
- Quins són els nivells (tractaments) de la variable explicativa? Són k = 5 diferents tipus de recobriment.

La pregunta formulada per l'investigador sobre si existeix relació entre el tipus de recobriment i la conductivitat dels components es pot expressar mitjançant el següent contrast d'hipòtesis:

H_0: Les mitjanes poblacionals de la conductivitat dels 5 tipus diferents de recobriment són iguals, és a dir, no hi ha diferències entre els recobriments respecte a la conductivitat:

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

H_1: Almenys una de les mitjanes poblacionals de la conductivitat dels 5 tipus diferents de recobriment és diferent de les altres.

$$\text{no}(\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5)$$

Per contrastar aquestes hipòtesis el procediment més adequat és realitzar una anàlisi de la variància.

- Feu una anàlisi descriptiva numèrica dels estadístics més importants de la variable *conduct* segons el tipus de recobriment. Aneu a `Estadístics, resums, resums numèrics ...` i feu un resum de *conduct* segons la variable *TipReco*. Empleneu la taula següent amb els resultats:

	Tipus de recobriment				
	1	**2**	**3**	**4**	**5**
Mitjana	145.00	147.00	131.50	129.25	145.25
Desviació estàndard	3.915780	4.242641	3.109126	2.061553	2.753785

- De forma totalment intuïtiva, creus que la mitjana poblacional de la variable *conduct* és la mateixa pels 5 tipus de recobriment? A la vista del valor de les mitjanes mostrals i les desviacions i tenint en compte que n = 4 per tots els grups, clarament es veu que les mitjanes 3 i 4 són menors que la resta. Per tant, hi ha indicis de diferència de *conduct* segons *TipReco*. És a dir, el recobriment afecta a la conductivitat.

- De forma totalment intuïtiva, diries que la variància de la variable *cond* és aproximadament la mateixa per a cada tipus de recobriment? Les variàncies (quadrat de les desviacions) semblen diferents perquè n'hi ha que valen aproximadament 16, altres 9 i altres 4. Tanmateix, caldria fer un contrast d'igualtat de variàncies per a confirmar-ho.

- Anàlogament al cas de dues mostres independents, com a anàlisi descriptiva gràfica podríem fer una diagrama de caixa múltiple de *conduct* en funció de *TipReco*. Tanmateix, atès que les mides de les mostres són molt petites (n = 4), també podem fer un gràfic de punts múltiple mitjançant el menú `Gràfics, Gràfic de franja ...` Feu-lo i comenteu-lo. El gràfic concorda amb les interpretacions dels resultats numèrics. El recobriments 3 i 4 donen resultats inferiors de conductivitat. A la vista del gràfic (amplituds) sembla que les variabilitats no són diferents entre les mostres, és a dir, suggereix que les variàncies són iguals.

Per tal de confirmar o no les nostres intuïcions és necessari aplicar procediments estadístics basats en models teòrics. En el nostre cas, el model a utilitzar és l'ANOVA:

`Estadístics, Mitjanes, Anova d' un factor ...`

Observareu que es crearà un model amb nom `AnovaModel.1`. Com que el nostre arxiu només conté dues variables, ja estan marcades per escollir-les. Deixeu les opcions tal com apareixen per defecte i polseu `D'acord`.

A partir de la sortida que proporciona R-Commander responeu:

- Estimació puntual de σ (arrel quadrada de *Mean Sq Residuals* o *Residual standard error*). $\hat{\sigma} = \sqrt{10.967} = 3.311646$.
- Valors de les sumes de quadrats SSTotal, SSTract i SSError.
  SSTotal = 1318.8, SSTract = 1154.3 i SSError = 164.5.
- Graus de llibertat (DF) associats a les anteriors sumes de quadrats:
  - Graus de llibertat de SSTotal. n – 1 = 19.
  - Graus de llibertat de SSTract. k – 1 = 4.
  - Graus de llibertat de SSError. n – k = 15.
- Valors de les mitjanes de quadrats MSTract i MSError.
  MSTract = 288.575 i MSError = 10.967.
- Valor de l'estadístic contrast F_{obs} = MSTract / MSError. $F_{obs} = 26.314$.
- Quina distribució segueix aquest estadístic de contrast? F-Fisher amb 4 graus de llibertat en el numerador i 15 graus de llibertat en el denominador.
- Quant val el p-valor del contrast? p-valor = $1.255 \cdot 10^{-6}$.
- Quina ha de ser la nostra decisió? El p-valor és inferior a un $\alpha = 0.05$, per tant, rebutgem la H_0 i afirmem que alguna mitjana poblacional és diferent.

Podem obtenir anàlisis de comparacions múltiples (grups 2 a 2) si en el menú `Anova d'un factor` activem l'opció `Comparacions dos a dos de les mitjanes`.

- Quins recobriments donen, en mitjana, conductivitats iguals? Quins diferents? Tant en els contrastos 2 a 2 com en els intervals de confiança 2 a 2 veiem que els recobriments 1, 2 i 5 són igual entre si i diferents dels recobriments 3 i 4, els quals són iguals entre ells.

Si anem al següent menú, obtindrem el gràfic que podem veure en la **figura 2**:

`Models, Gràfics, Gràfic dels efectes ...`

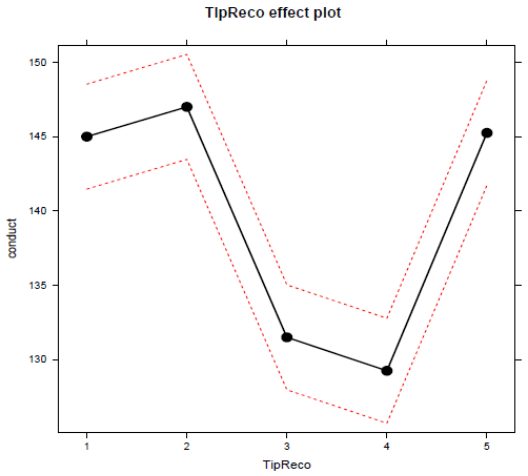


Figura 2: Gràfic d'efectes: mitjana de conductivitat segons el tipus de recobriment.

En aquest gràfic hi ha representades les mitjanes i una banda de confiança que permet avaluar l'efecte de la variable explicativa en la variable resposta.

Un model ANOVA es pot aplicar a les nostres dades si els residus es distribueixen segons una llei Normal, si la variància dels residus és la mateixa per a tots els tractaments i si aquests són independents entre si.

R-Commander disposa d'una opció del menú que permet fer una diagnosi gràfica del model:

```
Models, Gràfics, Gràfics bàsics de diagnòstic ...
```

Mitjançant el gràfic que es troba a l'esquerra en la fila superior avaluem si els errors són independents del valor de la mitjana de cada grup. El gràfic del seu costat ens serveix per avaluar la normalitat dels residus. Per últim, el gràfic que està a la dreta en la fila inferior ens informa de la homocedasticitat dels residus (variàncies iguals).

- Jutgeu, a partir d'aquests gràfics, si es verifiquen les condicions necessàries per poder aplicar el model ANOVA. En el primer gràfic observem que el patró dels residus és més o menys el mateix, independentment dels valors de mitjana dels grups. No observem cap tendència. Sembla que l'amplitud és una mica major per a valors grans de les mitjanes. En el gràfic de normalitat, els punts estan força a prop de la línia de normalitat, suggerint que la distribució dels residus és aproximadament normal. En el darrer gràfic observem que, pels recobriments 3 i 4, l'amplitud és una mica major que per els altres recobriments. Tanmateix, els patrons són força similars.

5. Construcció d'un diagrama Q-Q

El diagrama Q-Q és un procediment gràfic per avaluar heurísticament si un determinat conjunt de dades empíriques es pot ajustar per una determinada distribució teòrica de probabilitat. Veurem els casos de les distribucions normal, exponencial i Weibull. Tot i que R-Commander disposa d'una comanda automàtica per generar els diagrames Q-Q, primer de tot veurem quins passos s'han de seguir per fer-los manualment.

5.1. Diagrama Q-Q per passos

Estudiarem si les dades de l'arxiu `vel.llum` provenen d'una distribució normal. En primer lloc, recupereu les dades i calculeu un resum numèric (mitjana, mediana i desviació tipus) de les dades de la mostra anant a:

```
Estadístics, Resums, Resums numèrics ...
```

D'aquí podem extreure la següent informació:

- Mitjana. Mitjana = 299896.7
- Mediana. Mediana = 299930
- Desviació tipus. Desviació tipus = 111.9098

Feu també un anàlisi gràfic (histograma i diagrama de caixa) de les dades de la mostra anant a:

`Gràfics, Histograma ... i Gràfics, Caixa de dispersió ...`

Responeu a les següents preguntes:

- Hi ha dades anòmales? Identifiquem la dada de la fila 14 (299650) com a dada atípica.
- Quina és la forma de l'histograma? En l'histograma s'aprecia asimetria a l'esquerra (pot ser deguda a la presència de la dada atípica). Aquesta asimetria també apareix quan comparem la mediana i la mitjana (la mitjana pren un valor inferior a la mediana).

Anem a comprovar si les dades de la mostra poden provenir d'una N(μ = 299897, σ = 111,9098). En aquesta part treballarem des de la finestra d'instruccions.

1) Primer de tot hem d'ordenar les dades. Escriviu a la finestra d'instruccions:

```
vel.llum$ord=sort(vel.llum$v.llum)
```

Polseu `Executar` (no repetirem aquesta ordre a la resta de la secció).

2) Calculem el percentil que correspon a cada dada. Fem servir la funció `rank()`, que assigna un ordre a cada dada (dades repetides tindran el mateix ordre):

```
(rang=rank(vel.llum$ord))
```

Recordeu que el parèntesis al principi i al final de la comanda fa que es mostrin els resultats de la comanda.

3) Calculem els percentatges acumulats mitjançant la fórmula (rang – 0.5)/n:

```
(per.acu=(rang-0.5)/length(vel.llum$ord)
```

4) Mitjançant la inversa de la funció de distribució teòrica, trobem els valors teòrics que representen el mateix percentil en el model teòric. En R-Commander, per calcular els valor corresponents a les probabilitats donades per `per.acu`, fem servir el menú:

`Distribucions, Distribucions contínues, Distribució normal, Quantils normals ...`

On escrivim per.acu en la casella de les probabilitats, i els valors de $\mu = 299896,7$ i $\sigma = 111,9098$.

Si les dades es poden ajustar bé per una llei normal, aleshores els valors obtinguts haurien de ser similars als valors del vector v.llum.ord. Per a comparar-los farem un gràfic de dispersió. Abans de fer el gràfic necessitem crear una columna, que anomenarem vel.llum$teor, on carregarem els valors teòrics que hem obtingut. La manera més ràpida de fer-ho és anar a la finestra d'instruccions i executar la instrucció:

```
vel.llum$teor=qnorm(c(per.acu),mean=299896.7,sd=111.9098,
lower.tail=TRUE)
```

5) Per realitzar un diagrama de punts –dades vs valors teòrics– i valorar heurísticament el nivell d'ajust del núvol de punts a una recta ho farem amb un diagrama de dispersió. Primer cal refrescar el conjunt de dades des del menú:

```
Dades, Taula de dades activa, Refresca la taula ...
```

A continuació, anem al menú:

```
Gràfics, Diagrama de dispersió ...
```

Escollim *ord* com a variable del eix OX i *teor* com a variable de l'eix OY. De les opcions deixem només activada la que ens dibuixa la línia de mínims quadrats. El gràfic hauria de ser similar a la **figura 1**.

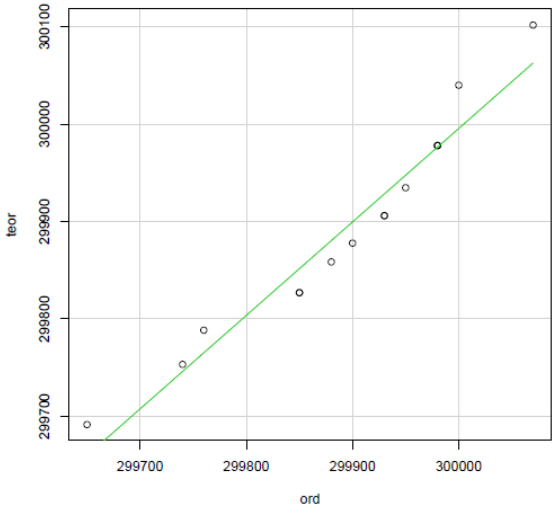


Figura 1: Gràfic Q-Q v.llum.

Si els punts del gràfic s'alineen formant una recta i aquesta recta passa per la bisectriu del quadrant, aleshores el gràfic està suggerint que les dades provenen d'una normal.

5.2. Diagrama Q-Q amb R-Commander

Amb R-Commander es pot fer el gràfic quantil-quantil d'una manera molt senzilla, sense necessitat de fer els passos anteriors. Si no teniu carregat l'arxiu `v.llum.rda`, feu-ho ara.

Per obtenir un gràfic Q-Q aneu a:

```
Gràfics, Gràfic quantil-quantil ...
```

Escolliu la variable (en quest cas *v.llum*) i marqueu la distribució normal. Per defecte, els paràmetres μ i σ són la mitjana i desviació tipus mostrals. La sortida ens mostra un gràfic similar a l'anterior, però amb diferències importants: la recta representa on haurien d'estar situats els punts per a què l'ajust fos perfecte (una banda al voltant, que limita on poden estar situats els punts per poder confirmar la normalitat de les dades) i els valors de l'eix OX són els valors teòrics de la llei normal estandarditzada N(0;1).

- Comenteu l'ajust per una llei Normal. En aquest cas, totes les dades es trobem dins la zona de normalitat i els punts semblen estar distribuïts al voltant de la recta sense seguir cap patró. En tot cas, s'aprecia que els punts a les cues estan tots per sota, mentre que els punts centrals semblen ajustar millor. Tanmateix, una mostra de mida 15 no permet jutjar gràficament l'ajust.

Fixeu-vos que en la finestra d'instruccions el gràfic Q-Q s'ha cridat utilitzant l'expressió `qqPlot(vel.llum$v.llum, dist="norm")`. El paràmetre `dist` representa la distribució a analitzar.
Anem a analitzar si les dades de la mostra de la velocitat de la llum poden ajustar-se per una llei exponencial. Aneu al menú:

```
Gràfics, Gràfic quantil-quantil ...
```

Escolliu la variable *v.llum* i marqueu l'opció `Altre`. En la casella `Especificar` escrivim `exp` i en la casella `Paràmetres` escrivim `1/mean(vel.llum$v.llum)`. En la finestra d'instruccions ha d'aparèixer l'ordre:

```
qqPlot(vel.llum$v.llum, dist="exp", 1/mean(vel.llum$v.llum))
```

- A la vista del gràfic Q-Q obtingut, comenteu l'ajust per una llei Exponencial. El gràfic suggereix que l'ajust no és gens bo, atès que tres dels punts de la cua de l'esquerra estan fora de les bandes i els punts de la cua de la dreta es separen molt de la recta.

- Repetiu l'anàlisi per una llei Weibull, tot especificant `weibull` i escrivint `shape=2, scale=4` en la casella dels paràmetres. L'ajust és millor que en el cas exponencial però pitjor que en el normal. Tres punts de la cua de l'esquerra cauen fora de les bandes.

6. Contrastos de bondat d'ajust

En els contrastos de bondat d'ajust, les hipòtesis de partida en tots els casos són:

H_0: La llei ajusta bé les dades.
H_1: La llei no ajusta bé les dades.

Abans de res carreguem el plug-in HH. Per fer-ho aneu al menú:

```
Eines, Carrega plugins del Rcmdr ...
```

Escolliu el plug-in RcmdrPlugin. HH. Acepteu i reinicieu el R-Commander.

Per executar alguns contrastos de normalitat cal tenir d'instal·lat i carregat el paquet nortest. Aneu a la finestra d'instruccions i executeu les comandes:

```
install.packages("nortest")

library("nortest")
```

D'una banda podem fer un contrast de normalitat mitjançant el contrast χ^2, que quantifica la diferència entre la distribució observada i la teòrica. Per fer un contrast χ^2 cal executar la següent comanda del paquet nortest en la finestra d'instruccions:

```
pearson.test(vel.llum$v.llum)
```

On vel.llum$v.llum és el vector que conté les dades de la velocitat de la llum de les que en volem contrastar la normalitat.

- Comenteu els resultats obtinguts. Pearson chi-square normality test: p-value = 0.1023. El p-valor és major de 0.05. Aleshores, amb un nivell de significació del $\alpha = 0.05$ podem dir que no tenim motius suficients per rebutjar que el model normal ajusti bé aquestes dades.

També podem fer el contrast de Kolmogorov-Smirnov (K-S), que verifica si una mostra pot provenir d'un determinat model. Aneu a la finestra d'instruccions i escriviu i executeu, respectivament, les ordres:

```
ks.test(vel.llum$v.llum, "pexp",1/mean(vel.llum$v.llum))

ks.test(vel.llum$v.llum, "pnorm",mean(vel.llum$v.llum),
sd(vel.llum$v.llum))

ks.test(vel.llum$v.llum, "pweibull",2,4)
```

- Què estem calculant amb aquestes ordres? Què representen els seus paràmetres? Estem realitzant el contrast K-S per a analitzar si els models exponencial, normal i Weibull ajusten bé les dades. El paràmetres de la comanda són els paràmetres de les corresponents distribucions: la lambda de l'exponencial, la mitjana i desviació de la normal i la forma-escala de la Weibull.

- Comenteu els resultats obtinguts. Els p-valors són, respectivament, 0.00001259, 0.8864, $1.872 \cdot 10^{-13}$. Per tant, en els casos de l'exponencial i la Weibull rebutgem la hipòtesi nul·la. En el cas de la normal no la rebutgem. Dels tres models, només el normal ajusta bé les dades.

Finalment, podem fer un contrast d'Anderson-Darling (A-D) per contrastar la normalitat, que es diferencia de l'anterior per donar un pes més important a les cues de la distribució. La comanda que hem d'utilitzar és:

```
ad.test(vel.llum$v.llum)
```

- Comenteu els resultats obtinguts. El p-valor és `p-value = 0.3058`, per tant, no tenim motius suficients per rebutjar que el model normal ajusti bé les dades.

TEMA 6: Relació lineal entre dues variables

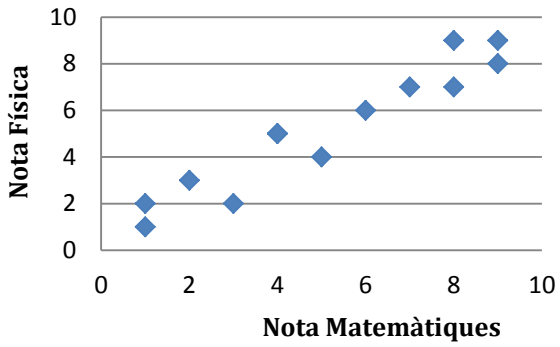
TEORIA

1. Diagrames de dispersió

Els diagrames de dispersió o núvols de punts són representacions gràfiques que serveixen per analitzar quin **tipus de relació** existeix **entre dues variables numèriques**. Es tracta de representar els valors que prenen aquestes variables numèriques sobre els individus d'una mostra.

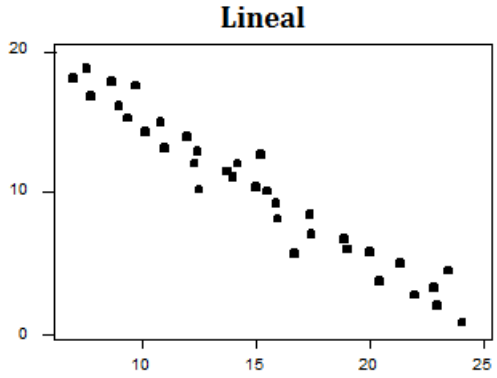
S'escriuen les unitats d'una de les variables sobre l'eix horitzontal i les de l'altra sobre l'eix vertical. Cada observació bivariant es representa mitjançant un punt:

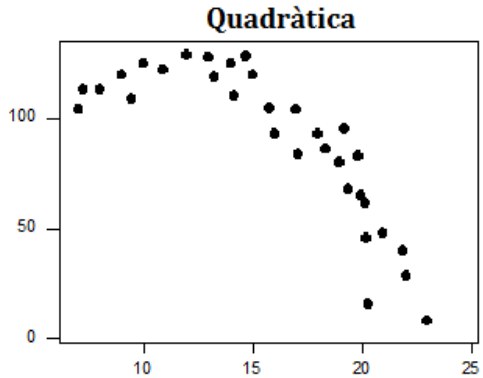
Nota Matemàtiques	8	2	4	7	9	8	1	3	5	6	1	9	4
Nota Física	7	3	5	7	8	9	1	2	4	6	2	9	5

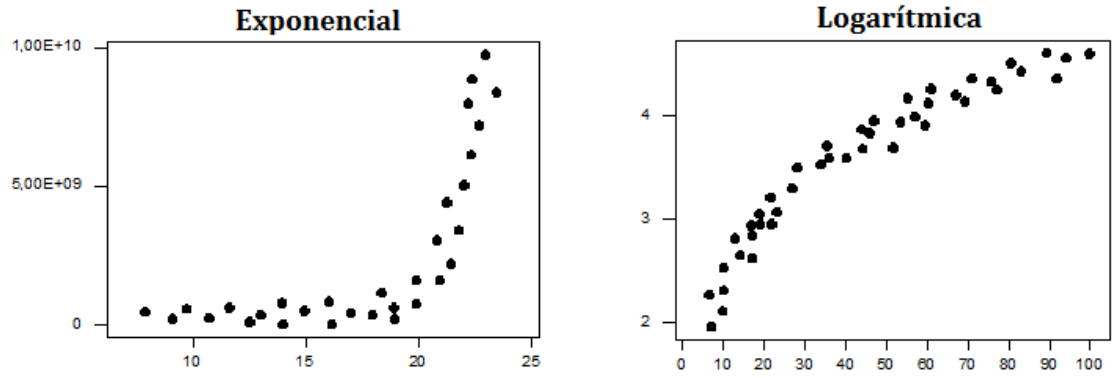


1.1. Tipus d'associacions entre dues variables

Existeixen diversos tipus d'associacions entre dues variables:







En aquest tema s'estudiarà la **relació lineal entre variables**.

2. Correlació lineal

La correlació estadística determina la **força** i **direcció** d'una **relació lineal** entre dues variables numèriques X i Y.

2.1. Covariància mostral, S_{XY}

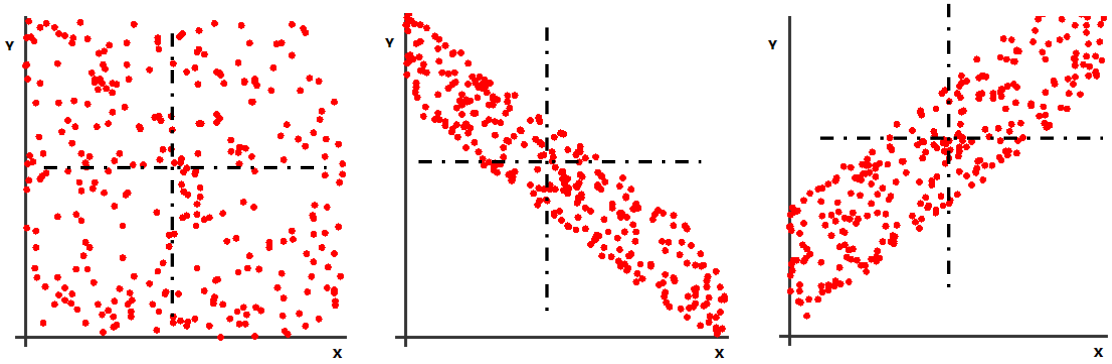
Donades n observacions bivariants de dues variables numèriques X i Y, la covariància mostral S_{XY} és la mitjana corregida dels productes centrats creuats:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

En el cas que X = Y, la covariància mostral és la variància mostral: $s_{XX} = s_X^2$

La covariància mostral ens permet afirmar que:

- Si $s_{XY} = 0$, X i Y no estan relacionades linealment.
- Si $s_{XY} \ll 0$, quan X augmenta Y tendeix a disminuir.
- Si $s_{XY} \gg 0$, quan X augmenta Y tendeix a augmentar.



Un inconvenient de s_{XY} és que el seu valor **depèn de la magnitud o escala de les variables X i Y**, per tant, no ens és útil per avaluar la intensitat d'una relació lineal.

2.2. Índex de correlació lineal de Pearson, r

El coeficient o índex de correlació lineal de Pearson r és igual a la covariància de les variables estandarditzades. Per tant, el seu valor **no depèn de la magnitud o escala de les variables X i Y**: $r = s_{Z_X Z_Y}$

Podem calcular r com:

$$r = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2}\cdot\sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Essent els valors que pot prendre: $-1 \leq r \leq 1$

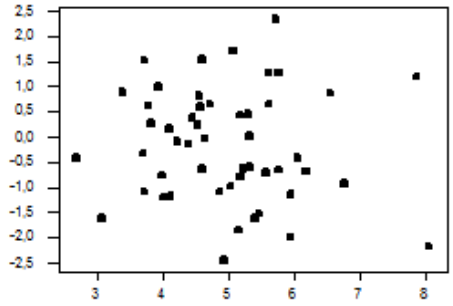
El coeficient de correlació lineal de Pearson ens permet afirmar, igual que la covariància mostral, que:

- Si $r = 0$, X i Y no estan relacionades linealment.
- Si $r \gg 0$, quan X augmenta Y tendeix a augmentar.
- Si $r \ll 0$, quan X augmenta Y tendeix a disminuir.

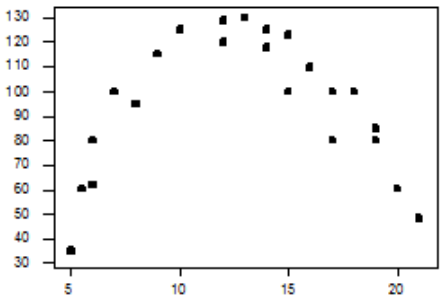
No obstant, a diferència de s_{XY}, r sí que mesura la intensitat d'una relació lineal, ja que no depèn de la magnitud de les variables. Per tant, podem afirmar que quan un diagrama de dispersió es pot ajustar bé per una recta aleshores $|r| \cong 1$.

En els següents gràfics es pot veure com varia el valor de r depenent del tipus de diagrama de dispersió:

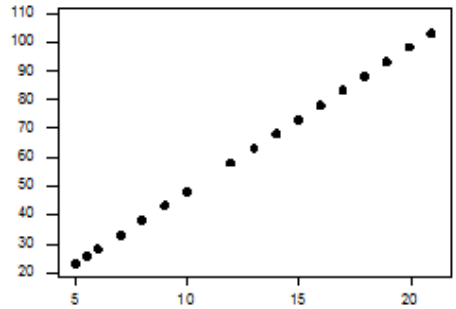
no hi ha cap mena de relació → r = –0.079



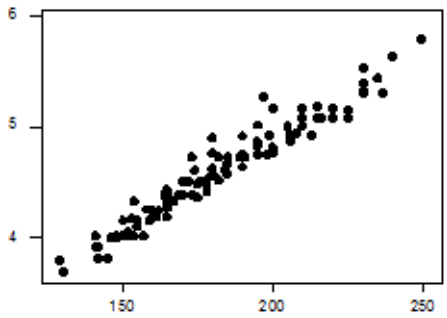
no hi ha una relació lineal → r = –0.079



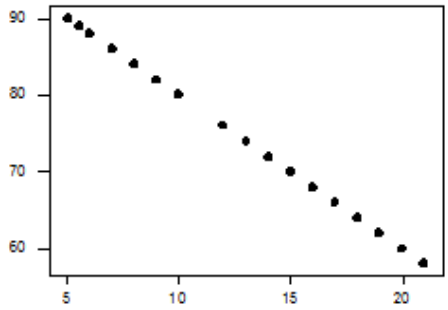
relació lineal positiva perfecta → r = 1



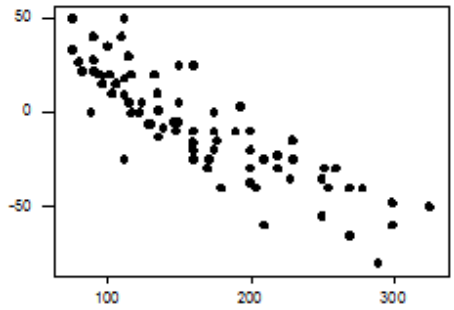
relació lineal positiva → r = 0.972



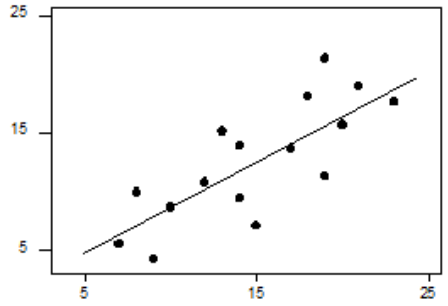
relació lineal negativa perfecta → r = –1

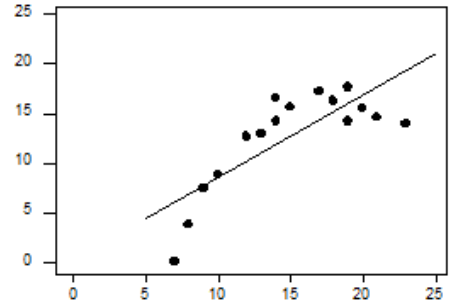


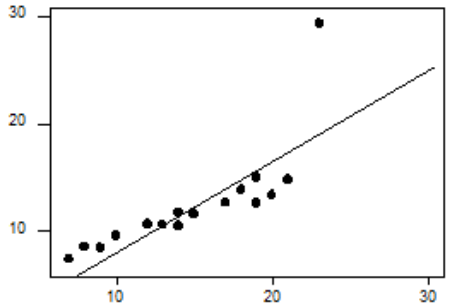
relació lineal negativa → r = –0.862

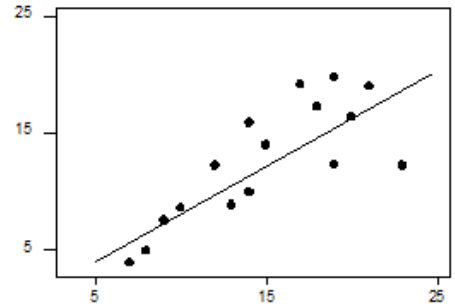


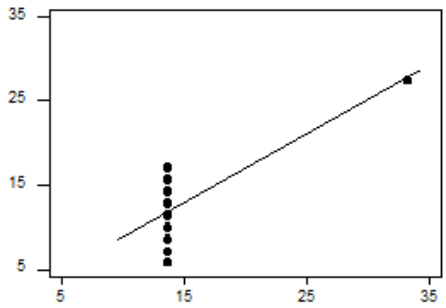
Un inconvenient de r és que el seu valor es veu molt influenciat per valors atípics. En podem veure un exemple en els següents diagrames de dispersió, en què en tots els casos r = 0.79:











Anàlogament com hem fet intervals o contrastos de mitjanes o de variàncies, per a tenir informació poblacional a partir de les dades mostrals o bé acceptar o rebutjar una determinada hipòtesi nul·la, podem fer el mateix amb la correlació d'una mostra de dades (r) vers la correlació de la població, que designarem per ρ.

Per a contrastar si hi ha relació o no entre les variables, contrastarem les hipòtesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Aquest contrast té com a estadístic de contrast:

$$\frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$$

i segueix una llei t-Student amb $\nu = n - 2$ graus de llibertat. Aquest ens permetrà calcular el p-valor i prendre una decisió.

- Si **p-valor** és suficientment gran, considerarem que no hi ha relació lineal entre les variables, és a dir que són linealment independents. Per això direm que "no tenim motius suficients per rebutjar la H_0".

- Si **p-valor** es suficientment petit considerarem que sí hi ha relació lineal entre les variables, i per tant direm que "tenim motius suficients per rebutjar la H_0 i per tant acceptem la H_1".

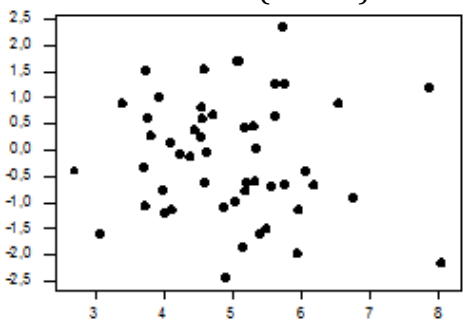
2.3. Coeficient de determinació, r^2

A partir de r podem calcular el coeficient de determinació lineal r^2, que ens diu la qualitat de l'ajust de la recta de regressió a les dades experimentals. Aquest estadístic dóna la proporció de la variació de la variable X, explicada per la relació lineal amb la variable Y.

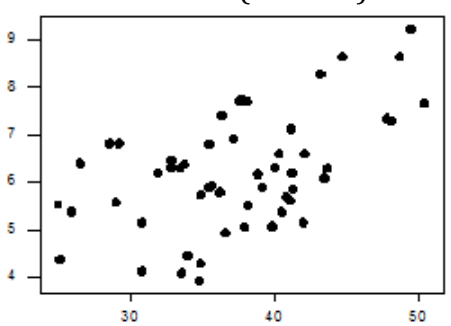
Com que $-1 \leq r \leq 1 \Rightarrow 0 \leq r^2 \leq 1$.

En els següents gràfics podem veure com varia r^2 depenent de la qualitat d'ajust de la recta de regressió al diagrama de dispersió format per cada una de les observacions bivariants de la mostra:

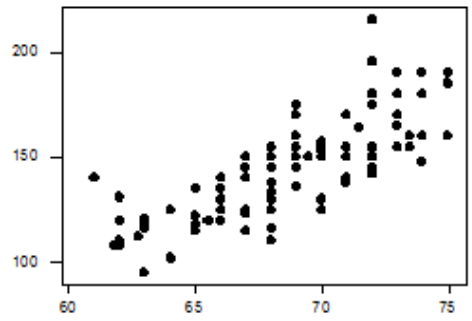
$r^2 = 0.0062$ (0.62%)



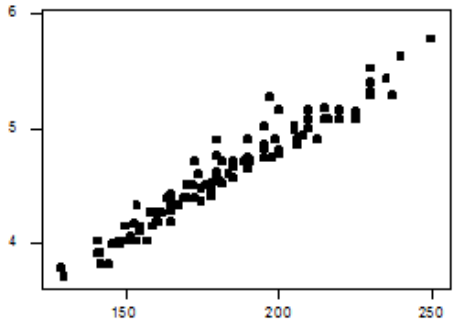
$r^2 = 0.2777$ (27.77%)



$r^2 = 0.6162$ (61.62%)



$r^2 = 0.9445$ (94.45%)



3. Regressió lineal

Si s'observa una forta relació entre dues variables X i Y, es pot utilitzar una funció que s'ajusti a les dades experimentals per representar aquesta relació. Com que en aquest tema estem parlant d'una relació de tipus lineal, la funció més adequada per representar-la és una recta, que s'haurà d'estimar a partir de les dades d'una mostra. Trobar el valor d'aquesta recta significa realitzar una **regressió lineal**, que ens pot servir per obtenir valors d'una de les variables quan no coneixem el valor de l'altra variable.

Abans de realitzar una regressió lineal entre dues variables, s'ha d'escollir quina és la **variable explicativa X** i quina la **variable resposta Y**, de manera que canvis en la variable de resposta Y s'han de poder explicar a partir de canvis en la variable explicativa X.

Cal remarcar que, tot hi haver-hi una associació entre X i Y, això no significa que X hagi de ser obligatòriament la causa de Y, ja que sovint hi intervenen o poden intervenir-hi altres variables. Per exemple, podria ser que una variable Z fos la causa de X i de Y o que X i Z fossin les causes de Y.

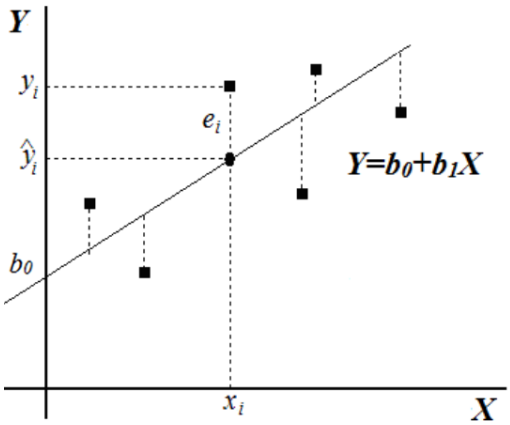
3.1. Càlcul de la recta de regressió

Partint d'una mostra de n observacions bivariants: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

El **mètode dels mínims quadrats** és el mètode que utilitzarem per ajustar n dades d'una mostra a una recta d'equació:

$$\widehat{\mathbf{Y}} = \mathbf{b_0} + \mathbf{b_1 X}$$

Fent un diagrama de dispersió com el que es pot veure a continuació, la recta de regressió es troba tenint en compte que la suma dels quadrats de les ordenades o distàncies verticals dels punts de la mostra a la recta ha de ser mínima: $\hat{y}_i$



- y_i: ordenada del punt (x_i, y_i) de la mostra.
- $\hat{y}_i = b_0 + b_1 x_i$: ordenada sobre la recta de regressió del punt d'abscissa x_i. Per tant, es tracta del **valor ajustat** o **estimat** de la variable Y per $X = x_i$.
- $e_i = \hat{y}_i - y_i$: distància vertical entre el punt (x_i, y_i) de la mostra i el punt $(x_i, \hat{y}_i)$ estimat o ajustat sobre la recta. Aquest valor s'anomena **error** o **residu** per $X = x_i$.

Segons el mètode dels mínims quadrats, el següent resultat en funció dels valors b_o i b_1, ha de ser mínim:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (b_0 + b_1 x_i - y_i)^2$$

Per altra banda, podem assignar les següents **propietats** a la recta de regressió:

- La recta de regressió passa sempre pel centre de gravetat $(\bar{x}, \bar{y})$ del núvol de punts.
- Es compleix que:

$$\sum_{i=1}^{n} e_i = 0 \Rightarrow \bar{e} = 0$$

4. Model de regressió lineal simple, MRLS

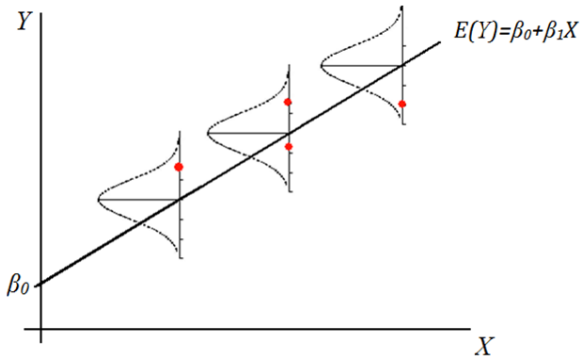
Un model de regressió lineal simple MRLS és aquell que estudia la relació lineal entre la variable resposta Y i la variable explicativa X d'una població que té definida la següent relació entre aquestes dues variables:

$$\mathbf{Y = \beta_0 + \beta_1 X + \varepsilon} \quad \text{on} \quad \boldsymbol{\varepsilon} \sim \mathbf{N(0; \sigma)}$$

El valor que pren Y és un valor obtingut com a la suma de dues parts:

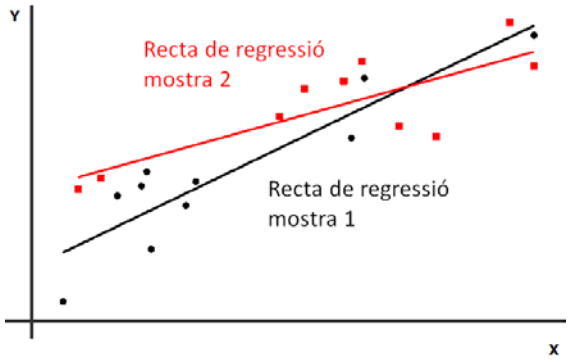
- Part sistemàtica $\beta_0 + \beta_1 X$, que ens dóna una funció entre Y i X anomenada **funció de regressió** de Y sobre X i que es correspon amb l'esperança de Y, és a dir, $\mathbf{E\{Y\} = \beta_0 + \beta_1 X}$.
- Part aleatòria ε, que segueix una llei Normal (per això el MLRS és normal) amb variància σ^2 constant i igual per a cada valor de X. Així, podem afirmar

que per X = x_i, els valors de Y es trobaran dins l'interval $(\beta_0 + \beta_1 x_i) \pm 3\sigma$ amb una probabilitat del 99.7%



Com que no és possible conèixer exactament ni el valor de β_0 (ordenada en l'origen de la població) ni el valor de β_1 (pendent de la població) ni σ, aquests paràmetres s'estimen a partir de b_0 (ordenada en l'origen de la recta de regressió) i b_1 (pendent de la recta de regressió), respectivament.

No obstant, els estimadors puntuals o coeficients de la recta de regressió varien segons la mostra obtinguda, és a dir, podem parlar d'una certa variabilitat mostral a l'ajustar els punts d'una mostra a una recta:



Així, el MRLS ens permet trobar una estimació de σ^2, així com les distribucions de probabilitat mostrals de l'estimador β_0 (b_0), l'estimador β_1 (b_1) i l'estimador σ^2.

4.1. Estimació de σ^2

L'estimador puntual de σ^2 és **MSE** o MSErr, que es coneix amb el nom de mitjana dels quadrats dels residus o variància residual S_R^2. MSE es calcula:

$$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{\sum e_i^2}{n-2} = \hat{\sigma}^2$$

On el valor SSE o SSErr s'anomena suma dels quadrats dels residus i podem dir que es tracta d'una mesura de la variabilitat dels e. Com que $\bar{e} = 0$, com hem explicat anteriorment:

$$\text{SSE} = \sum (e_i - \bar{e})^2 = \sum e_i^2$$

4.2. Estimació de β_1

L'estimador puntual de β_1 és b_1, que es calcula de la següent manera:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{s_{XY}}{s_X^2} = \frac{r \cdot s_X \cdot s_Y}{s_X^2} = r \cdot \frac{s_Y}{s_X}$$

L'esperança i la variància de b_1 valen, respectivament:

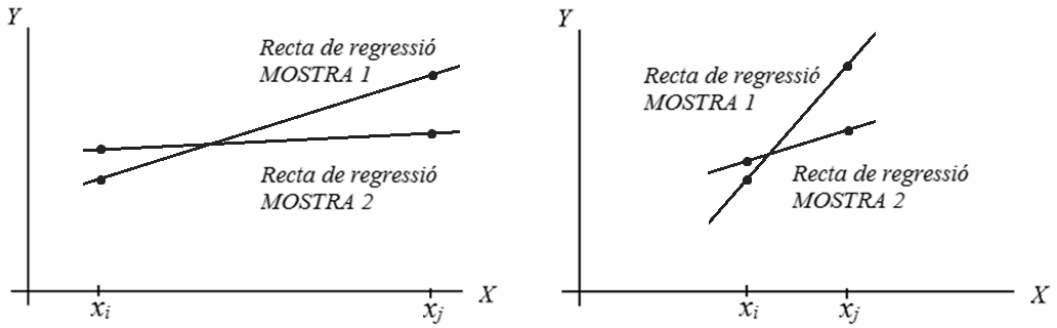
$$E\{b_1\} = \beta_1$$

$$var\{b_1\} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Tot i així, $var\{b_1\}$ no es pot calcular, ja que no es coneix amb exactitud el valor de σ, sinó que s'estima gràcies a $MSE = \hat{\sigma}^2$, per això podem estimar $var\{b_1\}$ per:

$$s^2\{b_1\} = \frac{MSE}{\sum(x_i - \bar{x})^2} = \frac{MSE}{(n-1)s_X^2}$$

Com es pot veure en els següents gràfics, la variància de b_1 és més petita com més separats entre ells estiguin els valors $x_1, x_2, ..., x_n$:



Pel que fa a la distribució de probabilitat de b_1, com que el MRLS és normal, b_1 també segueix una distribució normal. No obstant, com que $var\{b_1\}$ és desconeguda i s'ha d'estimar a partir de $s^2\{b_1\}$, b_1 segueix una llei t-Student amb $\nu = n - 2$ graus de llibertat:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}$$

D'aquesta manera, un $IC(\beta_1, 1 - \alpha)$ de β_1 ve donat per:

$$b_1 \pm t_{n-2,\alpha/2} \cdot s\{b_1\}$$

$$b_1 \pm t_{n-2,\alpha/2} \cdot \sqrt{\frac{MSE}{(n-1)s_X^2}}$$

- **Contrast per comprovar si $\beta_1 = 0$**

Per comprovar si el pendent β_1 de la població és nul o no, és a dir, si existeix alguna relació entre les variables X i Y, ens hem de plantejar la següent hipòtesi:

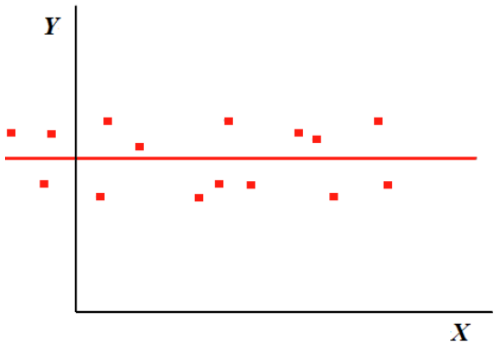
$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Aquest contrast té com a estadístic de contrast:

$$\frac{b_1 - \beta_1}{s\{b_1\}} = \frac{b_1 - 0}{s\{b_1\}} = \frac{b_1}{s\{b_1\}} \sim t_{n-2}$$

Fixat un nivell de significació α, rebutjarem H_0 si el p-valor < α, o el que és el mateix, si el valor de l'estadístic de contrast $|t_{obs}| > |t_{n-2,\alpha/2}|$. Això també és equivalent a veure si l'interval IC(β_1,1 – α) conté el valor zero.

En el cas que acceptéssim H_0 estaríem dient que no existeix cap tipus de relació entres les variables X i Y, com es pot veure en el diagrama següent:



4.3. Estimació de β_0

L'estimador puntual de β_0 és b_0. Aquest valor es calcula:

$$b_0 = \bar{y} - b_1\bar{x}$$

L'esperança i la variància de b_0 valen, respectivament:

$$E\{b_0\} = \beta_0$$

$$\text{var}\{b_0\} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)$$

Com en el cas de b_1, no es coneix amb exactitud el valor de σ, sinó que s'ha d'estimar gràcies a $\text{MSE} = \hat{\sigma}^2$. D'aquesta manera podem estimar $\text{var}\{b_0\}$ per:

$$s^2\{b_0\} = \text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right) = \text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right)$$

La distribució de probabilitat de b_0 funciona com en el cas anterior. Per tant, b_0 segueix una llei t-Student amb $\nu = n - 2$ graus de llibertat:

$$\frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{n-2}$$

D'aquesta manera, un $IC(\beta_0, 1-\alpha)$ de β_0 ve donat per:

$$b_0 \pm t_{n-2,\alpha/2} \cdot s\{b_0\}$$

$$b_0 \pm t_{n-2,\alpha/2} \cdot \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right)}$$

- **Contrast per comprovar si $\beta_0 = 0$**

Per comprovar si el MRLS passa per l'origen ($\beta_0 = 0$) ens hem de plantejar:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

Aquest contrast, com en el cas anterior, té com a estadístic de contrast:

$$\frac{b_0 - \beta_0}{s\{b_0\}} = \frac{b_0 - 0}{s\{b_0\}} = \frac{b_0}{s\{b_0\}} \sim t_{n-2}$$

Fixat un nivell de significació α, rebutjarem H_0 si el p-valor < α, o el que és el mateix, si el valor de l'estadístic de contrast $|t_{obs}| > |t_{n-2,\alpha/2}|$. Equivalentment podríem mirar si $IC(\beta_0, 1-\alpha)$ conté el zero.

5. Contrast de regressió: Anàlisi de la variància

Partint de la mateixa hipòtesi anterior, podem utilitzar un altre mètode per comprovar si el pendent de la població β_1 és nul o no:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

El que hem de fer és una partició de la variabilitat total (SSTot) de la variable Y en dues parts:

$$SSTot = SSReg + SSErr$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

El terme SSReg mesura la variabilitat de Y que queda explicada per la recta de regressió, mentre que SSErr mesura la variabilitat que no pot ser explicada per aquesta.

Podem calcular SSReg a partir de SSTot i SSErr:

$$SSReg = SSTot - SSErr = (n-1)s_Y^2 - (n-2)MSE$$

Ens interessa saber si SSReg és significatiu. Per això, si H_0 és certa, es compleix que el següent quocient segueix una distribució F de Fisher amb $\nu_1 = 1$ grau de llibertat i $\nu_2 = n - 2$ graus de llibertat:

$$\frac{SSReg}{MSErr} \sim F_{1,n-2}$$

Fixat un nivell de significació α, rebutjarem H_0 si el p-valor < α, o el que és el mateix, si el valor de l'estadístic de contrast F (F_{obs}) > $F_{1,n-2,\alpha}$.

Normalment es recullen els resultats del contrast en la **taula d'anàlisi de la variància**:

	Graus de llibertat (DF)	**Suma dels quadrats (SS)**	**Mitjana dels quadrats (MS)**	**Estadístic de contrast (F)**
Regressió	1	SSReg = MSReg	MSReg	$F_{obs} = \frac{MSReg}{MSErr}$
Error	n – 2	$SSErr = MSErr(n-2)$	MSErr	
Total	n – 1	$SSTot = SSReg + SSErr$		

5.1. MSE a partir de r^2

La fracció de la variació de la variable Y que queda explicada per la recta de regressió coincideix amb el quadrat del coeficient de correlació, és a dir, coincideix amb el coeficient de determinació r^2:

$$r^2 = \frac{SSReg}{SSTot}$$

Si a aquesta expressió hi substituïm el valor SSReg anterior i tenim en compte que $SSTot = (n-1)s_Y^2$:

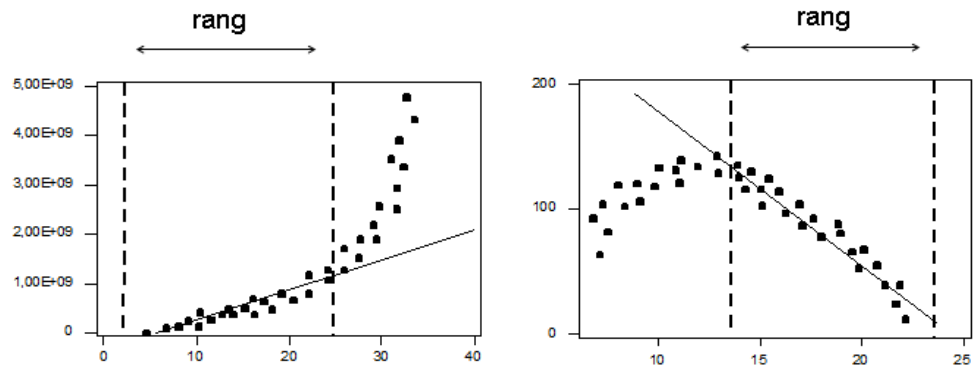
$$r^2 = \frac{(n-1)s_Y^2 - (n-2)MSE}{(n-1)s_Y^2}$$

$$MSE = \frac{s_Y^2(n-1)(1-r^2)}{n-2}$$

6. Prediccions

En realitzar una regressió lineal podem fer dos tipus de prediccions o estimacions: l'estimació de l'esperança E{Y} del MRLS i l'estimació d'un valor de la variable Y.

Cal anar en compte amb les **extrapolacions**: prediccions sobre E{Y} i Y per a valors de X que caiguin fora dels valors observats d'aquesta variable. En aquest cas, com es pot veure en els següents gràfics, podria ser que la relació entre les variables X i Y no fos lineal:



6.1. Predicció de E{Y}

Un MRLS ha de seguir una equació del tipus $Y = \beta_0 + \beta_1 X + \varepsilon$. No obstant, mai podem conèixer amb exactitud β_1 i β_0 i per tant, tampoc podem conèixer el valor de $E\{Y\} = \beta_0 + \beta_1 X$ donat un valor de X. Com que aquests dos coeficients s'estimen gràcies a b_0 i b_1, respectivament, l'estimador puntual de E{Y} és:

$$\hat{Y} = b_0 + b_1 X$$

En aquest cas, donat un valor de X l'esperança i la variància de $\hat{Y}$ valen:

$$E\{\hat{Y}\} = E\{Y\}$$

$$\text{var}\{\hat{Y}\} = \sigma^2\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

Com que tampoc es coneix exactament quin és el valor de σ, sinó que aquest s'estima gràcies a $\text{MSE} = \hat{\sigma}^2$, $\text{var}\{\hat{Y}\}$ s'estima per $s^2\{\hat{Y}\}$:

$$s^2\{\hat{Y}\} = \text{MSE}\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) = \text{MSE}\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{(n-1)s_X^2}\right)$$

La variància de $\hat{Y}$ és més petita com més separats entre ells estiguin els valors x_1, x_2, ..., x_n i és més gran com més allunyat estigui el valor de X de $\bar{x}$.

Com que es tracta d'un MRLS normal, l'estimador $\hat{Y}$ segueix una llei t-Student amb $\nu = n - 2$ graus de llibertat:

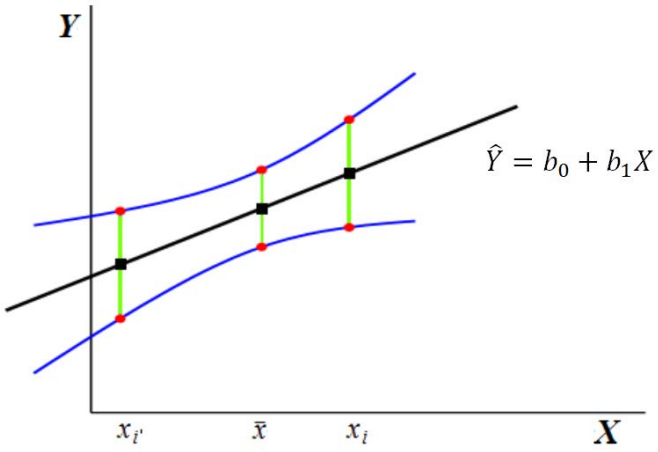
$$\frac{\hat{Y} - E\{Y\}}{s\{\hat{Y}\}} \sim t_{n-2}$$

D'aquesta manera, donat un valor de X, un IC$(1-\alpha)$ de E{Y} ve donat per:

$$\hat{Y} \pm t_{n-2,\alpha/2} \cdot s\{\hat{Y}\}$$

$$(b_0 + b_1X) \pm t_{n-2,\alpha/2} \cdot \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{(n-1)s_X^2}\right)}$$

Unint els extrems de cada IC per cada valor de X s'aconsegueixen les corbes que es poden veure en el gràfic. Podem observar que la precisió de l'estimació disminueix quan la variable X s'allunya de $\bar{x}$, ja que en aquest cas, com hem dit, la variància augmenta:



6.2. Predicció de Y

Partint d'un MRLS que segueix l'equació $Y = \beta_0 + \beta_1 X + \varepsilon$ on $\varepsilon \sim N(0; \sigma^2)$, donat un valor de X podem estimar quan valdrà un valor de la variable Y gràcies a l'estimador puntual:

$$\widehat{Y} = b_0 + b_1 X$$

En aquesta predicció, donat un valor X l'esperança val $E\{\widehat{Y}\} = Y$, mentre que la variabilitat consta de la suma de dues parts:

- La variabilitat deguda a l'estimació de E{Y}: $\sigma^2 \left(\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$
- La variabilitat intrínseca de la distribució de Y, és a dir, la variabilitat dels residus: σ^2

$$var\{\widehat{Y}\} = \sigma^2 \left(\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) + \sigma^2 = \sigma^2 \left(\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum(x_i-\bar{x})^2} + 1\right)$$

Com que σ s'estima gràcies a $MSE = \widehat{\sigma}^2$, $var\{\widehat{Y}\}$ s'estima per $\tilde{s}^2$:

$$\tilde{s}^2 = MSE\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{\sum(x_i-\bar{x})^2} + 1\right) = MSE\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{(n-1)s_X^2} + 1\right)$$

Igual que a l'apartat anterior, la variància és més petita com més separats entre ells estiguin els valors $x_1, x_2, ..., x_n$ i és més gran com més allunyat estigui la variable X de $\bar{x}$.

En aquest cas, com que el MRLS és normal, la distribució de probabilitat funciona com en els casos anteriors, és a dir, l'estimador $\widehat{Y}$ segueix una llei t-Student amb $\nu = n - 2$ graus de llibertat:

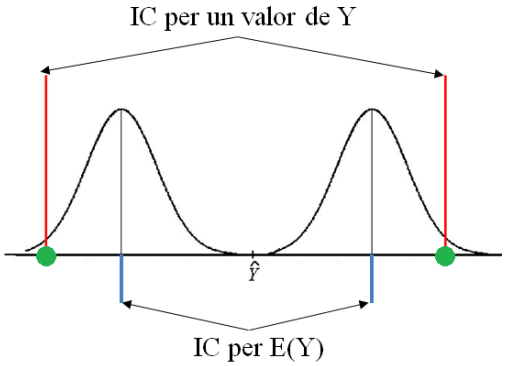
$$\frac{\widehat{Y} - Y}{\tilde{s}} \sim t_{n-2}$$

D'aquesta manera, donat un valor de la variable X un IC(1 – α) de Y ve donat per:

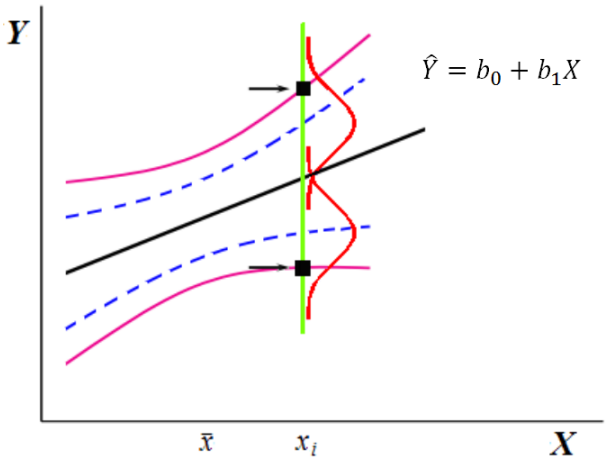
$$\widehat{Y} \pm t_{n-2,\alpha/2} \cdot \tilde{s}$$

$$(b_0 + b_1 X) \pm t_{n-2,\alpha/2} \cdot \sqrt{MSE\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{(n-1)s_X^2} + 1\right)}$$

El que ens diu aquest l'IC és que donat un valor de X, els diversos valors de Y es distribueixen com es pot veure en el gràfic. Així, podem trobar un valor de Y dins l'IC marcat de color vermell amb una probabilitat igual a 1 – α:



Podem observar, igual que a l'apartat anterior, que l'IC de Y (de color rosa al gràfic) augmenta quan la variable X s'allunya de $\bar{x}$, és a dir, la precisió de l'estimació disminueix al augmentar la variància per un valor de Y:



7. Gràfics dels residus

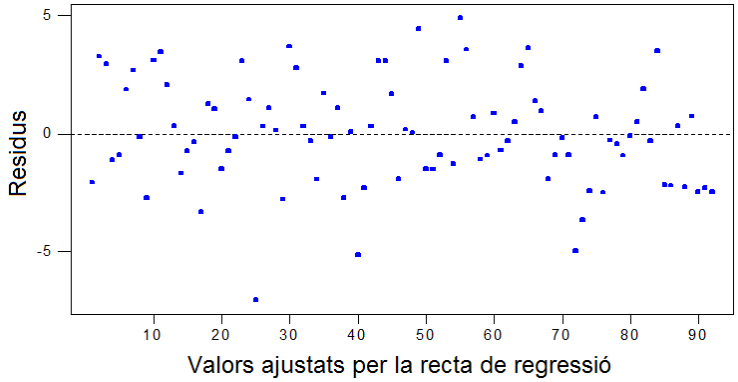
Els contrastos de significació i les estimacions explicades en apartats anteriors parteixen d'un MRLS normal. No obstant, un model d'aquest tipus ha de complir amb un seguit d'hipòtesis que s'ha de comprovar, abans de res, que verdaderament siguin certes. Això ho farem gràcies als **gràfics de residus**.

Els supòsits que s'han de verificar són els següents:

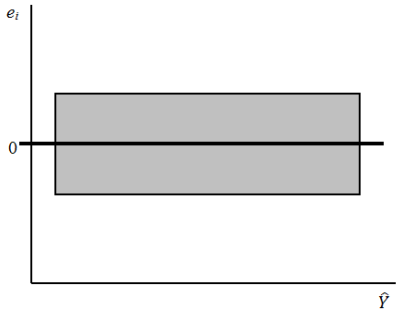
- **Homoscedasticitat**: la variància σ^2 dels residus s'ha de mantenir constant per a cada valor de X.
- **Independència**: no s'ha d'observar cap mena de relació entre els residus.
- **Normalitat**: els residus s'han de distribuir seguint una llei Normal.

7.1. Gràfic per comprovar l'homoscedasticitat i la independència

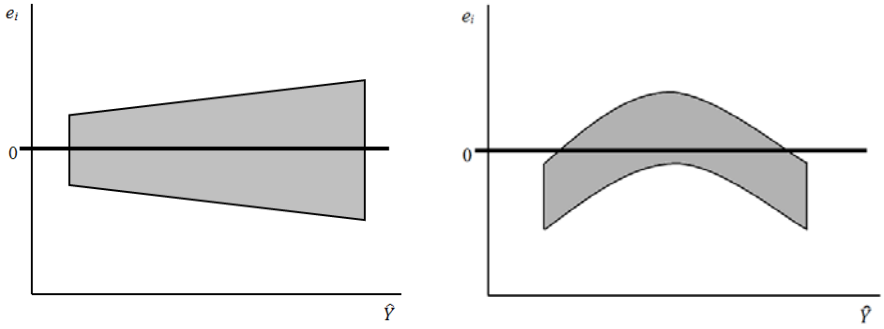
L'homoscedasticitat i la independència dels residus es comproven a través del següent gràfic, on a l'eix de les abscisses s'hi col·loquen els valors ajustats per la recta de regressió ($\hat{Y}$) i a l'eix de les ordenades els valors dels residus (e_i):



Per donar un gràfic com a correcte hauríem d'observar un núvol de punts sense cap mena de relació entre els residus i amb aproximadament la mateixa variabilitat a cada zona del gràfic. Esquemàticament hauríem de veure:



No obstant, a vegades no es compleixen totes les hipòtesis i podem trobar gràfics com els que es mostren a continuació:

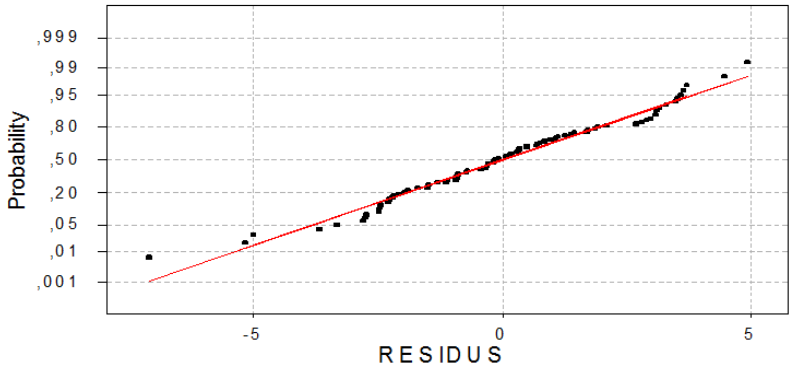


En el gràfic de l'esquerra es pot observar una manca d'homoscedasticitat, és a dir, la variància dels residus no es manté constant, sinó que augment a mesura que augmenten els valors de la variable X. Per altra banda, en el gràfic de la dreta es pot observar una mena de relació entre els residus, és a dir, no es compleix la hipòtesi d'independència dels residus:

7.2. Gràfic per comprovar la normalitat

Per comprovar la hipòtesi de normalitat dels residus ho farem a través dels diagrames de normalitat i contrastos. Així, si observem que els diferents punts del

gràfic es poden ajustar a una línia recta i el p-valor del contrast de normalitat és superior o igual a α, acceptarem que la hipòtesi de normalitat dels residus és certa:

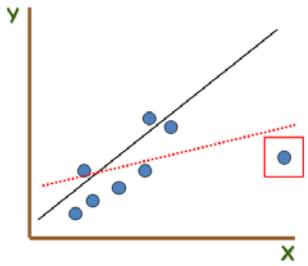


8. Dades atípiques i dades influents

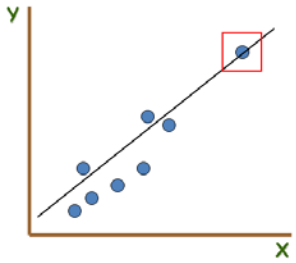
Una **dada atípica** és una dada molt allunyada de la resta d'observacions que es caracteritza per tenir, normalment, un residu gran. Per altra banda, una **dada influent** és una dada que provoca un canvi en la recta de regressió quan es treu del diagrama de dispersió.

Observem diferents casos amb els quals ens podem trobar:

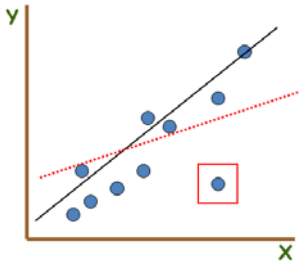
- **Dada atípica i influent:** en aquest cas, la dada marcada també és una dada atípica, però aquest cop sí que és influent. Aquesta dada modifica els coeficients de la recta de regressió de manera que obtenim una recta de regressió molt diferent a l'anterior.



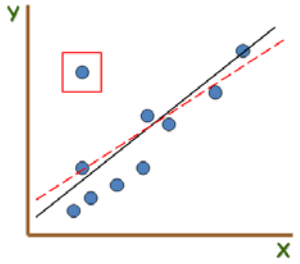
- **Dada atípica i no influent:** en el gràfic es pot observar que la dada marcada és una dada atípica, ja que no es troba dins dels valors observats de les variable X i Y. Aquesta dada no es considera una dada influent perquè està tan a prop de la recta de regressió que a penes la fa variar.



- **Dada no atípica i influent:** aquí, la dada marcada tampoc es considera un valor atípic, ja que igual que en el cas anterior pertany tant als valors observats de X com de Y. No obstant, sí que es tracta d'una dada influent perquè traient-la la recta de regressió es veu significativament modificada.



- **Dada no atípica i no influent:** la dada marcada, encara que es trobi lluny de la recta de regressió, no es considera una dada atípica, ja que tant pertany als valors observats de X com de Y. A més, la seva absència no fa variar significativament els coeficients de la recta de regressió.



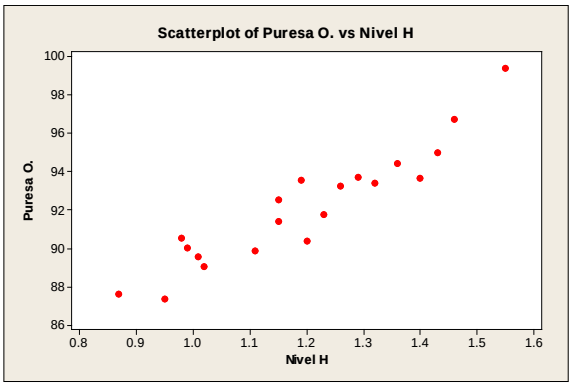
Exemple: *En un procés de destil·lació química, la puresa de l'oxigen produït (en %) depèn del percentatge d'hidrocarburs presents en el condensador principal. S'han pres 20 mostres d'aquest procés:*

Nivell H.	***Puresa O.***
0.99	*90.01*
1.02	*89.05*
1.15	*91.43*
1.29	*93.74*
1.46	*96.73*
1.36	*94.45*
0.87	*87.59*
1.23	*91.77*
1.55	*99.42*
1.40	*93.65*
1.19	*93.54*
1.15	*92.52*
0.98	*90.56*
1.01	*89.54*
1.11	*89.85*
1.20	*90.39*

1.26	*93.25*
1.32	*93.41*
1.43	*94.98*
0.95	*87.33*

1) Estudia, amb una confiança del 95%, si existeix relació lineal entre aquestes dues variables.

En primer lloc, podem realitzar un estudi gràfic, en aquest cas un diagrama de dispersió, per observar quin tipus de relació hi ha entre les variables. Per fer-ho, hem de reconèixer quina és la variable explicativa (X) i quina és la variable resposta (Y). En aquest cas, com que és el nivell d'hidrocarburs que fa variar la puresa de l'oxigen, escollim com a variable explicativa Nivell H. i com a variable resposta Puresa O:



Observem que sí que hi pot haver relació lineal entre les variables.

En segon lloc, realitzem un estudi numèric, és a dir, el càlcul de la recta de regressió (pendent i ordenada en l'origen) i el coeficient de determinació i/o el coeficient de correlació lineal de Pearson:

$b_0 = 74.28$
$b_1 = 14.95$

Recta de regressió: $\hat{Y} = 74.28 + 14.95X$

$r = 0.9367$
$r^2 = 0.8774 = 87.74\%$

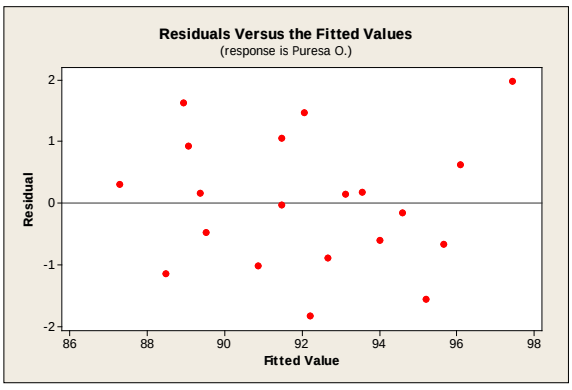
El coeficient de determinació ens confirma que la relació lineal entre les variables és forta, ja que s'apropa al 100% (relació lineal perfecta).

Ara hauríem de verificar que les hipòtesis del MRLS: homoscedasticitat, independència i normalitat dels residus siguin verdaderament certes. Aquest pas el fem a través dels gràfics de residus. Els gràfics de residus no fan servir les variables Nivell H. i Puresa O., per això hem de calcular dues noves columnes: V. Ajustats i Residus. A la columna V. Ajustats hi hem d'escriure els valors de Puresa O. calculats

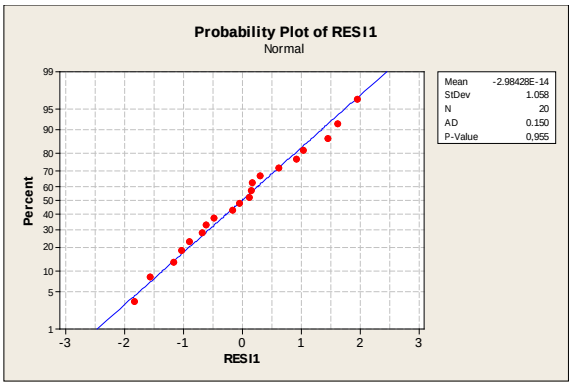
a través de la recta de regressió de l'apartat anterior, mentre que a la columna Residus hem de calcular la diferència Puresa O. – V. Ajustats.

Nivell H.	Puresa O.	V. Ajustats	Residus
0.99	90.01	89.0813	0.92868
1.02	89.05	89.5297	–0.47974
1.15	91.43	91.4729	–0.04292
1.29	93.74	93.5656	0.17444
1.46	96.73	96.1066	0.62337
1.36	94.45	94.6119	–0.16189
0.87	87.59	87.2876	0.30238
1.23	91.77	92.6687	–0.89871
1.55	99.42	97.4519	1.96809
1.40	93.65	95.2098	–1.55979
1.19	93.54	92.0708	1.46918
1.15	92.52	91.4729	1.04708
0.98	90.56	88.9318	1.62816
1.01	89.54	89.3803	0.15973
1.11	89.85	90.8750	–1.02502
1.20	90.39	92.2203	–1.83029
1.26	93.25	93.1171	0.13286
1.32	93.41	94.0140	–0.60399
1.43	94.98	95.6582	–0.67821
0.95	87.33	88.4834	–1.15342

Així, podem dibuixar el gràfic per comprovar l'homoscedasticitat i la independència dels residus, que en aquest cas ens permet donar aquestes dues hipòtesis com a correctes, ja que no s'observa cap mena de relació entre els punts dibuixats:



També podem dibuixar el gràfic per comprovar la normalitat dels residus:



Com que el contrast de normalitat ens dóna un p-valor = 0.955 ≥ α = 0.05, acceptem la hipòtesi que els residus segueixen una distribució de probabilitat normal.

El següent pas tracta de realitzar el contrast de regressió mitjançant l'anàlisi de la variància. Per fer-ho, necessitem omplir la taula corresponent:

$$\text{MSE} = \frac{s_Y^2(n-1)(1-r^2)}{n-2} = \frac{3.021^2 \cdot (20-1)(1-0.8774)}{20-2} = 1.181$$

$$\text{SSReg} = \text{SSTot} - \text{SSErr} = (20-1) \cdot 3.021^2 - (20-2) \cdot 1.181 = 152.1$$

	DF	**SS**	**MS**	**F**
Regressió	1	152.1	152.1	128.9
Error	18	21.25	1.181	
Total	19	173.35		

Busquem a les taules el valor $F_{1,18} = 4.41$. Com que aquest valor és menor a $F_{obs} = 128.9$, no acceptem H_0. Per tant, podem afirmar que $\beta_1 \neq 0$, és a dir, una altra vegada es confirma que existeix una relació entre les variables.

2) Si existeix relació lineal entre les variables Nivell H. i Puresa O., troba per Nivell H. = 1.35 un IC de l'esperança de Puresa O. i un IC de Puresa O. amb una confiança del 95%.

Busquem a les taules el valor $t_{18,0.025} = 2.101$.

IC(95%) de E{Y} per X = 1.35:

$$(b_0 + b_1 x_j) \pm t_{n-2,\alpha/2} \cdot \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x_j - \bar{x})^2}{(n-1)s_X^2}\right)}$$

$$(74.28 + 14.95 \cdot 1.35) \pm 2.101 \cdot \sqrt{1.181\left(\frac{1}{20} + \frac{(1.35 - 1.196)^2}{(20-1)\cdot 0.1893^2}\right)}$$

$$\text{IC}(95\%) = (93.80, 95.13)$$

IC(95%) de Y per X = 1.35:

$$(b_0 + b_1 x_j) \pm t_{n-2,\alpha/2} \cdot \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x_j - \bar{x})^2}{(n-1)s_X^2} + 1\right)}$$

$$(74.28 + 14.95 \cdot 1.35) \pm 2.101 \cdot \sqrt{1.181\left(\frac{1}{20} + \frac{(1.35 - 1.196)^2}{(20-1)\cdot 0.1893^2} + 1\right)}$$

$$\text{IC}(95\%) = (93.80, 95.13)$$

3) Si existeix relació lineal entre les variables, calcula amb una confiança del 95% l'IC del pendent β_1 i de l'ordenada en l'origen β_0 del MRLS.

Busquem a les taules el valor $t_{18,0.025} = 2.101$.

IC(95%) de β_1:

$$b_1 \pm t_{n-2,\alpha/2} \cdot \sqrt{\frac{\text{MSE}}{(n-1)s_X^2}}$$

$$14.95 \pm 2.101 \cdot \sqrt{\frac{1.181}{(20-1)\cdot 0.1893^2}}$$

$$\text{IC}(95\%) = (12.18, 17.72)$$

IC(95%) de β_0:

$$b_0 \pm t_{n-2,\alpha/2} \cdot \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_X^2}\right)}$$

$$74.28 \pm 2.101 \cdot \sqrt{1.181\left(\frac{1}{20} + \frac{1.196^2}{(20-1)\cdot 0.1893^2}\right)}$$

$$\text{IC}(95\%) = (70.93, 77.63)$$

PROBLEMES

1. Exercicis resolts

1.1. *En un sistema informàtic multiusuari s'està estudiant el temps de resposta Y de la unitat central en funció del nombre X d'usuaris que simultàniament estan utilitzant el sistema. S'ha recollit la següent informació:*

Usuaris, x_i	*1*	*2*	*3*	*4*	*5*
Temps, y_i	*0.22*	*0.59*	*1.01*	*1.36*	*1.42*

Es demana:

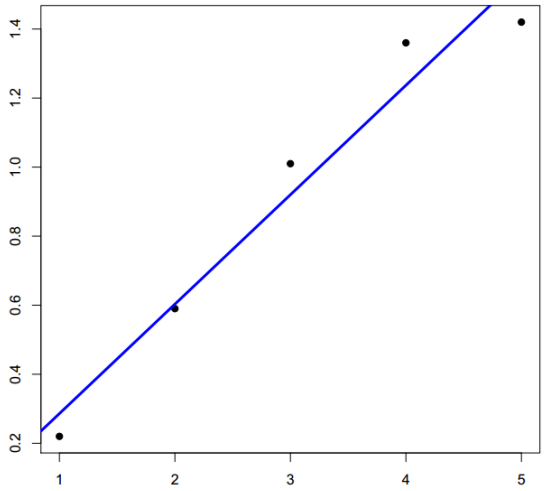
a) Estudieu numèricament i gràficament si existeix una relació lineal entre les dues variables.
b) Ajusteu un model lineal simple a l'anterior distribució. Doneu una interpretació del significat dels valors b_0 i b_1 obtinguts.
c) Trobeu un interval de confiança del 95% del pendent β_1 de la recta del model.
d) Estudieu, a un nivell de significació $\alpha = 0.10$, si es versemblant considerar que $\beta_0 = 0$.

a) El càlcul del coeficient de correlació lineal de Pearson i del coeficient de determinació ens serveixen per comprovar si realment existeix relació lineal entres les variables X i Y:

$$\mathbf{r = 0.978004}$$
$$\mathbf{r^2 = 0.956492}$$

Veient els resultats, podem afirmar que la qualitat de regressió és molt alta.

b) Si representem les dades a través d'un diagrama de dispersió, podem traçar la recta que s'ajusta millor a aquestes dades:



Calculem quant valen els coeficients b_1 i b_0 de la recta de regressió anterior:

$$\mathbf{b_1 = 0.317}$$
$$\mathbf{b_0 = -0.031}$$

En aquest cas, el coeficient b_1 (estimador de β_1) és un indicador de l'increment de temps per l'increment d'usuaris. Per altra banda, el coeficient b_0 (estimador de β_0) es pot interpretar com un temps de resposta base, quan no tenim usuaris.

La recta que obtenim és:

$$\hat{y} = -0.031 + 0.317x$$

c) Si el MRLS és normal, la distribució de b_1 també ho és. Per tant, la fórmula que hem d'aplicar per trobar un IC(95%) del pendent β_1 del MRLS plantejat és la següent:

$$b_1 \pm t_{n-2,\alpha/2} \cdot \sqrt{\frac{MSE}{(n-1)s_X^2}}$$

Necessitem trobar el valor MSE mitjançant la següent fórmula:

$$MSE = \frac{s_Y^2(n-1)(1-r^2)}{n-2}$$

Fem el càlcul de s_x i de s_Y:

$$s_X = 1.58114$$
$$s_Y = 0.512494$$

Fem el càlcul de MSE:

$$MSE = \frac{0.512494^2(5-1)(1-0.978004^2)}{5-2}$$

$$MSE = 0.0152367$$

Busquem a les taules el valor $t_{n-2,\alpha/2} = t_{3,0.025} = 3.182$ i calculem l'IC del pendent β_1 del model:

$$IC(95\%) = \left(0.317 - 3.182 \cdot \sqrt{\frac{0.0152367}{4 \cdot 1.58114^2}}, 0.317 - 3.182 \cdot \sqrt{\frac{0.0152367}{4 \cdot 1.58114^2}}\right)$$

$$\mathbf{IC(95\%) = (0.1928, 0.4412)}$$

d) Si el MRLS és normal, la distribució de b_0 també ho és. Per comprovar si $\beta_0 = 0$ ens plantegem les següents hipòtesis:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

Si H_0 és certa, t_{obs} seguirà una distribució t-Student amb n – 2 graus de llibertat:

$$t_{obs} = \frac{b_0}{s\{b_0\}} \sim t_{n-2}$$

Calculem el valor de t_{obs}:

$$t_{obs} = \frac{b_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)}}$$

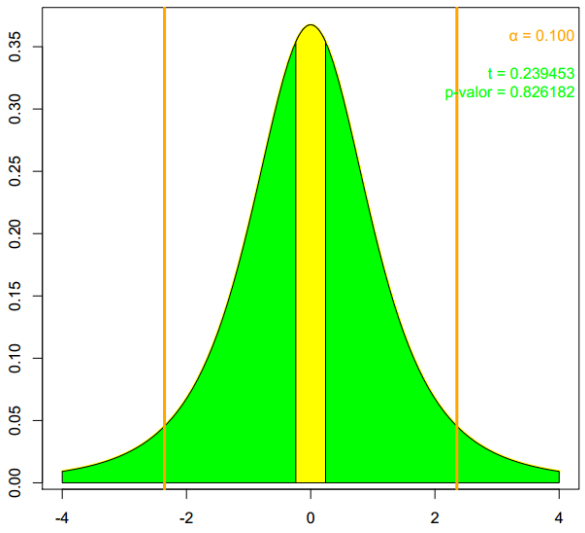
$$t_{obs} = \frac{-0.031}{\sqrt{0.0152367 \cdot \left(\frac{1}{5} + \frac{3^2}{(5-1) \cdot 1.581139^2}\right)}}$$

$$t_{obs} = -0.239453$$

Per $\alpha = 10\%$, busquem a les taules de la distribució t-Student el valor de $t_{n-2,\alpha/2}$:

$$t_{n-2,\alpha/2} = t_{3,0.05} = 2.353$$

Com que $|t_{obs}| = 0.239453 < t_{3,0.05} = 2.353$, acceptem H_0, és a dir, afirmem amb un 90% de confiança que és versemblant considerar que $\beta_0 = 0$.



1.2. *En un estudi s'investiga la relació entre la temperatura de la superfície d'una carretera (x) i la deformació del paviment (y). Després de realitzar 20 observacions (n = 20) s'han obtingut els següents resultats:*

$$\sum y_i = 8.86, \sum y_i^2 = 12.75, \sum x_i = 1478, \sum x_i^2 = 143215.8, \sum y_i x_i = 1083.67$$

Es demana:

a) Calculeu la recta de regressió i el coeficient de correlació. Interpreteu els resultats.
b) Contrasteu si la regressió és significativa.
c) Quin canvi s'espera en la deformació del paviment quan la temperatura de la superfície canvia 1 °F?
d) Calculeu un interval de confiança per a la deformació del paviment quan la temperatura de la superfície és de 85 °F.
e) Calculeu un interval de confiança per a la deformació mitjana del paviment quan la temperatura de la superfície és de 85 °F.

a) Per calcular els coeficients de la recta de regressió (β_1 i β_0) i el coeficient de correlació lineal de Pearson (r) necessitem conèixer quant valen s_{XY}, s_X i s_Y (el desenvolupament de s_Y és el mateix que per s_X):

$$s_{XY} = \frac{1}{n-1} \cdot \sum (x_i - \bar{x})(y_i - \bar{y}) =$$

$$= \frac{1}{n-1} \cdot \sum (y_i \cdot x_i - \bar{y} \cdot x_i - \bar{x} \cdot y_i + \bar{x} \cdot \bar{y}) =$$

$$= \frac{1}{n-1} \cdot \left(\sum (y_i \cdot x_i) - \frac{\sum y_i}{n} \cdot \sum x_i - \frac{\sum x_i}{n} \cdot \sum y_i + n \cdot \frac{\sum x_i}{n} \cdot \frac{\sum y_i}{n} \right) =$$

$$= \frac{1}{20-1} \cdot \left(1083.67 - \frac{8.86}{20} \cdot 1478 - \frac{1478}{20} \cdot 8.86 + 20 \cdot \frac{1478}{20} \cdot \frac{8.86}{20} \right)$$

$$s_{XY} = 22.5745$$

$$s_X^2 = \frac{1}{n-1} \cdot \sum (x_i - \bar{x})^2 =$$

$$= \frac{1}{n-1} \cdot \sum (x_i^2 - 2 \cdot \bar{x} \cdot x_i + \bar{x}^2) =$$

$$= \frac{1}{n-1} \cdot \left(\sum x_i^2 - 2 \cdot \frac{\sum x_i}{n} \cdot \sum x_i + n \cdot \bar{x}^2 \right) =$$

$$= \sqrt{\frac{1}{n-1} \cdot \left(\sum x_i^2 - 2 \cdot \frac{\sum x_i}{n} \cdot \sum x_i + n \cdot \left(\frac{\sum x_i}{n} \right)^2 \right)} =$$

$$= \sqrt{\frac{1}{20-1} \cdot \left(143215.8 - 2 \cdot \frac{1478}{20} \cdot 1478 + 20 \cdot \left(\frac{1478}{20}\right)^2\right)}$$

$$s_X = 42.2969$$

$$s_Y = \sqrt{\frac{1}{n-1} \cdot \left(\sum y_i^2 - 2 \cdot \frac{\sum y_i}{n} \cdot \sum y_i + n \cdot \left(\frac{\sum y_i}{n}\right)^2\right)} =$$

$$= \sqrt{\frac{1}{20-1} \cdot \left(12.75 - 2 \cdot \frac{8.86}{20} \cdot 8.86 + 20 \cdot \left(\frac{8.86}{20}\right)^2\right)}$$

$$s_Y = 0.681524$$

Calculem r, β_1 i β_0:

$$\mathbf{r} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{22.5745}{42.2969 \cdot 0.681524} = \mathbf{0.783120}$$

$$\mathbf{b_1} = r \cdot \frac{s_Y}{s_X} = 0.783120 \cdot \frac{0.681524}{42.2969} = \mathbf{0.0126183}$$

$$\mathbf{b_0} = \bar{y} - b_1\bar{x} = \frac{\sum y_i}{n} - b_1 \cdot \frac{\sum x_i}{n} = \frac{8.86}{20} - 0.0126183 \cdot \frac{1478}{20} = \mathbf{-0.489492}$$

La recta que obtenim és la següent:

$$\hat{y} = -0.489492 + 0.0126183x$$

El valor de r ens informa que la qualitat de la regressió no és molt bona. Per altra banda, el pendent de la recta de regressió β_1 és un indicador de la deformació per l'increment de temperatura. L'ordenada en l'origen de la recta de regressió β_0 ens dóna el valor de la deformació del paviment per una T = 0 °C.

b) Per calcular si la regressió es significativa hem de realitzar el contrast de β_1. Comencem plantejant-nos les següents hipòtesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Com que es tracta d'un MRLS normal, la distribució de β_1 també ho serà. Si H_0 és certa, l'estadístic del contrast (t_{obs}) ha de seguir una distribució t-Student amb n – 2 graus de llibertat:

$$t_{obs} = \frac{b_1}{s\{b_1\}} \sim t_{n-2}$$

$$t_{obs} = \frac{b_1}{\sqrt{\frac{MSE}{(n-1)s_X^2}}} \sim t_{n-2}$$

Observem que necessitem trobar quant val MSE:

$$MSE = \frac{s_Y^2(n-1)(1-r^2)}{n-2} = \frac{0.681524^2(20-1)(1-0.783120^2)}{20-2} = 0.189602$$

Calculem el valor de t_{obs}:

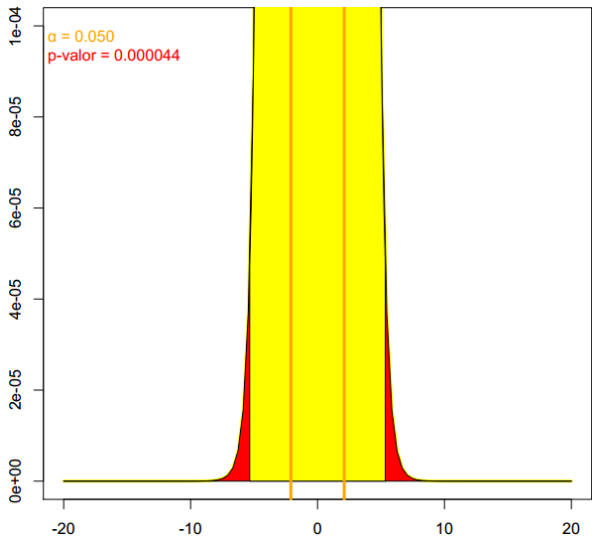
$$t_{obs} = \frac{0.0126183}{\sqrt{\frac{0.189602}{(20-1)\cdot 42.2969^2}}}$$

$$t_{obs} = 5.34276$$

Decidim que α = 5%, per tant, busquem a les taules de la distribució t-Student el valor de $t_{n-2,\alpha/2}$:

$$t_{n-2,\alpha/2} = t_{18,0.025} = 2.101$$

Com que $|t_{obs}| = 5.34276 > t_{18,0.025} = 2.101$, rebutgem H_0 i acceptem H_1, és a dir, afirmem amb un 95% de confiança que $\beta_1 \neq 0$, per tant, que la regressió és significativa.



Podem comprovar que, si fem el contrast de regressió a partir de l'anàlisi de la variància, la decisió final serà la mateixa.

En aquest cas, les hipòtesis que ens plantegem són les mateixes. Si H_0 és certa es compleix que el següent quocient segueix una distribució F de Fisher amb $\nu_1 = 1$ grau de llibertat i $\nu_2 = n - 2$ graus de llibertat:

$$F_{obs} = \frac{SSReg}{MSErr} \sim F_{1,n-2}$$

$$F_{obs} = \frac{(n-1)s_Y^2 - (n-2)MSE}{MSE} \sim F_{1,n-2}$$

Calculem el valor de F_{obs}:

$$F_{obs} = \frac{(20-1)\cdot 0.681524^2 - (20-2)\cdot 0.189602}{0.189602} = 28.5450$$

Decidim que $\alpha = 5\%$, per tant, busquem a les taules de la distribució F de Fisher el valor de $F_{1,n-2,\alpha}$:

$$F_{1,n-2,\alpha} = F_{1,18,0.05} = 4.41$$

Com que $F_{obs} = 28.5450 > F_{1,18,0.05} = 4.41$, rebutgem H_0 i acceptem H_1, igual que hem fet anteriorment.

c) Per cada grau de variació s'espera un canvi en la deformació de $b_1 = 0.0126183$. Aquest canvi serà en el mateix sentit que la variació.

d) Ens demanen que calculem un IC de Y per X = 85, per exemple del 95% de confiança ($\alpha = 5\%$). La fórmula que hem d'utilitzar és:

$$\hat{Y} \pm t_{n-2,\alpha/2} \cdot \tilde{s}$$

$$(b_0 + b_1 X) \pm t_{n-2,\alpha/2} \cdot \sqrt{MSE\left(\frac{1}{n} + \frac{\left(X - \frac{\sum x_i}{n}\right)^2}{(n-1)s_X^2} + 1\right)}$$

$$(-0.489492 + 0.0126183 \cdot 85) \pm 2.101 \cdot \sqrt{0.189602\left(\frac{1}{20} + \frac{\left(85 - \frac{1478}{20}\right)^2}{19 \cdot 42.2969^2} + 1\right)}$$

$$0.583064 \pm 0.939054$$

$$\mathbf{IC(95\%) = (-0.3560, 1.5221)}$$

e) Ens demanen que calculem un IC de E{Y} per X = 85, per exemple del 95% de confiança (α = 5%). La fórmula que hem d'utilitzar és:

$$\widehat{Y} \pm t_{n-2,\alpha/2} \cdot s\{\widehat{Y}\}$$

$$(b_0 + b_1 X) \pm t_{n-2,\alpha/2} \cdot \sqrt{MSE\left(\frac{1}{n} + \frac{\left(X - \frac{\sum x_i}{n}\right)^2}{(n-1)s_X^2}\right)}$$

$$(-0.489492 + 0.0126183 \cdot 85) \pm 2.101 \cdot \sqrt{0.189602\left(\frac{1}{20} + \frac{\left(85 - \frac{1478}{20}\right)^2}{19 \cdot 42.2969^2}\right)}$$

$$0.583064 \pm 0.211851; \ \mathbf{IC(95\%) = (0.3712, 0.7949)}$$

2. Exercicis proposats

2.1. *Per tal de preveure, en funció dels anys d'antiguitat, el nombre de reparacions per mes d'un determinat model de màquina, s'escull una mostra aleatòria de 8 d'aquestes màquines. S'han observat durant un període d'aproximadament un any i s'ha calculat el nombre de reparacions per mes. Les dades obtingudes són les següents:*

Reparacions / mes	*2.1*	*2.6*	*4.3*	*4.7*	*5.0*	*5.8*	*6.4*	*7.2*
Antiguitat (anys)	*2.3*	*2.5*	*3.4*	*4.1*	*4.7*	*5.1*	*5.8*	*6.0*

Es demana:

a) Estudieu numèricament i gràficament si existeix una relació lineal entre les dues variables.
b) Estimeu els paràmetres del model lineal.
c) Doneu una estimació puntual del nombre de reparacions per mes d'una màquina de 5 anys d'antiguitat.
d) Comproveu com, a un nivell de significació α = 0.10, es pot considerar que $\beta_0 = 0$.
e) Contrasteu si $\beta_1 = 0$.
f) Feu un interval de confiança de E{Y} per a x = 4.5.

Solució: a) r=0.984 b) $b_0 = -0.4353$; $b_1 = 1.2266$; c) 5.7 reparacions d) t=-1.093, p-valor=0.315 e) t=13.673, p-valor=0.000 (també tenim F=186.94 p-valor=0.000) f) (4.788, 5.381)

2.2. *Una empresa aeronàutica està dedicada exclusivament a la venda d'un determinat tipus de motor de propulsió. S'ha enregistrat durant cinc mesos el volum mensual Y de vendes de l'empresa, el preu unitari X_1 dels motors i la quantitat X_2 dedicada a la publicitat i activitats de màrqueting. Les dades recollides són les següents:*

Mes	*Juny*	*Juliol*	*Agost*	*Setembre*	*Octubre*
Y Milions	*40*	*50*	*55*	*30*	*45*
X_1 Milers	*300*	*220*	*210*	*330*	*250*
X_2 Milers	*600*	*500*	*800*	*750*	*550*

A partir d'aquestes dades, quina de les dues variables –X_1 o X_2– explica millor les variacions de la variable Y? Raoneu la vostra resposta.

Solució: Correlació linear Y, X_1 r=-0.976; Correlació lineal Y, X_2, r=-0.110 X_1

2.3. *S'està estudiant l'eficiència d'un nou software de manteniment de grans bases de dades en base al nombre de vegades Y que cal accedir al disc per tal de realitzar el manteniment. La variable Y depèn òbviament del nombre X de registres de la base de dades sobre la que es realitza el manteniment. Per tal d'estudiar la relació entre ambdues variables, s'ha avaluat sobre 15 bases de dades de diferents mides el nombre de vegades que ha estat necessari accedir al disc durant el període de manteniment d'aquestes bases. Els resultats obtinguts han estat els següents:*

Ident. BD	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
x_i milers	*350*	*200*	*450*	*50*	*400*	*150*	*350*	*300*
y_i milers	*36*	*20*	*45*	*5*	*40*	*18*	*38*	*32*

Ident. BD	***9***	***10***	***11***	***12***	***13***	***14***	***15***
x_i milers	*150*	*500*	*150*	*400*	*200*	*50*	*250*
y_i milers	*21*	*54*	*11*	*43*	*19*	*7*	*26*

Es demana:

a) Estudieu numèricament i gràficament si existeix una relació lineal entre les dues variables.
b) Busqueu la recta que millor s'ajusta a les dades d'aquesta distribució bivariant.
c) Estimeu, a un nivell de confiança del 95%, la mitjana del nombre de vegades d'accés al disc que es produiran durant el manteniment d'una base de dades de 220.000 registres.
d) Estimeu, a un nivell de confiança del 95%, el nombre de vegades d'accés al disc que es produiran durant el manteniment d'una base de dades de 220.000 registres.

Solució: b) $b_0 = 0.556$; $b_1 = 0.1010$; c) IC(95%) = (21.783, 24.627);
d) IC(95%) = (17.763, 28.647)

2.4. *Donades dues variables x i y, es vol estudiar si hi pot haver relació de tipus lineal entre elles. Es coneixen les següents dades:*

$$\sum y_i = 361, \sum y_i^2 = 9435, \sum x_i = 460, \sum x_i^2 = 15334, \sum y_i x_i = 11981, n = 14$$

Calculeu la recta de regressió i contrasteu si la regressió és significativa.

Solució: a) r=0.987 forta relació lineal y = 7.90 + 0.54x

2.5. *Un estudiant escriu en un examen que el model lineal de regressió es basa en la següent igualtat. Hi esteu d'acord?*

$$E\{Y_i\} = \beta_0 + \beta_1 X_i + \varepsilon \quad (i = 1, ..., n)$$

Solució: No, la relació que defineix un models lineal entre dues variables X i Y és $Y = \beta_0 + \beta_1 X + \varepsilon$ on $\varepsilon \sim N(0; \sigma)$. És la part sistemàtica $\beta_0 + \beta_1 X$ la que ens dona el valor de E{Y}.

2.6. *Considereu un Model de Regressió Lineal (normal). Suposem que els paràmetres del model són $\beta_0 = 200$, $\beta_1 = 5.0$ i $\sigma = 4$. Es demana:*

a) Com es distribueix la variable Y quan X = 10? I quan X = 40?
b) Trobeu un interval de confiança del 95% del valor de Y quan X = 40.
c) Calculeu P({Y | X = 10} > 254).

Solució: a) Per X = 10: N(250;4) i per X = 40: N(400;4);
b) IC(95%) = (392.16, 407.84); c) 0.1587

2.7. *A partir d'una mostra de 1000 dades bivariants (x_i, y_i) (i = 1, ..., 1000) s'ha ajustat un MRL (normal). L'equació de la recta de regressió obtinguda resulta ser igual a Y = 100 + 5X i la variància residual $s_R^2 = 10000$.*

Es demana:

a) Busqueu un IC(95%) per al valor de la variable Y quan X = 100.
b) Busqueu un IC(95%) per a la mitjana de la variable Y / X =100.

Solució: a) IC(95%) = (–100, 1300); b) IC(95%) = (400, 800)

2.8. *Se sap que la relació entre el rendiment R d'un procés químic i la temperatura T a què es realitza ve donada per un MRL:*

$$R_i = \beta_0 + \beta_1 T_i + \varepsilon$$

Hom vol estimar el valor del paràmetre β_1 a partir d'un disseny experimental que només pot incloure 6 experiments. Indiqueu quin dels 5 dissenys permet estimar β_1 amb més precisió:
A: Fer 3 experiments a temperatura T = 0 °C i 3 més a temperatura T = 100 °C.
B: Fer 1 experiment a cada una de les temperatures T = 0 °C, 20 °C, 40 °C, 60 °C, 80 °C i 100 °C.
C: Fer 2 experiments a temperatura T = 0 °C, 2 més a T = 50 °C i 2 més a T = 100 °C.
D: Fer 1 experiment a cada una de les temperatures T = 10 °C, 26 °C, 42 °C, 58 °C, 74 °C i 90 °C.
E: Escollir a l'atzar 6 temperatures entre 0 °C i 100 °C i fer 1 experiment a cada una de les temperatures obtingudes.

Solució: Si volem la màxima precisió, el disseny ens ha de donar la major variància possible de les x i això s'obté en el disseny A

2.9. *En una mina de carbó es vol estudiar fins a quin punt la predisposició a patir un accident laboral depèn del nombre d'hores que el treballador porta treballant. Sobre una mostra de 714 treballadors que han patit algun accident laboral s'ha controlat el nombre d'hores que portaven treballant en el moment de patir l'accident. Els resultats obtinguts es mostren en la següent taula:*

Nombre d'hores treballades	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Nombre d'accidents	*93*	*71*	*79*	*72*	*98*	*89*	*102*	*110*

A la vista d'aquestes dades, diríeu que la predisposició a patir un accident laboral depèn del nombre d'hores que porta treballant una persona?

Solució: r=0.676, r^2=45.7%, relació fluixa/moderada

2.10. *El temps per a codificar un registre k-byte mitjançant una certa tècnica de codificació s'ha mesurat en 24 observacions:*

Mida del registre	***Observacions***		
128	*382*	*375*	*393*
256	*850*	*805*	*824*
384	*1544*	*1644*	*1553*
512	*3035*	*3123*	*3235*
640	*6650*	*6839*	*6768*
768	*13887*	*14567*	*13456*
896	*28059*	*27439*	*27659*
1024	*50916*	*52129*	*51360*

Ajusteu un model lineal de regressió que permeti fer prediccions del temps per a codificar un registre en funció de la seva mida.

Solució: b_0=-15316.631, b_1=49.588, la recta de regressió no és molt adequada

2.11. *Un article publicat al Tappi Journal (març 1986) presenta dades sobre la concentració de licor verd Na_2S i la producció de paper d'una màquina. A partir de les següents dades es demana:*

Concentració licor verd	***Producció paper (Tm/dia)***
40	*825*
42	*830*
49	*890*
46	*895*
44	*890*
48	*910*
46	*915*
43	*960*
53	*990*
52	*1010*
54	*1012*
57	*1030*
58	*1050*

a) Dibuixeu un diagrama de dispersió adequat i estudieu si pot haver-hi una relació lineal entre les dues variables (considereu com a variable resposta, la concentració de licor verd).
b) Calculeu la recta de regressió i el coeficient de correlació.
c) Realitzeu el contrast de regressió.
d) Feu un interval de confiança per a la concentració mitjana de licor verd per a una producció de 950 Tm/dia.

Solució: b) $y = -16.509 + 0.069x$; $r = 0.894$; c) Acceptem H_1: $\beta_1 \neq 0$;d) IC(95%) = (45.72, 53.03)

2.12. *Es vol estudiar la relació que existeix entre el gruix (en cm) d'un material de plàstic i el seu cost (en euros). Es tenen les següents dades:*

Gruix X	*0.5*	*1*	*1.5*	*2*	*2.5*	*3*	*3.5*	*4*	*4.5*	*5*
Cost Y	*58*	*75*	*80*	*120*	*117*	*140*	*150*	*158*	*160*	*198*

Es demana:

a) Realitzeu el diagrama de dispersió. A partir de la gràfica obtinguda, es pot dir si el cost i el gruix poden tenir algun tipus de relació?
b) Calculeu la recta de regressió, així com el coeficient de correlació lineal de Pearson.
c) Quin serà el cost del material de 3.7 cm de gruix?

Solució: a) Sí, existeix certa relació lineal entre les variables;
b) $y = 125.6 + 28.582(x - 2.75)$; $r = 0.9786$; c) $y = 152.75$ €

PRÀCTIQUES

1. Relació lineal entre variables

Per a treballar les tècniques bàsiques de la regressió lineal utilitzarem el dataset `cotxes`. Recupereu l'arxiu `cotxes` del directori de treball. Aquest fitxer consta de 490 registres organitzat en 16 variables. Els registres corresponen a una mostra (no necessàriament representativa del mercat de l'automòbil) de 490 cotxes de diferents marques i tipus sobre els quals es varen mesurar 16 variables:

- *tipus*: Tipus de cotxe (D = Diesel; ND = No Diesel)
- *cili*: Cilindrada (en cc)
- *pote*: Potència (en cavalls)
- *rpmmax*: Nombre màxim de revolucions per minut
- *ncil*: Nombre de cilindres
- *long*: Longitud (en cm)
- *ampl*: Amplada (en cm)
- *altu*: Altura (en cm)
- *male*: Capacitat del maleter (en litres)
- *pes*: Pes (en kg)
- *npla*: Nombre de places
- *vmax*: Velocitat màxima (en km/h)
- *acce*: Acceleració (temps en segons de 0 a 100 km/h)
- *conpon*: Consum ponderat (litres de combustible en 100 km)
- *costkm*: Cost mitjà per quilòmetre recorregut (en pessetes)
- *preu*: Preu (en milers de pessetes)

Ens preguntem si es pot construir un MRL entre les variables *pote* i *vmax*.

Convé sempre començar fent una aproximació gràfica al problema. Hem d'anar a:

`Gràfics, Diagrama de dispersió ...`

A `Variable x` escolliu *pote* i a `Variable y`, *vmax*. Desmarqueu totes les opcions excepte `Línia de mínims quadrats` i polseu `D'acord`. El gràfic resultant serà similar al de la **figura 1**.

- Jutgeu intuïtivament, a partir del gràfic, la correlació entre ambdues variables (signe de la correlació, linealitat, etc). Les dues variables semblen relacionades perquè el gràfic mostra un patró on, per valors petits de *pote* li toquen valors petits de *vmax* (i el mateix per a valors mitjans i valors grans). Tanmateix, la forma del núvol de punts no sembla gaire lineal. A més, sembla que hi ha més concentració de punts en la cua de l'esquerra que no pas en la de la dreta

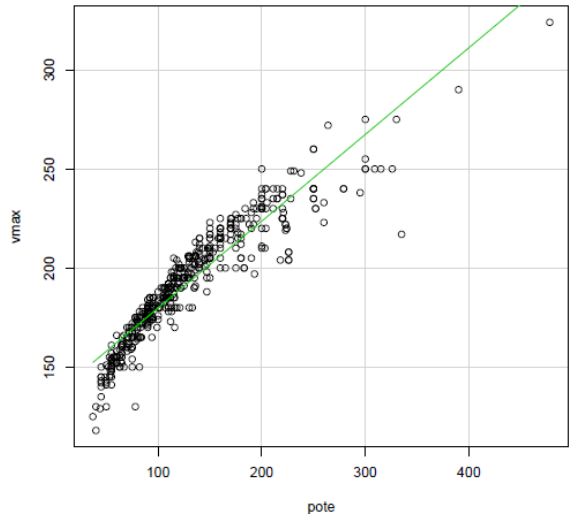


Figura 1: Núvol de punts *pote* i *vmax*.

Per tal d'avaluar numèricament la correlació d'aquestes dues variables farem :

`Estadístics, Resums, Matriu de correlacions ...`

A la finestra emergent, escollim les variables *pote* i *vmax* i marquem el coeficient de Pearson. També marquem l'opció que calcula el p-valor del contrast, que confirma si el coeficient de correlació és o no significativament diferent de zero:

$$H_0: \ \rho = 0 \text{ (variables independents linealment)}$$
$$H_1: \ \rho \neq 0 \text{ (variables dependents linealment)}$$

- Quant val el coeficient de correlació entre ambdues variables? Es significatiu? El seu signe és coherent amb el gràfic anterior? Valora el nivell de correlació. El coeficient val +0.93, amb un p-valor gairebé nul (malgrat R ensenya un 0 com a p-valor, sabem que el que realment passa és que és un valor molt i molt petit, que per arrodoniment el programa ensenya com a zero). El signe positiu és coherent amb el patró vist en el gràfic perquè un augment de *pote* fa esperar un augment de *vmax*. El nivell de correlació (0.93) és molt i molt alt, gairebé perfecte.

El contrast de correlació també es pot obtenir en el següent menú, que també calcula un interval de confiança del coeficient de correlació:

`Estadístics, Resums, Test de correlació,...`

- Quant val l'interval de confiança de r entre les variables *pote* i *vmax*?
  IC(95%): (0.9184753, 0.9421466)

2. El Model de regressió lineal

Ens plantegem construir un model de regressió lineal que ens permeti fer estimacions de la velocitat màxima (*vmax*) d'un vehicle a partir del coneixement de la seva potència (*pote*), és a dir, trobar un model del tipus:

$$vmax = \beta_0 + \beta_1 \cdot pote + \varepsilon \qquad \text{on} \qquad \varepsilon \sim N(0;\sigma)$$

Això ens porta a la necessitat d'estimar els tres paràmetres del model: l'ordenada en l'origen β_0 de la recta de regressió, el pendent β_1 de la recta de regressió i la variància σ^2 del model.

Les respectives estimacions d'aquests 3 paràmetres les simbolitzarem per b_0, b_1 i s^2 (variància residual), respectivament. Estimarem aquests paràmetres amb l'ajuda de R. Aneu al Menú:

`Estadístics, Ajustament de models, Regressió lineal ...`

Com a nom del model podeu escriure `ModelReg1`. Hem de tenir en compte que la variable explicada és *vmax* i la variable explicativa és *pote*. Polseu `D' acord`.

A partir d'ara, el model que acabem de definir és el model actiu (el teniu a la part superior dreta).

Observeu la finestra de resultats i contesteu a les següents preguntes:

- Quina és l'estimació b_0 de l'ordenada en l'origen β_0 de la recta de regressió (*Estimate Intercept*)? $b_0 = 136.1$
- Quina és l'estimació b_1 del pendent β_1 de la recta de regressió (*Estimate Sup. Par*)? $b_1 = 0.4378$.
- Quina és l'estimació de la variància del model o MSE (*Residual estandar error* elevat al quadrat)? $10.63^2 \approx 113$.
- Quant val el coeficient de determinació? Coef. Det. = 0.8673. I el de correlació? Coef. Correlació = sqrt(0.8673) = 0.9312894.
- Quina és, doncs, l'estimació de l'equació de la recta del model lineal?
  $vmax = 136.1 + 0.4378 \cdot pote$
- Aneu al menú `Models, Intervals de confiança ...` i calculeu els intervals de confiança per als coeficients del model.
  Per β_0 IC = (133.8911275, 138.2723870)
  Per β_1 IC = (0.4225841, 0.4530488)

D'aquesta manera hem calculat les estimacions dels tres paràmetres β_0, β_1 i σ^2 del model lineal o recta de regressió que mostra la **figura 1**.

3. Anàlisi de la variància

R també ens proporciona la taula d'anàlisi de la variància del model de regressió mitjançant la funció `aov()` (Anàlisi de la Variància). Recordeu de teoria que aquesta taula consisteix en descompondre la variabilitat total (SSTotal) de la variable resposta en la suma de la variabilitat explicada pel model de regressió (SSReg) més la variabilitat no explicada pel model (SSError), és a dir:

$$SSTotal = SSReg + SSError$$

Un model de regressió és tant més bo com més alta sigui la variabilitat explicada pel model de regressió en relació a la variabilitat total. El quocient SSReg / SSTot és el coeficient de determinació r^2. Si la regressió és bona, r^2 serà pròxim a 1.

Primer de tot hem de recuperar el nostre model anant a:

`Models, Selecciona el Modelo actiu ...`

Després, per fer l'anàlisi de la variància hem d'anar a:

`Models, Test d'hipòtesi, Taula de l'ANOVA ...`

Escolliu `Seqüencial` i polseu `D'acord`.

A la vista de la resposta donada per R, responeu a les següents qüestions:

- Quant val la variabilitat total de la variable resposta (Sum Sup. Par. + Sum Residuals)? SSTotal = 360504 + 55160 = 415664.
- Quant val la variabilitat explicada pel model de regressió (Sum Sup. Par.)? SSReg = 360504.
- Quin tant per cent de la variabilitat total és explicada pel model de regressió? SSReg / SSTotal = Coef. Det. = 360504 / 415664 = 0.8672967 = 86.72967%
- Quina variabilitat no aconsegueix explicar el model de regressió? 1 – 0.8672967 ≈ 13%
- El valor Sum Residuals / Df és la variància residual o MSE. Quant val? SSResid / Df = 55160 / 488 = 113.0328.

4. Gràfics dels residus

Tot i que d'entrada ens pot semblar que aquest model de regressió és prou bo (ja que té un coeficient de determinació alt), és sempre imprescindible analitzar els residus per veure si es compleixen les hipòtesis teòriques dels models de regressió. Aquestes hipòtesis ens diuen que els residus s'han de distribuir homoscedàsticament (variància constant) i aleatòriament al voltant del valor 0 i s'han d'ajustar a una llei Normal.

R-Commander ens proporciona els gràfics adients per a una comprovació intuïtiva de les hipòtesis anteriors. Heu d'anar al Menú:

`Models, Gràfics, Gràfics bàsics de diagnòstic ...`

Hi resulten quatre gràfics (**figura 2**), però solament els dos primers són del nostre interès.

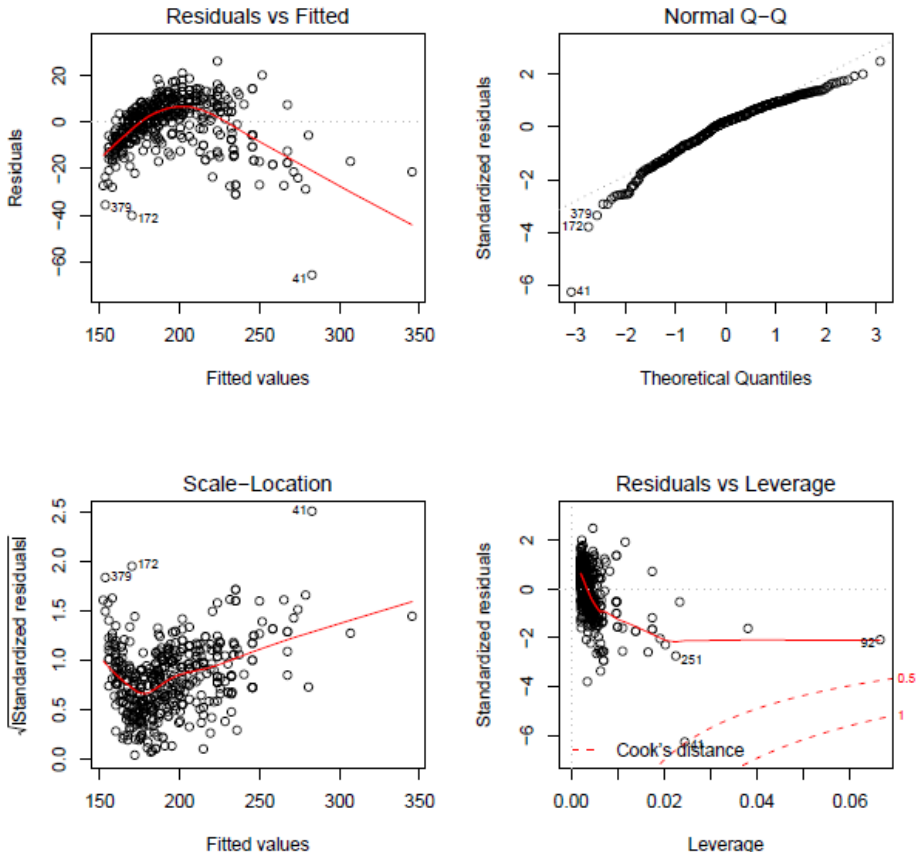


Figura 2: Gràfics de comprovació d'un model de regressió lineal.

Per comprovar que els residus es distribueixen de forma aleatòria i de variància constant es fa servir la primera gràfica: `Residuals vs Fitted` (Residus vs Valors ajustats). El núvol de punts ha de sortir sense cap tendència ni forma, en altres paraules, no hi ha d'haver cap correlació entre les variables. En cas contrari, no es verifica aquesta condició.

- A la vista del gràfic, comenteu si es verifica la condició. El gràfic mostra un patró (corba) que suggereix que el model no és apropiat. Uns residus apropiats haurien de donar un gràfic en forma rectangular: núvol horitzontal d'amplada constant.

Per comprovar si els residus es distribueixen segons una llei Normal s'utilitza un diagrama quantil-quantil (`Normal Q-Q`), el qual dibuixa un núvol de punts que si es troba alineat amb la recta que també es dibuixa ens està indicant que es verifica aquesta condició.

- A la vista del gràfic, comenteu si es verifica la condició. Les cues del diagrama semblen separar-se força de la recta. A més, la zona intermèdia de punts sembla estar primer per sota, després per sobre i al final una altra vegada per sota. L'ajust no sembla gaire bo. S'esperaria un núvol més ajustat i que els punts es repartissin per sobre i per sota sense cap patró.

Es pot fer una anàlisi més acurada dels residus si primer els afegim al conjunt de dades i després els analitzem com si fossin una variable més. Aneu al menú:

```
Models, Afegeix els estadístics de les observacions a les dades ...
```

Feu que s'afegeixin els residus i els valors ajustats. Visualitzeu les dades i comproveu que els valors de la variable *vmax* són iguals a la suma dels residus i els valors ajustats.

- Feu un gràfic de dispersió dels valors ajustats i els residus. Comenteu-lo. El gràfic de dispersió mostra una tendència, un patró, que no és horitzontal ni amb la mateixa amplada. Per tant, els residus no són independents dels valors ajustats.

- Analitzeu la normalitat dels residus (diagrama Q-Q i contrastos de normalitat). El gràfic Q-Q mostra molts punts fora de les bandes. A més, el núvol de punts té zones on està completament per sota o per sobre de la recta. Els contrastos de normalitat donen els següents valors:

  $pearson.test(cotxes\$residuals.RegModel.1) = 1.798 \cdot 10^{-6}$

  $ks.test(cotxes\$residuals.RegModel.1, \text{'pnorm'}, mean(cotxes\$residuals.RegModel.1), sd(cotxes\$residuals.RegModel.1)) = 0.00042$

  $ad.test(cotxes\$residuals.RegModel.1) = 2.539 \cdot 10^{-13}$

  Tots els resultats ens fan rebutjar el model de normalitat.

- Resumint, es pot considerar raonable el Model de Regressió Lineal entre *pote* i *vmax*? A la vista de l'anàlisi dels residus podem afirmar que el model no és apropiat.

5. Transformació de variables

Com haureu pogut comprovar, no es compleixen les hipòtesis teòriques dels models de regressió sobre els residus. Per tant, encara que el coeficient de determinació és prou gran, el model de regressió deixa bastant a desitjar. Per aquest motiu, anem a veure si utilitzant la variable transformada *log_pote* com a variable explicativa en comptes de la variable original *pote* aconseguim millorar el model de regressió.

Repetiu el mateix procés que hem fet abans i contesteu a les següents qüestions:

- Representeu el diagrama de dispersió amb la recta ajustada. Comenteu-lo. L'aspecte lineal del núvol ha millorat molt en relació a l'anterior.
  Quina és l'equació de la recta de regressió?
  $vmax = -93.504 + 60.179 \cdot log\_pote$
- Quant val s_R? $s_R = 8.413$.
- Quant val r^2? $r^2 = 0.9169$.
- Compareu-lo amb el coeficient de determinació del model anterior. Ara el r^2 és major que abans.
- Creieu que els residus són aproximadament homoscedàstics (variància constant)? S'ha millorat respecte el model anterior? El gràfic Residus vs Valors ajustats ha millorat molt i no sembla haver-hi una tendència marcada.
- Podem acceptar que els residus es distribueixen segons una llei Normal? El gràfic de normalitat és força dolent.

En resum, el fet de treballar amb el logaritme de la variable *pote* ens ha millorat sensiblement el model, tot i que les hipòtesis teòriques dels models lineals de regressió no es compleixen del tot.

6. Prediccions

Els models de regressió permeten fer prediccions de la variable resposta a partir de valors de la variable independent. El simple fet de substituir un valor de la variable independent en l'equació de la recta de regressió ens donarà el valor esperat de la variable resposta, que està a sobre de la mateixa recta de regressió. Aquesta és una estimació puntual. El més normal és donar intervals de confiança, que poden ser de dos tipus: de predicció per a un valor individual o de confiança per a la mitjana.

La funció d'R que permet fer prediccions és `predict()`. Si volem fer una predicció de la velocitat màxima d'un vehicle amb una potència de 200 CV hem d'escriure a la finestra d'instruccions:

```
predict(RegModel.1, data.frame(pote=200), interval='prediction')
```

El resultat és:

fit	lwr	upr
223.6451	202.7068	244.5833

Podem veure que a sota de fit trobem l'estimació puntual o valor mitjà, que resulta de substituir 200 a la recta de regressió. Per altra banda, `lwr` i `upr` designen els extrem de l'interval de predicció, és a dir, el vertader valor de la velocitat màxima d'aquest vehicle es trobarà entre 202.7068 km/h i 244.5833 km/h amb una

seguretat del 95%. Per defecte, el nivell de confiança es del 95%, però si volem podem afegir el paràmetre `conf. level=1 -α` per altres nivells de confiança.

Si volem un interval d'estimació per a la mitjana hem d'escriure:

```
interval= 'confidence'
```

Aquest interval serà més precís que l'anterior. El resultat és:

fit	lwr	upr
223.6451	222.2187	225.0714

També es poden fer estimacions anant al menú:

```
Models, PredictionIntervals... (HH)...
```

Si tenim activat el model amb la variable `log(pote)` (`RegModel.2`) i escrivim el valor 5.298317 (log(200)), activant les opcions `prediction interval for individual` o `confidence interval for mean` podem obtenir els intervals de les estimacions.

- Calculeu els dos intervals de les estimacions.
  $IC_{individu}$ = (208.7724, 241.9105)
  $IC_{mitjana}$ =(224.1961, 226.4868)

TEMA 7: La variable aleatòria numèrica discreta

TEORIA

1. Variable aleatòria discreta

Hem explicat que una **variable** X es diu que és **aleatòria** (v.a.) quan, a priori, no se sap exactament quin valor pren però si que es coneix quins valors numèrics pot arribar a prendre i la probabilitat que els adquireixi.

Direm que una v.a. X és **numèrica discreta** quan aquesta pot prendre valors enters, normalment nombres naturals. Per fer-nos-en un idea, el nombre de tares per m^2, els productes defectuosos per lot, el nombre de màquines avariades per dia, etc. són exemples de variables numèriques discretes. En tots aquests casos, si realitzem varies observacions veurem que els valors que prendran les variables seran 0, 1, 2, 3,... Freqüentment, l'amplitud de valors és petit i aquests es troben força repetits.

Una **v.a. discreta** X queda determinada quan es coneixen tots els possibles valors $x_1, x_2, ..., x_k$ que aquesta variable pot arribar a prendre, així com les probabilitats $p_1, p_2, ..., p_k$ que prengui aquests valors:

$$\begin{aligned} p_1 &= P(X = x_1) \\ p_2 &= P(X = x_2) \\ &\dots \\ p_k &= P(X = x_k) \end{aligned}$$

Sempre es complirà que:

$$p_1 + p_2 + \dots + p_k = 1$$

Exemple: *Si tenim tres monedes simètriques i les llancem totes tres a la vegada, quina és la probabilitat d'obtenir 0 cares al realitzar aquest llançament? I la probabilitat d'obtenir-ne 1, 2 i 3?*

Definim X com la variable "nombre total de cares que es poden obtenir al llançar 3 monedes simètriques". Observem que X només pot prendre els valors 0, 1, 2, 3.

Certament, si denotem C = cara i + = creu, tots els possibles esdeveniments que es poden donar són:

$$\Omega = \{CCC, CC+, C++, +++, ++C, +CC, C+C, +C+\}$$

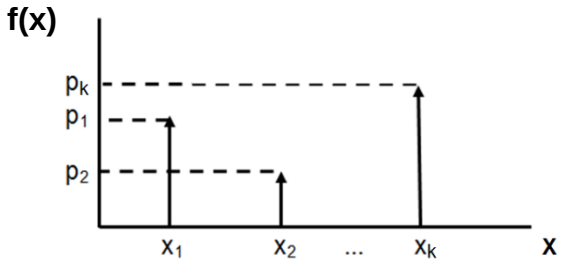
Aplicant la regla de Laplace, les probabilitats seran:

$$x_1 = 0 \rightarrow p_1 = P(X = 0) = 1/8 = 0.125$$
$$x_2 = 1 \rightarrow p_2 = P(X = 1) = 3/8 = 0.375$$
$$x_3 = 2 \rightarrow p_3 = P(X = 2) = 3/8 = 0.375$$
$$x_4 = 3 \rightarrow p_4 = P(X = 3) = 1/8 = 0.125$$

1.1. Funció de densitat, f(x)

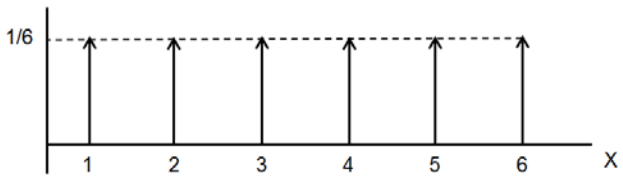
Com que hem dit que una v.a. discreta X pren els valors x_1, x_2, ..., x_k amb probabilitats respectives p_1, p_2, ..., p_k, la seva funció de densitat f(x) es defineix com la funció f tal que:

$$\mathbf{f(x_1) = p_1}$$
$$\mathbf{f(x_2) = p_2}$$
$$\mathbf{...}$$
$$\mathbf{f(x_k) = p_k}$$



En cas que $x \neq x_1$, $x \neq x_2$, ..., $x \neq x_k$ f(x) = 0, ja que la v.a. discreta X només pot prendre uns valors concrets.

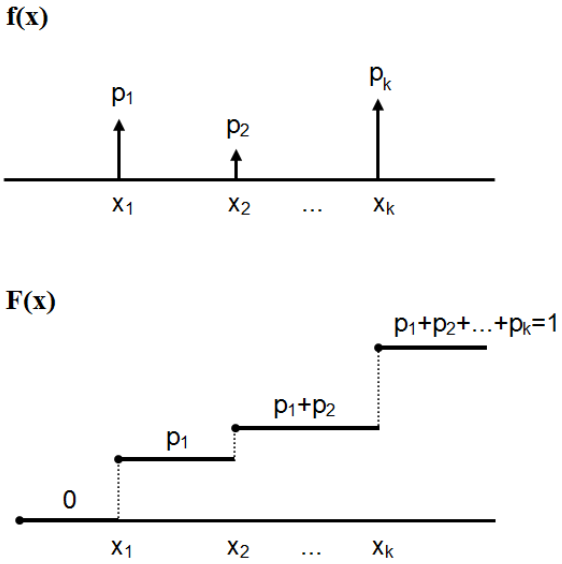
Exemple: *En el cas del llançament d'un dau no trucat (consta de 6 cares i cada una d'elles té la mateixa probabilitat de sortir) la funció de densitat per la v.a. X= "nombre que surt al llançar un dau" serà la següent:*



1.2. Funció de distribució, F(x)

La funció de distribució F(x) d'una v.a. discreta X es defineix com la funció de probabilitats acumulades fins al valor x, és a dir:

$$\mathbf{F(x) = \sum_{i=1}^{x} f(x_i) = P(X \leq x)}$$



F(x) es caracteritza per les següents **propietats**:

- Com es pot veure en el gràfic, $\mathbf{F(-\infty) = 0}$, mentre que $\mathbf{F(+\infty) = 1}$.
- Sempre és **creixent**, ja que les probabilitats de cada valor es van sumant.
- És **contínua per la dreta**.
- F(x) sempre es trobarà entre 0 i 1: $\mathbf{0 \leq F(x) \leq 1}$
- $\mathbf{P(X > x) = 1 - P(X \leq x) = 1 - F(x)}$

Exemple: *En el cas del llançament d'un dau, com en l'exemple anterior, quan val F(4.5)?*

Un dau només pot prendre els valors 1, 2, 3, 4, 5 i 6, per tant, la probabilitat de treure un valor més petit o igual a 4.5 serà la mateixa que la de treure un valor més petit o igual a 4:

$$F(4.5) = P(X \leq 4.5) = P(X \leq 4) = f(1) + f(2) + f(3) + f(4) = 4 \cdot 1/6 = 2/3$$

2. Estadístics d'una variable aleatòria discreta

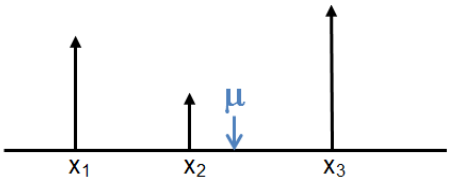
Com en el cas d'una v.a. contínua, els dos estadístics que estudiarem per una v.a. discreta X seran la seva **esperança** i la seva **variància**.

2.1. Esperança, μ = E{X}

L'esperança o valor esperat d'una v.a. discreta X és la suma dels valors x_1, x_2, ..., x_k que pren aquesta variable ponderats per les respectives probabilitats p_1, p_2, ..., p_k:

$$\mu = E\{X\} = x_1p_1 + x_2p_2 + \ldots + x_kp_k = \sum_{i=1}^{k} x_i \cdot f(x_i)$$

El valor μ o E{X} s'interpreta com el centre de la funció de densitat de la v.a. X:



Igual que per l'esperança d'una v.a. contínua, l'esperança d'una v.a. discreta X té de les següents **propietats**:

- L'esperança és lineal: $E\{a + bX\} = a + bE\{X\}$
- L'esperança de la suma de dues v.a. és la suma de les esperances: $E\{X + Y\} = E\{X\} + E\{Y\}$
- Si X i Y són dues v.a. independents, l'esperança del producte de les dues v.a. és el producte de les esperances: $E\{X \cdot Y\} = E\{X\} \cdot E\{Y\}$

Exemple: *Una empresa estima que la producció i venda d'un producte sense defectes li suposa un benefici de 100€, és a dir, si el producte fabricat no té defectes guanyem 100€. En cas contrari, si es fabrica un producte amb defectes i no es pot vendre perdem 100€. La fabricació d'aquest producte és molt delicada i s'estima que la probabilitat que el producte sigui defectuós és igual a 0.5135. Quin serà el guany mitjà de la fabricació d'aquest producte?*

Definim la v.a. discreta X com "guany del producte fabricat". Aquesta variable pot prendre dos valors, que amb les seves respectives probabilitats són:

$$\begin{aligned} x_1 = -100 \quad &\rightarrow \quad p_1 = P(X = -100) = 0.5135 \\ x_2 = +100 \quad &\rightarrow \quad p_2 = P(X = 100) = 1 - 0.5135 = 0.4865 \end{aligned}$$

Si calculem l'esperança d'aquesta variable podrem saber quin és el guany mitjà de la fabricació del producte:

$$\mu = E\{X\} = x_1p_1 + x_2p_2 = (-100) \cdot 0.5135 + 100 \cdot 0.4835 = -3\ €$$

Aquest resultat ens informa que, a la llarga, si continuem fabricant el producte perdrem, en mitjana, 3€ per unitat fabricada.

2.2. Variància (var{X} = σ²) i desviació (desv{X} = σ)

Si una v.a. discreta X pren els valors x_1, ..., x_k amb probabilitats respectives p_1, ..., p_k i hem definit μ com la seva esperança, la variància d'aquesta variable és igual a:

$$\text{var}\{X\} = \sigma^2 = (x_1 - \mu)^2 \cdot p_1 + (x_2 - \mu)^2 + \ldots + (x_k - \mu)^2 \cdot p_k = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot f(x_i)$$

Podem assignar les següents **propietats** a la variància d'una v.a. discreta X:

- La variància no és lineal: $\text{var}\{aX\} = a^2\text{var}\{X\}$ i $\text{var}\{X + a\} = \text{var}\{X\}$
- Si X i Y són dues v.a. independents, la variància de la suma o resta de les dues v.a. és la suma de les variàncies: $\text{var}\{X + Y\} = \text{var}\{X - Y\} = \text{var}\{X\} + \text{var}\{Y\}$
- $\text{var}\{X\} = E\{X^2\} - E^2\{X\} = \sum_{i=1}^{k} x_i^2 \cdot f(x_i) - E^2\{X\}$

Igual que fins ara, l'arrel quadrada positiva de la variància s'anomena **desviació estàndard** i mesura el nivell de dispersió dels valors de la v.a. X al voltant de la seva esperança μ:

$$\text{desv}\{X\} = +\sqrt{\text{var}\{X\}} = +\sqrt{\sigma^2} = \sigma$$

Les propietats de la variància expressades en termes de la desviació són:

- $\text{desv}\{aX\} = |a|\text{desv}\{X\}$ i $\text{desv}\{X + a\} = \text{desv}\{X\}$
- Si X i Y són dues v.a. independents, la desviació de la suma o resta de les dues v.a. és

$$\text{desv}\{X + Y\} = \text{desv}\{X - Y\} = +\sqrt{\text{var}\{X\} + \text{var}\{Y\}}$$

3. Llei Binomial: llei de les peces defectuoses en un lot

La llei Binomial parteix d'un **fenomen aleatori simple** amb només dos resultats possibles coneguts com: **èxit** i **fracàs** (o no èxit). Al resultat denominat èxit li suposarem una probabilitat **p** anomenada **probabilitat d'èxit**. Com que només existeixen dos resultats possibles, **1–p** serà la **probabilitat de fracàs**.

Ens interessa conèixer la probabilitat d'obtenir m èxits al realitzar n repeticions d'aquest fenomen aleatori simple, suposant que aquestes repeticions es realitzen independentment una de l'altra. Si definim la variable X com **"nombre total d'èxits en les n repeticions"**, aleshores X es tracta d'una **v.a. discreta** que segueix una llei Binomial: **X∼Bin(n; p)**. Observem que X pot prendre els valors 0, 1, 2, 3, ..., n.

3.1. Funció de densitat, f(x) i funció de distribució, F(x)

Sigui X∼Bin(n; p), tenint en compte que la llei Binomial ve definida per:

- n = nombre de repeticions del fenomen aleatori simple
- p = P(èxit) (d'un sol fenomen aleatori)

La **funció de densitat f(x)** per la v.a. X segueix la següent fórmula:

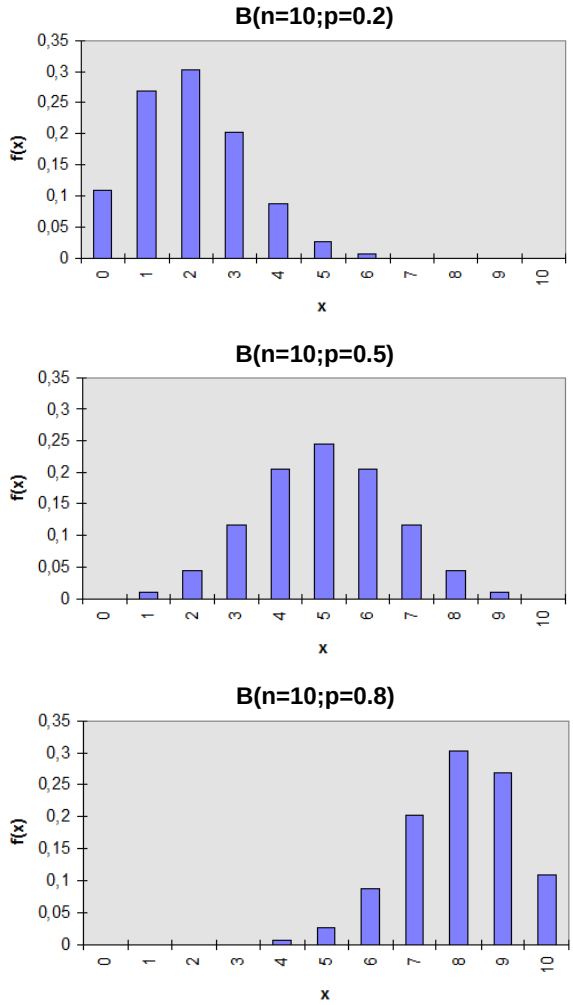
$$\text{P(obtenir m èxits)} = \text{P}(X = m) = \binom{n}{m} \cdot p^m \cdot (1-p)^{n-m}$$

on,

$$\binom{n}{m} = \frac{n!}{m!\,(n-m)!}$$

$0! = 1$
$m! = m \cdot (m-1) \cdot (m-2) \cdot \ldots \cdot 2 \cdot 1$
$(n-m)! = (n-m) \cdot (n-m-1) \cdot (n-m-2) \cdot \ldots \cdot 2 \cdot 1$

En els següents gràfics trobem representada la funció de densitat f(x) d'una llei Binomial per n=10 i p=0.2, 0.5 i 0.8:



Podem obtenir la funció de distribució F(x) acumulant (sumant) probabilitats. Disposem d'unes **taules** on fixat un valor de n i de p, trobem la probabilitat acumulada fins als valors 0, 1, 2 ...,m és a dir:

$$\text{F(m)= P}(X \leq m) = \sum_{i=0}^{m} \text{P}(X = i)$$

3.2. Esperança, μ = E{X} i variància, var{X} = σ^2

L'esperança d'una llei Binomial, que podem definir com el nombre mitjà d'èxits que hom espera obtenir després de n repeticions del fenomen aleatori, es calcula de la següent manera:

$$\mu = E\{X\} = n \cdot p$$

Per altra banda, la variància d'una llei Binomial correspon a:

$$var\{X\} = \sigma^2 = n \cdot p \cdot (1 - p)$$

Conseqüentment, la desviació estàndard, que és l'arrel quadrada de la variància, val:

$$desv\{X\} = \sigma = \sqrt{n \cdot p \cdot (1 - p)}$$

Exemple: *Partim d'un producte que s'agrupa en lots de 20 unitats i que presenta defectes en un 10% d'aquestes unitats. Si definim la v.a. X com "nombre d'unitats amb defectes en el lot", calculeu:*

a) Probabilitat d'obtenir 2 unitats defectuoses en un lot, P(X = 2).
b) Probabilitat d'obtenir alguna unitat defectuosa en un lot, P(X ≥ 1).
c) Probabilitat d'obtenir com a molt 5 unitats defectuoses en un lot, P(X ≤ 5).
d) Nombre esperat d'unitats defectuoses en un lot, μ.

Definim el f.a. simple com "triar a l'atzar un producte d'un lot de 20 unitats" i la probabilitat d'èxit com "probabilitat que el producte sigui defectuós". Aquesta probabilitat p val, segons l'enunciat, un 10% = 0.10. Per tant, com que n = 20, la llei Binomial amb la que treballarem és:

$$X \sim Bin(n = 20; p = 0.1)$$

a) Per calcular P(X = 2) podem aplicar directament de la funció de densitat de la llei Binomial:

$$P(X = 2) = \frac{20!}{2!\,(20-2)!} \cdot 0.1^2 \cdot (1 - 0.1)^{20-2} = 0.2852$$

Com que el valor de n=20 i de p=0.1 es troben a les taules també podríem haver fet el càlcul a partir del valor de les probabilitats acumulades que tenim disponibles a les taules:

$$P(X = 2) = P(X \leq 2) - P(X \leq 1) = 0.6769 - 0.3917 = 0.2852$$

b) Apliquem primer la propietat del complementari i seguidament fem el càlcul a través de les taules de la llei Binomial:

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X \leq 0) = 1 - P(X = 0) = 1 - 0.1216 = 0.8784$$

Com que $P(X \leq 0) = P(X = 0)$, per calcular aquesta part també podríem aplicar la fórmula de la funció de densitat de la llei Binomial.

c) Donat que es tracta d'una probabilitat acumulada, simplement hem de buscar el resultat a les taules de la llei Binomial:

$$P(X \leq 5) = 0.9887$$

També podríem haver fet el càlcul sumant les probabilitats individuals de cada un dels possibles valors:

$$P(X \leq 5) = P(X = 0) + P(X = 1) + P(X = 2) + \ldots + P(X = 5) = 0.9887$$

d) Apliquem la fórmula per calcular l'esperança μ:

$$\mu = E\{X\} = 20 \cdot 0.1 = 2 \text{ productes defectuosos}$$

4. Llei de Poisson: llei de les tares

Un procés de Poisson s'interessa pel nombre d'èxits que, de forma aleatòria, es produeixen en un **interval [0, T]**, ja sigui de temps, de superfície, de volum, etc. Es parteix de la hipòtesi que és possible dividir l'interval [0, T] en **subintervals** molt més petits, de manera que:

- La probabilitat de més d'un èxit en un subinterval val 0.
- L'èxit en un subinterval és independent de l'èxit en qualsevol altre subinterval.
- La probabilitat d'un èxit en un subinterval és la mateixa per a tots els subintervals i proporcional a la longitud d'aquest.

Si la mitjana d'èxits en l'interval [0, T] és igual a $\boldsymbol{\lambda > 0}$, podem afirmar que la **v.a. X**, que compta el nombre d'èxits que s'observen en aquest interval, segueix una **distribució de Poisson** amb paràmetre $\boldsymbol{\lambda}$: $\mathbf{X \sim Pois(\lambda)}$. Per tant, la v.a. X pot prendre els valors 0, 1, 2, 3...

Com a exemples de variables aleatòries que segueixen la llei de Poisson tenim el nombre de trucades que rep un telèfon per dia, el nombre de defectes per cm^2 en una superfície magnètica, el nombre d'e-mails enviats a una bústia electrònica per hora, etc.

4.1. Funció de densitat, f(x) i funció de distribució, F(x)

Sigui $X \sim Pois(\lambda)$, tenint en compte que la llei Poisson ve definida per:

- λ = mitjana d'èxits en un interval [0,T] (de temps, de superfície, de volum...)

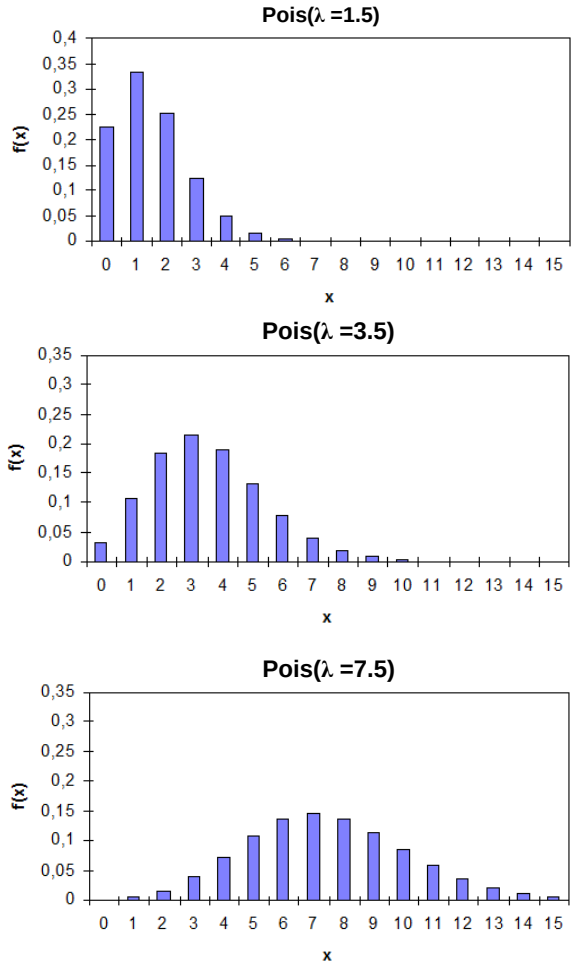
La **funció de densitat f(x)** per la v.a. X segueix la següent fórmula:

$$P(\text{obtenir k èxits}) = P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

$0! = 1$

$k! = k \cdot (k-1) \cdot (k-2) \cdot \ldots \cdot 2 \cdot 1$

En els següents gràfics trobem representada la funció de densitat d'una llei de Poisson per λ=1.5, 3.5, 7.5.



Podem obtenir la funció de distribució F(x) acumulant (sumant) probabilitats. Disposem d'unes **taules** on fixat un valor de de λ, trobem la probabilitat acumulada fins als valors 0, 1, 2 ...,k, és a dir:

$$F(k) = P(X \leq k) = \sum_{i=0}^{k} P(X = i)$$

4.2. Esperança, E{X} i variància, var{X}

Tant l'esperança com la variància d'una llei de Poisson es corresponen amb el paràmetre λ de la distribució:

$$E\{X\} = var\{X\} = \lambda,$$

per tant $desv\{X\} = \sqrt{\lambda}$.

4.3. Propietats

La llei de Poisson compleix dues propietats fonamentals:

- Si el nombre d'èxits en l'interval **[0, T]** té una distribució de Poisson amb paràmetre **λ**, el nombre d'èxits en l'interval **[0, aT]** tindrà una distribució de Poisson amb paràmetre **a · λ**.
- En un procés de Poisson amb una mitjana d'èxits igual a **λ**, la variable Y que mesura la llargada del període entre dos èxits consecutius té una distribució exponencial amb el mateix paràmetre **λ**: **Y∼Exp(λ)**.

5. Aproximacions entre lleis

Existeixen diferents aproximacions entre lleis de probabilitat que podrem utilitzar per calcular probabilitats d'un valor o probabilitats acumulades.

5.1. Aproximació d'una llei Binomial per una llei de Poisson

Partim d'una v.a. X que segueix una distribució Binomial amb paràmetres n i p:

$$X \sim Bin(n; p)$$

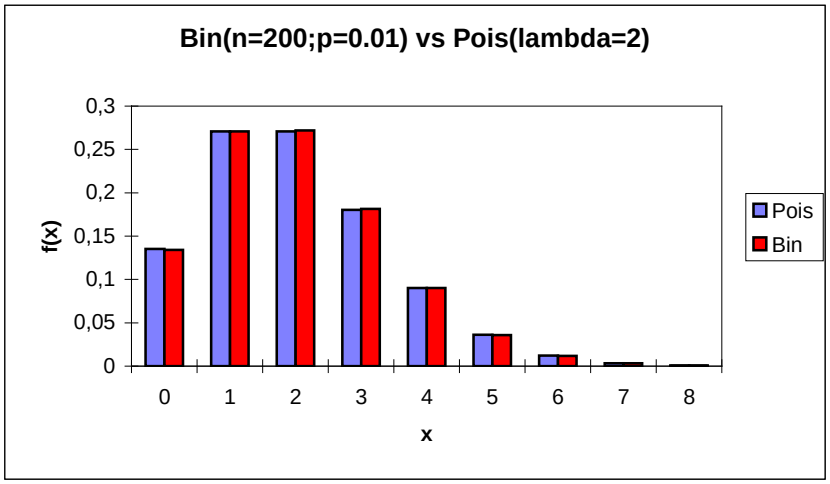
Si el nombre de repeticions del f.a. simple és gran ($\mathbf{n \geq 50}$) i la probabilitat d'èxit és molt petita ($\mathbf{p \leq 0.05}$), és a dir, el que anomenem èxit és un esdeveniment molt difícil que ocorri, podem aproximar X per una v.a. Y amb una distribució de Poisson amb paràmetre $\boldsymbol{\lambda = np}$. És a dir, que per $\mathbf{np < 5}$ el perfil de la funció de densitat de la v.a. X és molt semblant al de la v.a. Y:

$$\mathbf{X \approx Y \sim Pois(\lambda = np)}$$

i les probabilitats d'ambdues lleis no difereixen gaire:

$$P(X = m) = \binom{n}{m} \cdot p^m \cdot (1-p)^{n-m} \approx P(Y = m) = \frac{(np)^m}{m!} \cdot e^{-np}$$

En el següent gràfic podem comparar la densitat Bin(n=200,p=0.01) amb la densitat Pois($\lambda = \mathbf{2}$). Observem com les diferències són gairebé inapreciables.



Exemple: *Una empresa fabrica components electrònics a gran escala. La proporció de components defectuosos és del 0.22%. Si seleccionem a l'atzar una mostra de 1000 components de la producció i definim la v.a. X com "nombre de components defectuosos en el grup de 1000", calculeu:*

a) Probabilitat d'obtenir 3 components defectuosos, P(X = 3).
b) Probabilitat d'obtenir més de 2 components defectuosos, P(X ≥ 2).

Definim el f.a. simple com "triar a l'atzar un component d'un grup de 1000 components" i la probabilitat d'èxit com "probabilitat que el component sigui defectuós". Aquesta probabilitat p val, segons l'enunciat, un 0.22% = 0.0022. Per tant, com que n = 1000, la llei Binomial que segueix la nostra v.a. és:

$$X \sim \text{Bin}(n = 1000; p = 0.0022)$$

Com que n té un valor gran, p té un valor petit i $np = 2.2 < 5$, podem aproximar aquesta llei Binomial a la següent llei de Poisson:

$$Y \sim \text{Pois}(\lambda = 2.2)$$

a) Per calcular P(X = 3) podem aplicar la fórmula de la funció de densitat de la llei Binomial o bé de la llei de Poisson, ja que en aquest cas el resultat serà molt semblant:

Aplicant la llei Binomial:

$$P(X = 3) = \frac{1000!}{3!\,(1000-3)!} \cdot 0.0022^3 \cdot (1 - 0.0022)^{1000-3} = 0.1969$$

Aplicant la llei Poisson:

$$P(X = 3) = \frac{(1000 \cdot 0.0022)^3}{3!} \cdot e^{-1000 \cdot 0.0022} = 0.1966$$

b) Per calcular $P(X \geq 2)$, com que es tracta d'una probabilitat acumulada que no es troba a les taules de la llei Binomial (n és massa gran i p massa petita), podem utilitzar les taules de la llei de Poisson per $\lambda = 2.2$:

$$P(X \geq 2) \approx P(Y \geq 2) = 1 - P(Y < 2) = 1 - P(Y \leq 1) = 1 - 0.355 = 0.645$$

5.2. Aproximació d'una llei Binomial a una llei Normal

Partim d'una v.a. X que segueix una distribució Binomial amb paràmetres n i p:

$$X \sim Bin(n; p)$$

Si es compleix que $\mathbf{n \geq 30}$, $\mathbf{np \geq 5}$ i $\mathbf{n(1-p) \geq 5}$, el perfil de la funció de densitat de la v.a. X s'aproxima molt bé al d'una v.a. Y, que segueix una distribució normal amb paràmetres $\boldsymbol{\mu = np}$ i $\boldsymbol{\sigma = \sqrt{np(1-p)}}$:

$$\mathbf{X \approx Y \sim N(\mu = np; \sigma = \sqrt{np(1-p)})}$$

Donat que la v.a. **X** segueix una **llei discreta** mentre que la v.a. **Y** segueix una **llei contínua**, quan s'utilitza l'aproximació de X per Y s'ha d'aplicar l'anomenada **correcció per continuïtat**, és a dir:

$$P(X = a) \approx P(a - 0.5 \leq Y \leq a + 0.5)$$

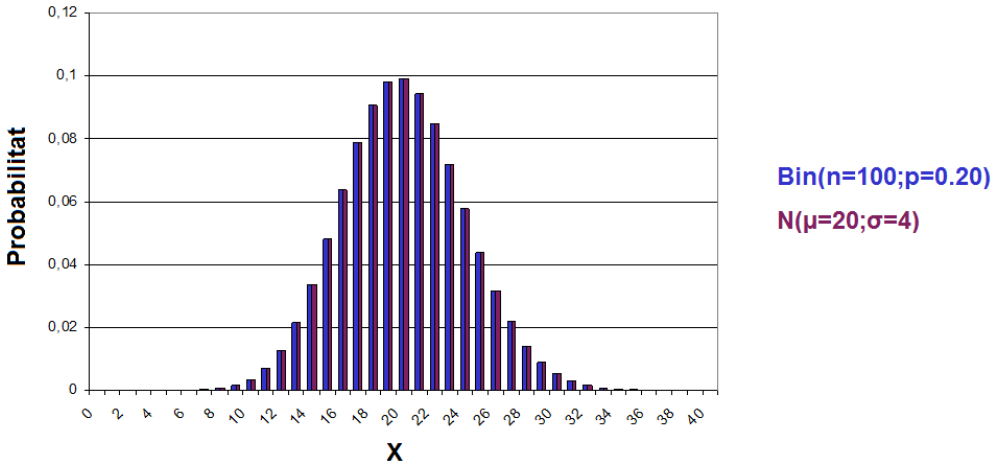
$$P(X \leq a) \approx P(Y \leq a + 0.5)$$

$$P(X < a) \approx P(Y \leq a - 0.5)$$

$$P(X \geq a) \approx P(Y \geq a - 0.5)$$

$$P(X > a) \approx P(Y \geq a + 0.5)$$

A través del següent gràfic podem veure amb quina precisió s'aproxima una llei Binomial per una llei Normal en les condicions explicades anteriorment. Les probabilitats de cada valor de la llei normal s'han calculat utilitzant la correcció per continuïtat.



5.3. Aproximació d'una llei de Poisson a una llei Normal

Partim d'una v.a. X que segueix una distribució de Poisson amb paràmetre λ:

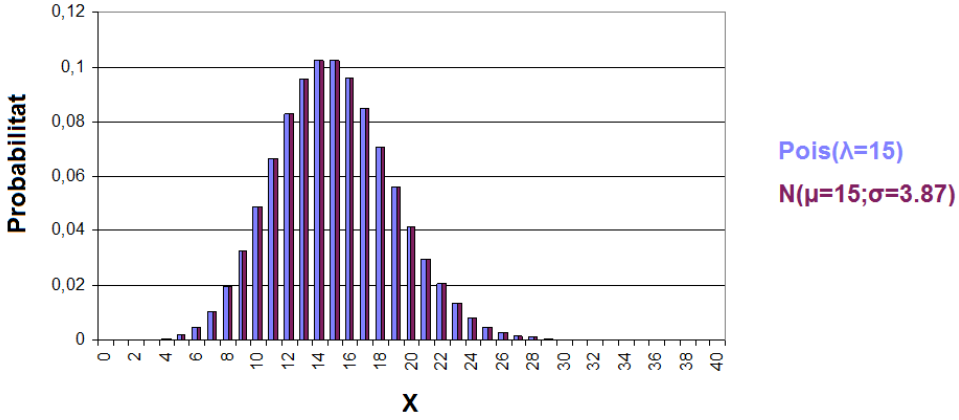
$$X \sim Pois(\lambda)$$

Si el nombre mitjà d'èxits per unitat és prou gran ($\boldsymbol{\lambda \geq 10}$), el perfil de la funció de densitat de la v.a. X s'aproxima molt al d'una v.a. Y, amb distribució normal de paràmetres $\boldsymbol{\mu = \lambda}$ i $\boldsymbol{\sigma = \sqrt{\lambda}}$:

$$\mathbf{X \approx Y \sim N(\mu = \lambda; \sigma = \sqrt{\lambda})}$$

Com que la v.a. **X** segueix una **llei discreta** i la v.a. **Y** segueix una **llei contínua**, caldrà aplicar la **correcció per continuïtat** explicada a l'apartat anterior en cas de voler aproximar les probabilitats de X per les de Y.

A través del següent gràfic podem veure amb quina precisió s'aproxima una llei de Poisson per una llei Normal en les condicions explicades anteriorment. Les probabilitats de cada valor de la llei normal s'han calculat utilitzant la correcció per continuïtat.



6. Estimació de proporcions, intervals de confiança

Quan es desconeix el valor d'un paràmetre *p* que representa la proporció d'unitats d'una població que presenten una determinada característica d'interès, utilitzem la proporció mostral $\hat{p}$ com a estimador.

Per calcular $\hat{p}$ cal escollir una mostra aleatòria de n unitats d'aquesta població i comptar quantes d'aquestes unitats presenten la característica d'interès. Sigui m el nombre d'aquestes unitats, aleshores l'estimador proporció mostral és

$$\hat{p} = \frac{m}{n}$$

Sabem que $\hat{p}$ és un bon **estimador puntual** de la proporció desconeguda (paràmetre) p. Sabem però que els dos valors no tenen perquè coincidir de valor.

6.1. Interval de confiança per una proporció

Suposem que desconeixem el valor d'una proporció p, el nostre objectiu és ara construir un interval $(\hat{p}-\varepsilon, \hat{p}+\varepsilon)$ de manera que puguem afirmar que el vertader valor del paràmetre p està dins d'aquest interval amb probabilitat $1-\alpha$. És a dir, de manera que si afirmem que el vertader valor de p està dins d'aquest interval $(\hat{p}-\varepsilon, \hat{p}+\varepsilon)$ només correm un risc α d'equivocar-nos. És el que **s'anomena interval de confiança de nivell $1-\alpha$ de la proporció p**. Recordeu del tema 4 que normalment prenem α igual a 0.01, 0.05 o 0.1.

Per construir aquest interval utilitzarem la següent aproximació: si la mida n de la mostra es prou gran aleshores la variable aleatòria proporció mostral $\widehat{P_n}$ s'ajusta prou bé a una llei normal.

$$\widehat{P_n} \approx N\left(\mu = p; \sigma = \sqrt{\frac{p(1-p)}{n}}\right)$$

A la pràctica assegurarem normalitat si es compleix que $n.p \geq 10$ i $n.(1-p) \geq 10$. En aquests casos, la v.a. estandarditzada es pot aproximar una llei normal estàndard:

$$Z = \frac{\widehat{P_n} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$$

Així doncs, per construir l'IC$(1-\alpha)$ només cal triar dos valors $-z_{\alpha/2}$ i $z_{\alpha/2}$ d'una llei $Z \sim N(0;1)$ que deixin entre ells una àrea (probabilitat) igual a $1-\alpha$. D'aquesta manera tindrem que aproximadament un $(1-\alpha)100\%$ de totes les possibles proporcions mostrals $\hat{p}$ calculades sobre mostres de mida n estaran dins l'interval

$$\left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right)$$

Això fa que, a partir de la proporció mostral $\hat{p}$ de la nostra mostra, l'IC(1-α) de p sigui igual a:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right)$$

Observem que a la pràctica és impossible calcular aquest interval donat que en la seva expressió hi apareix el valor de la proporció p desconeguda. Per aquesta raó cal prendre solucions aproximades. Tenim però diferents possibilitats:

- Aproximació **grollera**. Substituïm p per la seva estimació puntual $\hat{p}$, així obtenim l'IC(1-α) de p com

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

  Aquesta aproximació només s'admet si la mida n de la mostra és molt gran.

- Supòsit de **màxima indeterminació**. En aquest cas prenem p=0.5, aleshores el valor màxim que pot adquirir el producte p (1-p) que apareix en l'expressió de l'IC és igual a 0.5x0.5 = 0.25 [=1/4]. D'aquesta manera,

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{4n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{4n}} \right)$$

  és l'IC(1-α) de p calculat en el "pitjor dels casos". Si el valor real de p està molt allunyat de 0.5, aquest interval és més llarg del que realment hauria de ser.

- Cas que es disposa **d'informació històrica**. Malgrat desconèixer el valor real de la proporció p, en certs casos se sap a priori que el seu valor no pot superar un determinat valor que anomenarem p_h [<0.5]. És a dir, se sap que $p \leq p_h$. Això permet assegurar que $p(1-p) \leq p_h(1-p_h)$. D'aquesta manera, l'interval

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p_h(1-p_h)}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{p_h(1-p_h)}{n}} \right)$$

  és un IC(1-α) de p més ajustat que el calculat en el supòsit de màxima indeterminació [p=0.5].

6.2. Precisió de l'IC(1-α) de la p

La **precisió** d'un interval de confiança ve determinada pel valor

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Així doncs, per un nivell de confiança 1-α prefixat, la precisió depèn també del valor de la proporció p desconeguda així com de la mida n de la mostra.

Si volem saber quina ha de ser la mida n que ha de tenir la mostra per tal que l'IC(1-α) de p tingui una precisió ε prefixada només ens cal aïllar la n de la fórmula anterior:

$$n = \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 p(1-p)$$

i utilitzar les solucions aproximades

- Per màxima indeterminació: $n = \left(\frac{z_{\alpha/2}}{2\varepsilon}\right)^2$.
- Per informació històrica [$p \leq p_h < 0.5$]: $n = \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 p_h(1-p_h)$

Exemple: *Es vol fer una enquesta d'opinió abans d'unes eleccions i es tria una mostra aleatòria simple de n=1000 persones d'entre els 4 milions de votants potencials. Suposarem que la mostra ha estat triada correctament i per tant és representativa de la població de votants potencials. Ens informen que de les 1000 persones enquestades 80 han manifestat que pensen votar el partit A. Hom vol estimar, a un nivell de confiança del 95%, quina és la proporció real p de votants d'aquest partit A.*

Calculem la proporció mostral $\hat{p}=80/1000 = 0.08$ [≡8%].

Donat que ens demanen de calcular l'estimació de *p* a un nivell de confiança del 95%, tenim 1-α=0.95 i α=0.05. Busquem a les taules de la llei normal estàndard els dos valors $-z_{0.05/2}$ i $-z_{0.05/2}$ que deixen entre ambdós una probabilitat de 0.95:

$$-z_{0.025}=-1.960$$

$$+z_{0.025}=+1.960$$

Seguidament podem calcular l'IC aproximat utilitzant alguna de les aproximacions.

- Aproximació grollera:

$$\left(0.08 - 1.96\sqrt{\frac{0.08(1-0.08)}{1000}}\,,\, 0.08 + 1.96\sqrt{\frac{0.08(1-0.08)}{1000}}\right) = (0.063, 0.097)$$

- Màxima indeterminació [p=0.5]:

$$\left(0.08 - 1.96\sqrt{\frac{1}{4\cdot 1000}}\,, 0.08 + 1.96\sqrt{\frac{1}{4\cdot 1000}}\right) = (0.049, 0.111)$$

- Informació històrica. En aquest cas caldria tenir informació històrica addicional. Suposem que se sap d'altres eleccions que el partit *A* mai ha superat el 15% dels vots (p<0.15) i que hi ha motius per suposar que en aquestes eleccions tampoc ho farà. Aleshores podem considerar $p_h = 0.15$. Així, un IC(1-0.05) de *p* més ajustat és

$$\left(0.08 - 1.96\sqrt{\frac{0.15(1-0.15)}{1000}}\,, 0.08 + 1.96\sqrt{\frac{0.15(1-0.15)}{1000}}\right) = (0.058, 0.102)$$

7. Gràfics de control per atributs

El control per atributs s'aplica quan interessa controlar un atribut o una característica que pot tenir o no un producte (exemple: defecte/no defecte). Així doncs, els gràfics de control per atributs estan basats en una variable qualitativa. És important tenir establert un criteri per decidir si una unitat té o no la característica a estudi. Aquest ha de ser clar i s'ha de mantenir mentre es realitza el gràfic de control.

Els gràfics de control per atributs es basen en la distribució Binomial o la distribució de Poisson i s'utilitzen les aproximacions entre elles i les aproximacions a una llei normal per calcular les línies de control.

Veurem 4 tipus de gràfics de control per atributs: p, np, c, i u

7.1. Gràfic p

Les unitats d'un procés es classifiquen en defectuoses o no defectuoses, fora-dins de toleràncies, etc., Mitjançant aquest gràfic es controla la proporció p d'unitats en un d'aquests grups que produeix el procés en estat de control.

El procés a seguir és prendre mostres (al menys 20) de mida ni de forma consecutiva i a intervals de temps iguals i calcular la proporció $\overline{p_i}$ d'elements de cada mostra que compleixen la característica d'interès. Noteu que en els apartats d'estimació i contrastos per una proporció s'ha utilitzat la notació $\hat{p}$ per la proporció mostral. S'aconsella prendre mostres de mida suficientment gran per tal que puguin aparèixer, almenys, tres o quatre unitats amb la característica d'interès (ex: defecte).

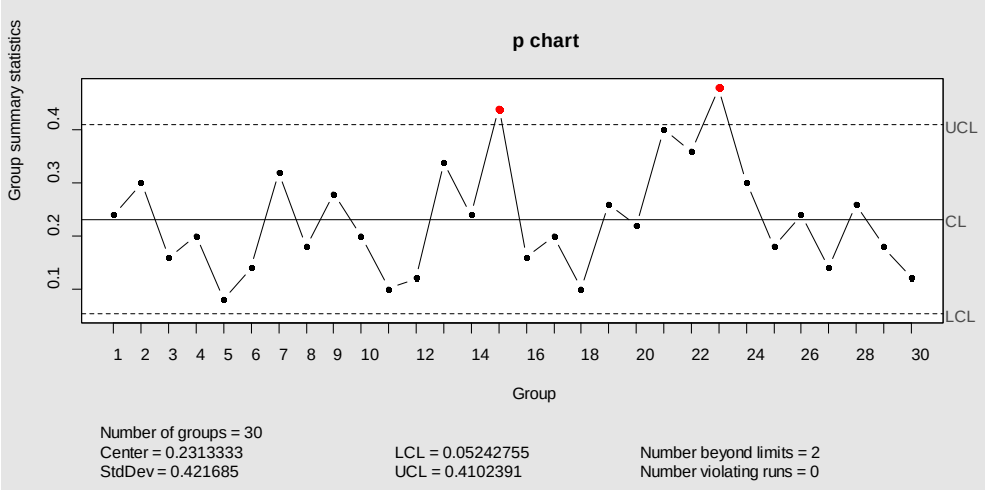
En el gràfic de control p es representen, en forma de diagrama temporal, les proporcions mostrals $\overline{p_1}, \overline{p_2}, \ldots, \overline{p_m}$ i les línies de control CCL, LCL i UCL que valdran:

CCL	$\bar{\bar{p}} = \frac{\sum n_i \bar{p_i}}{\sum n_i}$
LCL	$\bar{\bar{p}} - 3\sqrt{\frac{\bar{\bar{p}}(1-\bar{\bar{p}})}{n_i}}$
UCL	$\bar{\bar{p}} + 3\sqrt{\frac{\bar{\bar{p}}(1-\bar{\bar{p}})}{n_i}}$

Les línies de control es calculen utilitzant l'aproximació de la llei binomial per una normal quan la proporció es pot estimar en parts per cent. Si es tracta del control d'una característica poc freqüent (estimada en parts per mil), s'utilitza l'aproximació a la llei Poisson.

Si tenim informació històrica de p reemplacem les estimacions per la informació

Exemple: Gràfic p (dades orangejuice, paquet qcc, R)



7.2. Gràfic np

Aquest tipus de gràfic s'aplica al mateix tipus de processos que el gràfic p. La diferència està en que es controla el nombre d'unitats de cada mostra que compleixen una característica d'interès (exemple: defecte) enlloc de la proporció. Per altra banda, la mida de cada una de les mostres és fixada i gual a n. En aquest cas també s'aconsella prendre mostres de mida suficientment gran per a què puguin aparèixer, almenys, tres o quatre unitats amb la característica d'interès.

Observem doncs que en cada una de les mostres, el nombre d'unitats amb la característica d'interès es pot modelar amb una llei binomial $\text{Bin}(n; p)$ i per tant el nombre esperat d'unitats és np amb una desviació igual a $\sqrt{np(1-p)}$.

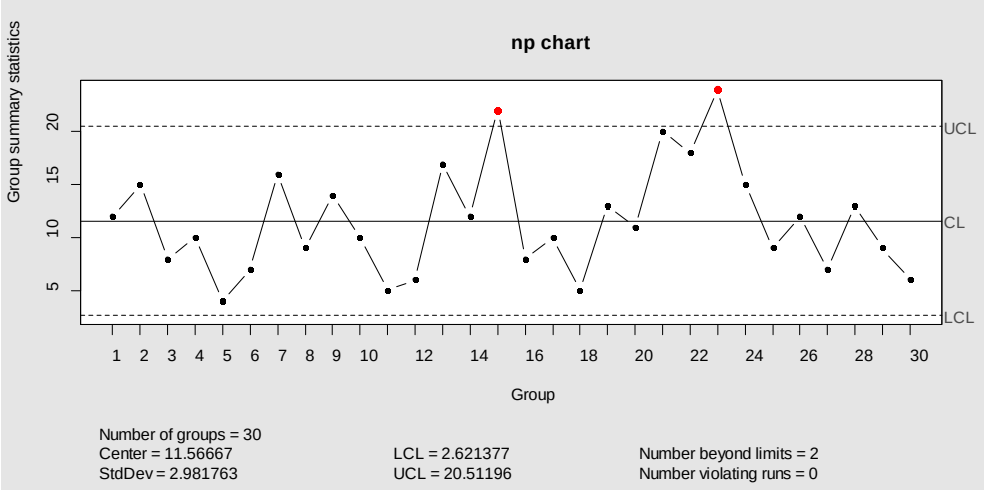
El procés a seguir és prendre m mostres (al menys 20) de mida n de forma consecutiva i a intervals de temps iguals i representar en forma de diagrama temporal el nombre d'elements amb la característica d'interès $\overline{d_1}, \overline{d_2}, \ldots, \overline{d_m}$ que obtenim en cada una de les mostres $\left(\overline{d_i} = n\overline{p_i}\right)$.

Les línies de control CCL, LCL i UCL valdran:

CCL	$\bar{\bar{d}} = \frac{\text{total defectuoses}}{\text{nombre mostres}} = \frac{\sum \overline{d_i}}{m} = \frac{\sum n\overline{p_i}}{m} = n\frac{\sum \overline{p_i}}{m} = n\bar{\bar{p}}$
LCL	$\bar{\bar{d}} - 3\sqrt{\bar{\bar{d}}\left(1 - \frac{\bar{\bar{d}}}{n}\right)}$
UCL	$\bar{\bar{d}} + 3\sqrt{\bar{\bar{d}}\left(1 - \frac{\bar{\bar{d}}}{n}\right)}$

Si es té informació històrica de p reemplacem les estimacions per la informació, on d=np.

Exemple: Gràfic np (dades orangejuice, paquet qcc, R)



7.3. Gràfic *c*

Aquest tipus de gràfic s'utilitza per controlar el nombre d'ocurrències que apareixen en una unitat de producte (exemple: nombre de tares per metre lineal de producte). S'aconsella escollir la unitat d'inspecció suficientment gran (exemple: 3 metres de cable, 1 metre quadrat de roba, etc.) per a què hi hagi un nombre esperat de com a mínim 10 ocurrències del fenomen a estudi.

La distribució de referència és la llei de Pois(λ), on λ representa el nombre mig d'ocurrències por unitat d'inspecció. Per λ suficientment gran aproximem la llei Poisson per la Normal.

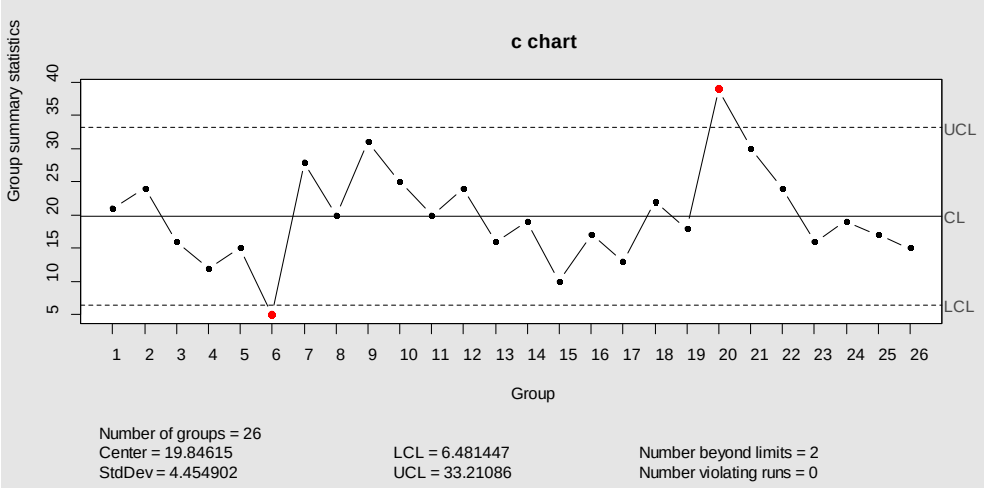
El procés a seguir és prendre m (≥ 20) unitats d'inspecció de forma consecutiva i a intervals de temps iguals i representar en forma de diagrama temporal el nombre d'ocurrències $\overline{\lambda_1}, \overline{\lambda_2}, \ldots, \overline{\lambda_m}$ que obtenim en cada unitat d'inspecció.

Les línies de control CCL, LCL i UCL valdran:

CCL	$\bar{\bar{\lambda}} = \frac{\sum \overline{\lambda_i}}{m}$
LCL	$\bar{\bar{\lambda}} - 3\sqrt{\bar{\bar{\lambda}}}$
UCL	$\bar{\bar{\lambda}} + 3\sqrt{\bar{\bar{\lambda}}}$

Si es té informació històrica de λ reemplacem les estimacions per la informació.

Exemple: Gràfic c (dades circuit, paquet qcc, R)



7.4. Gràfic *u*

S'aplica pel mateix tipus d'anàlisi que el gràfic c, però en aquest cas la unitat d'inspecció no és la mateixa en cada mostra. En aquest cas cal establir la unitat de mesura (exemple: metre, metre quadrat, minut, etc.).

La distribució de referència, que modela el nombre d'ocurrències per unitat d'inspecció és la llei de Poisson. Partim d'una llei Pois(λ), on λ representa el nombre d'ocurrències per unitat de mesura.

El procés a seguir és prendre una mostra d'unitats d'inspecció i mesurar el nombre de defectes en cada unitat c_i i la mida de la unitat n_i (unitats de mesura). Es representa en forma de diagrama temporal el nombre de defectes per unitat de mesura, és a dir

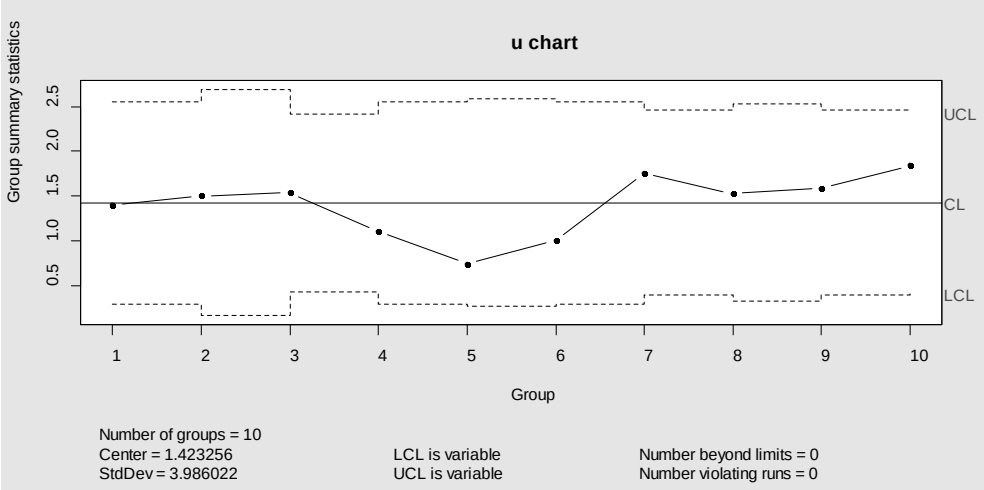
$$\bar{u}_i = \frac{c_i}{n_i}$$

Les línies de control CCL, LCL i UCL valdran:

CCL	$\bar{\bar{u}} = \frac{\sum n_i \bar{c}_i}{\sum n_i} = \frac{\sum c_i}{\sum n_i}$
LCL	$\bar{\bar{u}} - 3\sqrt{\frac{\bar{\bar{u}}}{n_i}}$
UCL	$\bar{\bar{u}} + 3\sqrt{\frac{\bar{\bar{u}}}{n_i}}$

Si es té informació històrica de λ reemplacem les estimacions per la informació, on λ=u.

Exemple: Gràfic u (dades dyedcloth, paquet qcc, R)



8. Contrastos per proporcions

El contrast d'un proporció ens permet comprovar si una determinada característica és present en una població en una proporció p_0. Podem plantejar els tres tipus de contrastos habituals:

Bilateral:	**Unilateral dret:**	**Unilateral esquerre:**
$H_0 : p = p_0$	$H_0 : p \leq p_0$	$H_0 : p \geq p_0$
$H_1 : p \neq p_0$	$H_1 : p > p_0$	$H_1 : p < p_0$

Per dur a terme el contrast caldrà extreure una mostra aleatòria de n unitats representativa de la població d'interès. Seguidament es calcula la proporció $\hat{p}$ d'unitats de la mostra que presenten la característica d'interès. En funció del

resultat de la comparació del valor de $\hat{p}$ amb p_0 hem de valorar si tenim motius suficients per desconfiar o no de la versemblança de la H_0.

Ens marquem primer un nivell de significació α que recordem representa la màxima probabilitat d'error de tipus I [rebutjar erròniament H_0] que estem disposats a cometre al fer aquest contrast. Recordem que habitualment α=0.01, 0.05 o 0.1.

Per valorar numèricament fins a quin punt la diferència entre la $\hat{p}$ observada sobre la mostra i la p_0 que figura a la H_0 és deguda a l'atzar podem calcular el p-valor a partir de l' estadístic de contrast

$$z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

i la distribució normal estàndard Z~N(0;1) si el valor de n és suficientment gran ($n.p_0 \geq 10$ i $n.(1-p_0) \geq 10$):

Bilateral:	$\text{p-valor} = P[\lvert Z\rvert > \lvert z_{obs}\rvert]$
Unilateral dret:	$\text{p-valor} = P[Z > z_{obs}]$
Unilateral esquerre:	$\text{p-valor} = P[Z < z_{obs}]$

Finalment apliquem la regla de decisió habitual:

- Si **p-valor ≥ α** considerarem que la diferència entre p_0 i $\hat{p}$ és deguda a l'atzar, i per tant direm que "No tenim motius suficients per rebutjar la H_0".

- Si **p-valor < α** considerarem que la diferència entre p_0 i $\hat{p}$ no és deguda a l'atzar, sinó que la H_0 és falsa. Per això direm que "Tenim motius suficients per rebutjar la H_0 i per tant acceptem la hipòtesi contrària H_1".

Exemple: *Una empresa consultora ha fet un estudi i afirma que el producte X és preferit pel 55% dels consumidors. Es pren una mostra de 1000 consumidors dels quals el 51% prefereixen l'esmentat producte. Jutgeu l'afirmació que realitza la consultora*

L'empresa consultora afirma que p=0.55, no obstant això, amb la mostra de mida 1000 obtenim $\hat{p} = 0.51$. Donat que volem jutjar l'afirmació de la consultora ens cal plantejar el següent contrast:

$$H_0 : p = 0.55$$
$$H_1 : p \neq 0.55$$

Prenem un nivell de significació α = 5%.
A partir de la proporció mostral calculem l'estadístic de contrast

$$z_{obs} = \frac{0.51 - 0.55}{\sqrt{\frac{0.55(1-0.55)}{1000}}} = -2.54$$

Seguidament podem calcular el p-valor utilitzant les taules de la normal estàndard. Donat que és un contrast bilateral, ens cal calcular la probabilitat que queda fora de l'interval (-2.54, 2.54). Com que la distribució normal és simètrica busquem la probabilitat que està directament tabulada i la multipliquem per dos.

p-valor$= P[|Z| > |-2.54|] = 2P[Z < -2.54] = 2 \cdot 0.0055 = 0.0110$

Donat que el p-valor és inferior al nivell de significació α (0.011 < 0.05), es conclou que tenim motius suficients per rebutjar la H_0 i acceptem la hipòtesi contrària H_1.

PROBLEMES

1. Exercicis resolts

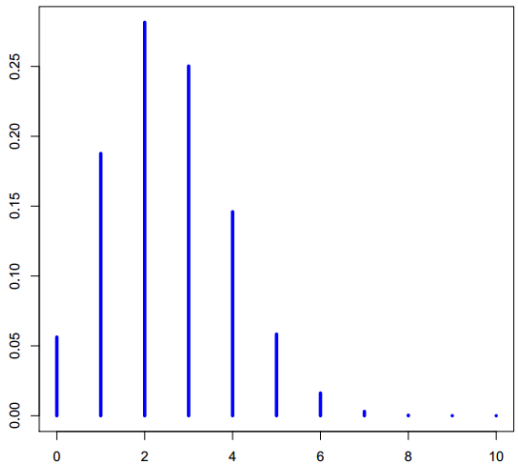
1.1. *Una prova tipus test consta de 10 preguntes, cada una de les quals admet 4 possibles respostes, de les quals només una és correcta. Un alumne contesta a l'atzar totes les preguntes del test.*

Es demana:

a) Quina probabilitat té de contestar correctament 5 preguntes?
b) Quina probabilitat té de contestar correctament totes les preguntes?
c) Quina probabilitat té de no contestar correctament cap de les preguntes?
d) Cada pregunta contestada correctament es puntua amb 1 punt positiu. Amb quina puntuació negativa cal valorar les preguntes contestades incorrectament si hom vol que l'esperança de la puntuació obtinguda per un alumne que contesta a l'atzar sigui igual a 0?

Definim la v.a. discreta X com "nombre de preguntes de la prova tipus test contestades correctament". Com que l'alumne contesta totes les preguntes tenim que n = 10 i com que contesta a l'atzar tenim que p = ¼= 0.25. Així doncs, podem dir que X segueix la següent llei Binomial:

$$X \sim Bin(n = 10; p = 0.25)$$



a) Per saber la probabilitat que X = 5 només cal aplicar la fórmula de la funció de densitat de la llei Binomial o bé fer el següent plantejament en termes de la funció de distribució, ja que estem parlant d'una v.a. discreta:

$$P(X = 5) = P(X \leq 5) - P(X \leq 4)$$

Busquem a les taules de la llei Binomial amb n=10 i p=0.25 la probabilitat acumulada per X = 5 i X = 4 que són respectivament 0.9803 i 0.9219 i resolem:

$$\mathbf{P(X = 5)} = 0.9803 - 0.9219 = \mathbf{0.0584}$$

b) Fem el mateix que en l'apartat anterior però ara buscant P(X = 10):

$$\mathbf{P(X = 10)} = \mathrm{P}(\mathrm{X} \leq 10) - \mathrm{P}(\mathrm{X} \leq 9) = 1.0000 - 1.0000 = \mathbf{0.0000}$$

c) Apliquem el mateix concepte per P(X = 0):

$$\mathbf{P(X = 0)} = \mathrm{P}(\mathrm{X} \leq 0) = \mathbf{0.0563}$$

d) Definim una nova v.a. discreta Y com "puntuació obtinguda en una pregunta". Aquesta variable pot prendre els valors +1, si la pregunta es contesta correctament, o un valor negatiu que anomenarem –r, si la pregunta es respon incorrectament. Si un alumne contesta a l'atzar, les probabilitats que Y prengui aquests valors són 1/4 i 3/5, respectivament:

$$\mathrm{P}(\mathrm{Y} = 1) = 0.25$$
$$\mathrm{P}(\mathrm{Y} = -\mathrm{r}) = 0.75$$

Ens demanen que E{Y} = 0, per tant només cal aplicar la definició de l'esperança d'una v.a. discreta i igualar-la a 0:

$$\mathrm{E}\{\mathrm{Y}\} = 1 \cdot 0.25 - \mathrm{r} \cdot 0.75 = 0$$

Finalment aïllem r i obtenim $\mathbf{r = 1/3.}$

1.2. *El nombre de trucades telefòniques que es reben en una centraleta cada 5 minuts s'ajusta a una distribució de Poisson de paràmetre λ = 3. Calculeu la probabilitat que:*

a) La centraleta rebi 6 trucades en 5 minuts.
b) La centraleta rebi 3 trucades en 10 minuts.
c) La centraleta rebi més de 15 trucades en 15 minuts.
d) La centraleta rebi 2 trucades en 1 minut.
e) La centraleta no rebi cap trucada en 5 minuts.

Definim la v.a. discreta X com "nombre de trucades telefòniques que es reben en una centraleta cada 5 minuts". Aquesta variable segueix la següent distribució:

$$\mathrm{X} \sim \mathrm{Pois}(\lambda = 3)$$

a) Per saber la probabilitat que X = 6 només cal aplicar la fórmula de la funció de densitat per la llei de Poisson o bé fer el següent plantejament en termes de la funció de distribució, ja que estem parlant d'una v.a. discreta:

$$\mathbf{P(X = 6)} = \mathrm{P}(\mathrm{X} \leq 6) - \mathrm{P}(\mathrm{X} \leq 5) = 0.966 - 0.916 = \mathbf{0.050}$$

b) Sabem que si el nombre d'èxits en l'interval [0, 5] segueix una distribució de Poisson amb paràmetre $\lambda = 3$, el nombre d'èxits en l'interval $[0, 2 \cdot 5] = [0, 10]$ seguirà una distribució de Poisson de paràmetre $2 \cdot \lambda = 2 \cdot 3 = 6$. Així doncs si ara X="nombre de trucades telefòniques que es reben en una centraleta cada 10 minuts", i volem calcular $P(X = 3)$ haurem d'aplicar el mateix concepte que en l'apartat anterior, però ara buscant a les taules les probabilitats acumulades pel nou paràmetre $\lambda = 6$.

$$\mathbf{P(X = 3)} = P(X \leq 3) - P(X \leq 2) = 0.151 - 0.062 = \mathbf{0.089}$$

c) Considerem ara la v.a X="nombre de trucades telefòniques que es reben en una centraleta cada 15 minuts". Observem que ara passem d'un interval [0, 5] a un interval $[0, 3 \cdot 5] = [0, 15]$. La nostra v.a. X seguirà una llei de Poisson de paràmetre $3 \cdot \lambda = 3 \cdot 3 = 9$. Així, per calcular $P(X > 15)$ podrem utilitzar les taules de probabilitats acumulades d'una llei de Poisson amb $\lambda = 9$:

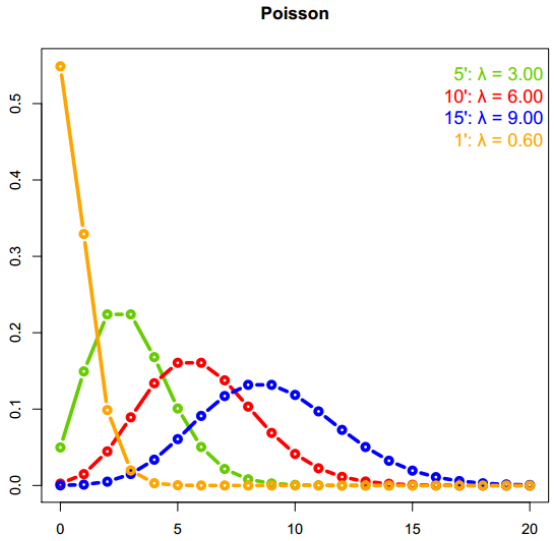
$$\mathbf{P(X > 15)} = 1 - P(X \leq 15) = 1 - 0.978 = \mathbf{0.022}$$

d) Considerem ara la v.a X="nombre de trucades telefòniques que es reben en una centraleta cada minut". Observem que ara passem d'un interval [0, 5] a un interval $[0, (1/5) \cdot 5] = [0, 1]$, és a dir, reduïm en 5 l'interval. Per tant la v.a. X seguirà una llei de Poisson amb el paràmetre $(1/5) \cdot 3 = 0.6$. Per calcular $P(X = 2)$ haurem de buscar a les taules les probabilitats acumulades d'una llei de Poisson amb paràmetre $\lambda = 0.6$:

$$\mathbf{P(X = 2)} = P(X \leq 2) - P(X \leq 1) = 0.977 - 0.878 = \mathbf{0.099}$$

e) Fem el mateix que en el primer apartat, és a dir, prenent X ="nombre de trucades telefòniques que es reben en una centraleta cada 5 minuts"$\sim Pois(\lambda = 3)$:

$$\mathbf{P(X = 0)} = P(X \leq 0) = \mathbf{0.050}$$



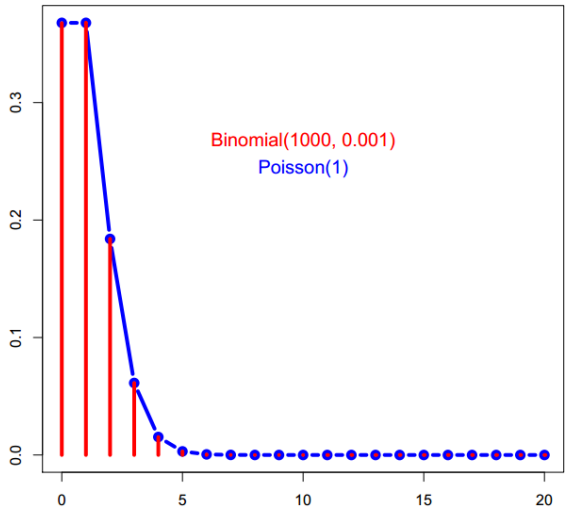
1.3. *Una companyia d'assegurances de vaixells de càrrega té assegurats un total de 1000 vaixells, el valor unitari dels quals és de 5 milions d'euros. Les clàusules de les assegurances cobreixen la pèrdua total d'un vaixell. La probabilitat que un vaixell tingui un sinistre total en el decurs d'un any és igual a 0.001. La legislació vigent obliga a la companyia a tenir un capital mínim destinat al pagament de sinistres que cobreixi els possibles sinistres amb probabilitat 0.999.*

Es demana:

a) Quin haurà de ser el capital anual mínim que haurà de tenir la companyia destinat al pagament de sinistres?
b) Si la companyia té unes despeses fixes anuals de 1.5 milions d'euros i vol tenir un benefici anual mitjà de 0.5 milions d'euros, quan haurà de cobrar a cada vaixell en concepte de prima d'assegurança?

Definim la v.a. discreta X com "nombre de sinistres totals per any entre els vaixells assegurats". X segueix una llei Binomial amb n=1000 i p=0.001. Donat que n és molt gran i p molt petita ($n \cdot p < 5$), podem aproximar aquesta distribució per una llei de Poisson de paràmetre $\lambda = n \cdot p = 1000 \cdot 0.001 = 1$:

$$X \sim \text{Bin}(n = 1000; p = 0.001) \sim \text{Pois}(\lambda = 1)$$



a) Calculem primer el nombre màxim de sinistres que es poden produir en un any amb probabilitat 0.999:

$$P(X \leq x) = 0.999$$

Busquem a les taules per $\lambda = 1$ i veiem que en el valor 5 hi ha una probabilitat acumulada de 0.999, per tant $x = 5$.

Sabent que es poden produir fins a 5 sinistres en un any amb una probabilitat igual a 0.999, calculem el capital anual mínim de què ha de disposar la companyia:

Capital mínim = 5 vaixells · 5 M€ = **25 M €**

b) Pel que fa als ingressos, si denotem per Q la quota anual que la companyia cobra a cada vaixell, els ingressos anuals de la companyia seran 1000·Q.
Pel que fa a les despeses, tenim unes despeses fixes de 1.5 milions d'euros. Cal també comptabilitzar les despeses que ocasionaran els sinistres que seran 5·X.
Així doncs,

$$\text{Benefici} = \text{Ingressos} - \text{Despeses} = 1000 \cdot Q - (1.5 + 5 \cdot X)$$

Volem E(Benefici)=0.5, per tant

$$0.5 = 1000 \cdot Q - 1.5 - 5 \cdot E(X)$$

Sabem que $E\{X\} = \lambda = 1$, per tant la quota anual que s'ha de cobrar a cada vaixell (Q) ha de ser:

$$0.5 = 1000 \cdot Q - 1.5 - 5$$

Aïllem Q i obtenim

$$\mathbf{Q} = 0.007 \text{ M} € = \mathbf{7000 €}$$

1.4. *La duració X (en h.) de les bombones de butà de 40 kg es distribueix segons una llei N(200h ; 20h). Calculeu la probabilitat que entre 4 bombones n'hi hagi 2, com a mínim, que durin entre 180h i 220h.*

Comencem utilitzant la llei normal per calcular la probabilitat que una bombona duri entre 180h i 220h:

$$P(180 \leq X \leq 220) =$$
$$= P\left(\frac{180-200}{20} \leq Z \leq \frac{220-200}{20}\right) = P(-1 \leq Z \leq 1) =$$
$$= P(Z \leq 1) - P(Z \leq -1) = 0.8413 - 0.1587 = 0.6826$$

Per calcular la probabilitat que de 4 bombolles n'hi hagi com a mínim 2 que durin entre 180h i 220h hem d'utilitzar una nova variable, en aquest cas una v.a. discreta Y que definim com "nombre de bombones d'entre 4 que duren entre 180h i 220h". Aquesta variable segueix la següent llei Binomial:

$$Y \sim Bin(n = 4; \lambda = 0.6826)$$

Calculem quant val $P(Y \geq 2)$:

$$P(Y \geq 2) = 1 - P(Y \leq 1)$$

Com que a les taules no trobem exactament el valor de p = 0.6826, fem el càlcul a partir de la fórmula de la funció de densitat de la llei Binomial:

$$P(Y \geq 2) = 1 - P(Y = 0) - P(Y = 1)$$

$$P(Y \geq 2) = 1 - \binom{4}{0} \cdot 0.6826^0 \cdot (0.3174)^4 - \binom{4}{1} \cdot 0.6826^1 \cdot (0.3174)^3$$

$$\mathbf{P(Y \geq 2) = 0.9025}$$

1.5. *El pes X (en kg.) de cada una de les caixes que circula per una cinta transportadora és una v.a. normal de μ = 30 kg i σ = 10 kg. Les caixes que circulen sobre la cinta són seleccionades aleatòriament. El número n de caixes que en un moment donat circulen simultàniament sobre la cinta ve donat per una llei Binomial de paràmetres n = 4 i p = 0.75. La cinta es para si el pes total de les caixes que transporta supera els 130 kg.*

Es demana:

a) Si en un determinat moment circulen 4 caixes, quina és la probabilitat que la cinta es pari?
b) Probabilitat que la cinta es pari.
c) Si la cinta s'ha parat, quina és la probabilitat que en la cinta hi hagi 3 caixes?

a) La v.a. X que definim com "pes d'1 caixa" segueix la següent llei Normal:

$$X \sim N(30\ kg\ ;\ 10\ kg)$$

Per tant, la v.a. X_4 que definim com "pes de 4 caixes" seguirà la següent distribució:

$$X + X + X + X \sim N\left(30 + 30 + 30 + 30; \sqrt{10^2 + 10^2 + 10^2 + 10^2}\right)$$

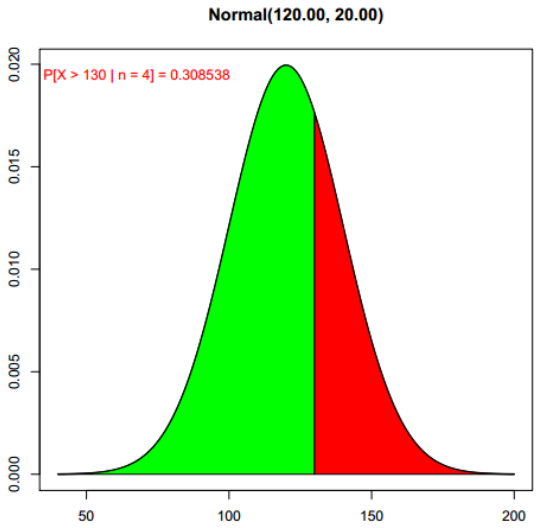
$$X_4 \sim N\left(4 \cdot 30; \sqrt{4 \cdot 10^2}\right)$$

$$X_4 \sim N(120\ kg\ ;\ 20\ kg)$$

La probabilitat que la cinta es pari és igual a la probabilitat que el pes de les 4 caixes superi els 130 kg i vindrà donada per:

$$\mathbf{P(X_4 > 130)} = 1 - P(X_4 \leq 130) =$$

$$= 1 - P\left(Z \leq \frac{130 - 120}{20}\right) = 1 - P(Z \leq 0.5) = 1 - 0.6915 = \mathbf{0.3085}$$



b) La probabilitat que la cinta es pari és diferent segons el nombre de caixes que es troben circulant simultàniament sobre la cinta. Segons el nombre de caixes 1, 2, 3 o 4, la distribució del pes serà diferent (amb distribució normal en tots els casos). Veiem quina és la probabilitat que la cinta es pari en aquests quatre casos (no tenim en compte la probabilitat que hi hagi 0 caixes sobre la cinta perquè sense pes en principi la cinta no s'ha de parar):

<u>Per 1 caixa</u>: $X \sim N(30; 10)$

$$P(X > 130) = 1 - P(X \leq 130) = 1 - P\left(Z \leq \frac{130 - 30}{10}\right) =$$

$$= 1 - P(Z \leq 10) = 1 - 1.0000 = 0.0000$$

Per 2 caixes: $X_2 \sim N\left(60; 10\sqrt{2}\right)$

$$P(X_2 > 130) = 1 - P(X_2 \leq 130) = 1 - P\left(Z \leq \frac{130 - 60}{10\sqrt{2}}\right) =$$

$$= 1 - P(Z \leq 4.95) = 1 - 1.0000 = 0.0000$$

Per 3 caixes: $X_3 \sim N\left(90; 10\sqrt{3}\right)$

$$P(X_2 > 130) = 1 - P(X_2 \leq 130) = 1 - P\left(Z \leq \frac{130 - 90}{10\sqrt{3}}\right) =$$

$$= 1 - P(Z \leq 2.31) = 1 - 0.9896 = 0.0104$$

Per 4 caixes (apartat a): $X_4 \sim N(120; 20)$

$$P(X_4 > 130) = 0.3085$$

Definim la v.a. discreta Y com "número de caixes que en un moment donat circulen simultàniament sobre la cinta". Segons l'enunciat aquesta variable segueix una distribució:

$$Y \sim Bin(n = 4; p = 0.75)$$

Com que a les taules de la llei Binomial el paràmetre p només arriba fins a 0.5, no ens és possible obtenir les probabilitats acumulades per p = 0.75. Per això definim la v.a. discreta complementària, amb paràmetre p = 1–0.75 = 0.25:

$$Z \sim Bin(n = 4; p = 0.25)$$

Calculem P(Y = 1), P(Y = 2), P(Y = 3) i P(Y = 4):

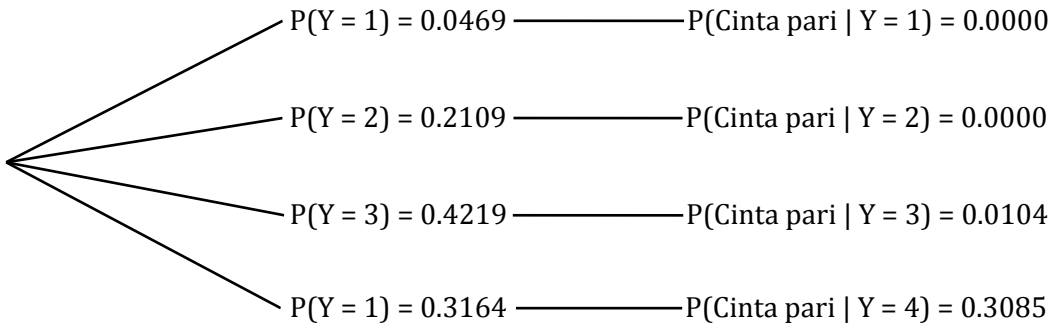
$$P(Y = 1) = P(Z = 4 - 1 = 3) = P(Z \leq 3) - P(Z \leq 2) = 0.9961 - 0.9492 = 0.0469$$

$$P(Y = 2) = P(Z = 4 - 2 = 2) = P(Z \leq 2) - P(Z \leq 1) = 0.9492 - 0.7383 = 0.2109$$

$$P(Y = 3) = P(Z = 4 - 3 = 1) = P(Z \leq 1) - P(Z \leq 0) = 0.7383 - 0.3164 = 0.4219$$

$$P(Y = 4) = P(Z = 4 - 4 = 0) = P(Z \leq 0) = 0.3164$$

Tenint en compte totes les probabilitats calculades, ja podem calcular quina es probabilitat que la cinta es pari:



P(Cinta pari) =

= P(Y = 1) · P(Cinta pari | Y = 1) + P(Y = 2) · P(Cinta pari | Y = 2) +

P(Y = 3) · P(Cinta pari | Y = 3) + P(Y = 4) · P(Cinta pari | Y = 4) =

= 0.0469 · 0.0000 + 0.2109 · 0.0000 + 0.4219 · 0.0104 + 0.3164 · 0.3085 = **0.1020**

c) Ens demanen que calculem P(Y = 3 | Cinta pari):

P(Y = 3 | Cinta pari) =

= P(Y = 3 ∩ Cinta pari) / P(Cinta pari) =

= P(Y = 3) · P(Cinta pari | Y = 3) / P(Cinta pari) =

= 0.4219 · 0.0104 / 0.1020 = **0.0430**

1.6. *Dins una campanya de màrqueting, en un sondeig d'opinió, es pregunta a una mostra de 100 persones quin dels dos productes –A o B– pensen comprar. El producte B l'elabora la companyia que ha encarregat la campanya i el producte A és el de la competència. Un total de 55 persones de la mostra declaren que comprarien el producte A mentre que els altres 45 prefereixen el B.*

Es demana:

a) Calcular, a un nivell de confiança del 95%, un interval d'estimació de la proporció de potencials compradors per a cadascun dels dos productes. Apliqueu el supòsit de màxima indeterminació.
b) Quina hauria de ser la mida mínima de la mostra perquè una proporció mostral igual a 0.55 pel producte A asseguri, a un nivell de confiança del 95%, que és més preferit que el producte B?

a) Calculem les proporcions mostrals pels dos productes

$$\widehat{p_A} = \frac{55}{100} = 0.55 \qquad \widehat{p_B} = \frac{45}{100} = 0.45$$

Donat que 1-α=0.95 i per tant $\alpha/2$=0.25, busquem a les taules de la llei normal estàndard els valors $\pm z_{\alpha/2}$

$$\pm z_{\alpha/2} = \pm z_{0.025} = \pm 1.960$$

Sabent que n=100, $\widehat{p_A} = 0.55$ i utilitzant el principi de màxima indeterminació, l'IC al 95% per p_A serà

$$\left(0.55 - 1.96\sqrt{\frac{1}{4 \cdot 100}}, 0.55 + 1.96\sqrt{\frac{1}{4 \cdot 100}}\right) = (\mathbf{0.452}, \mathbf{0.648})$$

Sabent que n=100, $\widehat{p_B} = 0.45$ i utilitzant el principi de màxima indeterminació, l'IC al 95% per p_B serà

$$\left(0.45 - 1.96\sqrt{\frac{1}{4 \cdot 100}}, 0.45 + 1.96\sqrt{\frac{1}{4 \cdot 100}}\right) = (\mathbf{0.352}, \mathbf{0.548})$$

b) Per assegurar, a un 95% de confiança, que el producte A és preferit al producte B, en fem prou exigint que l'extrem inferior d'un IC (95%) per p_A, calculat amb una mostra de mida n, sigui superior al 50%. Així doncs:

$$0.55 - 1.96\sqrt{\frac{1}{4 \cdot n}} > 0.50$$

Només cal aïllar la n de l'equació i obtenim n>384.16 i per tant podem afirmar que la mida mínima de la mostra ha de ser **385**.

1.7. *Les següents dades pertanyen al nombre de soldadures defectuoses trobades en 21 mostres consecutives de 500 juntes soldades cada una:*

106, 116, 164, 89, 99, 40, 112, 36, 69, 74, 42, 37, 25, 88, 101, 64, 51, 74, 71, 43, 80

Està sota control el procés?

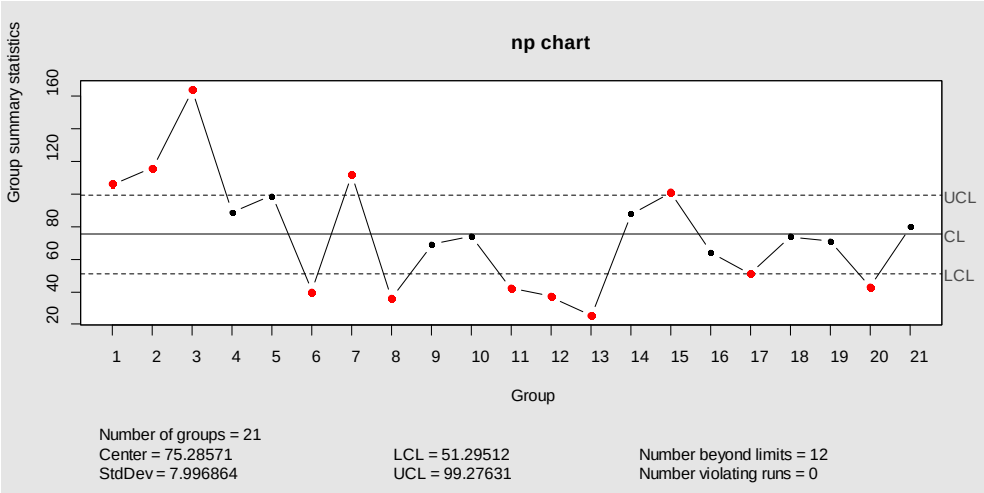
Utilitzarem el gràfic de control np.
Observem que tenim m=21 mostres de mida n=500. Així doncs, les línies de control valdran:

CCL= (106+116+⋯+80)/21 = 75.29

$$\text{LCL} = 75.29 - 3\sqrt{75.29\left(1 - \frac{75.29}{500}\right)} = 51.30$$

$$\text{UCL} = 75.29 + 3\sqrt{75.29\left(1 - \frac{75.29}{500}\right)} = 99.28$$

Seguidament només cal representar el nombre d'unitats defectuoses en cada una de les mostres en forma de diagrama temporal, juntament amb les línies de control.



Podem observar com el procés està totalment fora de control

1.8. *Una màquina A d'una empresa quan està centrada produeix, com a molt, un 1% de peces defectuoses. En un control de qualitat s'ha pres una mostra de 1500 peces fabricades amb aquesta màquina i ha resultat haver-hi 22 peces defectuoses. Feu el contrast adequat per a comprovar si la màquina està descentrada o no. Repetiu el mateix contrast en cas que el numero de defectuoses fos 30 i 50.*

En aquest cas resulta convenient plantejar un contrast unilateral per comprovar si la proporció de peces defectuoses és significativament superior al valor 0.01.

$$H_0 \ : \ p \leq 0.01$$
$$H_1 \ : \ p > 0.01$$

Farem aquest contrast prenent un nivell de significació $\alpha = 5\%$.
A partir de la proporció mostral $\hat{p} = 22/1500 = 0.015$ calculem l'estadístic de contrast

$$z_{obs} = \frac{0.015 - 0.01}{\sqrt{\frac{0.01(1-0.01)}{1500}}} = 1.82$$

Seguidament podem calcular el p-valor utilitzant les taules de la normal estàndard.

$$\text{p-valor} = P[Z > 1.82] = 1 - P[Z < 1.82] = 1 - 0.9656 = 0.0344$$

Donat que el p-valor és inferior al nivell de significació α ($0.0355 < 0.05$), es conclou que **tenim motius suficients per rebutjar la H_0** i acceptem la hipòtesi contrària H_1 que afirma que la màquina està fabricant més peces defectuoses de l'habitual.

En cas d'obtenir 30 i 50 peces defectuoses la conclusió serà la mateixa donat que tenim més defectes però amb una proporció mostral, un estadístic de contrast i un p-valor de

$$n = 30, \quad z_{obs} = 3.89, \quad p - valor = 0.000$$
$$n = 50, \quad z_{obs} = 9.08, \quad p - valor = 0.000$$

2. Exercicis proposats

2.1. *Un jugador de cartes extreu una carta a l'atzar d'una baralla espanyola (48 cartes). Si surt figura guanya 100€, si surt un as no guanya ni perd res i si surt qualsevol altra carta perd 25€. Quina és l'esperança de guany del jugador? I la variància?*

Solució: E{X} = 25/3 €; var{X} = 2847.2 €²

2.2. *La funció de distribució d'una v.a. X ve donada per:*

$$F(x) = f(x) = \begin{cases} 0, & x < 0 \\ 1/2, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

Busqueu la funció de masses de la v.a. X.

Solució: P(X = 0) = 1/2, P(X = 1) = 1/2

2.3. *Un magatzem majorista rep del proveïdor habitual una partida de 2000 bombetes. Per tal de controlar la qualitat de la partida s'escull una mostra a l'atzar de 20 bombetes. Si aquesta conté més d'una bombeta defectuosa, es decideix no acceptar la partida.*

Calculeu la probabilitat de:

a) Acceptar una partida que conté un 2% de bombetes defectuoses.
b) Rebutjar una partida que conté un 8 per mil de peces defectuoses.

Solució: a) 0.9401; b) 0.011

2.4. *Un magatzem de fruites comercialitza les llimones en capses de 200 unitats. La proporció de llimones malmeses és del 0.45%. Un eventual comprador, abans de fer l'encàrrec d'uns quants centenars de capses de llimones, decideix fer un control de qualitat que consisteix en escollir a l'atzar una de les capses i comprovar la qualitat de les llimones. Si no hi ha cap llimona en mal estat formalitza la compra. Si hi ha més de 2 llimones malmeses rebutja l'encàrrec. Si la capsa conté 1 o 2 llimones malmeses, escull a l'atzar una nova capça i, si aquesta conté menys de 2 llimones dolentes, formalitza la compra. En qualsevol altre cas, decideix no fer la compra. Calculeu la probabilitat que el comprador formalitzi la compra.*

Solució: 0.8162

2.5. *Un proveïdor ha de subministrar un lot de 100 peces a una indústria. Sap que el lot serà sotmès a un control abans que la indústria accepti el lot com a bo. El control consisteix en extreure del lot una mostra aleatòria simple de 3 peces. Si una o més*

peces de la mostra resulta defectuosa, la indústria rebutja el lot. En cas contrari, l'accepta.

a) Quin és el nombre màxim de peces defectuoses que pot contenir el lot per tal que la probabilitat que la indústria accepti el lot no sigui inferior a 0.95?
b) Suposeu ara que el control es realitza de la manera següent: s'extreu una primera mostra aleatòria simple de 5 peces. Si aquesta conté 2 o més peces defectuoses, la indústria rebutja el lot. Si no conté cap peça defectuosa, l'accepta. Si conté 1 peça defectuosa, extreu una altra mostra aleatòria simple de 5 peces. Si aquesta segona mostra no conté cap peça defectuosa, accepta el lot i, en cas contrari, el rebutja. Si el lot de 100 peces conté exactament 3 peces defectuoses, calculeu la probabilitat que la indústria acabi acceptant el lot.

Solució: a) x = 1; b) 0.9728

2.6. *La llargada de les peces fabricades per una determinada màquina s'ajusta a una distribució normal de mitjana 150 cm i desviació tipus 0.4 cm. Les peces es consideren acceptables si la seva llargada pertany a l'interval obert (149.2, 150.4).*

Es demana:

a) Si s'escull una mostra a l'atzar de 50 peces, calculeu la probabilitat que la mostra contingui exactament 44 peces acceptables.
b) Calculeu la probabilitat que hi hagi entre 38 i 45 peces acceptables, ambdós inclosos.

Solució: a) 0.0847; b) 0.8550

2.7. *Sigui X la variable aleatòria que indica el número de peces defectuoses produïdes per una màquina en un dia qualsevol de 8 h a 9 h del matí. Se sap que X té la següent funció de probabilitat:*

X	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
P(X = x)	*0.005*	*0.020*	*0.040*	*0.090*	*0.110*	*0.185*	*k*	*0.170*	*0.140*	*0.100*	*0.030*

Es demana:

a) Trobeu el valor de k.
b) Calculeu la funció de distribució.
c) Calculeu P(4 ≤ X ≤ 9).

Solució: a) k = 0.11; b) F(0) = 0.005, F(1) = 0.025, F(2) = 0.065, F(3) = 0.155, F(4) = 0.265, F(5) = 0.45, F(6) = 0.56, F(7) = 0.73, F(8) = 0.87, F(9) = 0.97, F(10) = 1; c) 0.815

2.8. *En un taller hi ha 10 màquines iguals. S'ha vist que les màquines estan avariades un de cada cinc dies de forma independent.*

a) Quina és la probabilitat que un cert dia hi hagi més de 6 màquines avariades?
b) Si la pèrdua diària ocasionada per tenir una màquina avariada és de 200 euros, calcula la pèrdua mitjana diària.

Solució: a) 0.0009; b) 400 €

2.9. *Del magatzem d'una fàbrica de plaques i peces metàl·liques s'ha escollit una mostra aleatòria de 400 cargols del model CM324. En la mostra s'ha observat que 260 dels cargols tenien alguna taca en la seva superfície. Trobeu l'interval del 95% de confiança per a la verdadera proporció de cargols amb alguna taca.*

Solució: IC(95%) = (60.3%, 69.7%)

2.10. *Es varen seleccionar 25 lots de 100 làmpades, cada un els quals tenien el nombre següent de làmpades defectuoses:*

3, 4, 6, 4, 0, 5, 2, 3, 0, 2, 3, 5, 3, 9, 1, 2, 4, 4, 1, 8, 4, 6, 5, 3, 2

Es demana:

a) Es pot acceptar, com afirma el fabricant, que aquest procés produeix en mitjana un 2% de làmpades defectuoses?
b) Amb les dades de la mostra calculeu uns nous límits de control per al nombre de làmpades defectuoses en un lot de mida 100.
c) Quina hauria de ser la proporció de làmpades defectuoses en mitjana que hauria d'especificar el fabricant?

Solució: a) Per d = 2, límits control gràfic np: (0, 6.2)
(valors 8 i 9, procés fora de control);
b) Per dades mostrals, límits control gràfic np: (0, 9.1187)
(procés sota control)

2.11. *Durant la fabricació de peces d'un aparell elèctric, s'han pres 33 mostres de 50 elements cada una cada 4 hores. S'ha registrat la següent quantitat d'elements defectuosos entre aquestes 50 peces:*

3, 3, 2, 0, 6, 1, 1, 1, 2, 1, 2, 3, 3, 0, 8, 0, 6, 5, 5, 0, 3, 3, 2, 1, 3, 4, 5, 3, 4, 5, 4, 6, 1

Utilitzeu aquestes dades per a establir uns valors de control per a la proporció d'elements defectuosos que produeix aquest procés, eliminant, si fos necessari, les mostres fora de control.

Solució: Límits control gràfic p: (0, 0.1574)

(valor 8/50 = 0.16 fora dels límits, l'eliminem);
Límits control gràfic p: (0, 0.1516)
(procés sota control)

2.12. *Dins un projecte de millora de la qualitat, una indústria tèxtil decideix controlar el nombre d'imperfeccions trobades en cada peça de roba. S'estima que en mitjana el nombre d'imperfeccions per cada peça de roba és de 12. Es demana:*

a) Calculeu la probabilitat que en una d'aquestes peces de roba es trobin entre 10 i 12 imperfeccions, ambdós valors inclosos.
b) Calculeu la probabilitat que en una d'aquestes peces de roba es trobin menys de 8 o més de 16 imperfeccions.
c) Inspeccionat un lot de 25 peces de roba, s'ha trobat els següents nombre de defectes en cada una d'elles:

13, 15, 9, 7, 12, 8, 4, 10, 3, 5, 8, 14, 10, 11, 14, 15, 7, 16, 8, 8, 9, 14, 17, 13, 9

Es manté 12 com a nombre d'imperfeccions en mitjana? Realitzeu el gràfic de control.

Solució: a) 0.334; b) 0.191; c) Interval històric amb $\lambda = 12$: (1.608, 22.392)
(tots els valors estan entre els límits de control)

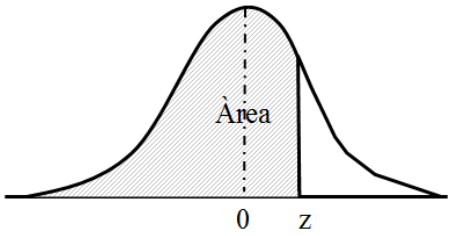
2.13. *Per tal d'estudiar el nombre d'accidents a les PYME (Pequeña Y Mediana Empresa) s'han fet 250 observacions en diferents empreses, en 30 de les quals hi ha hagut accidents.*

a) Si per estimar la proporció real utilitzem l'estimació per interval de confiança del 95%, quina hauria de ser la mida de la mostra si volem que la precisió de l'estimació no excedeixi de 0.05? (Utilitzeu el supòsit de màxima indeterminació).
b) Sabem, per estudis anteriors, que la proporció d'accidents a les PYME és del 16%. Tenim motius suficients per acceptar la hipòtesi que el percentatge d'accidents ha disminuït?
c) Calculeu el nombre mínim d'accidents que es poden admetre en una altra mostra de grandària 400 per no acceptar la hipòtesi nul·la de l'apartat anterior, amb un risc d'error del 5%.

Solució: a) 385; b) p-valor=0.0422 amb α=0.05 rebutgem H_0, c)51

TAULES ESTADÍSTIQUES

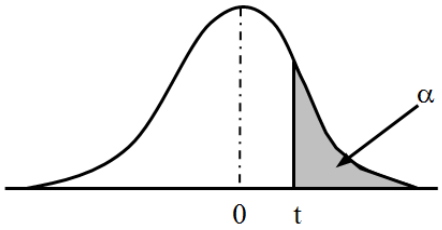
1. Funció de distribució NORMAL ESTÀNDARD



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262 .	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

2. Funció de distribució t-STUDENT



υ	α=0.1	α=0.075	α=0.05	α=0.04	α=0.03	α=0.025	α=0.02	α=0.01	α=0.005	α=0.0025	α=0.001	α=0.0005
1	3.078	4.165	6.314	7.916	10.579	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	1.886	2.282	2.920	3.320	3.896	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	1.638	1.924	2.353	2.605	2.951	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	1.533	1.778	2.132	2.333	2.601	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	1.476	1.699	2.015	2.191	2.422	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	1.440	1.650	1.943	2.104	2.313	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	1.415	1.617	1.895	2.046	2.241	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	1.397	1.592	1.860	2.004	2.189	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	1.383	1.574	1.833	1.973	2.150	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	1.372	1.559	1.812	1.948	2.120	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	1.363	1.548	1.796	1.928	2.096	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	1.356	1.538	1.782	1.912	2.076	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	1.350	1.530	1.771	1.899	2.060	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	1.345	1.523	1.761	1.887	2.046	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	1.341	1.517	1.753	1.878	2.034	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	1.337	1.512	1.746	1.869	2.024	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	1.333	1.508	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	1.330	1.504	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	1.328	1.500	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	1.325	1.497	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	1.323	1.494	1.721	1.840	1.988	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	1.321	1.492	1.717	1.835	1.983	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	1.319	1.489	1.714	1.832	1.978	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	1.318	1.487	1.711	1.828	1.974	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	1.316	1.485	1.708	1.825	1.970	2.060	2.167	2.485	2.787	3.078	3.450	3.725
30	1.310	1.477	1.697	1.812	1.955	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	1.303	1.468	1.684	1.796	1.936	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	1.299	1.462	1.676	1.787	1.924	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	1.296	1.458	1.671	1.781	1.917	2.000	2.099	2.390	2.660	2.915	3.232	3.460
70	1.294	1.456	1.667	1.776	1.912	1.994	2.093	2.381	2.648	2.899	3.211	3.435
80	1.292	1.453	1.664	1.773	1.908	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	1.290	1.451	1.660	1.769	1.902	1.984	2.081	2.364	2.626	2.871	3.174	3.390
120	1.289	1.449	1.658	1.766	1.899	1.980	2.076	2.358	2.617	2.860	3.160	3.373
200	1.286	1.445	1.653	1.760	1.892	1.972	2.067	2.345	2.601	2.839	3.131	3.340
500	1.283	1.442	1.648	1.754	1.885	1.965	2.059	2.334	2.586	2.820	3.107	3.310
∞	1.282	1.440	1.645	1.751	1.881	1.960	2.054	2.326	2.576	2.807	3.090	3.291

3. Gràfics de control

Mida del subgrup	Mitjanes			Desviacions tipus					Rangs						
	Factors límits de control			Factor línia central	Factors límits de control				Factors línia central		Factors límits de control				
	A	A_2	A_3	c_4	B_3	B_4	B_5	B_6	d_2	$1/d_2$	d_3	D_1	D_2	D_3	D_4
2	2.121	1.880	2.659	0.7979	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.575
4	1.500	0.729	1.628	0.9213	0	2.266	0	2.08S	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.115
6	1.225	0.483	1.287	0.9515	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.205	5.203	0.076	1.924
8	1.061	0.373	1.099	0.9650	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.387	5.307	0.136	1.864
9	1.000	0.337	1.032	0.9693	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.546	5.394	0.184	1.816
10	0.949	0.308	0.975	0.9727	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.812	5.534	0.256	1.744
12	0.866	0.266	0.886	0.9776	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.924	5.592	0.284	1.716
13	0.832	0.249	0.850	0.9794	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.026	5.646	0.308	1.692
14	0.802	0.235	0.817	0.9810	0.406	1.594	0.399	1.563	3.407	0.2935	0.762	1.121	5.693	0.329	1.671
15	0.775	0.223	0.789	0.9823	0.428	1.572	0.421	1.544	3.472	0.2880	0.755	1.207	5.737	0.348	1.652
16	0.750	0.212	0.763	0.9835	0.448	1.552	0.440	1.526	3.532	0.2831	0.749	1.285	5.779	0.364	1.636
17	0.728	0.203	0.739	0.9845	0.466	1.534	0.458	1.511	3.588	0.2787	0.743	1.359	5.817	0.379	1.621
18	0.707	0.194	0.718	0.9854	0.482	1.518	0.4/5	1.496	3.640	0.2747	0.738	1.426	5.854	0.392	1.608
19	0.688	0.187	0.698	0.9862	0.497	1.503	0.490	1.483	3.689	0.2711	0.733	1.490	5.888	0.404	1.596
20	0.671	0.180	0.680	0.9869	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.548	5.922	0.414	1.586
21	0.655	0.173	0.663	0.9876	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.606	5.950	0.425	1.575
22	0.640	0.167	0.647	0.9882	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.452	1.548
25	0.600	0.153	0.606	0.9896	0.565	1.435	0.559	1.420	3.931	0.2544	0.709	1.804	6.058	0.459	1.541
> 25	$3/\sqrt{n}$				$1-\frac{3}{\sqrt{2n}}$	$1+\frac{3}{\sqrt{2n}}$									

4. Corbes característiques

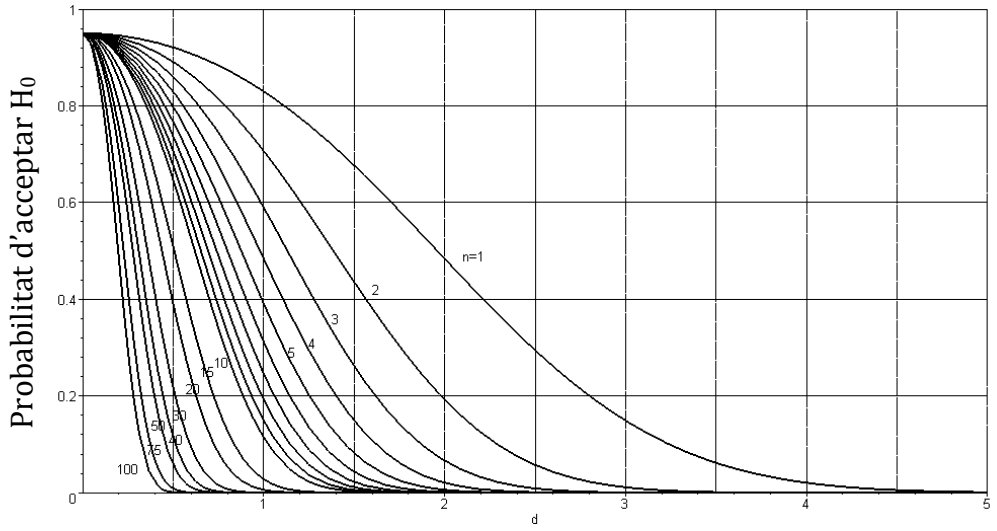
4.1. Basades en la distribució normal (σ^2 coneguda)

- **Contrast bilateral**

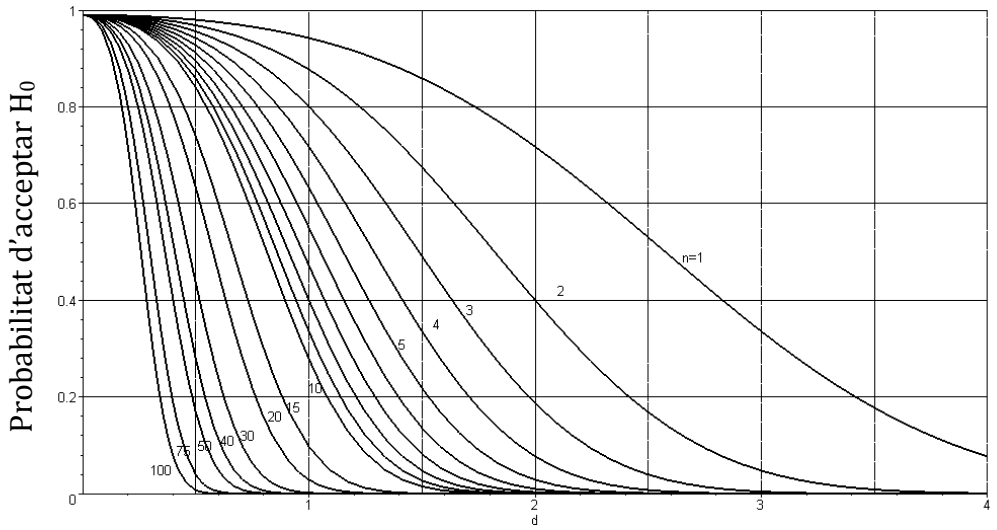
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

$$d = \frac{|\mu - \mu_0|}{\sigma}$$

α = 0.05



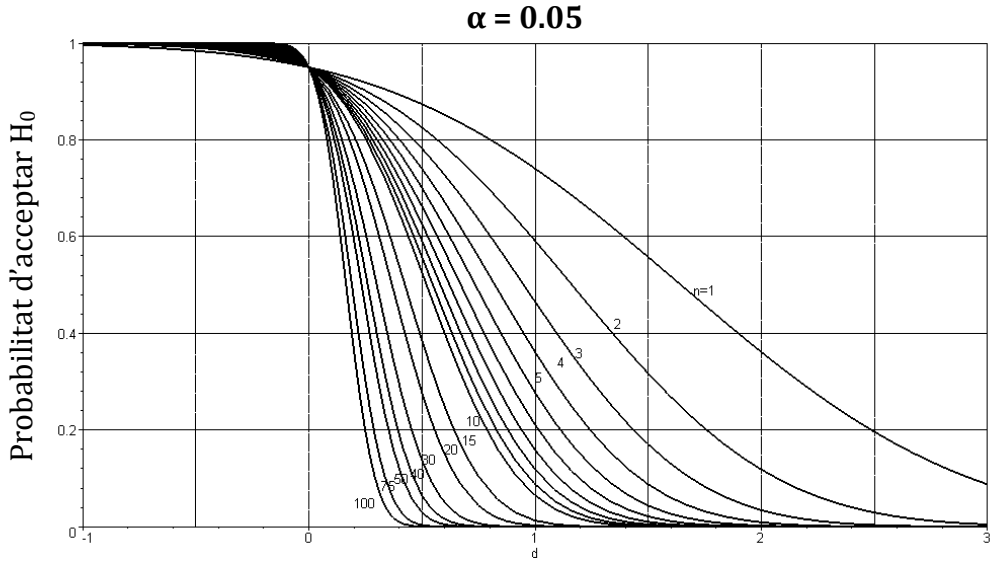
α = 0.01

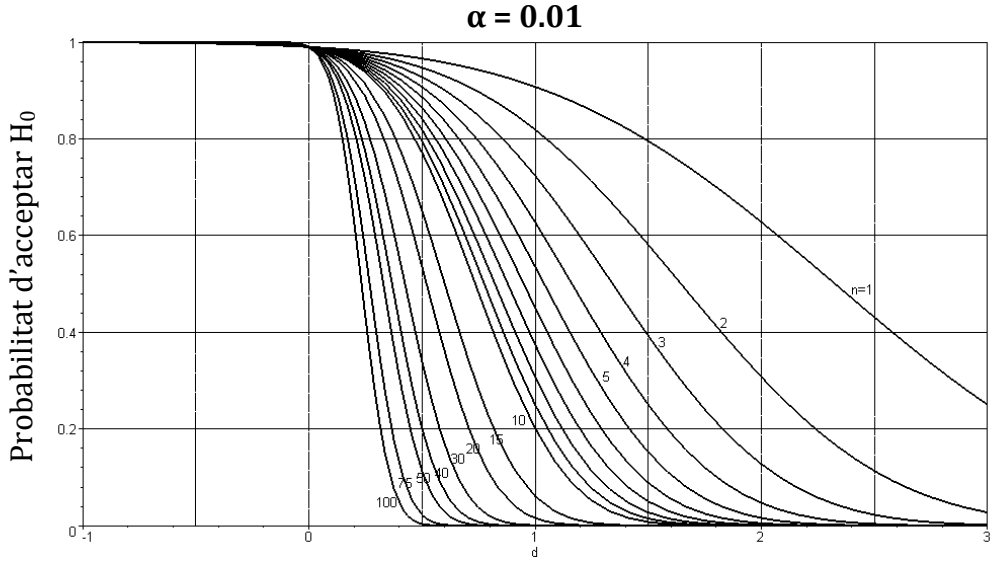


- **Contrast unilateral**

$$H_0 : \mu \leq \mu_0 \qquad H_0 : \mu \geq \mu_0$$
$$H_1 : \mu > \mu_0 \qquad H_1 : \mu < \mu_0$$

$$d = \frac{|\mu - \mu_0|}{\sigma} \qquad d = \frac{|\mu_0 - \mu|}{\sigma}$$





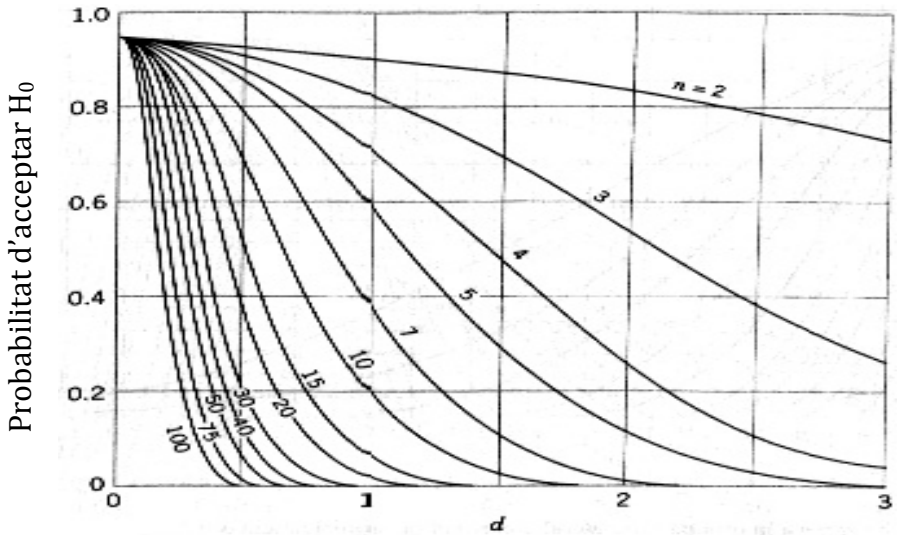
4.2. Basades en la distribució t-Student (σ^2 desconeguda)

- **Contrast bilateral**

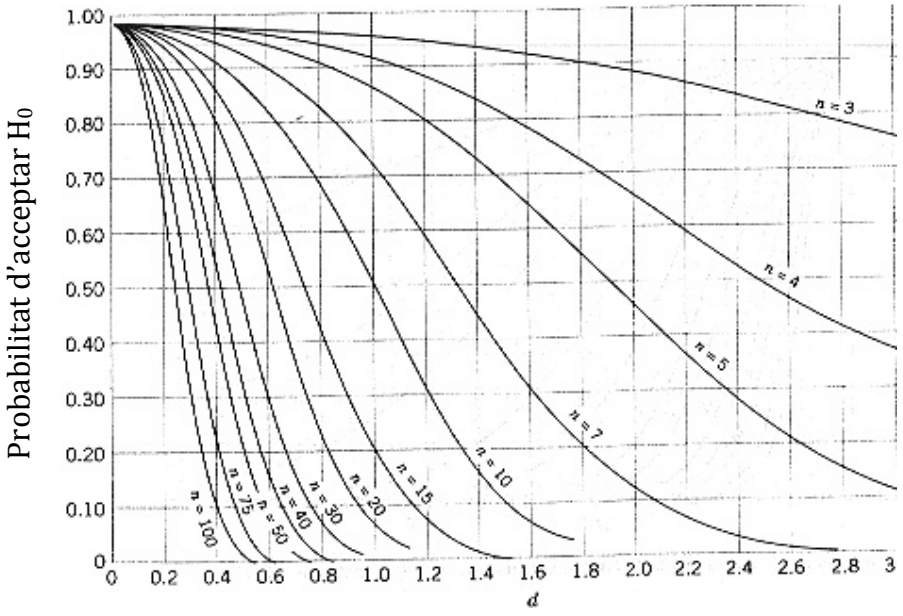
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

$$d = \frac{|\mu - \mu_0|}{s}$$

α = 0.05



α = 0.01

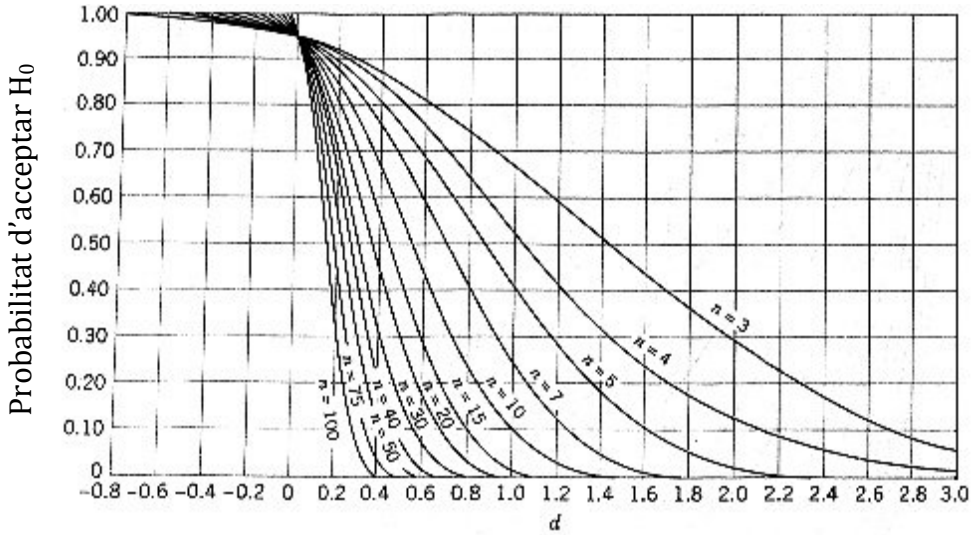


- **Contrast unilateral**

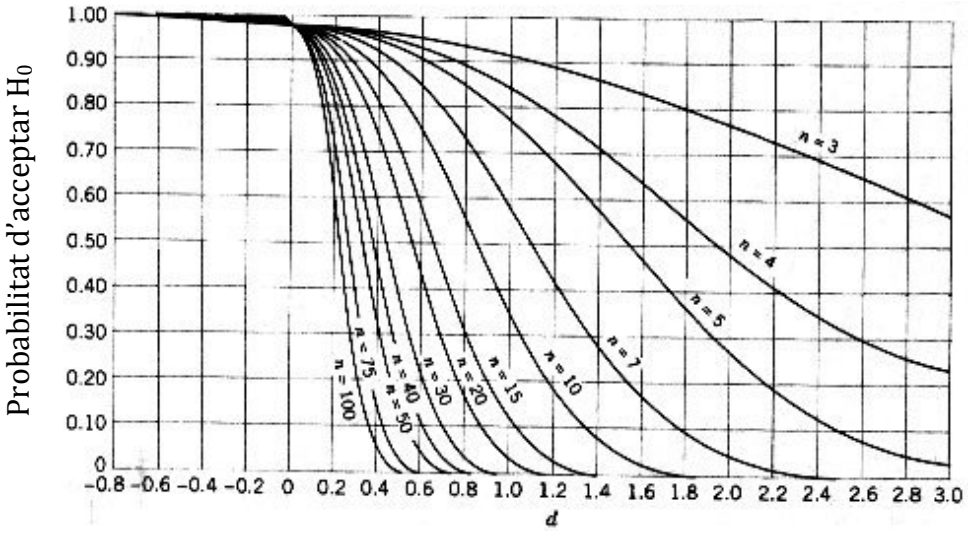
$$H_0 : \mu \leq \mu_0 \qquad H_0 : \mu \geq \mu_0$$
$$H_1 : \mu > \mu_0 \qquad H_1 : \mu < \mu_0$$

$$d = \frac{|\mu - \mu_0|}{s} \qquad d = \frac{|\mu_0 - \mu|}{s}$$

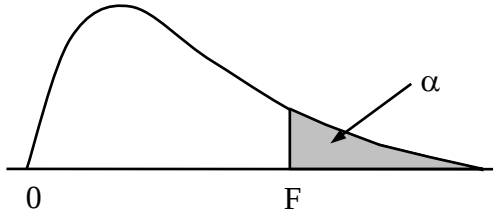
α = 0.05



α = 0.01



5. <u>Funció de distribució F de FISHER</u> ($\alpha = 0.05$)



Graus de llibertat del numerador υ_1

Graus de llibertat del denominador υ_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	45.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	1930	19.33	19.35	19.37	1933	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	894	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	436
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	334	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	307	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.51	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	268	2.57	249	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	..92	1.87	1.81
22	4.30	3.44	3.05	2.82	266	2.55	2.46	2.40	234	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	249	2.40	2.34	2.28	2.24	216	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	206	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	269	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.06	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	315	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	210	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Funció de distribució F de FISHER ($\alpha = 0.025$)

Graus de llibertat del numerador υ_1

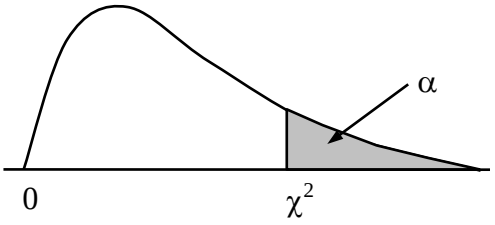
Graus de llibertat del denominador υ_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001.0	1006.0	1010.0	1014.0	1018.0
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	996	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5-70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.62	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3-51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	469	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.67	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.62	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	214	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	236	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.96	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	359	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.00	1.60	1.72	l.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	3.06	1.94	1.68	1.82	1.74	1.67	1.58	1.46
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

Funció de distribució F de FISHER ($\alpha = 0.01$)

Graus de llibertat del numerador υ_1

Graus de llibertat del denominador υ_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261.0	6287.0	6313.0	6339.0	6366.0
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.59
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	782	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	336	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.96	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.69	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	332	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

6. Funció de distribució XI-QUADRAT



υ	α = 0.995	α = 0.99	α = 0.975	α = 0.95	α = 0.05	α = 0.025	α = 0.01	α = 0.005
1	0.0000393	0.000157	0.000982	0.00393	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.0717	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.688	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672

7. Funció de distribució BINOMINAL

n	x	p									
		0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
2	**0**	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	**0**	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.9928	0.9720	0.9392	0.8960	0.8438	0.7840	0.7182	0.6480	0.5748	0.5000
	2	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	**0**	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125
	2	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
	3	1.0000	0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	**0**	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
	1	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875
	2	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000
	3	1.0000	0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125
	4	1.0000	1.0000	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	**0**	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.9672	0.8857	0.7765	0.6553	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
	2	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438
	3	0.9999	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6562
	4	1.0000	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906
	5	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	**0**	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
	2	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
	3	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
	4	1.0000	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
	5	1.0000	1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
	6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

n	x	p									
		0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
8	**0**	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
	2	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
	3	0.9996	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633
	4	1.0000	0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367
	5	1.0000	1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
	6	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
	7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	**0**	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
	2	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4	1.0000	0.9991	0.9944	0.9804	0.9511	0.9012	0-8283	0.7334	0.6214	0.5000
	5	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	**0**	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
	2	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
	3	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
	4	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
	5	1.0000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230
	6	1.0000	1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
	7	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
	8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
11	**0**	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005
	1	0.8981	0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0302	0.0139	0.0059
	2	0.9848	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327
	3	0.9984	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133
	4	0.9999	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744
	5	1.0000	0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5000
	6	1.0000	1.0000	0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256
	7	1.0000	1.0000	1.0000	0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867
	8	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9980	0.9941	0.9852	0.9673
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9978	0.9941
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9995
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

n	x	p									
		0.00	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
12	**0**	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
	1	0.8816	0.6590	0.4435	0.2749	0.1584	0.0850	0.0424	0.0196	0.0083	0.0032
	2	0.9804	0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193
	3	0.9978	0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.0730
	4	0.9998	0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938
	5	1.0000	0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872
	6	1.0000	0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128
	7	1.0000	1.0000	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062
	8	1.0000	1.0000	1.0000	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.9270
	9	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9972	0.9921	0.9807
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9968
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
13	**0**	0.5133	0.2542	0.1209	0.0550	0.0238	0.0097	0.0037	0.0013	0.0004	0.0001
	1	0.8646	0.6213	0.3983	0.2336	0.1267	0.0637	0.0296	0.0126	0.0049	0.0017
	2	0.9755	0.8661	0.6920	0.5017	0.3326	0.2025	0.1132	0.0579	0.0269	0.0112
	3	0.9969	0.9658	0.8820	0.7473	0.5843	0.4206	0.2783	0.1686	0.0929	0.0461
	4	0.9997	0.9935	0.9658	0.9009	0.7940	0.6543	0.5005	0.3530	0.2279	0.1334
	5	1.0000	0.9991	0.9924	0.9700	0.9198	0.8346	0.7159	0.5744	0.4268	0.2905
	6	1.0000	0.9999	0.9987	0.9930	0.9757	0.9376	0.8705	0.7712	0.6437	0.5000
	7	1.0000	1.0000	0.9998	0.9988	0.9944	0.9818	0.9538	0.9023	0.8212	0.7095
	8	1.0000	1.0000	1.0000	0.9998	0.9990	0.9960	0.9874	0.9679	0.9302	0.8666
	9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9975	0.9922	0.9797	0.9539
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9987	0.9959	0.9888
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	**0**	0.4877	0.2288	0.1028	0.0440	0.0178	0.0068	0.0024	0.0008	0.0002	0.0001
	1	0.8470	0.5846	0.3567	0.1979	0.1010	0.0475	0.0205	0.0081	0.0029	0.0009
	2	0.9699	0.8416	0.6479	0.4481	0.2811	0.1608	0.0839	0.0398	0.0170	0.0065
	3	0.9958	0.9559	0.8535	0.6982	0.5213	0.3552	0.2205	0.1243	0.0632	0.0287
	4	0.9996	0.9908	0.9533	0.8702	0.7415	0.5842	0.4227	0.2793	0.1672	0.0898
	5	1.0000	0.9985	0.9885	0.9561	0.8883	0.7805	0.6405	0.4859	0.3373	0.2120
	6	1.0000	0.9998	0.9978	0.9884	0.9617	0.9067	0.8164	0.6925	0.5461	0.3953
	7	1.0000	1.0000	0.9997	0.9976	0.9897	0.9685	0.9247	0.8499	0.7414	0.6047
	8	1.0000	1.0000	1.0000	0.9996	0.9978	0.9917	0.9757	0.9417	0.8811	0.7880
	9	1.0000	1.0000	1.0000	1.0000	0.9997	0.9983	0.9940	0.9825	0.9574	0.9102
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9989	0.9961	0.9886	0.9713
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9978	0.9935
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

n	x	p									
		0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
15	**0**	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000
	1	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005
	2	0.9638	0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037
	3	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176
	4	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592
	5	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509
	6	1.0000	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036
	7	1.0000	1.0000	0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000
	8	1.0000	1.0000	0.9999	0.9992	0.9958	0.9848	0.9578	0.9050	0.8182	0.6964
	9	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491
	10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9937	0.9824
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9963
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
16	**0**	0.4401	0.1853	0.0743	0.0281	0.0100	0.0033	0.0010	0.0003	0.0001	0.0000
	1	0.8108	0.5147	0.2839	0.1407	0.0635	0.0261	0.0098	0.0033	0.0010	0.0003
	2	0.9571	0.7892	0.5614	0.3518	0.1971	0.0994	0.0451	0.0183	0.0066	0.0021
	3	0.9930	0.9316	0.7899	0.5981	0.4050	0.2459	0.1339	0.0651	0.0281	0.0106
	4	0.9991	0.9830	0.9209	0.7982	0.6302	0.4499	0.2892	0.1666	0.0853	0.0384
	5	0.9999	0.9967	0.9765	0.9183	0.8103	0.6598	0.4900	0.3288	0.1976	0.1051
	6	1.0000	0.9995	0.9944	0.9733	0.9204	0.8247	0.6881	0.5272	0.3660	0.2272
	7	1.0000	0.9999	0.9989	0.9930	0.9729	0.9256	0.8406	0.7161	0.5629	0.4018
	8	1.0000	1.0000	0.9998	0.9985	0.9925	0.9743	0.9329	0.8577	0.7441	0.5982
	9	1.0000	1.0000	1.0000	0.9998	0.9984	0.9929	0.9771	0.9417	0.8759	0.7728
	10	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9938	0.9809	0.9514	0.8949
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9851	0.9616
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9991	0.9965	0.9894
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	] .0000	0.9999	0.9994	0.9979
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

n	x	p									
		0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
17	**0**	0.4181	0.1668	0.0631	0.0225	0.0075	0.0023	0.0007	0.0002	0.0000	0.0000
	1	0.7922	0.4818	0.2525	0.1182	0.0501	0.0193	0.0067	0.0021	0.0006	0.0001
	2	0.9497	0.7618	0.5198	0.3096	0.1637	0.0774	0.0327	0.0123	0.0041	0.0012
	3	0.9912	0.9174	0.7556	0.5489	0.3530	0.2019	0.1028	0.0464	0.0184	0.0063
	4	0.9988	0.9779	0.9013	0.7582	0.5739	0.3887	0.2348	0.1260	0.0596	0.0245
	5	0.9999	0.9953	0.9681	0.8943	0.7653	0.5968	0.4197	0.2639	0.1471	0.0717
	6	1.0000	0.9992	0.9917	0.9623	0.8929	0.7752	0.6188	0.4478	0.2902	0.1662
	7	1.0000	0.9999	0.9983	0.9891	0.9598	0.8954	0.7872	0.6405	0.4743	0.3145
	8	1.0000	1.0000	0.9997	0.9974	0.9876	0.9597	0.9006	0.8011	0.6626	0.5000
	9	1.0000	1.0000	1.0000	0.9995	0.9969	0.9873	0.9617	0.9081	0.8166	0.6855
	10	1.0000	1.0000	1.0000	0.9999	0.9994	0.9968	0.9880	0.9652	0.9174	0.8338
	11	1.0000	1 .0000	1.0000	1.0000	0.9999	0.9993	0.9970	0.9894	0.9699	0.9283
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9975	0.9914	0.9755
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9936
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9988
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	**0**	0.3972	0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000
	1	0.7735	0.4503	0.2241	0.0991	0.0395	0.0142	0.0046	0.0013	0.0003	0.0001
	2	0.9419	0.7338	0.4797	0.2713	0.1353	0.0600	0.0236	0.0082	0.0025	0.0007
	3	0.9891	0.9018	0.7202	0.5010	0.3057	0.1646	0.0783	0.0328	0.0120	0.0038
	4	0.9985	0.9718	0.8794	0.7164	0.5187	0.3327	0.1886	0.0942	0.0411	0.0154
	5	0.9998	0.9936	0.9581	0.8671	0.7175	0.5344	0.3550	0.2088	0.1077	0.0481
	6	1.0000	0.9988	0.9882	0.9487	0.8610	0.7217	0.5491	0.3743	0.2258	0.1189
	7	1.0000	0.9998	0.9973	0.9837	0.9431	0.8593	0.7283	0.5634	0.3915	0.2403
	8	1.0000	1.0000	0.9995	0.9957	0.9807	0.9404	0.8609	0.7368	0.5778	0.4073
	9	1.0000	1.0000	0.9999	0.9991	0.9946	0.9790	0.9403	0.8653	0.7473	0.5927
	10	1.0000	1.0000	1.0000	0.9998	0.9988	0.9939	0.9788	0.9424	0.8720	0.7597
	11	1.0000	1.0000	1.0000	1.0000	0.9998	0.9986	0.9938	0.9797	0.9463	0.8811
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9986	0.9942	0.9817	0.9519
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9846
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9990	0.9962
	15	1 .0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

n	x	p									
		0.05	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
19	**0**	0.3774	0.1351	0.0456	0.0144	0.0042	0.0011	0.0003	0.0001	0.0000	0.0000
	1	0.7547	0.4203	0.1985	0.0829	0.0310	0.0104	0.0031	0.0008	0.0002	0.0000
	2	0.9335	0.7054	0.4413	0.2369	0.1113	0.0462	0.0170	0.0055	0.0015	0.0004
	3	0.9868	0.8850	0.6841	0.4551	0.2630	0.1332	0.0591	0.0230	0.0077	0.0022
	4	0.9980	0.9648	0.8556	0.6733	0.4654	0.2822	0.1500	0.0696	0.0280	0.0096
	5	0.9998	0.9914	0.9463	0.8369	0.6678	0.4739	0.2968	0.1629	0.0777	0.0318
	6	1.0000	0.9983	0.9837	0.9324	0.8251	0.6655	0.4812	0.3081	0.1727	0.0835
	7	1.0000	0.9997	0.9959	0.9767	0.9225	0.8180	0.6656	0.4878	0.3169	0.1796
	8	1.0000	1.0000	0.9992	0.9933	0.9713	0.9161	0.8145	0.6675	0.4940	0.3238
	9	1.0000	1.0000	0.9999	0.9984	0.9911	0.9674	0.9125	0.8139	0.6710	0.5000
	10	1.0000	1.0000	1.0000	0.9997	0.9977	0.9895	0.9653	0.9115	0.8159	0.6762
	11	1.0000	1.0000	1.0000	1.0000	0.9995	0.9972	0.9886	0.9648	0.9129	0.8204
	12	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9969	0.9884	0.9658	0.9165
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9969	0.9891	0.9682
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9972	0.9904
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9978
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	**0**	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000
	1	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0021	0.0005	0.0001	0.0000
	2	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0121	0.0036	0.0009	0.0002
	3	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0444	0.0160	0.0049	0.0013
	4	0.9974	0.9568	0.8298	0.6296	0.4148	0.2375	0.1182	0.0510	0.0189	0.0059
	5	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.0553	0.0207
	6	1.0000	0.9976	0.9781	0.9133	0.7858	0.6080	0.4166	0.2500	0.1299	0.0577
	7	1.0000	0.9996	0.9941	0.9679	0.8982	0.7723	0.6010	0.4159	0.2520	0.1316
	8	1.0000	0.9999	0.9987	0.9900	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517
	9	1.0000	1.0000	0.9998	0.9974	0.9861	0.9520	0.8782	0.7553	0.5914	0.4119
	10	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881
	11	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483
	12	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9940	0.9790	0.9420	0.8684
	13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9935	0.9786	0.9423
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9936	0.9793
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9941
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

8. Funció de distribució POISSON

x	λ = E(X)									
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.449	0.407	0.368
1	0.995	0.982	0.963	0.938	0.910	0.878	0.844	0.809	0.772	0.736
2	1.000	0.999	0.996	0.992	0.986	0.977	0.966	0.953	0.937	0.920
3	1.000	1.000	1.000	0.999	0.998	0.997	0.994	0.991	0.987	0.981
4	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.996
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

x	λ = E(X)									
	1.1	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2.0**
0	0.333	0.301	0.273	0.247	0.223	0.202	0.183	0.165	0.150	0.135
1	0.699	0.663	0.627	0.592	0.558	0.525	0.493	0.463	0.434	0.406
2	0.900	0.879	0.857	0.833	0.809	0.783	0.757	0.731	0.704	0.677
3	0.974	0.966	0.957	0.946	0.934	0.921	0.907	0.891	0.875	0.857
4	0.995	0.992	0.989	0.986	0.981	0.976	0.970	0.964	0.956	0.947
5	0.999	0.998	0.998	0.997	0.996	0.994	0.992	0.990	0.987	0.983
6	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.997	0.995
7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999
8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

x	λ = E(X)									
	2.2	**2.4**	**2.6**	**2.8**	**3.0**	**3.2**	**3.4**	**3.6**	**3.8**	**4.0**
0	0.111	0.091	0.074	0.061	0.050	0.041	0.033	0.027	0.022	0.018
1	0.355	0.308	0.267	0.231	0.199	0.171	0.147	0.126	0.107	0.092
2	0.623	0.570	0.518	0.469	0.423	0.380	0.340	0.303	0.269	0.238
3	0.819	0.779	0.736	0.692	0.647	0.603	0.558	0.515	0.473	0.433
4	0.928	0.904	0.877	0.848	0.815	0.781	0.744	0.706	0.668	0.629
5	0.975	0.964	0.951	0.935	0.916	0.895	0.871	0.844	0.816	0.785
6	0.993	0.988	0.983	0.976	0.966	0.955	0.942	0.927	0.909	0.889
7	0.998	0.997	0.995	0.992	0.988	0.983	0.977	0.969	0.960	0.949
8	1.000	0.999	0.999	0.998	0.996	0.994	0.992	0.988	0.984	0.979
9	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0996	0.994	0.992
10	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

x	λ = E(X)									
	4.2	**4.4**	**4.6**	**4.8**	**5.0**	**5.2**	**5.4**	**5.6**	**5.8**	**6.0**
0	0.015	0.012	0.010	0.008	0.007	0.006	0.005	0.004	0.003	0.002
1	0.078	0.066	0.056	0.048	0.040	0.034	0.029	0.024	0.021	0.017
2	0.210	0.185	0.163	0.143	0.125	0.109	0.095	0.082	0.072	0.062
3	0.395	0.359	0.326	0.294	0.265	0.238	0.213	0.191	0.170	0.151
4	0.590	0.551	0.513	0.476	0.440	0.406	0.373	0.342	0.313	0.285
5	0.753	0.720	0.686	0.651	0.616	0.581	0.546	0.512	0.478	0.446
6	0.867	0.844	0.818	0.791	0.762	0.732	0.702	0.670	0.638	0.606
7	0.936	0.921	0.905	0.887	0.867	0.845	0.822	0.797	0.771	0.744
8	0.972	0.964	0.955	0.944	0.932	0.918	0.903	0.886	0.867	0.847
9	0.989	0.985	0.980	0.975	0.968	0.960	0.951	0.941	0.929	0.916
10	0.996	0.994	0.992	0.990	0.986	0.982	0.977	0.972	0.965	0.957
11	0.999	0.998	0.997	0.996	0.995	0.993	0.990	0.988	0.984	0.980
12	1.000	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.993	0.991
13	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.996
14	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

x	λ = E(X)									
	6.5	**7.0**	**7.5**	**8.0**	**8.5**	**9.0**	**9.5**	**10.0**	**11.0**	**12.0**
0	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.011	0.007	0.005	0.003	0.002	0.001	0.001	0.000	0.000	0.000
2	0.043	0.030	0.020	0.014	0.009	0.006	0.004	0.003	0.001	0.001
3	0.112	0.082	0.059	0.042	0.030	0.021	0.015	0.010	0.005	0.002
4	0.224	0.173	0.132	0.100	0.074	0.055	0.040	0.029	0.015	0.008
5	0.369	0.301	0.241	0.191	0.150	0.116	0.089	0.067	0.038	0.020
6	0.527	0.450	0.378	0.313	0.256	0.207	0.165	0.130	0.079	0.046
7	0.673	0.599	0.525	0.453	0.386	0.324	0.269	0.220	0.143	0.090
8	0.792	0.729	0.662	0.593	0.523	0.456	0.392	0.333	0.232	0.155
9	0.877	0.830	0.776	0.717	0.653	0.587	0.522	0.458	0.341	0.242
10	0.933	0.901	0.862	0.816	0.763	0.706	0.645	0.583	0.460	0.347
11	0.966	0.947	0.921	0.888	0.849	0.803	0.752	0.697	0.579	0.462
12	0.984	0.973	0.957	0.936	0.909	0.876	0.836	0.792	0.689	0.576
13	0.993	0.987	0.978	0.966	0.949	0.926	0.898	0.864	0.781	0.682
14	0.997	0.994	0.990	0.983	0.973	0.959	0.940	0.917	0.854	0.772
15	0.999	0.998	0.995	0.992	0.986	0.978	0.967	0.951	0.907	0.844
16	1.000	0.999	0.998	0.996	0.993	0.989	0.982	0.973	0.944	0.899
17	1.000	1.000	0.999	0.998	0.997	0.995	0.991	0.986	0.968	0.937
18	1.000	1.000	1.000	0.999	0.999	0.998	0.996	0.993	0.982	0.963
19	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.991	0.979
20	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.995	0.988
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.998	0.994
22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.997
23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999