TEMA 1

ANÀLISI DE DADES TRANSVERSALS

Part I

Estadística bàsica

Curs 2020/2021

Virgilio Pérez Giménez – virgilio.perez@uv.es

Vniversitat de València
Facultat d' Economia

Què és l’estadística?

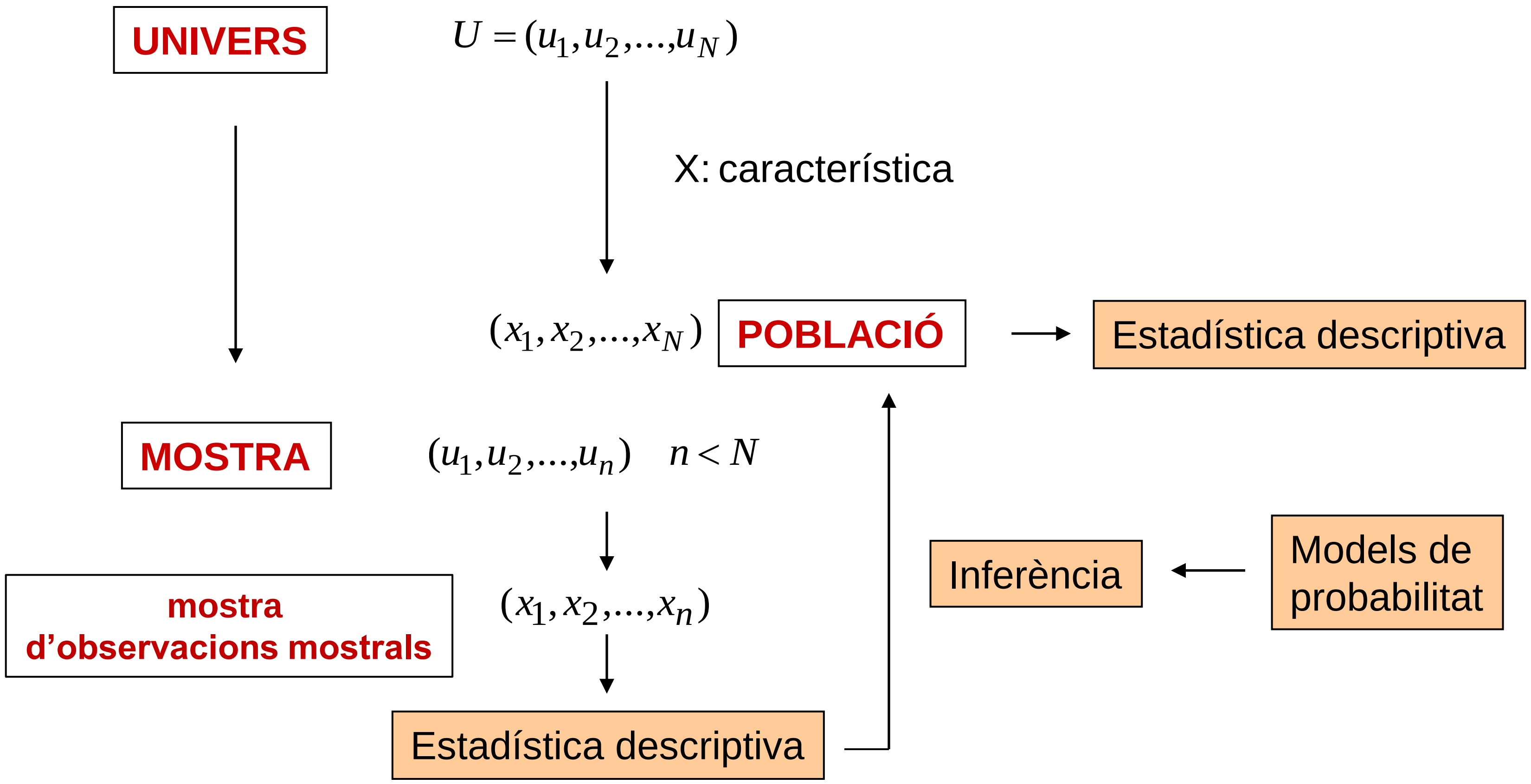
➢ Col·lecció de dades ordenades i sistemàtiques.

Veces matriculado	Registros	Porcentaje
1	50	75.76 %
2	12	18.18 %
5	2	03.03 %
3	1	01.52 %
4	1	01.52 %

➢ Ciència que tracta d’analitzar la regularitat del comportament col·lectiu, amb un doble objectiu:

- Fer descripcions.
- A partir d’aquestes, fer prediccions o estimacions.

Tota investigació científica ha de referir-se a un conjunt de persones o coses, denominat **col·lectiu** (univers):



En quins camps poden ser d'utilitat les tècniques estadístiques per a un economista?

Recursos humans:

Anàlisi dels resultats obtinguts en els tests d'aptituds.

Màrqueting:

Estudis de mercat dirigits al coneixement de la demanda, productes competidors, efectes de campanyes publicitàries i llançament de nous productes.

Producció:

El control de qualitat és un conjunt d'eines estadístiques eficaces per a millorar els processos de producció i reduir-ne els defectes.

Finances:

Les tècniques estadístiques poden ajudar un inversor a elaborar una anàlisi d'inversions per a seleccionar entre diferents productes financers i quantificar el grau d'incertesa de les operacions.

Les variables observades poden ser:

Qualitatives: No es poden mesurar, sinó que els seus caràcters, denominats *atributs*, es descriuen mitjançant paraules (estat civil, professió, nacionalitat…). Poden fer referència a:

- ***ordinals:*** susceptibles d'ordenació (qualificacions; motivació dels empleats; qualitat del servei prestat…)
- ***nominals:*** no susceptibles d'ordenació (sexe, estat civil, veredicte en un judici…)

Quantitatives: Els caràcters de les quals, denominats *variables,* són els que es descriuen mitjançant nombres (salaris, edat, vendes…):

- ***discretes:*** només poden prendre uns determinats valors i no és possible que prenguen un valor comprès entre dos consecutius (nombre de treballadors en una empresa, nombre de peces defectuoses…)
- ***contínues:*** poden prendre qualsevol valor definit per un interval (alçada, pes, salaris, temperatura…)

Els elements del col·lectiu observat poden ser:

- ✓ **Unidimensionals:** s'observa UNA única característica (variable). Per exemple, l'edat dels matriculats en aquest grup d'estadística.
- ✓ **Multidimensionals:** s'observen conjuntament DIVERSES característiques. Per exemple, l'edat i el sexe dels matriculats en aquest grup d'estadística.

- ✓ **Atemporals o de tall transversal:** estudi en un moment determinat del temps. Per exemple, els ingressos d'un grup d'empreses durant l'any passat.
- ✓ **Temporals o cronològiques:** evolució de la característica al llarg del temps. Per exemple, els ingressos d'una empresa durant els últims vint anys.

La informació s’ordena en una **DISTRIBUCIÓ DE FREQÜÈNCIES**

Qualificacions grup A

	n_i	f_i	N_i	F_i
Suspens	10	10/30 = 0.33	10	0.33
Aprovat	8	8/30 = 0.27	10+8 = 18	18/30 = 0.60
Notable	7	7/30 = 0.24	18+7 =25	25/30 = 0.84
Excel·lent	4	4/30 = 0.13	25+4 = 29	29/30 = 0.97
Matrícula d’honor	1	1/30 = 0.03	29+1 = **30**	30/30 = **1**
total	**N=30**	**1**		

Freqüència absoluta ordinària $\mathbf{n_i}$: nombre de vegades que es repeteix el valor de la variable o el nombre d’elements que pertanyen a una categoria.

Freqüència absoluta acumulada $\mathbf{N_i}$

$$\boldsymbol{N_i} = n_1 + n_2 + \cdots + n_i = \sum_{n=1}^{i} n_i$$

Freqüència relativa ordinària $\mathbf{f_i}$: proporció de vegades que es repeteix cada valor de la variable

$$\boldsymbol{f_i} = \frac{n_i}{N}$$

Freqüència relativa acumulada $\mathbf{F_i}$

$$\boldsymbol{F_i} = f_1 + f_2 + \cdots + f_i = \sum_{n=1}^{i} f_i = \frac{N_i}{N}$$

Qualificacions grup A

	n_i	f_i	N_i	F_i
Suspens	5	5/30 = 0.17	5	0.17
Aprovat	9	9/30 = 0.30	5+9 = 14	14/30 = 0.47
Notable	10	10/30 = 0.33	14+10 =24	24/30 = 0.8
Excel·lent	5	5/30 = 0.17	24+5 = 29	29/30 = 0.97
Matrícula d'honor	1	1/30 = 0.03	29+1 = **30**	30/30 = **1**
total	**N=30**	**1**		

Qualificacions grup B

	n_i	f_i	N_i	F_i
Suspens	15	15/52 = 0.29	15	15/52 = 0.29
Aprovat	11	11/52 = 0.21	15+11 = 26	26/52 = 0.50
Notable	15	15/52 = 0.29	26+15 =41	41/52 = 0.79
Excel·lent	10	10/52 = 0.19	41+10 = 51	51/52 = 0.98
Matrícula d'honor	1	1/52 = 0.02	51+1 = **52**	52/52 = **1**
total	**N=52**	**1**		

Qualificacions del grup A: **variable contínua**

	n_i	f_i	N_i	F_i
[0-5[	5	5/30 = 0.17	5	0.17
[5-7[	9	9/30 = 0.30	5+9 = 14	14/30 = 0.47
[7-9[	10	10/30 = 0.33	14+10 =24	24/30 = 0.8
[9-10[	5	5/30 = 0.17	24+5 = 29	29/30 = 0.97
10	1	1/30 = 0.03	29+1 = **30**	30/30 = **1**
total	**N=30**	**1**		

Expedients del grup A: **variable discreta**

	n_i	f_i	N_i	F_i
0	5	5/30 = 0.17	5	0.17
1	9	9/30 = 0.30	5+9 = 14	14/30 = 0.47
2	10	10/30 = 0.33	14+10 =24	24/30 = 0.8
3	5	5/30 = 0.17	24+5 = 29	29/30 = 0.97
4	1	1/30 = 0.03	29+1 = **30**	30/30 = **1**
total	**N=30**	**1**		

Representació gràfica de les variables

PICTOGRAMA

S’associa a cada categoria de la variable un dibuix que hi està relacionat, la grandària del qual és proporcional a la freqüència.

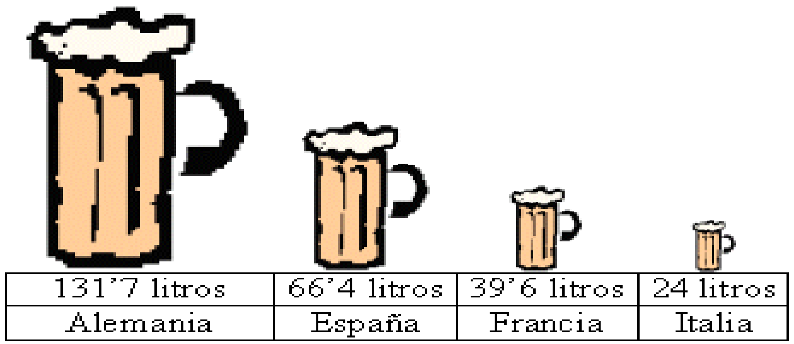
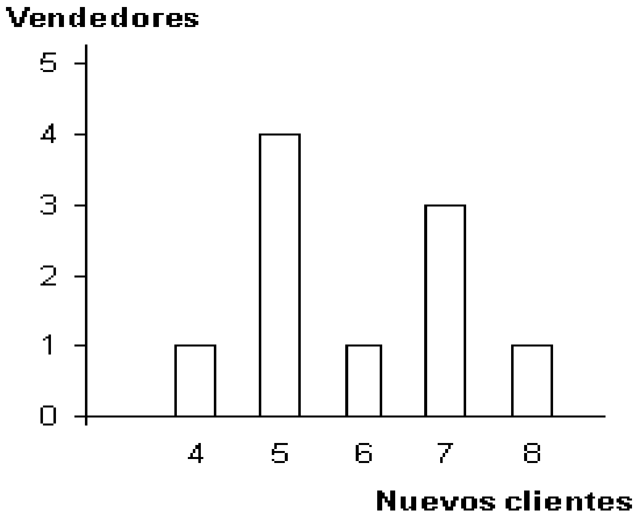


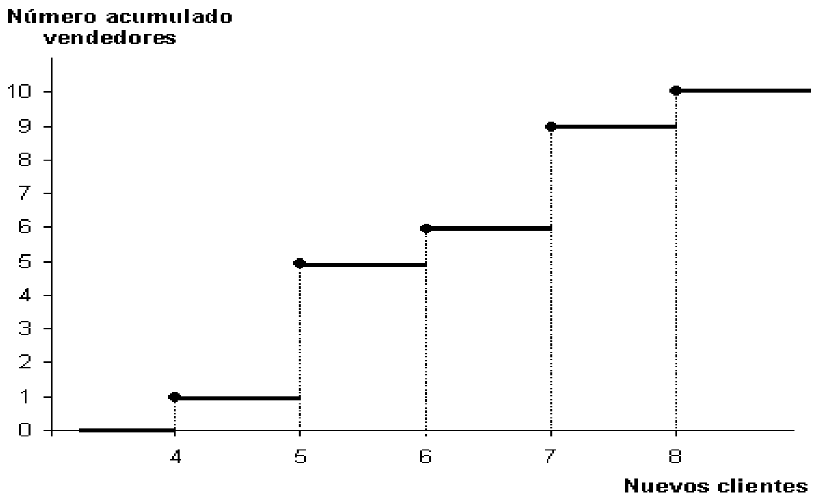
DIAGRAMA DE BARRES



Se situen a l’eix d’abscisses (eix X) els valors de la variable i a l’eix d’ordenades (eix Y) les freqüències ordinàries (absolutes o relatives). Així, s’associa una barra a cada x_i, l’altura de la qual és proporcional a la freqüència que li correspon.

DIAGRAMA EN ESCALA

En l’eix d’ordenades se situen les freqüències acumulades (N_i o F_i) i les barres s’uneixen mitjançant segments paral·lels a l’eix d’abscisses. Es denomina diagrama en escala per la forma en escala que adopta, on l’altura de cada graó és la freqüència ordinària (absoluta o relativa) associada al corresponent valor x_i.



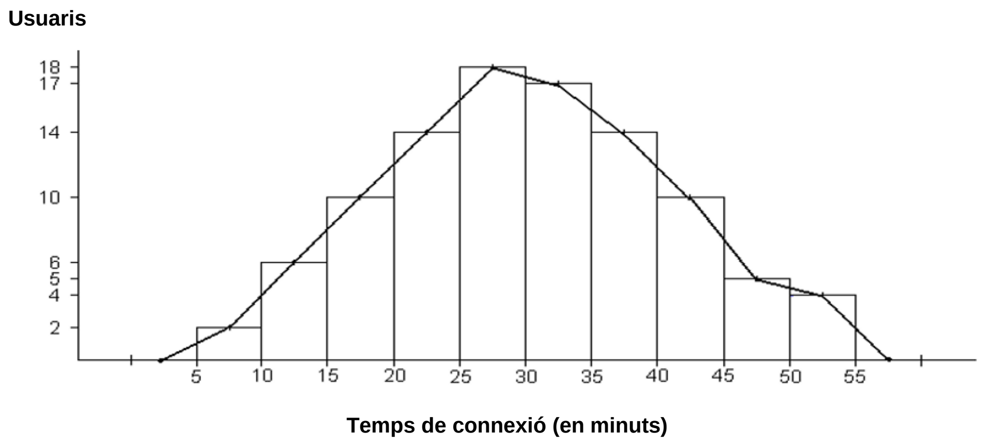
HISTOGRAMA:

A l'eix d'abscisses se situen els intervals (L_{i-1}, L_i), sobre cadascun dels quals es construeix un rectangle de base igual a l'amplitud de l'interval c_i i d'altura h_i :

$$h_i = \frac{n_i}{c_i} = \frac{f_i}{c_i}$$

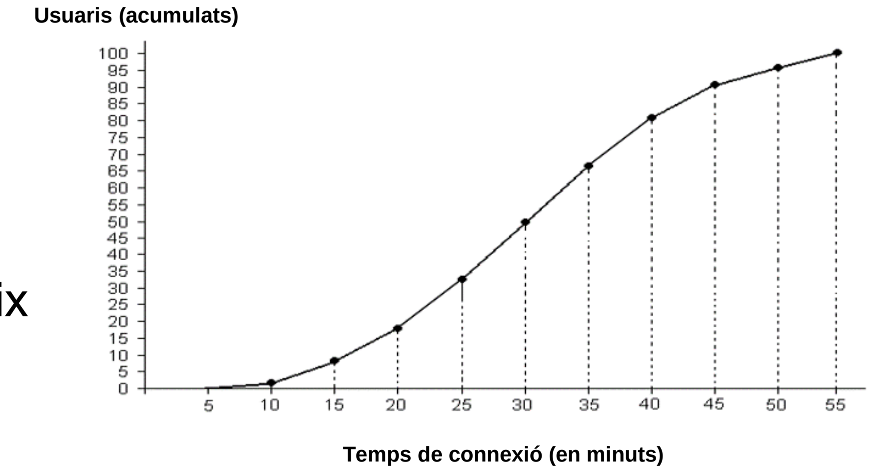
POLÍGON DE FREQÜÈNCIES

S'obté en unir, mitjançant línies rectes, els punts mitjans dels costats superiors dels rectangles de l'histograma:



POLÍGON ACUMULATIU DE FREQÜÈNCIES

A l'eix d'abscisses se situen els intervals (L_{i-1}, L_i) i a l'eix d'ordenades, les freqüències, absolutes o relatives, acumulades.



TEMA 1

ANÀLISI DE DADES TRANSVERSALS

Part II

Estadística bàsica

Curs 2020/2021

VNIVERSITAT DE VALÈNCIA
Facultat d' Economia

MESURES DE POSICIÓ: tenen per objectiu situar la distribució

- **mitjana aritmètica:** representa el valor mitjà de la distribució.
- **mediana:** és el valor que ocupa la posició central.
- **quantils:** divideixen la distribució en parts iguals.
- **moda:** el valor que més vegades es repeteix.

MESURES DE DISPERSIÓ: mesuren la variabilitat de les dades.

- **variància**
- **desviació típica**
- **coeficient de variació**

MESURES DE FORMA O PERFIL

- **simetria / asimetria**
- **apuntament / aplanament**

Mitjana aritmètica

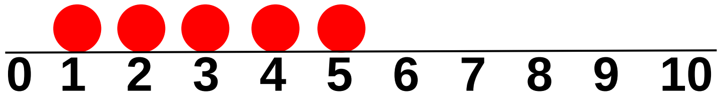
$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \overline{X} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i \qquad \overline{X} = \sum_{i=1}^{n} x_i f_i$$

Avantatges:

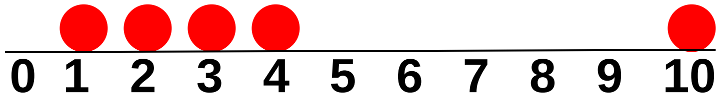
- Aprofita tota la informació disponible.
- És fàcil de calcular.
- Pren un valor únic.

Inconvenient:

Si la distribució té valors extrems, en utilitzar tota la informació, pot **distorsionar el resultat** i ser poc representatiu de la realitat.



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Exemple: una empresa disposa de 10 empleats. Durant el mes passat, el nombre de dies que cada empleat va estar malalt va ser:

3 , 0 , 5 , 6 , 1 , 0 , 11 , 6 , 0 , 4

Determina el nombre de dies que, per terme mitjà, va estar malalt un empleat.

Col·lectiu: els 10 empleats

Variable X: nombre de dies malalt el mes passat

$$\bar{x} = \frac{\sum_{i=1}^{N=10} x_i}{N} = \frac{3+0+5+6+1+0+11+6+0+4}{10} = \frac{36}{10} = 3,6$$

Quants valors diferents pren la variable X?

Es repeteixen valors?

$$\bar{x} = \frac{\sum_{i=1}^{I=7} x_i \cdot n_i}{N} = \frac{36}{10} = 3,6 \text{ dies per empleat}$$

x_i	n_i	$x_i n_i$
0	3	0
1	1	1
3	1	3
4	1	4
5	1	5
6	2	12
11	1	11
	Σ=10	Σ=36

Mitjana de mitjanes de r col·lectius o mitjana ponderada

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + \cdots N_r\bar{x}_r}{N_1 + N_2 + \cdots N_r} = \frac{\sum_{i=1}^{N_1} x_{1i} + \sum_{i=1}^{N_2} x_{2i} + \cdots \sum_{i=1}^{N_r} x_{ri}}{N_1 + N_2 + \cdots N_r}$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \rightarrow \quad N\bar{X} = \sum_{i=1}^{N} X_i$$

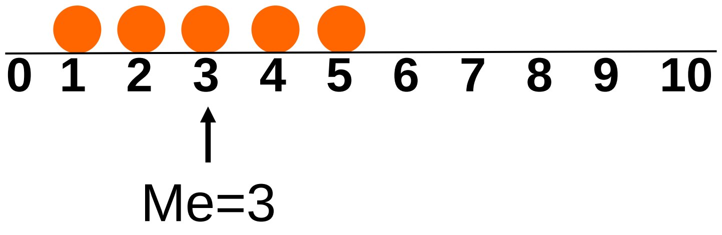
grups	nota mitjana	presentats
A	5.5	70
B	4	40
C	6	65

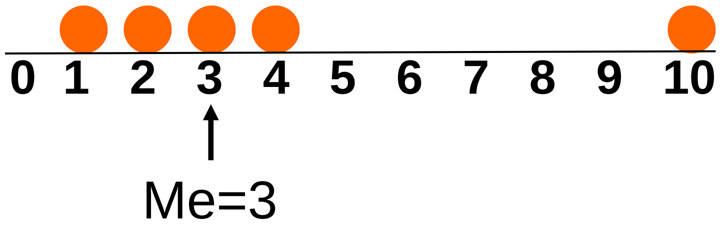
$$\bar{x} = \frac{70 \cdot 5.5 + 40 \cdot 4 + 65 \cdot 6}{70 + 40 + 65} = \frac{385 + 160 + 390}{175} = 5.34$$

Mediana, quantils i moda

➤ **Ordenats els valors**, la **mediana** es correspon amb el valor que ocupa el valor central.

➤ Utilitza menys informació que la mitjana aritmètica, perquè sols té en compte l'**ordenació**, però no la seua magnitud, de manera que no es veu alterada per valors extrems:





La mitjana és la mesura de tendència central que més s'usa, però quan hi ha **valors extrems,** la **mediana serà més convenient** com a mesura de posició.

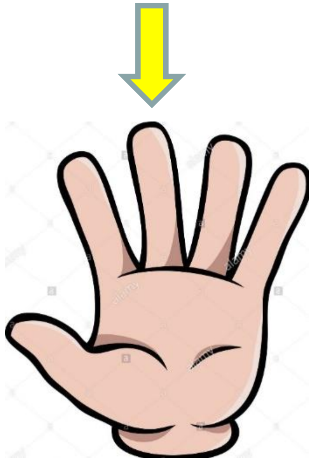
➤ Els **quantils** divideixen la mostra en parts iguals:

- en dues parts: mediana
- en quatre partes: quartils
- en deu parts: decils
- en cent parts: percentils

➤ La **moda** és el valor de la variable que més vegades es repeteix, i pot haver-hi més d'una moda.

Notes: 1 2 2.5 4 5 5 6 8 9 9 10

$\boxed{Me = 5}$



$N = 11 \longrightarrow$ Hi ha una posició central $\longrightarrow \dfrac{N+1}{2} = \dfrac{11+1}{2} = 6$

(nombre imparell)

Notes: 1 2 2.5 4 5 5 6 8 9 9 10 10

$$\boxed{Me = \frac{5+6}{2} = 5.5}$$

$N = 12 \longrightarrow$ No hi ha posició central $\longrightarrow \dfrac{N}{2} = \dfrac{12}{2} = 6$ i el valor següent

(nombre parell)

Càlcul dels QUANTILS quan la freqüència no és única

qualificacions	n_i	N_i
0	1	1
1	1	2
2	1	3
3	1	4
4.5	1	5
5	3	8
5.5	4	12
6	2	14
7	4	18
7.5	2	20
8	2	22
8.5	2	24
9	4	28
9.5	1	29
10	1	30

$\rightarrow$ $N/4 = 7.5$ $\rightarrow$ $q_1 = 5$

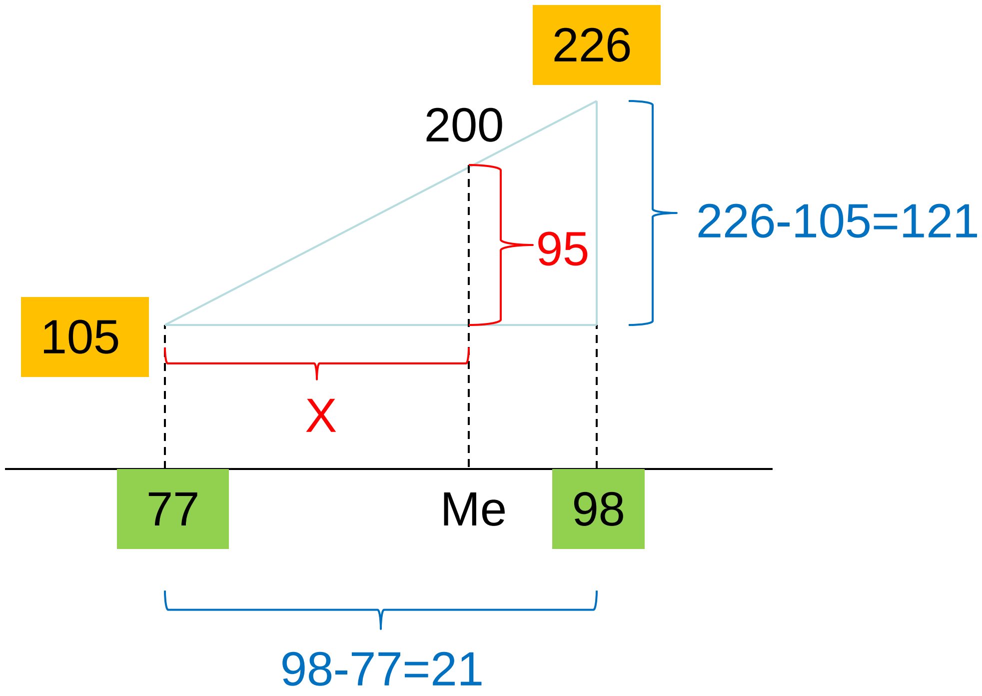
$\rightarrow$ $N/2 = 15$ y $N/2 + 1 = 16$ $\rightarrow$ $Me = 7$

$\rightarrow$ $3N/4 = 22.5$ $\rightarrow$ $q_3 = 8.5$

$\rightarrow$ $aN/100 = 28$ $\rightarrow$ $a \approx 93$ $\rightarrow$ $p_{93} \approx 9$

l_{i-1}	l_i	x_i	n_i	N_i
35	56	45,5	15	15
56	77	66,5	90	105
77	98	87,5	121	226
98	119	108,5	65	291
119	140	129,5	22	313
140	161	150,5	50	363
161	182	171,5	22	385
182	203	192,5	6	391
203	224	213,5	6	397
224	245	234,5	3	400
			400	

$$\frac{N}{2} = \frac{400}{2} = 200$$



Càlcul de la mediana si hi ha intervals.

De forma anàloga procedim si volem obtenir qualsevol quantil. Començaríem per calcular 35N/100 si volguérem, per exemple, el percentil 35.

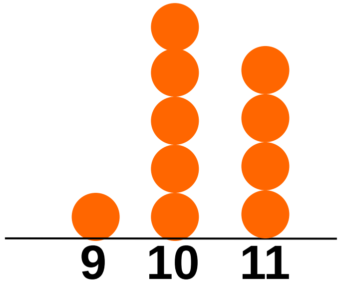
$$\frac{121}{21} = \frac{95}{X} \rightarrow 121X = 21 \cdot 95 \rightarrow X = \frac{21 \cdot 95}{121} = 16{,}49$$

$$\boxed{\mathrm{Me} = 77 + 16{,}49 = 93{,}49}$$

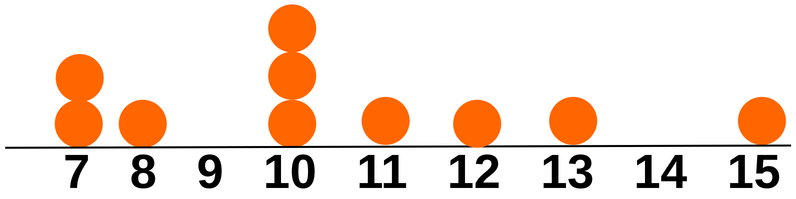
- Una mitjana resumeix tots els valors observats en un de sol, que els representa.
- La utilitat de la mitjana depèn, per tant, del seu poder de representació:
  - Si els valors observats de la variable estan molt concentrats al voltant de la mitjana, aquesta serà molt representativa.
  - Si els valors estan molt dispersos amb relació a la mitjana, aquesta serà poc representativa.

Exemple: ets el/la responsable de compres d'una empresa i saps que dos dels teus proveïdors tarden, per terme mitjà, deu dies a servir la comanda, però amb comportaments diferents. Quin dels dos prefereixes? Quin és més consistent?

proveïdor A



proveïdor B



- Com a conseqüència, una mitjana ha d'anar acompanyada d'una mesura de dispersió, variabilitat o oscil·lació de les observacions al voltant d'ella, que ens indicarà la seua representativitat.
- Les mesures de dispersió absolutes són la variància i la desviació típica.
- La mesura de dispersió relativa és el coeficient de variació.

- La solució que pareix simple quan parlem de dispersió absoluta és fer la mitjana de les desviacions de la variable respecte de la mitjana, però comprovem que sempre és zero:

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})n_i=\frac{1}{N}\sum_{i=1}^{N}x_i n_i-\frac{1}{N}\bar{x}\sum_{i=1}^{N}n_i=\bar{x}-\frac{1}{N}\bar{x}N=\bar{x}-\bar{x}=0$$

Aquesta és la propietat fonamental de la mitjana i és deguda al fet que les diferències positives es compensen amb les negatives.

Exemple: s'ha pres una mostra de persones que fumen, se'ls ha preguntat per l'edat en què varen començar a adquirir aquest hàbit. Les edats han sigut:

14, 16, 15, 18, 17, 19, 19, 18

Verifica que la suma de les desviacions respecte de la mitjana és zero.

- Per evitar aquest problema, una solució és elevar al quadrat aquestes diferències i així evitar les compensacions. El resultat és el que anomenem **variància**:

$$S_x^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 n_i=\frac{1}{N}\sum_{i=1}^{N}x_i^2 n_i-\bar{x}^2$$

L'última igualtat és la seua **expressió operativa**.

$$S_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 n_i = \frac{1}{N}\sum_{i=1}^{N} x_i^2 n_i - \overline{x}^2$$

Un **valor de variància alt** significa molta dispersió i, per tant, **poca representativitat de la mitjana**.

Un **valor de variància baix** significa poca dispersió i, per tant, **alta representativitat de la mitjana**.

x_i	n_i	$x_i n_i$	x_i-x	$(x_i$-x$)^2$	$(x_i$-x$)^2 n_i$	$x_i^2 n_i$
1	2	2	-2	4	8	2
2	5	10	-1	1	5	20
3	10	30	0	0	0	90
4	5	20	1	1	5	80
5	2	10	2	4	8	50
	N=24	Σ=72			Σ= 26	Σ= 242

$$\overline{x} = \frac{\sum_{i=1}^{I=5} x_i \cdot n_i}{N} = \frac{72}{24} = 3$$

$$S_X^2 = \frac{\sum_{i=1}^{I=5} (x_i - \overline{x})^2 \cdot n_i}{N} = \frac{26}{24} = 1,08\widehat{3}$$

$$S_X^2 = \frac{\sum_{i=1}^{I=5} x_i^2 \cdot n_i}{N} - \overline{x}^2 = \frac{242}{24} - 3^2 = 1,08\widehat{3}$$

- La **variància** té dos problemes fonamentals per a la seua interpretació:
  - La seua gran magnitud aparent, pel fet de ser **unitats al quadrat**.
  - La falta d'acotació superior, ja que pot tomar qualsevol valor superior o igual a zero.

- La primera dificultat se soluciona calculant l'arrel quadrada de la variància, i dona lloc a la **desviació típica.**

- La segona dificultat no té solució i la interpretació dependrà de l'experiència i el sentit comú.

- **Per a comparar dispersions** de dues o més distribucions amb diferents mitjanes o diferents unitats de mesura s'utilitza **el coeficient de variació de *Pearson*,** definit com:

$$CV(x) = g_0(x) = \frac{S_x}{\overline{x}}$$

estadístic adimensional

Com major és, major dispersió té la seua distribució.

Exercici

- Siga U un col·lectiu definit pels matriculats en el grup A d'un institut, la grandària del qual és N=47,
- Siga X la variable "notes en Socials" observada sobre el col·lectiu anterior,

Feu una descripció de la població "notes de socials del grup A" d'acord amb els estadístics següents :

$\bar{x} = 5.4$ Me=2 Mo=1 $s_x = 4.06$

- ✓ Si el valor que més vegades es repeteix és l’1, i la meitat dels estudiants tenen una nota inferior a 2, la conclusió és que hi haurà notes molt altes perquè la mitjana puga ser 5,4.

- ✓ De fet, una desviació típica de 4,06 punts indica que podria haver-hi qualificacions per damunt de 9 punts i pròximes a 1.

La distribució a partir de la qual s’han obtingut els estadístics és:

notes	presentats
1	13
2	11
9	11
10	12

TEMA 1

ANÀLISI DE DADES TRANSVERSALS

Part III

Estadística bàsica

Curs 2020/2021

Virgilio Pérez Giménez – virgilio.perez@uv.es

Vniversitat de València
Facultat d' Economia

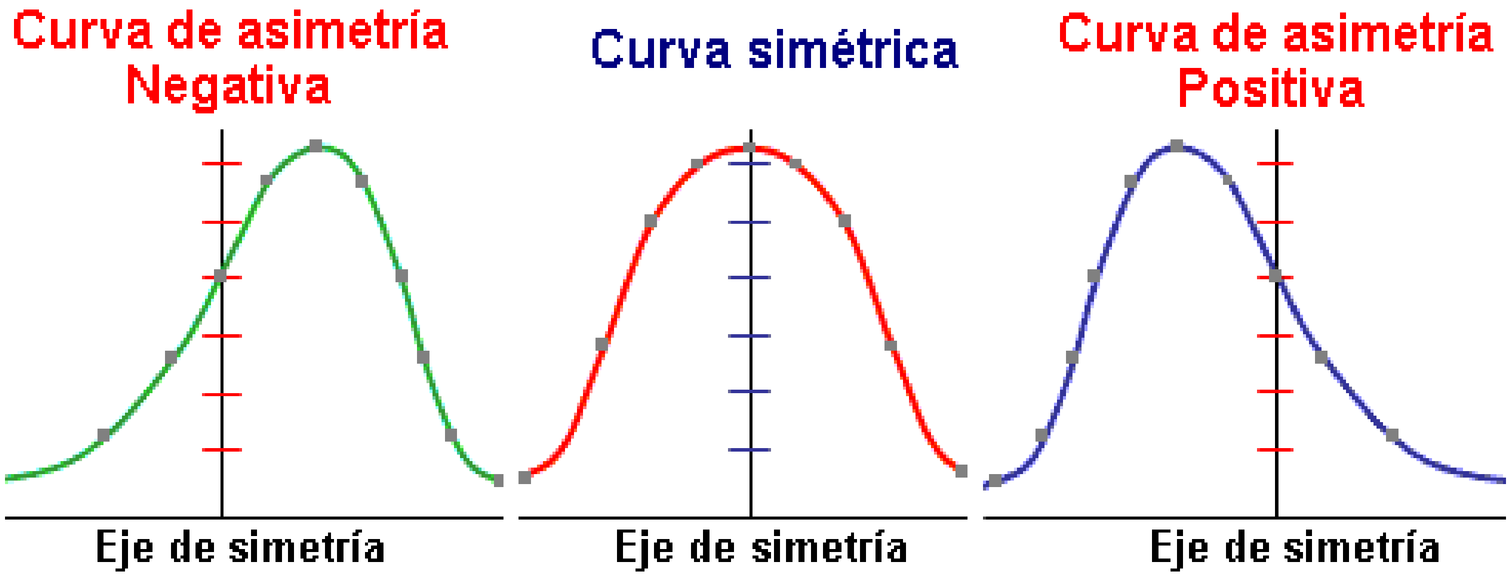
Mesures de forma

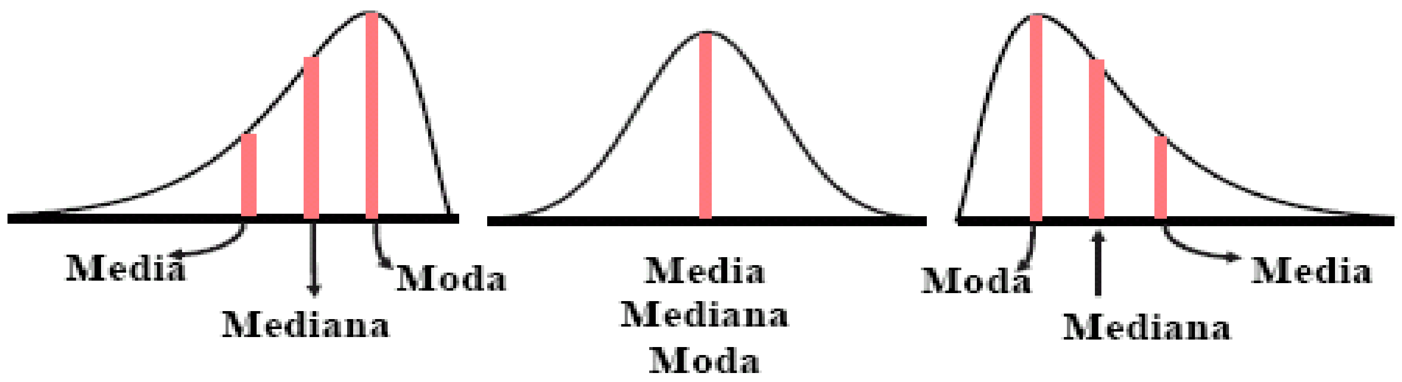
- Aquestes mesures descriuen la manera com les dades tendeixen a reunir-se d'acord amb la freqüència en què es troben en la distribució.

- La seua utilitat rau en la possibilitat d'identificar les característiques de la distribució sense necessitat de generar el gràfic. Les principals mesures són l'***asimetria*** i la ***curtosi***.

- Hi ha tres estats que indiquen com es distribueixen les dades respecte de l'eix de simetria:

  - **Asimetria positiva o a la dreta:** la major part dels valors es troben per sota de la mitjana (la cua de la **dreta** és més llarga).
  - **Asimetria negativa o a l'esquerra:** la major part dels valors es troben per damunt de la mitjana (la cua de l'**esquerra** és més llarga).
  - **Simetria:** aproximadament la mateixa quantitat a ambdós costats de la mitjana.





Asimétrica hacia la izquierda

$g_1(X) < 0$

Simétrica

$g_1(X) = 0$

Asimétrica hacia la derecha

$g_1(X) > 0$

$$g_1(X) = \frac{\dfrac{\sum_{i=1}^{I} (x_i - \bar{x})^3 \cdot n_i}{N}}{S_X^3}$$

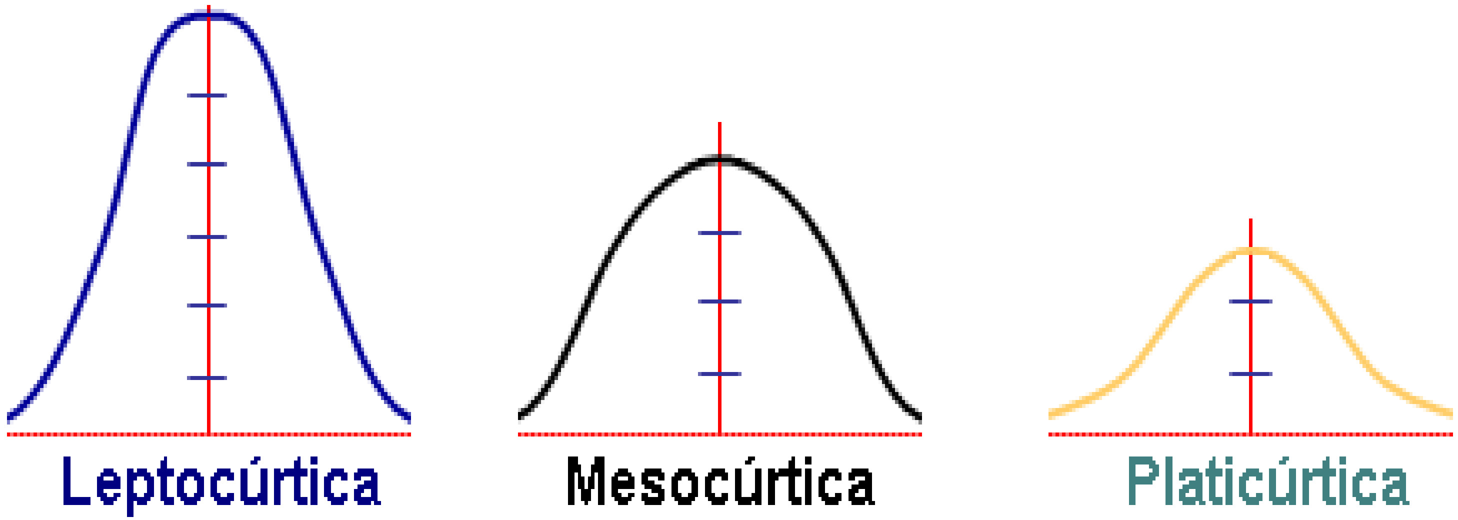
Coeficient d'asimetria de Fisher

- La **curtosi** determina el grau de concentració que presenten els valors a la regió central de la distribució.

- Hi ha tres estats d’apuntament/aplanament:
  - **Leptocúrtica:** gran concentració de valors.
  - **Mesocúrtica:** concentració normal de valors.
  - **Platicúrtica:** baixa concentració de valors.

$$g_2(X) = \frac{\dfrac{\sum_{i=1}^{I}(x_i - \overline{x})^4 . n_i}{N}}{S_X^4} - 3$$

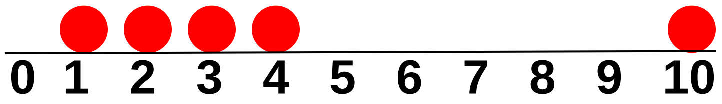
Coeficient de curtosi



$g_2(X) > 0$ $\qquad$ $g_2(X) = 0$ $\qquad$ $g_2(X) < 0$

Valors extrems

- Valors que disten de la mitjana i que poden distorsionar els resultats descriptius.
- Poden ser deguts a errors de mesurament o simplement valors que disten del comportament habitual en la distribució.
- Una vegada detectats, és important diferenciar si són producte de l'error o no.
- Una forma de detectar-los és comparar **el rang** de les dades (diferència entre el valor màxim i el mínim) i **el rang interquartílic** (diferència entre el tercer i el primer quartil). Si el rang és gran, hi haurà dispersió, però si a més el rang interquartílic és petit, hi haurà valors extrems.

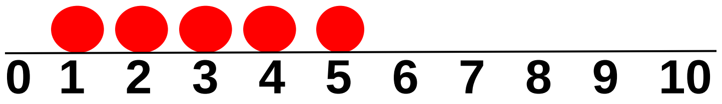


rang=10-1=**9**

N/4=5/4=1.25 $\rightarrow$ $q_1 \approx 2$
3N/4=3.75 $\rightarrow$ $q_3 \approx 4$
RI $\approx 4 - 2 = \mathbf{2}$

Xi	Ni
1	1
2	2
3	3
4	4
10	5

Yi	Ni
1	1
2	2
3	3
4	4
5	5



rang=5-1=**4**

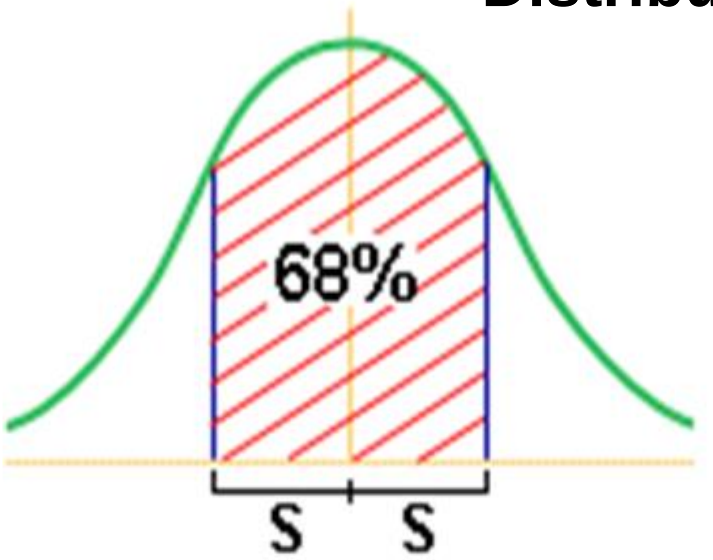
N/4=5/4=1.25 $\rightarrow$ $q_1 \approx 2$
3N/4=3.75 $\rightarrow$ $q_3 \approx 4$
RI $\approx 4 - 2 = \mathbf{2}$

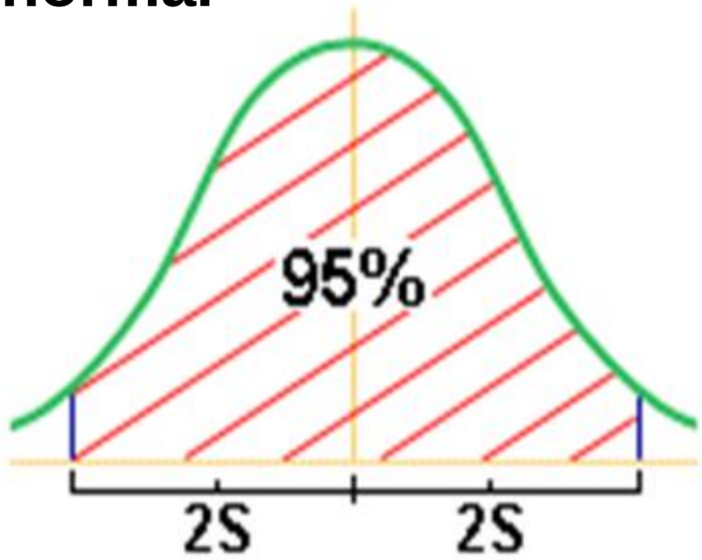
- Una distribució simètrica amb forma de campana i mesocúrtica és una **distribució normal**.

- El principal avantatge de la distribució normal es basa en el supòsit que el 95% dels valors es troben dins d'una distància de dues desviacions típiques de la mitjana aritmètica.

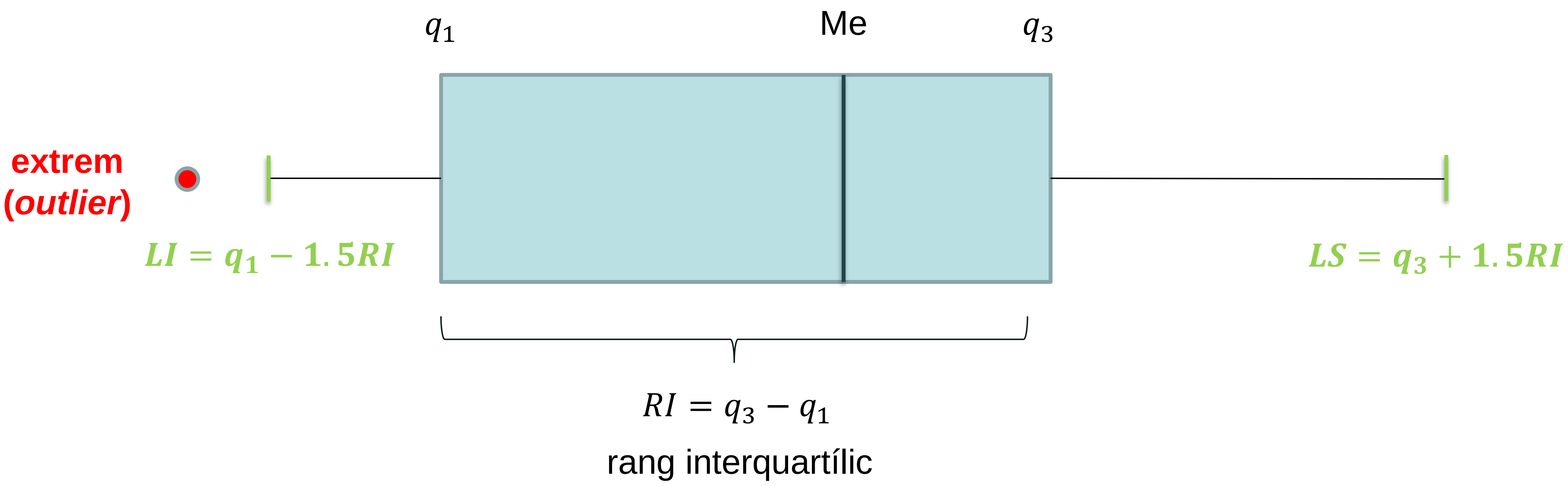
- Per això, en aquelles distribucions semblants a la normal, es considera com a **valor extrem** aquell que dista de la mitjana dues vegades la desviació típica.

Distribució normal





BOXPLOT O DIAGRAMA DE CAIXA I BIGOTIS



Aquest gràfic ens permet representar mesures de posició, i així tenir una visió de la dispersió a través del rang interquartílic, major rang, major dispersió; també ens permet representar valors extrems.

TRANSFORMACIONS LINEALS

- Hem de considerar tres escenaris en què, partint d'una variable "X", podem obtenir, per **transformació lineal**, una nova variable "Y".

$$y_i = a + bx_i$$

- Anomenem la variable "Y" **variable dependent** (o explicada), ja que depèn del que ocórrega a la variable "X".

- Anomenem la variable "X" **variable independent** o **variable explicativa**, ja que és la que explica el comportament de la variable "Y".

- **L'objectiu d'una transformació lineal és determinar les mesures de posició i de dispersió de la variable "Y" sense necessitat de recórrer calcular-la (distribució de freqüències).**

Si conec la mitjana, la mediana i la variància de la variable X, i hi ha relació lineal... → Puc calcular la mitjana, la mediana i la variància de la variable Y? → **SÍ**

TRANSFORMACIONS LINEALS

$$y_i = a + bx_i$$

$$\bar{y} = \frac{1}{N}\sum_i y_i = \frac{1}{N}\sum_i (a + bx_i) = \frac{1}{N}\sum_i a + \frac{1}{N}b\sum_i x_i = \frac{1}{N}Na + b\bar{x} = \boldsymbol{a + b\bar{x}}$$

Això mateix passa amb la mediana, la moda i qualsevol quantil.

$$\boldsymbol{S_y^2} = \frac{1}{N}\sum_i (y_i - \bar{y})^2 = \frac{1}{N}\sum_i (a + bx_i - a - b\bar{x})^2 = b^2\frac{1}{N}\sum_i (x_i - \bar{x})^2 = \boldsymbol{b^2 S_x^2}$$

TRANSFORMACIONS LINEALS

$$y_i = a + bx_i$$

Escenari 1

Per definició

Exemple:

- La variable “X” representa el nombre d’hores treballades per empleat.
- “b” és el preu per hora treballada.
- “a” és el cost fix per desplaçament.
- Això dona lloc a “Y” → facturació per empleat.

$$y_i = a + bx_i$$

Escenari 2

Per reducció d'escala

A fi de fer més manejable la variable, se'n redueix la magnitud. Per fer-ho, cal restar a la variable una quantitat i el resultat dividir-lo entre una altra quantitat:

$$y_i = \frac{x_i - a}{b} = \frac{1}{b}x_i - \frac{a}{b}$$

$$y_i = a + bx_i$$

Escenari 3 (cas particular de l'escenari 2)

Per tipificació

S'efectua per obtenir valors amb una mitjana de 0 i una desviació típica d'1, amb l'objectiu de poder comparar la posició relativa de dos valors de dues distribucions diferents. És el mateix obtenir una nota de 5 punts en "Incorporació a la universitat" que un 5 en "Matemàtiques I"? Lògicament no, ja que el comportament d'ambdues distribucions de notes és molt distint. Per poder comparar-les, cal **tipificar**.

$$y_i = \frac{x_i - \bar{x}}{S_x} = \frac{1}{S_x} x_i - \frac{\bar{x}}{S_x}$$

$$\bar{y} = \frac{1}{S_x}\bar{x} - \frac{\bar{x}}{S_x} = 0$$

$$S_y^2 = \left(\frac{1}{S_x}\right)^2 S_x^2 = 1$$

Exemple

Variable X: notes d'Incorporació

$$\bar{X} = 8{,}5$$
$$S_X = 0{,}5$$

Variable Y: notes de Matemàtiques I

$$\bar{Y} = 4{,}5$$
$$S_Y = 2{,}5$$

Té més valor un 5 en Incorporació o un 5 en Matemàtiques I?

$$z_i = \frac{5 - 8{,}5}{0{,}5} = -7$$

$$z_i = \frac{5 - 4{,}5}{2{,}5} = 0{,}2$$

0,2 > -7

Millor nota en termes relatius: 5 en Matemàtiques I

TEMA 1

ANÀLISI DE DADES TRANSVERSALS

Part IV

Estadística bàsica

Curs 2020/2021

Virgilio Pérez Giménez – virgilio.perez@uv.es

VNIVERSITAT DE VALÈNCIA
Facultat d' Economia

ÍNDEX DE GINI I CORBA DE LORENZ

- La **concentració** tracta de mesurar el major o menor **grau d'igualtat** en el repartiment del **valor total** de la variable entre els elements del col·lectiu.

- Per tant, només té interès calcular la concentració en aquelles variables en què el seu **total** té algun significat (renda, salaris, sectors econòmics, concentració humana). En general, variables de tipus socioeconòmic.

- Hi ha dues situacions extremes en la concentració:

$X_1 = X_2 = \cdots = X_n \longrightarrow$ màxima uniformitat
mínima concentració

$X_1 = \cdots = X_{n-1} = 0 \;\; y \;\; X_n \neq 0 \longrightarrow$ mínima uniformitat
màxima concentració

➢ Dos estadístics “p_i” y “q_i” ens serviran per determinar la **concentració** existent. Aquests es calculen de la manera següent :

x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
x_1	n_1	$x_1 n_1$	N_1	$x_1 n_1 = u_1$	N_1/N	u_1/u_r
x_2	n_2	$x_2 n_2$	N_2	$x_1 n_1 + x_2 n_2 = u_2$	N_2/N	u_2/u_r
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_r	$x_r n_r$	$N_r = N$	$\sum_{i=1}^{r} x_i n_i = u_r$	1	1

valors ordenats de menor a major

➢ **Exemple**: En una empresa treballen 40 empleats, agrupats en 4 categories professionals (cadascuna amb un sou diferent):

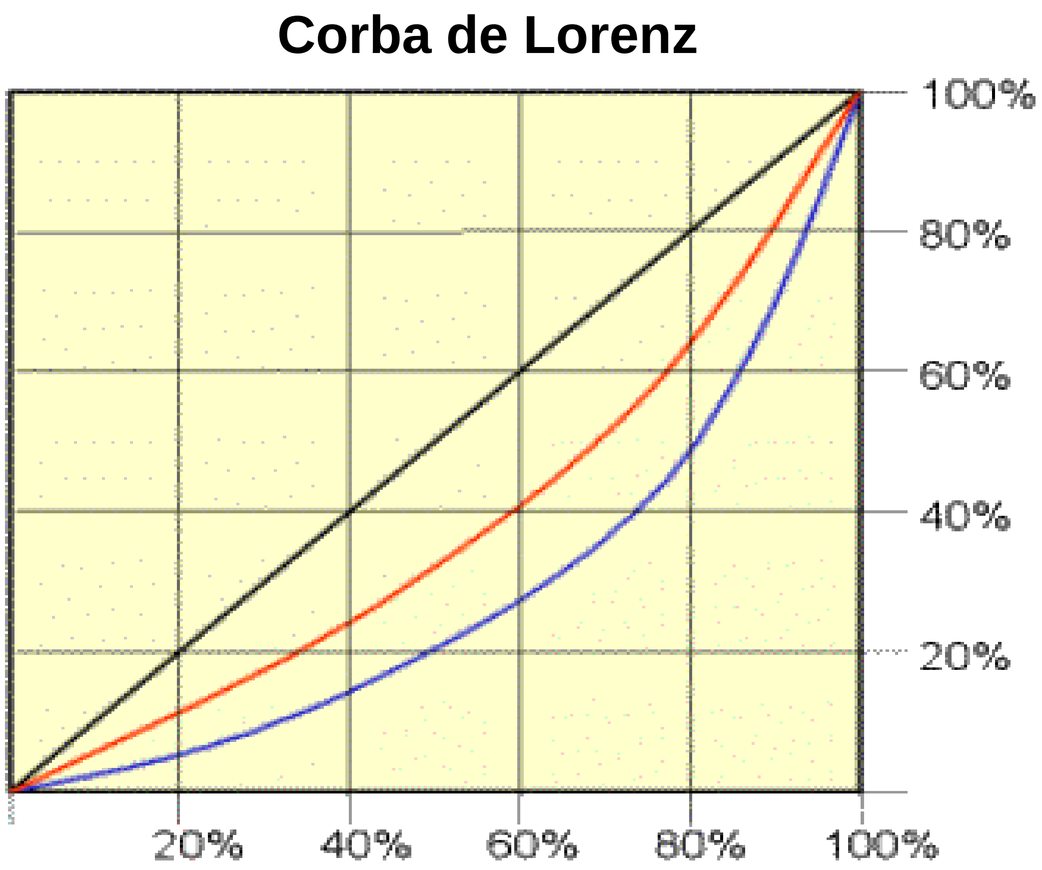
x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
600	25	15.000	25	15.000	0,625	0,517
800	10	8.000	35	23.000	0,875	0,793
1.000	4	4.000	39	27.000	0,975	0,931
2.000	1	2.000	40	29.000	1	1
	40	29.000				

$$\boxed{u_i = \sum x_i \cdot n_i}$$

$$\boxed{p_i = \frac{N_i}{N}}$$

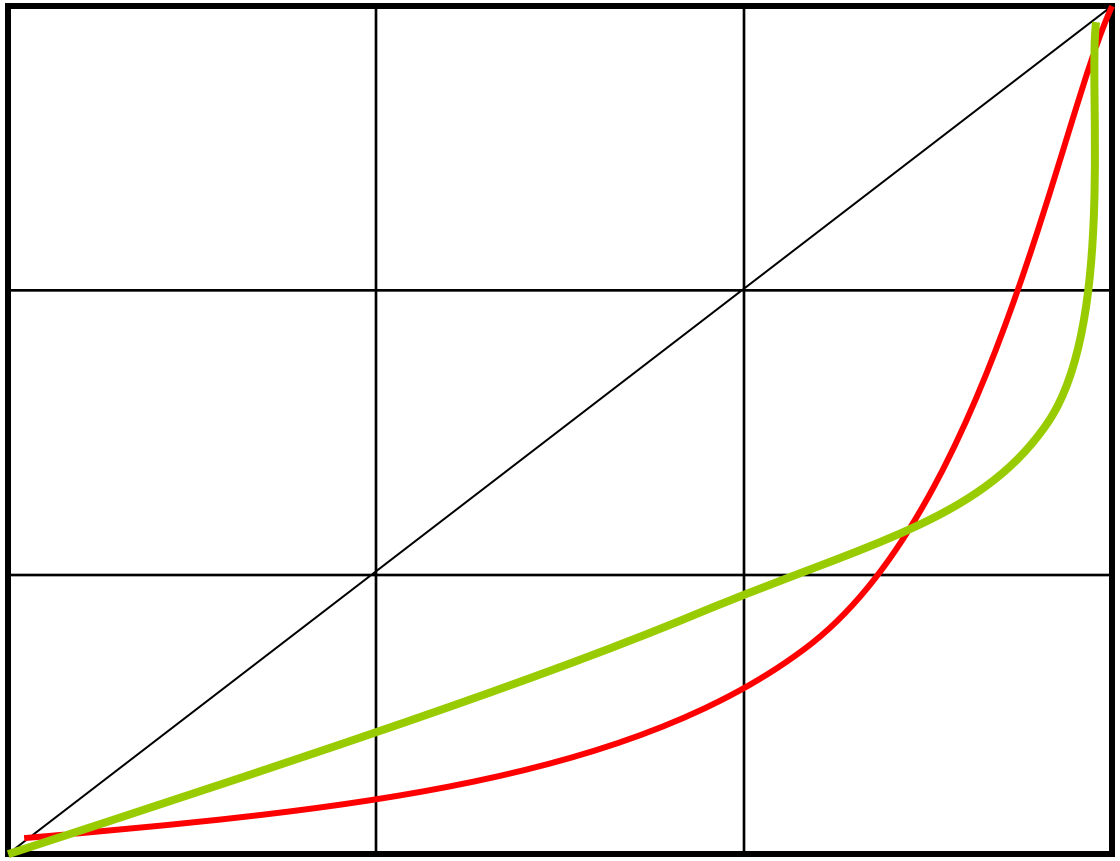
$$\boxed{q = \frac{u_i}{u_r}}$$

➢ Aquests dos estadístics poden representar-se gràficament:



➢ Si la corba està a prop de la bisectriu, la distribució és bastant uniforme. Si se n'allunya, la distribució és concentrada.

- El problema és que la corba no sempre permet comparar entre dos comportaments:



- La solució és obtenir una mesura quantitativa de la concentració, anomenada **ÍNDEX DE GINI:**

$$IG = \frac{\sum_{i=1}^{r-1}(p_i - q_i)}{\sum_{i=1}^{r-1} p_i} \quad , \quad 0 \le IG \le 1$$

- IG = 0, $p_i = q_i$ $\rightarrow$ mínima concentració i màxima uniformitat

- IG = 1, $q_1 = \ldots = q_{r-1} = 0$ $\rightarrow$ màxima concentració i mínima uniformitat

- Si l'**IG** d'una distribució és inferior a l'índex d'altra distribució, es diu que la concentració en la primera és inferior a la concentració en la segona.

➢ **Exemple**: En una empresa treballen 40 empleats, agrupats en 4 categories professionals (cadascuna amb un sou diferent):

x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
600	25	15.000	25	15.000	0,625	0,517
800	10	8.000	35	23.000	0,875	0,793
1.000	4	4.000	39	27.000	0,975	0,931
2.000	1	2.000	40	29.000	1	1
	40	29.000				

$$IG = \frac{\sum_{i=1}^{r-1}(p_i - q_i)}{\sum_{i=1}^{r-1} p_i} = \frac{(0{,}625 - 0{,}517) + (0{,}875 - 0{,}793) + (0{,}975 - 0{,}931)}{0{,}625 + 0{,}875 + 0{,}975} = 0{,}09$$

Sabent que $0 \leq IG \leq 1$

Un IG=0,09 és relativament baix, és a dir, ***“hi ha una forta equidistribució dels salaris a l’empresa i, per tant, hi ha una baixa concentració d’aquests”.***

Virgilio Pérez Giménez – virgilio.perez@uv.es

TEMA 2

ANÀLISI DE DADES MULTIDIMENSIONALS

Estadística bàsica

Curs 2020/2021

Vniversitat de València
Facultat d' Economia

- En aquest capítol plantegem l'anàlisi de dues variables observades sobre un col·lectiu, que donen lloc a una població bidimensional:

$$(u_1, u_2, \ldots, u_N) \quad \rightarrow \quad [(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)]$$

- L'interès d'observar més d'una variable és **establir relacions entre si**, si n'hi ha, per a poder realitzar prediccions.

- La forma de representar les dades bidimensionals és a través d'una taula, denominada ***taula de doble entrada,*** *o* ***taula de contingència,*** amb variables qualitatives, o una ***taula de correlació,*** amb variables quantitatives.

- A partir d'aquesta taula es pot obtenir la **distribució de freqüències conjuntes**, de **freqüències marginals** i de **freqüències condicionades**.

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_r	$n_{i\bullet}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1r}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2r}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ir}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_s	n_{s1}	n_{s2}	$\cdots$	n_{sj}	$\cdots$	n_{sr}	$n_{s\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet r}$	N

n_{ij} freqüències conjuntes

$n_{i\bullet}$, $n_{\bullet j}$: freqüències marginals

$$n_{i\bullet} = \sum_j n_{ij} \qquad n_{\bullet j} = \sum_i n_{ij} \qquad N = \sum_i \sum_j n_{ij} = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$$

Virgilio Pérez Giménez – virgilio.perez@uv.es

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_r	$n_{i\bullet}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1r}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2r}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ir}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_s	n_{s1}	n_{s2}	$\cdots$	n_{sj}	$\cdots$	n_{sr}	$n_{s\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet r}$	N

Distribució condicionada:

$x/y = y_j$	x_1	x_2	$\cdots$	x_i	$\cdots$	x_s	
$n_{i/j}$	n_{1j}	n_{2j}	$\cdots$	n_{ij}	$\cdots$	n_{sj}	$n_{\bullet j}$
$f_{i/j}$	$n_{1j}/n_{\bullet j}$	$n_{1j}/n_{\bullet j}$	$\cdots$	$n_{ij}/n_{\bullet j}$	$\cdots$	$n_{sj}/n_{\bullet j}$	1

Distribució condicionada:

$x/y=y_j$	x_1	x_2	$\cdots$	x_i	$\cdots$	x_s	
$n_{i/j}$	n_{1j}	n_{2j}	$\cdots$	n_{ij}	$\cdots$	n_{sj}	$n_{\bullet j}$
$f_{i/j}$	$n_{1j}/n_{\bullet j}$	$n_{1j}/n_{\bullet j}$	$\cdots$	$n_{ij}/n_{\bullet j}$	$\cdots$	$n_{sj}/n_{\bullet j}$	1

Si dues variables són **independents**:

$$f_{i/j}=f_{i\bullet} \quad \rightarrow \quad \frac{n_{ij}}{n_{\bullet j}}=\frac{n_{i\bullet}}{N} \quad \rightarrow \quad n_{ij}=\frac{n_{i\bullet}n_{\bullet j}}{N}$$

Vector de mitjanes

Matriu de variàncies-covariància

$$\vec{\mu} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{s} x_i n_{i\bullet}$$

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{r} y_j n_{\bullet j}$$

$$V = \begin{pmatrix} S_x^2 & S_{xy} \\ S_{yx} & S_y^2 \end{pmatrix}$$

$$S_x^2 = \frac{1}{N}\sum_{i=1}^{s}\left(x_i - \bar{x}\right)^2 n_{i\bullet} = \frac{1}{N}\sum_{i=1}^{s} x_i^2 n_{i\bullet} - \bar{x}^2$$

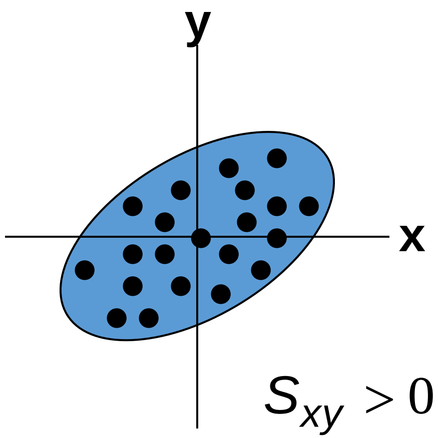
$$S_y^2 = \frac{1}{N}\sum_{j=1}^{r}\left(y_j - \bar{y}\right)^2 n_{\bullet j} = \frac{1}{N}\sum_{j=1}^{r} y_j^2 n_{\bullet j} - \bar{y}^2$$

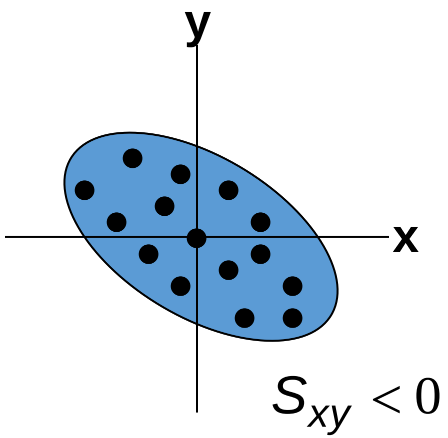
$$\boxed{S_{xy} = \frac{1}{N}\sum_{i=1}^{s}\sum_{j=1}^{r}\left(x_i - \bar{x}\right)\left(y_j - \bar{y}\right) n_{ij} = \frac{1}{N}\sum_{i=1}^{s}\sum_{j=1}^{r} x_i y_j n_{ij} - \bar{x}\bar{y} = S_{yx}}$$

La **covariància** és la variació conjunta, per terme mitjà, d'ambdues variables. Si és distinta de zero, hi ha una cooperació entre les variables i, per tant, no són independents.

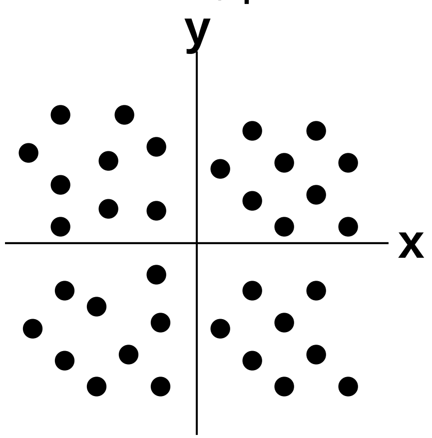
$S_{xy} > 0$ Hi ha una tendència positiva, és a dir, els majors valors de la variable X es corresponen amb els majors valors de la variable Y (com major siga la quantitat d'hores d'estudi, major proporció d'aprovats).

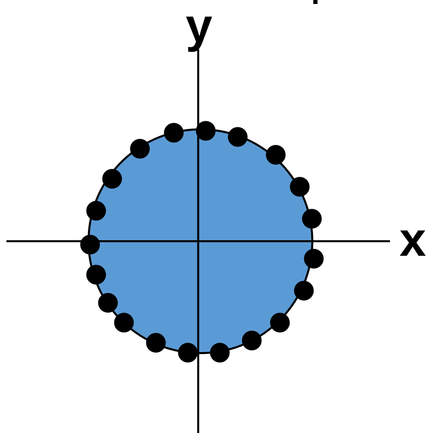
$S_{xy} < 0$ Hi ha una tendència contrària, és a dir, els majors valors de la variable X es corresponen amb els valors menors de la variable Y (com major siga renda per càpita en un país, menor mortalitat infantil).





$S_{xy} = 0$ No hi ha relació lineal, però n'hi pot haver d'un altre tipus.





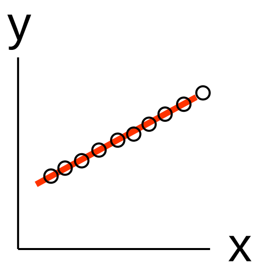
- La covariància, positiva o negativa, detecta una relació entre les variables de tipus **lineal**.

- **No obstant això, la magnitud de la covariància no aporta informació.**

- Com podem mesurar el **grau de relació lineal** que tenen les variables?

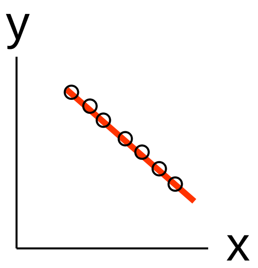
Amb el coeficient de correlació:

$$\boxed{r_{xy} = \frac{S_{xy}}{S_x S_y} \quad , \quad -1 \leq r_{xy} \leq 1}$$



$r_{xy} = 1$

Relació lineal perfecta amb pendent positiva



$r_{xy} = -1$

Relació lineal perfecta amb pendent negativa

$r_{xy} > 0$ A mesura que s'acosta a 1, millor relació lineal positiva

$r_{xy} < 0$ A mesura que s'acosta a -1, millor relació lineal negativa

$r_{xy} = 0$ **No hi ha relació lineal**, però n'hi pot haver d'un altre tipus.

Si dues variables són independents $\rightarrow$ $S_{xy} = r_{xy} = 0$ però no al revés.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})(y_j - \bar{y})n_{ij} = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})(y_j - \bar{y})\frac{n_{i\bullet}n_{\bullet j}}{N} =$$

$$= \frac{1}{N}\sum_{i=1}^{r}(x_i - \bar{x})\,n_{i\bullet}\frac{1}{N}\sum_{j=1}^{s}(y_j - \bar{y})\,n_{\bullet j} = 0 \times 0 = 0$$

A vegades podem crear o hem de crear una variable a partir d'unes altres:

$$W = a + bX + cY$$

$$\bar{W} = a + b\bar{X} + c\bar{Y}$$

$$S_w^2 = b^2 S_x^2 + c^2 S_y^2 + 2bcS_{xy}$$

$$W = a + bX$$

$$\bar{W} = a + b\bar{X}$$

$$S_w^2 = b^2 S_x^2$$

Virgilio Pérez Giménez – virgilio.perez@uv.es

TEMA 3
ANÀLISI DE REGRESSIÓ

Estadística bàsica

Curs 2020/2021

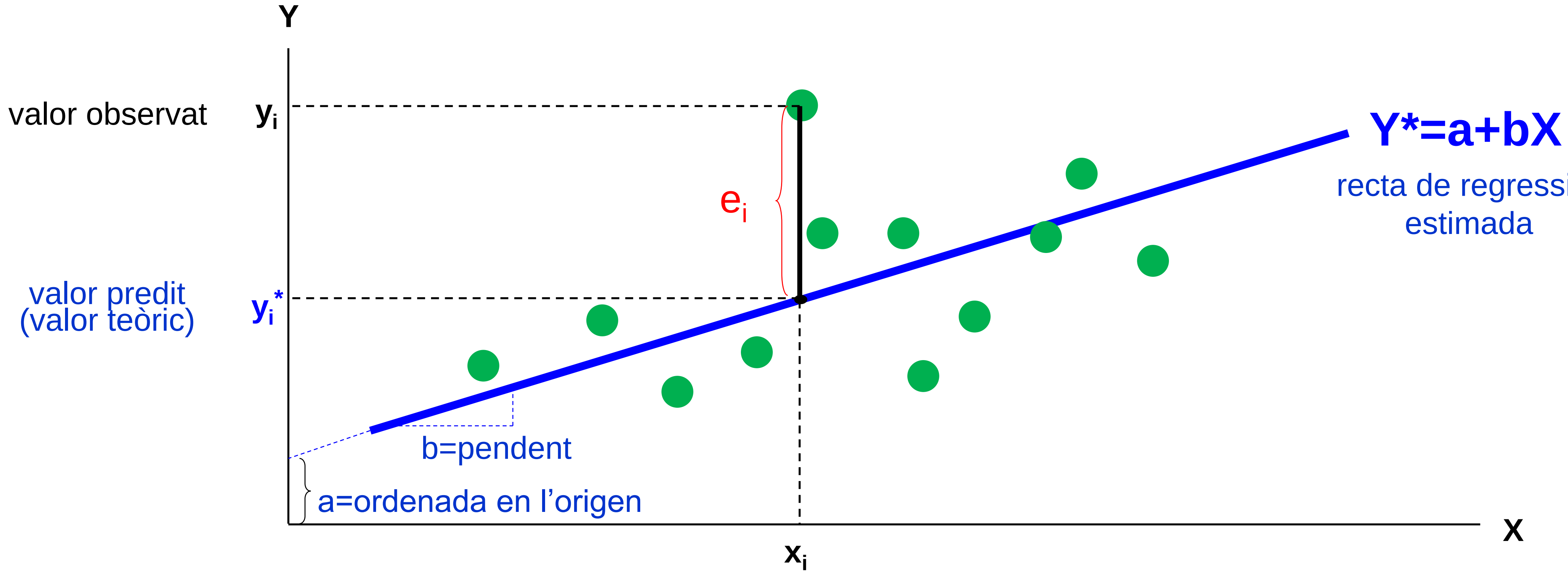
Universitat de València
Facultat d' Economia

- El coeficient de correlació lineal ens indica el grau de relació lineal que hi ha entre les variables, **però no la relació de causalitat.**
- Dues variables relacionades linealment, per definició, han de tenir un coeficient de correlació igual a 1 o -1. Per exemple, $Y = 20 + 15X$, en què la variable Y representa la facturació i la variable X les hores treballades.
- Si la variable X fora la despesa anual en màrqueting i la variable Y les vendes anuals obtingudes per una empresa, no s'obtindran les vendes exactes a partir de la inversió en màrqueting, però sí una **aproximació.**
- La **teoria de la regressió** s'encarrega de determinar l'estructura de dependència que millor expresse la relació entre les variables:

$\boldsymbol{Y^* = a + bX}$ **(lineal)** ; $Y^* = a + bX + cX^2$ (parabòlica) ; $Y *= a \cdot b^X$ (exponencial)

$\boldsymbol{Y^*}$seran els **valors teòrics** obtinguts a partir de la funció de dependència establerta. La diferència entre els valors teòrics i els reals, per a cada element, serà **l'error** comès amb l'ajust:

$$\boxed{y_i^* - y_i = e_i}$$



La recta que millor s'ajuste al nostre núvol de punts serà aquella en què els errors o residus siguen mínims.

- La **REGRESSIÓ LINEAL MÍNIM QUADRÀTICA** determina la relació funcional que minimitza els errors al quadrat:

$$\min \sum_i \left(y_i - y_i^*\right)^2 = \min \sum_i (y_i - (a + bx_i))^2 = h(a,b)$$

$$\left.\begin{aligned} \frac{\partial h(a,b)}{\partial a} &= 0 \\ \frac{\partial h(a,b)}{\partial b} &= 0 \end{aligned}\right\} \to \quad a = \bar{y} - b\bar{x} \qquad b = \frac{S_{xy}}{S_x^2}$$

ordenada en l'origen

coeficient de regressió

(divisió entre la covariància i la variància de la variable explicativa)

Y*=a + bX

- El coeficient de regressió indica la pendent de la recta de regressió, és a dir, l'impacte sobre la variable Y de l'increment unitari de la variable X.

- Anàlogament, si l'ajust és X* = c + d Y , l'ordenada en l'origen i el coeficient de regressió:

$$c = \bar{x} - d\bar{y} \quad \text{i} \quad d = \frac{S_{xy}}{S_y^2}$$

- Una vegada hem calculat l'ajust, hem de ser capaços de valorar la seua **bondat**, és a dir, la qualitat de l'ajust (si els valors teòrics s'acosten o no als reals).

- Per a quantificar la distància dels valors d'una variable X a la seua mitjana, utilitzem la variància:

$$S_x^2 = \frac{1}{N}\sum_i (x_i - \overline{x})^2 n_i$$

- Per a quantificar la distància dels valors teòrics Y* als reals Y, utilitzarem l'**Error Quadràtic Mitjà**:

$$ECM = \frac{1}{N}\sum_i \left(y_i - y_i^*\right)^2 n_i = \frac{1}{N}\sum_i (e_i)^2 n_i$$

- La propietat fonamental de la mitjana diu que: $\frac{1}{N}\sum_i (x_i - \overline{x}) n_i = 0$

- De la mateixa manera: $\frac{1}{N}\sum_i \left(y_i - y_i^*\right) n_i = 0 \rightarrow \frac{1}{N}\sum_i e_i n_i = \overline{e} = 0$

$$ECM = \frac{1}{N}\sum_i \left(y_i - y_i^*\right)^2 n_i = \frac{1}{N}\sum_i {e_i}^2 n_i = \frac{1}{N}\sum_i {e_i}^2 n_i - \overline{e}^2 = S_e^2$$

variància residual o variància no explicada pel model

- Si:
  - $\boxed{S_e^2 = 0 \quad \to \quad e_i = \bar{e} = 0}$ → **AJUST PERFECTE**
  - $\boxed{S_e^2 \neq 0}$ → **Hi ha errors i l'AJUST NO SERÀ PERFECTE.**

- Una altra manera de raonar la bondat de l'ajust és a través de la comparació de les variàncies dels valors reals i dels teòrics:

- Si:
  - $\boxed{S_{y^*}^2 = S_y^2}$ → **AJUST PERFECTE:** la variància explicada pel model coincideix amb la variància real dels valors originals.
  - $\boxed{S_{y^*}^2 \neq S_y^2}$ → **Hi ha errors i l'AJUST NO SERÀ PERFECTE.**

- Quan l'ajust no siga perfecte, haurem de calcular una mesura relativa que ens permeta valorar com de bo és l'ajust.

- Aquesta mesura és el **COEFICIENT DE DETERMINACIÓ:**

$$R^2 = \frac{S^2_{y*}}{S^2_y}$$

Recull quina part de la variabilitat real és explicada pel model $\boxed{0 \le R^2 \le 1}$

- El coeficient de determinació es pot expressar de diferents maneres:

$$R^2 = \boxed{\frac{S^2_{y*}}{S^2_y}} = \frac{S^2_y - S^2_e}{S^2_y} = \boxed{1 - \frac{S^2_e}{S^2_y}} = \frac{b^2 S^2_x}{S^2_y} = \boxed{\frac{S^2_{xy}}{S^2_x S^2_y} = r^2}$$

- Si:

$$\boxed{R^2 = 1 \rightarrow S^2_{y*} = S^2_y \rightarrow S^2_e = 0 \rightarrow r = \pm 1}$$

AJUST PERFECTE.

$$R^2 = 0 \begin{cases} S^2_{y*} = 0 \rightarrow y_i^* = \bar{y}^* = \bar{y} \\ S^2_e = S^2_y \\ r = 0 \end{cases}$$

NO HI HA AJUST LINEAL.

- La utilitat d’obtenir una relació funcional entre X i Y és calcular prediccions de Y a partir de valors de X.

- La fiabilitat de les prediccions dependrà de la bondat de l’ajust realitzat.

$$0 \leq R^2 \leq 1$$

pitjor ajust

millor ajust

Prediccions poc fiables

Prediccions fiables

REFLEXIONS

L'estadística ens diu que l'ajust $\mathrm{Y^* = a + bX}$ serà tan bo com l'ajust $\mathrm{X^* = c + dY}$, ja que en ambdós casos, la bondat ve determinada per R^2.

$$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

Per tant, no serà l'estadística la que decidisca sobre el millor ajust, sinó l'objectiu que es pretenga predir o la teoria econòmica que es vulga demostrar.

Calculat l'ajust $\mathrm{Y^* = a + bX}$, no podem obtenir $\mathrm{X^* = c + dY}$ buidant del primer.

Si hem obtingut l'ajust $\mathrm{Y^* = a + bX}$ a partir d'un interval de dades per a la variable X, podem obtenir estimacions de Y fiables, dins del rang d'aquest interval. Fora del rang, hem de suposar que el comportament lineal es manté.

Per exemple, a partir dels parells de valors (15,34), (16,31), (17,41), (19,32), (19,39), (22,43), (23,46), (24,55), (26,44), (26,53), (27,67), (28,45) podem estimar la recta de regressió $\mathrm{Y^* = 5.92 + 1.75X}$. Per a valors de X com, per exemple: 18, 20, 21 o 25, podríem estimar Y amb la fiabilitat donada pel coeficient de determinació. Però fora d'aquest rang, per al qual hem obtingut la recta, hauríem de suposar que la recta es manté.

Virgilio Pérez Giménez – virgilio.perez@uv.es

TEMA 4.1

ANÀLISI DE DADES TEMPORALS:

ÍNDEXS ECONÒMICS

Part I

Estadística bàsica

Curs 2020/2021

VNIVERSITAT DE VALÈNCIA
Facultat d' Economia

- Donada una sèrie d'observacions, ordenades en el temps i obtingudes en períodes de la mateixa duració (dies, setmanes, mesos, anys…):

$$y_0, y_1, \dots, y_t, \dots, y_T$$

El primer que ens interessarà d'aquestes sèries de dades són les seues **variacions intertemporals.** La variació més senzilla és la **VARIACIÓ ABSOLUTA:**

$$VA(y_t) = y_t - y_{t-n}$$

$VA(y_t) > 0 \longrightarrow$ evolució creixent

$VA(y_t) < 0 \longrightarrow$ evolució decreixent

$VA(y_t) = 0 \longrightarrow$ estancament

- Sovint es designa per $\Delta(y_t) = y_t - y_{t-1}$ la variació entre dos períodes consecutius.

- El problema d’aquest indicador és que observacions molt diferents en magnitud poden donar la mateixa variació absoluta i no és possible comparar-les:

$$y_{t-n} = 100 \quad , \quad y_t = 150$$

$$y_{t-n} = 1000 \quad , \quad y_t = 1050$$

$$VA(y_t) = 50$$

- Seria correcte utilitzar una <u>mesura de variació relativa</u>, com la **TAXA DE VARIACIÓ:**

$$TV(y_t) = \frac{VA(y_t)}{y_{t-n}} = \frac{y_t - y_{t-n}}{y_{t-n}} = \frac{y_t}{y_{t-n}} - 1$$

La taxa de variació així definida està en <u>tant per un</u>. Si es multiplica per 100, s’expressa en <u>tant per cent</u>.

- La taxa de variació pot ser positiva, negativa o nul·la. Si tots els valors de la sèrie són positius, el signe de la taxa dependrà del signe de la variació absoluta.

➢ En l’exemple anterior:

$$TV(y_t) = \frac{VA(y_t)}{y_{t-n}} = \frac{50}{100} = 0'5$$ o el que és equivalent: $$0'5 \cdot 100 = 50\%$$

És a dir, la variable entre el període “*t* ” i “*t-n*” s’ha incrementat un 50%.

$$TV(y_t) = \frac{VA(y_t)}{y_{t-n}} = \frac{50}{1000} = 0'05$$ o el que és equivalent: $$0'05 \times 100 = 5\%$$

És a dir, la variable entre el període “*t* ” i “*t-n*” s’ha incrementat un 5%.

➢ La **taxa de variació**, com tota mesura relativa, és **ADIMENSIONAL**, la qual cosa ens permet comparar l’evolució, no sols de sèries amb distinta magnitud, sinó també de sèries amb diferents unitats (dòlars i euros, per exemple).

- Si la sèrie fora decreixent, la taxa de variació seria:

$$y_{t-n} = 150 \qquad y_t = 100 \qquad \longrightarrow \qquad TV(y_t) = \frac{-50}{150} = -0'\hat{3}$$

Per tant, l'increment de 100 a 150 és d'un 50%, mentre que el decrement de 150 a 100 és d'aproximadament un 33%. No és el mateix augmentar 50 unitats des de 100 que reduir 50 unitats des de 150.

- A partir de la taxa de variació es dedueix el concepte de **FACTOR DE VARIACIÓ UNITÀRIA:**

$$TV(y_t) = \frac{y_t - y_{t-n}}{y_{t-n}} = \frac{y_t}{y_{t-n}} - 1 \quad \rightarrow \quad \boldsymbol{TV(y_t) + 1} = \frac{y_t}{y_{t-n}}$$

- Si es multiplica el valor de la sèrie en el moment "*t-n*" pel factor de variació unitària, s'obtindrà el valor de la sèrie en el moment "*t*".

- En l'exemple anterior, en el primer cas la taxa era de 50%. Per tant:

$$y_{t-n} = 100 \quad \rightarrow \quad 100 \times 1'50 = 150 = y_t$$

En el segon cas, la taxa de variació era de 5%. Per tant:

$$y_{t-n} = 1000 \quad \rightarrow \quad 1000 \times 1'05 = 1050 = y_t$$

- Com que la taxa de variació absoluta és positiva, la taxa de variació unitària és major que la unitat, però quan la variació absoluta siga negativa, la taxa de variació unitària serà inferior a la unitat:

$$y_{t-n} = 150 \ \ e \ \ y_t = 100 \quad \longrightarrow \quad TV(y_t) = \frac{-50}{150} = -0'\hat{3} \quad \longrightarrow \quad TV(y_t) + 1 = 0'\hat{6}$$

$$y_{t-n} = 150 \quad \rightarrow \quad 150 \times 0'\hat{6} = 100 = y_t$$

- El que acabem d'aplicar equival a calcular un preu amb IVA o calcular un preu final, donat el percentatge de rebaixa:

preu d'uns pantalons sense IVA 78 €
IVA del 21% → preu del pantaló amb IVA: $78 \times 1.21 = 94.38$ €

$78 \times 0.21 = 16.38 \longrightarrow 78 + 16.38 = 94.38$

preu d'uns pantalons 94.38 €
rebaixes del 30% → preu del pantaló rebaixat $94.38 \times 0.70 = 66.07$ €

$94.38 \times 0.30 = 28.31 \longrightarrow 94.38 - 28.31 = 66.07$

Virgilio Pérez Giménez – virgilio.perez@uv.es

Virgilio Pérez Giménez – virgilio.perez@uv.es

TEMA 4.1

ANÀLISI DE DADES TEMPORALS:

ÍNDEXS ECONÒMICS

Part II

Estadística bàsica

Curs 2020/2021

VNIVERSITAT DE VALÈNCIA
Facultat d' Economia

- El problema que ens plantegem ara és la **comparació d’una sèrie d’observacions respecte a una situació inicial o de referència**. Per això, considerem dos aspectes importants:

  - La fixació del **punt de referència** (ha de ser un período normal)

  - Si les magnituds a comparar són **simples** (formades per un sol element) o **compostes** (més d’un element). ***Per exemple:*** *evolució del preu de les taronges (simple) o evolució del preu dels cítrics (composta).*

- Un **NOMBRE ÍNDEX** és una mesura estadística que ens permet estudiar els canvis que es produeixen en una magnitud simple o composta respecte al temps o a l’espai.

  - Al període inicial se’l denomina **base o de referència** i al període que volem comparar, **període actual o corrent**.

  - Parlarem de **nombres índexs simples** (NIS) o **compostos** (NIC) segons si les magnituds a comparar són simples o compostes.

$$NIS = \frac{y_t}{y_0}$$

$$NIS = \frac{p_t}{p_0}$$

Índex de preus

- El kg de taronges costava **1 €** i ara costa **1,25 €.**

$$NIS = \frac{q_t}{q_0}$$

Índex de quantitats

- Abans en consumia **2 kg** a la setmana i ara en consumisc **3 kg** a la setmana.

$$NIS = \frac{p_t q_t}{p_0 q_0}$$

Índex de valors

- Comprar taronges em suposava una despesa setmanal de **2€** i ara em suposa **3,75 €**.

➢ La vertadera utilitat dels índexs radica a **sintetitzar la informació de diferents sèries en una única**, de manera que es pot estudiar l'evolució d'un sector o grup de magnituds en el seu conjunt.

- Existeixen diferents procediments per a la construcció dels **nombres índexs compostos:**

$$\frac{\sum_i y_{it}}{\sum_i y_{io}}$$

Aquest índex no pondera, considera tots els productes de la mateixa manera. No és possible calcular-lo quan les variables tenen unitats de mesura diferents. ***Per exemple***: *un índex de combustibles. El carbó es mesura en tones, el petroli en litres i el gas en metres cúbics*.

$$\frac{\sum_i \frac{y_{it}}{y_{io}}}{n}$$

Aquest índex soluciona el problema de tenir diferents magnituds, ja que és una **mitjana d'índexs**, però tampoc pondera perquè considera tots els elements de la mateixa manera. Pot ser útil quan la ponderació és complicada.

➢Els **nombres índexs** més reconeguts que solucionen els problemes de ponderació i de dimensionalitat són els següents:

$$L(p)=\frac{\sum_i p_{it}q_{i0}}{\sum_i p_{i0}q_{i0}}=\frac{\sum_i \frac{p_{it}}{p_{i0}}p_{i0}q_{i0}}{\sum_i p_{i0}q_{i0}}$$

Laspeyres de preus

$$L(q)=\frac{\sum_i q_{it}p_{i0}}{\sum_i q_{i0}p_{i0}}=\frac{\sum_i \frac{q_{it}}{q_{i0}}q_{i0}p_{i0}}{\sum_i q_{i0}p_{i0}}$$

Laspeyres de quantitats

$$P(p)=\frac{\sum_i p_{it}q_{it}}{\sum_i p_{i0}q_{it}}=\frac{\sum_i \frac{p_{it}}{p_{i0}}p_{i0}q_{it}}{\sum_i p_{i0}q_{it}}$$

Paasche de preus

$$P(q)=\frac{\sum_i q_{it}p_{it}}{\sum_i q_{i0}p_{it}}=\frac{\sum_i \frac{q_{it}}{q_{i0}}q_{i0}p_{it}}{\sum_i q_{i0}p_{it}}$$

Paasche de quantitats

Virgilio Pérez Giménez – virgilio.perez@uv.es

- En economia i empresa és molt important el coneixement d'índexs. Per això n'hi ha una gran quantitat: l'**índex de preus al consum** (**IPC**), índexs de producció i preus industrials, de producció i preus agrícoles, índexs comercials de preus al detall, a l'engròs, índexs d'exportacions, d'importacions, índexs d'ocupació, financers, borsaris (IBEX 35), etc.

- L'**IPC** actual té la base en l'any 2016, revisable cada 5 anys. És un índex de Laspeyres encadenat, en què les ponderacions no romanen fixes durant el període de vigència. És dinàmic, permet incloure nous productes en la cistella de la compra, cistella que recull els productes i serveis que representen el consum de la població resident en habitatges familiars a Espanya. La cistella de la compra la componen 479 articles, classificats en 12 grups (**INE**).

- Ha de tenir-se en compte que l'índex de Laspeyres és més econòmic de construir que l'índex de Paasche, ja que Laspeyres manté una ponderació constant (encara que l'IPC la renova cada cert temps) mentre que l'índex de Paasche requereix actualitzar el pes cada període, cosa que significa major informació, és a dir, major cost.

- Els problemes que comporta la construcció d'un índex són:
  - La selecció de les variables que conformen l'índex.
  - Els tipus de magnitud (preus o quantitats produïdes, venudes...).
  - La periodicitat (mensual, trimestral, anual…).
  - La fixació del període base o de referència, cosa que suposa plantejar també la seua **renovació** i per tant, el seu **enllaç** amb sèries anteriors.
  - Definir les ponderacions.
- Un índex és millor que un altre quan compleix una sèrie de propietats matemàtiques en què no entrarem. Ens centrarem en el seu maneig i no en la seua construcció. Una de les principals funcions d'un índex és la **DEFLACTACIÓ D'UNA SÈRIE**.
- Per entendre la deflactació cal conèixer els conceptes d'inflació i deflació.

- La **INFLACIÓ** és un fenomen econòmic consistent en una pujada generalitzada i persistent dels preus dels béns i serveis que es produeixen o es consumeixen.

- La **DEFLACIÓ** és el contrari.

- Tant la inflació com la deflació alteren el **valor dels diners i, per tant, el poder adquisitiu de les persones.**

- **DEFLACTAR** és passar d'un valor corrent a un valor constant. Aquesta acció ens permetrà analitzar **l'evolució REAL d'una sèrie de valors:**

$$\sum_i p_{it} q_{it} \xrightarrow{\text{deflactar}} \sum_i p_{i0} q_{it}$$

VALOR CORRENT

El valor de les "coses" a preus actuals

VALOR CONSTANT

El valor de las "coses" a preus constants

➢ **L'índex de preus de Paasche**, amb la mateixa cobertura que l'agregat de valors corrents, és el deflactor perfecte. És a dir, si es té un valor agregat de cítrics, l'índex de Paasche ha de contenir els mateixos cítrics:

$$\underbrace{\sum_i p_{it} q_{it}}_{\text{corrent}} \Big/ \frac{\sum_i p_{it} q_{it}}{\sum_i p_{i0} q_{it}} = \sum_i p_{it} q_{it} \frac{\sum_i p_{i0} q_{it}}{\sum_i p_{it} q_{it}} = \underbrace{\sum_i p_{i0} q_{it}}_{\text{constant}}$$

➢ En la pràctica, ni se solen tenir els mateixos agregats ni se sol utilitzar Paasche, ja que habitualment l'índex disponible és l'IPC, i és Laspeyres; d'aquesta manera s'aconsegueix una deflactació aproximada.

➢ És importantíssim, a l'hora d'analitzar una sèrie de valors, tenir-la en valors constants, per a eliminar la influència dels preus (deflació o inflació).

➢ Veurem la renovació i enllaç de les sèries directament amb les pràctiques.

Any	IPC (base 2001)	IPC (base 2006)
2001	IPC_{01}^{01} =100	
2002	IPC_{01}^{02}	
2003	IPC_{01}^{03}	¿?
2004	IPC_{01}^{04}	
2005	IPC_{01}^{05}	
2006	IPC_{01}^{06}	IPC_{06}^{06} =100
2007		IPC_{06}^{07}
2008	¿?	IPC_{06}^{08}
2009		IPC_{06}^{09}

L'any 2006 es calculen 2 IPC, un amb la cistella antiga i un altre amb la cistella nova.

enllaçar les sèries

TEMA 4.2

ANÀLISI DE DADES TEMPORALS:

SÈRIES TEMPORALS

Estadística bàsica

Curs 2020/2021

Virgilio Pérez Giménez – virgilio.perez@uv.es

Vniversitat de València
Facultat d' Economia

- L'anàlisi d'una sèrie temporal es du a terme per complir dos objectius:
  - Descriure l'evolució passada d'una variable a través del temps.
  - Realitzar prediccions sobre un futur més o menys pròxim sota la suposició què no es produiran canvis estructurals.

- Desenvoluparem el **MÈTODE CLÀSSIC** de descomposició d'una sèrie, que consisteix a descompondre el valor y_t d'una sèrie en un període t amb quatre components:

  - Tendència (T_t)
  - Cicle (C_t)
  - Estacionalitat (S_t)
  - **Variacions irregulars (I_t)**

$$\left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \quad \mathbf{Y_t = f\,(\,T_t\,,\,C_t\,,\,S_t\,,\,I_t)}$$

- Les 3 primeres són **VARIACIONS SISTEMÀTIQUES** (aquelles que ocorren amb regularitat i poden ser mesurades estadísticament i predir la seua ocurrència).

- La component **IRREGULAR** no es pot controlar ni predir.

- A continuació definirem cadascuna de les components:

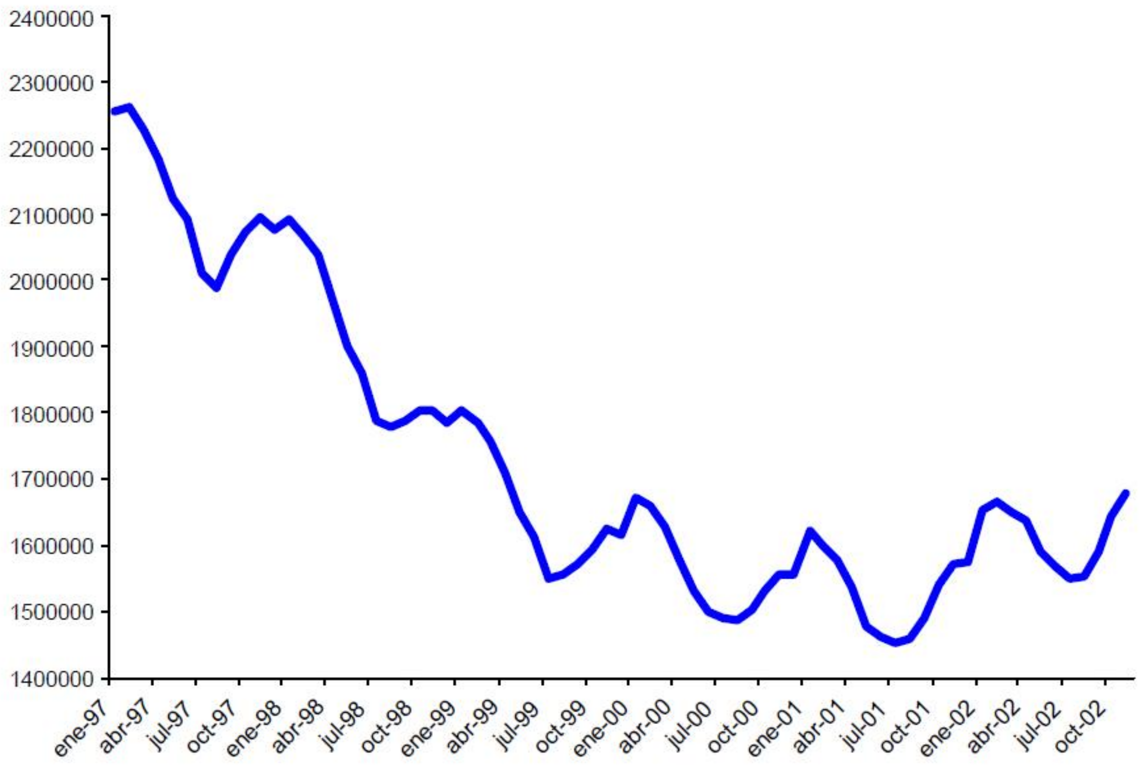
  - **Tendència:** informa sobre l'evolució de la sèrie a **llarg termini** (almenys deu anys). La tendència pot ser creixent, decreixent o estable.
  - **Cicle:** oscil·lacions que es produeixen en un període **superior a l'any**. Es deu principalment a l'alternança de períodes de prosperitat i de depressió. Hi ha qui opina que un cicle comprèn quatre fases: prosperitat, recessió, depressió i recuperació.
  - **Estacionalitat:** oscil·lacions a **curt termini**, en períodes **inferiors a l'any**. Variacions que es repeteixen quasi de la mateixa manera i amb la mateixa regularitat any rere any. Es deuen a causes climatològiques, festes, costums, rebaixes, etc.
  - **Component irregular:** provocada per factores fortuïts, estrictament aleatoris. Recullen tot allò que no consideren les altres components.

- Aquestes components es poden agregar a través d'un **esquema additiu**: $y_t = T_t + C_t + S_t + I_t$ o a través d'un **esquema multiplicatiu**: $y_t = T_t \times C_t \times S_t \times I_t$, entre altres. En economia és molt habitual utilitzar el multiplicatiu.

- Hi ha diferents mètodes per a analitzar **la tendència** d'una sèrie:

  - **Mètode gràfic:** és un mètode molt senzill, però imprecís i no garanteix fiabilitat. En qualsevol cas, la representació gràfica sempre és aconsellable abans d'iniciar l'anàlisi d'una sèrie temporal.

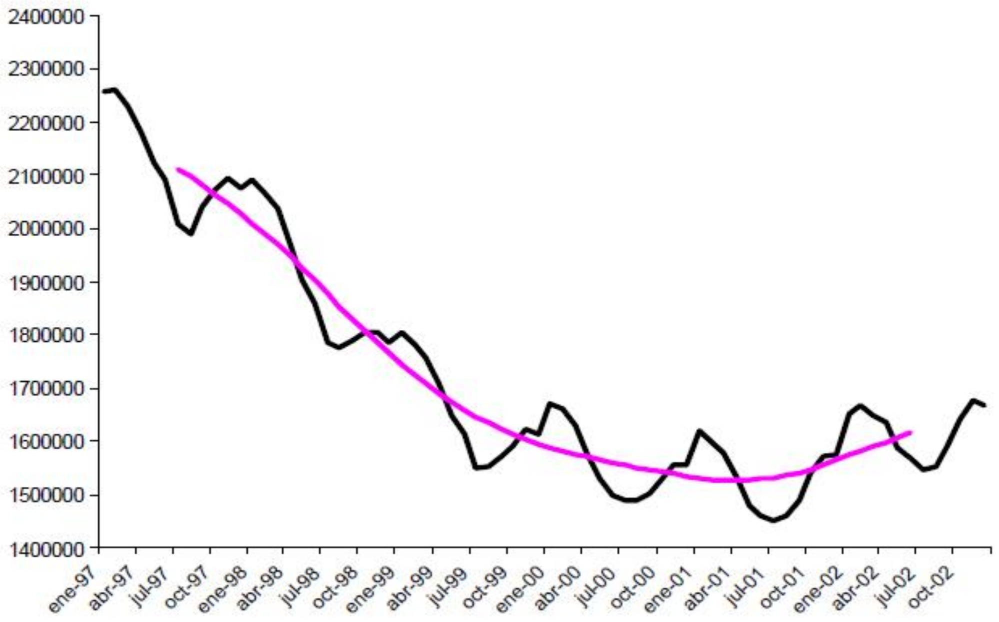
Evolució de la desocupació registrada a Espanya



▪ **Mitjanes mòbils:** és un mètode de suavització, que consisteix a calcular la mitjana de diferents observacions consecutives, mantenint-se el període constant, però desplaçant-se al llarg de la sèrie. Per exemple, si es consideren períodes de 3 observacions, seria:

$$\bar{y}_1 = \frac{y_1 + y_2 + y_3}{3},\ \bar{y}_2 = \frac{y_2 + y_3 + y_4}{3},\ \bar{y}_3 = \frac{y_3 + y_4 + y_5}{3}$$

Tendència de la desocupació registrada (mitjanes mòbils)



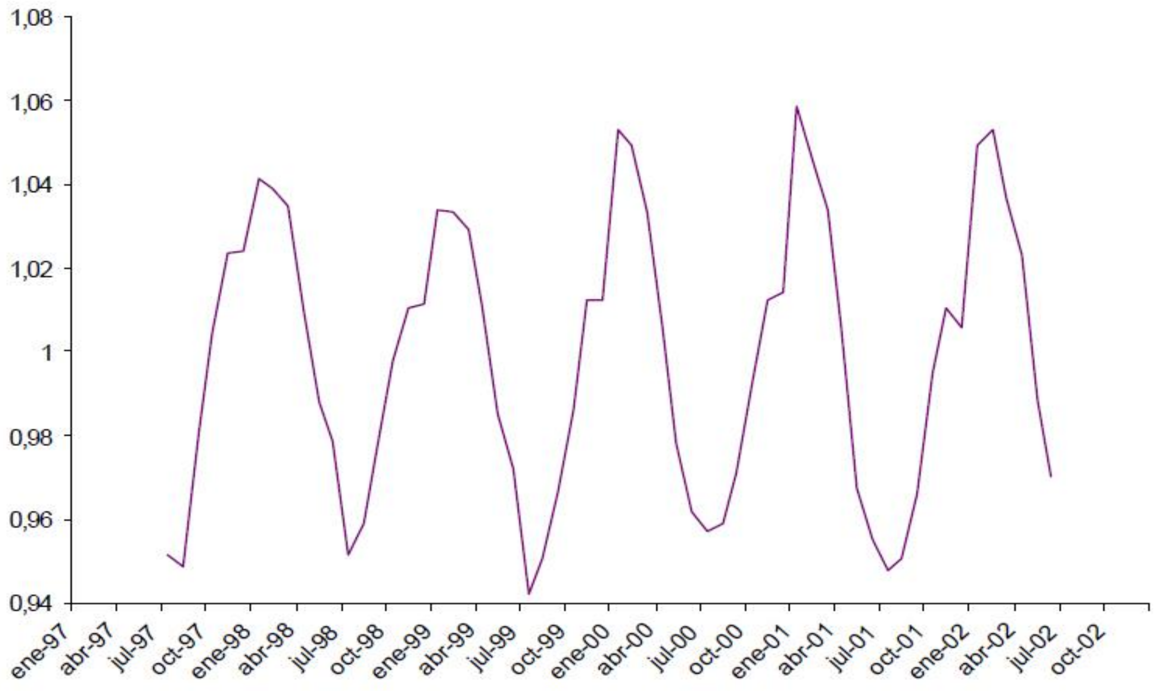
Si hi ha estacionalitat i és mensual, es prenen mitjanes de 12 observacions; si és trimestral, de 4; si és quadrimestral, de 3; i si és anual, se'n solen prendre 3 o 5 dades.

El problema d'aquest mètode és **la seua subjectivitat**: s'ha de decidir el nombre d'observacions, **la pèrdua d'informació** i considerar que **no permet realitzar prediccions**.

- Una vegada determinada la component tendència (cicle-tendència), i admetent un model multiplicatiu, es podria eliminar de la sèrie i obtenir l'estacionalitat i la component irregular:

$$y_t = \frac{TC_t \times S_t \times I_t}{TC_t} = S_t \times I_t$$

Sèrie de desocupació registrada, corregida de la component cicle-tendència



- **Mètode analític:** tracta d'obtenir una funció matemàtica que represente la tendència. Per això es considera la mateixa sèrie temporal com variable dependent i al mateix temps com a variable independent.

Hi ha nombroses funcions per a modelitzar la tendència, però ens centrarem en **la funció lineal:**

$$y_t^* = a + b\,t$$

En què $t = 0, 1, 2, \ldots, T$ és la variable *temps* (en anys) i $t = 0$ és l'origen de la sèrie.

Si $b > 0$, el pendent de la recta serà positiu i per tant la tendència serà creixent. Si $b < 0$, el pendent serà negatiu i la tendència serà decreixent.

L'obtenció de la tendència mitjançant el mètode analític té **dues avantatges**: la possibilitat de mesurar **la bondat de l'ajust** i l'obtenció de **prediccions**.

- **Altres mètodes d'allisatge:** dissenyats per a realitzar prediccions (a diferència del mètode de mitjanes mòbils). No són objecte d'aquest curs.

- Diferents mètodes determinen la **component estacional,** i tots requereixen sempre un pas previ que consisteix a eliminar la tendència per qualsevol dels mètodes anteriors.

- Ens centrarem en l'ús dels **ÍNDEXS DE VARIACIÓ ESTACIONAL**. Com qualsevol índex, ens informa sobre l'evolució, en aquest cas estacional, d'una sèrie.

- La característica principal d'aquests índexs, tal com s'elaboren, és la següent:

  - Si l'estacionalitat és mensual, n'hi haurà 12 índexs (la suma ha de ser un valor igual a 12).
  - Si l'estacionalitat és quadrimestral, n'hi haurà 3 índexs (la suma ha de ser un valor igual a 3).
  - Si l'estacionalitat és trimestral, n'hi haurà 4 índexs (la suma ha de ser un valor igual a 4).
  - I així successivament.

- No entrarem en la construcció de l'índex, però sí en el seu maneig per a realitzar prediccions en sèries temporals.

- El nostre model és el multiplicatiu, i com considerem les components tendència i cicle conjuntament, escriurem així el model:

$$y_t = T_t \cdot S_t \cdot I_t$$

- Per **realitzar prediccions**, el procediment serà:

1. **Mesurarem la tendència** a través de la recta de regressió:

$$\hat{T}_t = y_t^* = a + bt$$

En què la variable explicativa serà el temps *t* mesurat així: 0,1,2,3...

2. **Corregirem per l'estacionalitat**, multiplicant per l'IVE (Índex de Variació Estacional) corresponent:

$$\hat{y}_t = \hat{T}_t \cdot IVE_t$$

Per tant, si no hi haguera estacionalitat, l'IVE seria 1 en un model multiplicatiu, perquè no afecte la component tendència.

TEMA 5

MODELS DE PROBABILITAT UNIVARIANTS

Part I

Nocions de probabilitat

Estadística bàsica

Curs 2020/2021

Virgilio Pérez Giménez – virgilio.perez@uv.es

Vniversitat de València
Facultat d' Economia

- El terme **probabilitat** indica una mesura numèrica de la possibilitat, ocurrència o certitud d'un esdeveniment determinat.

- Els valors de la probabilitat sempre oscil·len des de 0 fins a 1:
  - Una probabilitat pròxima a 0 indica que l'esdeveniment és difícil que ocórrega, i igual a 0 indica que l'esdeveniment és **impossible**.
  - Una probabilitat pròxima a 1 indica que l'esdeveniment, quasi segur, ocorrerà, i igual a 1 que l'esdeveniment és **segur**.

- La probabilitat va sempre associada a un **experiment aleatori**, és a dir, una prova en la qual no es coneixen els resultats amb certesa, encara que sí el conjunt de possibles resultats, anomenat **espai mostral**.

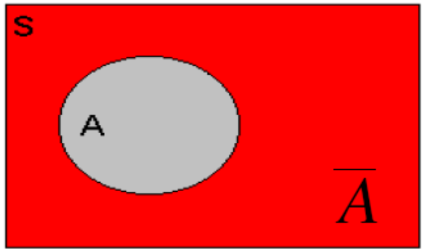
Exemple: "La suma de valors en el llançament de dos daus", on l'espai mostral seria totes les combinacions possibles, és a dir, $S = \{2, 3, 4, \ldots, 12\}$.

Exemple: "L'energia consumida per la ciutat de València durant un mes", on l'espai mostral seria $S = \{0, \ldots, +\infty\}$.

- **Succés** és qualsevol subconjunt de l'espai mostral. Per **exemple,** superar l'assignatura d'estadística bàsica, succés format per diversos elements (aprovat, notable, excel·lent o matrícula d'honor).
- Tota probabilitat ha de complir sempre **dos requisits**:
  - Els valors han d'estar compresos entre 0 i 1.
  - La suma de totes les probabilitats assignades als possibles resultats han de ser igual a 1.

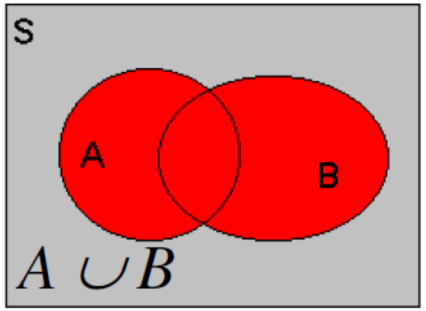
- **La probabilitat d'un succés** s'obté sumant les probabilitats de tots els elements que pertanyen al succés.
  - **Por exemple**, la probabilitat de cadascun dels costats d'un dau és $1/6$, ja que hi ha 6 costats, un per a cada nombre. Quina serà la probabilitat d'obtenir un nombre parell en llançar el dau?
    - El succés està format pels valors 2, 4 i 6.
    - Haurem de sumar la probabilitat de cadascun d'ells: $1/6 + 1/6 + 1/6 = 3/6 = 0{,}5$

- Vegem algunes **relacions bàsiques de probabilitat:**
  - $P(\bar{A}) = 1 - P(A)$, on $\bar{A}$ és un **succés complementari**.



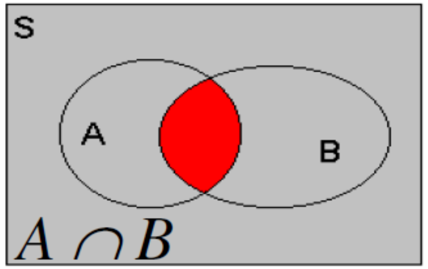
Si tinc un 30% de probabilitat que em contracten, tindré un 70% de probabilitat que no em contracten.

  - La **unió** de dos successos, A i B, és un altre succés que conté tots els elements que pertanyen a A, a B o a tots dos, i es representa amb $A \cup B$.



A: gent que té cotxe
B: gent que té moto
$A \cup B$: gent que té cotxe **o** moto (inclou la gent que té cotxe i moto)

  - La **intersecció** de dos successos, A i B, és un altre succés que conté tots els elements comuns d'A i B i es representa amb $A \cap B$.



$A \cap B$: gent que té cotxe **i** moto

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

➢A vegades la probabilitat d'un succés està influïda per l'ocurrència d'un altre succés relacionat amb l'anterior, i això dona lloc a una **probabilitat condicionada.**

- Per **exemple**, la probabilitat d'aprovar Estadística II (succés A) havent aprovat Estadística I (succés B), la qual cosa es representaria per $\boldsymbol{P(A|B)}$. S'ha d'interpretar com *la probabilitat que ocórrega el succés A sabent que ha ocorregut el succés B* i es calcula de la manera següent:

$$\left.\begin{array}{l} P(A|B) = \dfrac{P(A\cap B)}{P(B)} \\[2ex] P(B|A) = \dfrac{P(A\cap B)}{P(A)} \end{array}\right\} P(A\cap B) = P(B)P(A|B)$$

$$f_{i/j} = n_{ij}/n_{\bullet j}$$

➢Si els successos són **independents:**

$$\left.\begin{array}{l} P(A|B) = \dfrac{P(A\cap B)}{P(B)} = P(A) \\[2ex] P(B|A) = \dfrac{P(A\cap B)}{P(A)} = P(B) \end{array}\right\} P(A\cap B) = P(A)P(B)$$

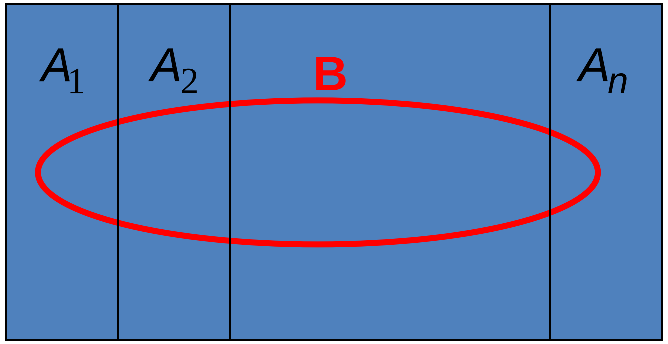
$$\frac{n_{ij}}{N} = \frac{n_{i.}}{N}\frac{n_{.j}}{N}$$

Virgilio Pérez Giménez – virgilio.perez@uv.es

➢**El teorema de Bayes** és un mètode que proporciona probabilitats **posteriors o *a posteriori*** a partir d'informacions **prèvies o *a priori***. Dit d'una altra manera, és un mètode de càlcul de probabilitats actualitzades. L'esquema és el següent:

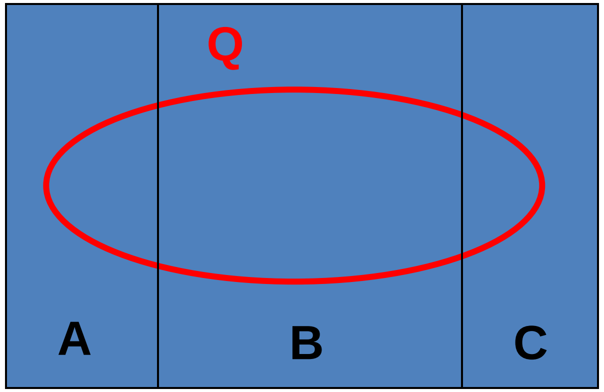
$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B)} =$$

$$= \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)}$$



En el Mercat Continu (MC) espanyol cotitzen 15 entitats del sector bancari, 35 empreses industrials i 10 empreses del sector tecnològic. La probabilitat que un banc dels que cotitzen en el Mercat Continu es declare en fallida és 0,01, la probabilitat que ho faça una empresa industrial és 0,02 i la d'una empresa tecnològica és de 0,1.

Havent-se produït una fallida, quina és la probabilitat que es tracte d'una empresa tecnològica?



A: cotitza en el MC i és del sector bancari

B: cotitza en el MC i és del sector industrial

C: cotitza en el MC i és del sector tecnològic

Q: cotitza en el MC i es declara en fallida

$$P(A)=\frac{15}{60} \quad ,P(B)=\frac{35}{60} \quad ,P(C)=\frac{10}{60} \qquad P(Q|A)=0.01 \quad ,P(Q|B)=0.02 \quad ,P(Q|C)=0.1$$

$$P(C|Q)=\frac{P(C\cap Q)}{P(Q)}=\frac{P(C\cap Q)}{P(A\cap Q)+P(B\cap Q)+P(C\cap Q)}=\frac{P(C)P(Q|C)}{P(A)P(Q|A)+P(B)P(Q|B)+P(C)P(Q|C)}=0.5411$$

TEMA 5

MODELS DE PROBABILITAT UNIVARIANTS

Part II

Variable aleatòria i distribució de probabilitat

Estadística bàsica

Curs 2020/2021

Virgilio Pérez Giménez – virgilio.perez@uv.es

Universitat de València
Facultat d' Economia

- Una **variable aleatòria** és una descripció numèrica d'un experiment aleatori, és a dir, d'un experiment en el qual no coneixem els resultats amb certesa.

Per exemple:

- Experiment aleatori: llançar dues monedes
- Espai mostral: $\Omega = \{CC, XX, CX, XC\}$
- Variable aleatòria: nombre de cares
- Camp de variació: $\Omega_X = \{0,1,2\}$

- Els valors d'una variable aleatòria, com que són incerts, han d'anar associats a una probabilitat d'ocurrència, que en l'exemple anterior serien:

$$P(X=0) = P(XX) = \frac{1}{4}$$

$$P(X=1) = P(CX \cup XC) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(X=2) = P(CC) = \frac{1}{4}$$

No centrarem la nostra anàlisi en variables aleatòries associades a jocs d'atzar, sinó en variables del tipus següent:

Experiment	**Variable aleatòria**	**Camp de variació**
Telefonar a 20 clients	Nombre de clients que realitzen la comanda	0, 1, 2, … , 20
Inspecció de qualitat de 50 vehicles	Nombre de vehicles defectuosos	0, 1, 2, … , 50
Funcionament d'una centraleta telefònica	Nombre de telefonades per hora	0, 1, 2, ….
Control del servei d'autobusos	Temps que tarden a completar el recorregut	$x \geq 0$
Qualitat docent	Percentatge d'aprovats	$0 \leq x \leq 1$

- En l'anàlisi descriptiva, hem parlat de **distribució de freqüències.** Teníem un total de **N** elements que es distribuïen entre els diferents elements del col·lectiu, cosa que donava lloc a les freqüències absolutes $\mathbf{n_i}$.

- Ara, parlem de **distribució de probabilitat** en què tenim una màxima probabilitat (1) que es distribueix entre els diferents valors de la variable aleatòria.

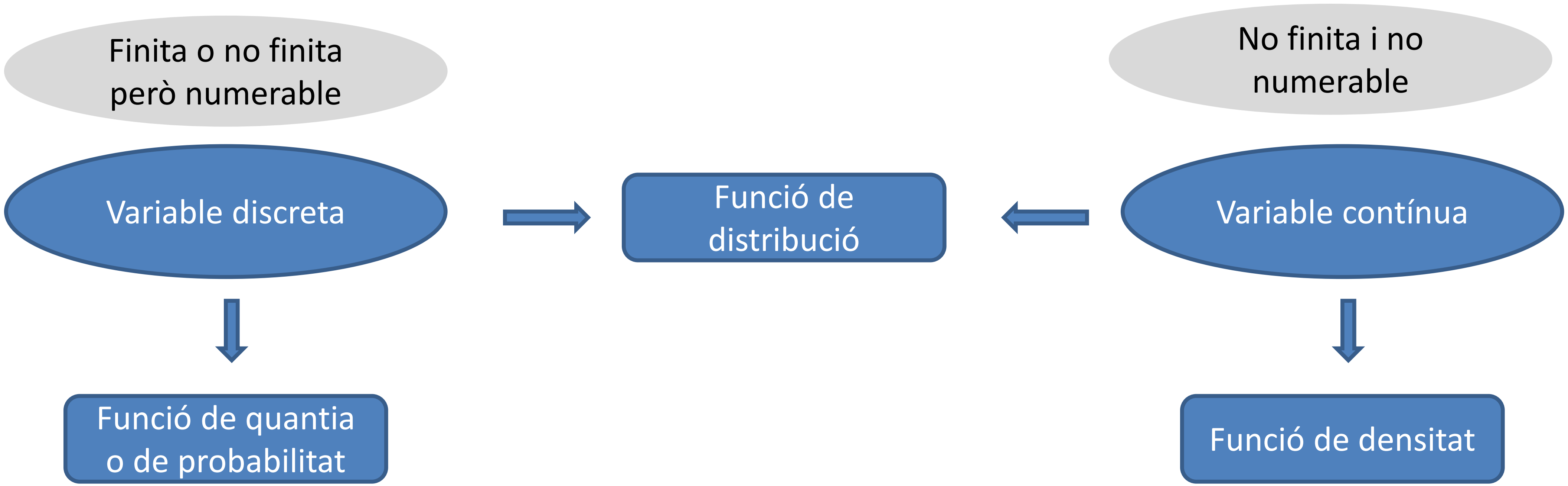
x_i	n_i
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_s	n_s
	N

Distribució de freqüències

x_i	p_i
x_1	p_1
x_2	p_2
$\vdots$	$\vdots$
x_s	p_s
	1

Distribució de probabilitat

- Abans teníem diferents maneres de presentar una **distribució de freqüències**, a través de **n_i, f_i, N_i o F_i**. Ara, també tenim diferents maneres de presentar una **distribució de probabilitat**, a través de la **funció de quantia** (o de probabilitat), de la **funció de densitat**, de la **funció de distribució** o de la funció característica. Aquestes dependran de si la variable aleatòria és de tipus discret o continu.



Experiment	**Variable aleatòria**	**Camp de variació**
Telefonar a 20 clients	Nombre de clients que realitzen la comanda	0, 1, 2, … , 20
Inspecció de qualitat de 50 vehicles	Nombre de vehicles defectuosos	0, 1, 2, … , 50
Funcionament d'una centraleta telefònica	Nombre de telefonades per hora	0, 1, 2, ….
Control del servei d'autobusos	Temps que tarden a completar el recorregut	$x \geq 0$
Qualitat docent	Percentatge d'aprovats	$0 \leq x \leq 1$

Virgilio Pérez Giménez – virgilio.perez@uv.es

DISTRIBUCIONS DE PROBABILITAT

- Funció de quantia o de probabilitat
- Funció de densitat
- Funció de distribució
- Funció característica

➢ **La funció de quantia o de probabilitat** es defineix com la probabilitat associada a cadascun dels valors de la variable aleatòria:

$$P : \Omega_X \rightarrow [0,1] \quad \text{de manera que:} \quad P(X = x_i) \geq 0 \qquad \sum_{x_i \in \Omega_X} P(X = x_i) = 1$$

❑ **Exemple.** La variable aleatòria "notes en estadística" té la distribució de probabilitat següent:

qualificació	**x_i**	**probabilitat**
Suspens	1	0.48
Aprovat	2	0.32
Notable	3	0.16
Excel·lent	4	0.04

➢ Si coneixem la funció de quantia o probabilitat, denotada per $P(x)$ o $P(X = x_i)$, **podrem calcular la probabilitat de qualsevol succés**.

❑ **Exemple:** la probabilitat de superar la matèria serà:

$$P(X > 1) = P(X = 2) + P(X = 3) + P(X = 4) = 1 - P(X \leq 1) = 1 - P(X = 1) = 1 - 0.48 = 0.52$$

- En les variables aleatòries de **tipus continu** no es pot parlar de la probabilitat en un punt, ja que existeixen infinits punts no numerables. Per tant, la probabilitat assignada a un punt és 0:

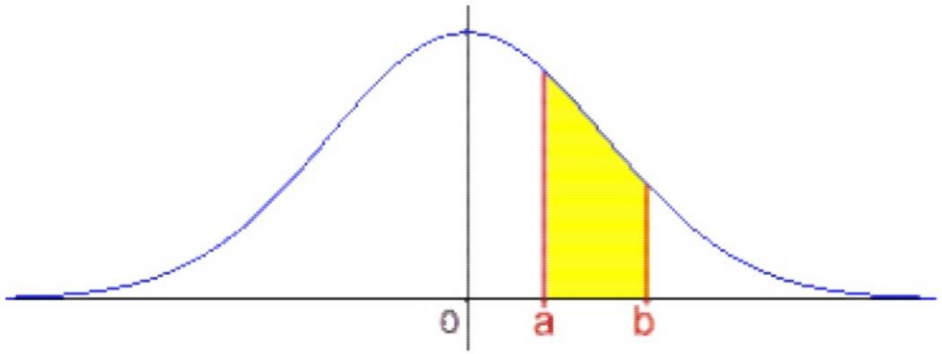
$$P(X = x_i) = 0$$

- Sí que es pot parlar de la probabilitat en un interval i s'obté de la maneresegüent :

$$P(a < X < b) = P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = \int_a^b f(x)dx$$

on ***f(x)*** és la **funció de densitat**, es compleix que: $\int_{\Omega_X} f(x)dx = 1$

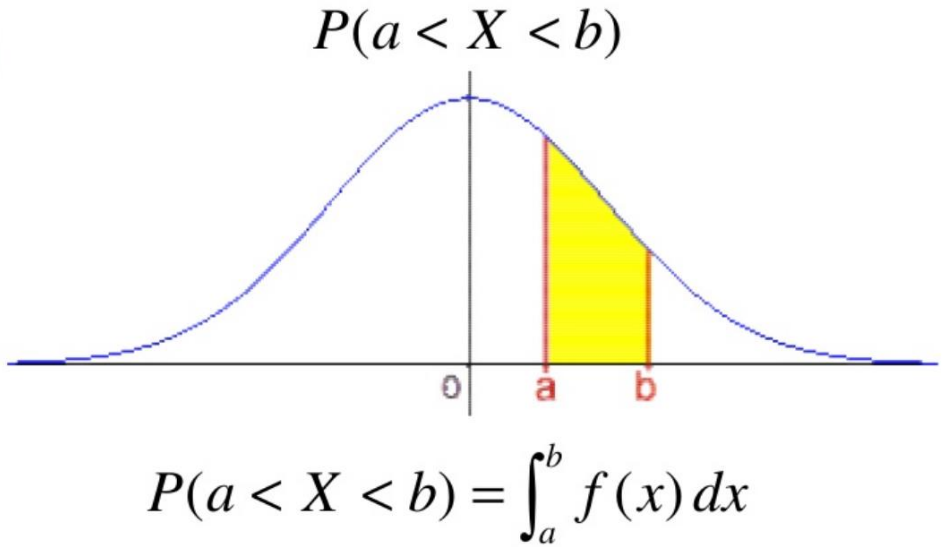
$$P(a < X < b)$$

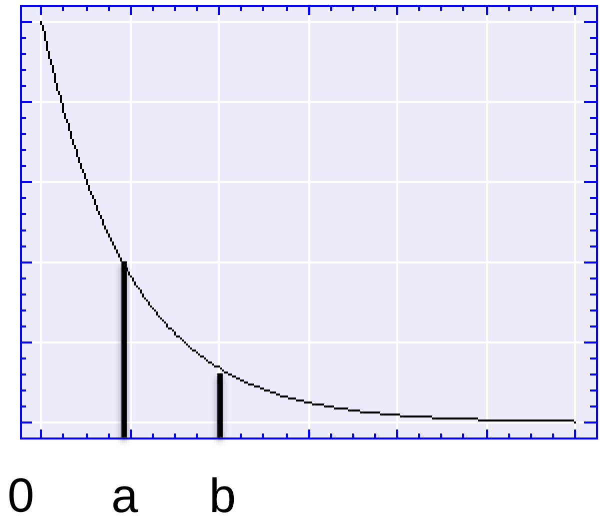


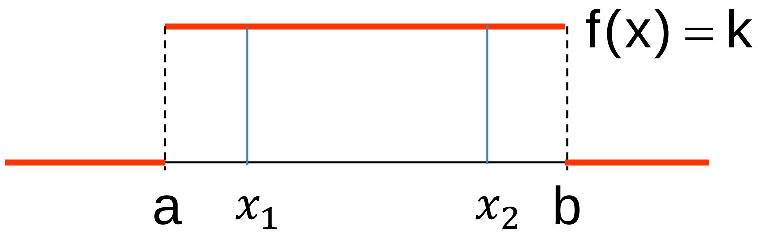
$$P(a < X < b) = \int_a^b f(x)\,dx$$

- Si es coneix la funció de densitat d'una variable aleatòria contínua, estarà completament especificada la seua distribució de probabilitat i **es podrà calcular la probabilitat de qualsevol succés**.

- En las variables aleatòries de tipus discret les probabilitats són **punts** i en las variables aleatòries de tipus continu les probabilitats són **àrees**.


$P(a < X < b)$
0
a
b
$P(a < X < b) = \int_a^b f(x)\,dx$


0
a
b


f(x) = k
a
x_1
x_2
b

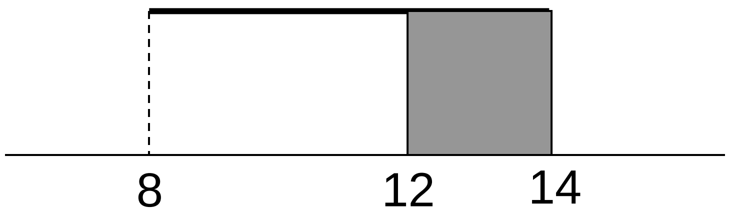
Virgilio Pérez Giménez – virgilio.perez@uv.es

- **Exemple.** Una centraleta telefònica rep trucades entre les 8 i les 14 hores. S'admet la funció de densitat següent:

$$f(x) = \begin{cases} \frac{1}{6} & 8 \le x \le 14 \\ 0 & \text{a la resta} \end{cases}$$

La probabilitat que la trucada es reba després de les 12 hores s'obtindria de la manera següent:

$$P(X > 12) = \int_{12}^{14} \frac{1}{6} dx = \frac{1}{6} [x]_{12}^{14} = \frac{1}{6}(14 - 12) = \frac{2}{6} = \frac{1}{3} = 0.33$$



➢ Igual que en les distribucions de freqüències definim les freqüències acumulades N_i o F_i, també podem definir probabilitats acumulades, representades per la **funció de distribució:**

$$F_x(x) = P(X \leq x_i) \rightarrow \begin{cases} \sum_{-}^{x_i} P(X = x_i) & \text{discretes} \\ \int_{-}^{x} f(x)dx & \text{contínues} \end{cases}$$

Una funció de distribució $F_x(x)$ sempre comença en 0, acaba en 1 i és creixent, ja que acumula probabilitats, que són quantitats positives.

❑ **Exemple.** Donada la variable aleatòria **“notes en estadística”**:

Qualificació	**x_i**	$\boldsymbol{P(X = x_i)}$	$\boldsymbol{F_X(x)}$
Suspens	1	0.48	0.48
Aprovat	2	0.32	0.80
Notable	3	0.16	0.96
Excel·lent	4	0.04	1

Virgilio Pérez Giménez – virgilio.perez@uv.es

- **Exemple.** Donada la variable aleatòria “moment en què es rep una trucada en una centraleta telefònica”:

$$f(x) = \begin{cases} \frac{1}{6} & 8 \leq x \leq 14 \\ 0 & \text{a la resta} \end{cases} \Rightarrow \boxed{P(X > 12) = \int_{12}^{14} \frac{1}{6} dx = \frac{1}{6}[x]_{12}^{14} = \frac{1}{6}(14-12) = \frac{2}{6} = \frac{1}{3}}$$

$$F_X(x) = \int_8^x \frac{1}{6} dx = \frac{1}{6}[x]_8^x = \frac{1}{6}(x-8) \Rightarrow F_X(x) = \begin{cases} 0 & si & x < 8 \\ \frac{1}{6}(x-8) & si & 8 \leq x < 14 \\ 1 & si & x \geq 14 \end{cases}$$

$$\Downarrow$$

$$\boxed{P(X > 12) = 1 - P(X < 12) = 1 - F_X(12) = 1 - \frac{1}{6}(12-8) = 1 - \frac{4}{6} = \frac{2}{6} = \frac{1}{3}}$$

Si fem la integral de la funció de densitat, obtenim la funció de distribució. Si derivem la funció de distribució, obtenim la funció de densitat.

$$\Rightarrow \frac{dF_X(x)}{dx} = f(x) = \frac{1}{6}$$

Virgilio Pérez Giménez – virgilio.perez@uv.es

TEMA 5

MODELS DE PROBABILITAT UNIVARIANTS

Part III

Esperança y variància d'una variable aleatòria

Estadística bàsica

Curs 2020/2021

VNIVERSITAT DE VALÈNCIA
Facultat d' Economia

ESTADÍSTICA DESCRIPTIVA

↓

Distribució de freqüències

↓

Calculem descriptius: mesures de posició, de dispersió i de forma

VARIABLES ALEATÒRIES

↓

Distribució de probabilitat

↓

Calculem descriptius: mesures de posició, de dispersió i de forma

Virgilio Pérez Giménez – virgilio.perez@uv.es

➢Com s'obté la mitjana d'una variable aleatòria? Amb l'operador **ESPERANÇA:**

$$E[X] = \mu_x \quad \begin{cases} \sum_{-}^{+} x_i P(X = x_i) & \text{discretes} \\ \int_{-}^{+} x f(x) dx & \text{contínues} \end{cases}$$

$$\bar{x} = \sum_i x_i f_i$$

mitjana aritmètica

➢Com s'obté la variància d'una variable aleatòria? Amb l'operador **ESPERANÇA:**

$$V[X] = \sigma_x^2 = E\left[(X - \mu_x)^2\right] \quad \begin{cases} \sum_{-}^{+} (x_i - \mu_x)^2 P(X = x_i) & \text{discretes} \\ \int_{-}^{+} (x - \mu_x)^2 f(x) dx & \text{contínues} \end{cases}$$

$$S_x^2 = \sum_i (x_i - \bar{x})^2 f_i$$

Què és la mediana?

La **mediana** és el valor de la variable aleatòria que divideix la distribució en dues parts iguals, és a dir, en un 50% de probabilitat per sota del valor i en un altre 50% per sobre.

- L’operador esperança té **3 propietats:**

$$E\left[\sum_i X_i\right]=\sum_i E[X_i]$$ L’esperança de la suma és la suma de l’esperança.

$$E[kX]=kE[X]$$ Una constant ix fora de l’operador esperança.

$$E[k]=k$$ L’esperança d’una constant és la constant mateixa.

- Considerant aquestes 3 propietats, es pot obtenir l’**expressió operativa de la VARIÀNCIA:**

$$\sigma_x^2=E\left[(X-\mu_x)^2\right]=E\left[X^2+\mu_x^2-2X\mu_x\right]=E\left[X^2\right]+\mu_x^2-2\mu_x E[X]=E\left[X^2\right]-\mu_x^2$$

$$S_x^2=\frac{1}{n}\sum_i x_i^2-\bar{x}^2$$

- **Exemple.** Donada la variable aleatòria **“notes en estadística”**:

$$\mu_x=E[X]=1\times 0.48+2\times 0.32+3\times 0.16+4\times 0.04=1.76$$

$$E\left[X^2\right]=1^2\times 0.48+2^2\times 0.32+3^2\times 0.16+4^2\times 0.04=3.84$$

$$\sigma_x^2=V[X]=E\left[X^2\right]-\mu_x^2=3.84-1.76^2=0.7424 \longrightarrow \sigma_x=\sqrt{0.7424}=0.86$$

- **Exemple.** Donada la variable aleatòria "moment en què es rep una trucada en una centraleta telefònica":

$$\mu_x = E[X] = \int_8^{14} x\frac{1}{6}dx = \frac{1}{6}\left[\frac{x^2}{2}\right]_8^{14} = \frac{1}{6}\left(\frac{14^2}{2} - \frac{8^2}{2}\right) = 11$$

$$E\left[X^2\right] = \int_8^{14} x^2\frac{1}{6}dx = \frac{1}{6}\left[\frac{x^3}{3}\right]_8^{14} = \frac{1}{6}\left(\frac{14^3}{3} - \frac{8^3}{3}\right) = 124$$

$$\sigma_x^2 = V[X] = E\left[X^2\right] - \mu_x^2 = 124 - 11^2 = 3 \longrightarrow \sigma_x = \sqrt{3} = 1.73$$

Virgilio Pérez Giménez – virgilio.perez@uv.es

TEMA 6

MODELS DE PROBABILITAT UNIVARIANTS ESPECÍFICS

Part I

Models discrets

Estadística bàsica

Curs 2020/2021

VNIVERSITAT DE VALÈNCIA
Facultat d' Economia

Virgilio Pérez Giménez – virgilio.perez@uv.es

Models específics discrets

Bernoulli, binomial i Poisson

Models específics continus

Uniforme, exponencial i normal

Bernoulli: variable aleatòria associada a un experiment amb dos resultats possibles :

$X = 1$ → èxit → P(X = 1) = p

$X = 0$ → fracàs → P(X = 0) = 1-p

Funció de quantia: $P(X = x_i) = p^{x_i}(1-p)^{1-x_i}$

$$E[X] = \sum_i x_i P(X = x_i) = 1 \cdot p + 0 \cdot (1-p) = p \quad \rightarrow \quad E[X] = p$$

$$E[X^2] = E[X] = p \quad \rightarrow \quad V[X] = E[X^2] - (E[X])^2 = p - p^2 \quad \rightarrow \quad V[X] = p(1-p)$$

$X \sim B[p]$ Significa que X segueix un model Bernoulli de paràmetre "p"

❑ Per exemple:

- S'extrauen peces a l'atzar d'un sistema continu de fabricació. Es classifiquen les peces en defectuoses o no defectuoses.
- Si se selecciona a l'atzar un estudiant d'estadística, té o no té aprovada Matemàtiques I?

Binomial: variable aleatòria associada a un experiment que consisteix en una successió de "n" intents o assajos idèntics, de manera que:

En cada intent només són possibles 2 resultats, **èxit o fracàs**, amb probabilitats "p" i "1-p", respectivament.

La probabilitat "p" es manté **constant** d'un intent a un altre i, en conseqüència, també "1-p".

Els intents o assajos són **independents**.

La variable aleatòria es defineix com el **nombre d'èxits en "n" proves**, en què:

X = 0, 1, 2, …., n

Funció de quantia: $P(X = x_i) = \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}$

$E[X] = np$

$V[X] = np(1-p)$

$X \sim Bi[n,p]$ Significa que X segueix un model binomial de paràmetres "n" i "p".

❑ Exemples:

- El nombre de peces defectuoses en una caixa de 10 peces.
- De 20 reclamacions presentades en una hora en l'Oficina del Consumidor un dia qualsevol, el nombre de reclamacions fetes al sector de la telefonia.
- La Direcció General de Trànsit defineix "setmana blanca" com aquella setmana en què no hi ha accidents de trànsit amb víctimes mortals i considera un període de 10 setmanes, el nombre de setmanes blanques.
- Nombre de períodes d'una hora en què no es recullen telefonades en una centraleta telefònica, en una jornada laboral de 8 hores.
- Etcètera.

❑ Suposem que en el primer exemple $\boldsymbol{X \sim Bi[n = 30, p = 0.3]}$:

$$P(X = 1) = \binom{10}{1} 0.3^1 0.7^9 = \frac{10!}{1!9!} 0.3^1 0.7^9 = 0.0403$$

$$P(X = x_i) = \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

$$P(X < 2) = P(X = 0) + P(X = 1) = \binom{10}{0} 0.3^0 0.7^{10} + \binom{10}{1} 0.3^1 0.7^9 = 0.1493$$

Poisson: variable aleatòria discreta i **no finita**, que representa la **quantitat de successos o ocurrències en un determinat interval de temps o espai**, complint les propietats següents:

1. La probabilitat d'una ocurrència és igual en dos intervals qualssevol d'igual amplitud.
2. L'ocurrència en un interval és independent de l'ocurrència en qualsevol altre interval.

$X = 0, 1, 2, \ldots.$

Funció de quantia: $$P(X = x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$E[X] = V[X] = \lambda$$

$X \sim P[\lambda]$ Significa que X segueix un model Poisson de paràmetre λ.

❑ Exemples:

- El nombre de telefonades rebudes en una centraleta telefònica durant una hora.
- El nombre de cotxes que passen per un determinat lloc.
- El nombre de defectes en 10 metros de tela.

❑ Suposem que en el primer exemple $X \sim P[\lambda = 3]$:

$$P(X = 0) = e^{-3}\frac{3^0}{0!} = e^{-3} = 0.0498$$

$$P(X = x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.0498 - e^{-3}\frac{3^1}{1!} = 0.8008$$

λ

Bernoulli

$X \sim B[p]$

$X = 0\,,1$

$P(X = x_i) = p^{x_i}(1 - p)^{\ x_i}$

$\mu_x = E[X] = p$

$\sigma_x^2 = V[X] = p(1 - p)$

Binomial

$X \sim Bi[n, p]$

$X = 0\,,1, \dots, n$

$P(X = x_i) = \binom{n}{x_i} p^{x_i}(1 - p)^{\ n - x_i}$

$\mu_x = E[X] = np$

$\sigma_x^2 = V[X] = np(1 - p)$

Poisson

$X \sim P[\lambda]$

$X = 0\,,1, \dots$

$P(X = x_i) = e^{-\lambda}\dfrac{\lambda^{x_i}}{x_i!}$

$\mu_x = E[X] = \lambda$

$\sigma_x^2 = V[X] = \lambda$

TEMA 6

MODELS DE PROBABILITAT UNIVARIANTS ESPECÍFICS

Part II

Models continus

Estadística bàsica

Curs 2020/2021

Virgilio Pérez Giménez – virgilio.perez@uv.es

Vniversitat de València
Facultat d' Economia

Virgilio Pérez Giménez – virgilio.perez@uv.es

Models específics discrets

Bernoulli, bInomial i Poisson

Models específics continus

Uniforme, exponencial i normal

Uniforme: la **funció de densitat** és constant al llarg de tot el camp de variació de la variable:

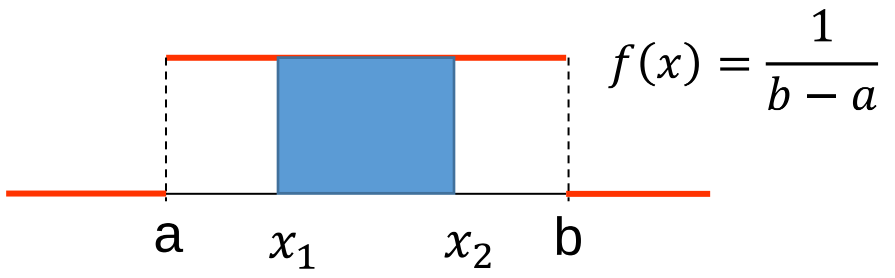
$$\int_a^b k\,dx = 1 \;\rightarrow\; k[x]_a^b = 1 \;\rightarrow\; k(b-a) = 1$$

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & \text{resta} \end{cases}$$

Qualsevol variable amb probabilitat igual en intervals iguals serà uniforme.

La probabilitat en un interval $[x_1, x_2]$, utilitzant la funció de densitat, es calcula:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} \frac{1}{b-a}\,dx = \frac{1}{b-a}[x]_{x_1}^{x_2} = \frac{1}{b-a}(x_2 - x_1)$$



$$X \sim U[a, b]$$

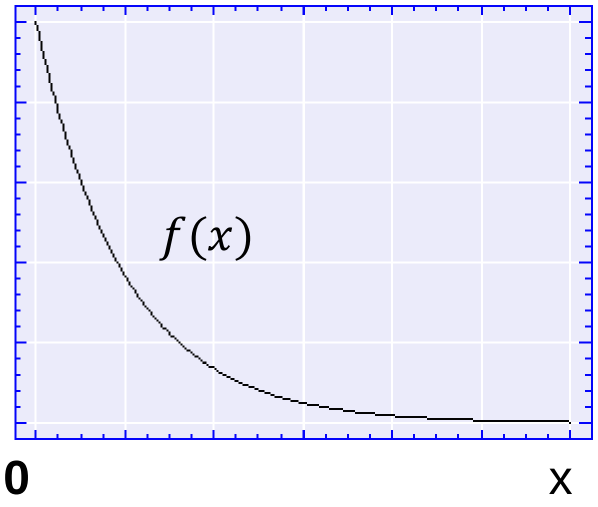
La variable aleatòria X segueix un model uniforme de paràmetres “a” i “b”.

$$E[X] = \frac{a+b}{2} \qquad V[X] = \frac{(a-b)^2}{12}$$

Exponencial: representa un comportament de probabilitat **decreixent** exponencialment, és a dir, decreix molt ràpidament.

Exemples: temps de duració d'una bateria; temps de tramitació d'un expedient administratiu; temps que un graduat tarda a trobar la seua primera ocupació, etcètera.

En general, aquestes variables representen temps d'espera, perquè, **a mesura que avança el temps, és menys probable que continuem esperant**.



$$f(x) = \begin{cases} \theta e^{-\theta x}, & x > 0 \\ 0, & \text{en la resta} \end{cases}$$

$$E[X] = \mu_x = \frac{1}{\theta}$$

$$V[X] = \sigma_x^2 = \frac{1}{\theta^2}$$

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{1}{\theta} x}, & x > 0 \\ 0, & \text{en la resta} \end{cases}$$

$$E[X] = \mu_x = \theta$$

$$V[X] = \sigma_x^2 = \theta^2$$

$X \sim Ex[\theta]$ La variable aleatòria X segueix un model exponencial de paràmetre θ.

$$F(x) = P(X \le x) = \int_0^x \theta e^{-\theta x} dx = \left[-e^{-\theta x}\right]_0^x = -e^{-\theta x} + 1$$

$$F(x) = \begin{cases} 0 & , \ x \le 0 \\ 1 - e^{-\theta x} & , \ x > 0 \\ 1 & , \ x \to \infty \end{cases}$$

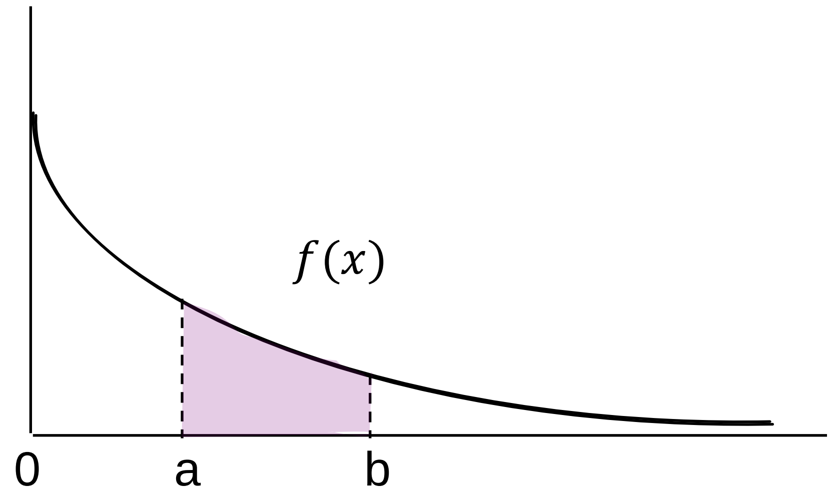
Anàlogament:

$$F(x) = \begin{cases} 0 & , \ x \le 0 \\ 1 - e^{-\frac{1}{\theta} x} & , \ x > 0 \\ 1 & , \ x \to \infty \end{cases}$$

La probabilitat en un interval es podrà calcular utilitzant la funció de densitat o la funció de distribució de la manera següent:

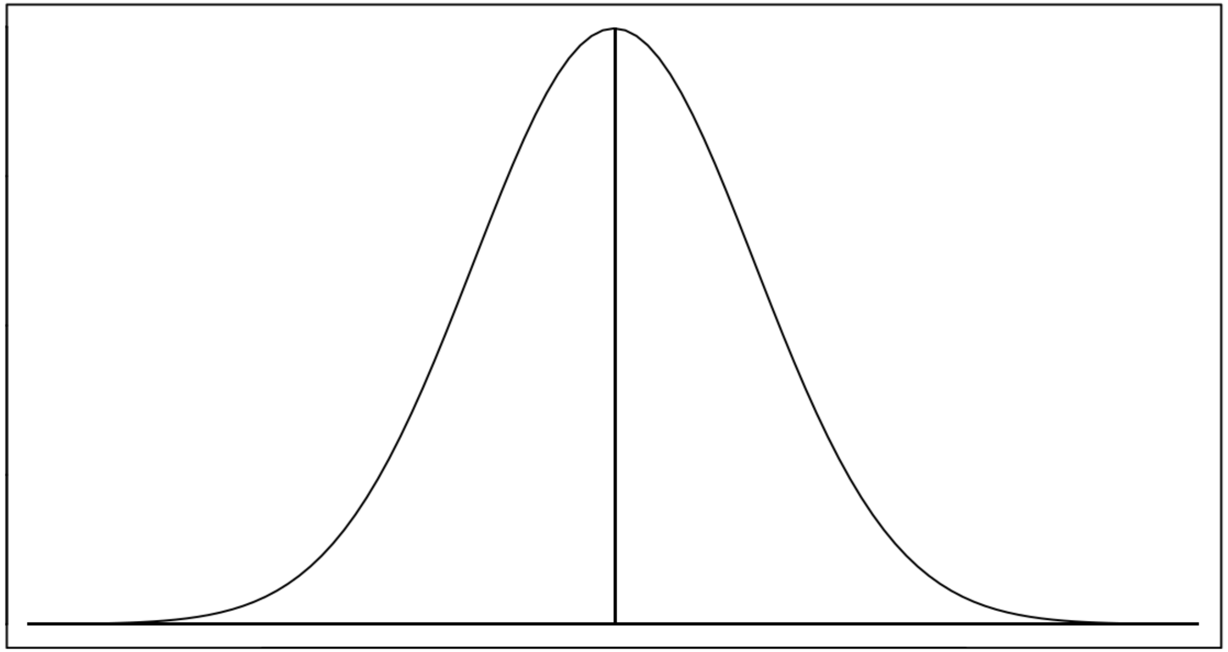
$$P(a < X < b) = \int_a^b \theta e^{-\theta x} dx = \left[-e^{-\theta x}\right]_a^b = e^{-\theta a} - e^{-\theta b}$$

$$P(a < X < b) = F(b) - F(a) = (1 - e^{-\theta b}) - (1 - e^{-\theta a}) = e^{-\theta a} - e^{-\theta b}$$



$X \sim E[\theta]$ Significa que X segueix un model exponencial de paràmetre θ.

Normal: és un model que concentra la major part de probabilitat al voltant de la mitjana de la variable aleatòria, i descendeix la probabilitat a mesura que ens n'allunyem, és a dir, els valors menys probables són els extrems.



$X \sim N[\mu_x, \sigma_x]$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Propietat de la distribució normal: tota transformació lineal d'una variable aleatòria que segueix un model normal, també és normal:

Si $X \sim N[\mu_x, \sigma_x^2]$ e $Y = a + bX$ llavors $Y \sim N[\mu_y, \sigma_y^2]$ sent:

$$\mu_y = E[Y] = E[a + bX] = a + bE[X] = a + b\mu_x$$

$$\sigma_y^2 = V[Y] = V[a + bX] = b^2\sigma_x^2$$

Cas particular: **TIPIFICACIÓ**

$$Z = \frac{X - \mu_x}{\sigma_x} \quad , \quad Z \sim N[\mu_z, \sigma_z^2] \quad \rightarrow \quad \left.\begin{array}{l} \mu_z = \dfrac{1}{\sigma_x}\mu_x - \dfrac{\mu_x}{\sigma_x} = 0 \\[2ex] \sigma_z^2 = \dfrac{1}{\sigma_x^2}\sigma_x^2 = 1 \end{array}\right\}$$

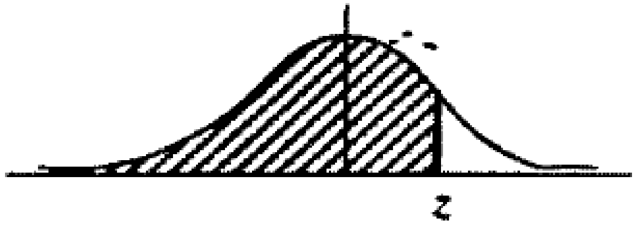
$$\mathbf{Z \sim \boldsymbol{N}[0, 1]}$$

Z és una variable aleatòria **normal** reduïda o **tipificada**.

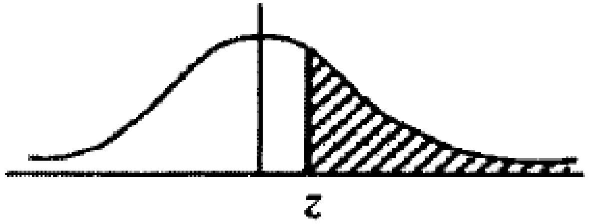
$Z \sim N[0,1]$ està **tabulada**; la imatge següent és un fragment de la taula:

TAULA DE LA DISTRIBUCIÓ NORMAL TIPIFICADA N[0, 1]

$F(z) = P(Z \leq z)$



$1 - F(z) = P(Z > z)$



z	$F(z)$	$1-F(z)$	z	$F(z)$	$1-F(z)$	z	$F(z)$	$1-F(z)$
0.00	0.5000	0.5000	**0.50**	0.6915	0.3085	**1.00**	0.8413	0.1587
0.01	0.5040	0.4960	**0.51**	0.6950	0.3050	**1.01**	0.8437	0.1563
0.02	0.5080	0.4920	**0.52**	0.6985	0.3015	**1.02**	0.8461	0.1539
0.03	0.5120	0.4880	**0.53**	0.7019	0.2981	**1.03**	0.8485	0.1515
0.04	0.5160	0.4840	**0.54**	0.7054	0.2946	**1.04**	0.8508	0.1492
0.05	0.5199	0.4801	**0.55**	0.7088	0.2912	**1.05**	0.8531	0.1469
0.06	0.5239	0.4761	**0.56**	0.7123	0.2877	**1.06**	0.8554	0.1446
0.07	0.5279	0.4721	**0.57**	0.7157	0.2843	**1.07**	0.8577	0.1423

Per al càlcul de qualsevol probabilitat sobre una variable normal, sempre haurem de **TIPIFICAR** primerament:

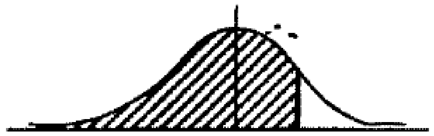
- **Exemple.** Siga $X \sim N[\mu_x = 10, \sigma_x = 2]$, calcula les probabilitats següents:

$$P(X > 11) = P\left(Z > \frac{11-10}{2}\right) = P(Z > 0.5) = 0.3085$$

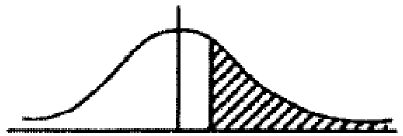
$$P(X < 9) = P\left(Z < \frac{9-10}{2}\right) = P(Z < -0.5) = P(Z > 0.5) = 0.3085$$

TAULA DE LA DISTRIBUCIÓ NORMAL TIPIFICADA N[0, 1]

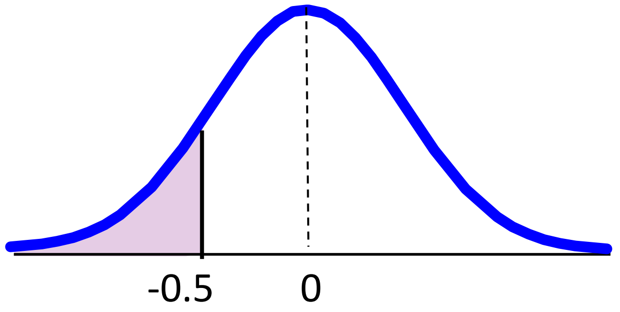
$F(z) = P(Z \le z)$

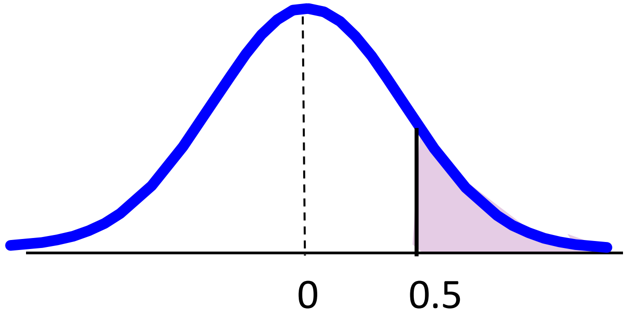


$1 - F(z) = P(Z > z)$



z	F (z)	1–F (z)	z	F (z)	1–F(z)	z	F (z)	1–F (z)
0.00	0.5000	0.5000	**0.50**	0.6915	0.3085	**1.00**	0.8413	0.1587
0.01	0.5040	0.4960	**0.51**	0.6950	0.3050	**1.01**	0.8437	0.1563
0.02	0.5080	0.4920	**0.52**	0.6985	0.3015	**1.02**	0.8461	0.1539
0.03	0.5120	0.4880	**0.53**	0.7019	0.2981	**1.03**	0.8485	0.1515
0.04	0.5160	0.4840	**0.54**	0.7054	0.2946	**1.04**	0.8508	0.1492
0.05	0.5199	0.4801	**0.55**	0.7088	0.2912	**1.05**	0.8531	0.1469
0.06	0.5239	0.4761	**0.56**	0.7123	0.2877	**1.06**	0.8554	0.1446
0.07	0.5279	0.4721	**0.57**	0.7157	0.2843	**1.07**	0.8577	0.1423





$$X \sim U[a,b)]$$

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & \text{resta} \end{cases}$$

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

$$V[X] = \frac{(a-b)^2}{12}$$

$$X \sim Ex[\theta]$$

$$f(x) = \begin{cases} \theta e^{-\theta x}, & x > 0 \\ 0, & \text{en la resta} \end{cases}$$

$$F(x) = \begin{cases} 0, & x \le 0 \\ 1 - e^{-\theta x}, & x > 0 \\ 1, & x \to \infty \end{cases}$$

$$E[X] = \mu_x = \frac{1}{\theta}$$

$$V[X] = \sigma_x^2 = \frac{1}{\theta^2}$$

$$X \sim N[\mu_x, \sigma_x]$$

$$Z \sim N[0,1]$$

$$f(x) = \frac{1}{\sigma_x \sqrt{2\Pi}} e^{-\frac{1}{2\sigma_x^2}(x-\mu_x)^2}$$

$$Z = \frac{X - \mu_x}{\sigma_x}$$