Curs de Bioestadística

Jaume Aguadé

UAB Universitat Autònoma de Barcelona

DEPARTAMENT DE **MATEMÀTIQUES**

Programa de l'assignatura

- Estadística descriptiva
  - Variables i freqüència
  - Representacions gràfiques
  - Mesures de tendència central i mesures de dispersió
  - Regressió lineal
- Fonaments de probabilitat
  - Conceptes bàsics de la probabilitat
  - Probabilitat condicionada. Independència.
- Variables aleatòries
  - Variables aleatòries discretes i contínues. Esperança i variància.
  - Distribucions binomial, Poisson, uniforme, exponencial, normal.
- Inferència estadística
  - Mostra i població. Estimació de paràmetres.
  - Intervals de confiança.
  - Tests d'hipòtesis.
  - Anàlisi de la variància.

Bibliografia recomanada

- **X. Bardina i M. Farré:** *Estadística descriptiva*. Manuals UAB, 2009.
- **R. Delgado:** *Probabilidad y estadística con aplicaciones*. 2018.

Estadística

És la ciència matemàtica que s'ocupa de la recollida i tractament de dades per tal de resumir-les, analitzar-les, extreure'n conclusions i arribar a poder fer prediccions. El cas més important és aquell en que aquestes dades presenten incertesa i/o es consideren aleatòries (="degudes a l'atzar").

Estadística descriptiva

S'ocupa de resumir les dades, representar-les gràficament, utilitzar indicadors descriptius (mitjana,...) etc.
No intervé la probabilitat.
Exemple: Un cens.

Inferència estadística

Pretén extrapolar les dades conegudes d'una mostra a tota una població molt més gran.
La probabilitat hi juga un paper clau.
Exemple: Una enquesta.

- Població, individus.
- Mostra. $N =$ mida de la mostra.
- Variables X, Y, ...
- Estadística univariant, estadística bivariant, estadística multivariant.
- Arrodoniment, notació científica. Percentatges.
- Sumatoris $(\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n)$

Variables

- numèriques o quantitatives
  - contínues: Temps, longitud, pes,...
  - discretes: Nombre d'individus, cèl·lules, votants d'una opció,...
- no numèriques o qualitatives: sexe, color dels ulls,...

Escales de mesura

- escala nominal.
- escala ordinal.
- escala numèrica.

Atenció

- Codificar amb xifres variables nominals o ordinals no les converteix en numèriques.
- Les mesures físiques com temps, pes, longitud, etc. sempre són contínues, encara que, evidentment, utilitzarem una certa unitat discreta per a mesurar-les (dies, grams, metres, etc.).

X una variable sobre una població. $x_1, \ldots, x_k$ els diferents valors que pren la variable X sobre una certa mostra de mida N.

Freqüència: Nombre de vegades que es dóna un cert valor de la variable, a la mostra.

- n_i = Freqüència absoluta del valor x_i
- f_i = Freqüència relativa del valor x_i. $f_i = n_i/N$

Taula de freqüències: Escrivim cada valor de X i la seva freqüència.

Amb una variable nominal la taula de freqüències és l'única cosa que podem fer-hi.

$$\sum_{i=1}^{k} n_i = N, \quad \sum_{i=1}^{k} f_i = 1$$

Es pregunta a 500 individus d'una mostra ($N = 500$) quin operador de telefonia mòbil utilitzen. Amb les respostes, fem aquesta taula de freqüències:

operador	n_i	f_i	p_i
movistar	281	0.562	56.2%
masmovil	18	0.036	3.6%
orange	75	0.150	15.0%
vodafone	118	0.236	23.6%
euskaltel	6	0.012	1.2%
altres	2	0.004	0.4%
	500	1.000	100%

Aquí X és una variable nominal que pren els valors
$X \in \{\text{movistar, masmovil, orange, vodafone, euskaltel, altres}\}$

Si els valors de X estan ordenats (escala ordinal o escala numèrica), podem considerar freqüències acumulades:

- $N_i = \sum_{r=1}^{i} n_r$. Freqüència acumulada.
- $F_i = \sum_{r=1}^{i} f_r = N_i/N$. Freqüència relativa acumulada.

$$n_1 = N_1 \leq N_2 \leq \cdots \leq N_k = N, \quad f_1 = F_1 \leq F_2 \leq \cdots \leq F_k = 1$$

Representem en una taula de freqüències el nombre de fills que han tingut 200 dones d'una certa mostra:

fills	n_i	N_i	F_i
0	13	13	0.07
1	72	85	0.43
2	91	176	0.88
3	12	188	0.94
més de 3	12	200	1
	200		

Quan X és contínua (o bé discreta amb molts valors), els valors de X s'agrupen en intervals:

Estructura d'edats dels habitants de Catalunya el 2018

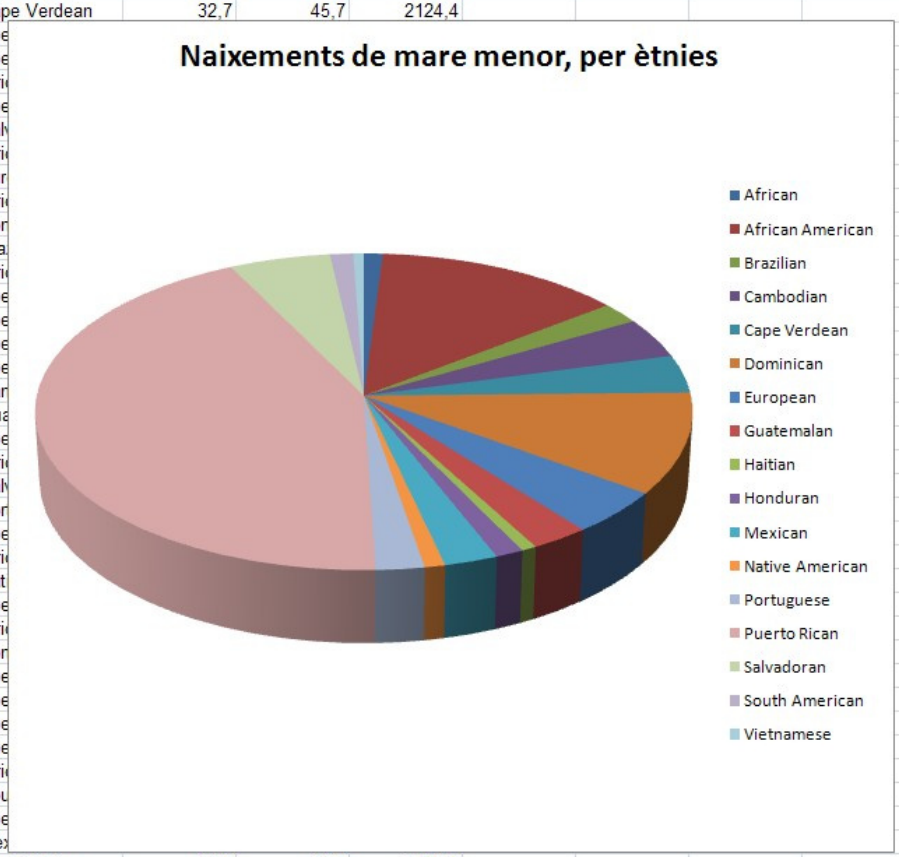
	I_i	n_i	f_i (%)	ℓ_i	x_i
de 0 a 24 anys	$[0, 25)$	$1,702,746$	26.2	25	12.5
de 25 a 49 anys	$[25, 50)$	$2,581,607$	39.8	25	37.5
de 50 a 74 anys	$[50, 75)$	$1,697,962$	26.2	25	62.5
75 anys i més	$[75, \infty)$	$507,537$	7.8		

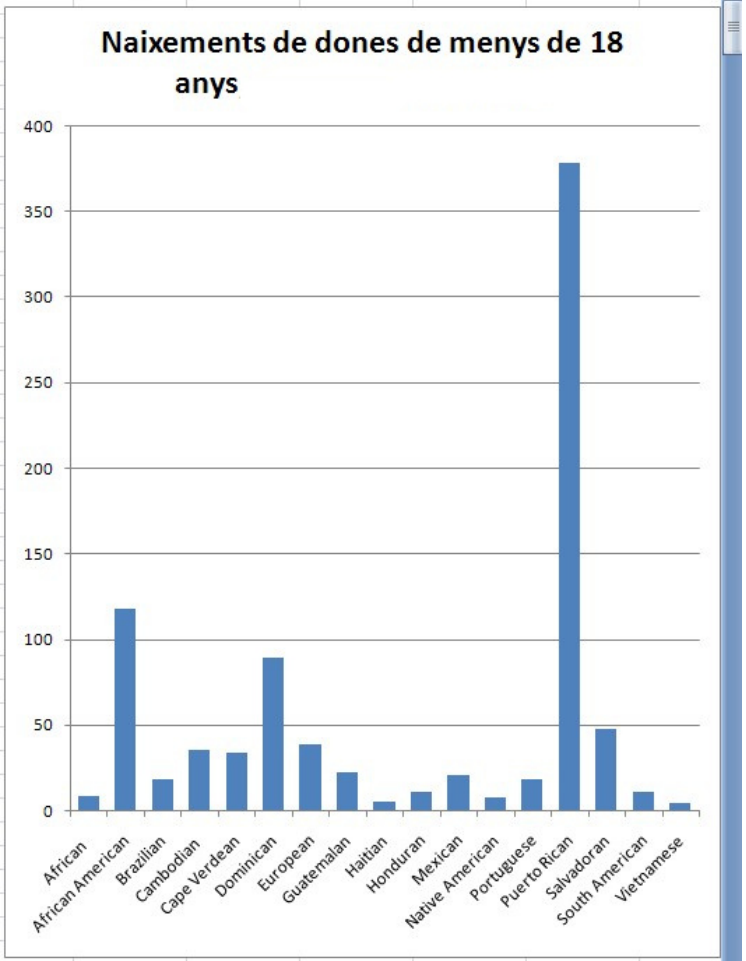
- intervals: $[L_i, L_{i+1})$.
- longitud dels intervals: $\ell_i = L_{i+1} - L_i$.
- marca de classe: $x_i = \dfrac{L_{i+1} + L_i}{2}$.

Massachusetts.xlsx [Modo de compatibilidad] - Microsoft Excel

ètnia	P	L	W
Puerto Rican	33,7	48,5	1836,3
Salvadoran	35,5	50,4	1932,5
Cambodian	33,7	48,9	1946,2
Cape Verdean	32,7	45,7	2124,4
Puerto Rican	33,8	48,2	2552,5
Puerto Rican	33,2	46,9	2555,4
Puerto Rican	36,1	53,3	2573,2
Puerto Rican	33,7	48,6	2581,4
Dominican	34,6	49,9	2582,0
African American	33,5	48,3	2588,8

Cuenta de ètnia	
ètnia	Total
African	9
African American	118
Brazilian	19
Cambodian	36
Cape Verdean	34
Dominican	90
European	39
Guatemalan	23
Haitian	6
Honduran	11
Mexican	21
Native American	8
Portuguese	19
Puerto Rican	378
Salvadoran	48
South American	11
Vietnamese	5
Total general	875





Histograma d'Excel, incorrecte

Columna1	
Media	4,147540984
Error típico	0,247100002
Mediana	3,5
Moda	3
Desviación estándar	2,729308728
Varianza de la muestra	7,449126135
Curtosis	-0,762999931
Coeficiente de asimetría	0,395357005
Rango	10
Mínimo	0
Máximo	10
Suma	506
Cuenta	122

[0,5)	4
[5,7)	6
[7,9)	8
[9,10]	10

Clase	*Frecuencia*
4	69
6	28
8	21
y mayor...	9

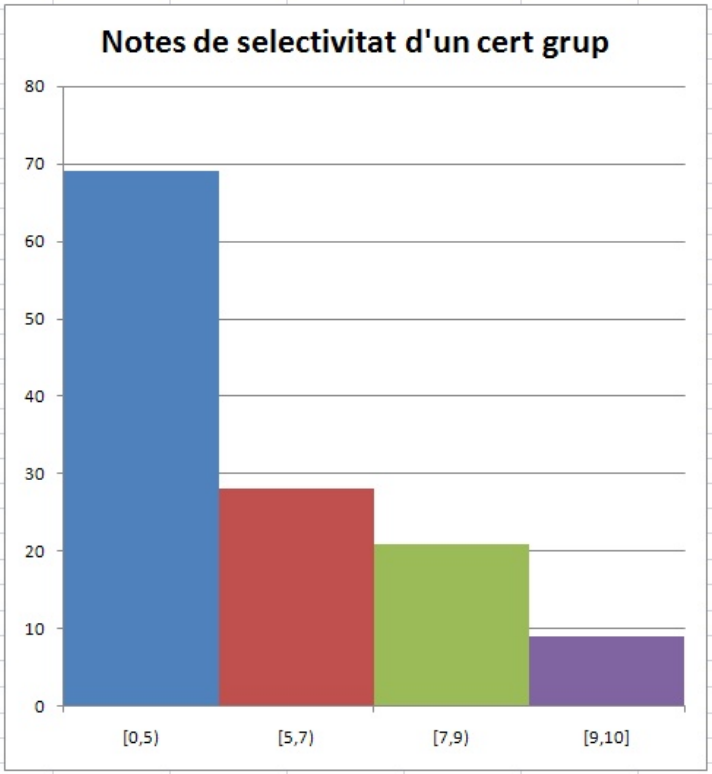


Diagrama de barres

- S'utilitza per a variables nominals.
- L'alçada de les barres és proporcional a la freqüència (absoluta o relativa).
- L'amplada és indiferent i ha de ser la mateixa per a totes les barres.
- No és lícit truncar l'eix vertical.

Histograma

- S'utilitza per a variables numèriques agrupades per intervals.
- La superfície dels rectangles és proporcional a la freqüència (absoluta o relativa). (Important si els intervals són d'amplades diferents).
- L'amplada de cada rectangle és la de l'interval corresponent. Els rectangles es toquen.
- No és lícit truncar l'eix vertical.

Cal saber fer:

- Taules de freqüència, a mà i amb un full de càlcul.
- Dibuixar, a mà i amb un full de càlcul, diagrames de barres i histogrames.

Exemple

En un estudi sobre els moviments naturals de les colònies d'una certa espècie d'aus, s'han obtingut aquests resultats sobre la distància (en km) entre colònies:

65.8	69.4	69.4	69.7	71.5	72.2	74.1	75.4	75.8	76.3
77.2	77.6	77.6	77.9	78.3	78.8	78.9	81.2	81.2	81.7
82.3	82.3	82.4	84.5	84.7	85.2	85.4	88.2	90.2	92.3

Agrupeu la mostra en intervals i construïu la taula de freqüències corresponent. Dibuixeu un histograma. Feu-ho a mà i amb un full de càlcul.

Moda (Mo)

- És el valor de freqüència màxima.
- Si treballem amb dades agrupades per intervals (de longitud igual): interval modal.
- Es pot utilitzar en tot tipus de variables, també nominals.
- Pot haver-hi més d'una moda (distribucions unimodals, bimodals,...)

És la suma de tots els valors, dividida pel nombre de valors:

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

- Només té sentit en l'escala numèrica.
- Hi intervenen tots els valors de X i està molt influïda pels valors extrems.
- Si les dades estan agrupades en intervals, utilitzarem les marques de classe x_i per obtenir un valor aproximat de $\overline{X}$.
- Comportament quan agrupem dues poblacions: una població de mida N_1 amb mitjana $\overline{X}_1$ i una població de mida N_2 amb mitjana $\overline{X}_2$. Si les agrupem, la mitjana total serà

$$\overline{X} = \frac{\overline{X}_1 N_1 + \overline{X}_2 N_2}{N_1 + N_2} \neq \frac{\overline{X}_1 + \overline{X}_2}{2} \quad \text{(en general)}$$

Més propietats de la mitjana

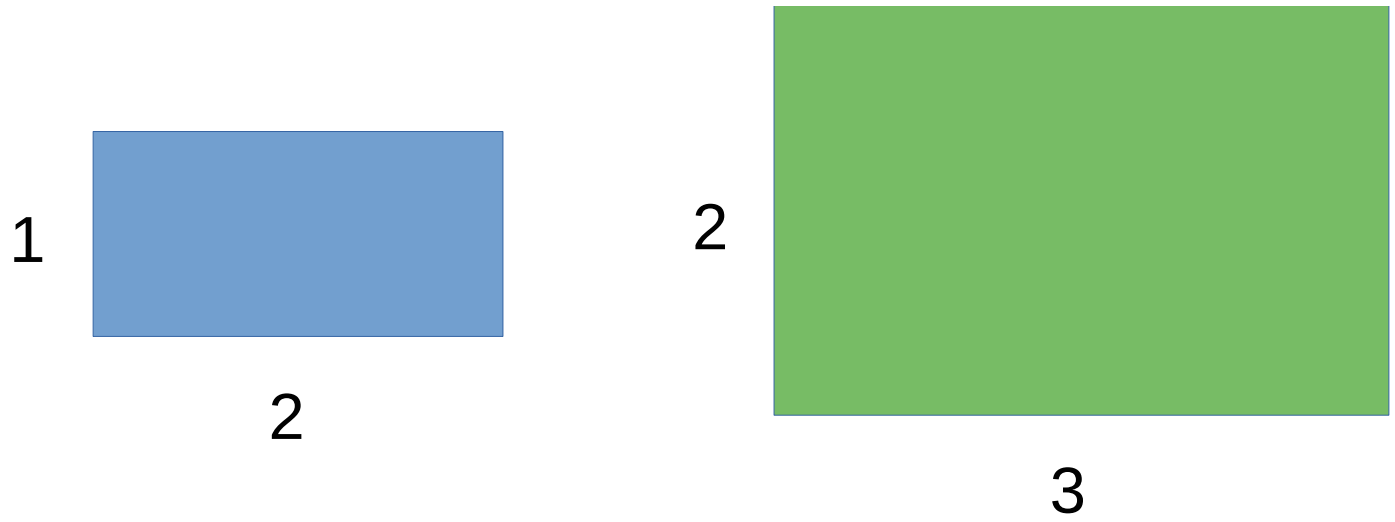
- La mitjana té les mateixes unitats que la variable.
- Si sumem dues variables, les mitjanes es sumen:

$$\overline{X+Y} = \overline{X} + \overline{Y}$$

- Si multipliquem la variable per una constant (exemple: canvi d'unitat de mesura), la mitjana queda multiplicada per aquesta mateixa constant:

$$\overline{aX} = a\,\overline{X}$$

- La mitjana és lineal: $\overline{aX+bY} = a\,\overline{X} + b\,\overline{Y}$
- Si a una variable li sumem una constant, a la mitjana també se li suma aquesta mateixa constant: $\overline{X+a} = \overline{X} + a$
- Però la mitjana no és multiplicativa: $\overline{X\,Y} \neq \overline{X}\ \overline{Y}$, en general.



$$X = \{2, 3\},\ \overline{X} = 2.5; \quad Y = \{1, 2\},\ \overline{Y} = 1.5; \quad \overline{X}\,\overline{Y} = 3.75$$

$$XY = \{2, 6\},\ \overline{XY} = 4.$$

Cal saber fer:

- Calcular la mitjana a mà, amb calculadora i amb un full de càlcul.
- Calcular la mitjana quan les dades estan agrupades per intervals.
- Calcular la mitjana a partir d'una taula de freqüències.

Exemple

Calculeu, a partir de la taula de freqüències, la mitjana del nombre de fills de les dones d'una certa mostra:

fills	n_i	N_i	f_i
0	13	13	0.07
1	72	85	0.38
2	91	176	0.48
3	12	188	0.06

La **Mediana (Md)** és aquell valor que deixa la meitat de les observacions per sota i l'altra meitat per sobre.

Exemples

- La mediana de 100, 27, 51, 3, 30 és **30**.
- La mediana de 3, 7, 3, 3, 5, 3, 9 és **3**.
- La mediana de 5, 2, 3, 5, 5, 2 és **4**. (Si el nombre d'observacions és parell, prenem la mitjana dels dos valors centrals.)
- Si tenim una taula de freqüències, també podem calcular la mediana fàcilment:

x_i	n_i	N_i
0	7	7
1	4	11
2	3	14
3	1	15
4	1	16

Md=1

x_i	n_i	N_i
0	7	7
1	4	11
2	3	14
3	1	15
4	1	16
5	6	22

Md=1.5

Càlcul de la mediana quan les dades estan agrupades en intervals

1. Igual que abans, es troba l'interval que conté la mediana $[L_i, L_{i+1}]$.
2. Es fa una interpolació lineal per trobar un punt d'aquest interval que sigui una bona estimació de la mediana:

$$Md = L_i + \frac{\ell_i(\frac{N}{2} - N_{i-1})}{n_i}$$

Generalitzacions de la mediana:

- Quartils: Q_1, $Q_2 = Md$, Q_3.
- Decils: $D_1, \ldots, D_9$.
- Centils: $C_1, \ldots, C_{99}$.

Exemple

Hem estudiat el temps de coagulació de la sang de 59 pacients i tenim la taula de freqüències en que el temps de coagulació (en minuts) s'ha agrupat per intervals.

temps	n_i	N_i	
$[0, 5)$	3	3	
$[5, 8)$	16	19	
$[8, 11)$	21	40	$\Leftarrow$ la mediana és aquí
$[11, 14)$	13	53	
$[14, \infty)$	6	59	

$$Md = 8 + \frac{3 \times (29.5 - 19)}{21} = 9.5$$

Propietats de la mediana

- Té sentit en l'escala ordinal.
- No depèn dels valors, sinó només de la seva ordenació.
- No es veu afectada pels valors extrems de la variable (en contra del que li passa a la mitjana).
- La mediana i la mitjana no s'han de confondre. Poden ser ben diferents entre sí. La diferència entre la mitjana i la mediana ens dóna informació sobre l'assimetria de la distribució.
- En moltes ocasions, s'invoca la mitjana quan el valor que ens interessa realment és la mediana. Per exemple, si demà en Bill Gates decidís venir a viure a Cerdanyola, la renda mitjana dels habitants de Cerdanyola creixeria molt, però la mediana de la renda no variaria.

Variància i desviació típica.

$$S^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{N} \qquad S = \sqrt{S^2}$$

També s'utilitzen la variància i la desviació típica corregides:

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i}{N-1} \qquad s = \sqrt{s^2}$$

A les calculadores, σ_n i σ_{n-1}.

$$S^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \overline{X})^2 n_i = \frac{1}{N}\sum_{i=1}^{k}(x_i^2 n_i - 2x_i\overline{X}n_i + \overline{X}^2 n_i) =$$

$$\frac{1}{N}\sum_{i=1}^{k} x_i^2 n_i - 2\overline{X}^2 + \overline{X}^2 = \overline{X^2} - \overline{X}^2$$

Propietats de la variància i la desviació típica

- Són nombres positius (o zero). S té les mateixes unitats que la variable.
- Si N és gran, hi ha poca diferència entre els valors corregits i sense corregir.
- Si a una variable li sumem una constant, la variància no varia: $\text{Var}(X + a) = \text{Var}(X)$.
- Si multipliquem una variable per una constant, la variància queda multiplicada pel quadrat d'aquesta constant: $\text{Var}(aX) = a^2 \text{Var}(X)$.
- La variància no és lineal. $\text{Var}(X + Y) \neq \text{Var}(X) + \text{Var}(Y)$ (en general).
- La variància d'una constant és zero.

Cal saber calcular la variància i la desviació típica amb calculadora i amb full de càlcul.

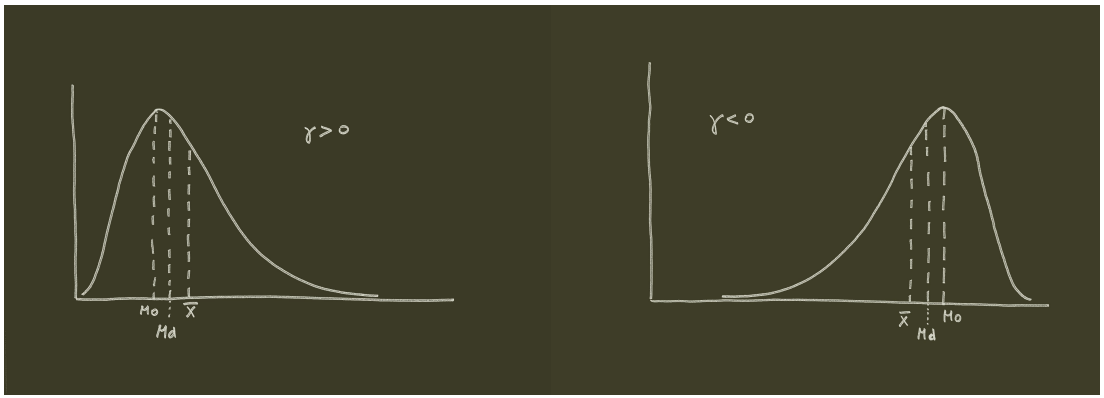
Altres mesures de dispersió

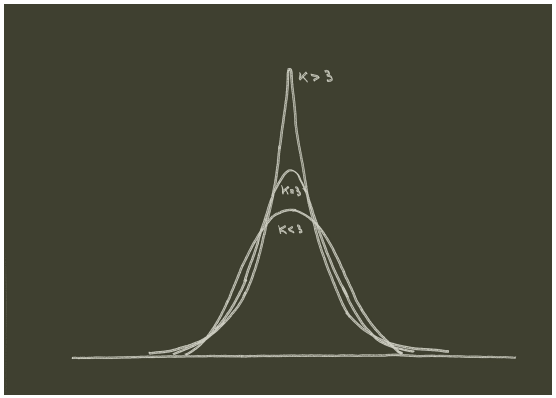
- El rang, diferència entre els valors màxim i mínim de la variable.
- El rang interquartílic, diferència entre Q_3 i Q_1.
- El coeficient de variació: $CV = S/\overline{X}$ (només per a variables positives).
- Els moments i els moments centrals

$$m_r = \frac{1}{N}\sum x_i^r, \quad \mu_r = \frac{1}{N}\sum (x_i - \overline{X})^r$$

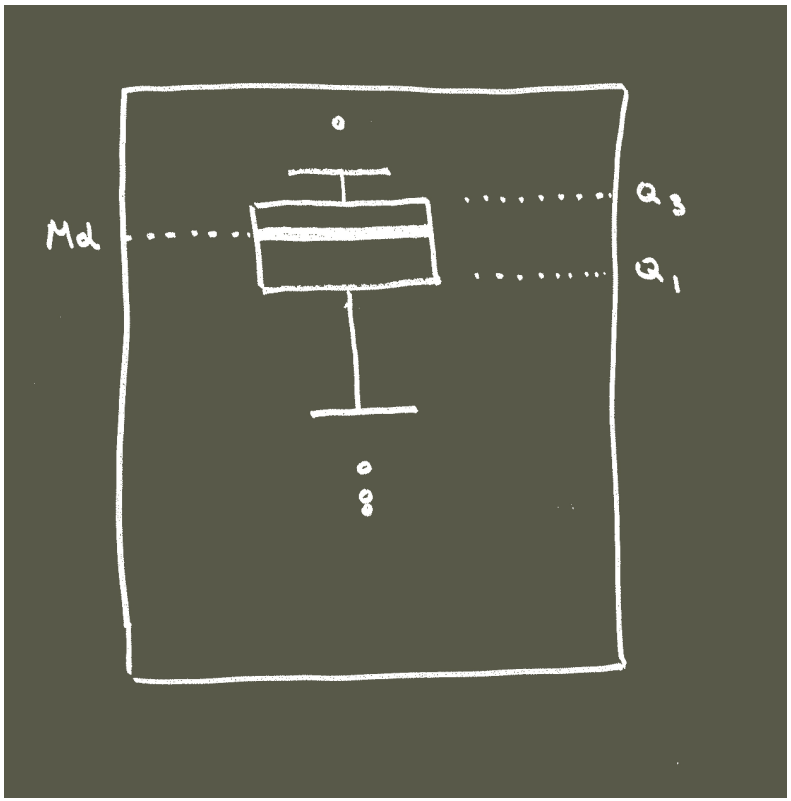
  Observem que $m_1 = \overline{X}$, $\mu_1 = 0$ i $\mu_2 = S^2$.
- El coeficient d'asimetria (de Fisher): $\gamma = \mu_3/S^3$. (Variància no nul·la.)
- El coeficient de curtosi (de Fisher): $K = \mu_4/S^4$. (Variància no nul·la.)

Ronald Fisher
1890–1962
biòleg i estadístic





La distribució normal té asimetria zero (és simètrica) i curtosi $K = 3$.



- L'escala de mesura de la variable és a l'eix vertical.
- La caixa central està delimitada pels quartils Q_1 i Q_3.
- La línia gruixuda a l'interior de la caixa és la mediana.
- Els petits cercles denoten outliers: observacions situades a més de 1.5 vegades el rang interquartílic de les vores de la caixa.
- Els "bigotis" o "patilles" de la caixa es dibuixen en l'última dada situada a menys de 1.5 vegades el rang interquartílic de les vores de la caixa.

Estadística bivariant

Dues variables X i Y referides a una mateixa població.

X pren valors $x_1, \ldots, x_k$, Y pren valors $y_1, \ldots, y_l$. Per a cada individu, tenim dos valors (X, Y). Podem estudiar cada variable per separat:

- Freqüències n_{ij}, freqüències marginals $n_{i\bullet}$, $n_{\bullet j}$.
- (Variables numèriques) mitjanes $\overline{X}$, $\overline{Y}$ i variàncies S_X^2, S_Y^2

També podem presentar les dades en forma de taula de contingència. Exemple: comarca (nominal) i afectació per processionària (ordinal).

	nul·la	lleu	notable	intensa	
Berguedà	60	18	57	21	156
Solsonès	50	10	41	30	131
	110	28	98	51	287

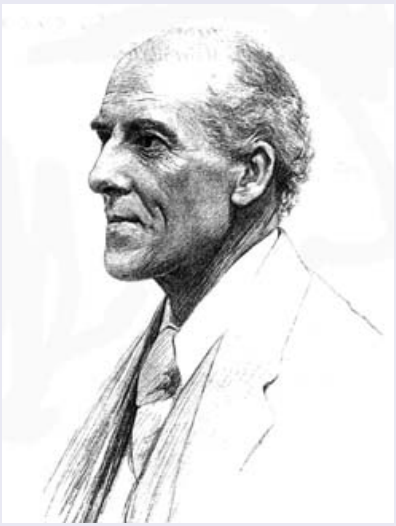
Podem dibuixar també un núvol de punts o diagrama de dispersió. (Practiqueu-ho amb un full de càlcul.)

La covariància

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum (x_i - \overline{X})(y_j - \overline{Y})n_{ij} = \overline{XY} - \overline{X}\ \overline{Y}$$

- Necessita l'escala numèrica i té unitats.
- Pot ser positiva o negativa.
- És simètrica respecte de X i Y: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
- La covariància mesura la no commutativitat de "mitjana" i "producte".
- Cal saber fer: calcular la covariància amb la calculadora i amb full de càlcul.

Karl Pearson 1857–1936

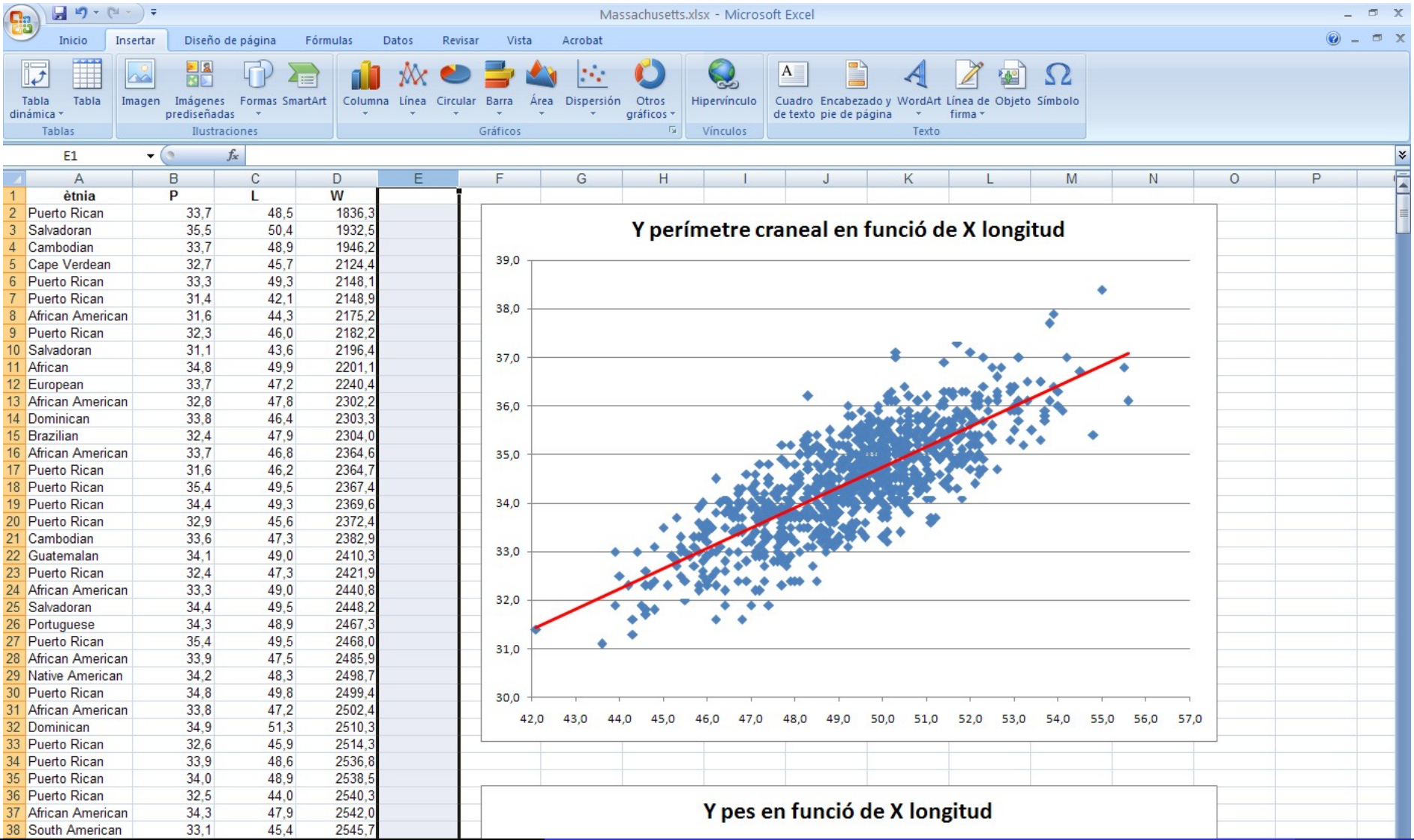


El coeficient de correlació de Pearson

$$r_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{S_X S_Y}$$

- És un nombre sense unitats, entre -1 i 1: $-1 \leq r_{X,Y} \leq 1$.
- És simètric respecte de X i Y: $r_{X,Y} = r_{Y,X}$.
- $r \gg 0$ indica una relació directa entre X i Y. $r \ll 0$ indica una relació inversa entre X i Y.
- $-0.5 < r < 0.5$: correlació lineal feble.
- r proper a ± 1: bona correlació lineal.
- Coeficient de determinació: $r^2_{X,Y}$. Indica quina part de la variabilitat de Y s'explica per la variabilitat de X.

Si el coeficient de correlació ens diu que hi ha una correlació lineal acceptable entre X i Y, podem aproximar el núvol de punts per una recta.



	ètnia	P	L	W
2	Puerto Rican	33,7	48,5	1836,3
3	Salvadoran	35,5	50,4	1932,5
4	Cambodian	33,7	48,9	1946,2
5	Cape Verdean	32,7	45,7	2124,4
6	Puerto Rican	33,3	49,3	2148,1
7	Puerto Rican	31,4	42,1	2148,9
8	African American	31,6	44,3	2175,2
9	Puerto Rican	32,3	46,0	2182,2
10	Salvadoran	31,1	43,6	2196,4
11	African	34,8	49,9	2201,1
12	European	33,7	47,2	2240,4
13	African American	32,8	47,8	2302,2
14	Dominican	33,8	46,4	2303,3
15	Brazilian	32,4	47,9	2304,0
16	African American	33,7	46,8	2364,6
17	Puerto Rican	31,6	46,2	2364,7
18	Puerto Rican	35,4	49,5	2367,4
19	Puerto Rican	34,4	49,3	2369,6
20	Puerto Rican	32,9	45,6	2372,4
21	Cambodian	33,6	47,3	2382,9
22	Guatemalan	34,1	49,0	2410,3
23	Puerto Rican	32,4	47,3	2421,9
24	African American	33,3	49,0	2440,8
25	Salvadoran	34,4	49,5	2448,2
26	Portuguese	34,3	48,9	2467,3
27	Puerto Rican	35,4	49,5	2468,0
28	African American	33,9	47,5	2485,9
29	Native American	34,2	48,3	2498,7
30	Puerto Rican	34,8	49,8	2499,4
31	African American	33,8	47,2	2502,4
32	Dominican	34,9	51,3	2510,3
33	Puerto Rican	32,6	45,9	2514,3
34	Puerto Rican	33,9	48,6	2536,8
35	Puerto Rican	34,0	48,9	2538,5
36	Puerto Rican	32,5	44,0	2540,3
37	African American	34,3	47,9	2542,0
38	South American	33,1	45,4	2545,7

La recta de regressió de Y respecte de X és:

$$y = bx + a; \quad b = \frac{\text{Cov}(X, Y)}{S_X^2}, \quad a = \overline{Y} - b\overline{X}$$

Propietats de la recta de regressió

- És la recta que minimitza la diferencia entre el valor observat de Y per un X donat i el valor de Y calculat amb l'equació de la recta.
- Passa pel punt $(\overline{X}, \overline{Y})$ i té pendent igual a b.
- Només té valor predictiu per a valors de X propers a $\overline{X}$ i valors de r propers a ± 1.
- Cal triar quina variable prenem com a independent: La recta de regressió de Y respecte de X no és la mateixa que la recta de regressió de X respecte de Y.

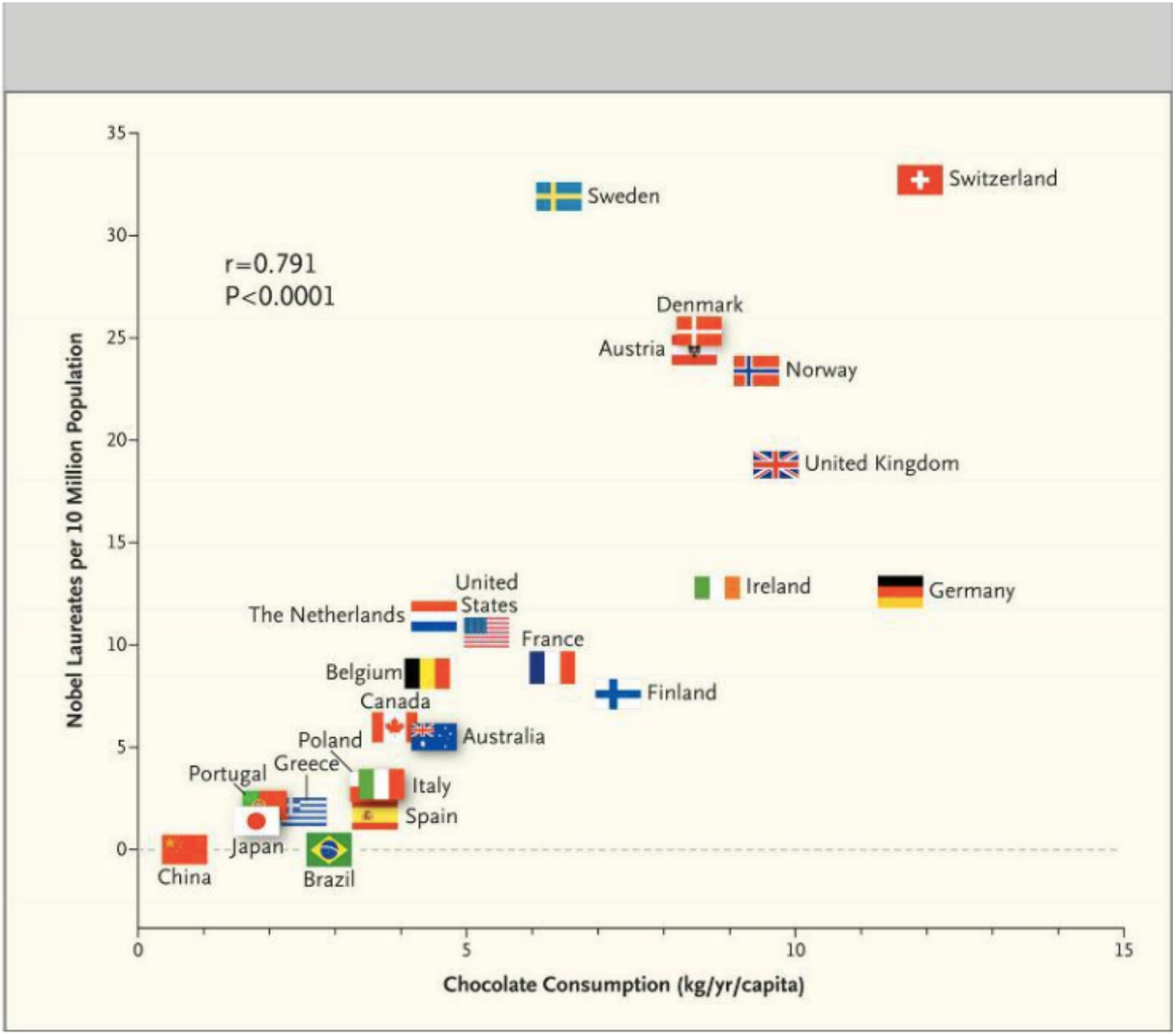
En un estudi clàssic es va estudiar la correlació entre el consum de tabac i el càncer de pulmó els anys 60 als USA. La mostra és de 15 estats. X = consum mitjà de cigarretes per habitant i any, en milers. Y = morts per càncer de pulmó per cent-mil habitants i any. Els resultats observats van ser aquests:

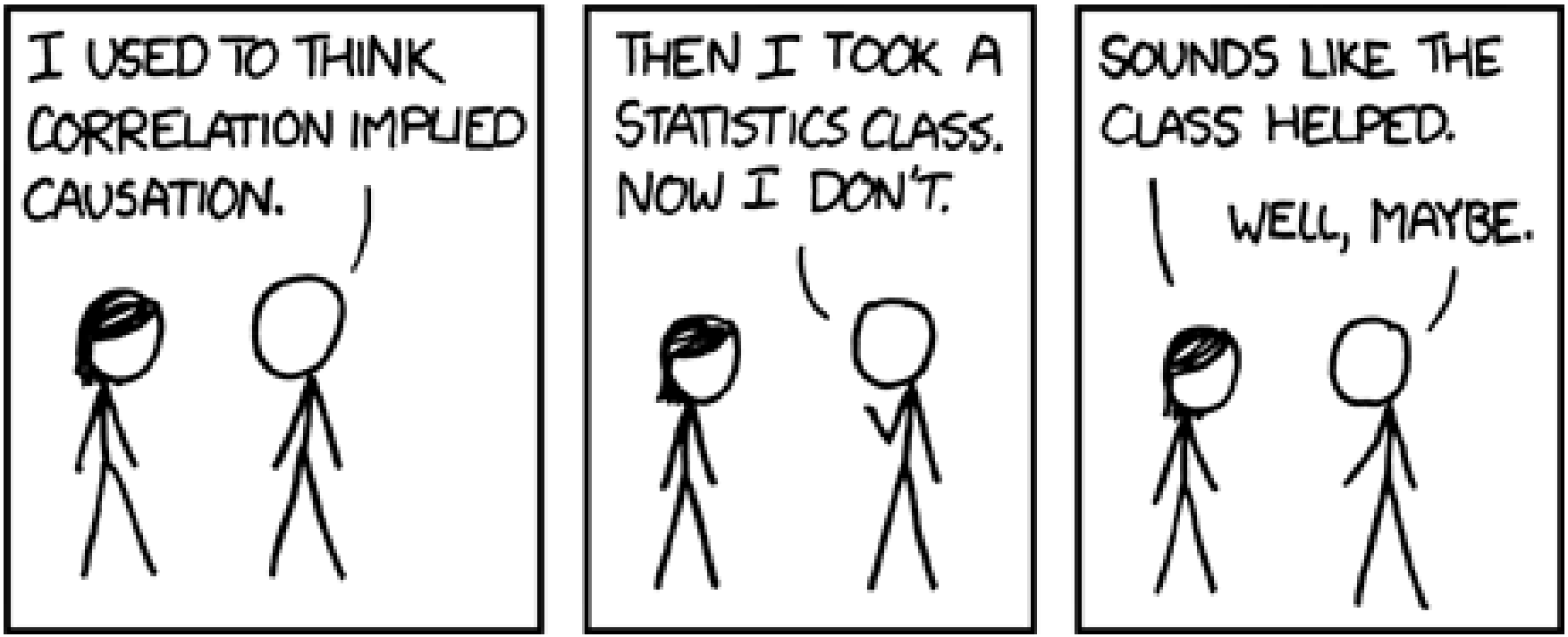
X	1.8	2.6	1.8	2.9	3.1	3.4	4.0	2.8
Y	17.0	19.8	16.0	22.1	22.8	24.6	27.3	23.6
X	2.0	2.8	2.6	2.2	2.2	2.3	2.2	
Y	13.6	22.8	20.3	16.6	16.8	17.7	25.5	

- Podeu afirmar que el nombre de morts per càncer de pulmó depèn linealment del consum de tabac? (Sí, perquè $r = 0.81$)
- Si en un estat el consum de tabac fos de 2.5, quin nombre de morts per càncer de pulmó podríem esperar?($y = 20.0$)
- Si en un estat el nombre de morts per càncer de pulmó és de 21.5, quin consum de tabac podem esperar? ($x = 2.71$)

Resoleu-ho amb calculadora i amb full de càlcul.

Correlació no implica causalitat


r=0.791
P<0.0001
Nobel Laureates per 10 Million Population
Chocolate Consumption (kg/yr/capita)
0
5
10
15
20
25
30
35
Sweden
Switzerland
Denmark
Austria
Norway
United Kingdom
Ireland
Germany
United States
The Netherlands
France
Belgium
Finland
Canada
Australia
Poland
Greece
Portugal
Italy
Spain
Japan
China
Brazil


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Curs de Bioestadística

Jaume Aguadé

Capítol 2: Probabilitat

UAB Universitat Autònoma de Barcelona

DEPARTAMENT DE **MATEMÀTIQUES**

- Els experts estimen en un 60% la probabilitat de trobar nous jaciments de petroli a la costa de Tarragona.
- Amb 6 daus és més probable una doble parella que una parella.
- La probabilitat de pluja demà, a Cerdanyola, és del 30%.
- Crec que tinc una probabilitat del 50% d'aprovar aquest test.
- És molt probable que aquest cap de setmana em quedi a casa.
- Si fumes, és probable que tinguis problemes cardíacs.
- No és impossible, però és molt improbable.
- L'atzar no existeix: tot està determinat.
- Tot és fruit de l'atzar, tot és aleatori. (Random)
- És molt probable que la causa de l'extinció dels dinosaures fos la caiguda d'un meteorit.
- Si una dona té dos fills, hi ha tres possibilitats: dos nois, dues noies o un de cada. Per tant, la probabilitat de tenir-ne un de cada és un terç. (O no??)
- Quina és la probabilitat que demà es faci de dia? I que el Barça guanyi la lliga la temporada que ve?

Dues interpretacions fonamentals de probabilitat:

- Freqüència dels diversos resultats d'un experiment aleatori.
- Grau de confiança que un fet es produeixi, a partir de l'evidència de què es disposa (Thomas Bayes, 1702–1761).

Experiment aleatori

És un experiment ben definit, indefinidament repetible, del qual coneixem quins són els resultats possibles, però no podem predir quin serà el resultat que es donarà en cada nova execució de l'experiment.

Probabilitat matemàtica: idea general

En un experiment aleatori, la probabilitat d'un cert resultat és la freqüència d'aquest resultat en el límit, quan repetim l'experiment un nombre de vegades que tendeix a infinit.

- Experiment aleatori
- Espai mostral Ω: el conjunt de tots els possibles resultats de l'experiment.
- Un esdeveniment és qualsevol conjunt de resultats possibles.

Exemple

Si llancem un dau, l'espai mostral és

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Un esdeveniment pot ser *"que surti un 3"* o *"que surti un nombre parell"* o *"que no surti ni un 1 ni un 6"*, etc.
L'espai mostral és finit.

Exemple

Llancem una moneda fins que surti *"cara"* i prenem nota de quantes vegades l'hem haguda de llançar. Aquí l'espai mostral és infinit numerable:

$$\Omega = \{1, 2, 3, 4, \ldots, \infty\}.$$

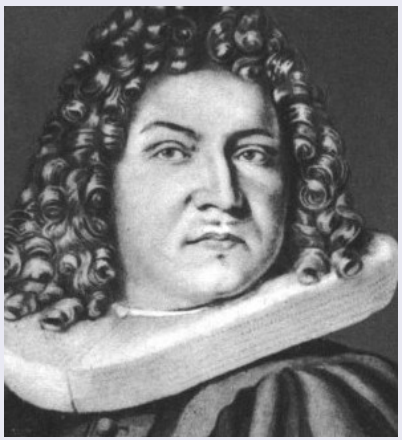
Esdeveniments: *"l'hem llançada entre quatre i sis vegades"*, *"l'hem llançada més de deu vegades"*, *"l'hem llançada un nombre primer de vegades"*, etc.

Tambe ens interessa estudiar experiments aleatoris on l'espai mostral és infinit no numerable. Per exemple:

- Escollim un individu a l'atzar i li mesurem el contingut de glucosa a la sang, en mg/dL. El resultat és un nombre real positiu: $\Omega = [0, \infty)$.
- Mesurem, cada deu segons, la temperatura de l'aire que entra per un cert sistema de ventilació. $\Omega = \mathbb{R}$.

Si Ω és no numerable, hi ha dificultats matemàtiques importants que fan inviable acceptar com a esdeveniment qualsevol conjunt de resultats. Cal que el conjunt de resultats sigui *mesurable* perquè se'l pugui considerar un esdeveniment. A la pràctica, tots els conjunts que trobarem són mesurables.

Jakob Bernoulli
1654–1705



- Conjunts A, B, C,..., elements $x \in A$, $y \notin C$. Subconjunts $A \subset B$.
- $A = \{a, b, c, d, \ldots\}$.
- Dos conjunts són iguals si tenen exactament els mateixos elements.
- $X = \{x \in Y \,|\, x \text{ compleix la propietat } P\}$.
- Conjunt buit: $\varnothing = \{\}$, el conjunt que no té cap element.
- Unió: $A \cup B$. Intersecció: $A \cap B$. $(A \cap B) \subset A \subset (A \cup B)$
- Complement: A^c. $(A^c)^c = A$.
- Lleis de Morgan: $(A \cup B)^c = A^c \cap B^c$. $(A \cap B)^c = A^c \cup B^c$.
- Conjunts disjunts: A i B són disjunts si no tenen cap element en comú, és a dir, si $A \cap B = \varnothing$.
- Unions i interseccions infinites: $\bigcup_{i=1}^{\infty} A_i$, $\bigcap_{i=1}^{\infty} B_i$
- Diagrames de Venn.

Hi ha una correspondència entre el llenguatge de conjunts i el llenguatge d'esdeveniments:

- $A \cup B \iff$ "A o B".
- $A \cap B \iff$ "A i B".
- $A^c \iff$ "no A".
- $\varnothing \iff$ "l'esdeveniment impossible".
- $\Omega \iff$ "l'esdeveniment segur".
- $A \cap B = \varnothing \iff$ "A i B són esdeveniments incompatibles".
- $A = \{a\} \iff$ "A és un esdeveniments elemental"

És a dir, l'esdeveniment impossible és "l'experiment no dóna cap resultat"; l'esdeveniment segur és "l'experiment dóna algun resultat d'entre tots els possibles".
Dos esdeveniments són incompatibles si mai no poden succeir a l'hora. Per exemple, si llancem un dau, els esdeveniments $A =$"surt un nombre parell" i $B =$"surt un nombre més gran que quatre" **no** són incompatibles, mentre que $A =$"surt un nombre parell" i $C =$"surt un cinc" sí que són incompatibles.

Quan els matemàtics no saben (o no volen) definir un concepte, utilitzen el mètode axiomàtic:

Definició axiomàtica de probabilitat

En un experiment aleatori, la probabilitat és una funció que assigna a cada esdeveniment A un nombre real entre 0 i 1 (ambdós inclosos), $P(A)$, anomenat la probabilitat d'aquest esdeveniment. Aquesta funció ha de complir dues propietats:

1. La probabilitat de l'esdeveniment segur és igual a 1: $P(\Omega) = 1$.
2. Si tenim una quantitat finita o numerable d'esdeveniments $A_1, A_2, \ldots$, incompatibles dos a dos,

$$P(\bigcup_i A_i) = \sum_i P(A_i).$$

Es dedueixen fàcilment de les dues propietats bàsiques:

- $P(\text{no } A) = 1 - P(A)$.
- $P(\varnothing) = 0$: l'esdeveniment impossible té probabilitat zero.
- Si $A \subset B$, aleshores $P(A) \leq P(B)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. En particular, si A i B són incompatibles, $P(A \text{ o } B) = P(A) + P(B)$.

- No hi ha cap fórmula que permeti calcular, en general, $P(A \text{ i } B)$ a partir de $P(A)$ i $P(B)$.
- Els diagrames de Venn permeten interpretar la probabilitat com una mena de "àrea".
- fixeu-vos en la distinció entre "Un esdeveniment impossible" i "Un esdeveniment de probabilitat zero". L'esdeveniment impossible té probabilitat zero, però hi pot haver esdeveniments de probabilitat zero que no són impossibles. (Exemple: penseu en l'experiment de llançar una moneda fins que surti "cara".)

La probabilitat d'un 6 quan llancem un dau és 1/6. Per què? Perquè hi ha sis possibles valors i només un m'interessa ("és favorable"). En aquests casos, per calcular la probabilitat n'hi ha prou amb dividir el nombre de casos favorables entre tots els casos possibles. Però aquest resultat només serà cert si estem suposant que el dau és perfecte i, per això, tots els valors són igualment probables. En direm Probabilitat clàssica.

Probabilitat clàssica

Parlarem de Probabilitat clàssica si:

1. Ω és finit. Només un nombre finit de resultats possibles.
2. Tots els esdeveniments elementals tenen la mateixa probabilitat.

Si això passa, la probabilitat d'un esdeveniment A es calcula així:

$$P(A) = \frac{\text{nombre d'elements de } A}{\text{nombre d'elements de } \Omega}.$$

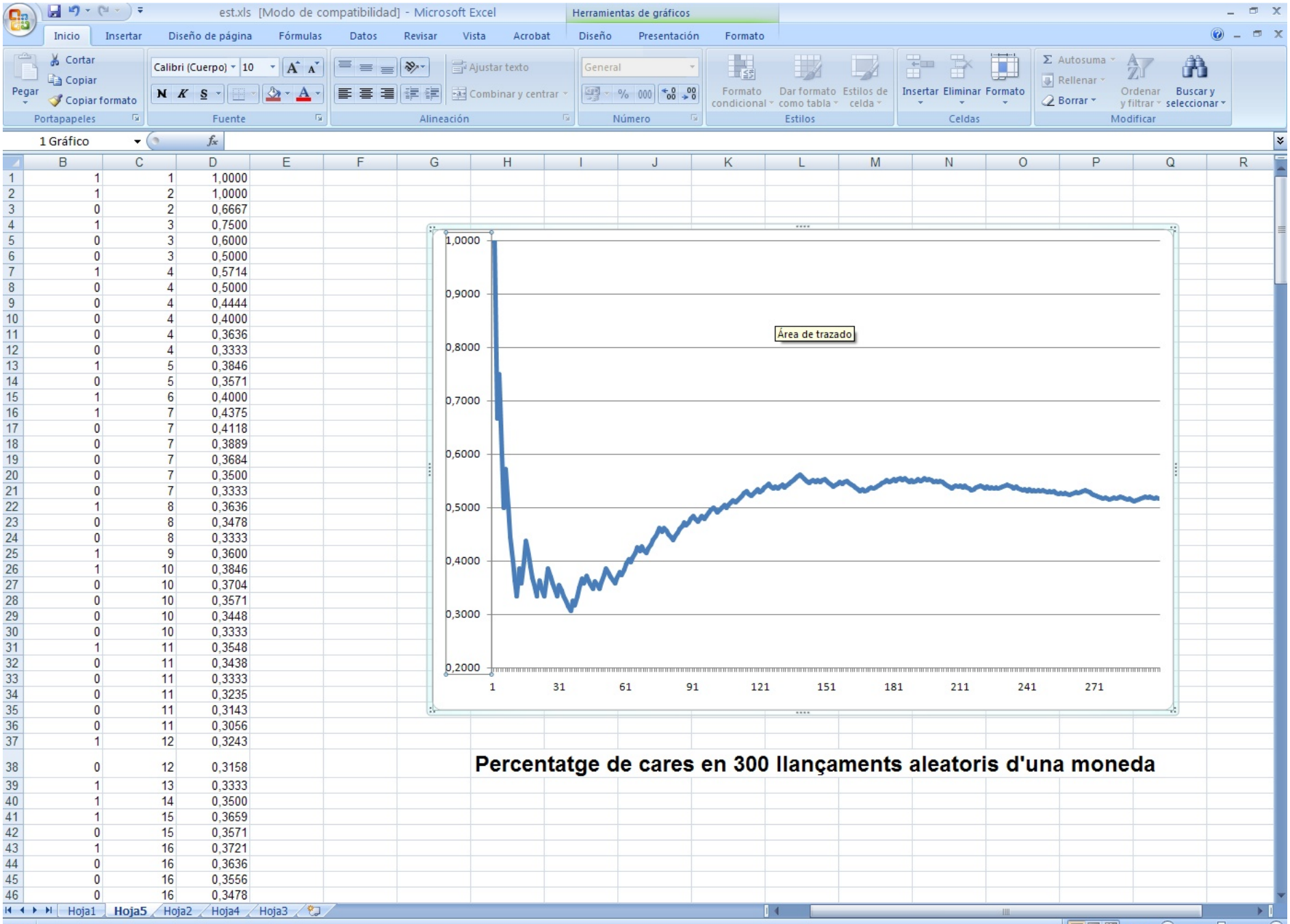
Direm que dos esdeveniments són independents si el fet que un d'ells succeeixi "no influeix" en la probabilitat que l'altre succeeixi.

- Llancem un dau dues vegades. El fet que en el primer llançament surti, diguem, un 3, no influeix en que en el segon llançament surti, diguem, un 5. Els esdeveniments "treure un tres en el primer llançament" i "treure un cinc en el segon llançament" són independents.
- Escollim dues cartes d'una baralla i ens les quedem. L'esdeveniment "treure un as com a primera carta" i l'esdeveniment "treure un tres com a segona carta" no són independents. (Per què?)

Definició matemàtica d'independència

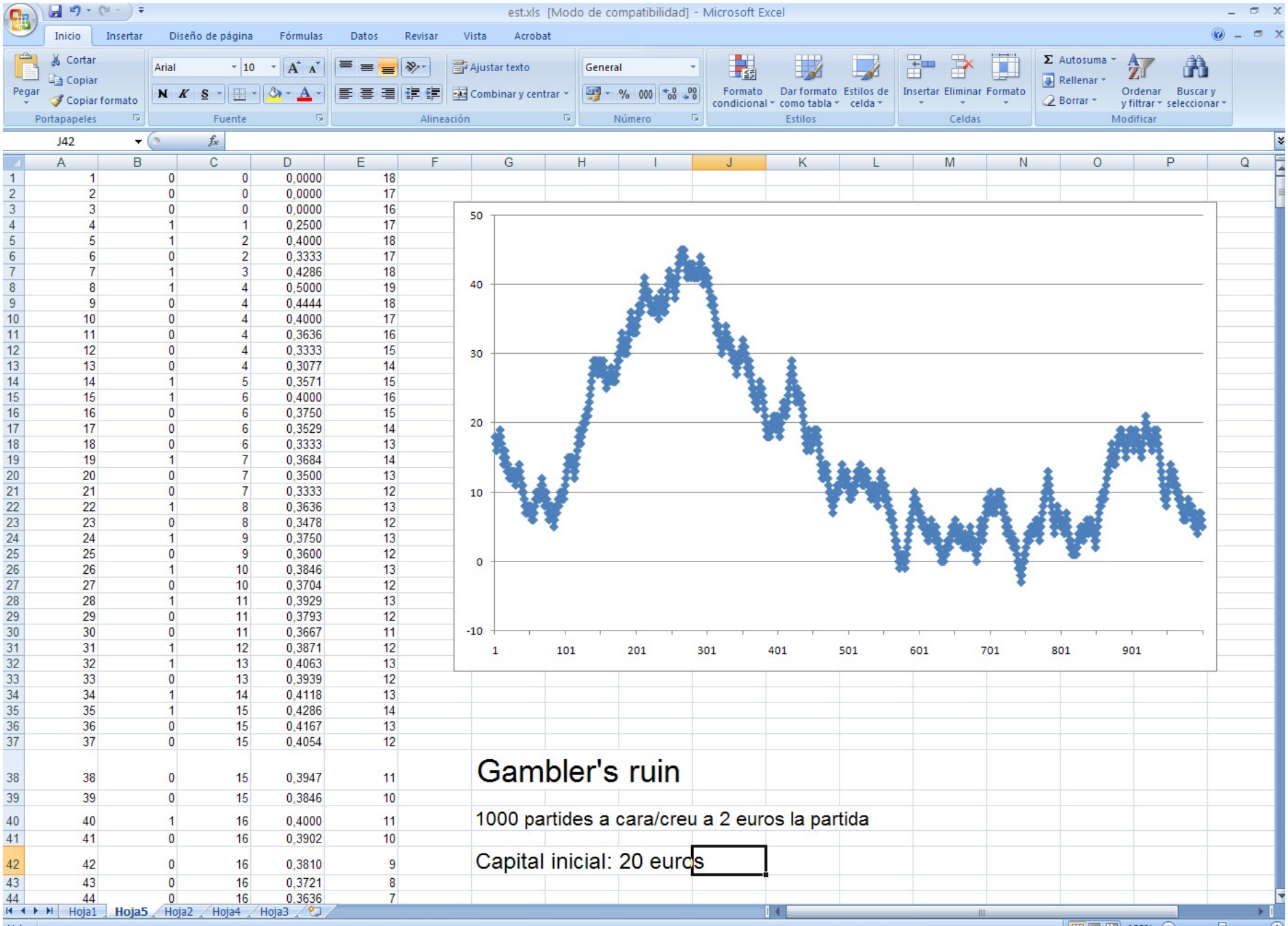
Direm que A i B són independents si $P(A \text{ i } B) = P(A) \cdot P(B)$.

(Penseu per què aquesta definició tradueix la idea intuïtiva d'independència.)



	B	C	D
1	1	1	1,0000
2	1	2	1,0000
3	0	2	0,6667
4	1	3	0,7500
5	0	3	0,6000
6	0	3	0,5000
7	1	4	0,5714
8	0	4	0,5000
9	0	4	0,4444
10	0	4	0,4000
11	0	4	0,3636
12	0	4	0,3333
13	1	5	0,3846
14	0	5	0,3571
15	1	6	0,4000
16	1	7	0,4375
17	0	7	0,4118
18	0	7	0,3889
19	0	7	0,3684
20	0	7	0,3500
21	0	7	0,3333
22	1	8	0,3636
23	0	8	0,3478
24	0	8	0,3333
25	1	9	0,3600
26	1	10	0,3846
27	0	10	0,3704
28	0	10	0,3571
29	0	10	0,3448
30	0	10	0,3333
31	1	11	0,3548
32	0	11	0,3438
33	0	11	0,3333
34	0	11	0,3235
35	0	11	0,3143
36	0	11	0,3056
37	1	12	0,3243
38	0	12	0,3158
39	1	13	0,3333
40	1	14	0,3500
41	1	15	0,3659
42	0	15	0,3571
43	1	16	0,3721
44	0	16	0,3636
45	0	16	0,3556
46	0	16	0,3478

Percentatge de cares en 300 llançaments aleatoris d'una moneda



	A	B	C	D	E
1	1	0	0	0,0000	18
2	2	0	0	0,0000	17
3	3	0	0	0,0000	16
4	4	1	1	0,2500	17
5	5	1	2	0,4000	18
6	6	0	2	0,3333	17
7	7	1	3	0,4286	18
8	8	1	4	0,5000	19
9	9	0	4	0,4444	18
10	10	0	4	0,4000	17
11	11	0	4	0,3636	16
12	12	0	4	0,3333	15
13	13	0	4	0,3077	14
14	14	1	5	0,3571	15
15	15	1	6	0,4000	16
16	16	0	6	0,3750	15
17	17	0	6	0,3529	14
18	18	0	6	0,3333	13
19	19	1	7	0,3684	14
20	20	0	7	0,3500	13
21	21	0	7	0,3333	12
22	22	1	8	0,3636	13
23	23	0	8	0,3478	12
24	24	1	9	0,3750	13
25	25	0	9	0,3600	12
26	26	1	10	0,3846	13
27	27	0	10	0,3704	12
28	28	1	11	0,3929	13
29	29	0	11	0,3793	12
30	30	0	11	0,3667	11
31	31	1	12	0,3871	12
32	32	1	13	0,4063	13
33	33	0	13	0,3939	12
34	34	1	14	0,4118	13
35	35	1	15	0,4286	14
36	36	0	15	0,4167	13
37	37	0	15	0,4054	12
38	38	0	15	0,3947	11
39	39	0	15	0,3846	10
40	40	1	16	0,4000	11
41	41	0	16	0,3902	10
42	42	0	16	0,3810	9
43	43	0	16	0,3721	8
44	44	0	16	0,3636	7

Combinatòria: La ciència de comptar quants elements hi ha en un cert conjunt.

1. **La regla del producte**: Si tinc 5 samarretes, 2 jerseis i 3 pantalons, em puc vestir de $5 \times 2 \times 3 = 30$ maneres diferents.
2. **Permutacions**: De quantes maneres es poden ordenar n objectes? Resposta: $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$ ("factorial de n"). Observació: $0! = 1! = 1$
3. **Nombres combinatoris**: Tenim un conjunt de n elements i n'hem d'escollir k. De quantes maneres ho podem fer? Resposta:

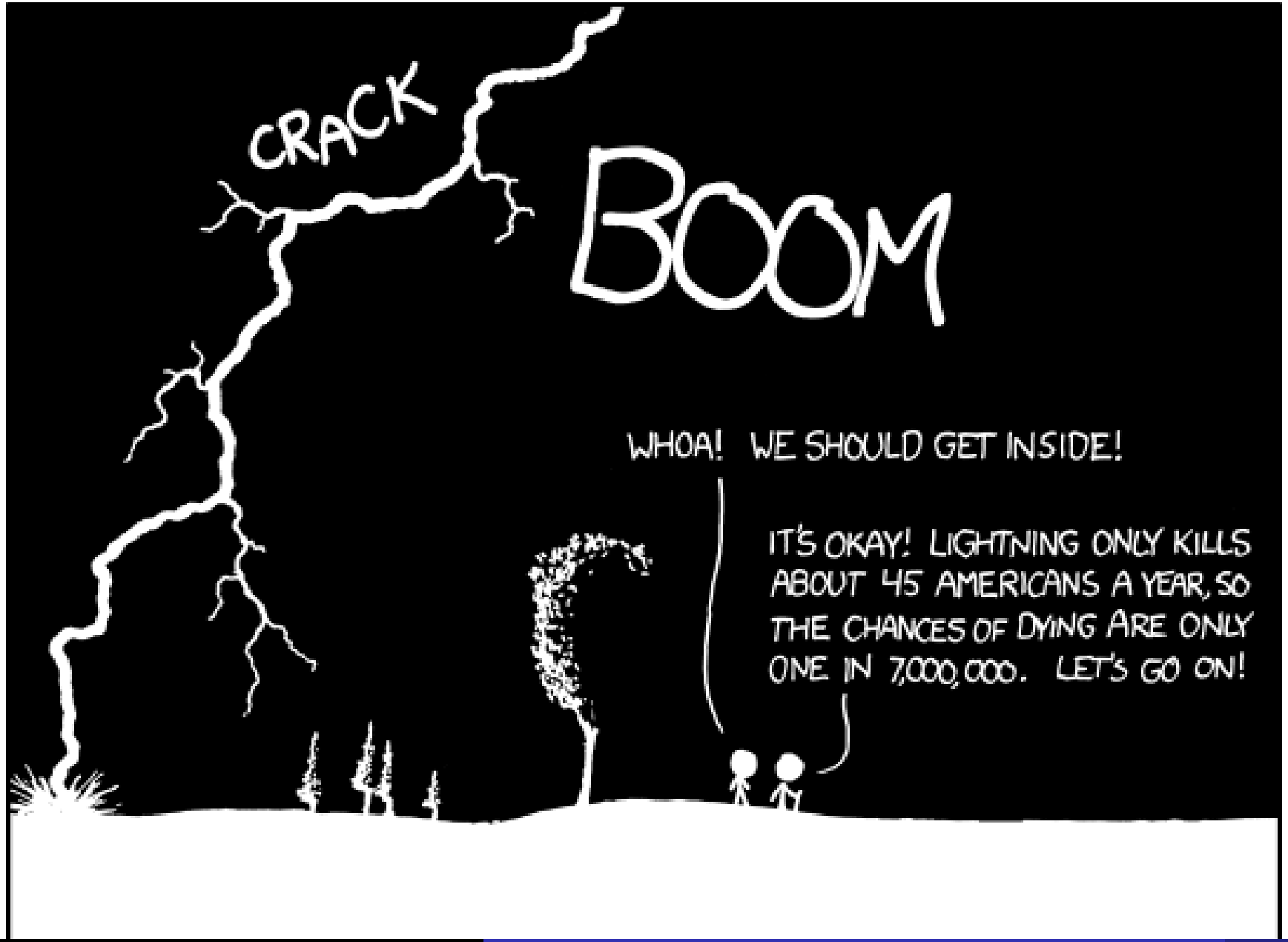
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}$$

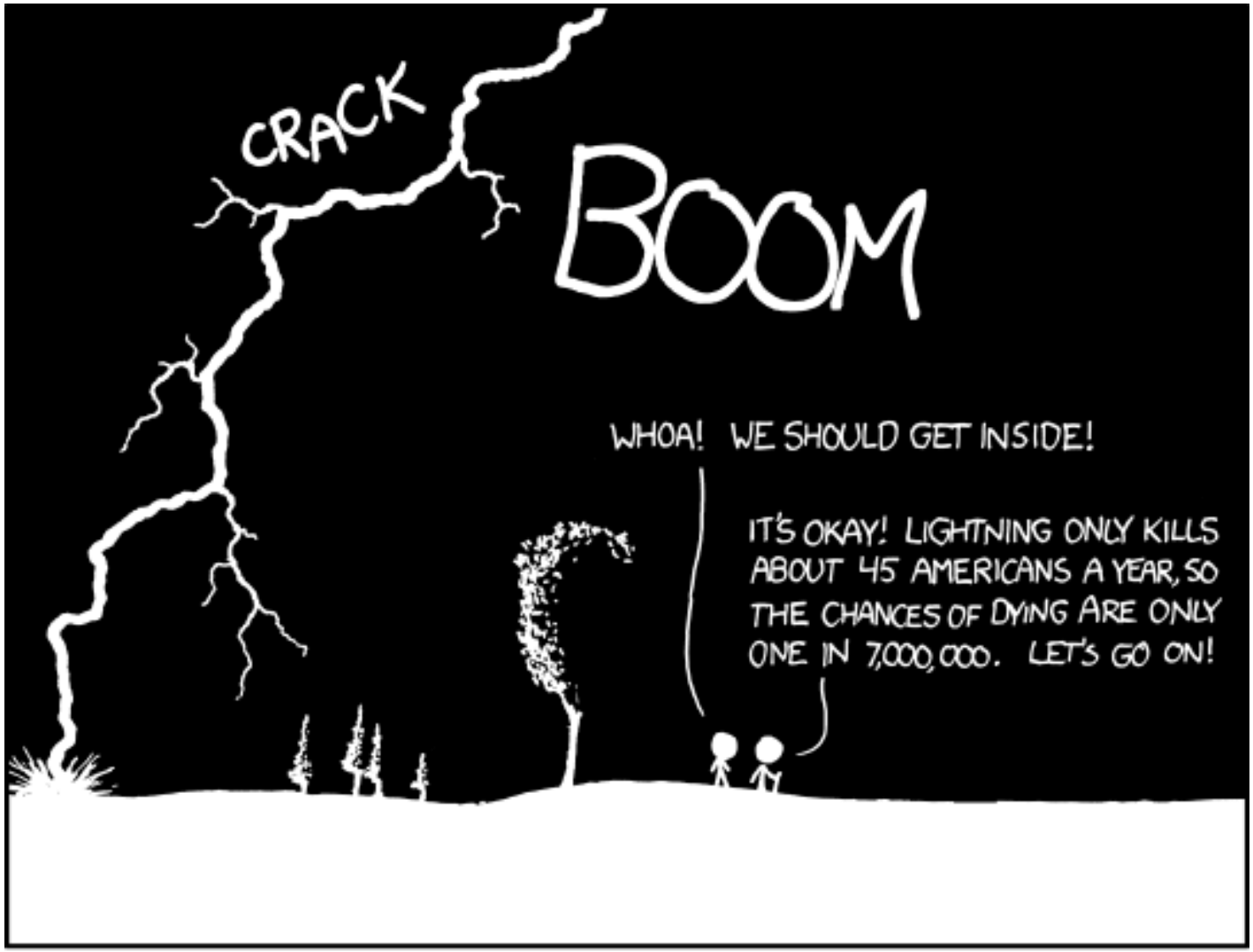
("n sobre k". En anglès: *"n choose k"*).
Observació: $\binom{n}{0} = 1$, $\binom{n}{1} = n$, $\binom{n}{k} = \binom{n}{n-k}$.

1. Una dona té dos fills. Quina és la probabilitat que siguin de sexe diferent?
2. Hi ha 23 persones en una festa. Quina és la probabilitat que n'hi hagi dues (o més) que facin anys el mateix dia?
3. Tenim sis parells de guants i fem sis parelles oblidant que els guants tenen mà. Quina és la probabilitat que no haguem fet cap parella correcta?
4. Llancem un dau sis vegades. Quina és la probabilitat que no surti cap valor repetit?
5. El 1654, un famós jugador, el Chevalier de Méré, va proposar a Fermat i a Pascal aquest problema: Dos jugadors decideixen jugar partides d'un cert joc d'atzar (equitatiu) fins que un d'ells n'hagi guanyat sis, moment en que es quedarà amb l'aposta. Quan un jugador ha guanyat 4 partides i l'altre 3, el joc s'ha d'interrompre per força major. Quina és la manera justa de repartir l'aposta entre els dos jugadors?

1. Si un paquet de cartes està ben remenat, quina és la probabilitat que els quatre asos estiguin junts?
2. Calculeu les probabilitats de les jugades amb 5 daus de poker.
3. Un avió pot volar amb la meitat dels seus motors, però no amb menys. Què és més segur, un avió de dos motors o un de quatre motors?
4. *(El problema de les dues capses de llumins de Banach)*. Un matemàtic compra dues capses de 20 llumins i se'n posa una a cada butxaca. Quan necessita un llumí, l'agafa d'una butxaca o l'altra, a l'atzar. En un moment donat, vol un llumí i el busca a la capsa de la butxaca de la dreta, però aquesta capsa és buida. Quina és la probabilitat que a la capsa de l'altra butxaca hi hagi exactament 10 llumins?
5. Capturem 100 truites d'un llac, les marquem i les tornem al llac. Passat un temps, capturem 100 truites i en trobem 7 de marcades. Quina és la probabilitat que això passi si el llac conté n truites?. Feu una estimació sobre el valor de n.

Probabilitat condicionada


CRACK
BOOM
WHOA! WE SHOULD GET INSIDE!
IT'S OKAY! LIGHTNING ONLY KILLS ABOUT 45 AMERICANS A YEAR, SO THE CHANCES OF DYING ARE ONLY ONE IN 7,000,000. LET'S GO ON!



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Definició

La probabilitat condicionada $P(A|B)$ és la probabilitat que succeeixi l'esdeveniment A quan ja sabem que l'esdeveniment B ha succeït. Es calcula així:

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

Exemple: El 5% dels homes i el 0.25% de les dones d'una certa població són daltònics. Si la població té un 52% de dones, quina és la probabilitat que un individu a l'atzar sigui daltònic?
D = daltònic; M = masculí; F = femení. Volem calcular $P(D)$.

$$\begin{aligned} P(D) &= P((D \cap M) \cup (D \cap F)) = P(D \cap M) + P(D \cap F) \\ &= P(D|M)P(M) + P(D|F)P(F) \\ &= 0.05 \times 0.48 + 0.0025 \times 0.52 = 0.0253 \end{aligned}$$

Fórmula de les probabilitats totals

Si descomponem l'espai mostral en esdeveniments $A_1, A_2, A_3, \ldots$ disjunts dos a dos, aleshores

$$P(A) = P(A|A_1)P(A_1) + P(A|A_2)P(A_2) + P(A|A_3)P(A_3) + \cdots$$

Si sabem la probabilitat de A condicionada a B i volem saber la probabilitat de B condicionada a A, fem això:

Una fórmula molt útil

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

Exemple: En la situació de l'exemple anterior, si escollim una persona a l'atzar i observem que és daltònica, quina és la probabilitat que sigui una dona?

$$P(F|D) = P(D|F)\,\frac{P(F)}{P(D)} = 0.0025 \times \frac{0.52}{0.0253} \approx 0.0514$$

En un concurs televisiu hi participen parelles mare–filla. Avui, concursen l'Aïna i la seva filla Noa. L'Aïna ens explica que té en total dos fills. Quina és la probabilitat que l'altre fill de l'Aïna sigui de sexe masculí?

Resposta 1 (incorrecta)

L'altre fill pot ser un noi o una noia i els dos esdeveniments són igualment probables (aprox.). Per tant, la resposta és $P = 0.5$.

Resposta 2

La informació que tenim sobre l'Aïna és que té dos fills i un és femení. Hem de calcular la probabilitat condicionada $P(\text{un fill masculí} \mid \text{dos fills i un és femení})$. Posem $m =$"un fill masculí"; $F =$"dos fills i un és femení". Tenim:

$$P(m|F) = \frac{P(m\,i\,F)}{P(F)} = \frac{\#\{\text{noi-noia, noia-noi}\}}{\#\{\text{noi-noia, noia-noi, noia-noia}\}} = \frac{2}{3}$$

Combinant la fórmula de les probabilitats totals i la "fórmula molt útil", obtenim la famosa fórmula de Bayes:

La fórmula de Bayes

Si descomponem l'espai mostral en esdeveniments disjunts dos a dos $B = A_1, A_2, A_3, \ldots$, aleshores

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|A_1)P(A_1) + P(A|A_2)P(A_2) + P(A|A_3)P(A_3) + \cdots}$$

Exemple: En una pregunta hi ha $m = 5$ opcions de resposta i només una és correcta. L'estudiant té una probabilitat $p = 0.5$ de saber la resposta. Si respon correctament, quina és la probabilitat que sabés la resposta?

$K =$ "sap la resposta"; $C =$ "respon correctament". $P(K|C) =?$

$$P(K|C) = \frac{P(C|K)P(K)}{B(C|K)P(K) + P(C|K^c)P(K^c)} = \frac{mp}{1 + (m-1)p} = \frac{5}{6}$$

Exemple: En una anàlisi de sang ens detecten una certa malaltia estranya (afecta només una persona de cada 100,000) i aleatòria. Ens informen que el test que ens han fet té una gran fiabilitat: dóna només un 1 per mil de falsos positius i un 0.5 per mil de falsos negatius. Quina és la probabilitat que, realment, tinguem aquesta malaltia estranya?

Resposta: $S =$ "estic sa"; $M =$"estic malalt"; $+ =$ "el test ha donat positiu". $P(M|+) =?$

$$P(M|+) = \frac{P(+|M)P(M)}{P(+|M)P(M) + P(+|S)P(S)}$$

$$= \frac{0.9995 \times 0.00001}{0.9995 \times 0.00001 + 0.001 \times 0.99999} \approx 0.0099$$

És a dir, la probabilitat és, aproximadament, d'una entre cent. (Sorpresa??)

Una altra manera de resoldre aquest exercici és fer una taula de contingència amb una població hipotètica de, per exemple, mil milions de persones:

	+	−	
M	9,995	5	10,000
S	999,990	998,990,010	999,990,000
	1,009,985	998,990,015	1,000,000,000

$$P(M|+) = \frac{9995}{1009985} \approx 0.0099$$

- *El famós problema Monty Hall*. En un concurs hi ha tres portes i darrera d'una d'elles hi ha un premi. Les altres dues no donen premi. El concursant tria una porta. El presentador, abans d'obrir la porta escollida, obre una de les dues altres portes i mostra que no té premi. Aleshores, dóna al concursant l'opció de canviar de porta.
  Quina és la millor estratègia per al concursant: canviar de porta o no canviar? O potser és indiferent? Per què?
- Tres cartolines iguals. Una és negra per les dues cares, l'altra és vermella per les dues cares i la tercera té una cara de cada color. Triem una cartolina a cegues i la deixem sobre la taula. Mostra color vermell. S'admeten apostes sobre el color de l'altra cara. Per quin color apostaries? Per què? És indiferent?

Curs de Bioestadística

Jaume Aguadé

Capítol 3. Variables aleatòries

- Obrim un document per una pàgina a l'atzar i comptem quants errors ortogràfics hi ha.
- Comprem una bombeta i mirem quantes hores funciona abans de fondre's.
- Truquem un servei d'ambulància i mesurem el temps que triga a arribar.
- Escollim a l'atzar una setmana de l'any passat i prenem nota de quants accidents hi va haver en aquesta setmana en un tram concret d'una carretera.
- Analitzem una dosi d'un cert medicament i mesurem la quantitat de principi actiu que conté.
- Escollim a l'atzar un pack de 12 ous i mirem quants n'hi ha de trencats.
- Capturem un ocell en migració i mesurem el seu pes.
- En un dia a l'atzar, mirem quantes persones han viatjat en una certa ruta aèria.
- Escollim un passatger a l'atzar i mirem amb quina antelació ha arribat a l'aeroport.

- Llancem dos daus i designem per X la suma dels seus valors.

Observem:

1. Realitzem un experiment aleatori: llançar dos daus i mirar quins valors surten. Per tant, tenim espai mostral Ω, esdeveniments i probabilitat.
2. A partir del resultat de l'experiment evaluem una certa funció numèrica X: la suma dels dos valors.

Definició de variable aleatòria

Una variable aleatòria és una funció

$$X : \Omega \longrightarrow \mathbb{R}$$

on Ω és un espai mostral i, per tant, tenim una funció de probabilitat P sobre Ω.

Variables discretes

Si X pren només una quantitat finita o numerable de valors, direm que X és discreta.
Exemple: En la llista anterior, distingiu quines variables són discretes i quines no.

- Llancem dos daus i designem per X la suma dels seus valors.

Ens podem preguntar coses com aquestes:

- Quina és la probabilitat que X valgui 3? $P(X = 3) =?$
- Quina és la probabilitat que X valgui menys de 7? $P(X < 7) =?$
- Quina és la probabilitat que X estigui entre 4 y 10, excloent 4 i incloent 10? $P(4 < X \leq 10) =?$

Funció de distribució i funció de probabilitat

Si X és una variable aleatòria discreta, conèixer la funció de distribució (resp. probabilitat) de X ("la distribució de X") és conèixer els valors $P(X \leq k)$ (resp. $P(X = k)$ per tot k.

Exemple:

$$P(4 < X \leq 8) = P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8)$$

- Llancem dos daus i designem per X la suma dels seus valors.

La funció de probabilitat de X és:

$$P(X=2)=P(X=12)=1/36 \quad P(X=5)=P(X=9)=4/36$$
$$P(X=3)=P(X=11)=2/36 \quad P(X=6)=P(X=8)=5/36$$
$$P(X=4)=P(X=10)=3/36 \quad P(X=7)=6/36$$

- Llancem una moneda tres vegades i designem per Y el nombre total de cares que obtenim.

$$P(Y=0)=P(Y=3)=1/8 \quad P(Y=1)=P(Y=2)=3/8$$

- Llancem una moneda fins que surti cara i designem per Z el nombre de vegades que l'hem llançada.

$$P(Z=k)=\frac{1}{2^k} \qquad P(Z=\infty)=0$$

Definició d'esperança

Si una variable aleatòria X pren valors $\{x_i\}$, la seva esperança és:

$$E(X) = \mu_X = \sum x_i \, P(X = x_i)$$

Exemples: Si X, Y, Z són com en els exemples anteriors;

$$E(X) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + \cdots + 12 \times \frac{1}{36} = 7$$

$$E(Y) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 3/2$$

$$E(Z) = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + \cdots = \sum_{k=1}^{\infty} \frac{k}{2^k} = 2$$

$E(X)$ representa el valor que obtindríem, en mitjana, si evaluéssim X un nombre molt gran de vegades.

Propietats de l'esperança

$E(X+Y) = E(X) + E(Y)$

$E(aX+b) = aE(X) + b$

En general,

$E(XY) \neq E(X) \cdot E(Y)$

Joc just

Un joc just és aquell en que el guany net té esperança zero. Per què? Hi ha algun joc d'atzar just a la vida real?

- Considerem un cert fenòmen aleatori que pot succeïr amb probabilitat p (Exemple: escollim un individu al atzar i li preguntem si és fumador; marquem un número de telèfon i mirem si respon; considerem un cert producte financer i mirem si demà puja el seu valor;...) Sigui X la variable aleatòria que val 1 si el fenòmen succeeix i 0 si no succeeix. L'esperança de X és:

$$E(X) = 1 \times P(X=1) + 0 \times P(X=0) = 1 \times p + 0 \times (1-p) = p$$

Direm que X és una variable de tipus Bernoulli.

Definició de variància

Si una variable aleatòria X pren valors $\{x_i\}$ i té esperança μ_X, la seva variància és:

$$\mathrm{Var}(X) = \sigma_X^2 = \sum (x_i - \mu_X)^2 \, P(X = x_i)$$

Propietats

$\mathrm{Var}(X) = E(X^2) - E(X)^2$
$\mathrm{Var}(aX + b) = a^2 \, \mathrm{Var}(X)$
En general,
$\mathrm{Var}(X+Y) \neq \mathrm{Var}(X)+\mathrm{Var}(Y)$

Exemples

- $X =$ resultat de llançar un dau. $E(X) = 7/2$, $E(X^2) = 91/6$, $\mathrm{Var}(X) = 35/12$.
- $X =$ variable Bernoulli. $X = X^2$, $E(X) = E(X^2) = p$, $\mathrm{Var}(X) = p - p^2 = p(1-p) = pq$.

X i Y seran independents si els valors que pren X no afecten la probabilitat dels valors que pot prendre Y. Mes exactament:

Definició de variables aleatòries independents

X i Y són independents si, per tots els valors k, k', els esdeveniments $X \leq k$ i $Y \leq k'$ són independents. És a dir:

$$P\big((X \leq k) \text{ i } (Y \leq k')\big) = P(X \leq k)\, P(Y \leq k')$$

La Covariància és $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$. Si X i Y són independents, aleshores:

- $\text{Cov}(X, Y) = 0$.
- $E(XY) = E(X)E(Y)$.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$:

$$\text{Var}(X+Y) = E((X+Y)^2) - E(X+Y)^2 = E(X^2) + E(Y^2) + 2E(XY) -$$

$$E(X)^2 - E(Y)^2 - 2E(X)E(Y) = \text{Var}(X) + \text{Var}(Y)$$

Una variable aleatòria contínua és aquella que pren una quantitat no numerable de valors i la probabilitat de $X \leq a$ es calcula com la integral d'una funció. Més exactament:

Definició de variable aleatòria contínua

X és contínua si existeix una funció f, anomenada funció de densitat de X tal que

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

Recordeu que la integral d'una funció no és altra cosa que l'àrea sota de la funció.

Important

En una variable contínua, els valors individuals tenen sempre probabilitat zero:

$$P(X = k) = 0$$

Només tenen (poden tenir) probabilitat positiva els intervals:

$$P(X \leq a), \quad P(X > b), \quad P(c < X < d), \ldots$$

En particular, en una variable contínua

$$P(X \leq a) = P(X < a)$$

$$P(X \geq a) = P(X > a)$$

(Això no és cert en una discreta!)

Per a variables contínues, l'esperança i la variància es defineixen utilitzant la integral en lloc del sumatori:

Esperança

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

Variància

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\, dx$$

També: $\text{Var}(X) = E(X^2) - E(X)^2$.

Curs de Bioestadística

Jaume Aguadé

Capítols 4 i 5: Variables aleatòries discretes i contínues

Un experiment de Bernoulli és un experiment aleatori amb dos únics resultats:

- èxit, $X = 1$.
- fracàs, $X = 0$.

En cada cas, caldrà fixar què entenem per èxit i què entenem per fracàs. La variable aleatòria X té aquestes propietats:

Variable de tipus Bernoulli

$$E(X) = p, \quad \text{Var}(X) = pq$$

- Llancem una moneda i mirem si surt cara.
- Escollim una persona que acaba d'emetre el vot en un referèndum i li preguntem si ha votat "sí".
- Fem una anàlisis de sang a un pacient i mirem si el nivell de triglicèrids està en un cert interval.
- Un jugador de bàsquet llança un tir lliure i mirem si puntua.
- Fan un control d'alcoholèmia a un conductor escollit a l'atzar i ens fixem en si és sancionat.

Llancem un dau tres vegades i comptem el nombre total de sisos.

Distribució binomial

X té distribució binomial si X compta el nombre total d'èxits en n experiments de Bernoulli independents.

$$X \sim B(n, p)$$

$X \sim B(3, 1/6)$. Nombre de resultats: $6 \times 6 \times 6$. D'aquests:

- Cap sis. $X = 0$. $5 \times 5 \times 5$ vegades.
- Un sis. $X = 1$. $1 \times 5 \times 5 + 5 \times 1 \times 5 + 5 \times 5 \times 1$ vegades.
- Dos sisos. $X = 2$. $1 \times 1 \times 5 + 1 \times 5 \times 1 + 5 \times 1 \times 1$ vegades.
- Tres sisos. $X = 3$. $1 \times 1 \times 1 = 1$ vegades.

$$P(X = 0) = \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}, \quad P(X = 1) = 3 \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6}$$

$$P(X = 2) = 3 \times \frac{5}{6} \times \frac{1}{6} \times \frac{1}{6}, \quad P(X = 3) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6}$$

Funció de probabilitat

$$X \sim B(n,p); \qquad P(X=k) = \binom{n}{k} p^k q^{n-k},\ k = 0,1,2,\ldots,n$$

Esperança i variància

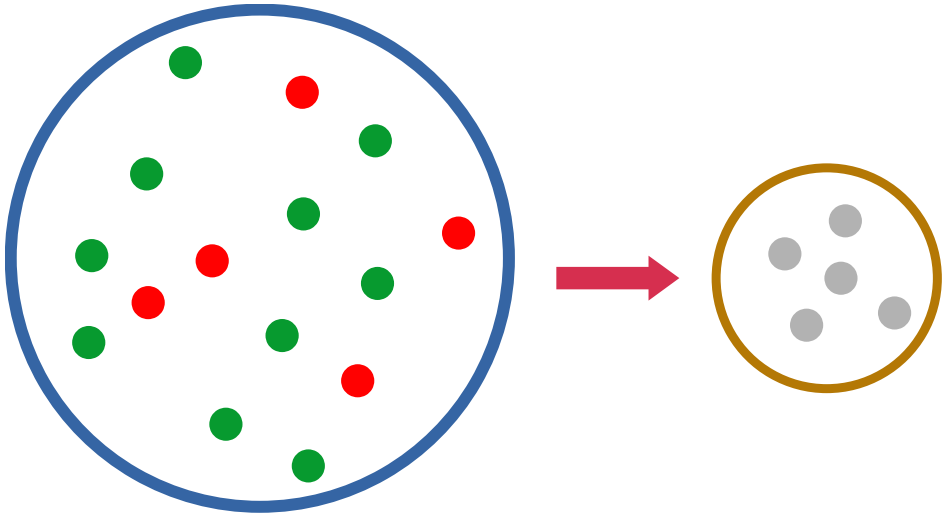
$$X \sim B(n,p); \qquad E(X) = np; \qquad \mathrm{Var}(X) = npq$$

Cal saber fer: Calcular probabilitats d'una distribució binomial amb calculadora i amb un full de càlcul.

Important: La independència dels experiments (reposició/no reposició, tamany de la població,...)

1. Llancem una moneda cinc vegades. Probabilitat de treure meys de dues cares. Més de tres cares. Més cares que creus.
2. Examen test amb 10 preguntes i cinc respostes per pregunta. Si contestem a l'atzar, probabilitat d'aprovar.
3. Una operació té un percentatge d'èxit del 80%. Probabilitat que, dels cinc propers pacients, exactament quatre superin l'operació amb èxit.
4. Un 0.1% dels xips que produïm són defectuosos. Si els venem en paquets de 20, probabilitat que un paquet no en tingui més de 2 de defectuosos.
5. La incidència de la grip entre els nostres empleats és del 15%. Probabilitat que algun membre d'un departament concret de 5 persones agafi la grip.
6. Un sistema (electrònic o humà) amb 4 components pots seguir funcionant si té un mínim de dos components en bon estat. Si la probabilitat que un component falli és de 0.4, probabilitat que el sistema funcioni.

Si la població és petita i no hi ha reposició, la distribució binomial no és aplicable. Cal utilitzar la distribució hipergeomètrica.



Tenim 10 boles verdes i 5 boles vermelles. En triem 5 a l'atzar. Quina és la probabilitat que n'haguem triat exactament 3 de vermelles? Depèn de com les haguem triat:

- Amb reposició: triem una bola i la retornem a la bossa; i això ho fem 5 vegades.
- Sense reposició: triem 5 boles, sense retornar-les a la bossa.

Si hi ha reposició: distribució binomial $B(5, 1/3)$ i

$$P(X = 3) = \binom{5}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2 \approx 16.5\%$$

Si no hi ha reposició: distribució hipergeomètrica i

- Nombre de maneres d'escollir 5 boles d'entre 15: $\binom{15}{5}$.
- Nombre de maneres d'escollir 3 boles vermelles d'entre 5: $\binom{5}{3}$.
- Nombre de maneres d'escollir 2 boles verdes d'entre 10: $\binom{10}{2}$.

$$P(X = 3) = \frac{\binom{5}{3}\binom{10}{2}}{\binom{15}{5}} \approx 15\%.$$

Distribució hipergeomètrica

Una població de mida N amb K "èxits" (i, per tant $N-K$ "fracassos"). Escollim aleatoriament n objectes, sense reposició. La probabilitat que haguem triat exactament k èxits és

$$P(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Si N és gran, no té gaire importància si hi ha reposició o no n'hi ha i podem utilitzar la distribució binomial.

Siméon Denis Poisson 1781–1840

1837: *Recherches sur la probabilité des jugements en matière criminelle et en matière civile.*

La distribució de Poisson

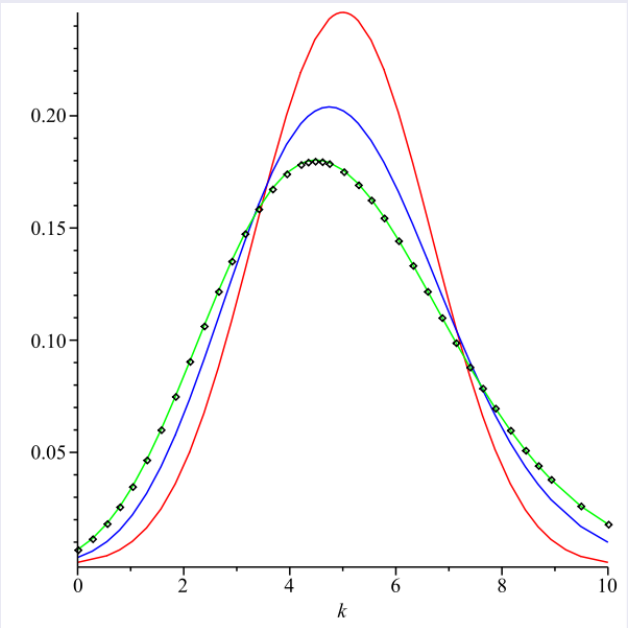
Si en una binomial $B(n, p)$, n és molt gran i p és molt petit, la distribució només depèn de l'esperança $\lambda = np$. S'en diu una distribució de Poisson $\text{Pois}(\lambda)$.

Exemple: Una alteració genètica no hereditària té una prevalença de 1 entre 50,000. El nombre de casos en una població de 100,000 habitants vindrà donat per

$$B(10^5, 2 \times 10^{-5}) \approx \text{Pois}(2)$$

Es considera que es té una aproximació acceptable si $n \geq 20$ i $p \leq 0.05$.

- En una carretera hi ha, en promig, 3 accidents per setmana. Probabilitat que la propera setmana hi hagi algun accident.
- Si el 4% dels nens són esquerrans i les aules són de 25 nens, quantes aules no tindran cap nen esquerrà?
- En una ciutat de 400,000 habitants hi ha una tasa de suicidi d'un per cada cent mil habitants i any. Probabilitat que en un any hi hagi més de 7 suicidis.
- Una agència d'assegurances rep cada dia, en mitjana, 5 comunicats de sinistres amb danys personals. Quina proporció de dies tindran menys de 3 comunicats?
- En una regió es produeixen, en mitjana, 2.2 grans inundacions cada 50 anys. Quina és la probabilitat que en els propers 50 anys hi hagi 3 grans inundacions?
- La guàrdia urbana d'una localitat posa 8 denúncies d'aparcament per hora, en mitjana. Probabilitat de que en una hora hi hagi més de 10 denúncies.



Comparació de la Poisson (verd) amb $\lambda = 5$ i la binomial per $n = 10$ (vermell) i $n = 20$ (blau).

Funció de probabilitat

$$X \sim \text{Pois}(\lambda); \quad P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

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

$$\frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} =$$

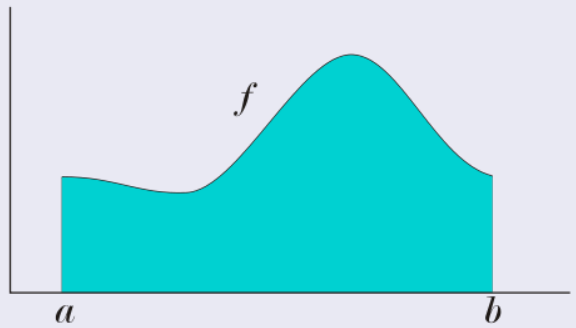
$$\frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{\lambda^k}{k!}\frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} \approx$$

$$\frac{\lambda^k}{k!}e^{-\lambda}$$

$$X \sim \text{Pois}(\lambda) \Rightarrow E(X) = \text{Var}(X) = \lambda$$

Recordem:

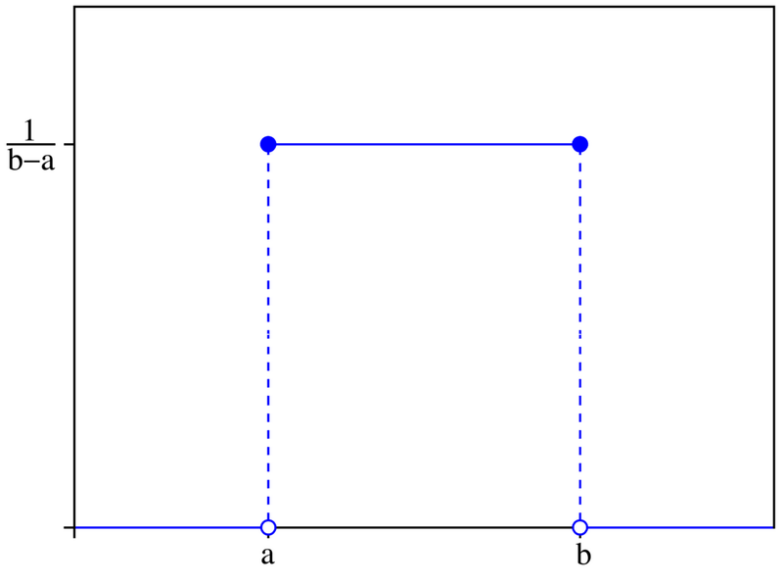
Una Distribució contínua té una funció de densitat $f(x)$ i la probabilitat que $a < X < b$ es mesura com **l'àrea sota de** $y = f(x)$ **entre** a **i** b:



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

- L'àrea total sota de $y = f(x)$ ha de ser 1.
- Els valors individuals tenen probabilitat zero: $P(X = a) = 0$.
- No cal distingir entre $<$ i $\leq$ ni entre $>$ i $\geq$

Els tren surten als quarts d'hora exactes. Si arribo entre les 7 i les 7:30, en un moment uniformement distribuït, quina és la probabilitat que no m'hagi d'esperar més de 5 minuts?



$$P(u < X < v) = \frac{v-u}{b-a}$$

La distribució uniforme apareix quan una variable aleatòria té igual probabilitat d'estar prop de qualsevol valor d'un interval $[a, b]$.

mitjana i variància

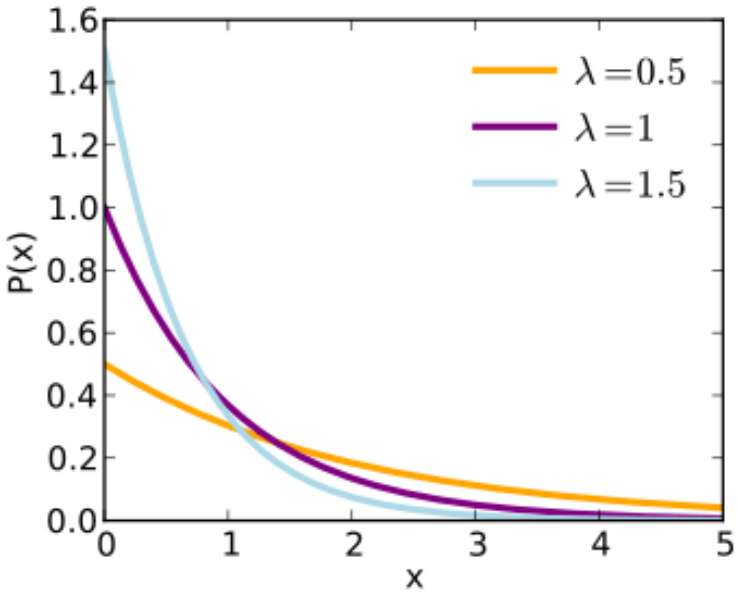
$$E(X) = \frac{a+b}{2}; \qquad \text{Var}(X) = \frac{(b-a)^2}{12}$$

(Nombres aleatoris: valors d'una variable aleatòria uniformement distribuida entre 0 i 1.)

La distribució exponencial

La distribució exponencial és una distribució contínua que descriu el temps entre dos successos de Poisson (=successos aleatoris que es produeixen independentment i a velocitat mitjana constant).

- Temps entre l'arribada de dos clients a una oficina, o entre dos ingressos a urgències.
- Temps entre dos terratrèmols, entre la caiguda de dos llamps en una zona.
- Temps entre dues peticions a una pàgina web.
- Temps entre dos accidents aeris.
- Temps entre dues caigudes d'un sistema informàtic.
- Temps entre dues emissions de partícules per una substància radioactiva.
- Temps entre dues trucades telefòniques.
- Distància entre dos animals atropellats en una carretera.



La distribució exponencial no té memòria:

$$P(X > s + t \,|\, X > t) = P(X > s)$$

Aquesta propietat és característica de l'exponencial.

Funció de densitat

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

Funció de distribució

$$X \sim \text{Exp}(\lambda)$$

$$P(X < u) = 1 - e^{-\lambda u}$$

$$P(X > u) = e^{-\lambda u}$$

$$E(X) = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

La nit del 17 de novembre es la millor per observar la pluja d'estels dels Leònids. Per aquest any la predicció és de 12 observacions per hora. Estem observant el cel i fa estona que no veiem cap meteor. Algú comenta que 12 meteors per hora és un cada 5 minuts i, per tant, si esperem 5 minuts, *segur* que en veiem algun. Algú replica: *no, si esperem 5 minuts més només tenim un 63% de probabilitat de veure'n algun*. Per què?

Fem la hipòtesi que l'aparició d'un meteor és un fenomen de Poisson i, per tant, el temps entre dues observacions ve donat per la distribució exponencial. Com que l'exponencial "no té memòria", no importa el temps que fa que no veiem cap meteor. Si mesurem el temps en hores, $\lambda = 12$, 5 minuts són $1/12$ hores i

$$P(X < 1/12) = 1 - e^{-\lambda/12} = 1 - e^{-1} \approx 63\%$$

Curs de Bioestadística

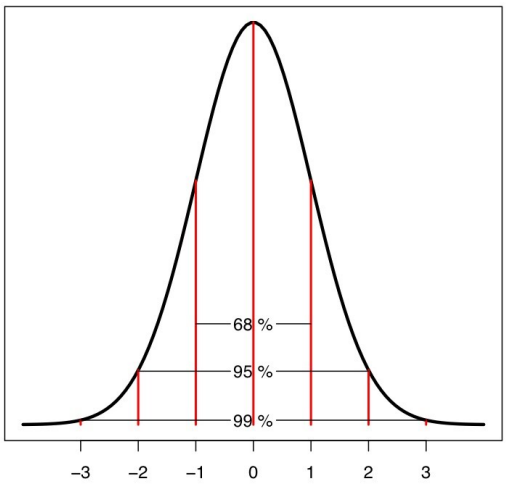
Jaume Aguadé

Capítol 6: La distribució normal

C. F. Gauss
1777–1855

L'u de gener del 1801, Giuseppe Piazzi descobreix Ceres i determina la seva posició durant 40 dies, fins que ja no es veu perquè el Sol l'oculta. Gauss (tenia 24 anys) utilitza el mètode dels mínims quadrats (inventat per ell quan tenia 18 anys) per calcular on seria Ceres quan sortís de la zona del Sol.

El 1809, Gauss observa que la justificació del mètode dels mínims quadrats es basa en suposar que els errors experimentals es distribueixen segons una corba interessant, que s'anomena "campana de Gauss".



$$h(x) = e^{-x^2}, \qquad \int_{-\infty}^{\infty} e^{-x^2} = \sqrt{\pi}$$

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

Funció de densitat de la Distribució Normal Estàndard $N(0,1)$.

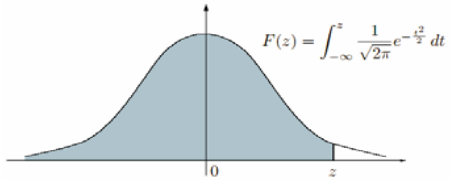
$N(0,1)$

- És una distribució contínua amb funció de densitat $\frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$.
- Té esperança $\mu = 0$ i variància $\sigma^2 = 1$.
- És una distribució unimodal simètrica.
- El 68% de les observacions es separen menys de 1 de la mitjana. El 95% de les observacions es separen menys de 2 de la mitjana. El 99.7% de les observacions es separen menys de 3 de la mitjana.
- Com en tota distribució contínua, la probabilitat es calcula com l'àrea sota de la corba. Per exemple, si $Z \sim N(0,1)$,

$$P(1.5 < Z < 2.1) = \frac{1}{\sqrt{2\pi}} \int_{1.5}^{2.1} e^{-\frac{x^2}{2}} \approx 0.05$$

- Aquestes integrals es calculen: (a) amb una taula de la normal estàndard; (b) amb un programa estadístic.

La llei normal $N(0,1)$



Negatius	-0,09	-0,08	-0,07	-0,06	-0,05	-0,04	-0,03	-0,02	-0,01	0,00
-3,9	,0000	,0000	,0000	,0000	,0000	,0000	,0000	,0000	,0000	,0000
-3,8	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001
-3,7	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001
-3,6	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0001	,0002	,0002
-3,5	,0002	,0002	,0002	,0002	,0002	,0002	,0002	,0002	,0002	,0002
-3,4	,0002	,0003	,0003	,0003	,0003	,0003	,0003	,0003	,0003	,0003
-3,3	,0003	,0004	,0004	,0004	,0004	,0004	,0004	,0005	,0005	,0005
-3,2	,0005	,0005	,0005	,0006	,0006	,0006	,0006	,0006	,0007	,0007
-3,1	,0007	,0007	,0008	,0008	,0008	,0008	,0009	,0009	,0009	,0010
-3,0	,0010	,0010	,0011	,0011	,0011	,0012	,0012	,0013	,0013	,0013
-2,9	,0014	,0014	,0015	,0015	,0016	,0016	,0017	,0018	,0018	,0019
-2,8	,0019	,0020	,0021	,0021	,0022	,0023	,0023	,0024	,0025	,0026
-2,7	,0026	,0027	,0028	,0029	,0030	,0031	,0032	,0033	,0034	,0035
-2,6	,0036	,0037	,0038	,0039	,0040	,0041	,0043	,0044	,0045	,0047
-2,5	,0048	,0049	,0051	,0052	,0054	,0055	,0057	,0059	,0060	,0062
-2,4	,0064	,0066	,0068	,0069	,0071	,0073	,0075	,0078	,0080	,0082
-2,3	,0084	,0087	,0089	,0091	,0094	,0096	,0099	,0102	,0104	,0107
-2,2	,0110	,0113	,0116	,0119	,0122	,0125	,0129	,0132	,0136	,0139
-2,1	,0143	,0146	,0150	,0154	,0158	,0162	,0166	,0170	,0174	,0179
-2,0	,0183	,0188	,0192	,0197	,0202	,0207	,0212	,0217	,0222	,0228
-1,9	,0233	,0239	,0244	,0250	,0256	,0262	,0268	,0274	,0281	,0287
-1,8	,0294	,0301	,0307	,0314	,0322	,0329	,0336	,0344	,0351	,0359
-1,7	,0367	,0375	,0384	,0392	,0401	,0409	,0418	,0427	,0436	,0446
-1,6	,0455	,0465	,0475	,0485	,0495	,0505	,0516	,0526	,0537	,0548
-1,5	,0559	,0571	,0582	,0594	,0606	,0618	,0630	,0643	,0655	,0668
-1,4	,0681	,0694	,0708	,0721	,0735	,0749	,0764	,0778	,0793	,0808
-1,3	,0823	,0838	,0853	,0869	,0885	,0901	,0918	,0934	,0951	,0968
-1,2	,0985	,1003	,1020	,1038	,1056	,1075	,1093	,1112	,1131	,1151
-1,1	,1170	,1190	,1210	,1230	,1251	,1271	,1292	,1314	,1335	,1357
-1,0	,1379	,1401	,1423	,1446	,1469	,1492	,1515	,1539	,1562	,1587
-0,9	,1611	,1635	,1660	,1685	,1711	,1736	,1762	,1788	,1814	,1841
-0,8	,1867	,1894	,1922	,1949	,1977	,2005	,2033	,2061	,2090	,2119
-0,7	,2148	,2177	,2206	,2236	,2266	,2296	,2327	,2358	,2389	,2420
-0,6	,2451	,2483	,2514	,2546	,2578	,2611	,2643	,2676	,2709	,2743
-0,5	,2776	,2810	,2843	,2877	,2912	,2946	,2981	,3015	,3050	,3085
-0,4	,3121	,3156	,3192	,3228	,3264	,3300	,3336	,3372	,3409	,3446
-0,3	,3483	,3520	,3557	,3594	,3632	,3669	,3707	,3745	,3783	,3821
-0,2	,3859	,3897	,3936	,3974	,4013	,4052	,4090	,4129	,4168	,4207
-0,1	,4247	,4286	,4325	,4364	,4404	,4443	,4483	,4522	,4562	,4602
0,0	,4641	,4681	,4721	,4761	,4801	,4840	,4880	,4920	,4960	,5000

Positius	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
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	,9279	,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
3,5	,9998	,9998	,9998	,9998	,9998	,9998	,9998	,9998	,9998	,9998
3,6	,9998	,9998	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999
3,7	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999
3,8	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999
3,9	1	1	1	1	1	1	1	1	1	1

Si $Z \sim N(0,1)$, calculeu (amb la taula de la normal, amb R i amb un full de càlcul):

- $P(-0.36 < Z < 0.97)$
- $P(Z < 1.62)$
- $P(Z > 2.4)$
- z tal que $P(Z < z) = 0.95$
- z tal que $P(Z > z) = 0.05$
- El 3% dels valors de Z seran més petits que...
- Quin percentatge de valors de Z superarà 2.6?
- Quin interval simètric respecte de zero contindrà el 80% dels valors de Z?
- Quin interval simètric respecte de zero deixarà fora el 12% dels valors de Z?

La normal estàndard $Z \sim N(0,1)$ té mitjana 0 i desviació típica 1.
Si fem

$$X = \sigma Z + \mu$$

X té una distribució normal de mitjana μ i desviació típica σ.
Escriurem $X \sim N(\mu, \sigma^2)$.

Estandardització

$$Z \sim N(0,1) \iff X \sim N(\mu, \sigma^2)$$

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

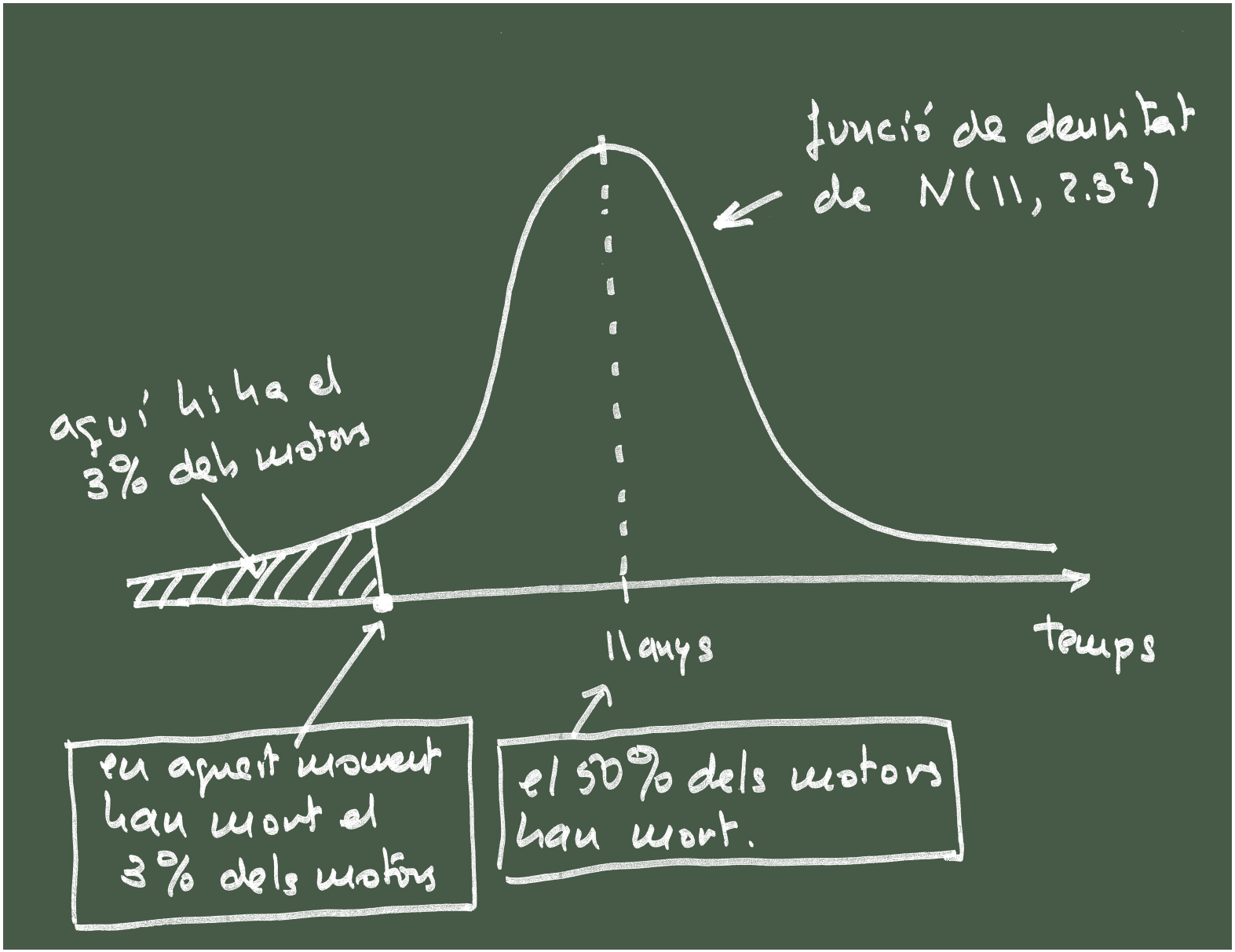
D'aquesta manera podem reduir càlculs sobre una normal qualsevol $N(\mu, \sigma^2)$ a càlculs sobre la normal estàndard $N(0,1)$. I viceversa.

Un fabricant de frigorífics està disposat a reemplaçar en garantia, com a màxim, un 3% dels motors dels seus aparells. Ha observat que els seus motors tenen una vida mitjana de 11 anys, amb una desviació estàndard de 2.3 anys. Quin període de garantia ha d'oferir?

- $X =$ temps de vida (en anys) d'un motor. Podem fer la hipòtesis que X és una variable aleatòria normal, $X \sim N(11, 2.3^2)$
- Si x és el temps de garantia, volem $P(X < x) = 0.03$
- Per calcular el valor de x, tipifiquem:

$$Z \sim N(0,1), \quad 0.03 = P(X < x) = P(Z < \frac{x - 11}{2.3})$$

- A la taula de la normal veiem $P(Z < -1.881) \approx 0.03$
- Per tant, $x = 2.3 \times (-1.881) + 11 \approx 6.7$ anys


funció de densitat
de $N(11, 2.3^2)$
aquí hi ha el
3% dels motors
temps
11 anys
en aquest moment
han mort el
3% dels motors
el 50% dels motors
han mort.

La distribució normal apareix en molts llocs a la ciència, la medicina, l'economia, les biociències, les ciències socials,...

- Els errors de mesura.
- El logaritme de les mesures dels organismes vius.
- La velocitat de les molècules d'un gas ideal.
- Variables binomials $B(n, p)$ quan np i nq són molt grans.
- Variables Poisson amb λ molt gran.
- Resultats d'una avaluació feta a molta gent.
- Canvis de preu d'un valor financer.

El Teorema Central del Límit

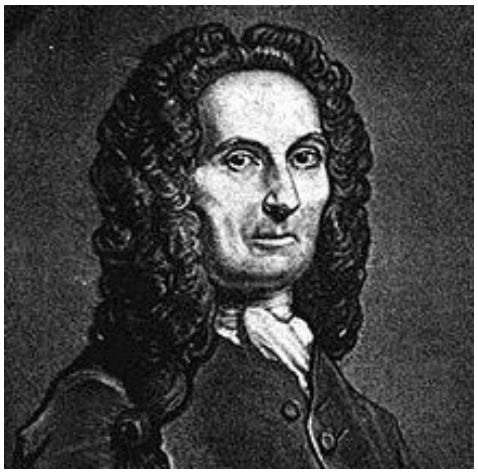
Si $X_1, X_2, \ldots, X_n$ són variables aleatòries independents amb la mateixa distribució, la mateixa mitjana i la mateixa variància (diferent de zero), aleshores $X_1 + X_2 + \cdots + X_n$ és aproximadament normal, si n és gran.

Aquest teorema dóna una explicació teòrica de l'aparició de la distribució normal en tants i tants àmbits.

Llancem una moneda 1000 vegades. Calculeu la probabilitat que surtin menys de 475 cares.

$X \sim B(1000, 0.5)$, per tant, $P(X < 475) =$

$$\binom{1000}{0}(0.5)^{1000} \times (0.5)^0 + \binom{1000}{1}(0.5)^{999} \times (0.5)^1 +$$
$$\binom{1000}{2}(0.5)^{998} \times (0.5)^2 + \cdots + \binom{1000}{474}(0.5)^{526} \times (0.5)^{474}$$



Abraham de Moivre
1667–1754

de Moivre es va adonar que la funció de Gauss podia servir per obtenir aquest resultat de manera més simple. Posteriorment, el teorema central del límit explica el resultat de de Moivre.

Aproximació de la binomial per la normal

$$B(n, p) \approx N(np, npq) \text{ si } np,\ nq \geq 5$$

Llancem una moneda 1000 vegades. Calculeu la probabilitat que surtin menys de 475 cares.

$X \sim B(1000, 0.5)$ l'aproximem per $Y \sim N(500, 250)$. Per tant,

$$P(X < 475) = P(X < 474.5) \approx P(Y < 474.5) =$$

$$P\left(Z < \frac{474.5 - 500}{\sqrt{250}}\right) = P(Z < -1.61) \approx 5.37\%$$

- Observeu la correcció de continuïtat que millora l'aproximació (important si np no és molt gran).
- Experimenteu amb un full de càlcul (o amb R) l'aproximació de la binomial per la normal.

Al llarg del curs aniran apareixent altres distribucions contínues relacionades amb la distribució normal. Per exemple:

- La distribució chi-quadrat amb n graus de llibertat, χ^2_n, és una distribució positiva de mitjana n i variància $2n$ que s'utilitza, per exemple, per fer el test d'independència de Pearson. Es defineix com la distribució de $Z_1^2 + \cdots + Z_n^2$ on cada Z_i és una normal estàndard independent de les altres.
- La distribució t de Student amb n graus de llibertat és una distribució que s'assembla a la normal i que utilitzarem en el test per a la mitjana d'una població normal, amb mostra petita.
- La distribució $F(n, m)$ de Fisher amb n i m graus de llibertat, que és una distribució positiva que s'utilitza en anàlisi de la variància.

Ja estudiarem amb més detall aquestes distribucions quan arribi el moment d'utilitzar-les.

Curs de Bioestadística

Jaume Aguadé

Capítol 7. Estimació de paràmetres

L'objectiu de la inferència estadística és obtenir informació sobre les característiques d'una variable X sobre una població (gran) a partir dels valors de la variable mesurats sobre una mostra (petita). Per poder fer inferència estadística de manera correcta, cal que el mostreig (anglès: *sampling*) sigui aleatori: Tots els individus de la població han de tenir la mateixa probabilitat de ser escollits per formar part de la mostra.

- Volem estimar l'alçada mitjana dels catalans de 18 anys (població = les alçades de tots els catalans de 18 anys) i mesurem les alçades d'una mostra aleatòria de mida 150.
- Volem estimar la intenció de vot de tota una població i estudiem la intenció de vot d'una mostra aleatòria de 700 persones d'aquesta població.
- Volem estimar l'eficiència d'un nou fàrmac quan l'administrem a tota una població i estudiem l'eficiència en una mostra aleatòria de 25 pacients.
- Volem estimar la vida útil d'un cert mecanisme i estudiem la vida útil d'una mostra aleatòria de 30 d'aquests mecanismes.

- X una variable numèrica sobre una població.
- X té una certa distribució i uns certs paràmetres (mitjana i variància són els més importants).
- Els paràmetres de X són, en general, desconeguts, però tenen un valor concret (són nombres).
- Prenem una mostra aleatòria de mida n de la població.
- Cada valor de la mostra $X_1, \ldots, X_n$ és una variable aleatòria.
- Com que suposem que la població és molt gran i la mostra és aleatòria, les variables $X_1, \cdots, X_n$ són independents.
- A partir dels valors $X_1, \ldots, X_n$ de la mostra, podem calcular tota mena d'estadístics. Per exemple, la mitjana mostral

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

o la variància mostral

$$S^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2$$

- Els estadístics (exemples: mitjana mostral i variància mostral) són variables aleatòries. Tenen valor diferent amb cada nova mostra que escolim.
- Per tant, aquests estadístics tenen també distribució, mitjana, variància...
- El nostre objectiu és: a partir dels valors dels estadístics (coneguts), estimar els valors dels paràmetres (desconeguts).

Paràmetres

- Són nombres fixos.
- Depenen de tota la població.
- Són desconeguts.
- Els principals són μ, σ^2

Estadístics

- Són variables aleatòries.
- Depenen d'una mostra.
- Són coneguts.
- Els principals són $\overline{X}$, S^2

Administrem a 15 pacients amb nivells de colesterol elevats un nou fàrmac, durant un mes. Després del tractament, els mesurem el nivell de colesterol. Alguns pacients no han experimentat cap canvi. Altres han augmentat el colesterol, altres l'han reduït. Observem que, en mitjana, en aquests 15 pacients el colesterol s'ha reduït en un 12%.

- Creus que, si administrem aquest fàrmac a tota la població amb colesterol elevat, el colesterol se'ls reduirà, en mitjana, exactament el 12%?
- Podem pensar que la reducció de colesterol a tota la població serà "propera" al 12%? Què vol dir "propera"?
- Creus que, si escollim una nova mostra de 15 pacients també se'ls reduirà el colesterol exactament el 12%?
- Podem pensar que la reducció de colesterol a una nova mostra serà "propera" al 12%?
- Quins són la població, la mostra, els paràmetres, els estadístics?

Estimació puntual

Donem un valor que creiem que serà "proper" al valor desconegut del paràmetre.

Exemple: el nostre fàrmac contra el colesterol donarà, aproximadament, una disminució del 12% en el nivell del colesterol, en mitjana.

Interval de confiança

Donem un interval $(a - \epsilon, a + \epsilon)$ que conté el paràmetre desconegut, amb una certa probabilitat γ.

Exemple: el nostre fàrmac contra el colesterol donarà una disminució d'entre el 10% i el 18% en el nivell del colesterol, amb un nivell de confiança de $\gamma = 95\%$.

- X una variable sobre una població amb mitjana poblacional μ i variància poblacional σ^2.
- $X_1, \ldots, X_n$ una mostra aleatòria de mida n.
- $X_1, \ldots, X_n$ són variables aleatòries independents de mitjana μ.
- La mitjana mostral és $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$.
- $\overline{X}$ és una variable aleatòria. Per tant, té esperança i variància.
- $E(\overline{X}) = \frac{1}{n}\left[E(X_1) + \cdots + E(X_n)\right] = \mu$.
- $\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\left[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\right] = \frac{\sigma^2}{n}$.

$E(\overline{X}) = \mu$ i $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$ ens diuen que $\overline{X}$ és un bon estimador de μ:

- És un estimador no esbiaixat perquè la seva mitjana és el valor del paràmetre.
- És un estimador consistent perquè, quan la mida de la mostra creix, la seva variància tendeix a zero.

- X una variable sobre una població amb mitjana poblacional μ i variància poblacional σ^2.
- $X_1, \ldots, X_n$ una mostra aleatòria de mida n. Són variables aleatòries independents de mitjana μ i variància σ^2.
- La variància mostral és $S^2 = \frac{(X_1-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n-1}$.
- S^2 és una variable aleatòria. Per tant, té esperança i variància.
- $E(S^2) = \sigma^2$. Per tant, la variància mostral S^2 és un estimador no esbiaixat de la variància poblacional σ^2.
- $\mathrm{Var}(S^2)$ no es pot calcular sense tenir més informació sobre X.

El misteri de les dues variàncies

Per què hi ha dues variàncies, segons es divideixi per n o per $n-1$?

- Si tenim una variable aleatòria X, la seva variància és la variància poblacional

$$\sigma^2 = \frac{1}{N}\sum(X_i - \mu)^2 = E(X^2) - E(X)^2$$

- Si prenem una mostra de X i volem estimar la variància poblacional de X, σ^2, hem d'utilitzar la variància mostral $S^2 = \frac{1}{n-1}\sum(X_i - \mu)^2$ perquè és aquesta la que és un estimador de σ^2. Si dividíssim per n en lloc de per $n-1$ obtindríem un resultat lleugerament diferent que, encara que sembli estrany, NO és el millor estimador de σ^2.
- Tanmateix, si la mida de la mostra ($= n$) és prou gran, hi ha poca diferència entre dividir per n o per $n-1$.

R. A. Fisher
1890–1962

Si la població és normal, podem dir més coses sobre $\overline{X}$ i sobre S^2.

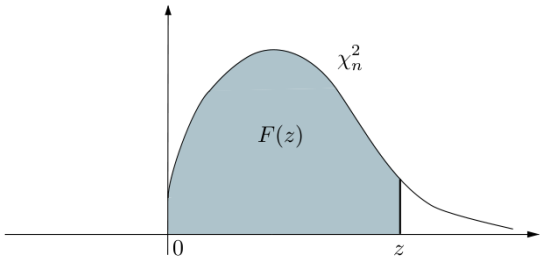
Que la població sigui normal vol dir que tenim $X \sim N(\mu, \sigma^2)$. Aleshores, es compleix això:

- $\overline{X}$ és també normal. Per tant, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- $\overline{X}$ i S^2 són variables aleatòries independents.
- S^2 és una variable aleatòria de tipus chi-quadrat amb $n-1$ graus de llibertat. Amb més exactitud:

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

Això es coneix com a teorema de Fisher.

La llei khi-quadrat



$F(z)\rightarrow$ $n\downarrow$	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
1	0.0002	0.0010	0.0039	0.0158	2.7055	3.8415	5.0239	6.6349
2	0.0201	0.0506	0.1026	0.2107	4.6052	5.9915	7.3778	9.2103
3	0.1148	0.2158	0.3518	0.5844	6.2514	7.8147	9.3484	11.3449
4	0.2971	0.4844	0.7107	1.0636	7.7794	9.4877	11.1433	13.2767
5	0.5543	0.8312	1.1455	1.6103	9.2364	11.0705	12.8325	15.0863
6	0.8721	1.2373	1.6354	2.2041	10.6446	12.5916	14.4494	16.8119
7	1.2390	1.6899	2.1673	2.8331	12.0170	14.0671	16.0128	18.4753
8	1.6465	2.1797	2.7326	3.4895	13.3616	15.5073	17.5345	20.0902
9	2.0879	2.7004	3.3251	4.1682	14.6837	16.9190	19.0228	21.6660
10	2.5582	3.2470	3.9403	4.8652	15.9872	18.3070	20.4832	23.2093
11	3.0535	3.8157	4.5748	5.5778	17.2750	19.6751	21.9200	24.7250
12	3.5706	4.4038	5.2260	6.3038	18.5493	21.0261	23.3367	26.2170
13	4.1069	5.0088	5.8919	7.0415	19.8119	22.3620	24.7356	27.6882
14	4.6604	5.6287	6.5706	7.7895	21.0641	23.6848	26.1189	29.1412

Suposeu (encara que no sigui del tot cert) que les notes de selectivitat segueixen un model teòric normal amb una mitjana de 6,4 punts i una desviació estàndard de 0,9 punts.
Si escollim 18 alumnes a l'atzar,
(a) Quina és la probabilitat que la mitjana de les seves notes sigui superior a 6,6?
(b) Quina és la probabilitat que la desviació típica de les notes d'aquests 18 alumnes sigui superior a 1,2?

X = nota de selectivitat: $X \sim N(6.4, 0.9^2)$.
$\overline{X}$ = nota mitjana d'una mostra de 18 alumnes: $\overline{X} \sim N(6.4, 0.045)$.
S^2 = var. de les notes d'una mostra de 18 alumnes: $S^2 \sim \frac{0.81}{17}\chi^2_{17}$.

$$P(\overline{X} > 6.6) = P\left(Z > \frac{6.6 - 6.4}{\sqrt{0.045}}\right) \approx 17\%$$

$$P(S^2 > 1.2^2) = P(\chi^2_{17} > \frac{1.2^2 \times 17}{0.81}) \approx 2.5\%$$

W. S. Gosset
(*"Student"*)
1876–1937

Si $X \sim N(\mu, \sigma^2)$, sabem que $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$. Per tant,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

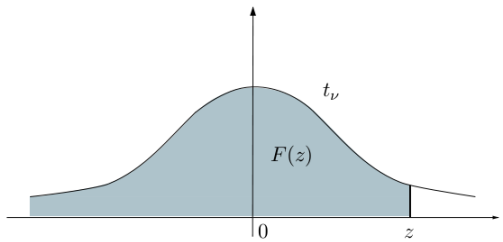
Problema: En general, σ és desconegut!
El que sí que podem calcular és

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

però aquesta variable no té distribució normal.
Té una distribució que s'assembla força a la normal: La distribució t de Student amb $n-1$ graus de llibertat.

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

La llei t de Student



$F(z)\rightarrow$ / $\nu\downarrow$	0.900	0.950	0.975	0.990	0.995	0.999
1	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076
7	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853
8	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247
12	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296
13	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520
14	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874

24	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852
40	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317
120	1.2886	1.6577	1.9799	2.3578	2.6174	3.1595
∞	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902

Donar un interval de confiança per a la mitjana mostral μ (per exemple) és:

- Prendre una mostra $X_1, \ldots, X_n$ aleatòria i calcular la seva mitjana mostral $\overline{X}$.
- Trobar un interval simètric al voltant de $\overline{X}$:

$$(\overline{X} - \epsilon, \overline{X} + \epsilon),$$

- de manera que es pugui afirmar, amb un nivell de confiança γ, que

$$\mu \in (\overline{X} - \epsilon, \overline{X} + \epsilon)$$

  amb probabilitat γ.
- Aquest nivell de confiança és, normalment, del 95% o del 99%, a la pràctica.

És a dir, estem dient que $P(\overline{X} - \epsilon < \mu < \overline{X} + \epsilon) = \gamma$

Exemple: Analitzem $n = 25$ aus per estudiar l'efecte d'un cert pesticida en la disminució del seu pes i afirmem, amb un 95% de confiança, que aquesta disminució es troba entre un 8% i un 12%.

- Si augmentem el nivell de confiança γ, creix també la mida de l'interval.
- Si disminuïm el nivell de confiança γ, disminueix també la mida de l'interval.

Per què?

Lògic! Si augmentem el nivell de confiança, vol dir que volem estar més segurs de que el que diem és cert, i aleshores, hem de donar un interval més gran. I viceversa, si acceptem d'equivocar-nos en més ocasions, podem donar intervals més petits.

Estudiarem aquests intervals de confiança:

- Per a la mitjana.
  - amb variància poblacional coneguda;
  - amb variància poblacional desconeguda.
- Per a la variància.
- Per a la proporció.
- Per a la recta de regressió.

Fonament teòric

$$X \sim N(\mu, \sigma^2) \implies \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Les fórmules

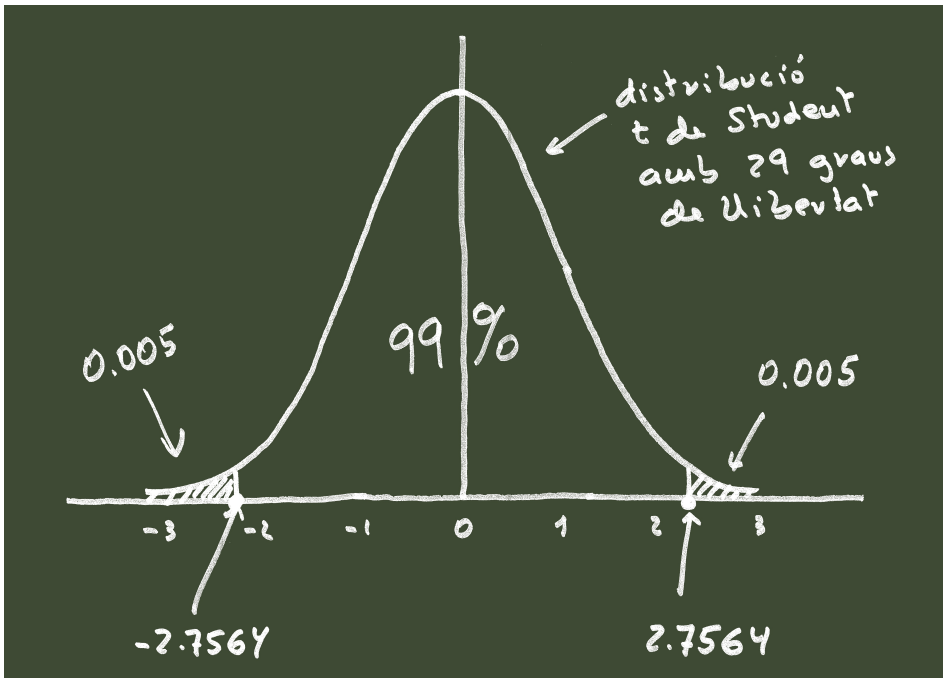
$$\epsilon = \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}$$

$$\epsilon = \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}$$

- Z_β significa aquell valor de Z que deixa per sota una àrea igual a β.
- $\alpha = 1 - \gamma$ és el risc d'equivocar-nos.
- La primera fórmula és per quan coneixem σ.
- L'interval és, sempre, $(\overline{X} - \epsilon, \overline{X} + \epsilon)$.

Si X no és normal, o no sabem si ho és o no, pel teorema central del límit, si n és gran, podem utilitzar, de manera aproximada, aquestes mateixes fórmules.

Suposem que volem estimar el temps d'absorció intestinal al 50% d'una certa substància i suposem que aquest temps és una variable aleatòria normal de mitjana (desconeguda) μ i variància (desconeguda) σ^2. Observem aquest temps d'absorció en 30 casos i tenim una mitjana (mostral) de $\overline{X} = 12.3$ minuts i una variància (mostral) de $S^2 = 2.7$ minuts^2. Volem un interval de confiança al 99% per a μ.



$\gamma = 0.99$; $\alpha = 0.01$;
$t_{29,0.005} = -2.7564$;

$$\epsilon = \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \approx 0.83$$

Per tant, $\mu \in (11.47, 13.13)$ amb un nivell de confiança del 99%.

1. En una ciutat s'ha mesurat la concentració de diòxid de carboni a una zona propera a l'aeroport. S'han obtingut aquests resultats (parts per milió): 102.2, 98.4, 104.1, 101.0, 102.2, 100.4, 98.6, 88.2, 78.8, 83.0, 84.7, 94.8, 105.1, 106.2, 111.2, 108.3, 105.2, 103.2, 99.0, 98.8. Trobeu un interval de confiança al 95% per a la concentració mitjana de monòxid de carboni a aquesta zona.
2. En un estudi sobre els mamífers d'una certa reserva natural, volem conèixer el pes mitjà de les guineus adultes d'aquest hàbitat. Si volem obtenir un interval de confiança d'aquest valor mitjà amb nivell de confiança del 90% i amb una amplitud d'una unitat, quants individus haurem de pesar? Se suposa que una bona estimació de σ és 1.2 kg.

Fonament teòric

$$X \sim N(\mu, \sigma^2) \quad \Longrightarrow \quad \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

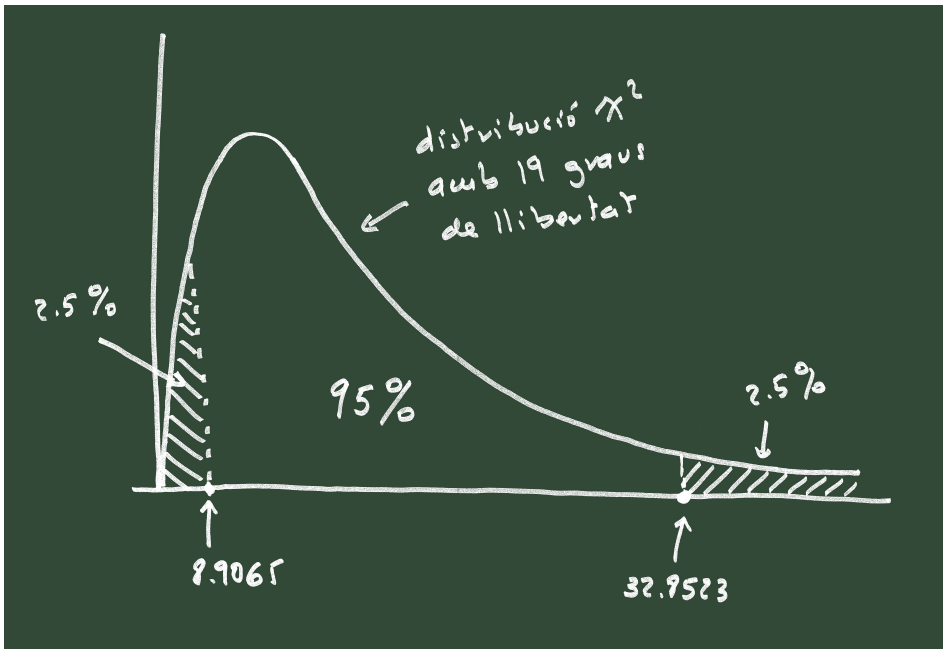
La fórmula

$$\sigma^2 \in \left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \right)$$

amb probabilitat $\gamma = 1 - \alpha$.

- $\chi^2_{n-1,\beta}$ significa aquell valor de χ^2_{n-1} que deixa per sota una àrea igual a β.
- $\alpha = 1 - \gamma$ és el risc d'equivocar-nos.

En un cert fàrmac, interessa controlar que la quantitat de principi actiu en cada dosi sigui el més uniforme possible. Estudiem 20 dosis d'aquest fàrmac i observem una variància mostral de $S^2 = 0.0018$ (en unes certes unitats). Volem un interval de confiança al 95% per a la desviació típica σ.



$\gamma = 0.95$; $\alpha = 0.05$;

$\chi^2_{19,0.025} = 8.9065$;

$\chi^2_{19,0.975} = 32.8523$

Per tant, $\sigma \in (0.032, 0.062)$ amb un nivell de confiança del 95%.

Volem trobar un interval de confiança per a una proporció: proporció de maletes perdudes, proporció d'intenció de vot d'un partit polític, proporció de pacients que mostraran una certa reacció a un fàrmac...

- En una població hi ha una proporció p (desconeguda) d'individus amb una certa característica.
- Prenem una mostra de mida n. A la mostra hi haurà una certa proporció $\hat{p}$ d'individus amb la característica i una proporció complementària $\hat{q} = 1 - \hat{p}$ d'individus sense la característica.
- $n\hat{p}$ és una variable aleatòria de tipus binomial. L'aproximem per una normal

$$n\hat{p} \approx N(np, npq)$$

- Per tant,

$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \approx N(0, 1)$$

- No és possible calcular aquest interval exactament (de fet, no és un interval).

Dues opcions de càlcul aproximat de l'interval de confiança per a una proporció:

Aproximem $\frac{pq}{n} \approx \frac{\hat{p}\hat{q}}{n}$.
Obtenim:

$$\epsilon = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Observem que $pq \leq \frac{1}{4}$:

$$\epsilon = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{4n}}$$

Aquest mètode ens permet determinar la mida de la mostra en funció de la longitud de l'interval que volem:

$$n \geq \left(\frac{Z_{1-\frac{\alpha}{2}}}{2\epsilon} \right)^2$$

Volem estimar la proporció d'habitants d'un municipi que estarien a favor de l'ampliació d'una zona de vianants. S'enquesten 400 persones i 140 es manifesten a favor. Treballem amb un nivell de confiança del 95%.

$$\hat{p} = \frac{140}{400} = 0.35$$

$$\epsilon = Z_{0.025}\sqrt{\frac{\hat{p}\hat{q}}{n}} \approx 0.0467$$

L'interval de confiança al 95% és $(0.303, 0.397)$

Quantes persones hauríem d'enquestar si volem un error màxim del 2%?

$$n \geq \left(\frac{1.96}{2 \times 0.02}\right)^2 = 2401$$

Recordem que havíem estudiat la recta de regressió. Repassem-ho:

- Una variable independent X i una variable Y (presumptament) dependent linealment de X.
- Una mostra de n valors (x_i, y_i).
- A partir d'aquí trobem una recta $y = a + bx$ tal que les diferències entre $a + bx_i$ i y_i són mínimes. Els coeficients a, b es determinen així:

$$b = \frac{\text{Cov}(X, Y)}{S_X^2}, \quad a = \overline{Y} - b\overline{X}$$

- En tot això no hi intervé la probabilitat.

El que volem fer ara és:

1. Entenem a i b com estimacions puntuals d'uns paràmetres desconeguts.
2. Volem trobar intervals de confiança per a a i b, amb una certa confiança α fixada: $a \pm \epsilon_1$, $b \pm \epsilon_2$.

Procedim d'aquesta manera:

1. Considerem els errors $e_i := y_i - (a + bx_i)$.
2. Considerem $\hat{\sigma}^2 = (\sum e_i^2)/(n-2)$.
3. L'interval de confiança per a b és

$$b \pm t_{1-\frac{\alpha}{2}}^{n-2} \sqrt{\frac{\hat{\sigma}^2}{nS_X^2}}.$$

4. L'interval de confiança per a a és

$$a \pm t_{1-\frac{\alpha}{2}}^{n-2} \sqrt{\frac{\hat{\sigma}^2}{n}\left(1 + \frac{\overline{X}^2}{S_X^2}\right)}.$$

Interval de predicció

- Suposem que tenim n observacions (x_i, y_i) i determinem la recta de regressió $y = a + bx$.
- Suposem que tenim un valor x i volem utilitzar la recta de regressió per predir el valor de y que li pertocaria.
- Evidentment, el valor és $y := a + bx$ però, quin seria l'interval de confiança per a y (=Interval de predicció)?

$$y \pm t^{n-2}_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{n}\left(n + 1 + \frac{(x - \overline{X})^2}{S_X^2}\right)}$$

Curs de Bioestadística

Jaume Aguadé

Capítol 8. Tests d'hipòtesis

Tenim una població X amb una certa distribució i volem contrastar una certa hipòtesi sobre els paràmetres d'aquesta distribució.

1. El fabricant afirma que les seves bombetes de baix consum tenen una vida mitjana de 2000 hores. En una mostra de 103 d'aquestes bombetes trobem una vida mitjana de 1832 hores. Hi ha prou evidència per a refutar l'afirmació del fabricant?
2. La Unió Europea ha fixat en 0.1 μg/l el límit màxim d'acrilamides a l'aigua potable. Si en 23 mostres independents d'aigua d'una certa població hem trobat una mitjana de 0.25 μg/l, podem afirmar que l'aigua d'aquesta població no compleix, en mitjana, la normativa europea?
3. Comparem una certa gamma A de productes d'ús cutani amb una nova gamma B que s'afirma que és hipoal·lergènica. A una mostra de 100 individus tractats amb els productes A s'han donat 10 casos d'al·lèrgia i a una mostra de 79 individus tractats amb els productes B s'han donat 5 casos d'al·lèrgia. Hi ha prou evidència per afirmar que la gamma B produeix realment menys al·lèrgies que la gamma A?

1. En una certa marca d'injectables, és crític que la desviació típica en la quantitat de principi actiu en cada injectable sigui molt petita. Si una mostra aleatòria de 50 injectables ha donat una desviació de 0.08, podem acceptar que la desviació típica dels injectables no superarà 0.02?
2. En un estudi sobre l'efecte de l'oxitocina en la pressió sanguínia, s'ha injectat aquesta hormona a 10 persones i s'han registrat les seves pressions sistòliques abans i després de la injecció. A partir d'aquest registre, com podem decidir si l'oxitocina afecta la pressió sanguínia?
3. En una enquesta recent, 54 de 200 enquestats afirma que té un extintor a casa. Quan es va fer aquesta enquesta fa uns anys, 30 de 150 enquestats deia que tenia un extintor. Podem afirmar que el percentatge de llars amb extintor ha crescut?

Hipòtesis

Hipòtesi nul·la H_0: És la hipòtesi que acceptarem si no hi ha prou evidència en contra. ("Presumpció d'innocència!")
Hipòtesi alternativa H_1: És la hipòtesi que contradiu H_0 i que acceptarem si hi ha prou evidència estadística a favor seu.

Quines són H_0 i H_1 en els exemples anteriors?
Observeu que H_0 és la hipòtesi "conservadora" i H_1 és la hipòtesi "revolucionària". No juguen un paper simètric. H_1 no conté "=".

Tipus d'errors

Error de tipus I: Acceptar H_1 quan H_1 és fals.
Error de tipus II: No acceptar H_1 (és a dir, quedar-nos amb H_0) quan H_1 és cert.

Intentarem controlar els errors de tipus I perquè entenem que són els més perillosos.

Com que no podrem controlar simultàniament els dos tipus d'errors, controlarem la probabilitat de cometre un error de tipus I:

Nivell de significació

El nivell de significació α d'un test estadístic és la probabilitat de cometre un error de tipus I. Normalment, es treballa amb valors $\alpha = 0.05, 0.01, 0.001$ etc., en funció del perill que representi cometre un error de tipus I.

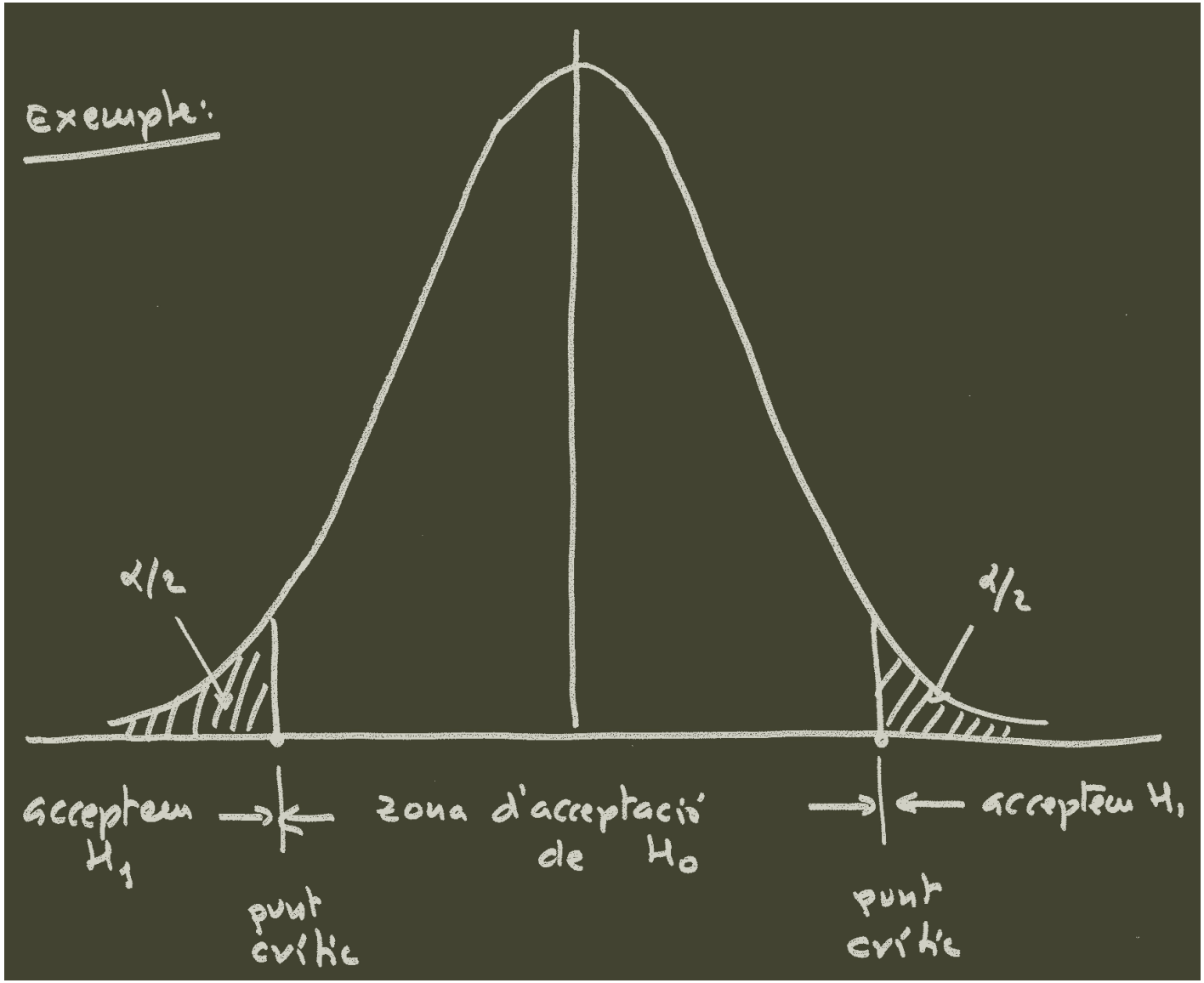
Un test té sempre dues conclusions possibles:

- Acceptem H_1, és a dir, creiem que H_1 és cert i tenim una probabilitat α d'equivocar-nos.
- Rebutgem H_1 i ens quedem amb H_0 perquè si acceptéssim H_1 amb l'evidència que tenim, la probabilitat d'equivocar-nos seria massa gran (superior a α).

(Penseu en el cas d'un judici: innocent, culpable)

En línies generals, un test estadístic es desenvolupa en aquestes fases:

1. Calculem un cert estadístic θ, és a dir, un cert valor que s'obté a partir de les dades de la mostra. θ depèn del tipus de test que estem fent: test per a la mitjana, test per a la variància, etc.
2. Sabem la distribució teòrica Θ d'aquest estadístic. Per exemple, podem saber que Θ té una distribució normal.
3. En aquesta distribució teòrica determinem, a partir de α, la zona d'acceptació de H_0 i la zona d'acceptació de H_1, que estan separades pels valors crítics de Θ.
4. Mirem a quina de les dues zones es troba θ i arribem a una conclusió.


Exemple:
α/2
α/2
acceptem H_1
zona d'acceptació de H_0
acceptem H_1
punt crític
punt crític

Test per a la mitjana d'una població normal $X \sim N(\mu, \sigma^2)$ amb σ coneguda.

- Suposem que volem contrastar una hipòtesi alternativa sobre μ del tipus $H_1 : \mu \neq \mu_0$ contra la hipòtesi nul·la $H_0 : \mu = \mu_0$. Tenim fixat un nivell de significació del test igual a α.
- Prenem una mostra de mida n i calculem la seva mitjana mostral $\overline{X}$.
- Acceptarem que $\mu \neq \mu_0$ si $\overline{X}$ està "molt lluny" de μ. En cas contrari, ens quedarem amb H_0.
- Per decidir què vol dir "molt lluny", utilitzarem que $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ i, per tant, l'estadístic a utilitzar és:

Estadístic a utilitzar en un Z-test

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Tests unilaterals

Si H_1 és del tipus $\mu < \mu_0$ o bé $\mu > \mu_0$, tenim un test unilateral o test amb una cua: la superfície α estarà tota situada a un únic costat de la corba de Gauss i hi haurà un únic valor crític. Si H_1 és del tipus $\mu \neq \mu_0$, tenim un test bilateral o test amb dues cues.

El p-valor

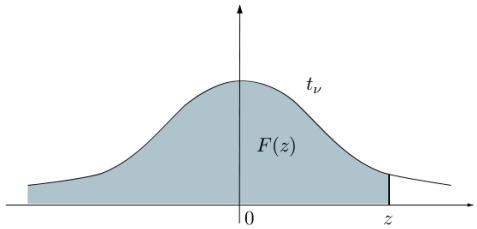
En lloc de fixar el nivell de significació α, també podem calcular el valor de l'estadístic z_0 i la probabilitat que Z superi aquest valor. D'aquesta probabilitat s'en diu el p-valor del test. És el nivell de significació que ens duria a no saber decidir entre H_0 i H_1.

El t-test

Si no coneixem σ, utilitzarem la variància mostral S en lloc seu, però aleshores, haurem d'utilitzar, en lloc de la normal Z, la distribució t de Student amb $n-1$ graus de llibertat:

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t^{n-1}$$

La llei t de Student



$F(z)\rightarrow$ / $\nu\downarrow$	0.900	0.950	0.975	0.990	0.995	0.999
1	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076
7	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853
8	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247
12	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296
13	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520
14	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874

Suposem que fem un test sobre una hipòtesi H_1. Calculem l'estadístic de contrast i obtenim un cert valor Θ.

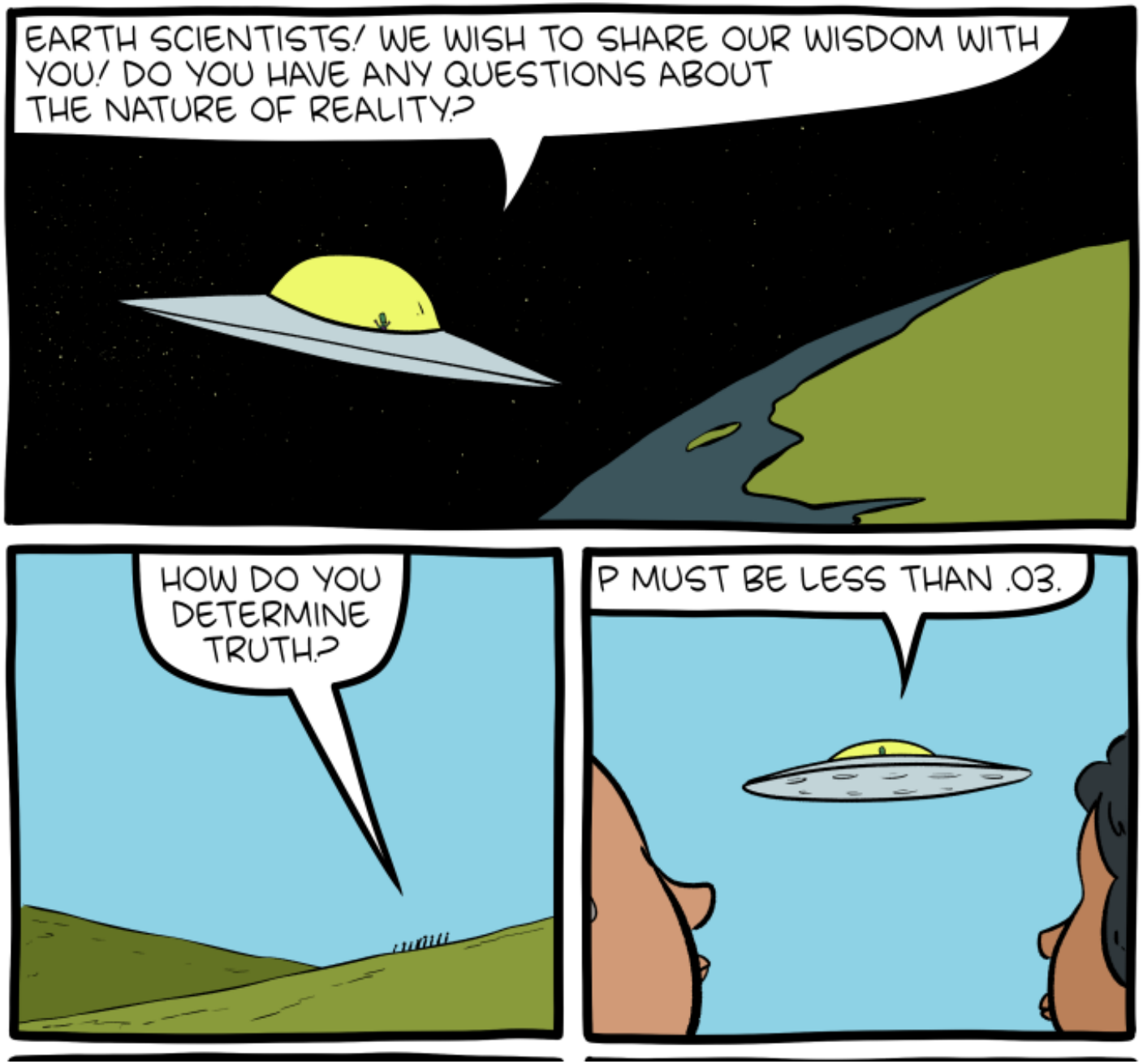
El p-valor

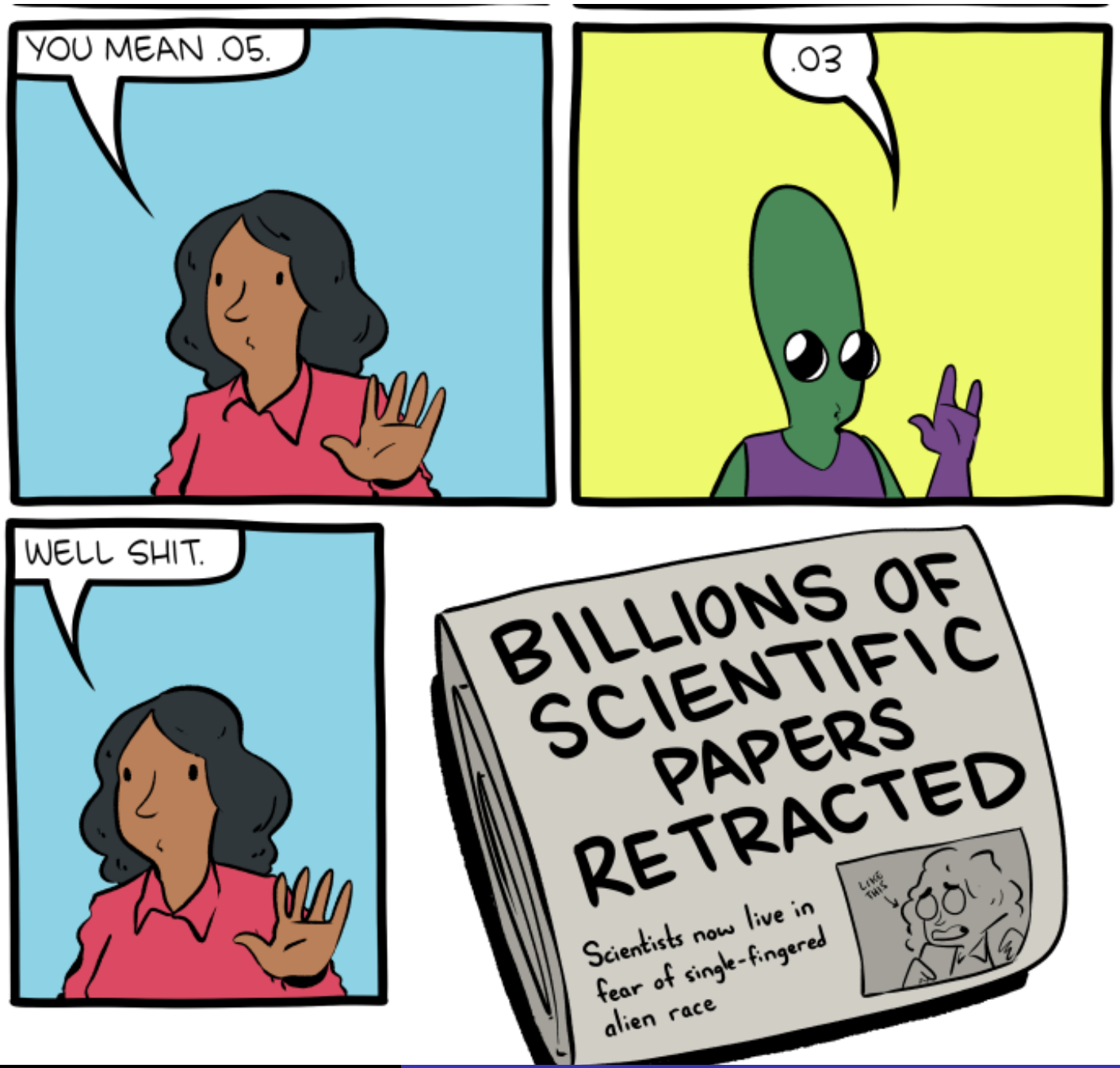
El p-valor del test és la probabilitat d'obtenir aquest valor Θ, o un de més extrem, suposant que la hipòtesi H_0 fos certa.

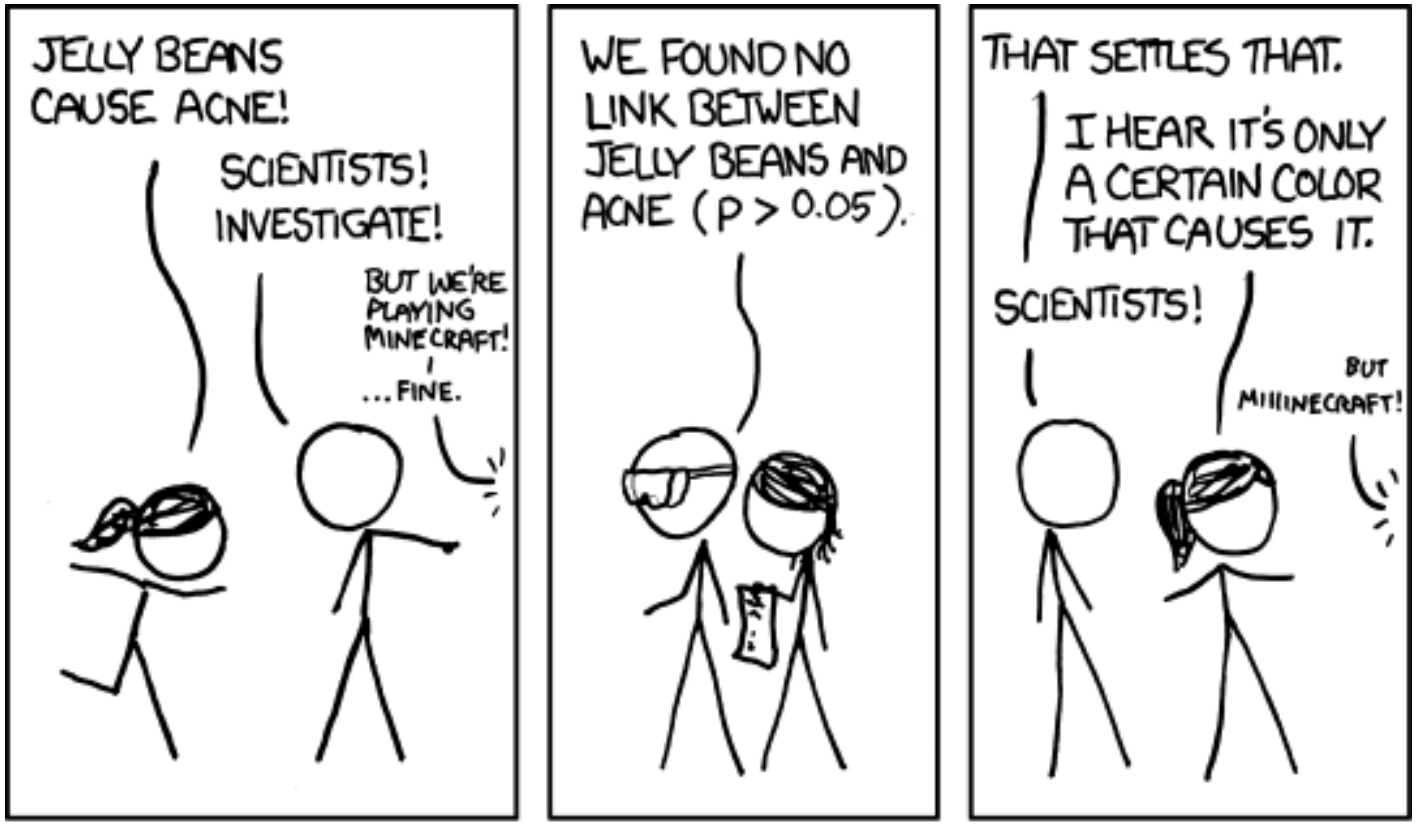
- Com més petit és el p-valor, més improvable és que la hipòtesi nul·la sigui certa.
- Com més petit és el p-valor, més confiança podem tenir que la hipòtesi H_1 sigui certa.
- Segons el context, es considera que un p-valor $p < 0.05$ dóna plausibilitat a la hipòtesi H_1.
- En altres contextos, exigirem $p < 0.01$, $p < 0.001$, etc. per validar una hipòtesi H_1.

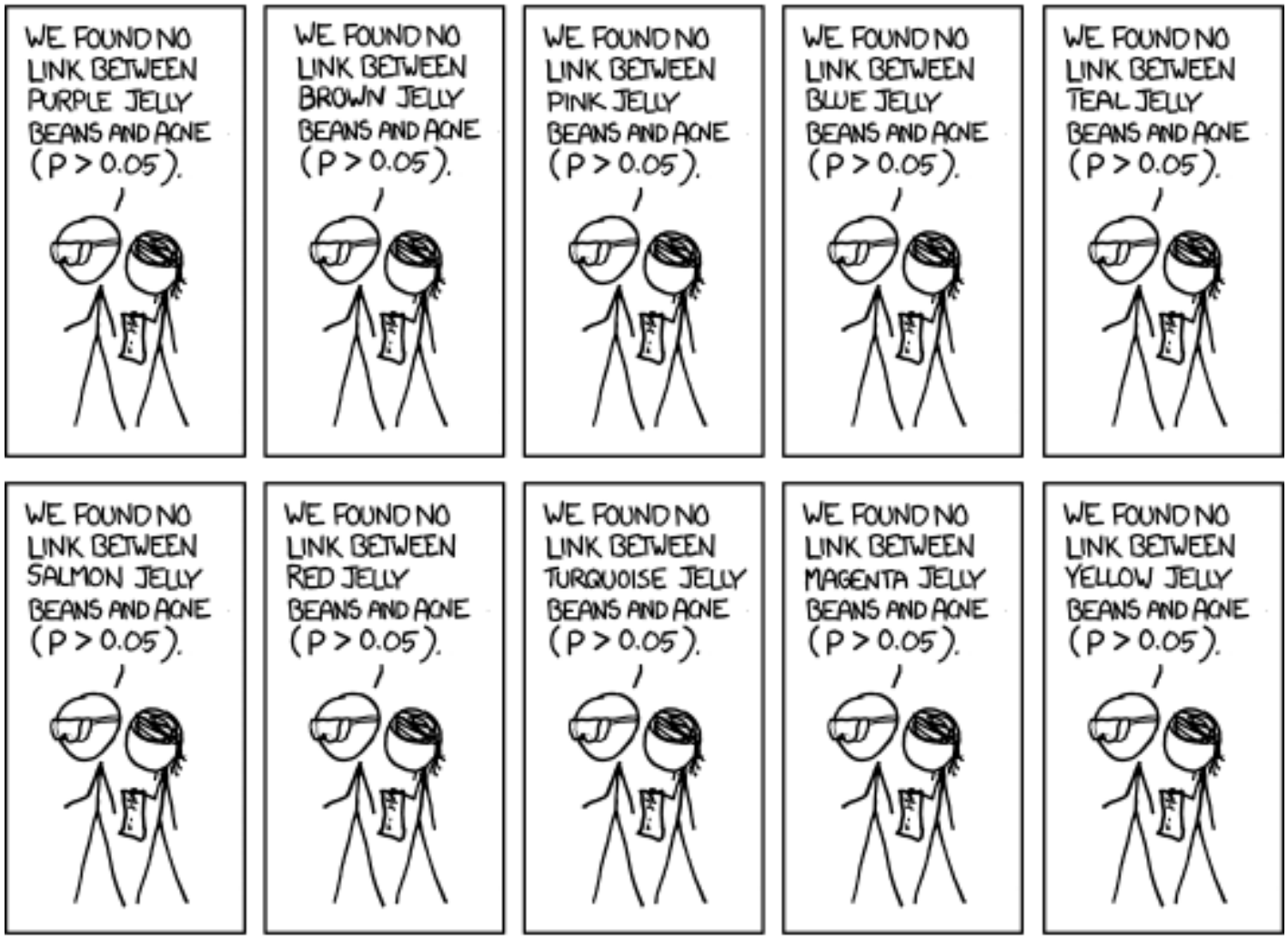
Llancem una moneda 100 vegades. No tenim cap motiu per dubtar que cara i creu tenen la mateixa probabilitat de sortir (hipòtesi nul·la).

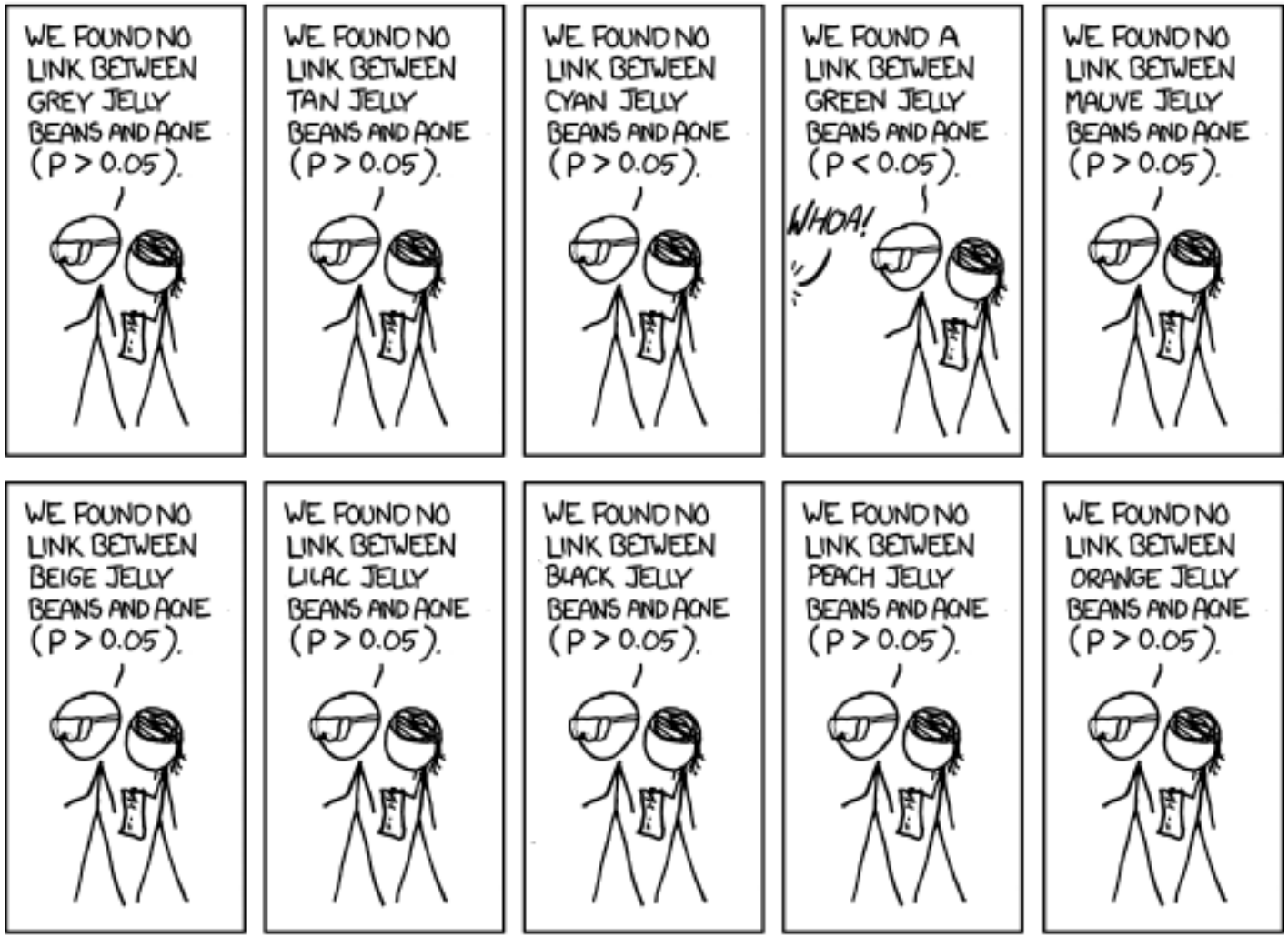
- Observem que, per exemple, obtenim 60 cares i 40 creus. Potser dubtarem que la moneda és equitativa? Depèn.
  - No és impossible llançar 100 vegades una moneda equitativa i obtenir 60 cares i 40 creus...
  - ... però no és gaire probable. Si fem els càlculs (distribució binomial!) veiem que la probabilitat d'una diferència de 20 o més entre cares i creus és força petita: $p = 0,05689$. Això és el p-valor del test. Com que $p > 0.05$, en principi no afirmarem que la moneda està desequilibrada (però podem començar a sospitar).
- Si, en canvi, obtenim 80 cares i 20 creus, sí que tenim motius per refusar la hipòtesi que la moneda està ben equilibrada: la probabilitat que hi hagi una diferència de 60 o més entre el nombre de cares i el de creus és petitíssima. $p = 0,000000001$. No és impossible que surtin 80 cares, però la probabilitat que això passi amb una moneda justa és tant petita que, científicament (no matemàticament) afirmarem que la moneda no està equilibrada.

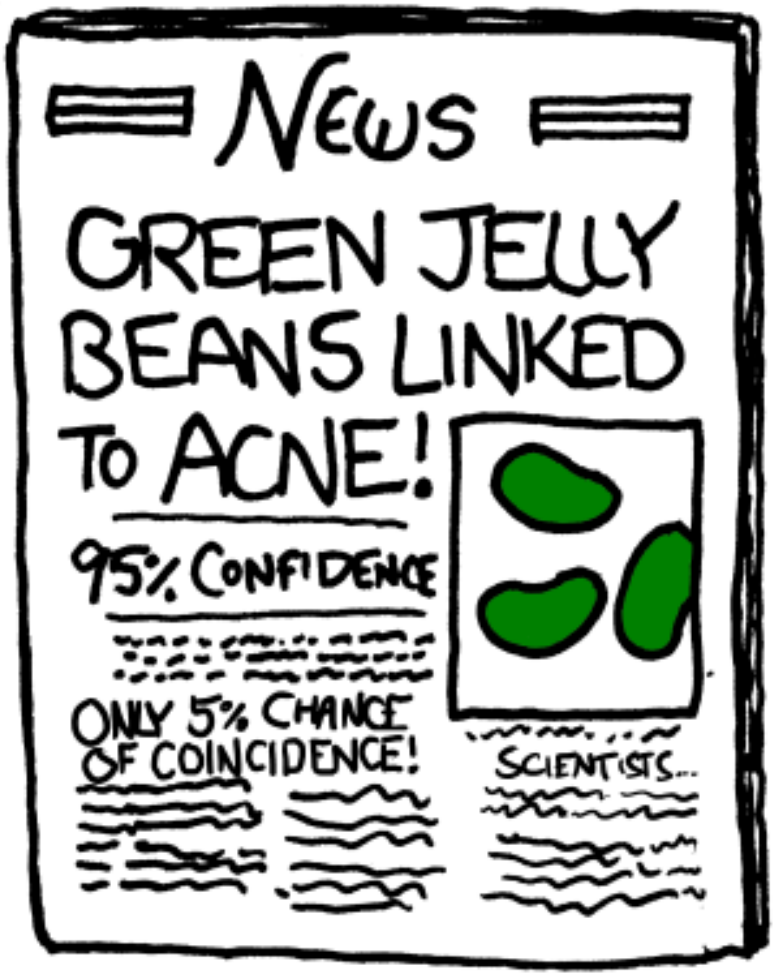

EARTH SCIENTISTS! WE WISH TO SHARE OUR WISDOM WITH YOU! DO YOU HAVE ANY QUESTIONS ABOUT THE NATURE OF REALITY?
HOW DO YOU DETERMINE TRUTH?
P MUST BE LESS THAN .03.


YOU MEAN .05.
.03
WELL SHIT.
BILLIONS OF SCIENTIFIC PAPERS RETRACTED
Scientists now live in fear of single-fingered alien race
LIKE THIS


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ($p > 0.05$).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE ($P < 0.05$).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE ($P > 0.05$).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE ($P > 0.05$).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Suposem que estem estrudiant dues variables X, Y i volem saber si hi ha una bona correlació lineal entre elles. Prenem una mostra de mida n

$$(x_1, y_1), \ldots, (x_n, y_n)$$

i calculem el coeficient de correlació r.
Sabem que si $|r|$ és proper a 1, hi ha bona correlació. Però... què vol dir "bona"? Es pot quantificar com és de "bona"?
Sí, es pot quantificar amb el p-valor:

p-valor per a r

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t^{n-2}.$$

En un estudi s'afirma que l'al·ligàtor del Yangtze (*Alligator sinensis*) adut té un pes mitjà de 40 kq, però nosaltres creiem que el pes mitjà és, en realitat, inferior. Volem contrastar aquesta afirmació amb un nivell de significació del 5%. Escollim una mostra de 12 espècimens i mesurem el seu pes. Obtenim (kg):

36.1, 40.2, 33.8, 38.5, 42.0, 35.8, 37.0, 41.0, 36.8, 37.2, 33.0, 36.0

Mostra

$\overline{X} = 37.28$
$S = 2.73$
$n = 12$

Hipòtesis

$H_0 : \mu = 40$
$H_1 : \mu < 40$

Estadístic de contrast

$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = -3.44$. Comparem amb la distribució t de Student amb 11 graus de llibertat. El test és unilateral amb una única cua a l'esquerra. El valor crític és -1.7959.

Conclusió

-3.44 es situa a la zona d'acceptació de H_1. Per tant, acceptem que el pes mitjà dels al·ligàtors del Yangtze és inferior a 40 kg, amb significació del 5%, i neguem la validesa de l'estudi anterior. Quin és el p-valor?

Cas de dades aparellades

Si tenim dues variables X_1, X_2 sobre uns mateixos individus i volem fer un test de comparació de mitjanes, podem considerar una nova variable $D = X_1 - X_2$ i fer un test ordinari per a la mitjana de D (t-test).

Per estudiar si la pràctica d'exercici físic fa reduir el ritme cardíac prenem vuit voluntaris que no fan exercici físic i els mesurem el ritme cardíac abans i després de seguir un programa d'exercici físic d'un mes de durada. Hem obtingut:

individu	1	2	3	4	5	6	7	8
ritme abans del programa	74	86	98	102	78	84	79	70
ritme després del programa	70	85	90	110	71	80	69	74

Hi ha prou evidència per concloure que l'exercici físic comporta una reducció del ritme cardíac, amb un nivell de significació del 5%?

individu	1	2	3	4	5	6	7	8
ritme abans del programa	74	86	98	102	78	84	79	70
ritme després del programa	70	85	90	110	71	80	69	74
$D = X_a - X_d$	4	1	8	-8	7	4	10	-4

Mostra

$\overline{D} = 2.75$
$S = 6.16$
$n = 8$

Estadístic de contrast

$t = \frac{\overline{D}}{S/\sqrt{n}} = 1.26$. Comparem amb la distribució t de Student amb 7 graus de llibertat. Test unilateral amb una única cua a la dreta. El valor crític és 1.89.

Hipòtesis

$H_0 : \overline{d} = 0$
$H_1 : \overline{d} > 0$

Conclusió

1.26 es situa a la zona d'acceptació de H_0. Per tant, amb un nivell de significació del 5% no tenim prou evidència per acceptar que l'exercici físic baixi el ritme cardíac.

Variàncies conegudes

Utilitzem l'estadístic $\dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Variàncies desconegudes iguals

Utilitzem l'estadístic

$$\frac{\overline{X_1} - \overline{X_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t^{n_1+n_2-2}; \quad S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2$$

Variàncies desconegudes, mostres grans ($n_1 > 30$, $n_2 > 30$)

Podem fer un test aproximat:

$$\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1)$$

Estem estudiant l'efecte d'una dieta pobre en calci sobre la concentració de plom a la sang i els teixits en ratolins. En el grup de control format per 50 animals, hem trobat una concentració de plom amb $\overline{X_1} = 5.2$ i $S_1 = 1.1$ (en unes certes unitats) mentre que el grup experimental amb una dieta pobre en calci, format per 40 animals, la concentració observada de plom verifica $\overline{X_2} = 6.1$ i $S_2 = 1.3$. Podem concloure que la manca de calci fa créixer la concentració de plom? (nivell de significació $\alpha = 0.01$)

Com que les mostres són grans, encara que no coneguem les variàncies poblacionals, podem utilitzar aquest estadístic:

$$\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{5.2 - 6.1}{\sqrt{\frac{1.21}{50} + \frac{1.69}{40}}} = -3.49$$

que hem de comparar amb la distribució normal standard. H_1 és $\mu_2 > \mu_1$ i H_0 és $\mu_2 = \mu_1$. És un test unilateral amb una cua a l'esquerra.El valor crític és $Z_{0.01} = -2.32$. Per tant, -3.49 cau a la zona d'acceptació de H_1.

Poblacions no necessàriament normals

Si les mostres són prou grans, el teorema central del límit ens diu que la mitjana mostral és aproximadament normal. Per tant, podem utilitzar el $Z-$test com si la població fos normal.

Si X és normal, $X \sim N(\mu, \sigma^2)$, sabem que $(n-1)S^2/\sigma^2$ té una distribució χ^2_{n-1}. Per tant:

Test per a la variància

$H_0 : \sigma^2 = \sigma_0^2$; $H_1 : \sigma^2 \neq \sigma_0^2$. Estadístic:

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

També hi ha tests unilaterals.

La màquina que controla la quantitat de fil que hi ha en una bobina treballa correctament si la desviació estàndard de la quantitat de fil en cada bobina és inferior a 0.15 cm. En una mostra de 20 bobines hem observat una variància mostral de 0.025 cm^2. Podem concloure que la màquina no està treballant correctament?

Aquest test utilitza la distribució $F(n, m)$ de Fisher que té dos graus de llibertat, n i m.
$X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$. Prenem una mostra de X_1 de mida n_1 i una mostra de X_2 de mida n_2. Aleshores:

Test de comparació de variàncies

$H_0 : \sigma_1^2 = \sigma_2^2$; $H_1 : \sigma_1^2 \neq \sigma_2^2$. Estadístic:

$$\frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

També hi ha tests unilaterals.

A l'hora de calcular els valors crítics de la distribució de Fisher, és útil aquesta propietat:

$$F(n, m)_\alpha = \frac{1}{F(m, n)_{1-\alpha}}$$

Volem veure si el nou model de màquina omplidora d'ampolles (N) presenta menys variabilitat que la màquina antiga (A). Prenem 22 mostres d'ampolles omplertes per la màquina A i observem una variància de 0.0018, mentre que en una mostra de 25 ampolles de la màquina B s'observa una variància de 0.0008. Feu un test amb $\alpha = 0.01$.

Ara X és una població Bernoulli amb una proporció desconeguda p d'èxits. Volem fer un test per a p. Prenem una mostra de mida n i trobem una proporció mostral $\widehat{p}$.

Test per a una proporció (aproximat)

$H_0 : p = p_0$; $H_1 : p \neq p_0$. (També unilateral) Estadístic:

$$\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0,1)$$

Test de comparació de proporcions (aproximat)

$H_0 : p_1 = p_2$; $H_1 : p_1 \neq p_2$. (També unilateral) Estadístic:

$$\frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1} + \frac{\overline{p}(1-\overline{p})}{n_2}}} \approx N(0,1), \quad \overline{p} = \frac{n_1\widehat{p_1} + n_2\widehat{p_2}}{n_1 + n_2}$$

Hem desenvolupat una nova crema solar (N) i volem contrastar que produeix menys reaccions al·lèrgiques que la crema antiga (A). En una mostra de 100 individus que han usat la crema A s'han donat 22 casos d'al·lèrgia. En 90 individus que han usat la crema N s'han donat 10 casos d'al·lèrgia. Hi ha prou evidència (amb un nivell de significació del 5%) per afirmar que la nova crema és menys al·lergènica que l'antiga?

Usem l'estadístic:

$$\frac{\widehat{p_A} - \widehat{p_N}}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1} + \frac{\overline{p}(1-\overline{p})}{n_2}}} = \frac{0.22 - 0.11}{\sqrt{0.168 \times 0.83(0.01 + 0.011)}} \approx 2.0025$$

$H_1 : p_N < p_A$, $H_0 : p_N \geq p_A$: test unilateral amb una cua a la dreta.

Valor crític: $Z_{0.95} = 1.6449 < 2.0025$. Per tant, el valor de l'estadístic de contrast cau a la zona d'acceptació de H_1 i podem concloure que la nova crema produeix menys reaccions al·lèrgiques que l'antiga (amb un nivell de significació del 5%).

El famós Test d'independència de la χ^2 va ser introduït per Karl Pearson el 1900 i és una de les eines més utilitzades en estadística. Consisteix en això:

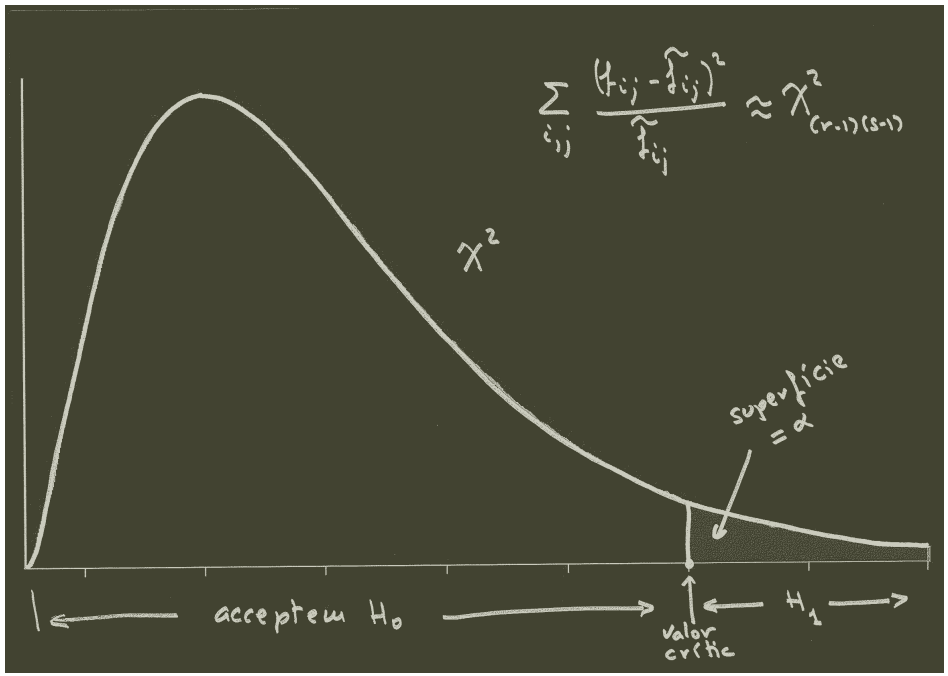
- Dues variables aleatòries nominals X i Y. X pren r valors i Y pren s valors.
- H_0 és "X i Y són independents". H_1 és "X i Y no són independents".
- Prenem una mostra de mida n. Per a cada individu tenim el valor de X i el valor de Y.
- Fem una taula de freqüències observades (taula de contingència).
- Calculem una taula de freqüències esperades.
- Utilitzem un estadístic que ens mesura la discrepància entre la taula de freqüències observades i la taula de freqüències esperades. Acceptarem H_1 si aquesta discrepància és gran (test unilateral amb cua a la dreta).

Freqüències

f_{ij} = freqüència de $X = i$, $Y = j$.
$f_{i\bullet}$ = freqüència de $X = i$.
$f_{\bullet j}$ = freqüència de $Y = j$.
$\widetilde{f}_{ij} = \frac{f_{i\bullet} \times f_{\bullet j}}{n}$ = freqüència esperada de $X = i$, $Y = j$.

Estadístic de contrast

$$\sum_{i,j} \frac{(f_{ij} - \widetilde{f}_{ij})^2}{\widetilde{f}_{ij}} \approx \chi^2_{(r-1)(s-1)}$$



Hipòtesis

H_0 : X i Y són independents.
H_1 : X i Y no són independents.

Cal que la mida de la mostra sigui gran i que les freqüències esperades no siguin massa petites, $\widetilde{f}_{ij} \geq 5$.

Un supermercat vol estudiar si hi ha relació entre l'edat dels clients i la seva satisfacció amb els productes de marca blanca. Una enquesta dóna aquests resultats:

	molt satisfet	satisfet	poc satisfet
menor de 40 anys	140	100	15
40 o més anys	140	50	20

1. Fem una taula de contingència i calculem les freqüències marginals. Calculem la taula de les freqüències esperades.

140	100	15	255
140	50	20	210
280	150	35	465

153.55	82.26	19.19
126.45	67.74	15.81

2. Calculem l'estadístic de contrast $D = 19.36$ i el comparem amb el valor crític de χ^2 amb 2 graus de llibertat. Si $\alpha = 0.05$, aquest valor crític val 5.99. Per tant, acceptem H_1: la satisfacció amb els productes de marca blanca no és independent de l'edat del client.

El cas més simple d'Anàlisi de la variància té aquest aspecte:

- Hi ha diferències significatives en el contingut en greixos de les hamburgueses de les cinc cadenes principals de menjar ràpid?
- S'analitzen tres metodologies d'ensenyament i ens preguntem si l'elecció d'una d'aquestes metodologies pot tenir una influència en l'efectivitat de l'ensenyament.
- L'encarregat d'un supermercat estudia la influència de l'elecció de quatre possibles localitzacions d'un producte sobre el volum de vendes d'aquest producte.
- Abans de decidir-nos a adquirir un determinat producte volem contrastar si hi ha diferències significatives entre el rendiment d'aquest producte en funció del fabricant.
- En un estudi clínic ens preguntem si hi ha diferències significatives entre els efectes de cinc marques d'un mateix fàrmac.

L'esquema general d'Anàlisi de la variància d'un factor és:

1. Tenim m variables aleatòries normals $X_i \sim N(\mu_i, \sigma^2)$. (Observeu que fem la hipòtesi de que les variàncies són totes iguals.) Cada valor de m és un nivell.
2. Prenem, de cada variable, una mostra de mida n: $X_{i,1}, \ldots, X_{i,n}$.
3. Fixem un cert nivell de significació α.
4. Volem contrastar la hipòtesi nul·la

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_m$$

Què són cada una d'aquestes coses en els exemples anteriors?
Observeu que l'ANOVA pot determinar que hi ha diferències significatives entre els nivells, però NO serveix per dir que un nivell és "millor" que els altres.

$x_{i,j}$ = valor de l'observació $j-$èsima del nivell i.
$N = mn$ és el nombre total d'observacions.
$\overline{x_{\bullet\bullet}} = \frac{\sum x_{i,j}}{N}$ és la mitjana total de totes les observacions.
$\overline{x_{i\bullet}} = \frac{\sum x_{i,j}}{n}$ és la mitjana de totes les observacions del nivell i.
El criteri que utilitzarem per decidir si podem refusar la hipòtesi nul·la és aquest:

- Dintre de cada nivell, hi haurà una certa variabilitat interna, deguda a l'atzar.
- També hi haurà una variabilitat entre nivells deguda a l'atzar i també, potser, a que la hipòtesi nul·la no és certa.
- Si la variabilitat entre nivells, comparada amb la variabilitat interna dels nivells, és molt gran, tindrem motius per acceptar la hipòtesi alternativa H_1

- Com a mesura de la "variabilitat interna" prenem

$$S_e^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} (x_{ij} - \overline{x_{i\bullet}})^2 \right)$$

- Com a mesura de la "variabilitat entre nivells" prenem

$$S_F^2 = \sum_{ij} (\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}})^2 = n \sum_{i} (\overline{x_{i\bullet}} - \overline{x_{\bullet\bullet}})^2$$

Utilitzem l'estadístic

$$\frac{S_F^2/(m-1)}{S_e^2/(N-m)} \sim F(m-1, N-m)$$

i acceptem la hipòtesi alternativa si aquest estadístic és "gran", és a dir, més gran que el valor crític de la distribució F corresponent al nivell de significació α.

Estudiem l'eficàcia de dos fàrmacs A i B i dos tipus de dieta C i D en la reducció de la hipertensió. Escollim 25 individus hipertensos i els distribuïm aleatòriament en 5 grups. Als quatre primers grups els apliquem els tractaments A, B, C i D i al cinquè grup no li apliquem cap tractament. Volem fer una anàlisi de la variància ($\alpha = 0.01$) per contrastar si hi ha diferències significatives en l'eficàcia dels diversos tractaments.

Les pressions arterials sistòliques dels 25 individus, després d'un mes de tractament, són:

A	158, 146, 160, 171, 155
B	147, 152, 143, 155, 160
C	172, 158, 167, 160, 175
D	163, 170, 158, 162, 170
E	180, 173, 175, 182, 181

- $n = m = 5$. $N = 25$.
- $S_F^2 = 2010.64$. Graus de llibertat $m - 1 = 4$.
- $S_e^2 = 894.40$. Graus de llibertat $N - m = 20$.
- Estadístic de contrast: $\frac{S_F^2/(m-1)}{S_e^2/(N-m)} = 11.24$
- Comparem amb el valor crític de la distribució $F(4, 20)$ per $\alpha = 0.01$, que val 4.43.
- Conclusió: Acceptem que hi ha diferències significatives en l'efectivitat dels diversos tractaments.