Estadística

Resum idees bàsiques

Joan del Castillo
2008

Estadística Descriptiva

- Conjunt de tècniques per organitzar, simplificar i resumir la informació continguda en un conjunt de dades.
- Les dades poden provenir de variables quantitatives o de variables categòriques.

Dades quantitatives

- Dades continues, les que varien de forma *continua*, com la temperatura, la resistència elèctrica, el nivell de radiació, etc.

- Dades discretes, les que només poden prendre valors enters com per exemple el nombre d'accessos a una determinada pàgina web, etc.

Dades qualitatives

Poden ser:

- Nominals: són una simple etiqueta com per exemple la marca, l'estat d'un fusible (fos/no fos), etc.
- Ordinals: en les que existeix una relació de ordre entre els possibles valors, per ex. el resultat d'una exploració qualitativa de la temperatura d'un aparell (fred, tebi, calent).

Recollida d'informació

Codi	Edat (anys)	Gènere	Pes	Alçada	Fumador
1	20	f	61	170	1
2	20	f	65	171	0
3	19	f	55	166	0
4	20	f	63	170	1
5	18	f	56	170	1
6	18	f	59	160	0
7	16	f	58	165	0
8	20	f	57	169	1
9	20	f	58	165	1
10	20	f	50	153	1
11	18	m	90	171	1
12	18	m	60	168	1
13	19	m	61	172	1
14	22	m	72	187	0
15	21	m	65	170	0

Primeres nocions

- Matriu de dades
  - Casos (files) i Variables (columnes).
  - Mida de la mostra = nombre de files = n.
- Classes de variables:
  - Discretes i Contínues.
  - Qualitatives i Quantitatives.
- Variables discretes:
  - Gènere, tabac, color del cabell, nombre de parts,...
- Variables contínues:
  - Colesterol en sang, pes, alçada,...

Descripció de variables:

- Variables contínues i quantitatives:
  - Exemples: Colesterol en sang, pes, alçada,...
  - Descripció: mitjanes, variància i desviacions.
- Variables discretes:
  - Exemples: Gènere, tabac, color del cabell, nombre de parts,...
  - Descripció: freqüències relatives.

Descripció. Variables discretes

- Notació
  - n = Tamany de mostra.
  - A una propietat, x, y = Variables.
- Freqüencies absolutes: $F(A)$
  - Nombre de casos amb una propietat.
- Freqüencies relatives: $f(A)$
  - Nombre de casos amb una propietat dividit pel nombre total de casos.

Enumeració dels elements

$$H = \{11,12,13,14,15\}$$
$$D = \{1,2,3,4,5,6,7,8,9,10\}$$
$$F = \{1,4,5,8,9,10,11,12,13\}$$
$$NF = \{2,3,6,7,14,15\}$$

$$\Omega = \{1,2,3,...,14,15\}$$

Unió i Intesecció de conjunts

- Intersecció:
  - Homes i fumadors $H \cap F = \{11,12,13\}$
  - Dones i fumadores $D \cap F = \{1,4,5,8,9,10\}$
- Unió:
  - Homes o fumadors
    $H \cup F = \{1,4,5,8,9,10,11,12,13,14,15\}$
  - Homes o no fumadors
    $H \cup NF = \{2,3,6,7,11,12,13,14,15\}$

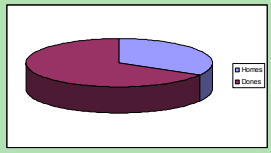
Freqüencies relatives

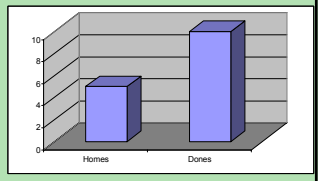
$$f(H) = \frac{F(H)}{F(\Omega)} = \frac{F\{11,12,13,14,15\}}{15} = \frac{5}{15} = 0.33$$

$$f(NF) = \frac{F(NF)}{F(\Omega)} = \frac{F\{2,3,6,7,14,15\}}{15} = \frac{6}{15} = 40\%$$

Representació gràfica

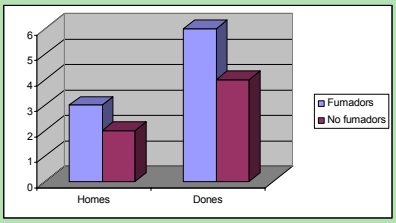
- Variables discretes (marginals).





Representació gràfica

- Dues variables discretes



Descripció de dades quantitatives

- Mesures de localització.
  - Mitjana i mediana.
  - Quantils.
- Mesures de dispersió.
  - Desviació estàndard. Variància.
  - Rang. Rang interquartil.
- Mesures de forma.
  - Biaix (sesgo).
  - Curtosi.

Descripció:Variables Quantitatives

- Mitjana $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variància $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Desviació estàndard

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

Mediana

Com es calcula: Mediana (m): $f\{x_i \le m\} = f\{x_i \ge m\}$

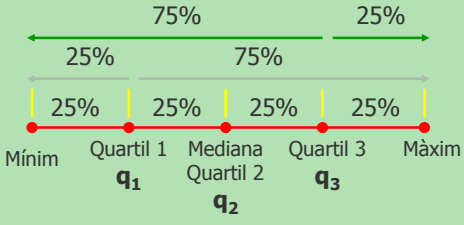
- Ordenar les dades de més petit a més gran.

  -Si el nombre de dades és senar, la mediana és la dada que està en el mig.

  - Si el nombre és parell, és la mitjana dels dos valors centrals.

La mediana és "robusta" davant la presència de outliers.

Quartils



Rang

És la diferència entre els valors observats més gran i més petit: $R = x_{máx} - x_{mín}$

- conserva les unitats de mesura de les dades originals.
- valors petits indiquen una menor dispersió.
- es veu extraordinàriament afectada pels valors extrems (outliers).
- és molt útil per detectar valors extrems.

Descriptiva variables contínues

- Posició
- Dispersió
- Forma
- Extrems

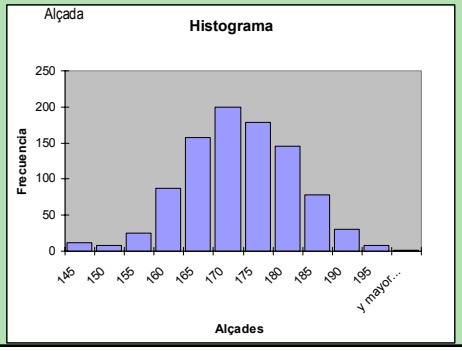
Alçada (cm)	
Media	170.35
Mediana	170
Moda	170
Desviación estándar	9.80
Varianza de la muestra	96.12
Curtosis	5.20
Coeficiente de asimetría	-0.96
Rango	102
Mínimo	97
Máximo	199
Suma	158594
Cuenta	931

Histograma

- Classifiquem les observacions

Classes	Frecuencia
145	12
155	33
165	244
175	379
185	223
195	39
y mayor...	1

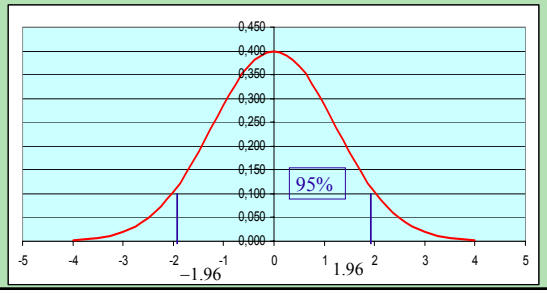
Representació de les freqüències



La llei dels errors

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

- Distribució normal de probabilitats



Distribucions Normals $N(\mu,\sigma^2)$

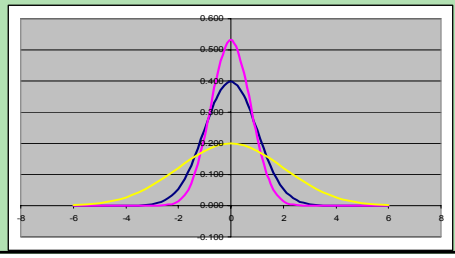
- Són les que tenen funció de densitat

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

- on $\mu=\overline{x}, \quad \sigma=S$
- Si $\mu=0, \quad \sigma=1$ aleshores $N(0,1)$

Tendència i errors de dispersió

- Tendència central $\mu=0.$
- Dispersió $\sigma=0.75,\ 1,\ 2.$



Classificació i histogrames

- Sota condicions "normals"
  - El 95% de les observacions d'una variable contínua es troben separades com a màxim dues desviacions de la mitjana.

$$|x - \bar{x}| \leq 1.96\, S$$

Que és un fet estrany ?

- Estrany, estadísticament:
  - **Allò que passa un 1 cop de cada 20**.
  - $p = 1/20 = 0.05$
- Molt estrany:
  - **Allò que passa un 1 cop de cada 100.**
  - $p = 1/100 = 0.01$
- Si no és estrany en direm "**normal**".

Estudi de dues variables

- Dues variables contínues:
  - Recta de regressió.
- Dues variables discretes:
  - Taules de contingència.
  - Test Xi-quadrat de Pearson.
- Una discreta i una contínua:
  - Proves t de comparació de grups.
  - Anàlisi de la variància.

Estadística bivariant

- Mitjanes $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Variàncies

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Covariància

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Covariància i correlació

- Covariància

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Coeficient de correlació i determinació

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

No tenen unitats

Recta de regressió

- Equació de la recta

$$y = a + b\,x$$

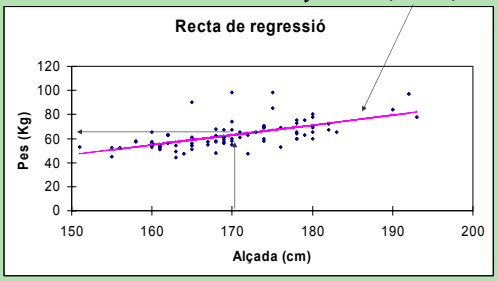
Xm	168,47
Ym	62,00
Sx	7,17
Sy	9,27
Sxy	36,43

- Coeficients: $b = \frac{S_{xy}}{S_x^2}, \quad a = \bar{y} - b\,\bar{x}$
- Coeficient de correlació y de determinació

$$r = \frac{S_{xy}}{S_x S_y} \rightarrow 0 \leq r^2 \leq 1$$

Relació pes i alçada

$y = -76{,}9 + 0{,}82\,x$



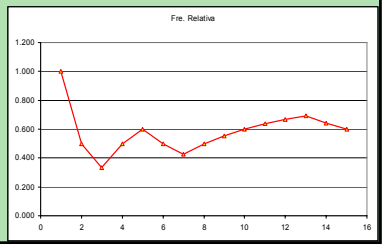
Probabilitat $f_n(A) \to P(A)$

- Lleis de la probabilitat:
  - 1 $P(A) \geq 0, \quad P(\Omega) = 1, \quad P(\varnothing) = 0.$
  - 2 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - 3 $P(\overline{A}) = 1 - P(A)$
  - 4 $P(A \cap C) = P(A|C)P(C)$
  - 5 $P(A \cap B) = P(A)P(B)$ si A i B son independents

Freqüències relatives

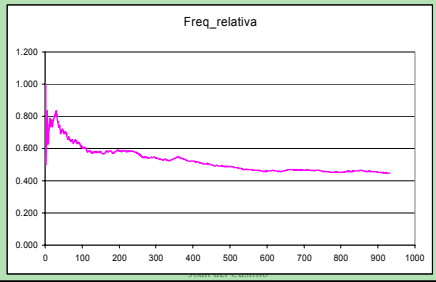
Codi	Fumad	Acum	F_Relat
1	1	1	1.000
2	0	1	0.500
3	0	1	0.333
4	1	2	0.500
5	1	3	0.600
6	0	3	0.500
7	0	3	0.429
8	1	4	0.500
9	1	5	0.556
10	1	6	0.600
11	1	7	0.636
12	1	8	0.667
13	1	9	0.692
14	0	9	0.643
15	0	9	0.600

Evolució de la freqüencia relativa del fumadors (Tabac)



Límit de les freqüències

- Freqüència de fumadors en una mostra de 931 persones



Hi ha relació gènere-Tabac ?

	Homes	Dones	Total
Fumador	217	200	417
No_Fuma	244	270	514
Total	461	470	931

$$f(F|H) = \frac{f(H \cap F)}{f(H)} = \frac{217}{461} = 0.471$$

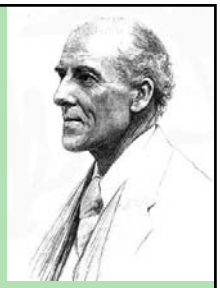
$$f(F|D) = \frac{f(F \cap D)}{f(D)} = \frac{200}{470} = 0.426$$

Comparem les dues taules

Observats	Homes	Dones	Total
Fumador	217	200	417
No_Fuma	244	270	514
Total	461	470	931

Esperats	Homes	Dones	Total
Fumador	206.48	210.52	417
No_Fuma	254.52	259.48	514
Total	461	470	931

Test X^2_n de Pearson



Un dels 20 descobriments més importants del Segle XX

$$\sum_{i,j=1}^{2} \frac{\left(E_{ij}-O_{ij}\right)^2}{E_{ij}} \le 3.84 = 1.96^2$$

En el 95% dels casos, si hi ha independència

No hi ha evidència de diferencies

Observat	Esperat	X^2
217	206.48	0.54
244	254.52	0.43
200	210.52	0.53
270	259.48	0.43
p-valor =	0.1657	1.92

$$\sum_{i,j=1}^{2} \frac{\left(E_{ij}-O_{ij}\right)^2}{E_{ij}} = 1.92^{ns} \le 3.84$$

Podem controlar l'Atzar

- Sabem mesurar la variabilitat produïda per atzar.
- Podem fixar un nivell de confiança per a les nostres afirmacions.
  - En biologia el nivell habitual és: 95% - 99%.
- És possible saber la veritat i és fàcil detectar errors i falsedats.

Proves diagnòstiques

	Malaltia	Salud
Diag. Positiu	Ver. Positiu	Fals Positiu
Diag. Negatiu	Fals Negatiu	Ver. Negatiu

Sensibilitat $P\left(DP \middle| M\right)$ per descartar (no detectar greu)

Especificitat $P\left(DN \middle| S\right)$ confirmar (diagnostic fals greu)

Estadística

Variables aleatòries discretes

Joan del Castillo

2008

Variables aleatòries

- Una variable aleatòria, X, representa el resultat numèric d'un experiment aleatori.
  - El resultat d'un dau.
  - L'alçada d'una persona, escollida al atzar.
  - Els valors 1 o 0 segons si la persona fuma.
- Analitzarem models simples, per tal d'introduir els conceptes d'esperança i variància d'una població.

Esperança d'un model arbitrari

- Valors i probabilitats: $X = \begin{cases} v_1 & p_1 \\ v_2 & p_2 \\ \dots & \dots \\ v_n & p_n \end{cases}$
- Valor esperat:

$$E[X] = v_1 \cdot p_1 + v_2 \cdot p_2 + \dots + v_n \cdot p_n$$

- Interpretació: $\overline{x}_n \to E[x]$

Propietats de les mitjanes

- La mitjana de la suma és suma de mitjanes.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i + y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Si multipliquem les dades per una constant, la mitjana també es multiplica.

$$\frac{1}{n}\sum_{i=1}^{n} a\, x_i = a\,\frac{1}{n}\sum_{i=1}^{n} x_i = a\,\overline{x}$$

Propietats de l'Esperança $\overline{x}_n \to E[x]$

- L'esperança de la suma és la suma d'esperances.

1. $E[X+Y] = E[X] + E[Y]$

- Si multipliquem per una constant, la esperança també es multiplica.

2. $E[a\,X] = a\,E[X]$

Distribució de Bernoulli

- X pot prendre els valors 1, 0 amb probabilitats p i (1-p).

$$X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$$

- Quan p = ½, és el model del llançament d'una moneda.

Valor esperat d'una Bernoulli

- Suposem X amb distribució:

$$X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$$

- El valor esperat és:

$$E[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

Distribució Binomial

- Anomenem B el nombre d'èxits, en n proves. La probabilitat d'obtenir k èxits és:

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

- on

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

Valor esperat d'una Binomial

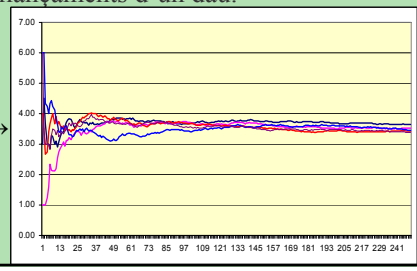
- Suposem X amb distribució: $X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$
- Sigui $B = \sum_{i=1}^{n} X_i$
- El valor esperat de B és:

$$E[B] = n\,E[X_i] = n\,p$$

Evolució de la mitjana $\bar{x}_n \to E[x] = 3.5$

- Calculem la mitjana en 5 sèries de 250 llançaments d'un dau.

3.5 →



↕ $V[X]$

Variància d'un model arbitrari

- Valors i probabilitats: $X = \begin{cases} v_1 & p_1 \\ v_2 & p_2 \\ \dots & \dots \\ v_n & p_n \end{cases}$

Interpretació de la variància:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \to V[X]$$

Definició:

$$V[X] = E[X^2] - E[X]^2$$

Càlcul de la variància $V[X] = E[X^2] - E[X]^2$

$$X = \begin{cases} v_1 & p_1 \\ v_2 & p_2 \\ \dots & \dots \\ v_n & p_n \end{cases} \qquad X^2 = \begin{cases} v_1^2 & p_1 \\ v_2^2 & p_2 \\ \dots & \dots \\ v_n^2 & p_n \end{cases}$$

$$E[X] = v_1 \cdot p_1 + v_2 \cdot p_2 + \dots + v_n \cdot p_n$$

$$E[X^2] = v_1^2 \cdot p_1 + v_2^2 \cdot p_2 + \dots + v_n^2 \cdot p_n$$

Variància d'una Bernoulli

$$X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases} \qquad X^2 = \begin{cases} 1^2 & p \\ 0^2 & 1-p \end{cases}$$

- Valor esperat:

$$E[X] = 1 \cdot p + 0 \cdot (1-p) = E[X^2] = p$$

- Variància:

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

Propietats de les variàncies

La variancia es pot calcular com la mitjana dels quadrats menys el quadrat de la mitjana

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

Si multipliquem les dades per una constant, la variancia es multiplica pel quadrat.

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

Propietats de les variàncies

- La variància de la suma de variables és la suma de variàncies més dues vegades la covariància.

$$\frac{1}{n}\sum_{i=1}^{n}\left((x_i+y_i)-(\bar{x}+\bar{y})\right)^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2+$$
$$+\frac{2}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

- Per a variables independents:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\to 0$$

Propietats de la Variància

- La variància de la suma o diferència de variables independents és suma de variàncies.

1. $V[X\pm Y]=V[X]+V[Y]$

- Si multipliquem per una constant, la variància es multiplica pel quadrat.

2. $V[a\,X]=a^2V[X]$

Moments d'una Bernoulli

$$X=\begin{cases}1 & p\\ 0 & 1-p\end{cases}\qquad X^2=\begin{cases}1^2 & p\\ 0^2 & 1-p\end{cases}$$

- Valor esperat:

$$E[X]=1\cdot p+0\cdot(1-p)=E\left[X^2\right]=p$$

- Variància:

$$V[X]=E\left[X^2\right]-E[X]^2=p-p^2=p(1-p)$$

Moments d'una Binomial

- Suposem X amb distribució: $X=\begin{cases}1 & p\\ 0 & 1-p\end{cases}$
- Sigui $\boxed{B=\sum_{i=1}^{n}X_i}\qquad B\approx B(n,p)$
- El valor esperat i la variància de B són:

$$\begin{cases}E[B]=n\,E[X_i]=n\,p\\ V[B]=n\,V[X_i]=n\,p\,(1-p)\end{cases}$$

Mostra i Població

$\boxed{f_n(A)\to P(A)}$

- Mitjana

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

- Esperança

$$E[x]=\sum v_k p_k$$

- Variància mostral

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

- Variància teòrica

$$V[X]=E\left[X^2\right]-E[X]^2$$

• Llei dels grans nombres: $\begin{cases}\bar{x}_n\to E[x]\\ S_n^2\to V[x]\end{cases}$

Funcions de distribució

- Sigui X una variable qualsevol. Anomenem funció de distribució (funció de probabilitat acumulada) a

$$F(x)=\mathrm{P}\{X\le x\}$$

- Aleshores calculem probabilitats

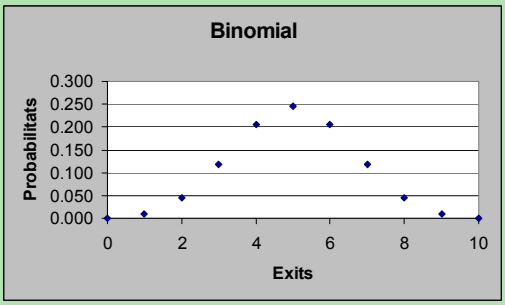
$$\mathrm{P}\{a<X\le b\}=F(b)-F(a)$$

Distribució Binomial

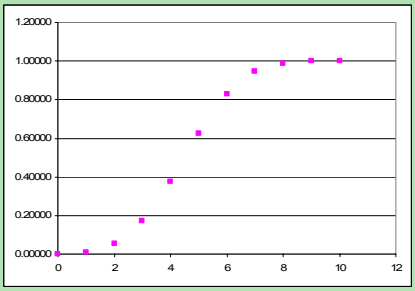
- Nombre d'èxits en n = 10 proves (p = 0.5)

Valors	P(k)	F(k)
0	0.001	0.001
1	0.010	0.011
2	0.044	0.055
3	0.117	0.172
4	0.205	0.377
5	0.246	0.623
6	0.205	0.828
7	0.117	0.945
8	0.044	0.989
9	0.010	0.999
10	0.001	1.000

Probabilitats Binomial (n = 10)



Probabilitats acumulades



$$F(x) = \mathrm{P}\{X \le x\}$$

Propietats funció de distribució

1. $F(x) = \mathrm{P}\{X \le x\}$, es creixent.
2. $\lim_{x\uparrow} F(x) = 1$
3. $\lim_{x\downarrow} F(x) = 0$
4. $\mathrm{P}\{a < X \le b\} = F(b) - F(a)$

Variables aleatòries contínues

- Si X és una variable contínua, la funció de distribució és derivable

$$F(x) = \mathrm{P}\{X \le x\}$$

- Anomenen funció de densitat a la derivada:

$$f(x) = F'(x)$$

- La densitat representa el límit dels histogrames.

Estadística

Distribució Normal

Joan del Castillo

2008

Variables aleatòries contínues

- Si X és una variable contínua, la funció de distribució és derivable

$$F(x) = \mathrm{P}\{X \le x\}$$

- Anomenen funció de densitat a la derivada:

$$f(x) = F'(x)$$

- La densitat representa el límit dels histogrames.

Variables continues

- Són equivalents $F'(x) = f(x)$
- i $P\{X \le x\} = F(x) = \int_{-\infty}^{x} f(t)\,dt$
- Les dues ens diuen que

$$\mathrm{P}\{a < X \le b\} = \int_a^b f(x)\,dx = F(b) - F(a)$$

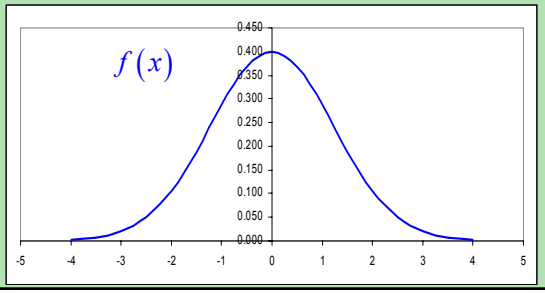
Propietats funció de densitat

1. $f(x) \ge 0.$
2. $\lim_{x\uparrow} f(x) = \lim_{x\downarrow} f(x) = 0.$
3. $\mathrm{P}\{a < X \le b\} = \int_a^b f(x)\,dx$
4. $\int_{-\infty}^{\infty} f(x)\,dx = 1.$

Normal estàndard

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

- Distribució normal de probabilitats



Estandardització

- Proposició: Utilitzant variables normals, sempre ens podem reduir a la normal N(0,1)

$$X \approx N(\mu, \sigma^2) \Leftrightarrow \boxed{Z = \frac{X-\mu}{\sigma}} \approx N(0,1)$$

- Anomenarem $\phi(x)$ a la funció de distribució de la N(0,1).

Exemple

$$Z \approx N(0,1)$$

- Calculem: $\mathrm{P}\{160 < X \le 170\}$
  - on $\mu = 170.35, \quad \sigma = 9.80$

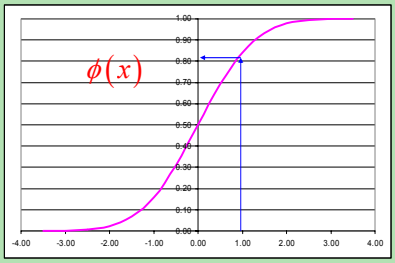
$$= P\left\{\frac{160-\mu}{\sigma} < \frac{X-\mu}{\sigma} \le \frac{170-\mu}{\sigma}\right\} = P\{-1.06 < Z \le -0.04\}$$

$$= \int_{-1.06}^{-0.04} f_z(x)\,dx \quad = \phi(-0.04) - \phi(-1.06)$$

$$= 0.484 - 0.146 \quad = 0.338$$

Funció de distribució de N(0,1)

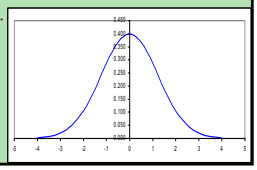
x	f(x)	F(x)
-3.50	0.001	0.0002
-3.00	0.004	0.0013
-2.50	0.018	0.0062
-2.00	0.054	0.0228
-1.50	0.130	0.0668
-1.00	0.242	0.1587
-0.50	0.352	0.3085
0.00	0.399	0.5000
0.50	0.352	0.6915
1.00	0.242	0.8413
1.50	0.130	0.9332
2.00	0.054	0.9772
2.50	0.018	0.9938
3.00	0.004	0.9987
3.50	0.001	0.9998



La distribució normal

- Apareix en molts fenòmens reals. Histogrames.
- Encara que alguna variable no segueixi la distribució normal, les sumes es comporten segons la distribució normal.
- Moltes característiques són el resultat de la suma de petits factors independents.
- Valors esperats d'una normal

$$X \approx N(\mu,\sigma^2) \Rightarrow \begin{cases} E[X]=\mu \\ V[X]=\sigma^2 \end{cases}$$



Teorema central del límit

- 1. Combinacions lineals de normals són normals.

$$a\ X + b$$

- 2. La suma de normals independents és normal.

$$X_1 + X_2 + \ldots + X_n$$

- 3. TCL: La suma de variables independents de qualsevol tipus és aproximadament normal.

Les rèpliques disminueixen l'error

- X és una variable on
  - μ és el valor mitjà $E[x]=\mu$
  - σ desviació de la mostra $Sd[x]=\sigma$
- Fem la mitjana de n observacions $\overline{x}=\frac{1}{n}\sum x_i$
  - El valor esperat és $E[\overline{x}]=\mu$
  - Desviació de la mitjana $V[\overline{x}]=\frac{\sigma^2}{n}$ $\quad Sd[\overline{x}]=\frac{\sigma}{\sqrt{n}}$

Teorema central del límit

- Sigui X una variable qualsevol amb
  - μ el valor esperat $E[x]=\mu$
  - σ^2 la variància $V[x]=\sigma^2$
  - Fem la mitjana de n observacions $\overline{x}=\frac{1}{n}\sum x_i$
- Teorema
  - La distribució de $\overline{x}$ és $\overline{x} \approx N(\mu,\frac{\sigma^2}{n})$

Apliquem el TCL a la Binomial

- Sigui B una variable Binomial.
  - Sabem que és suma de Bernoullis $B=\sum_{i=1}^{n} X_i$
  - L'esperança és $E[B]=n\,p=\mu$
  - La variància és $V[B]=n\,p(1-p)=\sigma^2$
- Teorema. $B(n,p) \approx N(\mu,\sigma^2)$
  - on $\mu=n\,p,\quad \sigma^2=n\,p(1-p)$
  - sempre que $n\,p(1-p)\geq 10$

Si $n\,p(1-p)\geq 5$, amb correccions.

Exemple

- Hem fet una enquesta a 80 persones. Si la mostra ha estat escollida a l'atzar, quina és la probabilitat d'entrevistar més de 25 dones ?
- El model correspon a una Binomial $B \approx B(n, p)$ amb $n = 80, \quad p = 0.5$

$$n\,p\,(1-p) = 80 \cdot 0.5 \cdot 0.5 = 20 \geq 10$$

Podem aproximar per una $X \approx N(\mu, \sigma^2)$

$\mu = n\,p = 40, \quad \sigma^2 = n\,p\,(1-p) = 20$

Exemple Podem aproximar per una $X \approx N(\mu, \sigma^2)$

$$P\{B > 25\} \approx P\{X > 25\}, \qquad \mu = 40, \quad \sigma^2 = 20$$

$$P\{X > 25\} = 1 - P\{X \leq 25\} = 1 - (1 - 0.999) = 0.999$$

$$P\{X \leq 25\} = P\left\{\frac{X-\mu}{\sigma} \leq \frac{25-\mu}{\sigma}\right\} = P\{Z \leq -3.35\}$$

$$= \phi(-3.35) = 1 - \phi(3.35) = 1 - 0.999$$

Valors crítics amb error 0.05

- Calculeu z_1, z_2 per a $Z \approx N(0,1)$ de manera que:

$$\phi(z_1) = P\{Z \leq z_1\} = 0.025$$

$$\phi(z_2) = P\{Z \leq z_2\} = 0.975$$

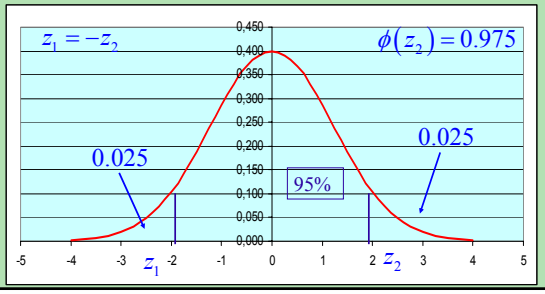
- Aleshores

$$P\{z_1 < Z \leq z_2\} = \phi(z_2) - \phi(z_1) = 0.95$$

La llei dels errors

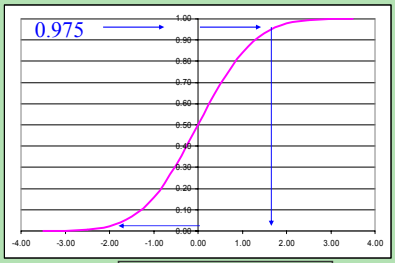
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

- Distribució normal de probabilitats



Valors crítics de la N(0,1) $\phi^{-1}(\alpha) = z_\alpha$

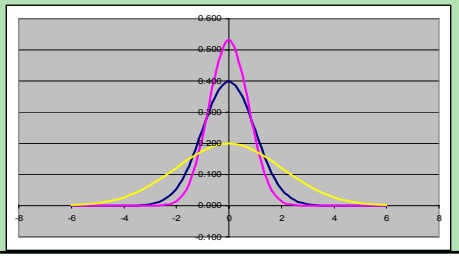
x	f(x)	F(x)
-3.50	0.001	0.0002
-3.00	0.004	0.0013
-2.50	0.018	0.0062
-2.00	0.054	0.0228
-1.50	0.130	0.0668
-1.00	0.242	0.1587
-0.50	0.352	0.3085
0.00	0.399	0.5000
0.50	0.352	0.6915
1.00	0.242	0.8413
1.50	0.130	0.9332
2.00	0.054	0.9772
2.50	0.018	0.9938
3.00	0.004	0.9987
3.50	0.001	0.9998



$$\phi^{-1}(0.025) = -1.96 \qquad \phi^{-1}(0.975) = 1.96$$

Normal segons la variància

- Tendència central $\mu = 0.$
- Dispersió $\sigma = 0.75,\ 1,\ 2.$



Valors crítics

- Suposem una $B \approx B(300, 0.5)$
  - És prou gran per aproximar per la Normal.
  - No cal fer correcció de Yates.
- Volem determinar b_1, b_2 tals que

$$P\{b_1 \le B \le b_2\} = 0.95$$

Aproximem per la Normal

- La Binomial $B \approx B(300, 0.5)$ es pot aproximar per la normal amb

$$\mu = n\,p = 150, \quad \sigma^2 = n\,p(1-p) = 75$$

$$B \approx B(300, 0.5) \quad \Rightarrow \quad X \approx N(150, 75)$$

Estandarditzem

$$Z \approx N(0,1)$$

- Determinar: $\mathrm{P}\{b_1 \le X \le b_2\} = 0.95$
  - on $\mu = 150, \quad \sigma^2 = 75$

$$= P\left\{\frac{b_1-\mu}{\sigma} < \frac{X-\mu}{\sigma} \le \frac{b_2-\mu}{\sigma}\right\} = P\{z_1 < Z \le z_2\} = 0.95$$

$$z_1 = \frac{b_1-150}{\sqrt{75}} = -1.96 \qquad z_2 = \frac{b_2-150}{\sqrt{75}} = 1.96$$

$$b_1 = 133, \quad b_2 = 167$$

Conseqüències de la simetria

- La simetria de la N(0,1) ens dona:

$$\phi(-x) = 1 - \phi(x)$$

- Pels valors crítics, la simetria ens dona:

$$\phi^{-1}(\alpha) = -\phi^{-1}(1-\alpha) \quad \Rightarrow \quad z_\alpha = -z_{1-\alpha}$$

Valors crítics N(0,1)

$$z_p = -z_{1-p}$$

- Els valors crítics més usuals són:

p-valor	Zp
0.025	-1.960
0.975	1.960
0.05	-1.645
0.95	1.645
0.005	-2.576
0.995	2.576
0.01	-2.326
0.99	2.326
0.10	-1.282
0.90	1.282

$$\phi^{-1}(p) = z_p$$

Aprox. Normal de les freqüències

- Per a la freqüència relativa $f = B/n$ tenim l'aproximació:

$$f \approx N\left(p, p(1-p)/n\right)$$

Sempre que: $n\,p(1-p) \ge 10$

L'Estandarització ens dona:

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

Precisió en les enquestes

- Suposem que fem 50 enquestes amb probabilitat 0.4 d'observar un esdeveniment ($\alpha = 0.05$).

$$\begin{cases} E[f] = 0.4 \\ V[f] = (0.4 \times 0.6)/50 = 0.0048 \\ error = 1.96 \times \sqrt{0.048} = 0.1358 \end{cases}$$

$$\text{inter.freq.} = (0.4 - 0.136, 0.4 + 0.136) = (0.26, 0.54)$$

Precisió en les enquestes

- Suposem que fem 2000 enquestes amb probabilitat 0.4 d'observar un esdeveniment ($\alpha = 0.05$).

$$\begin{cases} E[f] = 0.4 \\ V[B] = (0.4 \times 0.6)/2000 = 0.00012 \\ error = 1.96 \times \sqrt{0.00012} = 0.0215 \end{cases}$$

$$\text{inter.freq.} = (0.4 - 0.022, 0.4 + 0.022) = (0.38, 0.42)$$

Intervals per a freqüències

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

Amb confiança $(1\text{-}\alpha)$, o error α

$$z_{\alpha/2} \le \sqrt{n}\frac{f-p}{\sqrt{p(1-p)}} \le z_{1-\alpha/2}$$

$$\sqrt{n}\frac{f-p}{\sqrt{p(1-p)}} \le z_{1-\alpha} \qquad z_{\alpha} \le \sqrt{n}\frac{f-p}{\sqrt{p(1-p)}}$$

Intervals per a proporcions

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

Amb confiança $(1\text{-}\alpha)$, o error α

$$-z_{1-\alpha/2} \le \frac{p-f}{\sqrt{f(1-f)}/\sqrt{n}} \le z_{1-\alpha/2}$$

$$\frac{p-f}{\sqrt{f(1-f)}/\sqrt{n}} \le z_{1-\alpha} \qquad -z_{1-\alpha} \le \frac{p-f}{\sqrt{f(1-f)}/\sqrt{n}}$$

Distribucions derivades de N(0,1)

- Distribució Xi-quadrat. χ^2_n
- Distribució d'Student. t_n
- Distribució de Fisher. $F(m,n)$

Distribució Xi-quadrat

- Suposem que tenim n normals N(0,1) independents:

$$Z_1, Z_2, ..., Z_n \approx N(0,1)$$

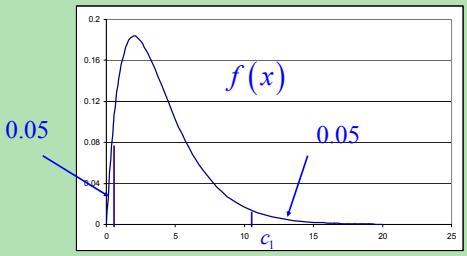
- Considerem:

$$Ch = Z_1^2 + Z_2^2 + ... + Z_n^2$$

- Direm que Ch té distribució: χ^2_n

$$Ch \approx \chi^2_n$$

Densitat d'una distribució χ^2_n



Valors crítics

Distribució d'Student

- Suposem que tenim n+1 normals N(0,1) independents:

$$Z_0, Z_1, Z_2, ..., Z_n \approx N(0,1)$$

- Considerem:

$$\begin{cases} Z_0, \\ Ch = Z_1^2 + Z_2^2 + ... + Z_n^2 \end{cases}$$

Distribució d'Student t_n

- Direm que

$$T = \frac{Z_0}{\sqrt{Ch/n}}$$

té distribució t_n d'Student amb *n graus de llibertat:*

$$T \approx t_n$$

Distribució F de Fisher

- Suposem que tenim:

$$\begin{cases} Ch(m) = Z_1^2 + Z_2^2 + ... + Z_m^2 \approx \chi^2_m \\ Ch(n) = Z_1^2 + Z_2^2 + ... + Z_n^2 \approx \chi^2_n \end{cases}$$

- Aleshores:

$$F = \frac{Ch(m)/m}{Ch(n)/n}$$

té distribució

$$F \approx F(m,n)$$

Taules de la $F(m,n)$ de Fisher

m →

n ↓

$n_\downarrow$	$m_\rightarrow$	1	2	3	4	5	6	7	8	9	10
	$z = 0.95$										
1		161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2		18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3		10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4		7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5		6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6		5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7		5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8		5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9		5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10		4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11		4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12		4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13		4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14		4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60

Observacions

- 1. Una χ^2_n és sempre positiva.
- 2. Una t_n és simètrica. $\dfrac{Z_0}{\sqrt{Ch/n}}$
- 3. El quadrat d'una t_n és una $F(1,n)$

$$\left(\frac{Z_0}{\sqrt{Ch(n)/n}}\right)^2 = \frac{Z_0^2}{Ch(n)/n} = \frac{Ch(1)/1}{Ch(n)/n}$$

Estadística

Inferència

Joan del Castillo

2008

Intervals de confiança

$$x \approx N(\mu, \sigma^2)$$

- Intervals per als paràmetres (desconeguts).
- Interval per a la mitjana amb variància coneguda.
- Teorema de Fisher.
- Interval de confiança per a la variància.
- Interval de confiança per a la mitjana, sense conèixer la variància.

Teorema de Fisher

- 1. $\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} \approx N(0,1) \quad \Leftrightarrow \bar{x} \approx N(\mu, \dfrac{\sigma^2}{n})$
- 2. $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum (x_i-\bar{x})^2}{\sigma^2} \approx \chi^2_{n-1}$
- 3. $\dfrac{\bar{x}-\mu}{S/\sqrt{n}} \approx t_{n-1}$

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

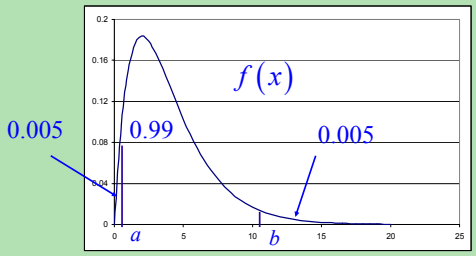
Exemple:

- Suposem que en una mostra de pesos de 25 persones la desviació resulta S = 12.5, entre quins valors està la veritable σ, amb una confiança del 99% ?

$$\alpha = 0.01, \quad \Rightarrow \quad \frac{\alpha}{2} = 0.005, \quad 1-\frac{\alpha}{2} = 0.995$$

$$\chi^2_{24}(0.005) = 9.89 \quad \rightarrow \quad a = 9.89$$
$$\chi^2_{24}(0.995) = 45.56 \quad \rightarrow \quad b = 45.56$$

Densitat d'una distribució χ^2_n



Valors crítics

Intervals de confiança per a σ^2

- Fisher (2) ens diu: $\dfrac{\sum (x_i-\bar{x})^2}{\sigma^2} \approx \chi^2_{n-1}$

Fixem un marge d'error: $\alpha \quad (0.05-0.01)$

Tindrem un nivell de confiança de $(1-\alpha)$

Busquem els valors crítics: $a \le \chi^2_{n-1} \le b$

$$a \le \frac{\sum (x_i-\bar{x})^2}{\sigma^2} \le b$$

Intervals de confiança per a σ^2

- Utilitzant: $\boxed{a \le \chi^2_{n-1} \le b}$
- Tenim: $\dfrac{\sum (x_i - \bar{x})^2}{b} \le \sigma^2 \le \dfrac{\sum (x_i - \bar{x})^2}{a}$
- Equivalentment:

$$\sqrt{\frac{\sum (x_i - \bar{x})^2}{b}} \le \sigma \le \sqrt{\frac{\sum (x_i - \bar{x})^2}{a}}$$

Exemple

$$a = \chi^2_{24}(0.005) = 9.89$$
$$b = \chi^2_{24}(0.995) = 45.56$$

- Tenim que: $P\{9.89 \le \chi^2_{24} \le 45.56\} = 0.99$
- Aleshores: $\dfrac{\sum (x_i - \bar{x})^2}{45.56} \le \sigma^2 \le \dfrac{\sum (x_i - \bar{x})^2}{9.89}$
- Finalment $9.0724 \le \sigma \le 19.472$

Interval per a μ amb σ desconeguda

- Fisher (3) ens diu: $\sqrt{n}\dfrac{\bar{x} - \mu}{S} \approx t_{n-1}$
- Fixem un marge d'*error*: $\alpha \quad (0.05 - 0.01)$
- Tindrem un *nivell de confiança* de

$$(1-\alpha) \quad (0.95 - 0.99)$$

- Busquem els valors crítics:

$$a \le \sqrt{n}\frac{\bar{x} - \mu}{S} \le b$$

Intervals de confiança per a μ

- Volem trobar (a, b) que compleixin:

$$P\{a \le t_{n-1} \le b\} = (1-\alpha)$$

- Per simetria: $a = -b$ on $\boxed{b = t_{1-\alpha/2} = t^{n-1}_{1-\alpha/2}}$

$$a = t_{\alpha/2} = -t_{1-\alpha/2}$$

- Finalment: $\boxed{\bar{x} - t_{1-\alpha/2}\dfrac{S}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2}\dfrac{S}{\sqrt{n}}}$

Contrastos d'hipòtesis

- Intervals de confiança o contrastos ?
- Contrastos d'hipòtesis.
  - Contrastos per a mitjanes.
  - Contrastos per a variàncies.
  - Regió d'acceptació i regió crítica.
  - Contrastos unilaterals i bilaterals..

Exemple

- Es tenen dades sobre 100.000 individus (servei militar) on es conclou que l'alçada mitjana era μ = 170 i la desviació σ = 10.
- Es vol decidir si ha augmentat l'alçada mitjana de la població, amb una confiança del 95%.
- Disposem d'una mostra de n = 25 individus, amb una mitjana de $\bar{x} = 172.3$

Contrastos d'hipòtesis $\bar{x} \approx N(\mu, \frac{\sigma^2}{n})$

- El plantejament del problema porta a considerar dues hipòtesis:

$$\begin{cases} H_0 : \mu = 170 \\ H_1 : \mu > 170 \end{cases}$$

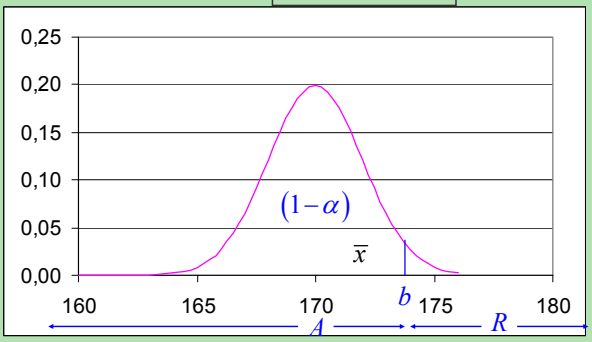
  - La primera o nul·la és la que pretenem rebutjar.
  - La segona, o alternativa, és la que creiem certa.
  - No contemplem la possibilitat que: $\mu < 170$
  - Ens basarem en la distribució de la mitjana:

Contrastos d'hipòtesis

- *Acceptarem* $H_0 : \mu = 170$ sempre que $\bar{x}$ sigui menor que un cert valor: $\bar{x} \le b$
- Acceptarem $H_1 : \mu > 170$ quan $\bar{x} > b$.
- En el segon cas direm que *rebutgem la hipòtesi nul·la*.
- La *regió d'acceptació* $A = (-\infty, b)$ serà:

$$P\{\bar{x} \in A\} = P\{\bar{x} \le b\} = (1-\alpha) = 0.95$$

Densitat de: $\bar{x} \approx N(\mu, \frac{\sigma^2}{n})$



Contrastos d'hipòtesis

- La *regió d'acceptació* (de H_0) serà de la forma: $A = (-\infty, b)$

$$P\{\bar{x} \in A\} = P\{\bar{x} \le b\} = (1-\alpha) = 0.95$$

- La regió de *rebuig* (de H_0) o *regió crítica* serà de la forma: $R = \bar{A} = (b, \infty)$

Determinació del punt crític: $b = ?$

- Volem trobar b de manera que:

$$P\{\bar{x} \in A\} = P\{\bar{x} \le b\} = (1-\alpha) = 0.95$$

- Estandarditzant: $Z = \sqrt{n}\,\frac{\bar{x}-\mu}{\sigma} \approx N(0,1)$

$$P\left\{\sqrt{n}\,\frac{\bar{x}-\mu}{\sigma} \le \sqrt{n}\,\frac{b-\mu}{\sigma}\right\} = (1-\alpha) = 0.95$$

Determinació del punt crític: b

- Cal resoldre: $\sqrt{n}\,\frac{b-\mu}{\sigma} = z_{1-\alpha} = 1.645$

$$b = \mu + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 170 + 1.645\frac{10}{\sqrt{25}} = 173.3$$

$A = (-\infty, b) = (-\infty, 173.3)$ $\quad \bar{x} = 172.3 \in A$

Acceptem $H_0 : \mu = 170$

No hi ha prou evidència per rebutjar Ho

Contrastos per a mitjanes $\bar{x} \approx \mu$

- Es poden presentar les situacions:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 \end{cases}$$

Regions d'acceptació

- Les regions d'acceptació poden ser:

$$\begin{cases} A = (-\infty, b) & \text{Una cua.} \\ A = (a, b) & \text{Dues cues (bilateral).} \\ A = (a, \infty) & \text{Una cua.} \\ A = (0, b) & \text{Una cua, distribució positiva.} \end{cases}$$

Contrast sobre variàncies

- En una mostra de 30 individus hem calculat una desviació estàndard de S = 12,5. Volem decidir si els individus pertanyen a una població amb $\sigma = 10$ o a un altra amb $\sigma = 15$.

Considerem: $\begin{cases} H_0 : \sigma = 10 \\ H_1 : \sigma = 15 \end{cases}$ $(\alpha = 0.05)$

Regió d'acceptació ? $A = (0, b) = \{S \leq b\}$

Contrast sobre variàncies

- Estadístic: $\dfrac{(n-1)S^2}{\sigma^2} \approx \chi^2_{n-1}$
- Regió d'acceptació: $A = (0, b) = \{S \leq b\}$
- Equació per determinar b: $P\{S^2 \leq b^2\} = 0.95$
- Estandarditzem... $P\left\{\dfrac{(n-1)S^2}{\sigma^2} \leq \dfrac{(n-1)b^2}{\sigma^2}\right\} =$

Contrast sobre variàncies $\dfrac{(n-1)S^2}{\sigma^2} \approx \chi^2_{n-1}$

- Estandardització (suposant $H_0 : \sigma = 10$):

$$P\left\{\frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b^2}{\sigma^2}\right\} = (1-\alpha) = 0.95$$

- Busquem a les taules: $ch_{1-\alpha} = ch^{n-1}_{1-\alpha} = \dfrac{(n-1)b^2}{\sigma^2}$

$$P\{\chi^2_{29} \leq ch_{1-\alpha}\} = 0.95,$$

Contrast sobre variàncies

- De les taules: $ch_{1-\alpha} = \dfrac{(n-1)b^2}{\sigma^2}$

$$ch_{1-\alpha} = 42.56, \quad b^2 = \frac{ch_{1-\alpha}\,\sigma^2}{n-1} = 146.8$$

- Regió d'acceptació: $A = (0, b) = \{S^2 \leq b^2\}$

$P\{S^2 \leq 146.8\} = 0.95, \qquad P\{S \leq 12.1\} = 0.95$

$S = 12.5 \notin A$ Rebutgem $H_0 : \sigma = 10$

Tenim un 95% de confiança que σ=15

Comparació de dos grups

- Contrastos població-mostra.
- Comparació de dues mitjanes:
  - Mostres independents, variàncies conegudes.
  - Mostres independents, variàncies iguals.
  - Dades aparellades.
- Comparació de dues variàncies.

Comparació mitjanes de pesos

- Observacions:

	1	2	3	4	5	6	7	8	9	10	11	12
Grup-f	65	67	67	52	62	50	70	57				
Grup-m	75	70	49	73	75	110	86	70	60	58	82	54

- Model:

$$\begin{cases} Gr.1: x_1, x_2, ..., x_{n_x} \approx N(\mu_x, \sigma_x^2) \\ Gr.2: y_1, y_2, ..., y_{n_y} \approx N(\mu_y, \sigma_y^2) \end{cases}$$

Contrastos d'hipòtesis

- El plantegem el problema com a:

$$\begin{cases} H_0: \mu_y = \mu_x \\ H_1: \mu_y > \mu_x \end{cases}$$

  - Primer suposarem conegudes σ_y i σ_x.
  - Segon, suposarem $\sigma_y = \sigma_x$ desconegudes.
  - Tercer, ens plantejarem contrastar: $H_0: \sigma_y = \sigma_x$

Distribució de l'estadístic $\overline{y} - \overline{x}$

- Calculem:

$$E[\overline{y} - \overline{x}] = E[\overline{y}] - E[\overline{x}] = \mu_y - \mu_x = \delta$$

$$V[\overline{y} - \overline{x}] = V[\overline{y}] + V[\overline{x}] = \frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}$$

- Estandarditzant:

$$Z = \frac{(\overline{y} - \overline{x}) - \delta}{\sqrt{\frac{\sigma_y^2}{n_y} + \frac{\sigma_x^2}{n_x}}} \approx N(0,1)$$

Exemple-2

- Observacions:

	1	2	3	4	5	6	7	8	9	10	11	12
Grup-f	65	67	67	52	62	50	70	57				
Grup-m	75	70	49	73	75	110	86	70	60	58	82	54

- Contrasteu: $\begin{cases} H_0: \mu_y = \mu_x \\ H_1: \mu_y \neq \mu_x \end{cases}$ suposant $\begin{cases} \sigma_x = 8 \\ \sigma_y = 10 \end{cases}$ $(\alpha = 0.05)$
  - Calculeu: $\begin{cases} \overline{y} = 71.8 \\ \overline{x} = 61.3 \end{cases}$ $\delta = 10$

Comparació mitjanes de pesos

Prueba z para medias de dos muestras

	Variable 1	*Variable 2*
Media	71.83	61.25
Varianza (conocida)	100	64
Observaciones	12	8
Diferencia hipotética de las medias	0	
z	2.619	
P(Z<=z) una cola	0.004	
Valor crítico de z (una cola)	1.645	
Valor crítico de z (dos colas)	0.009	
Valor crítico de z (dos colas)	1.960	

Valor esperat de S^2

- Del Teorema de Fisher:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum (x_i - \bar{x})^2}{\sigma^2} \approx \chi^2_{n-1}$$

- Aleshores:

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = E\left[\chi^2_{n-1}\right] = n-1$$

$$\boxed{E\left[S^2\right] = \sigma^2}$$ Estimador no-esbiaixat

Comparació de variàncies

- Sabem del Teorema de Fisher:

$$Ch_x = \frac{(n_x-1)S_x^2}{\sigma^2} \approx \chi^2_{n_x-1}, \quad Ch_y = \frac{(n_y-1)S_y^2}{\sigma^2} \approx \chi^2_{n_y-1}$$

- Aleshores:

$$\frac{S_x^2}{S_y^2} = \frac{Ch_x/(n_x-1)}{Ch_y/(n_y-1)} \approx F_{n_x-1,n_y-1}$$

Exemple: Comparació de Variàncies

- Observacions:

	1	2	3	4	5	6	7	8	9	10	11	12
Grup-f	65	67	67	52	62	50	70	57				
Grup-m	75	70	49	73	75	110	86	70	60	58	82	54

- Contrasteu: $(\alpha = 0.05)$

$$\begin{cases} H_0 : \sigma_y = \sigma_x \\ H_1 : \sigma_y > \sigma_x \end{cases}$$

  - Calculeu:

$$\begin{cases} \bar{y} = 71.8 \\ \bar{x} = 61.3 \end{cases} \qquad \begin{cases} S_x = 7.44 \\ S_y = 16.40 \end{cases}$$

Comparació de Variàncies

Prueba F para varianzas de dos muestras

	Variable 1	*Variable 2*
Media	71.83	61.25
Varianza	269.06	55.36
Observaciones	12	8
Grados de libertad	11	7
F	4.86	
P(F<=f) una cola	0.02	
Valor crítico para F (una cola)	3.60	

Dades aparellades

- Colesterol en sang abans i després d'un tractament basat en dieta i esport.

Paci.	Abans (Y)	Després (X)	Dif.
1	201	200	1
2	231	236	-5
3	221	216	5
4	260	233	27
5	228	224	4
6	237	216	21
7	326	296	30
8	235	195	40
9	240	207	33
10	267	247	20
11	284	210	74
12	201	209	-8

$$\begin{cases} \text{Augmenta}: & 2 \\ \text{Disminueix}: & 10 \end{cases}$$

$$d = Y - X$$

W.Daniel, 1987

Dades aparellades

Considerem la diferencia: $d_i = y_i - x_i$,

$$\sqrt{n}\,\frac{\bar{d} - \mu_d}{S_d} \approx t_{n-1}$$

$$\mu_d = \mu_y - \mu_x,$$

$$\bar{d} = \bar{y} - \bar{x}, \qquad S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2$$

Contrastem mitjana zero

- Volem contrastar en base a $d = Y - X$

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d > 0 \end{cases}$$

- Una població.
- Contrast sobre la mitjana.
- Variancia desconeguda.

$$\sqrt{n}\frac{\overline{x}-\mu}{S} \approx t_{n-1} \quad \Rightarrow \quad \boxed{\sqrt{n}\frac{\overline{d}-\mu_d}{S_d} \approx t_{n-1}}$$

Variàncies conegudes

- Suposant conegudes σ_x, σ_y .

$$\boxed{Z = \frac{(\overline{y}-\overline{x})-\delta}{\sqrt{\frac{\sigma_y^2}{n_y}+\frac{\sigma_x^2}{n_x}}} \approx N(0,1)}$$

Variàncies iguals, desconegudes

- Expressió final del Test

$$\boxed{T = \frac{(\overline{y}-\overline{x})-\delta}{S\sqrt{\frac{1}{n_y}+\frac{1}{n_x}}} \approx t_{n_x+n_y-2}}$$

on $\boxed{S^2 = \frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{(n_x+n_y-2)}}$

Comparació de variàncies

- Distribució del quocient:

$$\boxed{\frac{S_x^2}{S_y^2} \approx F_{n_x-1,n_y-1}}$$

$$S_x^2 = \frac{1}{n_x-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \qquad S_y^2 = \frac{1}{n_y-1}\sum_{i=1}^{n}(y_i-\overline{y})^2$$

Estadística

Anàlisi de la variància

Joan del Castillo

2008

Anàlisi de la variància

- Model: $\begin{cases} Gr.1 : y_{11}, y_{12}, ..., y_{1n} \approx N(\mu_1, \sigma^2) \\ Gr.2 : y_{21}, y_{22}, ..., y_{2n} \approx N(\mu_2, \sigma^2) \\ ... \\ Gr.I : y_{I1}, y_{I2}, ..., y_{In} \approx N(\mu_I, \sigma^2) \end{cases}$
- Un procediment per comparar mitjanes.
- Hipòtesis del model: $\boxed{y_{ij} = \mu_i + e_{ij}}$ $\quad e_{ij} = N\left(0, \sigma^2\right)$
  - Variables normals.
  - Variàncies iguals.
  - Independència.

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_I = \mu \\ H_1 : \text{Alguna } \mu_i \neq \mu \text{ diferent.} \end{cases}$$

Concentracions arterials d'epinefrina

- Es va mesurar en 15 animals de laboratori, dividits en tres grups sotmesos a tres tipus diferents d'anestèsia, la concentració arterial d'epinefrina en plasma sanguini (unitats en 10 nanograms).

Anestèsia-1	9	12	10	8	15
Anestèsia-2	20	21	23	17	30
Anestèsia-3	6	5	8	16	7

  - Es pregunta si hi ha diferències entre els tres grups al 99%.

Calculeu: $SST = \sum_{ij}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2 = (N-1)S_t^2$

$N = I \cdot n = 3 \times 5 = 15$

						Suma	Promedio	Varianza
Anestèsia-1	9	12	10	8	15	54	10.8	7.7
Anestèsia-2	20	21	23	17	30	111	22.2	23.7
Anestèsia-3	6	5	8	16	7	42	8.4	19.3
						Promedio	13.8	16.9

- Volem contrastar:

$\bar{y}_{\cdot\cdot} \approx \mu \quad S^2 \approx \sigma^2$

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_I = \mu \\ H_1 : \text{Alguna } \mu_i \neq \mu \text{ diferent.} \end{cases}$$

Càlculs Anova:

$$S_e^2 = \frac{1}{I-1}\sum_i (y_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$S_t^2 = \frac{1}{N-1}\sum_{ij}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$$

$$\text{Gr.1} : y_{11}, y_{12}, ..., y_{1n} \Rightarrow \bar{y}_{1\cdot}, S_1^2$$
$$\text{Gr.2} : y_{21}, y_{22}, ..., y_{2n} \Rightarrow \bar{y}_{2\cdot}, S_2^2$$
$$...$$
$$\text{Gr.I} : y_{I1}, y_{I2}, ..., y_{In} \Rightarrow \bar{y}_{I\cdot}, S_I^2$$

$$\begin{cases} \bar{y}_{i\cdot} = \frac{1}{n}\sum_j y_{ij} \\ S_i^2 = \frac{1}{n-1}\sum_j \left(y_{ij} - \bar{y}_{i\cdot}\right)^2 \end{cases}$$

$$\begin{cases} \bar{y}_{\cdot\cdot} = \frac{1}{I}\sum_i y_{i\cdot} = 13.8 \\ S^2 = \frac{1}{I}\sum_i S_i^2 = 16.9 = \frac{1}{I(n-1)}\sum_{i,j}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2 = \frac{SSE}{I(n-1)} \end{cases}$$

Descomposició de la variància

$$\sum_{ij}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2 = \sum_{ij}\left(y_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2 + \sum_{ij}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

$$SST = SSA + SSE$$

$$\begin{cases} SSA = \text{Tractament (Entre)} = n\sum_i \left(y_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2 = n(I-1)S_e^2 \\ SSE = \text{Error (Dintre)} = \sum_{ij}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2 = I(n-1)S^2 \\ SST = \text{Total} = \sum_{ij}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2 = (N-1)S_t^2 \end{cases}$$

Construïu la taula de anova

$$MS = SS/gl, \quad F = MSA/MSE$$

ANÀLISIS DE VARIANZA

variaciones	*SS*	*g.l.*	*MS*	*F*	*Fc*
Tractament (A)	SSA	I - 1	MSA	F	F(I-1,I(n-1))
Error	SSE	I (n-1)	MSE		
Total	SST	n.I - 1			

$$SSA = n\sum_i \left(y_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2 = n(I-1)S_e^2 \qquad S_e^2 = \frac{1}{I-1}\sum_i \left(y_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$$

$$SSE = \sum_{ij}\left(y_{ij} - y_{i\cdot}\right)^2 = I(n-1)S^2 \qquad S^2 = \frac{1}{I}\sum_i S_i^2$$

$$SST = \sum_{ij}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2 = (N-1)S_t^2 \qquad S_t^2 = \frac{1}{N-1}\sum_{ij}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$$

Calculeu directament

$$S_e^2 = \frac{1}{I-1}\sum_i (y_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \quad 54.36$$

$$SSA = \text{Entre} = n\sum_i \left(y_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2 = n(I-1)S_e^2 \quad = 5 \cdot 2 \cdot 54.36 = 543.6$$

$$SSE = \text{Dintre} = \sum_{ij}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2 = I(n-1)S^2 \quad = 3 \cdot 4 \cdot 16.9 = 202.8$$

$$SST = \text{Total} = \sum_{ij}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2 = (N-1)S_t^2 \quad = 14 \cdot 53.3 = 746.4$$

Taules de la $F(m,n)$ de Fisher

m

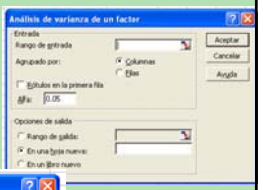
n

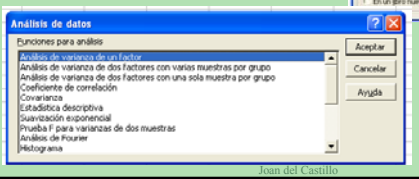
$\alpha = 0.95$

$n \downarrow$ \ $m \rightarrow$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	250.1	251.1
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.47
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.72
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.77
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.34
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.04
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.83
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.66
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.57	2.53
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.43
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.38	2.34
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.31	2.27

Anova en Excel

Anestèsia-1	Anestèsia-2	Anestèsia-3
9	20	6
12	21	5
10	23	8
8	17	16
15	30	7





Anàlisi de la variància un factor

$MS = SS / gl \qquad F = MSA / MSE$

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
Anestèssia-1	5	54	10.8	7.7
Anestèssia-2	5	111	22.2	23.7
Anestèssia-3	5	42	8.4	19.3

SS MS

ANÁLISIS DE VARIANZA

variaciones	*S.cuadrados*	*gl*	*Promedio*	*F*	*p-valor*	*V. crítico*
Entre grupos	543.6	2	271.8	16.08	0.0004	3.89
Dentro de los	202.8	12	16.9			
Total	746.4	14				

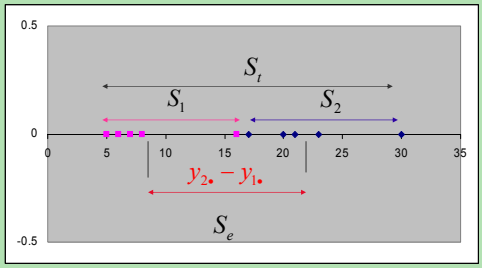
$SST = SSA + SSE$

$16.08 > 3.89 \quad \Leftrightarrow \quad F > V.C$

Rebutjem la igualtat de mitjanes

Interpretació anova

Anestèsia-1	Anestèsia-2	Anestèsia-3
	20	6
	21	5
	23	8
	17	16
	30	7



$$S^2 = \frac{1}{I}\sum_i S_i^2, \qquad S_i^2 = \frac{1}{N-1}\sum_{ij}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2, \qquad S_e^2 = \frac{1}{I-1}\sum_i \left(y_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2$$