MANUAL D'EINES ESTADÍSTIQUES BÀSIQUES PER A USUARIS

Concepción Arenas

"*Manual d'eines estadístiques bàsiques per a usuaris*", pretén com el seu nom indica ser una guia d'estadística bàsica per als usuaris. En aquest sentit és una publicació activa on es van incloent i/o ampliant tots aquells temes d'interès. De cada eina estadística es dona una idea clara i fonamental de perquè serveix i de sota quines condicions és correcte aplicar-la. Alguns capítols inclouen un petit exemple amb les corresponents instruccions en llenguatge R per tal de mostrar com realitzar els càlculs mitjançant aquesta eina informàtica.

ÍNDEX

Anosim

Cluster: quin mètode de cluster cal fer servir? Quants grups hi ha?

Com escriure fórmules matemàtiques amb word?

Mitjana ± desviació estàndard? o Mitjana± error estàndard de la mitjana?

Mds and stress

Outliers/valors extrems

Alguns test basics, paramètrics i no paramètrics

Potència d'un test

Taules de contingència: test de khi-quadrat, test exacte de Fisher i test de Mcnemar

Puntuacions Z-score

ANOSIM

Test multivariant no paramètric per a detectar diferències en la composició de les comunitats entre grups.

El p-valor del test s'obté per permutacions.

L'estadístic és: $R = \frac{r_{b-r_w}}{\frac{1}{4}n(n-1)}$ on
r_b és la mitjana dels ranks de les distàncies entre grups
r_w és la mitjana dels ranks de les distàncies dintre de grups

Els valors de *R* varien de -1 a 1.
H_0: R≈0 (no hi ha diferència significativa entre grups).
H_1: R>0 (si hi ha diferències significatives entre grups, no acostuma a tenir valors negatius)

Càlcul del p-valor: es fan permutacions aleatòries DINTRE dels grups. Es calcula R^* i es mira si $R^* \geq R$. Es repeteix tantes vegades com sigui possible (mínim 1000).
p-valor = proporció de R^* que son $R^* \geq R$

Si hi ha més de dos grups cal fer comparacions múltiples amb la corresponent correcció: α=0,05/nombre de tests.

Cal com a mínim 10 rèpliques per grup (per tenir una potència adequada)
Per tal de detectar petits canvis cal un gran nombre de mostres
ANOSIM és sensible a la heterogeneïtat de variàncies entre grups

Clarke, K. R. (1993). Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* **18**: 117-143.

Warwick, R.M., Clarke, K.R. and Suharsono (1990). A statistical analysis of coral community responses to the 1982-83 El Niño in the Thousand Islands, Indonesia. *Coral Reefs* **8**: 171-179.

CLUSTER: quin mètode de cluster cal fer servir? quants grups hi ha?

Els mètodes de cluster són adequats quan es busquen possibles grups d'unes dades concretes. No hi ha cap mètode de cluster que sigui "sempre" millor que un altre. La seva eficàcia depèn de la naturalesa de les dades, de les relacions entre elles i de la distància o similaritat utilitzada. Per tal de decidir quin fer servir, caldrà doncs utilitzar algun índex que ens indiqui quin dels procediments de cluster descriu millor la relació existent entre les dades d'estudi. A la literatura hi ha bastants d'aquest índexs, un dels més coneguts és el coeficient de correlació cophenetica (Rohlf and Fisher, 1968; Rohlf and Sokal, 1981). Si fem servir aquest índex triarem aquell mètode amb un valor de correlació més alt, ja que estarem mesurant la correlació entre les distancies originals de les nostres dades i l'obtinguda pel mètode de cluster.

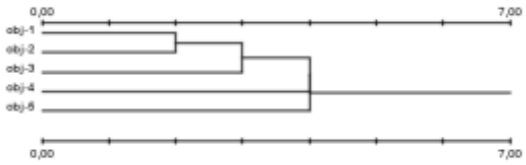
Un petit exemple.

Considerem la següent matriu de distàncies i dos mètodes de clúster, el single linkage i el UPGMA.

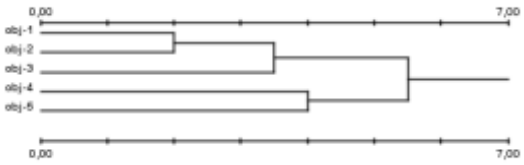
	obj-1	obj-2	obj-3	obj-4	obj-5
obj-1	1				
obj-2	2	0			
obj-3	3	4	0		
obj-4	4	5	6	0	
obj-5	5	6	7	4	0

Quin mètode descriu millor les possibles agrupacions?

Single linkage:



UPGMA:



Com el coeficient de correlació cophenetica de Pearson pel mètode single linkage és 0.781 i pel UPGMA és 0.840, ens hauríem de quedar amb aquest últim.

En un anàlisi cluster, moltes vegades és l'usuari que decideix en funció del dendrograma quants clústers o grups hi ha en les seves dades. Però aquesta decisió pot estar influenciada per idees preconcebudes. Cal doncs decidir el nombre de clústers a partir d'un índex. A Dudoit and Fridlyand (2002) es pot trobat una bona revisió dels

índexs existents. El més habitual és el denominat silhouette (Rousseeuw, 1987) malgrat aquest índex sempre considera que com a mínim hi ha dos clusters. Un altre índex que té en compte que pugui no existir estructura de cluster és el INCA (Irigoien and Arenas, 2007; Irigoien et al., 2012). Existeixen programes informàtics per a calcular aquests índexs.

Dudoit S. and Fridlyand J. (2002). A prediction-based resampling method for estimating the number of clusters in a data set. *Genome Biology* **3**: 36.1–36.21.

Irigoien I. and Arenas C. (2007). **INCA:** New statistic for estimating the number of clusters and identifying atypical units. *Statistics in Medicine* **27**: 2948-2973.

Irigoien I., Sierra B. and Arenas, C. (2012). ICGE: an R package for detecting relevant clusters and atypical units in gene expression. *BMC Bioinformatics* **13**: 30.

Rohlf F.J. and Fisher D.L. (1968). Test for hierarchical structure in random data sets. *Systematic Zoology* **17**: 407-412.

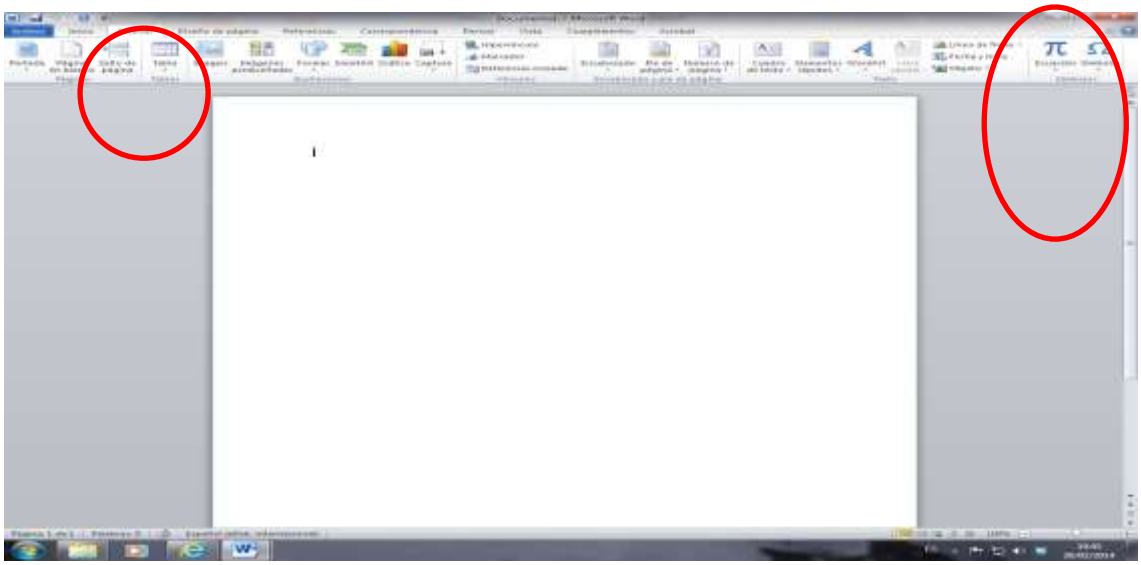
Rohlf F.J. and Sokal R.R. (1981). Comparing numerical taxonomic studies. *Systematic Zoology* **30**: 459-490.

Rousseeuw P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics* **20**: 53–65.

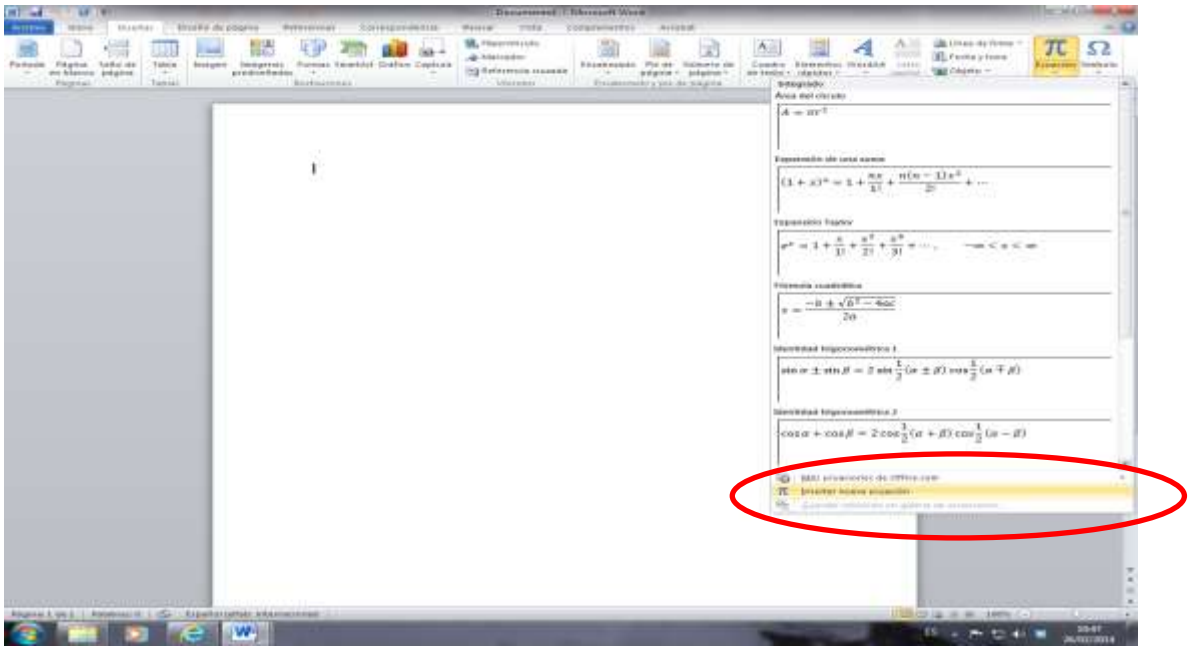
COM ESCRIURE FÓRMULES MATEMÀTIQUES AMB WORD?

Word té un editor d'equacions molt fàcil de fer servir. Només cal seguir els següents passos:

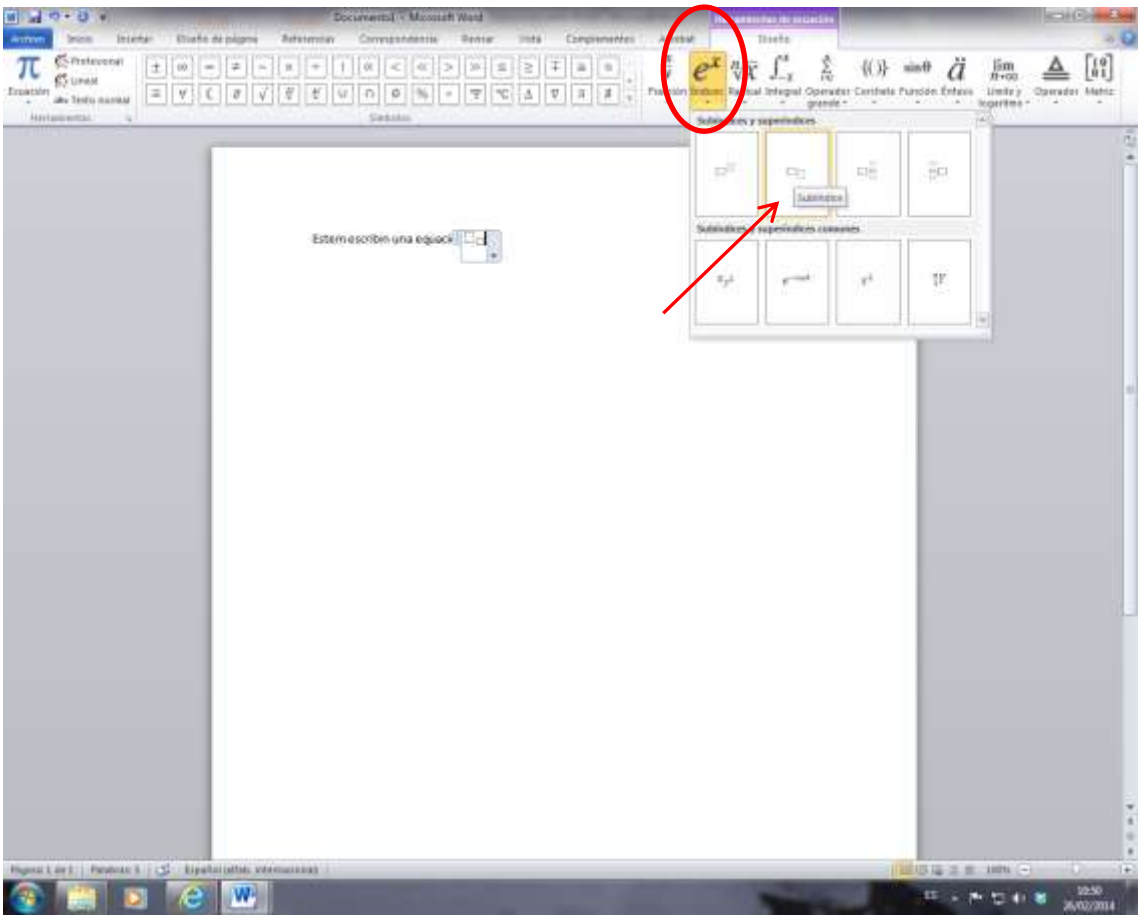
1. Anar a *Insertar -> Ecuacion*



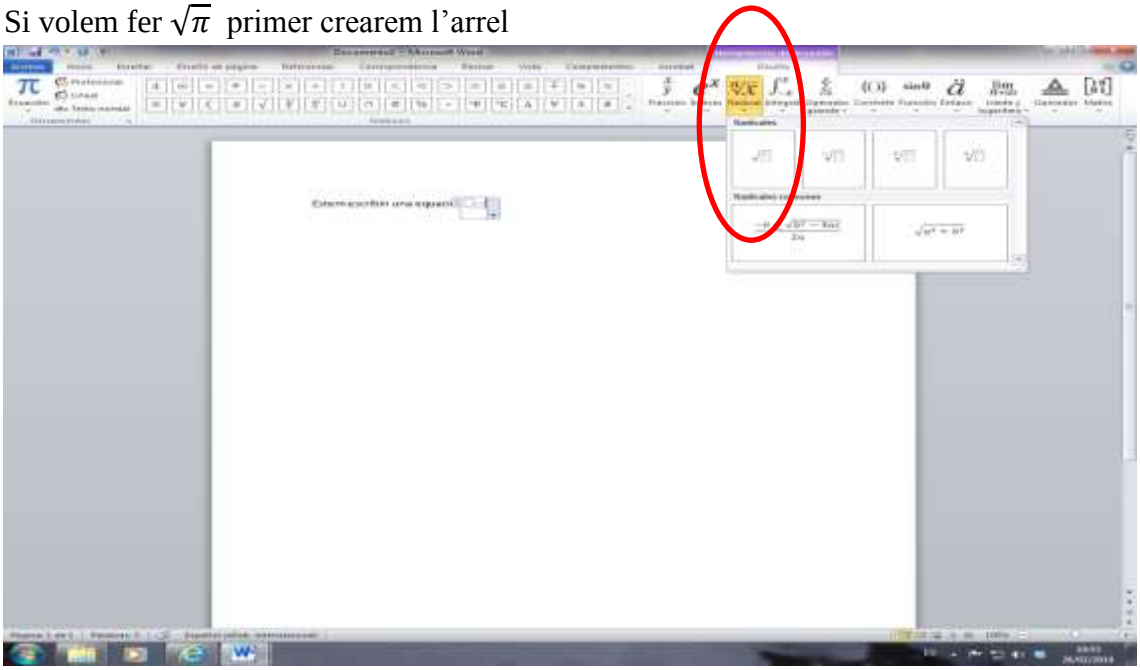
2. Escollir *Insertar nueva ecuacion*



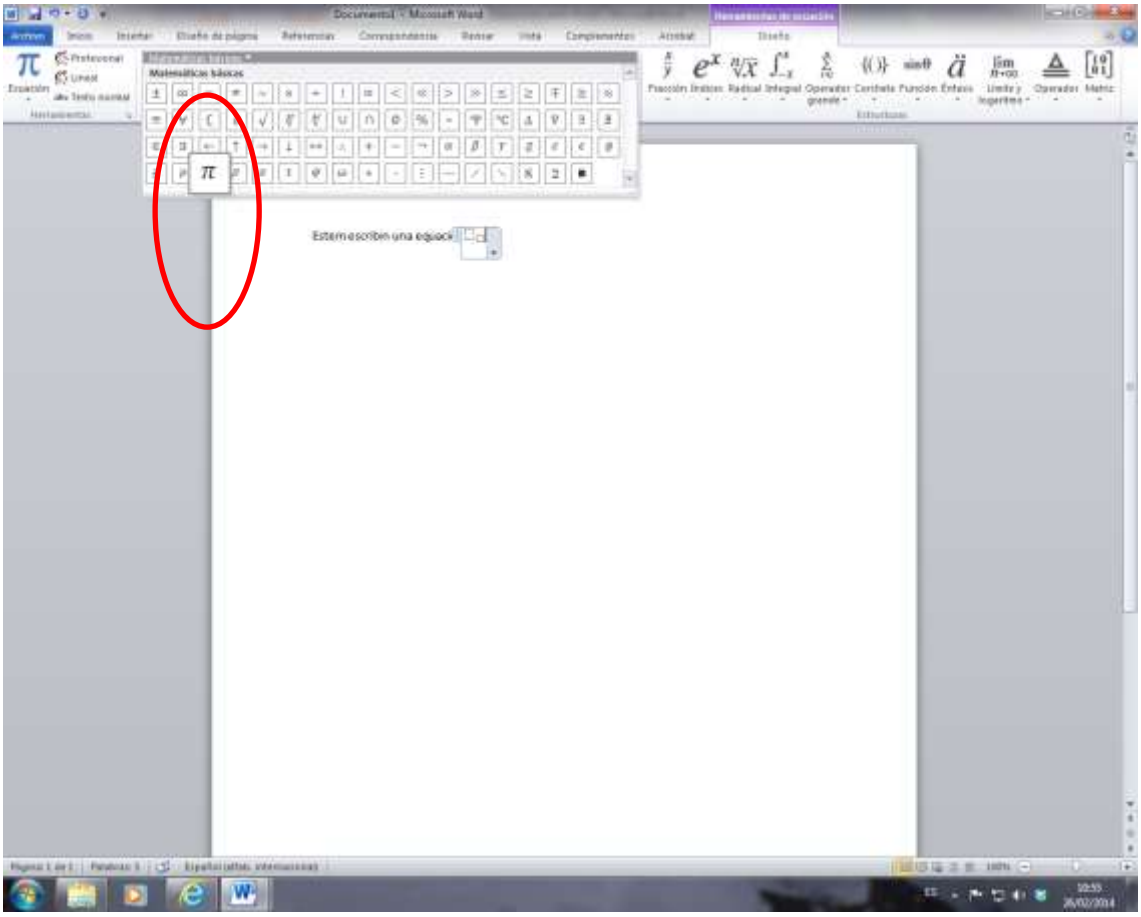
3. Podem començar a escriure en l'espai que s'obre. Si volem fer un subíndex del tipus x_{ij} triarem l'opció que indica la figura



4. Si volem fer $\sqrt{\pi}$ primer crearem l'arrel



I després afegirem el π



MITJANA ± DESVIACIÓ ESTÀNDARD? O MITJANA± ERROR ESTÀNDARD DE LA MITJANA?

Donats uns valors mostrals, quin índex cal posar en les taules o gràfiques, la desviació estàndard o l'error estàndard de la mitjana? Malgrat pugui semblar que donen la mateixa informació, conceptualment són molt diferents.

La desviació estàndard (SD) representa la variabilitat de les dades respecte de la mitjana mostral.

L'error estàndard de la mitjana (*Estándar Error of the Mean*, SEM) representa la precisió de la mitjana mostral com a estimador de la mitjana poblacional.

Entre si estan relacionats: $SEM = \frac{SD}{\sqrt{n}}$, on *n* representa la grandària mostral. L'error estàndard és inversament proporcional a la mida de la mostra, quan més mostra, més petit serà l'error estàndard.

Un petit exemple:

Suposem les següents dades mostrals, 50, 58, 55 i 60.

El valor de la mitjana és $m = \frac{50+58+55+60}{4} = 55{,}75.$
El valor de la desviació estàndard és
$SD = \sqrt{\frac{(50-55{,}75)^2+(58-55{,}75)^2+(55-55{,}75)^2+(60-55{,}75)^2}{4}} = 3{,}767.$
El valor del error estàndard de la mitjana és $SEM = \frac{3{,}767}{\sqrt{4}} = 1{,}883.$

Agafem ara la mostra 50, 58, 55, 60, 50, 58, 55, 60, 50, 58, 55 i 60.
En aquest cas la mitjana i la variabilitat de les dades no canvia.

El valor de la mitjana és $m = 55{,}75.$
El valor de la desviació estàndard és $SD = 3{,}767.$

Però si que canvia error estàndard de la mitjana .
El valor del error estàndard de la mitjana és $SEM = \frac{3{,}767}{\sqrt{12}} = 1{,}087.$

La variabilitat de les dades és manté, però l'error en l'estimació de la mitjana disminueix perquè tenim més mostra.

Si el que es vol indicar en un gràfic és el valor de la mitjana mostral i la variabilitat dels valors mostrals, caldria fer servir la desviació estàndard.
Si el que es vol indicar en un gràfic és el valor de la mitjana mostral i la seva precisió com a estimador de la mitjana poblacional, caldria fer servir l'error estàndard.

MDS and STRESS

MDS és un mètode multivariant per a representar en dimensió reduïda objectes a partir de les distàncies entre ells.

STRESS és una mesura per tal de saber si la representació gràfica en dimensió reduïda és prou bona. És un valor positiu i quan més gran pitjor és la representació.

La mesura més comuna del STRESS és la de Kruskal (Kruskal,1964). Ve donat per la fórmula

$$STRESS = \left[\frac{\sum_{ij}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{ij} d_{ij}^2}\right]^{1/2}$$

on, d_{ij} són les distàncies originals, i $\hat{d}_{ij}$ són les distàncies euclidianes calculades sobre la representació en dimensió reduïda.

Aleshores, segons els seus valors la representació gràfica és:

0,2 pobre

0,1 regular

0,05 bona

0,025 excel·lent

0 perfecta

Kruskal, J. B. (1964). Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika* **2**: 115-129.

OUTLIERS/VALORS EXTREMS

Donats uns valors observats, de vegades pot haver-hi outliers o observacions extremes. Es molt perillós, sense fer un bon estudi o sense conèixer molt bé el tipus de dades, eliminar aquells valors que puguin semblar "a ull" més diferents de la resta. A més aquesta apreciació "a ull" ve condicionada per el nombre d'observacions.

Per exemple, amb les següents dades

0,85
0,86
0,83
0,88
0,84
0,89
0,80
0,75
0,62

Podria semblar que el valor de 0,62 és un valor extrem.

Però si només disposem de les últimes 4 dades

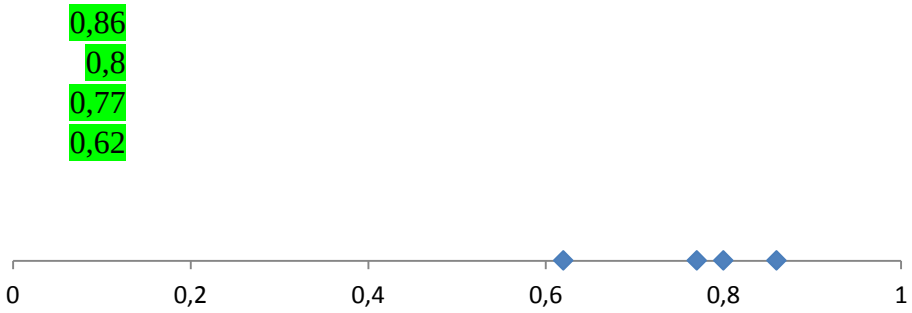
0,89
0,80
0,75
0,62

ja no és tan clar que ho sigui.

Podríem tenir la situació contraria. Per exemple, amb les següents dades

0,78
0,69
0,73
0,68
0,64
0,86
0,8
0,77

no observem cap valor extrem. Però si disposem només de les últimes 4 dades



podríem pensar que el valor 0,62 ho és.

El fet de tenir o no més valors mostrals pot estar influint en la nostra decisió "a ull".

- Cal deixar "parlar" les dades i no manipular-les sota criteris personals i/o no demostrables.
- En el cas de tenir realment un valor extrem, cal fer-se la pregunta, "per què aquesta dada és un valor extrem?" De vegades aquest valors, malgrat són poc probables, ens donen informació molt valuosa del nostre estudi.

ALGUNS TEST BASICS, PARAMÈTRICS I NO PARAMÈTRICS

	Test paramètric	Test no paramètric
Una mostra de mida n	Sota normalitat *Test t-Student d'una mostra*	*Test dels signes de Wilcoxon* Fer servir la distribució exacta si $n < 10$ *Test dels signes-rang de Wilcoxon* Fer servir la distribució exacta si $n < 25$
Dues mostres a) Amb dades APARELLADES mostres de mida n	Sota normalitat *Test t-Student de dades aparellades*	*Test dels signes de Wilcoxon* Fer servir la distribució exacta si $n < 10$ *Test dels signes-rang de Wilcoxon* Fer servir la distribució exacta si $n < 25$
b) Amb mostres INDEPENDENTS una de mida n_1 una de mida n_2	Sota normalitat i igualtat de variàncies *Test t-Student de dues mostres independents*	*Test U de Mann-Whitney* Fer servir la distribució exacta si $n_1 < 10$ i $n_2 < 10$

NORMALITAT: la normalitat de les dades es pot comprovar amb el test de Shapiro-Wilks

IGUALTAT DE VARIÀNCIES: la igualtat de variàncies per dues mostres amb distribució normal es pot comprovar amb el test *F* de Fisher-Snedecor de igualtat de variàncies.

POTÈNCIA D'UN TEST

En un test d'una hipòtesi H_0 (hipòtesi nul·la) versus una hipòtesi H_1 (hipòtesis alternativa) es poden cometre dos errors.

- Error tipus I => la hipòtesi H_0 és la certa, però arribem a la conclusió de que és falsa
- Error tipus II => la hipòtesi H_1 és la certa, però arribem a la conclusió de que és falsa

Caldrà doncs minimitzar la probabilitat d'aquests dos errors, P(Error tipus I) i P(Error tipus II). Normalment es treballa amb 1- P(Error tipus II) que és el que s'anomena Potència del test. Per tant cal:

- ❖ Minimitzar P(Error tipus I)
- ❖ Maximitzar la potència del test

COM ES CONTROLEN AQUESTES DUES PROBABILITATS?

Els tests estan construïts (pels estadístics/matemàtics) de forma que:

a) Es garanteix que:

> la probabilitat d'equivocar-nos si rebutgem H_0 quan és certa, és més petita o igual que el nivell de significació α fixat pel experimentador (normalment $\alpha = 0{,}05$).

Per tant sempre esta garantit que, P(Error tipus I)$\leq \alpha$.

Si es disminueix el valor de α (per exemple $\alpha = 0{,}01$) cal tenir en compte que com són "vasos comunicants" automàticament la potència del test disminueix.

b) La potència del test ve donada per les propietats que verifiqui el mètode estadístic (cal doncs conèixer per cada test aquestes propietats) però l'experimentador pot augmentar sempre la potència del test augmentant la grandària de la mostra,

> més mostra => més informació => més potència del test

NOTA: per mostres petites normalment els test tenen un valor de potència petit.

NOTA: els test no paramètrics són menys potents que els paramètrics.

NOTA: de vegades es fan estudis pilot per tal de determinar quina ha de ser la grandària de la mostra si es vol garantir un valor de potència determinat.

TAULES DE CONTINGENCIA: TEST DE KHI-QUADRAT,

TEST EXACTE DE FISHER I TEST DE McNEMAR

Aquest tres test s'utilitzen per a estudiar si dues variables qualitatives estan o no relacionades. Les dades es disposen en una taula MxN de contingència (de freqüències) on M i N representa el nombre de categories per les variables d'estudi.

Test de khi-quadrat: cal que la grandària mostral sigui gran i que el valor esperat de cada cel·la (sota la hipòtesi nul·la) sigui superior a 5 (alguns autors diuen que basta amb que ho verifiquin el 80% de les cel·les). Utilitza una aproximació asimptòtica.

Test exacte de Fisher: es pot utilitzar malgrat la mostra sigui petita i no utilitza cap aproximació.

Test de McNemar: s'utilitza quan una mateixa característica es mesura més d'una vegada sobre els mateixos individus i per tant no hi ha independència entre les mesures. En aquesta situació normalment es vol comparar les mesures realitzades en dos instants diferents (com abans i després d'un tractament)

Exemple amb instruccions del llenguatge R

1) Taula de contingència per a estudiar les diferències de la prevalença d'obesitat entre tres grups d'edat.

	OBESITAT	
	Si	NO
De menys de 20 anys	1	4
De 20 fins 60 anys	7	2
De més de 60 anys	4	8

#test de khi-quadrat; si l'aproximacio no es possible dona missatge d'error

#introduccio manual de les dades

```
Taula_contigencia_1 <-
matrix(c(1, 7, 4, 4, 2, 8),
      nrow = 3,
      dimnames =
      list(c("Menys de 20", "De 20 a 60","Més de 60"),
          c("Si", "No")))
Taula_contigencia_1
```

```
# test de khi-quadrat

Xsq <- chisq.test(Taula_contigencia_1)
Xsq$observed   # dona els valors observats
Xsq$expected   # dona els valors esperats sota hipòtesi nul·la
```

Amb les mateixes dades,

```
#test exacte de Fisher

fisher.test(Taula_contigencia_1,alternative = "two.sided")
```

2) Taula de contingència de 20 pacients intervinguts quirúrgicament als que se'ls valora el dolor després de la cirurgia però abans de prendre cap analgèsic i passada 1h després de l'administració d'un analgèsic.

	Dolor després del tractament	
Dolor abans tractament	Si	No
Si	1	11
No	2	6

```
#introduccio manual de les dades

Taula_contigencia_2 <-
matrix(c(1, 2, 11, 6),
      nrow = 2,
      dimnames =
      list(c("Si despres", "No despres"),
          c("Si amb trac", "No amb trac")))
Taula_contigencia_2

#McNemar test

mcnemar.test(Taula_contigencia_2)
```

links on s'explica les opcions de les següents funcions de R:

chisq.test : http://127.0.0.1:14428/library/stats/html/chisq.test.html

fisher.test : http://127.0.0.1:14428/library/stats/html/fisher.test.html

mcnemar.test :

http://stat.ethz.ch/R-manual/R-patched/library/stats/html/mcnemar.test.html

PUNTUACIONS Z-SCORE

Un z-score s'utilitza per a indicar quantes desviacions estàndard s'allunya una mesura del valor de la mitjana. Així, donats valors $x_1, \dots, x_n$, el z-score associat a una dada x_i, es calcula fent: $z_i = \frac{x_i - m}{s}$, on m i s són, respectivament, els valors de la mitjana i de la desviació estàndard de $x_1, \dots, x_n$.

♣ Un z-score menor (major) que 0 representa un element menor (major) que la mitjana.
♣ Un z-score igual a 4 (-4) representa un element que és 4 desviacions estàndard superior (inferior) a la mitjana.

De vegades s'associa el terme z-score amb la distribució normal reduïda N(0,1). Ja que donada una distribució normal N(μ,σ) si es tipifica $\frac{N(\mu,\sigma)-\mu}{\sigma}$ s'obté una N(0,1) i aquest procés de tipificació és el mateix que el que s'utilitza per trobar z-scores.

Exemple amb instruccions del llenguatge R

```
# entrada dades

dades<-c(12.3,65,84.3,55.21,78,63)

#calcul mitjana i desviacio

m <- mean(dades)

s <- sd(dades)   #la calcula dividint per n-1 (n numero de dades)

#calcul z-score

z_score <- (dades-m)/s

z_score
```

links on s'explica les opcions de les següents funcions de R:

mean: http://127.0.0.1:23954/library/base/html/mean.html

sd: http://127.0.0.1:23954/library/stats/html/sd.html