Estadística avançada
A2 Anàlisi descriptiu i inferencial

Noemi Lorente Torrelles

14 de noviembre, 2018

Contents

1 Introducció **3**

2 Càrrega de dades **4**

3 Estadística descriptiva **5**
3.1 Valors centrals . . . 5
3.2 Dispersió . . . 9
3.3 Càlcul manual de la dispersió . . . 9
3.4 Histograma . . . 10
3.5 Dades categòriques . . . 11

4 Estadística inferencial **15**
4.1 Interval de confiança . . . 15
4.2 Analitzar la capacitat pulmonar de les dones . . . 17
4.2.1 Escriure la hipòtesi nul · la i alternativa . . . 17
4.2.2 Mètode . . . 17
4.2.3 Calcular l'estadístic de contrast, el valor crític i el valor p. . . . 17
4.2.4 Interpretar el resultat . . . 19
4.3 Comparació entre fumadors i no fumadors . . . 20
4.3.1 Hipòtesi nul · la i alternativa . . . 20
4.3.2 Mètode . . . 20
4.3.3 Càlcul . . . 20
4.3.4 I al 99 % de confiança? Refeu els càlculs . . . 20
4.3.5 Interpretació . . . 20
4.4 Després de 5 anys . . . 21
4.4.1 Calculeu si hi ha diferències significatives en el grup de fumadors entre la capacitat pulmonar inicial i la capacitat pulmonar al cap de 5 anys. Realitzeu els passos necessaris. . . . 21
4.4.2 Realitzeu el mateix càlcul però ara pels no fumadors . . . 21
4.4.3 Interpreteu els resultats obtinguts en els dos contrastos . . . 21

5 t.value <- (mean(capacitatPC)−3.30)/(sd(*capacitat*PC) / sqrt(length(capacitat))) 22

6 p.value = 2*pt(-abs(t.value), df=length(data)-1) 22

7 You need the abs() function because otherwise you run the risk of getting p-values bigger than 1 22

8 (when the mean of the data is bigger than the given mean)! 22

9 Referències 23

1 Introducció

En aquesta activitat, usarem el fitxer resultant de l'activitat anterior degudament preprocessat. Aquest fitxer emmagatzema les dades d'una investigació mèdica sobre la capacitat pulmonar de vàries persones, amb l'objectiu d'estudiar si els hàbits de salut i els hàbits com a fumadors influencien la capacitat pulmonar.

Per a realitzar l'estudi es va recollir una mostra de 300 persones. A cada persona, se li va preguntar a través d'un qüestionari el seu gènere, hàbits d'esport, si era fumadora, i en cas que ho fos, quants cigarrets al dia de promig fumava i els anys que feia que fumava. A més, es va mesurar la capacitat pulmonar de cada persona a partir d'un test d'aire expulsat, des d'on es va prendre com a capacitat pulmonar la mesura FEF (forced expiratory flow), que és la velocitat de l'aire sortint del pulmó durant la porció central d'una espiració forçada. Es mesura en litres/segon.

Altres dades personals recollides són: l'alçada, pes i ciutat on viu. S'inclou en el fitxer una columna addicional "PC5Y" que és la capacitat pulmonar de cada persona mesurada al cap de 5 anys de realitzar el primer test. S'assumeix que la persona no ha canviat les seves condicions personals significativament en aquest temps.

En aquesta activitat usarem el fitxer de fumadors "net", és a dir, després del preprocés realitzat. Un cop el fitxer està preparat per a l'anàlisi, aplicarem anàlisis propis de l'estadística descriptiva i inferencial. Us proporcionem el fitxer Fumadores_clean_5Y.csv per a que tots treballeu amb el mateix fitxer de dades, independentment del resultat obtingut en l'activitat 1.

Nota important a tenir en compte per a lliurar l'activitat: * És necessari lliurar el fitxer Rmd i el fitxer de sortida (PDF o html). El fitxer de sortida ha d'incloure el codi i el resultat de la seva execució (pas a pas). S'ha de respectar la numeració dels apartats de l'enunciat. * No realitzeu llistat dels conjunts de dades, donat que aquests poden ocupar varies pàgines. Si voleu comprovar l'efecte d'una instrucció sobre les dades, podeu usar la funció *head* que mostra les 10 primeres files del conjunt de dades.

2 Càrrega de dades

Carregueu el fitxer de dades ***Fumadores_clean_5Y.csv*** *i valideu que els tipus de dades s'interpreten correctament.*

Abans de carregar el fitxer, he visualitzat el seu contingut amb *gedit* i he pogut comprobar que s'utilitza la coma "," com a separador de camps. Així m'asseguro que la lectura del fitxer es realitza de forma correcta.

```
# carrego el fitxer amb read.table, separador=',', decimal='.'
capacitat <- read.table("Fumadores_clean_5Y.csv", header=TRUE, sep=",", na.strings="NA",
                        dec=".", strip.white=TRUE, stringsAsFactors = FALSE)
```

Amb la funció **str** puc veure l'estructura interna del data frame *capacitat*. Veig que té 300 observacions, 10 variables i es mostra el nom de les variables.

```
# la funció str mostra l'estructura interna del data frame capacitat
str(capacitat)
```

```
## 'data.frame':    300 obs. of  10 variables:
##  $ Sex   : chr  "M" "F" "M" "M" ...
##  $ Sport : chr  "E" "E" "S" "N" ...
##  $ Years : int  25 18 0 25 0 0 33 0 0 5 ...
##  $ Cig   : int  10 32 0 14 0 0 15 0 0 12 ...
##  $ PC    : num  2.58 1.56 3.75 2.76 3.49 ...
##  $ City  : chr  "Barcelona" "Terrassa" "La Bisbal" "Blanes" ...
##  $ Weight: int  65 65 69 70 72 64 69 71 72 73 ...
##  $ Age   : int  49 35 38 55 55 42 55 44 45 35 ...
##  $ Height: int  171 166 175 176 178 165 175 177 178 179 ...
##  $ PC5Y  : num  2.53 1.44 3.73 2.67 3.49 ...
```

Com R no ha assignat correctament el tipus apropiat a les variables qualitatives nominals *Sex*, *Sport* i *City*, cal fer la conversió de caràcter a factor.

```
# Canvio a variable factor la variable Sex
capacitat$Sex <- as.factor(capacitat$Sex)
```

```
# Canvio a variable factor la variable Sport
capacitat$Sport <- as.factor(capacitat$Sport)
```

```
# Canvio a variable factor la variable City
capacitat$City <- as.factor(capacitat$City)
```

Per conèixer el tipus de variable utilitzo la funció **class**. A continuació, mostro el tipus de variable de cada variable. Per mostrar els tipus de les variables en una taula, utilitzo la funció **kable** .

```
# Recupero el tipus de variable
tipus <- sapply(capacitat,class)
kable(data.frame(Variable=names(tipus),Classe=as.vector(tipus)), align='l',
      caption="Tipus de les variables")
```

Table 1: Tipus de les variables

Variable	Classe
Sex	factor
Sport	factor
Years	integer
Cig	integer
PC	numeric
City	factor
Weight	integer
Age	integer
Height	integer
PC5Y	numeric

3 Estadística descriptiva

En primer lloc, estudiarem el valors valors centrals i dispersió d'algunes variables de la mostra. Seguiu els passos que s'especifiquen a continuació.

3.1 Valors centrals

Calcular la mitjana, mediana i els cinc nombres (de Tukey) de la capacitat pulmonar de la mostra.

Els cinc nombres de Tukey són el mínim, el màxim, el primer quartil Q1, el tercer quartil Q3 i el punt del mig Q2 o mediana. Amb la funció *summary* puc calcular els cinc nombres de Tukey i la mitjana.

```
# Amb la funció summary genero els cinc nombres de Tukey i la mitjana de la variable PC
summary(capacitat$PC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.557   2.909   3.554   3.331   3.793   4.466
```

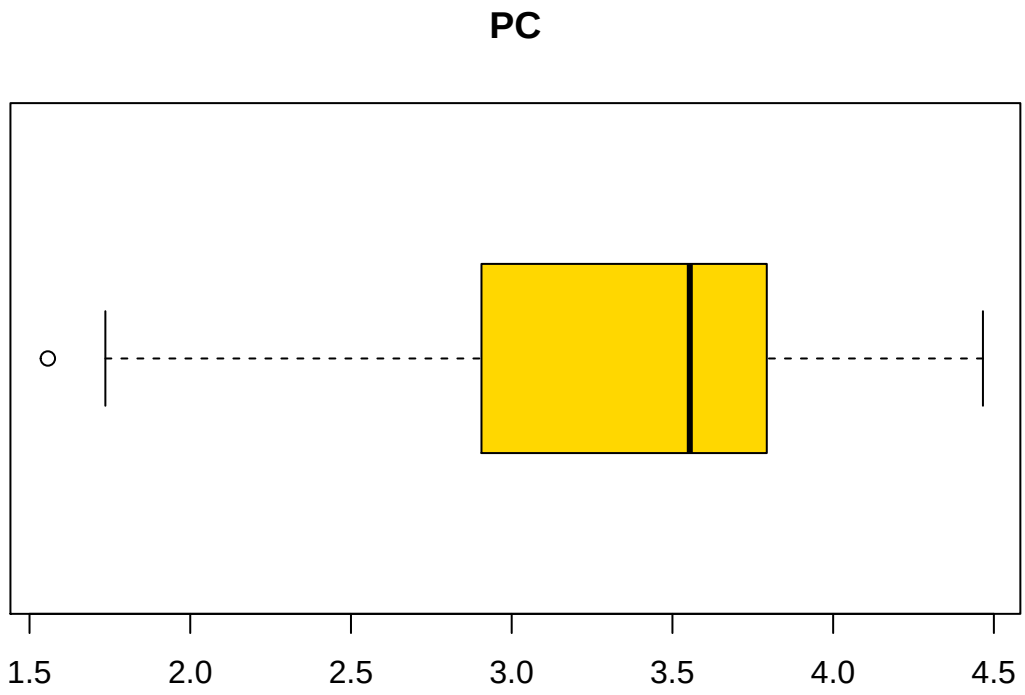
```
summary(capacitat$PC5Y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.444   2.744   3.543   3.289   3.796   4.472
```

Visualitzar-ho en un diagrama de caixa (boxplot). Es detecten valors extrems (outliers) en el diagrama?

Visualitzo el diagrama de caixa amb la funció *boxplot* i veig que hi ha un outlier. Per saber quin és aquest valor atípic de la variable *PC* utilizo el valor out de **boxplot.stats**.

```
# Genero boxplot de la variable PC
par(mfrow=c(1,1))
boxplot(capacitat$PC, main="PC", horizontal=TRUE, col="gold")
```




```
# Per saber quin son els valors atípics de la variable PC
boxplot.stats(capacitat$PC)$out
```

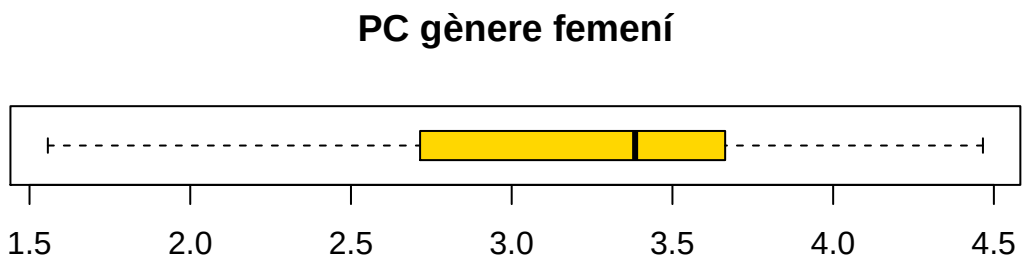
```
## [1] 1.557
```

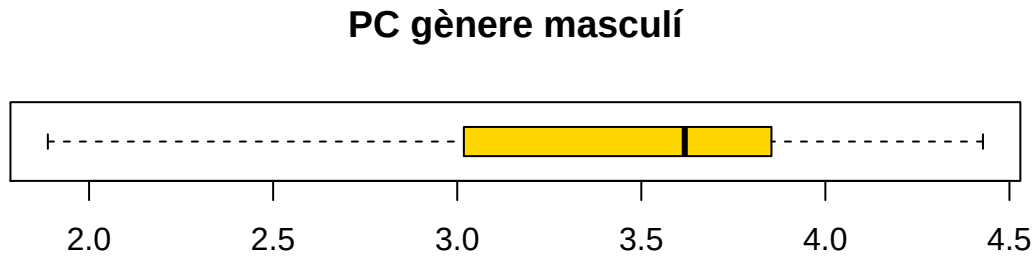
A continuació, mostrar un diagrama de caixa de la capacitat pulmonar pel gènere femení i pel masculí.

Visualitzant les dades de capacitat pulmonar, separant les dades per gènere, ja no es mostra cap outlier. Separant les dades per gènere, el outlier amb valor 1.557 es considera el valor mínim de les dades de capacitat pulmonar del gènere femení.

```
#  Genero boxplot de la variable PC separant les dades per gènere
par(mfrow=c(2,1))
idxF <- capacitat[capacitat$Sex=='F',]
idxM <- capacitat[capacitat$Sex=='M',]
```

```r
boxplot(idxF$PC, main="PC gènere femení", horizontal=TRUE, col="gold")
boxplot(idxM$PC, main="PC gènere masculí", horizontal=TRUE, col="gold")
```

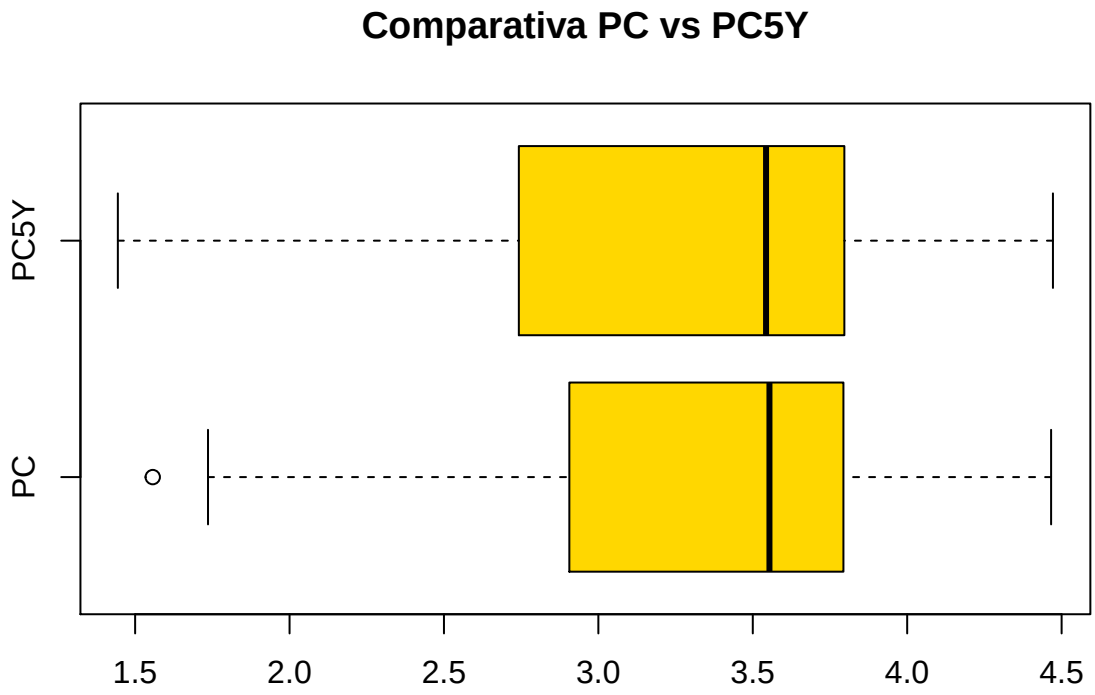




Mostreu, finalment, un diagrama de caixa que compari el valor de PC original i al cap de 5 anys. Interpreteu els resultats.

Comparant la capacitat pulmonar original (PC) i al cap de 5 anys (PC5Y), observo que: - La mediana, el 3er quartil i el valor màxim són similars. - El valor mínim i el 1er quartil són menors en la capacitat pulmonar al cap de 5 anys. - El rang interquartílic és més gran al cap de 5 anys, les dades estan més disperses.

```
# Per saber quin son els valors atípics de la variable PC
par(mfrow=c(1,1))
boxplot(capacitat$PC,capacitat$PC5Y, main="Comparativa PC vs PC5Y",
        names=c("PC","PC5Y"), horizontal=TRUE, col="gold")
```



3.2 Dispersió

Calculeu la dispersió de la capacitat pulmonar usant les mesures: variància, desviació típica i rang interquartílic.

Per calcular la dispersió de la capacitat pulmonar utilitzo les funcions següents:

- La funció **var** per al càlcul de la *variància*.
- La funció **sd** per al càlcul de la *desviació estàndard o típica*.
- La funció **IRQ** per al càlcul del *rang interquartílic (RIC)*.

```
# Calculo la variància
var(capacitat$PC)
```

```
## [1] 0.3937751
```

```
# Calculo la desviació estàndard
sd(capacitat$PC)
```

```
## [1] 0.627515
```

```
# Calculo el rang interquartílic
IQR(capacitat$PC)
```

```
## [1] 0.88475
```

3.3 Càlcul manual de la dispersió

Calculeu la desviació típica de la capacitat pulmonar manualment i compareu el resultat amb la funció corresponent d'R.

La desviació típica es defineix com l'arrel quadrada positiva de la variància: $s = \frac{\sqrt{\sum_{i=1}^{N}(x_i^2 - \overline{x}\check{s})}}{N}$

La diferència entre els dos càlculs és molt poca, varia la precisió a partir del tercer decimal, tal com podem veure al comparar la desviació estandard calculada amb la funció *sd* (0.6275150) i la desviació estandard calculada manualment (0.6264683).

La funció *sd* de R utilitza $n - 1$ en el denominador i en el càlcul manual he utilitzat n.

En estadística, la **correció de Bessel** és utilitzar $n - 1$ enlloc de n en la formula de la variancia mostral i desviació estandard mostral, on n és el nombre d'observacions de la mostra. Aquest mètode corregeix el biaix en l'estimació de la variància poblacional i en l'estimació de la desviació estandard poblacional. Tanmateix, la correcció sovint augmenta l'error quadrat mitjà en aquestes estimacions, tal com he pogut observa.

Per què es divideix entre n-1? Si és coneix la mitjana mostral i tots menys un dels valors, es pot calcular quin ha de ser aquest darrer valor. Per tant, en estadística es diu que hi ha n-1 graus de llibertat.

```
# Calculo manualment de la desviació típica
desviacio_manual <- function (x) sqrt(sum((x - mean(x))^2) / length(x))
```

```
comparar <- c(sd(capacitat$PC), desviacio_manual(capacitat$PC))
comparar
```

```
## [1] 0.6275150 0.6264683
```

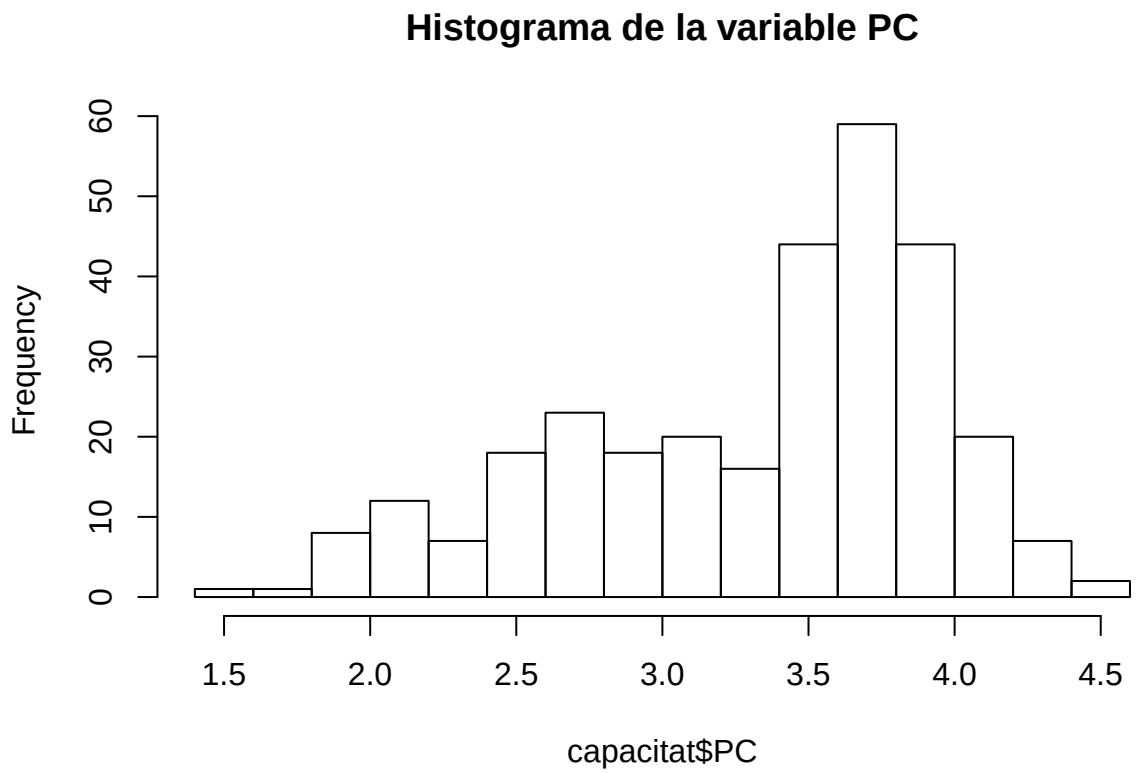
3.4 Histograma

Representar un histograma de PC de la mostra. Si és necessari, configureu els paràmetres de l'histograma per a que es vegi amb una bona precisió.

Per interpretar un *histograma* em fixo en la simetria; si no hi ha simetria, en les cues; en el nombre de pics que té (unimodal, bimodal); si hi ha classes buides i dades extremes, que separen la població.

Observo que la variable **PC** presenta assimetria a l'esquerra, cua a l'esquerra; només té un pic, és unimodal; hi ha persones en totes les classes creades per a la capacitat pulmonar, des del mínim 1.557 al màxim 4.466.

Amb el paràmetre *breaks* indico el nombre de cel · les del histograma i així aconsegueixo més precisió.



3.5 Dades categòriques

En les variables Sex, Sport i City, realitzeu el resum dels valors i dibuixeu un diagrama circular que mostri la proporció de casos cada tipus.

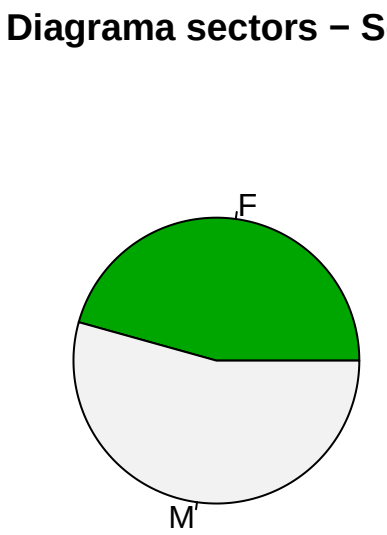
Per a variables qualitatives, la funció **summary** mostra les freqüencies absolutes dels valors de les variables. Si té molts valors, només mostra les freqüències d'alguns dels valors.

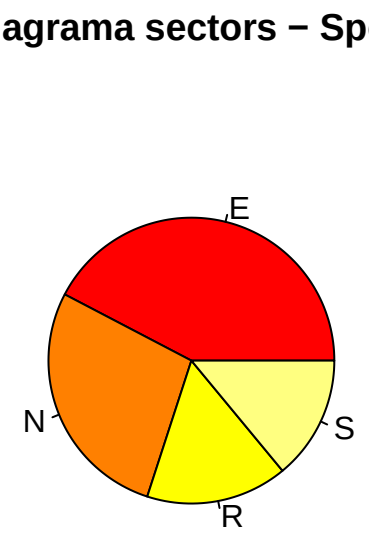
Table 2: Estadística descriptiva - Variables qualitatives

Sex	Sport	City
F:137	E:127	Barcelona:102
M:163	N: 83	Terrassa : 42
NA	R: 48	Valls : 15
NA	S: 42	Tarragona: 14
NA	NA	Lleida : 13
NA	NA	Sitges : 13
NA	NA	(Other) :101

– Per a les variables **Sex** i **Sport** utilitzo un *diagrama de sectors* perquè tenen pocs valors diferents. La variable **Sex** està gairebé repartida per igual entre homes i dones. En la variabe **Sport** es veu com gairebé el 50% no realitzen esport (N), aproximadament un 25% algunes vegades (S) i l'altre 25% entre regularment (R) i cada dia (E).

????????????????? https://www.theanalysisfactor.com/r-tutorial-part-14/





– Tot i que es demana representar la varialbe **City** amb un diagrama de sectors, com aquesta variable té més valors diferents, també utilitzo el *diagrama de barres* per a la seva representació. La ciutat de Barcelona és on viuen més persones de l'estudi, seguit de Terrassa. La resta de persones estan bastant repartides entre les altres 18 ciutats.

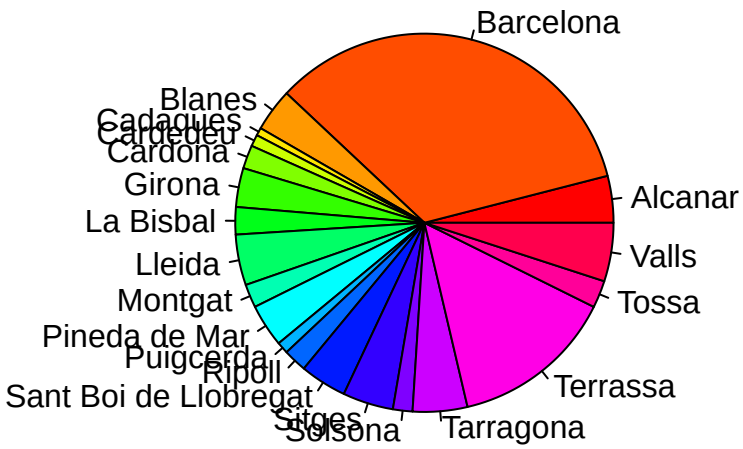


Diagrama sectors − Variable City
Barcelona
Blanes
Cadaques
Cardedeu
Cardona
Girona
La Bisbal
Lleida
Montgat
Pineda de Mar
Puigcerda
Ripoll
Sant Boi de Llobregat
Sitges
Solsona
Tarragona
Terrassa
Tossa
Valls
Alcanar

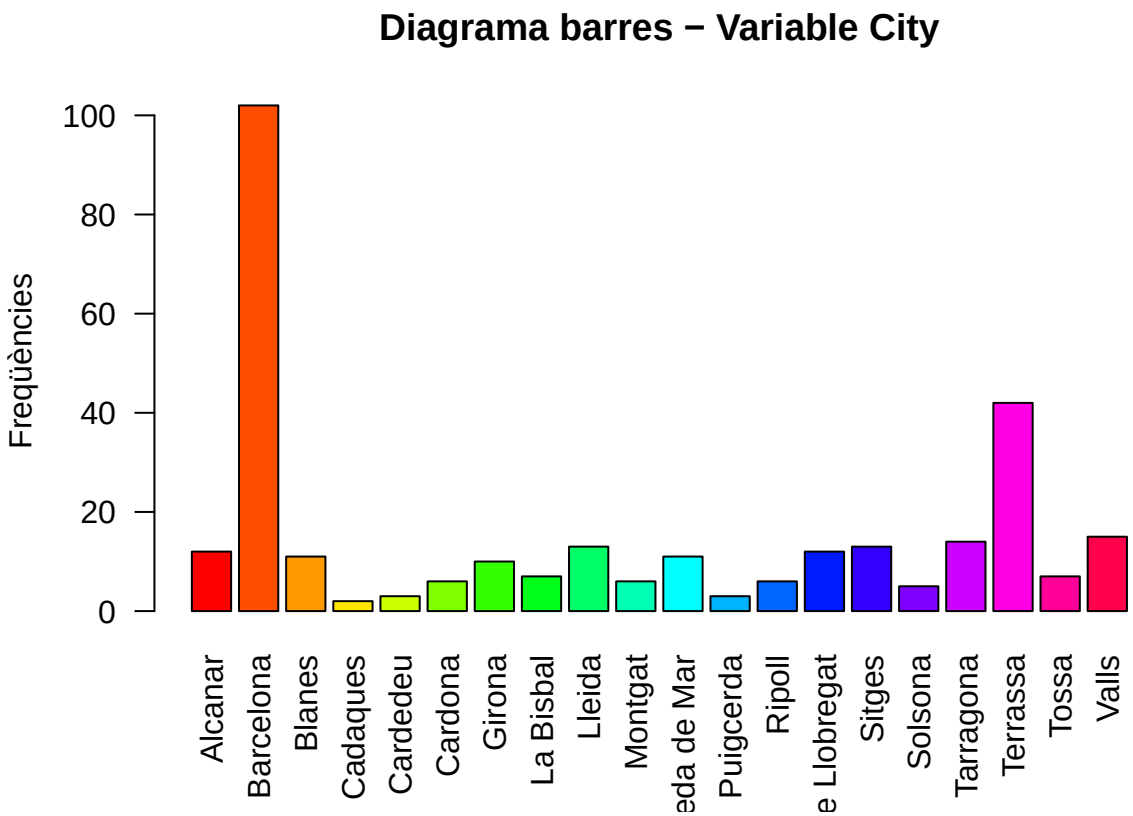


Diagrama barres − Variable City
Freqüències
100
80
60
40
20
0
Alcanar
Barcelona
Blanes
Cadaques
Cardedeu
Cardona
Girona
La Bisbal
Lleida
Montgat
eda de Mar
Puigcerda
Ripoll
e Llobregat
Sitges
Solsona
Tarragona
Terrassa
Tossa
Valls

4 Estadística inferencial

4.1 Interval de confiança

Calcular l'interval de confiança del 97 % de la capacitat pulmonar de la població.

Nota: S'han de realitzar els càlculs manualment. No es poden usar funcions d'R que calculin directament l'interval de confiança com t.test o similar. Sí que podeu usar funcions com qnorm, pnorm, qt i pt.

Suposem que la variable *PC* segueix una llei normal de mitjana μ (desconeguda) i desviació típica σ coneguda. Per calcular l'interval de confiança del 97% faig el següent:

1. Fixo el nivell de confiança **alpha** $(1 - \alpha = 1 - 0.97 = 0.03)$

```
alpha<- 1-0.97
```

2. Calculo l'**error estàndard** de la mitjana: $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ La desviació és igual a l'arrel quadrada de la variancia.

```
n <- length(capacitat$PC)
variancia <- var(capacitat$PC)
error_std <- sqrt(variancia) / sqrt(n)
error_std
```

```
## [1] 0.0362296
```

3. Calculo el **valor crític**, que és aquell punt $z_{\alpha/2}$ tal que $P(Z >= z_{\alpha/2}) = \frac{\alpha}{2}$, on Z és una variable N(0,1).

La funció *qnorm* retorna el quantil del valor donat en una distribució normal.??????????????????

$z_{\alpha/2} = z_{0.03/2}$

```
quantil <- qnorm(1 - alpha/2)
quantil
```

```
## [1] 2.17009
```

4. Calculo el **marge d'error**: $z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$

```
marge_error <- quantil * error_std
marge_error
```

```
## [1] 0.0786215
```

5. L'**interval de confiança** és el següent: (3.25, 3.41)

```
lim_inf <- mean(capacitat$PC) - marge_error
lim_inf
```

```
## [1] 3.252369
```

```
lim_sup <- mean(capacitat$PC) + marge_error
lim_sup
```

```
## [1] 3.409611
```

4.2 Analitzar la capacitat pulmonar de les dones

Assumim que coneixem la capacitat pulmonar mitjana de la població, que és igual a 3.30. Podem dir que la capacitat pulmonar de les dones és inferior a la mitjana poblacional, amb un nivell de confiança del 95 %? Per a respondre a aquesta pregunta, seguiu els passos que s'indiquen.

Nota: S'han de realitzar tots els càlculs manualment. No es poden usar funcions R que calculin directament el contrast com t.test o similar. Sí que podeu usar funcions com: qnorm, pnorm, qt i pt.

4.2.1 Escriure la hipòtesi nul · la i alternativa

La hipotesi nul · la és: la mitjana de la capacitat pulmonar de les dones és 3.30

$H_0 : \mu = 3.30$

La hipotesi alternativa és: la mitjana de la capacitat pulmonar de les dones és inferior a 3.30.

$H_1 : \mu < 3.30$

4.2.2 Mètode

Indiqueu quin és el mètode més apropiat per a fer aquesta anàlisi, en funció de les característiques de la mostra i l'objectiu de l'anàlisi.

Quan la mostra és prou gran, la solució ens ve donada per un dels resultats fonamentals de l'estadística: el teorema del límit central. Aquest teorema indica que si una mostra és prou gran (n>30), sigui quina sigui la distribució de la variable d'interès, la distribució de la mitjana mostral serà aproximandament una normal. A més, la mitjana serà la mateixa que la de la variable d'interès, i la desviació típica de la mitjana mostral serà aproximadament l'error estàndard.

En principi, no sabem si les dades provenen d'una distribució normal, però com la mostra és força gran (>30), podem aplicar un contrast d'hipòtesis. Amb el contrast d'hipòtesis, podem suposar una certa hipòtesi i decidir, a partir de les nostres observacions, si tenim prou evidències per a poder-la rebutjar. Les hipotesis s'expressen en termes d'algun paràmetre de la distribució de les dades que estudiem.

El nivell de significació α d'un contrast és l'error màxim de tipus I que estem disposats a assumir, és a dir, rebutjar H_0 quan aquesta és certa.

Com l'objectiu és assegurar un nivell de confiança del 95% ($p = 0.95$), implica que el nivell de significació és del 5% ($\alpha = 1 - p = 0.05$). Per tant, no rebutjarem la hipòtesi nul · la si $\alpha < 0.05$.

4.2.3 Calcular l'estadístic de contrast, el valor crític i el valor p.

L'estadístic de contrast z és una funció de la mostra de la qual en coneixem la distribució sota la hipòtesi nul · la.

Com tenim una mostra de n=300 persones escollides a l'atzar, aleshores sota la hipòtesi nul · la (és a dir, $\mu = 3.30$), defineixo la variable $z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ com estadístic de contrast.

1. Calculo la mitjana de la capacitat pulmonar de les dones:

```
# Calculo la mitjana PC de les dones
meanF <- mean(capacitat[capacitat$Sex=='F',]$PC)
meanF
```

```
## [1] 3.215241
```

2. Com la desviació típica poblacional és desconeguda, cal fer una estimació de la desviació típica amb l'anomenada desviació típica mostral, on es divideix per $n-1$, enlloc de n, tal com s'explica en la pàgina 8 dels apunts de Teorema del límit central. Per tant, utilitzo la funció *sd* per calcular la desviació típica poblacional.

```
# Calculo la desviació poblacional
desv_poblacional <- sd(capacitat$PC)
desv_poblacional
```

```
## [1] 0.627515
```

3. Calculo l'estadístic de contrast z:

```
z <- (meanF - 3.3) / (desv_poblacional / sqrt(300))
z
```

```
## [1] -2.3395
```

4. Calculo el valor crític z_α.

Com el contrast que realitzem és unilateral, és a dir, només es compara en una direcció, per a determinar el valor crític cal imposar que: α, $P(Z < z_\alpha) = \alpha$, on Z és la distribució de l'estadístic de contrast.

```
valor_critic <- qnorm(0.05)
valor_critic
```

```
## [1] -1.644854
```

Sota la hipòtesi nul · la, està distribuït com una normal estandard. Per $\alpha = 0.05$, aleshores $z_\alpha = -1.65$

La regla de decisió és la següent:

- Acceptarem H_0 si $z >= z_\alpha$
- Rebutjarem H_0 si $z < z_\alpha$

Com z és menor que z_α (-2.3395 < 1.65), rebutjo H_0.

5. Calculo el p-valor.

En molts casos, per a resoldre un contrast d'hipòtesis, no calcularem el valor crític, sinó que utilitzarem l'anomenat *p-valor*.

El *p-valor* és la probabilitat del resultat de l'estadístic de contrast observat o d'un de més allunyat quan la hipòtesi nul · la és certa, és a dir, el *p-valor* és el nivell de significació més petit que ens permet de rebutjar la hipòtesi nul · la.

Per tant, si el *p-valor* és inferior al nivell de significació α, rebutjarem la hipòtesi nul · la. Si el *p-valor* és superior o igual al nivell de significació α, acceptarem la hipòtesi nul · la.

La funció *pnorm* permet calcular el p-valor. Emprem el paràmetre *lower.tail=TRUE* per a determinar el càlcul de la probabilitat de la cua de l'esquerra.

```
# Calculo el p-valor P(Z < -2.3395) amb cua cap a l'esquerra, cua inferior
pnorm(-2.3395, lower.tail=TRUE)
```

```
## [1] 0.009654786
```

Com p-value = 0.009 és menor que el nivell de significació α (0.05), rebutjo la hipòtesi nul · la H_0.

4.2.4 Interpretar el resultat

Amb les dues maneres de resolució de contrast d'hipòtesi arribo a la mateix conclusió, es rebutja la hipotesi nul · la H_0, i això implica, acceptar la hipotesi alternativa, és a dir, la capacitat pulmonar de les dones és inferior a la mitjana poblacional 3.30 amb un nivell de confiança del 95%.

4.3 Comparació entre fumadors i no fumadors

Ens preguntem si la capacitat pulmonar dels fumadors és inferior a la capacitat pulmonar dels no fumadors. Apliqueu un test d'hipòtesis que testegi aquesta hipòtesi amb un 95 % de confiança i interpreteu el resultat.

Nota: S'han de realitzar tots els càlculs manualment. No es poden usar funcions R que calculin directament el contrast com t.test o similar. Sí que podeu usar funcions com: qnorm, pnorm, qt i pt.

Seguiu els passos que s'indiquen a continuació.

4.3.1 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i alternativa.

4.3.2 Mètode

Expliqueu el mètode que aplicareu per a realitzar aquest contrast i justifiqueu-lo.

contrast de dues mostres

4.3.3 Càlcul

Realitzeu el càlcul. A l'igual que anteriorment, no podeu usar funcions d'R o llibreries que calculin directament el contrast. Heu de realitzar el càlcul manualment. Podeu usar funcions del tipus qnorm, pnorm, qt, pt.

4.3.4 I al 99 % de confiança? Refeu els càlculs

4.3.5 Interpretació

Interpreteu els resultats obtinguts.

4.4 Després de 5 anys

Després de 5 anys es mesura de nou la capacitat pulmonar de les mateixes persones de l'estudi. La columna PC5Y incorpora la capacitat pulmonar dels mateixos subjectes al cap de 5 anys. Ens preguntem si la capacitat pulmonar ha canviat significativament, amb un nivell de confiança del 95 % en el cas dels fumadors i en el cas dels no fumadors. Responeu a les preguntes següents.

Nota: no podeu usar funcions d'R que ja calculin el contrast directament, a l'igual que en els exercicis previs.

4.4.1 Calculeu si hi ha diferències significatives en el grup de fumadors entre la capacitat pulmonar inicial i la capacitat pulmonar al cap de 5 anys. Realitzeu els passos necessaris.

_ Escriviu la hipòtesi nul · la i alternativa, el mètode que escolliu i els càlculs._

Nota: Si definiu una funció que realitzi aquest càlcul, podreu usar-la en aquest apartat i en el següent, on cal repetir el mateix càlcul per a unes altres dades.

4.4.1.1 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i alternativa.

4.4.1.2 Mètode

Indiqueu el mètode que usareu i la seva justificació.

4.4.1.3 Càlcul

Realitzeu els càlculs necessaris.

4.4.2 Realitzeu el mateix càlcul però ara pels no fumadors

4.4.3 Interpreteu els resultats obtinguts en els dos contrastos

PER ESBORRAR

$N(\mu, \sigma^2)$. # Student t-Test t.test(x=capacitat$PC, mu=3.3, conf.level=0.95)

5 t.value <- (mean(capacitat$PC) − 3.30)/(sd(capacitat$PC) / sqrt(length(capacitat)))

6 p.value = 2*pt(-abs(t.value), df=length(data)-1)

7 You need the abs() function because otherwise you run the risk of getting p-values bigger than 1

8 (when the mean of the data is bigger than the given mean)!

9 Referències

- Rmarkdown cheat sheet
  https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf
- Rmarkdown: The Definitive Guide
  https://bookdown.org/yihui/rmarkdown/pdf-document.html
- RDocumentation kNN
  https://www.rdocumentation.org/packages/VIM/versions/4.7.0/topics/kNN
- RDocumentation write.table
  https://www.rdocumentation.org/packages/utils/versions/3.5.1/topics/write.table
- RDocumentation kable
  https://www.rdocumentation.org/packages/knitr/versions/1.20/topics/kable

https://en.wikipedia.org/wiki/Bessel%27s_correction

https://www.theanalysisfactor.com/r-tutorial-part-14/

http://web.ift.uib.no/Teori/KURS/WRK/TeX/symALL.html

https://stats.stackexchange.com/questions/25956/what-formula-is-used-for-standard-deviation-in-r