PAC4 Primavera 2021

UOC

Les PACs es basaran en una base de dades obtinguda a partir del repositori de microdades del "Banc Mundial" a https://microdata.worldbank.org/index.php/catalog/424/get-microdata

Conté indicacions, entre d'altres de

1. *City* = Nom de la ciutat
2. *Country* = País
3. *Population2000* = Població de la ciutat a l'any 2000.
4. *PM10Concentration1999 = PM10 concentrations (micro grams per cubic meter) in residential areas of cities larger than 100,000*, l'any 1999
5. *Region* = Classificació en regió geogràfica
6. *IncomeGroup* = Classificació segons nivell d'ingressos del país.

Per importar les dades podem usar la següent instrucció i per comprovar que tot funcioni mostrem els 3 primers registres:

```
dadesPM10<-read.table("AirPollution2000WB_UOC2.csv", header=TRUE,
          sep=";",na.strings="NA",
          fileEncoding = "UTF-8", quote = "\"",
          colClasses=c(rep("character",4),rep("numeric",2),
                       rep("character",2)))
```

Us pot ser útil consultar el següent material:

1. Mòdul Intervals de confiança.
2. Activitats Resoltes del Repte 3 (Intervals de confiança).
3. Procureu usar les funcions pròpies de R per fer els càlculs a no ser que s'indiqui el contrari.

NOM:

PAC4

Un cop importades les dades, amb la mateixa base de dades i suposant que les dades corresponen a una mostra

Pregunta 1 (25%)

Trobeu un interval de confiança de la concentració de partícules *PM10Concentration1999* a les ciutats de l'Índia

1. Amb un nivell de confiança del 95%.
2. Amb un nivell de confiança del 90%.
3. Compareu els intervals trobats i comenteu quin és més gran i el perquè.
4. Compareu els intervals amb la mitjana de la concentració de partícules de totes les ciutats de la mostra.

Solució

1. Calculem

```
attach(dadesPM10)
t.test(PM10Concentration1999[Country=="India"],conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  PM10Concentration1999[Country == "India"]
## t = 41.978, df = 318, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  72.16306 79.26013
## sample estimates:
## mean of x
##   75.7116
```

   L'interval de confiança demanat serà: $(72.1631, 79.2601)$.

2. En aquest cas

```
t.test(PM10Concentration1999[Country=="India"],conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  PM10Concentration1999[Country == "India"]
## t = 41.978, df = 318, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
```

```
##  72.73624 78.68696
## sample estimates:
## mean of x
##   75.7116
```

L'interval de confiança demanat serà: $(72.7362, 78.687)$.

3. L'interval de confiança del 95% és més llarg perquè la probabilitat de contenir la veritable mitjana és més alta que en el del 90%.

4. La mitjana és

```
mean(PM10Concentration1999)
```

```
## [1] 51.10099
```

molt per sota dels intervals de confiança, el que ens informa que la contaminació a les ciutats de l'Índia és força superior a la mitjana general.

Pregunta 2 (25%)

Trobeu un interval de confiança al 80% per a la mitjana de la població l'any 2000 de les ciutats de la Índia i un altre per a les de Japó. Quina conclusió podem extreure sobre la mida de les ciutats als dos països? (en particular fixeu-vos en les mitjanes i els intervals de confiança).

Solució

```
t.test(Population2000[Country=="Japan"],conf.level = 0.8)
```

```
##
##  One Sample t-test
##
## data:  Population2000[Country == "Japan"]
## t = 5.9698, df = 216, p-value = 9.644e-09
## alternative hypothesis: true mean is not equal to 0
## 80 percent confidence interval:
##  289724.2 448736.5
## sample estimates:
## mean of x
##  369230.3
```

```r
t.test(Population2000[Country=="India"],conf.level = 0.8)
```

```
##
##  One Sample t-test
##
## data:  Population2000[Country == "India"]
## t = 7.12, df = 318, p-value = 7.228e-12
## alternative hypothesis: true mean is not equal to 0
## 80 percent confidence interval:
##  472473.1 680417.1
## sample estimates:
## mean of x
##  576445.1
```

La mitjana de la població de les ciutats de l'Índia és força més alta i els intervals de confiança són disjunts cosa que confirma que les ciutats a l'Índia són efectivament més poblades que les del Japó.

Pregunta 3 (25%)

Volem estudiar la proporció de ciutats que corresponen a països d'ingressos baixos ("Low income"); per fer-ho:

1. Calculeu un interval de confiança per aquesta proporció del 88% usant la funció *prop.test* amb l'opció *correct=FALSE*.
2. Calculeu el mateix interval seguint les fórmules de les notes d'estudi (pot haver petites diferències al resultat).

Solució

1. Amb *prop.test* obtenim

```r
LI<-table(IncomeGroup)["Low income"]
T<-nrow(dadesPM10)
LI
```

```
## Low income
##         99
```

```r
T
```

```
## [1] 3218
```

```r
PB<-LI/T
PB
```

```
## Low income
## 0.03076445
```

```r
prop.test(LI, T, alternative='two.sided', p=.5, conf.level=.88, correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  LI out of T, null probability 0.5
## X-squared = 2834.2, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 88 percent confidence interval:
##  0.02637260 0.03586073
## sample estimates:
##          p
## 0.03076445
```

2. Si ho calculem manualment hem d'aplicar la següent fórmula

$$\left(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

necessitem ara

```r
za2<--qnorm((1-0.88)/2)
SPB<-sqrt(PB*(1-PB)/T)
II<-PB-za2*SPB
II
```

```
## Low income
##  0.0260317
```

```r
IS<-PB+za2*SPB
IS
```

```
## Low income
##  0.0354972
```

amb la qual cosa l'interval queda com $(0.0260317, 0.0354972)$, pràcticament igual que amb *prop.test*.

Pregunta 4 (25%)

Simularem ara intervals de confiança de la variable *PM10Concentration1999*. Concretament, de manera similar al que apareix al document d'activitats resoltes corresponents als intervals de confiança:

1. Generarem 1000 mostres aleatòries de mida 200 obtingudes de la variable *PM10Concentration1999* amb reemplaçament (vegeu *sample*). Per poder comprovar els resultats que obtingueu usarem la instrucció *set.seed(n)* abans de generar les mostres on n serà el nombre de lletres del vostre primer cognom; per exemple si el primer cognom és "Sasdtdmdswrtas" usarem *set.seed(14)*. Escriviu les cinc primeres columnes de les dues primeres mostres amb la instrucció *head*.
2. A continuació calcularem un interval de confiança per a cada mostra, al nivell del 95%, suposant que no coneixem la desviació típica de la variable. Mostreu els tres primers intervals.
3. Finalment comptarem quants dels intervals contenen la mitjana de la variable.
4. En acabar comenteu els resultats.

Solució

1. Preparem les mostres usant *set.seed(14)*

```
set.seed(14)
NUMMOSTR<-1000
MOSTR<-c()
for (i in 1:NUMMOSTR)
  {MOSTR<-cbind(MOSTR,sample(PM10Concentration1999,200,rep=TRUE))}
head(MOSTR[1:2,1:5],2)
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]   22   36   25   21   27
## [2,]   15   23   93  100   86
```

```
print("o bé")
```

```
## [1] "o bé"
```

```
head(MOSTR[1:5,1:2],5)
```

```
##      [,1] [,2]
## [1,]   22   36
## [2,]   15   23
## [3,]   68   21
## [4,]   22   60
## [5,]   86   74
```

2. Generem els intervals de confiança amb *t.test*:

```
INTCONF<-c()
for (i in 1:NUMMOSTR) {INTCONF<-rbind(INTCONF,t.test(MOSTR[,i],
                                  conf.level = 0.95)[[4]])}
head(INTCONF,3)
```

```
##          [,1]     [,2]
## [1,] 46.94095 57.77905
## [2,] 43.19951 52.02049
## [3,] 46.98851 58.75149
```

3. Comptem ara quants intervals contenen la mitjana de la variable:

```
MIT<-mean(PM10Concentration1999)
MIT
```

```
## [1] 51.10099
```

```
CONT<-0
for (i in 1:NUMMOSTR) {
  if(INTCONF[i,1]<=MIT && INTCONF[i,2]>=MIT)
  {CONT<-CONT+1}}
CONT
```

```
## [1] 949
```

4. Haurem d'esperar que el 95% dels intervals (950) continguin la mitjana i en realitat són 949, un valor força aproximat o igual, depenent del valor introduït a *set.seed*.