PAC5 Primavera 2021 - Solució

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:

```
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. Mòduls Contrast d'hipòtesi i Contrast de dues mostres
2. Activitats resoltes del Repte 4

NOM:

PAC5

Pregunta-1 (25%)

Contrasteu amb un nivell de significació del 5% si la concentració mitjana de PM10 de l'any 1999 a les ciutats dels Estats Units és inferior a 25. Indiqueu les hipòtesis nul · la i alternativa. A partir de la sortida de R indiqueu el valor de l'estadístic de contrast, el p-valor i la conclusió a la que arribeu. Suposeu que les observacions corresponen a una mostra i que la variable considerada és normal. Atenció: feu servir la funció de R que toqui, és a dir, no feu els càlculs manualment amb les fòrmules de les notes d'estudi.

```
t.test(dadesPM10$PM10Concentration1999[dadesPM10$Country == "United States of America"],
       mu = 25, conf.level = 0.95, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  dadesPM10$PM10Concentration1999[dadesPM10$Country == "United States of America
## t = -2.0834, df = 207, p-value = 0.01922
## alternative hypothesis: true mean is less than 25
## 95 percent confidence interval:
##      -Inf 24.80598
## sample estimates:
## mean of x
##   24.0625
```

Es tracta d'un contrast sobre la mitjana de la concentració de PM10 de l'any 1999 a les ciutats dels Estats Units.

Hipotesi nul · la: $H_0 : \mu = 25$, hipòtesi alternativa: $H_1 : \mu < 25$

Estadístic de contrast: $t = -2.0834$ que, sota hipòtesi nul · la certa, correspon a una observació d'una distribució t de Student amb 207 graus de llibertat.

El p-valor=0.01922. Donat que $0.01922 < 0.05$ rebutgem la hipòtesi nul · la i concloem que la mitjana de la concentració de PM10 de l'any 1999 és inferior a 25.

Pregunta-2 (25%)

Contrasteu amb un nivell de significació del 5% si la concentració mitjana de PM10 de l'any 1999 és més baixa a les ciutats dels Estats Units que a les de Xina. Indiqueu les hipòtesis nul · la i alternativa. A partir de la sortida de R indiqueu el valor de l'estadístic de contrast, el p-valor i la conclusió a la que arribeu. Suposeu que les observacions corresponen a mostres i que les variables considerades són normals i amb variàncies iguals. Atenció: feu servir la funció de R que toqui, és a dir, no feu els càlculs manualment amb les fòrmules de les notes d'estudi.

```
dadesUSA <- dadesPM10$PM10Concentration1999[dadesPM10$Country == "United States of Ameri
dadesChi <- dadesPM10$PM10Concentration1999[dadesPM10$Country == "China"]
t.test(dadesUSA, dadesChi, conf.level = 0.95, var.equal = TRUE, alternative = "less")
```

```
##
##  Two Sample t-test
##
## data:  dadesUSA and dadesChi
## t = -40.377, df = 586, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -54.38213
## sample estimates:
## mean of x mean of y
##  24.06250  80.75789
```

Es tracta d'un contrast sobre la diferència de mitjanes de la concentració de PM10 de l'any 1999 entre les ciutats dels Estats Units i les ciutats de Xina.

Hipòtesi nul · la: $H_0 : \mu_U = \mu_C$, hipòtesi alternativa: $H_1 : \mu_U < \mu_C$

Estadístic de contrast: $t = -40.377$ que, sota hipòtesi nul · la certa, correspon a una observació d'una distribució t de Student amb 586 graus de llibertat.

El p-valor és inferior a $2.2 \cdot 10^{-16}$. Donat que el p-valor és pràcticament nul rebutgem la hipòtesi nul · la i concloem que la mitjana de la concentració de PM10 de l'any 1999 a les ciutats dels Estats Units és inferior a la mitjana de les ciutats de Xina.

Pregunta-3 (25%)

Contrasteu amb un nivell de significació de l' 1% si la proporció de ciutats corresponents a països d'ingressos alts és major que el 30%. Indiqueu les hipòtesis nul · la i alternativa. A partir de la sortida de R indiqueu el p-valor i la conclusió a la que arribeu. Suposeu que les observacions corresponen a una mostra. Atenció: feu servir la funció de R que toqui i a més a més feu els càlculs manualment amb les fòrmules de les notes d'estudi.

```
totalCity <- length(dadesPM10$City)
totalCityIngHigh <- length(dadesPM10$City[dadesPM10$IncomeGroup == "High income"])
totalCity           # total ciutats
```

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

```
totalCityIngHigh    # total ciutats ingressos alts
```

```
## [1] 1095
```

```
prop.test(totalCityIngHigh, totalCity, p = 0.30, correct = TRUE,
          alternative = "greater", conf.level = 0.99)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  totalCityIngHigh out of totalCity, null probability 0.3
## X-squared = 24.663, df = 1, p-value = 3.414e-07
## alternative hypothesis: true p is greater than 0.3
## 99 percent confidence interval:
##  0.3209729 1.0000000
## sample estimates:
##         p
## 0.3402735
```

Es tracta d'un contrast sobre la proporció.

Hipòtesi nul · la: $H_0 : p = 0.30$, hipòtesi alternativa: $H_1 : p > 0.30$

El p-valor=$3.414e-07$. Donat que $3.414e-07 < 0.01$ rebutgem la hipòtesi nul · la i concloem que la proporció de ciutats corresponents a països d'ingressos alts és major que el 30%.

Observació:

En la instrucció prop.test es pot incloure l'argument "correct=FALSE". En l'ajuda de "prop.test", es comenta que amb aquesta opció no es té en compte la correcció de continuïtat de Yates. Tal i com es diu en les notes d'estudi, podem considerar com correctes els dos resultats, que són molt semblants:

```
prop.test(totalCityIngHigh, totalCity, p = 0.30, correct = FALSE,
          alternative = "greater", conf.level=0.99)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  totalCityIngHigh out of totalCity, null probability 0.3
## X-squared = 24.854, df = 1, p-value = 3.091e-07
## alternative hypothesis: true p is greater than 0.3
## 99 percent confidence interval:
##  0.3211259 1.0000000
## sample estimates:
##         p
## 0.3402735
```

Si ho fem manualment fent servir el R pels càlculs però fent servir les fórmules de les notes d'estudi obtenim uns resultats que no són exactament els mateixos, degut principalment a l'estimació de la proporció poblacional.També obtenim un p-valor quasi nul i arribem a la mateixa conclusió.

```
pm <- totalCityIngHigh/totalCity
pm  # proporció mostral
```

```
## [1] 0.3402735
```

```
sp <- sqrt(0.30*(1-0.30)/totalCity)
sp  # error estàndard
```

```
## [1] 0.008078238
```

```
z <- (pm-0.30)/sp
z   # estadístic de contrast
```

```
## [1] 4.985427
```

```
pv <- pnorm(z, lower.tail = FALSE)
pv # p-valor
```

```
## [1] 3.091261e-07
```

Pregunta-4 (25%)

Contrasteu amb un nivell de significació del 10% si la proporció de ciutats corresponents a països d'ingressos alts és diferent als països de l'Àsia de l'Est i Pacífic que als d'Europa i Àsia Central. Indiqueu les hipòtesis nul · la i alternativa. A partir de la sortida de R indiqueu el p-valor i la conclusió a la que arribeu. Interpreteu l'interval de confiança que os proporciona R sobre la diferència de proporcions. Suposeu que les observacions corresponen a una mostra. Atenció: feu servir la funció de R que toqui, és a dir, no feu els càlculs manualment amb les fòrmules de les notes d'estudi.

```
totalCityEAP <-
    length(dadesPM10$City[dadesPM10$Region == "East Asia & Pacific"])
totalCityEAC <-
    length(dadesPM10$City[dadesPM10$Region == "Europe & Central Asia"])
totalCityEAPHI <-
    length(dadesPM10$City[dadesPM10$Region == "East Asia & Pacific" &
                              dadesPM10$IncomeGroup == "High income"])
totalCityEACHI <-
    length(dadesPM10$City[dadesPM10$Region == "Europe & Central Asia" &
                              dadesPM10$IncomeGroup == "High income"])
totalCityEAP
```

```
## [1] 839
```

```
totalCityEAC
```

```
## [1] 871
```

```
totalCityEAPHI
```

```
## [1] 284
```

```
totalCityEACHI
```

```
## [1] 487
```

```
prop.test(c(totalCityEAPHI, totalCityEACHI), c(totalCityEAP, totalCityEAC),
          correct = TRUE, conf.level = 0.90)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(totalCityEAPHI, totalCityEACHI) out of c(totalCityEAP, totalCityEAC)
## X-squared = 83.131, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.2603709 -0.1808876
```

```
## sample estimates:
##    prop 1    prop 2
## 0.3384982 0.5591274
```

Es tracta d'un contrast d'hipòtesis sobre la diferència de proporcions.

Hipòtesi nul · la: $H_0 : p_{EAP} = p_{EAC}$, hipòtesi alternativa: $H_1 : p_{EAP} \neq p_{EAC}$

El p-valor és inferior a $2.2 \cdot 10^{-16}$. Donat que el p-valor és pràcticament nul rebutgem la hipòtesi nul · la i concloem que la proporció de ciutats corresponents a països d'ingressos alts és diferent als països d'Àsia de l'Est i Pacífic que als d'Europa i Àsia Central. L'interval de confiança sobre la diferència de proporcions $(-0.2603709, -0.1808876)$ no conté el zero, per tant, les proporcions són diferents. El fet que l'interval sigui negatiu ens informa de que la proporció de països amb ingressos alts és inferior a la zona d'Àsia de l'Est i Pacífic.

Observació:

Al igual que en la pregunta 3, en la instrucció prop.test es pot incloure l'argument "correct=FALSE". En l'ajuda de "prop.test", es comenta que amb aquesta opció no es té en compte la correcció de continuïtat de Yates. Tal i com es diu en les notes d'estudi, podem considerar com correctes els dos resultats, que són molt semblants:

```
prop.test(c(totalCityEAPHI, totalCityEACHI), c(totalCityEAP, totalCityEAC),
          correct = FALSE, conf.level = 0.90)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(totalCityEAPHI, totalCityEACHI) out of c(totalCityEAP, totalCityEAC)
## X-squared = 84.02, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.2592009 -0.1820576
## sample estimates:
##    prop 1    prop 2
## 0.3384982 0.5591274
```