PEC5 Primavera 2023 - Solución

UOC

Las PECs se basarán en una base de datos obtenida a partir del repositorio de microdatos del "Banco Mundial" en https://microdata.worldbank.org/index.php/catalog/424/get-microdata

Contiene indicaciones, entre otros de

1. *City* = Nombre de la ciudad
2. *Country* = País
3. *Population2000* = Población de la ciudad en 2000
4. *PM10Concentration1999* = "PM10 concentrations (micro gramos por cubic meter) in residential areas of cities larger than 100,000", en 1999
5. *Region* = Clasificación en región geográfica
6. *IncomeGroup* = Clasificación según nivel de ingresos del país

Para importar los datos podemos usar la siguiente instrucción:

```
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)))
```

Os puede ser útil consultar el siguiente material:

1. Módulos Contraste de hipótesis y Contraste de dos muestras
2. Actividades resueltas del Reto 4

NOMBRE:

PEC5

Pregunta-1 (25%)

Contrastad con un nivel de significación del 5% si la concentración media de PM10 del año 1999 en las ciudades de Estados Unidos es inferior a 25. Indicad las hipótesis nula y

alternativa. A partir de la salida de R indicad el valor del estadístico de contraste, el p-valor y la conclusión a la que llegáis. Suponed que las observaciones corresponden a una muestra y que la variable considerada es normal. Atención: utilizad la función de R que toque, es decir, no hagáis los cálculos manualmente con las fórmulas de las notas de estudio.

```
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
```

Se trata de un contraste sobre la media de la concentración de PM10 del año 1999 en las ciudades de Estados Unidos.

Hipótesis nula: $H_0 : \mu = 25$, hipótesis alternativa: $H_1 : \mu < 25$

Estadístico de contraste: $t = -2.0834$ que, bajo hipótesis nula cierta, corresponde a una observación de una distribución t de Student con 207 grados de libertad.

El p-valor=0.01922. Dado que $0.01922 < 0.05$ rechazamos la hipótesis nula y concluimos que la media de la concentración de PM10 del año 1999 es inferior a 25.

Pregunta-2 (25%)

Contrastad con un nivel de significación del 5% si la concentración media de PM10 del año 1999 es más baja en las ciudades de Estados Unidos que en las de China. Indicad las hipótesis nula y alternativa. A partir de la salida de R indicad el valor del estadístico de contraste, el p-valor y la conclusión a la que llegáis. Suponed que las observaciones corresponden a muestras y que las variables consideradas son normales y con varianzas iguales. Atención: utilizad la función de R que toque, es decir, no hagáis los cálculos manualmente con las fórmulas de las notas de estudio.

```
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
```

Se trata de un contraste sobre la diferencia de medias de la concentración de PM10 del año 1999 entre las ciudades de Estados Unidos y las ciudades de China.

Hipótesis nula: $H_0 : \mu_U = \mu_C$, hipótesis alternativa: $H_1 : \mu_U < \mu_C$

Estadístico de contraste: $t = -40.377$ que, bajo hipótesis nula cierta, corresponde a una observación de una distribución t de Student con 586 grados de libertad.

El p-valor es inferior a $2.2 \cdot 10^{-16}$. Dado que el p-valor es prácticamente nulo rechazamos la hipótesis nula y concluimos que la media de la concentración de PM10 del año 1999 en las ciudades de Estados Unidos es inferior a la media de las ciudades de China.

Pregunta-3 (25%)

Contrastad con un nivel de significación del 1% si la proporción de ciudades correspondientes a países de ingresos altos es mayor que el 30%. Indicad las hipótesis nula y alternativa. A partir de la salida de R indicad el p-valor y la conclusión a la que llegáis. Suponed que las observaciones corresponden a una muestra. Atención: utilizad la función de R que toque con la opción "correct=TRUE" y además haced los cálculos manualmente con las fórmulas de las notas de estudio.

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

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

```
totalCityIngHigh    # total ciudades ingresos altos
```

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

```r
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
```

Se trata de un contraste sobre la proporción.

Hipótesis nula: $H_0 : p = 0.30$, hipótesis alternativa: $H_1 : p > 0.30$

El p-valor=$3.414e-07$. Dado que $3.414e-07 < 0.01$ rechazamos la hipótesis nula y concluimos que la proporción de ciudades correspondientes a países de ingresos altos es mayor que el 30%.

Si lo hacemos manualmente utilizando R para los cálculos pero usando las fórmulas de las notas de estudio obtenemos unos resultados que no son exactamente los mismos, debido principalmente a la estimación de la proporción poblacional. También obtenemos un p-valor casi nulo y llegamos a la misma conclusión.

```r
pm <- totalCityIngHigh/totalCity
pm  # proporción muestral
```

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

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

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

```r
z <- (pm-0.30)/sp
z   # estadístico de contraste
```

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

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

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

Pregunta-4 (25%)

Contrastad con un nivel de significación del 10% si la proporción de ciudades correspondientes a países de ingresos altos es diferente en los países de Asia del Este y Pacífico que en los de Europa y Asia Central. Indicad las hipótesis nula y alternativa. A partir de la salida de R indicad el p-valor y la conclusión a la que llegáis. Interpretad el intervalo de confianza que os proporciona R sobre la diferencia de proporciones. Suponed que las observaciones corresponden a una muestra. Atención: utilizad la función de R que toque con la opción "correct=TRUE", es decir, no hagáis los cálculos manualmente con las fórmulas de las notas de estudio.

```r
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
```

```r
totalCityEAC
```

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

```r
totalCityEAPHI
```

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

```r
totalCityEACHI
```

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

```r
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
```

Se trata de un contraste de hipótesis sobre la diferencia de proporciones.

Hipótesis nula: $H_0 : p_{EAP} = p_{EAC}$, hipótesis alternativa: $H_1 : p_{EAP} \neq p_{EAC}$

El p-valor es inferior a $2.2 \cdot 10^{-16}$. Dado que el p-valor es prácticamente nulo rechazamos la hipótesis nula y concluimos que la proporción de ciudades correspondientes a países de ingresos altos es diferente en los países de Asia del Este y Pacífico que en los de Europa y Asia Central. El intervalo de confianza sobre la diferencia de proporciones $(-0.2603709, -0.1808876)$ no contiene el cero, por tanto, las proporciones son diferentes. El hecho que el intervalo sea negativo nos informa de que la proporción de países con ingresos altos es inferior en la zona de Asia del Este y Pacífico.