Práctica 5. t-test y test ANOVA

Carlos Blom-Dahl Ramos

Introducción

Durante esta práctica realizaremos un estudio estad?stico sobre el *Análisis de los efectos del medio ambiente sobre la salud y el desarrollo del feto en embarazos gemelares.*

Se sabe que la exposición materna a tóxicos medioambientales puede repercutir en la salud y el desarrollo del feto. Numerosos estudios en animales y humanos apoyan la hipótesis [1, 2] de que las consecuencias adversas de estas exposiciones en la salud (neuroconductual, inmunológica, cardiovascular...) pueden alcanzar incluso la madurez. Además, el crecimiento fetal es un marcador importante de salud y mortalidad perinatal.

Por otra parte, los recién nacidos resultantes de embarazos gemelares monocigóticos suponen una oportunidad ?nica para explorar, de manera eficiente, el papel relativo de factores genéticos en la asociación entre exposición medioambiental y la salud fetal [3].

En este proyecto se pretende estudiar los efectos adversos de la exposición a contaminantes ambientales durante el embarazo sobre la salud y el desarrollo del feto y del recién nacido en embarazos gemelares monocigóticos (mismo sexo) y también el posible papel protector de la dieta en estas asociaciones

La base de datos del estudio lleva por nombre *datwin19.xls* y consta de 108 registros ficticios, cada uno representando a un embarazo gemelar. El estudio respecto a diseño, tóxicos examinados, estilos de vida, dieta y variables sociodemográficas ha sido inspirado por el estudio real INMA (Infancia y Medio Ambiente [4]). Para más información sobre este proyecto se puede visitar la página web (http://www.proyectoinma.org/). Los datos sobre antropometría al nacimiento fueron simulados de acuerdo a las distribuciones empíricas descritas en más de 10 tablas de antropometría gemelar publicadas en artículos científicos en el periodo 1994-2016.

Las variables recogidas en la base se detallan en el archivo Excel anexo que lleva por título *codebooktwins.xls*. Bibliograf?a 1. Barker DJ, Bull AR, Osmond C, et al. Fetal and placental size and risk of hypertension in adult life. BMJ 1990;301:259-62. 2. Gluckman PD. Epigenetics and metabolism in 2011: epigenetics, the life-course and metabolic disease. Nat Rev Endocrinol 2011;8:74-6. 3. Tong C, et al. Protocol for a longitudinal twin birth cohort study to unravel the complex interplay between early-life environmental and genetic risk factors in health and disease: the Chongqing Longitudinal Twin Study (LoTiS). BMJ Open.2018;8(2):e017889. 4. Guxens M, et al; INMA Project. Cohort Profile: the INMA–INfancia y Medio Ambiente–(Environment and Childhood) Project. Int J Epidemiol. 2012;41(4):930-40.

Vamos a cargar los datos:

```
library(readxl)
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

```
datwins19 <- read_excel("./Datos-20221102/datwins19.xlsx")
```

Análisis estadístico de una muestra

Los test $t-$Student son los tests más famosos en estadística. Se utilizan en contrastes de hipótesis asociados a la media poblacional de una muestra, de dos muestras emparejadas y de dos muestras independientes procedentes de poblaciones estadísticas normales con desviaciones típicas desconocidas. Recordemos que el comando utilizado en R es el *t.test*, podéis encontrar su documentación en https://www.rdocumentation.org /packages/stats/versions/3.6.1/topics/t.test

Actualmente, los embarazos gemelares han aumentado debido a los tratamientos de reproducción asistida y a una mayor edad de las madres primíparas, debido a que cada vez se tienen hijos más tarde. En general, la incidencia de este tipo de embarazos sobre la población es 1 de cada 80 embarazos. Siendo un 25% de los embarazos gemelares, monocigóticos. El desarrollo de los embarazos gemelares monocigóticos ocurre cuando un solo óvulo es fecundado por un espermatozoide dando lugar a un embrión. Posteriormente, este embrión se divide en dos durante las primeras 2 semanas después de la concepción. Los gemelos monocigóticos son también llamados gemelos idénticos ya que ambos son del mismo sexo y son semejantes entre sí física y psíquicamente. Además, tanto su serología (su sangre tiene el mismo ADN y componentes) como sus deformidades son también idénticas. Como parte del estudio "Análisis de los efectos del medio ambiente sobre la salud y el desarrollo del feto en embarazos gemelares" se pretende estudiar la relación peso-sexo de los gemelos monocigóticos, as? como las diferencias de peso existentes en los bebés prematuros según su sexo.

Se habla de bebés prematuros de acuerdo con el criterio de su edad gestacional. Si nacen entre la semana 25 y la 28 se habla de bebés extremadamente prematuros, entre la semana 28 y la 32 de grandes prematuros, y entre la semana 33 y la 36, de prematuros moderados o tardíos. Aunque la duración media de un embarazo único es de 40 semanas, la duración media para un embarazo de gemelos es de 37 semanas.

Los datos necesarios para realizar el estudio se encuentran en el fichero *datwins19.xls*, concretamente, las variables de interés serán:

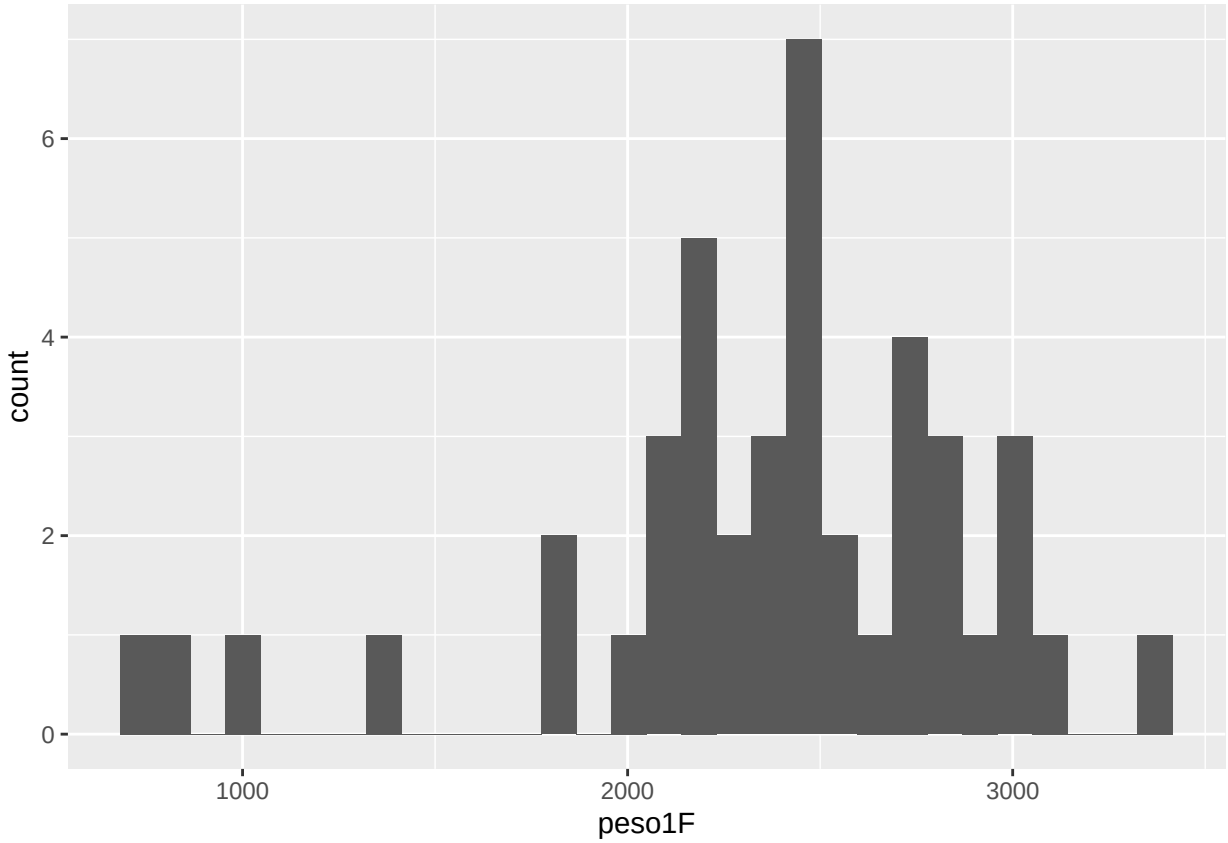
- sexo
- peso1 y peso2: peso al nacer del gemelo con peso mayor y menor, respectivamente
- sges: semanas de gestación

Ejercicio 1 Calcula el intervalo de confianza para el peso medio de las niñas de mayor peso. ¿Podrías considerar que hay evidencia de que, en el caso de las niñas, el peso del gemelo con mayor peso (variable peso1) no está por debajo de los 2100 gramos ni por encima de los 2700 gramos? Interpreta el intervalo de confianza y calcula la normalidad de la variable peso1.

```
#Como el tamaño de la muestra es mayor que 30, podemos construir un intervalo de confianza para saber e
t.test(peso1F,alernative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  peso1F
## t = 27.421, df = 42, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2177.867 2523.900
## sample estimates:
## mean of x
##  2350.884
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Respuesta: pese a que hemos obtenido un p-valor pequeño en la prueba de Shapiro-Wilk, es bien sabido que esta prueba es muy sensible a grandes desviaciones con respecto a la media. Además, el test t que hemos empleado al comienzo para calcular el intervalo de confianza no es muy sensible a la normalidad de los datos si la muestra es relativamente grande, en este caso, mayor que 30. Como nuestra muestra es de tamaño 43 para el caso de estudio, consideramos los resultados obtenidos válidos.

Por lo tanto, al 95% de confianza podemos afirmar que la media del peso del gemelo con mayor peso, en el caso de las niñas, no está por debajo de los 2100 gramos ni por encima de los 2700 gramos.

Ejercicio 2 En base a los datos del estudio, ¿se puede afirmar que, en embarazos prematuros de niñas, el peso al nacer del gemelo con mayor peso no alcanza los dos kilos con un nivel de significatividad del 5%?

```
#Como es en niñas prematuras, hay que calcular el subconjunto de niñas y prematuras
peso1FP<-subset(datwins19$peso1, datwins19$sexo=="girl" & datwins19$prematuro=="Si")
length(peso1FP)
```

```
## [1] 20
```

```
summary(peso1FP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1350    2170    2336    2406    2683    3362
```

```
#como el tamaño de la muestra es menor que $30$ no podemos suponer normalidad y aplicar el TCL y tendre
shapiro.test(peso1FP)
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  peso1FP
## W = 0.98031, p-value = 0.938
```

```
#como el p_valor es mayor que 0.05 no rechazamos $H_0$ y por tanto $X\sim N(\mu,\sigma$
t.test(peso1FP, alternative = "less", mu=2000, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  peso1FP
## t = 4.0652, df = 19, p-value = 0.9997
## alternative hypothesis: true mean is less than 2000
## 95 percent confidence interval:
##      -Inf 2578.265
## sample estimates:
## mean of x
##    2405.7
```

Respuesta: como el p-valor del t.test es 0.9996699, no existe evidencia estadística suficiente para rechazar H_0 *y asumir como cierta la hipótesis alternativa, es decir, no podemos asumir,en embarazos prematuros de niñas, que el peso al nacer del gemelo con mayor peso no alcanza los dos kilos con un nivel de significatividad del* 5%.

Análisis estadístico de muestras independientes

El análisis estadístico de k muestras independientes, desde el punto de vista de la Inferencia Estadística, comporta la obtención de intervalos de confianza y/o la resolución de contrastes de hipótesis, referentes a la diferencia de localización de las poblaciones de las que provienen las k muestras analizadas (k???2).

El objetivo del estudio es: Comparar el valor de una misma variable en k grupos (poblaciones) diferentes: Experimentos con k muestras independientes La creación del archivo adecuado puede depender del tipo de experimento: - Si las muestras son emparejadas cada pareja se considera un caso (fila) por lo que crearemos k columnas de tipo numérico, en la primera introduciremos el dato del primer elemento de la pareja, en la segunda el del segundo elemento y así hasta el k-ésimo elemento. - Si las muestras son independientes cada observación es un caso (fila) y tenemos que señalar a qué grupo, o a cuál de las k muestras, pertenece cada medida observada. Tendremos, por tanto, dos columnas: una numérica con los valores observados de la variable de interés y otra columna, preferiblemente de tipo carácter, en la que indicaremos el grupo al que pertenece cada observación.

Ejercicio 3 ¿Existe evidencia de que la media de peso del gemelo con mayor peso es superior a la del gemelo con menor peso?

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
# Como tenemos 66 muestras, supondremos la normalidad de los datos y pasaremos a
# aplicar el t.test. Primero corrobaremos que las muestras están emperejadas con
# el test Chi-cuadrado

chisq.test(datwins19$peso1, datwins19$peso2)
```

```
## Warning in chisq.test(datwins19$peso1, datwins19$peso2): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  datwins19$peso1 and datwins19$peso2
```

```
## X-squared = 8424, df = 8100, p-value = 0.005907
```

```
# Al obtener un p-valor inferior a 0.05, rechazamos la hipótesis nula del test de
# que las muestras sean independientes

# Comprobaremos con el test de Levene si podemos asumir la igualdad de varianzas
juntos<-gather(datwins19, key="Tipo",value="Peso", peso1, peso2)
juntos$Tipo<-as.factor(juntos$Tipo)
leveneTest(juntos$Peso, group=juntos$Tipo)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.0104 0.9191
##       206
```

```
# Como el p-valor obtenido es alto, no rechazamos la hipótesis nula es
# decir, asumimos la igualdad de las varianzas.

t.test(datwins19$peso1, datwins19$peso2, paired=TRUE, var.equal=TRUE, alternative="greater")
```

```
##
##  Paired t-test
##
## data:  datwins19$peso1 and datwins19$peso2
## t = 85.054, df = 103, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  83.41709      Inf
## sample estimates:
## mean of the differences
##                85.07733
```

Respuesta: hemos obtenido un p-valor de $1.5237995 \times 10^{-97}$*, lo cual nos permite rechazar la hipótesis nula y aceptar la hipótesis alternativa, que en este caso era que la media de peso del gemelo con mayor peso es superior a la del gemelo con menor peso.*

Ejercicio 4 ¿Afecta el sexo al peso del gemelo con mayor peso?

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Eliminamos las observaciones que no contienen información sobre el sexo
juntos$sexo<-as.factor(juntos$sexo)
masc<-juntos%>%drop_na(sexo)%>%filter(Tipo=="peso1" & sexo=="boy")
```

```r
fem<-juntos%>%drop_na(sexo)%>%filter(Tipo=="peso1" & sexo=="girl")

# Supondremos que las muestras ahora son independientes, ya que se
# trata de bebés diferentes nacidos en partos distintos gestados por
# madres diferentes.

juntos_sexo<-juntos%>%filter(Tipo=="peso1")%>%select(Peso, sexo)
leveneTest(juntos_sexo$Peso, group=juntos_sexo$sexo)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.9859 0.1618
##       101
```

```r
# Como el p-valor obtenido es alto, no rechazamos la hipótesis nula es
# decir, asumimos la igualdad de las varianzas.

t.test(masc$Peso, fem$Peso, paired=FALSE, var.equal=TRUE, alternative="two.sided")
```

```
##
##  Two Sample t-test
##
## data:  masc$Peso and fem$Peso
## t = 0.62328, df = 101, p-value = 0.5345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -167.8746  321.6972
## sample estimates:
## mean of x mean of y
##  2427.795  2350.884
```

Respuesta: hemos obtenido un p-valor de t.test(mascPeso, femPeso, paired=FALSE, var.equal=TRUE, alternative="two.sided")$p.value, por lo que no podemos rechazar la hipóitesis nula, es decir, no tenemos evidencia estadística suficiente como para afirmar que existan diferencias en el peso del bebé con mayor peso en función del sexo.

Ejercicio 5 ¿Podemos afirmar que existe una diferencia significativa en el peso de la gemela con mayor peso según sean muy prematuras, prematuras o no prematuras?

Para poder dar respuesta a este objetivo necesitaremos disponer de una variable categórica `CatSGES` que indique el tipo de embarazo y trabajar únicamente con los embarazos de gemelas.

```r
#Si queremos poner las etiquetas a la variable sges, hay que convertir la variable original a un factor
datwins19$CatSGES[datwins19$sges<=33] <-1
```

```
## Warning: Unknown or uninitialised column: `CatSGES`.
```

```r
datwins19$CatSGES[datwins19$sges>33 & datwins19$sges<=37] <-2
datwins19$CatSGES[datwins19$sges>37] <- 3
datwins19$CatSGES<-factor(datwins19$CatSGES, labels = c("muy prematuro", "prematuro", "no prematuro"))
table(datwins19$CatSGES)
```

```
##
## muy prematuro     prematuro  no prematuro
##            10            49            45
```

```r
#Normalidad:
tapply(datwins19$peso1,datwins19$CatSGES,shapiro.test)
```

```
## $`muy prematuro`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.94008, p-value = 0.5539
##
##
## $prematuro
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.9849, p-value = 0.7772
##
##
## $`no prematuro`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.96674, p-value = 0.2197
```

```
#igualdad varianzas
#install.packages("car")
library("car")
leveneTest(datwins19$peso1,datwins19$CatSGES)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2   0.156 0.8558
##       101
```

```
test<-aov(peso1~CatSGES, data=datwins19)
summary(test)
```

```
##              Df   Sum Sq  Mean Sq F value Pr(>F)
## CatSGES       2 22871247 11435624   72.03 <2e-16 ***
## Residuals   101 16034682   158759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Respuesta: como hemos obtenido un p-valor inferior a 0.05, podemos determinar que tenemos evidencia estadística suficiente para afiermar que las medias de los tres grupos NO son iguales.