PEC4: Constrastes de Hipótesis.

MANUEL ROJAS GARCÍA

```
# Cargamos la base de datos como indica el enunciado.

library(dplyr)
library(ggplot2)
load("C:/Users/Manuel/Desktop/UOC/SEMESTRE 3 (Sep 2023 - Feb
2024)/Estadistica/PEC 4/BBDD_COVID2.RData")
head(dat)

##         SEX PATIENT_TYPE AGE OBESITY HIPERTENSION DIABETES PNEUMONIA
ICU
## 1 Masculino          Yes  45      No           No       No       Yes
Yes
## 2 Masculino          Yes  29     Yes          Yes      Yes       Yes
No
## 3  Femenino          Yes  36      No           No      Yes        No
No
## 4  Femenino          Yes  36      No           No       No       Yes
No
## 5 Masculino          Yes  61      No          Yes      Yes        No
No
## 6 Masculino          Yes  44     Yes          Yes       No        No
No
##   INTUBED DIED
## 1     Yes   No
## 2      No  Yes
## 3     Yes  Yes
## 4      No   No
## 5      No   No
## 6      No   No
```

El reciente problema de salud pública generado por la COVID-19 ha retado a la comunidad cientifica a identificar los factores de riesgo para el desarrollo de COVID severo. La base de datos que se os presenta contiene una gran cantidad de información anonimizada relacionada con el paciente hospitalizado por COVID-19, incluidas características sociodemográficas y clínicas previas a la infección por SARS-CoV-2. Los datos públicos proceden del siguiente enlace de Kaggle. En esta práctica se trabaja con una selección aleatoria de pacientes hospitalizados (n = 3000) y variables de interés (p = 10).

Las variables que se encuentran en el dataset son las siguientes:

• *SEX* : Sexo del paciente (Femenino/Masculino).

• *PATIENT_TYPE* : Hospitalización del paciente (Yes/No).

• *AGE* : Edad del paciente.

• *OBESITY* : Paciente con diagnóstico de obesidad (Yes/No).

• *HIPERTENSION* : Paciente con diagnóstico de hipertensión (Yes/No).

• *DIABETES* : Paciente con diagnóstico de diabetes (Yes/No).

• *PNEUMONIA* : Paciente con inflamación pulmonar (Yes/No).

• *ICU* : Ingreso en unidad de cuidados intesivos (Yes/No).

• *INTUBED* : Requerimiento de intubación (Yes/No).

• *DIED* : fallecimiento (Yes/No).

Os puede ser útil consultar el siguiente material:

- Apuntes de contraste de hipótesis
- Apuntes de contraste de dos muestras

Hay que entregar la práctica en fichero pdf o html (exportando el resultado final a pdf o html por ejemplo). Se recomienda generar el informe con Rmarkdown que genera automáticamente el pdf/html a entregar. Se puede utilizar el fichero .Rmd, que disponéis en la PEC, como plantilla para resolver los ejercicios.

Problema 1

a) Según estudios previos, la edad media de los pacientes hospitalizados por COVID-19 es 53 años. Realiza el contraste de hipótesis sobre si la media de edad de nuestra cohorte es diferente a la media de edad teórica. ¿Podemos rechazar la hipótesis nula de igualdad a un nivel de significación del 0.05?

Función pt - https://www.statology.org/working-with-the-student-t-distribution-in-r-dt-qt-pt-rt/

```
# Creamos la variable con las edades
edad <- dat$AGE
# Media teórica de la hipótesis nula
mu_teorica <- 53
# Desviación estándar
sd_Edad <- sd(edad)
# Tamaño muestra
n <- length(edad)
# Valor t
Valort <- (mean(edad) - mu_teorica) / (sd_Edad / sqrt(n))
# P-valor
Pvalor <- 2 * pt(-abs(Valort), df = n - 1)

cat("El valor t es:", Valort, "El valor P:", Pvalor)
```

```
## El valor t es: -0.1326328 El valor P: 0.8944927

# Realizar el contraste de hipótesis de forma automática
resultado_test <- t.test(dat$AGE, mu = 53)
resultado_test$p.value

## [1] 0.8944927

resultado_test

##
##  One Sample t-test
##
## data:  dat$AGE
## t = -0.13263, df = 2999, p-value = 0.8945
## alternative hypothesis: true mean is not equal to 53
## 95 percent confidence interval:
##  52.26344 53.64322
## sample estimates:
## mean of x
##  52.95333
```

Los valores son los mismos, el resultado es correcto y debemos aceptar la hipotesis nula ya que Pvalor (0.8945) es mayor al nivel de significación (0.05). Además 52,95 está dentro del rango (52.26-53.64)

b) Otros estudios estiman que la prevalencia de diabetes en pacientes hospitalizados por COVID-19 es del 30%. ¿Hay evidencia suficiente para considerar que la proporción de diabéticos en la cohorte es diferente a la teórica? Realiza el contraste y razona la respuesta.

```
# Sumatorio de diabeticos
summary(dat)

##      SEX            PATIENT_TYPE            AGE         OBESITY
HIPERTENSION
##  Length:3000        Length:3000        Min.   :  0.00   Yes: 583
Yes:1027
##  Class :character   Class :character   1st Qu.: 42.00   No :2417   No
:1973
##  Mode  :character   Mode  :character   Median : 55.00
##                                        Mean   : 52.95
##                                        3rd Qu.: 66.00
##                                        Max.   :102.00
##  DIABETES   PNEUMONIA   ICU       INTUBED        DIED
##  Yes: 910   Yes:1763   Yes: 269   Yes: 522   Length:3000
##  No :2090   No :1237   No :2731   No :2478   Class :character
##                                             Mode  :character
##
##
##
```

```r
diabeticos <- sum(dat$DIABETES == "Yes")
tamañomuestra <- nrow(dat)
proporcion <- 0.30

phat <- diabeticos / tamañomuestra
phat

## [1] 0.3033333

test_resultado <- prop.test(x = diabeticos, n = tamañomuestra, p =
proporcion, alternative = "two.sided")
test_resultado

##
##  1-sample proportions test with continuity correction
##
## data:  diabeticos out of tamañomuestra, null probability proporcion
## X-squared = 0.14325, df = 1, p-value = 0.7051
## alternative hypothesis: true p is not equal to 0.3
## 95 percent confidence interval:
##  0.2869798 0.3201950
## sample estimates:
##         p
## 0.3033333
```

Como el valor p es 0.7051 damos como buena la hipotesis nula de la proporcion de diabeticos, aún así hemos calculado de manera manual y automática dando un valor de 0.3033333, es decir el 30% que también está dentro del rango [0.28-0.32]

c) Existe evidencia de que los pacientes diabéticos tienen un mayor riesgo de requirimiento de soporte ventilatorio invasivo (INTUBED). ¿Qué contraste de hipótesis plantearías para comparar el riesgo entre grupos?

• Hipótesis nula (H0): Cantidad de pacientes intubados en el grupo de diabéticos es igual a los del grupo de no diabéticos. • Hipótesis alternativa (H1): Cantidad de pacientes intubados en el grupo de diabéticos es diferente de los del grupo de no diabéticos.

```r
# Calculamos los datos, diabeticos, diabeticos e intubados contra no
diabeticos y si entubados
diabeticos <- sum(dat$DIABETES == "Yes")
diabeticosintubado <- sum(dat$DIABETES == "Yes" & dat$INTUBED == "Yes")
nodiabetico <- sum(dat$DIABETES == "No")
nodiabeticonointubado <- sum(dat$DIABETES == "No" & dat$INTUBED == "Yes")

# Proporciones muestrales
pdiabetico <- diabeticosintubado  / diabeticos
pnodiabetico <- nodiabeticonointubado  / nodiabetico

cat("Proporción entubados en diabéticos:", pdiabetico, "Proporción
entubados en no diabéticos:", pnodiabetico)
```

```
## Proporción entubados en diabéticos: 0.2065934 Proporción entubados en
no diabéticos: 0.1598086
```

Hacemos los cálculos automáticamente con una tabla de contingencia con chi al cuadrado

https://r-coder.com/tabla-contingencia-r/

```
# Crear una tabla de contingencia con los datos de yes or no
tabla_contingencia <- table(dat$DIABETES, dat$INTUBED)
# Realizar un test de proporciones chi-cuadrado
chi <- prop.test(tabla_contingencia, conf.level = 0.99)
chi

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tabla_contingencia
## X-squared = 9.3321, df = 1, p-value = 0.002252
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.005730069 0.087839520
## sample estimates:
##    prop 1    prop 2
## 0.2065934 0.1598086
```

Como se comprueba los resultados son iguales y los cálculos son correctos.

d) Presenta la tabla de contigencia entre las variables diabetes y soporte ventilatorio invasivo. Además realiza el contraste de hipótesis planteado en el apartado anterior con un nivel de signifiación del 0.05. (Nota: usa el parámetro correct = FALSE)

Utilizo las variables del ejercicio anterior, indicando el parámetro y la significación indicada en el ejercicio.

```
nivelsignificancia <- 0.05
confianza <- 1 - nivelsignificancia

chi2 <- prop.test(tabla_contingencia, correct = FALSE, conf.level =
confianza)
chi2

##
##  2-sample test for equality of proportions without continuity
correction
##
## data:  tabla_contingencia
## X-squared = 9.6549, df = 1, p-value = 0.001888
## alternative hypothesis: two.sided
## 95 percent confidence interval:
```

```
##  0.01614612 0.07742347
## sample estimates:
##    prop 1    prop 2
## 0.2065934 0.1598086
```

e) De acuerdo con el valor p, ¿Podemos decir que existe aumento estadísticamente significativo en la proporción de pacientes con requerimiento de soporte ventilatorio invasivo en el grupo de pacientes con diabetes? Razona la respuesta. Nota: fijamos el nivel de significación al 0.05.

Seguimos utilizando la tabla de contingencia anterior y obtenemos el valor p.

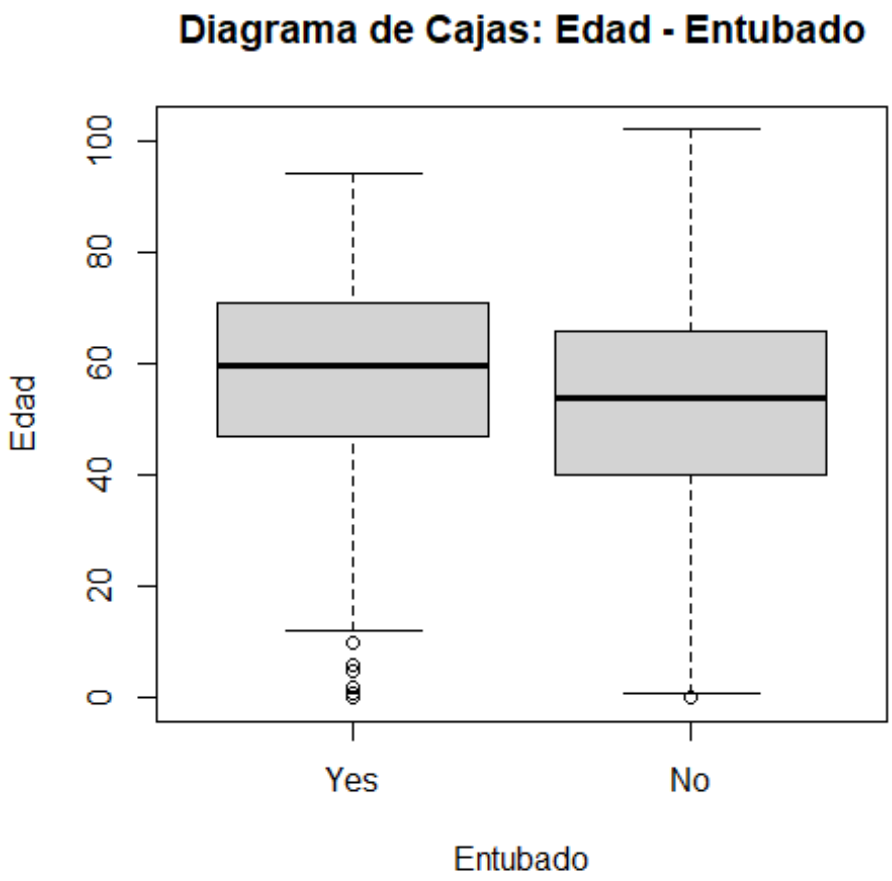
```
valorp <- chi2$p.value
cat("Valor p:", valorp)

## Valor p: 0.001888499
```

Dado que 0.001888499 < 0.05 se rechaza la hipótesis nula. Hay un aumento de pacientes que necesitan ser intubados si padecen diabetes

f) Es conocido que la edad es un factor determinante en el riesgo de intubación. Realiza una comparativa gráfica de la distribución de la edad de los pacientes en función de si requirieron intubación o no. Discute los resultados. (Nota: puedes realizar la comparativa gráfica mediante diagramas de cajas o histogramas).

https://r-coder.com/boxplot-en-r/

```
boxplot(AGE ~ INTUBED, data = dat,
        main = "Diagrama de Cajas: Edad - Entubado",
        xlab = "Entubado", ylab = "Edad")
```



Podemos determinar que cuanta más edad, más posiblidad de tener que ser entubado como se muestra en la gráfica de cajas superior. También podemos determinar la cantidad de valores atipicos que se obtienen de YES, se debería comprobar si los datos de menores a 15 años son correctos.

g) Estima mediante un intervalo de confianza al 95% la diferencia de media de edad entre los pacientes que fueron intubados respecto aquellos que no. Comenta el resultado. Nota: asume varianzas iguales.

```
#Realizamos los calculos de edad y entubados y edad y no intubados y sus
medias

edadintubados <- dat$AGE[dat$INTUBED == "Yes"]
edadnointubados <- dat$AGE[dat$INTUBED == "No"]
totalintubados <- length(edadintubados)
totalnointubados <- length(edadnointubados)
mediaintubados <- mean(edadintubados)
medianointubados <- mean(edadnointubados)


# Varianza (aunque solo usaremos una como indica el ejercicio)
varintubados <- var(edadintubados)
```

```r
varnointubados <- var(edadnointubados)

# Diferencia de medias
diferenciamedias <- mediaintubados - medianointubados

# Error estándar
errorestandar <- sqrt((varintubados / totalintubados) + (varnointubados /
totalnointubados))

# Grados de libertad
gradoslibertad <- ((varintubados / totalintubados + varnointubados /
totalnointubados)^2) /
                  ((varintubados^2 / (totalintubados^2 * (totalintubados
- 1))) +
                   (varnointubados^2 / (totalnointubados^2 *
(totalnointubados - 1))))

# Valor t intervalo de confianza del 95%
valort <- qt(0.975, df = gradoslibertad)

# Límites del intervalo de confianza
limiteinferior <- diferenciamedias - valort * errorestandar
limitesuperior <- diferenciamedias + valort * errorestandar


cat("Diferencia de medias:", diferenciamedias, "Intervalo de confianza al
95%:", limiteinferior, " - ", limitesuperior)
```

```
## Diferencia de medias: 5.31426 Intervalo de confianza al 95%: 3.572658
-  7.055862
```

Realizamos los calculos de manera automatica

```r
testresultado <- t.test(edadintubados, edadnointubados)
testresultado
```

```
##
##  Welch Two Sample t-test
##
## data:  edadintubados and edadnointubados
## t = 5.9897, df = 789.04, p-value = 3.192e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.572658 7.055862
## sample estimates:
## mean of x mean of y
##  57.34291  52.02865
```

```r
diferenciamedias <- mean(edadintubados) - mean(edadnointubados)
```

```
limite_inferior <- testresultado$conf.int[1]
limite_superior <- testresultado$conf.int[2]


cat("Diferencia de medias:", diferenciamedias, "Intervalo de confianza al
95%:", limite_inferior, " - ", limite_superior)
```

```
## Diferencia de medias: 5.31426 Intervalo de confianza al 95%: 3.572658
-  7.055862
```

Sabemos que la diferencia de media es 5.31 y el intervalo de diferencias de las personas intubadas a no intubadas tiene un rando entre 3.572-7.055 aproximadamente al 95%. Como el intervalo no contiene el 0 y que es positivo, podemos concluir que la diferencia es estadisticamente significativa.

h) En clave contraste de hipótesis, ¿podemos rechazar la hipótesis nula de igualdad de medias entre grupos? razona la respuesta.

```
mediaintubados <- mean(edadintubados)
medianointubados <- mean(edadnointubados)
mediaintubados
```

```
## [1] 57.34291
```

```
medianointubados
```

```
## [1] 52.02865
```

```
diferenciamedias <- mediaintubados - medianointubados
diferenciamedias
```

```
## [1] 5.31426
```

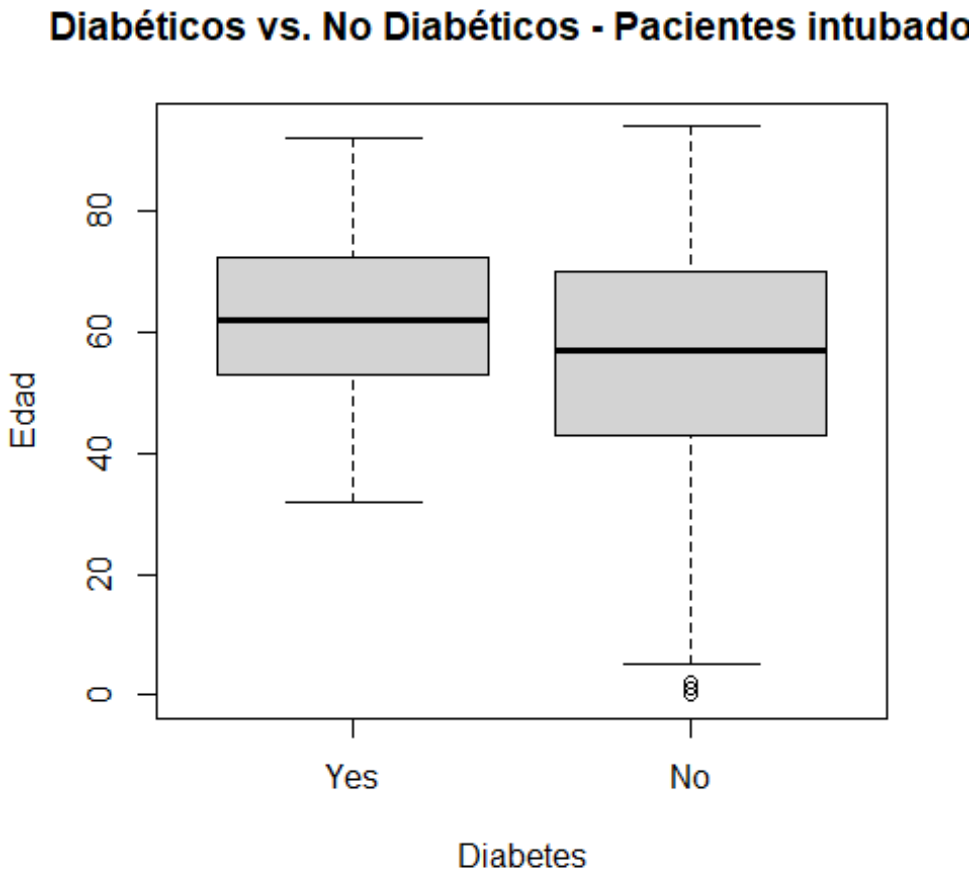
Como el intevalo de confianza no incluye el valor 0 y el la diferencia de medias es positiva (5,31) hay que rechazar la hipótesis nula. Hay una gran diferencia entre edades de pacientes intubados y los que no fueron intubados

A partir de este momento se decide estudiar en profundidad a los pacientes que requirieron soporte ventilatorio invasivo (INTUBED). Para ello, nos centraremos únicamente en esta subploblación.

i) Entre los pacientes que requirieron soporte ventilatorio invasivo, contrasta si la proporción de mortalidad fue diferente entre los pacientes con diabetes y los que no. Comenta los resultados.

```
#Creamos la variable que usaremos a continuacion del subgrupo de
intubados.
subgrupointubados <- dat[dat$INTUBED == "Yes", ]

subdiabeticos <- sum(subgrupointubados$DIABETES == "Yes")
subnodiabeticos <- sum(subgrupointubados$DIABETES == "No")
```

```r
#Creamos un diagrama de cajas para tener un primer analisis de diabeticos
y no diabeticos intubados

boxplot(AGE ~ DIABETES, data = subgrupointubados,
        main = "Diabéticos vs. No Diabéticos - Pacientes intubados",
        xlab = "Diabetes", ylab = "Edad")
```

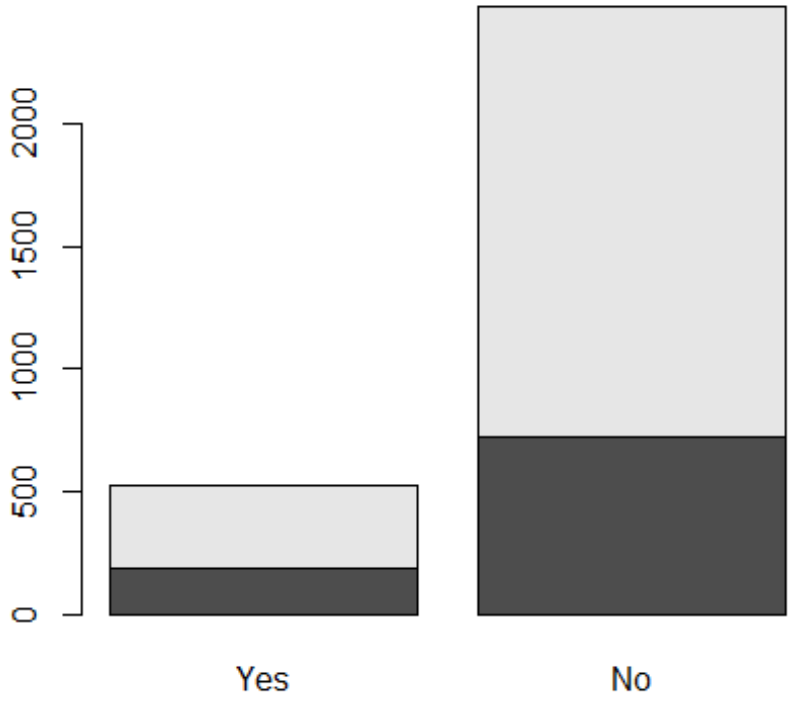


https://rpubs.com/osoramirez/111403

```r
# Creamos la tabla de contingencia para pacientes intubados y diabéticos
tablacontingencia <- table(subgrupointubados$DIABETES,
subgrupointubados$DIED)
chi <- chisq.test(tablacontingencia)
chi

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tablacontingencia
## X-squared = 6.8402, df = 1, p-value = 0.008913
```

El valor p es muy por debajo de 0.05, eso indica que hay que rechazar la hipótesis nula. Un paciente con diabetes dentro de este subgrupo implica un mayor riesgo de muerte en comparación de aquellos pacientes que fueron intubados y no tenian diabetes. Incluyo un barplot a continuación donde la evidencia se hace más visible

```
barplot(tabla_contingencia)
```



j) Tras el alta hopitalaria los pacientes supervivientes sufren secuelas pulmonares graves. Está demostrado que la capacidad de difusión del monóxido de carbono (DLCO) es una medida sensible al daño pulmonar y su funcionalidad. Interesa saber si seis meses después del alta hospitalaria los pacientes han mejorado este parámetro. Para ello, se les hace una medición el día del alta hopitalaria y seis meses después. Plantea el contraste de hipótesis que realizarías para contrastar si los pacientes mejoraron significativamente. (nota: Mayor valor de DLCO mejor estado pulmonar).

Debemos plantear un contrastr de hipoteisis sobre si los pacientes mejoran en la capacidad de difusión de monoxido de carbono seis meses después de su alta. Para ello podemos usar un contraste de medias H0 - No hay diferencia de DLCO entre el día de alta y 6 meses después. μ1 H1 - Hay diferenciaa de DLCO entre el día de alta y 6

meses después. μ2 H0: μ1 - μ2 = 0 H1: μ1 - μ2 < 0 Realizariamos un testresultado <- t.test(media1, media2)

Si el valor p es menor a 0.05 rechazamos la H0.