PEC 3

UOC

Introducción

En un estudio sobre ingresos en un hospital, se recogió información detallada de cada una de las estancias (motivos, características del paciente,etc.)

Los datos se pueden encontrar en formato csv, texto plano separado por comas, "ingresos_hospital_clean.csv".

Las variables de la base de datos son:

- id_ingreso: Identificador único para cada ingreso.
- fecha_ingreso: Fecha en la que el paciente fue ingresado.
- fecha_egreso: Fecha en la que el paciente fue dado de alta.
- edad: Edad del paciente.
- género: Género del paciente.
- diagnóstico: Diagnóstico principal para el ingreso.
- coste: coste total del ingreso.
- pagado: Indica si el ingreso ha sido pagado (Sí o No).
- tipo_tratamiento: Tipo de tratamiento proporcionado (por ejemplo, medicación, cirugía, fisioterapia, etc.).
- seguro: Tipo de seguro (Público, Privado, etc.).
- nivel_urgencia: Nivel de urgencia del caso (Baja, Media, Alta).

Os puede ser útil consultar el siguiente material:

1. Módulo de Intervalos de confianza.
2. Actividades resueltas del Reto 3 (Intervalos de confianza).
3. Procurad utilizar las funciones propias de R para hacer los cálculos a no ser que se diga lo contrario.

El informe final se librará en formato pdf o html (exportando el resultado final en pdf o html por ejemplo). Se recomienda generar el informe con Rmarkdown que genera automáticamente el pdf/html a librar.

Esta PEC se tiene que realizar de forma estrictamente individual, quedando totalmente prohibido el uso de herramientas de IA. Cualquier indicación de copia será sancionada con un suspenso (D) por todas las partes implicadas y la posible evaluación negativa de la asignatura de forma íntegra.

```
##        id        fecha_ingreso      fecha_egreso            edad
##  Min.   :  1.0   Length:351         Length:351         Min.   :18.00
##  1st Qu.: 88.5   Class :character   Class :character   1st Qu.:31.00
##  Median :176.0   Mode  :character   Mode  :character   Median :47.00
##  Mean   :176.0                                         Mean   :47.28
##  3rd Qu.:263.5                                         3rd Qu.:64.00
##  Max.   :351.0                                         Max.   :80.00
##     género          diagnóstico        tipo_tratamiento       coste
##  Length:351         Length:351         Length:351         Min.   :2002
##  Class :character   Class :character   Class :character   1st Qu.:3931
##  Mode  :character   Mode  :character   Mode  :character   Median :5823
##                                                           Mean   :5980
##                                                           3rd Qu.:8200
##                                                           Max.   :9976
##     pagado             seguro          nivel_urgencia
##  Length:351         Length:351         Length:351
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
head(dat)
```

```
##   id fecha_ingreso fecha_egreso edad género diagnóstico tipo_tratamiento coste
## 1  1    2023-07-23   2023-08-06   54      M    COVID-19     Fisioterapia  4461
## 2  2    2023-09-11   2023-09-26   74      M    Neumonía          Cirugía  8685
## 3  3    2023-04-17   2023-04-18   54      M    COVID-19     Fisioterapia  8144
## 4  4    2023-03-05   2023-03-20   48      F    COVID-19          Cirugía  5395
## 5  5    2023-10-13   2023-10-18   29      M  Bronquitis        Monitoreo  6712
## 6  6    2023-07-03   2023-07-18   42      F   Infección          Cirugía  8197
##   pagado  seguro nivel_urgencia
## 1     Sí Público           Baja
## 2     Sí Privado           Alta
## 3     No Privado           Alta
## 4     No Público           Alta
## 5     No Privado          Media
## 6     Sí Público           Alta
```

NOMBRE: Juan Luis Acebal Rico

El objetivo principal de este estudio es evaluar la influencia de la gravedad en el estado de los pagos y evaluar si hay alguna otra característica relevante para los seguros.

Pregunta 1 (30%)

Encontrar un intervalo de confianza para la media del coste por ingreso con un nivel de confianza del 95% para:

a) (10%) Nivel urgencia baja

```r
# Filtro por urgencia baja
coste_baja <- subset(dat, nivel_urgencia == "Baja")$coste

# Intervalo de confianza del 95% para la media del coste con nivel de urgencia baja,
#es decir, podemos decir que estoy seguro al 95% de que, la media verdadera del coste
#de ingreso para nivel de urgencia bajo, estará entre los valores
#del intervalo 5762.464 a 6619.761.
t.test(coste_baja, conf.level = 0.95)$conf.int
```

```
## [1] 5762.464 6619.761
## attr(,"conf.level")
## [1] 0.95
```

b) (10%) Nivel urgencia alta

```r
#filtro
coste_alta <- subset(dat, nivel_urgencia == "Alta")$coste

# Intervalo de confianza del 95% para la media del coste con nivel de urgencia alta,
#es decir, podemos decir que estoy seguro al 95% de que, la media verdadera del coste
#de ingreso para nivel de urgencia bajo, estará entre los valores
#del intervalo 5542.651 a 6458.647.
t.test(coste_alta, conf.level = 0.95)$conf.int
```

```
## [1] 5542.651 6458.647
## attr(,"conf.level")
## [1] 0.95
```

c) (10%) ¿Qué conclusión podemos extraer sobre el coste según el nivel de urgencia? (en particular fijaos en las medias y los intervalos de confianza).

Se superponen los intervalos de confianza entonces es complicado sacar conclusiones ya que no hay diferencias significativas entre los costes de los ingresos en función de la urgencia. Los intervalos de confianza dicen donde esta la media con una seguridad, y para ambos intervalos se solapan, entonces no podria decir conforme al nivel de estudio hecho. Eso quiere decir que no podemos afirmar que el coste de los ingresos sea diferente en función de la urgencia solamente con estos datos.

Pregunta 2 (40%)

Queremos estudiar la proporción de personas ingresadas con una edad inferior a 30 años.

a) (10%) Calculad un intervalo de confianza del 90% para dicha proporción mediante la función prop.test con la opción correct=FALSE.

```r
menor30 <- sum(dat$edad < 30)
total <- nrow(dat)
prop.test(menor30, total, conf.level = 0.90, correct = FALSE)$conf.int
```

```
## [1] 0.1719961 0.2427714
## attr(,"conf.level")
## [1] 0.9
```

Intervalo de confianza que estamos seguros al 90% de personas menores a 30 años ingresadas, es decir, la proporción de personas menores a 30 años ingresadas estará entre 0.1719961 y 0.2427714 y estoy seguro al 90% de lo que digo.

b) (10%) Calculad el mismo intervalo siguiendo las fórmulas de las notas de estudio (puede haber pequeñas diferencias en el resultado). Usad R para hacer las operaciones y para calcular el valor crítico.

```
p_hat <- menor30 / total

z <- qnorm(0.95)
error <- z * sqrt(p_hat * (1 - p_hat) / total)
p_hat
```

```
## [1] 0.2051282
```

```
z
```

```
## [1] 1.644854
```

```
error
```

```
## [1] 0.03545153
```

```
c(p_hat - error, p_hat + error)
```

```
## [1] 0.1696767 0.2405797
```

Aqui entiendo que es 0.95 ya que buscamos el punto de corte para un intervalo de confianza del 90% que se encuentra en 0.95 ya que dejamos 0.05 a la izquerda y 0.05 a la derecha (antes y despues)

c) (20%) Si queremos calcular el intervalo de confianza igual que en el apartado anterior (90%) pero con un margen de error inferior a 0.01, calculad cuál tiene que ser la medida mínima de la muestra si tomamos la muestra trabajada en el apartado anterior como información previa.

```
E <- 0.01

n_min <- (z^2 * p_hat * (1 - p_hat)) / E^2
n_min
```

```
## [1] 4411.406
```

Uso p_hat del apartado anterior y vemos que da 4411.406, es decir, necesitamos una muestra de 4412 personas para tener un margen de error inferior a 0.01. (no podemos usar da 4411.406 personas, es decir personas enteras, sino 4412 personas ya que si usamos 4411 tendriamos mas margen de error, y el enunciado dice inferior a 0.01)

Pregunta 3 (30%)

Los seguros creen que el coste por ingreso es más elevado de lo que suponían inicialmente. Creen que el coste medio, independientemente de la urgencia es superior a 5000. La persona que ha realizado el análisis no ha especificado correctamente el código en R del test, ha realizado lo siguiente:

```r
t.test(dat$coste)
```

```
##
##  One Sample t-test
##
## data:  dat$coste
## t = 46.513, df = 350, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5727.281 6233.009
## sample estimates:
## mean of x
##  5980.145
```

A partir de estos resultados responded las siguientes preguntas:

a) (20%) A partir del intervalo de confianza, podemos concluir que la media es superior a 5000? y con el p.valor? Razona la respuesta.

Ahora bien, el enunciado dice que, "La persona que ha realizado el análisis no ha especificado correctamente el código en R del test", quizas se refiere a que al no dar un valor de referencia, R toma por defecto el valor de 0, por lo que el test se ha realizado para saber si la media no es 0, y no si la media es superior a 5000. Y tenemos un P-value realmente bajo, es decir, con 16 ceros despues del punto decimal, seria aproximadamente p-value 0.00000000000000022. Todo incorrecto creo yo, ya que no se ha realizado con test unilateral sino bilateral, es decir, nosotros queremos saber si es mayor que un valor especifico, es decir nos interesa la direccion. En un test bilateral, no nos interesa la direccion, sino si es diferente de un valor especifico. Por tanto, el test no es correcto y no podemos sacar conclusiones de este test, ya que es evidente que al ser un test bilateral con mu=0, nos va a dar significativamente mayor siempre.

```r
t.test(dat$coste, mu = 5000, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  dat$coste
## t = 7.6235, df = 350, p-value = 1.172e-13
## alternative hypothesis: true mean is greater than 5000
## 95 percent confidence interval:
##  5768.108      Inf
## sample estimates:
## mean of x
##  5980.145
```

Si quiero saber si la media es superior a 5000, y la hipotesis alternativa es que sea mayor a 5000 la media. El intervalo de confianza es 5768.108 e infinito y es un 95 % de confianza (estamos seguros al 95%) de la media esta entre el intervalo y por tanto superior a 5000. Tambien puedo concluir que, dado que todo el intervalo es por encima de 5000, la media es superior a 5000 tambien. Por ultimo, ademas tenemos un P-value realmente bajo, que nos da una probabilidad de si la hipotesis nula fuera cierta (es decir, una media menor a 5000), la probabilidad de tener una media muestral tan o mas alta es del 0.00000000001172%, es decir practicamente cero. Por tanto, podemos rechazar la hipotesis nula y la conclusion es de que la media es superior a 5000 tanto la evidencia del p-value como del intervalo de confianza.

b) (10%) Cuántos pacientes había en la muestra?

df+1, es decir, 351, ya que el df es el número de grados de libertad y en este caso es 350, pero el número de pacientes es 351. df no seria dataframe sino grados de libertad, que seria df=n-1

```
nrow(dat)
```

```
## [1] 351
```