PEC1: Inferencia estadística 2022-23 semestre 2

bas, magí

2023-05-09

Ejercicio 1

Se supone que una enfermedad se hereda vía un gen dominante y que uno de los dos progenitores tiene la enfermedad y el otro no. Esto implica que la probabilidad de que un descendiente herede la enfermedad es de 1/2.

a) ¿Cuál es la probabilidad de que en una familia con dos descendientes ambos estén afectados de la enfermedad?

```
P=1/2

P_dos_descendientes = P*P

print(P_dos_descendientes)
```

```
## [1] 0.25
```

b) ¿Cuál es la probabilidad de que en una familia con dos descendientes sólo 1 este afectado por la enfermedad?

Tenemos que plantear dos escenarios posibles e independientes entre si.

1. El primero hereda el gen
2. El segundo hereda el gen

```
P_solo_uno = (P*(1-P))+((1-P)*P)

print(P_solo_uno)
```

```
## [1] 0.5
```

c) ¿Cuál es la probabilidad de que si el primer descendiente tenga la enfermedad la tenga el segundo?

La misma, 0.5, ya que son eventos independientes. Que el primero tenga la enfermedad no afecta a la probabilidad del segundo.

Imagina que la herencia del gen es recesiva es decir la probabilidad de que un progenitor transfiera la enfermedad es de 1/4

d) ¿Cuál es la probabilidad de que los dos descendientes tengan la enfermedad?

```
p= 1/4

p_dos_descendientes= p*p

print(p_dos_descendientes)
```

```
## [1] 0.0625
```

Imagina que la herencia del gen está relacionada con el sexo, es decir sólo ocurre en los descendientes masculinos. Así en un niño la probabilidad de recibir la herencia es 1/2 mientras que en una niña es 0.

e) En una familia con un niño y una niña, ¿Cuál es la probabilidad de que sólo uno de los dos tenga la enfermedad?

La probabilidad de que solo uno la tenga es 1, ya que solo el niño puede tener la enfermedad.

f) Se dispone de una pareja de niños que tienen la enfermedad y se asume que la probabilidad de que dos hijos estén enfermos es la calculada en los ejercicios anteriores. ¿Cuál de los tres modos de herencia (dominante, recesiva o relacionada con el sexo) es el más probable en esta familia si la distribución de los tipos de herencia es la misma?

```
p_dos_niños= 1/2*1/2
tabla1= data.frame(P_Gen_A = P_dos_descendientes,
                   P_Gen_a = p_dos_descendientes,
                   P_Gen_sex = p_dos_niños)

tabla1
```

```
##   P_Gen_A P_Gen_a P_Gen_sex
## 1    0.25  0.0625      0.25
```

Las dos herencias más probables serian tanto la dominante como la relacionada con el sexo.

Ejercicio 2

Se sabe que el 70% de los ingresos hospitalarios en neurología están relacionados con un Ictus. En el servicio de neurología hay 40 personas enfermas hospitalizadas.

a) ¿Cuál es la probabilidad de que todas ellas estén ingresadas por ictus?

```
P_ictus= .7

P_todas_ictus= .7^40
# P_todas_ictus= 1-P_ninguna_ictus= 0.3^40
print(P_todas_ictus)
```

```
## [1] 6.366806e-07
```

b) ¿Cuál es la probabilidad de que exactamente 10 personas no estén ingresadas por ictus?

No es lo mismo calcular la probabilidad de que 'almenos 10 personas' esten ingresadas por ictus que calcular la probabilidad de que 'exactamente 10 personas' esten ingresadas por ictus.

Para ello debemos emplear la distribución binomial que toma en cuenta la probabilidad de èxito y fracaso de cada ensayo, así como el número de ensayos y el número de éxitos que se desean obtener.

```
dbinom(10, 40, 0.3)
```

```
## [1] 0.1128173
```

11.28% de probabilidades de que exactamente 10 personas no esten ingresadas por ictus.

c) ¿Cuál es la probabilidad de que el número de personas ingresadas por ictus sea estrictamente mayor de 30?

Podemos calcular la probabilidad acumulativa de la siguiente manera:

```r
1-pbinom(30, 40, .7)
```

```
## [1] 0.1959254
```

Para calcular la probabilidad de que el número de personas ingresadas por ictus sea estrictamente mayor de 30, necesitamos calcular la probabilidad de la cola superior, es decir, la probabilidad de que haya más de 30 personas ingresadas por ictus.

Se va a abrir una nueva planta de neurología en un nuevo hospital de 20 camas.

Genera una muestra aleatoria para obtener la ocupación por ictus de la nueva planta un día suponiendo en la población el 70% de casos tienen ictus.

```r
set.seed(34234)
muestra= rbinom(20,1, .7)
muestra
```

```
##  [1] 1 1 0 1 1 0 0 0 0 1 0 0 1 1 0 0 1 1 1 0
```

d) Estima el % de casos de ictus ingresados en la nueva planta a partir de la muestra y calcula su intervalo de confianza al 99%

```r
# rbinom(n, size, prob)|Genera n muestras de tamaño "size" con una probabilidad de éxito "prob"
```

```r
porcentaje_muestra= mean(muestra)*100
cat(porcentaje_muestra, '%')
```

```
## 50 %
```

```r
binom.test(sum(muestra), 20, conf.level = .99)
```

```
##
##  Exact binomial test
##
## data:  sum(muestra) and 20
## number of successes = 10, number of trials = 20, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.5
## 99 percent confidence interval:
##  0.2177475 0.7822525
## sample estimates:
## probability of success
##                    0.5
```

El intervalo de confianza al 99% para el porcentaje de casos de ictus en la nueva planta sería de 21.77% a 78.23%, según los resultados del test de binomio exacto realizado en R. Esto significa que hay un 99% de confianza de que el verdadero porcentaje de casos de ictus en la población cae dentro de este rango.

e) Si en lugar de la muestra de 1 día obtuviéramos 365 muestras, 1 por cada día del año y en cada caso calculamos el porcentaje camas ocupadas con ictus y su intervalo de confianza al 99%. ¿En cuántos de estos 365 intervalos esperas que esté el verdadero valor del 70% de camas ocupadas por ictus?

Esperaría que 361 de las 365 muestras, es decir, el 99% de las muestras tuvieran el verdadero valor del porcentaje de camas ocupadas por ictus en el año

g) Efectúa la simulación e indica cuántos intervalos contienen verdaderamente el 70% de casos de ictus en su interior.

```r
set.seed(361)
muestra_365= rbinom(365, 20, .7)
n_muestras= 365
```

```r
tamaño_size= 20
p= 0.7
conf_level= 0.99
n_contain= 0

for (i in 1:n_muestras) {
  muestra= rbinom(tamaño_size, 1, p)
  conf_int= binom.test(sum(muestra), tamaño_size, conf.level = conf_level)$conf.int
  if (p >= conf_int[1] && p <= conf_int[2]) {
    n_contain = n_contain + 1
  }
}

n_contain
```

```
## [1] 364
```

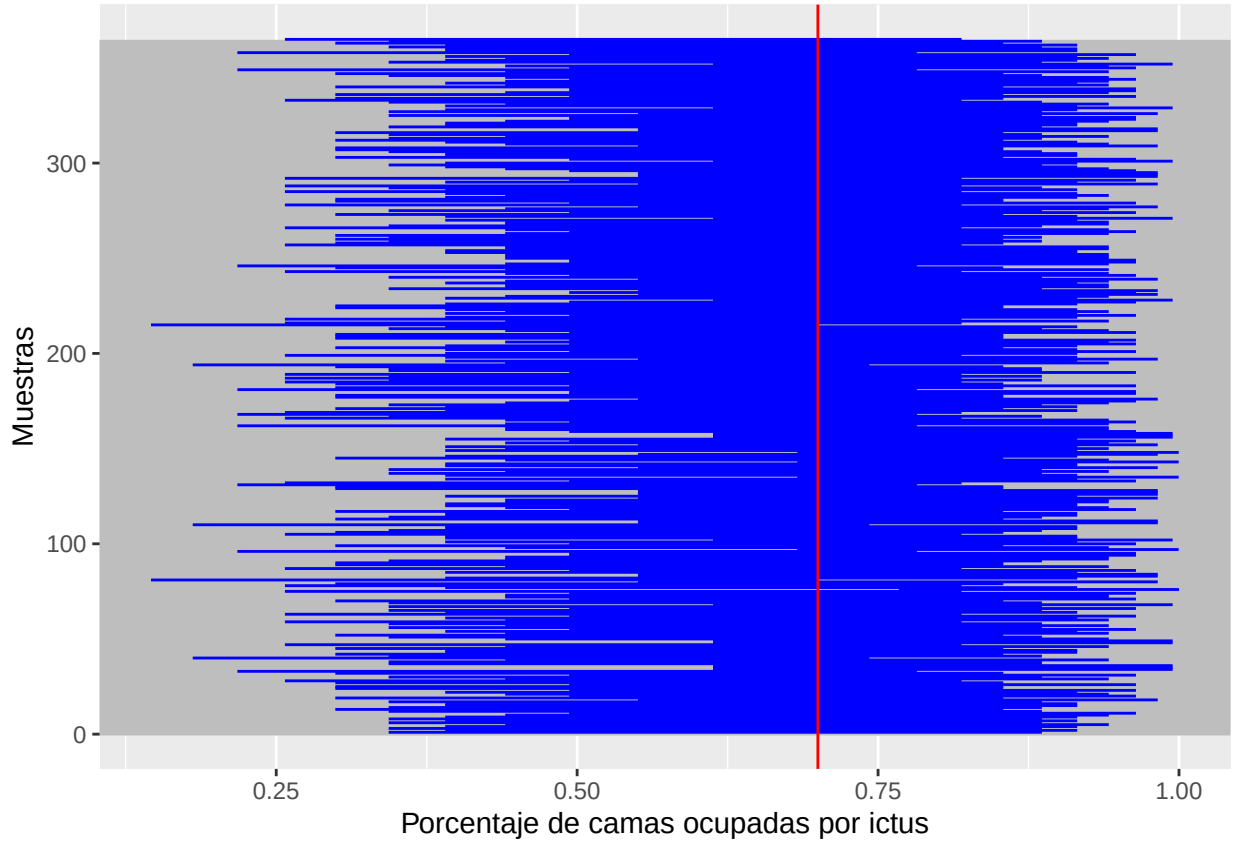
Et esta seed el número de muestras que contienen el verdadero valor es de 364, lo que esta por encima del 99% de confianza establecido. Podemos visualizarlo de la siguiente forma:

```r
library(ggplot2)
set.seed(361)
muestra_365 <- rbinom(365, 20, 0.7)
conf_level <- 0.99
p <- 0.7

df <- data.frame(muestra = muestra_365,
                 lower = rep(0, length(muestra_365)),
                 upper = rep(0, length(muestra_365)))
for (i in 1:length(muestra_365)) {
  conf_int <- binom.test(muestra_365[i], 20, conf.level = conf_level)$conf.int
  df$lower[i] <- conf_int[1]
  df$upper[i] <- conf_int[2]
}

ggplot(df, aes(x = muestra)) +
  geom_hline(yintercept = 0:364, color = "grey") +
  geom_segment(aes(x = lower, xend = upper, y = 1:length(muestra_365), yend = 1:length(muestra_365)),
               size = 0.5, color = "blue") +
  geom_vline(xintercept = p, color = "red") +
  xlab("Porcentaje de camas ocupadas por ictus") +
  ylab("Muestras")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Ejercicio 3

La distribución de algunas variables ecográficas en pacientes con una afección cardíaca o normal sigue una distribución normal. Así por ejemplo la dispersión de la onda P (PWD) sigue una distribución normal(gaussiana) de media 31.8ms con una desviación estándar de 5.30 en los normales y de 44.7ms con una desviación estándar de 4.2 en los afectados cardíacos. Si se considera un punto de corte de 40 para indicar que un sujeto por encima de este valor tiene una afección cardíaca.

Población sana => N~(31.8, 5.30)

Afectados cardíacos => N~(44.7, 4.2)

```
# rnorm(n, mean, std)| Genera 1 muestras de tamaño n de una normal con media=mean y desviación típica=s
```

```
# pnorm(q, mean, sd) | Calcula la probabilidad acumulada hasta q, de media= mean y desviación estándar=
```

a) ¿Qué porcentaje de la población normal serán clasificados como cardíacos con el punto de corte de 40?

```
1- pnorm(40, 31.8, sd=5.3)
```

```
## [1] 0.06091115
```

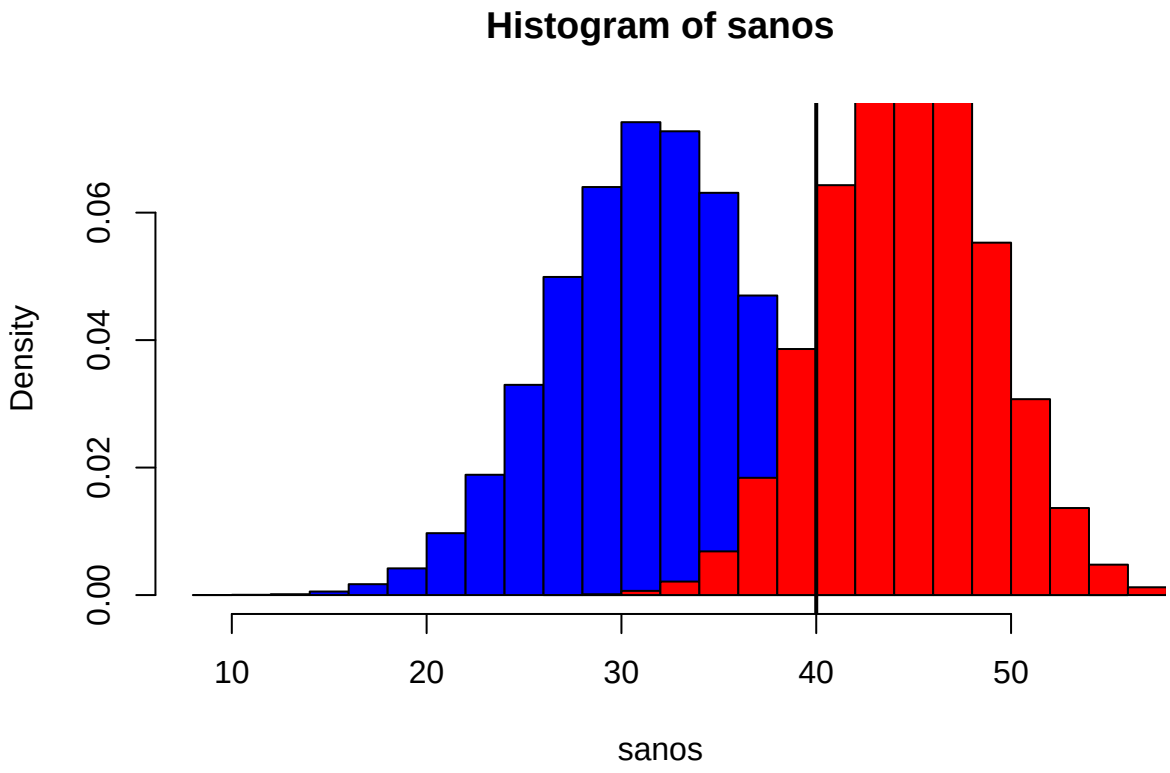
Un 6% de población sana seria clasificada con afección cardíaca

```
# Generar los datos para la distribución normal en pacientes normales
sanos= rnorm(100000, 31.8, 5.3)
```

```r
# Generar los datos para la distribución normal en pacientes con afección cardíaca
cardiacos= rnorm(100000, mean = 44.7, sd = 4.2)

# Graficar las distribuciones normales
hist(sanos, prob = TRUE, col = "blue")
hist(cardiacos, prob = TRUE, col = "red", add = TRUE)

# Dibujar una línea vertical en el punto de corte
abline(v = 40, col = "black", lwd = 2)
```



b) ¿Qué porcentaje de población cardíaca será clasificada como normal con el punto de corte situado en 40?

```r
pnorm(40, 44.7, sd=4.2)
```

```
## [1] 0.1315599
```

Un 13,16% de la población cardíaca seria classificada como sana.

c) Si el 10% de la población que es estudiada tiene una afección cardíaca. ¿Qué porcentaje de la población total tiene el nivel de PWD por encima de 40?

En este caso podemos utilizar el teorema de Bayes.

```r
prop_muestra= .1
sanos= pnorm(40, 31.8, sd=5.3)
cardiacos= pnorm(40, 44.7, 4.2)
proporcion_poblacion= (prop_muestra*cardiacos)+ (0.9*sanos)
proporcion_poblacion
```

```
## [1] 0.858336
```

El 85% de la población total tiene un nivel de PWD por encima de 40

d) Si elegimos a un individuo de toda la población con un nivel de PWD por encima de 40 ¿qué probabilidad tiene realmente de tener una afección cardíaca?

P(cardiaco | PWD > 40) = P(PWD > 40 | cardiaco) * P(cardiaco) / P(PWD > 40)

```
# Probabilidad de tener un nivel de PWD por encima de 40
p_pwg <- 0.1 * (1 - pnorm(40, 44.7, 4.2)) + 0.9 * (1 - pnorm(40, 31.8, 5.3))

# Probabilidad de tener una afección cardíaca dado que se tiene un nivel de PWD por encima de 40
p_c_given_pwg <- 0.1 * (1 - pnorm(40, 44.7, 4.2)) / p_pwg

p_c_given_pwg
```

```
## [1] 0.6130279
```

La probabilidad es de un 61.30%

e) Extrae una muestra de 300 pacientes con afección cardíaca y estima el valor medio del PWD. Calcula el intervalo de confianza al 95% e interpreta los resultados. ¿Qué diferencias encontrarías si el tamaño de la muestra fuera de 30 con afección cardíaca?

```
set.seed(78110)
# rnorm(n, mean, sd)|Genera una muestra de tamaño n con las medias y desviaciones típicas por las que s
muestra_300= rnorm(300, 47, 4.2)
t.test(muestra_300, conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  muestra_300
## t = 200.98, df = 299, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  46.70720 47.63091
## sample estimates:
## mean of x
##  47.16905
```

El intervalo de confianza para esta muestra es de 46.7 - 47.63.

Si el tamaño de la muestra fuera de 30 con afección cardíaca, el intervalo de confianza sería más amplio y, por lo tanto, menos preciso que el obtenido con una muestra de 300 pacientes.