Intervalos de Confianza

Monitoria de Estadística

Febrero de 2020

Profesor: Edwin Santiago Alférez

1. Introducción

Uno de los objetivos principales de la Estadística es el obtener información sobre parámetros desconocidos de una población a partir de una muestra aleatoria. La estimación intervalar es una de las formas más usadas para obtener esta información. Por ejemplo, si tenemos una población normal con media o varianza deconocida, o ambas, quizás no podemos conocer el valor exacto de estos parámetros a partir de un conjunto de datos, pero si podemos dar una idea de cómo serán estos parámetros. Los intervalos de confianza nos proporcionan un rango de valores sobre el cual el parámetro está incluido. Como este intervalo depende de la muestra usada, se garantiza que una significante proporción de los intervalos calculados contienen al valor real del parámetro. En esta sesión se utilizán métodos clásicos para la estimación intervalar de la media de una población. Además, se decribirán las funciones propias de R que sirven para calcular directamente este intervalo.

2. Intervalos de confianza

Para ilustrar el procedimiento para el cálculo de intervalos de confianza de la media poblacional, se utilizará un conjunto de datos de un grupo de estudiantes de estadística en una universidad australiana. Este conjunto de datos, llamado "survey" pertenece al package MASS que está incluido en la instalación básica de R pero debe cargarse con anterioridad de la siguiente manera:

```
library(MASS) # Load the MASS package
```

Se puede visualizar los primeros datos de survey (es un dataframe) de la siguiente forma:

```
head(survey)
```

```
##      Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke Height
## 1 Female   18.5   18.0 Right  R on L    92    Left Some Never 173.00
## 2   Male   19.5   20.5  Left  R on L   104    Left None Regul 177.80
## 3   Male   18.0   13.3 Right  L on R    87 Neither None Occas     NA
## 4   Male   18.8   18.9 Right  R on L    NA Neither None Never 160.00
## 5   Male   20.0   20.0 Right Neither    35   Right Some Never 165.00
## 6 Female   18.0   17.7 Right  L on R    64   Right Some Never 172.72
##        M.I    Age
## 1   Metric 18.250
## 2 Imperial 17.583
## 3     <NA> 16.917
## 4   Metric 20.333
## 5   Metric 23.667
## 6 Imperial 21.000
```

Para obtener más detalles del conjunto de datos del survey, se puede consultar la documentación de R.

2.1 Estimación puntual de la media poblacional

Para cualquier muestra aleatoria podemos calcular siempre su media muestral. A pesar de que normalmente esta media no tiene el mismo valor de la media de la población actual, sirve como una buena estimación puntual. Por ejemplo, a partir del conjunto de datos "survey", que es una muestra de la población estudiantil, podemos calcular su media muestral y usarla como una estimación del correspondiente parámetro de la población (media poblacional). Usando los datos cargados, una estimación puntual de la media de todos los estudiantes está dada por:

```
mean(survey$Height, na.rm = TRUE) # skip missing value
```

```
## [1] 172.3809
```

Como algunos de los estudiantes encuestados no contestaron todas las preguntas, existen algunas variables con parámetros faltantes, por lo tanto, debemos filtrar estos valores anexando la opción `na.rm` con el argumento *TRUE* a la función `mean`.

2.2. Estimaciòn intervalar de la media de una población con varianza conocida

Después de encontrar una estimación puntual de la media poblacional, necesitaremos un modo de cuantificar su precisión. Aquí discutimos el caso en que la varianza poblacional σ^2 se asume conocida. Denotamos el $100\left(1-\frac{\alpha}{2}\right)$ percentil de la distribución normal estándar como $z_{\alpha/2}$. Para cualquier muestra aleatoria, los 2 puntos finales del intervalo estimado con un nivel de confianza de $(1-\alpha)\%$ están dados por:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

En el ejemplo, asumimos que la desviación estándar σ de la altura de todos los estudiantes es 9.48 cm (desviación poblacional). Para encontrar el margen del error y una estimación intervalar con un nivel de confianza del 95%, primero filtramos los valores que faltan en `survey$Height` con la función `na.omit`, y lo guardamos en `height.response`:

```
height.response = na.omit(survey$Height)
```

A continuación se calcula el error estándar (desviación estándar de la media muestral):

```
n = length(height.response)
sigma = 9.48
SE = sigma/sqrt(n); SE
```

```
## [1] 0.6557453
```

Debido a que el intervalo está centrado en $\bar{x}$, el 95% de confianza implica el 97.5th percentil de la distribución normal en la cola superior. Por lo tanto, $z_{\alpha/2}$ está dado por `qnorm(0.975)`. Lo multiplicamos por el error estándar SE y obtenemos el margen de error.

```
E = qnorm(0.975)*SE; E
```

```
## [1] 1.285237
```

Finalmente, a la media muestral le sumamos y le restamos este valor para obtener los extremos del intervalo.

```r
xbar = mean(height.response)
IC = xbar + c(-E,E); IC
```

```
## [1] 171.0956 173.6661
```

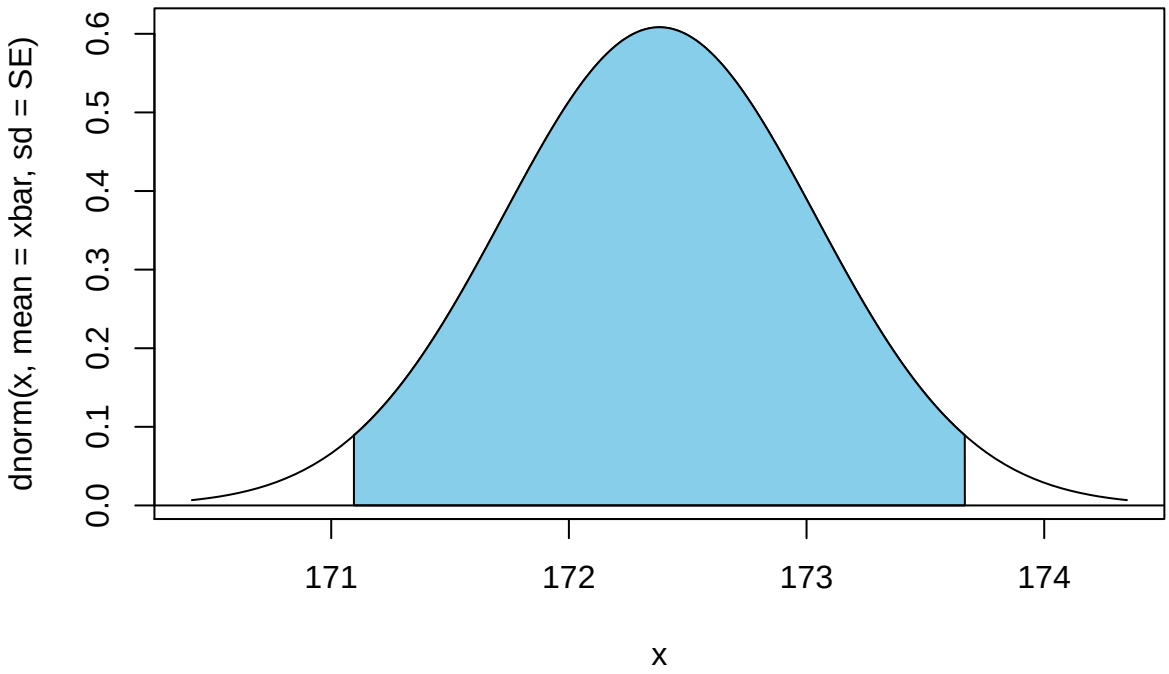
Conclusión: Asumiendo la desviación estándar poblacional σ como 9.48, el margen de error para la media de la altura de los estudiantes con 95% de confianza es 1.2852 centímetros, y el intervalo para la media poblacional es entre 171.10 y 173.67 centímetros.

TIP

Para representar gráficamente el intervalo de confianza para la media poblacional, se puede ejecutar el siguiente código:

```r
# Gráfica de la función de densidad de la media muestral
curve(dnorm(x,mean=xbar,sd=SE),from=xbar-3*SE, to=xbar+3*SE)

# Gráfica de la región del intervalo
cord.x=c(IC[1],seq(IC[1],IC[2],0.01),IC[2])
cord.y=c(0,dnorm(seq(IC[1],IC[2],0.01),mean=xbar,sd=SE),0)
polygon(cord.x,cord.y,col="skyblue")
abline(h=0)
```



Hasta el momento hemos usado la fórmula general para el cálculo del intervalo de confianza, sin embargo, podemos aplicar la función `z.test` del paquete "TeachingDemos". No es un paquete que se instala por defecto en R, este debe ser instalado y cargado previamente:

```
library(TeachingDemos) # load TeachingDemos package
IC = z.test(height.response, sd=sigma, conf.level=0.95); IC
```

```
##
##  One Sample z-test
##
## data:  height.response
## z = 262.88, n = 209.00000, Std. Dev. = 9.48000, Std. Dev. of the
## sample mean = 0.65575, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  171.0956 173.6661
## sample estimates:
## mean of height.response
##                172.3809
```

2.3 Estimación intervalar de la media de una población con varianza desconocida a partir de una muestra pequeña

Por otro lado, si la varianza poblacional σ^2 se asume como desconocida, para una muestra aleatoria pequeña, los puntos finales del intervalo estimado con nivel de confianza de $(1-\alpha)\%$ están dados por: $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$, donde $t_{\alpha/2}$ es el percentil $100(1-\alpha/2)$ de la distribución t-student con $n-1$ grados de libertad y s es la desviación estándar de la muestra. En el ejemplo dado, para encontrar el error marginal y el intervalo estimado al 95% de confianza utilizando únicamente los 10 primeros elementos de la muestra, primero hemos de filtrar los valores que faltan en `survey$Height` con la función `na.omit`, y guardarlo en `height.response`.

```
height.response = na.omit(survey$Height)
height.response = height.response[1:10]
```

Entonces calculamos el error estándar estimado ya que la deviación de la población σ es desconocida y se estima a partir de la desviación de la muestra s.

```
n = length(height.response)
s = sd(height.response)
SE = s/sqrt(n); SE
```

```
## [1] 2.874441
```

De la misma manera que el caso anterior, hay dos colas de la distribución t-student, por lo tanto, el 95 de nivel de confianza implica el 97.5th percentil de la distribución t-student en la cola superior. De esta manera, $t\alpha/2$ está dado por `qt(0.975,df=n-1)`. Lo multiplicamos por el error estándar estimado SE y obtendremos el margen del error.

```
E = qt(.975, df=n-1)*SE; E
```

```
## [1] 6.502436
```

A continuación se lo restamos y sumamos a la media muestral para determinar los límites del intervalo de confianza.

```
xbar = mean(height.response)
IC = xbar + c(-E,E); IC
```

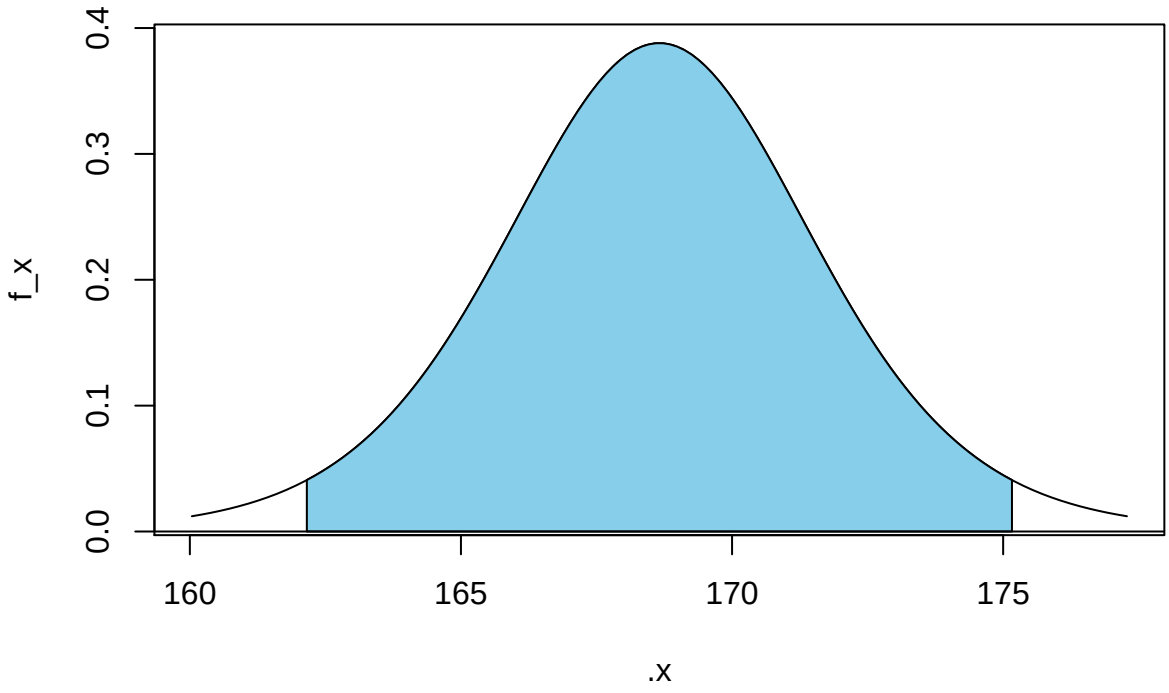
```
## [1] 162.1576 175.1624
```

Conclusión: Sin asumir la desviación estándar poblacional σ, el margen del error para la media de la altura de los estudiantes con 95% de confianza es de 6.502 centímetros y el intervalo de confianza para la media de la población está entre 162.16 y 175.16 centímetros.

TIP

Para representar gráficamente el intervalo de confianza para la media poblacional, se puede ejecutar el siguiente código:

```
.x = seq(xbar-3*SE,xbar+3*SE,length=100)
f_x = dt(seq(-3,3,length=100),df=n-1)
plot(.x,f_x,type="l")
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dt(seq((IC[1]-xbar)/SE,(IC[2]-xbar)/SE,length=100),df=n-1),0)
polygon(cord.x,cord.y,col="skyblue")
abline(h=0)
```



De igual forma, R contiene una función que calcula este intervalo directamente. La función se denomina `t.test` y pertenece al paquete "stats" que normalmente está integrado en la instalación básica de R.

```
t.test(height.response)
```

```
##
##  One Sample t-test
##
## data:  height.response
## t = 58.676, df = 9, p-value = 6.111e-13
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  162.1576 175.1624
## sample estimates:
## mean of x
##    168.66
```

2.4 Estimación intervalar de la media de una población con varianza desconocida a partir de una muestra grande

Ya que la distribución t-student con un grado de libertad alto (n > 30) se puede considerar aproximadamente igual a una distribución normal, si la varianza poblacional σ^2 es desconocida, para una muestra de tamaño

suficiente, los puntos finales del intervalo estimado con un nivel de confianza (1 - α) % están dados por:

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

, donde $z_{\alpha/2}$ es el percentil $100(1-\alpha/2)$ de la distribución normal y s es la desviación estándar de la muestra. En el ejemplo dado, al no conocer la desviación estándar poblacional σ de la altura de los estudiantes, para encontrar el margen de error y el intervalo de confianza estimado con 95% de confianza, primero debemos filtrar los valores que faltan en `survey$Height` con la función `na.omit`,y guardarlo en `height.response`.

```
height.response = na.omit(survey$Height)
```

Después calculamos el error estándar estimado ya que la desviación de la población σ es desconocida y se estima a partir de la desviación de la muestra s.

```
n = length(height.response)
s = sd(height.response)
SE = s/sqrt(n); SE
```

```
## [1] 0.6811677
```

Como en los casos anteriores, las dos colas de la distribución t-student, sugieren que el 95 de nivel de confianza implica el 97.5th percentil de la distribución t-student en la cola superior. Así pues, $t_{\alpha/2}$ está dado por `qnorm(0.975)`. Lo multilpicamos por el error estándar estimado SE y obtendremos el margen de error.

```
E = qnorm(.975)*SE; E
```

```
## [1] 1.335064
```

Lo añadimos con la media muestral, y encontramos el intervalo de confianza.

```
xbar = mean(height.response)
IC = xbar + c(-E,E); IC
```

```
## [1] 171.0458 173.7159
```

Conclusión: Sin conocer la desviación estándar poblacional σ, el margen de error para la media de la altura de los estudiantes al 95% de confianza es de 1.335 centímetros y el intervalo de confianza para la media de la población está entre 171.05 y 173.73 centímetros.

Para calcular este intervalo, también podemos utilizar las funciones `t.test` del paquete "stats" o `z.test` del paquete "TeachingDemos".

```
t.test(height.response)
```

```
##
##  One Sample t-test
##
## data:  height.response
## t = 253.07, df = 208, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  171.0380 173.7237
## sample estimates:
## mean of x
##  172.3809
```

```r
z.test(height.response,sd=sd(height.response))
```

```
##
##  One Sample z-test
##
## data:  height.response
## z = 253.07, n = 209.00000, Std. Dev. = 9.84753, Std. Dev. of the
## sample mean = 0.68117, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  171.0458 173.7159
## sample estimates:
## mean of height.response
##                172.3809
```

2.5 Tamaño de la muestra

La calidad de una muestral se puede mejorar incrementando el tamaño de la misma. Para determinar el tamaño de la muestra necesario para cumplir con los requerimientos del intervalo de la media poblacional estimado con nivel de confianza (1 - α) % está dado por:

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2}$$

, donde E es el margen de error, σ^2 la varianza poblacional y $z_{\alpha/2}$ es el $100(1-\alpha/2)$ percentil de la distribución normal estándar. En el ejemplo, conociendo que la desviación estándar poblacional σ de la altura de los estudiantes es 9.48, para encontrar el tamaño necesario de la muestra para conseguir un margen de error de 1.2 centímetros al 95% de confianza, se puede ejecutar el siguiente código.

```r
z_alpha_2 = qnorm(.975)
sigma = 9.48
E = 1.2
z_alpha_2^2 * sigma^2/ E^2
```

```
## [1] 239.7454
```

Conclusión: Basándonos en que conocemos la desviación estándar poblacional, la muestra ha de tener un tamaño mínimo de 240 observaciones para conseguir un margen de error de 1.2 centímetros al 95% de confianza.

2.6 ¿Qué quiere decir el nivel de significancia?

Se ha definido en todo momento que el intervalo estimado tiene un nivel de confianza de (1 - α) % . Por otra parte, se puede apreciar que si se conoce la varianza poblacional, la longitud del intervalo $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ solo depende del tamaño de la muestra n. Por lo tanto, si realizamos otra muestra del mismo tamaño, la longitud del intervalo se mantiene, pero la posición del mismo no, ya que depende de $\bar{x}$ y este nuevo intervalo puede o no incluir el valor real de la media de la población. Teniendo en cuenta lo anterior, se puede definir que el nivel de confianza del intervalo es la proporción de intervalos estimados que incluirán el valor real del parámetro, es decir, existe un $100\alpha\%$ de probabilidad de que el intervalo estimado no incluya el valor real. Para verificarlo, suponemos que la altura media de la población del ejemplo es 172 y su desviación estándar es 9.48. Si realizamos 100 muestras de 20 observaciones cada una, simulamos los valores de las 100 medias muestrales, calculamos los 100 intervalos de confianza del 95% ($\alpha = 0.05$) y los representamos gráficamente. Finalmente, resaltaremos aquellos intervalos estimados que no contienen el valor real de la media de la población.

```r
mu = 172 ; sigma = 9.48 ; n=20 ; alpha = 0.05
xbar = rnorm(100,mean=mu, sd=sigma/sqrt(n))
SE = sigma/sqrt(n); SE
```

```
## [1] 2.119792
```

```r
E = qnorm(1-alpha/2)*SE; E
```

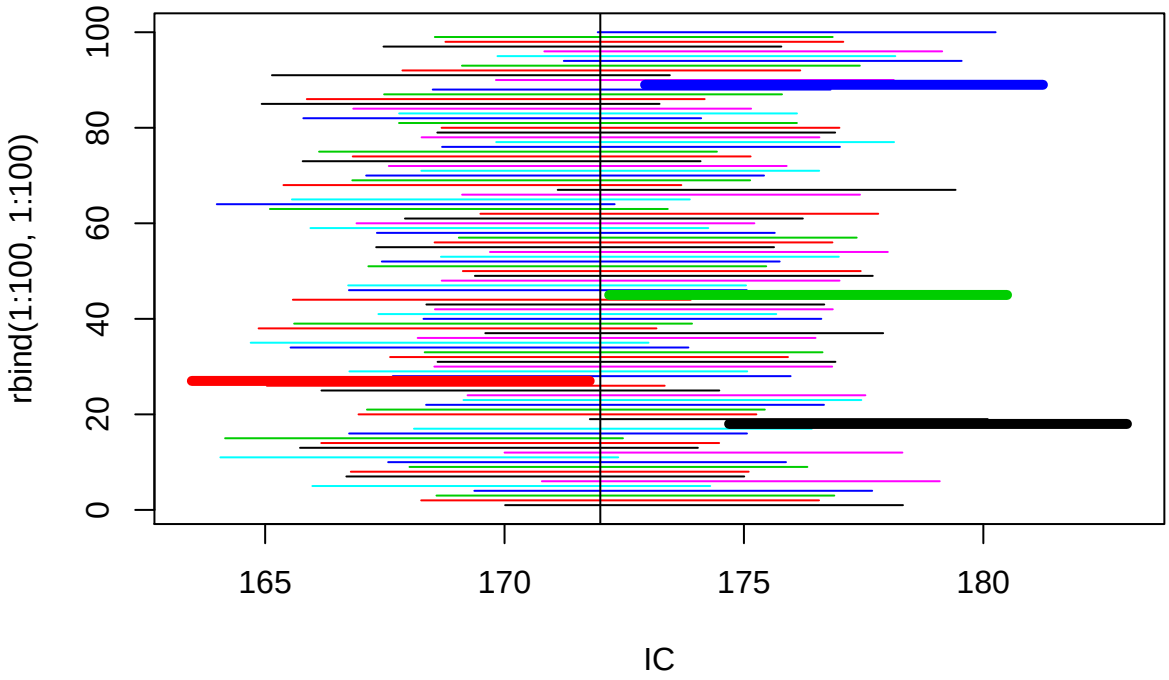
```
## [1] 4.154717
```

```r
IC = rbind(xbar - E, xbar + E); IC
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,] 170.0121 168.2588 168.5780 169.3690 165.9861 170.7798 166.6962
## [2,] 178.3215 176.5682 176.8874 177.6784 174.2955 179.0893 175.0056
##          [,8]     [,9]    [,10]    [,11]    [,12]    [,13]    [,14]
## [1,] 166.7893 168.0151 167.5687 164.0641 170.0001 165.7284 166.1722
## [2,] 175.0987 176.3245 175.8782 172.3736 178.3096 174.0378 174.4816
##         [,15]    [,16]    [,17]    [,18]    [,19]    [,20]    [,21]
## [1,] 164.1634 166.7540 168.1084 174.6901 171.7817 166.9498 167.1266
## [2,] 172.4728 175.0634 176.4179 182.9996 180.0911 175.2592 175.4360
##         [,22]    [,23]    [,24]    [,25]    [,26]    [,27]    [,28]
## [1,] 168.3616 169.1420 169.2273 166.1760 165.0352 163.4688 167.6652
## [2,] 176.6710 177.4514 177.5367 174.4855 173.3447 171.7783 175.9747
##         [,29]    [,30]    [,31]    [,32]    [,33]    [,34]    [,35]
## [1,] 166.7594 168.5332 168.6015 167.6101 168.3319 165.5307 164.6980
## [2,] 175.0688 176.8426 176.9109 175.9195 176.6414 173.8401 173.0074
##         [,36]    [,37]    [,38]    [,39]    [,40]    [,41]    [,42]
## [1,] 168.1858 169.5980 164.8633 165.6058 168.3065 167.3657 168.5461
## [2,] 176.4953 177.9074 173.1727 173.9152 176.6159 175.6751 176.8555
##         [,43]    [,44]    [,45]    [,46]    [,47]    [,48]    [,49]
## [1,] 168.3677 165.5789 172.1853 166.7501 166.7338 168.6871 169.3811
## [2,] 176.6772 173.8884 180.4947 175.0596 175.0432 176.9966 177.6905
##         [,50]    [,51]    [,52]    [,53]    [,54]    [,55]    [,56]
## [1,] 169.1293 167.1557 167.4367 168.6717 169.6957 167.3186 168.5388
## [2,] 177.4387 175.4651 175.7461 176.9811 178.0052 175.6280 176.8482
##         [,57]    [,58]    [,59]    [,60]    [,61]    [,62]    [,63]
## [1,] 169.0453 167.3343 165.9459 166.9085 167.9205 169.4953 165.0997
## [2,] 177.3547 175.6437 174.2554 175.2179 176.2299 177.8047 173.4092
##         [,64]    [,65]    [,66]    [,67]    [,68]    [,69]    [,70]
## [1,] 163.9903 165.5594 169.1140 171.1104 165.3824 166.8200 167.1096
## [2,] 172.2998 173.8689 177.4235 179.4198 173.6918 175.1294 175.4190
##         [,71]    [,72]    [,73]    [,74]    [,75]    [,76]    [,77]
## [1,] 168.2618 167.5826 165.7840 166.8294 166.1262 168.6954 169.8254
## [2,] 176.5712 175.8920 174.0935 175.1389 174.4357 177.0049 178.1349
##         [,78]    [,79]    [,80]    [,81]    [,82]    [,83]    [,84]
## [1,] 168.2671 168.5956 168.6851 167.7962 165.7970 167.7983 166.8403
## [2,] 176.5765 176.9050 176.9945 176.1056 174.1064 176.1077 175.1497
##         [,85]    [,86]    [,87]    [,88]    [,89]    [,90]    [,91]
## [1,] 164.9278 165.8699 167.4849 168.5003 172.9350 169.8198 165.1423
## [2,] 173.2372 174.1794 175.7943 176.8097 181.2444 178.1292 173.4517
##         [,92]    [,93]    [,94]    [,95]    [,96]    [,97]    [,98]
```

```
## [1,] 167.8668 169.1111 171.2374 169.8522 170.8309 167.4720 168.7653
## [2,] 176.1763 177.4205 179.5468 178.1616 179.1403 175.7814 177.0748
##         [,99]   [,100]
## [1,] 168.5440 171.9468
## [2,] 176.8534 180.2562
```

```
matplot(IC,rbind(1:100,1:100),type="l",lty=1)
abline(v=mu)
out=which(!(IC[1,]<mu & mu<IC[2,]))
matplot(IC[,out],rbind(out,out),type="l",lty=1,add=T,lwd=5)
```



TAREA

Instrucciones

- Favor entregar la tarea en un archivo `ZIP` conteniendo tanto el desarrollo en Rmarkdown (archivo .Rmd) como la compilación en HTML.
- Cree una cuenta en www.kaggle.com, un sitio web donde encontrará grandes cantidades de fuentes de datos, así como recursos sobre R, python, y ciencia de datos en general.
- En este ejercicio se utilizará el dataset disponible en https://www.kaggle.com/kmader/colorectal-histology-mnist. Este dataset contiene imágenes histopatológicas de pacientes con cáncer colorrectal. En particular, hay 8 tipos diferentes de tejidos, descritos de forma codificada por la columna `label`. La codificación realizada se muestra en la siguiente tabla:

Código	Etiqueta
1	TUMOR
2	STROMA
3	COMPLEX
4	LYMPHO
5	DEBRIS
6	MUCOSA
7	ADIPOSE
8	EMPTY

Para este ejercicio utilizaremos sólo el dataset que contiene 5000 imágenes de escala de grises de 64 píxeles. Específicamente, baje sólo el archivo *hmnist_28_28_L.csv* (que pesa 4342 KB comprimido en ZIP). No es necesario bajar todo el dataset completo (que pesa 991MB). Cada fila corresponde a una imagen, mientras que cada columna corresponde al valor de un píxel. Considere que cada imagen se ha desdoblado por columnas (es decir, en cada fila se ha puesto en serie la primera columna, luego la segunda columna y así sucesivamente). Realice todos sus cálculos en el lenguaje `R`.

Ejercicios

1. Realice un análisis descriptivo (usando también gráficas descriptivas como histogramas, boxplots, etc.) del valor del píxel central de la imagen. Considere los siguientes subconjuntos de datos.

    a. Todas los tipos de tejidos.
    b. Los tejidos con cáncer (TUMOR).
    c. Los tejidos sin cáncer (que no son TUMOR).
    d. Cada uno de los restantes tipos de tejido (STROMA, COMPLEX, etc.).

2. Grafique una imagen de cada tipo en R, a partir del dataset estudiado.

3. Calcule un intervalo de confianza (95% de confianza) para la media de la variable del píxel central para los siguientes casos. Para este cálculo no utilice librerías adicionales, evalúe cada elemento del intervalo usando las funciones disponibles en la distribución estándar de R. En particular, considere las funciones `qt` y `qnorm` para el cálculo de cuantiles.

    a. Los tejidos con cáncer.
    b. Los tejidos sin cáncer.
    c. Cada uno de los restantes tipos de tejido.

4. Repita el punto anterior con las siguientes indicaciones:

    a. Con una muestra aleatoria de 200 imágenes.
    b. Con una muestra aleatoria de 20 imágenes.
    c. Con 100 muestras aleatorias cada una con un tamaño de 200 imágenes. Grafique el histograma de las medias y los intervalos de confianza para cada muestra.
    d. Con 100 muestras aleatorias cada una con un tamaño de 20 imágenes. Grafique el histograma de las medias y los intervalos de confianza para cada muestra.

5. Vuelva y realice todo el análisis de los puntos anteriores pero utilizando como variable, la media de los valores de píxel de cada imagen.