.UBAfiuba
FACULTAD DE INGENIERÍA

Trabajo Práctico

[91.03] Estadística Aplicada I
2do cuatrimestre - 2020

Alumna	**Padrón**	**Correo electrónico**
Cabeza, Milagros	103759	mcabeza@fi.uba.ar

1. Introducción

El presente informe reúne la documentación de la solución del trabajo práctico de la materia Estadística Aplicada I (91.03), que consiste en la simulación de un experimento aleatorio.

A continuación, una descripción de la Variable Aleatoria a estudiar:

***"Registros de máxima presión"**: Una industria química posee 4 plantas, y un total de 15 reactores recientemente adquiridos en los que se realiza un proceso de síntesis de ácido clorhídrico. Para dimensionar el sistema de control, se registró la máxima presión de operación (en Bar) durante un día para distintos reactores tomados al azar de las 4 plantas, generando 100 datos en total. Los reactores de las 4 plantas trabajaron en condiciones controladas durante el experimento y se asume que no hay diferencias entre las plantas.*

Se llamará, a lo largo del informe, X: "Registros de máxima presión" [Bar]

2. Modelización de la Variable Aleatoria

Se define a una **muestra aleatoria**: para X una Variable Aleatoria con distribución de probabilidad $f_X(x)$, se efectúan n observaciones x_i de Variables Aleatorias X_i observables. Para que la muestra sea aleatoria se deben cumplir dos condiciones:

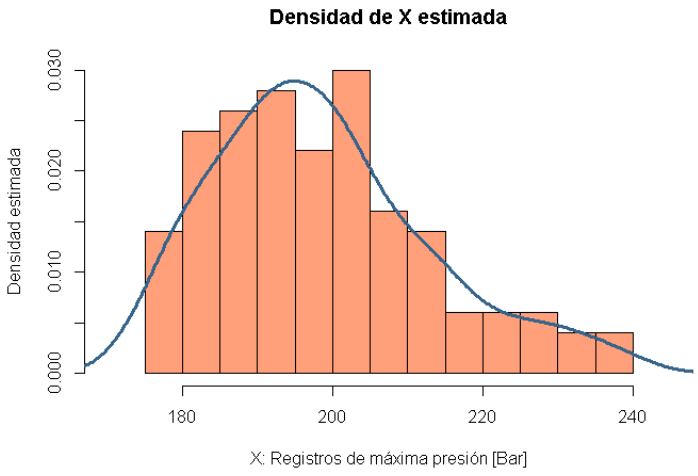
- Las X_i son independientes.
- Las X_i son Variables Aleatorias idénticamente distribuidas.

Obteniendo así una sucesión de Variables Aleatorias $\{X_1, X_2, ..., X_n\}$ independientes e idénticamente distribuidas, de las cuáles se obtendrán valores en cada observación.

La muestra se realizó en reactores tomados al azar entre 4 plantas, como los reactores trabajaron en condiciones controladas y no hay diferencias entre las plantas, se puede asegurar que la muestra fue aleatoria, dado que cumple que son independientes e idénticamente distribuidas.

2.0.1. Histograma

A partir de los 100 datos obtenidos en la muestra aleatoria, se realizó un histograma:



2.0.2. Cálculo de la media y la varianza muestral

Se utilizaron estadísticos insesgados para realizar la estimación puntual: $\overline{X}$ para la media y S^2 para la varianza. Siendo estos estadísticos las variables aleatorias dadas por:

- $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
- $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{(n-1)}$

De forma tal que $E(\overline{X}) = \mu$ y $E(S^2) = \sigma$, lo que los clasifica como insesgados.

Luego se realizó el cálculo del valor aleatorio que toman los estadísticos $\overline{X}$ y S^2 correspondientes a los datos obtenidos:

$\overline{x} = 198{,}6$
$s^2 = 207{,}8691$

Se puede observar que la media estimada es un punto del intervalo en el que, a simple vista, puede estar la media en el histograma realizado.

2.0.3. Cálculo del coeficiente de asimetría de Fisher

Se llevó a cabo el cálculo del coeficiente de asimetría de Fisher o momento centrado de tercer orden $\alpha_3 = \frac{(x-\mu)^3}{[var(x)]^{\frac{3}{2}}}$. Al obtener el resultado, se pueden presentar 3 diferentes casos:

- $\alpha_3 > 0$: la variable presenta una asimetría positiva.
- $\alpha_3 = 0$: la variable es simétrica.
- $\alpha_3 < 0$: la variable presenta una asimetría negativa.

Para la Variable Aleatoria X, $\alpha_3 = 0{,}7015147$, lo cual indica que X presenta asimetría positiva. Es decir, son más frecuentes los valores más bajos de máxima presión que los más altos. Esto se puede observar en el histograma, la mayor acumulación de probabilidad se encuentra en los valores más bajos de máxima presión.

2.0.4. Cálculo de los cuartiles y boxplot

Se definen a los cuartiles como aquellos valores que acumulan a su izquierda los valores de probabilidad: {0, 0.25, 0.5, 0.75, 1}.

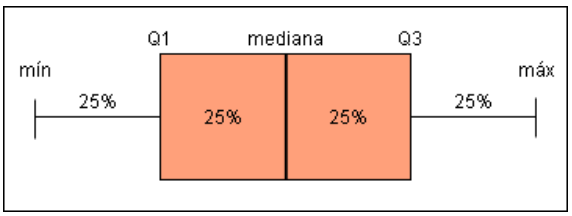
A continuación, una tabla con los valores que toma X para cada cuartil:

Cuartil	**x**
0	175.900
0.25	188.475
0.5	196.850
0.75	206.325
1	238.400

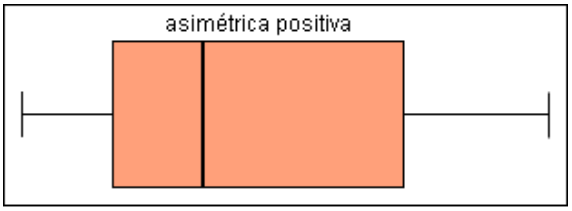
Se puede observar que el valor que acumula 0.5 de probabilidad (coincidente con la mediana), es levemente menor que la media muestral, lo cual hace referencia a la asimetría positiva de la variable, antes vista en el cálculo del coeficiente de asimetría.

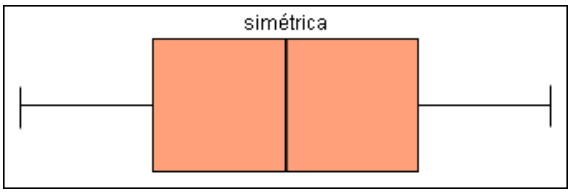
Para resumir y visualizar los cuartiles y la asimetría de la variable, se realiza un diagrama boxplot. Se dividen los datos en 4 secciones, donde cada sección representa el 25 % de los datos.

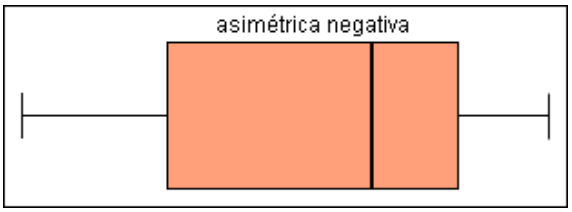
Los extremos del diagrama representan el mínimo (cuartil 0) y máximo (cuartil 1). Los extremos de la caja representan Q1 (cuartil 0.25) y Q3 (cuartil 0.75). La línea dentro de la caja representa la mediana (cuartil 0.5).



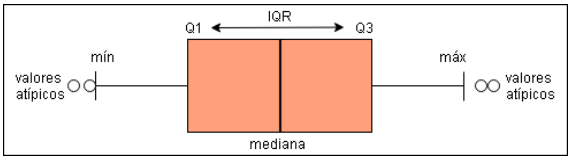
Según la asimetría de la variable, las secciones a los costados de la caja tendrán distinta longitud. La mediana se desplazará dentro de la caja del centro hacia la derecha para asimétricas negativas, y del centro hacia la izquierda para asimétricas positivas.



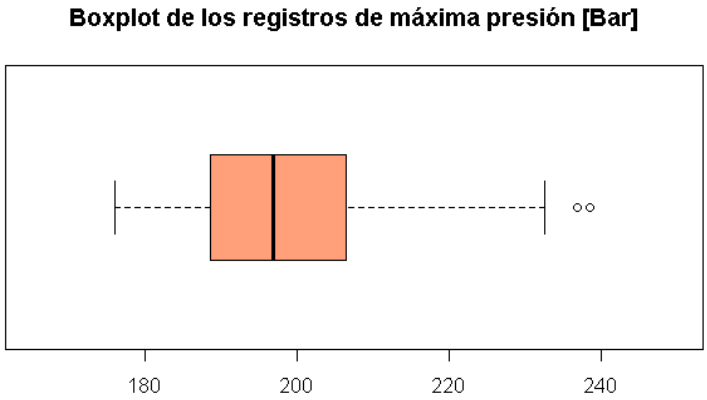




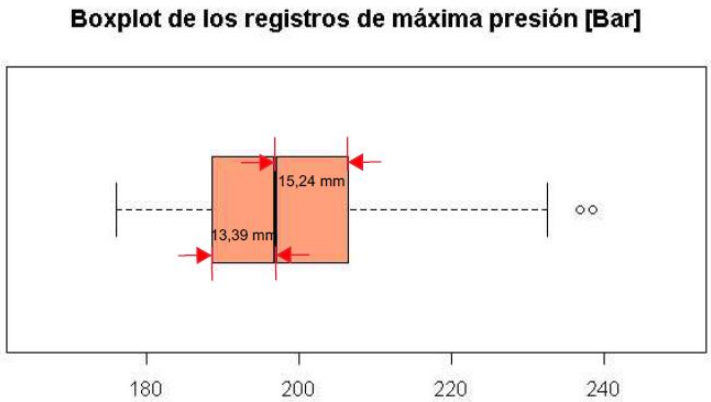
La longitud de la caja es el rango intercuartílico $IQR = Q3 - Q1$, es una medida de la variabilidad. Representa la cantidad de dispersión en la mitad central de los datos.
Si existieran valores atípicos, se pueden representar fuera del diagrama con puntos, son observaciones numéricamente muy distantes de los límites superior e inferior.

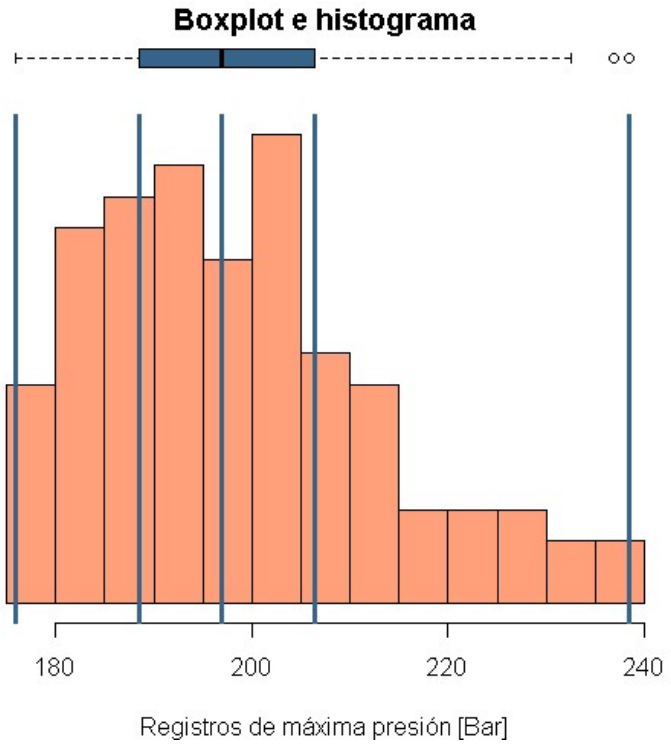


A continuación se observa el boxplot correspondiente a la variable X.
Se puede observar su asimetría positiva: la mayor densidad de probabilidad (es decir la caja, que muestra un 50 % de probabilidad) se encuentra corrida del centro hacia la izquierda; además, la mediana se encuentra desplazada hacia la izquierda, lo cual quiere decir que para valores más bajos ya se acumuló el 50 % de la probabilidad; también se observa su asimetría positiva en las secciones a los costados de la caja, la de la izquierda es de menor longitud que la de la derecha, aunque se acumula la misma probabilidad de 25 % en ambas secciones, en los valores más bajos de máxima presión se acumuló en un intervalo más pequeño.
Además, se puede observar que por fuera del diagrama, se representan con puntos los valores atípicos, alejados del límite superior.
El rango intercuartílico $IQR = Q3 - Q1 = 17{,}85$. A simple vista, la caja es angosta, lo cuál indica baja dispersión. Esto se con-dice con que los reactores hayan trabajado en condiciones controladas durante el experimento.



Para facilitar la visualización de la relación entre el boxplot y el histograma, se realiza un gráfico donde se hacen coincidir exactamente los cuartiles con el boxplot, como se muestra a continuación. Además, se agrega un boxplot acotado, donde se ve cómo la mediana está levemente desplazada del centro hacia la izquierda.





2.0.5. Propuesta de modelos de distribución

Teniendo en cuenta los aspectos empíricos (asimetría, forma del histograma) y la naturaleza de la Variable Aleatoria, se proponen modelos de distribución de probabilidad que puedan ajustar los datos. Los modelos propuestos serán desarrollados a continuación.

Distribución Lognormal

Se elige esta distribución de probabilidad por dos razones: la primera es su asimetría positiva, coincidente con los aspectos empíricos; la segunda es su naturaleza, la distribución Lognormal toma valores de grupos de individuos en un tiempo fijo y no los valores de un individuo a lo largo del tiempo, por lo cual se la llama variable de *corte trasversal*. En este caso, la variable toma valores de presión máxima de distintos valores en un tiempo fijo.

•Estimación de parámetros para la distribución Lognormal

Para la estimación de parámetros se utiliza el *método de estimación por máxima verosimilitud.*

Para $X \sim LN(m, D^2)$, se realiza el planteo de la función de verosimilitud:

$L(m, D^2) = \prod_{i=1}^{n} f(x_i)$

$L(m, D^2) = (\frac{1}{\sqrt{2\pi D^2 x}})^2 e^{-\frac{1}{2D^2} \sum_{i=1}^{n} (log(x_i)-m)^2}$

Planteo de la función de log-verosimilitud:
[1]

$log(L(m, D^2)) = \ell(m, D^2)$

[1] log es logaritmo natural

$\ell(m, D^2) = -\frac{n}{2}log(D^2) - \sum_{i=1}^{n} log(x_i) - \frac{n}{2}log(2\pi) - \frac{1}{2D^2}\sum_{i=1}^{n}(log(X_i) - m)^2$

Para estimar m, se busca el máximo de la función de verosimilitud, derivando respecto de m e igualando a 0:

$\frac{\partial \ell}{\partial m} = \frac{1}{D^2}\sum_{i=1}^{n}(log(x_i) - m)$

$\frac{1}{D^2}\sum_{i=1}^{n}(log(X_i) - \hat{m}) = 0$

$\sum_{i=1}^{n} log(X_i) - \hat{m}n = 0$

$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} log(X_i)$

Para estimar D^2, se busca el máximo de la función de verosimilitud, derivando respecto de D e igualando a 0. Como $D > 0$ y como D^2 es una función creciente para $D > 0$, el máximo que se consigue derivando respecto de D es el mismo que para D^2:

$\frac{\partial \ell}{\partial D} = -\frac{n}{\sqrt{D^2}} + \sum_{i=1}^{n}(log(X_i) - \hat{m})^2(D^2)^{-\frac{3}{2}}$

$-\frac{n}{\sqrt{\hat{D}^2}} + \sum_{i=1}^{n}(log(X_i) - \hat{m})^2(\hat{D^2})^{-\frac{3}{2}} = 0$

$\hat{D^2} = \frac{\sum_{i=1}^{n}(log(X_i) - \hat{m})^2}{n}$

A partir de los valores calculados anteriormente de $\overline{x} = 198{,}6$ y de $s^2 = 207{,}8691$, se calcularon los valores de los estimadores puntuales $\hat{m}$ y $\hat{D}^2$

$\hat{m} = 5{,}288753$
$\hat{D}^2 = 0{,}005017238$

•Función de verosimilitud y log-verosimilitud para la distribución Lognormal

Función de verosimilitud:

La función de verosimilitud da la densidad de probabilidad conjunta de haber obtenido la muestra que se obtuvo para una determinada distribución. Siendo X_i independientes e idénticamente distribuidas, se calcula como:

$L(m, D^2) = \prod_{i=1}^{n} f(x_i)$

$L(m, D^2) = \prod_{i=1}^{n} \frac{1}{x_i\sqrt{D^2 2\pi}} e^{-\frac{1}{2}(\frac{log(x_i)-m}{\sqrt{D^2}})^2}$

Cálculo de log-verosimilitud para la distribución Lognormal:

$\ell(m, D^2) = log(L(m, D^2)) = \sum_{i=1}^{n} log(f(x_i))$

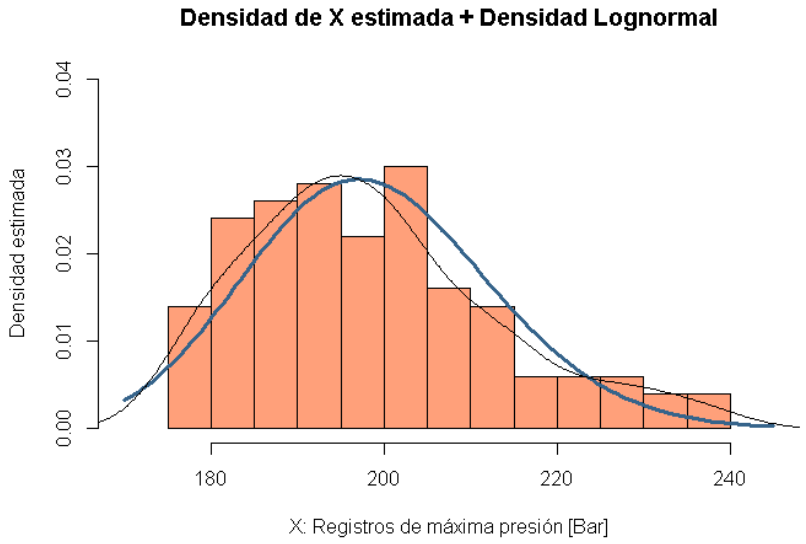
$\ell(m, D^2) = \sum_{i=1}^{n} log(\frac{1}{x_i\sqrt{D^2 2\pi}} e^{-\frac{1}{2}(\frac{log(x_i)-m}{\sqrt{D^2}})^2})$

Para los valores de $\hat{m} = 5{,}288753$ y $\hat{D}^2 = 0{,}005017238$ estimados anteriormente, se realiza el cálculo de la log-verosimilitud.

$\ell(\hat{m}, \hat{D}^2) = \sum_{i=1}^{n} log(\frac{1}{x_i\sqrt{\hat{D}^2 2\pi}} e^{-\frac{1}{2}(\frac{log(x_i)-\hat{m}}{\sqrt{\hat{D}^2}})^2})$

$\ell(\hat{m}, \hat{D}^2) = -406{,}0253$

A continuación se muestra un gráfico del histograma, donde se incluye la densidad estimada (en color negro) y superpuesta la densidad de la Lognormal (en color azul).



Distribución Gumbel del máximo

Las razones por la que esta distribución fue elegida para modelar a la variable X fueron: por un lado, su naturaleza, dado que esta distribución se utiliza para modelar variables extremas, en este caso para modelar el extremo máximo, de los registros de presión se tomaron para la muestra solo los máximos; la otra razón que determinó su elección fue su asimetría positiva.

•Estimación de parámetros para la distribución Gumbel del Máximo

Para la estimación de parámetros se utiliza el *método de estimación por momentos*. A partir de los valores previamente calculados de $\overline{x} = 198{,}6$ y de $s^2 = 207{,}8691$, se calcularon los valores de los estimadores puntuales para los parámetros $\hat{\beta}$ (parámetro de escala) y $\hat{\theta}$ (parámetro de posición).

Se plantea el sistema de ecuaciones:

$s^2 = \frac{\pi}{\sqrt{6}}\hat{\beta}$

$\overline{x} = \hat{\theta} + 0{,}5772157\hat{\beta}$

Del cuál se obtuvieron los siguientes valores aleatorios de los estimadores puntuales $\hat{\beta}$ y $\hat{\theta}$:

$\hat{\beta} = 11{,}241$
$\hat{\theta} = 192{,}11$

•Función de verosimilitud y log-verosimilitud para la distribución Gumbel del Máximo

Función de verosimilitud:

Siguiendo el mismo razonamiento que para la distribución Lognormal, se plantea la función de verosimilitud de la Gumbel del Máximo como:

$L(\theta, \beta) = \prod_{i=1}^{n} f(x_i)$

$L(\theta, \beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-e^{-\frac{x_i-\theta}{\beta}} + \frac{x_i-\theta}{\beta}}$

Cálculo de log-verosimilitud para la distribución Lognormal:

$\ell(\theta, \beta) = log(L(\theta, \beta)) = \sum_{i=1}^{n} Ln(f(x_i))$

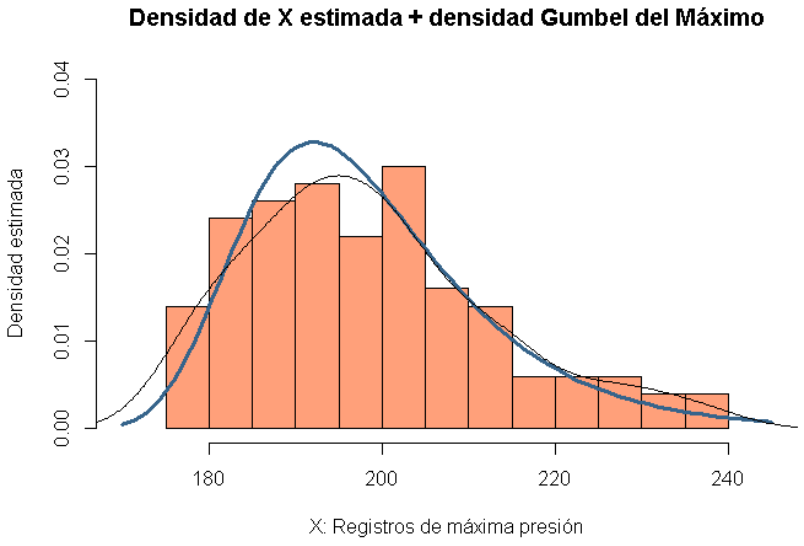
$\ell(\theta, \beta) = \sum_{i=1}^{n} log(\frac{1}{\beta} e^{(-e^{-\frac{x_i-\theta}{\beta}} + \frac{x_i-\theta}{\beta})})$

Para los valores de $\hat{\theta}$ = 192,111 y $\hat{\beta}$ = 11,241 estimados anteriormente, se realiza el cálculo de la log-verosimilitud.

$\ell(\hat{\theta}, \hat{\beta}) = \sum_{i=1}^{n} log(\frac{1}{\hat{\beta}} e^{(-e^{-\frac{x_i-\hat{\theta}}{\hat{\beta}}} + \frac{x_i-\hat{\theta}}{\hat{\beta}})})$

$\boxed{\ell(\hat{\theta}, \hat{\beta}) = -402{,}8769}$

A continuación se muestra un gráfico del histograma, donde se incluye la densidad estimada (en color negro) y superpuesta la densidad de la Gumbel del Máximo (en color azul).



Como se especificó anteriormente, la función de verosimilitud da la densidad conjunta de probabilidad de haber obtenido la muestra que se obtuvo. Para facilitar los cálculos, se plantea la función de log-verosimilitud.
Al ser la verosimilitud una densidad de probabilidad, toma valores mayores a 0; al calcular su

logaritmo, el valor de la log-verosimilitud estará contenido entre $(-\infty, \infty)$. Por lo tanto, como el máximo de la función de verosimilitud coincide con el de la log-verosimilitud (por ser esta última una función creciente), los valores de mayor verosimilitud son los de mayor log-verosimilitud.

2.0.6. Comparación de modelos desde la log-verosimilitud

Si se realiza un análisis sobre los resultados de las log-verosimilitudes para cada distribución:

Distribución	Log-verosimilitud
Lognormal	-406.0253
Gumbel del Máximo	-402.8769

Se observa que, si bien son valores próximos, el mayor valor corresponde a la log-verosimilitud de la distribución Gumbel del Máximo. Como conclusión, entre estas dos distribuciones, la que mejor la modela es la de Gumbel del Máximo.

3. Estimación de la media por intervalos de confianza

Para la estimación de la media se realiza la técnica del bootstrapping. Consiste en tomar a la muestra original como una "población" y llevar a cabo la realización de re-muestras sacadas con reposición de la muestra original. En general se toman tantas re-muestras como elementos tenga la muestra original. El mismo se utiliza en este caso para construir intervalos de confianza para la media, el método es más preciso que si se calcularan los intervalos de confianza directamente de la muestra original.
El procedimiento consiste en tomar los datos iniciales, y generar 100 muestras aleatorias con reposición de tamaño n. De esa forma se generan 100 re-muestras, a cada una se le construye un intervalo de confianza del 90 % para la media.
Como la re-muestra es lo suficientemente grande y se puede aproximar a la media a una distribución normal por TCL, y como la varianza es desconocida, se realiza el cálculo del intervalo de confianza utilizando el estadístico t de la siguiente manera:

La media de cada re-muestra de tamaño n se calcula como:

$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Para n lo suficientemente grande, como los elementos de la muestra provienen de un vector aleatorio de variables independientes e idénticamente distribuidas, se puede aproximar $\overline{X}$ a una distribución normal por el Teorema Central del Límite, se tendrá $\overline{X} \sim_{aprox} N(\mu, \frac{\sigma}{\sqrt{n}})$, por lo que se estandariza:

$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

Se utiliza el estimador S^2 de la varianza, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{x})^2$ y se calcula el desvío $S = \sqrt{S^2}$ de cada re-muestra. Como el estadístico S^2 está condicionado al valor que toma $\overline{X}$, tiene n-1 grados de libertad, por lo que para la estimación de la media tendremos $\nu = n - 1$ grados de libertad. Se calcula el estadístico t como:

$t_\nu = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$

En este caso, se plantean intervalos de confianza del 90 % para la estimación de la media, por lo que se calcula como:

$P(t_{\nu=n-1;\frac{\alpha}{2}} \leq t_\nu = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}} \leq t_{\nu=n-1;1-\frac{\alpha}{2}}) = 1-\alpha$

$P(-t_{\nu=n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} + \overline{x} \leq \mu \leq t_{\nu=n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} + \overline{x}) = 1-\alpha$

$IC_{(1-\alpha)}(\mu) = [\overline{x} - t_{\nu=n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}; \overline{x} + t_{\nu=n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$

En este caso, $\alpha = 0{,}1$, por lo que el intervalo de confianza para cada una de las 100 muestras se calcula como:

$IC_{(0,9)}(\mu) = [\overline{x} - t_{\nu=n-1;0,95} \frac{s}{\sqrt{n}}; \overline{x} + t_{\nu=n-1;0,95} \frac{s}{\sqrt{n}}]$

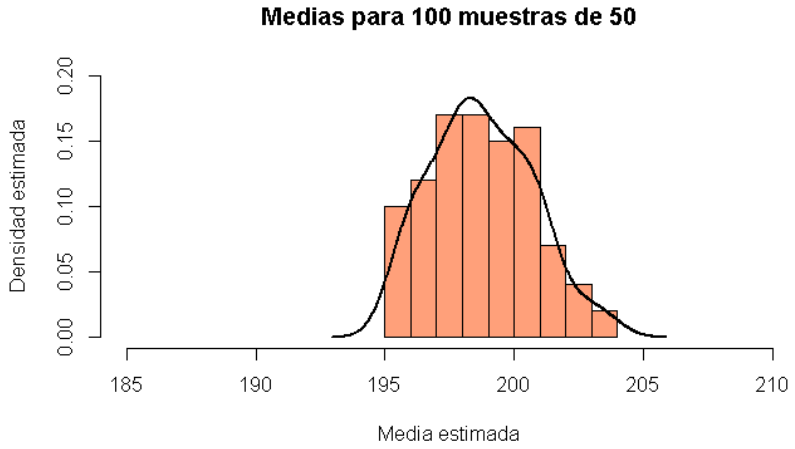
Donde $\overline{x}$ es el valor que toma la media para cada re-muestra, s es el valor que toma el desvío para cada re-muestra, $t_{\nu=n-1;0,975}$ es el fractil 0,975 de una t-student con $n-1$ grados de libertad y n el tamaño de cada re-muestra.

Por lo que, finalmente, se tendrán 100 valores aleatorios de medias $\{\overline{x}_1, \overline{x}_2, ..., \overline{x}_{100}\}$, donde los $\overline{x}_i$ son las medias calculadas para cada re-muestra, provienen de variables aleatorias independientes e idénticamente distribuidas $\overline{X}_i$, que se aproximan a una distribución normal por Teorema Central del Límite.

3.1. Bootstrapping para n=50

Se toman re-muestras de tamaño n y se realiza el procedimiento antes descripto.

Se puede observar que, por lo antes descripto, la distribución de las medias $\{\overline{x}_1, \overline{x}_2, ..., \overline{x}_{100}\}$, por ser n=50 un número lo suficientemente grande, se aproxima a una distribución normal por Teorema Central del Límite. Se construye el histograma de las medias de los 100 intervalos, como cada una de las medias de las re-muestras provienen de distribuciones normales idénticas aproximadas por TCL, se puede observar cómo la densidad tiende a la de una distribución normal:

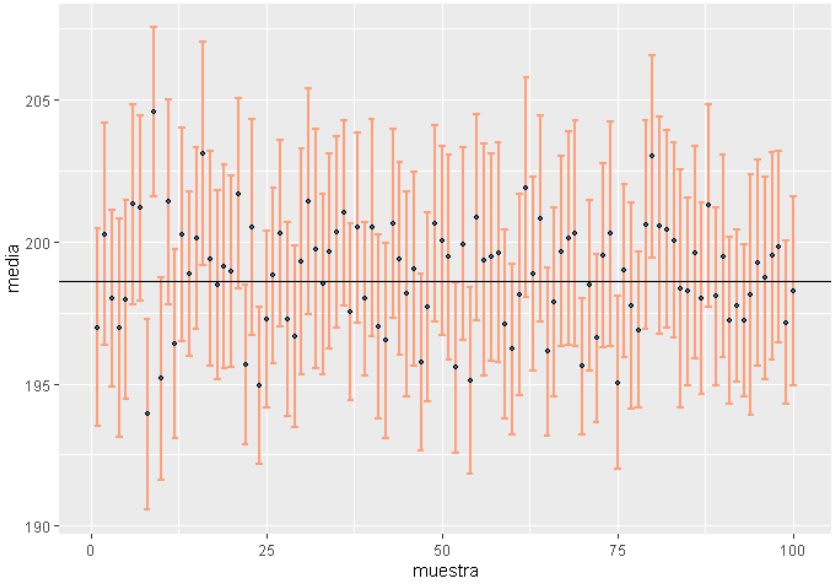


Se realiza el cálculo del desvío de las medias (S_{medias}) y se lo compara contra el desvío muestral de la muestra original sobre la raíz de n ($\frac{S}{\sqrt{50}}$), el cuál corresponde al desvío de la distribución normal aproximada por TCL:

$\hat{\beta} = S_{medias} = 1{,}975186$
$\hat{\theta} = \frac{S}{\sqrt{50}} = 2{,}038966$

Se puede observar que los desvíos son muy similares, debido a la aproximación realizada por TCL.

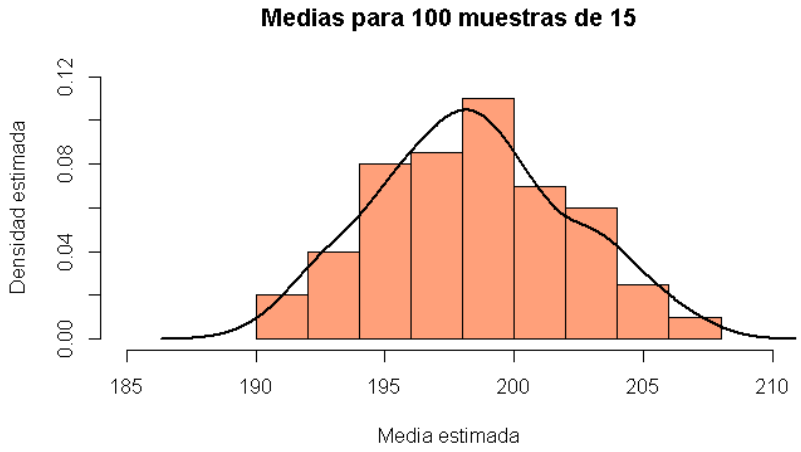
Se resume la información de los 100 intervalos de confianza en el siguiente gráfico, donde se marca con una línea horizontal la media muestral de los datos originales, de esa manera se calcula cuántos intervalos de confianza contienen a la media muestral:



De los 100 intervalos la cantidad que contienen a la media son 90, lo cual tiene sentido, dado que los intervalos tienen una confianza del 90 % de contener a la media muestral.

3.2. Bootstrapping para n=15

Se lleva a cabo el mismo procedimiento que para n=50, pero tomando re-muestras de 15 elementos.
Análogo a lo ocurrido con n=50, se observa que la distribución de las medias $\{\overline{x}_1, \overline{x}_2, ..., \overline{x}_{100}\}$, para n=15, se aproxima a una distribución normal por Teorema Central del Límite. Si bien n=15 no pareciera ser un número lo suficientemente grande, se realiza de todas formas esta aproximación; como los valores correspondientes a cada una de las 100 las medias provenientes de distribuciones idénticas son aproximadas a una normal por TCL, se puede observar cómo la densidad en el histograma tiende a esta distribución. A continuación, el histograma:



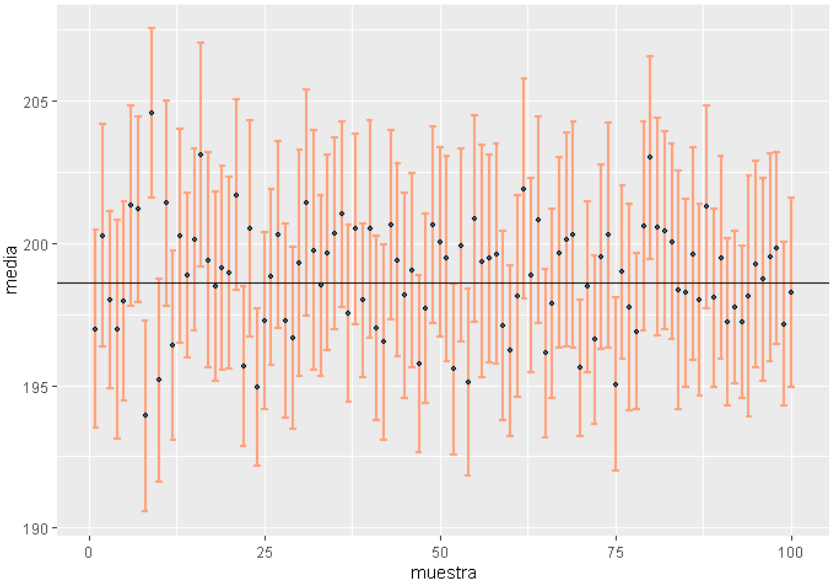
Se realiza la comparación del desvío de las medias S_{medias} contra el desvío muestral de la muestra original sobre la raíz de n ($\frac{S}{\sqrt{15}}$), el cuál corresponde al desvío de la distribución normal aproximada por TCL:

$\hat{\beta} = S_{medias} = 3{,}735876$
$\hat{\theta} = \frac{S}{\sqrt{15}} = 3{,}722625$

Se puede observar que los desvíos son altamente similares, debido a la aproximación realizada por TCL.

Una observación interesante es que, debido a que disminuyó la cantidad de muestras, dado que el desvío de las normales aproximadas se calcula como $\frac{\sigma}{\sqrt{n}}$, el desvío para n=15 es mayor que el desvío para n=50, esto se puede ver claramente en los cálculos y en el histograma, donde se aprecia que para n=15 la dispersión es mucho mayor.

Se resume la información de los 100 intervalos de confianza en el siguiente gráfico, donde se marca con una línea horizontal la media muestral de los datos originales, de esa manera se calcula cuántos intervalos de confianza contienen a la media muestral:



Se puede observar cómo los intervalos se encuentran más dispersos que en el gráfico de n=50, esto se debe a que el número de muestras tomado es mucho menor (analizado previamente en el histograma). De todas formas, de los 100 intervalos la cantidad que contienen a la media son 91, lo cual se aproxima bastante bien, dado que los intervalos tienen una confianza del 90 % de contener a la media muestral.

3.3. Conclusión

Para sacar una conclusión de los resultados obtenidos con ambos métodos, se recurre a la definición de *robustez estadística*.
La estadística robusta es una propuesta de métodos alternativos para buscar resultados que no sean afectados por valores atípicos. El hecho de que haya valores que varíen levemente de la hipótesis propuesta no afecta a la estimación.
En este caso, aún habiendo generado re-muestras de un tamaño mucho menor, se siguen superponiendo una cantidad muy próxima al 90 % de los intervalos de confianza, esto se debe a la robustez del método. El propósito de estas re-muestras es calcular intervalos de confianza más precisos en una situación en la que otra prueba estadística se hubiese basado en una distribución específica para los datos.

4. Anexos

Cuadro 1: Datos de la muestra

Registros de máxima presión									
197.3	192.1	189.3	222.2	209.5	200.8	194.5	190.6	207.6	236.9
204.1	193.6	225.4	192.7	232.6	183.5	182.3	187	217.3	198.4
199	181.8	181.4	206.7	205.8	187.8	188.1	202	201.2	184.1
200.8	195.4	194.3	213.7	212.9	189.1	175.9	189.7	200.5	188.8
220.7	195.9	200.3	214.9	203.7	197.3	187.6	188.5	200.2	205.6
208.3	178.4	176.5	192.9	192.3	196.6	194	184.4	188.1	213.3
182.9	178.8	197.7	238.4	176.1	215.4	178.6	197.1	180.1	206.2
189.1	182.6	201.8	188.4	201.1	194.1	192.3	198	200.7	227.4
203.1	229.8	207.6	184.1	212	193.5	193.1	211.4	180.8	195.7
201.1	202.8	215.2	189.7	222.3	212.5	176.3	230.1	194.2	181.7

```
#-----------TRABAJO PRACTICO E.A.I-----------

#----------PARTE A----------

datos<-as.matrix(datos[,1])

x<-as.call(datos)

rosa<-"lightsalmon"
azul<-"steelblue4"

n<-100

hist(datos,freq=FALSE,breaks=15,col=rosa,main='Densidad de X estimada',ylab='
    Densidad estimada',
     xlab = 'X: Registros de m xima  presin [Bar]',xlim=c(170,245))
lines(density(datos),lwd=3,col=azul)

#calculo la media y la varianza muestral
media_muestral<-mean(datos, na.rm=FALSE)
varianza_muestral<-var(datos)

#calculo el coeficiente de asimetra de fisher
library(moments)
coeficiente_asimetria_fisher<-skewness(datos)

#calculo los cuartiles
cuartiles<-quantile(datos, prob=c(0,0.25,0.5,0.75,1))

#realizo el boxplot
boxplot(datos, horizontal=TRUE,ylim=c(165,250),main='Boxplot de losregistros de
    m xima  presin [Bar]',
        col=rosa)

datos_ordenados<-sort(c(datos))

maximo<-max(c(datos))
minimo<-min(c(datos))

#realizo el boxplot sobre el histograma

layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(1,8))

par(mar=c(0, 3.1, 1.1, 2.1))
```

```
boxplot(datos , main="Boxplot e histograma",horizontal=TRUE , ylim=c(165,250), xaxt
    ="n" , col=azul , frame=F)
par(mar=c(4, 3.1, 1.1, 2.1))
hist(datos , breaks=15 , col=rosa  , main="" , xlab="Registros de m xima presin [
    Bar]",xlim=c(165,250))
abline(v=quantile(datos,0),lwd=3,col=azul)
abline(v=quantile(datos,0.25),lwd=3,col=azul)
abline(v=quantile(datos,0.5),lwd=3,col=azul)
abline(v=quantile(datos,0.75),lwd=3,col=azul)
abline(v=quantile(datos, 1),lwd=3,col=azul)

#estimo los parmetros de la lognormal

parametro_m<-((1/n)*sum(log(datos)))

parametro_D2<-((1/n)*sum((log(datos)-parametro_m)^(2)))

parametro_D<-sqrt(parametro_D2)

#calculo la logverosimilitud de la lognormal

logverosimilitud_ln<-sum(dlnorm(datos, meanlog = parametro_m, sdlog = parametro_D,
    log = TRUE))

#ploteo la densidad estimada y la densidad lognormal

hist(datos,freq=FALSE,breaks=15,col=rosa,main='Densidad de X estimada + Densidad
    Lognormal',
     ylab='Densidad estimada',xlab = 'X: Registros de m xima presin [Bar]',xlim=c
    (170,245),ylim=c(0,0.04))

curve(dlnorm(x, meanlog = parametro_m, sdlog = parametro_D, log = FALSE),lwd=3,add=
     TRUE, col=azul)

lines(density(datos),lwd=1.8,col='black')

#estimo los parmetros de la gumbel del mximo

parametro_beta<-((sqrt(6)/pi)*sqrt(varianza_muestral))

parametro_theta<-(media_muestral-0.5772157*parametro_beta)


#calculo la logverosimilitud de la gumbel del mximo

library("ordinal")

logverosimilitud_gumbel<-sum(dgumbel(datos, location = parametro_theta, scale =
    parametro_beta, log = TRUE, max = TRUE))

#ploteo la densidad estimada y la densidad gumbel del mximo

hist(datos,freq=FALSE,breaks=15,col=rosa, main='Densidad de X estimada + densidad
    Gumbel del Mximo',
     ylab='Densidad estimada', xlab = 'X: Registros de m xima presin',xlim=c
    (170,245),ylim=c(0,0.04))

curve(dgumbel(x, location = parametro_theta, scale = parametro_beta, log = FALSE,
              max = TRUE),lwd=3,add= TRUE, col=azul)

lines(density(datos),lwd=1.8,col='black')

#estimo los parmetros de la gamma

parametro_alpha<-(((media_muestral)^(2))/varianza_muestral)

parametro_gammabeta<-(varianza_muestral/media_muestral)
```

```
101
102 #calculo la logverosimilitud de la gamma
103
104 logverosimilitud_gamma<-sum(dgamma(datos, shape=parametro_alpha,
105                                    scale = parametro_gammabeta, log = TRUE))
106
107 #ploteo la densidad estimada y la densidad de la gamma
108 hist(datos,freq=FALSE,breaks=15,col=rosa,main='Densidad de X estimada + densidad
    Gamma',ylab='Densidad estimada',
109      xlab = 'X: Registros de  m xima  presin',xlim=c(170,245),ylim=c(0,0.04))
110
111 curve(dgamma(x, shape=parametro_alpha, scale = parametro_gammabeta, log=FALSE),lwd
    =3,add= TRUE,col=azul)
112
113 lines(density(datos),lwd=1.8,col='black')
114
115
116 #----------PARTE B----------
117 set.seed(190)
118 #MUESTRA DE 50
119 #calculo los intervalos de confianza de cada muestra y genero un data frame
120
121 t50<-qt(0.95,df=49)
122
123 simulaciones50<-list()
124 medias_simulaciones50<-c()
125 s_simulaciones50<-c()
126 error_simulaciones50<-c()
127 li_simulaciones50<-c()
128 ls_simulaciones50<-c()
129
130 for(i in 1:100){
131   simulacion50<-sample(c(datos), 50, replace = TRUE, prob = NULL)
132   simulaciones50[[i]]<-simulacion50
133   medias_simulaciones50[i]<-mean(simulacion50)
134   s_simulaciones50[i]<-sd(simulacion50)
135   error_simulaciones50[i]<-t50*(s_simulaciones50[i]/sqrt(50))
136   li_simulaciones50[i]<-medias_simulaciones50[i]-error_simulaciones50[i]
137   ls_simulaciones50[i]<-medias_simulaciones50[i]+error_simulaciones50[i]
138 }
139
140 int_conf_simulaciones50<-data.frame(muestra=c(1:100),
141                                     media=medias_simulaciones50,
142                                     lower=li_simulaciones50,
143                                     upper=ls_simulaciones50)
144
145 #ploteo los intervalos de confianza
146
147 library(ggplot2)
148 ggplot() +
149   geom_errorbar(data=int_conf_simulaciones50,mapping=aes(x=muestra,ymin=li_
      simulaciones50,ymax=ls_simulaciones50), width=1, size=1, color=rosa) +
150   geom_point(data=int_conf_simulaciones50, mapping=aes(x=muestra, y=media), size=1,
       shape=21, fill=azul)+
151   geom_hline(yintercept = media_muestral)
152
153 #calculo la proporcin de veces que aparece la media en las muestras
154
155 contador50<-0
156 for(i in 1:100){
157   if((media_muestral>=li_simulaciones50[i]&media_muestral<=ls_simulaciones50[i])){
158     contador50=contador50+1
159   }
160 }
161
162
163 #genero un histograma con las medias
```

```r
hist(medias_simulaciones50,freq=FALSE,col=rosa, main='Medias para 100 muestras de 50',
     ylab='Densidad estimada',xlab='Media estimada',ylim=c(0,0.20))
lines(density(medias_simulaciones50),lwd=2)


#MUESTRA DE 15
#calculo los intervalos de confianza de cada muestra y genero un data frame

t15<-qt(0.95,df=14)

simulaciones15<-list()
medias_simulaciones15<-c()
s_simulaciones15<-c()
error_simulaciones15<-c()
li_simulaciones15<-c()
ls_simulaciones15<-c()

for(i in 1:100){
  simulacion15<-sample(c(datos), 15, replace = TRUE, prob = NULL)
  simulaciones15[[i]]<-simulacion15
  medias_simulaciones15[i]<-mean(simulacion15)
  s_simulaciones15[i]<-sd(simulacion15)
  error_simulaciones15[i]<-t15*(s_simulaciones50[i]/sqrt(15))
  li_simulaciones15[i]<-medias_simulaciones15[i]-error_simulaciones15[i]
  ls_simulaciones15[i]<-medias_simulaciones15[i]+error_simulaciones15[i]
}

int_conf_simulaciones15<-data.frame(muestra=c(1:100),
                                    media=medias_simulaciones15,
                                    lower=li_simulaciones15,
                                    upper=ls_simulaciones15)

#ploteo los intervalos de confianza

library(ggplot2)
ggplot() +
  geom_errorbar(data=int_conf_simulaciones15,mapping=aes(x=muestra,ymin=li_
    simulaciones15, ymax=ls_simulaciones15), width=1, size=1, color=rosa) +
  geom_point(data=int_conf_simulaciones15, mapping=aes(x=muestra, y=media), size=1,
     shape=21, fill=azul)+
  geom_hline(yintercept = media_muestral)

#calculo la proporcin de veces que aparece la media en las muestras

contador15<-0
for(i in 1:100){
  if((media_muestral>=li_simulaciones15[i]&media_muestral<=ls_simulaciones15[i])){
    contador15=contador15+1
  }
}

#genero un histograma con las medias
hist(medias_simulaciones15,freq=FALSE,col=rosa, main='Medias para 100 muestras de 15',
     ylab='Densidad estimada',xlab='Media estimada')
lines(density(medias_simulaciones15),lwd=2)
```

Listing 1: Código en R del trabajo práctico