



# TRABAJO PRÁCTICO

Estadística Aplicada I [91.03] - 1C 2022
Curso lunes
**Tutora:** María de los Ángeles Sarquís

**Co-Tutor:** Miguel Remy Jung

**Alumnos**

Barreto, Ramiro Emilio - *105652*
rbarreto@fi.uba.ar

Breghelli, Francisco - *94696*
fbreghelli@fi.uba.ar

Cardozo, Iván - *105687*
icardozo@fi.uba.ar

Castro, Tomás Facundo - *105845*
tcastro@fi.uba.ar

Melone, Sebastián- *105603*
smelone@fi.uba.ar

Facultad de Ingeniería, Universidad de Buenos Aires

# Índice

# 1. Introducción

El presente informe reúne la documentación de la solución del Trabajo Práctico de Estadística Descriptiva e Inferencia Estadística de la materia Estadística Aplicada I.

La Inferencia Estadística es un conjunto de métodos y técnicas que permiten inducir, a partir de los datos empíricos proporcionados por una muestra, cuál es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad [1]. Para esto, manipulamos datos numéricos considerados valores observados de una o más variables aleatorias, realizando estimaciones puntuales o por intervalos y pruebas de hipótesis. Nuestra tarea es, aplicando procedimientos estadísticos, extraer conclusiones e interpretar los resultados obtenidos.

Para conocer características de una población, se extraen subconjuntos de la misma, denominadas muestras, las cuales deben ser obtenidas aleatoriamente para garantizar representatividad y no incurrir en sesgos de selección, ni de medición. La aleatoriedad de la muestra nos asegura, también, que sigue la misma distribución que la población.

En este Trabajo Práctico, los datos no existen todavía; y es nuestra responsabilidad indicar la forma más apta de generarlos. Solamente se sabe la distribución que siguen las poblaciones a estudiar.

Por lo expresado anteriormente, utilizando R, simulamos 1000 valores de cada variable aleatoria y las adoptamos como nuestras poblaciones. A partir de ellas, obtuvimos muestras representativas de tamaño 50, para luego estudiar su comportamiento mediante las distintas herramientas presentadas a continuación.

# 2. Simulación y muestreo

Si bien en la realidad no se conocen todos los datos de la población y es por ello que se realiza un muestreo, a fines prácticos de este trabajo simularemos dos variables aleatorias (VA) para obtener nuestras poblaciones. De acuerdo con lo pedido ambas debían tener el mismo tamaño, 1000, con sus respectivas distribuciones.

Siendo X la VA de distribución Normal e Y la VA de distribución Gamma, al ser el grupo 5:

$$X \sim \mathcal{N}(\mu = 500, \sigma = 20)$$

y en cuanto a la VA Y, su parámetro de escala es 100 y el de forma viene dado por la parte entera del promedio del último dígito de los padrones de los integrantes del grupo. Con lo cual, ya que:

$$\frac{2+6+7+5+3}{5} = 4{,}6,$$

entonces

$$\alpha = 4$$

y

$$Y \sim \Gamma(\alpha = 4, \beta = 100).$$

Simulando en R obtuvimos las poblaciones que se muestran en el archivo de Excel.

## 2.1. Muestras en poblaciones reales

Según su definición la población es el conjunto de datos sobre el cual queremos realizar un estudio; pueden ser finitas o infinitas. Es claro que en la realidad uno no conoce a la población en su totalidad, este es el motivo por el cual uno toma algún subconjunto de los elementos de la población bajo estudio (muestra) para poder sacar conclusiones del conjunto de datos total. Es importante definir cuidadosamente a la población objetivo para que no difiera de la población muestreada.

Una vez definida la población es fundamental que la muestra tomada sea representativa. Es decir, que sea de un tamaño adecuado y que se haya extraído mediante procesos aleatorios. Por lo tanto, las observaciones deben ser VA independientes, esto quiere que se deben considerar todas las observaciones tomadas. En todo caso, si se tiene un outlier, en vez de descartarla se la estudia para encontrar la causa de su valor extraño. Además, tienen que estar idénticamente distribuidas, implicando que no cambian las condiciones (esto es garantizado por la estabilidad del proceso y los instrumentos de medición). Todo esto llevaría que sus características reflejen el comportamiento de la población objeto de estudio.

Si no hubiéramos tenido ninguna información previa de los datos númericos, solo se podrían realizar inferencias válidas a partir de observaciones repetidas del experimento que genera tales resultados. Cabe aclarar que las conlusiones extraídas de este subconjunto serían válidas únicamente si fueron obtenidas aleatoriamente.

Obtener una muestra de dichas características implica evitar los factores que atentan contra la representatividad. El sesgo de selección, por ejemplo, se da cuando omitimos parte de la población objetivo (tanto por conveniencia, como por mal juicio o por error). Debemos también cuidarnos del sesgo de medición, que es consecuencia de los mecanismos o instrumentos por los cuales se obtiene la información (estos pueden inducir a un cierto valor por algún motivo).

# 3. Técnicas gráficas y métricas

Dada una población finita compuesta por N unidades experimentales, cada una de las cuales tiene asignado un valor numérico, $X_1, X_2, \ldots, X_N$, un muestreo aletorio consiste en elegir una muestra de n observaciones al azar entre las N. A partir de las poblaciones definidas en la sección 2, obtuvimos dos muestras, una de cada una, de tamaño 50. Las mismas se encuentran en el archivo de Excel.

A estos subconjuntos de datos les aplicaremos técnicas de Estadística Descriptiva para obtener conclusiones.

## 3.1. Métricas de tendencia central y de dispersión

Las principales métricas de tendencia central son la media, la mediana y la moda. Cada una de ellas busca representar en un único valor un concepto distinto del conjunto de valores.

La media, también conocida como esperanza, es el promedio de la variable aleatoria:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

La mediana es el valor de la variable que corresponde al fractil $x_{0,5}$, es decir, el que acumula 50 % de probabilidad:

$$F_X(mediana) = G_X(mediana) = \int_{-\infty}^{mediana} f_X(x) dx = 0,5$$

La moda, en cambio, es el valor alrededor del cual se tiene el intervalo con la mayor probabilidad de ocurrencia. Puede no existir o puede haber varias. Se la calcula encontrando el/los valor/es $\omega$ que maximice/n la función de densidad:

$$\text{máx}[f_X(x)] = \frac{df_X(\omega)}{dx} = 0$$

Finalmente, el desvío muestral, $\sigma$, mide la dispersión de los valores respecto del valor medio:

$$\sigma = \sqrt{E[(X - \mu)^2]}$$

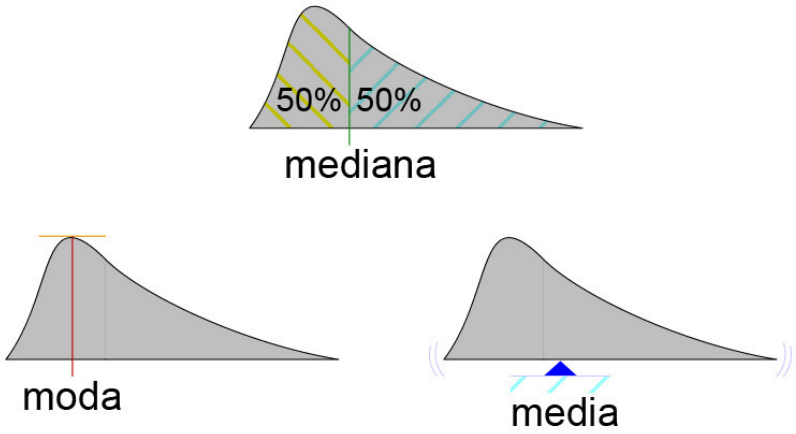


Figura 1: Principales métricas de tendencia central.

Ahora bien, la media muestral se calcula con el estadístico $\bar{X}$ que es un estimador de la media poblacional:

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

En cuanto al desvío muestral, se lo puede estimar según:

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

.

Nótese que, como la media es desconocida se pierde un grado de libertad (se dispone de n-1 grados de libertad).

Ambos estimadores, $S$ y $\bar{X}$ son consistentes, pues son insesgados y su varianza tienda a cero conforme n crece. Esto se puede demostrar fácilmente ya que las VA $X_1, X_2, \ldots, X_N$ son independientes y están idénticamente distribuidas.

Ordenando la muestra en forma creciente, la obtención de la mediana se reduce a:

$$x_{0,5} = \begin{cases} x_{k+1}, \text{ si } n = 2k+1 \text{ (impar)} \\ \frac{x_k + x_{k+1}}{2}, \text{ si } n = 2k \text{ (par)} \end{cases}$$

.

Para el cáculo de la moda recurrimos a los histogramas **3.2**, de ellos pudimos observar los intervalos con mayor frecuencia en cada muestra. Como se obtuvo la muestra de una VA continua, no íbamos a tener un valor que se repitiera más de una vez (justamente ese evento tiene probabilidad cero).

En base a lo explicado, obtuvimos los siguientes valores para las muestras en cuestión:

| **VALORES ESTIMADOS** | **VALORES REALES** |
|---|---|
| $\bar{x}_{muestral} = 496,99$ | $\mu = 500$ |
| $x_{0,5} = 498,51$ | $x_{0,5} = 500$ |
| moda $\in [490, 495]$ | moda $= 500$ |
| $\sigma_{muestral} = 17,551$ | $\sigma = 20$ |

Cuadro 1: Muestra de la población con distribución Normal.

| **VALORES ESTIMADOS** | **VALORES REALES** |
|---|---|
| $\bar{y}_{muestral} = 361,63$ | $\mu = 400$ |
| $y_{0,5} = 315,82$ | $y_{0,5} = 367,21$ |
| moda $\in [200, 250]$ | moda $= 300$ **A.1** |
| $\sigma_{muestral} = 195,89$ | $\sigma = 200$ |

Cuadro 2: Muestra de la población con distribución Gamma.

Los valores similares estimados son cercanos a los poblaciones en ambos casos. Puede verse además que de los valores muestrales ya se observa que $Moda \simeq Mediana \simeq Media$ para X y $Moda < Mediana < Media$ para Y, cumpliendo con sus respectivas simetrías y asimetrías.

### 3.1.1. Momentos centrales adimensionales

La utilización de momentos centrales adimensionales resulta de gran utilidad a la hora realizar comparaciones entre modelos con ciertas características. Al independizarnos de escalas y unidades, es posible obtener conclusiones del comportamiento de variables que pueden describir fenómenos muy distintos. De manera genérica, el momento central adimensional de orden k es:

$$\alpha_k = \frac{E[(x-\mu)^k]}{\sigma^k}$$

Las variables aleatorias suelen tener un sesgo o asimetría. El mismo muestra el desplazamiento que tienen los valores de la variable respecto del valor medio. Para medirlo se utiliza el momento central adimensional de orden 3, también conocido como **coeficiente de asimetría**:

$$\alpha_3 = \frac{E[(x-\mu)^3]}{\sigma^3}$$

Según el grado y tipo de asimetría que posea la distribución se obtendrán distintos valores del coeficiente:

$$\begin{cases} \alpha_3 < 0 \Rightarrow \text{asimetría negativa} \\ \text{simétrica} \Rightarrow \alpha_3 = 0 \text{ (no vale el recíproco)} \\ \alpha_3 > 0 \Rightarrow \text{asimetría postiva} \end{cases}$$

Una característica de interés para analizar y comparar distribuciones es el grado de apuntamiento de la función de densidad, o **curtosis**. Permite medir el grado de concentración que tienen los valores en la zona central de la distribución. Se utiliza el momento central adimensional de orden 4,

$$\alpha_4 = \frac{E[(x-\mu)^4]}{\sigma^4}$$

y según los valores que tome el mismo se puede clasificar en las siguientes categorías

$$\begin{cases} \alpha_4 > 3, & \text{leptocúrtica} \\ \alpha_4 = 3, & \text{mesocúrtica} \\ \alpha_4 < 3, & \text{platicúrtica} \end{cases}$$

Nuevamente, los valores estimados son relativamente similares a los poblacionales y cumplen con las condiciones esperadas ($\alpha_3 \simeq 0$ por ser simétrica, $\alpha_3 > 0$ por tener asimetría positiva). Como se aprecia en las siguientes tablas:

| VALORES ESTIMADOS | VALORES REALES |
|---|---|
| $\alpha_3 = -0,18796$ | $\alpha_3 = 0$ |
| $\alpha_4 = 2,5018$ | $\alpha_4 = 3$ |

Cuadro 3: Muestra de la población con distribución Normal.

| VALORES ESTIMADOS | VALORES REALES |
|---|---|
| $\alpha_3 = 1,1043$ | $\alpha_3 = 1$ |
| $\alpha_4 = 4,6223$ | $\alpha_4 = 4,5$ |

Cuadro 4: Muestra de la población con distribución Gamma.

## 3.2. El histograma de densidad de frecuencias

El histograma es una herramienta utilizada para estimar la función de densidad de variables aleatorias a partir de una muestra de datos. Estos datos se dividen en k clases o categorías buscando un nivel de síntesis que permita obtener información clara y representativa de la distribución en análisis. Se suele estimar k de varias maneras, entre las cuales se encuentran

$$\begin{cases} k = \sqrt{n} \\ k = 1 + \frac{ln(n)}{ln(2)} \text{ (fórmula de Sturges)} \end{cases}$$

siendo n la cantidad de observaciones.

Un valor k muy alto puede dificultar la identificación de patrones de comportamiento, mientras que uno muy bajo puede causar que la información tenga demasiada síntesis lo cual también complica la interpretación de la información.

Con respecto al ancho h de las clases, el mismo puede o no ser igual para todas. Una manera de obtenerlo es

$$h = \frac{\text{máx - mín}}{k}$$

en donde máx es el mayor valor observado y mín el menor.

Una vez que se definen las clases se obtienen las frecuencias absolutas , la cantidad de observaciones de cada clase, y las frecuencias relativas ($f_{relativas} = \frac{f_{absolutas}}{n}$). Nótese que se debe cumplir que la sumatoria de las frecuencias absolutas coincide con el número de observaciones mientras que las relativas dan 1:

$$\begin{cases} \sum_{i=1}^{k} f_i = n \\ \sum_{i=1}^{k} f_{r_i} = 1 \end{cases}$$

Finalmente, las alturas de cada clase resultan $\frac{f_{ri}}{h_i}$, con las cuales es posible armar el histograma.

Con el entorno R obtuvimos los respectivos histogramas de nuestras variables. Si bien esta cantidad de datos no resulta de un tamaño ideal para aproximar las densidades, aún así es posible observar las tendencias de las distribuciones.

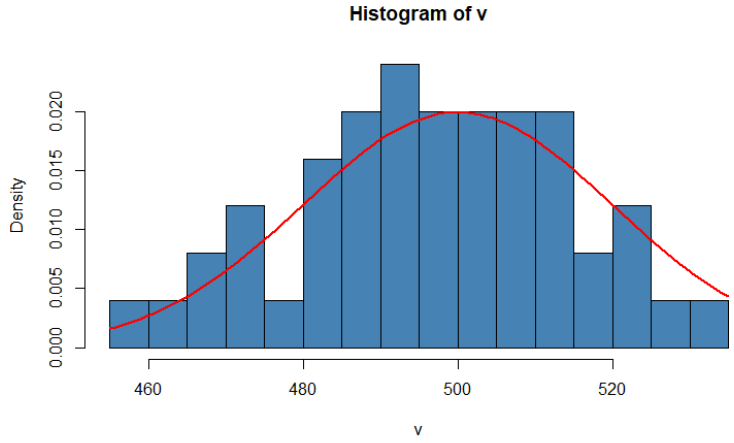


Figura 2: Histograma de la muestra de la población Normal.

Se puede observar la simetría de la variable, así como también el hecho de que la densidad es mesocúrtica, pues $\alpha_4 \simeq 3$.

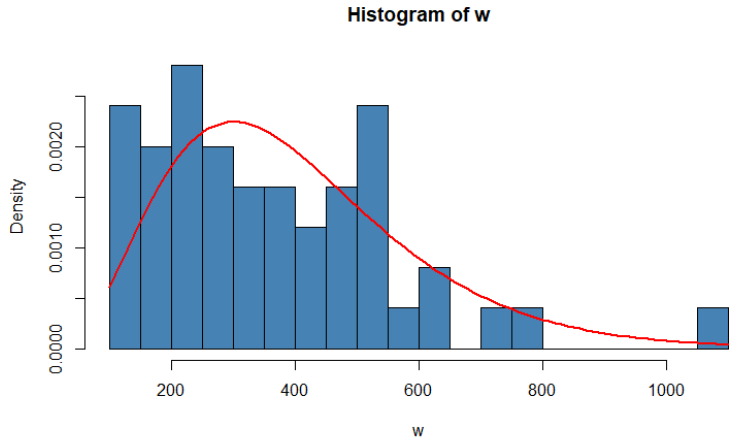


Figura 3: Histograma de la muestra de la población Gamma.

En este caso se distingue la asimetría positiva y además que la densidad es leptocúrtica, lo que es coherente con $\alpha_4 = 4,6223 > 3$

## 3.3. Box Plot

El Box Plot (también conocido como Diagrama de Cajas y Bigotes) es método estandarizado para representar gráficamente una serie de datos numéricos a través de sus cuartiles. Sirve principalmente para visualizar el centro, dispersión, grado y tipo de simetría (positiva o negativa) de la distribución de una variable como así también para identificar los outliers (puntos inusuales y/o extremos). Este se puede interpretar como la vista superior de un histograma.

A continuación, se muestra un ejemplo del mismo:

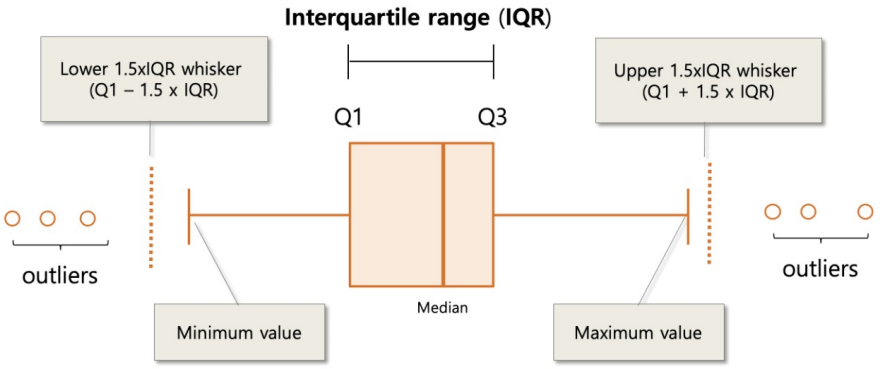


Figura 4: Box Plot genérico.

Los cuartiles $Q_1$, $Q_2$, $Q_3$ y $Q_4$ son fractiles que acumulan $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ y 1 de la frecuencia a su izquierda, respectivamente. El segundo cuartil ($Q_2$) corresponde con la mediana (valor de la varialeble que se encuentra en el centro, es decir que acumula 0,5 de probabilidad).

Luego, el rango intercuartílico (IQR) corresponde a la resta entre $Q_3$ y $Q_1$ y nos muestra donde se encuentra de forma "central" el 50 % de la muestra.

Cabe aclarar que la longitud de los bigotes de una idea de la dispersión de la muestra y el bigote más largo es el queda la asimetría. En este ejemplo, la distribución tendrá una leve asimetría positiva.

El mínimo valor que se muestra en el Box Plot corresponde al dato más chico contenido dentro del tramo ($Q_1 - 1,5 \cdot \text{IQR}$). Esto quiere decir que dicho valor no necesariamente es el mínimo valor de toda la muestra. Análogamente con el máximo.

En nuestro caso obtuvimos que:

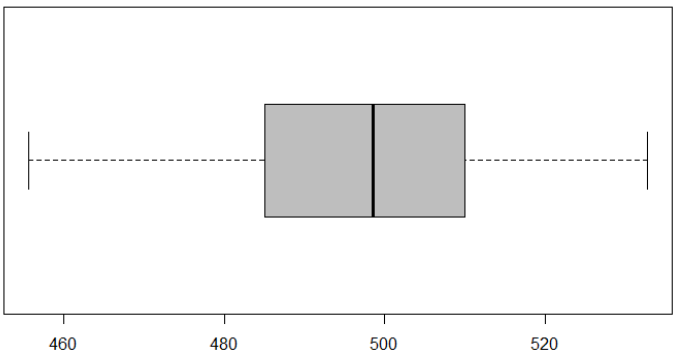


Figura 5: Box Plot de la muestra de población Normal.

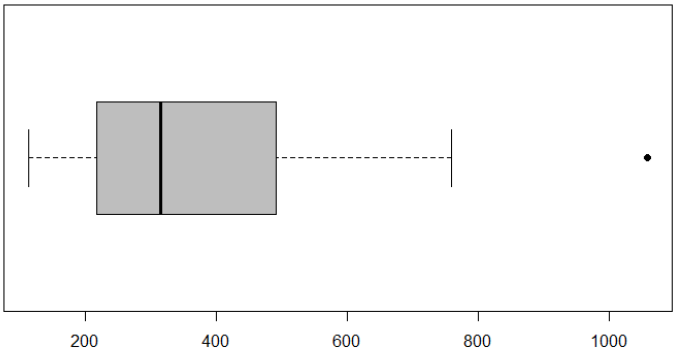


Figura 6: Box Plot de la muestra de población Gamma.

Tal como observamos en los cuadros, los valores estimados se asemejan en su gran mayoría a los valores reales. Se cumplen los valores esperados que dan forma y modelan las variables. Esto permite hacer un buen análisis y una buena aproximación en las situaciones donde no poseemos todos los valores reales. Al momento de graficar los boxplots, en las muestras de la población Normal, no observamos outliers y vemos que el bigote izquierdo es ligeramente más extenso que el derecho. En el boxplot de la gamma, se observa la asimetría positiva y un outlier.

## 3.4. Outliers

Cuando se trata con puntos extraños es fundamental no eliminarlos sin más, pues estos nos pueden aportar mucha información sobre la muestra. Es por esto que es muy importante identificarlos, y posteriormente estudiarlos. Se busca obtener el motivo de su valor particular y en base a ello obtener conclusiones.

## 3.5. Q-Q plot contra una distribución Normal

El Q-Q plot es una técnica gráfica con la cual es posible contrastar conjuntos de datos y analizar cuán cerca está la distribución de los mismos respecto a alguna distribución teórica. También se utiliza para comparar dos conjuntos de datos con el objetivo de determinar si los mismos provienen la misma población. En este trabajo se realizó la contrastación de las dos muestras contra una distribución Normal.

El procedimiento es el siguiente:

1. Se ordenan los datos de la muestra en forma creciente.
2. Se calculan las probabilidades acumuladas correspondientes a cada dato.
3. A partir de estas probabilidades se calculan los fractiles teóricos asociados a la distribución en cuestión.
4. Se grafican los fractiles teóricos vs los datos de la muestra.

Una vez hecho el gráfico será posible observar el comportamiento de los datos y determinar si los mismos provienen o no la distribución propuesta. Cuanto más alineados estén los puntos graficados, mas cerca están los datos de la muestra de provenir de esa distribución.

Para la primera muestra proveniente de la población Normal

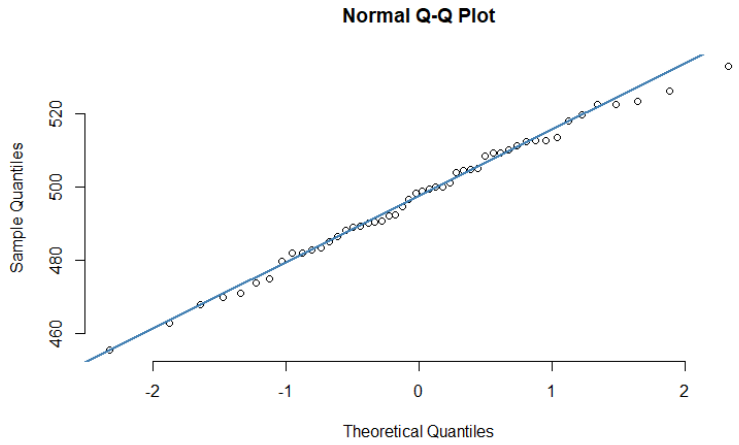


Figura 7: Q-Q plot de la muestra de población Normal.

Es posible observar que los datos están bien alineados, situación esperable puesto que la muestra fue tomada de la distribución Normal. Obviamente los puntos no coinciden completamente con la línea de tendencia por tener una muestra de tamaño 50.

Para la muestra proveniente de la población Gamma, vemos que el Q-Q plot es considerablemente distinto

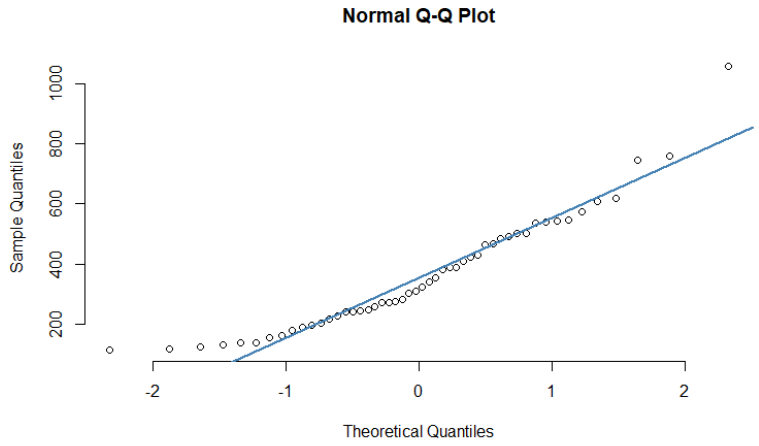


Figura 8: Q-Q plot de la muestra de población Gamma.

En este caso los datos ya no están del todo alineados, lo que implica que la muestra no proviene de una distribución Normal. Además se puede apreciar como se va separando cada vez más en las puntas dejando en evidencia su asimetría positiva.

## 3.6. Q-Q plots sin outliers

Para ambas situaciones los gráficos son prácticamente idénticos. Esto se debe a que los outliers identificados y removidos fueron muy pocos, y por lo tanto no se generó un cambio significativo en la muestra. Aún más, en la muestra normal no obtuvimos ningún valor extraño (para esta simulación, en otras simulaciones sí habíamos obtenido).

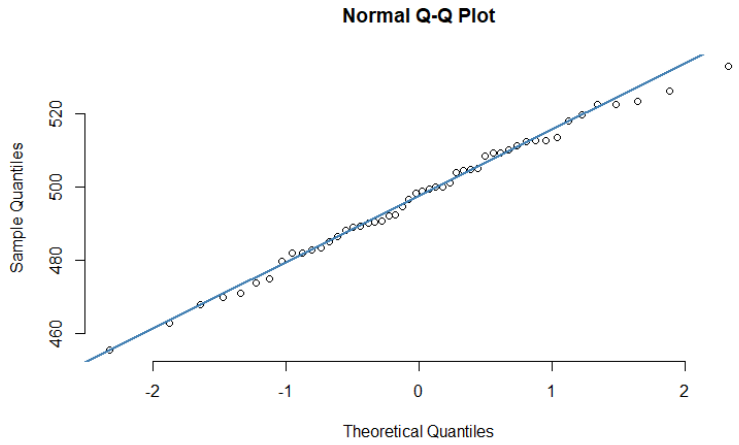


Figura 9: Q-Q plot sin outliers de la muestra de población Normal.

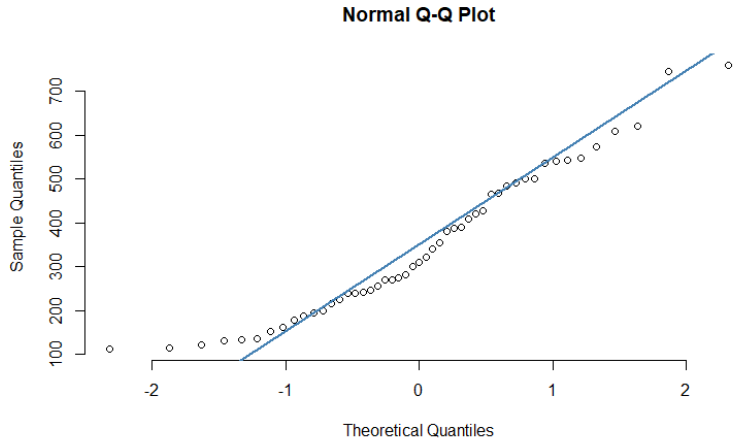


Figura 10: Q-Q plot sin outliers de la muestra de población Gamma.

### 3.7. Consistencia entre las poblaciones y las muestras

Como se mencionó previamente, los valores obtenidos de la muestra son coherentes con los valores poblaciones. Desde ya que esto es función de la muestra y habrá muestras en donde las estimaciones no sean tan buenas. Nosotros por ejemplo obtuvimos un coeficiente de curtósis de 2,5 para la muestra Normal cuando debía ser de 3. Nos llevamos que, si la muestra es representativa, aún para subconjuntos de datos que no parecen tan grandes ($n = 50$ en este caso), se pueden extrapolar conclusiones significativas para la población en general.

## 4. Teoría detrás de los Intervalos de Confianza

La Inferencia Estadística es una rama de la estadística que consiste en un proceso inductivo que se encarga de extraer conclusiones a partir de una muestra (subconjunto de la población) que luego se extrapola y generaliza para toda la población.

Cabe aclarar que el razonamiento inferencial es menos certero que uno deductivo. El primero puede tener errores y es por eso que trabajamos con niveles de confianza (NC). Con el nivel de confianza impuesto se busca minimizar la probabilidad de obtener conclusiones equivocadas.

Una de las ramas de la Inferencia Estadística es la Estimación, se pueden estimar tanto parámetros como distribuciones.

La Estimación de parámetros puede realizarse por 3 técnicas distintas:

- Estimación puntual
- Estimación por intervalo
- Test de hipótesis

En este informe abordaremos la estimación de parámetros únicamente por estimación puntual y por intervalo.

**Estimación Puntual**

Un estimador es un estadístico (una función de la muestra) usado para estimar un parámetro desconocido de la población.

En el desarrollo del informe, trabajamos con una variable aleatoria Normal. Su media muestral sigue distribución normal por la propiedad reproductiva; esta indica que la combinación lineal de variables aleatorias normales independientes siguen distribución Normal:

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

Luego, trabajaremos con una variable aleatoria Gamma. En este caso, su media muestral seguirá distribución Normal únicamente cuando tengamos una muestra de tamaño suficientemente grande ya que esto nos habilitaría a aplicar el Teorema Central del Limite (TCL).

El Teorema Central del Límite indica que, si S es una variable aleatoria que corresponde a la suma de n variables aleatorias independientes con varianza finita; cuando n es lo suficientemente grande (a fines prácticos $n \geq 30$) y no hay ninguna preponderancia de una de las VA o de su varianza respecto de las demás, la variable S converge a una distribución Normal.

En nuestro caso particular, las VA son independientes y están idénticamente distribuidas, luego nuestro estimador de la media muestral para la variable aleatoria Gamma es (con n grande):

$$\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

Al utilizar muestras de tamaño 5, esta aproximación justamente no es válida. En este caso, es erróneo decir que el estimador de la media muestral converge a una distribución Normal ya que se suman solo 5 variables aleatorias Gamma (cuando en rigor necesito, por lo menos, 30).

## 4.1. Intervalos de confianza para la media de poblaciones Normales

Una manera de realizar una estimación de un parámetro poblacional es mediante la técnica de intervalos de confianza. Al realizar una estimación puntual no es posible conocer el error que se está cometiendo, es así como mediante los intervalos de confianza se puede acotar dicho error, obteniendo el valor máximo y mínimo que puede tomar el parámetro que se está estimando, para un dado nivel de confianza. El nivel de confianza (NC) es la probabilidad con la que se desea que el intervalo contenga al verdadero valor del parámetro y $\alpha$ es el nivel de riesgo.

$$P(\widetilde{A} < \theta < \widetilde{B}) = NC = 1 - \alpha$$

Dado que la estimación es aleatoria, los límites y el error cometido también lo serán. Es así como para obtener los límites de confianza es necesario utilizar estadísticos que contengan en su expresión al parámetro que se desea estimar, pero que el mismo no aparezca en su distribución,

es decir, que no se necesite conocer el valor real para calcular probabilidades. A estos estadísticos se los conoce como pivotes.

Al utilizar el estimador $\bar{X}$, que como se ha mencionado sigue distribución normal, es posible realizar la estandarización para obtener el pivote y estimar la media con varianza conocida en una población Normal.

**Deducción**

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_1 - \frac{\alpha}{2}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq Z_1 - \frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$P(Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq Z_1 - \frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}} - \bar{X}) = 1 - \alpha$$

$$P(-Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} + \bar{X} \geq \mu \geq -Z_1 - \frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}} + \bar{X}) = 1 - \alpha$$

$$P(Z_1 - \frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}} + \bar{X} \geq \mu \geq Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} + \bar{X}) = 1 - \alpha$$

$$P(\bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_1 - \frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

### 4.2. Intervalos de Confianza para la media de poblaciones Normales con desvío desconocido

Para esta situación es necesaria la utilización de un estimador distinto que funcione como pivote, dado que al desconocerse la varianza esta deberá ser estimada como se realizó con la media, utilizando el estimador del desvió muestral ya mencionado previamente.

$$z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \Rightarrow t_v = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$$

$$P(t_v;\frac{\alpha}{2} \leq \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \leq t_v;1-\frac{\alpha}{2}) = 1 - \alpha$$

$$P(t_v;1-\frac{\alpha}{2} \cdot \frac{S}{\sqrt{n}} - \bar{X} \leq \mu \leq t_v;1-\frac{\alpha}{2} \cdot \frac{S}{\sqrt{n}} - \bar{X}) = 1 - \alpha$$

### 4.3. Estimación de la media en poblaciones no Normales

Si se desea estimar la media de una población con la técnica de intervalos de confianza será necesario realizar estimaciones y recurrir al Teorema Central del Límite para hacer uso de los pivotes recién mencionados. Para un tamaño de muestra suficientemente grande ($n > 30$), dado que el estimador $\bar{X}$ es una combinación lineal de variables aleatorias independientes con esperanza y varianza finita, el mismo tiende en distribución a una variable aleatoria normal. De esta manera es posible obtener los intervalos de confianza conociendo o no la varianza con los mismos pivotes que para las poblaciones normales.

## 5. Análisis de la población Normal

Los intervalos de confianza para la media (cuando el desvió es tanto conocido como desconocido) corresponden a un límite superior y uno inferior ambos centrados en la media muestral $\bar{X}$. Es por esto que suele expresarse como: $\bar{X} \pm EA$ en donde EA es el error absoluto y se calcula como: $EA = \frac{LS-LI}{2}$

Cabe destacar que tanto el límite superior como el inferior son funciones de la muestra ya que depende del $\bar{X}$ observado que depende de cada muestra tomada.

Acá se puede ver que para disminuir el error (es decir el ancho del intervalo de confianza) lo que debemos es aumentar el tamaño de muestra (que en la práctica puede implicar mayores costos, más tiempo dedicado al muestreo, entre otros factores). De esta forma conseguiremos un intervalo más preciso.

En su contrapartida, si queremos un nivel de confianza más elevado, y mantenemos el tamaño de muestra, aumentará el error absoluto por tener intervalos más anchos. Esto se puede apreciar aquí (como n no varía el desvío del estimador tampoco y por lo tanto la única opción es agrandar los intervalos):

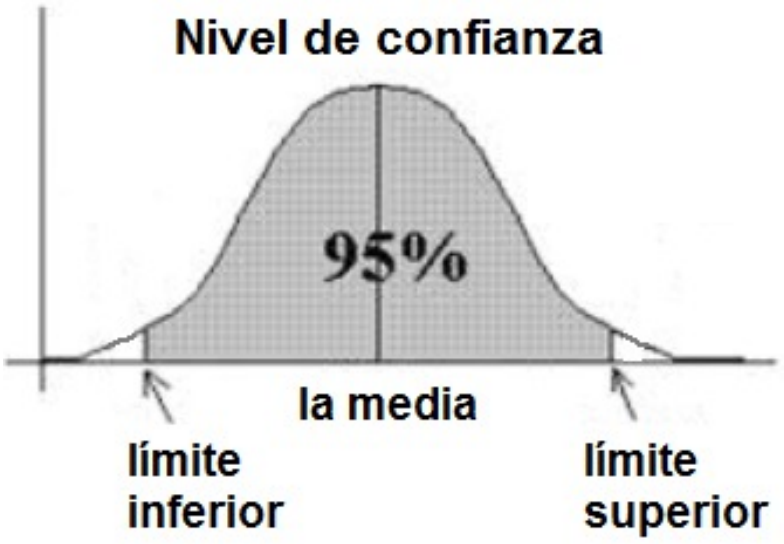


Figura 11: Nivel de confianza gráficamente.

El nivel de confianza representa el porcentaje de veces que el parámetro estimado estará comprendido dentro del intervalo de confianza hallado.

## 5.1. 100 muestras de tamaño 5 - Desvío conocido

En el código de programación, calculamos el límite superior e inferior de cada una de las 100 muestras de tamaño 5 de una distribución Normal con desvío conocido. El pivote utilizado es el siguiente:

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(\mu = 0, \sigma = 1)$$

A partir de los mismos, calculamos el nivel de confianza empírico, es decir, cuántos de estos intervalos contienen a la media poblacional y obtuvimos un valor de 0.91.

Es importante resaltar que el nivel de confianza empirico, al ser función de la muestra, puede ir variando de acuerdo a las muestras tomadas.

Además, utilizando las siguientes expresiones calculamos la longitud de cada intervalo, así como su error relativo:

$$\text{Longitud del intervalo} = LS - LI$$

$$\text{Error relativo} = \frac{EA}{\bar{X}}$$

De esta forma obtuvimos que la longitud promedio de los intervalos es de 35.061 y el error relativo promedio es 3,5007 %.

## 5.2. 100 muestras de tamaño 5 - Desvío desconocido

Luego, hicimos lo mismo, pero suponiendo que tanto la media como el desvío son desconocidos. Entonces a la media la estimamos con $\bar{X}$ y al desvío con el S utilizando la expresión para n-1 grados de libertad ya que la media no se conoce. Para esta oportunidad, el pivote utilizado es el siguiente:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{v=n-1}$$

En este caso, obtuvimos que el nivel de confianza empírico fue de 0.91, la longitud promedio es de 48.412 y el error relativo promedio es igual a 4,8316 %.

Es lógico que tanto la longitud promedio de los intervalos como el error relativo promedio aumenten respecto al caso donde el desvío era conocido ya que debimos utilizar la t de Student y esta distribución pierde cierta precisión con respecto a la Normal Estándar. Esto puede visualizarse en la siguiente gráfica.

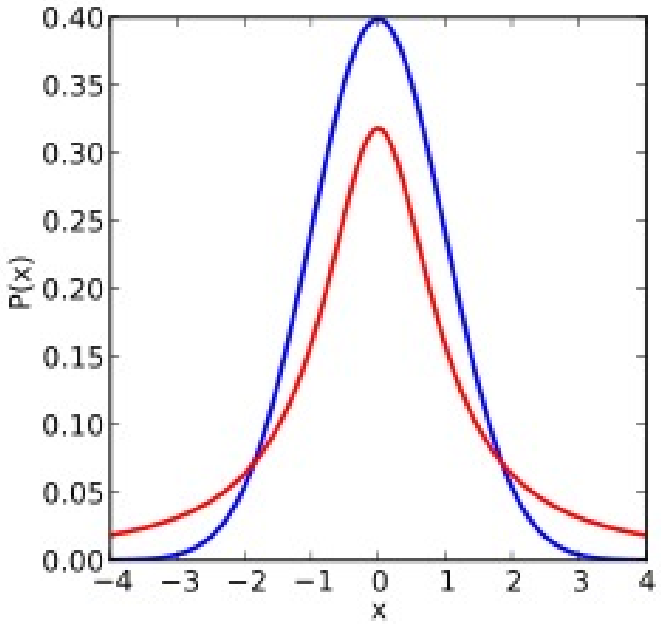


Figura 12: Comparación de la Normal estándar **azul** y la t de Student **roja**

Se puede ver que para un mismo nivel de confianza (visto como un área central dentro de la gráfica tal como se ejemplifico previamente), la Normal estándar presenta un intervalo más pequeño (ya que esta concentra más probabilidad en torno de la media) a diferencia de la T de Student.

## 5.3. 100 muestras de tamaño 30 - Desvío conocido

Ahora lo que hacemos en volver a repetir todo lo hecho pero cambiando el tamaño de muestra a 30. Entonces tendremos 100 muestras de tamaño 30 cada una.

De esta forma, considerando que el desvío es conocido obtuvimos un nivel de confianza empírico de 0.95, una longitud promedio del intervalo de 14.314 y un error relativo promedio es igual a 1,4305 %.

### 5.4. 100 muestras de tamaño 30 - Desvío desconocido

Luego, considerando que el desvío es desconocido obtuvimos un nivel de confianza empírico de 0.95, una longitud promedio del intervalo de 14.946 y un error relativo promedio igual a 1,4936 %. Nuevamente se cumple lo analizado. Al trabajar con la t de Student, puesto que desconocemos el desvío, la precisión disminuye haciendo que la longitud de los intervalos aumente al igual que su error relativo.

### 5.5. Comparación entre ambas muestras - Población Normal

Como era de esperar, tanto la longitud promedio de los intervalos como el error relativo promedio disminuyen al aumentar el tamaño de muestra, es decir que puede acotar a la media en intervalos más pequeños (tengo mayor precisión). Al aumentar en 25 unidades el tamaño de muestra (pasamos de tamaño 5 a tamaño 30), podemos notar las siguientes variaciones:

- Si el desvío es conocido (utilizando distribución Normal estándar)

$$\Delta\text{Longitud promedio} = \frac{Long_{n=30} - Long_{n=5}}{Long_{n=5}} \cdot 100 = \frac{14{,}314 - 35{,}061}{35{,}061} \cdot 100 = -59,174\,\%$$

$$\Delta\text{Error relativo promedio} = \frac{e.r._{n=30} - e.r._{n=5}}{e.r._{n=5}} \cdot 100 = \frac{1{,}4305\,\% - 3{,}5007\,\%}{3{,}5007\,\%} \cdot 100 = -59{,}137\,\%$$

- Si el desvío es desconocido (utilizando distribución t de Student)

$$\Delta\text{Longitud promedio} = \frac{Long_{n=30} - Long_{n=5}}{Long_{n=5}} \cdot 100 = \frac{14{,}946 - 48{,}412}{48{,}412} \cdot 100 = -69{,}127\,\%$$

$$\Delta\text{Error relativo promedio} = \frac{e.r._{n=30} - e.r._{n=5}}{e.r._{n=5}} \cdot 100 = \frac{1{,}4936\,\% - 4{,}8316\,\%}{4{,}8316\,\%} \cdot 100 = -69{,}087\,\%$$

En conclusión, y tal como predijimos en nuestro análisis cualitativo, al aumentar a uno 6 veces mayor se logra disminuir aproximadamente un 59 % tanto la longitud promedio de los intervalos como el error relativo promedio cuando el desvio es conocido. Cuando el desvío es desconocido, sucede algo similar pero una disminución aproximada del 69 %.

## 6. Análisis de la población Gamma

Para construir estos intervalos de confianza se mantuvo la suposición de que $\bar{X}$ sigue distribución normal con la aproximación por TCL. Fue posible observar, para los distintos tamaños de muestra, el error en el que se incurre en cada caso.

### 6.1. 100 muestras de tamaño 5 - Desvío conocido

Como se explicó previamente, aproximar por TCL es erróneo ya que, para poder aplicar dicha aproximación, necesito un número considerable de sumandos. En este caso, el nivel de confianza emperico es de 0.96, la longitud promedio de los intervalos es igual a 350.61 y el error relativo promedio vale 46.864 %

### 6.2. 100 muestras de tamaño 5 - Desvío desconocido

En este caso el nivel de confianza empírico obtenido fue de 0,93. La longitud promedio de los intervalos resultó de 474,39 y el error relativo promedio fue de 60,507 %.

### 6.3. 100 muestras de tamaño 30 - Desvío conocido

Al tomar 100 muestras de tamaño 30 de una población gamma con un desvío conocido, realizamos un analisis para hallar el nivel de confianza empírico, obteniendo 0.93 como resultado. La longitud promedio del intervalo, resultó ser de 143.14, presentando un error relativo promedio del 17.995 %.

### 6.4. 100 muestras de tamaño 30 - Desvío desconocido

En este caso el nivel de confianza empírico obtenido fue de 0,93. La longitud promedio de los intervalos resultó de 149.22 y el error relativo promedio fue de 18.585 %.

### 6.5. Comparación entre ambas muestras - Población Gamma

Puede verse el notable aumento en el error relativo promedio en comparación con los intervalos de la distribución Normal. Esto era esperable puesto que este tamaño de muestra no resulta lo suficientemente grande para que la aproximación por el Teorema Central del Límite sea aceptable.

Para el caso de un $n = 30$ la muestra resulta de un tamaño mayor, por lo que la aproximación por TCL es más adecuada lo cual puede verse reflejado en el hecho de que los errores relativos son significativamente menores que para $n = 5$. Sin embargo, siguen siendo valores bastante altos, en especial en comparación con los obtenidos anteriormente en donde $\bar{X}$ tenía exactamente distribución Normal por propiedad reproductiva.

Con respecto a los niveles de confianza empíricos, para todos los casos estos tuvieron valores similares, rondando el nivel de confianza propuesto teóricamente. No se esperaba que esto fuese distinto, puesto que los intervalos de confianza fueron armados con esa condición, que el 95 % de los mismos contuvieran al parámetro a estimar.

La principal diferencia en cada estimación radica en la longitud de los intervalos. Observando las expresiones de los límites de confianza, para tamaños de muestra chicos los intervalos son mas amplios ya que para garantizar el 95 % de confianza se deben establecer límites menos acotados. Asimismo, por lo explicado de la t de Student y la Normal, cuando el desvío es desconocido las longitudes de los intervalos y los errores relativos son mayores.

- Si el desvío es conocido (utilizando distribución Normal estándar)

$$\Delta\text{Longitud promedio} = \frac{Long_{n=30} - Long_{n=5}}{Long_{n=5}} \cdot 100 = \frac{143{,}14 - 350{,}61}{350{,}61} \cdot 100 = -59,174\,\%$$

$$\Delta\text{Error relativo promedio} = \frac{e.r._{n=30} - e.r._{n=5}}{e.r._{n=5}} \cdot 100 = \frac{17{,}995\,\% - 46{,}864\,\%}{46{,}864\,\%} \cdot 100 = -61{,}602\,\%$$

- Si el desvío es desconocido (utilizando distribución t de Student)

$$\Delta\text{Longitud promedio} = \frac{Long_{n=30} - Long_{n=5}}{Long_{n=5}} \cdot 100 = \frac{149{,}22 - 474{,}39}{474{,}39} \cdot 100 = -68{,}545\,\%$$

$$\Delta\text{Error relativo promedio} = \frac{e.r._{n=30} - e.r._{n=5}}{e.r._{n=5}} \cdot 100 = \frac{18{,}585\,\% - 60{,}507\,\%}{60{,}507\,\%} \cdot 100 = -69{,}285\,\%$$

## 7. Conclusión

Luego de realizar todo el informe, podemos notar que las técnicas gráficas y las métricas proporcionadas por la Estadística Descriptica son muy útiles para obtener una rápida idea o visualización sobre nuestra variable y, además, permite transmitir con facilidad y eficacia la información recopilada a otras personas.

Cabe destacar que en la mayoría de los casos prácticos no sabremos qué distribución siguen las distintas observaciones que tomemos y probablemente, estas no correspondan exactamente a alguna distribución. Lo que sí podremos hacer es estimar algunos de sus parámetros en base a las observaciones tomadas y en algunos casos, podremos decir que las observaciones tienden o siguen aproximadamente alguna distribución.

El trabajo además nos permitió tener un acercamiento a un caso práctico y nos permitió ver que, al aumentar el tamaño de la muestra, logramos aumentar la precisión de nuestros intervalos de confianza (achicando los intervalos). Sin embargo, no siempre podremos aumentar el tamaño de muestra tanto como queramos ya que en la práctica nos encontraremos con distintas limitaciones, costos, tiempos, etcétera.

En la vida profesional deberemos determinar cuán chico y con cuánto nivel de confianza eligiremos el intervalo teniendo en cuenta la relación entre costo y beneficio que ello implica. En adición, tendremos que aportar nuestro conocimiento y nuestra experiencia para decidir si tomamos un desvío conocido o no, las conclusiones que extraemos del Q-Q plot, cómo tomamos la muestra para asegurar que sea representativa, entre otras muchas cosas más.

# A. Anexo

## A.1. Moda de una VA con distribución Gamma

La función de densidad de una VA con distribución Gamma de parámetros $\alpha$ y $\beta$ es:

$$f_Y(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta},\ y > 0; \alpha, \beta > 0$$

.

Tomando *log* de $f_Y(y)$, nos queda

$$\log f(y) = \log\left(\frac{1}{\beta^\alpha \Gamma(\alpha)}\right) + (\alpha - 1)\log y - \frac{y}{\beta}.$$

.

Derivando $log(f_Y(y))$ e igualando a cero se obtiene

$$\begin{aligned} & \frac{d \log f(y)}{dy} = 0 \\ \Rightarrow \quad & 0 + \frac{\alpha - 1}{y} - \frac{1}{\beta} = 0 \\ \Rightarrow \quad & x = \beta(\alpha - 1). \end{aligned}$$

.

Además,

$$\frac{d^2 \log f(y)}{dy^2} = -\frac{(\alpha - 1)}{y^2} < 0.$$

Ergo, la densidad $f_Y(y)$ es máxima cuando $y = \beta(\alpha - 1)$. Por lo tanto, la moda de Y es $\beta(\alpha - 1) = 100 \cdot (4 - 1) = 300$.

# Referencias

[1] Roberto Mariano García. *Inferencia Estadística y Diseño de Experimentos.* Eudeba, 2004.