Universidad Nacional De Colombia
Facultad de Ciencias
Pregrado en Estadística

Parcial III de Inferencia Estadística (Parte 2)
Distribución de Ingresos por Género

9 de septiembre de 2024

Jorge Andrés Sánchez Duarte y Daniel Mauricio Vanegas Oliveros
josanchezdu@unal.edu.co dvanegaso@unal.edu.co

1. Introducción

Como se indica en la página instructiva sobre el caso de estudio, la base de datos 'créditos.txt' contiene información personal y financiera de personas y compañías a las que un banco muy reconocido otorgó algún tipo de crédito en Colombia durante 2017. El objetivo de este caso de estudio es revisar la existencia de alguna brecha de salarial o de ingresos entre hombres y mujeres a través de distintas técnicas (ajustes a distribuciones o bootstrap). Esta base de datos está anonimizada, así no tenemos identificación alguna; pero de todas formas, para el objetivo del estudio, se podrá prescindir de las mismas.

2. Análisis Exploratorio

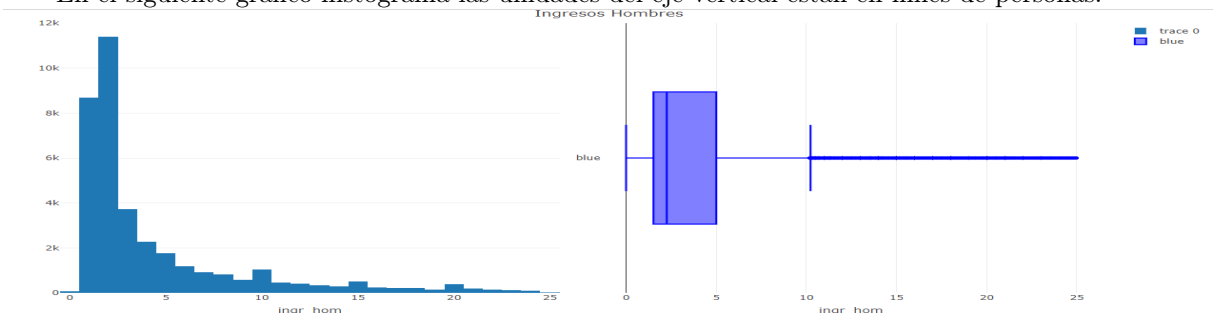
Previo a realizar análisis de inferencia, toca mostrar cómo se comporta el conjunto de datos. Para el análisis transformamos la escala de medición de los ingresos, dividiendo todos los valores por un 1000000.; por lo que, las medidas quedaron medidas en millones de pesos. Además, en el estudio se descartaron clientes con ingresos superiores a 25 millones.

2.1. Graficación

Para suplementar la exploración se adjuntan diagramas de tipo histograma y boxplot para ambas poblaciones donde el eje horizontal representa, en millones de pesos, los ingresos de las personas y para el diagrama de tipo histograma el eje vertical representa la cantidad de personas que pertenecen a cada intervalo de ingresos.

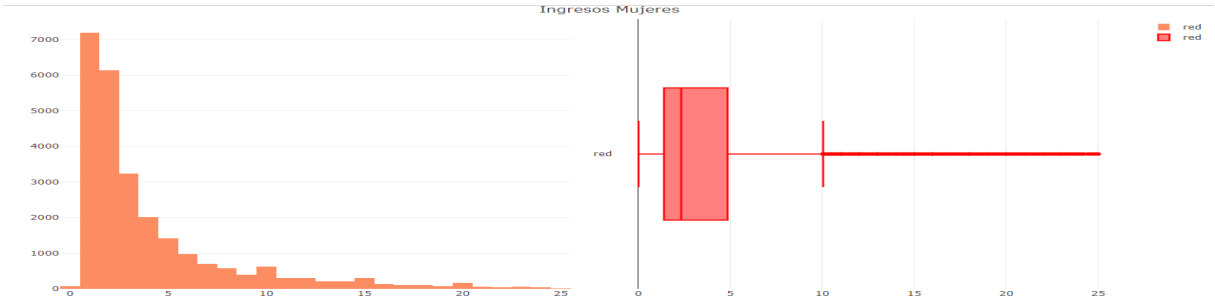
2.1.1. Histograma y Diagrama de Caja para Ingresos en Hombres

*En el siguiente gráfico histograma las unidades del eje vertical están en miles de personas.



Donde se alcanza a evidenciar, a partir del boxplot, que efectivamente existe una cantidad importante de atípicos. Tambien cabe a notar que la mayoría de los hombres tienen como ingresos aproximadamente 2 millones de pesos mensuales. También cabe a resaltar que la distribución de ingresos aparenta tener otras tres modas de 10, 15 y 20 millones.

2.1.2. Histograma y Diagrama de Caja para Ingresos en Mujeres



Se alcanza a ver que en este caso, el intervalo de ingresos más repetido en las mujeres es de 1 a 2 millones de pesos, correspondientes a aproximadamente 1.35 salarios mínimos del año en el que se realizó el muestreo

[?]. Se mantiene tambien, en menor medida respecto a los hombres, las múltiples modas que aparenta tener la distribución de los ingresos.

2.2. Tabla Resumen

A continuación mostramos una tabla donde se agrupa la información recogida sobre los ingresos entre hombres y mujeres encuestados (las mediciones se encuentran en millones de pesos):

	Media	Desv.Est.	Mín.	Cuartil 1	Mediana	Cuartil 3	Máx.	Coef.Var(%)
Hombres	4.31	4.73	0.003	1.51	2.26	5.00	24.93	1.10
Mujeres	3.99	4.22	0.02	1.40	2.33	4.86	24.99	1.06

2.3. Conclusiones de la Exploración

Es clara la fuerte presencia de datos atípicos, dado a que ambos histogramas y boxplots mostraron una muy notable acumulación de datos cercanos a 1 o 2 millones de pesos. Esto a su vez explica por qué la media y la mediana se muestran tan distintas en ambos conjuntos de datos.

Cabe a resaltar también que existen indicios de diferencia a favor de los hombres en cuanto a los ingresos comparados a las mujeres. Esta diferencia se presenta en ambos cuartil uno y cuartil dos (mediana) de manera ligera, donde a pesar de notarse, no es tan marcada como se muestra en el cuartil 3; además fijándonos en los máximos es visible que esta diferencia se mantiene en los valores más extremos. Así, se puede observar que para la mitad de los valores existen diferencias significativas entre los intresos entre hombres y mujeres, pero estas se disparan hasta casi el doble una vez alcanzado el 75 % de los datos revisados.

3. Distribución y Modelo Lognormal

3.1. Introducción

La distribución lognormal es comúnmente utilizada para modelar fenómenos naturales e ingresos (), así que será ideal modelar nuestro conjunto de datos con esta distribución. Para ponernos más técnicos esta distribución corresponde a la exponenciación de una distribución normal que también será tomada en cuenta posteriormente en este informe.

3.2. Parámetros Estimados

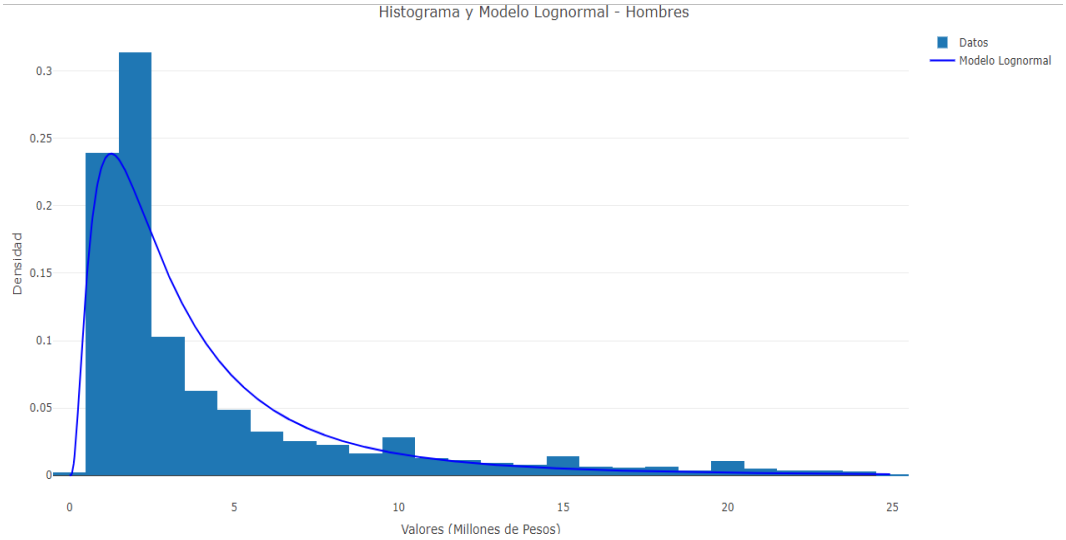
En caso de una revisión sobre los parámetros y su obtención mediante el método de la log-verosimilitud, consultar el primer anexo:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i) \quad \text{y} \quad \hat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(\ln(X_i)-\mu)^2$$

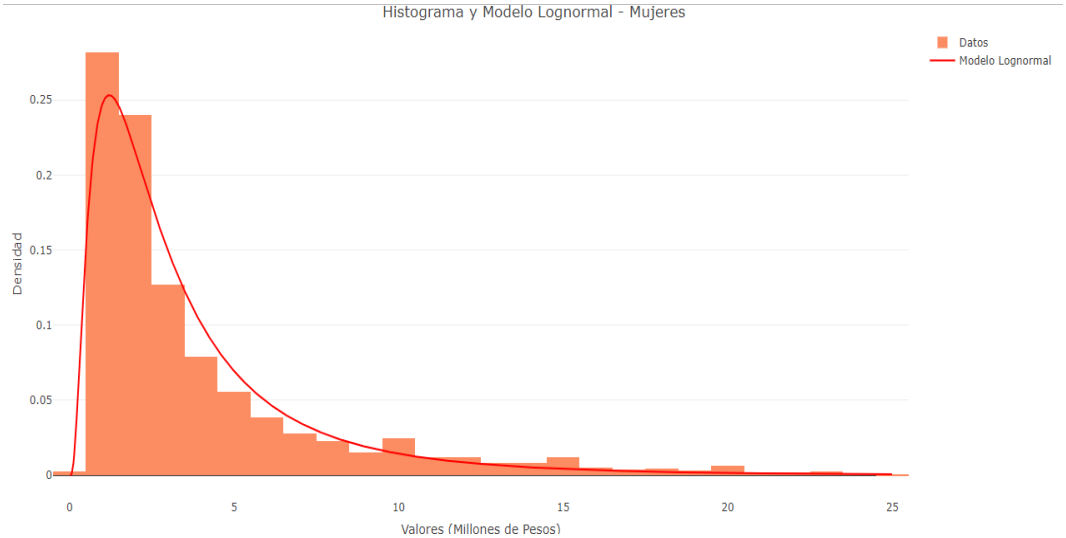
3.3. Ajuste a Datos

Aplicamos el modelo lognormal a los datos proporcionados. De esta manera mostraremos un histograma con el ajuste de distribución lognormal sobrepuesto.

3.3.1. Modelo Lognormal Ajustado a Hombres



3.3.2. Modelo Lognormal Ajustado a Mujeres



Así, queda claro que en este caso de estudio, tal cual como se viene comentando desde la introducción, la distribución lognormal logra ajustarse de manera correcta a los ingresos de ambas poblaciones. Por lo tanto, en la siguiente sección, procedemos a revisar las diferencias de promedio muestrales las cuales nos servirán para resolver los cuestionamientos y objetivos del estudio.

4. Inferencia sobre la Diferencia de Promedios

4.1. Análisis con Modelo Lognormal

Tomando en cuenta las expresiones 4 y 5 contenidas en el anexo 1.3; definimos a $\hat{\theta}_{\text{MLE}}$ como el promedio muestral esperado, que es calculado a partir del principio de invarianza de los MLE's ya calculados y mostrados con anterioridad.

$$\hat{\theta}_{\text{MLE}} = e^{\hat{\mu}_{\text{MLE}} + \frac{\hat{\sigma}^2_{\text{MLE}}}{2}}$$

con la fórmula correspondiente al intervalo de confianza de la diferencia:

$$\mathrm{IC}_{95\,\%}(\theta^{\mathrm{H}} - \theta^{\mathrm{M}}) = \left(\hat{\theta}^{\mathrm{H}}_{\mathrm{MLE}} - \hat{\theta}^{\mathrm{M}}_{\mathrm{MLE}}\right) \pm \mathrm{Z}_{0,975}\sqrt{\mathrm{Var}\left(\hat{\theta}^{\mathrm{H}}_{\mathrm{MLE}}\right) + \mathrm{Var}\left(\hat{\theta}^{\mathrm{M}}_{\mathrm{MLE}}\right)}$$

al computar los datos con tal de conseguir la estimación requerida conseguimos la siguiente tabla:

	Estimación	Desv.Est.	C.V	M.E.	IC Inf	IC Sup
Ajuste	0.252	0.034	13.53 %	0.067	0.185	0.319

Por lo que asumiendo la versatilidad en el uso de la distribución lognormal para hacer inferencia sobre los ingresos, se puede suponer, desde este modelo inicial, la existencia de una brecha en los ingresos de ambas poblaciones.

4.2. Análisis con Modelo Normal

La distribución normal es una de las más utilizadas generalmente en la estadística, esto por el teorema central del límite, el cual le brinda una grán versatilidad y aplio campo de aplicaciones.[Blanco, 2023] En caso de consulta sobre la obtención de intervalos, revisar el anexo 2, cabe a mencionar que, por haber hallado heterocedasticidad en el análisis explotatorio, usamos la estimación de diferencias a valores esperados bajo normalidad con varianzas desconocidas. Realizando el mismo procedimiento anterior notamos que la diferencia de promedios se logra evidenciar con la siguiente tabla:

	Estimación	Desv.Est.	C.V(%)	M.E.	IC Inf	IC Sup
Ajuste	0.320	0.036	11.3 %	0.071	0.249	0.391

Donde queda claro que en este caso, así sea tal vez uno de los métodos más simples de análisis por su sencilla aplicación, existen tambien diferencias significativas al analizar la aproximación mediante inferencia de promedios y su respectivo intervalo de confianza.

4.3. Análisis con Modelo Boostrap

El bootstrap es un algoritmo estadístico enfocado en la creación de intervalos de confianza sin necesidad de hacer asunciones previas sobre el comportamiento de los datos o el ajuste a un modelo específico. Consiste en hacer remuestreo sobre la muestra original[Joseph, 2023]. Al ser una revisión sobre la diferencia de promedios, vamos a realizar una estimación de diferencias a partir de los vectores de promedios, que son imprescindibles a la hora de realizar el algoritmo, para así poder aplicar este método al estudio. A continuación colocamos los intervalos de confianza obtenidos en los tres métodos de bootstrap más comunes: (para verificar cada uno con su formulación revisar en [Sosa, 2024]) o en el tercer apéndice.

4.3.1. Bootstrap por Percentiles

	Estimación	Desv.Est.	C.V	M.E.	IC Inf	IC Sup
Bootstrap por Percentiles	0.320	0.036	11.3 %	0.070	0.248	0.390

4.3.2. Bootstrap Empírico

	Estimación	Desv.Est.	C.V	M.E.	IC Inf	IC Sup
Bootstrap Empírico	0.320	0.036	11.3 %	0.073	0.251	0.393

4.3.3. Bootstrap por Normalidad

	Estimación	Desv.Est.	C.V	M.E.	IC Inf	IC Sup
Bootstrap por Normalidad	0.320	0.036	11.3 %	0.071	0.249	0.392

5. Resumen y Conclusiones

Concluimos con la siguiente tabla mostrando todos los ajustes hechos a la diferencia de promedios durante el estudio:

Ajustes	Estimación	Desv.Est.	C.V	M.E.	IC Inf	IC Sup
Lognormal	0.252	0.034	13.53 %	0.067	0.185	0.319
Normal	0.320	0.036	11.31 %	0.071	0.249	0.391
Bootstrap Percentiles	0.320	0.036	11.30 %	0.070	0.248	0.390
Bootstrap Empírico	0.320	0.036	11.30 %	0.073	0.251	0.393
Bootstrap Normal	0.320	0.036	11.30 %	0.071	0.249	0.392

Es claro que así los cinco métodos muestren la existencia de una brecha de ingresos o salarial, toca resaltar la fuerte diferencia en las estimaciones del modelo lognormal con respecto a los demás. Hay que tomar en cuenta una cosa; todos estos datos de ingresos claramente son de sesgo positivo, y ambas estimaciones normal y bootstrap asumen la existencia de valores negativos, lo que puede dar lugar a que haya comportamientos incorrectos en las estimaciones. Digamos, en el ejemplo de bootstrap de [Sosa, 2024] ocurre que se hace simulación de una distribución chi cuadrado y bootstrap, así sea aumentanto el tamaño de la muestra simulada, estima un valor esperado por encima del valor real. Posiblemente ocurrió lo mismo aquí dado a que el único ajuste a una distribución de sesgo positivo fue la lognormal. Así, asumimos que el ajuste lognormal es el más realista de todos, además que, como se aclaraba desde su introducción en [Pavlovic, 2024], esta distribución es muy usada para este uso particular.

De esta manera se puede declarar que, a partir de la información proporcionada en el caso de estudio, con 95 % de confiabilidad en cinco modelos de inferencia distintos basados en intervalos de confianza, y con evidencia empírica, que existe una clara diferencia significativa entre los ingresos de hombres y mujeres, a favor de los hombres, a los que se les ha otorgado algún tipo de crédito por parte del banco consultado.

Anexos

1 Intervalos por Distribución LogNormal

1.1 Estimación Puntual de Parámetros

Por definición, se tiene que para $X \sim \text{LogNormal}(\mu, \sigma^2)$; $f_X(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\ln(x)-\mu)^2}$. De esta manera, definimos la verosimilitud $\text{L}(\mu, \sigma^2)$ así:

$$\begin{aligned}\text{L}(\mu, \sigma^2) &= \prod_{i=1}^{n} f_{X_i}(X_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{X_i\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\ln(X_i)-\mu)^2} \\ &= \frac{1}{\left(\prod_{i=1}^{n} X_i\right)\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln(X_i)-\mu)^2}\end{aligned}$$

Así, la log-verosimilitud es:

$$\begin{aligned}\ell(\mu, \sigma^2) = \ln(\text{L}(\mu, \sigma^2)) &= \ln\left(\frac{1}{\left(\prod_{i=1}^{n} X_i\right)\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln(X)-\mu)^2}\right) \\ &= \ln\left(\frac{1}{\sqrt{2\pi}\prod_{i=1}^{n} X_i}\right) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln(X_i)-\mu)^2\end{aligned}$$

• Para el cálculo de $\hat{\mu}_{\text{MLE}}$, maximizamos la función de log-verosimilitud $\ell(\mu, \sigma^2)$:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{\sum_{i=1}^{n}\ln(X_i)}{\sigma^2} - \frac{n\mu}{\sigma^2} \quad \therefore \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$$

Veamos que para $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$, usando el criterio de segunda derivada;

$$\frac{\partial^2 \ell(\mu, \sigma^2)}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

Por lo que, al tenerse definida $\sigma^2 > 0$, se concluye que $\hat{\mu}$ es un máximo local. Al verificarse los límites de frontera:

$$\lim_{\mu\to-\infty} \text{L}(\mu, \sigma^2) = -\infty \quad \text{y} \quad \lim_{\mu\to\infty} \text{L}(\mu, \sigma^2) = -\infty$$

Porque dentro de la función de verosimilitud solo hay diferencias de cuadrados. Por lo tanto, se concluye que $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$ es un máximo total. Por lo que:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\ln(X_i) \tag{1}$$

• Ahora, para hallar $\hat{\sigma}^2_{\text{MLE}}$ se aplica nuevamente el mismo proceso de maximización, de tal manera que:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(\ln(X_i)-\mu)^2}{2(\sigma^2)^2} \quad \therefore \quad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(\ln(X_i)-\mu)^2$$

Posteriormente, calculamos la segunda derivada parcial respecto a σ^2, de tal manera que

$$\frac{\partial^2 \ell(\mu, \sigma^2)}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^{n}(\ln(X_i)-\mu)^2}{(\sigma^2)^3}$$

De esta manera, evaluando $\hat{\sigma}^2$ en la segunda derivada:

$$\left.\frac{\partial^2 \ell(\mu, \sigma^2)}{\partial (\sigma^2)^2}\right|_{\sigma^2=\hat{\sigma}^2} = \left.\left(\frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^{n}(\ln(X_i)-\mu)^2}{(\sigma^2)^3}\right)\right|_{\sigma^2=\hat{\sigma}^2} = \frac{n}{2(\hat{\sigma}^2)} - \frac{1}{(\hat{\sigma}^2)^3}n(\hat{\sigma}^2) = -\frac{2n}{(\hat{\sigma}^2)^2}$$

Y al ser $\hat{\sigma}^2$ una suma de cuadrados, se tiene que $\hat{\sigma}^2 < 0$ y por ende, es un máximo local. Al comprobarse los límites de frontera:

$$\lim_{\sigma^2 \to 0^+} \mathrm{L}(\mu, \sigma^2) = 0 \quad \text{y} \quad \lim_{\sigma^2 \to \infty} \mathrm{L}(\mu, \sigma^2) = 0$$

Así, queda demostrado que $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(\ln(X_i) - \mu)^2$ es un máximo total, por lo que:

$$\hat{\sigma^2}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(\ln(X_i) - \mu)^2 \tag{2}$$

1.2 Información de Fisher

Para hallar la información esperada de fisher usaremos el siguiente teorema:

$$I_n = -E\left(\frac{\partial^2}{\partial\theta^2}\ln\prod_{i=1}^{n} f_X(X_i;\theta)\right) = -E\left(\frac{\partial^2 \ell(\theta)}{\partial\theta^2}\right)$$

De tal manera que para la información de fisher correspondiente a μ:

$$I_{\mu,n} = -E\left(\frac{\partial^2 \ell(\mu,\sigma^2)}{\partial\mu^2}\right) = -E\left(-\frac{n}{\sigma^2}\right)$$

$$I_{\mu,n} = \frac{n}{\sigma^2}$$

Ahora, para la información de fisher correspondiente a σ^2, de manera previa recordemos que por definición, para una variable aleatoria Y_i con distribución $\mathcal{N}(\mu,\sigma^2)$, se define $X_i = e^{Y_i}$ con distribución LogNormal(μ,σ^2). Por lo que para X se tiene que $\ln(X_i) = \ln(e^{Y_i}) = Y_i$. De esta manera vemos que, asumiendo independencia entre las mediciones de la muestra:

$$\begin{aligned}
I_{\sigma^2,n} = -E\left(\frac{\partial^2 \ell(\mu,\sigma^2)}{\partial(\sigma^2)^2}\right) &= -E\left(\frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^{n}(\ln(X_i)-\mu)^2}{(\sigma^2)^3}\right) \\
&= -\frac{n}{(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\left(\sum_{i=1}^{n} E\left(\ln(X_i)^2\right) - 2\mu\sum_{i=1}^{n} E\left(\ln(X_i)\right) + n\mu^2\right) \\
&= -\frac{n}{(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\left(\sum_{i=1}^{n} E\left(Y_i^2\right) - 2\mu\sum_{i=1}^{n} E\left(Y_i\right) + n\mu^2\right) \\
&= -\frac{n}{(\sigma^2)^2} + \frac{1}{(\sigma^2)^3}\left(n\mu^2 + n\sigma^2 - 2n\mu^2 + n\mu^2\right)
\end{aligned}$$

$$I_{\sigma^2,n} = \frac{n}{2(\sigma^2)^2}$$

De esta manera, la información esperada de fisher queda en la siguiente matriz a partir de (3) y (4):

$$I_n = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{bmatrix}$$

Y posteriormente para hallar la información observada de fisher, evaluamos en la matriz los el MLE de σ^2:

$$\hat{I}_n = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{bmatrix}\Bigg|_{\sigma^2=\hat{\sigma}^2_{\mathrm{MLE}}} = \begin{bmatrix} \frac{n}{\hat{\sigma}^2_{\mathrm{MLE}}} & 0 \\ 0 & \frac{n}{2(\hat{\sigma}^2_{\mathrm{MLE}})^2} \end{bmatrix} \quad \therefore \quad \hat{I}_n^{-1} = \begin{bmatrix} \frac{\hat{\sigma}^2_{\mathrm{MLE}}}{n} & 0 \\ 0 & \frac{2(\hat{\sigma}^2_{\mathrm{MLE}})^2}{n} \end{bmatrix} \tag{3}$$

1.3 Estimación de Intervalos

Para hallar los estimadores de la diferencia de promedios en la lognormal, procedemos a utilizar los MLE's ya encontrados, su respectivo principio de invarianza, y el método delta, con $g(\mu, \sigma^2) = e^{\mu+\frac{\sigma^2}{2}}$:

$$E(X) = \theta = e^{\mu+\frac{\sigma^2}{2}} \quad \therefore \quad g(\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}}) = \hat{\theta}_{\text{MLE}} = e^{\hat{\mu}_{\text{MLE}}+\frac{\hat{\sigma}^2_{\text{MLE}}}{2}} \tag{4}$$

Luego, aplicando la definición del método delta:

$$\nabla(g(\mu, \sigma^2)) = \nabla(e^{\mu+\frac{\sigma^2}{2}}) = \begin{bmatrix} \frac{\partial g}{\partial \mu} \\ \frac{\partial g}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} e^{\mu+\frac{\sigma^2}{2}} \\ \frac{1}{2}e^{\mu+\frac{\sigma^2}{2}} \end{bmatrix} \quad \therefore \quad \nabla(\hat{\theta}_{\text{MLE}}) = \hat{\theta}_{\text{MLE}} \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}$$

Así, por definición; como la varianza del promedio a partir del método delta se define así,

$$\text{Var}(\hat{\theta}_{\text{MLE}}) = \nabla(g)^T \hat{I}^{-1} \nabla(g)$$

se tiene que:

$$\text{Var}(\hat{\theta}_{\text{MLE}}) = (\hat{\theta}_{\text{MLE}}) \begin{bmatrix} 1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{\hat{\sigma}^2_{\text{MLE}}}{n} & 0 \\ 0 & \frac{2(\hat{\sigma}^2_{\text{MLE}})^2}{n} \end{bmatrix} (\hat{\theta}_{\text{MLE}}) \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}$$

$$\text{Var}(\hat{\theta}_{\text{MLE}}) = \frac{1}{n}(\hat{\theta}_{\text{MLE}})^2 \left(\hat{\sigma}^2_{\text{MLE}} + \frac{1}{2}(\hat{\sigma}^2_{\text{MLE}}) \right)$$

Por lo que, de esta manera, se concluye que el promedio muestral se distribuye de la siguiente manera:

$$\frac{\hat{\theta}_{\text{MLE}} - \theta}{\sqrt{\text{Var}(\hat{\theta}_{\text{MLE}})}} \xrightarrow{\text{d}} \mathsf{Normal}\,(0, 1)$$

Ahora, como queremos calcular el valor de la diferencia entre promedios $\hat{\theta}^{\text{H}}_{\text{MLE}} - \hat{\theta}^{\text{M}}_{\text{MLE}}$ de ambas poblaciones muestreadas, definimos de la siguiente manera:

$$\text{ME} = z_{1-\alpha/2} \sqrt{\text{Var}\left(\hat{\theta}^{\text{H}}_{\text{MLE}} - \hat{\theta}^{\text{M}}_{\text{MLE}} \right)}$$

Pero como $\hat{\theta}^H$ y $\hat{\theta}^M$ son poblaciones independientes:

$$\text{Var}\left(\hat{\theta}^{\text{H}}_{\text{MLE}} - \hat{\theta}^{\text{M}}_{\text{MLE}} \right) = \text{Var}\left(\hat{\theta}^{\text{H}}_{\text{MLE}} \right) + \text{Var}\left(\hat{\theta}^{\text{M}}_{\text{MLE}} \right)$$

Por lo que de esta manera, se podría describir por intervalo, de manera asintótica, el estumador de $\hat{\theta}^H - \hat{\theta}^M$ con la siguiente expresión:

$$\frac{\left(\hat{\theta}^{\text{H}}_{\text{MLE}} - \hat{\theta}^{\text{M}}_{\text{MLE}} \right) - (\theta^{\text{H}} - \theta^{\text{M}})}{\sqrt{\text{Var}\left(\hat{\theta}^{\text{H}}_{\text{MLE}} \right) + \text{Var}\left(\hat{\theta}^{\text{M}}_{\text{MLE}} \right)}} \xrightarrow{\text{d}} \mathsf{Normal}\,(0, 1)$$

Así, procedemos a calcular los intervalos con un intervalo de confianza del 95 %:

$$\text{IC}_{95\,\%}(\theta^{\text{H}} - \theta^{\text{M}}) = \left(\hat{\theta}^{\text{H}}_{\text{MLE}} - \hat{\theta}^{\text{M}}_{\text{MLE}} \right) \pm \text{Z}_{0{,}975} \sqrt{\text{Var}\left(\hat{\theta}^{\text{H}}_{\text{MLE}} \right) + \text{Var}\left(\hat{\theta}^{\text{M}}_{\text{MLE}} \right)} \tag{5}$$

2 Intervalos por Distribución Normal

Para dar con los intervalos usando este modelo, citamos directamente la página del curso [Sosa, 2024] correspondiente a intervalos de confianza, en la sección 7.3:
Se usa la siguiente variable pivotal:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_X}{n_X} + \frac{S^2_Y}{n_Y}}} \sim \mathsf{t}_\nu$$

Con ν de la siguiente manera (usamos la función round para calcularlo en este caso):

$$\nu \approx \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{s_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{s_Y^2}{n_Y}\right)^2}{n_Y - 1}}$$

Y así queda el intervalo de confianza:

$$\mathsf{Pr}\left((\bar{X} - \bar{Y}) - \mathsf{t}_{\nu,1-\alpha/2}\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} < \mu_X - \mu_Y < (\bar{X} - \bar{Y}) + \mathsf{t}_{\nu,1-\alpha/2}\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}\right) = 1 - \alpha$$

3 Intervalos por Bootstrap

Nuevamente citando [Sosa, 2024], tenemos los siguientes intervalos con bootstrap:

Algoritmo

1. Obtener una realización $x_1, \ldots, x_n$ de una muestra aleatoria $X_1, \ldots, X_n$ de tamaño n de una distribución $F_X(\theta)$ con parámetro θ.
2. Establecer un estimador $T = T(X_1, \ldots, X_n)$ del parámetro θ.
3. Tomar una muestra aleatoria $x_1^*, \ldots, x_n^*$ con reemplazo de tamaño n de la muestra original $x_1, \ldots, x_n$. La muestra $x_1^*, \ldots, x_n^*$ se denomina remuestra.
4. Calcular el estadístico T a partir de la remuestra, es decir, calcular $t^* = t^*(x_1^*, \ldots, x_n^*)$.
5. Almacenar el valor de t^*.
6. Repetir M veces los pasos 3., 4. y 5. (e.g., $M = 1000$).

Asumiendo $t^*_{\alpha/2}$ y $t^*_{1-\alpha/2}$ como estimadores se tienen los siguientes intervalos de confianza:

Intervalo de confianza basado en percentiles

$$\mathsf{IC}_{100(1-\alpha)\,\%}(\theta) = \left(t^*_{\alpha/2}, t^*_{1-\alpha/2}\right)$$

Intervalo de confianza empírico

$$\mathsf{IC}_{100(1-\alpha)\,\%}(\theta) = \left(2t - t^*_{1-\alpha/2}, 2t - t^*_{\alpha/2}\right)$$

Intervalo de confianza Normal

$$\mathsf{IC}_{100(1-\alpha)\,\%}(\theta) = \left(t - \mathsf{z}_{1-\alpha/2}\, s_{t^*}, t + \mathsf{z}_{1-\alpha/2}\, s_{t^*}\right)$$

Referencias

[Blanco, 2023] Blanco, L. B. (2023). *Probabilidad Teoría y Práctica*. Editorial, 3 edition.

[Joseph, 2023] Joseph, T. (2023). What is bootstrapping statistics? Built In.

[Pavlovic, 2024] Pavlovic, M. (2024). Log-normal distribution - a simple explanation - towards data science. Medium.

[Sosa, 2024] Sosa, J. (2024). Intervalos de confianza.