Instituto Tecnológico de Buenos Aires

Trabajo Práctico

El desperdicio de alimentos

Estadística Aplicada 11.67

Comisión C

Grupo 5

Integrantes:

Camilo Barbero - 63150

Ignacio Simón - 63172

Profesor titular:
Pedro Camilo Cosatto Ammann

Ayudante a cargo:
Ariel Waisburg

Fecha de entrega:
10 de Septiembre de 2023

Índice

Trabajo Práctico 1: Estadística Descriptiva

Búsqueda del set de datos

La base de datos elegida para analizar es “Food Waste Dataset”, extraído del sitio web Kaggle pero desarrollado por las Naciones Unidas. En este se muestran las cantidades de desperdicio de alimento, en el que se especifican tanto por toneladas y kg por capital de desperdicios de comida que generan los diferentes actores de la sociedad (hogares, minoristas y servicios de comida) en cada uno de los países del mundo, entre otros territorios. Los datos por país (y territorio) fueron integrados por las Naciones Unidas en un único dataset a partir de diversas fuentes primarias y estudios de medición provenientes de organizaciones nacionales de cada país.

A pesar de contar con varias columnas, se decidió por filtrar la base de datos para conservar únicamente las columnas “Household estimate (tonnes/year)”, “Retail estimate (tonnes/year)” y “Food service estimate (tonnes/year)” que serán las variables cuantitativas. Del mismo modo, se eligió como variable categórica de interés a “Confidence in estimate” que cuenta con cuatro niveles de confianza: “High Confidence”, “Medium Confidence”, “Low Confidence” y “Very Low Confidence”.

Es importante destacar que, el dataset es una muestra de corte transversal ya que los datos fueron extraídos en un mismo momento por lo que al pertenecer a una misma población están idénticamente distribuidas las observaciones. Además, son independientes entre sí porque los desperdicios de un actor en un país no tienen relación con los de otros actores en un mismo país y tampoco con el mismo actor en otros países.

Para simplificar el trabajo, se modificaron los nombres de las columnas para referenciar a las variables de interés de manera más sencilla por: “Hogares”, “Minoristas”, “Servicios” y “Confianza” respectivamente. Las variables se definen de la siguiente manera:

Hogares: Cantidad estimada de desperdicio de comida de los hogares del país medida en toneladas en el año 2021.

Minoristas: Cantidad estimada de desperdicio de comida de vendedores minoristas del país medida en toneladas en el año 2021.

Servicios: Cantidad estimada de desperdicio de comida de servicios/proveedores de comida del país medida en toneladas en el año 2021.

Confianza: Nivel de confianza establecido por las Naciones Unidas acerca de los datos cuantitativos proporcionados por el país.

Análisis descriptivo

	Mínimo	Máximo
Hogares	850	91.646.213
Minoristas	138	22.424.705
Servicios	276	65.377.741

Tabla 1

Al realizar un gráfico de la función de distribución empírica $\widehat{F}(x)$ para cada una de las variables numéricas Hogares, Minoristas y Servicios, se pueden destacar diversas observaciones.

$$\widehat{F}(x_0) = (\frac{1}{n}) \cdot \sum_{i=1}^{n} 1(X_i \leq x_0)$$

En primer lugar, al observar los gráficos (*Figura 1*), se puede detectar cierta continuidad en la muestra. Dado que, no hay visibles puntos pesados o "escalones" lo cual indicaría una discreción en las variables si fuera lo contrario, así como el hecho de que las tres tratan de mediciones en toneladas, se pueden considerar las tres variables continuas. En segundo lugar, se pueden registrar algunos puntos más separados en los tres gráficos, lo cual permite suponer la existencia de outliers que aumentan en gran medida el rango de las tres variables.

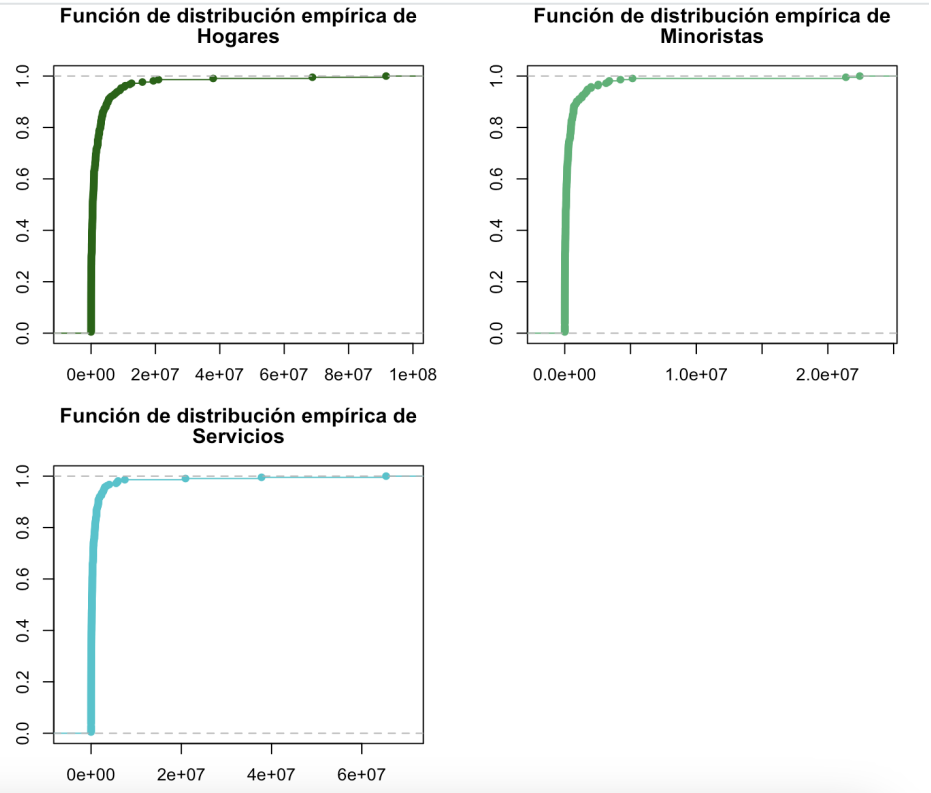

Figura 1

Debido a la continuidad de las variables, se decidió por graficar la función de densidad ya que un histograma sería más apropiado para una variable discreta.

En los gráficos de densidad (*Figura 2*), se puede apreciar cómo para cada una de las variables se concentran los valores de la muestra al principio de su rango, mientras que una cantidad menor de datos toman valores extremadamente altos, lo cual confirma la existencia de datos atípicos. Esto es más reconocible al analizar más a fondo los datos atípicos con gráficos como el boxplot o un beeswarm. Como resultado, estas variables siguen una fuerte inclinación hacia la izquierda por lo que se puede deducir que tienen asimetría positiva.

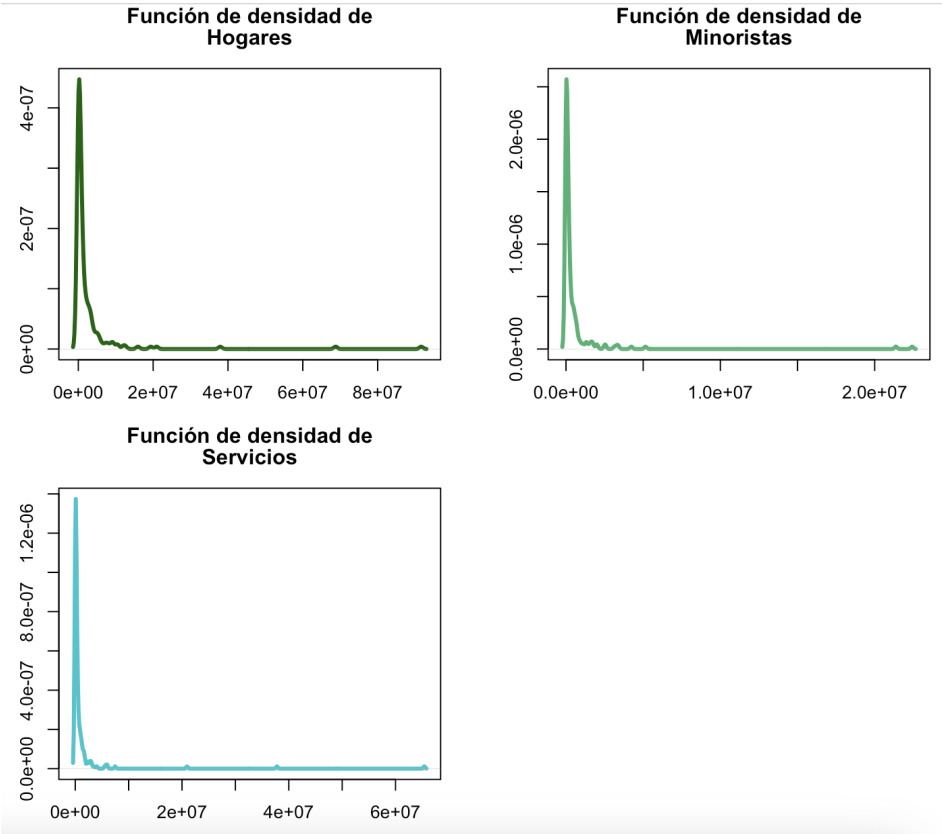

Figura 2

A continuación se calcularon los valores de la media y mediana de cada una de las variables.

Mediciones/Variables	**Hogares**	**Minoristas**	**Servicios**
Media ($\bar{x}$)	2.658.896	552.045,4	1.138.859
Mediana ($Me = x_{0,5}$)	520.508	100.650	188.466

Tabla 2

A partir de los datos de la *Tabla 2*, se puede interpretar que como para las tres variables la mediana se encuentra bastante más a la izquierda de la media, las variables tienen una asimetría a la derecha lo cual confirma lo que se vio previamente en los gráficos de la densidad.

También, se buscaron los valores de los cuantiles de todas las variables y se graficaron los boxplots. (*Tabla 3* y *Figura 4*).

Mediciones/Variables	Hogares	Minoristas	Servicios
Cuantil 0.25 (Q1 = $x_{0.25}$)	67.385	12.243	22.013
Cuantil 0.75 (Q3 = $x_{0.75}$)	2.119.455	364.585	644.496
Rango inter-cuartilico (IQR)	2.042.999	344.306,5	608.370,5

Tabla 3

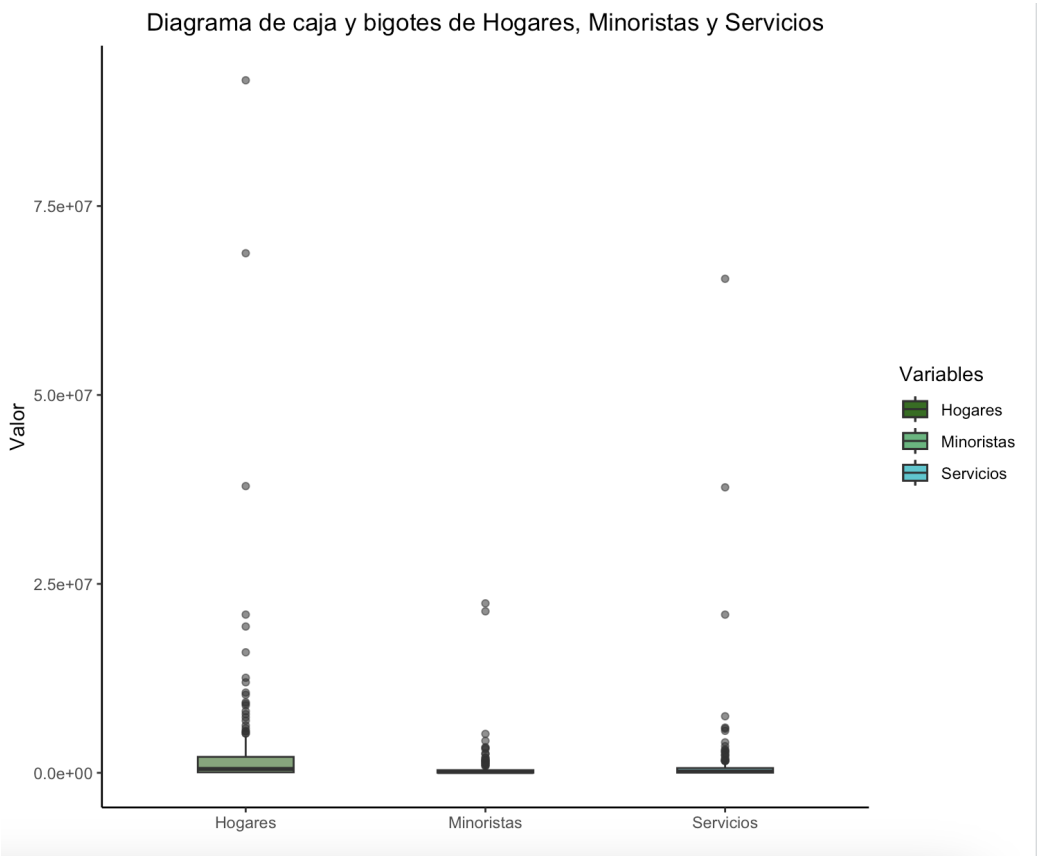

Figura 4

A partir del cálculo de las medidas de tendencia central, las de posición así como el rango intercuartílico, se puede confeccionar un diagrama de caja y bigote. A pesar de haber varias metodologías para indicar y analizar los valores atípicos, si se consideran como atípicos aquellos datos que sean menores al bigote mínimo $(Q1 - IQR * 1,5)$ y mayores al bigote máximo $(Q3 + IQR * 1,5)$ del boxplot, entonces se pueden observar como hay bastantes puntos que están altamente dispersos de la concentración de puntos. Al mismo tiempo, a pesar de no verse tan claro en

las tres variables, se nota la asimetría positiva mencionada previamente debido a que la mediana no está en el centro del rango intercuartílico (media) pero más a la izquierda de ella. Esto también indica como la media podría estar altamente influenciada por estos datos atípicos.

Mediciones/Variables	Hogares	Minoristas	Servicios
Desviación Estándar (*S*)	8.596.906	2.195.578	5.380.459
Coeficiente de Variación (CV)	3,233	3,977	4,724
Coeficiente de Asimetría (γ_1)	7,743	8,70099	9,675
Coeficiente de Curtosis (g_2)	71,524	83,975	105,876

Tabla 4

Por un lado, en la *Tabla 4* se puede notar cómo lo que se explicó anteriormente sobre la asimetría positiva de las variables se plasma con el coeficiente de asimetría siendo positivo en los tres casos. Esto se ve claramente en el gráfico de densidad al encontrarse la curva a la izquierda.

Por otro lado, la desviación estándar es grande, lo cual indica que hay mucha distancia entre en los datos y la media, posiblemente debido a datos inflados como los outliers, que modifican sustancialmente el valor de la media y consecuentemente al desvío.

Por lo que se refiere al coeficiente de variación, el cual indica cuantas media hay en el desvío, se puede analizar como hay una dispersión mayor con respecto a la media, lo cual confirma una alta variabilidad en los datos. La variable Servicio es la que presenta mayor variabilidad.

Por último, el coeficiente de curtosis, el cual indica según el signo de su valor el aplastamiento o apuntamiento de la curva indicando una mayor concentración de los datos en la zona central o no, permitiendo un buen índice contraste entre la densidad de la muestra con una distribución normal. Este es positivo y de grado alto en las tres variables. Por lo tanto, se puede afirmar que las variables presentan distribuciones leptocúrticas. Esto es debido a sus picos pronunciados y colas pesadas las cuales indican una frecuencia más usual de datos extremos a diferencia de una distribución normal.

Para resumir, todas las variables al tratarse de toneladas de desperdicio producidas por diferentes actores sociales, presentan alta variabilidad, distribuciones leptocúrticas así como una gran dispersión con respecto a la media y una asimetría hacia la derecha.

Como se mencionó repetidas veces, hay una presencia de datos atípicos extremos los cuales a pesar de poder ser visualizados en el boxplot, convendría el uso de un Bee Swarm donde se pueden inspeccionar con mayor facilidad. Puesto a que, el método del rango intercuartílico está más adaptado a los datos en cuestión a diferencia de la regla fija de 2 o 3 desvíos estándar de la media para buscar los outliers, se decidió por utilizar el primer método para calcular los outliers. Por esta razón, como las distribuciones de los datos parecieran no ser normales, el uso del desvío estándar como regla fija no tomaron en cuenta la fuerte asimetría positiva de los datos producida por los valores atípicos extremos presentes. De esta manera, se evita la asunción de una distribución normal en los datos y la falsa identificación de valores atípicos.

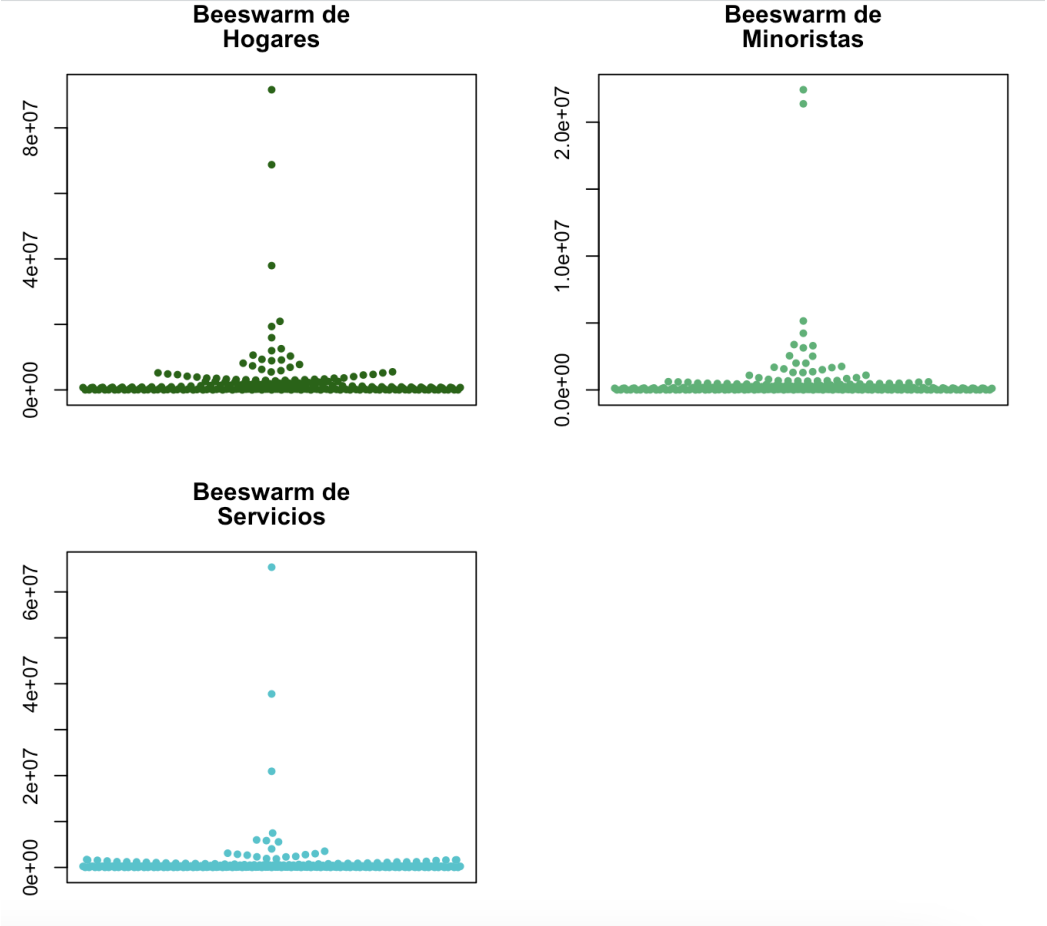

Figura 5

Mediciones/ Variables	Hogares	Hogares (s/outliers)	Minoristas	Minoristas (s/outliers)	Servicios	Servicios (s/outliers)
Media ($\bar{x}$)	2.658.896	*930.780*	552.045,4	*156.573,5*	1.138.859	*283.387.7*
Mediana ($Me = x_{0,5}$)	520.508	*449.895*	100.650	*74.216*	188.466	*130.061.5*
Cuantil 0.25 (Q1 = $x_{0.25}$)	67.385	*54.765*	12.243	*8305*	22.013	*16.075*
Cuantil 0.75 (Q3 = $x_{0.75}$)	2.119.455	*1.391.729*	364.585	*238.248*	644.496	*426.966*
Rango inter-cuartilico (IQR)	2.042.999	*1.317.496*	344.306.5	*225.200,5*	608.370,5	*405.304,5*
Desviación Estándar (S)	8.596.906	*1.185.806,2*	2.195.578	*198247,1*	5.380.459	*362.527,4*
Coeficiente de Variación (CV)	3,233	*1,274*	3,977	*1,266*	4,724	*1,279*
Coeficiente de Asimetría (γ_1)	7,743	*1,5054*	8,70099	*1,495*	9,675	*1,566*
Coeficiente de Curtosis (g_2)	71,524	*4,387*	83,975	*4,336*	105,876	*4,646*

Tabla 5

Para comprender mejor el efecto de los outliers, es importante observar el cambio en las medidas de tendencia central así como de posición y dispersión en la *Tabla 5*. Al momento de filtrar los outliers, las tres variables presentan veintitrés o veinticuatro outliers por variable, probablemente correspondiente a los mismos países (observaciones) que generan inusualmente alto desperdicio alimentario. No obstante, a pesar de la importancia de ver el efecto de los datos atípicos, es importante no excluirlos completamente ya que esto alteraría los resultados de la muestra.

En primer lugar, se confirma lo dicho anteriormente acerca de cómo los datos atípicos inflan extremadamente la media y consecuentemente la desviación estándar. Al excluirlos de las tres variables, la variabilidad de los datos baja significativamente, la curva de la densidad se desplaza un poco a derecha ya que disminuye el coeficiente de asimetría a pesar de mantenerse a la derecha. Esto

lleva a que los tres coeficientes sean bastante similares entre las variables, probablemente debido a ser las tres medidas de desperdicio alimentario.

En segundo lugar, el coeficiente de curtosis es en el que se observa una disminución dramática ya que al excluir los datos atípicos, se minimizan las colas pesadas llevando así a una distribución mucho más parecida a la de una normal. Esto se puede ver particularmente en la *Figura 6* con la variable de Servicios.

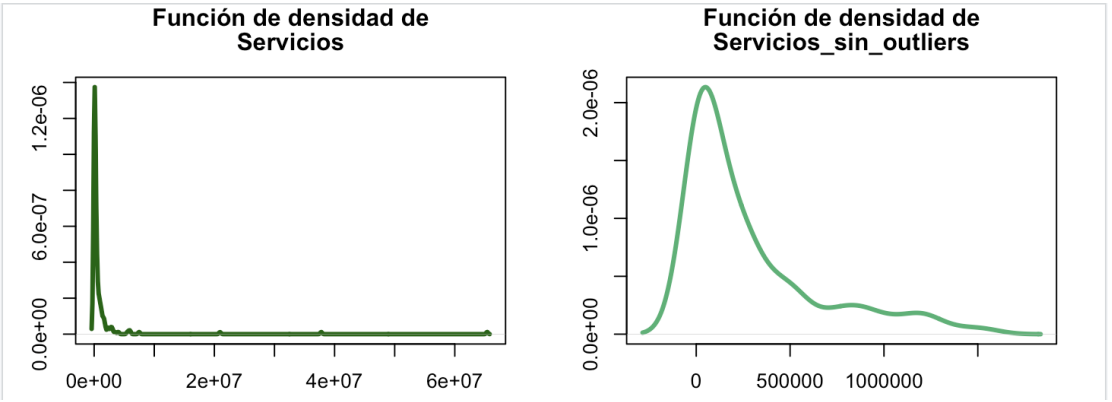

Figura 6

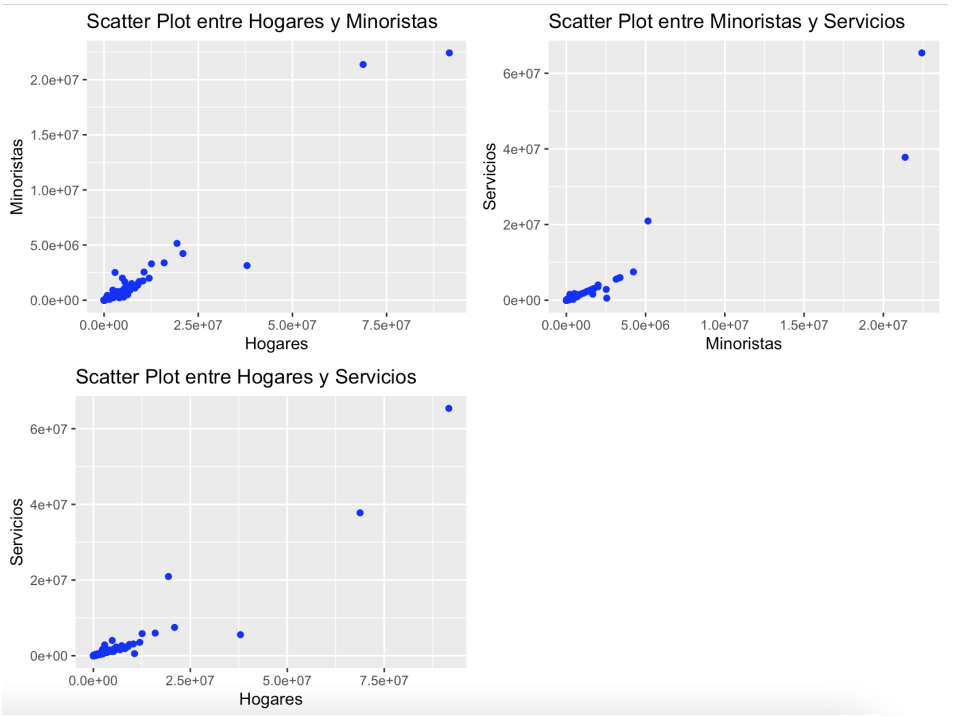

Figura 7

Variables	Hogares	Minoristas	Servicios
Hogares	**1**	0,964	0,945
Minoristas	0,964	**1**	0.962
Servicios	0,945	0,962	**1**

Tabla 6

Para entender la relación entre las variables numéricas, se realizaron scatterplots en la *Figura 7* y una matriz de correlación en la *Tabla 6*. Se puede analizar como hay una correlación de Pearson fuerte entre todas las variables, al ser mayor a 0,5. Aunque hay que tener en cuenta que todas se tratan de cantidad de desperdicio alimentario en grandes grupos. Sin embargo, hay que interpretar estos coeficientes con precaución ya que el coeficiente de correlación de Pearson es sensible a la presencia de datos atípicos, lo cual ocurre en las tres variables de la muestra.

En relación con la variable categórica Confianza, la cual toma los valores: "Very Low Confidence", "Low Confidence", "Medium Confidence" y "High Confidence", se confeccionó una comparación gráfica (*Figura 8*) entre la distribución de la variable Servicios y las distribuciones de la misma pero en distintos niveles de confianza. Cabe destacar que, se limitó el rango de los datos (eje x) hasta 25 millones de toneladas para visualizar las distribuciones positivamente asimétricas más claramente.

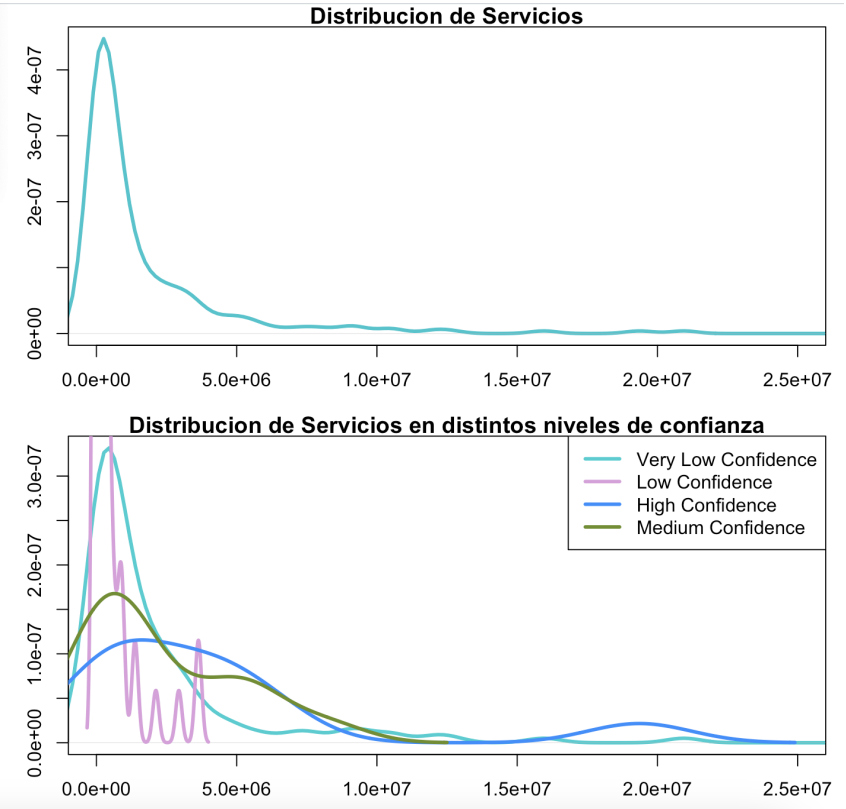

Figura 8

Por un lado, se puede observar como hay un mayor achatamiento en la curva de densidad a mayor nivel de confianza, y mayor apuntamiento a menor nivel de confianza. Sin embargo, esto no es estrictamente correcto ya que para “Low Confidence” hay un apuntamiento tan alto de la curva, aun mayor que “Very Low Confidence”, que supera el rango del eje y.

Por otro lado, hay una similitud entre la distribución de la variable y la de la variable con los diferentes valores de “Confianza”. En esta se constata cómo la distribución de la variable con “Very Low Confidence” es similar a la distribución general de Servicios. Puesto que, la mayor parte de la muestra presenta muy baja confianza en los datos proporcionados, se podría inferir que los valores que toma la variable con “Very Low Confidence” afectan en gran medida a la distribución de la variable en general, generando ese apuntamiento alto al inicio.

Trabajo Práctico 2: Estimadores puntuales

Estimadores puntuales de la mediana

Luego del análisis hecho a partir de la variable aleatoria elegida Servicios y como su distribución cambia notoriamente en distintos niveles de confianza, se eligió por utilizar la variable condicionada en "Very Low Confidence". Teniendo en cuenta que esta categoría ocupaba la mayor cantidad de observaciones en la muestra, así como la manera en la que se distribuía empíricamente, se decidió por utilizar la variable definida como:

X: Cantidad estimada de desperdicio de comida de servicios/proveedores de comida del país medida en toneladas en el año 2021 con un nivel de confianza Muy Bajo.

A partir de la variable de estudio elegida, se pueden estimar diversos parámetros de interés, en este caso la mediana, a través de distintos estimadores:

$$q = Me = x_{0.5}$$

$$\widehat{q} = \widehat{Me} = \widehat{x_{0.5}}$$

Con respecto al parámetro de interés, se eligieron tres estimadores de la mediana. Por un lado, se calculó a través de su forma no paramétrica, con el uso de la función de distribución acumulada empírica. Esto permite observar al momento de calcular el error estándar, entender que tan representativa es la muestra obtenida de la población entera, o si en realidad el estimador tiene una utilidad limitada.

$$\left(x_1, x_2, \ldots, x_{130}\right) \sim iid \sim F$$

Por otro lado, al tratarse de una variable que trata mediciones positivas y debido a que previamente se observó que la distribución empírica tiene una asimetría altamente positiva con colas pesadas, un estimador paramétrico que podría ser utilizado es del modelo de Log-Normal. De forma semejante, debido a la alta presencia de datos extremadamente inflados, al tratarse de un contexto socio-económico donde hay un pequeño porcentaje de países (observaciones) como China que concentran una porción significativa de la riqueza, desechando así numerosas toneladas de desperdicio, se decidió hacer uso del modelo de Pareto de Tipo I. Este modelo a pesar de no ser tan aplicable universalmente como el Gamma, entre otros que también presentan asimetría positiva, se considera particularmente provechoso para este contexto de colas pesadas ya que es muy utilizado en el estudio de la distribución de la riqueza y los ingresos, lo cual puede tener semejanzas con la distribución de desperdicio de comida en diferentes países.

$$\left(x_1,\ x_2,\ \ldots,\ x_{130}\right)\ \sim\ iid\ \sim\ Ln(m,\ D)$$

$$\left(x_1,\ x_2,\ \ldots,\ x_{130}\right)\ \sim\ iid\ \sim\ P(\theta,\ \beta)$$

Estimadores	**Parámetros**	$\widehat{q} = \widehat{x_{0.5}}$
No paramétrico $X \sim F$	-	$\widehat{q} = F^{-1}(0.5) = min\{x_i \widehat{F}(x_i) \geq 0,5\}$
Modelo de Log-normal $X \sim Ln(m,\ D)$	$\widehat{m_{MV}} = \left(\frac{1}{130}\right) \cdot \sum_{i=1}^{130} ln(x_i)$ $\widehat{D_{MV}} = \sqrt{\left(\frac{1}{130}\right) \cdot \sum_{i=1}^{130} (ln(x_i) - \widehat{m_{MV}})^2}$	$\widehat{q} = F^{-1}(0.5) = e^{\widehat{m_{MV}} + z_{0,5} \cdot \widehat{D_{MV}}}$
Modelo de Pareto $X \sim P(\theta,\ \beta)$	$\widehat{\theta_{MV}} = min\{x_1,\ x_2,\ \ldots,\ x_{130}\}$ $\widehat{\beta_{MV}} = 130 \cdot \left(\sum_{i=1}^{130} ln(x_i - \widehat{\theta_{MV}})\right)^{-1}$	$\widehat{q} = F^{-1}(0.5) = \widehat{\theta_{MV}} \cdot \left((0,5)^{1/\widehat{\beta_{MV}}}\right)^{-1}$

Tabla 7

En primer lugar, en la *Tabla 7* se plantearon los cálculos necesarios previos al cálculo de $\widehat{q}$; el cual en el caso de los estimadores paramétricos requieren de los parámetros correspondientes de los modelos que también deben ser estimados. Como método de estimación se escogió el uso de la máxima verosimilitud ya que optimiza los parámetros lo más posible y permite teóricamente tener un error estándar menor que con otros métodos como el de ajuste de momentos.

Estimadores	**Parámetros**	$\widehat{q} = \widehat{x_{0.5}}$
No paramétrico $X \sim F$	-	$\widehat{q_{np}} = 279.293$
Modelo de Log-normal $X \sim Ln(m,\ D)$	$\widehat{m_{MV}} = 12,259$ $\widehat{D_{MV}} = 2,0749$	$\widehat{q_{ln}} = 210.822,84$
Modelo de Pareto $X \sim P(\theta,\ \beta)$	$\widehat{\theta_{MV}} = 213$ $\widehat{\beta_{MV}} = 0,154$	$\widehat{q_{pa}} = 28.799,61$

Tabla 8

Se puede observar en la *Tabla 8* como hay una gran diferencia entre los tres estimadores, particularmente entre el modelo Pareto respecto a los demás. Es razonable deducir que debido a que el Modelo Pareto fue construido para tener en mente la existencia de observaciones altamente infladas, aunque su distribución y parámetros es sensible a estos outliers de naciones potencia en comparación

a los otros estimadores, ya que busca capturarlos. No obstante, antes llegar a ciertas conclusiones hay que tener en cuenta el error estándar, el cual proporciona información acerca de qué tan preciso es el estimador en alcanzar el valor q real o poblacional.

Análisis del error

En cuanto al error estándar de los estimadores de la muestra, debido a la falta de información poblacional como su distribución real y consecuentemente la falta de una formula analitica para el cálculo del error estándar, este se tuvo estimar a través del método bootstrap en el que se extrajeron 1000 muestras aleatorias de tamaño 130 con reposición a partir de la muestra original, permitiendo así una simulación de cómo se podría distribuir la población. A partir de este procedimiento que se realizó tres veces correspondiente a cada estimador, se obtuvo el desvío estándar de las 1000 repeticiones para estimar el error.

Se puede observar en la *Tabla 9* como hay una alta variación entre los errores de los estimadores. El Modelo de Pareto es el que tiene menor error estándar, seguido por el modelo Log-normal el cual está bastante cerca al del estimador no paramétrico. El no paramétrico tiene el mayor error estándar probablemente porque al ser una muestra que representa una parte pequeña de la población hay mucha variabilidad de muestra en muestra. De manera que, se podría suponer que el estimador de Pareto con máxima verosimilitud es el que más se aproxima a la mediana real.

Estimadores	$\widehat{q} = \widehat{x_{0.5}}$	**se**$(\widehat{q})$
No paramétrico $X \sim F$	$\widehat{q_{np}} = 279.293$	$se(\widehat{q_{np}}) = 48.513,77$
Modelo de Log-normal $X \sim Ln(m,\ D)$	$\widehat{q_{ln}} = 210.822,84$	$se(\widehat{q_{ln}}) = 38.075,15$
Modelo de Pareto $X \sim P(\theta,\ \beta)$	$\widehat{q_{pa}} = 28.799.61$	$se(\widehat{q_{pa}}) = 12.004,33$

Tabla 9

Ajuste de las distribuciones

Por último, con lo que respecta a qué modelo de distribución de los elegidos es el que se ajusta mejor al histograma de la muestra, ambos se acercan bastante a la distribución de la muestra en la *Figura 9* debido al apuntamiento en sus curvas y la visible asimetría hacia la derecha. Sin embargo, los cálculos obtenidos, así como las curvas del gráfico parecen respaldar la suposición de que el modelo de Pareto es el que mejor se ajusta a los datos y una estimación de parámetros reales. Siendo

más específico, el error estándar a través del método Bootstrap fue sustancialmente menor, además de que los modelos asumen diferentes suposiciones.

En cuanto al modelo Log-normal, este asume una transformación de la normal lo cual no parece tan lógico en un ambiente socioeconómico donde los valores de variables como el desperdicio de comida no se distribuyen simétricamente alrededor de una línea de tendencia central como la media. Por el contrario, el modelo de Pareto toma no solo en cuenta la asimetría pero también la existencia de pocos pero datos extremadamentes altos, que en este caso pertenecen a pocos países que generan altos niveles de producción y desperdicio, mientras que la gran mayoría de que pertenecen a los desperdicios de Servicios con muy baja confianza según las Naciones Unidas, suelen ser países más pobres productivamente. Pero es importante resaltar que, el modelo Pareto es sensible a los datos atípicos y por ello su curva podría no ser tan alta como el histograma.

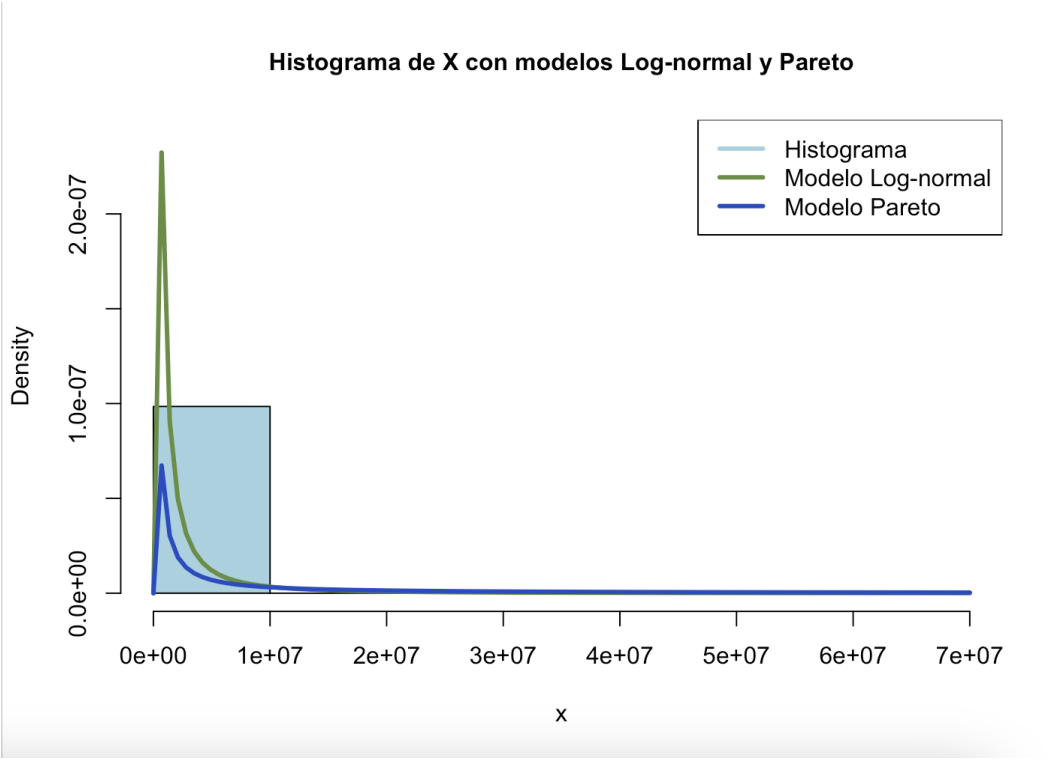

Figura 9

Bibliografía

ARVIDSSON, J. (2019, March 9). *Food Waste*. Kaggle. Retrieved August 20, 2023, from https://www.kaggle.com/datasets/joebeachcapital/food-waste/code

Forbes, H., Quested, T., & O'Connor, C. (2021, March 4). *UNEP Food Waste Index Report 2021 | UNEP*. UN Environment Programme. Retrieved August 20, 2023, from https://www.unep.org/resources/report/unep-food-waste-index-report-2021