ISSN 2796-8952

CUADERNOS DEL GRUPO DE ESTADÍSTICA PARA EL ESTUDIO DEL LENGUAJE

Volumen I

Julia Carden (ed.)
Ezequiel Koile (ed.)
Analí Taboh (ed.)
Federico Alvarez
Santiago Gualchi
Mercedes Güemes
Laura Tallon
Charo Zacchigna

GESEL

Corrección y edición
Julia Carden
Ezequiel Koile
Analí Taboh

Cuadernos del Grupo de Estadística para el Estudio del Lenguaje, I volumen / Federico Alvarez; Santiago Gualchi; Mercedes Güemes; Laura Tallon y Charo Zacchigna.

1era edición.
25 de mayo, 217; primer piso.
Ciudad Autónoma de Buenos Aires, Argentina.
Publicación digital, diciembre 2021.
ISSN: 2796-8952

CUADERNOS DEL GRUPO DE ESTADÍSTICA PARA EL ESTUDIO DEL LENGUAJE

Volumen I

Federico Alvarez, Santiago Gualchi , Mercedes Güemes, Laura Tallon y Charo Zacchigna.

Editado por Julia Carden, Ezequiel Koile y Analí Taboh.

ISSN 2796-8952

Prólogo

Sobre este cuaderno

El Cuaderno del Grupo de Estadística para el Estudio del Lenguaje - Volumen I es una obra colectiva. Los distintos capítulos que lo integran son el resultado de una recopilación bibliográfica llevada a cabo dentro del marco del subsidio FiloCyT "Métodos cuantitativos en el estudio del lenguaje". Cada capítulo fue elaborado para acompañar una reunión del equipo en la que se trataron los distintos temas con mayor profundidad, a modo de presentación oral. Este volumen no pretende ser una obra acabada sobre los temas de estadística que se abordan, sino una guía para la lectura de bibliografía más compleja y completa sobre el tema. Por este motivo, estos capítulos deben ser acompañados de la lecturas que se sugieren en las referencias.

Estos cuadernos fueron concebidos para aportar material de consulta sobre estadística aplicada al estudio del lenguaje que, al menos en español, sigue siendo un área de vacancia bibliográfica. El objetivo es que proporcionen información, acceso a nuevos materiales y una explicación breve y simple sobre los puntos esenciales que deben ser conocidos para quienes se introduzcan en el estudio cuantitativo del lenguaje y no posean una base firme en matemática.

Capítulo III

Estadística descriptiva

Laura Tallon y Charo Zacchigna

Antes de realizar cualquier análisis cuantitativo es imprescindible conocer la naturaleza del conjunto de datos y ciertas características básicas. Como su nombre lo indica, la estadística descriptiva es la rama de la estadística que se centra en organizarlos y describirlos, de manera de tener un pantallazo general de los datos a analizar. En el presente capítulo se trabajan nociones básicas por donde comenzar el estudio cuantitativo, tales como: los tipos de variables, la distribución, la importancia de visualizar los datos, diversas medidas de tendencia y dispersión, valores de normalización y correlación entre variables.

1. Introducción

La estadística descriptiva generalmente constituye el primer paso del análisis cuantitativo de un conjunto de datos. El objetivo es transformar la información a un número menor y más manejable de datos, dejando de lado los detalles y el ruido, para describir las propiedades básicas y generales de los datos. No se busca hacer inferencias sobre una población más grande que la muestra estudiada, sino que se utiliza como material exploratorio del que pueden surgir preguntas que conduzcan a una hipótesis sobre una población más grande. La estadística descriptiva, entonces, se circunscribe solo a los datos de la muestra estudiada.

1.1. Variables

El primer paso en el análisis es identificar los tipos de variables en estudio, ya que esto definirá los métodos y tipos de medidas estadísticas a utilizar.

-
    - **Variable de razón**: es una variable numérica que toma valores de una escala, cuyos intervalos pueden ser medidos (se puede decir que la diferencia entre 2 sílabas y 4 sílabas es igual a la diferencia entre 5 y 7 sílabas, y podemos afirmar que una palabra de 6 sílabas tiene el doble que una palabra de 3). Además la escala tiene un cero definido que marca la inexistencia (no existe una palabra de 0 sílabas) y por supuesto no cuenta con valores negativos. Otros ejemplos son los tiempos de reacción y la duración de emisiones, pausas, fijaciones, etc.
    - **Variable de intervalo**: a diferencia de la anterior, la escala de esta variable tiene un cero arbitrario, cuyo valor no significa la inexistencia de dicha variable, por lo tanto esta escala también puede tomar valores negativos. Un caso que ejemplifica este tipo de variables es el año de nacimiento, o los datos de una encefalografía en estudios de potenciales relacionados con eventos.
- **Variables ordinales**: son aquellas que tienen tres o más categorías y pueden ser listadas en un orden natural (primero, segundo, tercero o bajo, medio, alto). Con estas variables no se puede saber si los intervalos entre los valores son iguales. Es decir, la diferencia que hay entre bajo y

medio no necesariamente es la misma diferencia que hay entre medio y alto. Un ejemplo puede ser el curso escolar cuando estudiamos grupos de niños con diferentes niveles educativos, los valores de un juicio de aceptabilidad con escala de likert, o el nivel de conocimiento del idioma (alto/medio/bajo).

- **Variables nominales**: son aquellas que tienen tres o más categorías sin un orden natural. Al no estar cuantificadas no pueden realizarse operaciones matemáticas con ellas. Un ejemplo de variable nominal puede ser la clase de palabras y los puntos o modos de articulación.

- **Variables dicotómicas**: son un caso particular de variables nominales que solo tiene dos categorías opuestas. Pueden ser la presencia/ausencia de un rasgo y suelen cuantificarse como 0 y 1. Esta cuantificación puede usarse para aplicar algunos procedimientos de las variables continuas. Un ejemplo puede ser el rasgo sonoro/sordo de los fonemas consonánticos, la presencia/ausencia de pausas, verbos regulares/irregulares o finitos/no finitos.

1.2. Distribución

En segundo lugar, es importante observar la manera en que se distribuyen los datos recolectados. La distribución hace referencia a la probabilidad de cada valor de una variable. En el caso de las variables continuas, una distribución típica que se pueden tomar las variables es la llamada **distribución normal** (Figura 1), también conocida como "campana de Gauss" por su forma, que está definida por dos parámetros: la media y el desvío estándar.

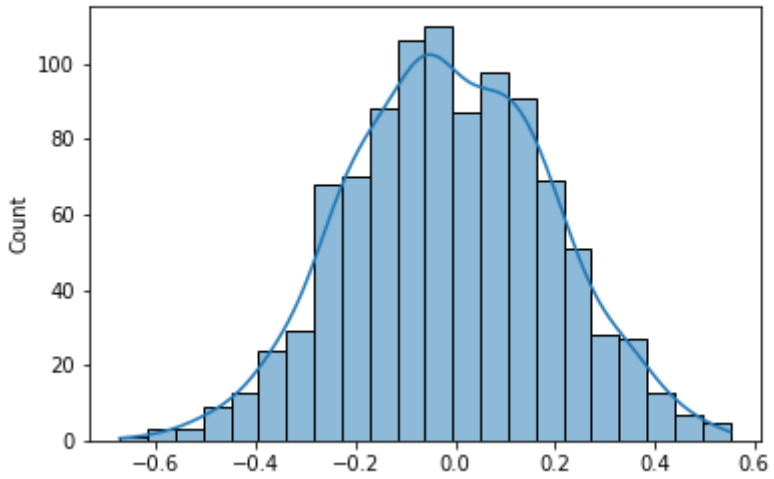


Figura 1. Se grafica un histograma con datos superpuestos con la curva de su distribución normal.

Por supuesto, los datos reales no tienen porqué estar distribuidos como se espera, pero ciertos métodos estadísticos usan esta distribución como base para el análisis. Un método que asume una distribución específica para la población estudiada, se denomina **método paramétrico**. Existen también métodos que no asumen una distribución específica, como los **métodos no-paramétricos** (los más apropiados para datos con variables ordinales o nominales).

Por esto es necesario conocer la distribución antes de aplicar cualquier tipo de método estadístico. Esto puede hacerse graficando la información en un eje cartesiano o representándola en una tabla de frecuencias.

2. Medidas de tendencia central

Estas medidas hacen referencia al valor más representativo de una variable. Hay distintas concepciones de qué es lo más representativo y esto se refleja en distintas formas de medir la tendencia central. En una distribución normal los valores de estas medidas se asemejan (excepto la media geométrica).

- **Media** ($\overline{X}$/μ): también conocida como promedio, se obtiene sumando todos los valores de la variable y dividiendo por la cantidad total de observaciones. Una variable solo tiene una media, que se ve muy afectada por la presencia de ***outliers*** (un valor cuya magnitud es muy diferente a la del resto de los valores de una muestra, ver Figura 3).[1]

$$\overline{X} = \frac{\Sigma X_i}{n}$$

- **Media geométrica**: es la raíz *N*-ésima del producto de todos los valores, siendo *N* la cantidad total de observaciones. Al igual que la anterior, una variable solo tiene una media geométrica, y se ve muy afectada por la presencia de *outliers*.[2]

$$\sqrt[n]{\Pi x_i}$$

- **Mediana**: si se ordenan todos los valores de una variable, desde el más pequeño hasta el más grande, la mediana será el valor que demarca la mitad de los datos. Al tomar por ejemplo la Figura 2, la mediana es 50,37; si empezamos a contar desde el valor más pequeño, al llegar a la mediana habremos contado la mitad de los casos. Si tenemos un número par de valores, la mediana se calcula sacando el promedio de los dos valores centrales. Se utiliza para variables ordinales, continuas y de razón. Una variable solo tiene una mediana. Es menos sensible a los *outliers* que la media.

- **Moda**: Es el valor más común en una distribución, el que más se repite. Una variable puede tener más de una moda. No se ve afectada por la presencia de *outliers*.

[1] **Σ** es una notación matemática que permite representar sumas de varios sumandos. Se conoce como suma o sumatoria. Por ejemplo, $\mathbf{\Sigma}X_i$ significa que se sumarán todos los valores de X; $\mathbf{\Sigma}(X_i/2)$ significa que a cada valor de X se lo dividirá por dos y luego se sumarán los resultados de cada división: $((X_i/2) + (X_i+2/2) + (....) + (X_n/2))$

[2] Π es una notación matemática que permite representar la multiplicación de una cantidad arbitraria. Se conoce como productoria o multiplicatoria. Por ejemplo, ΠX_i significa que se multiplicaran todos los valores de X; $\Pi(X_i/2)$ significa que a cada valor de X se lo dividirá por dos y luego se multiplicarán los resultados de cada división: $((X_i/2) . (X_{i+2}/2) . (....) . (X_n/2))$

ISSN: 2796-8952

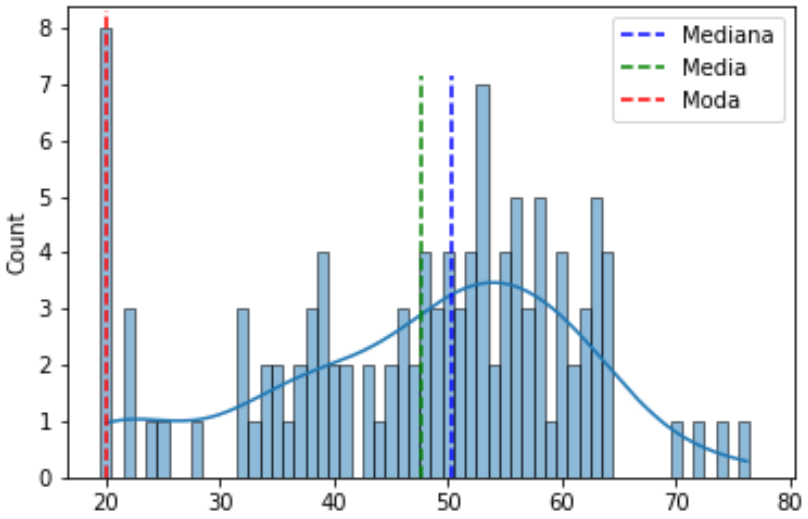


Figura 2. Moda, mediana y media. El gráfico muestra que las medidas de tendencia central no siempre coinciden. En este caso, la moda toma el valor 20, la mediana el valor 50,37 y la media o promedio el valor de 47,57.

Si el conjunto de datos presenta muchos *outliers* es asimétrico, es preferible reportar la mediana. También se utiliza esta medida para variables ordinales, mientras que para variables nominales solo puede utilizarse la moda.

ISSN: 2796-8952

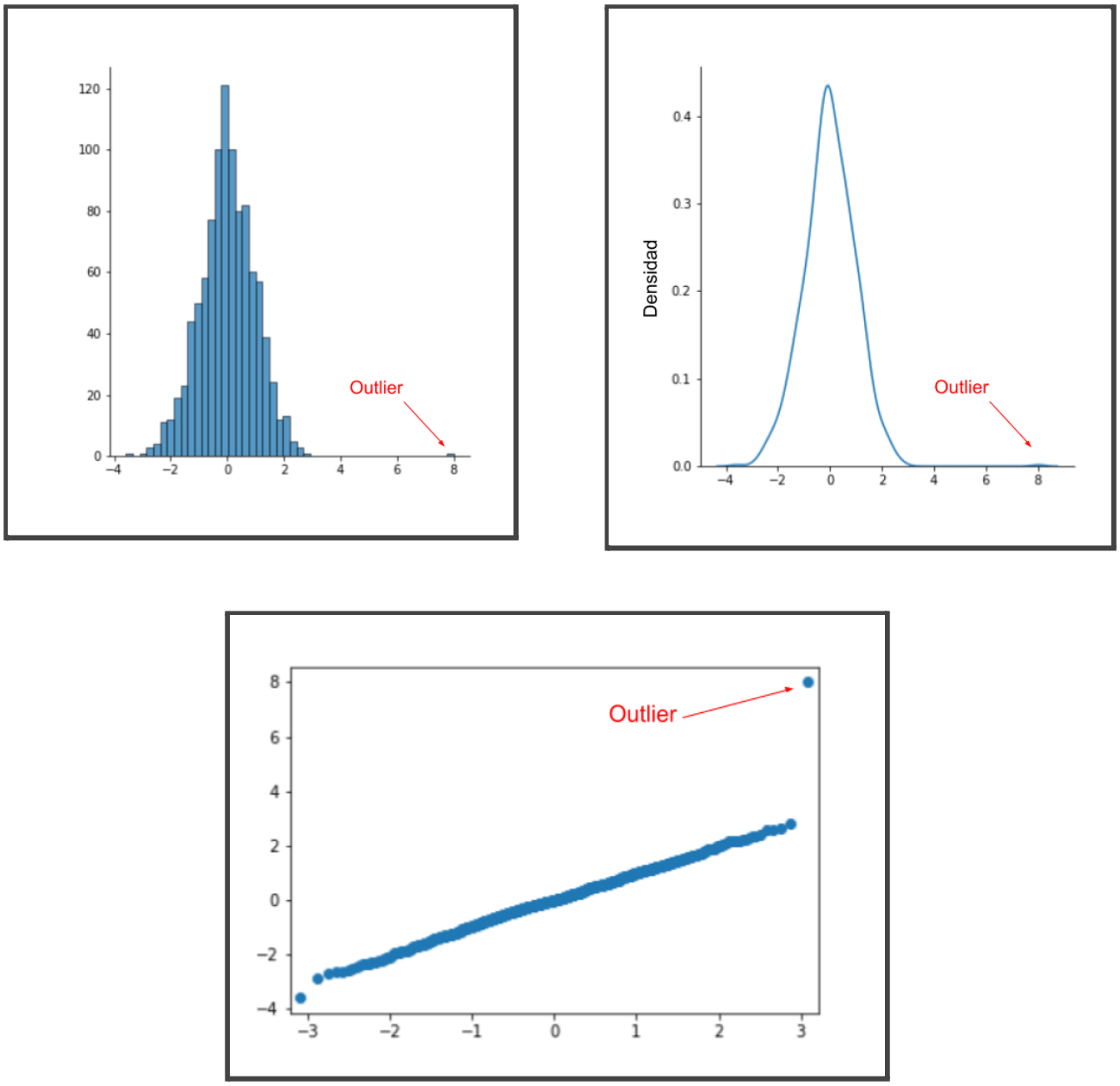


Figura 3. Outlier o "valor atípico": valor cuya magnitud es significativamente diferente a la del resto de los valores de una muestra. Usualmente se utilizan 2 o 3 desvíos estándar por debajo o por encima de la media para delimitar cuáles son los outliers.
La Figura muestra cómo se ven los outliers en un histograma, un gráfico de densidad y un gráfico cuantil-cuantil (qqplot) hechos ad hoc.

ISSN: 2796-8952

3. Medidas de dispersión

3.1. Medidas de dispersión dimensionales

Antes de comenzar con los tipos de medidas es importante conocer los conceptos de percentil y cuartil. El **percentil** es una medida de posición que indica, una vez ordenados los datos de menor a mayor en un grupo de observaciones, el valor por debajo del cual se encuentra un porcentaje dado de observaciones. Por ejemplo, si el 50% de los datos se encuentran por debajo de 34, el percentil 50 será 34. La forma de calcularlo es una regla de tres simple:

$$N = 100\%$$

$$I = P\%$$

Siendo N el total de los datos, P el percentil buscado e I el valor del percentil P_i. Si el valor de I es decimal, se redondea para arriba. A los percentiles 25, 50 (la mediana), 75 y 100 se los denominan **cuartiles**.

Al hacer estadística descriptiva resulta necesario medir cómo se comporta la variable estudiada. La medida de tendencia central no es suficiente por sí misma para describir su distribución (ver §7). Las medidas de dispersión representan qué tanto varían los valores con respecto a la tendencia central. Existen varias medidas, que toman las mismas unidades que la variable:

- **Rango**: es la diferencia entre el valor máximo y el valor mínimo de una variable. Es muy sensible a *outliers*.
- **Rango intercuartil** (**IQR**): si se calcula la diferencia entre el tercer (75) y el primer cuartil (25), se obtiene lo que se llama "rango intercuartil" (IQR por sus siglas en inglés). La diferencia entre los cuartiles muestra la heterogeneidad de los datos (mientras más difieran los cuartiles, más heterogénea es la muestra). Utilizando el IQR se disminuye la influencia de los *outliers*.
- **Desviación estándar** (**SD o σ**): es la medida más utilizada y se obtiene sacando la raíz cuadrada de la varianza. La **varianza** (**σ^2**) se calcula, primero, obteniendo para cada valor de la variable el cuadrado de su distancia con la media, luego realizando la sumatoria de estos cuadrados, y finalmente dividiendo todo por N. Entonces:

$$varianza = \frac{\sum_{i=1} (x_i - \overline{x})^2}{n}$$

$$\sigma = \sqrt{varianza}$$

  Cuando la muestra es grande hay que cambiar el divisor por n-1. A esto se le llama "corrección de Bessel" y se usa para no sobreestimar la muestra con respecto a la población. Dos distribuciones pueden tener medias similares, pero desvíos muy diferentes (ver §7).
- **Error estándar** (**SE** por sus siglas en inglés): se define como la desviación estándar de las medias de muestras igualmente grandes de una misma población. La media de una muestra no es lo mismo que la media de toda la población, a menos que la muestra sea perfectamente representativa. Una forma de saber qué tan representativa es la media de la muestra en relación con toda la población es computar la media de varias muestras semejantes (tomadas al azar, pero del mismo tamaño) y

ISSN: 2796-8952

calcular la desviación estándar de todos esos promedios. A este cálculo se lo conoce como error estándar:

$$SE = \sqrt{\frac{var}{n}} = \frac{\sigma}{\sqrt{n}}$$

Mientras más grande es el error, menos representativa es la estimación para el resto de la población. Y mientras más grande sea *N* (el tamaño de la muestra), más pequeño será el error estándar. Si al comparar las medias de dos muestras aproximadamente iguales los errores estándar se superponen, las medias no son significativamente diferentes. Ahora bien, que los errores no se superpongan no significa que las medias sean diferentes significativamente. La forma de calcular la diferencia entre medias es la siguiente:

$$SE_{(\text{diferencia entre medias})} = \sqrt{SE_{m1}^{\ 2} + SE_{m2}^{\ 2}}$$

Esta medida sólo es útil cuando se aplica a una distribución normal o cuando la muestra es igual o mayor a 30.

Al analizar una variable usualmente se elige la media y el desvío estándar, o bien la mediana y el IQR. Sin embargo, siempre es mejor graficar la muestra y luego utilizar todas o varias de las medidas de tendencia central y dispersión.

3.2. Medidas de dispersión adimensionales

Las medidas nombradas hasta el momento se expresan en la misma unidad de la variable, es decir, si la variable toma valores de F0 (Hz) o segundos, su desvío estándar también será un valor de F0 o segundos. También existen medidas adimensionales que son útiles para comparar datos con diferentes unidades de medida:

- **Coeficiente de dispersión cuartil**: medida no-paramétrica adimensional que se obtiene al dividir el IQR por la suma del primer y tercer cuartil:

$$[(\text{cuartil}_3 - \text{cuartil}_1) / (\text{cuartil}_1 + \text{cuartil}_3)]$$

- **Coeficiente de variación**: un problema con la desviación estándar es que su tamaño depende de la media. Cuando los valores y, por lo tanto, también la media se incrementan, también aumenta la desviación estándar. Es por ello que no se pueden comparar desviaciones estándar de distribuciones con diferentes medias si no se normalizan primero. En cambio, al dividir la desviación estándar por su media se consigue el coeficiente de variación, una medida paramétrica que no se ve afectada por el aumento de los valores de la distribución.

$$Cv = \frac{\sigma}{\bar{X}}$$

4. Otras medidas

- **Asimetría** (***skewness***): una distribución puede tener mayor cantidad de valores por encima o por debajo de la media, lo que en un gráfico se percibe como una cola considerablemente más larga que

la otra (Figura 4). Para conseguir esta medida se debe calcular el "coeficiente de asimetría de Fisher", a partir de la sumatoria de la diferencia de cada valor con la media elevada al cubo y esto dividirlo por el cubo de la desviación estándar. Si la distribución tiene muchos valores por encima de la media, cuando estos se potencien al cubo van a resultar en grandes números positivos (asimetría derecha). Si la distribución tiene muchos valores por debajo de la media, el resultado va a ser negativo (asimetría izquierda). Esta es una medida adimensional, no tiene unidades. Una distribución simétrica posee una asimetría con valor 0 (por ende, una distribución normal posee una asimetría con valor 0, pero es importante resaltar que una asimetría con valor 0 no implica que la distribución sea normal).

$$\text{Asimetría} = \frac{\sum_{i=1} (x_i - \bar{x})^3}{\sigma^3}$$

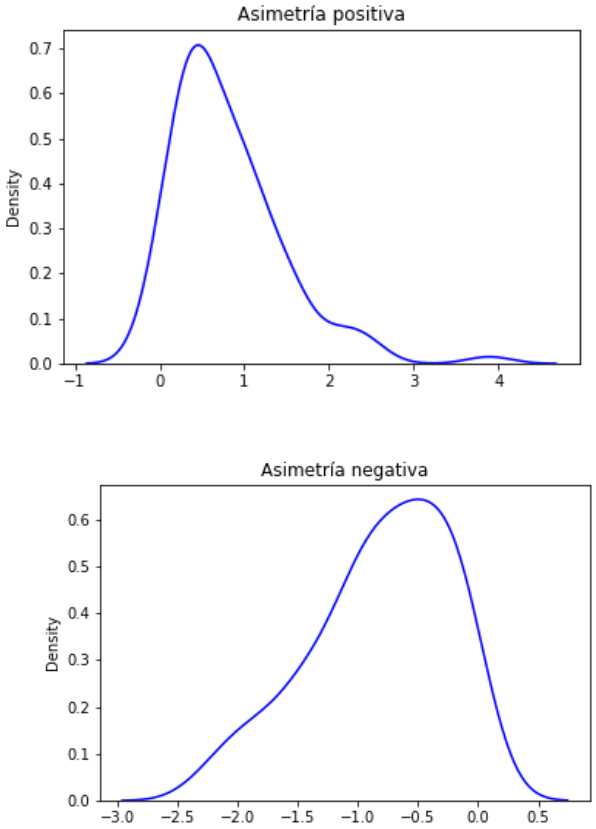


Figura 4. Representación de distribuciones asimétricas.

- **Curtosis**: se utiliza para saber si la distribución tiene un pico puntiagudo (leptocúrtica), redondeado (platicúrtica) o similar a una distribución normal (mesocúrtica) (Figura 5). La fórmula es igual que para la asimetría, pero se reemplaza el cubo por la cuarta potencia en numerador y denominador. Hay que restarle 3 para hacer una corrección, llamada "exceso de curtosis", y así se obtienen

valores positivos para una distribución leptocúrtica y negativos para una distribución platicúrtica. No tiene unidades. En una distribución mesocúrtica tendría un valor de 0.

$$\text{Curtosis} = \frac{\sum_{i=1} (x_i - \bar{x})^4}{\sigma^4} - 3$$

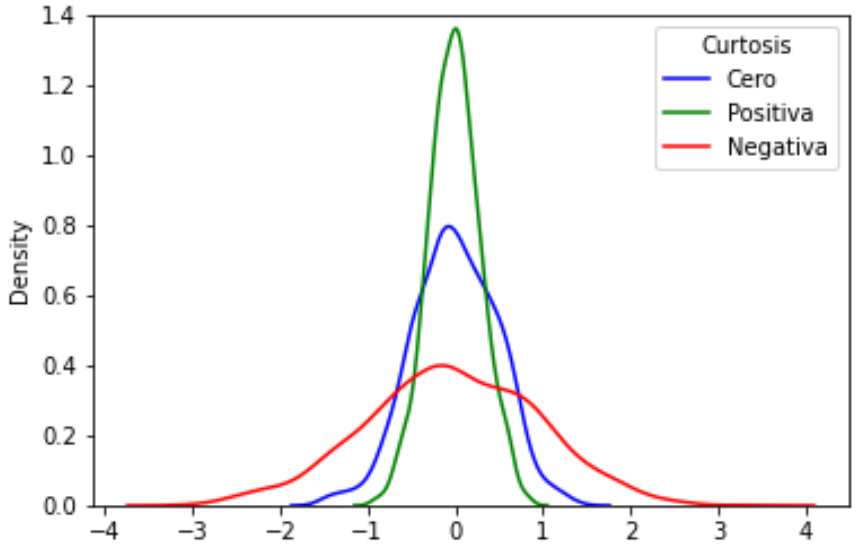


Figura 5. Distribuciones según sus valores de Curtosis. En el gráfico puede verse una curva leptocúrtica (la superior, en color verde), una curva mesocúrtica (la intermedia, en color azul) y una curva platicúrtica (la inferior, en color rojo).

- **Normalización**: no se pueden comparar valores o medidas que surgen de diferentes escalas, para esto es necesario normalizar o relativizar estos valores. Dos formas de hacerlo son:
  - **Centrado**: de cada valor individual se resta la media de la escala. Gráficamente, consistiría en centrar la media en 0.

$$C = X_i - \bar{x}$$

  - **Estandarización**: es la conversión de valores individuales en una forma estándar llamada "valores-*z*" (*z-scores*) o "valores-*t*" (*t-scores*).
- ***z-scores***: esto indica a cuántos desvíos estándar se encuentra un valor en relación a la media, y se utiliza cuando el tamaño de la muestra es igual o mayor a 30. El valor-*z* de *X* es la diferencia de *X* y la media, dividido por la desviación estándar de la población.

$$Z_{xi} = \frac{x_i - \bar{x}}{\sigma}$$

ISSN: 2796-8952

- ***t-scores***: esta forma estandarizada se utiliza cuando no se conoce la desviación estándar de la población y el tamaño de la muestra es menor a 30. En este caso, se calcula la diferencia entre el valor y la media, y se divide por el error estándar.

$$T_{xi} = \frac{x_i - \bar{x}}{\frac{\sigma}{\sqrt{n}}}$$

- **Intervalo de confianza**: permite saber qué tan bien se puede generalizar el resultado de una muestra a su población. Las siguientes fórmulas se basan en el error estándar, por lo que si los datos no se distribuyen normalmente o la muestra es demasiado pequeña, conviene usar otros métodos para computar el I.C.

  - **I.C. de la media**: provee un intervalo de valores aproximados, alrededor de los cuales se asume que no hay diferencia significativa entre este y la media poblacional. Para calcularlo debemos tener en cuenta el tamaño de la muestra:
    Si la muestra es mayor a 30, los límites del intervalo se calculan sumando y restando, respectivamente, a la media muestral el error estándar multiplicado cierta cantidad de veces. Esta última cantidad, representada en la fórmula como $z_{1\text{-}p/2}$, varía de acuerdo al porcentaje definido para el intervalo de confianza. Generalmente se utiliza un intervalo del 95%, en cuyo caso se multiplicará el error estándar por 1,96[3].

$$CI = (SE * Z_{\frac{1-p}{2}}) \pm \bar{X}$$

    El intervalo de confianza de 95% nos indica que si recolectáramos datos de infinitas muestras diferentes, el valor real de la media poblacional estaría dentro del intervalo de confianza en el 95% de esas repeticiones.
    Si se comparan los intervalos de confianza de dos muestras y estos no se superponen, las medias de las muestras son significativamente diferentes.

5. Estadística con dos variables: medidas de correlación

5.1. Medidas de correlación paramétricas

Cuando existe una dependencia entre dos variables se dice que están asociadas o que existe una correlación. Por ejemplo, en una correlación lineal, si *X* aumenta una cierta cantidad, *Y* aumentará o disminuirá una cierta cantidad proporcionalmente. Siempre que una variable nos ayude a predecir otra se puede hacer el camino inverso. Si *X* nos ayuda a predecir *Y, Y* nos ayudará a predecir *X*. Esto, sin embargo, no implica causalidad.

- **Correlación de Pearson** (***r***): antes de calcularla, tenemos que graficar las variables a fin de asegurarnos de que la relación entre estas sea lineal. Esta correlación se define como la covarianza

[3]El valor 1,96 es un valor constante para los intervalos de confianza del 95% que surge de partir del producto de restarle a 1 la probabilidad (*p*) que se quiere para el intervalo de confianza (en este caso 0,95 para un intervalo de 95%) y dividirlo por dos. La *z* refiere a *z-score*, debido a que se asume una distribución-z, es decir, una distribución normal cuya media es 0 y su desvío estándar 1.

ISSN: 2796-8952

de las dos variables, dividida por el producto de sus desviaciones estándar, y siempre toma un valor entre -1 y 1. En tanto que es un método paramétrico, la correlación de Pearson funciona mejor cuando las dos variables tienen una distribución normal. No tiene unidades y es sensible a los *outliers*.

- **Covarianza**: para cada dato se mide la diferencia entre su valor de *X* y la media de *X*; esto se multiplica por la diferencia entre su valor de *Y* y la media de *Y*. La covarianza es el promedio del producto de estas dos diferencias[4]:

$$\mathrm{Cov}_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

Entonces:

$$r = \frac{Cov_{xy}}{\sigma_x * \sigma_y}$$

El signo (+/-) de *r* define la dirección de la correlación, mientras que su valor absoluto define la fuerza de la correlación. Cuando *r* es igual a 0 no hay correlación entre las variables. La Tabla 1 muestra la interpretación de distintos valores de *r*.

Coeficiente de correlación *(r)*	**Etiqueta de correlación**	**Tipo de correlación**
0,7 < r ≤ 1	Muy alta	Un valor positivo de r representa una correlación positiva*: "cuando* X *aumenta…* Y *aumenta…"* o *"cuando* X *disminuye…* Y *disminuye…".*
0,5 < r ≤ 0,7	*Alta*	
0,2 < r ≤ 0,5	*Intermedia*	
0 < r ≤ 0,2	*Baja*	
r ≈ 0	*No hay correlación estadística (H_0)*	
0 >r ≥ -0,2	*Baja*	Un valor negativo manifiesta una correlación negativa*: "cuando* X *aumenta…* Y *disminuye…"* o *"cuando* X *disminuye…* Y *aumenta…".*
-0,2 >r ≥ -0,5	*Intermedia*	
-0,5 >r ≥ -0,7	*Alta*	
-0,7 >r ≥ -1	*Muy alta*	

Tabla 1. Interpretación de los valores de un coeficiente de correlación. Tomado de Gries (2013), traducción nuestra.

[4]En la fórmula el denominador es *n*-1 porque se hace una corrección, así que estrictamente no sería el promedio que se suele utilizar cotidianamente.

ISSN: 2796-8952

- **Coeficiente de determinación** (***r-squared***): sirve para resumir cómo se ajusta un modelo a los datos, es decir, qué tanto de la varianza de la variable dependiente es explicada por la variable independiente.
- **Regresión lineal**: otra forma de investigar la correlación es tratando de predecir valores de la variable dependiente con base en la independiente.[5]

5.2. Medidas de correlación no paramétricas

Cuando la relación entre variables no es lineal, es mejor utilizar alguna de las siguientes medidas no paramétricas (que también son más apropiadas para variables ordinales).

- **Rho (ρ) de Spearman**: mide la "monotonía" de una asociación. En una relación perfectamente monótona, si una variable aumenta la otra va a aumentar o disminuir consistentemente (pero no ambas). Si ambas se mueven en la misma dirección, rho = 1; si se mueven en direcciones opuestas, rho = -1. Se calcula utilizando el mismo método que *r* (covarianza dividida por el producto del desvío estándar), pero los datos primero son transformados en rangos. En este caso, el concepto de **rango** se refiere al orden de los valores de la escala (primero, segundo, etc). Los métodos no paramétricos usualmente involucran rangos, convirtiendo variables continuas a una escala ordinal. Esto hace más débil al método (se necesitan más observaciones), pero es más fuerte frente a *outliers*, asimetrías o distribuciones multimodales.
- **Tau (τ) de Kendall**: si unimos dos puntos cualesquiera de los datos con una línea recta, la *tau* de Kendall es la probabilidad de que esta línea tenga una pendiente positiva menos la probabilidad de que tenga una pendiente negativa. Este resultado va a ser entre -1 y 1; la *Tau* de Kendall tiende a ser menor que la *Rho* de Spearman, pero bastante similar.

6. Análisis de variables categóricas

Hasta ahora, muchas de las medidas descritas solo pueden aplicarse a variables continuas, pero no a variables categoriales, dado que precisan de valores numéricos.

- Variables nominales: no es posible calcular la media o la mediana de una variable nominal. Lo mismo sucede con el desvío estándar. La moda es el valor más frecuentemente utilizado. Para dar la dispersión de una variable nominal se puede usar el **índice de dispersión**, que es cercano a cero si la mayoría de los datos pertenece a una sola categoría y es igual a 1 si se distribuye en todas las categorías equitativamente. Si n es el total de observaciones, K el número de categorías y f describe la cantidad de datos en cada categoría, entonces el índice de dispersión es:

$$\frac{K*(n^2-\sum(f^2))}{n^2*(k-1)}$$

[5]La regresión lineal será abordada con mayor detalle en el siguiente volumen de estos cuadernos.

ISSN: 2796-8952

- Variables ordinales: se puede utilizar la mediana y algunas medidas relacionadas, como el IQR. Sin embargo, a menos que la variable tenga muchas categorías, esto no es muy útil.
- Variables binarias o dicotómicas: se puede representar las variables con 1 y 0, y calcular un tipo de media usando esos números (promedio de la cantidad de 1 o 0). A esto se le llama **media de una proporción** o "*p*". Para medir la dispersión se puede calcular el desvío estándar, pero el resultado no es independiente de la media, por esto el desvío estándar de una proporción no es muy útil.

Otra medida de correlación es el **Coeficiente *phi*** que determina que una variable independiente y una dependiente son intercambiables. Se utiliza para tablas de contingencia[6] si al menos una de las variables es nominal, o ambas son dicotómicas. El resultado es un número entre -1 y 1 cuya interpretación es similar a la de la *r* de Pearson.

$$phi = \sqrt{p * (1 - p)}$$

En relación a la correlación de dos variables, Kendall y Spearman son apropiadas para variables ordinales, o en el caso de una ordinal y una continua.

7. La importancia de graficar la distribución

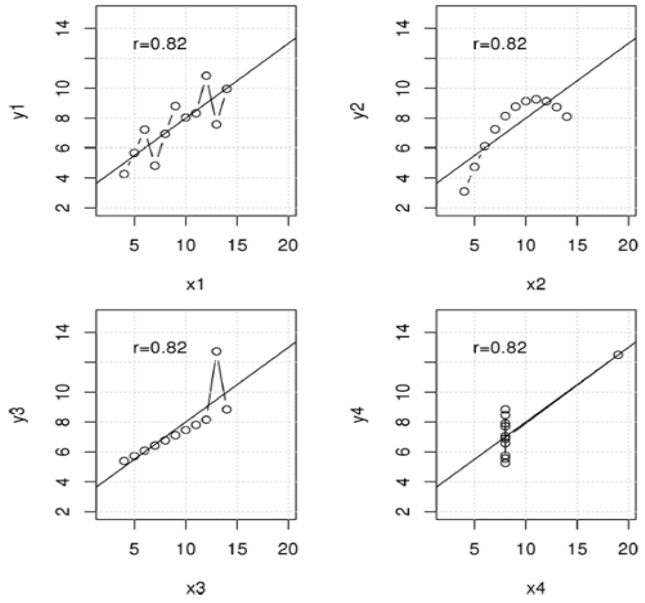


Figura 6. Diferencias de distribución, tomado de Gries (2013, p. 155).

Finalmente, la media, la varianza y la línea de regresión pueden ser las mismas para dos conjuntos de datos, pero esto no significa que su distribución sea la misma. En los gráficos superiores de la Figura 6

[6] Tabla de doble entrada donde se detallan las frecuencias, es decir, la cantidad de apariciones de una variable (en relación con otra). Se emplean para registrar y analizar la asociación entre dos o más variables. Por ejemplo:

	Respuestas correctas	Respuestas incorrectas	Total
Grupo A	35	15	50
Grupo B	22	28	50
Total	57	43	100

ISSN: 2796-8952

tenemos una situación donde *X* e *Y* están relacionados en una forma curvilínea, por lo que utilizar el método de correlación lineal no tiene mucho sentido. En los dos gráficos inferiores, los *outliers* tienen una gran influencia en el valor de *r* y la regresión lineal. A simple vista podemos notar que estas cuatro distribuciones son diferentes, a pesar de presentar el mismo valor de *r* (0,82). Estos cuatro gráficos ejemplifican lo indispensable que es inspeccionar visualmente los datos.

8. Referencias

Gries, S. T. (2013). Descriptive statistics. En *Statistics for linguistics with R* (2da ed., pp. 102-156). Walter de Gruyter GmbH.

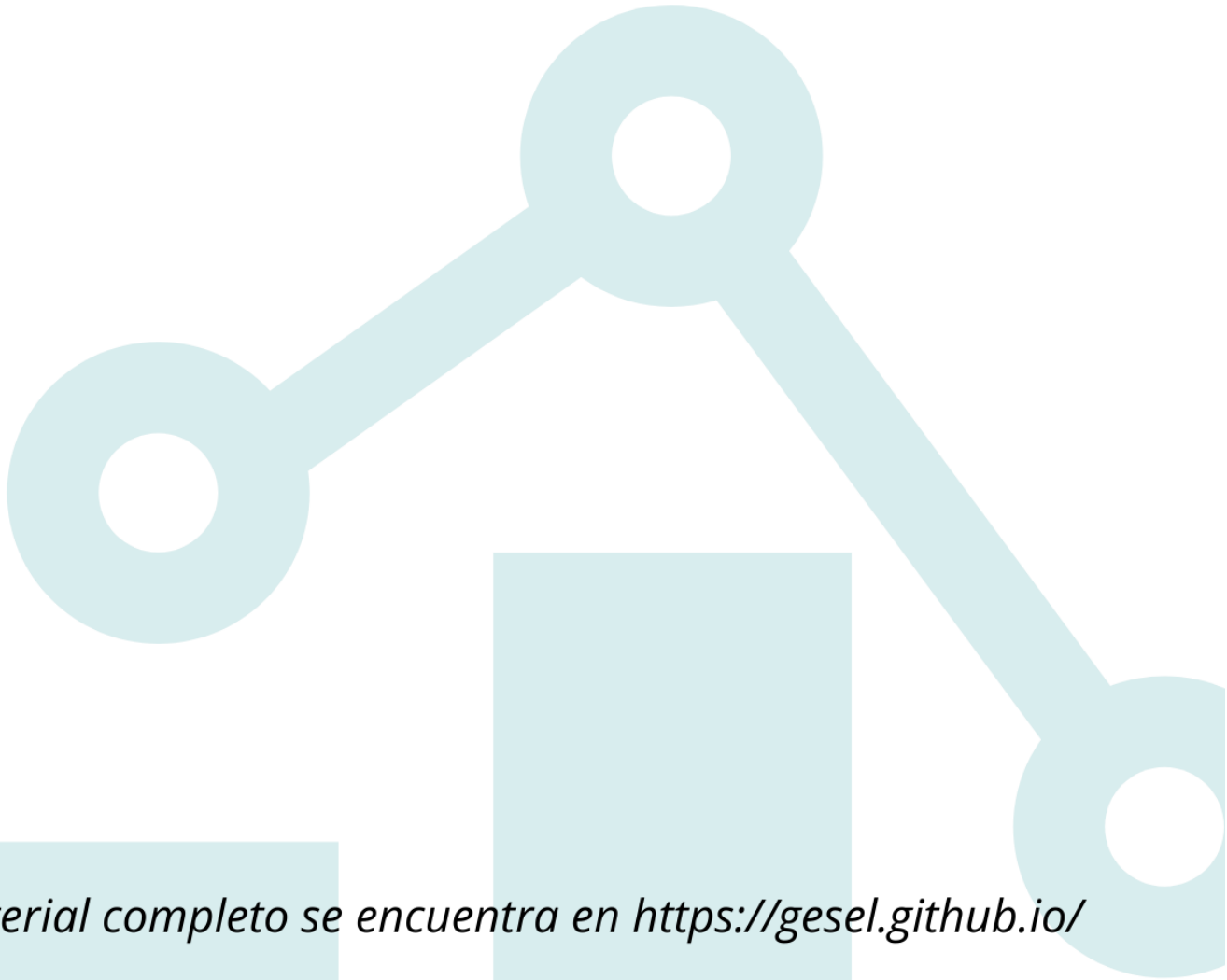
9. Bibliografía sugerida

Gries, S. T. (2013). Descriptive statistics. En *Statistics for linguistics with R* (2da ed., pp. 102-156). Walter de Gruyter GmbH.

Johnson, D. E. (2013). Descriptive statistics. En R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 288–315). Cambridge University Press.

Levshina, N. (2015). “Chapter 3. Descriptive statistics for quantitative variables” en Levshina, N. *How to do Linguistics with R. Data exploration and statistical analysis* (pp. 41-68). John Benjamins Publishing Company.

ISSN: 2796-8952



El material completo se encuentra en https://gesel.github.io/