Estadística descriptiva

1. INTRODUCCIÓN

La estadística descriptiva generalmente constituye el segundo paso en un análisis cuantitativo, transforma la información en un número menor y más manejable, abstrayendo 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 estudiada sino que se utiliza como material exploratorio del que pueden surgir preguntas que conduzcan a una hipótesis.

El primer paso en el análisis es identificar el/los tipo(s) de variable(s) que estamos tomando, ya que esto definirá los métodos y tipos de medidas estadísticas que utilizaremos. El segundo paso es examinar su distribución.

1.1. Variables

- *Variables continuas*: son aquellas que tienen infinitas posibilidades de valores dentro de un rango. En las variables verdaderamente continuas dos datos nunca toman el mismo valor. La frecuencia fundamental, formantes, tiempos de reacción, frecuencias lexicales son algunos ejemplos.

a. *Variable de razón:* es una variable que toma valores de una escala de X, cuyos intervalos pueden ser medidos (se puede decir que la diferencia entre $5 y $6, es igual a la diferencia entre $45 y $46). Además la escala X tiene un cero definido que marca la inexistencia de X ($0 significaría que no tenés dinero), y por supuesto no cuenta con valores negativos.
b. *Variable de intervalo:* es una variable que toma valores de una escala cuyos intervalos pueden ser medidos. Esta escala tiene un cero arbitrario, cuyo valor no significa la inexistencia de dicha variable, por ejemplo: 0º no significa que no haya temperatura, por lo tanto esta escala también puede tomar valores negativos (podemos considerar aquí los siglos, toman A.C como negativos y D.C como positivos. Un siglo cero no implicaría la inexistencia del tiempo).

- *Variables discretas*: Una variable discreta es aquella que solo toma cierto número de valores. Por ejemplo, la edad de un sujeto (si sólo se consideran los años) o la cantidad de respuestas correctas/incorrectas, o de palabras. Suelen utilizarse como la variable independiente.

a. *Variables ordinales*: son aquellas que 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.

b. *Variables nominales*: tienen dos o más categorías sin un orden natural. Al no estar cuantificadas no pueden realizarse operaciones matemáticas con ellas, y además no se puede asignarles un orden. (Un ejemplo sería las clases de palabras sustantivo, verbo, adjetivo, adverbio).

c. *Variables binarias*: tienen dos categorías opuestas. Puede ser presencia/ausencia de un rasgo y suelen cuantificarse como 0 y 1. Esta cuantificación puede usarse para aplicar algunos procedimientos de las variables continuas.

- *Variable numérica:* es una variable a la que se le otorga un número y puede ser medida. También se las llama variables cuantitativas. Un ejemplo sería la cantidad de errores o aciertos en una prueba.

1.2. Distribución

Hace referencia a la probabilidad de los valores de una variable. A veces se espera que se aproxime a una distribución matemática. Una de estas distribuciones se llama *distribución normal* (Figura 1) y se aplica a las variables

continuas. También se la conoce como "campana de Gauss" por su forma y está definida por dos parámetros: la media y el desvío estándar.

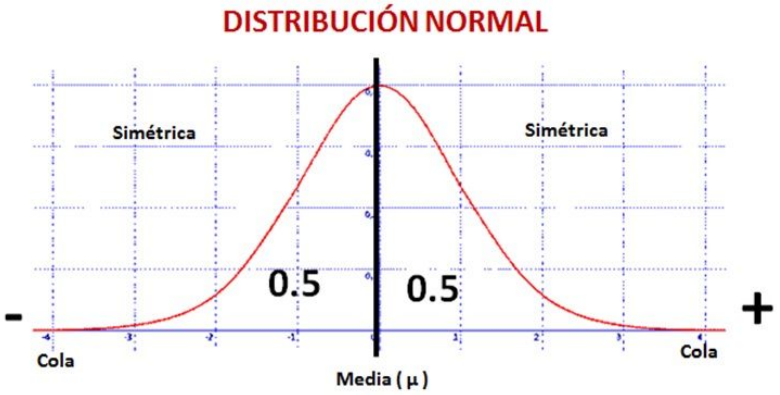

Figura 1: gráfico de una distribución normal.

Por supuesto, los datos reales no tienen porqué estar distribuidos como se espera, pero ciertos métodos estadísticos usan esta distribución como base para analizar datos, como en el caso de las estadísticas paramétricas. Existen también métodos llamados "robustos" que no asumen la distribución de los datos, como los métodos no-paramétricos (los más apropiados para datos con variables ordinales o nominales).

Por esto es necesario conocer la distribución de nuestros datos 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

Estás 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).

a. *Media* ($\overline{X}$): se obtiene sumando todos los valores de la variable y dividiendo por N (siendo N la cantidad de total de observaciones) . Una variable solo tiene una media, que se ve muy afectada por la presencia de outliers (un valor que cuya magnitud es muy diferente a la del resto de los valores de una muestra, ver figura 3).

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

b. *Media geométrica*: es la raíz N del producto de todos los valores. Una variable solo tiene una media geométrica. Se ve muy afectado por la presencia de outliers.

$$\sqrt[n]{x_1 * x_2 * \ldots x_n}$$

c. *Mediana*: Se obtiene ordenando todos los valores de una variable, desde el más chico al más grande, la mediana se ubica donde se llega a la mitad de los casos. Al tomar por ejemplo la figura 2, la media se ubica en $47,500, si empezamos a contar desde el valor más pequeño (0), al llegar a la media habremos contado la mitad de los casos. Si los valores son pares, la mediana se calcula sacando el promedio de los dos valores centrales. Se utiliza para variables ordinales. Una variable solo tiene una mediana. No se ve muy afectada por la presencia de outliers.

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

Si la data tiene muchos outliers, o está sesgada, se prefiere la mediana. También se utiliza la mediana para variables ordinales. Para variables nominales se define la moda.

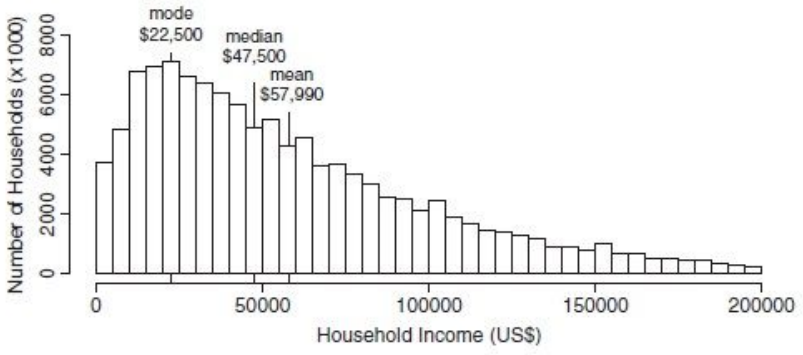

Figura 2: Moda, mediana y media. Tomado de Podesva-Sharma 2013.

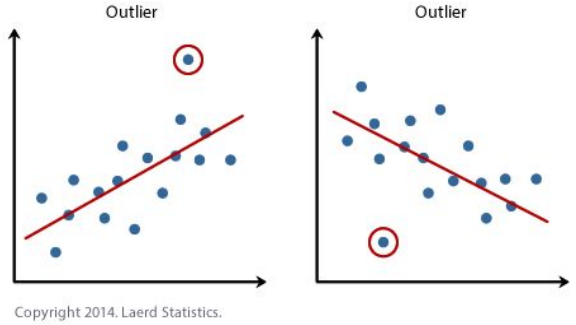

Figura 3: Outliers.

3. MEDIDAS DE DISPERSIÓN

Al hacer estadística descriptiva resulta necesario medir cómo se comporta la variable. Las medidas de dispersión nos dicen qué tanto varían los valores con respecto a la tendencia central. Hay varias medidas que toman las mismas unidades que la variable:

a. *Rango*: Es el valor máximo menos el valor mínimo de una variable. Es sensible a outliers.

b. *Rango intercuartil (IQR)*: El percentil es una medida de posición que indica, una vez ordenados los datos de menor a mayor, el valor por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo de observaciones. Por ejemplo, el percentil 20º es el valor bajo el cual se encuentran el 20 % de las observaciones.. 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 es el valor del percentil P (es decir, que el P% de mis datos va a ser *menor* al valor de I). Si el valor de I es decimal, se redondea para arriba.

Cuando esos porcentajes son el 25, 50 (equivalente a la mediana), 75 y 100 se les llama cuartiles. Si se hace la diferencia del tercer y primer cuartil, se obtiene lo que se llama "rango intercuartiles" (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.

c. *Desviación estándar*: es la medida más utilizada, se obtiene sacando la raíz cuadrada de la varianza. La *varianza* es la sumatoria del cuadrado de las distancias de cada valor con de la media, todo dividido por N. Entonces:

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

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

Cuando la muestra es grande tenemos que cambiar el divisor por n-1 en la varianza y en la desviación estándar, a esto se le llama "corrección de Bessel" y se usa porque no queremos sobreestimar nuestra muestra con respecto a la población. Dos distribuciones pueden tener medias similares, pero desvíos muy diferentes.

Al analizar una variable continua usualmente se elige entre la media y el desvío estándar o la mediana y el rango intercuartil. Sin embargo, siempre es mejor graficar la muestra y luego utilizar todas o varias de las medidas de tendencia central y dispersión que venimos trabajando.

Las medidas de dispersión descritas hasta ahora están todas expresadas en la misma unidad de la variable. También hay medidas adimensionales que son útiles para comparar datos con diferentes unidades de medida. Un ejemplo de estas es la medida paramétrica del coeficiente de variación. Una medida no-paramétrica adimensional es el coeficiente de dispersión cuartil, que se obtiene de:

[(3 cuartil- 1 cuartil)/ (1 cuartil+ 3 cuartil)]

d. *El coeficiente de variación:* Un problema con la desviación estándar es que su tamaño depende de la media. Cuando los valores, y por lo tanto la media, son incrementados también aumenta la desviación estándar. Es por ello que no se pueden comparar desviaciones estándar de distribuciones con diferentes medias sino las normalizamos primero. En cambio, al dividir la desviación estándar por su media se consigue el coeficiente de variación, el cual no se ve afectado por el aumento de los valores de la distribución.

e. *El error estándar:* Se define como la desviación estándar de las medias de muestras igualmente grandes de una misma población. La media de una muestra no es el mismo que el promedio de toda la población, a menos que la muestra sea perfectamente representativa. Una forma de saber qué tan representativo es el promedio de tu muestra en relación con toda la población es computar el promedio de varias muestras semejantes (tomadas al azar, pero del mismo tamaño) y calcular la desviación estándar de todos esos promedios. Este es el error estándar:

$$SE_{mean} = \sqrt{\frac{var}{n}} = \frac{SD}{\sqrt{n}}$$

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 muestras) más chico será el error estándar. Cuando comparás las medias de dos muestras aproximadamente iguales y 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.

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

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

4. OTRAS MEDIDAS

a. *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. Para conseguir esta medida se debe calcular la diferencia del valor con el promedio al cubo y dividirlo por el cubo de la desviación estándar. Si la distribución tiene muchos valores por encima de la media, cuando estos se potencien 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 tendría un valor 0 (por ende una distribución normal también sería 0, pero que sea 0 no implica que sea una distribución normal).

$$asimetría = \frac{(x_i - \bar{x})^3}{sd^3}$$

b. *Kurtosis*: Se utiliza para saber si se tiene un pico puntiagudo (leptocúrtica) o redondeado (platicúrtica). La fórmula es igual que para la asimetría pero se reemplaza el cubo por la cuarta potencia en numerador y denominador. Hay que restarle tres para hacer una corrección y así se obtiene valores positivos para una distribución leptocurtica y negativos para una distribución platicúrtica. No tiene unidades. Una distribución simétrica tendría un valor 0.

$$\textbf{Kurtosis=}\ \frac{(x_i - \bar{x})^4}{sd^4}$$

c. *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*: se resta de cada valor individual la media de la escala. Gráficamente, consistiría en centrar la media en 0.

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

- *Estandarización:* la estandarización es la conversión de valores individuales en una forma estándar llamada "valores-z" (z-scores) o "valores-t" (t-scores).

  a- *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 un valor es la diferencia del valor y la media, dividido por la desviación estándar de la población.

$$\textbf{Z-score}_{xi} = \frac{\bar{X} - x_i}{sd}$$

  b- *T-scores*: esta forma estandarizada se utiliza cuando no se conoce la desviación estándar de la población, sino que se calcula la diferencia entre el valor y la media y se lo divide por la desviación estándar de la muestra, que a su vez fue dividida por la raíz de N.

$$\textbf{T-score}_{xi} = \frac{\bar{X} - x_i}{sd/\sqrt{n}}$$

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

- *I.C. de la media aritmética*: Provee un intervalo de valores aproximados de la media de la muestra, alrededor de los cuales asumiremos que no hay diferencia significativa con la media muestral. De la expresión "diferencia significativa" se sigue que el intervalo es típicamente definido como 1 menos el nivel de significancia (p=0.05), por ejemplo: 1-0.05=0.95. Para sacar este intervalo se utiliza el valor del error estándar, el cual se multiplica por el T-score. Luego se resta y se suma la media, de modo de conseguir los dos valores del intervalo.

$$\textbf{CI=}\ \bar{X} \pm t * SE$$

Así se consigue un rango de valores del cual podes estar 95% seguro que contiene el verdadero valor de un parámetro en la población.

Si se compara el intervalo de confiabilidad de dos muestras y estos no se superponen las medias de las muestras son significativamente diferentes.

- *CI de los porcentajes:* se multiplica el z-score por el error estándar y luego se le suma y resta ese porcentaje del que se quiere conseguir el intervalo.

$$CI_a = a \pm Z*SE$$

5. ESTADÍSTICA CON DOS VARIABLES: MEDIDAS DE CORRELACIÓN

Cuando existe una dependencia entre dos variables se dice que estas están asociadas o que existe una correlación. Por ejemplo en una asociación lineal, si X aumenta una cierta cantidad, Y aumentará una cierta cantidad. 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.

a. *Correlación de Pearson (r)*: Antes de calcularla, tenemos que graficar la variable a fin de asegurarnos de que sea lineal. Esta correlación se define como la covarianza 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 y esto se multiplica por la diferencia de su valor de Y y de la media de Y. La Covarianza es el promedio de estos productos:

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

Entonces:

$$r = \frac{Cov_{xy}}{sd_x*sd_y}$$

El signo (+/-) define la dirección, mientras que el número absoluto define la fuerza de la correlación. Cuando r=0 entonces no hay correlación entre las variables.

Table 18. Correlation coefficients and their interpretation

Correlation coefficient	Labeling the correlation	Kind of correlation
$0.7 < r \leq 1$	very high	positive correlation:
$0.5 < r \leq 0.7$	high	the more/higher …, the more/higher …
$0.2 < r \leq 0.5$	intermediate	the less/lower …, the less/lower …
$0 < r \leq 0.2$	low	
$r \approx 0$	no statistical correlation (H_0)	
$0 > r \geq -0.2$	low	negative correlation:
$-0.2 > r \geq -0.5$	intermediate	the more/higher …, the less/lower …
$-0.5 > r \geq -0.7$	high	the less/lower …, the more/higher …
$-0.7 > r \geq -1$	very high	

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

b. *Coeficiente de determinación (*s r-squared): sirve para resumir cómo se ajusta un modelo, es decir qué tanto de la varianza de la variable dependiente es explicada por la variable independiente.

c. *Regresión lineal:* Otra forma de investigar la correlación es tratando de predecir valores de la variable dependiente con base en la independiente.

En su forma más simple es tratar de dibujar una línea recta que represente la "nube de puntos" (scattercloud) de la mejor manera. Esta recta está definida por una ecuación con dos parámetros: una intersección A (Y cuando x=0), y una pendiente B.

La ecuación regresiva es más útil para el rango de valores que cubren los valores observados, va a ser menos confiable para valores no observados. La ecuación también hace predicciones sin sentido, porque las hace en base a números no a la realidad (ejemplo: puede calcular los TR para palabras de -9 letras).

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

d. *Rho (ρ) de Spearman:* Mide la "monotonía" de una asociación. En una relación perfectamente proporcional, 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 y 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. Los rangos sólo miran el orden de los números, no su valor. 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 y sesgos, o distribuciones multimodales.

e. *Tau (τ) de Kendall:* Si unimos dos puntos cualquiera 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; Tau de Kendall tiende a ser menor que la Rho de Spearman, pero bastante similar.

6. ANÁLISIS DE VARIABLES CATEGORIALES

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

a. *Nominales:* No es posible calcular la media o la mediana de una variable nominal con tres o más categorías, lo mismo sucede con el desvío estándar. La moda es el valor más frecuente.

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, *K* el número de categorías y *F* es el vector de la cantidad por cada categoría, entonces el índice de dispersión es:

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

b. *Ordinales:* se puede utilizar la mediana y algunas medidas relacionadas, como el rango intercuartiles. Sin embargo, a menos que la variable tenga muchas categorías, esto no es muy útil.

c. *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 pueden tomar esos 1 y 0 y calcular una desviación estándar, pero el resultado no es independiente de la media, por esto el desvío estándar de una proporción no es muy útil.

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

Otra medida de correlación es el C*oeficiente phi* que determina que una variable independiente y una dependiente son intercambiables. Se utiliza para tablas de contingencia[1] 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 al de *r* de Pearson.

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

7. LA IMPORTANCIA DE GRAFICAR LA DISTRIBUCIÓN

Finalmente, las medias y varianzas, junto con la línea de regresión pueden ser los mismos para dos variables, pero esto no significa que su distribución sea la misma.
En los gráficos superiores de la figura 4 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 tus datos.

[1] *"Contingency tables (also called crosstabs or two-way tables) are used in statistics to summarize the relationship between several categorical variables. A contingency table is a special type of* ***frequency distribution table****, where two variables are shown simultaneously."* (Stephanie, en Statistics How To)

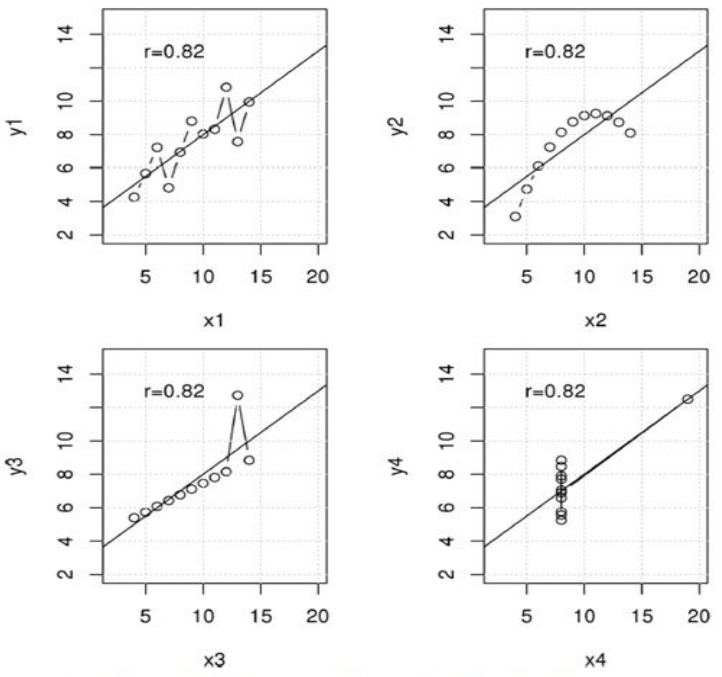

Figure 38. The sensitivity of linear correlations: the Anscombe data

Figura 4. Diferencias de distribución, tomado de Gries 2013.

8. BIBLIOGRAFÍA

- Gries, S. T. (2013) "Descriptive statistics" *Statistics for linguistics whit R.*
- Johnson D. E. (2017) "Descriptive statistics" en Podesva, R. J. & Sharma, D. *Research methods in linguistics.*
- Stephanie (2013), "Contingency Table: What is it used for?" en: https://www.statisticshowto.datasciencecentral.com/what-is-a-contingency-table/