Introducción a la Estadística descriptiva y exploratoria

Mg. Ciro Ivan Machacuay Meza

Facultad de Economía

1 de marzo de 2025

Índice de contenidos

1.- Definición de Estadística

Es la ciencia que proporciona un conjunto de métodos, técnicas o procedimientos para recopilar, organizar (clasificar, agrupar), presentar y analizar datos con el fin de describirlos o de realizar generalizaciones válidas sobre los mismos. Así, se tiene la siguiente división:

- **Descriptiva**: Conjunto de métodos estadísticos que se relacionan con el resumen y descripción de los datos, como tablas, gráficos y el análisis mediante algunos cálculos.
- **Inferencial** : Conjunto de métodos con los que se hacen las generalizaciones o la inferencia sobre una población utilizando una muestra. La inferencia puede contener conclusiones que pueden ser no ciertas en forma absoluta, por lo que es necesario que éstas sean dadas con una medida de confiabilidad, que es la *probabilidad*. Estas dos partes de la estadística no son mutuamente excluyentes, ya que para utilizar los métodos de la inferencia estadística, se requiere conocer los métodos de la estadística descriptiva.

2.- Definición de Población

En forma general, en Estadística se denomina población a un conjunto de elementos que consiste de personas, objetos, etc., en los que se puede observar o medir una o más características de naturaleza cualitativa o cuantitativa. A cada elemento se le denomina **unidad elemental** o **unidad estadística**.

La población es definida por una tarea o investigación estadística a realizarse. Y como la medición de la característica especificada por la investigación se hace a cada unidad elemental, en este caso se define a la población como la totalidad de valores posibles de una característica particular especificada por la investigación estadística. En este sentido, la población consiste en un conjunto de datos estadísticos que se reúnen de acuerdo con la formulación de una investigación estadística o con la definición de la población específica.

Por ejemplo, los empleados de una empresa en un día laborable, constituyen una población en la que cada empleado (unidad estadístico) tiene muchas características a ser observadas como género, estado civil, lugar de procedencia, grado de instrucción, etc.

Se denomina **parámetro** a una medida descriptiva que resume una característica definida en la población, tal como la media (μ) o la varianza (σ^2), calculada a partir de los datos observados de toda la población.

Por el número de elementos que la componen, la población se clasifica en finita o infinita. La población es **finita** si tiene un número finito N de elementos. En caso contrario la población es **infinita.** En la práctica, una población finita con un número grande de elementos se considera como una población infinita.

Después de definir la tarea o investigación estadística a realizar, se debe decidir entre investigar a toda la población o sólo una parte de ella. El primer procedimiento es denominado **censo** y el segundo es llamado **muestreo**.

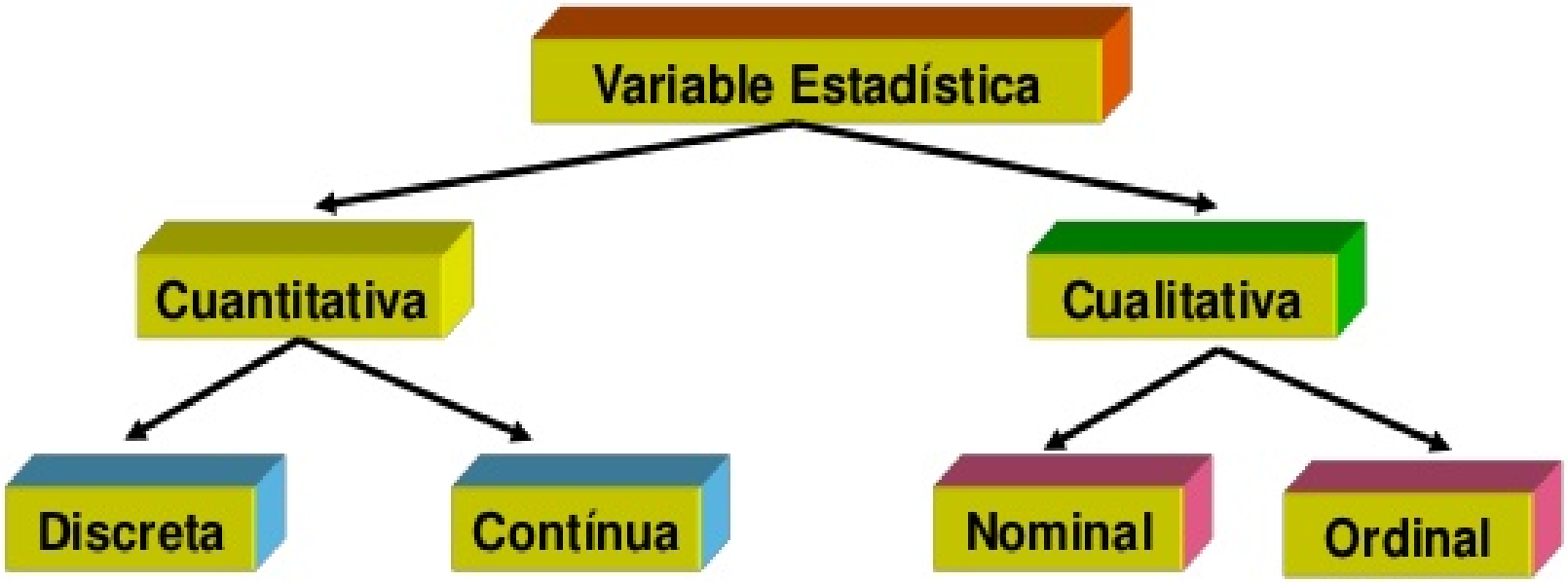
3.- Definición de Muestra

Parte de la población seleccionada de acuerdo con un plan o regla para obtener información acerca de la población de la cual proviene. La muestra debe ser seleccionada de manera que sea representativa de la población.

Se denomina **estadística, estadísgrafo, estadístico** a una medida descriptiva que resuma una característica definida en una muestra aleatoria, tal como la media $(\overline{x})$ o la varianza (s_n^2), calculada a partir de los datos observados en la muestra. Es importante tener en cuenta, si el análisis estadístico se está haciendo con una muestra o con una población. En ambos casos las medidas descriptivas se calculan del mismo modo. Para diferenciarlos, los parámetros se denotan por letras griegas.

4.- Variables estadísticas

La característica que se observa o mide en las unidades estadísticas de una población tiene diversos valores de naturaleza cualitativa o cuantitativa. Por ejemplo, la característica *género* tiene dos valores (masculino, femenino); sin embargo, el *peso* tiene infinitos valores.
Se denomina *variable estadística* a una característica definida en una población y que asume por lo menos dos valores. Estos pueden ser de cualidad o de cantidad.



Las variables se clasifican en :

- **Cualitativa :** Es la característica cuyos valores son cualidades. Estos valores están en el nivel de escala nominal u ordinal. Por ejemplo: género, profesión, estado civil, grado de instrucción, etc. Con las cualidades, aun cuando estén codificados, no se pueden aplicar estadísticos de resumen (medidas de tendencia central, dispersión).

- **Cuantitativo :** Es la característica cuyos valores son numéricos. Por ejemplo, temperatura, números de hijos, ingresos mensuales, tiempo de vida útil, etc. A su vez, estas se clasifican en :
  a. **Discretas :** Es aquella entre cuyos valores posibles no admite otros. Por ejemplo, el número de hijos admite posibles valores como $0, 1, \ldots,$ pero no entre ellos (1,25).
  b. **Continuas :** Es aquella que puede tomar cualquier valor entre dos valores de la variable, por ejemplo : salario, tiempo, peso, volumen, longitud. La variable continua es descrita por : *"Si a y b son números reales tales que $a < b$, entonces existe un número real c tal que $a < c < b$. Entonces, entre a y c existe otro."*

Los valores de las variables, además de ser cualidad o cantidad, definen niveles de medición de las unidades estadísticas. Las escalas de medición corresponde a los distintos niveles de valores que las variables estadísticas asignan a las unidades estadísticas en estudio. Los niveles de las escalas de medición son:

- **Nominal**
- **Ordinal**
- **De intervalos**
- **De razón**

- **Nominal** : Se dice que los valores de una variable estadística están en el nivel de escala nominal si estos **sólo clasifican a las unidades estadísticas en iguales o diferentes**. Tales valores cualitativos son como etiquetas que la variable asigna a las unidades estadísticas haciéndolas iguales entre sí o diferentes. Si se asignara números a estos valores cualitativos, con estos no es posible realizar operaciones aritméticas.

  Por ejemplo, la variable género, definida en una población de personas, asigna a éstas los valores *masculino* y *femenino*. Los valores de estas variables están en escala nominal por que sólo *diferencia a las personas por su género y a todas las que están en un mismo género las hace iguales*. Si se le asignara un *cero* al género masculino y *uno* al género femenino, no es posible obtener resúmenes estadísticos de la variable. Solo se puede decir que el símbolo *0* es distinto al símbolo *1*. No es válido decir que $1 > 0$ o $0 < 1$.

  El método estadístico con datos obtenidos en escala nominal consiste básicamente en obtener el número (o porcentaje) de casos en cada modalidad y obtener la moda (valor de mayor frecuencia).

- **Ordinal** : Se dice que los valores de una variable estadística están en el nivel de escala ordinal si están en escala nominal y si además ordenan a las unidades estadísticas por la característica definida que se observa. Los valores cualitativos de una variable en escala ordinal son los resultados de un criterio para ordenar a las unidades estadísticas. Si se asignaron números a tales valores, con estos, no es posible realizar operaciones matemáticas. Solo son válidas las relaciones de igualdad, de no igualdad y de orden.
  Por ejemplo, la variable *nivel socioeconómico*, definida en una población de hogares, contiene un criterio que genera sus valores : bajo, medio, alto. Los diferencia a los hogares por su ingreso económico o las hace iguales a todos los hogares en una misma categoría, además estas unidades se ordenan por su nivel de ingreso monetario, es decir, un hogar en el nivel bajo tiene un ingreso menor que un hogar en el nivel medio y este a su vez un ingreso menor de un hogar en el nivel alto.

- **Escala de intervalos** : Una escala de intervalos es una escala ordinal que asigna a las unidades estadísticas valores numéricos que son mediciones realizadas con respecto a un cero arbitrario (o móvil). Ese cero no es real o absoluto, pues no mide la ausencia total de las características que se observa en la unidad estadística.
  Por ejemplo, la variable *altitud* definida en una población de ciudades, tiene valores numéricos que son mediciones hechas de alturas con respecto al nivel del mar. El nivel mar es un *cero elegido arbitrariamente*. Este cero no significa ausencia total de altura. Otro ejemplo es la calificación de pruebas (conocimientos, aptitud). La calificación cero no significa ausencia total de las característica que se mide en las unidades estadísticas. Con los valores se puede comparar la diferencia de las mediciones de dos unidades estadísticas con otra diferencia. Esto es, si x_1, x_2 y x_3 son valores de X se verifica:

$$\frac{x_3 - x_1}{x_2 - x_1} = c$$

Con los valores de esta escala son válidas pues, las relaciones de igualdad, de no igualdad y de orden. Además, son válidos las operaciones de adición y sustracción entre los valores de la escala, y la multiplicación y división entre las diferencias de dos valores de la escala. Pero, no es válida la multiplicación y división entre los valores mismos de la escala.

- **Escala de razón** : Es una escala de intervalo que asigna a las unidades estadísticas valores numéricos que son mediciones realizadas con respecto a un cero real. Este cero significa ausencia total de la característica que se observa. Los valores de esta escala se obtienen en general, por mediciones que son conteos (variables discretas) o por mediciones continuas, tales como longitud, peso, volumen, tiempo, unidades monetarias, etc. Además, con estos valores se pueden comparar cuántas veces la medida de una unidad estadística es igual a la medida de otra unidad estadística. Sea x_1 y x_2 dos valores de X se tiene:

$$\frac{x_2}{x_1} = c$$

Son válidas las relaciones de igualdad, de no igualdad, de orden y todas las operaciones matemáticas.

Medidas de tendencia central

Los datos organizados en una distribución de frecuencias destacan sus características más esenciales, como marcas de clases, centro, forma de distribución (asimétrica, simétrica), etc. Sin embargo, los indicadores que describen a los datos en forma más precisa, deben calcularse. Estos indicadores resumen los datos en medidas descriptivas que se refieren a la centralización o posición, a la dispersión o variación, a la simetría, y a la curtosis de los datos. Las medidas de tendencia central, denominados también promedios, ubican el centro de los datos, como la media aritmética, la media geométrica, la media armónica y la mediana.

1.- Mediana

Es el número, que separa a la serie de datos *ordenados* (en forma creciente o decreciente) en dos partes de igual número de datos. La mediana es el percentil 50 de los datos observados no agrupados o agrupados por intervalos. Así, la mediana es la medida promedio que depende del número de orden de los datos y no de los valores de los mismos, por lo que, no afectan los valores aislados grandes o pequeños.

Para realizar el cómputo de la mediana de n valores no agrupados de una variable cuantitativa X se siguen los siguientes pasos:

- Se ordenan los datos en forma creciente.
- Se ubica el valor central. Si n es impar, la mediana es el dato ordenado del centro. Pero si n es par, la mediana es la semisuma de los dos valores ordenados centrales.

Algunas propiedades de la mediana son:

- La mediana solo depende del número de datos ordenados y no del valor de los datos. Por lo tanto, no es sesgada por algún valor aislado grande o pequeño.
- La mediana puede ser calculada para distribuciones de frecuencia con intervalos de diferente amplitud, siempre que se pueda determinar el límite inferior L_i del intervalo que contiene a la mediana.
- La mediana puede ser calculada para variables con valores en escala ordinal.

2.- Moda

La moda de una serie de datos es el valor M_o que se define como el dato que ocurre con mayor frecuencia. En la distribución de frecuencias por intervalos la moda se ubica en el intervalo que tiene la mayor frecuencia.
La moda de una función cualquiera es el valor de la variable en el que existe un máximo absoluto (o dos o más máximos relativos iguales).
La moda no siempre existe, y si existe, no siempre es única.
El empleo de la moda como medida promedio puede estar justificado cuando se quiere señalar el valor más común de una serie de datos o se precise rápidamente de una medida promedio y no haya tiempo de calcular otras. Por ejemplo, los comerciantes se stockean con productos que están de moda.

Si se presenta información en datos agrupados por intervalos, el cálculo de la moda se realiza con la siguiente ecuación:

$$M_0 = L_i + \frac{d_1}{d_1 + d_2} \cdot A$$

Donde L_i es el límite inferior del intervalo modal, $d_1 = f_i - f_{i-1}$, $d_2 = f_i - f_{i+1}$ y A es la amplitud del intervalo modal.

3.- Media aritmética

Es el valor numérico que se obtiene dividiendo la suma total de los valores observados de una variable entre el número de observaciones. Para valores de una variable X observados de una muestra, la media aritmética será denotada por $\overline{X}$, siendo esta:

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

Algunas propiedades de la media aritmética son:

- La suma total de n valores cuya media es $\overline{x}$ es igual a $n\overline{x}$.
- Si a la variable X se le realiza una transformación lineal $Y = aX + b$, se puede verificar que $\overline{y} = a\overline{x} + b$.
- La suma algebraica de las desviaciones de n datos x_i con respecto a su media $\overline{x}$ es igual a cero. Es decir:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

- La suma de los cuadrados de las desviaciones de n datos con respecto a su media es mínima. Para datos no tabulados se tiene:

$$\sum_{i=1}^{n}(x_i - c)^2 = \text{ mínima}, \quad \text{si} \quad c = \overline{x}$$

En efecto, se tiene que:

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

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

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

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

- Si los valores de la variable X se observaron en k grupos de tamaños respectivos $n_1, n_2, \ldots,$ entonces la media global es:

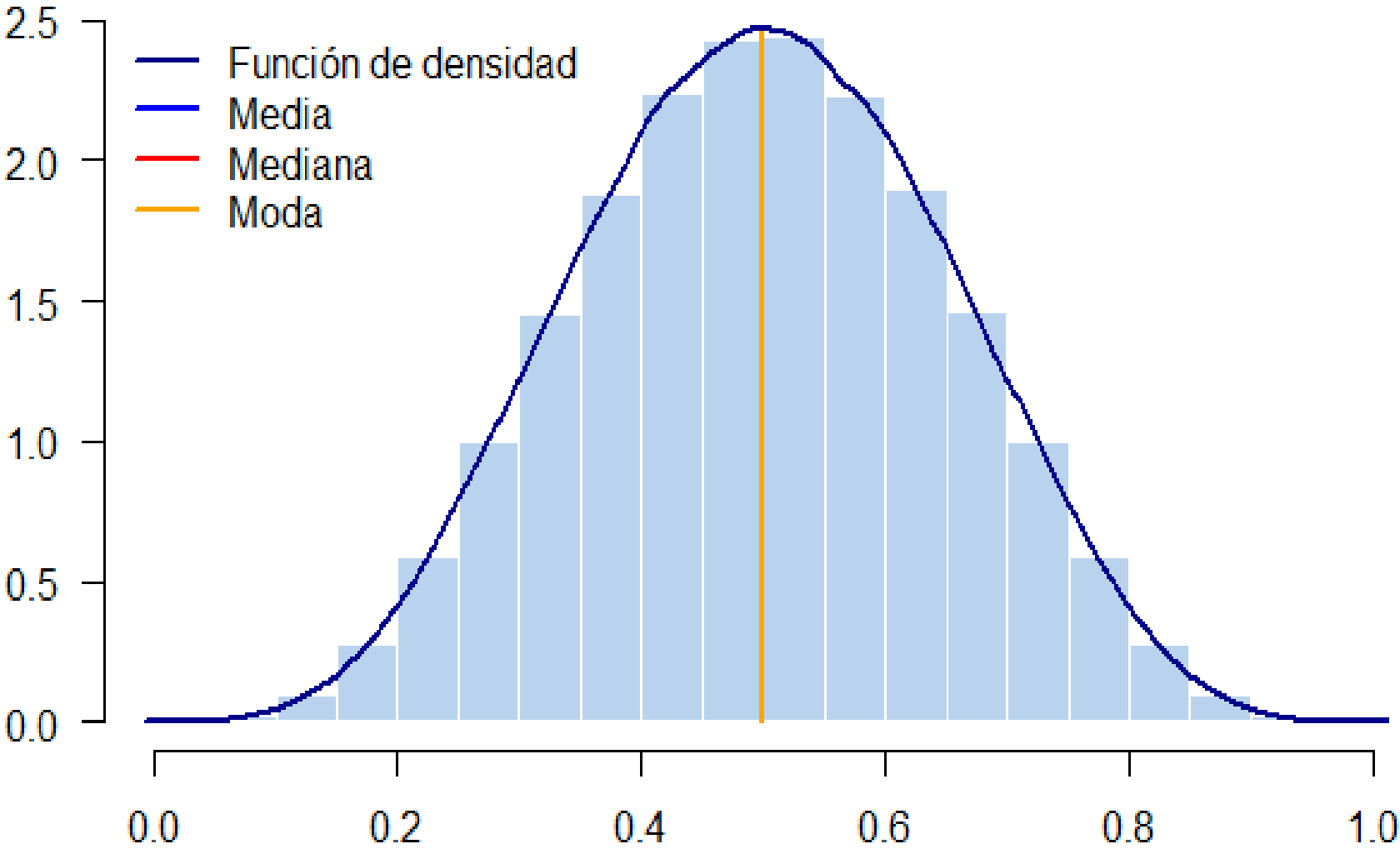
$$\overline{x} = \frac{n_1 \cdot \overline{x_1} + n_2 \cdot \overline{x_2} + \cdots + n_k \cdot \overline{x_k}}{\sum_{i=1}^{k} n_i}$$

- La media de los valores $x_1, x_2, \ldots, x_k$ ponderada por los pesos $w_1, \ldots, w_k$ se define por la siguiente relación:

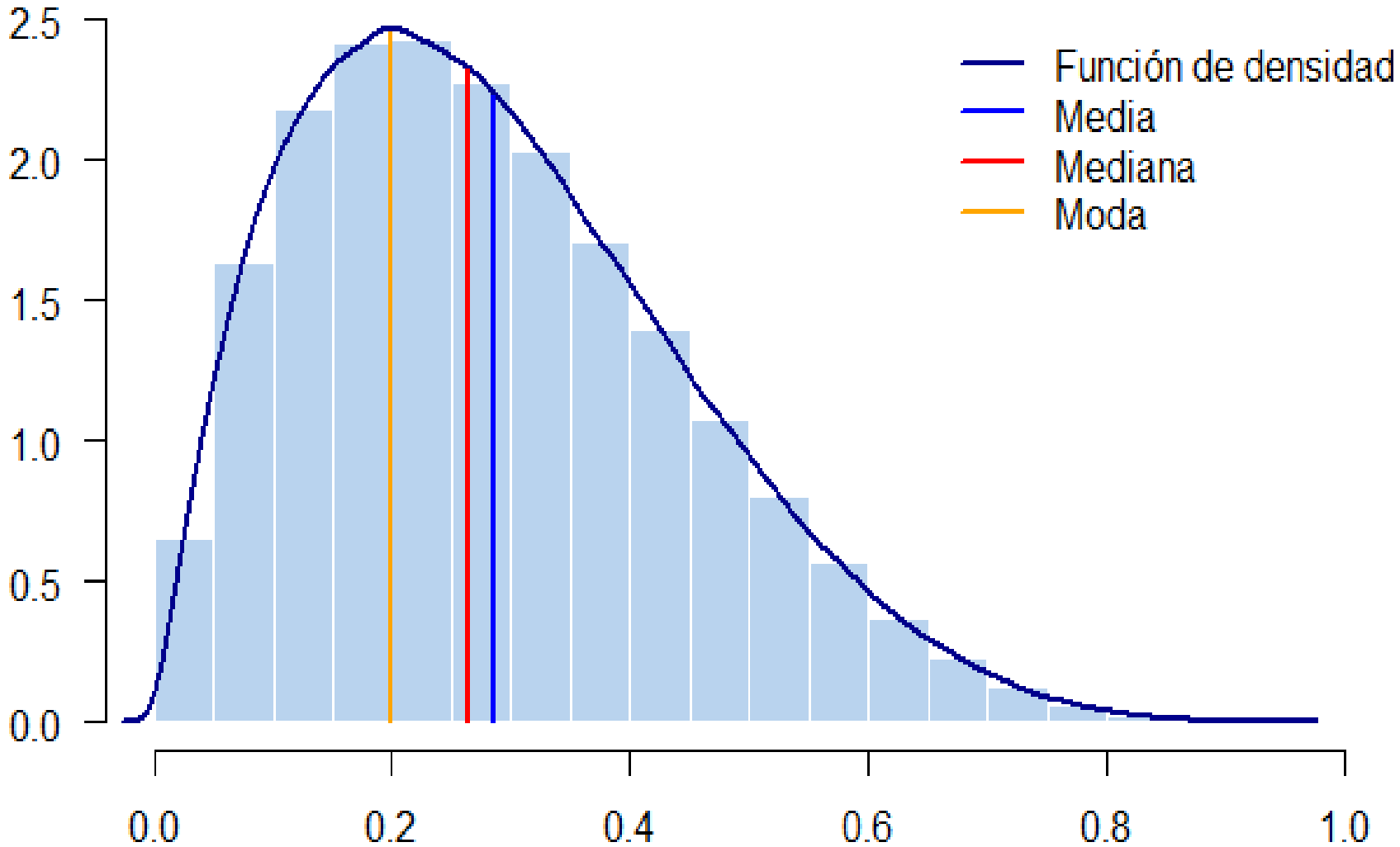
$$\overline{x} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_k \cdot x_k}{w_1 + w_2 + \cdots + w_k} = \frac{\sum_{i=1}^{k} w_i \cdot x_i}{\sum_{i=1}^{k} w_i}$$

- La media aritmética depende de todos los valores observados, en consecuencia, es afectada o sesgada por valores extremos discordantes o atípicos (extremadamente grandes o pequeños).
- La media aritmética puede ser calculada también en distribución de frecuencias por intervalos de amplitud diferentes, siempre que puedan determinarse los puntos medios (marcas de clase) de los intervalos.
- Existe una relación entre la media, mediana y moda.

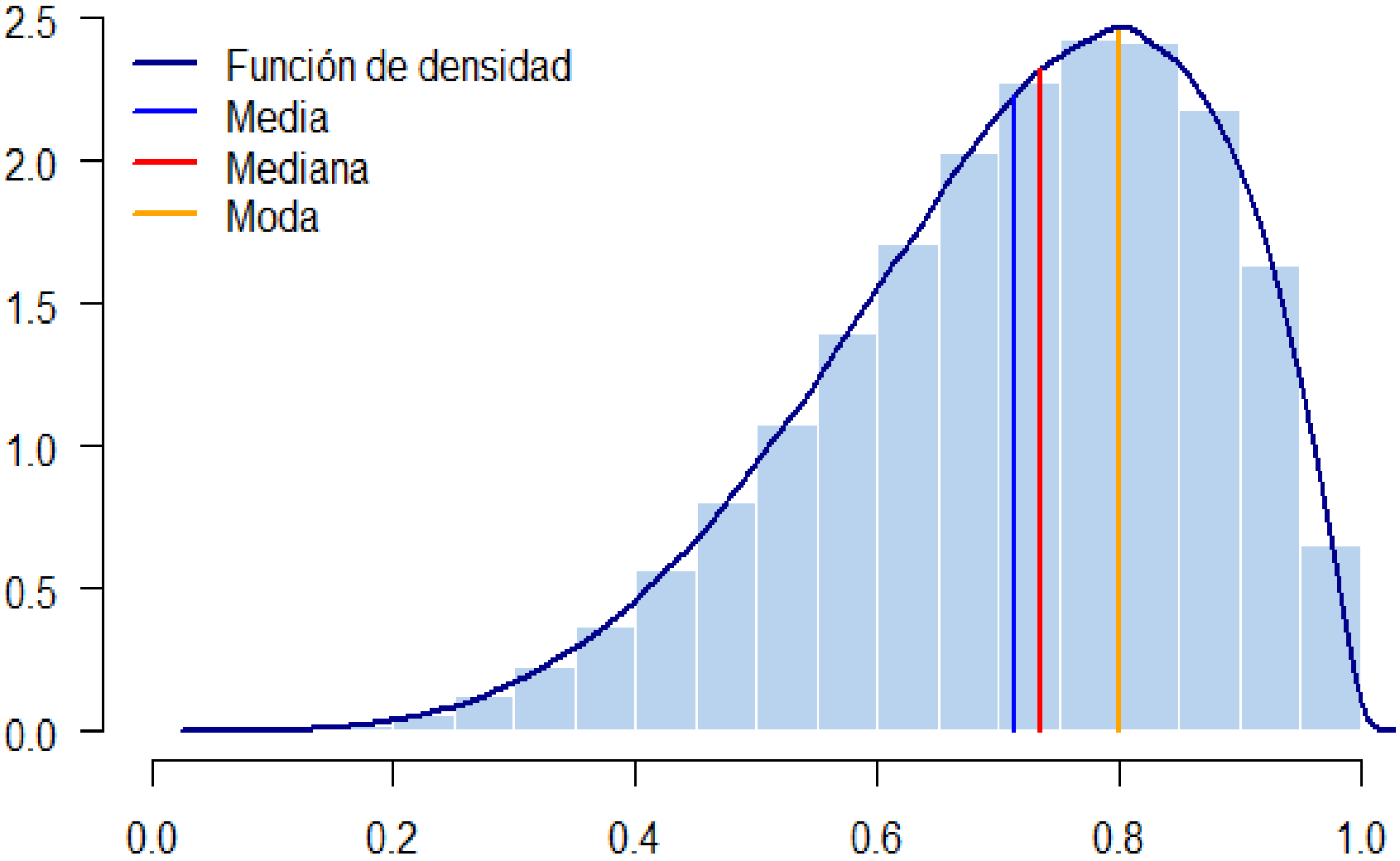
1. Si la distribución de frecuencias de los datos es **simétrica**, entonces, la media, mediana y moda tienen el mismo valor ($\overline{x} = Me = Mo$).



2) Si la distribución de frecuencias de los datos es **asimétrica de cola derecha**, entonces, la moda es menor que la mediana y esta a su vez es menor que la media ($Mo < Me < \overline{x}$).



3) Si la distribución de frecuencias de los datos es **asimétrica de cola izquierda**, entonces, la media es menor que la mediana y esta a su vez es menor que la moda ($\overline{x} < Me < Mo$).



4.- Media recortada

Es una medida de tendencia central que se diseñó para que no esté afectada por datos atípicos. La media recortada se calcula al arreglar los valores de la muestra en orden, "recortar" un número igual a partir de cada extremo y calcular la media de los restantes. Si se "recorta" el p % de los datos de cada extremo, la media recortada resultante se denomina "media recortada un p %". Las más comunes son las medias recortadas al 5, 10 y 20 %.

5.- Cuartil

Los cuartiles dividen la muestra tanto como sea posible en cuartos. Una muestra tiene tres de aquéllos. Existen diferentes formas de calcular cuartiles, pero todas dan aproximadamente el mismo resultado. El método más simple cuando se calcula manualmente es el siguiente: Sea n el tamaño de la muestra. Ordene los valores de la muestra del más pequeño al más grande. Para encontrar el primer cuartil, calcule el valor $0{,}25(n+1)$.

6.- Percentiles

El p-ésimo percentil de una muestra, para un número p entre 0 y 100, divide a la muestra tanto como sea posible, el p % de los valores de la muestra es menor que el p-ésimo percentil y el $(100 - p)$ % son mayores. Hay muchas maneras para calcular los percentiles; con todas se obtienen resultados similares. Ordene los valores de la muestra del más pequeño al más grande y después calcule la cantidad $(p/100)(n + 1)$, donde n es el tamaño de la muestra. Si esta cantidad es un entero, el valor de la muestra en esta posición es el p-ésimo percentil. Por otro lado, promedie los dos valores de la muestra en cualquier lado. Observe que el primer cuartil es el 25avo. percentil, la mediana es el 50avo. percentil y el tercer cuartil es el 75avo. percentil.

Los percentiles con frecuencia se usan para interpretar puntajes de exámenes estandarizados. Por ejemplo, si a una estudiante se le informa que su puntaje en un examen de ingreso a la universidad está en el 64avo. percentil, esto significa que 64 % de los estudiantes que presentaron el examen obtuvo puntajes inferiores.

Medidas de dispersión

Las medidas de tendencia central no son suficientes para describir un conjunto de valores de alguna variable estadística. Los promedios determinan el centro pero no indican acerca de cómo están situados los datos respecto del centro. Por ejemplo, se necesita una medida de dispersión con la finalidad de ampliar la descripción de los datos o de comparar dos o más series de datos. También se necesita una medida del grado o nivel de la asimetría o deformación en ambos lados del centro de una serie de datos, con el fin de describir la forma de la distribución de los datos.
Se necesita una medida que nos permita comparar el apuntamiento o curtosis de distribuciones simétricas con respecto a la distribución simétrica normal.

1.- Rango o recorrido de variable

Denotado por R es el número que resulta de la diferencia del valor máximo (x_{max}) menos el valor mínimo (x_{min}) de una serie de datos observados de la variable X. Es muy inestable, ya que solo depende de los valores extremos.

2.- Rango intercuartílico

Denotado por RI, es el número que resulta de la diferencia del cuartil 3 menos el cuartil 1 de los datos, es decir : $RI = Q_3 - Q_1$.
Es una medida que excluye el 25 % superior (cuarto superior) y el 25 % inferior (cuarto inferior) dando un rango dentro del cual se encuentra el 50 % central de los datos observados, y a diferencia del rango, no se encuentra afectada por los valores extremos.
Si el rango intercuartil es muy pequeño entonces describe alta uniformidad o pequeña variabilidad de los valores centrales.
El rango semiintercuartílico (RSI) es la mitad del rango intercuartílico. Si la distribución de frecuencias de los datos es simétrica, entonces, los cuartiles Q_1 y Q_3 son equidistantes de la mediana Q_2. En este caso, el RI es equivalente a $Q_2 \pm RSI$. Por lo tanto, $Q_2 \pm RSI$ contiene exactamente el 50 % de los datos. Si la distribución es casi simétrica, se concluye que el intervalo $Q_2 \pm RSI$ contiene aproximadamente el 50 % de los datos.

3.- Varianza y desviación estándar

La varianza se define como la media aritmética de los cuadrados de las diferencias de los datos con respecto a su media aritmética.

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

La varianza es una medida que cuantifica el nivel de dispersión o de variabilidad de los valores de una variable cuantitativa con respecto a su media aritmética. Si los datos tienden a concentrarse alrededor de su media, la varianza será pequeña. Si los valores tienden a distribuirse lejos de su media, la varianza será grande.

La desviación estándar es la raíz cuadrada de la varianza, es decir :

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

Algunas propiedades de la varianza son:

- La varianza es un número real no negativo y viene expresada en mediciones cuadráticas. Mientras que la desviación estándar está expresada en las mismas unidades que los datos observados.

- Dada la media $\overline{x}$ y la varianza s_n^2 de n datos de una variable X, la suma total de los cuadrados de los valores es igual a $n \times (s_n^2 + \overline{x}^2)$.
- Si X se transforma en $Y = aX + b$, la varianza de Y está dada por : $Var(Y) = a^2 \times Var(X)$. De igual forma, para la desviación estándar de Y se tiene: $s_n(Y) = |a| \times s_n(X)$.
- La varianza y la desviación estándar se calculan también en distribución de frecuencias de intervalos de amplitud diferentes, siempre que puedan determinarse las marcas de clase. Por otra parte, estas medidas dependen de todos los datos y son sensibles a los cambios.

4.- Coeficiente de variación

Es una medida de dispersión relativa (libre de unidades de medición) que se define como el cociente entre desviación estándar y media :

$$CV = \frac{s}{\overline{x}}$$

Se utiliza para comparar la variabilidad de dos o más series de datos que tengan medias iguales o diferentes o que tengan unidades de medida iguales o diferentes. Menor CV implica mayor homogeneidad de los datos.

- Si dos o más grupos de datos tienen medias aritméticas iguales, entonces, es más dispersa o de mayor variabilidad la serie que tiene mayor valor en : R,RI, s^2, s o CV. Si hay marcada asimetría, es preferible comparar la variabilidad con el rango intercuartil.
- Si dos o más series de datos, no tienen medias iguales o no tienen las mismas unidades de medición, entonces es más homogénea o de menor variabilidad la serie que tenga menor coeficiente de variación, sin importar su forma de asimetría.
- Cuando se necesiten comparar valores observados que pertenecen a distintas distribuciones de datos o difieres en el tipo de unidad de medida, entonces se estandarizan los valores observados :

$$Z = \frac{X - \overline{x}}{s_n}$$

5.- Índices de asimetría

Una distribución de frecuencias es simétrica si son iguales las frecuencias de sus valores equidistantes del valor central.

El índice de asimetría de Pearson se define como:

$$As = \frac{\overline{x} - Mo}{s}$$

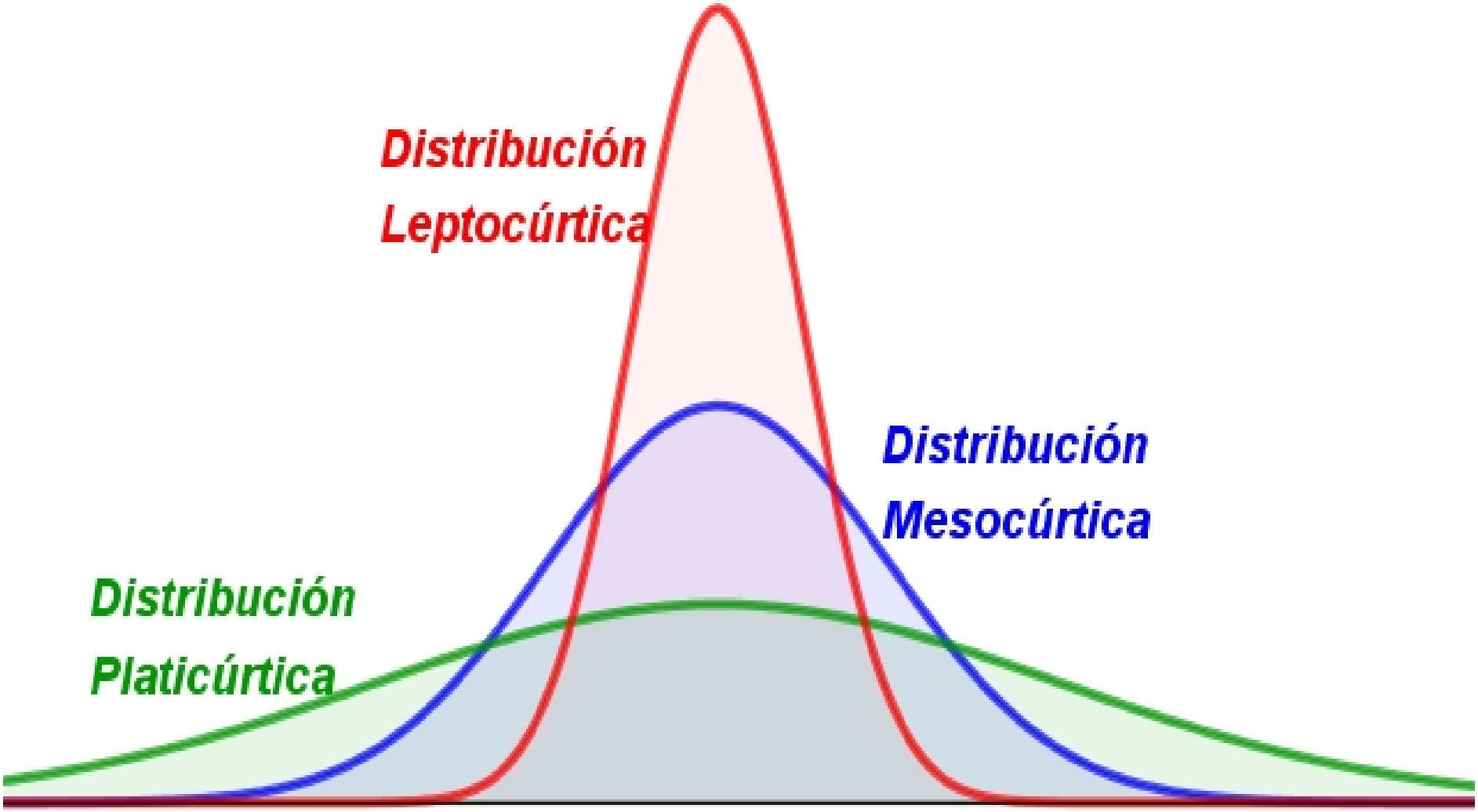
Si $As = 0$, la distribución es simétrica, si $As > 0$, la simetría es positiva o sesgada hacia la derecha; y, si $As < 0$, la simetría es negativa o sesgada hacia la izquierda. Aplicando la definición de momentos se tiene:

$$As = \frac{nM_3}{(n-1)(n-2)s_n^3}$$

De donde $M_3 = \sum_{i=1}^{n} (X_i - \overline{x})^3$. Este índice es utilizado por los paquetes estadísticos para determinar la asimetría de distribuciones.

6.- Curtosis

Es la propiedad de una distribución de frecuencias por la cual se compara la dispersión de los datos observados cercanos al valor central con la dispersión de los datos cercanos a ambos extremos de la distribución.



Una curva es leptocúrtica cuando su curtosis es mayor que la normal ($K > 3$). Mientras que, cuando la curtosis es menor que la normal ($K < 3$) se denomina platicúrtica.

La ecuación de curtosis es:

$$K = \frac{1}{N} \frac{\sum_{i=1}^{N} (X_i - \overline{x})^4}{s_n^4}$$

Utilizando la definición de momentos se tiene:

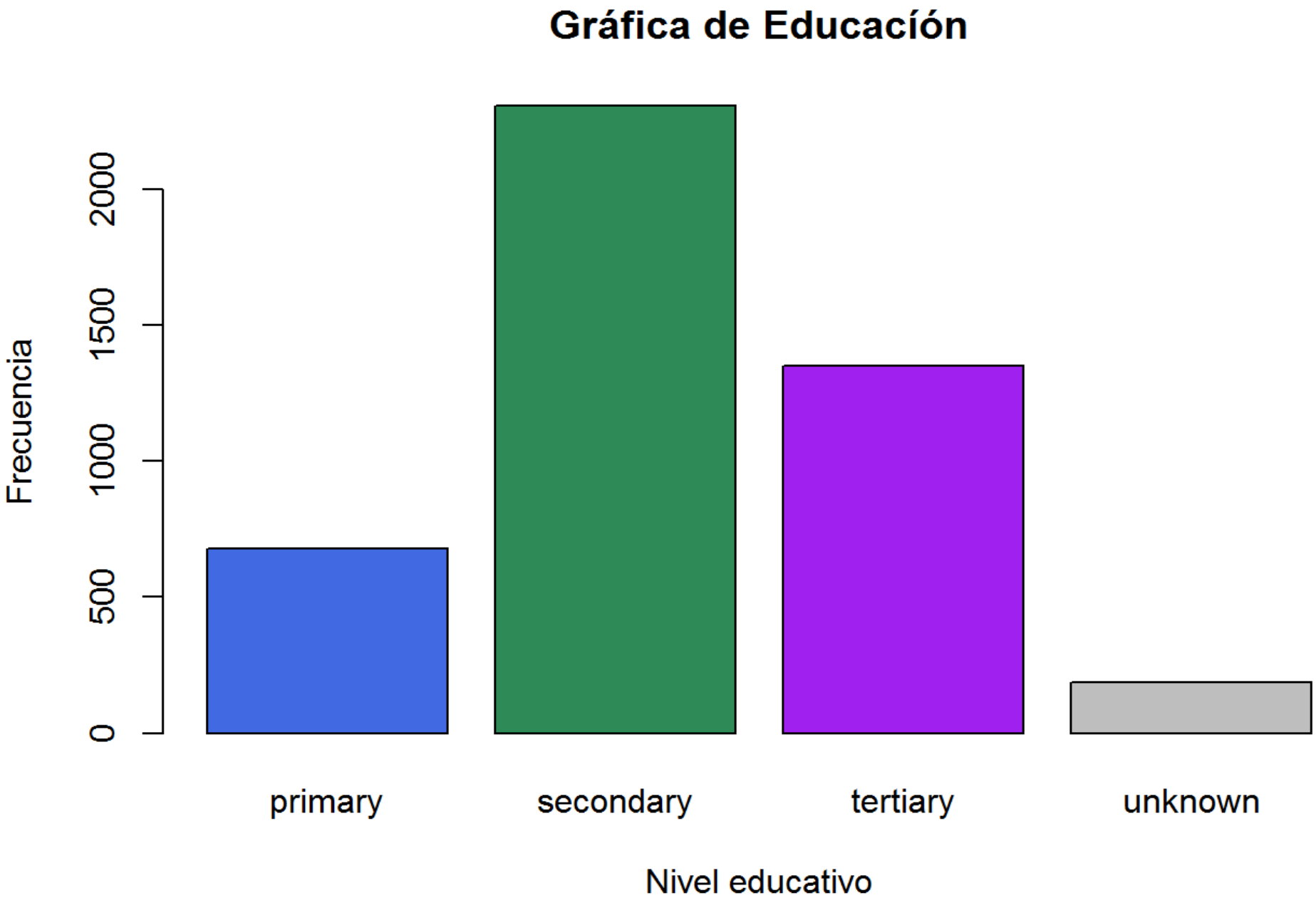
$$K = \frac{n(n+1) M_4 - 3M_2^2 (n-1)}{(n-1)(n-2)(n-3) s_n^4}$$

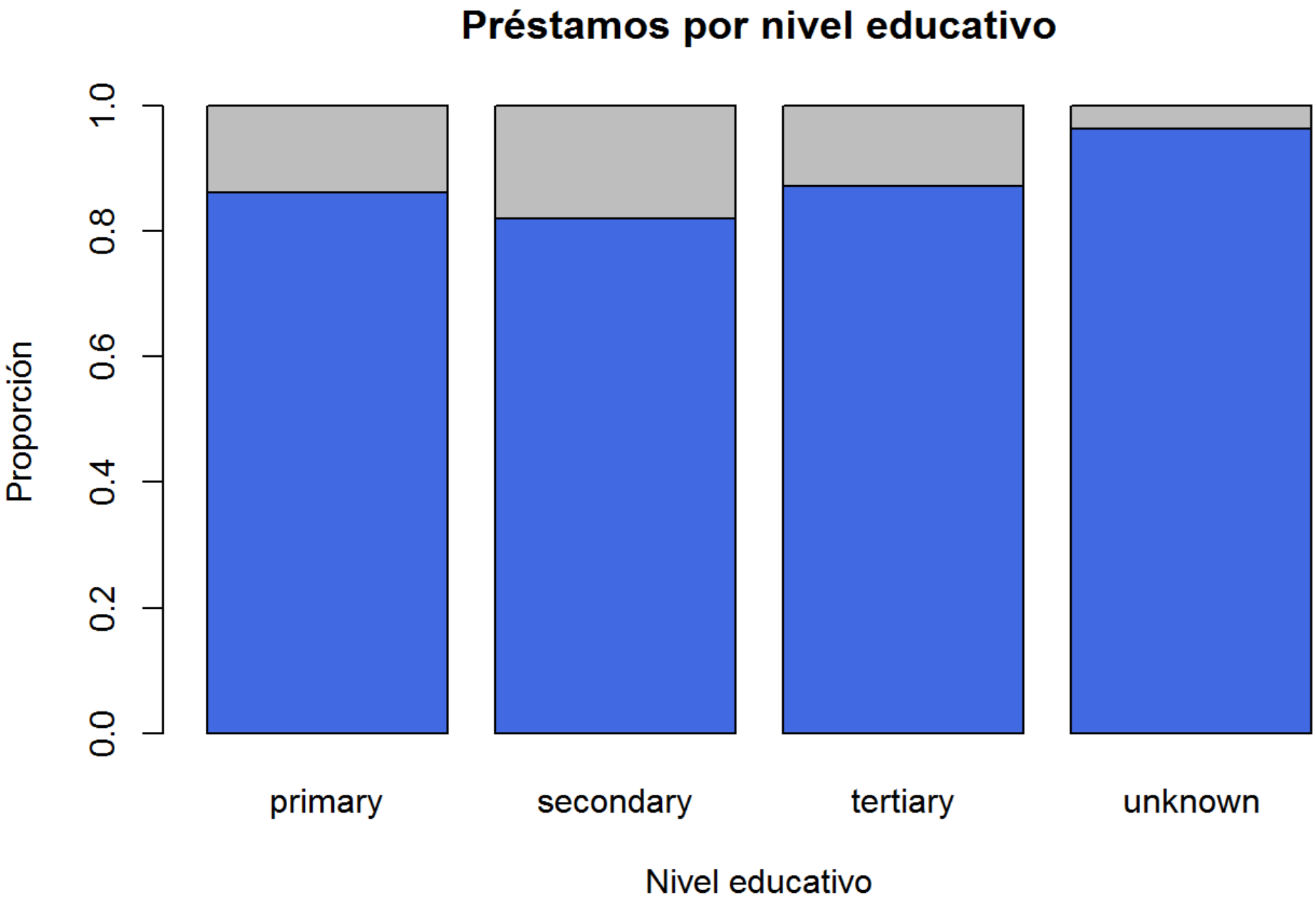
De donde : $M_4 = \sum_{i=1}^{n} (X_i - \overline{x})^4$ y $M_2 = \sum_{i=1}^{n} (X_i - \overline{x})^2$.

Finalmente, la curtosis es una medida de variabilidad utilizada para caracterizar la morfología de una distribución. De esta forma se pueden comparar distribuciones simétricas con el mismo promedio e igual dispersión (dada por la desviación estándar). Disponer de medidas de variabilidad asegura que los promedios sean confiables y ayuda a controlar las variaciones de la distribución.

Gráfico de barras

Un gráfico de barras es una forma de resumir un conjunto de datos por categorías. Muestra los datos usando varias barras de la misma anchura, cada una de las cuales representa una categoría concreta. La altura de cada barra es proporcional a una agregación específica (por ejemplo, la suma de los valores de la categoría que representa). Las categorías podrían ser desde grupos de edad a ubicaciones geográficas. Si se aplica al crear el análisis, el gráfico de barras puede mostrar información adicional en líneas de referencia o varios tipos distintos de curvas. Estas líneas o curvas podrían, por ejemplo, mostrar si los puntos de los datos se adaptan bien a un ajuste de curva polinómica determinado, o resumir un conjunto de puntos de datos de muestra ajustándolos a un modelo que describirá los datos y mostrará una curva o una línea recta sobre la visualización. La curva normalmente cambia su aspecto en función de los valores que se hayan filtrado del análisis.


Gráfica de Educacíón
Frecuencia
0
500
1000
1500
2000
primary
secondary
tertiary
unknown
Nivel educativo


Préstamos por nivel educativo
Proporción
0.0
0.2
0.4
0.6
0.8
1.0
primary
secondary
tertiary
unknown
Nivel educativo

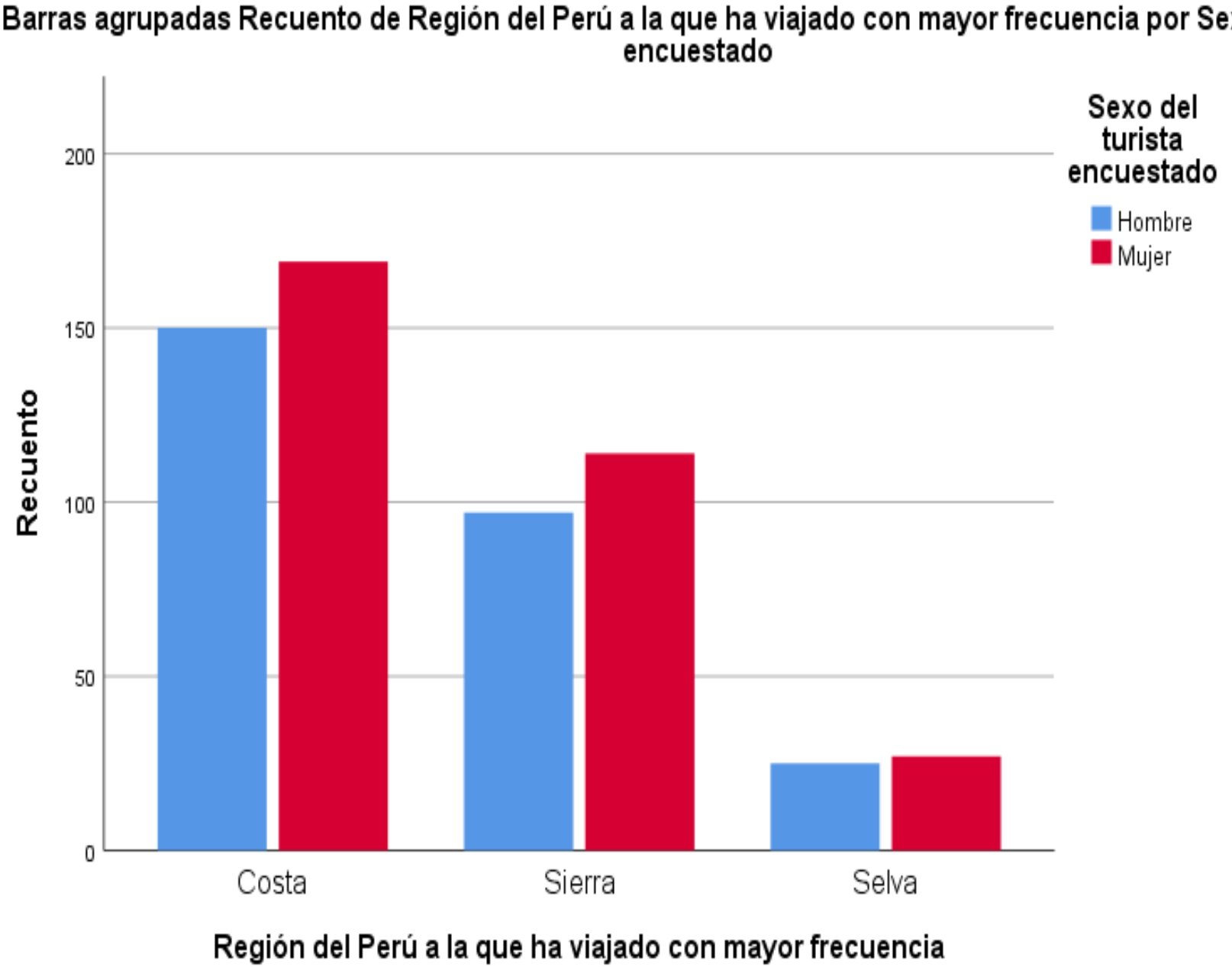

Barras agrupadas Recuento de Región del Perú a la que ha viajado con mayor frecuencia por Sexo del turista encuestado
Sexo del turista encuestado
Hombre
Mujer
Recuento
200
150
100
50
0
Costa
Sierra
Selva
Región del Perú a la que ha viajado con mayor frecuencia

Gráfico circular

Un gráfico circular o gráfica circular, también llamado gráfico de pastel, gráfica de pizza, gráfico de tarta, gráfico de torta o gráfica de 360 grados, es un recurso estadístico que se utiliza para representar porcentajes y proporciones. El número de elementos comparados dentro de una gráfica circular suele ser de más de cuatro. Se utilizan en aquellos casos donde interesa no solamente mostrar el número de veces que se dan una característica o atributo de manera tabular sino más bien de manera gráfica, de tal manera que se pueda visualizar mejor la proporción en que aparece esa característica respecto del total.

A pesar de su popularidad, se trata de un tipo de gráfico poco recomendable debido a que nuestra capacidad perceptual para estimar relaciones de proporción o diferencias entre áreas de sectores circulares es mucho menor que, por ejemplo, entre longitudes o posiciones, tal y como sucede en otras gráficas.

Recepcion de turistas

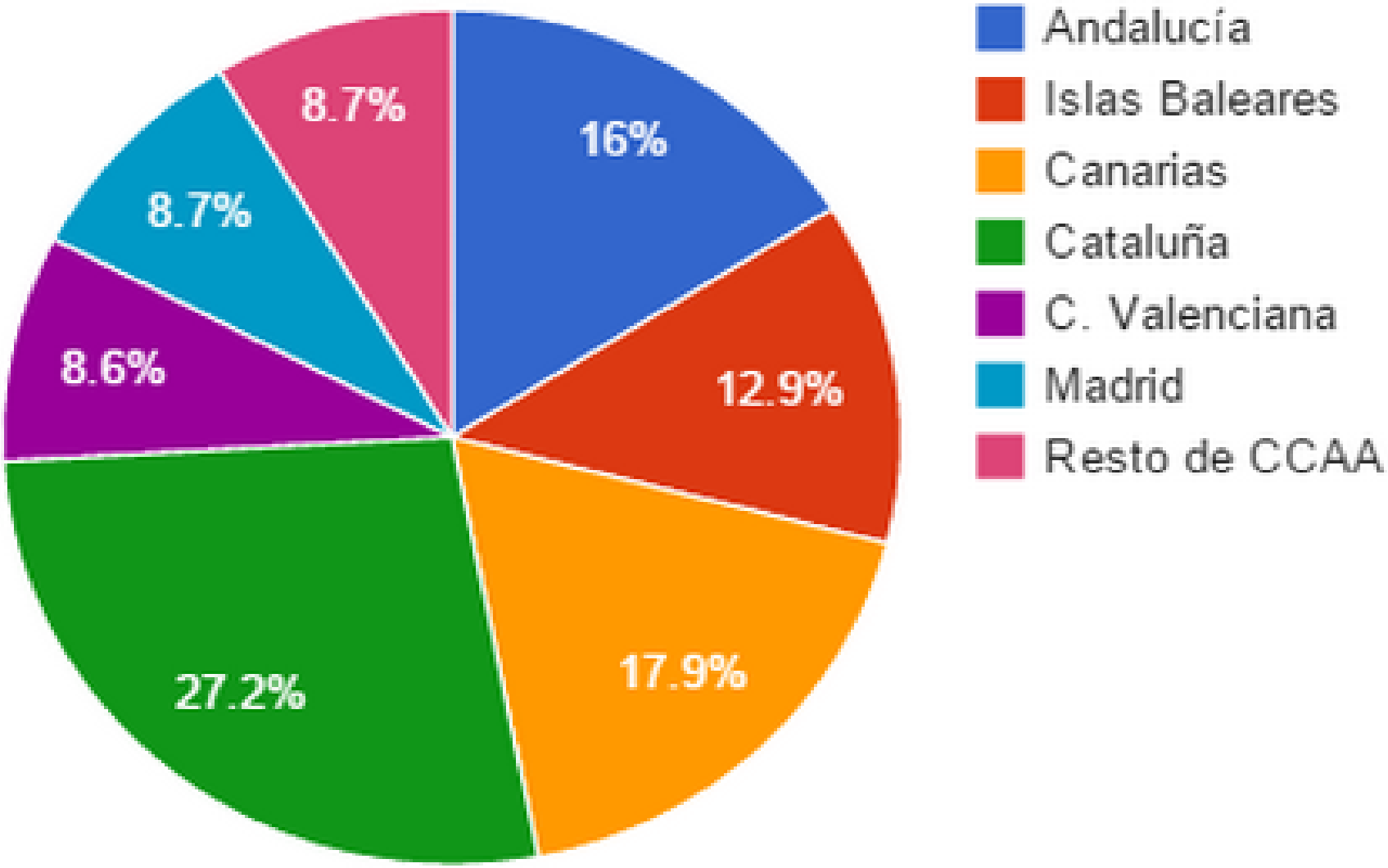
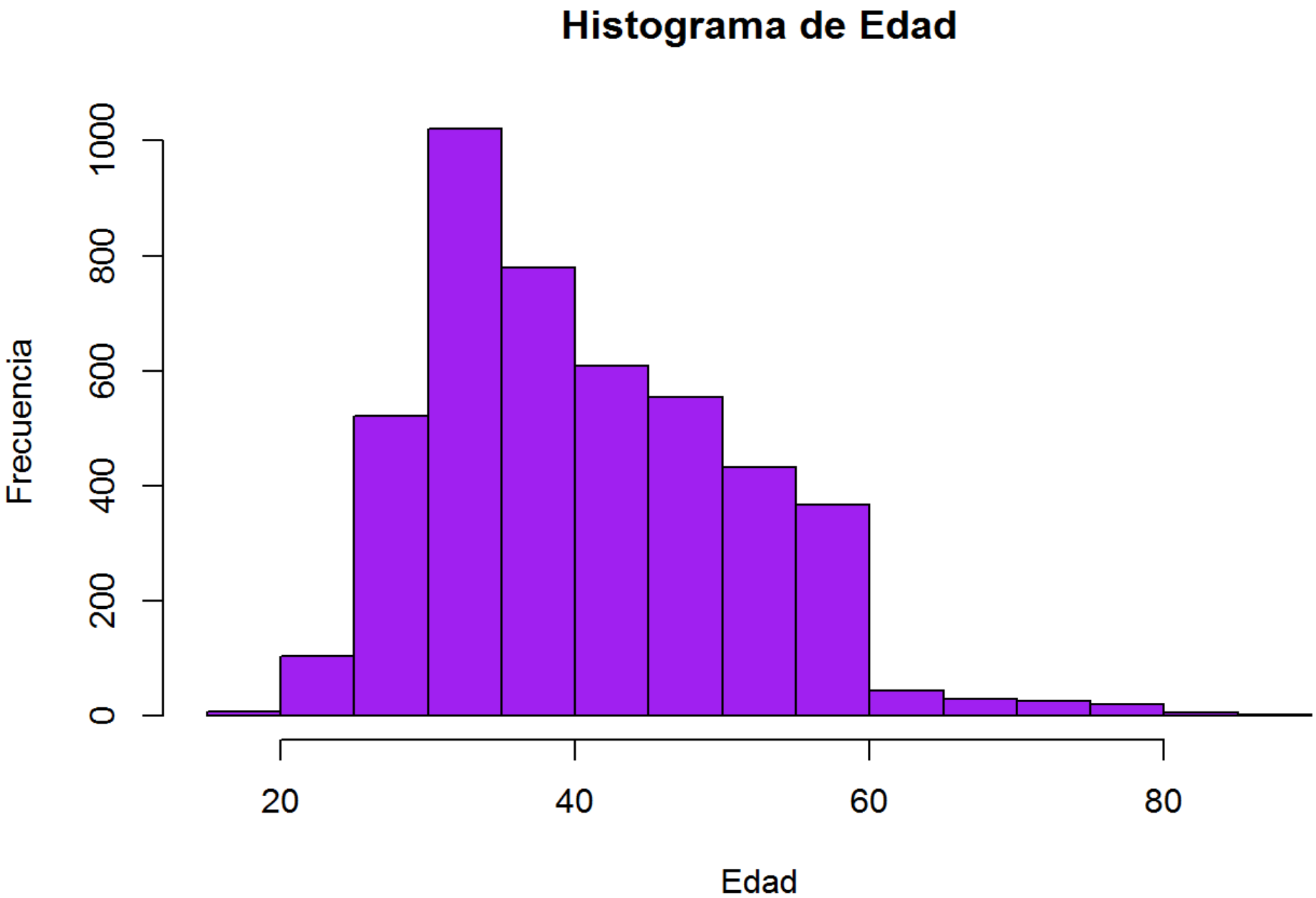


Gráfico de histograma

Es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. Sirven para obtener una "primera vista"general, o panorama, de la distribución de la población, o de la muestra, respecto a una característica, cuantitativa y continua (como la longitud o el peso). De esta manera ofrece una visión de grupo permitiendo observar una preferencia, o tendencia, por parte de la muestra o población por ubicarse hacia una determinada región de valores dentro del espectro de valores posibles (sean infinitos o no) que pueda adquirir la característica. Así pues, podemos evidenciar comportamientos, observar el grado de homogeneidad, acuerdo o concisión entre los valores de todas las partes que componen la población o la muestra, o, en contraposición, poder observar el grado de variabilidad, y por ende, la dispersión de todos los valores que toman las partes, también es posible no evidenciar ninguna tendencia y obtener que cada miembro de la población toma por su lado y adquiere un valor de la característica aleatoriamente sin mostrar ninguna preferencia o tendencia.


Histograma de Edad
Frecuencia
0
200
400
600
800
1000
20
40
60
80
Edad

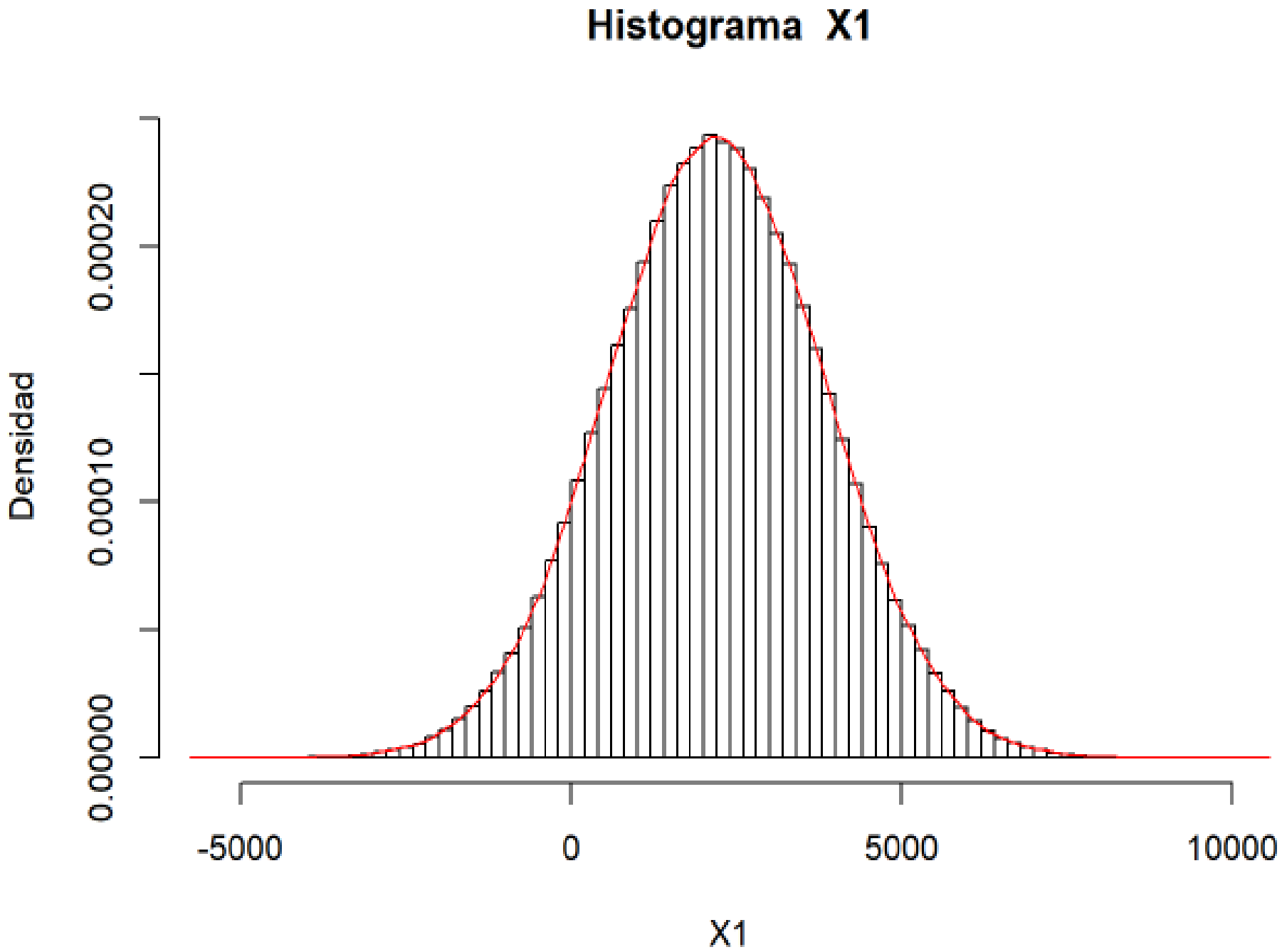
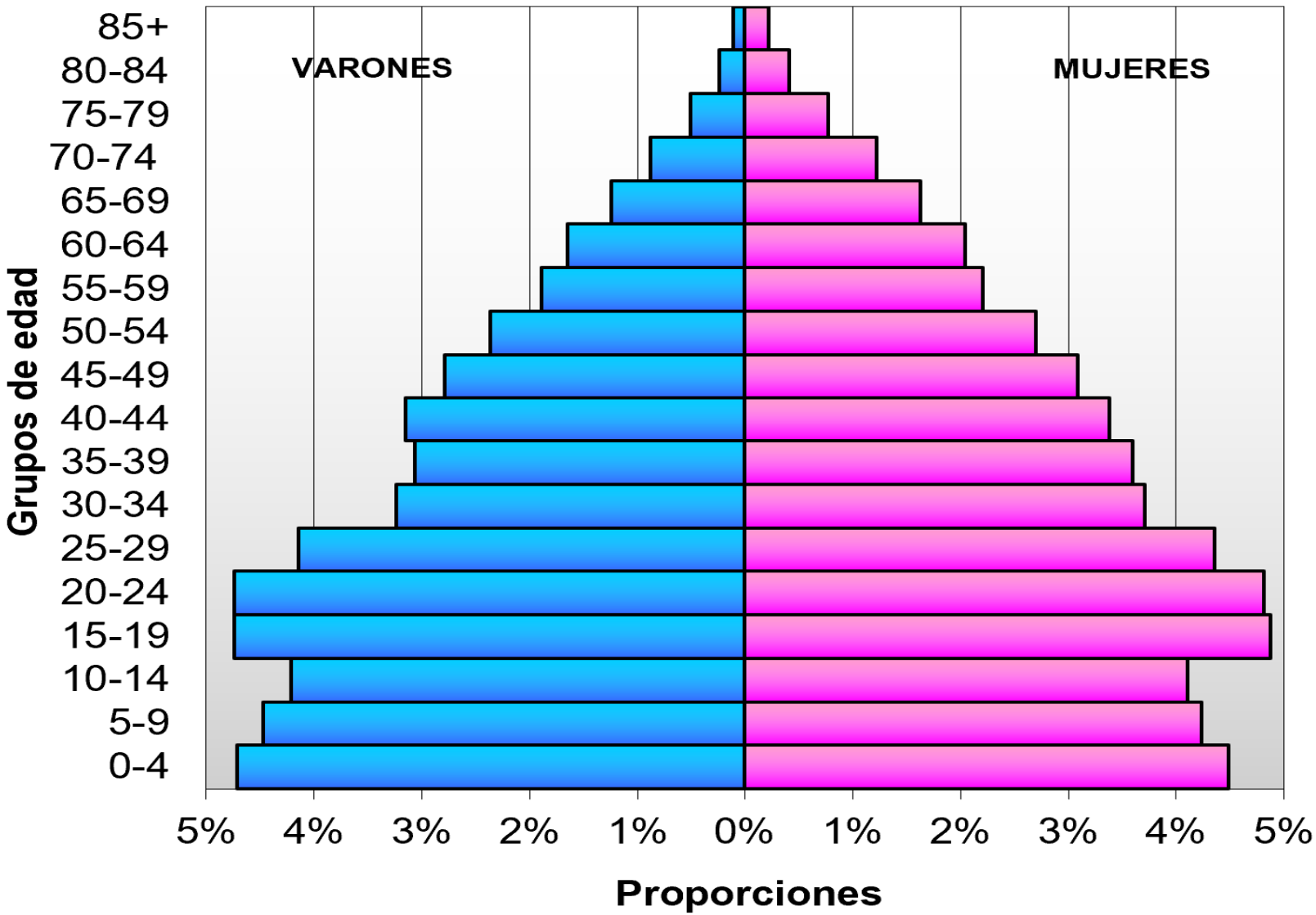

Histograma X1
Densidad
0.00000
0.00010
0.00020
-5000
0
5000
10000
X1

Gráfico de pirámide poblacional

La pirámide de población es una representación gráfica de las características de una población perteneciente a una localidad, ciudad o país, en un momento en el tiempo. Se muestra con barras en posición horizontal, la longitud de cada barra tiene una relación directamente y proporcional al número de individuos de la población. Permite visualizar de manera sencilla la proporción de edad, sexo, envejecimiento, entre otros. Es decir, gracias a su construcción que es posible observar el crecimiento, la estructura la distribución, así como la movilidad de una población y simplificar su correlación con la dinámica social, económica así como ambiental que prevalece.Así mismo, permite mostrar una tendencia; facilitando así su interpretación y la toma de decisiones con respecto a la planeación territorial, así como políticas de salud pública. Es posible que se realicen comparativos en distintas ciudades o países para conocer la evolución demográfica en cada uno, con respecto al tiempo en que se lleva a cabo la construcción de la pirámide.

Pirámide de población de España, año 1950

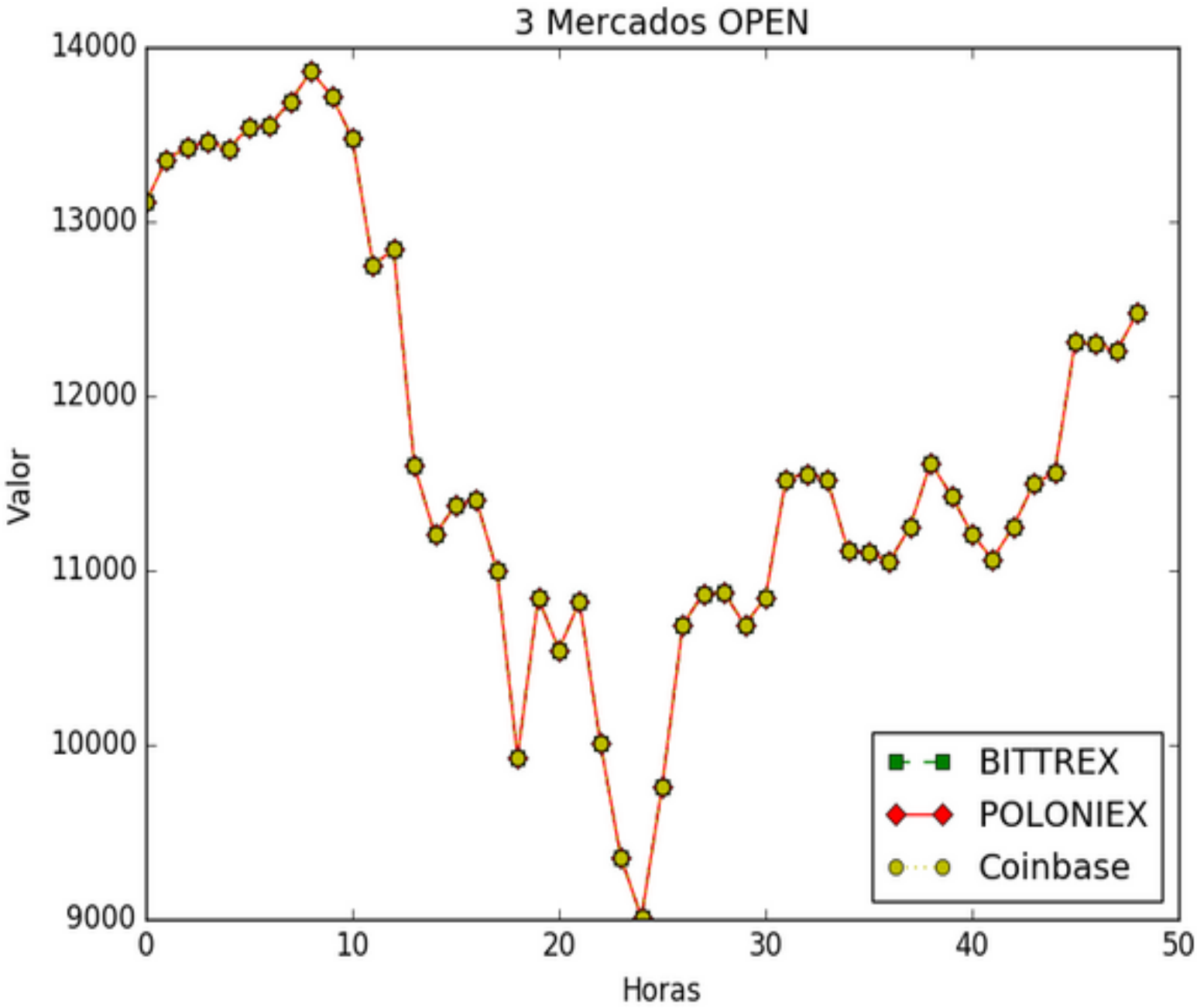


Fuente: Instituto Nacional de Estadística. Censo de 1950

Gráfico de línea

Los gráficos de líneas se utilizan para mostrar el valor cuantitativo en un intervalo o intervalo de tiempo continuo. Se usa con mayor frecuencia para mostrar tendencias y relaciones (cuando se agrupan con otras líneas). Los gráficos de línea también ayudan a dar un «panorama general» en un intervalo, para ver cómo se ha desarrollado durante ese período. Los gráficos de líneas se representan dibujando primero los puntos de datos en una cuadrícula cartesiana, y luego conectando una línea entre estos puntos. Típicamente, el eje Y tiene un valor cuantitativo, mientras que el eje X tiene una escala de categoría o secuenciada. Los valores negativos se pueden mostrar debajo del eje X.

Los gráficos de líneas se componen de un rango continuo de fechas o números en el eje x, y de un valor numérico asociado en el eje y. El eje x de un gráfico de líneas muestra una variable continua, como el tiempo o la distancia, y dibuja una línea que visualiza el cambio de valores entre cada intervalo de tiempo o distancia consecutivos. Cada intervalo está marcado con un punto correspondiente a un valor numérico medido por el eje y.


3 Mercados OPEN
14000
13000
12000
11000
10000
9000
Valor
0
10
20
30
40
50
Horas
BITTREX
POLONIEX
Coinbase

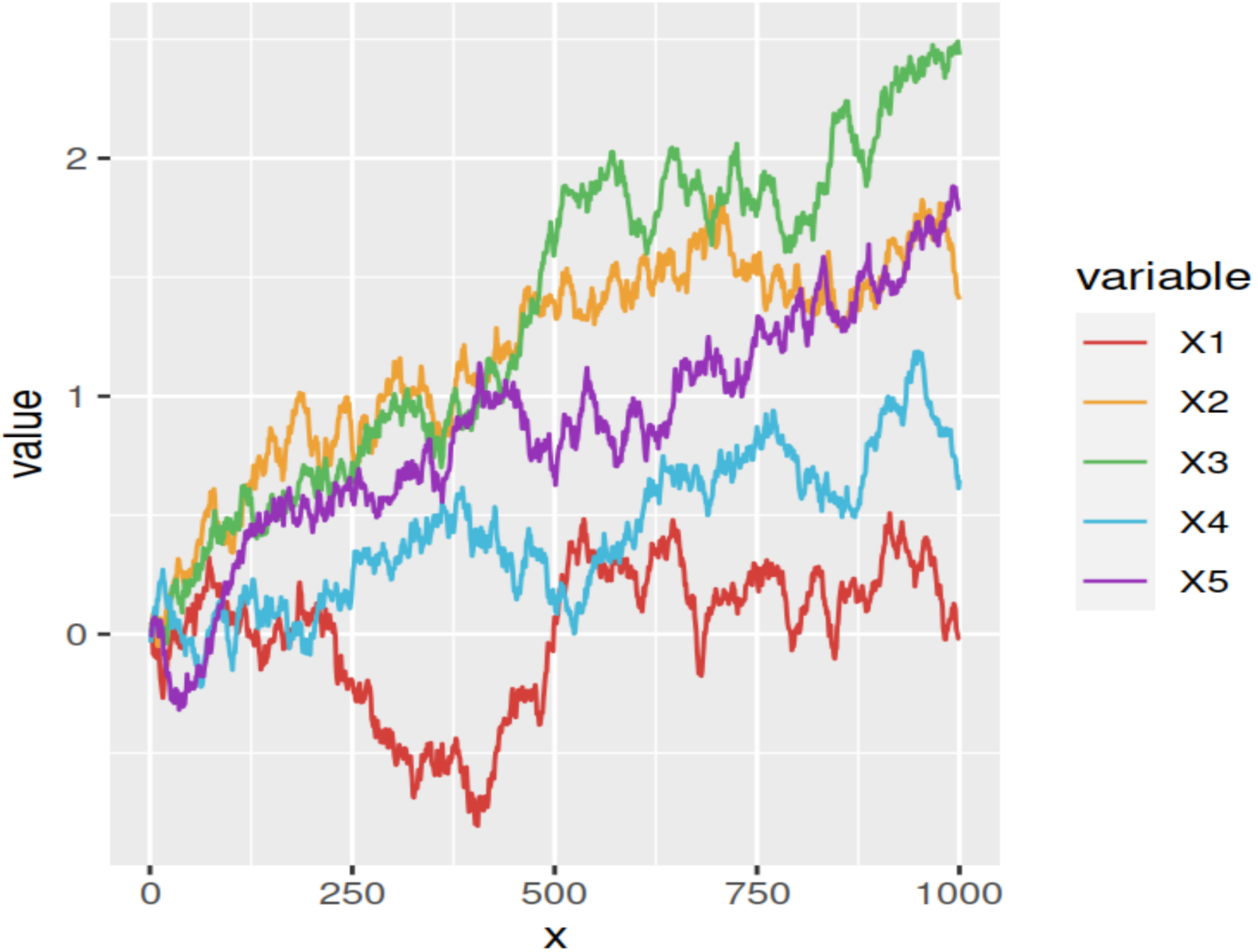
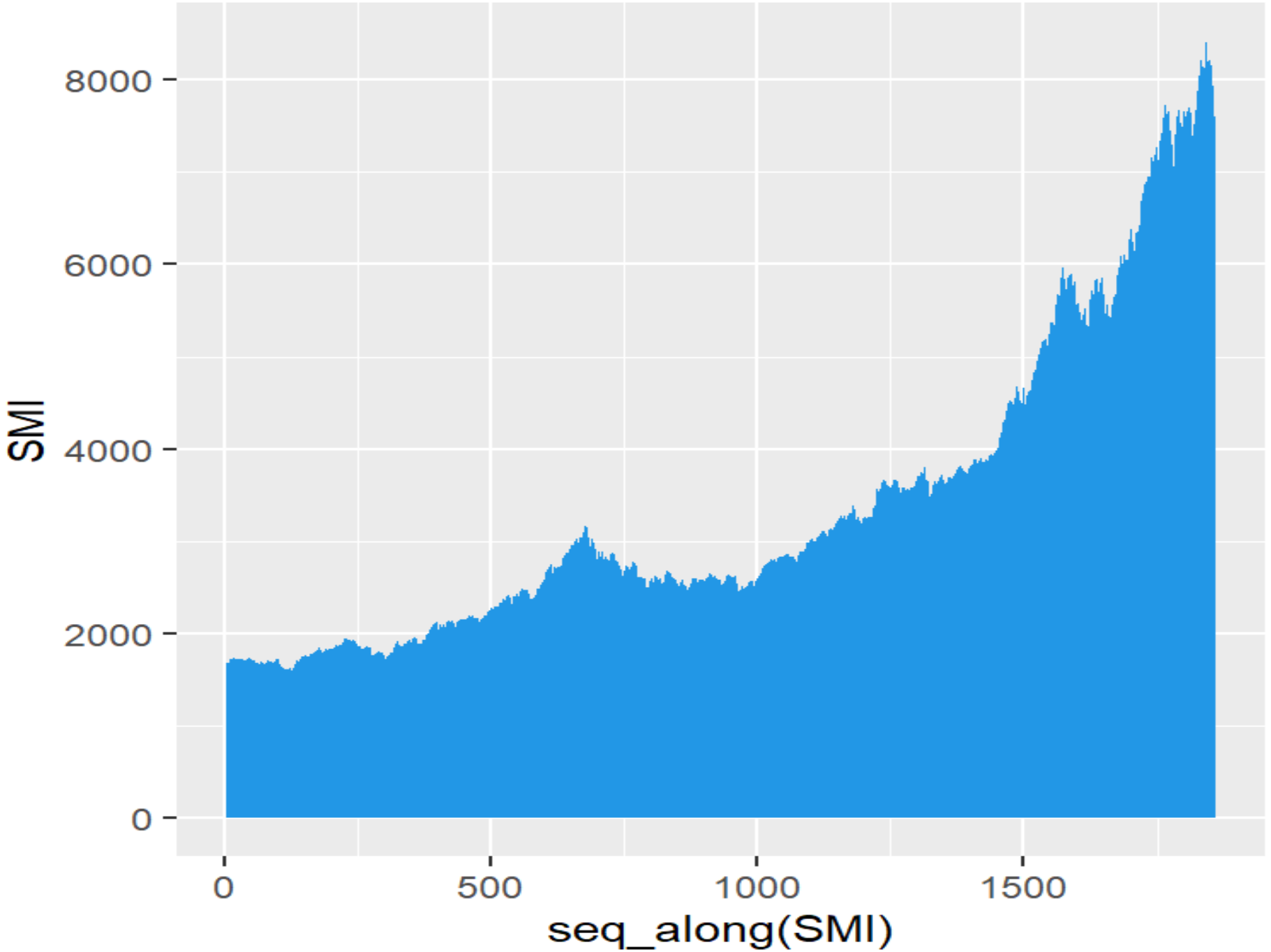

value
2
1
0
0
250
500
750
1000
x
variable
X1
X2
X3
X4
X5

Gráfico de área

Un gráfico de áreas representa el cambio en una o más cantidades a lo largo del tiempo. Es similar a un gráfico de líneas. Tanto en los gráficos de área como en los gráficos de líneas, los puntos de datos se trazan y luego se conectan mediante segmentos de línea para mostrar el valor de una cantidad en varios momentos diferentes. Sin embargo, los gráficos de áreas son diferentes de los gráficos de líneas porque el área entre el eje XY la línea se rellena con color o sombreado.

Un gráfico de áreas apiladas muestra cuánto contribuye cada parte a la cantidad total. Por ejemplo, el propietario de una cadena de tiendas de comestibles puede querer hacer un gráfico que muestre las ganancias obtenidas por cada una de sus tiendas y las ganancias totales obtenidas por todas las tiendas juntas. Un gráfico de áreas apiladas sería perfecto para representar este tipo de datos.

Aunque los gráficos de áreas se utilizan con mayor frecuencia para mostrar tendencias generales en los datos a lo largo del tiempo, también puede hacer un gráfico de áreas apiladas que muestre explícitamente la contribución exacta de cada cantidad al total.


8000
6000
4000
2000
0
SMI
0
500
1000
1500
seq_along(SMI)

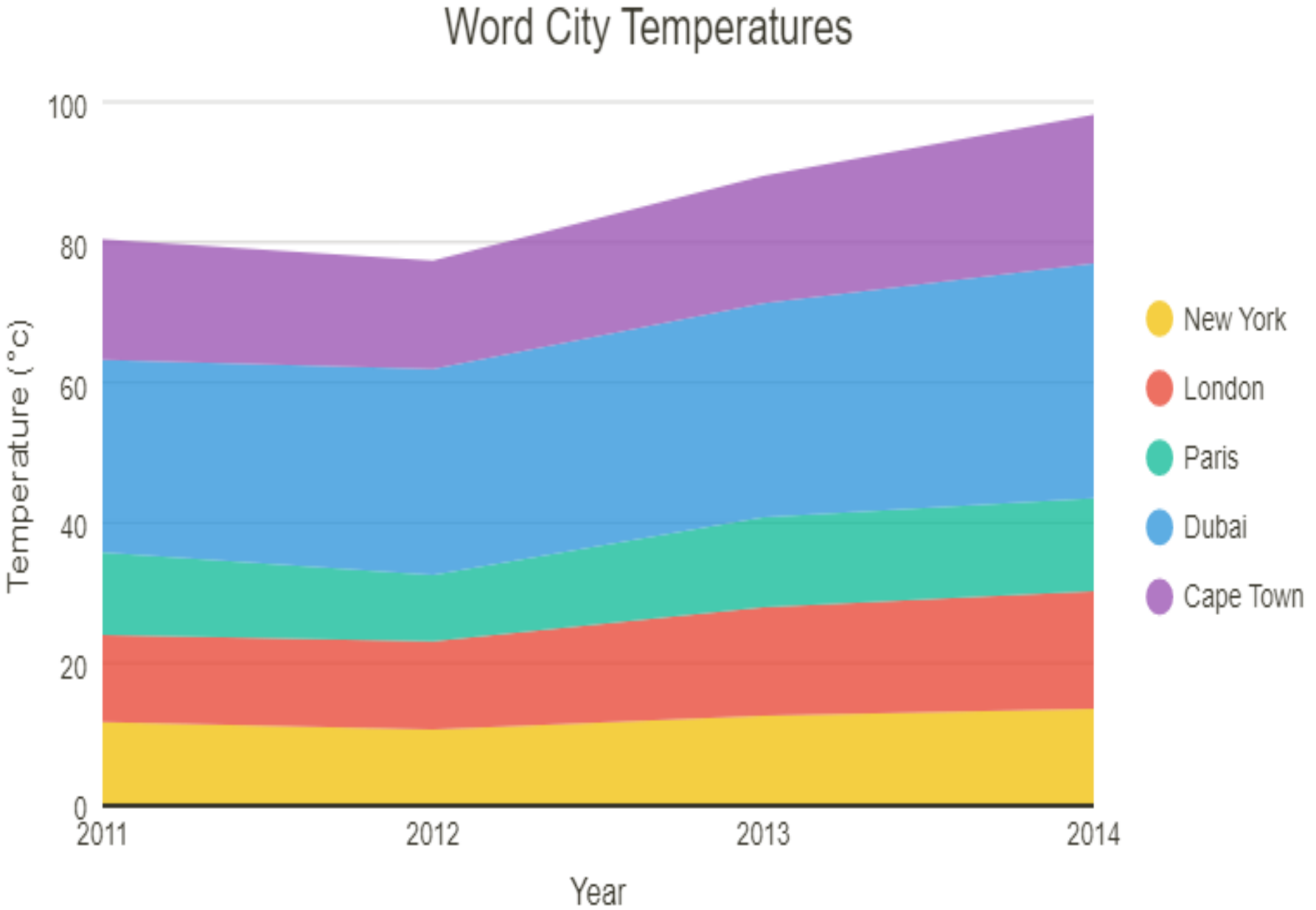
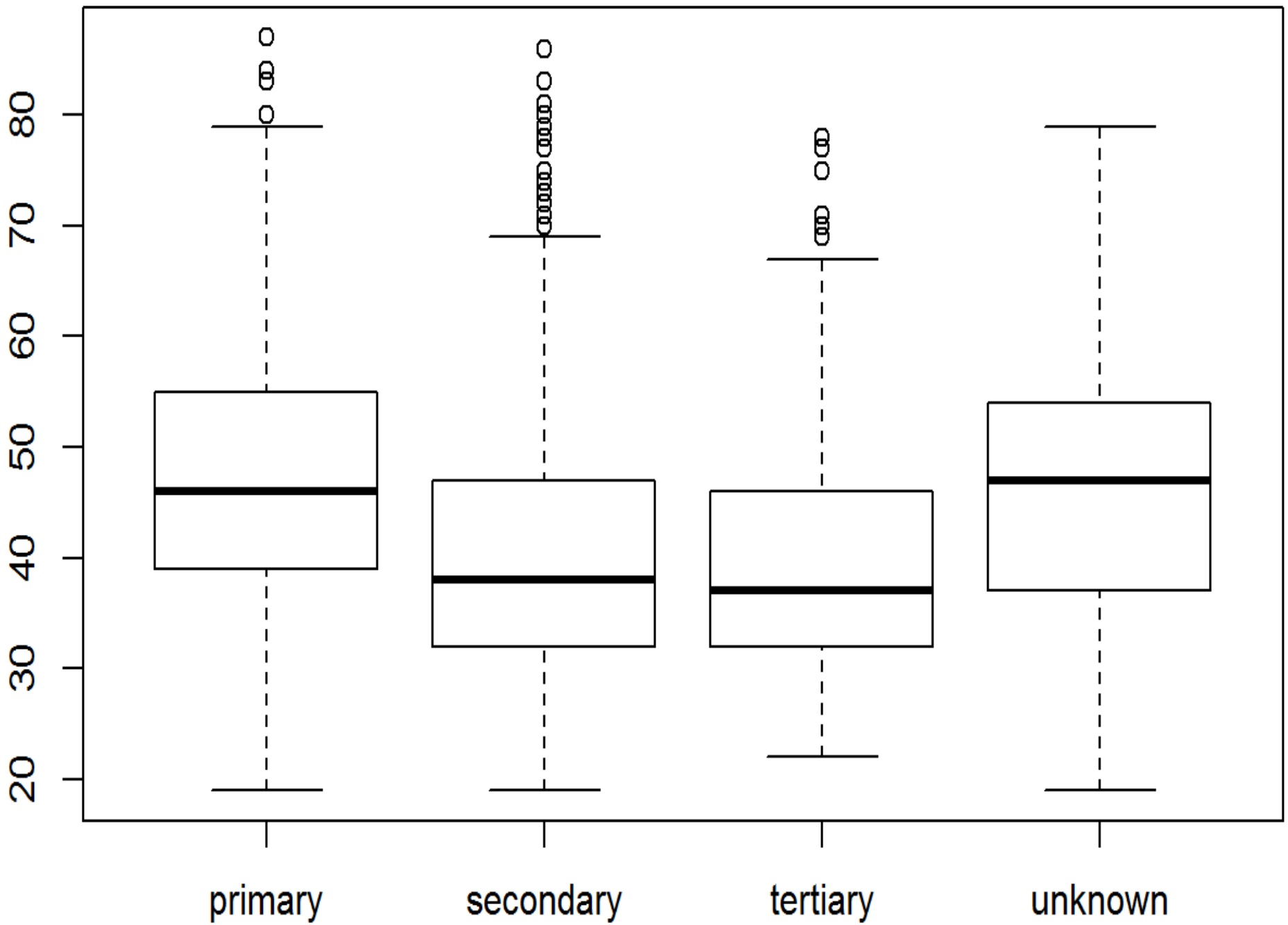

Word City Temperatures
100
80
60
40
20
0
Temperature (°c)
2011
2012
2013
2014
Year
New York
London
Paris
Dubai
Cape Town

Gráfico de cajas y bigotes

Los diagramas de Caja-Bigotes (boxplots o box and whiskers) son una presentación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y simetría. Para su realización se representan los tres cuartiles y los valores mínimo y máximo de los datos, sobre un rectángulo, alineado horizontal o verticalmente.
Una gráfica de este tipo consiste en una caja rectangular, donde los lados más largos muestran el recorrido intercuartílico. Este rectángulo está dividido por un segmento vertical que indica donde se posiciona la mediana y por lo tanto su relación con los cuartiles primero y tercero(recordemos que el segundo cuartil coincide con la mediana). Esta caja se ubica a escala sobre un segmento que tiene como extremos los valores mínimo y máximo de la variable. Las lineas que sobresalen de la caja se llaman bigotes. Estos bigotes tienen tienen un límite de prolongación, de modo que cualquier dato o caso que no se encuentre dentro de este rango es marcado e identificado individualmente


20
30
40
50
60
70
80
primary
secondary
tertiary
unknown

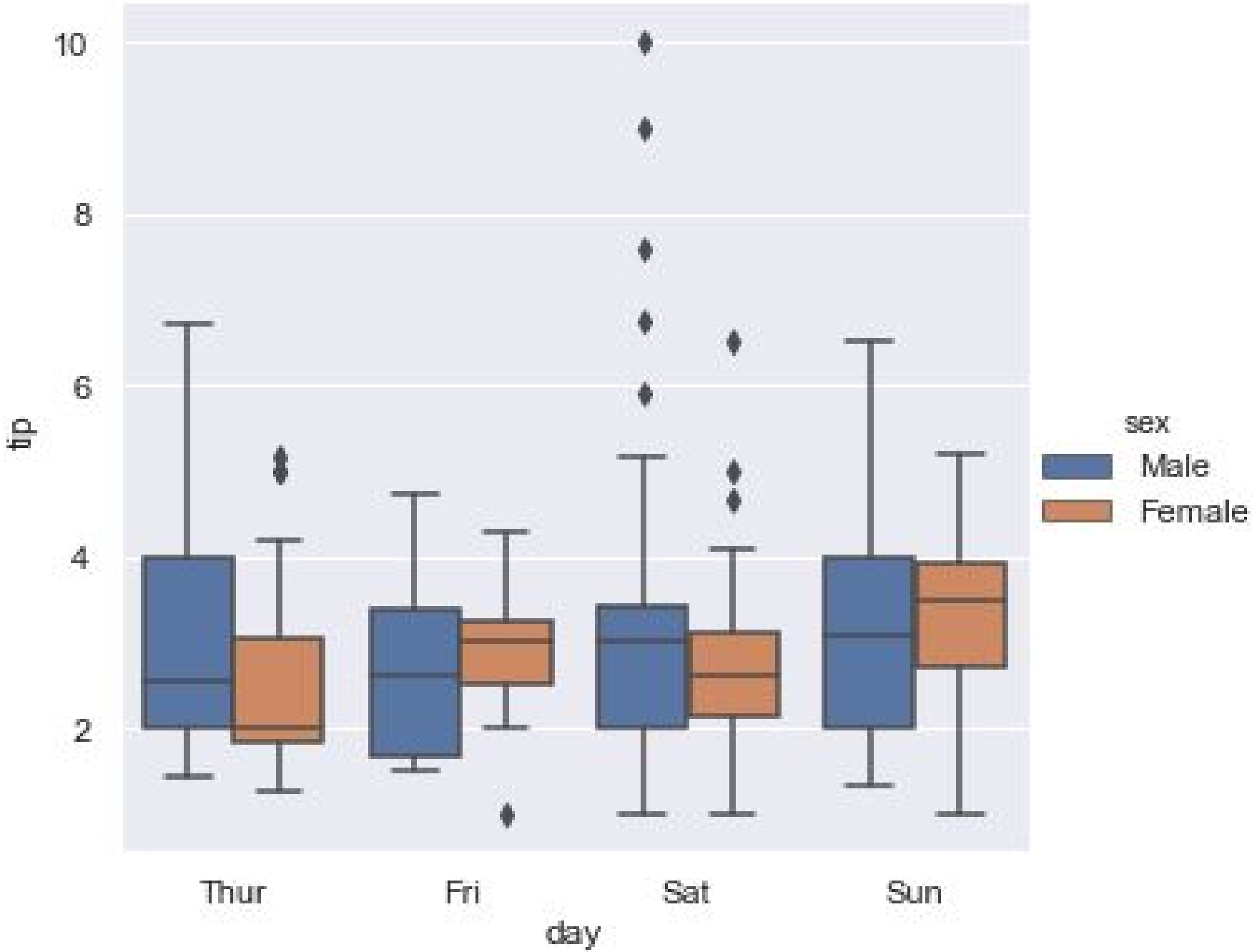
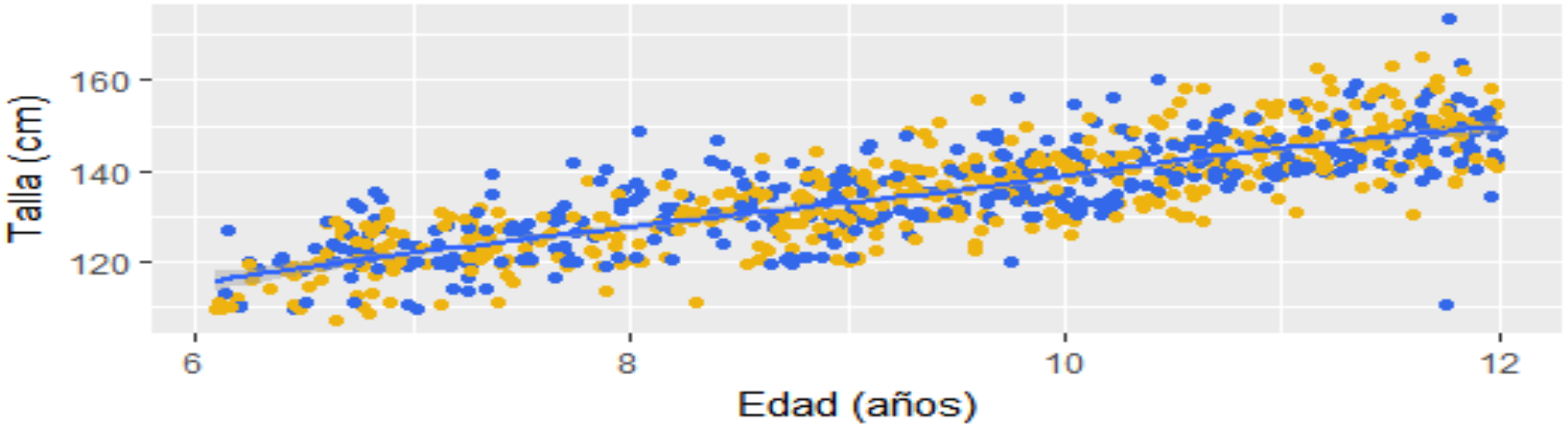
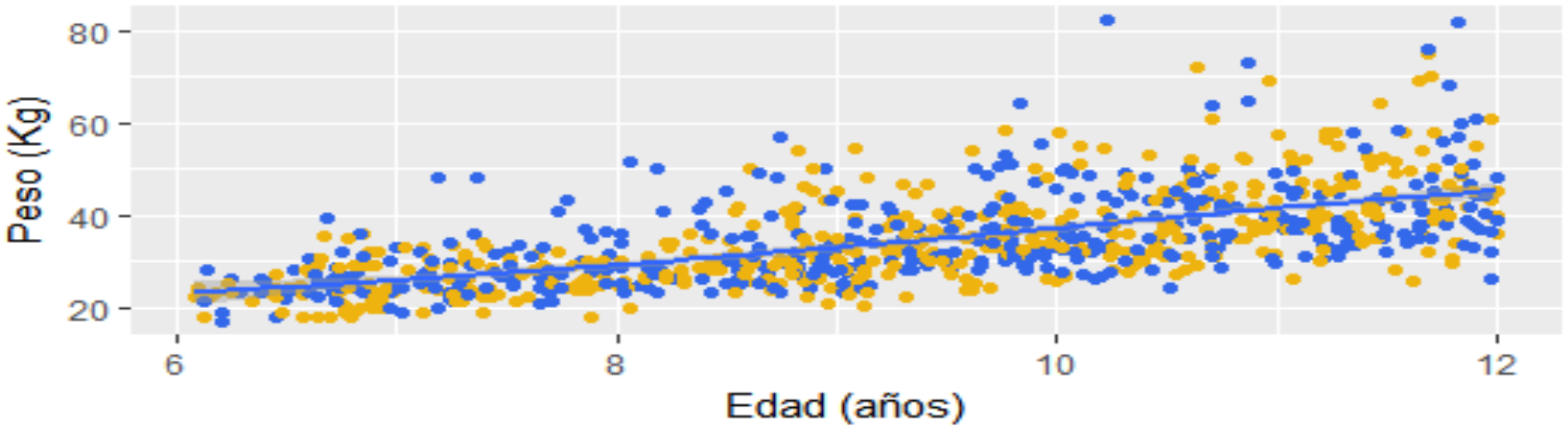
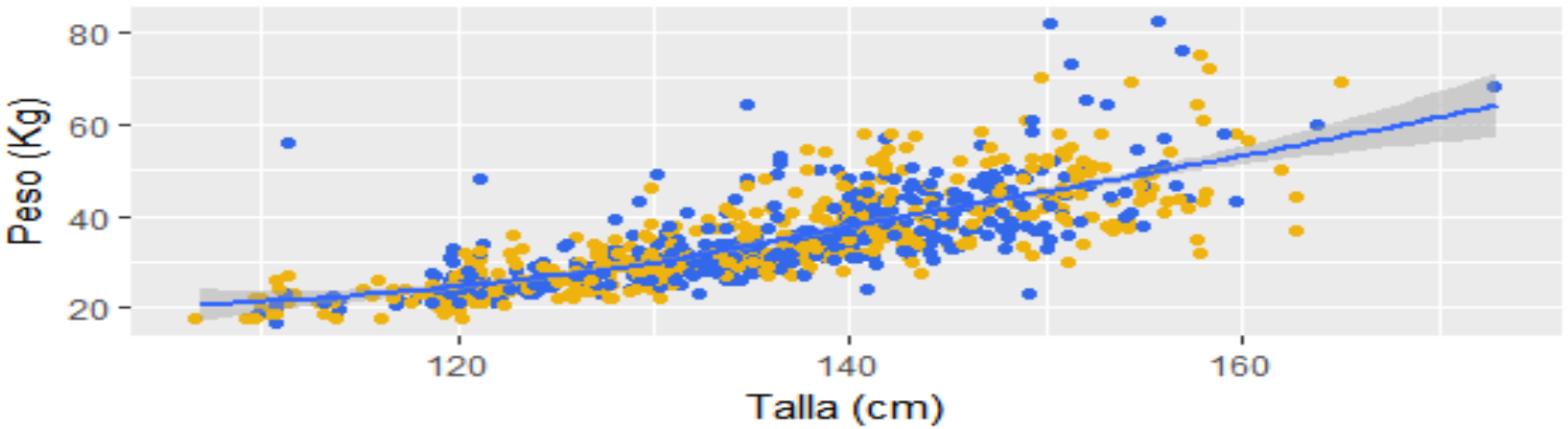

10
8
6
4
2
tip
Thur
Fri
Sat
Sun
day
sex
Male
Female

Gráfico de dispersión

Un diagrama de dispersión o gráfica de dispersión o gráfico de burbujas es un tipo de diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos.
Se emplea cuando una o varias variables está bajo el control del experimentador. Si existe un parámetro que se incrementa o disminuye de forma sistemática por el experimentador, se le denomina parámetro de control o variable independiente y habitualmente se representa a lo largo del eje horizontal (eje de las abscisas). La variable medida o dependiente usualmente se representa a lo largo del eje vertical (eje de las ordenadas). Si no existe una variable dependiente, cualquier variable se puede representar en cada eje y el diagrama de dispersión mostrará el grado de correlación (no causalidad) entre las dos variables.
Un diagrama de dispersión puede sugerir varios tipos de correlaciones entre las variables con un intervalo de confianza determinado. La correlación puede ser positiva (aumento), negativa (descenso), o nula (las variables no están correlacionadas). Se puede dibujar una línea de ajuste (llamada también "línea de tendencia") con el fin de estudiar la correla-


Talla (cm)
160
140
120
6
8
10
12
Edad (años)
Sexo
F
M


Peso (Kg)
80
60
40
20
6
8
10
12
Edad (años)
Sexo
F
M


Peso (Kg)
80
60
40
20
120
140
160
Talla (cm)
Sexo
F
M

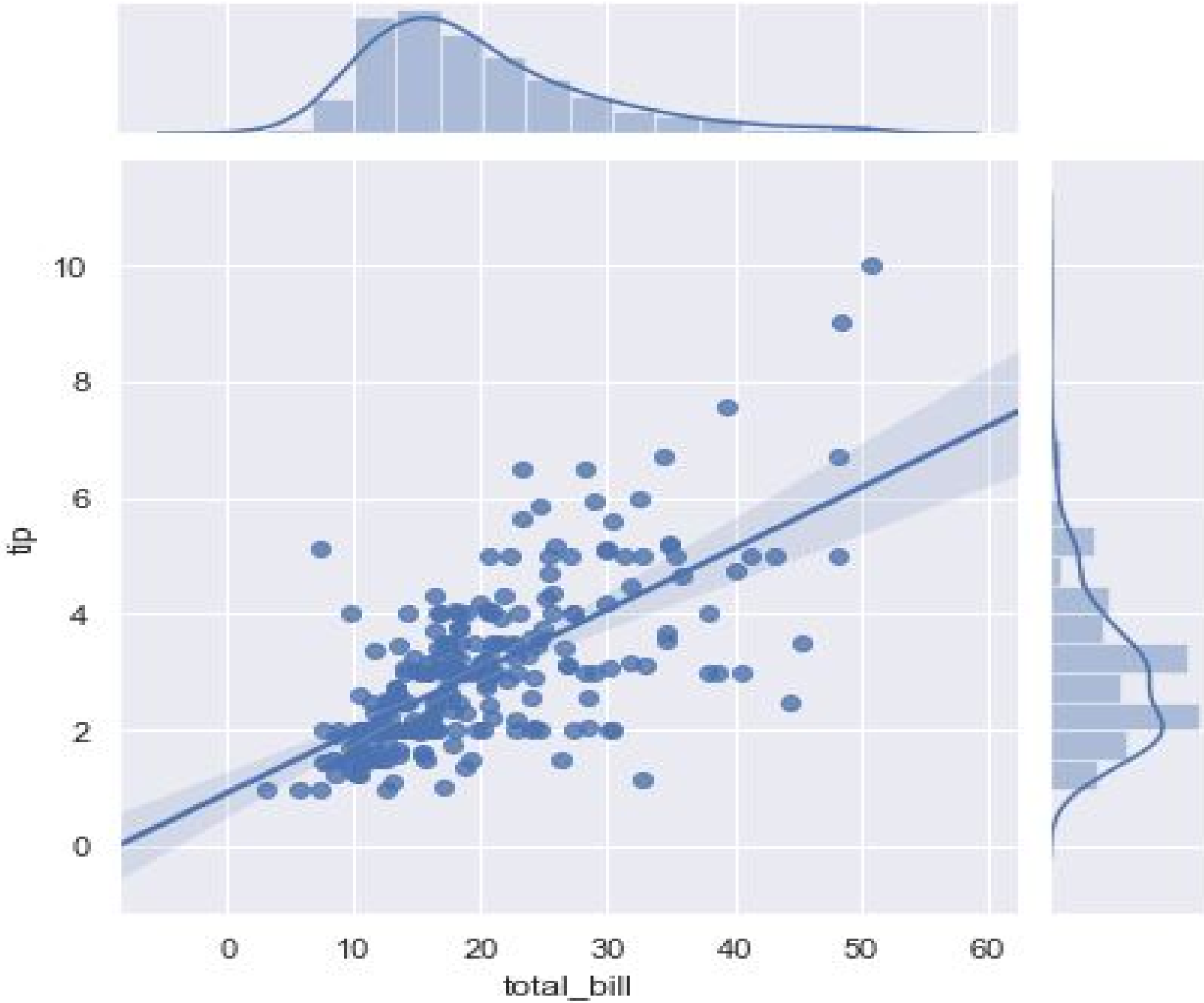
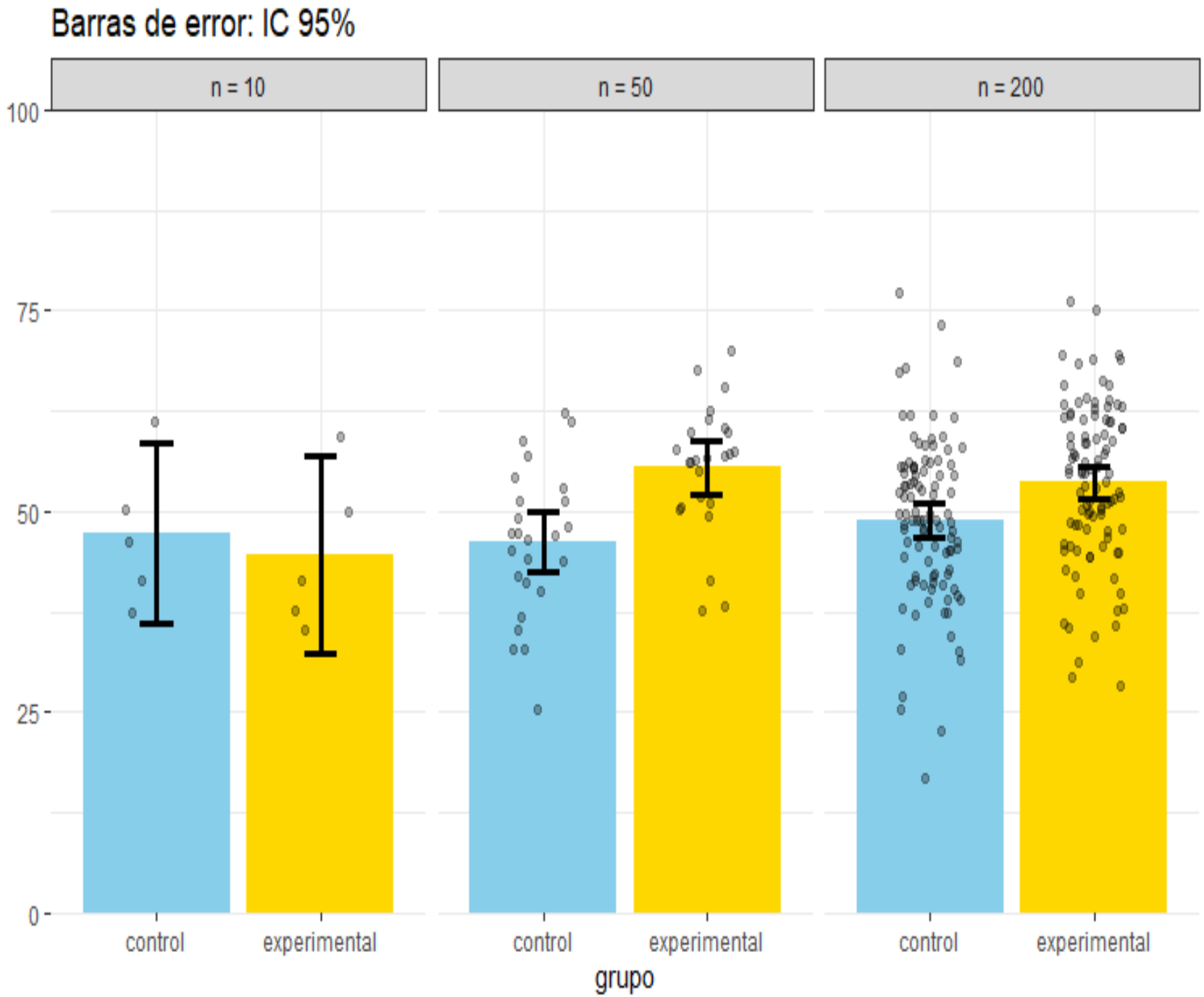

10
8
6
4
2
0
tip
0
10
20
30
40
50
60
total_bill

Gráfico de barra de error

Las barras de error son representaciones gráficas de la variabilidad de los datos, y se usan en gráficos para indicar el error o la incertidumbre en una determinada medida. Dan una idea general de lo precisa que es una medición o, a la inversa, a qué distancia del valor indicado puede estar el valor verdadero (sin errores) del elemento medido. Representan una incertidumbre utilizando una desviación típica, un error estándar o un intervalo de confianza particular (por ejemplo, un intervalo del 95 %). Estas cantidades no expresan necesariamente valores coincidentes, por lo que debe indicarse explícitamente en el gráfico o en el texto de apoyo cuál es el indicador del error utilizado.

Se pueden usar para comparar visualmente dos cantidades, e implícitamente, si se cumplen determinadas condiciones, permiten determinar a simple vista si las diferencias son de significación estadística. Las barras de error también pueden sugerir la bondad de ajuste de una función dada, es decir, la exactitud con la que la función describe los datos. Es habitual que los artículos científicos en el campo de las ciencias experimentales incluyan barras de error en todos los gráficos.


Barras de error: IC 95%
n = 10
n = 50
n = 200
100
75
50
25
0
control
experimental
control
experimental
control
experimental
grupo