CONDICIONES PARA GARANTIZAR UNA ADECUADA REPRESENTACIÓN GRÁFICA DE DATOS

Jamil Hernán Camayo Guachetá

Cristhian Fabián Solarte Borrero

Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la Educación

Programa de Matemáticas

Popayán

2024

CONDICIONES PARA GARANTIZAR UNA ADECUADA REPRESENTACIÓN GRÁFICA DE DATOS

Trabajo presentado como requisito parcial para optar al grado de Matemático

Jamil Hernán Camayo Guachetá

Cristhian Fabián Solarte Borrero

Director

Dr. Yilton Riascos Forero

Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la Educación

Programa de Matemáticas

Popayán

2024

Nota de aceptación

_____

_____

_____

_____

Director: _____

Dr. Yilton Riascos Forero

Jurado: _____

Dr. Luis Felipe Narváez

Jurado: _____

Mg. Jhon Alejandro Delgado

Lugar y fecha de sustentación: Popayán, 6 de Junio de 2024

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Resumen

Se realiza la descripción de algunos gráficos estadísticos para datos unidimensionales y multidimensionales. Los gráficos se clasifican según la dimensionalidad de los datos y la naturaleza de la variable. Se describen conceptos técnicos básicos y avanzados en la elección y construcción de gráficos con el fin de procurar una adecuada representación. Finalmente se proporcionan algunos ejemplos con errores comunes, la mayoría relacionados con las escalas, estandarizaciones y otros errores conceptuales.

Abstract

The description of some statistical graphs for unidimensional and multidimensional data is made. The graphs are classified according to the dimensionality of the data and the nature of the variable. Basic and advanced technical concepts are described in the selection and construction of graphics for the purpose of adequate representation. Finally, some examples with common errors are provided, most related to scales, standardizations and other conceptual errors.

Palabras clave: variable estadística, datos estadísticos, dimensionalidad de los datos, gráficos estadísticos, gráficos inadecuados.

Key words: statistical variable, statistical data, dimensionality of data, statistical graphs, inadequate graphs.

Introducción.

Una de las dificultades que más se presenta a los investigadores a la hora de reportar información cuantitativa es la forma en que se pretende hacerlo. La estadística ofrece herramientas adecuadas para resolver estas dificultades, como son las tablas y gráficos para la representación de datos, permitiendo "comprimir la información" ya que se pueden organizar, clasificar y ordenar los datos, teniendo en cuenta la escala de medición en la que fueron medidos.

En el análisis descriptivo de datos una de las herramientas más utilizadas es la representación gráfica. La ventaja de una buena representación gráfica de datos lleva a identificar características y situaciones atípicas en su comportamiento, por ende, proporciona un primer acercamiento para realizar el análisis de este comportamiento, lo cual lleva a que se obtenga más claridad de la información y posteriormente se puedan tomar mejores decisiones.

Cuando el análisis estadístico involucra muchas variables, en muchos casos estas pueden visualizarse fácilmente de forma individual, con representación de gráficos básicos, por ejemplo, histogramas, gráfico de cajas, gráficos de barras o pictogramas; pero no siempre es tan fácil. Cuando se tienen datos multivariados o en altas dimensiones, se requieren representaciones gráficas más complejas como correlogramas, gráfico de "amebas", gráfico de Andrews, de estrellas, de caras de Chernoff, etc. Lo que hace indispensable el conocimiento de estas herramientas para que, a través del uso de herramientas informáticas y programas estadísticos sofisticados, no se comentan errores que puedan comprometer la veracidad de la información.

De esta forma, la decisión de realizar una representación gráfica de datos genera, además de inquietudes para el investigador, dudas en relación a las características de los métodos de traficación, que exigen mayor conocimiento y cuidado al momento de determinar su uso.

Planteamiento del problema, objetivos y desarrollo.

En la mayoría de los textos se plantean procedimientos para manipular datos mediante tablas de frecuencias, estadísticas descriptivas y su visualización gráfica, pero muchos de estos no explican aspectos técnicos relacionados con el significado de las gráficas ni el porqué de ellas, lo que en muchos casos lleva a que el investigador o analista de datos elija equivocadamente el tipo de gráfica adecuada para representar las distribuciones de sus datos, sus indicadores o funciones.

Esta problemática trae como consecuencia debilidades en el trabajo investigativo, así como la aceptación de este para la publicación de los resultados, conllevando pérdida de recursos y esfuerzos en este tipo de procesos.

Pensando en este tipo de situaciones y observando la falta de explicaciones que apoyen la toma de este tipo de decisiones, se plantea como propuesta de trabajo describir *aspectos técnicos de algunos métodos para la representación gráfica que, a nivel univariado y multivariado, presenta la estadística.*

Se espera con esta monografía aportar información que permita apoyar el trabajo de investigadores e interesados en el procesamiento de datos al momento de utilizar la representación gráfica de datos que ofrece la ciencia estadística.

Justificación.

La ausencia de información conceptual sobre la representación gráfica de datos a llevado generar una algoritmización en el uso de esta herramienta apoyada en los software informáticos, observándose errores en la presentación e interpretación de los mismos en los diferentes medios informativos y divulgativos, que particularmente afecta la comunidad de académicos en la Universidad del Cauca, lo que hace necesaria la realización de un trabajo que sintetice y ponga a su alcance un documento que resulte de ayuda al momento de utilizar estas herramientas.

Igualmente, se espera que estos resultados permitan al lector ampliar la comprensión en la interpretación de la información estadísticamente que permanentemente es presentada para dar a conocer el estado de fenómenos de interés común.

Objetivo general.

Describir aspectos técnicos de algunos tipos de representación gráfica que, a nivel univariado y multivariado, presenta la estadística.

Objetivos específicos.

- Describir y conceptualizar algunos tipos de gráficos para datos de una variable aleatoria (discreta, continua).
- Describir y conceptualizar algunos tipos de gráficos para datos de dos variables aleatorias (cruzadas).
- Describir y conceptualizar algunos tipos de gráficos para datos multivariados.

Desarrollo.

Para dar solución a esta propuesta se presenta una metodología basada en 4 etapas que se describen a continuación:

1) Estudiar, describir y conceptualizar algunos tipos de gráficos para datos de una variable aleatoria (discreta, continua).
2) Estudiar, describir y conceptualizar algunos tipos de gráficos para datos de dos variables aleatorias
3) Estudiar, describir y conceptualizar algunos tipos de gráficos para datos multivariados.
4) Redactar y sistematizar conclusiones de las etapas anteriores.

Preliminares.
¿Qué son los datos? Para responder estos, se plantean brevemente conceptos esenciales para justificar algunas representaciones gráficas en un problema estadístico.

Población y muestra estadística.
De manera general, población es la totalidad de los elementos de interés de los cuales se desea obtener información y hacia los cuales se extenderán las conclusiones.

Esta definición causa dificultades para los estudiantes y usuarios neófitos de la estadística, debido a que hace alusión a elementos concretos (personas, casas, artículos, etc.) cuando son los datos los elementos a los que refiere la estadística. En este sentido, resulta práctico asumir la postura de *Azorin (1972, pág. 3)* que propone el término *universo* para indicar "un conjunto de elementos, seres u objetos" y *población* para indicar "un conjunto de números obtenidos midiendo o contando cierta característica de los mismos".

En este sentido, se define la muestra estadística como un subconjunto representativo de la población y es el resultado de implementar un método de muestreo estadístico.

Calot llama *unidades estadísticas o individuos* a los elementos componentes de la población, donde la población son un conjunto de elementos los cuales pueden ser personas u objetos físicos o abstractos.

Variable estadística.
Cada uno de los elementos del conjunto universo, también llamados individuos puede describirse mediante uno o varias características.

De manera general, una ***variable estadística*** es una característica de una población de interés para el cual su resultado no se puede anticipar. Esta se caracteriza por poder medirse o contarse. Las variables estadísticas pueden ser de naturaleza cualitativa o cuantitativa**.**

Los posibles resultados que presenta una característica se define como ***modalidades*** de la característica. Cada individuo presenta una y solamente una de las modalidades del carácter, lo que significa que las modalidades son incompatibles. Por ejemplo, la característica género tiene dos modalidades: masculino y femenino.

El número de modalidades puede varias de acuerdo a la información que se desea obtener. Por ejemplo, el carácter estado matrimonial puede tener dos modalidades: soltero, casado; Pero, también tres modalidades: soltero, casado, divorciado. Ahora, si consideramos la temperatura de una ciudad, las modalidades de cierto modo son infinitas, tenemos infinitos resultados posibles y un tratamiento a esto es trabajar con clases.

Variable cualitativa.
Se dice que una variable es cualitativa si sus diversas modalidades o valores no son medibles "numéricamente" (el resultado es un valor categórico), pero si en un sentido exclusivo de establecer diferencias para comparar u ordenar.

Por ejemplo, algunas características de este tipo para el personal de una empresa son: genero, estado civil, procedencia, estrato socioeconómico, etc.

Variable cuantitativa.

Se dice que una variable es cuantitativa si sus diversas modalidades o valores tienen como resultado valores numéricos (los cuales están en un discreto o continuo).

- **Variable estadística discreta.** Es aquella cuya naturaleza hace que el conjunto de valores posibles que puede tomar la variable sea finito o infinito numerable.
  Por ejemplo, miembros de una familia, número de hijos, número de bombillos defectuosos, etc.
- **Variable estadística continua.** Es aquella, cuya naturaleza hace que exista un intervalo de puntos, los cuales son valores que puede tomar la variable.
  Por ejemplo, la temperatura, velocidad de un auto, estatura, etc.

Vector estadístico.

A un individuo u, que pertenece a un colectivo estadístico de interés, se le miden una serie de variables estadísticas $X_1, X_2, X_3, \ldots, X_P$. El arreglo $(X_1, X_2, X_3, \ldots, X_P)$ se dice vector estadístico de p componentes o vector estadístico p-dimensional.

Por ejemplo, en el caso del personal de una empresa telefónica, se pueden considerar los siguientes caracteres: Edad, estado civil, genero, estatura, número de hijos, salario, antigüedad, etc. La producción de una fábrica de muebles, los caracteres serian: Tipo de mueble, color, material, etc.

Datos estadísticos y escalas de medición.

Un dato es la unidad estadística básica y es el resultado de medir una o varias características a un elemento que hacer parte de un colectivo de interés. Un dato por sí solo no me dice nada, pero un conjunto de datos sí. Los datos obtenidos en una investigación son la materia prima para la estadística.

La dimensionalidad de los datos juega un papel importante en el análisis de datos. Hablar de la dimensión de los datos, podemos referirnos como número de características o variables estadísticas que se están considerando.

La preparación de los datos para un análisis conlleva organizarlos (tablas o matrices), identificar datos faltantes y limpiar los datos para que sean de calidad y representativos. Que, junto a herramientas matemáticas, en especial herramientas del álgebra lineal, son la base para los modelamientos y tratamientos de datos multidimensionales.

Escalas de medición.

La medición puede definirse como la asignación de números a objetos y eventos de acuerdo con ciertas reglas. En general, cuando analizamos datos necesitamos conocer el contexto y el procedimiento de medición que los generó. La medición puede ser precisa, o poco precisa, pero siempre tendrá asociadas interpretaciones comparativas. Además, es importante clasificar los datos para tener un mejor entendimiento del fenómeno.

Los datos siempre estarán asociados a conceptos específicos llamados variables; éstas son los conceptos de referencia más importantes en la investigación ya que los datos son el

resultado de mediciones sobre estas variables o características. *Orlandoni (2010, pág. 245)* plantea que según *Stevens (1957),* las propiedades del sistema numérico asociadas con las escalas de medición son la identidad, magnitud, igual intervalo y cero absoluto:

1- *Identidad*: Cada número tiene un significado particular.
2- *Magnitud:* Los números tienen un orden inherente ascendente o descendente.
3- *Intervalos iguales*: Las diferencia entre números en cualquier punto de la escala son las mismas (la diferencia entre 10 y 20 es la misma que entre 100 y 110).
4- *Cero absoluto*: El punto cero en la escala de medición representa la ausencia de la propiedad que se estudia.

Hay una clasificación general de tipos de datos que se refiere a las escalas de medición propuestas por el psicólogo Stevens (citado por *Orlandoni, 2010*), la cual es casi universalmente aceptada. Los datos están referidos siempre a una de estas escalas. Siguiendo a *Orlandoni (2010, págs. 245-246),* se presentan las definiciones de estas escalas.

Escala Nominal*:* En esta escala las unidades observacionales (UO) se agrupan en clases excluyentes según determinada propiedad, con lo que se define una partición sobre el conjunto de tales unidades. Los números se usan como identificadores o nombres. Cuando se estudia el desempleo de un país y se incluye la variable sexo, se codifica masculino como 1 y femenino como 2, por ejemplo; los números 1 y 2 representan categorías de datos: son simples identificadores y son completamente arbitrarios. La operación matemática permitida es el conteo.

Escala Ordinal*:* Surge a partir de la operación de ordenamiento; en esta escala se habla de primero, segundo, tercero. No se sabe si quien obtiene el primer puesto está cerca o lejos del segundo puesto. Los valores de la escala representan categorías o grupos de pertenencia, con cierto orden asociado, pero no una cantidad mensurable. La escala ordinal tiene las propiedades de identidad y magnitud. Los números representan una cualidad que se está midiendo, y expresan si una observación tiene más de la cualidad medida que otra UO. La distancia entre puntos de la escala no es constante: no se puede determinar la distancia entre las categorías, solo es interpretable el orden entre sus valores. Ejemplos: situación socioeconómica, nivel educativo.

Escala de Intervalos: Esta escala representa magnitudes, con la propiedad de igualdad de la distancia entre puntos de escala de la misma amplitud. Aquí puede establecerse orden entre sus valores, hacerse comparaciones de igualdad, y medir la distancia existente entre cada valor de la escala. El valor cero de la escala no es absoluto, sino un cero arbitrario: no refleja ausencia de la magnitud medida, por lo que las operaciones aritméticas de multiplicación y división no son apropiadas. Cumple con las propiedades de identidad, magnitud e igual distancia. La igual distancia entre puntos de la escala significa que puede saberse cuantas unidades de más tiene una UO comparada con otra, con relación a cierta característica analizada. Por ejemplo, en la escala de temperatura centígrada puede decirse que la distancia entre 25°C y 30°C es la misma que la existente entre 20°C y 25°C, pero no puede afirmarse que una temperatura de 40°C equivale al doble de 20°C en cuanto a intensidad de calor se refiere, debido a la ausencia de cero absoluto. Así, los valores numéricos en la escala de

temperatura centígrada se pueden expresar en valores de la escala *Fahrenheit* (F) mediante la ecuación.

$$\mathrm{C} = -17.778 + \frac{5}{9}\mathrm{F}$$

***Escala de Razón**:* Corresponde al nivel de medición más completo. Tiene las mismas propiedades que la escala de intervalos, y además posee el cero absoluto. Aquí el valor cero no es arbitrario, pues representa la ausencia total de la magnitud que se está midiendo. Con esta escala se puede realizar cualquier operación lógica (ordenamiento, comparación) y aritmética. A iguales diferencias entre los números asignados corresponden iguales diferencias en el grado de atributo presente en el objeto de estudio. Ejemplos: longitud, peso, distancia, ingresos, precios.

Dimensionalidad de los datos.

La dimensionalidad se refiere a la cantidad de características que tiene un conjunto de datos. Por ejemplo, un conjunto de datos de atención médica puede tener múltiples variables como la presión arterial, el peso y el nivel de colesterol. Dado un vector estadístico p-dimensional, los resultados son datos de dimensión p.

Datos estadísticos unidimensionales.

Para un universo de elementos o individuos de interés, medir una característica, el resultado son datos unidimensionales. Los datos unidimensionales dependiendo de la escala a la que correspondan pueden referirse como discretos o continuos. Los nominales u ordinales al ser de naturaleza cualitativa pueden referirse como discretos, los valores que pueden tomar forman un conjunto finito. Los datos que se refieren como continuos toman valores de conjuntos de carácter continuo como intervalos de números reales.

Datos bidimensionales.

Para un individuo de interés medir dos características simultáneamente (vector estadístico de dimensión dos), se obtiene el par (a, b) como un dato bidimensional, cada componente es el valor de una variable estadística diferente.

Datos multidimensionales.

Los datos multidimensionales son los resultados de medir un vector estadístico (de dimensión mayor o igual a tres) a un colectivo de interés. Un dato p-dimensional es (x_1, x_2, ..., x_p).

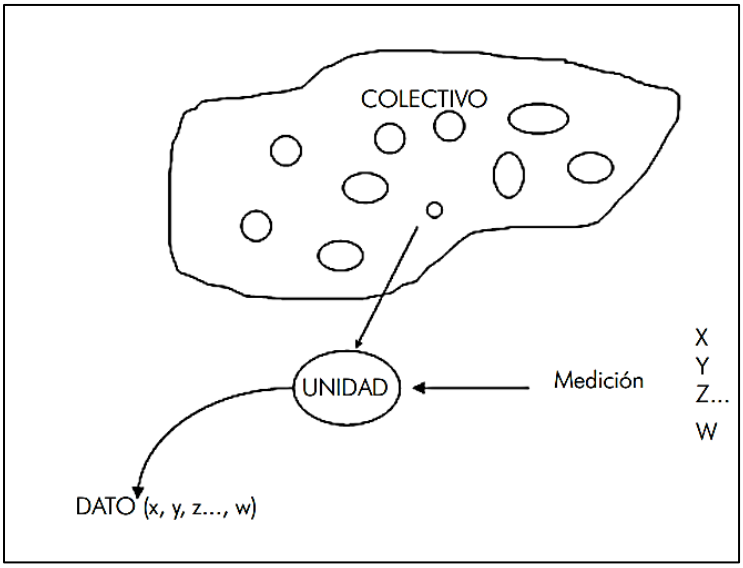


Figura 1. Elementos en un problema estadístico. Fuente: Behar-Ojeda.

Matriz de datos.

La organización de datos para un análisis multivariado se realiza generalmente en forma de una matriz con n filas, cada fila conteniendo las observaciones registradas sobre un mismo individuo, y p columnas, cada una representando una variable estadística. Cada fila es una observación multivariada.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Figura 2. Matriz de datos de n-filas y p-columnas. Fuente: Behar-Ojeda.

Nota. La matriz de datos multivariados se puede abordar de dos formas: desde el conjunto de individuos o desde las variables, el espacio fila o de individuos tiene dimensión p y el espacio columna o de variables tiene dimensión n. Las técnicas multivariadas se dirigen sobre alguno de estos dos espacios o sobre ambos simultáneamente.

Es importante considerar si algunas variables pesan sobre otras, debido a las distintas escalas que pueden ser medidas y esto puede afectar en los resultados de los análisis. La estandarización de los datos permite que cada variable tenga igual peso o significancia, mejorando la eficiencia en el manejo de los mismos. La elección del método dependerá del tipo de datos y del objetivo del análisis, algunos métodos son:

- Escalamiento estándar: Consiste en restar la media de la variable y dividir por la desviación estándar. De esta forma, los datos quedan escalados con media cero y una desviación estándar de uno.

- Escalamiento mín-máx: Se transforman los valores de la variable para que se encuentren en un rango específico, generalmente entre 0 y 1. Se calcula restando el valor mínimo y dividiendo por la diferencia entre el valor máximo y el valor mínimo.

Distancia euclídea entre dos observaciones.

La distancia entre dos observaciones X y Y *cuantitativas* de dimensión p, notado d(X,Y) es:

$$d(X,Y) := \left(\sum_{i=1}^{p} (x_i - y_i)^2 \right)^{1/2}$$

De la matriz de datos, se obtiene la matriz de distancias para analizar similitudes o diferencias entre las observaciones.

Datos atípicos.

En estadística, los valores atípicos, también llamados observaciones extremas u ***outliers***, son valores que se alejan del resto de conjunto de datos, pueden ser el resultado de errores en la entrada de datos, mala medición o inexactas, verdaderas referencias producto de eventos raros que se producen en circunstancias excepcionales y puede reflejar "desfase" o corrimiento en la media, la varianza o la correlación.

Para los datos unidimensionales caracterizados como outliers tiene gran importancia señalar que lo "raro", supone un criterio de lo que es "normal". Un punto puede ser raro, si se supone que la distribución de la cual proviene es Gaussiana (campana de Gauss), pero puede no serlo.

Para los datos multidimensionales, los casos atípicos deben analizarse en el conjunto de todas las variables consideradas. Puede ocurrir que una variable tenga valores extremos eliminables, pero al considerar las otras variables en el análisis, el investigador puede decidir no eliminarlos. Lo indica que un vector de observaciones puede ser un *outliers* debido a que alguna de sus componentes lo es. Por ende, es necesario detectarlos, los buenos resultados dependerán de aceptarlos, eliminarlos o ajustarlos, por ejemplo, en modelos de aprendizaje automático, los datos atípicos pueden afectar la precisión del modelo.

Algunas técnicas comunes para detectar los datos atípicos:

Pruebas gráficas: En general cualquier gráfico que represente los datos.
Pruebas numéricas: La prueba Z, el test de Dixon o el test de Grubs, cuyos p-valores detectan los valores atípicos, con análisis de regresión, análisis de componentes principales, la distancia (observación-dato de medias) euclídea, distancia *D2* de *Mahalanobis*.

Distribución estadística y tabla de frecuencia.
Los datos tomados bajo un proceso estadístico que no ha tenido ningún tipo de tratamiento se le llama *muestra bruta* de la variable estadística y pueden seguir o no, un comportamiento que pueden ser escritos de alguna manera. La forma como se comportan los datos se puede entender como la distribución estadística de los datos. Los datos muchas veces presentan distribuciones que se aproximan a distribuciones teóricas: normal, exponencial, poisson, binomial, etc.

Por ejemplo, un conjunto de datos que presenta un comportamiento normal, puede significar que la variable o característica que se está estudiando en una población sigua una distribución normal. La importancia cuando se desea identificar algunos parámetros desconocidos: los conceptos inferenciales proceden a determinar la media poblacional, varianza poblacional, etc.

Tabla de frecuencia de una y dos variables.
Una forma de representar el comportamiento o distribución de los datos unidimensionales y bidimensionales es mediante el uso de tablas estadísticas o tablas de distribución de frecuencia, esto permite representar información de forma resumida. Las distribuciones más simples son aquellas que aceptan una sola característica.

Tabla de frecuencias para una variable cualitativas.
Consideremos un conjunto de datos de tamaño n, descritos por la característica C, donde las distintas modalidades(categorías) son C_1, C_2, C_3, …, C_i, …, C_m.

Designemos por n_i el número de individuos que presentan la modalidad C_i, se dice entonces que n_i es la frecuencia absoluta de la modalidad C_i.

La proporción:

$$f_i = \frac{n_1}{n}$$

es la frecuencia relativa de la modalidad C_i.

Las modalidades son incompatibles y permiten:

$$\sum_{i=1}^{m} n_i = n \qquad y \qquad \sum_{i=1}^{m} f_i = 1$$

Modalidades de la característica C	***Frecuencia absoluta de cada modalidad***	***Frecuencia relativa de cada modalidad***
C_1	n_1	f_1
C_2	n_2	f_2
⋮	⋮	⋮
C_i	n_i	f_i
⋮	⋮	⋮
C_m	n_m	f_m
Total	N	1

Figura 3. Representación general de frecuencias de una variable cualitativa. Fuente: Elaboración propia.

Tabla de frecuencias para una variable discreta.

Es una extensión a la tabla anterior, donde las modalidades al ser numéricas tienen orden y magnitud, además, con un agregado: frecuencias acumuladas.

Considere la muestra de tamaño n dada como: $x_1, x_2, \dots, x_n$.

Considere las m modalidades como: $x_1, x_2, \dots, x_m$.

La frecuencia absoluta del dato x_i, se denota por n_i y la relativa por f_i.

Es natural preguntarse sobre el número de datos que son menores o iguales a x_i, esto se denomina *frecuencia absoluta acumulada hasta* x_i, se denota por N_i.

Si $x_1, x_2, \dots, x_m$ están ordenada en forma creciente, entonces:

$$N_i = n_1 + n_2 + n_3 + \cdots + n_i$$

Si la frecuencia absoluta acumulada la expresamos como una fracción o porcentaje de toda la muestra, aparece lo que se denomina *frecuencia relativa acumulada*, denotada por F_i.

$$F_i = \frac{N_i}{n} = f_1 + f_2 + \cdots + f_i$$

Valor de la variable	***Frecuencia absoluta***	***Frecuencia relativa***	***Frecuencia acumulada absoluta***	***Frecuencia acumulada relativa***
x_1	n_1	f_1	N_1	F_1
⋮	⋮	⋮	⋮	⋮
x_i	n_i	f_i	N_i	F_i
⋮	⋮	⋮	⋮	⋮
x_m	n_m	f_m	N_m=n	F_m=1
Total	N	1		

Figura 4. Representación general de frecuencias de una variable discreta. Fuente: Elaboración propia.

Propiedades.

Considere una muestra de n datos con m categorías, que ordenados en forma creciente son $x_1, x_2, \dots, x_m$.

Entonces:

1) $0 \leq n_i \leq n$; i=1, 2, …, m
2) $n_1 + n_2 + \dots + n_m = n$
3) $0 \leq f_i \leq 1$; i=1, 2, …, m
4) $f_1 + f_2 + \dots + f_m = 1$
5) $N_j = n_1 + n_2 + \cdots + n_j$
6) $N_m = n$
7) $n_1 = N_1 \leq N_2 \leq \cdots \leq N_m = n$
8) $F_j = f_1 + f_2 + \cdots + f_j$
9) $f_1 = F_1 \leq F_2 \leq \cdots \leq F_m = 1$

A continuación, se redefinen las frecuencias acumuladas como funciones sobre todos los números reales.

1) $N(x) :=$ número de datos que son menores o iguales a x.

$$N(x) = \begin{cases} 0 & \text{si } x < x_1 \\ N_j & \text{si } x_j \leq x < x_{j+1} \quad j = 1, 2, \dots, m \\ n & \text{si } x \geq x_m \end{cases}$$

2) $F(x) :=$ fraccion (o porcentaje) de datos que son menores o iguales a x.

$$F(x) = \begin{cases} 0 & \text{si } x < x_1 \\ F_j & \text{si } x_j \leq x < x_{j+1} \quad j = 1, 2, \dots, m \\ 1 & \text{si } x \geq x_m \end{cases}$$

Las funciones son monótonas no decrecientes.

Si $x_1 \leq x_2$, entonces $N(x_1) \leq N(x_2)$ *y* $F(x_1) \leq F(x_2)$.

La función F(x) es conocida como *función empírica de distribución acumulada*, para indicar que se obtuvo con base a una muestra, pretendiendo con ella aproximarse a la distribución acumulada de la población, es decir a la función de distribución acumulada de probabilidad.

Tabla de frecuencia para una variable continua.

Para este caso, la tabla de frecuencias es un poco más compleja, tiene otros conceptos como: intervalos de clases, marca de clases, densidad. Suponga una muestra de n datos, la mayoría

de los datos son distintos o seguramente todos. No es de interés conocer la frecuencia que toma un valor de la variable. Bajo estas circunstancias, se procede a agrupa la información en los llamados ***intervalos de clase.***

A partir de los n datos se proceden los pasos:

1) Identificar los extremos: elemento mínimo y máximo como mín y máx
2) Se define el rango de la muestra:

$$\text{Rango} = \text{máx} - \text{mín}$$

3) Definir el número de intervalos: m. Por ejemplo, *regla de Sturges.*
4) Determinar los límites de los m intervalos de clases. $(L_{i-1}, L_i]$.
5) Los intervalos pueden ser de longitudes distintas; Cuando sea posible debe procurarse que todos los intervalos sean de igual longitud.

$$C \approx \frac{\text{Rango}}{m}$$

6) Marca de clase: Es el representante de los datos que pertenecen al intervalo correspondiente, se presenta por:

$$x_i' = \frac{L_{i-1} + L_i}{2}$$

Para determinar la frecuencia asociada a cada intervalo, deben contarse la cantidad de datos que pertenezcan a cada uno y se retoman las definiciones de frecuencias n_i, f_i, N_i, F_i abordadas para la variable discreta, al igual que las propiedades.

Clase $\mathbf{n_0 . i}$	***Limites intervalo de clase*** $\mathbf{(L_{i-1}, L_i]}$	***Marca de clase*** $\mathbf{x_i'}$	***Frecuencia absoluta*** $\mathbf{n_i}$	***Frecuencia relativa*** $\mathbf{f_i}$	***Frecuencia acumulada absoluta*** $\mathbf{N_i}$	***Frecuencia acumulada relativa*** $\mathbf{F_i}$
1						
2						
⋮	⋮	⋮	⋮	⋮	⋮	⋮
M					N	1
Total			n	1		

Figura 5. Representación general de frecuencias de una variable continua. Fuente: Elaboración propia.

El número de intervalos y las longitudes está condicionado a los objetivos que pretenda el investigador (conveniencia del problema). Cuando los datos se agrupan en intervalos de clase, se produce pérdida de información, puesto que no se dispone de los datos en forma individual sino una caracterización más global, Por otro lado, si se construyen demasiados intervalos se hace compleja y su presentación poco comprensible. Se recomienda un número de intervalos mayor que cinco y menor que veinte. No deben existir intervalos de clase que no contengan datos.

Concepto de densidad.

La densidad o densidad de frecuencia relativa de cada intervalo puede entenderse como la concentración de datos y consiste en expresar la fracción de los datos que hay por unidad de intervalo de clase.

Supongamos que los datos en cada intervalo están uniformemente distribuidos, se puede definir la densidad, notada como: f_i^* en el i-*ésimo* intervalo.

$$f_i^* = \frac{f_i}{C_i}$$

la densidad puede redefinirse como una función para todo número real x, llamada *función empírica de densidad*. Así,

$$f^*(x) = \begin{cases} 0 & \text{si } x \leq L_0 \,,\ x > L_m \\ \dfrac{f_i}{C_i} & \text{si } L_{i-1} < x \leq L_i \ , i = 1, \dots, m \end{cases}$$

También se puede construir en forma general, para cualquier x, el porcentaje (o fracción) de datos que son menores o iguales que x. Se denota por F(x) y se conoce como *función empírica de distribución acumulativa.*

Sea x que pertenece al intervalo $(L_{i-1}, L_i]$ el cual tiene una longitud C_i y una frecuencia relativa f_i, e interesa conocer la frecuencia relativa acumulada hasta x.

Suponga uniformidad en la distribución de los datos en cada intervalo, se puede plantear el porcentaje

$$\frac{f_i}{C_i}(x - L_{i-1})$$

Por lo tanto, el acumulado es

$$F(x) = F(L_{i-1}) + \frac{f_i}{C_i}(x - L_{i-1})$$

Se define la *función empírica de distribución acumulativa*:

$$F(x) = \begin{cases} 0 & \text{si} \quad x \leq L_0 \\ F(L_{i-1}) + \dfrac{f_i}{C_i}(x - L_{i-1}) & \text{si} \quad L_{i-1} < x \leq L_i \ , \quad i = 1, \dots, m \\ 1 & \text{si} \quad x \geq L_m \end{cases}$$

Con $f_i^* = \frac{f_i}{C_i}$

$$F(x) = \begin{cases} 0 & \text{si} \quad x \leq L_0 \\ F(L_{i-1}) + f_i^*(x - L_{i-1}) & \text{si} \quad L_{i-1} < x \leq L_i \ , \quad i = 1, \dots, m \\ 1 & \text{si} \quad x \geq L_m \end{cases}$$

La acumulativa representa el área bajo la función empírica de densidad. Si se desea estimar el porcentaje de datos que hay entre "a" y "b", se define $H(a, b)$ como:

$$H(a, b) = F(b) - F(a)$$

De la tabla de frecuencias y de las funciones $f^*(x)$, $F(x)$ se construyen representaciones gráficas: el histograma, ojiva, polígono de frecuencias.

La palabra empírica hace referencia que la función proviene de una muestra y pretende ajustarse a la función de densidad de probabilidad que indica el comportamiento de la variable en la población.

Tabla de doble entrada.

A continuación, se dará una herramienta para datos bidimensionales, donde las dos variables a considerar son de naturaleza cualitativa o cuantitativa.

Cuando se recopilan datos a menudo son de dos caracteres, por ejemplo: deporte y género, peso y estatura, y el interés es identificar las relaciones que guardan, en este sentido la importancia de la distribución de frecuencias, considerando conjuntamente los valores (o categorías) de las variables, permitiendo identificar el comportamiento de los datos y responde preguntas como: ¿Qué tan probable es que individuo sea hombre y prefiera voleibol como su deporte favorito?

El primer paso es organizar los datos mediante una tabla llamada *tabla de doble entrada*: tiene forma matricial donde filas y columnas reflejan las modalidades (o clases de intervalos) de las variables correspondientes. Básicamente es una tabla de frecuencias que presenta la distribución (distribución conjunta de frecuencias) de un conjunto de datos bidimensionales.

Es de interés explorar el grado de asociación que tienen dos características sobre los elementos de cierta población. Los métodos y herramientas matemáticas que permiten lograr describir la asociación o el grado de asociación entre dos características está ligado a la naturaleza de las variables, por ende, es de importancia reconocer el tipo de variables que se están involucrando en una investigación estadística. Algunos textos usan el termino: *tablas de contingencia* para referirse a las tablas de doble entrada donde ambas variables son cualitativas.

Sea un conjunto de n datos: (x_1, y_1), (x_2, y_2), (x_3, y_3), …, (x_n, y_n), determinados por el vector estadístico (X, Y).

En general se usará la siguiente notación: X_1, X_2, ..., X_i, ..., X_m representan las "m" categorías a considerar para clasificar los elementos de la muestra en lo que respecta a la variable X. Estas categorías pueden corresponder a nombres si se trata de escala nominal de las variables cualitativas, puede coincidir con los valores que toma la variable X si es discreta o pueden representar intervalos de clase si X es una variable continua.

Análogamente Y_1, Y_2, ... Y_j, ..., Y_s. representan las "s" categorías a considerar para clasificar los elementos de la muestra con respecto a la variable Y.

Las dos variables a considerar dan siguientes situaciones, cada una con sus respectivas tablas doble entrada:

1) Ambas variables (X,Y) son discretas.
2) Ambas variables (X,Y) son continuas.
3) Una variable discreta X y la otra continua Y.
4) Una variable es cualitativa X y la otra discreta Y.
5) Una variable es cualitativa X y la otra continua Y.
6) Ambas variables (X,Y) son cualitativas.

Ambas variables son discretas.

Similar al caso de una variable, se organizan los datos mediante el conteo y se construye la tabla doble entrada.

$n =$ Tamaño de la muestra.

$n_{ij} =$ N° elementos que pertenecen en forma simultámea a las categorías X_i y Y_j

$n_{i.} =$ N° de elementos de la muesra que pertenecen a la categoría X_i

$n_{.j} =$ N° de elementos de la muesra que pertenecen a la categoría Y_j

$f_{ij} = \frac{n_{ij}}{n}$, fracción (o %) de elementos que pertenecen a las categorías X_i y Y_j

X \ Y	Y_1	Y_2	...	Y_j	...	Y_s	
X_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1s}	$f_{1\cdot}$
X_2	f_{21}	f_{21}	...	f_{2j}	...	f_{2s}	$f_{2\cdot}$
:	:	:	:	:	:	:	:
X_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{is}	$f_{i\cdot}$
:	:	:	:	:	:	:	:
X_m	f_{m1}	f_{m2}	...	f_{mj}	...	f_{ms}	$f_{m\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$	...	$f_{\cdot j}$	...	$f_{\cdot s}$	1.00

Figura 6. Distribución conjunta de frecuencias relativas de las variables X y Y. Fuente: Behar-Yepes.

Tiene en sus márgenes dos frecuencias, estas distribuciones se conocen como *distribuciones marginales de frecuencia relativa*, son las distribuciones por separado de las variables, o bien de X, o bien de Y.

Se introduce el concepto de distribución conjunta de frecuencias acumuladas (absolutas o relativas), que puede denotarse como:

N_{ij} : número de elementos menores o iguales simultáneamente a las categorías de X_i, Y_j.

F_{ij} : porcentaje (o fracción) de N_{ij}.

X / Y	Y_1	Y_2	…	Y_s
X_1	f_{11}	$f_{11} + f_{12}$		
X_2	$f_{11} + f_{21}$	$f_{11} + f_{12} + f_{21} + f_{22}$		
⋮			⋮	
X_m				1

Figura 7. Distribución conjunta de frecuencias relativas acumuladas. Fuente: Elaboración propia.

Si redefinimos las conjuntas acumulativas para cualquier par (x, y) de $\mathbb{R}^2$, se obtiene la llamada *función empírica de distribución conjunta acumulativa.*

$N(x, y) = N°$ de elementos cuya característica X es menor o igual que x, su característica Y es menor o igual que y

$F(x, y) = \frac{N(x,y)}{n}$ fracción (o %)de elementos para los cuales $X \leq x$ y $Y \leq y$

Propiedades. Son una extensión al caso de una dimensión.

Sea

$$X_1 < X_2 < \cdots < X_m$$
$$Y_1 < Y_2 < \cdots < Y_s$$

si $x < X_1, y < Y_1$ entonces $F(x,y) = 0$

si $x \geq X_m, y \geq Y_s$ entonces $F(x,y) = 1$

si $x < x^*, y < y^*$ entonces $F(x,y) < F(x^*, y^*)$

Ambas variables son continuas.

En este caso, cada modalidad no representa un número, sino un intervalo de clase. Para cada variable se construyen los intervalos de clase como en el caso unidimensional, además de un trabajo similar cuando ambas variables son discretas como propiedades y tablas de la distribución conjunta de frecuencia. También aparece el concepto de densidad y se habla de función empírica de densidad conjunta de las variables X y Y.

Primero se habla de ***regiones o áreas*** que son determinada por los intervalos de clases de cada variable. Las áreas de estos rectángulos se pueden interpretar como las "intersecciones de los intervalos de clases".

Suponga que los datos están uniformemente distribuidos. Se define la densidad conjunta como la densidad por unidad de área, notada f^*.

$$f_{ij}^* = \frac{f_{ij}}{A_{ij}}$$

- A_{ij} es el área, el cual es producto de la longitud del intervalo de clase i- *ésimo* por el intervalo de clase j-*ésimo.*
- $f_{ij} = \frac{n_{ij}}{n}$ es la frecuencia conjunta relativa en la posición i, j.

Redefiniendo f_{ij}^* para cualquier punto de $\mathbb{R}^2$, se obtiene la llamda *función empírica de densidad conjunta* de X y Y.

$$f^*(x,y) = f_{ij}^*; \ (x,y) \text{ en } A_{ij}.$$

$$f^*(x,y) = 0; \ \text{Otro caso}$$

Se define para (x,y) en el plano la *función empírica de distribución conjunta acumulada para las variables* X y Y, notada por F(x,y); Como una generalización para el caso de variable continua, se obtiene el cálculo del volumen correspondiente de la región $(-\infty, x)\mathbf{X}(-\infty, y)$.

Caso discreta X *y continua* Y.

Para X las modalidades se constituyen por valores discretos, en cambio para Y se deben construir intervalos de clase. De esta manera se pueden clasificar y contar los datos de la muestra para construir la tabla doble entrada y representar la distribución conjunta de frecuencias relativas o absolutas para (X_i, Y_j); Además, las propiedades siguen siendo válidas.

Por otra parte, tiene sentido hablar de la función empírica de densidad de Y, más no de X puesto que es discreta; No sería muy adecuado referirse a la función empírica de densidad conjunta de (X,Y); Pero, por conveniencia, se va a usar el nombre de función empírica de densidad conjunta f*(x,y), haciendo la precisión de su significado y tratamiento.

La convenida función empírica de densidad conjunta, resulta de fraccionar la frecuencia relativa f_{ij} por unidad de intervalo de Y_j

$$f_{ij}^{*} = \frac{f_{ij}}{C_j}$$

C_j = longitud del intervalo Y_j.

f_{ij}^{*} es una densidad por unidad lineal y no por área.

Se define para cualquier par (x,y) del plano la *funcion empirica de densidad conjunta para las variables* X y Y,

$$f^{*}(x,y) = \begin{cases} f_{ij}^{*} & \text{si } (x,y) \epsilon (X_i \cap Y_j);\ i = 1,2,\dots,m;\ j = 1,2,\dots,s \\ 0 & \text{Otro caso.} \end{cases}$$

Es posible definir para (x,y) en el plano la *función empírica de distribución conjunta acumulada* para las variables X y Y, notado F (x, y).

Caso discreta X *y cualitativa* Y.

Las tablas son semejantes cuando ambas variables son discretas. Solo que las categorías para la cualitativa y numérica para la discreta; Para este caso, no se define la función acumulativa conjunta, algunas propiedades siguen siendo válidas.

Caso continua X *y cualitativa* Y.

Solo contamos con las tablas de frecuencias conjunta y algunas propiedades. No hay función de densidad conjunta, tampoco función acumulativa conjunta debido a la variable cualitativa.

Ambas variables son cualitativas.

En particular, la tabla doble entrada se llama ***tabla de contingencia*** y conserva la misma estructura que las tablas anteriores y algunas propiedades; Donde, las modalidades que toma ambas variables son categorías. Las funciones empíricas como la acumulada conjunta, la densidad conjunta, se definen para los casos en donde ambas variables son de naturaleza cuantitativa, ya que, al ser funciones, estas deben recibir puntos del plano cartesiano y devolver un punto de la recta real.

Algunas medidas numéricas descriptivas.

Se determina algunos estadísticos, los cuales son valores representativos de los datos que describen comportamientos "globales", como la centralidad, dispersión, cuartiles, o correlación para el caso de dos variables, entre otras. El tratamiento de datos conlleva un estudio descriptivo que paralelamente usan representaciones gráficas, dando pie al análisis de los datos.

Media aritmética.

También conocida como promedio aritmético, valor promedio y, en teoría de la probabilidad, como valor esperado o esperanza matemática; Es considerada, en estadística, la principal medida de tendencia central porque, además de cumplir todas las características propuestas por *Yule*, posee las mejores características algebraicas (*Riascos*, 2013).

Simbólicamente se representa por la letra griega $\boldsymbol{\mu}$ cuando se habla de la inclusión de todos los posibles valores (como población) de la variable involucrada en el fenómeno observado o $\overline{X}$ si se trata de una parte de ellos (muestra). Se define estadísticamente como el valor alrededor del cual oscilan o tienden a concentrarse los datos.

Atendiendo al tipo de variable estadística, el cálculo se realiza de forma diferente; así, para el caso de una variable de tipo discreta, el valor corresponderá a la suma de los datos dividida por el total de ellos, o como la integral, en el rango definido, del producto de valores la variable por su correspondiente valor de la función de densidad f(x), si se trata de una variable de tipo continua.

Simbólicamente

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{para el caso de una variable de tipo discreta.}$$

$$\overline{X} = \int_{-\infty}^{\infty} xf(x)dx \quad \text{para el caso de una variable de tipo continua.}$$

Estas fórmulas se adecúan a las características que presente la colección de datos; por ejemplo, si los valores aparecen en una tabla de frecuencia, el cálculo se realiza multiplicando cada valor de la variable x_i por su frecuencia f_i observada.

Simbólicamente

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{m} x_i f_i \quad \text{donde m es el total de categorías y n el total de datos.}$$

Esta forma de la ecuación de la media aritmética se conoce en la literatura como ***media ponderada o promedio ponderado.***

Varianza.

Si bien la media aritmética es la más importante de las medidas de tendencia central, la varianza es considerada la medida de mayor prioridad para la estadística, debido a que permite el estudio de la dispersión de los datos alrededor de la media.

Matemáticamente se define de acuerdo al tipo de variable estadística así:

La varianza de una muestra de mediciones $x_1, x_2, \ldots, x_n$ es la suma del cuadrado de las diferencias entre las mediciones y su media, dividida entre n–1. Simbólicamente, la varianza muestral es

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

Cuando los datos se presentan agrupados en una tabla de frecuencias de m categorías, el cálculo de la varianza se ajusta de la siguiente forma:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{m}(x_i - \overline{X})^2 f_i$$

Desviación estándar.

Si bien la varianza resulta ser la medida más importante en la estadística, su interpretación directa no resulta sencilla al realizar un análisis descriptivo. Por esta razón, se cuenta con un indicador que resulta de la transformación de la varianza y se denomina la Desviación Estándar e indica la dispersión lineal de los datos alrededor de la media aritmética y se representa de la siguiente manera:

$$S = \sqrt{S^2}$$

La covarianza y el coeficiente de correlación lineal.

La descripción de dos variables cuantitativas se logra con los diagramas de dispersión y se complementa con la covarianza y el coeficiente de correlación. Con estos dos conceptos se pretende detectar o conocer la fuerza de asociación estadística entre dos variables en la dirección de una línea recta.

Covarianza.

Sea (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) una muestra de tamaño n, que proviene de observar las características X y Y. Es decir, del vector (X,Y*).*

La covarianza entre las variables X y Y, denotada por Cov(X,Y) (o S_{XY} para una muestra) se define como:

$$S_{XY} = \sum_{i=1}^{n} \frac{(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

la covarianza es afectada por los cambios de escala, esto hace que su magnitud dependa de las unidades en que se midan las variables X y Y, lo cual no es bueno cuando se trata de conocer si la covarianza es "grande" o no para obtener una idea sobre el grado de relación lineal entre las variables.

Coeficiente de correlación de Pearson.

Una forma de corregir ese inconveniente que presenta la covarianza dio origen al denominado coeficiente de correlación lineal (o coeficiente de ***correlación de Pearson***) entre X y Y es denotado por γ_{XY} y se define como

$$\gamma_{XY} = \frac{S_{XY}}{S_X \,.\, S_Y}; \qquad -1 \leq \gamma_{XY} \leq 1$$

Donde, S_X , S_Y desviación estandar de X y Y respectivamente.

Esto facilita la interpretación de la fuerza de asociación de la relación entre las dos variables. Además, la covarianza puede ser negativa, a diferencia de la varianza que no lo es.

Veamos que la covarianza entre X y X es la varianza de X:

$$S_{XX} = \sum_{i=1}^{n} \frac{(x_i - \bar{X})(x_i - \bar{X})}{n-1}$$

$$= \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n-1}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$$

$$= {S_X}^2$$

Nota. Si no hay asociación implica incorreción, pero el recíproco no siempre es cierto; Recordemos que la incorreción se refiere a la ausencia de relación ***lineal***. Por ejemplo, dos variables X y Y pueden estar relacionadas por una función matemática no lineal con correlación lineal cero.

Estudiar la asociación depende de la naturaleza de las variables, se usa el *coeficiente de V Cramer* o el *coeficiente de contingencia*, basados en el estadístico ji-*cuadrado o el coeficiente de Spearman* y para variables cuantitativas, usa, por ejemplo, el *coeficiente de correlación de Pearson*. También, con un análisis gráfico de datos categóricos, es posible determinar asociaciones existentes entre pares de variables.

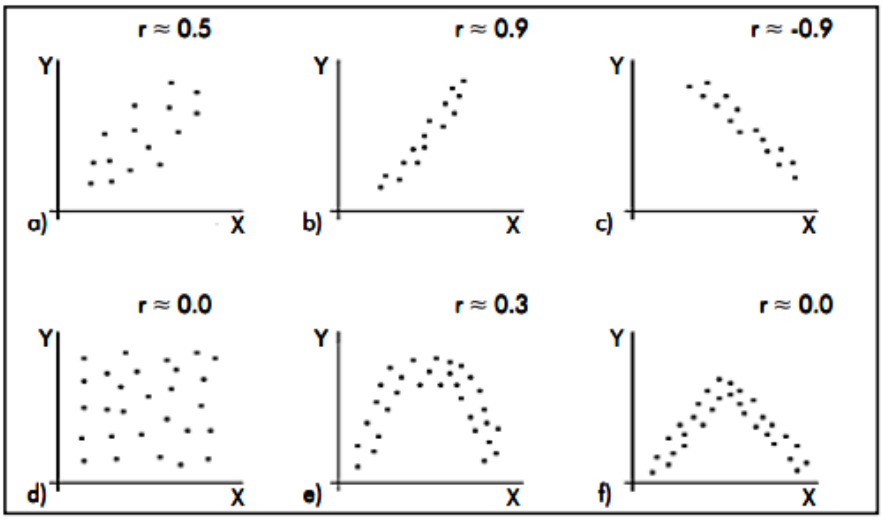


Figura 8. Algunos valores aproximados del coeficiente de correlación de Pearson. Fuente: Behar-Ojeda.

Matriz de varianzas y covarianzas.

Se considera un vector estadístico de dimensión p con componentes X_1, X_2, …, X_p cuantitativas. La matriz de varianzas y covarianzas es simétrica con varianzas en la diagonal.

$$\Sigma = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \dots & \mathrm{cov}(X_1, X_p) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \dots & \mathrm{cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_p, X_1) & \mathrm{cov}(X_p, X_2) & \dots & \mathrm{var}(X_p) \end{pmatrix}$$

A partir de la anterior matriz se obtiene la matriz de correlaciones, conteniendo unos en la diagonal.

Algunos estadísticos de orden.

En forma general, se presentan los conceptos de estadísticos de orden, los cuales son valores que provienen de los datos cuya escala es de razón o intervalo. Son estadísticos de orden, el ***dato mínimo***, ***el máximo***, la mediana, los cuartiles o los deciles.

La mediana es el dato o valor que divide los datos ordenados en dos partes iguales, cada parte con *50 %* de los casos.

Los cuartiles son valores que dividen los datos en cuatro partes; El ***cuartil primero***, a veces denotado por **Q_1**, es el valor (o dato) que por debajo tiene *25 %* de los casos, **Q_2** es la mediana, **Q_3** es el valor o dato abajo del cual se encuentra *75 %* de los casos.

Rango intercuartílico (*RIC*) dado como

$$\mathbf{RIC} = Q_3 - Q_1$$

El **RIC** da una medida de la variación de los datos con respecto al centro: concentrando el *50 %* de los datos. Calcular los estadísticos de orden depende: si el número de datos es par o impar o si los datos están agrupados o no agrupados.

Gráficos estadísticos.

Los gráficos estadísticos son recursos visuales que permiten representar y comunicar la información, se constituye en un instrumento estadístico para analizar los datos en forma eficiente.

En un proceso estadístico, los datos son la materia prima, son los resultados de medir algunos elementos de la población o colectivo de interés, la fuente de procedencia de la información, y la forma en que se obtuvo juega un papel importante, aporta a la validez del proceso estadístico, comprendiendo la calidad de los datos y el tratamiento de los datos con herramientas numéricas y gráficas adecuadas.

El investigador debe ser consiente de los riegos que puede traer al trabajar con datos de dudosa calidad, a pesar de usar las herramientas adecuadas en el análisis de datos, los resultados no serían confiables.

En este capítulo se expondrán algunas representaciones gráficas con las consideraciones técnicas para su construcción y la forma adecuada de aplicarlas. Todo esto, partiendo de la identificación de la naturaleza de la o las variables y la dimensión de los datos.

Elementos de un gráfico estadístico.

Todo gráfico debe tener un cuerpo (figura principal) acompañado de título, código o número de gráfico, leyendas, etiquetas, fuente; En lo posible esta información debe ser clara y simple.

Gráficos para datos unidimensionales.

Existen gráficos para los datos numéricos y para los datos categóricos. Considerando la naturaleza de la variable de estudio (cualitativa o cuantitativa) y los valores obtenidos en la tabla de frecuencia (T.F) se dará paso a clasificar algunas representaciones gráficas.

Algunas representaciones gráficas para datos unidimensionales.	***Se usa***
Barras	T.F
Escalonado	T.F
Pastel	T.F
Diagrama de anillo	T.F
Radar	T.F
Pictograma	T.F
Puntos	-
Histograma y Ojiva	T.F
Polígono de frecuencias	T.F
Dispersión unidimensional	-
Cajas y alambres	-
Tallos y hojas	-
Pareto	T.F
Áreas rectangulares	T.F

*Tabla 1. A*lgunas R. gráficas para datos unidimensionales. Fuente: Elaboración propia.

Muchas de las representaciones gráficas, se derivan directamente de la tabla de frecuencias, otras se construyen de forma más técnica como el gráfico de cajas y alambres.

El lenguaje gráfico tiene un papel esencial en la organización, descripción y análisis de datos, al ser un instrumento de transnumeración y consiste en obtener una nueva información, al cambiar de un sistema de representación a otro. Por ejemplo, al pasar de un listado de datos a un histograma, el alumno puede percibir el valor de la *moda*, que antes no era visible en la muestra bruta.

Otros gráficos importantes que no se obtienen de una tabla de frecuencias, sino que directamente en la construcción involucran los datos. Lo que significa que la visualización presenta cada dato, en un sistema referencial mediante algún símbolo (punto) o directamente usando los valores de la muestra; Por ejemplo, el diagrama de tallo y hojas, o el diagrama de cajas y alambres, que se construye con algunos estadísticos de orden.

El objetivo de estas visualizaciones, al igual que en los casos anteriores, es poder identificar la forma de la distribución o patrones que presentan los datos.

Gráfico de barras.

También llamado diagrama de barras, gráfico de frecuencias o diagrama de frecuencias.

Esta representación gráfica es muy usada por su simplicidad y facilidad de interpretación; Se emplea para comparar gráficamente las frecuencias de categorías o la moda, asociadas a una variable cualitativa que puede ser de escala nominal u ordinal, aunque también podría utilizarse para una variable cuantitativa discreta. Consideraciones:

- El diagrama de barras, muestra la distribución de frecuencias absolutas o relativas; Consiste en un despliegue bidimensional en el eje de la abscisa van los valores que

toma la variable, levanta (ordenada) en cada valor una barra o línea vertical de longitud igual a la frecuencia correspondiente.

- El gráfico de frecuencias absolutas y relativas sólo cambia en la escala del eje de las ordenadas. La frecuencia se puede expresar en escala porcentual.
- Las barras pueden ser horizontales o verticales. Todas las barras deben tener el mismo ancho y no deben superponerse las unas con las otras, la separación entre barras es constante. El ancho es un concepto estético.
- La ventaja de las barras horizontales es que permite considerar una variable con varias modalidades, un ejemplo son los gráficos de población o pirámides de población.

Ejemplo 1.

En un estudio realizado en el Meta y que está relacionado con el turismo, una de las preguntas de la encuesta aplicada tenía el siguiente contenido: ¿cuál es el principal motivo por el que usted visita al departamento del Meta? Señale solo una opción. Variable cualitativa: *Motivo para visitar el departamento del Meta.*

1	4	4	2	5	3	3
2	4	5	1	6	2	3
2	6	1	1	6	3	4
2	6	2	1	6	3	2
3	6	2	1	3	4	1
3	6	2	1	3	6	2
3	6	3	4	3	6	3
5	3	3	4	5	6	3
5	3	3	4	5	5	3
5	3	4	4	5	5	6
1	3	4	4	4	4	
1	3	1	4	4	4	
1	2	1	5	4	3	
1	2	1	5	3	2	
4	2	5	5	3	2	

N.	Modalidad.	$\mathbf{n_i}$	$\mathbf{f_i}$
1	Paisaje	15	0.15
2	Gastronomía	15	0.15
3	Diversión	25	0.25
4	Calor humano	19	0.19
5	Clima	14	0.14
6	Descanso	12	0.12

Tabla 2. Datos de la encuesta y frecuencias. Fuente: Obando Bastidas, J. A. y Castellanos Sánchez.

Si la variable es cuantitativa ***discreta***, se tiende a cambiar las barras por segmentos de línea. Una diferencia importante es que, al ser una variable numérica, se tiene orden y magnitud, la suma de las longitudes de las barras o segmentos es 1 o 100 %.

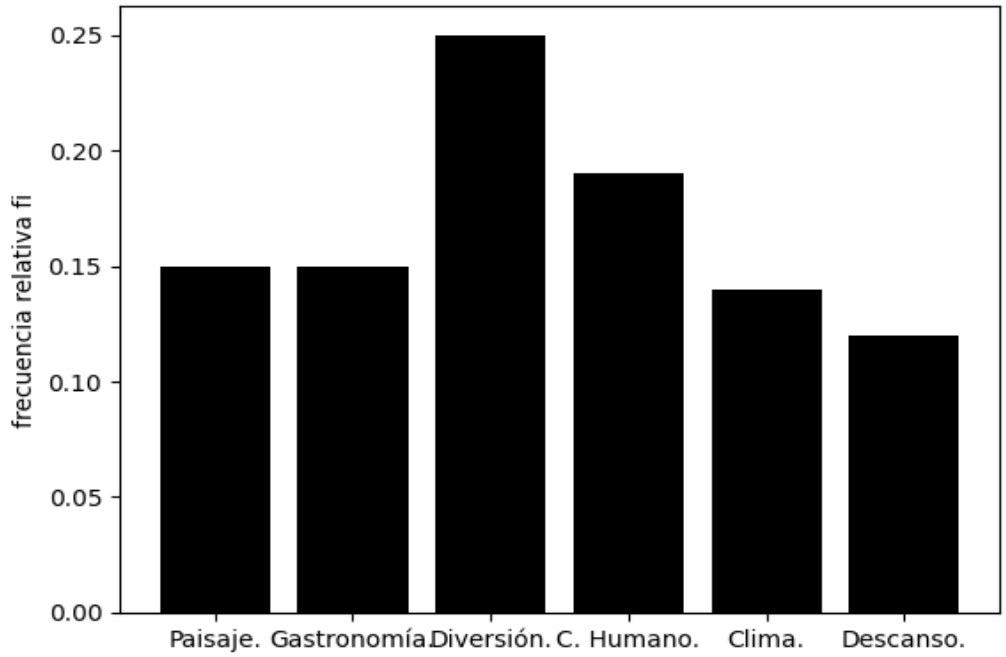


Figura 9. Gráfico de barras para la variable: Motivo para visitar el departamento del Meta. Fuente: Elaboración propia.

Ejemplo 2.

Se toma información sobre la variable que cuenta: *Número de clientes que llegan a un banco en una hora pico, observando una muestra de 25 períodos de un minuto*.

Datos: 8, 6, 7, 9, 8, 7, 8, 10, 4, 10, 8, 7, 9, 8, 7, 6, 5, 10, 7, 8, 5, 6, 8, 10, 11.

Valor observado x_i	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia acumulada absoluta N_i	Frecuencia acumulada relativa F_i
$x_1 = 4$	1	0.04	1	0.04
$x_2 = 5$	2	0.08	3	0.12
$x_3 = 6$	3	0.12	6	0.24
$x_4 = 7$	5	0.20	11	0.44
$x_5 = 8$	7	0.28	18	0.72
$x_6 = 9$	2	0.08	20	0.80
$x_7 = 10$	4	0.16	24	0.96
$x_8 = 11$	1	0.04	25	1
Total	25	1.00		

Tabla 3. Tabla de frecuencias del número de clientes que llegan a un banco en un minuto de la hora pico. Fuente: Behar-Yepes.

De la tabla de frecuencias se derivan: Gráfico de frecuencias (líneas) y el gráfico escalonado.

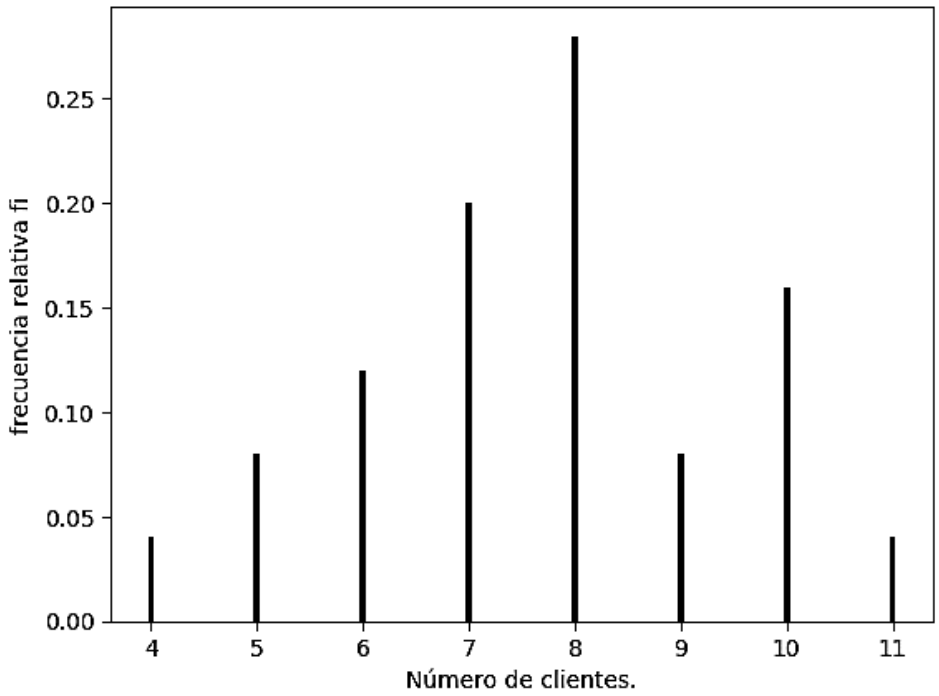


Figura 10. Gráfico de frecuencias relativas del número de clientes que llegan a un banco en un minuto, en la hora pico. Fuente: Elaboración propia.

Gráfico escalonado.

El gráfico escalonado es la gráfica de la función empírica de distribución de frecuencias acumulativas de una variable estadística discreta. La gráfica consiste en llevar a un plano cartesiano las funciones N(x) y F(x). Consideraciones:

- El gráfico de frecuencias acumuladas absolutas difiere del gráfico de frecuencias acumuladas relativas sólo en la escala del eje de las ordenadas, ambas gráficas son válidas.
- Como puede notarse el gráfico corresponde a una función escalonada, lo cual indica que sólo hay datos en los puntos de discontinuidad, cuya frecuencia está representada por el valor del salto correspondiente. Cerrado a la izquierda y abierto a la derecha en cada "escalón". La gráfica es continua a la derecha.
- Estimar el porcentaje de datos que son menores o iguales que cierto valor, este gráfico escalonado permite responder gráficamente a esta situación.

Del ejemplo 2, se define la función empírica de distribución acumulativa.

$$F(x) = \begin{cases} 0 & \text{Si} \quad x < 4 \\ 0{,}04 & \text{Si} \quad 4 \leq x < 5 \\ 0{,}12 & \text{Si} \quad 5 \leq x < 6 \\ 0{,}24 & \text{Si} \quad 6 \leq x < 7 \\ 0{,}44 & \text{Si} \quad 7 \leq x < 8 \\ 0{,}72 & \text{Si} \quad 8 \leq x < 9 \\ 0{,}80 & \text{Si} \quad 9 \leq x < 10 \\ 0{,}96 & \text{Si} \quad 10 \leq x < 11 \\ 1 & \text{Si} \quad x \geq 11 \end{cases}$$

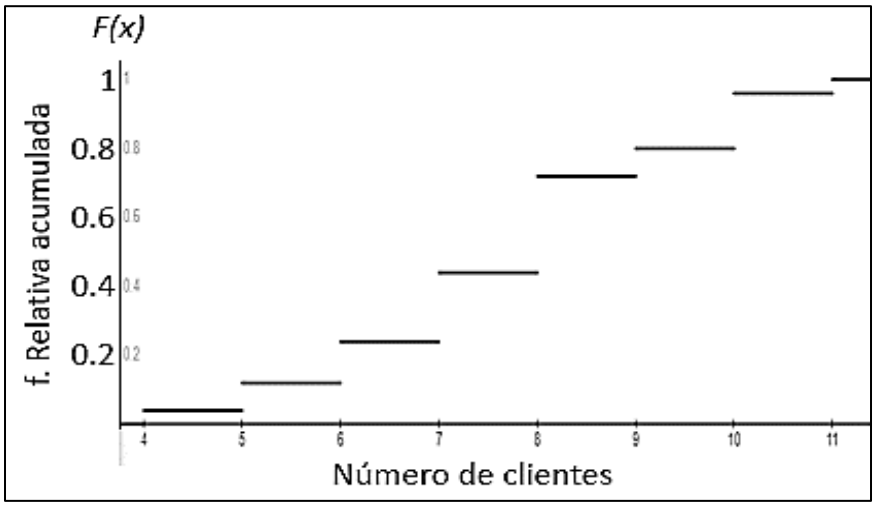


Figura 11. Gráfico escalonado de frecuencias (relativa) acumuladas para la variable número de clientes que llegan a un banco en un minuto, en la hora pico. Fuente: Elaboración propia.

En este ejemplo, se puede deducir que: Hay 20 periodos que a lo sumo tienen 8 personas o para periodos de a lo sumo 8 personas hay un acumulado de 20 periodos o que el 80 % de los periodos tiene a lo sumo 9 clientes.

Diagrama de sectores.

Llamado también gráfico de pastel, gráfico de pizza, gráfico de 360°, diagrama circular, entre otros. Este se le atribuye al economista escoces *William Playfair* al igual que el gráfico de barras y de líneas que los presento hace más de 200 años.

Es una representación gráfica derivado de la tabla de frecuencias, en general para tablas de variables cualitativas; Este gráfico presenta la misma información que un diagrama de barras, permiten ver la distribución interna de los datos en forma de porcentajes sobre un total.

Consiste en fraccionar un círculo, las porciones representan de manera proporcional las frecuencias porcentuales (o relativas) de cada modalidad de la variable. Consideraciones:

- Se realiza un círculo arbitrario.
- Se relaciona cada modalidad de la variable con un ángulo θ del círculo.
- Sea m_i una modalidad con frecuencia relativa f_i, entonces el ángulo θ_i es:

$$\theta_i = f_i * 360°$$

- Cada porción del círculo se etiqueta con el valor porcentual de la modalidad respectiva.
- Se usan distintos tramas o colores para diferenciar las modalidades.
- Los sectores deben ordenarse de mayor a menor, siguiendo el sentido de las agujas del reloj, de izquierda a derecha.
- Los sectores no deben ser muy pequeños, si eso se presenta se deben agrupar en una nueva categoría llamada "Otros" a fin de que la porción sea más visible.
- Es recomendable que los sectores incluyan el porcentaje, siempre que el gráfico no quede muy saturado.
- La leyenda que identifica los sectores se coloca preferentemente a la derecha del gráfico, si el espacio lo permite, se podrá ubicar en la parte inferior.

Del ejemplo 1 se construye los ángulos. Variable: *Motivo para visitar el departamento del Meta.*

Motivo	$\mathbf{n_i}$	$\mathbf{f_i}$	**%**	$\boldsymbol{\theta_i}$
Paisaje	15	0.15	15 %	54°
Gastronomía	15	0.15	15 %	54°
Diversión	25	0.25	25 %	90°
Calor humano	19	0.19	19 %	68.4°
Clima	14	0.14	14 %	50.4°
Descanso	12	0.12	12 %	43.2°
Total	**100**	**1**	**100 %**	**360°**

Tabla 4. Tabla de frecuencias y ángulos para la variable: Motivo para visitar el departamento del Meta. Fuente: Elaboración propia.

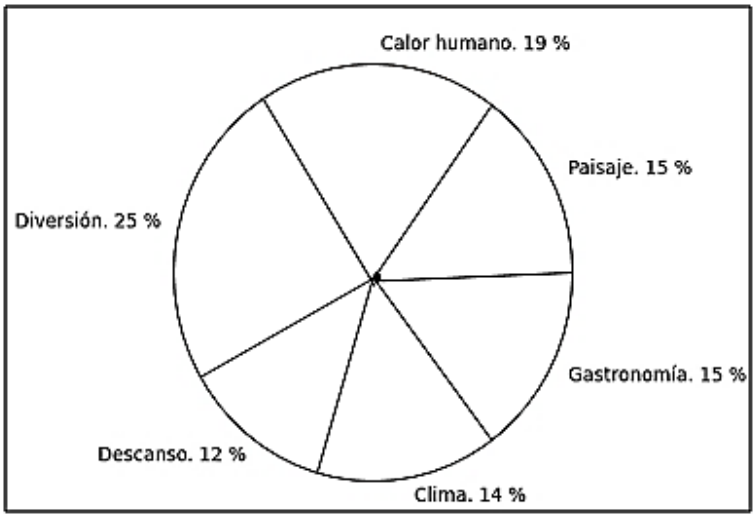


Figura 12. Diagrama de sectores para la variable: Motivo para visitar el departamento del Meta. Fuente: Elaboración propia.

En este ejemplo, se tiene que el 25 % de los turistas visitan el Meta por diversión. Una interpretación más general estará amarrada al objetivo de la investigación que generó la pregunta. Nótese que en este caso solo se dice qué respuesta fue más frecuente y en qué porcentaje difiere de las otras.

Diagrama de anillo (simple).

Otra de las formas de representar las frecuencias de variables cualitativas (si el caso lo amerita también es válido para variables numéricas discretas) lo constituye el diagrama de anillo. Son pequeños sectores en forma de un aro o anillo, divido en forma proporcional de acuerdo con el valor de dicha frecuencia. Consideraciones:

- Los sectores se construyen de la misma forma que se hizo para un diagrama circular.
- Se elimina un círculo concéntrico para darle la forma de anillo.
- Aunque también se pueden representar las frecuencias absolutas o relativas, normalmente en el gráfico de anillo se ponen los porcentajes.
- Se puede añadir una leyenda para indicar qué significa cada color de la gráfica. El color también puede indicar una segunda variable y se debe especificar.
- Cuando hay muchos sectores diferentes en el diagrama, se puede complicar la lectura del gráfico. En tal caso se recomienda agrupar los sectores más pequeños en un único sector llamado "*Otros"*.
- El diagrama de anillo es anillo es básicamente una versión del diagrama circular con algunas ventajas, con uno de anillo se facilita para hacer comparaciones entre grupos de datos. Extendiéndose a datos bidimensionales con diagrama de anillos múltiples.

Para el ejemplo anterior, de variable: *Motivo para visitar el departamento del Meta.* La representación gráfica es:

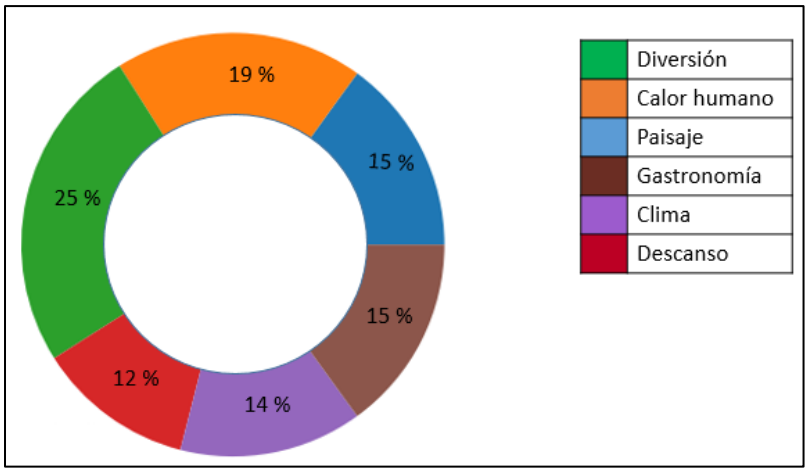


Figura 13. Diagrama de anillo simple para la variable: Motivo para visitar el departamento del Meta. Fuente: Elaboración propia.

Gráfico de radar.

También llamado gráfico de araña, gráfico tela de araña, es una herramienta que se construye a partir de la tabla de frecuencias.

Útil para mostrar visualmente los valores de una frecuencia relativa o absoluta. Proporciona la misma información de un diagrama de barras, circular o anillo. Consideraciones:

- Se hace con base a polígonos regulares donde el número de vértices indica el número de categorías de la variable cualitativa.
- El número de polígonos regulares internos equidistantes, representa la escala de la frecuencia.
- Se ubican las frecuencias de las categorías mediante puntos.
- Se deben unir con segmento de recta, formando un polígono irregular.
- El polígono irregular indica las frecuencias.

Esta representación se puede usar como base y dar pie a un gráfico para representar varias variables. Usualmente lo llaman gráfico radial, gráfico polar o gráfico de radar de varias variables.

Para el ejemplo con la variable: *Motivo para visitar el departamento del Meta.* La representación gráfica es:

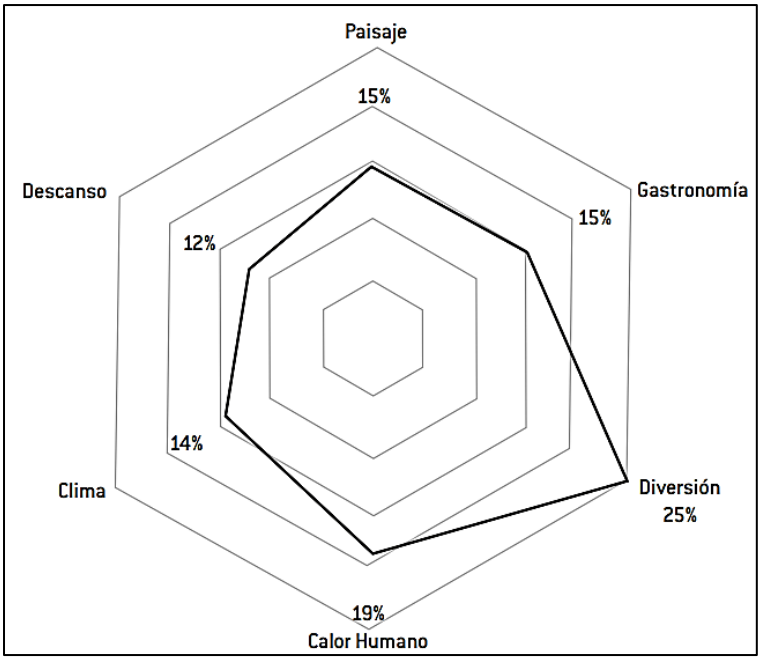


Figura 14. Gráfico de radar para la variable: Motivo para visitar el departamento del Meta. Fuente: Obando Bastidas, J. A. y Castellanos Sánchez-(2021).

Pictograma.

En un pictograma se usan imágenes o símbolos que son alusivas a la variable o al tema de estudio. Para representar una cantidad específica y su tamaño es proporcional a la frecuencia de las modalidades de la variable. Se debe usar el mismo patrón de medida para identificar las frecuencias o indicar las unidades de medida que tiene las figuras.

Este tipo de gráficos suele usarse en los medios de comunicación, para que sean comprendidos por el público no especializado, sin que sea necesaria una explicación

compleja. El diagrama circular o el de barras también son muy usados por los medios de comunicación. Consideraciones:

- Identificar la imagen alusiva al tema.
- El escalamiento de los dibujos debe ser tal que el área de cada uno de ellos sea proporcional a la frecuencia de la modalidad que representa o puede verse como una versión del gráfico de barras.
- Para representar cantidades que tengan valores decimales las imágenes se visualizan "fraccionadas".
- Frecuentemente se usan pictogramas para las frecuencias de variables cualitativas dicotómicas (presentan solo dos modalidades) o variables de tres modalidades, dado que una variable con muchas modalidades no es conveniente utilizar un pictograma, presenta saturación; Para esto, se pueden usar otros diagramas como barras o de sectores.

Ejemplo 3.

La producción de café de un país A es de 650 millones de quintales y de un país B de 250 millones de quintales.

País	***Producción***	***%***
A	650	72.2 %
B	250	27.8 %
Total	900	100 %

Tabla 5. Frecuencia %. Fuente: Gerard Calot.

Usando la siguiente representación pictórica se presta a confusión porque visualmente no se distingue si las frecuencias absolutas son proporcionales a las longitudes, áreas o volúmenes.

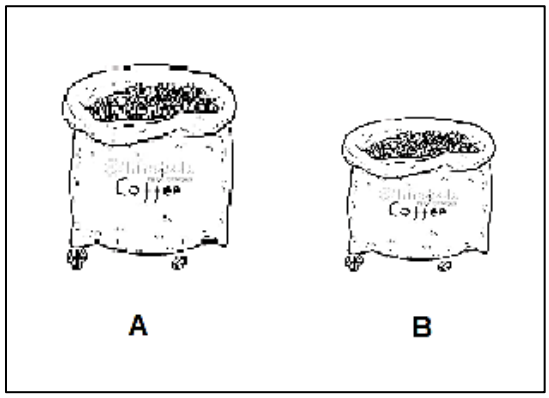


Figura 15. Pictograma (incompleto) de la variable: producción de café de dos países. Fuente: Gerard Calot.

Ahora bien, sin están las áreas en proporción con las frecuencias entones es importante en la figura dar visualmente un indicio del patrón de medida.

Para nuestro ejemplo la relación es de $\frac{5}{13}$, lo que indica que por 5 partes de B son 13 partes de A. luego, el pictograma adecuado es:

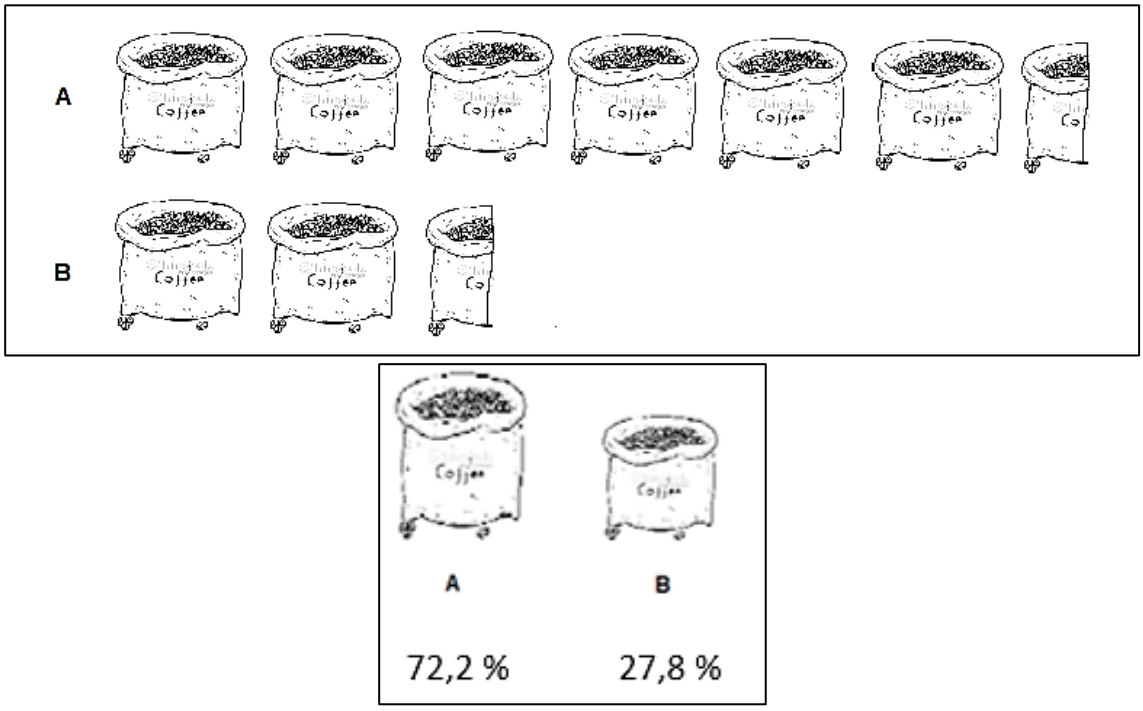


Figura 16. Pictogramas de la variable: producción de café de dos países. Fuente: Gerard Calot.

Diagrama de puntos.

Es un despliegue unidimensional que ubica los datos sobre un eje de amplitud igual al rango de valores, más dos unidades de medida (una a cada lado del eje).

El diagrama de puntos es de gran utilidad y de suma sencillez para tener una primera visión de la distribución de una variable discreta donde se cuenta con pocos datos.

Presenta los rangos de variación de los datos, así como una primera imagen de la forma de la distribución o zonas de mayor concentración. Asimismo, da la posibilidad de identificar de manera preliminar valores atípicos e información sobre la localización, dispersión, extremos y brechas. Su uso se recomienda para cuando el número de datos es pequeño ($n \leq 30$). Cuando hay grandes cantidades de datos, el gráfico de puntos da pie al diagrama de dispersión unidimensional o en caso contrario se recomienda otras gráficas más adecuadas como el gráfico de barras, el histograma o cajas y alambres.

Este diagrama es también útil para hacer comparación de una característica en distintos grupos; Por ejemplo, comparar el ingreso semanal de hombres y mujeres. Consideraciones:

- Ordenar los datos de menor a mayor.
- Identificar el rango.

- Hacer un eje vertical (o horizontal) de longitud: Rango más dos unidades; Una a cada lado del eje.
- Sobre el eje disponer los datos, ubicándolos uno sobre el otro para los valores repetidos.
- El gráfico de puntos puede verse como una versión a uno de barras o como una versión preliminar a un diagrama de dispersión unidimensional.

Ejemplo 4.

En el cuadro siguiente se presentan las mediciones de 24 matrimonios seleccionados en un sector de clase media alta. Con la entrevista se obtuvieron datos sobre las variables discretas.

$\mathbf{X_1}$*: Ingreso semanal del hombre.* $\mathbf{X_2}$*: Ingreso semanal de la mujer.*

Pareja	Ingreso-Hombre	Ingreso-Mujer
1	231	201
2	233	193
3	245	206
4	231	191
5	243	203
6	234	190
7	240	203
8	231	195
9	234	195
10	224	193
11	255	210
12	252	211
13	245	210
14	255	211
15	267	223
16	213	184
17	211	180
18	216	184
19	220	185
20	217	180
21	210	179
22	209	177
23	210	178
24	218	182

Tabla 6. Datos de los matrimonios. Fuente: Behar-Ojeda.

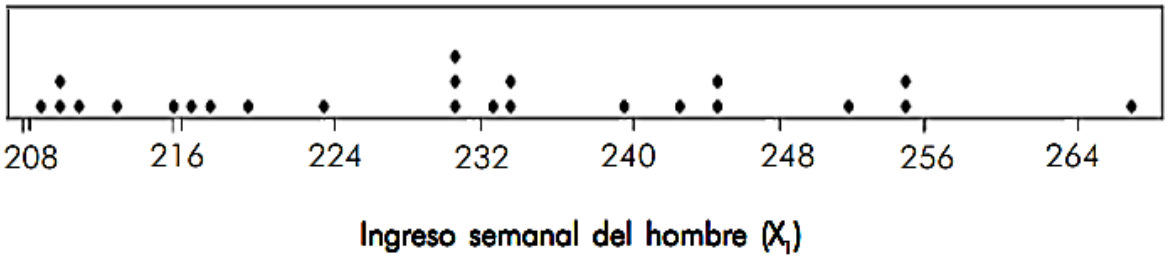


Figura 17. Diagrama de puntos para variable X_1. Fuente: Behar-Ojeda.

Histograma.

Es quizá la representación gráfica para datos de variable continua que más se conoce. En todos los cursos de estadística se enseña a elaborar una tabla de distribución de frecuencias y a partir de ella construir un histograma. Consideraciones:

- Es un despliegue bidimensional (abscisa y ordenada) y consiste en graficar la función empírica de densidad.
- La escala de valores define la escala de la abscisa y se identifican los intervalos de clases. En daca intervalo de clase se levanta un rectángulo:
  Base: Intervalo de clase. ***Altura***: Función empírica de densidad, define la escala de la ordenada.
- Los rectángulos son contiguos compartiendo los límites de clase y el área es igual a la frecuencia relativa del mismo.

Un histograma se debe utilizar cuando los datos son de escala de intervalo o razón; Se recomienda para problemas con grandes cantidades de datos (n>50). El propósito del histograma, es mostrar la distribución de los datos. Un histograma revela la cantidad de variación propia de un proceso.

La decisión central en la elaboración de un histograma está en la definición del tamaño y número de clases que determina una buena o mala representación gráfica de los datos. Para tal decisión hay desde recomendaciones generales hasta fórmulas y con el apoyo de un software, el analista puede realizar varias versiones de un histograma y quedarse con aquél que mejor represente los datos.

Ejemplo 5.

Los siguientes datos corresponden a una variable continua que mide los: *tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital.*

13.1, 7.1, 14.8, 19.0, 10.2, 18.0, 19.8, 15.0, 17.3, 10.8, 22.3, 14.5, 17.1, 14.9, 12.0, 14.0, 18.4, 10.2, 15.8, 16.5, 15.0, 17.6, 4.2, 13.4, 21.2, 14.7, 13.8, 21.0, 14.3, 11.1, 18.9, 8.3, 16.6, 11.2, 20.2, 14.4, 13.5, 18.2, 12.4, 17.0, 26.7, 15.5, 22.0, 12.9, 17.9, 7.4, 18.0, 19.8, 16.0, 21.2

Clase N°i	Intervalo de Clase	x_i	n_i	f_i	N_i	F_i
1	4.15 - 7.15	5.65	2	0.04	2	0.04
2	7.15 - 11.15	9.15	5	0.1	7	0.14
3	11.15 - 13.15	12.15	6	0.12	13	0.26
4	13.15 - 16.15	14.65	15	0.3	28	0.56
5	16.15 - 18.15	17.15	9	0.18	37	0.74
6	18.15 - 21.15	19.65	8	0.16	45	0.9
7	21.15 - 27.15	24.15	5	0.1	50	1
Total			50	1		

Tabla 7. Frecuencias de la variable: Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Se define la función empírica de densidad:

$$f^*(x) = \begin{cases} 0 & \text{Si} \quad x < 4.15 \text{ ó } x > 27.15 \\ \frac{0.04}{3} \equiv 1.33\ \%/\text{min} & \text{Si} \quad 4.15 < x \leq 7.15 \\ \frac{0.10}{4} \equiv 2.50\ \%/\text{min} & \text{Si} \quad 7.15 < x \leq 11.15 \\ \frac{0.12}{2} \equiv 6\ \%/\text{min} & \text{Si} \quad 11.15 < x \leq 13.15 \\ \frac{0.30}{3} \equiv 10\ \%/\text{min} & \text{Si} \quad 13.15 < x \leq 16.15 \\ \frac{0.18}{2} \equiv 9\ \%/\text{min} & \text{Si} \quad 16.15 < x \leq 18.15 \\ 5.33\ \%/\text{min} & \text{Si} \quad 18.15 < x \leq 21.15 \\ 1.66\ \%/\text{min} & \text{Si} \quad 21.15 < x \leq 27.15 \end{cases}$$

La representación gráfica de esta función da pie al histograma:

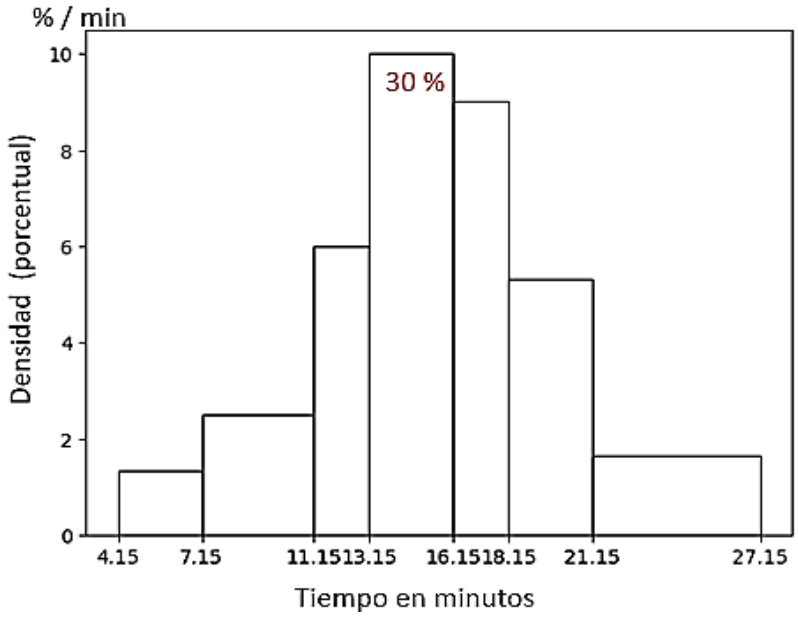


Figura 18. Histograma de la variable: Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Nota: Al definir la función se identifica su equivalente en porcentaje; Permitiendo modificar la escala de la ordenada y facilitando la graficación. Este cambio no altera la forma del histograma y la única consecuencia es que el área total da 100 % y no 1.

Según el histograma, el intervalo (13.15, 16.15] tiene la mayor concentración de los datos (10 % /min) y el área del rectángulo corresponde a la frecuencia porcentual que es del 30 %.

Esta representación gráfica permite:

- Resumir grandes cantidades de datos cuantitativos.
- "Sesgo": El gráfico puede ser asimétrico hacia la izquierda o hacia la derecha si su cola se extiende más hacia uno de esos lados, o ser simétrico si ambas colas son una imagen una de la otra. No espere que cada distribución tenga forma de "campana" (distribución normal). Algunos procesos son sesgados por naturaleza.
- Observar si la distribución es multimodal al presentar varios máximos locales.
- Permite tener una idea de su variabilidad.
- Permite comparar grupos de datos.

Pueden darse falsas modas debido a ejecutar el histograma con la frecuencia relativa o absoluta.

La ojiva.

Esta representación gráfica se deriva de la tabla de frecuencia de una variable continua. Al igual que el caso de la función empírica de densidad, se define la *función empírica de distribución acumulada* y la gráfica de esta función se le llama *Ojiva*, dando lugar a una curva continua y monótona no decreciente. La *Ojiva* muestra información de los datos que se

encuentran por encima o por debajo de un determinado valor en un conjunto de datos. Permite:

- Estimar el porcentaje de datos que son menores o iguales que cierto valor.
- Estimar el porcentaje de datos que hay entre "a" y "b".
- Relación con los percentiles: se identifican “fácilmente” por ejemplo los cuartiles.

Se define la *función empírica de distribución acumulada* F(x):

$$F(x) = \begin{cases} 0 & \text{Si} \quad x \leq 4.15 \\ \frac{0.04}{3}(x - 4.15) & \text{Si} \quad 4.15 < x \leq 7.15 \\ 0.04 + \frac{0.10}{4}(x - 7.15) & \text{Si} \quad 7.15 < x \leq 11.15 \\ 0.14 + \frac{0.12}{2}(x - 11.15) & \text{Si} \quad 11.15 < x \leq 13.15 \\ 0.26 + \frac{0.30}{3}(x - 13.15) & \text{Si} \quad 13.15 < x \leq 16.15 \\ 0.56 + \frac{0.18}{2}(x - 16.15) & \text{Si} \quad 16.15 < x \leq 18.15 \\ 0.74 + \frac{0.16}{3}(x - 18.15) & \text{Si} \quad 18.15 < x \leq 21.15 \\ 0.90 + \frac{0.10}{6}(x - 21.15) & \text{Si} \quad 21.15 < x \leq 27.15 \\ 1 & \text{Si} \quad x > 27.15 \end{cases}$$

De la función F(x) se observa que, en cada intervalo, F(x) representa un segmento de la recta cuya pendiente es la densidad del intervalo respectivo.

Para facilitar la graficación de F(x) se cambia la escala de la ordenada a porcentual, esto no cambia la forma de la ojiva y la consecuencia es que la ojiva se acota superiormente por 100 % y no 1.

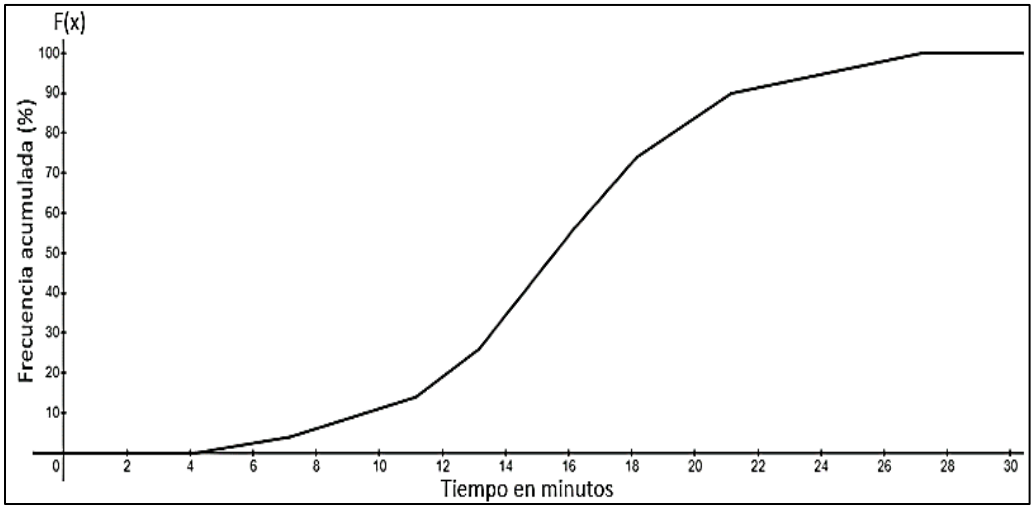


Figura 19. Ojiva de la variable: Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Estimar el porcentaje de datos que son menores o iguales que 15 minutos, se calcula:

$$F(15) = 0.26 + \frac{0.30}{3}(15 - 13.15)$$

$$= 0.26 + 0.185$$

$$= 0.445$$

Por lo tanto, el 44.5 % de los pacientes son atendidos en 15 minutos o menos.

El porcentaje de datos entre 15 min y 20 min es:

$$F(20) - F(15) = 0.8386 - 0.445$$

$$= 0.3936.$$

Esto significa que, aproximadamente el 39.4 % de los pacientes son atendidos en el "filtro" en un tiempo entre 15 y 20 minutos.

Polígono de frecuencias.

Esta representación gráfica se deriva de la tabla de frecuencias de una variable continua, consiste en identificar en el plano bidimensional la marca de clases, sobre el cual se levanta un punto con altura la *densidad* correspondiente; Luego se unen los puntos por segmentos de recta, logrando así una forma geométrica o polígono. Consideraciones:

- Se emplean los polígonos cuando es necesario graficar o resaltar distintas distribuciones conjuntas o bien una clasificación cruzada de una variable cuantitativa continua, junto con otra variable cualitativa o cuantitativa discreta, todo dentro de un mismo gráfico.
- la curvatura del polígono, y su punto más alto es siempre el de mayor frecuencia del conjunto.

- cuenta con la virtud de ser apreciable a simple vista. Por esta razón es sumamente empleado dentro de las ciencias sociales y ciencias económicas, permitiendo así establecer comparaciones útiles entre los distintos resultados de un mismo proceso.
- Para representar el polígono de frecuencias en el primer y último intervalo, suponemos que adyacentes a ellos existen otros intervalos de la misma amplitud y frecuencia nula, y se unen por una línea recta los puntos del histograma que corresponden a sus marcas de clase. Estos intervalos no contienen datos y solo se utiliza para que el gráfico toque el eje x. Obsérvese que, de este modo, el polígono de frecuencias tiene en común con el histograma, es que el área del polígono y el área del histograma coinciden.

Otra forma de obtenerlo es directamente del histograma, usando la marca de clases y uniendo los puntos con segmentos de recta.

Como ***ejemplo***, se usa el histograma anterior para obtener el polígono de frecuencias.

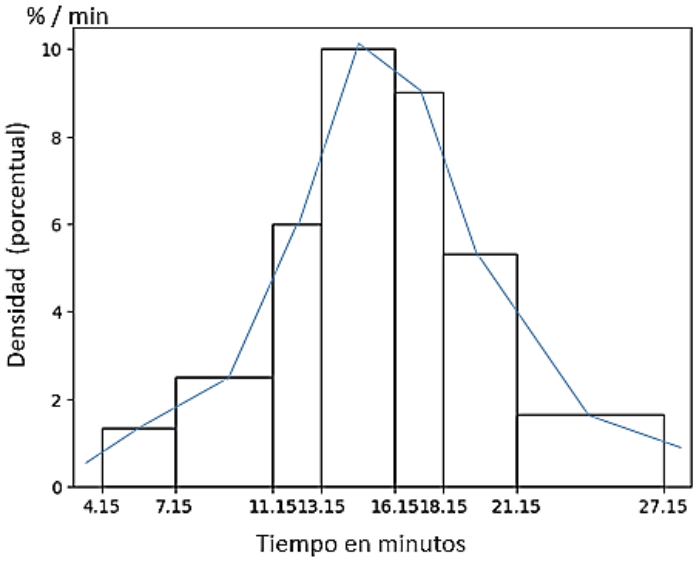


Figura 20. Histograma y polígono de frecuencias de la variable: Tiempos de atención (en min) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Diagrama de dispersión unidimensional.

Es una presentación visual de la muestra bruta en un despliegue unidimensional. Cada dato se representa mediante un punto respecto a su valor. Consideraciones:

- Usado para datos de carácter continuo.
- Permite identificar de cierta manera la distribución de los datos.
- Se identifican posibles datos atípicos.
- Zonas de mayor concentración.

Se toman los datos del ***ejemplo 5*** para hacer el diagrama de dispersión unidimensional.

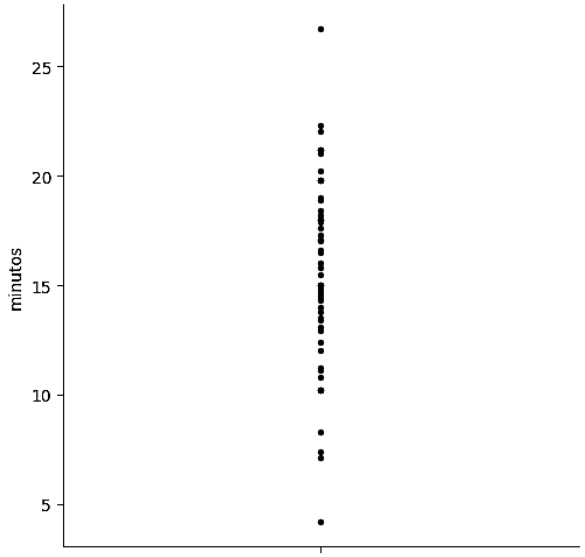


Figura 21. Diagrama de dispersión unidimensional de: Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Muchos puntos seguramente esta superpuestos por otros, al aplicar un poco de “ruido” se mejora el resultado en cuanto a la representación gráfica: identificando con más facilidad el comportamiento de los datos, zonas de mayor y menor concentración o posibles datos atípicos.

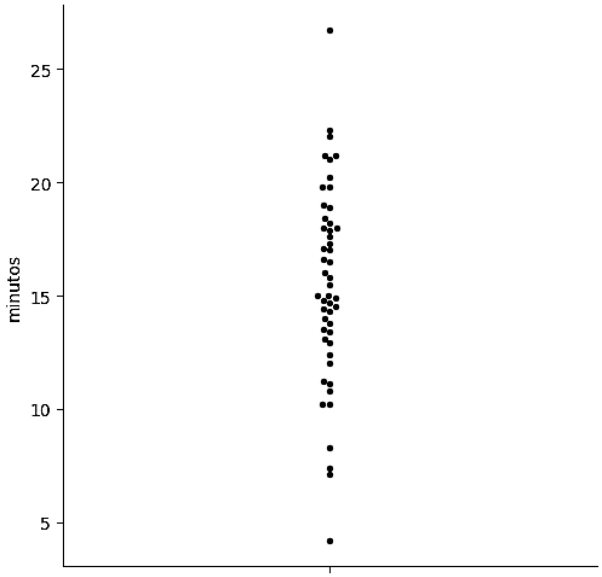


Figura 22. Mejora del diagrama de dispersión unidimensional de: Tiempos de atención (en min) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Diagrama de Cajas y alambres (Box-plot).

Es un tipo de representación gráfica para variables cuantitativas; Llamado originalmente como cajas y bigotes. Constituye una síntesis buena de la distribución de los datos y su sencillez la hace útil; Se construye a partir de estadísticos de orden.

- Dato máximo.	Máx.
- Dato mínimo.	Mín.
- Primer cuartil.	Q_1.
- Segundo cuartil (mediana).	Q_2.
- Tercer cuartil.	Q_3.
- RIC.	Q_3-Q_1
- "cerco interno inferior"	Q_1 - 1.5*RIC.
- "cerco interno superior"	Q_3 + 1.5*RIC.
- "cerco externo inferior"	Q_1 - 3*RIC.
- "cerco externo superior"	Q_3 + 3*RIC.

Tabla 8. Estadísticos para el diagrama de cajas y alambres. Fuente: Elaboración propia.

Propósitos del diagrama de cajas y alambres:

- Para identificar la localización de los datos alrededor de la mediana.
- Para hacerse una muy buena idea de la dispersión de los datos, basándose en la longitud de la caja (rango intercuartílico), siempre, la caja corresponde al 50 % de los datos.
- Además, se aprecia el rango de los datos, el cual corresponde a la distancia entre las observaciones más extremas.
- Permite una buena idea sobre el grado de asimetría de una distribución (se compara la proporción de la caja que queda a la izquierda de la mediana, con la que queda a la derecha al igual que la longitud de los alambres respectivos).
- En el diagrama pueden identificarse dos tipos de datos atípicos: *OUTSIDES* como posibles puntos atípicos que están fuera de los cercos internos, pero dentro de los externos, y *OUTLIERS* como posibles puntos muy atípicos que están por fuera de los cercos externos.
- Si un punto está por fuera de los cercos externos no podrá ser ni el mínimo ni máximo en el diagrama.
- Una utilidad grande de los diagramas de caja y alambres, es comparar en una misma escala, varias poblaciones o grupos, a través de sus distribuciones. En este caso se construye un diagrama para cada distribución y se dibujan en una misma escala (sobre un mismo plano), lo cual permite fácilmente hacerse una idea de las semejanzas y las diferencias de los rasgos más importantes de las distribuciones. En la parte de datos bidimensionales se trabajarán este tipo de representaciones gráficas, donde se pueden involucran varias variables (conjuntamente) mediante algún tipo de codificación.

Consideraciones:

- Se traza un eje referencial, cuya escala está determinado por los valores de los datos.
- Con base al eje referencial se marcan los cuartiles y demás estadísticos.
- la caja se construye entre los cuartiles Q_1 y Q_3, con un ancho arbitrario.
- Dentro de la caja se marca Q_2.
- Los alambres que salen de Q_1 y Q_3, van hasta el dato más próximo al cerco interno sin cruzar el cerco. Así,

$M = máx\{X_i: X_i \leq \text{cerco interno superior}\}$

$m = mín\{X_i: X_i \geq \text{cerco interno inferior}\}$

Se toman los datos del ***ejemplo 5*** para hacer el diagrama de cajas y alambres. Datos ordenados:

4.2, 7.1, 7.4, 8.3, 10.2, 10.2, 10.8, 11.1, 11.2, 12, 12.4, 12.9, 13.1, 13.4, 13.5, 13.8, 14, 14.3, 14.4, 14.5, 14.7, 14.8, 14.9, 15, 15, 15.5, 15.8, 16, 16.5, 16.6, 17, 17.1, 17.3, 17.6, 17,9, 18, 18, 18.2, 18.4, 18.9, 19, 19.8, 19.8, 20.2, 21, 21.2, 21.2, 22, 22.3, 26.7

Dato mínimo	4.2
Dato máximo	26.7
Cuartil 1: Q_1	13.1
Cuartil 2: Q_2	15.25
Cuartil 3: Q_3	18.2
Rango intercuartílico: RIC	5.1
Cerco interno inferior = Q_1-1.5*RIC	5.45
Cerco interno superior = Q_3+1.5*RIC	25.85

Tabla 9. Estadísticos de orden (datos sin agrupar): Tiempos de atención (min) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

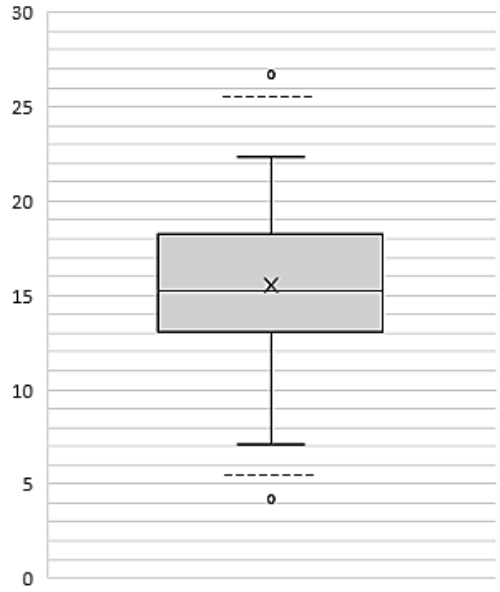


Figura 23. Diagrama de cajas y alambres para los datos sin agrupar: Tiempos de atención (min) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Nota: En el diagrama se observa el valor medio mediante con un valor aproximado de *15.5*.

Dato mínimo	4.2
Dato máximo	26.7
Cuartil 1: Q_1	12.98
Cuartil 2: Q_2	15.55
Cuartil 3: Q_3	18.35
Rango intercuartílico: RIC	5.37
Cerco interno inferior	4.92
Cerco interno superior	26.40

Tabla 10. Estadísticos de los datos agrupados: Tiempo de atención en urgencias. Fuente: Elaboración propia.

Una modificación para el gráfico de cajas para establecer algún comportamiento en cuanto a la centralidad de la distribución es el amuescamiento, que depende de un intervalo para la mediana. Cuando la muestra es lo suficientemente grande se puede realizar una representación gráfica de los intervalos de confianza (IC) al 95 % e indican si existen diferencias significativas entre las medianas (m) de dos variables. Estos IC 95 % se representan mediante muescas calculadas como

$$\mathrm{m} \pm 1.57\frac{\mathrm{RIC}}{\sqrt{\mathrm{n}}}$$

Si las muescas de las cajas de ambas variables no se solapan significará que existen diferencias significativas entre las medianas de ambas variables. Este tipo de interpretación gráfica tan solo se recomienda para muestras grandes en las que las muescas ocupan una posición clara dentro del RIC.

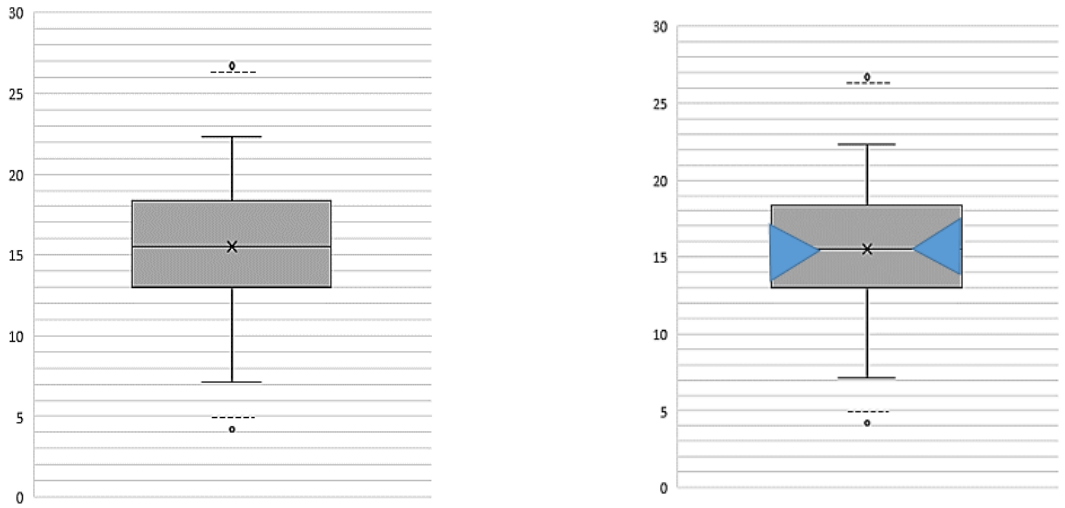


Figura 24. Diagrama de cajas y alambres sin y con amuescamiento para los datos agrupados: Tiempo de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Observaciones (el diagrama de cajas y alambres varía poco respecto al anterior):

- Para este caso de datos agrupados, el “cerco externo” estaría esta entre los puntos -3.13 y 34.46, fuera de este no hay ningún dato.
- Todos los puntos quedaron dentro de los dos cercos (internos), lo cual no ocurre siempre, por esta razón los puntos interiores más cercanos al cerco son el mínimo y

el máximo de los datos y definen la longitud de los alambres que van pegados a la caja.
- En ambos diagramas se observa que los datos están más concentrados entre Q_1 y Q_2 que entre Q_2 y Q_3, lo cual es una muestra de cierto grado de asimetría.

Diagrama de tallos y hojas.

El diagrama de tallos y hojas es una técnica gráfico–numérica que permite organizar y explorar características de la distribución de un conjunto de datos cuantitativos. Este diagrama no sólo cuenta con la totalidad de los datos, sino también la forma de la distribución que se construye con los propios números.

Consideraciones:

- Se recomienda ordenar los datos de mayor a menor, esto permite identificar valores máximo y mínimo, de igual manera permite más claridad en el número de tallos.
- Se seleccionan los tallos, que son clases de valores. Generalmente son los dígitos a la izquierda de las cifras.
- La adecuada elección de los tallos es fundamental para la adecuada representación.
- Se traza una línea vertical y a la izquierda se escriben los tallos.
- Se ordenan y apilan las hojas a la derecha de los tallos conforme se revisan los datos.

Las estadísticas de orden se pueden identificar en el diagrama de tallos y hojas, de igual forma se le puede agregar mayor información, combinándolo con una tabla de distribución de frecuencias o identificando los cuartiles como en el tallo que se encuentran ubicados. Por otro lado, la forma de la distribución es perfectamente percibida; se pueden identificar además otros indicadores de centralidad como el "***tallo modal***", es decir, el que mayor frecuencia presenta. Otro aspecto importante es aumentar el número de tallos, esto se hace al fraccionar cada tallo en dos partes o más, esto hace que la visualización presente una distribución más detallada de los datos.

El diagrama de tallos y hojas tiene limitaciones; Para grandes cantidades de datos ($n>200$) hace difícil el manejo y la disposición de los dígitos, incluso a través de un paquete estadístico. Ante esta circunstancia es recomendable utilizar otra representación gráfica por el histograma o el gráfico de cajas y alambres.

Además:

- Un diagrama tiene un propósito similar al del histograma y se utiliza para evaluar rápidamente las propiedades de distribución de una muestra:
  Identificación de un valor típico o representativo, valores atípicos, grado de dispersión en torno al valor típico, presencia de brechas en los datos, grado de simetría en la distribución, número y localización de modas.
- Permite reconstruir la muestra con los datos originales mediante una *escala* establecida en el diagrama. Con un histograma realmente no es posible saber los datos originales; Pero, con un diagrama de tallo y hojas se obtiene la muestra.

Consideremos los datos del ***ejemplo 5***.

13.1, 7.1, 14.8, 19.0, 10.2, 18.0, 19.8, 15.0, 17.3, 10.8, 22.3, 14.5, 17.1, 14.9, 12.0, 14.0, 18.4, 10.2, 15.8, 16.5, 15.0, 17.6, 4.2, 13.4, 21.2, 14.7, 13.8, 21.0, 14.3, 11.1, 18.9, 8.3, 16.6, 11.2, 20.2, 14.4, 13.5, 18.2, 12.4, 17.0, 26.7, 15.5, 22.0, 12.9, 17.9, 7.4, 18.0, 19.8, 16.0, 21.2

Tallos	Hojas						
4	2						
7	1	4					
8	3						
10	2	2	8				
11	1	2					
12	0	4	9				
13	1	4	5	8			
14	0	3	4	5	7	8	9
15	0	0	5	8			
16	0	5	6				
17	0	1	3	6	9		
18	0	0	2	4	9		
19	0	8	8				
20	2						
21	0	2	2				
22	0	3					
26	7						

Escala: 7|4 → 7.4

Figura 25. Diagrama de tallos y hojas de Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Diagrama de Pareto.

El Diagrama de Pareto, también llamado ***curva 80-20*** es un sencillo método de análisis gráfico que permite clasificar entre las causas más importantes de un problema y las que lo son menos.

Se deriva de la tabla de frecuencias de variables cualitativas y consiste en ordenar intencionalmente las modalidades de mayor a menor frecuencia (es una representación jerárquica). Permite localizar el problema principal o seleccionar la causa más importante, donde potencialmente el éxito puede ser mayor. Es decir, se presenta el ***Principio de Pareto***, conocido como Ley *80–20* o "pocos vitales, muchos triviales". Este principio reconoce que unos pocos elementos (20 %) generan la mayor parte del efecto (80 %), el resto de los elementos generan muy poco del efecto total. Consideraciones:

- Ordene en forma decreciente la frecuencia, y de izquierda a derecha sobre un eje horizontal se ubican las diferentes categorías. De ser necesario se crea la categoría denominada "otros", la cual es colocada al extremo derecho de la clasificación.
- Hay dos ejes verticales para la escala, la absoluta en el eje izquierdo y porcentual en el derecho. En cada categoría se levanta una barra cuya altura es la frecuencia absoluta. El eje vertical derecho contiene la escala en porcentaje para la representación de una ***curva acumulativa*** de categorías en relación al orden de las modalidades (no es función).

- De la barra más alta, y moviéndose de izquierda a derecha a través de las categorías se trazan segmentos de línea para la acumulación de categorías. Esto pueden contestar preguntas como: ¿Cuánto del total está representado por las tres primeras categorías?
- Utilice el sentido común; Los eventos más frecuentes no son siempre los más importantes, por ejemplo: dos accidentes fatales requieren más atención que cien lesiones menores.
- Cuando los datos son cuantitativos, las modalidades se deben categorizar, es decir hacer una trasformación de la variable cuantitativa a una variable cualitativa, esto permite ordenar y por ende usar la visualización de Pareto. Por ejemplo, para una variable continua, los intervalos de clases que tiene la forma:

  (c, b]

  Se categoriza como “c hasta b”.

Considere el ***ejemplo*** 5: *Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital.*

La variable inicial es continua, y categorizando se obteniendo una nueva variable cualitativa de *Categorías: Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital.*

T. categoria	*f. absoluta*	*acumulación*
(13.15,16.15]	15	30%
(16.15,18.15]	9	48%
(18.15,21.15]	8	64%
(11.15,13.15]	6	76%
(7.15,11.15]	5	86%
(21.15,27.15]	5	96%
(4.15,7.15]	2	100%

Tabla 11. Frecuencia de la variable Categorías: Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

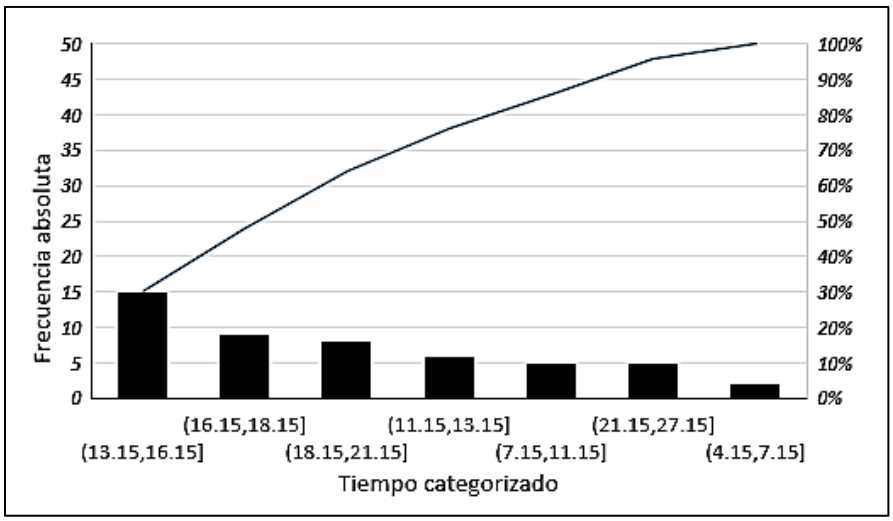


Figura 26. Diagrama de Pareto para la variable categórica: Tiempos de atención (en min) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Diagrama de áreas rectangulares.

Es una representación gráfica jerárquica, al igual que el diagrama de Pareto, prioriza las modalidades (de una variable cualitativa) que más impacto generan en el fenómeno; Es decir, ordena las frecuencias de mayor a menor frecuencia.

Los gráficos de rectángulos son una buena opción para comparar proporciones dentro de la jerarquía. Además, es óptimo para comparar (no más de dos) conjuntos de datos conservando un orden jerárquico. Para varios grupos, se recomienda otros gráficos, por ejemplo, el de proyección solar que es básicamente una versión extendida del diagrama de áreas rectangulares. Consideraciones:

- Se construye a partir de una tabla de frecuencias ordenadas jerárquicamente y geométricamente en dividir un rectángulo en rectángulos más pequeños, donde cada área de cada rectángulo es proporcional a la frecuencia respectiva, generalmente se trabaja con la frecuencia porcentual.
- Es posible caracterizarlos con tramas o colores para identificar las diferencias entre categorías; Pero, también es posible que indique una segunda variable y se debe especificar.

Considere como ***ejemplo*** la anterior *variable categórica*: *Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital.*

T. categoria	f. absoluta	%
(13.15,16.15]	15	30%
(16.15,18.15]	9	18%
(18.15,21.15]	8	16%
(11.15,13.15]	6	12%
(7.15,11.15]	5	10%
(21.15,27.15]	5	10%
(4.15,7.15]	2	4%

Tabla 12. Frecuencias de la variable categórica: Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

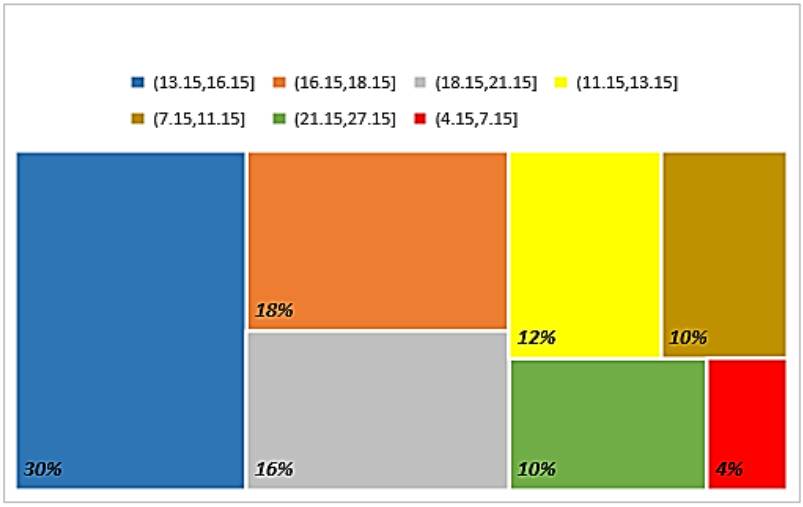


Figura 27. Diagrama de áreas rectangulares de la variable categórica: Tiempos de atención (en minutos) de pacientes en el "filtro" del servicio de urgencias de un hospital. Fuente: Elaboración propia.

Gráficos para datos bidimensionales.

Para este tipo de datos, algunas representaciones gráficas se construyen a partir de la tabla de doble entrada como una extensión al caso unidimensional, por ejemplo, el estereograma puede considerarse como una extensión del histograma; Otras visualizaciones presentan directamente los datos en un sistema cartesiano, por ejemplo, el diagrama de dispersión bidimensional o el diagrama de tallo y hojas; También es común el uso de gráficos comparativo. Por ejemplo, la comparación de histogramas, cajas y alambres, barras, barras apiladas, etc. Tantos como modalidades se tenga y según la naturaleza de la variable las representaciones cambian. La representación gráfica pone en evidencia:

- La distribución de los datos de forma global, es decir, el conjunto delas frecuencias absolutas de la tabla de doble entrada.
- Cada una de las distribuciones condicionadas según un carácter en función de las modalidades del otro.
- Frecuencias absolutas marginales.

Para los datos bidimensionales se tienen los seis posibles casos:

Caso 1.	Discreto – Discreto
Caso 2.	Continuo – Continuo
Caso 3.	Continuo – Discreto
Caso 4.	Discreto – Cualitativo
Caso 5.	Continuo – Cualitativo.
Caso 6.	Cualitativo – Cualitativo

Tabla 13. Casos para dos variables estadísticas. Fuente: Elaboración propia.

Caso 1. Discreto-discreto.

Gráfico de frecuencia conjunta.

A partir de las tablas de doble entrada (con ambas variables discretas) y se hace la representación gráfica de la distribución conjunta de frecuencias absolutas o relativas, de las variables X y Y. Consiste en identificar las modalidades de las dos variables en el plano cartesiano y sobre el cual se levanta un segmento de recta de longitud igual a la frecuencia conjunta. Básicamente es una versión extendida del diagrama de frecuencias de datos discretos unidimensionales.

Ejemplo 6.

De cierta población en estudio se sacó una muestra de 50 familias con el propósito de observar las variables:

X: Número de personas que componen la familia.

Y: Número de personas que producen algún ingreso.

Los datos de (X,Y):

(6,1), (1,1), (3,1), (4,2), (6,1), (1,1), (3,1), (4,2), (5,2), (5,1), (5,4), (6,1), (2,1), (3,2), (4,3), (6,2), (2,1), (3,2), (4,2), (3,2), (4,2), (4,3), (3,3), (4,3), (4,4), (4,4), (4,4), (4,2), (2,1), (6,2), (6,3), (4,4), (2,1), (5,1), (5,5), (4,4), (3,2), (2,2), (6,4), (6,5), (6,4), (6,2), (6,3), (6,2), (6,2), (5,2), (5,4), (5,1), (5,4), (5,4).

Y / X	1	2	3	4	5	
1	0.04	0.00	0.00	0.00	0.00	0.04
2	0.08	0.02	0.00	0.00	0.00	0.10
3	0.04	0.08*	0.02	0.00	0.00	0.14*
4	0.00	0.10	0.06	0.10	0.00	0.26
5	0.06	0.04	0.00	0.08	0.02	0.20
6	0.06	0.10	0.04	0.04	0.02	0.26
	0.28	0.34*	0.12	0.22	0.04	1.00

Tabla 14. Distribución conjunta de frecuencia relativa de las variables X y Y. Fuente: Behar-Yepes.

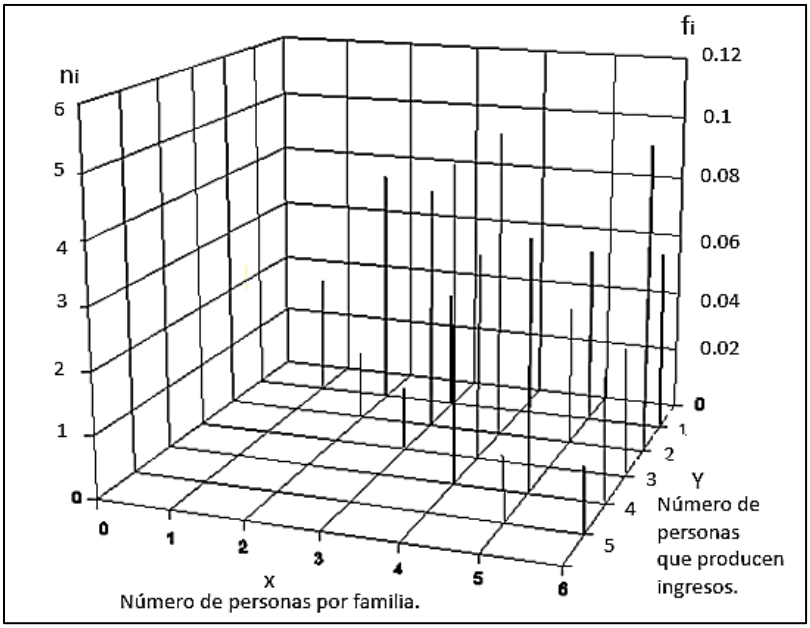


Figura 28. Gráfico de frecuencia conjunta de las variables X y Y. Fuente: Behar-Yepes.

Observaciones:

- El gráfico de frecuencias absolutas difiere del gráfico de frecuencias relativas sólo en la escala del eje de las ordenadas. Por comodidad se trabajar con porcentajes, facilitando identificar datos de más frecuencia.
- La suma de las longitudes de recta da la totalidad.
- Para las condicionales o marginales de la tabla, la gráfica un caso unidimensional.

De la tabla acumulativa conjunta, es posible definir y construir la gráfica de la función empírica de distribución acumulativa conjunta; Pero, su visualización escalonada puede ser confusa.

X \ Y	1	2	3	4	5
1	0.04	0.04	0.04	0.04	0.04
2	0.12	0.14	0.14	0.14	0.14
3	0.16	0.26	0.28	0.28	0.28
4	0.16	0.36	0.44	0.54	0.54
5	0.22	0.46	0.54*	0.72	0.74
6	0.28	0.62	0.74	0.96	1.00

Tabla 15. Distribución conjunta. Fuente: Elaboración propia.

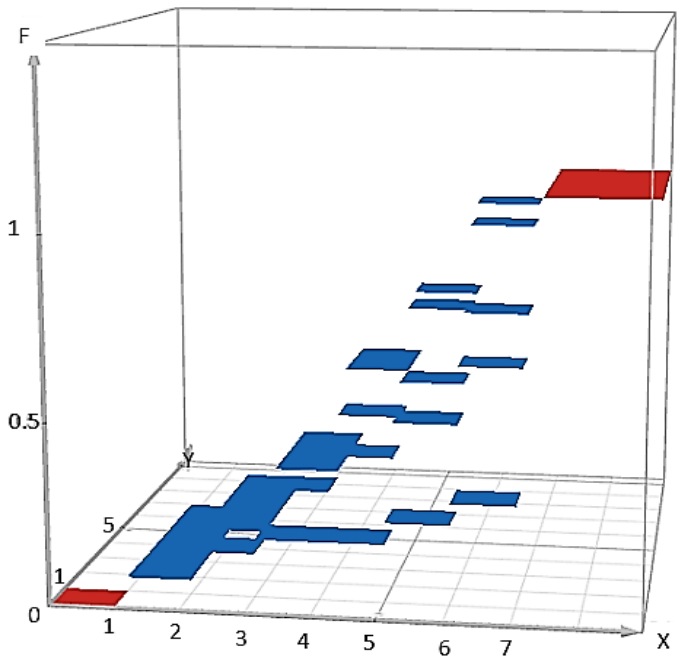


Figura 29. Distribución conjunta acumulativa de las variables X y Y. Fuente: Elaboración propia.

Diagrama de dispersión bidimensional.

Es una forma de representar visualmente la muestra bruta o la matriz de datos cuantitativos en el sistema cartesiano. También llamado nube de dispersión, gráfico de dispersión.

Consiste en representar cada dato mediante un punto geométrico en el plano cartesiano. A cada eje se le asigna una variable, la escala de los ejes está dado por los valores de las variables respectivas.

El diagrama de dispersión permite:

- Presentar cada dato mediante un punto geométrico en el plano cartesiano.
- Identificar de cierta manera la distribución o patrón de los datos.
- Se identifican posibles datos atípicos.
- Zonas de mayor y menor concentración.
- Identificar, clasificar y comparar grupos de datos.
- Dar una idea de la relación o grado de relación que guardan las variables, que posteriormente el investigador verificara con el coeficiente de correlación de Pearson.
- Representar información codificada de otras variables (cuantitativas o cualitativas) mediante formas, tamaños o colores. Además, se puede extender al espacio tridimensional (diagrama de dispersión tridimensional).

Considere la muestra bruta del ***ejemplo*** anterior para el diagrama de dispersión bidimensional.

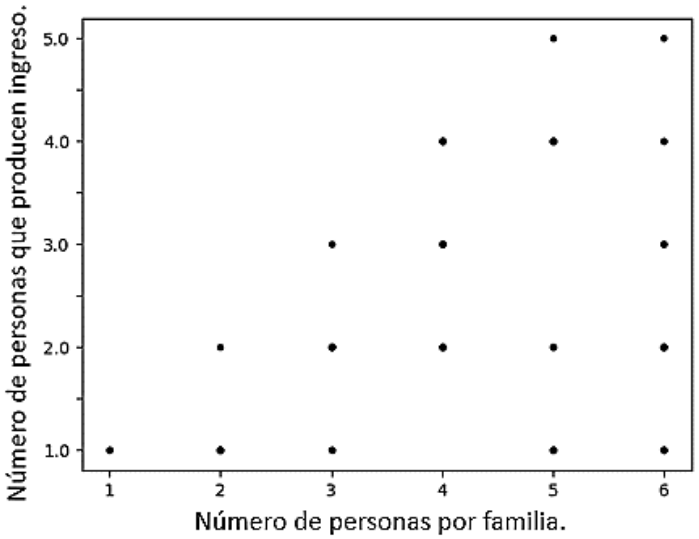


Figura 30. Diagrama de dispersión de las variables X y Y. Fuente: Elaboración propia.

Este diagrama de dispersión un poco simple visualmente, ya que no se aprecia que puntos (o zonas) presentan mayor frecuencia, para esto se hace un ajuste a cada punto, cambiando su tamaño para indicar la frecuencia absoluta. Se aclara que considerar el tamaño de los puntos no representa una tercera variable, caso contrario para datos en altas dimensiones.

El objetivo de esta nueva visualización es representar de forma más clara la distribución global. El *gráfico de frecuencias conjunta*, está en el espacio tridimensional y la nueva representación está en el espacio bidimensional, haciendo más clara la representación.

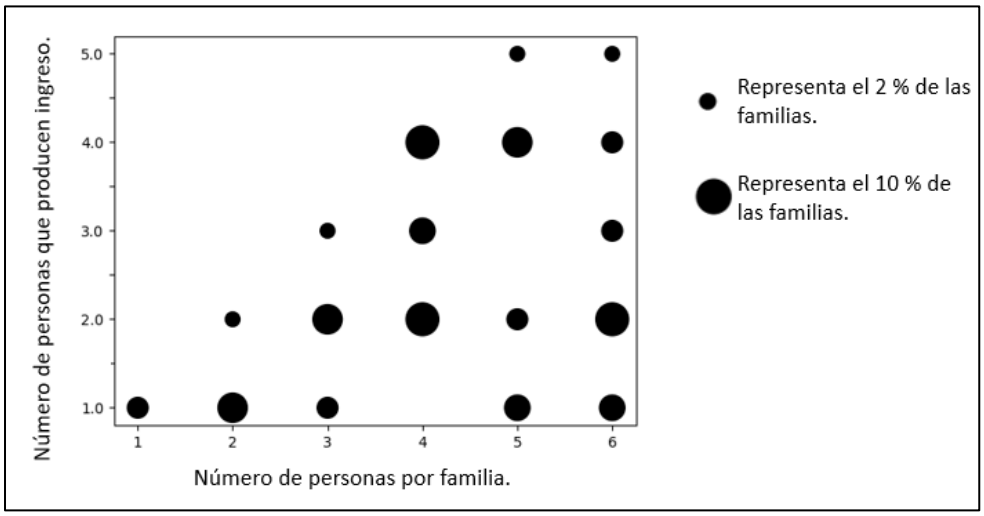


Figura 31. Diagrama de dispersión con la distribución porcentual de las variables X y Y. Fuente: Elaboración propia.

Gráfico de distribuciones condicionadas.

En esta representación se estandarizan las condicionales (o por filas o por columnas) y se representan en el plano:

- La abscisa: Se ubican las condiciones.
- La ordenada: Se ubican los porcentajes acumulados.

La visualización puede incluir barras o puntos para identificar los distintos porcentajes acumulados en cada condición; También, se pueden agregar líneas punteadas (o segmentos de recta cuando son variables continuas) para unir los puntos., permitiendo ver patrones o tendencias en los datos.

	1	2	3	4	5	Marginal X
1	14.3	0.0	0.0	0.0	0.0	**4.0**
2	28.6	5.9	0.0	0.0	0.0	**10.0**
3	14.3	23.5	16.7	0.0	0.0	**14.0**
4	0.0	29.4	50.0	45.5	0.0	**26.0**
5	21.4	11.8	0.0	36.4	50.0	**20.0**
6	21.4	29.4	33.3	18.2	50.0	**26.0**
Total %	**100**	**100**	**100**	**100**	**100**	**100**

Tabla 16. Distribución condicional de las familias según el número de personas que la conforman en función del número de personas que generan ingresos. Fuente: Elaboración propia.

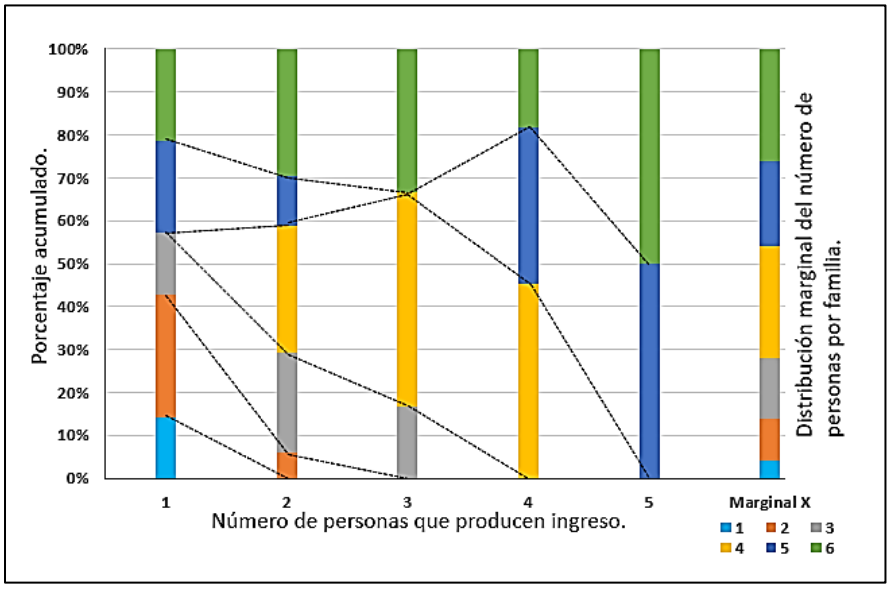


Figura 32. Barras apiladas. Fuente: Elaboración propia.

De forma análoga se hace la representación gráfica para las: *Distribuciones condicionas de las familias según el número de personas que generan ingresos en función del número de personas que la conforman*.

Caso 2. Continuo-continuo.

Gráfico de densidad conjunta (Estereograma).

Partiendo de las tablas de doble entrada (con ambas variables continuas) se construye y se hace la representación gráfica de la función empírica de densidad conjunta, dando pie a un

conjunto de paralelepípedos rectangulares, el cual se denomina como gráfico de densidad conjunta.

Con base al plano cartesiano el proceso consiste en identificar los intervalos de clases de cada variable en los ejes del plano, lo cual generan regiones rectangulares en el plano y sobre estas se levantan paralelepípedos rectangulares de altura igual a la densidad conjunta respectivamente. Básicamente es una versión extendida del histograma, con la diferencia que en lugar de hablarse de área se habla de volumen. Lo cual significa que el volumen de un prisma representa la frecuencia relativa (o porcentaje de datos) que pertenecen a la región definida por la base del mismo por tal razón al calcular el volumen total del gráfico debe arrojar como resultado 1 ó 100 %.

Esta representación gráfica permite:

- Estimar el porcentaje de datos que pertenecen a cualquier región del plano, tan sólo calculando el volumen que se levanta sobre la mencionada región.
- Identificar de cierta manera la distribución o patrón de los datos.
- Resumir grandes cantidades de datos cuantitativos.
- Identificar si la distribución de los datos se comporta como una normal bivariada.

Ejemplo 7.

En un estudio realizado en la región del *Omait* en el cual la población de interés estaba constituida por las fincas que cultivan maíz, se tomó al azar una muestra de 200 fincas de las cuales se registra las variables:

X: Área cultivada en hectáreas (Ha).

Y: Producción anual de maíz en toneladas (Ton).

Con base en los 200 datos, se construyó los intervalos de clase de X y Y, generando la siguiente tabla doble entrada.

Y / X	(0 ; 25] Y_1	(25 ; 60] Y_2	(60 ; 180] Y_3	(180 ; 250] Y_4	(250 ; 350] Y_5	
(0 ; 10] X_1	0.170	0.150	0.070	0.010	0	0.40
(10 ; 40] X_2	0.115	0.060	0.100*	0.020	0.005	0.30*
(40 ; 90] X_3	0.065	0.040	0.120	0.020	0.005	0.25
(90 ; 150] X_4	0	0	0.010	0.025	0.015	0.05
	0.35	0.25	0.30	0.075	0.025*	1.00

Tabla 17. Distribución conjunta de frecuencia relativa de las variables X y Y. Fuente: Behar-Yepes.

El objetivo es construir la *función empírica de densidad conjunta* de X y Y.

La tabla de densidad conjunta: frecuencia relativa conjunta entre el área de la región (A_{ij}) definida por los intervalos de clase en X y Y:

Y / X	(0 ; 25] Y_1	(25 ; 60] Y_2	(60 ; 180] Y_3	(180 ; 250] Y_4	(250 ; 350] Y_5
(0 ; 10] X_1	0.00068000	0.00042857	0.00005833	0.00001428	0
(10 ; 40] X_2	0.00015333	0.00005714	0.00002777	0.00000952	0.00000166
(40 ; 90] X_3	0.00005200	0.00002285	0.00002000	0.00000571	0.00000100
(90 ; 150] X_4	0	0	0.00000138	0.00000595	0.00000250

Tabla 18. Densidad empírica conjunta para las variables X y Y. Fuente: Behar-Yepes.

Se define la función empírica de densidad y su gráfica:

$$f^*(x,y) = \begin{cases} f_{ij}^*; & \text{Si } (x,y) \in (X_i \cap Y_i), \quad i = 1,2,3,4; j = 1,2,3,4,5. \\ 0; & \text{Otro caso.} \end{cases}$$

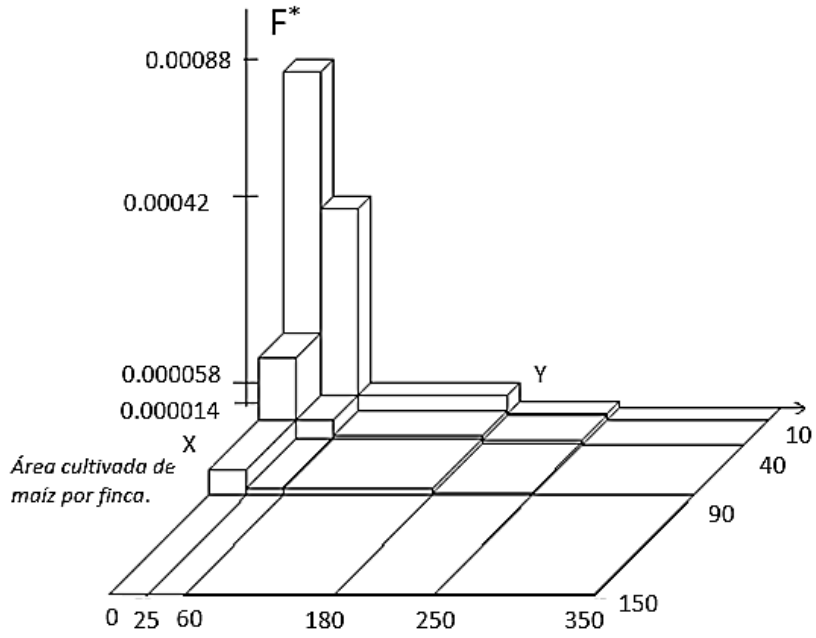


Figura 33. Gráfico de densidad conjunta de las variables “área cultivada" y "producción anual de maíz”. Fuente: Behar-Yepes.

El porcentaje de fincas que tienen áreas de cultivo de maíz entre 30 Ha y 60 Ha y producen anualmente entre 100 Ton. y 300 Ton, se reduce a calcular el volumen correspondiente, el cual es aproximadamente 0.08 (8 %).

El mayor defecto de la representación por estereograma es la complejidad de su realización práctica. Además, imposibilita representar algunos paralelepípedos al quedar ocultos por otros situados más adelante. Entonces se opta por el diagrama de dispersión cuya visualización brindar información de la distribución y posiblemente evidenciando asociación.

Partiendo de la muestra bruta, el diagrama de dispersión se obtiene directamente; Al no contar con los datos, se hace una ***aproximación***: En cada rectángulo de las clases i en X, j en Y se ubican n_{ij} puntos uniformemente distribuidos.

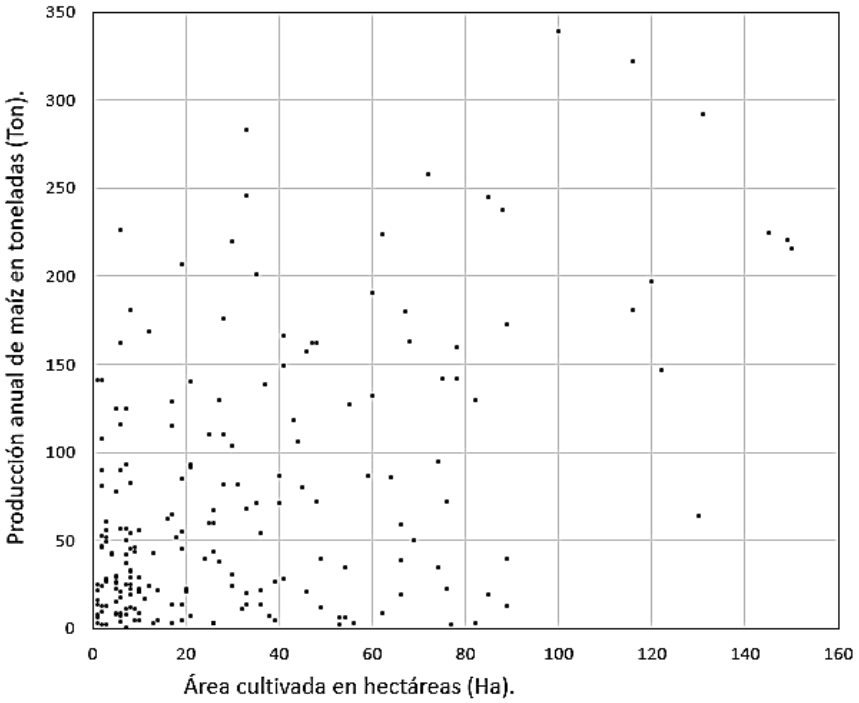


Figura 34. Distribución de las fincas que cultivan maíz según área cultivada y producción anual. Fuente: Elaboración propia.

Observaciones.

La gráfica de esta función F(x,y) es la ampliación del concepto de Ojiva (se omite) y para la representación gráfica de las *distribuciones condicionadas* se hace análogamente como en caso discreto-discreto. Se considera las marcas de clases, los porcentajes acumulados y segmentos de recta.

Caso 3. Discreto-continuo.

Gráfico laminar.

Se aclara que en este caso solo la variable continua cuenta con densidad y para evitar nuevos conceptos de densidad, se seguirá hablando del termino de función empírica de densidad conjunta.

Partiendo de las tablas de doble entrada (discreto-continuo) se construye y se hace la representación gráfica de la función empírica de densidad conjunta, dando pie a un conjunto

de “láminas” formando un histograma por cada como modalidad de la variable discreta; El resultado es un gráfico de densidad laminar.

Con base al plano cartesiano, el proceso consiste en identificar los intervalos de clases de la variable continua y las modalidades de la variable discreta en los ejes respectivos, lo cual generan segmentos de recta en el plano y sobre estos se levantan rectángulos (o láminas) de altura igual a la densidad respectiva. Esto significa que el área de una lámina representa la frecuencia relativa (o porcentaje) para el segmento de recta, por tal razón al calcular el área total del gráfico debe arrojar como resultado *1 ó 100 %*.

Esta representación gráfica permite:

- Estimar el porcentaje de datos que pertenecen a cualquier región del plano, tan sólo calculando el área de las láminas que se levanta sobre la mencionada región.
- Identificar de cierta manera la distribución o patrón de los datos.
- Resumir grandes cantidades de datos cuantitativos.

Ejemplo 8.

Se tomó una muestra de *500* hogares en los cuales se observó las características:

X: número de personas que constituyen el hogar. (discreta)

Y: ingreso del hogar (en miles de pesos). (continua)

El objetivo es construir la *función empírica de densidad “conjunta”* de X y Y.

Y / X	(50 ; 75] Y_1	(75 ; 125] Y_2	(125 ; 200] Y_3	(200 ; 300] Y_4	(300 ; 550] Y_5	
X_1=1	0.072	0.030	0.024	0.018	0.006	0.15
X_2=2	0.076	0.040	0.046	0.028	0.010	0.20
X_3=3	0.172	0.120	0.050	0.044	0.014	0.40
X_4=5	0.030	0.060	0.080	0.060	0.020	0.25
	0.35	0.25	0.20	0.15	0.05	1.00

Tabla 19. Distribución conjunta de frecuencia relativa de las variables X y Y. Fuente: Behar-Yepes.

Como Y es continua, la convenida función empírica de densidad conjunta f_{ij}^* es una densidad por unidad lineal:

$$f_{ij}^* = \frac{f_{ij}}{C_j}$$

Y / X	(50 ; 75] Y_1	(75 ; 125] Y_2	(125 ; 200] Y_3	(200 ; 300] Y_4	(300 ; 550] Y_5
X_1=1	0.00288000	0.00060000	0.00032000	0.00018000	0.00002400
X_2=2	0.00304000	0.00080000	0.00061333	0.00028000	0.00004000
X_3=3	0.00688000	0.00240000	0.00066666	0.00044000	0.00005600
X_4=5	0.00120000	0.00120000	0.00106666	0.00060000	0.00008000

Tabla 20. Densidad empírica conjunta para las variables X y Y. Fuente: Behar-Yepes.

Se define la función empírica de densidad "conjunta" y su gráfica:

$$f^*(x,y) = \begin{cases} f_{ij}^* & \text{Si } (x,y) \in (X_i \cap Y_i),\ i = 1,2,3,4; j = 1,2,3,4,5. \\ 0 & \text{Otro caso.} \end{cases}$$

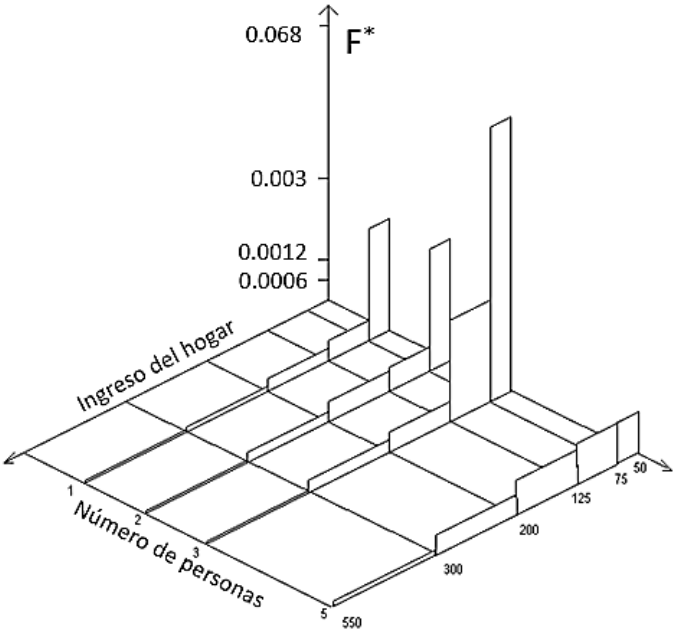


Figura 35. Gráfico de láminas de las variables X y Y. Fuente: Behar-Yepes.

Otra versión de este gráfico sin perder la escala: consiste en "acostar" cada histograma sobre el plano, etiquetando el % (*Gerard Calot (1988) pág. 243)*. También, se puede resumir aún más la representación con un diagrama de dispersión. Por cada modalidad de la variable discreta se tiene puntos colineales o con ruido.

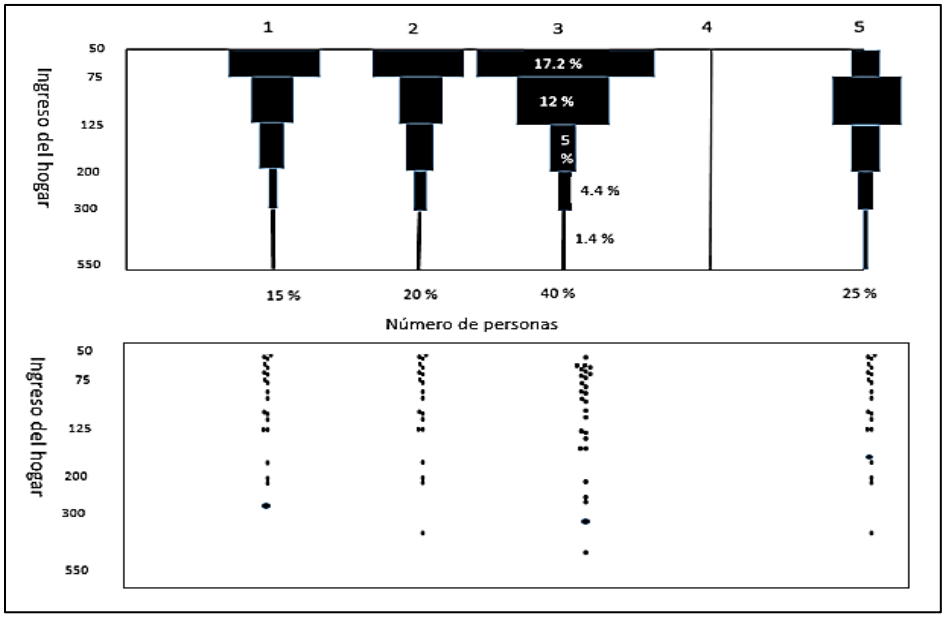


Figura 36. Versión comparativa: Representación gráfica de continua vs discreto. Fuente: Elaboración propia.

Nota. Se omite la gráfica de F(x,y).

Caso 4. Discreto-cualitativo.

Al replicar los procesos anteriores para visualizar datos bidimensionales en el cual una variable cuantitativa y otra cualitativa, se presenta un problema con eje que corresponde a la variable cualitativa ya que no es numérica. Por tanto, se debe recurrir a otro tipo de técnica como las ***comparaciones gráficas***.

La técnica consiste en representar gráficamente varios conjuntos de datos usando el mismo tipo de gráfico con la *misma escala* y cada conjunto se identifican por las distintas categorías de la variable cualitativa.

Quien determina el gráfico base es la variable cuantitativa, esto significa que para el caso *discreto-cualitativo* el gráfico a comparar es el diagrama de puntos, gráfico de frecuencias, escalonado, el diagrama de cajas y alambres ya que también se emplea para variables discretas.

Se debe tener cuidado con el número de categorías toma la variable cualitativa y el gráfico base que se va a utilizar, la visualización se puede saturar si no se elige el gráfico correcto. Lo idóneo es usar comparaciones donde la visualización no este saturado o muy cargado.

Este tipo de representación permite:

- Resumir y comparar grupos de datos de una misma característica en una misma escala.
- Compara las distribuciones de cada grupo e identificar diferencias o patrones que pueden tener.
- Identificar gráficamente algunos estadísticos como la media, mediana o moda.

Cajas comparativas.

Con este tipo de representación se puede observar si hay diferencias entre sus distribuciones en cuanto a simetría, tendencia y dispersión. De hecho, se puede decir que en los análisis comparativos es donde mayor potencial adquiere esta herramienta.

Datos del ***ejemplo 4***, ingresos de los matrimonios.

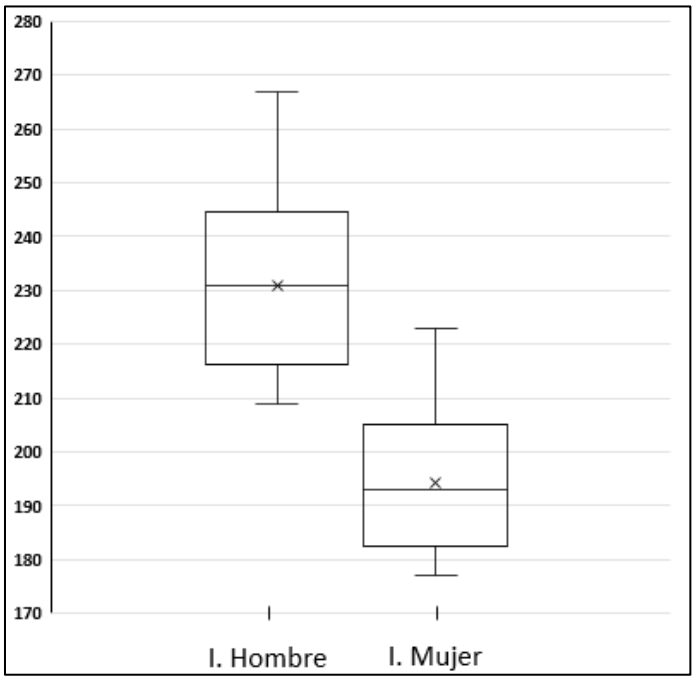


Figura 37. Gráfico de cajas y alambres para la comparación del ingreso: hombres y mujer. Fuente: Elaboración propia.

Observe que:

- No se identifican datos atípicos.
- Sesgadas (alambre superior).
- Las medias expresan que los hombres tienen mayores ingresos que las mujeres.

Diagrama de puntos comparativos.

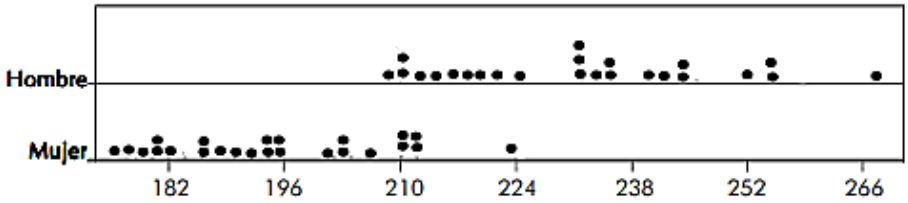


Figura 38. Diagrama de puntos para la comparación del ingreso semanal para hombres y mujeres. Fuente: Behar-Ojeda.

Para este diagrama se observa que la media de las mujeres está por debajo que la media de los hombres, indicando que el ingreso es mayor.

Gráfico de barras comparativo (líneas).

Ejemplo 9.

A 60 estudiantes del curso preuniversitario se miden varias características con el objetivo de identificar el estudiante típico; Una es el "género" como carácter cualitativo y la talla del calzado.

Género	N. Calzado	Género	N. Calzado	Género	N. Calzado
m	41	f	36	f	44
m	42	f	37	f	40
m	43	f	38	f	38
m	42	f	38	f	35
m	43	f	38	f	37
m	42	f	36	f	38
m	41	f	37	f	37
m	41	f	37	f	37
m	42	f	35	f	37
m	43	f	37	f	38
m	42	f	38	f	36
m	45	f	40	f	37
m	42	f	39	f	38
m	41	f	38	f	39
m	43	f	37	f	36
m	35	f	37	f	36
m	36	f	40	f	35
f	37	f	37	f	40
f	36	f	38	f	36
f	37	f	46	f	41

Tabla 21. Datos de los 60 estudiantes. Fuente: Carmen Batanero.

Género/Talla	*35*	*36*	*37*	*38*	*39*	*40*	*41*	*42*	*43*	*44*	*45*	*46*
Masculino	1	1	0	0	0	0	4	6	4	0	1	0
Femenino	3	7	14	10	2	4	1	0	0	1	0	1

Tabla 22. Frecuencia absoluta conjunta. Fuente: Elaboración propia.

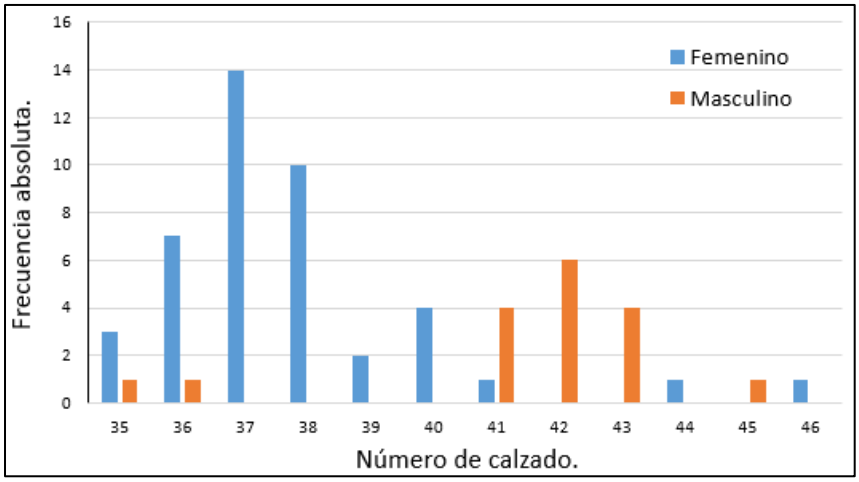


Figura 39. Comparación del número de calzado entre femenino y masculino. Fuente: Elaboración propia.

Fácilmente se identifican algunas características como las modas de cada grupo (moda: 37 en femenino y 42 en masculino); Además, se identifica que hay más mujeres que hombres y que los hombres tiene el calzado más grande.

Caso 5. Continuo-cualitativo.

La metodología es similar al anterior, la diferencia radica en el tipo de gráfico. Esto significa que quien determina el gráfico base es la variable continua; Gráficos a comparar: histograma, cajas y alambres, polígono de frecuencias, dispersión unidimensional, ojiva, etc. Al comparar histogramas, cada grupo presenta sus respectivos intervalos de clases y no necesariamente el número de clases deben ser iguales.

Ejemplo 10.

A 66 estudiantes del curso de estadística descriptiva se miden dos características. El género como variable cualitativa y estatura (cm) como variable continua.

Género	Estatura (Cm)	Género	Estatura (Cm)	Género	Estatura (Cm)	Género	Estatura (Cm)
m	178	f	162	f	161	m	176
m	176	f	162	f	166	m	176
m	175	f	170	f	166	m	176
m	170	f	170	f	156	m	176
m	170	f	165	f	165	m	176
m	180	f	163	f	167	m	172
m	175	f	167	f	160	f	180
m	174	f	167	f	160	f	181
m	178	f	160	f	161		
m	185	f	164	f	169		
m	175	f	163	f	158		
m	183	f	175	f	155		
m	185	f	173	f	163		
m	174	f	161	f	168		
m	179	f	162	f	163		
m	164	f	172	f	165		
m	170	f	165	f	155		
f	161	f	160	f	174		
f	159	f	164	f	162		
f	169	f	191	f	172		

Tabla 23. Datos registrados: Estatura (masculino y femenino). Fuente: Elaboración propia.

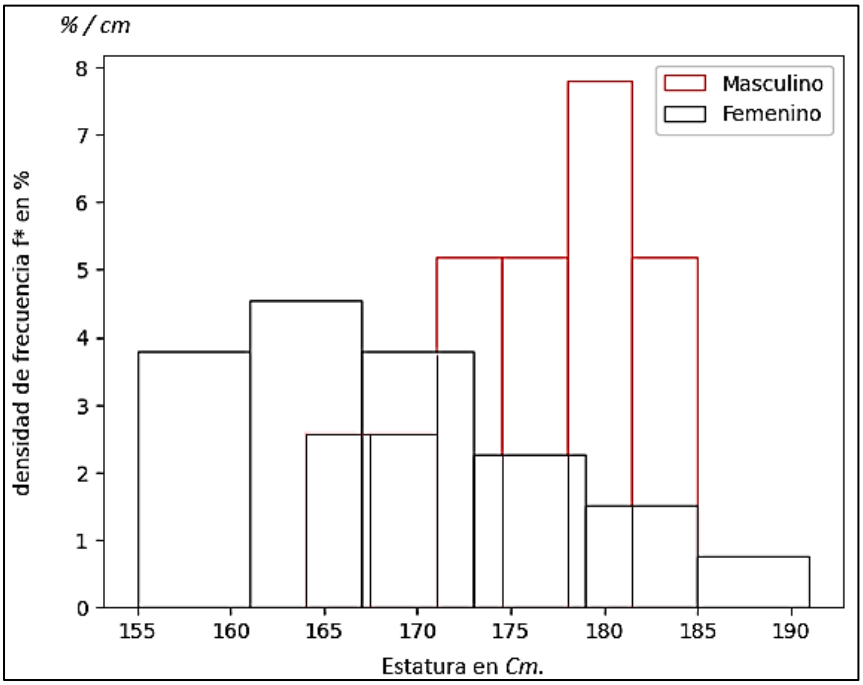


Figura 40. Histogramas comparativos para la estatura entre masculino y femenino. Fuente: Elaboración propia.

El histograma para el género masculino se toma como dato mínimo 164 cm, máximo como 185 cm, rango 21, seis clases y todas de longitud 3.5 cm.

El histograma para el género femenino se toma como dato mínimo 155 cm, máximo como 191 cm, rango 36, 6 clases y todas de longitud 6 cm.

Los histogramas al traslaparse no deben perder el objetivo principal que es representar las distribuciones. En esta representación gráfica se observa que los hombres tienden a ser más altos que las mujeres. En la distribución de las mujeres hay un sesgo a la derecha. Se refleja las diferencias entre las clases modales.

Polígonos de frecuencia comparativos.

Esta representación es óptima cuando hay más de dos grupos a comparar, ya que al usar los histogramas la representación estaría saturada. Vea datos del ***ejemplo 10***.

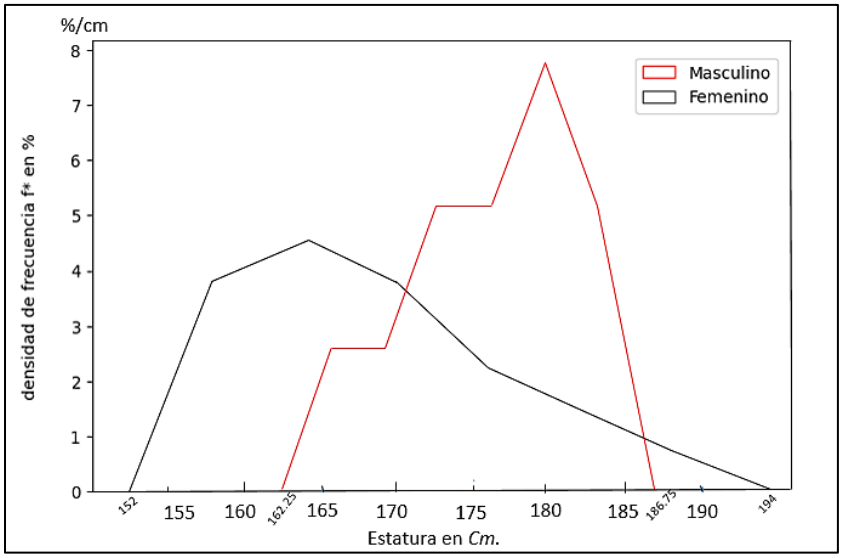


Figura 41. Poligonos de frecuencias para la comparación de la estatura según el género. Fuente: Elaboración propia.

Diagramas de dispersión unidimensional comparativos.

Para los datos del ***ejemplo 10***, se realiza una representación mediante los diagramas de dispersión unidimensionales, identificando y comparando las distribuciones en cada categoría del género.

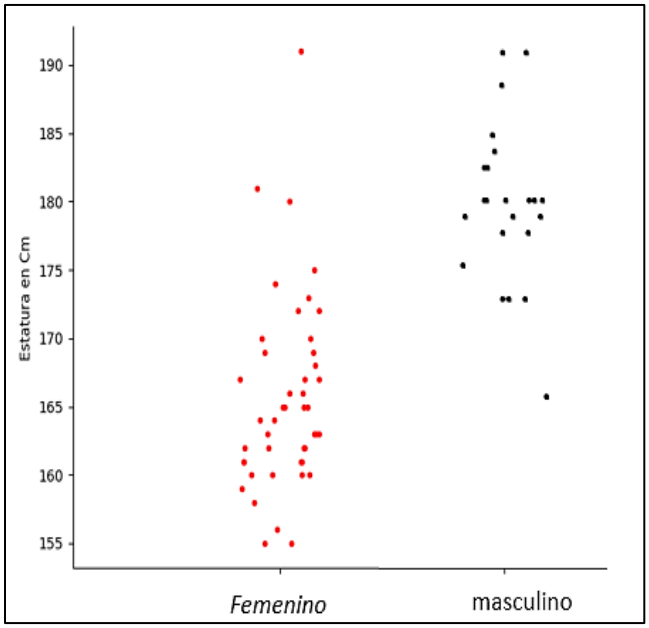


Figura 42. Diagramas dispersión unidimensionales comparativos de estatura entre estudiantes masculinos y femeninos. Fuente: Elaboración propia.

Tallos y hojas comparativo.

Continuando con el ***ejemplo 10***, se aplicará el diagrama de tallos y hojas. En esta representación se aprecia los tallos modales, la forma de la distribución en cada grupo. Además, se identifica la relación que hay entre los histogramas comparativos, por ejemplo, el sesgo a la "derecha" para el grupo femenino.

Hojas	Tallo	Hojas
9 8 6 5 5	15	
4 4 3 3 3 3 2 2 2 2 1 1 1 1 0 0 0 0	16	4
9 9 8 7 7 7 6 6 5 5 5 5	16	
4 3 2 2 0 0	17	0 0 0 2 4 4
5	17	5 5 5 6 6 6 6 6 8 8 9
5 1 0	18	0 3
5	18	5 5
1	19	
femenino		masculino

Escala:16|4 → 164.

Figura 43. Diagrama de tallos y hojas comparativos para la estatura (Cm) de estudiantes según el género. Fuente: Elaboración propia.

Caso 6. Cualitativo-cualitativo.

Las tablas de contingencia ayudan tener más control de las visualizaciones, permitiendo identificar patrones en la distribución conjunta, condicionales o marginales. En las representaciones se emplean gráficos simples de datos categóricos.

Algunos ejemplos de datos de este tipo son:

- Procedencia vs género.

- Color de ojos vs color de pelo.
- Estrato socioeconómico vs nivel educativo.
- Etc.

En el análisis de datos hay herramientas en cargadas de estudiar la relación entre dos variables cualitativas; El análisis de correspondencia simple (ACS) permite determinar la existencia de asaciones entre las modalidades de dos variables cualitativas. Las representaciones gráficas permiten un primer acercamiento para identificar estas asaciones, por ejemplo, el gráfico de *perfiles*, que es básicamente un gráfico de barras apilado 100 %.

- *Perfiles fila:* Presenta las distribuciones condicionales por fila.
- *Perfiles columna:* Presenta las distribuciones condicionales por columna.

Barras comparativas son representaciones gráficas básicas para datos de dos variables cualitativas, tienen como fin comparar las distribuciones condicionadas por filas o columnas y se clasifican en tres tipos: *Agrupadas, apiladas, apiladas 100 %.*

Barras agrupadas, apiladas y apiladas 100 %.

Ejemplo 12.

A 60 estudiantes del curso preuniversitario se miden características con el objetivo de identificar el estudiante típico; Además, el *género* y *deporte favorito*, ambas como caracteres cualitativos:

m: masculino, f: femenino, 1: Voleibol. 2: Futbol. 3: Baloncesto.

Género.	Deporte.	Género.	Deporte.	Género.	Deporte.
m	1	f	2	f	1
m	2	f	2	f	1
m	2	f	1	f	3
m	1	f	2	f	1
m	3	f	1	f	2
m	2	f	3	f	2
m	2	f	2	f	2
m	3	f	2	f	3
m	2	f	2	f	1
m	3	f	1	f	3
m	3	f	2	f	1
m	3	f	2	f	3
m	3	f	2	f	1
m	2	f	2	f	1
m	2	f	2	f	2
m	2	f	1	f	1
m	2	f	2	f	2
f	2	f	1	f	2
f	2	f	2	f	2
f	3	f	3	f	2

Tabla 24. Registros de los 60 estudiantes. Fuente: Carmen Batanero.

	m	f	
voleibol.	2	13	**15**
futbol.	9	23	**32**
baloncesto.	6	7	**13**
	17	**43**	**60**

Tabla 25. Distribución conjunta: deporte y género. Fuente: Elaboración propia.

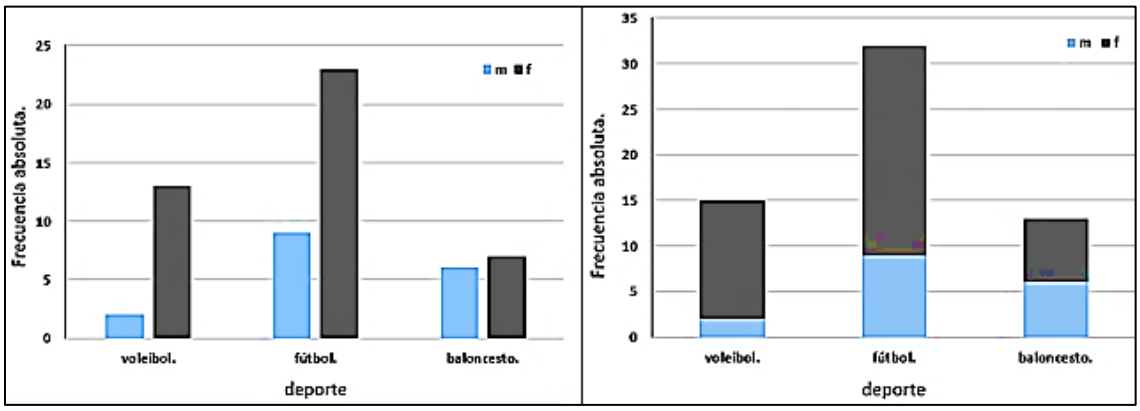


Figura 44. Barras agrupadas y apiladas: deporte según el género. Fuente: Elaboración propia.

	m	f	
voleibol.	13%	87%	**100%**
fútbol.	28%	72%	**100%**
baloncesto.	46%	54%	**100%**
Todos los deportes.	28%	72%	**100%**

Tabla 26. Perfiles fila: deporte según género. Fuente: Elaboración propia.

	m	f	Género
voleibol.	12%	30%	25%
fútbol.	53%	53%	53%
baloncesto.	35%	16%	22%
	100%	**100%**	**100%**

Tabla 27. Perfiles columna: género según deporte. Fuente: Elaboración propia.

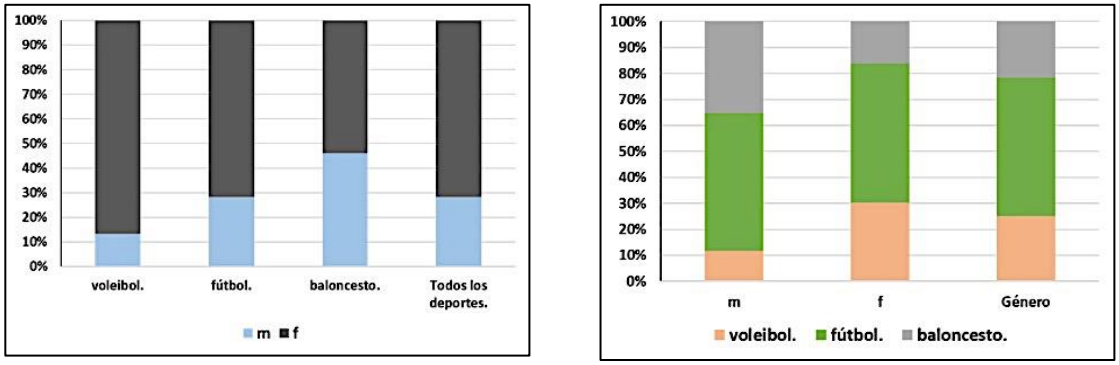


Figura 45. Barras apiladas 100 %: perfiles fila, columna y marginales. Fuente: Elaboración propia.

Rectángulos apilados 100 %.

Es posible agregar más información al gráfico de los perfiles permitiendo representar sobre un mismo gráfico la distribución global (n_{ij}) y una familia de distribuciones condicionadas (por filas o por columnas, pero no las dos simultáneamente).

Se presenta la frecuencia absoluta conjunta n_{ij} por el área de un rectángulo cuya base es proporcional a $n_{.j}$ ($n_{i.}$) y cuya altura es proporcional a la frecuencia condicionada f en i dado j (o f en j dado i).

El resultado es un gráfico que permite evidenciar:

- Las frecuencias absolutas marginales $n_{.j}$ (base delos rectángulos).
- Las frecuencias condicionadas (altura de los rectángulos).
- La frecuencia absoluta de las tablas de doble entrada n_{ij} (área de los rectángulos).

Ejemplo 13.

Distribución de viviendas según la época de construcción y la categoría socio-profesional del cabeza de familia (*Censo general de la población; marzo 1962. Sondeo del 1/20*).

Nota. Rigurosamente el tiempo es de carácter continuo. Debido a la gran amplitud de las clases consideradas, se supone aquí que cada clase define una época de construcción de la modalidad de un carácter cualitativo.

	Antes de 1871	1871 a 1914	1915 a 1948	Despúes de 1948	Total.
Agricultores	873340	410040	158380	74620	**1516380**
Asalariados agrícolas.	233060	100160	48600	27280	**409100**
Patronos de la industria y del comercio.	415380	413000	280520	195380	**1304280**
Profecionales liberales y cuadros superiores.	87120	175660	148760	204440	**615980**
Cuadros medios.	144560	247860	210640	293180	**896240**
Empleados.	231760	322700	237800	249180	**1041440**
Obreros.	1118440	1177820	954500	956040	**4206800**
Personal de servicio.	112560	123260	72400	43720	**351940**
Otras categorías.	73240	77960	65360	95960	**312520**
Personas no activas.	1398840	1232120	932340	322220	**3885520**
Total.	**4688300**	**4280580**	**3109300**	**2462020**	**14540200**

Tabla 28. Distribución conjunta: época de construcción y categoría socio-profesional. Fuente: Gerard Calot.

	Antes de 1871	1871 a 1914	1915 a 1948	Despúes de 1948
Agricultores	57.6%	84.6%	95.1%	100.0%
Asalariados agrícolas.	57.0%	81.5%	93.3%	100.0%
Patronos de la industria y del comercio.	31.8%	63.5%	85.0%	100.0%
Profecionales liberales y cuadros superiores.	14.1%	42.7%	66.8%	100.0%
Cuadros medios.	16.1%	43.8%	67.3%	100.0%
Empleados.	22.3%	53.2%	76.1%	100.0%
Obreros.	26.6%	54.6%	77.3%	100.0%
Personal de servicio.	32.0%	67.0%	87.6%	100.0%
Otras categorías.	23.4%	48.4%	69.3%	100.0%
Personas no activas.	36.0%	67.7%	91.7%	100.0%
Todas las categorias socio-profecionales.	32.2%	61.7%	83.1%	100.0%

Tabla 29. Perfiles fila: categoría social según época. Fuente: Gerard Calot.

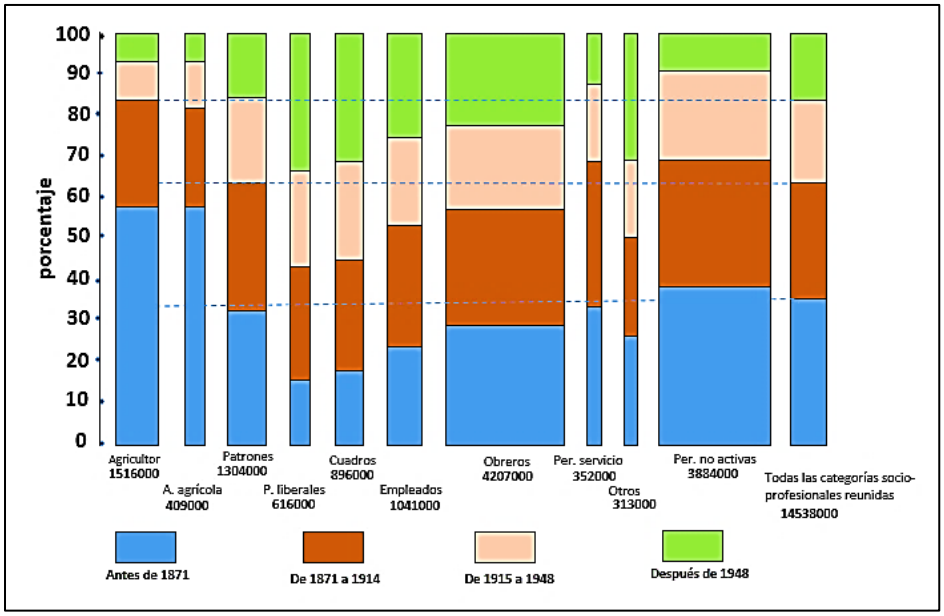


Figura 46. Gráfico de rectángulos apiladas 100 % para la distribución de viviendas según la época de construcción y la categoría socio-profesional del cabeza de familia. Fuente: Gerard Calot.

Análogamente se construye los perfiles columna y su representación gráfica; Pero la visualización en este ejemplo es menos satisfactoria: Es preferible compara las categorías socio-profesionales entre sí (por ejemplo, para apreciar en qué medida cada una tiene acceso a las viviendas nuevas) antes que compara las épocas de construcción. Además, el número de modalidades de la categoría socio-profesional es muy superior al número de modalidades de la época de construcción. Así, el gráfico anterior es mucho más claro y será el adecuado para una representación gráfica.

Gráfico de araña para dos variables cualitativas.

Se identifica la variable con mayor categoría y se construye un diagrama de araña, para cada modalidad de la segunda variable se construye su respectivo polígono.

Se aplica cuando alguna de las dos variables tiene como mínimo tres categorías. Se permite la escala de frecuencia conjunta absoluta o relativa.

De los datos del ***ejemplo 12*** se tiene.

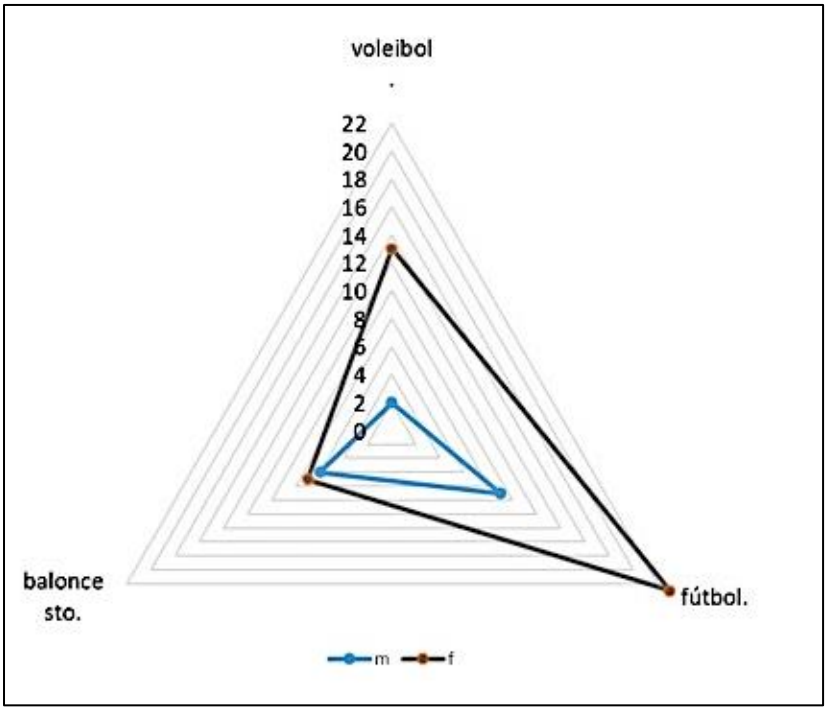


Figura 47. Gráfico de radar (f. absoluta) para las variables género-deporte. Fuente: Elaboración propia.

Diagrama de áreas rectángulas comparativas.

Este diagrama es la ampliación para el caso unidimensional, conservando el orden ***jerárquico*** de las categorías y permitiendo una rápida y fácil comparación gráfica de la distribución de cada grupo de datos mediante el tamaño de los rectángulos. Comparar demasiados grupos (más de dos) podría saturar el gráfico, esto implica recurrir a otras técnicas gráficas como el *gráfico de proyección solar*. Considere el ***ejemplo 12.***

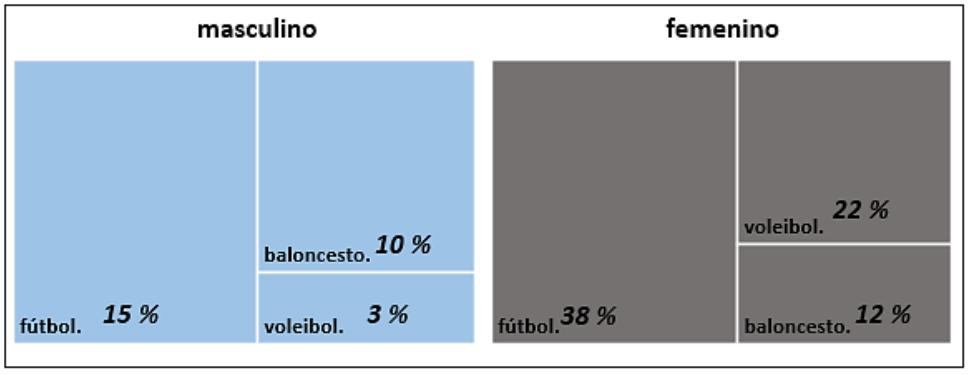


Figura 48. Diagrama de áreas rectangulares comparativas para las variables género-deporte. Fuente: Elaboración propia.

Gráfico de sectores comparativos.

Un gráfico equivalente al gráfico de *barras apiladas 100 %* es mediante la comparación de gráficos de sectores del mismo radio. Si deseamos que los sectores sean de mayor relevancia, entonces influirá el radio de cada círculo, este puede indicar un orden jerárquico.

Cuando uno de los caracteres cualitativos presenta solo dos modalidades, se puede utilizar una representación por círculos o semicírculos, de gran relevancia que contenga la distribución global y una familia de distribuciones condicionales. Consideraciones:

- A cada una de las dos modalidades se le asocia un semicírculo compartiendo base.
- Los *radios* son proporcionales a la raíz cuadrada de la frecuencia absoluta *marginal*.
- Los *ángulos* son proporcionales a las frecuencias *condicionadas*.
- Las *áreas* de los sectores son proporcionales a las frecuencias absolutas conjunta n_{ij}.

Ejemplo 14.

Considera la tabla de doble entrada que muestra información de la distribución de *5254000* franceses de 65 años y más por sexo y estado matrimonial (*calculo INSEE, 1. ° enero 1960*).

	Soltero	Casado	Viudo	Divorciado	Total
Masculino	112	1352	439	35	1938
Femenino	350	1002	1902	62	3316
Total	462	2354	2341	97	5254

Tabla 30. Distribución conjunta de las variables sexo y estado matrimonial en miles de personas. Fuente: Gerard Calot.

	Soltero	Casado	Viudo	Divorciado	Total
Masculino	10.4	125.6	40.8	3.3	180
Femenino	19.0	54.4	103.2	3.4	180

Tabla 31. Ángulos correspondientes a la distribución de los perfiles filas. Fuente: Elaboración propia.

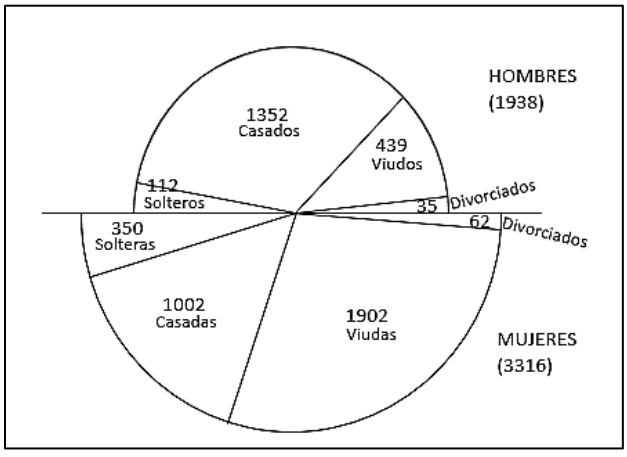


Figura 49. Distribuciones según estado matrimonial para los hombres y las mujeres. Fuente: Elaboración propia.

Gráfico de anillos múltiple.

Otra versión al gráfico anterior es el gráfico de anillos múltiples que es una extensión del gráfico de anillo simple. Se usa para representar varios grupos de datos mediante anillos concéntricos, donde cada anillo corresponde a un grupo de datos distinta, permitiendo extraer conclusiones de manera fácil y rápida. La ventaja respecto al gráfico de sectores, es que visualmente es más atractivo y resume varias modalidades sin saturar la representación gráfica. Considere los datos del ***ejemplo 12***.

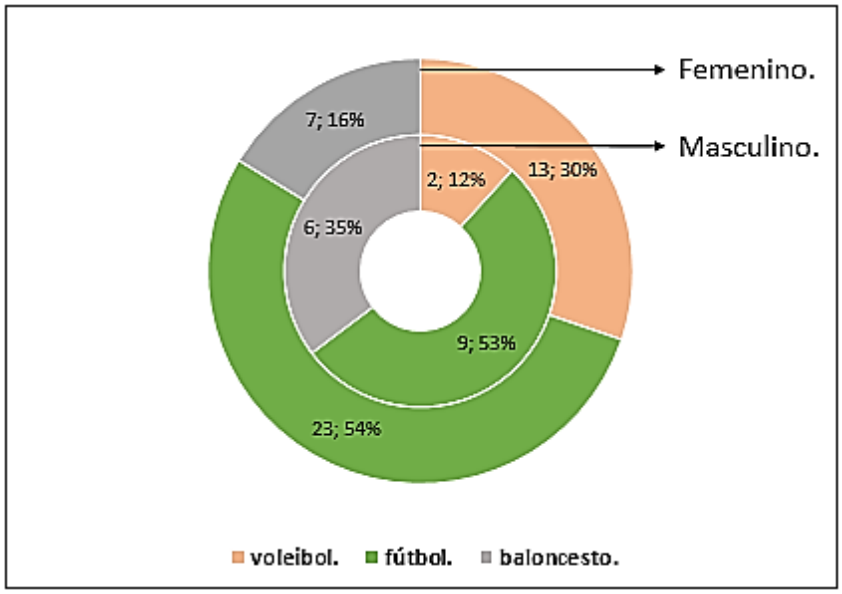


Figura 50. Diagrama de anillo múltiple para las variables género-deporte. Fuente: Elaboración propia.

Gráfico de proyección solar.

El gráfico de proyección solar es ideal para mostrar datos jerárquicos en relacionan a los anillos exteriores con los interiores. El tramado está dado por el primer anillo de mayor jerarquía, proyectándolo para los demás anillos los cuales solo van aumentando las divisiones del color. Las etiquetas son esenciales para diferenciar y comparar los sectores.

Figura 51. Gráfico de proyección solar para las ventas por país y ciudades. Fuente: https://tutorialexcel.com/los-graficos-de-jerarquia-en-excel/

Gráficas para datos multidimensionales.

A continuación, se presenta una serie de técnicas gráficas que permiten visualizar datos en altas dimensiones, cuando el principal objetivo es identificar patrones de agrupación o niveles de discriminación entre individuos o correlaciones entre las variables. Algunos gráficos estadísticos multivariantes, son gráficos trabajados anteriormente que permiten extenderse codificando información para representar más de dos variables. Nuevas técnicas como *dispersión tridimensional*, *Caras de Chernoff y Curvas de Andrews* son usados en varios campos de investigación. Estos gráficos en general responden a ¿qué variables son las dominantes en cada unidad de observación elegida?, ¿se pueden establecer similitudes entre estas unidades o relaciones?, ¿se pueden establecer conjuntos (*clusters*) ?, ¿existen datos atípicos?

Ejemplo 15.

A 60 estudiantes del curso preuniversitario se miden varias características con el objetivo de identificar el estudiante típico. Se miden características cualitativas y cuantitativas, obteniendo la siguiente matriz (60x8) de datos.

X_1: Genero, X_2: Deporte favorito, X_3: Color de ojos, X_4: Color cabello. X_5: N°. Calzado, X_6: Peso, X_7: Estatura, X_8: Longitud de brazos.

Género	Deporte	C. ojos	C. cabello	N. calzado	Peso (kg)	Estatura (cm)	Log. Brazos (cm
m	1	o	o	41	62	178	181
m	2	o	o	42	69	176	179
m	2	o	o	43	74	175	179
m	1	o	o	42	68	170	172
m	3	o	o	43	72	184	185
m	2	c	c	42	74	180	182
m	2	c	c	41	66	175	177
m	3	c	c	41	82	174	180
m	2	o	c	42	68	178	180
m	3	c	c	43	71	185	187
m	3	c	c	42	68	175	172
m	3	o	c	45	74	183	178
m	3	o	o	42	68	185	185
m	2	o	o	41	69	172	172
m	2	o	o	43	81	184	188
m	2	c	c	35	65	179	171
m	2	o	o	36	65	164	158
f	2	c	c	37	59	161	160
f	2	o	o	36	50	159	153
f	3	c	c	37	62	169	165
f	2	o	o	36	56	162	158
f	2	o	o	37	58	162	163
f	1	o	o	38	52	170	171
f	2	o	o	38	60	170	168
f	1	c	c	38	60	165	161
f	3	o	o	36	55	163	160
f	2	o	o	37	60	167	165
f	2	c	o	37	50	167	165
f	2	c	c	35	52	160	157
f	1	o	o	37	53	164	160
f	2	o	c	38	58	163	166
f	2	o	o	40	74	175	178
f	2	o	o	39	63	173	180
f	2	o	c	38	60	161	164
f	2	o	o	37	53	162	162
f	1	o	c	37	64	172	175
f	2	o	o	40	65	165	165
f	1	o	o	37	46	160	158
f	2	c	c	38	58	164	166
f	3	c	o	46	86	191	180
f	1	o	o	44	70	161	185
f	1	o	c	40	64	166	171
f	3	o	o	38	64	166	155
f	1	o	o	35	70	156	152
f	2	c	c	37	51	165	160
f	2	c	o	38	62	167	159
f	2	o	c	37	58	160	160
f	3	o	o	37	55	160	154
f	1	o	o	37	57	161	155
f	3	c	o	38	57	169	164
f	1	o	o	36	68	158	150
f	3	c	o	37	50	155	155
f	1	o	o	38	58	163	162
f	1	c	o	39	66	168	168
f	2	c	c	36	50	163	161
f	1	c	c	36	60	165	160
f	2	c	c	35	50	155	155
f	2	c	c	40	62	174	179
f	2	o	c	36	58	162	160
f	2	c	c	41	63	172	171

Tabla 32. Matriz de datos. Fuente: Carmen Batanero.

Diagrama de dispersión tridimensional.

Es una extensión al diagrama de dispersión bidimensional, permitiendo investigar datos atípicos o valores de respuesta. Considere X_6: Peso, X_7: Estatura, X_8: Longitud de brazos del ***ejemplo 15.***

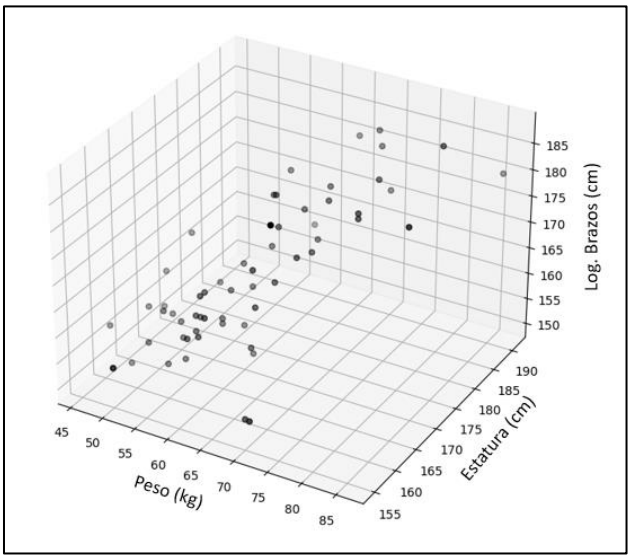


Figura 52. Diagrama de dispersión para las características: Peso, estatura, log. brazos. Fuente: Elaboración propia.

Diagrama de dispersión con variables codificadas.

Representar más de tres características lleva a representarlas de manera codifica usando *glifos*, donde el tamaños o tonalidades pueden representar otras características. La variable a codificar puede ser de naturaleza cuantitativa o cualitativa. Considere X_1: Genero, X_2: Deporte, X_6: Peso, X_7: Estatura, X_8: Longitud de brazos del ***ejemplo 15.***

La característica *género* se codificará mediante el color, el *deporte* se codifica con glifos y se tendrá una representación gráfica para datos de dimisión cinco.

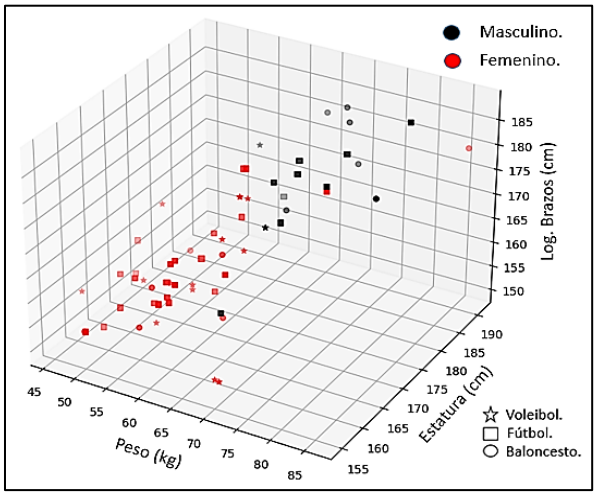


Figura 53. Diagrama de dispersión codificado para las variables: X_1, X_2, X_6, X_7, X_8. Fuente: Elaboración propia.

Diagramas comparativos multidimensional.

Son extensiones de gráficos de datos unidimensionales que de manera adecuada pueden representar datos en distintas dimensiones. Por ejemplo, son de gran utilidad los diagramas de caja y alambres para comparar rasgos de las distribuciones de varios conjuntos de datos.

Considere X_1: Genero, X_2: Deporte, X_7: Estatura del ***ejemplo 15.***

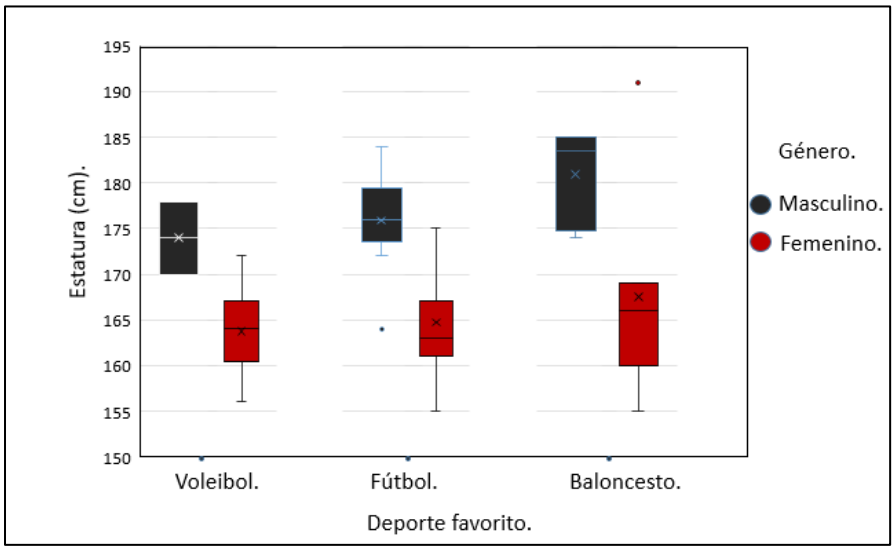


Figura 54. Diagrama de cajas y alambres comparativos de la estatura según el deporte y el género. Fuente: Elaboración propia.

Se observa claramente que los hombres son más altos que las mujeres en cada deporte y algunos puntos atípicos.

Gráfico de araña para tres variables cualitativas.

Los polígonos comparan frecuencias, donde, el grosor, el color o la continuidad indican variables.

Considere X_1: Genero, X_2: Deporte, X_4: Color cabello del ***ejemplo 15.***

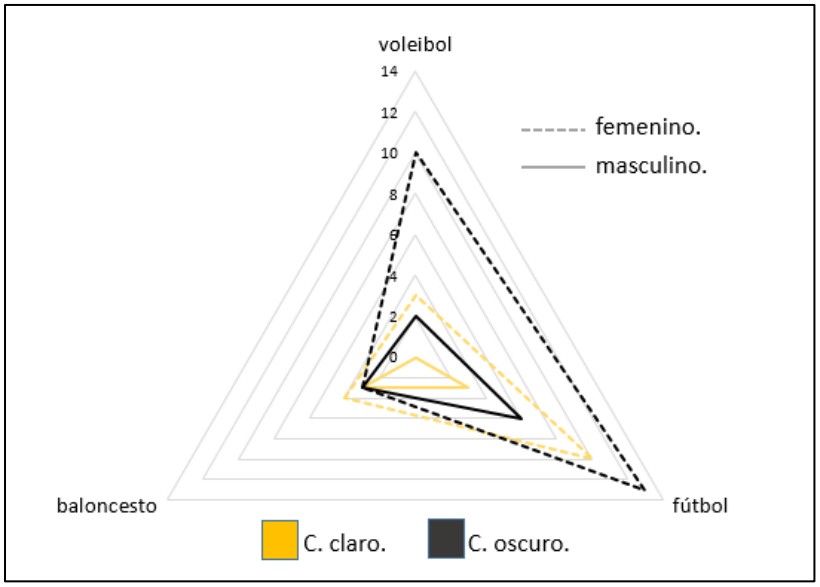


Figura 55. Gráfico de araña para las características: X_1, X_2 y X_4 de la matriz de datos. Fuente: Elaboración propia.

Diagrama de escalera.

Para una matriz de datos de n observaciones con p variables cuantitativas y se desea visualizar los patrones de asociación entre variables, se realizan diagrama de dispersión para todos los pares de variables, se sugiere *p<10*, obteniéndose un diagrama de escalera (arreglo triangular).

Permite tener una visualización en conjunto de las asociaciones entre pares de variables, incluso información de las distribuciones de cada variable como histogramas o información de variables categóricas usando alguna codificación.

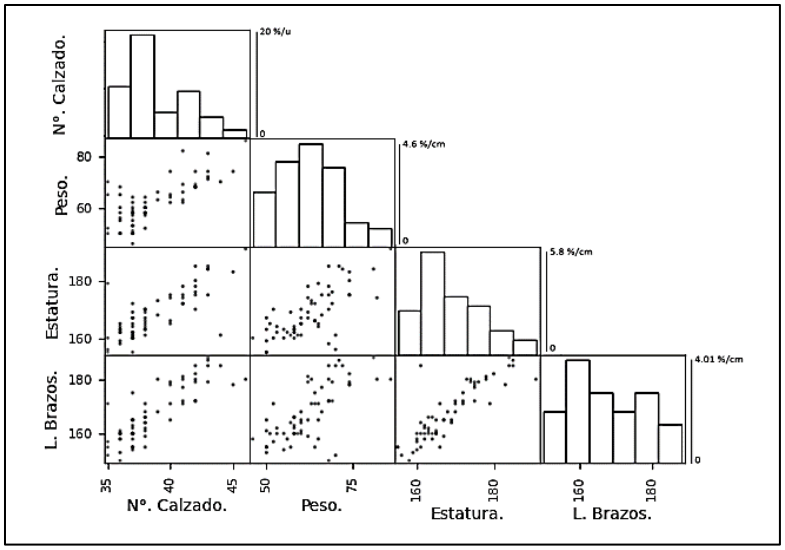


Figura 56. Diagrama de escalera para las variables: X_5, X_6, X_7, X_8. Fuente: Elaboración propia.

En la diagonal cada histograma tiene una escala distinta ya que cada uno es de una variable diferente; No se está estableciendo comparaciones entre grupos de datos.

En el diagrama de escalera también es posible incorporar variables de carácter discreto, esto implica que en la diagonal del diagrama de escalera también se identifiquen gráficos de frecuencia para variable discreta. Mediante el coeficiente de correlación de Pearson se tiene el grado de relación, identificándose comportamientos de asociación lineal.

	Calzado	Peso	Estatura	L. Brazo
Calzado	1	0.97	0.98	0.98
Peso	0.97	1	0.98	0.97
Estatura	0.98	0.98	1	0.98
L. Brazo	0.98	0.97	0.98	1

Tabla 33. Correlaciones. Fuente: Elaboración propia.

Gráfico de radar multivariante.

Se asemeja al gráfico anterior, pero se construye en base a que cada variable (cuantitativa) en estudio es un polígono. Las unidades de observación identifican los vértices o rayos. Los valores individuales de cada variable, debidamente estandarizados, se representan sobre los rayos mencionados y la conexión de ellos genera el polígono. La longitud de un determinado rayo es proporcional a la magnitud de la variable para el subconjunto con respecto al máximo de la variable a través de todos los subconjuntos.

Ejemplo 16.

El Sistema Público de Salud Chileno cubre el territorio nacional y está estructurado en 29 organizaciones denominadas Servicios de Salud; Se le miden 7 características cuantitativas a cada uno de los servicios (*año 2010*).

Variables: Dotación, Días Cama Disponibles, Días Cama Ocupados, Días de Estada, Promedio Camas Disponibles, Número de Egresos y Número de Egresos Fallecidos.

Las variables fueron estandarizadas con valores entre *-2 y 3*, para evitar que una variable "domine" debido a que sus valores son (escala) miles de veces mayores que los de las otras variables.

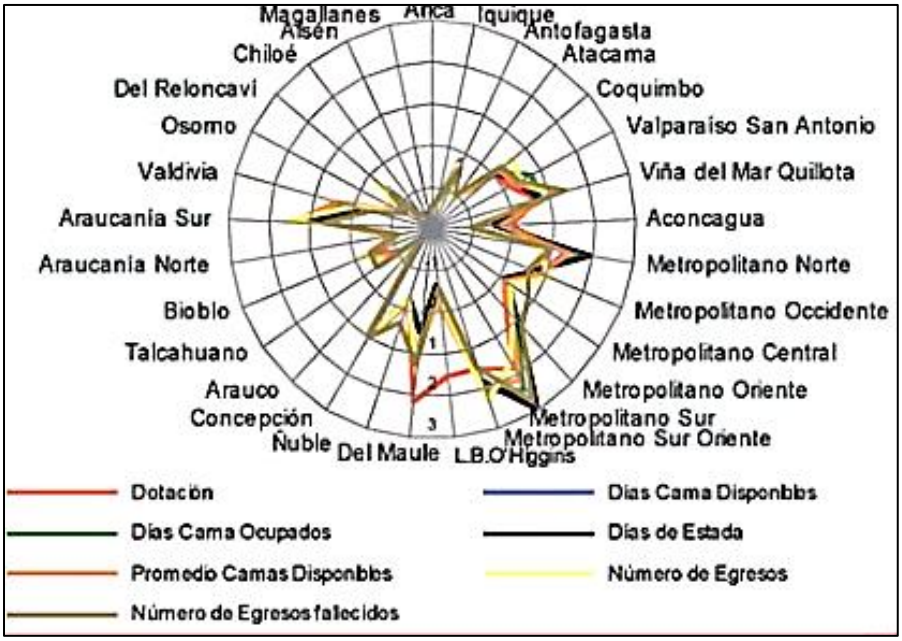


Figura 57. Gráfico de radar: 29 centros de servicio de salud del sistema público chileno, año 2010. Fuente: Irene Schiattino-Claudio Silva.

Gráfico de estrellas.

Contraste al gráfico de radar, el gráfico de estrella consiste en asignar a cada observación una estrella o polígono con tantos rayos como variables cuantitativas queramos representar. Cada estrella se identifica por separado, esto permite agrupar las observaciones según las similitudes y adecuados para evidenciar valores atípicos.

Ejemplo 17.

A 16 automóviles se le miden las siguientes características: *Precio, Kilometraje (MPG), Registro de reparación de 1978 (1 = peor, 5 = mejor), Registro de reparación de 1977 (1 = peor, 5 = mejor), Espacio para la cabeza, Espacio en el asiento trasero, Espacio del maletero, Peso, Largo.* Luego se estandarizan los datos.

El *Cadillac Seville* es uno de los autos más caros, obtiene un rendimiento de gasolina por debajo del promedio (pero no entre los peores), tiene un promedio registro de reparación, y tiene una amplitud y un tamaño de promedio a superior al promedio.

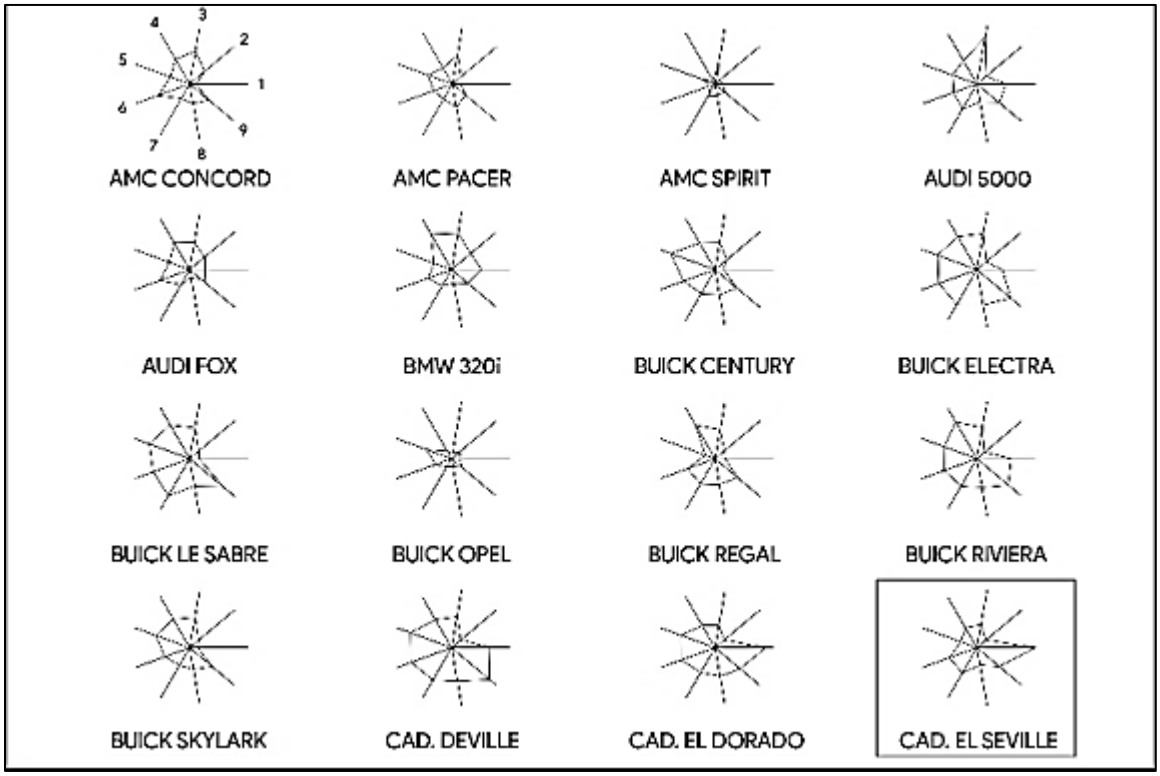


Figura 58. Gráfico de estrellas para el análisis de 16 autos de 1979. Fuente: https://commons.wikimedia.org/wiki/File:Star_Plot_of_16_cars.jpg

A menudo resulta útil combinar la representación gráfica con método de reducción de dimensión de los datos.

El análisis de componentes principales (ACP) es un método de reducción de dimensión de los datos. La idea es conseguir *p* variables ortogonales nuevas (***componentes principales***) y cada una de ellas es una combinación lineal de las *m* variables originales, de forma que expliquen la mayor cantidad de variabilidad del conjunto de datos original. ACP pertenece a los métodos de aprendizaje no supervisado. Ayuda a la representación de los datos en dos o tres dimensiones evidenciando agrupamientos de datos.

Ejemplo 18.

Suponga 28 modelos de coches de 1977, se miden cinco variables y se estandarizan los datos con una media de cero y varianza uno. Por ejemplo, usando ACP se reducen la dimensionalidad de los datos pasando de cinco a dos, cuya variabilidad sea representativa de los datos originales: tomando como medida de diferenciación las distancias euclídeas entre las observaciones estandarizadas. Finalmente se crear una gráfica en 2D significando una pérdida de información, al trazar las estrellas, se incorpora toda la información de alta dimensión de los datos. El propósito de este método gráfico-numérico es imponer cierta regularidad a la variación de los datos, para que los patrones entre los glifos sean más fáciles de ver.

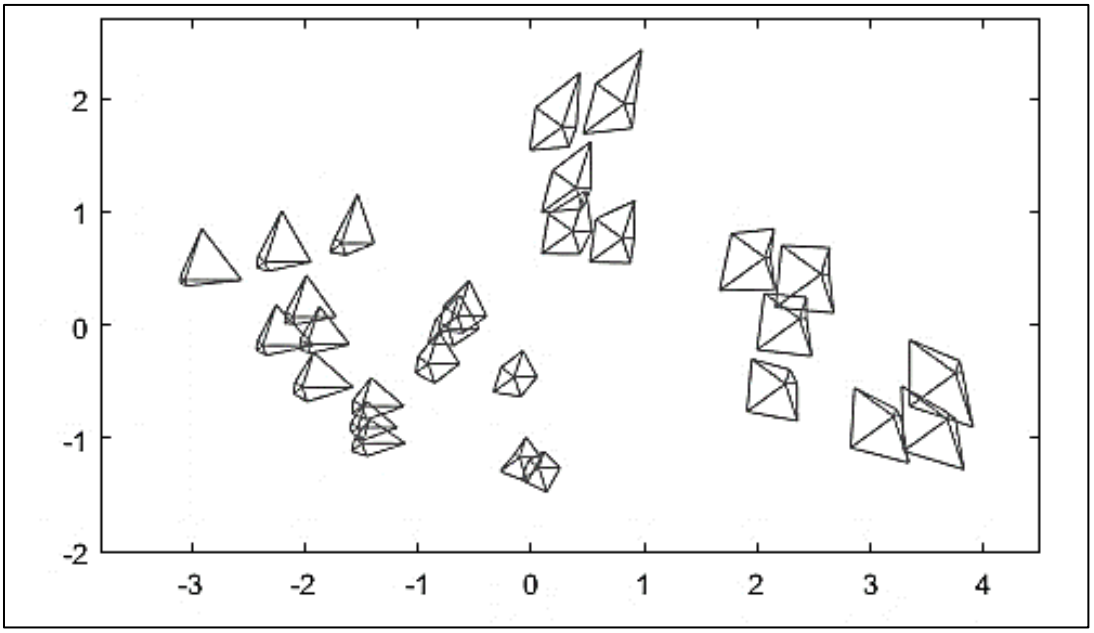


Figura 59. Gráfico de estrellas y reducción de dimensión con ACP. Fuente: Visualización de datos multivariados - MATLAB & Simulink Example - MathWorks América Latina.

Gráfica de líneas paralelas.

También llamado gráfico de coordenadas paralelas, es un despliegue bidimensional para representar datos numéricos o categóricos; Está formado por un conjunto de líneas paralelas equidistantes que identifican a las variables. y otro que identifica a cada individuo. A partir de este gráfico se puede identificar datos atípicos o patrones de asociación entre variables; El gráfico resulta ideal si las escalas para cada variable son las mismas, en caso de contrario se estandariza. También, permite un número considerable de observaciones y variables, pero, es claro que puede resultar difícil de leer. El gráfico de estrellas y de líneas paralelas son equivalentes, presentan la misma información, el orden de las variables influye en los patrones.

Ejemplo 19.

Se consideran 13 muestras de leche y se mide porcentualmente las siguientes variables.

X_1: Grasa, X_2: Ceniza, X_3: Solidos totales, X_4: Proteínas.

X_1	X_2	X_3	X_4
4.8	1.14	14.45	4.707
5.7	0.94	16.7	4.6322
5.8	1.07	14.66	4.5512
4.8	1.13	16.74	4.5101
4.8	1.01	16.99	4.6344
5.4	1.01	17.05	4.8012
5.1	0.95	14.19	4.7103
5.25	0.85	16.68	4.7847
5.6	1.06	15.03	4.8038
5.2	1.01	16.33	4.9851
5.1	1.06	13.66	4.998
5.3	0.92	16.19	4.5359
5.75	0.94	16.7	5.0332

Tabla 34. Matriz de datos de tamaño 13x4 muestra de leche. Fuente: Q. F. B. Rosa Guadalupe Herrera Lee.

Suponga una nueva variable, X_5: Jornada (modalidades: mañana y tarde).

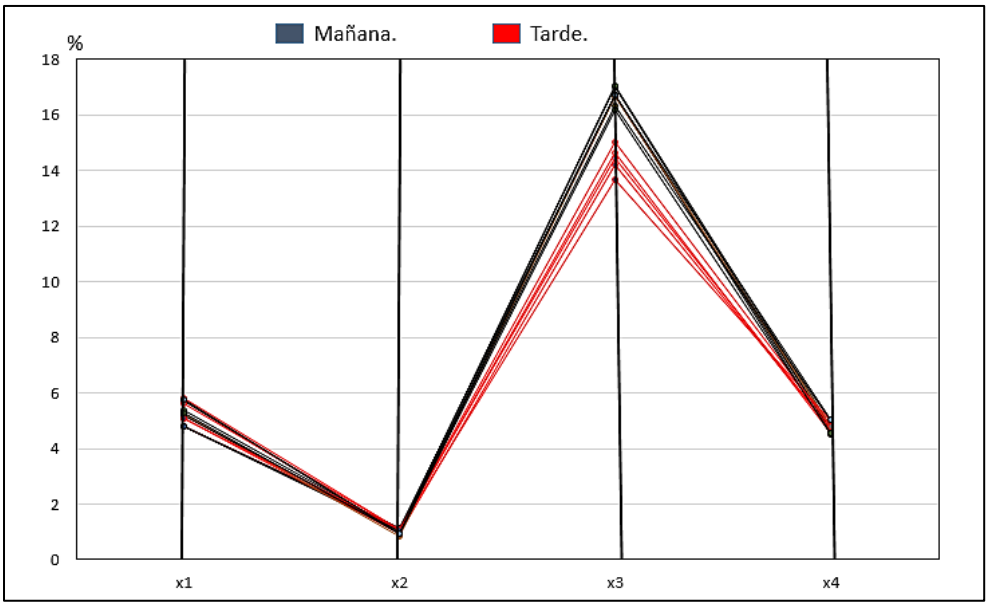


Figura 60. Gráfico de líneas paralelas para las variables: X_1, X_2, X_3, X_4 y X_5. Fuente: Propia.

Caras de Chernoff.

Otra técnica ingeniosa para representar gráficamente datos de altas dimensiones se da por *Herman Chernoff;* Propone transformar los datos en rostros humanos; Asocia a cada variable una característica del rostro, como longitud de la nariz, tamaño de los ojos, forma de los ojos, ancho de la boca, entre otras. Esta visualización representa los valores de los datos de cada observación en rasgos faciales.

La visualización permite identificar patrones en los datos, donde se miden varias variables al mismo tiempo, obteniendo una clasificación visual. Generalmente los datos son estandarizados para que estén entre cero y uno.

Para generar este tipo de representación gráfica, es primordial el uso de software estadístico; Algunos softwares ofrecen la posibilidad de producir un despliegue de hasta 20 variables.

1. Altura de la cara.
2. Forma de la frente.
3. Posición vertical de los ojos.
4. Tamaño de los ojos.
5. Inclinación de los ojos.
6. Inclinación de la ceja izquierda.
7. Inclinación de la ceja derecha.
8. Largo de la ceja izquierda.
9. Largo de la ceja derecha.
10. Posición del iris izquierdo.
11. Posición del iris derecho.
12. Longitud de la nariz.
13. Sonrisa en la boca.
14. anchura de la boca.
15. Curvatura de la boca.
16. Altura ceja izquierda.
17. Altura ceja derecha.

Ejemplo 20.

Considere la siguiente matriz de datos que contiene los registros del coeficiente intelectual (Ci) de 25 niños, peso al nacer y edad de la madre. Los datos se estandarizaron. Datos originales*: Díaz - Morales (2012, pág. 8).*

ci	peso	edad
0.735	0.516	0.458
0.162	0.469	0.583
0.647	0.694	0.625
0.559	0.573	0.125
0.382	0.280	0.542
1.000	0.373	0.542
0.838	0.641	0.417
0.456	0.608	0.792
0.676	0.444	0.708
0.500	0.330	0.500
0.191	0.359	0.417
0.603	0.447	0.292
0.397	0.223	0.375
0.000	0.230	0.250
0.221	0.516	0.292
0.500	0.579	0.208
0.426	0.406	0.542
0.515	0.570	0.292
0.309	0.599	0.292
0.382	0.364	0.250
0.294	0.418	0.792
0.618	0.584	0.167
0.588	0.363	0.917
0.926	0.000	1.000
0.147	1.000	0.000

Tabla 35. Matriz (25x3) de datos estandarizados entre 0 y 1. Fuente: Elaboración propia.

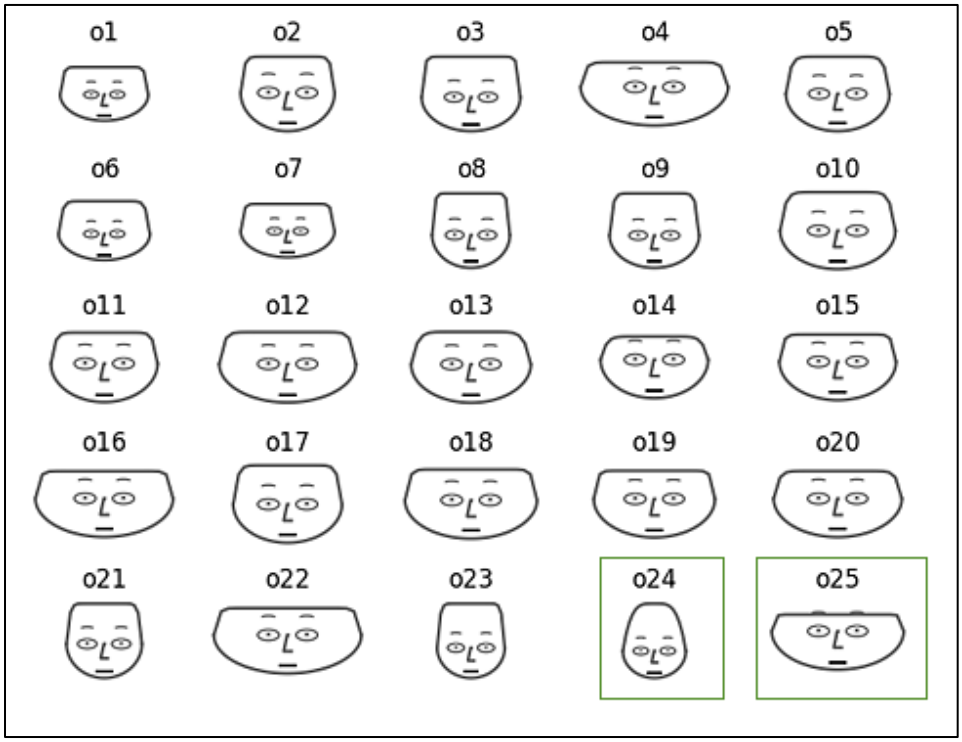


Figura 61. Caras de Chernoff. Fuente: Elaboración propia.

En general se observan ciertas diferencias, las más notorias son la altura de la cara, forma de la frente y altura de las cejas, así, las dos últimas observaciones son posibles datos atípicos.

El rango de variabilidad se establece de manera que la estructura global mantenga las características básicas de una cara. El orden de las variables influye favorablemente para identificar conjuntos de datos mediante caras semejantes, dando a entender que los valores de las variables son "cercanos".

Llevando esta representación a otro nivel, resulta ser optimo al considerar la posición de la cara en un plano cartesiano mediante ACP.

Curvas de Andrew.

Andrews D.F plantea un método que considera un número grande de variables numéricas y que permite identificar agrupaciones en los *n* datos p-dimensional, consiguiendo determinar observaciones atípicas. El método está basado en una *serie finita de Fourier* que representa datos multivariantes en dos dimensiones.

Considere una matriz de datos de tamaño nxp y consideremos $\mathbf{x_i} = (\mathbf{x_{i1}}, \mathbf{x_{i2}}, \mathbf{x_{i3}}, \dots, \mathbf{x_{ip}})$ la observación i-*ésima*. Para $\mathbf{x_i}$ se le asigna una función definida por:

$$f_{\mathbf{x_i}}(t) = \frac{\mathbf{x_{i1}}}{\sqrt{2}} + \mathbf{x_{i2}}\,\mathrm{sen}(t) + \mathbf{x_{i3}}\cos(t) + \mathbf{x_{i4}}\,\mathrm{sen}(2t) + \mathbf{x_{i5}}\cos(2t) + \cdots$$

$$-\boldsymbol{\pi} \leq \mathbf{t} \leq \boldsymbol{\pi}$$

Las variables determinan los coeficientes de la *serie de Fourier*.

La gráfica de la función es una curva suave que corresponde a un dato *p*-dimensional, el valor de cada componente afecta la frecuencia, la amplitud y la periodicidad de la función, dando una representación única para cada observación.

Esta representación de datos tiene la propiedad de preservar la ***media.***

Propiedad 1.

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{x_i}(t).$$

Demostración.

Suponga $\bar{x} = (\bar{x}_1,\ \bar{x}_2,\ \bar{x}_3,\ \dots,\ \bar{x}_p)$ es la observación de medias, entonces mostremos que su representación en curvas de Andrews coincide con la función media de las n curvas.

Note que,

$$f_{\bar{x}}(t) = \frac{\bar{x}_1}{\sqrt{2}} + \bar{x}_2\,\mathrm{sen}(t) + \bar{x}_3\cos(t) + \bar{x}_4\,\mathrm{sen}(2t) + \bar{x}_5\cos(2t) + \cdots ; -\boldsymbol{\pi} \leq \mathbf{t} \leq \boldsymbol{\pi}$$

Donde

$$\bar{x}_j = \frac{x_{1j} + x_{2j} + x_{3j} + \cdots + x_{nj}}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \ \text{ con } \ j = 1, 2, \dots, p$$

Reemplazando $\bar{x}_j$ para todo $j = 1, 2, \dots, p$ en la serie finita de Fourier se tiene que

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\ \frac{x_{i1}}{\sqrt{2}} + x_{i2}\,\mathrm{sen}(t) + x_{i3}\cos(t) + x_{i4}\,\mathrm{sen}(2t) + x_{i5}\cos(2t) + \cdots\right]$$

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[f_{x_i}(t)\right] \qquad \blacksquare$$

La distancia entre dos curvas, se define como

$$\|f_X(t) - f_Y(t)\|_{L_2}^2 := \int_{-\pi}^{\pi} [f_X(t) - f_Y(t)]^2\, dt$$

Otra propiedad interesante es de preservar proporcionalidad en la distancia. Lo que significa que la distancia entre dos funciones es proporcional a la distancia euclidiana entre dos observaciones. Así que, dos observaciones X y Y, con valores similares corresponden a curvas próximas.

Propiedad 2.

$$\|f_X(t) - f_Y(t)\|_{L_2}^2 = \pi\|X - Y\|^2$$

Demostración.

Sean $f_X(t)$ y $f_Y(t)$ series de Fourier tal que:

$$f_{\mathbf{X}}(t) = \frac{x_1}{\sqrt{2}} + x_2\,\mathrm{sen}(t) + x_3\cos(t) + x_4\,\mathrm{sen}(2t) + x_5\cos(2t) + \cdots$$

y

$$f_{\mathbf{Y}}(t) = \frac{y_1}{\sqrt{2}} + y_2\,\mathrm{sen}(t) + y_3\cos(t) + y_4\,\mathrm{sen}(2t) + y_5\cos(2t) + \cdots$$

Donde

$$X = (x_1, x_2, x_3, \ldots, x_p)$$

$$Y = (y_1, y_2, y_3, \ldots, y_p).$$

Supongamos que

$$(f_{\mathbf{X}}(t) - f_{\mathbf{Y}}(t))^2 = (f_{\mathbf{X}}(t))^2 - 2 * f_{\mathbf{X}}(t) * f_{\mathbf{Y}}(t) + (f_{\mathbf{Y}}(t))^2 \qquad (1)$$

Expandiendo a la serie finita de Fourier y mediante propiedades de sumatorias, se hacen tres partes:

Parte 1.

$$(f_X(t))^2 = \left(\frac{x_1}{\sqrt{2}} + x_2\,\mathrm{sen}(t) + x_3\cos(t) + x_4\,\mathrm{sen}(2t) + x_5\cos(2t) + \cdots\right)^2 \qquad (2)$$

Ahora, resolviendo el cuadrado en (2) y asociando términos se tiene que

$$(f_X(t))^2 = \left(\frac{x_1}{\sqrt{2}}\right)^2 + 2\frac{x_1}{\sqrt{2}}\left[\sum_{i=1}^{p} x_{2i}\,\mathrm{sen}(it) + \sum_{i=1}^{p} x_{2i+1}\cos(it)\right] +$$

$$\left[\sum_{i=1}^{p} x_{2i}\,\mathrm{sen}(it) + \sum_{i=1}^{p} x_{2i+1}\cos(it)\right]^2$$

$$= \frac{x_1^2}{2} + \left(2\frac{x_1}{\sqrt{2}}\right)\left[\sum_{i=1}^{p}(x_{2i}\text{sen}(it) + x_{2i+1}\cos(it))\right] +$$

$$\sum_{i=1}^{p}[x_{2i}^2 \sin^2(it) + 2x_{2i}x_{2i+1}\text{sen}(it)\cos(it) + x_{2i+1}^2 \cos^2(it)]$$

Nota: Para el caso $\left(f_Y(t)\right)^2$ es análogo a $\left(f_X(t)\right)^2$.

Parte 2.

$$f_X(t) * f_Y(t) = \left(\frac{x_1}{\sqrt{2}} + x_2\,\text{sen}(t) + x_3\cos(t) + x_4\,\text{sen}(2t) + x_5\cos(2t) + \cdots\right) *$$

$$\left(\frac{y_1}{\sqrt{2}} + y_2\,\text{sen}(t) + y_3\cos(t) + y_4\,\text{sen}(2t) + y_5\cos(2t) + \cdots\right) \quad (3)$$

Resolviendo el producto en (3) se tiene:

$$f_X(t) * f_Y(t) = \frac{x_1y_1}{2} + \frac{x_1y_2}{\sqrt{2}}\text{sen}(t) + \frac{x_1y_3}{\sqrt{2}}\cos(t) + \cdots +$$

$$\frac{x_2y_1}{\sqrt{2}}\text{sen}(t) + \frac{x_2y_2}{1}\sin^2(t) + x_2y_3\text{sen}(t)\cos(t) + x_2y_4\text{sen}(t)\text{sen}(2t) + \cdots$$

Con los resultados de (2) y (3) e integrando en ambos lados en (1), desde [-π , π] se tiene:

$$\int_{-\pi}^{\pi}[f_X(t) - f_Y(t)]^2\,dt = \int_{-\pi}^{\pi}[f_X(t)]^2\,dt - 2\int_{-\pi}^{\pi}[f_X(t) * f_Y(t)]\,dt + \int_{-\pi}^{\pi}[f_Y(t)]^2\,dt \qquad (4)$$

Ahora se desarrollan las tres integrales en (4).

Primera integral:

$$\int_{-\pi}^{\pi}\left(f_X(t)\right)^2 dt = \int_{-\pi}^{\pi}\frac{x_1^2}{2}dt + \left(2\frac{x_1}{\sqrt{2}}\right)\left[\int_{-\pi}^{\pi}\sum_{i=1}^{p}(x_{2i}\text{sen}(it) + x_{2i+1}\cos(it))dt\right] +$$

$$\int_{-\pi}^{\pi} \sum_{i=1}^{p} [x_{2i}^2 \sin^2(it) + 2 * x_{2i} * x_{2i+1} * \text{sen}(it)\cos(it) + x_{2i+1}^2 * \cos^2(it)]\, dt$$

$$= \pi x_1 - 2 * \frac{x_1}{\sqrt{2}}(0) + x_{2i}^2(0) + \cdots + \sum_{i=1}^{p} \pi x_{2i}^2 + \sum_{i=1}^{p} \pi x_{2i+1}^2$$

$$= \pi \sum_{i=1}^{p} x_i^2$$

Segunda integral:

$$\int_{-\pi}^{\pi} (f_X(t) * f_Y(t))dt$$

$$= \int_{-\pi}^{\pi} \left(\frac{x_1 y_1}{2} + \frac{x_1 y_2}{\sqrt{2}} \text{sen}(t) + \frac{x_1 y_3}{\sqrt{2}} \cos(t) + \cdots + \frac{x_2 y_1}{\sqrt{2}} \text{sen}(t) + \frac{x_2 y_2}{\sqrt{2}} \sin^2(t) + x_2 y_3 \text{sen}(t)\cos(t) + x_2 y_4 \text{sen}(t)\text{sen}(2t) + \cdots \right) dt$$

Por la linealidad de la integral definida y usando el teorema fundamental del cálculo se tiene que:

$$= \pi x_1 y_1 + \frac{x_1 y_1}{\sqrt{2}}(0) + \cdots + 0 + \sum_{i=1}^{p} \pi x_{2i}\, y_{2i} + \sum_{i=1}^{p} \pi x_{2i+1}\, y_{2i+1}$$

$$= \pi \sum_{i=1}^{p} x_i\, y_i$$

Tercera integral: análoga a la primera integral.

$$\int_{-\pi}^{\pi} \left(f_{\mathbf{Y}}(t)\right)^2 dt = \pi \sum_{i=1}^{p} y_i^2$$

Por lo tanto, se tiene que:

$$\int_{-\pi}^{\pi} [f_X(t) - f_Y(t)]^2 \, dt = \pi \sum_{i=1}^{p} x_i^2 - 2\pi \sum_{i=1}^{p} x_i \, y_i + \pi \sum_{i=1}^{p} y_i^2$$

$$= \pi \sum_{i=1}^{p} (x_i^2 - 2x_i y_i + y_i^2)$$

$$= \pi \sum_{i=1}^{p} (x_i - y_i)^2$$

$$= \pi \|X - Y\|^2 \qquad \blacksquare$$

Esta herramienta matemática para representar datos en altas dimensiones presenta ciertas consideraciones como el orden de las variables, lo cual puede mejorar la visualización.

Por ejemplo, sea $A = (1,0,0)$, $B = (0,1,0)$, $C = (0,0,1)$ tres observaciones de dimensión tres.

Las funciones correspondientes son $f_A(t) = \frac{1}{\sqrt{2}}$, $f_B(t) = \text{sen}(t)$, $f_C(t) = \cos(t)$.

Estas curvas son bastante distintas, ya que las tres observaciones son los vectores unitarios de $\mathbb{R}^3$. Por tanto, el orden de las variables juega un papel importante para la interpretación, las últimas variables tendrán una pequeña contribución a la curva. Por ello se recomienda que las primeras columnas de la matriz de datos estén asociadas a las variables más importantes en la investigación, se sugiere utilizar el orden dado por el método de ACP.

Otro aspecto importante es estandarizar para evitar que algunas variables dominen a otras y preservar la proporcionalidad en las distancias de las observaciones. Considere la matriz de datos del ***ejemplo 20.***

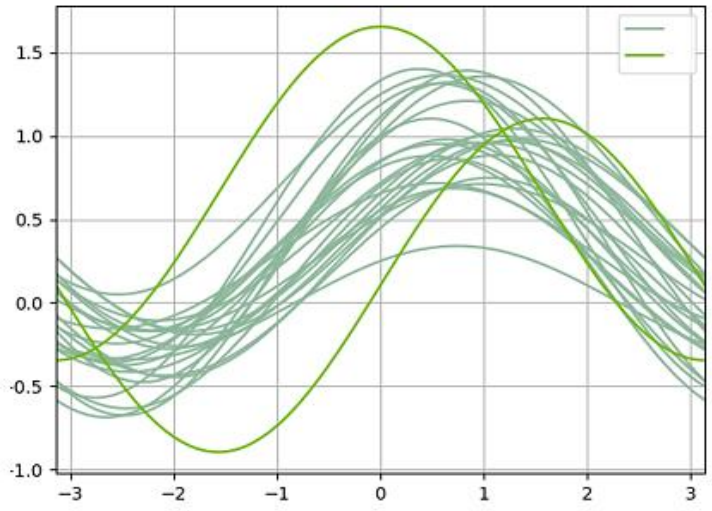


Figura 62. Curvas de Andrews para el ejemplo 20. Fuente: Elaboración propia.

Suponga que se considera una nueva variable con modalidades: Menor a 30 años y mayor o igual a 30 años. Ahora los datos son de dimensión 4 y la visualización es:

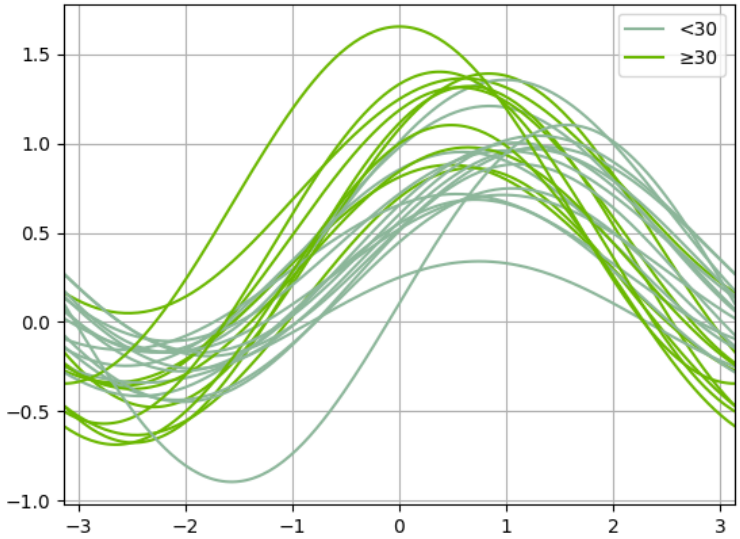


Figura 63. Curvas de Andrews considerando la nueva variable cualitativa. Fuente: Elaboración propia.

Cartogramas.

Cuando se tienen datos asociados a la localización de un territorio, el tratamiento de los datos es especial y para este tipo de datos se encarga el *análisis de datos espaciales.* La representación gráfica para este tipo de datos son los cartogramas que pueden entenderse como ¨mapas¨ de un territorio que representa información de elementos observables localizables. Los cartogramas pueden presentar datos de una o más dimensiones de forma directa o codificada.

Unidades geográficas: Son unidades estadísticas con la particularidad de estar sometidas espacialmente a un determinado territorio.

Es común hacer delimitaciones con el fin de facilitar la identificación del territorio y de esa forma facilitar la localización de las unidades geográficas. Por ejemplo, los mapas geográficos o mapas de división territorial asignan a cada unidad geográfica un área (variable cuantitativa) proporcional al área de la superficie real. Algunas de las variables de interés que se miden a estas unidades son variables geofísicas, variables socio-demográficos, variables económicas, etc.

En la representación de los datos puede incluir colores, sombreados, tramas, glifos para indicar proporciones, movimientos, desplazamientos, rutas o incluso otros gráficos que identifican la distribución geográfica de los datos.

Los cartogramas pueden identificar datos puntuales mediante (x, y) en el mapa como iglesias, sitios turísticos, etc. Identificar una comunidad es una ubicación zonal y no puntual.

El análisis de datos espaciales está constituido por diversas técnicas que permiten explorar los datos, detectar patrones, formular hipótesis que se refieren a los casos o subconjuntos de casos que son inusuales dada su localización en el mapa y responder preguntas como: ¿Dónde se encuentran en el mapa los casos atípicos observados en el histograma? o para identificar asociaciones.

Ejemplo 21.

Número de casos de coronavirus para diferentes países.

	Casos	Muertes
EE.UU.	824.065	44.996
España	204.178	21.282
Italia	183.957	24.648
Alemania	148.453	5.086
Reino Unido	129.044	17.337
Francia	117.324	20.796
Turquía	95.591	2.259
Irán	84.802	5.297
China	83.864	4.636
Rusia	52.763	456
Brasil	43.368	2.761
Bélgica	40.956	5.998
Canadá	39.405	1.915
Holanda	34.139	3.916
Suiza	28.063	1.478
Portugal	21.379	762

Tabla 36. Propagación global de coronavirus, 22 abril 2020. Fuente: https://www.bbc.com/mundo/noticias-51693616

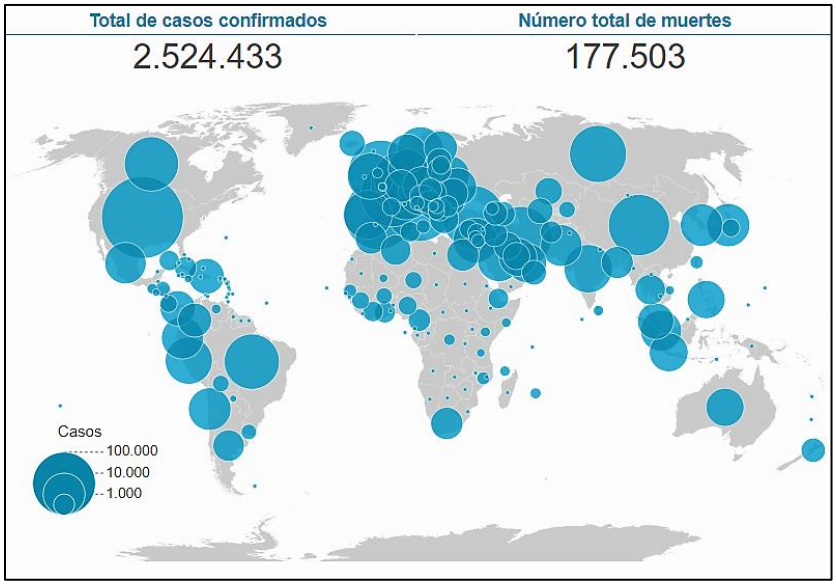


Figura 64. Cartograma con glifo circular para casos de coronavirus. Fuente: https://www.bbc.com/mundo/noticias-51693616

La variable cuantitativa está representada por el área de círculos de colores situados sobre el país al que pertenecen, siendo el área del círculo proporcional al valor de la variable.

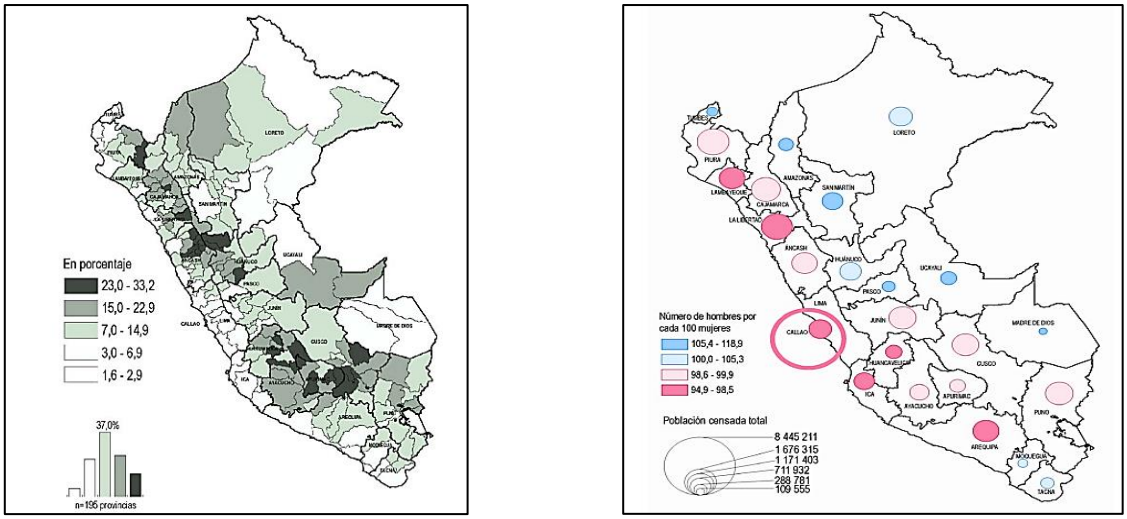


Figura 65. Izquierda: Tasa de analfabetismo, según provincia, 2007, Perú. Derecha: índice de masculinidad, población, según departamento, 2007, Perú. Fuente: Instituto nacional de estadística e informática (censos de población y vivienda, 2007).

Series temporales.

Las series temporales presenta la evolución en el tiempo de una o varias características para un fenómeno de interés que generalmente son cuantitativas. Las mediciones que se realizan dependen de un continuo: El tiempo, naturalmente surge la pregunta: ¿Podemos registrar para cada tiempo t la información que se va a estudiar? el tiempo es un parámetro continuo y por tanto las mediciones conforman un conjunto infinito imposible de almacenar.

Realizar una medición para variables de carácter continuo cada segundo, hora o día, proporciona una cantidad finita de observaciones, y aunque no se tenga registro de cada instante de tiempo t se sabe que los datos están definidos para los intervalos que hay entre cada uno de ellos. Por tanto, la variable tiempo toma valores discretos y el dato como tal es bidimensional.

Los datos son representados en un despliegue bidimensional de forma discreta, para facilitar el comportamiento de la variable de interés se unen las observaciones, haciendo que la gráfica sea una consecución de semirrectas interpoladas entre cada par de observaciones.

Se define ***serie temporal*** como una secuencia de *n* observaciones (n *datos*) ordenados y equidistantes cronológicamente sobre una característica o sobre varias características de ***una unidad*** observable en diferentes momentos.

En este texto se usará la siguiente notación.

Notación serie temporal univariable

$$y_1, y_2, y_3, \dots, y_t, \dots, y_n.$$

Donde,

y_t es la observacion $t - \acute{e}sima$ de la serie temporal.

Las n observaciones pueden registrarse en un vector columna:

$$\mathbf{y} = [y_1, y_2, y_3, \dots, y_n]^{\mathbf{T}} \text{ de orden n}\mathbf{x}1.$$

La serie temporal multivariable

$$\mathbf{y_1}, \mathbf{y_2}, \mathbf{y_3}, \dots, \mathbf{y_t}, \dots, \mathbf{y_n}.$$

Donde,

$\mathbf{y_t} = [y_{t1}, y_{t2}, y_{t3}, \dots, y_{tp}]^{\mathbf{T}}$ es el $t - \text{ésimo}$ vector de observación de tamaño $p \geq 2$.

Los n vectores de observaciones pueden registrarse dando lugar a una matriz **M** de tamaño nxp.

$$\mathbf{M} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

Donde,

y_{tj} es la observación en el momento t de la caracteristica j con $1 \leq j \leq p$

En *el análisis de series temporales* es común suavizar las series, brindando herramientas ventajosas, por ejemplo, conocer información entre cada par de puntos sin aumentar las observaciones o para las mediciones que no todas fueron tomadas en el mismo instante. Además, permite predecir el comportamiento de la variable mediante distintos modelos matemáticos para actuar con antelación y tomar decisiones.

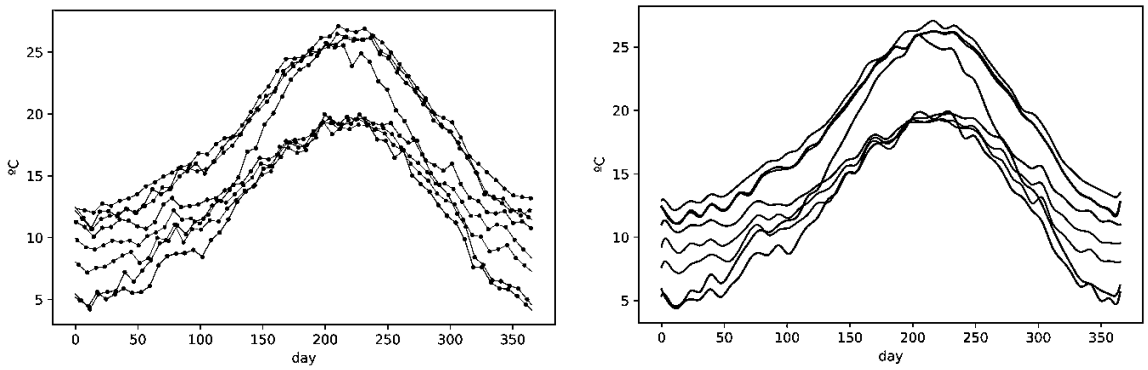


Figura 66. Representación discreta y representación en bases de funciones de la media de la temperatura semanal entre 1980-2009 realizadas por ocho estaciones meteorológicas en España. Fuente: Pablo P. Manso.

Una serie temporal se puede descomponer en las siguientes componentes:

- **i.** ***Tendencia*** (lineal o no lineal): La ***tendencia*** es el movimiento de los datos hacia arriba o hacia abajo a largo plazo en el tiempo.
- **ii.** ***Estacionalidad***: La ***estacionalidad*** se identifica como el patrón que muestran los datos en intervalos regulares.
- **iii.** ***Variación cíclica***: se identifica como el patrón que muestra los datos en ciertos intervalos de tiempo, es decir, que refleja las fluctuaciones periódicamente, pero no necesariamente regular. La ***diferencia entre estacionalidad y variación cíclica*** es que la estacionalidad ocurre a intervalos de tiempo conocidos y los intervalos de tiempo en los que ocurre la variación cíclica no se pueden determinar con precisión.
- **iv.** ***Variación irregular***: El comportamiento de los datos son esporádicos que no muestran una periocidad reconocible. Son resultado de hechos no previsibles, pero identificables a posteriori como huelgas, catástrofes, etc.

La asociación de estas componentes hace que las series de tiempo puedan ser aditivas o multiplicativas. La *serie de tiempo aditiva* es igual a la suma de las 4 componentes. La *serie de tiempo multiplicativa* es igual al producto de las 4 componentes.

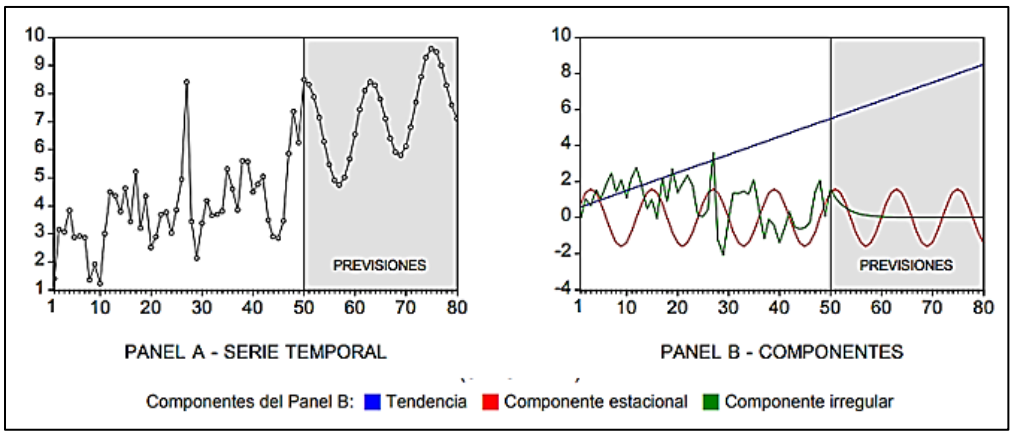


Figura 67. Descomposición de una serie temporal y su previsión. Fuente: Introducción al análisis de series temporales. José Alberto Mauricio.

Representar variables con escalas diferentes bajo la misma referencia de tiempo es común hacerlo por separado para observar la relación de cada una de ellas con el tiempo. Para la superposición los datos deben estar estandarizados previamente.

Ejemplo: Datos diarios en un periodo de ocho años de una estación meteorológica ubicada en *Aranjuez (Madrid).*

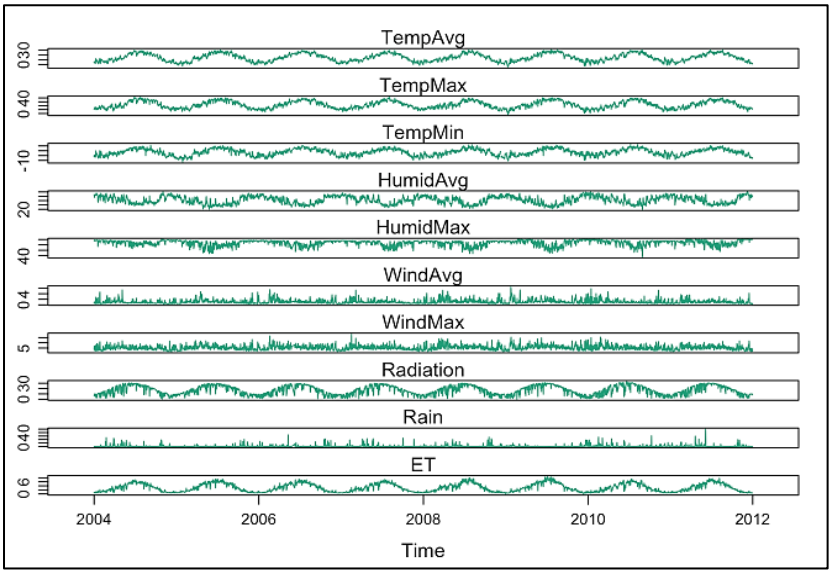


Figura 68. Representación gráfica de series temporales: 10 variables que miden condiciones meteorológicas. Fuente:https://rstudio-pubs-static.s3.amazonaws.com/587089_84766cb0bd5c4c7790469a3f4aebd51c.html

Otra forma de representar múltiples series temporales de igual escala es mediante un sistema de celdas con su respectivo tono. El objetivo es mostrar el comportamiento de la colección de variables como un todo y así examinar cómo cambia un gran número de series temporales a lo largo del tiempo, permitiendo analizarlas por separado y entre sí, observando fácilmente los comportamientos extraordinarios, los patrones o clasificación por grupos. La referencia es el promedio diario a largo plazo.

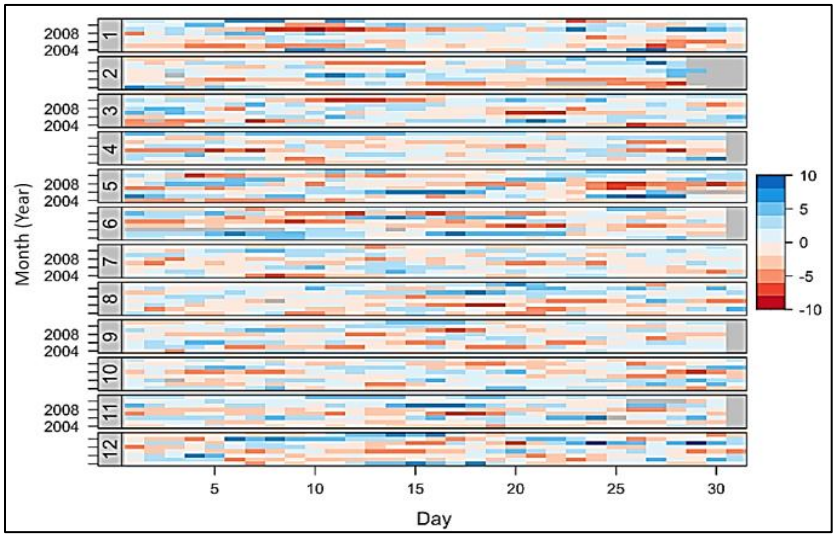


Figura 69. Temperatura: 96 series temporales con referencia el promedio diario a largo plazo. Fuente: https://rstudio-pubs-static.s3.amazonaws.com/587089_84766cb0bd5c4c7790469a3f4aebd51c.html

También, para datos temporales el tiempo puede concebirse como una variable de agrupamiento, condicionamiento o complementaria.

El tiempo como una variable de agrupamiento.

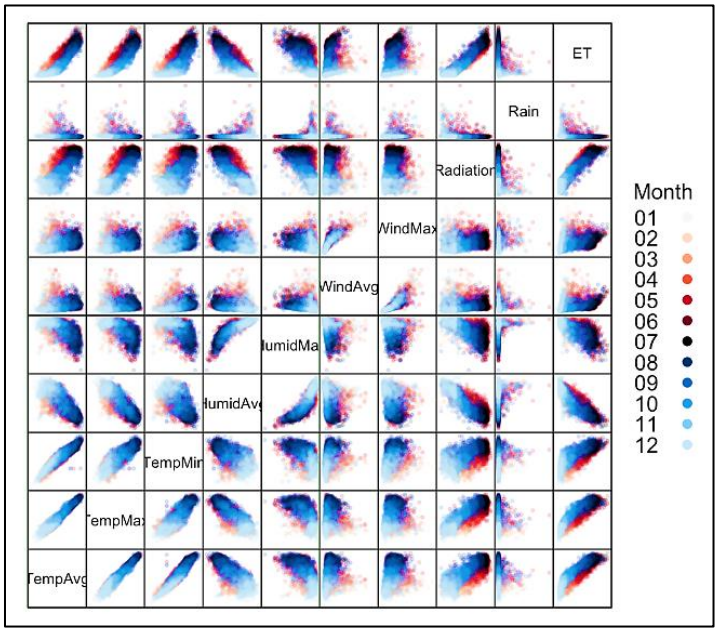


Figura 70. Diagrama de escalera para series temporales: Relación de 10 variables según el mes. Fuente: https://rstudio-pubs-static.s3.amazonaws.com/587089_84766cb0bd5c4c7790469a3f4aebd51c.html

El tiempo como variable condicionante.

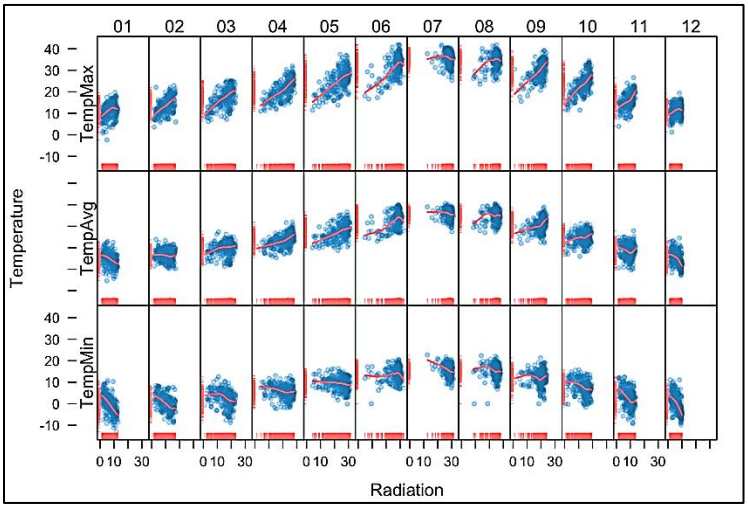


Figura 71. Relación de 4 variables según el mes. Fuente: https://rstudio-pubs-static.s3.amazonaws.com/587089_84766cb0bd5c4c7790469a3f4aebd51c.html

El tiempo como una variable complementaria.

Evolución (2000 a 2014) de la relación para: El ingreso nacional bruto (INB) y las emisiones de dióxido de carbono (CO2). Datos abiertos del Banco Mundial.

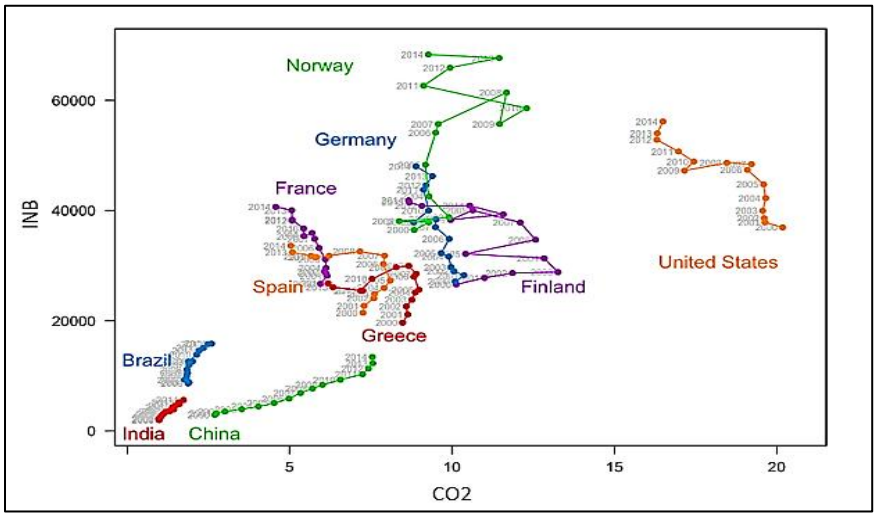


Figura 72. Representando 20 series temporales. Fuente: https://rstudio-pubs-static.s3.amazonaws.com/587089_84766cb0bd5c4c7790469a3f4aebd51c.html

Errores en la construcción de un gráfico estadístico.

Considerando los elementos que debe tener todo gráfico estadístico como cuerpo, titulo, fuente, etc. Generalmente, la figura principal o cuerpo, es la que presenta mayor cantidad de falencias en cuanto a las proporciones.

Desproporción en la escala de los ejes: ejes "cargados", amplitud de ángulos, radios, coherencia en la escala de colores.

Desproporción en la figura principal: pictogramas, barras, sectores, líneas, área, volumen.

También los errores están dados por falta de identificación de la naturaleza de las variables, la dimensionalidad de los datos o cantidad de datos. Construir un histograma con pocos datos o un histograma para una variable cuantitativa discreta, variables de escala ordinal como numéricas.

Ejemplos de representaciones gráficas con falencias (en medios de comunicación).

El texto de *estadística en acción* muestra un pictograma tomado del periódico *El País, 22 de noviembre de 2005, pág. 15*. No sólo no ayuda a entender la información contenida, sino que más bien complica la comprensión de los datos que, además, no mantienen ninguna proporcionalidad con la imagen representada. Se propone una forma de representar usando barras y no un pictograma por cuestiones de claridad.

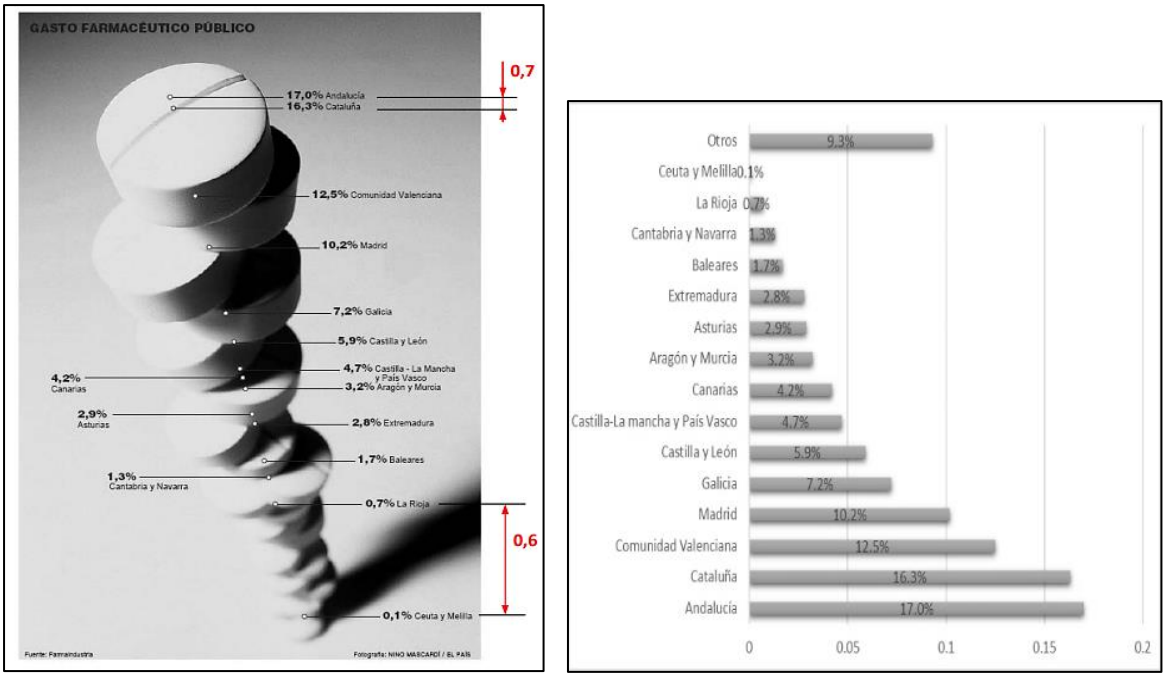


Figura 73. Gasto farmacéutico- España (incorrecto y propuesto). Fuente: Estadística en acción- Propia.

Otro ejemplo, propuesto en el texto *estadística en acción,* son las comparaciones de actividades relacionadas con el "consumo cultural" de jóvenes y adultos. Lado izquierdo muestra "lo que no hacen los jóvenes" es realmente difícil de interpretar. ¿Quién va más al teatro, los jóvenes o los adultos? Estaría más claro si lo hicieran como en la parte inferior, explicando qué es lo que sí hacen.

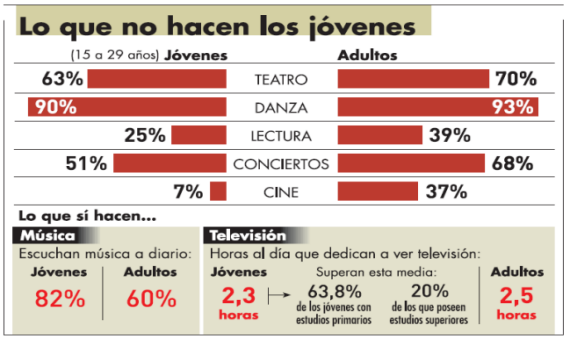


Figura 74. "Consumo cultural" de jóvenes y adultos (erróneo). Fuente: Estadística en acción.

La siguiente representación, no ayuda a entender, ni a captar con más rapidez el significado de los datos, sino que el hecho de mezclar datos y dibujos sin ningún sentido complica la lectura y la comprensión de la información.

Figura 75. Pictograma (erróneo). Fuente: Estadística en acción.

El texto *Estadística en acción* propone el siguiente ejemplo donde las escalas mantengan su proporcionalidad. La falta de proporcionalidad puede provocar que un pequeño incremento parezca mayor de lo que realmente es, o disimular la importancia de otro mayor. En la prensa, seguramente esta práctica está más orientada a buscar imágenes creativas que a confundir al lector, pero en anuncios publicitarios esta deformación del gráfico puede tener interés más allá de la pura estética. Se muestra la evolución del tipo oficial fijado por el Banco Central Europeo (BCE). Sólo se han representado aquellos meses en los cuales el BCE aumentó o disminuyó los tipos de interés. Una representación como la propuesta es más fiel a la realidad: se muestran las variaciones únicamente cuando se producen, se representa todo el periodo estudiado y los cambios, en vez de dibujarlos progresivos, se hacen puntuales (tal como ocurre con los tipos de interés).

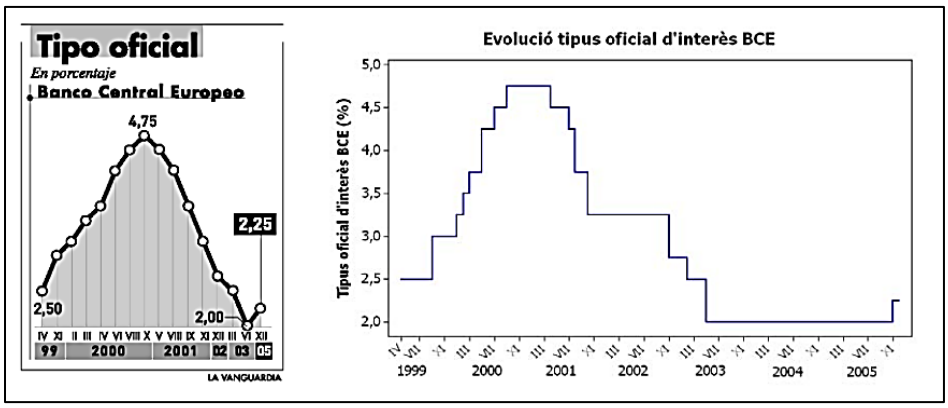


Figura 76. Evolución tipo oficial BCE (Gráfico incorrecto izquierdo-correcto derecho). Fuente: Estadística en acción.

Todas las representaciones deben realizarse con los mismos criterios que se quieren comparar. En la siguiente figura del texto *estadística en acción,* se hace la comparación de las audiencias de televisión en Cataluña y en España de un conjunto de partidos de fútbol. De acuerdo con los datos que aparecen a la figura, en porcentaje, las audiencias son mayores en Cataluña que en el resto de España cuando juega el Barcelona, y no es así cuando juega el Real Madrid. El gráfico compara frecuencia absoluta y, aunque también da el dato del porcentaje, del gráfico propiamente dicho no se extrae, de forma fácil, ninguna información relevante. A la derecha se incluye una representación gráfica que compara los porcentajes de audiencia donde se ve claramente que cuando juega el Barça la audiencia es mayor en Cataluña en términos relativos. Al hacer comparaciones de diferentes tamaños de datos es siempre mejor hacerlo en porcentajes.

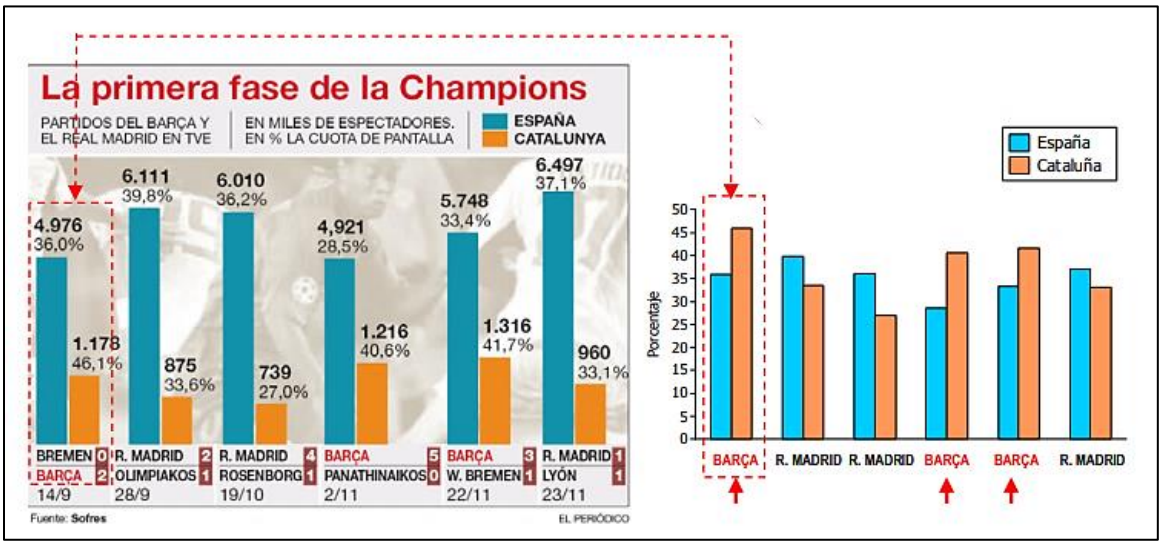


Figura 77. Comparaciones de audiencia partidos de futbol-Gráfico incorrecto izquierdo-correcto derecho). Fuente: Estadística en acción.

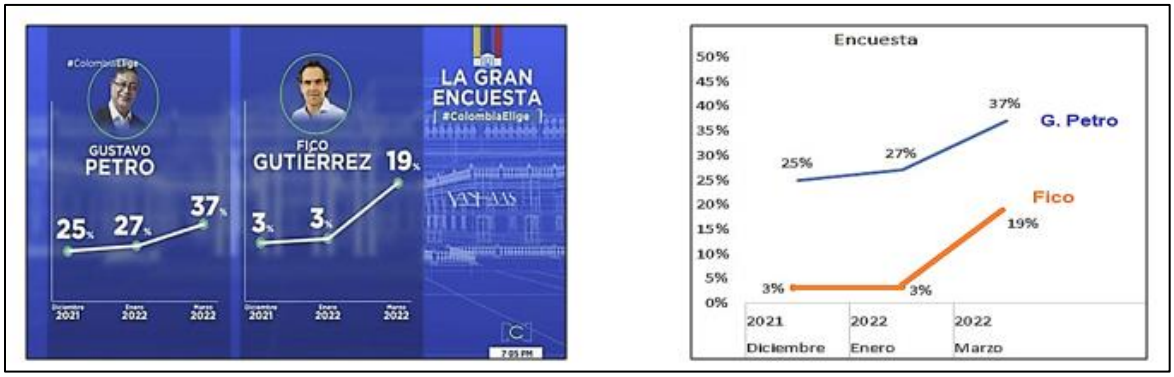


Figura 78. Comparación en la evolución de favorabilidad elecciones presidenciales-2022. (Gráfico incorrecto izquierdo-correcto derecho). Fuente: RCN-Propia.

Aparentemente el candidato Gutiérrez subió la tendencia superando al candidato Petro. El error está en la comparación con dos escalas distintas los dos candidatos colombianos.

Formas de saturación en los gráficos.

Consiste en evidenciar gráficos cargados o saturados de información que los convierte en representaciones no idóneas, además, son difíciles de interpretar y generan confusión al lector. En sí, no son errores, dado que las fallas no son técnicas, pero se convierten en falencias, ya que no es lo idóneo en cuanto a una buena visualización.

La naturaleza: si consideramos una sola variable es importante identificar su naturaleza. Si es cualitativa, los gráficos a considerar son: barras, sectores, pictogramas, Pareto, radar, áreas, anillos. Pero, si es cuantitativa, entonces los gráficos a considerar son puntos, líneas, histogramas, cajas, tallos y hojas. Ahora si son consideramos dos variables, entonces se pueden usar combinaciones de los anteriores gráficos con el fin de establecer comparaciones, identificar distribuciones y patrones, así, los gráficos como el diagrama de dispersión, estereograma, laminas o gráficos codificados son los indicados. Cuando se tiene tres o más variables, se tienen demasiadas combinaciones según la naturaleza de cada variable, esto indica que hacer gráficos de esta manera se convierta en un proceso tedioso y poco optimo; entonces aparecen técnicas más avanzadas para representar múltiples variables.

El número de variables:

Una característica importante en los gráficos es identificar la cantidad de variables que pueden soportar sin colapsar visualmente.

Por ejemplo, para el gráfico de sectores, es posible identificar tres variables, mediante la amplitud, color, textura, pero el número de variables optimo es una.

Para el de barras, es posible identificar agregar líneas para identificar tendencias, la comparabilidad de las barras, el tramado o textura de las barras, en fin, son varias las variables que se pueden resumir usando como base el diagrama de barras. De este modo es recomendable representar solo dos variables. Además, un factor importante cuando se usas barras es la cantidad de modalidades a considerar.

Los gráficos de rectángulos son una buena opción para comparar proporciones dentro de la jerarquía. Además, es óptimo para comparar (no más de dos) conjuntos de datos conservando un orden jerárquico.

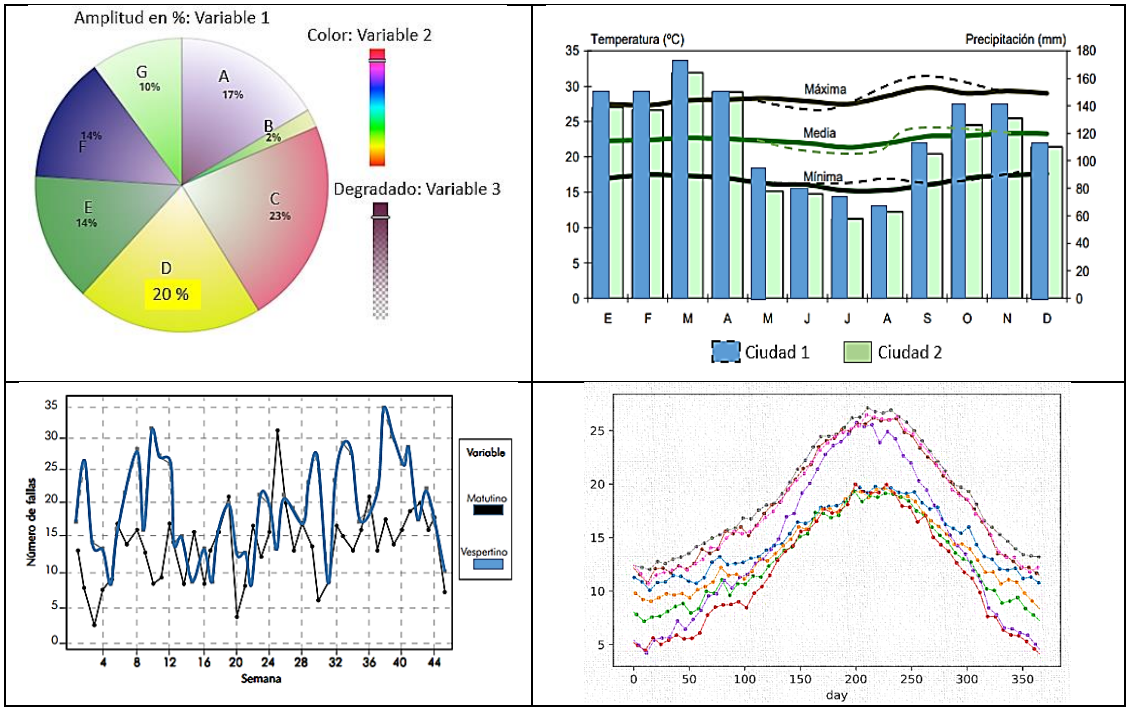


Figura 79. Gráficos saturados por la cantidad de variables. Fuente: Elaboración propia.

Los gráficos de líneas también son útiles para datos de series temporales y someter un mismo sistema cartesiano más cuatro series temporales, puede convertirse en una representación cargada. Si lo que se desea es analizar los elementos de una serie temporal, lo óptimo es representar cada serie por separado; Pero, si lo que buscamos es patrones en conjunto de variables, entonces lo conveniente es representarlas en mismo sistema coordenado.

Las modalidades:

En general, variables con demasiadas modalidades hacen que los gráficos por muy simple que sean tienden a saturase, lo óptimo es encapsular modalidades que son de bajas frecuencias bajo una nueva modalidad que usualmente se cataloga como “Otros”. Otra solución es codificación de variables. Por ejemplo, gráfico de radar, que agrega demasiadas modalidades, puede saturase y aún más si se considera una segunda o tercera variable. Lo óptimo es una representación mediante radares comparativos y reducir las modalidades, pero, no menor a tres, ya que el radar no se puede construir.

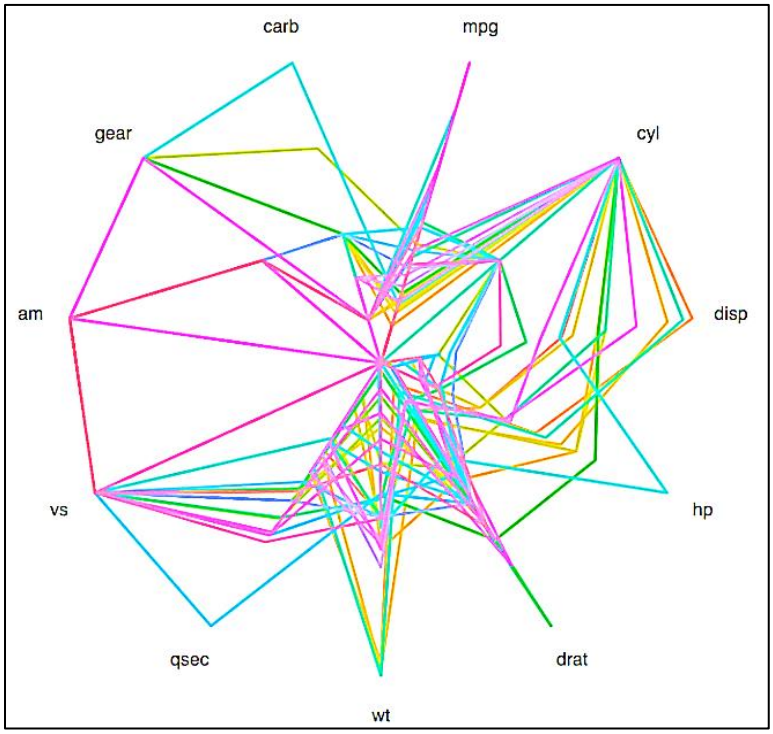


Figura 80. Gráfico de radar saturado. Fuente: Elaboración propia.

Puede utilizar un gráfico de líneas para resumir variables categóricas, en cuyo caso es similar a un gráfico de barras y son óptimos cuando el número de modalidades es grande.

Cantidad de datos:

Es un factor influyente que se debe tener en cuenta para no cargar un gráfico. Por ejemplo, en general, un diagrama de dispersión tridimensional es algo complicado de identificar la distribución en conjunto de las tres variables, si codifica tres variables cualitativas, esto hace que colapsar si se tiene una cantidad de datos considerable y por la perspectiva tridimensional lo hace más difícil de entender. Entonces, lo conveniente es usarlo cuando son pocos datos, codificar una sola variable o pasar a una dispersión bidimensional. Lo óptimo cuando se tiene datos multidimensionales y se desea estables patrones en conjunto de variables, es usar curvas de Andrews, estrellas, caras de Chernoff con pocos datos. estrellas. Otro aspecto importante es el escalado de los datos e identificar variables más relevantes.

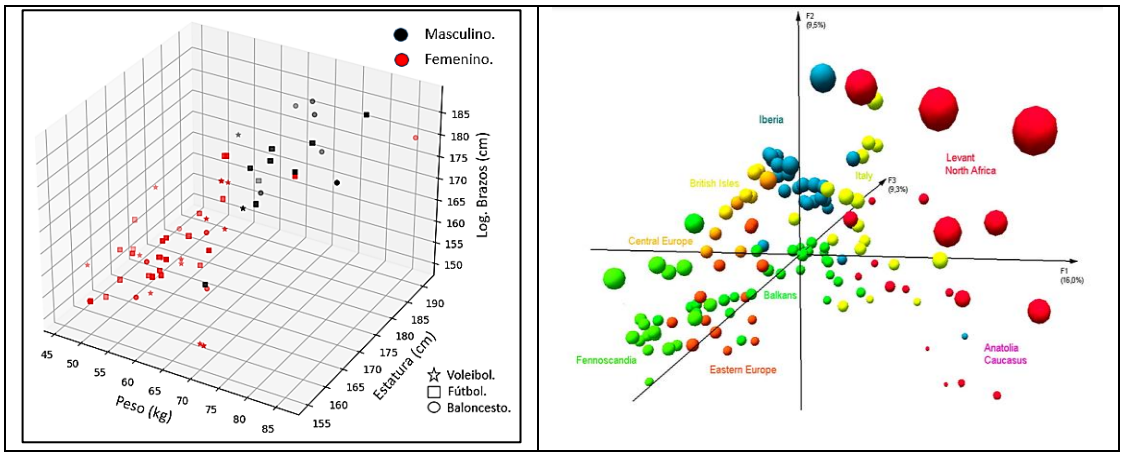


Figura 81. Gráficos saturados por la cantidad de datos. Fuente: Francalacci P, Morelli L, Useli A y Sanna D.

Lo dispersión tridimensional pueden guardar varias características, esto es tedioso para la lectura, lo ideal es considerar dispersiones bidimensionales con a lo sumo dos variables codificadas.

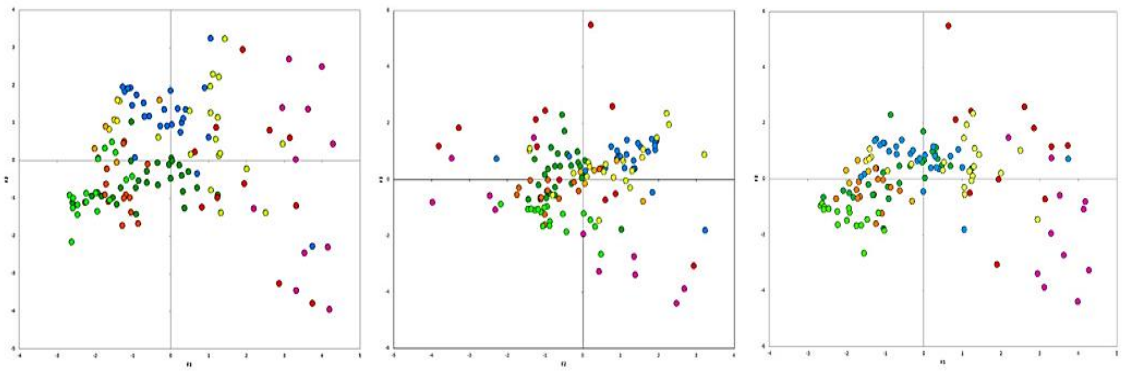


Figura 82. Dispersión con una variable codificada. Fuente: Francalacci P, Morelli L, Useli A y Sanna D.

Para los datos no están escalados, se corre el riesgo de variables dominantes, en este ejemplo, si hay una clara evidencia de dos datos atípicos. El otro caso se tiene una mala estandarización, por lo cual no se evidencia patrones en la visualización.

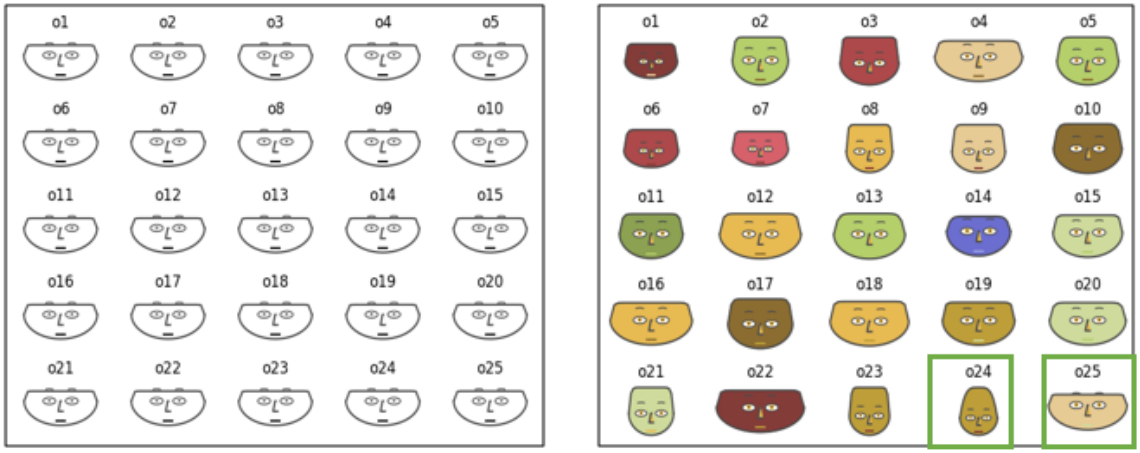


Figura 83. Caras de Chernoff para datos no estandarizados y estandarizados. El color identifica agrupamientos. Fuente: Elaboración propia.

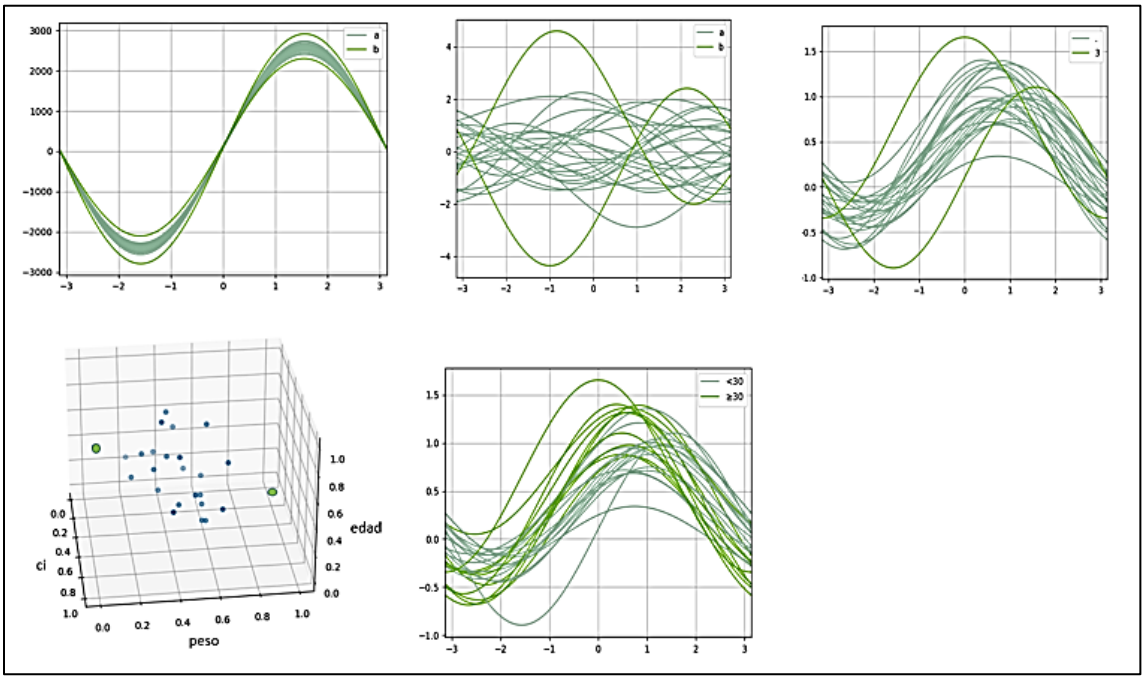


Figura 84. Curvas de Andrews y dispersión escalados con datos atípicos. Fuente: Elaboración propia.

Condiciones para una adecuada representación gráfica de datos.

Partiendo de que los datos están listos y son de calidad, se debe considera la dimensión de la matriz de datos, es decir la cantidad de datos y el número de variables, la naturaleza de las variables, las modalidades o localizaciones. Además, determinar preguntas puntuales para cumplir los objetivos de los gráficos.

Elegir la ayuda visual incorrecta o utilizar de forma predeterminada el tipo más común de visualización de datos podría confundir al espectador o dar lugar a una interpretación errónea de los datos.

Un gráfico ayuda a comparar diferentes valores como frecuencias o algunos estadísticos, comprender cómo las diferentes partes impactan el todo o analizar tendencias, reconocer datos que se desvían e identificar las relaciones entre variables.

Los diferentes tipos de gráficos utilizan diferentes tipos de datos: cuantitativos o cualitativos. Además, la dimensionalidad de los datos es una condición para elegir los gráficos.

Algunas preguntas para encontrar el tipo de gráfico correcto.

¿Quiere comparar valores?

Los gráficos de frecuencias son perfectos para comparar uno o varios conjuntos de datos y pueden mostrar fácilmente los valores bajos y altos de los conjuntos de datos, comparar distribuciones. Para crear gráficos comparativos, utilice estos tipos de gráficos:

Barras, sectores, anillos, líneas, histogramas, polígonos de frecuencia, tallos y hojas, cajas y alambres, rectángulos o áreas.

¿Quiere mostrar la composición?

Utilice este tipo de gráfico para mostrar cómo las partes individuales forman el todo. Por ejemplo, frecuencias condicionales para mostrar la composición: Anillos, barra apilada, Área, rectángulos o áreas.

¿Quiere entender la distribución de tus datos?

Los gráficos de distribución le ayudan a comprender los valores o datos atípicos, la tendencia, variabilidad, sesgos en los datos. Gráficos para mostrar la distribución: Dispersiones, línea, barras, áreas, cajas y alambres, tallos y hojas, histograma, estereograma, láminas.

¿Quiere comprender tendencias en su conjunto de datos?

Los tipos de gráficos óptimos que funcionan bien para comprender el comportamiento de un conjunto de datos durante un tiempo específico: Línea, barras, radar.

¿Quiere comprender mejor la relación entre variables?

Los gráficos de relaciones pueden mostrar cómo una variable se relaciona con una o muchas variables diferentes: Dispersiones, anillos, líneas.

¿Quiere comprender grupos y patrones en conjunto de variables?

Diagrama de escalera, curvas de Andrews, caras de Chernoff, estrellas, líneas paralelas o radar.

Por otra parte, si los datos se someten a una ubicación, los cartogramas son los óptimos que combinados con gráficos simples pueden ser de gran ayuda a comprender datos multidimensionales.

Conclusiones.

En el texto se recopilaron distintas representaciones gráficas que se trabajan en la estadística para datos de baja dimensión o altas dimensiones y se destaca la importancia que tienen estas visualizaciones para los procesos que tiene el análisis de datos, ayudando a tener claridad en el uso adecuado de las representaciones gráficas para conocer y analizar mejor la realidad apoyando la toma de decisiones. En la actualidad podemos encontrar varias aplicaciones de estos gráficos, en el ámbito de percepción de consumidores, calidad de servicio, actitud hacia políticas de protección del medio ambiente, clasificación de productos, ensayos clínicos, ciencias sociales, etc.

Un problema básico de los datos es la dimensionalidad, las representaciones gráficas para comprender los datos son mucho más difíciles que los análisis univariado, porque su interpretación correcta depende del conocimiento de los procedimientos y conceptos para su construcción. Por ejemplo, saber conceptos de la geometría multidimensional con el fin de lograr la interpretación correcta de las salidas gráficas y de los índices numéricos que las acompañan.

Por ende, se presentan herramientas ingeniosas para solventar esta dificultad con métodos gráficos que van desde representaciones básicas como histogramas, barras, cajas y alambres, dispersiones, gráficos comparativos, hasta caras de Chernoff, curvas de Andrew, gráfico de estrellas, diagrama de escalera, etc. Además, los datos no tienen única forma de visualizarlos, favoreciendo una investigación, ya que alguna visualización presenta información que en otras visualizaciones no es evidente.

Errores comunes en la representación gráfica

Utilizar polígonos de frecuencias con variables cualitativas, o diagrama de barras para representar datos que debieran representarse en un diagrama de dispersión, elegir una escala inadecuada, omitir escalas en alguno de los ejes, interpretar los histogramas como gráficos de barras o de dos variables, histogramas construidos con la frecuencia absoluta y no con la función empírica de densidad, confusión de la naturaleza de las variables y sus valores, o información innecesaria; Para representaciones multivariadas, la dificulta es la dimensionalidad de los datos y una errada estandarización, además de involucrar demasiadas variables llevando a saturar la visualización.

Algunas conclusiones particulares:

- En el contexto de las variables categóricas, los diagramas para las frecuencias acumuladas no tienen sentido.
- Los datos pierden valor si están referidas a un concepto ambiguo. Es más grave cuando los términos utilizados sugieren un significado, pero en realidad tienen otro.
- No ignorar la variabilidad, en muchos casos no es suficiente con la media para describir un conjunto de dato.
- Necesidad de hacer una representación gráfica de los datos cuando esta ayuda a su lectura e interpretación, no cuando la dificulta.
- En cuanto a la presencia de datos multivariados atípicos en un conjunto de datos, tratarlos es más complicado que en el caso univariado. Uno ***primer*** problema es que

estos datos pueden distorsionar no sólo las medidas de localización y escala sino las de asociación u orientación. Un ***segundo*** problema es que es más difícil caracterizar y descubrir que un dato univariado atípico. Un ***tercer*** problema es que un dato multivariado por el hecho mismo de ser un vector conformado por datos univariados, la atipicidad puede deberse a un error extremo en alguna de sus componentes o a la ocurrencia de errores sistemáticos en varias o todas sus componentes.

- Estandarizar para eliminar el efecto de la escala de medición y establecer comparaciones. Este tipo de gráficas comparativas facilitan la lectura sobre centralidad, variabilidad, simetría, presencia de observaciones atípicas e incluso asociación entre variables, en un conjunto de datos.
- Codificación de variables para representar variables cuantitativas o cualitativas, permitiendo involucrar mayor información en una representación gráfica.

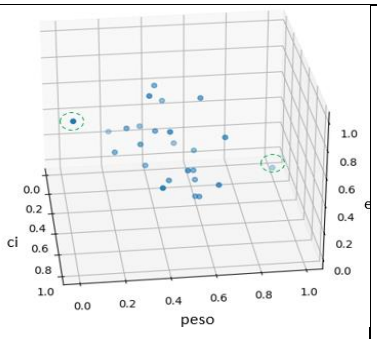
Finalmente se puede decir que, para una óptima representación, es fundamental el conocimiento de algunos conceptos y gráficos básicos de la estadística descriptiva. Esto con el fin de tener una mayor comprensión y dominio de las representaciones de gráficos de datos multidimensionales, aunque son técnicas más avanzadas, muchas de sus construcciones se componen de gráficos básicos. Otro aporte importante de este trabajo es brindar cierto nivel de comprensión que requiere tanto investigadores como lectores y que tengan la capacidad de identificar representaciones erróneas que en muchos casos son hechas de manera intencional.

Para el cumplimiento de los objetivos, la información y datos fueron recopilados en distintas fuentes, como libros, internet, revistas y para la manipulación de los datos se usó Python como lenguaje de programación, el programa Excel y algunos programas libres en línea para representar gráficamente funciones matemáticas. En el desarrollo del documento no se presentaron métodos matemáticos o gráfico-numéricos rigurosos que presenta el tratamiento de datos; Pero, se anexa una breve demostración para las curvas de Andrews conservando el formalismo matemático para la representación gráfica.

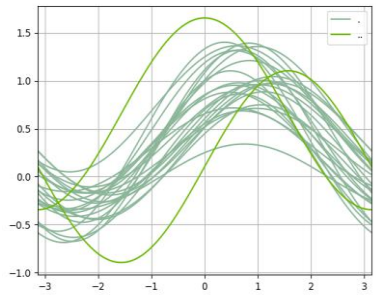
Códigos en Python.

Representaciones gráficas en Python. Datos del ejemplo 20 etiquetados para identificar dos datos atípicos.

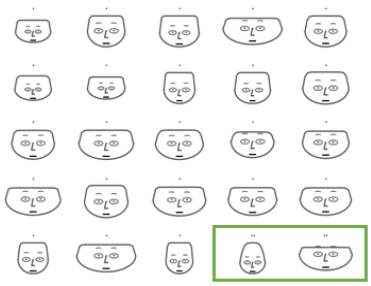
ci	peso	edad	etiqueta
0.735	0.516	0.458	.
0.162	0.469	0.583	.
0.647	0.694	0.625	.
0.559	0.573	0.125	.
0.382	0.280	0.542	.
1.000	0.373	0.542	.
0.838	0.641	0.417	.
0.456	0.608	0.792	.
0.676	0.444	0.708	.
0.500	0.330	0.500	.
0.191	0.359	0.417	.
0.603	0.447	0.292	.
0.397	0.223	0.375	.
0.000	0.230	0.250	.
0.221	0.516	0.292	.
0.500	0.579	0.208	.
0.426	0.406	0.542	.
0.515	0.570	0.292	.
0.309	0.599	0.292	.
0.382	0.364	0.250	.
0.294	0.418	0.792	.
0.618	0.584	0.167	.
0.588	0.363	0.917	.
0.926	0.000	1.000	..
0.147	1.000	0.000	..




```
import pandas as pd
import matplotlib.pyplot as plt
# Cargar datos de excel: Ruta
df=pd.read_excel(r"C:\Users\TECNO\Desktop\datosflore
s.xlsx", "Hoja10")
# ver datos de excel
print(df.head())
# Variables
x=df["ci"].tolist()
y=df["peso"].tolist()
z=df["edad"].tolist()
# Generar gráfico 3D
fig = plt.figure(figsize=(10,10))
ax = fig.add_subplot(projection = '3d')
ax.scatter(x, y, z, color="blue", s=40)
plt.xlabel("ci")
# ver gráfico 3D
plt.show()
```

<table>
<tr>
<td>  </td>
<td>

```
import pandas as pd
import matplotlib.pyplot as plt
from pandas.plotting import andrews_curves
df=pd.read_excel(r"C:\Users\TECNO\Desktop\datosflore
s.xlsx", "Hoja10")
#ver datos
print(df.head())
#ver figura
plt.figure()
andrews_curves(df, "etiqueta")
plt.show()
```

</td>
</tr>
<tr>
<td>  </td>
<td>

```
from ChernoffFace import *
import matplotlib.cm
import pandas as pd
df=pd.read_excel(r"C:\Users\TECNO\Desktop\datosflore
s.xlsx", "Hoja10")
#ver datos
print(df.head())
fig = chernoff_face(data=df)
fig.tight_layout()
matplotlib.pyplot.show()
```

</td>
</tr>
</table>

Figura 85. Código Python para algunas representaciones. Fuente: Elaboración propia.

Referencias bibliográficas.

- Arteaga, P. (2009). *ANÁLISIS DE GRÁFICOS ESTADÍSTICOS ELABORADOS EN UN PROYECTO DE ANÁLISIS DE DATOS*. (tesis de máster). UNIVERSIDAD DE GRANADA. Granada, España.
- Balzarini M., Bruno C., Córdoba M. y Teich I. (2015). Herramientas en el Análisis Estadístico Multivariado. Escuela Virtual Internacional CAVILA. Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba. Córdoba, Argentina.
- Batanero, C. y Godino, J. D. (2001). Análisis de datos y su didáctica. Granada, España: Servicio de Reprografía de la Facultad de Ciencias Universidad de Granada.
- Behar, R. y Ojeda, M. (2006). Estadística, Productividad y Calidad. Veracruz, México: Secretaría de Educación de Veracruz.
- Behar, R. y Yepes. M. (2007). Estadística, Un Enfoque Descriptivo. Cali, Colombia: Talleres Gráficos De Impresora FERIVA S.A.
- Calot, G. (1988). Curso de estadística descriptiva. Madrid, España: Paraninfo S.A.
- Canavos, G. (1988) Probabilidad y Estadística Aplicaciones y Métodos, Naucalpan de Juárez, México: Interamericana de México.
- Cook, D. y Swayne, D. F. (2007). Interactive and Dynamic Graphics for Data Analysis. New York, USA: Springer Science+Business Media.
- Díaz Monroy, L. G. (2007). Estadística multivariada: inferencia y métodos. Bogotá, Colombia: Proceditor ltda.
- Díaz Monroy, L. G. y Morales Rivera, M. A. (2012). Análisis estadístico de datos multivariados. Bogotá, Colombia: Coordinación de Publicaciones, Facultad de Ciencias.
- Francisco Azorin, (1972). Curso de Muestreo y Aplicaciones. Madrid: Aguilar.
- García Marín, A. (2022). *Visualización de Datos Univariantes y Multivariantes, Catalogo de Técnicas* (tesis de pregrado). Universidad Politécnica de Madrid.
- Grande Atienza, E. (2021). *Análisis exploratorio de datos. Una variable (*tesis de pregrado*).* Universidad de Valladolid. Valladolid, España.
- Obando-Bastidas, J. y Castellanos Sánchez, M. (2021). Gráficos estadísticos: guía práctica para estadística descriptiva. Universidad Cooperativa de Colombia, Facultad de Ciencias Económicas, Administrativas y Contables, Contaduría Pública, Villavicencio. Disponible en: https://doi.org/10.16925/gcgp.32
- Orlandoni, G. (2010). Escalas de medición en Estadística. TELOS. Revista de Estudios Interdisciplinares en Ciencias Sociales, 12(2), 243-247.
- Pablo Pérez Manso. (2019). Análisis de datos funcionales: representación en bases y regresión funcional en scikit-fda (tesis de pregrado). UNIVERSIDAD AUTÓNOMA DE MADRID. Madrid, España.
- Pardo, C. E. (2020). Estadística descriptiva multivariada. Bogotá, Colombia: Universidad Nacional de Colombia.
- PÉREZ LÓPEZ, C. (2004). Técnicas de Análisis Multivariante de Datos. Madrid, España: Pearson prentice hall.

- RUBIO DONET, J. L. (2018). *DETECCIÓN DE DATOS MULTIVARIADOS ATÍPICOS CON SERIES FINITAS DE FOURIER (*tesis de maestría*).* Universidad Nacional Agraria la Molina. Lima, Perú.
- Vigo Ruiz, J. M. (2016). *Comprensión de gráficos estadísticos por alumnos de formación profesional básica* (tesis de máster). Universidad de Granada. Granada, España.
- Yilton Riascos, (2013). El pensamiento estadístico asociado a las medidas de tendencia central: un estudio psicogenético sobre la media aritmética, la mediana y la moda. Cali: Universidad del Valle.

Otras referencias.

- Acevedo Bohórquez, I. y Velásquez Ceballos, E. (2008) Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales. Recuperado de: https://repository.eafit.edu.co/server/api/core/bitstreams/e172a144-1e8f-4fca-91bc-80541200e57f/content
- Batanero, C. (2001). Didáctica de la Estadística. Recuperado de: https://www.ugr.es/~batanero/pages/ARTICULOS/didacticaestadistica.pdf
- Castaño, E. Introducción al análisis de datos multivariados en ciencias sociales. Recuperado de: https://www.inec.gob.pa/IASI/docs/announcements/documentos/MemoriasCursillos/4%20Casta%C3%B1o_An%C3%A1lisis%20de%20datos%20multivariados.pdf
- Francalacci P, Morelli L, Useli A y Sanna D. (2010). The History and Geography of the Y Chromosome SNPs in Europe: an update. Recuperado de: https://www.researchgate.net/publication/46220557_The_History_and_Geography_of_the_Y_Chromosome_SNPs_in_Europe_an_update
- Irene Schiattino y Claudio Silva. (2013). Representación gráfica de información multivariante. Aplicación al sistema de salud de Chile (2010). Recuperado de: https://r.search.yahoo.com/_ylt=AwrNYNlVhnBmDTQ5mACrcgx.;_ylu=Y29sbwNiZjEEcG9zAzIEdnRpZAMEc2VjA3Ny/RV=2/RE=1718679254/RO=10/RU=https%3a%2f%2frevistas.uchile.cl%2findex.php%2fRCSP%2farticle%2fdownload%2f27126%2f28758/RK=2/RS=qBswPLg58dxmN90KRMn7YWGr38A-
- Mauricio, J. A. (2007). Introducción al Análisis de Series Temporales. Recuperado de: https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IAST-Libro.pdf
- Peña, D. (2002). Análisis de Datos Multivariantes. Recuperado de: https://www.researchgate.net/publication/40944325_Analisis_de_Datos_Multivariantes
- Pere Grima. (2008). Estadística en acción: Qué es y para qué sirve la estadística a través de casos prácticos basados en proyectos final de carrera. Barcelona. España: BARCELONA DIGITAL. Recuperado de: https://upcommons.upc.edu/handle/2117/7915?show=full

- Sánchez S, E. (2013) ELEMENTOS DE ESTADÍSTICA Y SU DIDÁCTICA A NIVEL BACHILLERATO. Recuperado de: https://drive.google.com/file/d/1XCDwF_GWNFmESuCe2fJiCrIrXNQuDnjY/view
- GUÍA PARA LA PRESENTACIÓN DE GRÁFICOS ESTADÍSTICOS (2009). Instituto Nacional de Estadística e Informática. Lima, Perú. Recuperado de: https://www.inei.gob.pe/media/MenuRecursivo/metodologias/libro.pdf