

**FORMACIÓN MÉDICA****Columnas, tartas y un italiano ilustre. Representación gráfica de variables cualitativas.**

**Artículo original:** Blum RH, Muret-Wagstaff SL, Boulet JR, Cooper JB, Petrusa ER. Simulation-based assessment to reliably identify key resident performance attributes. *Anesthesiology* 2018; 128:821-31. ([HTML](#))

*Molina Arias M.*

*Hospital Infantil Universitario La Paz. Madrid. España.*

**Resumen**

La estadística descriptiva es aquella rama de la estadística que utilizamos habitualmente para obtener una primera aproximación a los resultados de nuestro estudio. Con frecuencia se recurre a la representación gráfica de las variables para conseguir una síntesis rápida más sencilla de comprender. Se revisan los tres gráficos más utilizados para la representación de las variables cualitativas.

**Introducción**

La estadística descriptiva es aquella rama de la estadística que utilizamos habitualmente para obtener una primera aproximación a los resultados de nuestro estudio. Con frecuencia se recurre a la representación gráfica de las variables para conseguir una síntesis rápida más sencilla de comprender. Se revisan los tres gráficos más utilizados para la representación de las variables cualitativas.

Cuando uno lee el título de esta entrada puede preguntarse con qué estúpida ocurrencia voy a machacar hoy a la

sufrida concurrencia, pero no temáis, lo único que vamos a hacer es poner en valor ese famoso aforismo que dice que una imagen vale más que mil palabras. ¿Os he aclarado algo? Supongo que no.

Como todos sabemos, la estadística descriptiva es aquella rama de la estadística que utilizamos habitualmente para obtener una primera aproximación a los resultados de nuestro estudio, una vez que lo hemos terminado.

Lo primero que hacemos es describir los datos, para lo cual realizamos tablas de frecuencias y utilizamos medidas diversas de centralización y dispersión. El problema con estos parámetros es que, aunque representan verdaderamente la esencia de los datos, a veces es difícil proporcionar con ellos una visión sintética y comprensiva. Es en estos casos en los que podemos recurrir a otro recurso, que no es otro que la representación gráfica de los resultados del estudio. Ya sabéis, una imagen vale más que mil palabras, o eso dicen.

Hay multitud de tipos de gráficos para ayudarnos a comprender mejor la representación de los datos, pero hoy nos vamos a limitar a aquellos que tienen que ver con las variables cualitativas o categóricas.

Recordad que las variables cualitativas representan atributos o categorías de la variable. Cuando la variable no incluye ningún sentido de orden, se dice que es cualitativa nominal, mientras que si se puede establecer cierto orden entre las categorías diríamos que es cualitativa ordinal. Por ejemplo, la variable “fumador” sería cualitativa nominal si tiene dos posibilidades: “sí” o “no”. Sin embargo, si la definimos como “ocasional”, “poco fumador”, “moderado” o “muy fumador”, ya existe cierta jerarquía y hablamos de variable cualitativa ordinal.

El primer tipo de gráfico que vamos a considerar a la hora de representar una variable cualitativa es el gráfico de sectores, mucho más conocido como gráfico de tarta. Este consiste en una circunferencia cuya área representa el total de los datos. Así, a cada categoría se le asigna un área que será directamente proporcional a su frecuencia. De esta forma, las categorías más frecuentes tendrán áreas mayores, de modo que de un vistazo podemos hacernos una idea de cómo se distribuyen las frecuencias en las categorías.

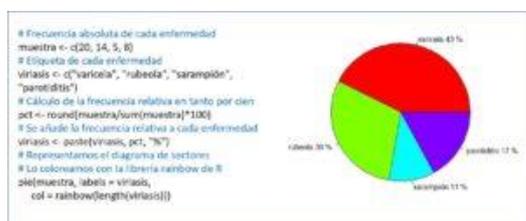


Figura 1. Gráfico de sectores

Hay tres formas de calcular el área de cada sector. La más sencilla es multiplicar la frecuencia relativa de cada categoría por 360°, obteniendo los grados de ese sector.

La segunda es utilizar la frecuencia absoluta de la categoría, según la siguiente regla de tres:

$$\text{Frecuencia absoluta} / \text{Frecuencia total de datos} = \text{Grados del sector} / 360^\circ$$

Por último, la tercera forma consiste en utilizar las proporciones o porcentajes de las categorías:

$$\% \text{ de la categoría} / 100\% = \text{Grados del sector} / 360^\circ$$

Las fórmulas son muy sencillas, pero, de todas formas, no habrá necesidad de recurrir a ellas porque el programa con el que hagamos el gráfico lo hará por nosotros. La instrucción en R es `pie()`, tal como podéis ver en la figura 1, en la que os muestro una distribución de niños con enfermedades exantemáticas y cómo se representaría el gráfico de sectores.

El gráfico de sectores está diseñado para representar variables categóricas nominales, aunque no es raro ver tartas representando variables de otros tipos. Sin embargo, y en mi humilde opinión, esto no es totalmente correcto.

Por ejemplo, si hacemos un gráfico de sectores para una variable cualitativa ordinal estaremos perdiendo la información sobre la jerarquía de las variables, por lo que resultaría más correcto utilizar un gráfico que permita ordenar las categorías de menos a más. Y este gráfico no es otro que el diagrama de barras, del que hablaremos a continuación.

El diagrama de sectores será especialmente útil cuando haya pocas variables. Si hay muchas, la interpretación deja de ser tan intuitiva, aunque siempre podemos completar el gráfico con una tabla de frecuencias que nos ayude a interpretar mejor los datos. Otro consejo es tener mucho cuidado

con los efectos en 3D a la hora de dibujar las tartas. Si nos pasamos de elaborados, el gráfico perderá claridad y será más difícil de leer.

El segundo gráfico que vamos a ver es, ya lo hemos mencionado, el gráfico de barras, el óptimo para representar las variables cualitativas ordinales. En el eje horizontal se representan las diferentes categorías y sobre él se levantan unas columnas o barras cuya altura es proporcional a la frecuencia de cada categoría. También podríamos utilizar este tipo de gráfico para representar variables cuantitativas discretas, pero lo que no es muy correcto hacer es usarlo para las variables cualitativas nominales.

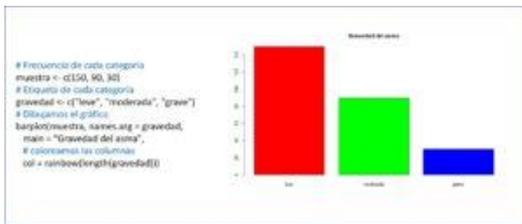


Figura 2. Diagrama de barras

El diagrama de barras es capaz de expresar la magnitud de las diferencias entre las categorías de la variable, pero ahí está, precisamente, su punto débil, ya que es fácilmente manipulable si modificamos las escalas de los ejes. Por eso hay que tener cuidado al analizar este tipo de gráficos para evitar que nos engañen con el mensaje que el autor del estudio pueda querer transmitir.

Este gráfico también es sencillo de hacer con la mayor parte de los programas estadísticos y hojas de cálculo. La función en R es `barplot()`, como veis en la figura 2, que representa la gravedad de una muestra de niños asmáticos.

Con lo visto hasta ahora, algunos pensaréis que el título de esta entrada es un poco engañoso. En realidad, la cosa no va de columnas y tartas, sino de

barras y sectores. Además, ¿quién es el italiano ilustre? Pues aquí sí que no engaño a nadie, porque el personaje fue las dos cosas, italiano e ilustre, y me estoy refiriendo a Vilfredo Federico Pareto.

Pareto fue un italiano que nació a mediados del siglo XIX en París. Esta pequeña contradicción se debe a que su padre estaba entonces exiliado en Francia por ser uno de los seguidores de Giuseppe Mazzini, que estaba entonces empeñado en la unificación italiana. De todas formas, Pareto vivió en Italia desde los 10 años de edad, convirtiéndose en un ingeniero con amplios conocimientos matemáticos y humanistas y que contribuyó de manera decisiva al desarrollo de la microeconomía. Hablaba y escribía con fluidez en francés, inglés, italiano, latín y griego, y se hizo famoso por multitud de contribuciones como la distribución de Pareto, la eficiencia de Pareto, el índice de Pareto y el principio de Pareto. Para representar este último inventó el diagrama de Pareto, que es el que le trae hoy aquí entre nosotros.

El diagrama de Pareto (también conocido en economía como curva cerrada o distribución A-B-C) organiza los datos en orden descendente de izquierda a derecha, representados por barras, asignando así un orden de prioridades. Además, el diagrama incorpora una línea curva que representa la frecuencia acumulada de las categorías de la variable. Esto permitía inicialmente explicar el principio de Pareto, que viene a decir que hay muchos problemas sin importancia frente a unos pocos que sí son importantes, con lo que resultaba muy útil para la toma de decisiones.

Como es fácil de comprender, esta priorización hace que el diagrama de Pareto sea especialmente útil para representar variables cualitativas

ordinales, superando al diagrama de barras al dar información sobre el porcentaje acumulado al ir agregando las categorías de la distribución de la variable. El cambio de pendiente de esta curva nos informa también del cambio en la concentración de datos, que depende de la variabilidad en que los sujetos de la muestra se reparten entre las distintas categorías.

Por desgracia, R no dispone de una función simple para representar diagramas de Pareto, pero podemos obtenerlo fácilmente con el script que tenéis en la figura 3. El resultado podéis verlo en la figura 4.

```
# Construir la tabla con las frecuencias de cada categoría
tabaquismo <- c(500, 145, 170, 200, 25)
grado <- c("no fumador", "ocasional", "leve", "moderado", "intenso")
datos <- data.frame(tabaquismo, grado)
# Reordenar en orden descendente
datos <- datos[order(datos$tabaquismo, decreasing = TRUE), ]
# Convertir los grados de tabaquismo en factor
datos$grado <- factor(datos$grado, levels = datos$grado)
# Calcular la frecuencia acumulada total de las variables
datos$acumulada <- cumsum(datos$tabaquismo)
# Construir el diagrama de Pareto
barplot(datos$tabaquismo, names.arg = datos$grado,
        main = "Diagrama de Pareto del grado de tabaquismo", # Título del gráfico
        ylab = "Frecuencias absolutas", # Título del eje y
        ylim = c(0, sum(datos$tabaquismo)*1.05), # Límites del eje y
        col = rainbow(length(datos$grado))) # Lo coloreamos
# Dibujamos la línea de frecuencia acumulada
lines(1:length(datos$tabaquismo), datos$acumulada, type = "b", pch=16, col="blue")
# Dibujamos una caja alrededor del gráfico
box()
```

Figura 3. Script diagrama de Pareto

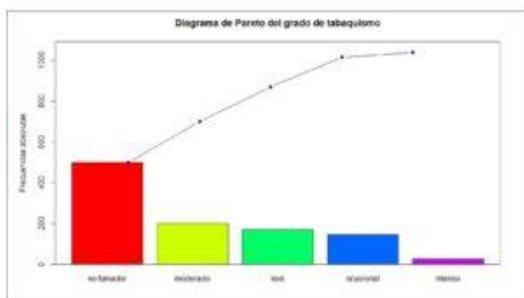


Figura 4. Diagrama de Pareto

Y aquí lo vamos a dejar por hoy. Antes de decir adiós quiero avisaros que no debéis confundir las barras del diagrama

de barras con las del histograma ya que, aunque pueden parecerse desde el punto de vista gráfico, ambos representan cosas muy diferentes. En un diagrama de barras solo se representan los valores de las variables que hemos observado al hacer el estudio. Sin embargo, el histograma va mucho más allá ya que, en realidad, encierra la distribución de frecuencias de la variable, por lo que representa todos los valores posibles que existen dentro de los intervalos, aunque no hayamos observado ninguno de forma directa. Permite así calcular la probabilidad de que se represente cualquier valor de la distribución, lo que es de gran importancia si queremos hacer inferencia y estimar valores de la población a partir de los resultados de nuestra muestra. Pero esa es otra historia...

## Bibliografía

- Turbany J, Barrios M, Carrera MJ. Estadística descriptiva univariante y bivariante. En: Peró M, Leiva D, Guardia J, Solanas A, eds. Estadística aplicada a las ciencias sociales mediante R y R-Commander. Ibergarceta Publicaciones SL. Madrid, 2012: 165-228.
- Ortega Páez E, Ochoa Sangrador C, Molina Arias M. Representación gráfica de variables. Evid Pediatr. 2019;15:13. (HTML)

---

### Correspondencia al autor

Manuel Molina Arias  
 mma1961@gmail.com  
 Servicio de Gastroenterología.  
 Hospital Infantil Universitario La Paz.  
 Madrid. España.

---

Aceptado para el blog en noviembre de 2019