

Unidad 2

- Tipos de variables.
- Escalas de medición.
- Análisis exploratorio de datos.
 - ✓ Resumen de datos en forma gráfica.
 - ✓ Medidas de tendencia central o de posición.
 - ✓ Medidas de dispersión.

La descripción en el análisis estadístico

La **estadística descriptiva** abarca los métodos para resumir la información recolectada (que puede provenir tanto de una población como de una muestra). El propósito principal es reducir la información de manera tal de que se distorsione o se pierda la menor cantidad de información posible.

Veremos cómo organizar y resumir la información mediante:

- Tablas
- Gráficos
- Medidas resumen

Las **variables** son propiedades, atributos o características que forman parte del problema y a través de las cuales podremos explorarlo, describirlo o explicarlo. Las variables toman distintos *valores*.

Un **dato** es el *valor* que adopta una *variable* medida en una *unidad de análisis*.

Tipos de variables

Los valores que toman las variables (*observaciones*) pueden ser números (por ej.: ingreso, en \$, de un jefe de hogar) o bien cada observación puede pertenecer a una categoría (por ej.: Sí/No para la variable «Trabaja»).

- Una variable se dice **categórica** si cada observación pertenece a una de un conjunto de categorías.
- Una variable se dice **cuantitativa** si sus observaciones toman valores numéricos que representan diferentes magnitudes de dicha variable.

Al definir una variable cuantitativa, decimos que los valores numéricos deben *representar diferentes magnitudes*.

Esto es así ya que las variables cuantitativas miden “cuánto de algo” (es decir, *cantidades o magnitudes*).

Con variables cuantitativas podemos calcular medidas de resumen numéricas, como promedios.

Sin embargo, algunas variables numéricas, como códigos de áreas, no son consideradas cuantitativas porque no representan una “cantidad o magnitud”.

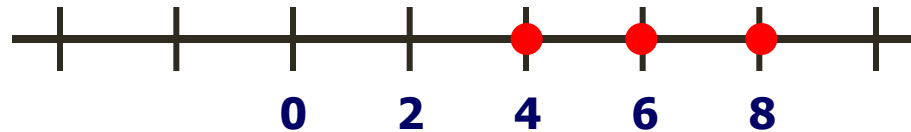
Promediarlos o sumarlos *no tendría ningún sentido*.

Las variables cuantitativas pueden clasificarse a su vez:

- Una variable se dice **cuantitativa discreta** si sus valores posibles forman un conjunto de números enteros (0, 1, 2, 3...). Por lo general, se trata de conteos («número de...»). Puede decirse que cualquier variable con un número finito de valores posibles es discreta.
- Una variable se dice **cuantitativa continua** si sus valores posibles forman un intervalo. Es decir, puede tomar infinitos valores posibles (tiempo, distancia, peso, altura, etc.)

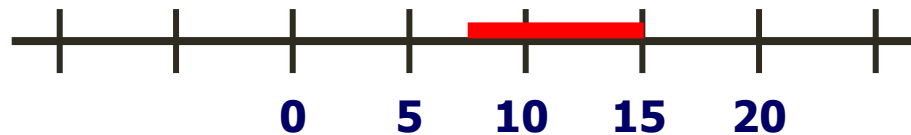
Variables cuantitativas discretas:

los posibles valores de la variable son puntos separados sobre la recta de números Reales.



Variables aleatorias continuas:

el conjunto de posible valores forma un intervalo sobre la recta de números Reales.



Escala de medida

Escala de razón

- Considera intervalos constantes; existe un cero absoluto que representa ausencia de atributo. El origen es único. Se puede determinar igualdad de razones.



Escala de intervalo

- Considera intervalos constantes, pero no un cero absoluto. Se puede determinar igualdad de intervalos o diferencias (se puede calcular la distancia exacta entre 2 elementos)
(*Tiempo, temp.*)



Escala ordinal

- Los datos pueden ser ordenados pero no medidos. No se conoce la distancia entre una categoría y la siguiente.



Escala nominal

- Los datos son clasificados no por una medida numérica sino mediante alguna cualidad o atributo. Sólo se pueden determinar relaciones de igualdad.

Análisis descriptivo según tipo de variable

Las medidas de resumen y los gráficos describen las características principales de una variable.

- Para las variables cuantitativas:
las características a describir son el centro y la variabilidad o dispersión de los datos.
- Para las variables categóricas:
una característica fundamental a describir es el número relativo de observaciones en las distintas categorías o niveles.

Ejemplo motivador

Modificaciones en la vía aérea de pacientes embarazadas en trabajo de parto

Se cuenta con datos correspondientes a un estudio observacional realizado sobre 100 pacientes en trabajo de parto en cierto Hospital de la ciudad de Rosario a fines de 2015.

El objetivo del estudio fue valorar la calificación de Mallampati en las mujeres en trabajo de parto y si existía un incremento en el puerperio inmediato en comparación con el trabajo de parto.

En la mujer embarazada, las complicaciones derivadas del manejo de la vía aérea representan la principal causa de morbimortalidad de origen anestésico.

La evaluación de la vía aérea antes del acto anestésico es fundamental.

La clasificación de Mallampati es una herramienta simple para evaluar la posible dificultad para la intubación. Está definida por cuatro niveles: I, II, III, IV, donde el primero indica que no hay ninguna dificultad y el IV indica la situación más desfavorable.

Se evaluó la vía aérea con la escala de Mallampati al ingreso y en las primeras 2 hs. posparto.

Además se registraron otras variables como la edad, peso, índice de masa corporal (IMC), dilatación cervical al ingreso, gestas y partos previos y duración del trabajo de parto.

Algunas de las preguntas que se podrían plantear acerca de los datos podrían ser:

- ¿Qué características presentan las embarazadas bajo estudio?
- ¿Cómo fue su distribución de acuerdo a su edad?
- ¿Eran todas primerizas las embarazadas evaluadas o no?
- ¿Cuántos partos previos habían tenido quienes no lo eran?
- ¿Cómo fue la distribución de la escala de Mallampati en la primera y en la segunda medición?
- ¿Hay alguna mujer con valores inusuales en alguna o algunas variables?

Para poder responder estas preguntas es necesario organizar los datos de una manera adecuada mediante presentaciones tabulares y gráficas.

Tablas de frecuencias

El primer paso al realizar un resumen numérico de una variable es observar los valores que tomó la variable y contar con qué frecuencia se presentaron.

Para una variable categórica, dado que cada observación cae en una de sus categorías, podemos utilizar **proporciones o porcentajes** para resumir el número de observaciones en cada una de las categorías.

En este caso, una forma común de presentar los datos es a través de una **tabla de distribución de frecuencias**.

La **proporción** de observaciones que cae en cierta categoría es la frecuencia (el conteo) de observaciones en dicha categoría dividido por el total de observaciones. El **porcentaje** es la proporción $\times 100$. Las proporciones y porcentajes también son llamadas **frecuencias relativas**.

Una **tabla de frecuencias** es una tabla que muestra los valores que toma una variable (las posibles categorías), junto con el número de observaciones para cada valor (**frecuencias absolutas**) y/o con las frecuencias relativas.

...continuando con ejemplo motivador

Primerizas	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa porcentual
Sí	59	0.59	59
No	41	0.41	41
Total	100	1.00	100

En R...

```
attach(datos)

(n=length(PRIMERIZA))
(tablaFrecAbs = table(PRIMERIZA))
(tablaFrecRel = tablaFrecAbs/n)
(tablaFrecRelPct = (tablaFrecAbs/n)*100)
```

Gráficos para variables categóricas

Los gráficos más utilizados para mostrar información correspondiente a variables categóricas son:

- Gráficos de sectores o de “torta”.
- Gráficos de barras.
- Gráfico de Pareto.

Gráficos de sectores o de “torta”

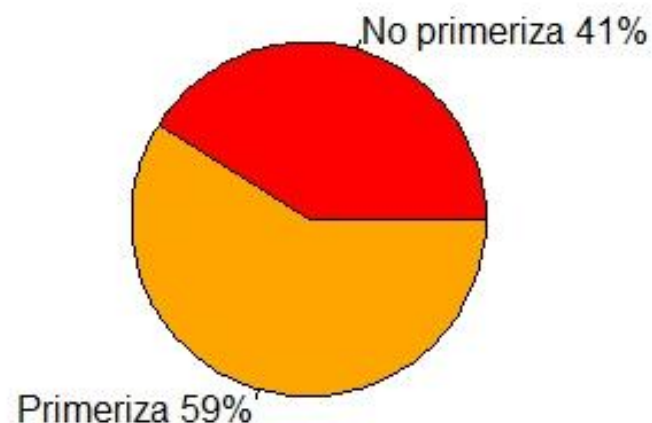
Se utiliza un círculo para representar todo el conjunto de datos, mientras que los "sectores" del círculo representan las posibles categorías.

El tamaño de un sector para una categoría en particular es proporcional a la correspondiente frecuencia o frecuencia relativa.

Estos gráficos son útiles cuando se cuenta con pocas categorías.

Tener en cuenta que su uso es adecuado cuando, además, las categorías son **nominales**.

..continuando con el ejemplo motivador



En R...

```
frec_abs<-table(PRIMERIZA)
porcentajes<-(frec_abs/n)*100

lbls <- c("No primeriza", "Primeriza")
lbls <- paste(lbls, porcentajes)
lbls <- paste(lbls,"%", sep="")
pie(frec_abs,labels = lbls, col=c("red", "orange"))
```

Gráficos de barras

Cada categoría en la distribución de frecuencias se representa por una barra o rectángulo y el gráfico se construye de forma tal que el área de la barra sea proporcional a la frecuencia o frecuencia relativa correspondiente.

La altura de las barras se corresponde con la frecuencia (o frecuencia relativa) de observaciones en cada categoría.

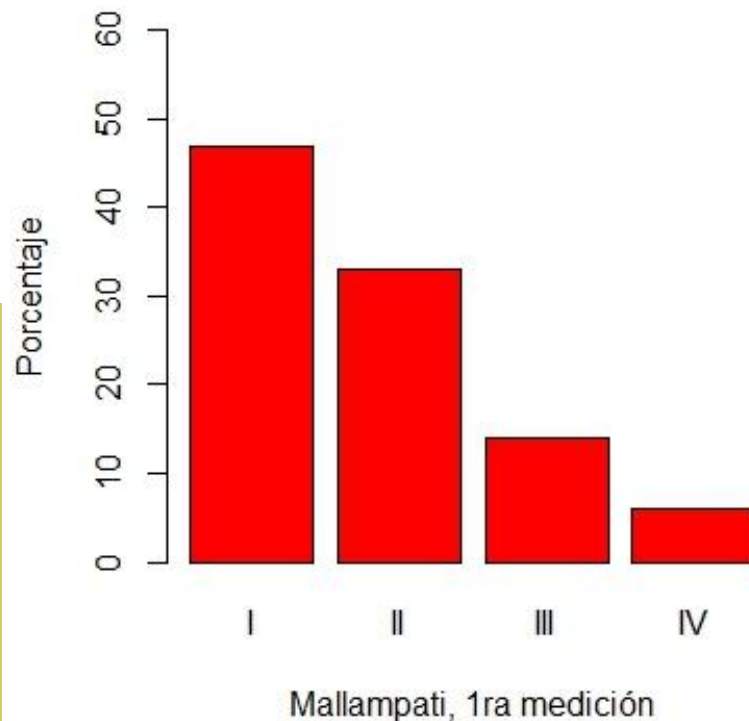
Todas las barras deben tener el mismo ancho, el cual debe ser mayor que la separación existente entre ellas. No se recomienda usar colores diferentes para cada barra (puede engañar visualmente)

Cuando la variable a graficar es **categórica nominal**, lo adecuado es que las barras se grafiquen de forma horizontal.

Modificaciones en la vía aérea de pacientes en trabajo de parto...

En R...

```
barplot(  
  tablaFrecRelPct,  
  xlab="Mallampati, 1ra medición",  
  ylab="Porcentaje", ylim=c(0,60),  
  col=c("red"),  
  names=c("I", "II", "III", "IV")  
)
```



Gráficos de barras

Notar que la variable grafica corresponde a una variable ordinal, sus categorías presentan un orden natural.

En este caso no sería apropiado trabajar con un gráfico de sectores, aún cuando las categorías son pocas, ya que se desaprovecharía la información que provee la ordinalidad de la variable.

Si la variable fuera nominal, sería conveniente ordenar las barras de acuerdo a su frecuencia, ya sea descendente o ascendente para facilitar la lectura y las conclusiones.

Gráficos de barras comparativas

Los gráficos de barras también pueden utilizarse para presentar una comparación visual entre 2 o más grupos.

Esto se realiza construyendo 2 o más gráficos de barras usando los mismos ejes verticales y horizontales.

Si los tamaños muestrales para los distintos grupos son diferentes es aconsejable utilizar las frecuencias relativas para poder realizar comparaciones entre ellos.

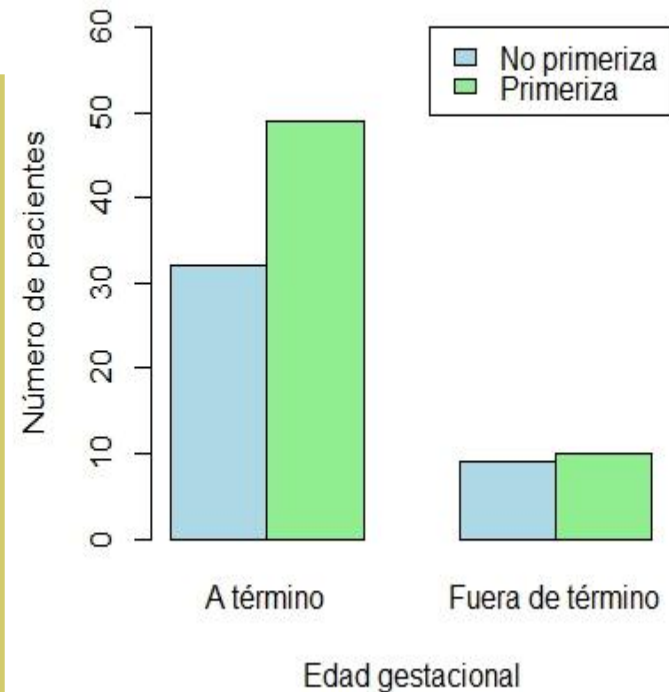
...continuando con ejemplo motivador

En R...

```
conteo <- table(PRIMERIZA,  
                EDAD_GEST_2)
```

```
barplot(  
  conteo, xlab="Edad gestacional",  
  names=c("A término", "Fuera de  
término"),  
  ylab="Número de pacientes",  
  col=c("lightblue", "lightgreen"),  
  ylim=c(0,60), beside=TRUE )
```

```
legend("topright",  
  legend = c("No primeriza", "Primeriza"),  
  fill = c("lightblue", "lightgreen"))
```



Gráficos de sectores vs. gráficos de barras

Ambos gráficos son simples de construir y de interpretar.

En algunas situaciones ambos son factibles, pero puede ocurrir que sea más claro el gráfico de barras ya que si dos sectores son muy similares en tamaño puede no ser claro cuál es mayor.

La diferencia es más clara cuando se compara la altura de las barras y éstas están ordenadas de acuerdo a su magnitud.

Como ya se dijo, las barras son preferibles cuando son muchas las categorías a comparar.

Gráfico o diagrama de Pareto*

Se trata de un gráfico de barras con categorías ordenadas por su frecuencia, desde la más alta hasta la más baja.

Se utiliza frecuentemente en aplicaciones de negocios o de control de calidad para identificar la mayoría de los resultados más comunes. Por ejemplo, identificar los productos con las ventas más altas o los tipos más frecuentes de defectos en cierto producto.

Este diagrama ayuda a mostrar el **principio de Pareto**, el cual postula que un pequeño número de categorías contienen a la mayoría de las observaciones (*“separa las pocas causas vitales de las muchas triviales”*).

* *En honor al economista italiano Vilfredo Pareto (1848–1923) quien propuso su uso.*

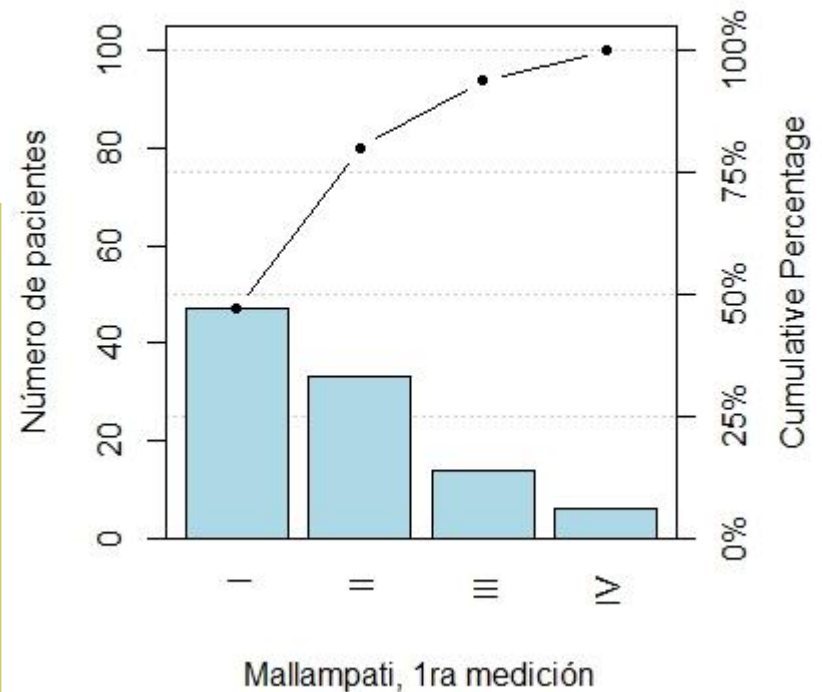
...continuando con ejemplo motivador

En R...

```
install.packages("qcc")  
library(qcc)
```

```
frec_abs <- table(MAL_1)
```

```
pareto.chart(frec_abs,  
names=c("I", "II", "III", "IV"),  
xlab="Mallampati, 1ra medición",  
ylab="Número de pacientes",  
main="", col="lightblue")
```



(sólo para ejemplificar, no sería un caso habitual donde se presente este gráfico)

Gráficos para variables cuantitativas

Entre los gráficos adecuados para mostrar información correspondiente a variables cuantitativas se encuentran:

- Gráficos de bastones.
- Gráficos de puntos o “*dot plots*”.
- Diagramas de tallo y hoja.
- Histogramas.

Gráfico de bastones

Estos gráficos son los adecuados para representar datos numéricos discretos, los cuales generalmente se originan en procesos de conteo.

En tales casos, cada observación es un número entero. Como en el caso de datos categóricos, una distribución de frecuencias lista cada valor posible (en forma individual o agrupados en intervalos), la frecuencia asociada y la frecuencia relativa.

Gráfico de bastones

En un sistema de coordenadas cartesianas se representan en el eje de las abscisas u horizontal los distintos valores que asume la variable discreta en estudio y en el eje de las ordenadas o vertical se construye una escala adecuada para representar la frecuencia correspondiente a cada uno de esos valores.

Sobre cada valor de la variable, se levanta una línea o bastón igual a la frecuencia de la categoría en cuestión.

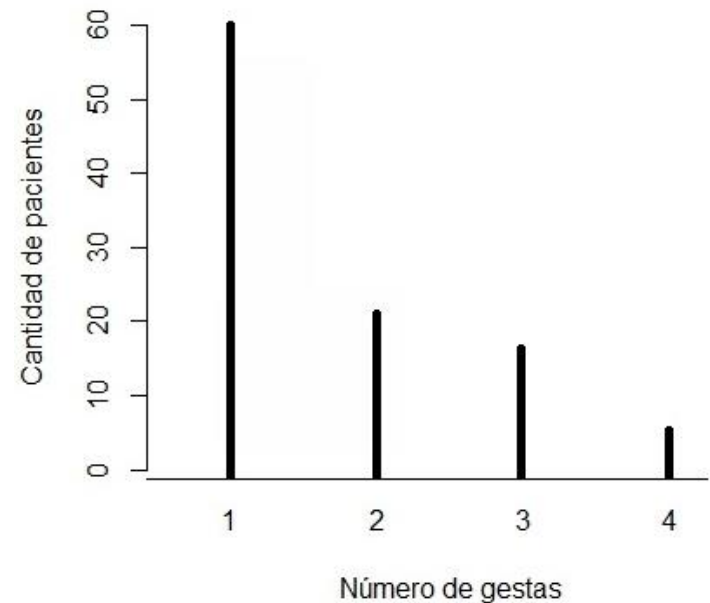


Gráfico de puntos

Es una manera simple de presentar datos numéricos cuando el conjunto de datos es razonablemente pequeño ya que se muestran todas las observaciones individuales. A partir de este gráfico, podemos reconstruir (al menos en forma aproximada) todos los datos de la muestra.

Cada observación se representa por un punto sobre la ubicación correspondiente a su valor en una escala horizontal.

Cuando un valor se presenta en más de una ocasión, los puntos se apilan verticalmente. Es decir que el número de puntos sobre un valor en la recta numérica, representa la frecuencia de ocurrencia de dicho valor.

Otro ejemplo

Peso y género de 15 potrillos recién nacidos

Potrillo	Género	Peso	Potrillo	Género	Peso	Potrillo	Género	Peso
1	F	129	6	M	113	11	M	108
2	M	119	7	F	95	12	F	95
3	M	132	8	F	104	13	M	117
4	F	123	9	M	104	14	F	128
5	F	112	10	F	93	15	F	127

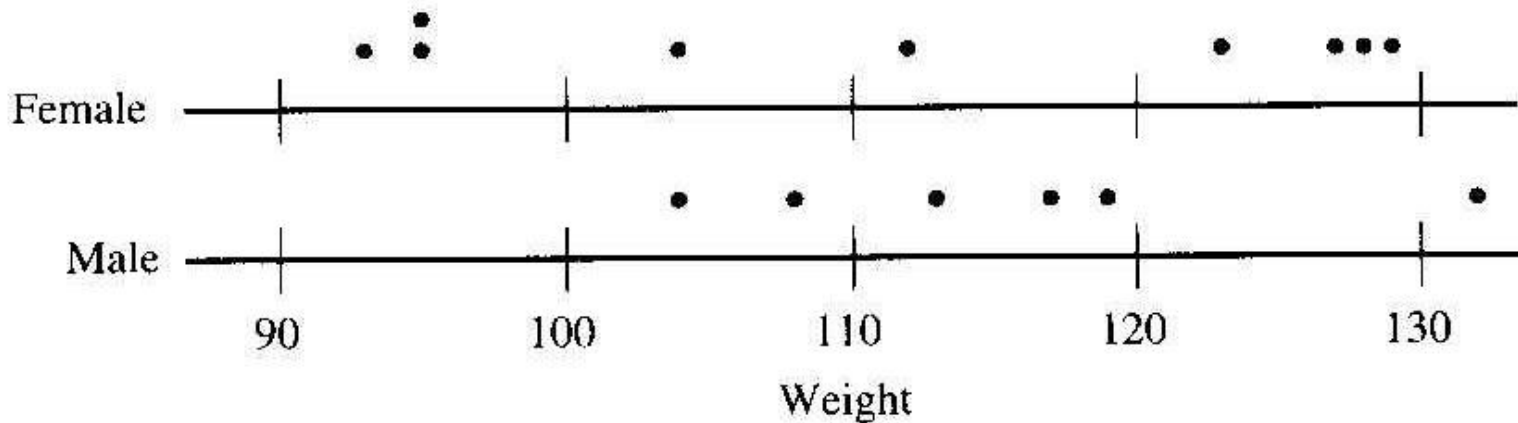


Diagrama de tallo y hoja

Un diagrama de tallo y hoja es una forma efectiva y compacta de resumir información numérica. Cada número en el conjunto de datos se divide en dos partes llamadas el tallo y la hoja.

- Tallo: es la primer parte del número y consiste del primer dígito (o primeros dígitos).
- Hoja: es la última parte del número y consiste de el o los dígitos finales.

Por ej.: el número 213 puede ser dividido en un tallo igual a 2 y una hoja igual a 13 o en un tallo igual a 21 y una hoja igual a 3.

Por último se utilizan los tallos y hojas resultantes para construir el diagrama.

Diagrama de tallo y hoja

A través de estos diagramas simples es posible obtener información acerca de diversas características importantes de un conjunto de datos tales como la forma y la dispersión.

Los diagramas de tallo y hoja pueden ser muy útiles para obtener una idea de los valores más comunes en un conjunto de datos y cuan dispersos están los datos.

También es posible detectar valores que se encuentran muy alejados del resto de las observaciones en conjunto de datos. Dichos valores se denominan valores "**extremos**" o "**outliers**".

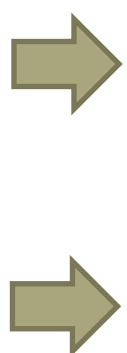
En general, se recomienda que estos diagramas contengan entre 5 y 20 tallos.

En el análisis de la variable peso (ejemplo 1)

En R...

`stem(PESO)`

The decimal point is 1 digit(s) to the right of the |



```
5 | 555
6 | 0000022233344555666888899
7 | 00000000000001122222233333455555556677788
8 | 0001122234458899
9 | 0002234555568
10 |
11 | 5
12 | 0
```

Diagrama de tallo y hoja

Se utilizan comas para separar las hojas cuando tienen 2 o más dígitos. Una alternativa es eliminar algunos dígitos de las hojas, siempre que este truncamiento no origine una gran pérdida de información con respecto a la forma o dispersión de los datos.

En ciertas ocasiones la elección natural de los tallos produce un diagrama en el cual muchas observaciones se concentran en unos pocos tallos.

Se puede obtener un diagrama más informativo dividiendo las hojas en un tallo determinado en dos grupos: las que comienzan con 0,1,2,3 ó 4 y las que comienzan con 5,6,7,8 ó 9.

Luego, cada tallo se lista dos veces al construir el diagrama.

Modificaciones en la vía aérea de pacientes en trabajo de parto...

En R...

`stem(TALLA)`

The decimal point is 1 digit(s) to the left of the |

```
15 | 00001224
15 | 55557888888999
16 | 000000000000111122222333334
16 | 55555555555555666667777888888999
17 | 00000011222333
17 | 5555
```

Ejemplo: ingesta diaria de proteínas (en gramos de proteínas por kilogramo de peso) para 20 atletas.

1.4	2.2	2.7	1.5	2.3	1.7	2.3	1.5	1.8	2.8
1.8	1.9	2.0	2.3	1.5	1.9	1.7	1.8	1.6	3.0

```

1 | 457588959786
2 | 2733803
3 | 0
Stem: Ones
Leaf: Tenths

```

```

1L | 4
1H | 57588959786
2L | 23303
2H | 78
3L | 0
Stem: Ones
Leaf: Tenths

```

Histogramas

Con un gráfico de puntos o uno de tallo y hoja es simple reconstruir el conjunto original de datos porque se muestran las observaciones individuales. Esto es poco práctico para grandes conjuntos de datos.

Una forma más versátil de graficar los datos es agruparlos de alguna manera que permita resumir ocasionando la menor pérdida de información.

Un **histograma** es un gráfico que utiliza barras para mostrar las frecuencias o frecuencias relativas de los posibles resultados de una variable cuantitativa.

en un *histograma* el “area” de cada barra es proporcional a la “frecuencia relativa” del intervalo correspondiente!!! No confundir con los gráficos de barra que se utilizan para representar variables cualitativas.

Histogramas

Una variable continua es natural que asuma muchos valores distintos lo que hace necesario dividir el rango de valores posibles para dicha variable en intervalos más pequeños, que faciliten su interpretación.

Puede ser necesario recurrir a este método también en el caso de una variable discreta, si ésta toma un gran número de valores posibles.

En cualquiera de los casos, se forma una tabla de distribución de frecuencias para los intervalos elegidos y se grafican las frecuencias (absolutas o relativas) para dichos intervalos.

Histogramas

La dificultad para construir distribuciones de frecuencias e histogramas surge por el hecho de que no existen categorías naturales.

Para variables tales como tiempo de reacción (s) o rendimiento de combustible (km/l), será necesario definir categorías arbitrarias.

Por ejemplo una variable cuantitativa continua que asume muchos valores distintos en un rango que desde valores mayores a 45 y menores que 49.5 se pueden definir los siguientes intervalos:



Los intervalos se denominan **intervalos de clase** o "**clases**".

Histogramas

Utilizar muy pocos intervalos "anchos" amontonará los datos, mientras que muchos intervalos "angostos" pueden distribuir los datos sobre demasiados intervalos con pocas observaciones.

Ninguno de los dos casos anteriores va a brindar una idea de cómo se distribuyen los datos y se pueden perder algunas características especiales de la distribución de los datos.

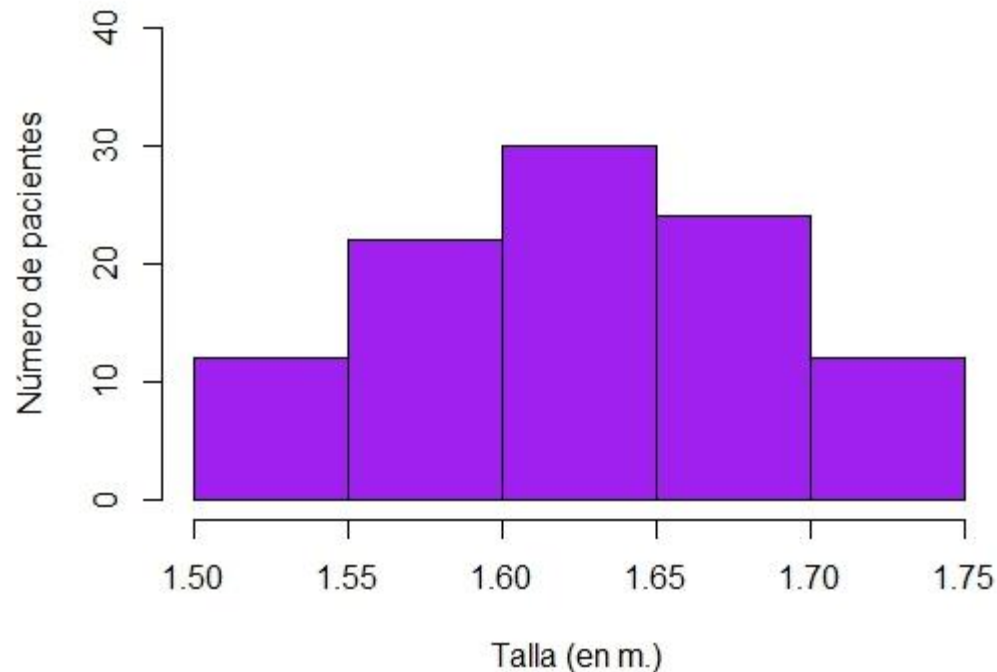
Si bien no existen reglas para seleccionar el número o la longitud de los intervalos, usualmente se obtiene un valor aproximado de la cantidad de intervalos mediante: $\sqrt{n}=m$.

La amplitud de los intervalos puede calcularse en forma aproximada como: $(\text{máx.}-\text{mín.})/m$.

.....continuando con ejemplo motivador

```
hist(TALLA, main="", xlab="Talla (en m.)",  
ylab="Número de pacientes", ylim=c(0,40), col="purple")
```

En R...



Una forma de presentar la información resumida es a través de la Tabla de distribución de frecuencias

Variable: Talla (en mts) de las mujeres....

Clases	f_j	h_j	$h_j\%$	F_j	H_j	$H_j\%$

Histogramas

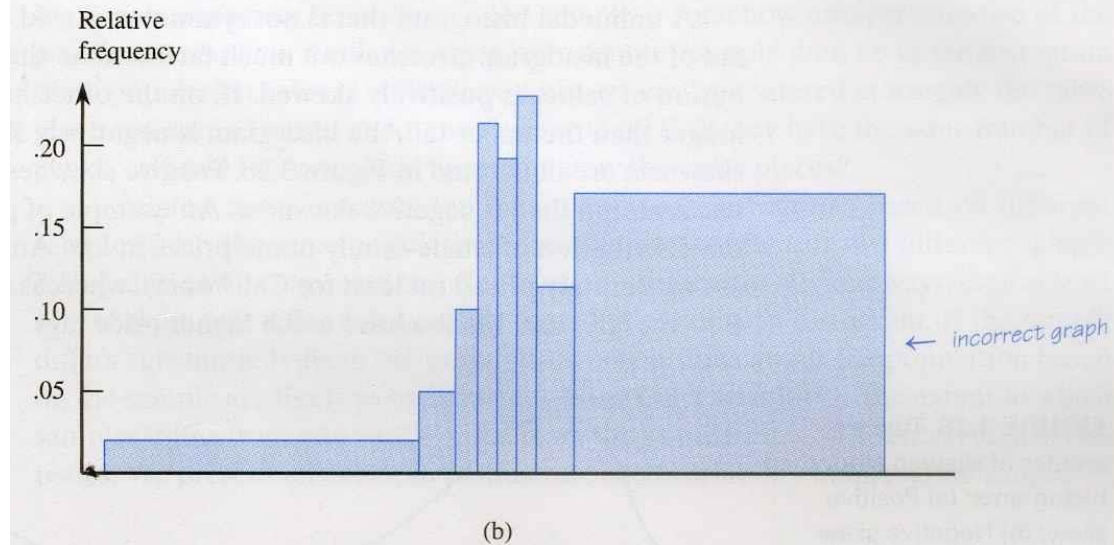
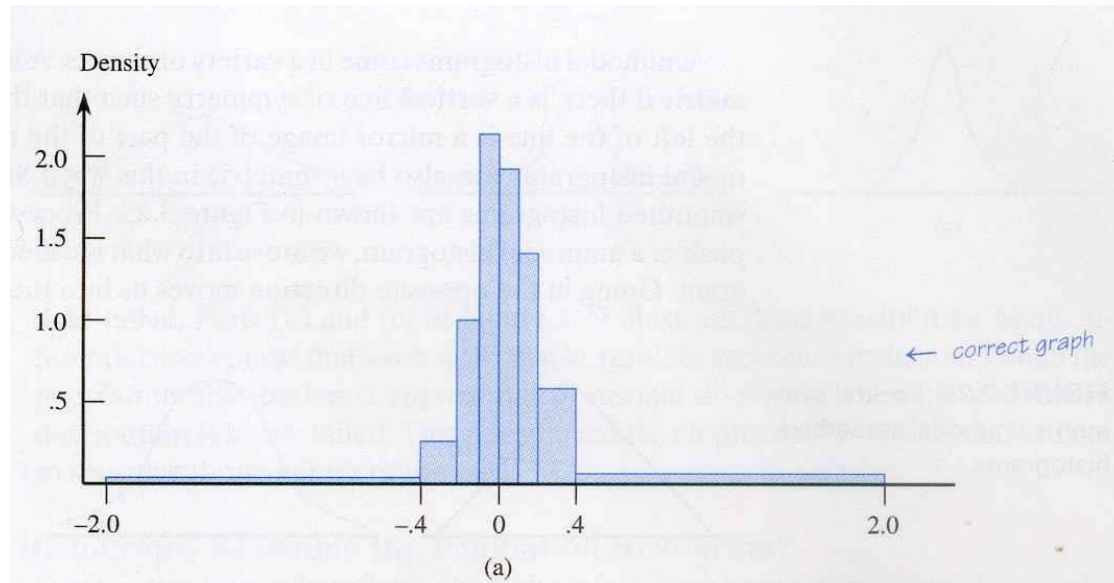
Intervalos de clase de ancho desigual

En algunas oportunidades, una buena alternativa es utilizar pocas clases (relativamente anchas) en los extremos e intervalos angostos en el centro de la distribución.

Es decir, **intervalos de clase de distinta amplitud**.

En este caso, en el eje vertical debe utilizarse la "**densidad**" de la clase. Esto asegura que el área de cada rectángulo sea proporcional a la frecuencia relativa correspondiente.

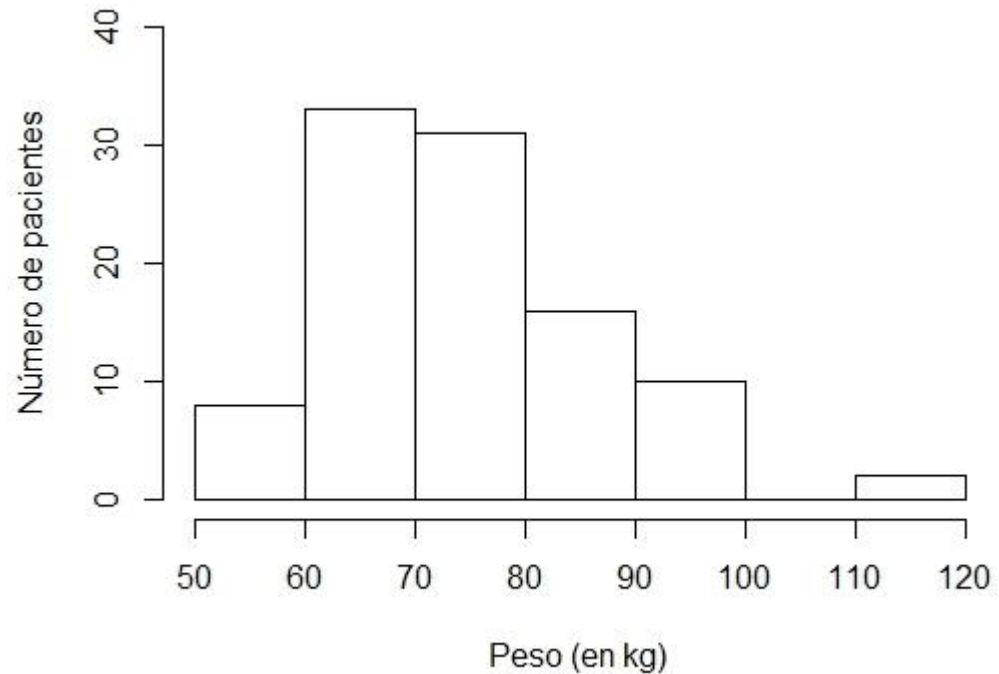
$$\text{densidad} = \text{altura del rectángulo} = \frac{\text{frecuencia relativa}}{\text{ancho de la clase}}$$



En ejemplo 1

```
hist(PESO, main="", xlab="Peso (en kg)",  
      ylab="Número de pacientes", ylim=c(0,40))
```

En R...

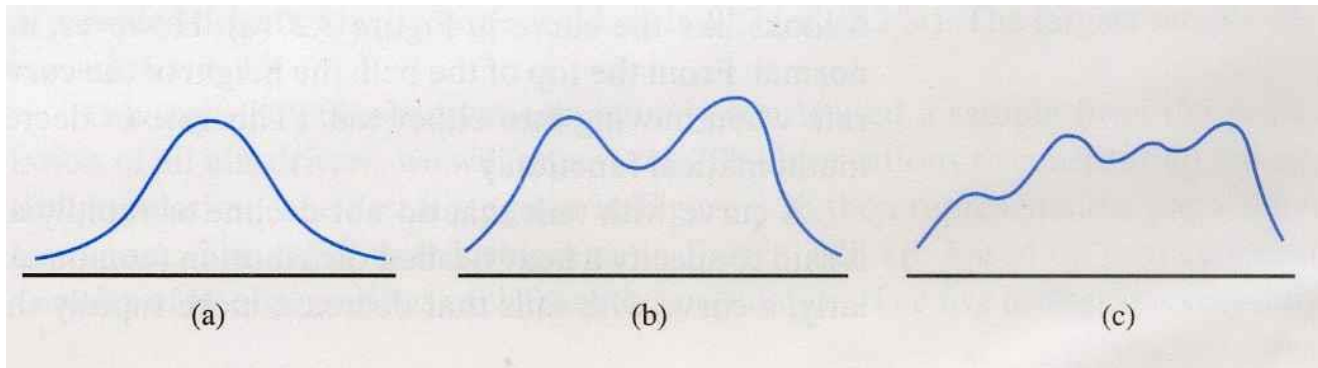


*Sería recomendable
trabajar con un intervalo
de 100 a 120, con
amplitud diferente al resto.*

Formas de los histogramas

La forma general de un histograma es una característica importante. Una caracterización de la forma general se relaciona con el número de "picos".

Un histograma es **unimodal** si tiene un único pico, **bimodal** si tiene dos picos y **multimodal** si tiene más de dos picos.

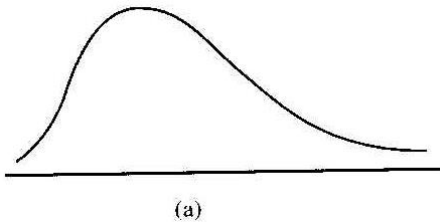


(b) los datos corresponden a 2 grupos diferentes de individuos u objetos.

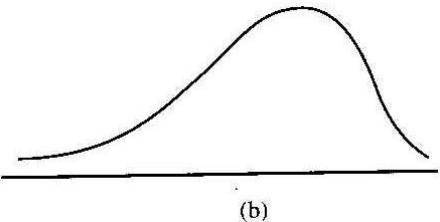
Formas de los histogramas

Un histograma es **simétrico** si existe una línea vertical de simetría tal que la parte del histograma a la izquierda de dicha línea es un reflejo de la parte ubicada a la derecha.

Un histograma unimodal que no es simétrico es **asimétrico**.



- Si la cola superior del histograma se prolonga más que la cola inferior, la distribución de los valores es **asimétrica positiva** o **a la derecha**.



- Si la cola inferior es mayor que la cola superior, la distribución es **asimétrica negativa** o **a la izquierda**.

Medidas descriptivas para variables categóricas

Entre las medidas descriptivas se pueden distinguir:

- Las frecuencias relativas para las distintas categorías.
- La moda o categoría modal.
Es la categoría que presenta la frecuencia más alta.

Frecuencias relativas

Las medidas resumen para datos categóricos son las frecuencias relativas para las distintas categorías.

Cuando sólo existen dos categorías (variable **dicotómica**), es común asignar a una de las categorías el rótulo "éxito" y reportar solamente la proporción de éxitos en la muestra.

La **proporción muestral de éxitos** se define como:

$$\hat{p} = \frac{\text{número de éxitos en la muestra}}{n}$$

donde "éxitos" representa una de las dos categorías de la variable dicotómica.

...continuando con ejemplo motivador

Primerizas	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa porcentual
Sí	59	0.59	59
No	41	0.41	41
Total	100	1.00	100

La muestra contiene 59 mujeres primerizas, por lo tanto:

$$\hat{p} = 59 / 100 = 0.59$$

Medidas descriptivas para variables cuantitativas

Entre las medidas descriptivas se pueden distinguir:

- Medidas de posición central.
Las dos medidas más populares para describir el centro de una distribución son la media (o promedio) y la mediana.
- Medidas de dispersión o variabilidad.

Medidas de posición central

Sea:

X = variable bajo análisis.

n = número de observaciones muestrales (**tamaño muestral**).

x_1 = primera observación muestral.

x_2 = segunda observación muestral.

x_n = n -ésima (última) observación muestral.

Media muestral

La **media muestral** se define como la sumatoria de todos los valores muestrales dividida el tamaño muestral.

Se puede interpretar como el punto de balance de la distribución.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Un problema potencial de la media como medida de posición central es que su valor puede estar muy influido por la presencia de observaciones extremas (*outliers*; aquellas observaciones que son inusualmente pequeñas o inusualmente grandes).

...continuando con el ejemplo motivador

En R...

```
> mean(TALLA)  
[1] 1.6347
```

```
> mean(PESO)  
[1] 75.46
```

Los valores del peso (kg) son todos enteros, sin embargo la media se reporta con decimales. Es común utilizar más dígitos de precisión decimal cuando se reporta la media.

Mediana muestral

Luego de que los datos son ordenados de menor a mayor, la mediana es el valor central, el cual divide al conjunto de datos en dos partes iguales.

Mediana muestral {
Si n es **impar** es el valor central del conjunto de datos.
Si n es **par** es el promedio de los dos valores centrales.

Modificaciones en la vía aérea de pacientes en trabajo de parto...

En R...

```
> median(TALLA)
[1] 1.645

> median(PESO)
[1] 73
```

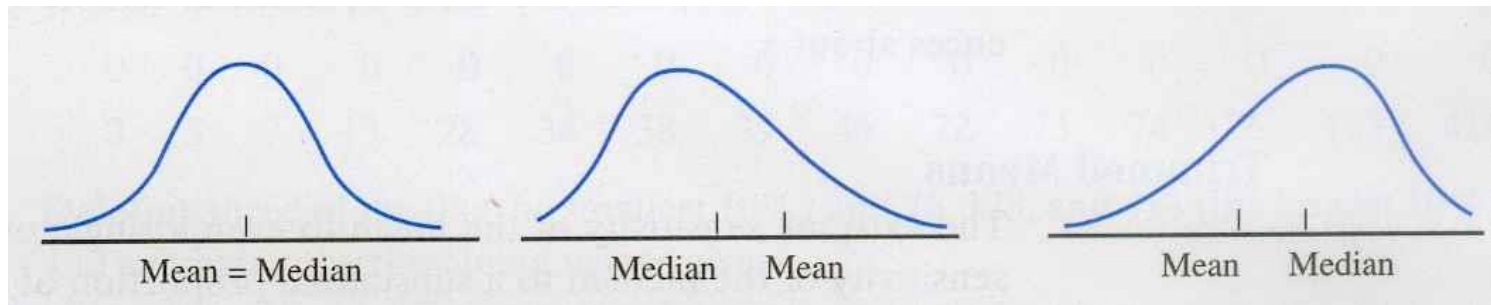
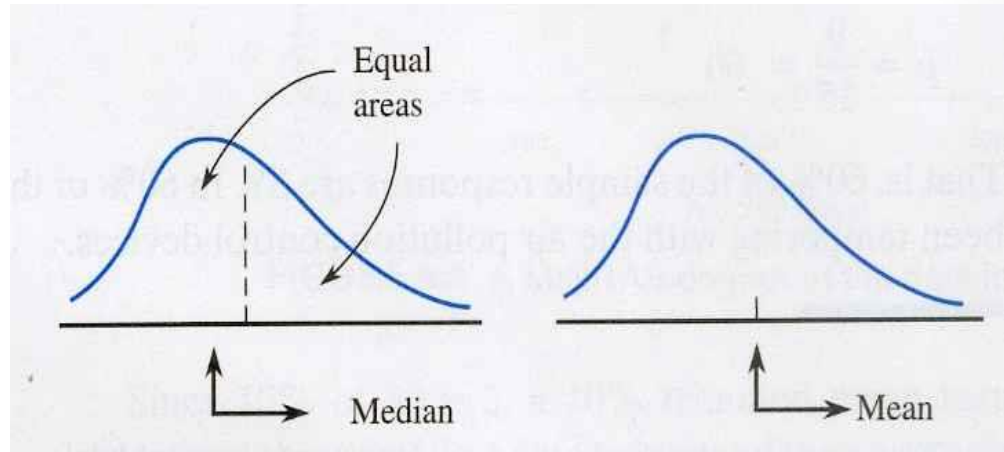
Dado que $n=100$, la mediana es el promedio de las observaciones en las posiciones centrales (50° y 51°) en el conjunto de datos ordenados (*se puede ver en el diagrama de tallo y hoja, por ej.*).

Comparar medias y medianas para ambos casos, ver cuál es más representativo, observar histograma o diagrama de tallo y hoja.

Media vs mediana

Si existen valores extremos, la media se desplaza hacia dichos valores mientras que la mediana es insensible a ellos, por lo tanto es recomendable informar la mediana ya que representará mejor lo que es “típico”.

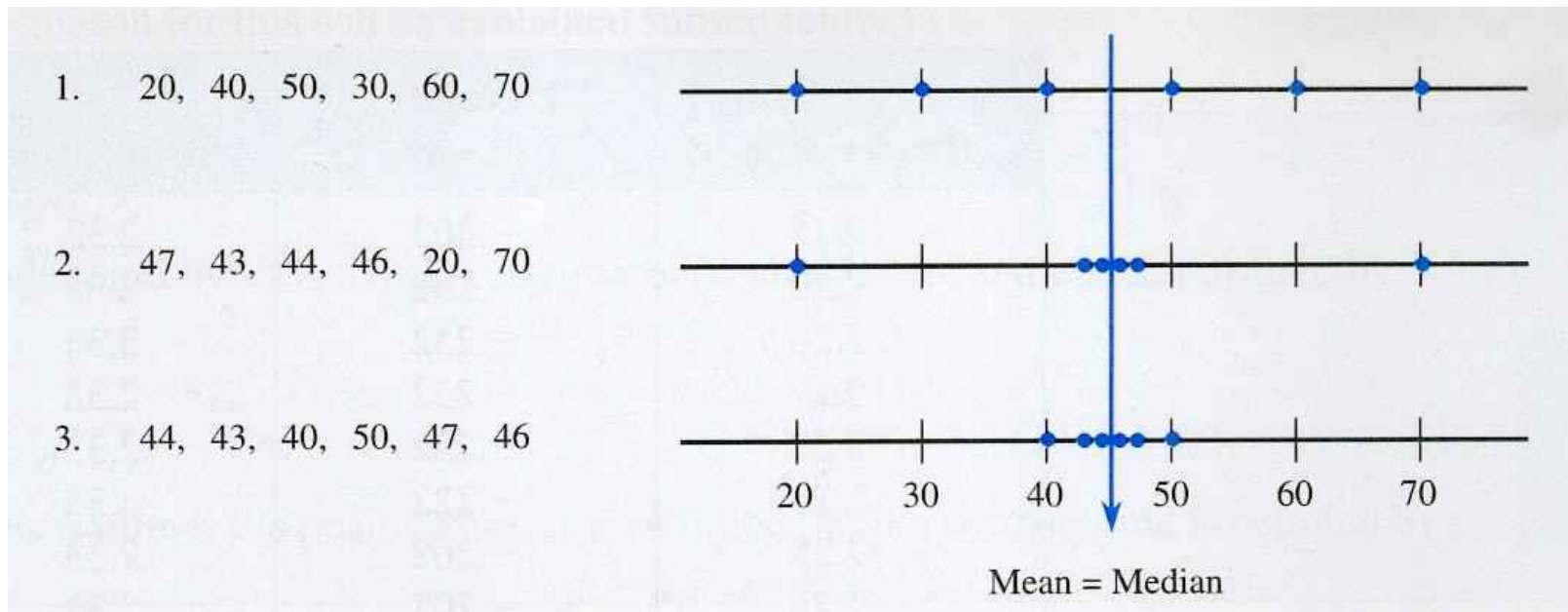
Si la distribución es aproximadamente simétrica o si los datos son discretos con pocos valores distintos se prefiere la media ya que utiliza los valores numéricos de todas las observaciones.



Medidas de variabilidad

Una medida de posición central solo brinda información parcial.

Es importante describir también la dispersión de los valores alrededor del centro.



Medidas de variabilidad

La medida de variabilidad más simple es el **rango** = máx.-mín.

En general, mayor variabilidad se reflejará en un rango mayor.

Sin embargo, la variabilidad no sólo depende de los valores extremos; es una característica del conjunto de datos completo, cada observación contribuye.

En la figura anterior, las dos primeras muestras tienen el mismo rango pero existe menos dispersión en la segunda.

Medidas de variabilidad

Las medidas más comunes de variabilidad describen el grado en el cual las observaciones muestrales se alejan de la media muestral.

Substrayendo la media muestral de cada observación se obtiene un conjunto de **desviaciones de la media**.

Una desviación es positiva si los valores de x exceden \bar{x} y es negativa si los valores de x son menores que \bar{x} .

Generalmente, a mayor variabilidad en la muestra, mayor es la magnitud (ignorando el signo) de las desviaciones.

Medidas de variabilidad

Una forma de resumir las desviaciones de la media en una única medida podría ser promediarlas a través de toda la muestra.

Sin embargo, salvo por efectos de redondeo, la suma de las desviaciones de la media siempre es 0 y por lo tanto su promedio es 0, y no puede ser utilizada como medida de variabilidad.

La forma de prevenir que desviaciones negativas y positivas se compensen es elevándolas al cuadrado antes de combinarlas.

Variancia y desvío estándar muestral

La **variancia muestral**, s^2 , es la suma de las desviaciones de la media al cuadrado dividida por $(n - 1)$.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

El **desvío estándar muestral**, s , es la raíz cuadrada positiva de la variancia muestral.

Una gran variabilidad en la muestra se reflejará en un valor relativamente grande de s^2 o s .

El desvío estándar se expresa en la misma unidad de medida que la variable original.

..continuando con el ejemplo motivador

En R...

```
> min(TALLA)
[1] 1.5
> max(TALLA)
[1] 1.75
> range(TALLA) [1] 1.50 1.75
> (rango<-max(TALLA)-min(TALLA))
[1] 0.25
```

```
> var(TALLA)
[1] 0.003742333
> sd(TALLA)
[1] 0.06117461
```



```
dif <- TALLA - mean(TALLA)
var <- sum((dif)^2)/n
var
```


Variancia y desvío estándar muestral

Un desvío estándar puede ser interpretado informalmente como la magnitud de una desviación "típica" de la media.

En el ejemplo, una desviación típica de la media 130.38 es 11.47. Algunas observaciones estarán más cerca de 130.38 y otras más alejadas.

Notar que no se señaló si este valor representa una gran o una pequeña variabilidad. Por ahora, se utilizará sólo con fines comparativos y no como una medida de variabilidad absoluta.

Si en otra muestra de rangos de movimiento se obtiene $s=9.1$, se podría concluir que existe más variabilidad en la primera muestra que en la segunda.

Rango intercuartil

Al igual que la media, el valor de s resulta muy afectado por la presencia de observaciones extremas.

El **rango intercuartil** es una medida de variabilidad resistente a los efectos de los *outliers*.

Se basa en cantidades denominadas **cuartiles**.

El **1º cuartil (Q1)** separa el 25% inferior del conjunto de datos del 75% superior.

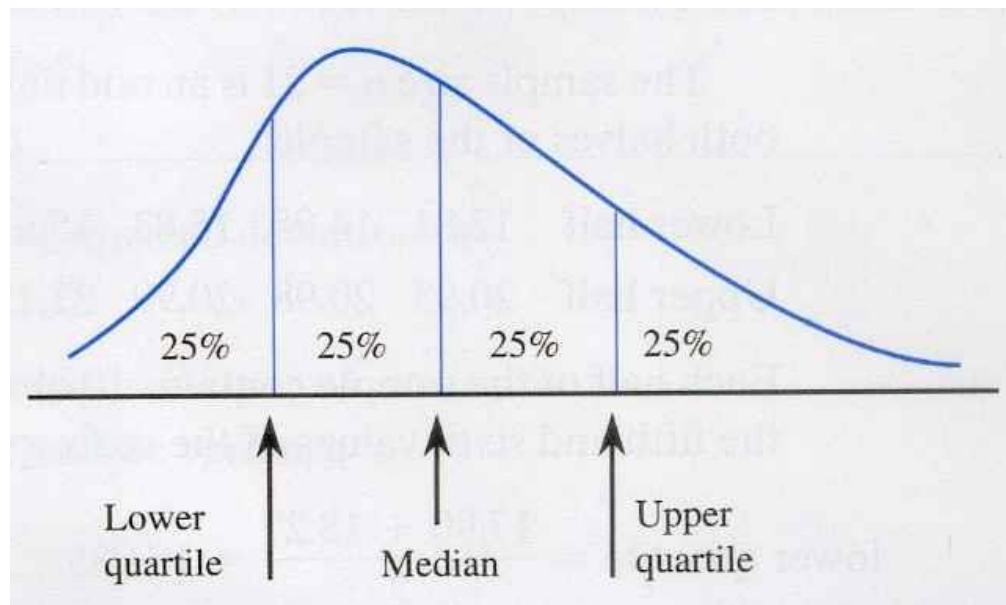
El **2º cuartil (Q2)** es la **mediana** la cual separa el 50% inferior del 50% superior.

El **3º cuartil (Q3)** separa el 25% superior del conjunto de datos del 75% inferior.

Cuartiles y percentiles

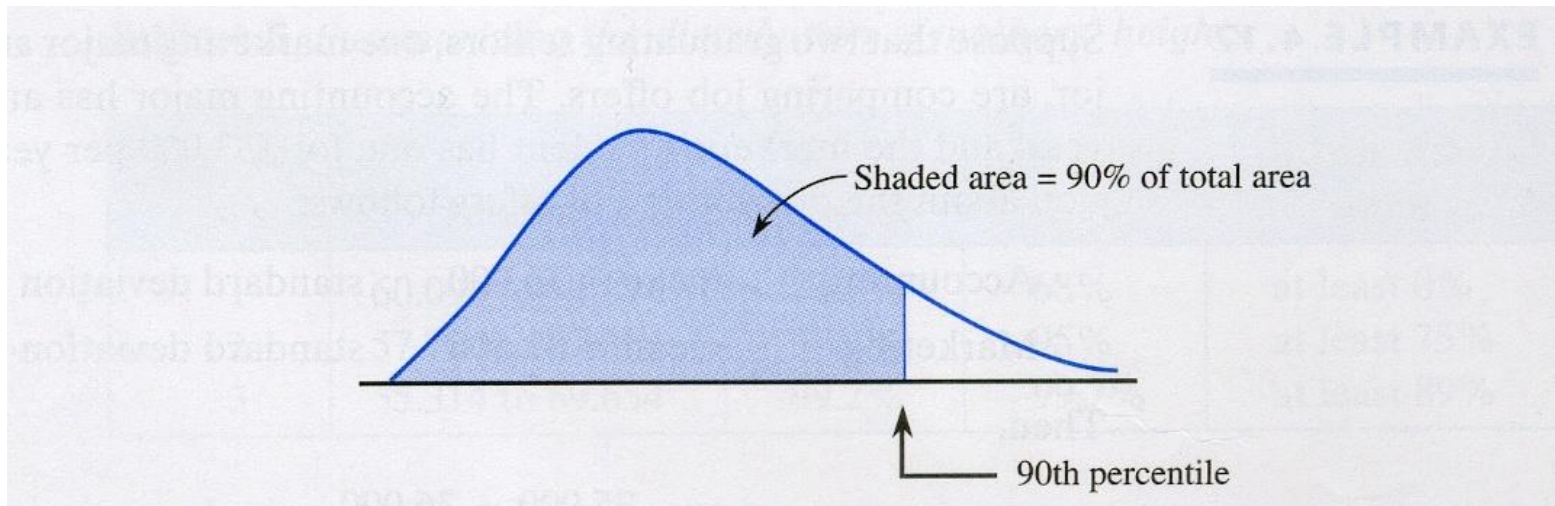
Los cuartiles de una muestra se obtienen dividiendo las n observaciones ordenadas en una mitad inferior y una mitad superior; si n es impar la mediana se excluye de ambas mitades.

Los dos cuartiles extremos son las medianas de las dos mitades.



Cuartiles y percentiles

Para cualquier número r entre 0 y 100, el **percentil del $r\%$** es un valor tal que el $r\%$ de las observaciones son menores o iguales que dicho valor.



Rango intercuartil

El rango intercuartil se define como:

$$RI = 3^{\circ} \text{ cuartil} - 1^{\circ} \text{ cuartil}$$

La naturaleza resistente del RI se deriva del hecho de que hasta el 25% de las observaciones más pequeñas de la muestra y hasta el 25% de las observaciones más grandes pueden hacerse más extremas sin que el RI resulte afectado.

Algunos cuantiles de interés para la variable talla del ejemplo motivador:

```
> quantile(TALLA)
0% 25% 50% 75% 100%
1.500 1.600 1.645 1.680 1.750
> IQR(TALLA)
[1] 0.08
> quantile(TALLA, c(0.10,0.90))
10% 90%
1.550 1.711
```

En R...

```
> quantile(TALLA, c(0.10,0.90))
10% 90%
1.550 1.711

> summary(TALLA)
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.500 1.600 1.645 1.635 1.680 1.750
```

Boxplot o Diagrama de caja

Un boxplot es un diagrama que permite resumir un conjunto de datos brindando mas detalle que las medidas de posición y dispersión pero menos detalle que un diagrama de tallo y hoja.

El boxplot es una representación compacta que provee información acerca del centro, la dispersión y la simetría o asimetría de los datos.

Se basa en la mediana y el rango intercuartil, de manera que sólo se requiere que la variable a graficar esté medida en una escala ordinal.

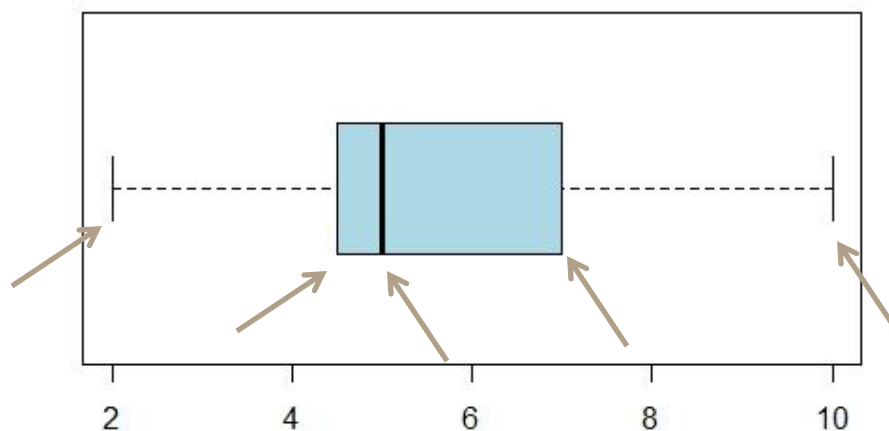
Algunas medidas descriptivas y boxplot de la variable Duración del TP (en hs)

```
> summary(DUR_TP_HS)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	4.500	5.000	5.765	7.000	10.000

En R...

Duración del trabajo de parto (hs)



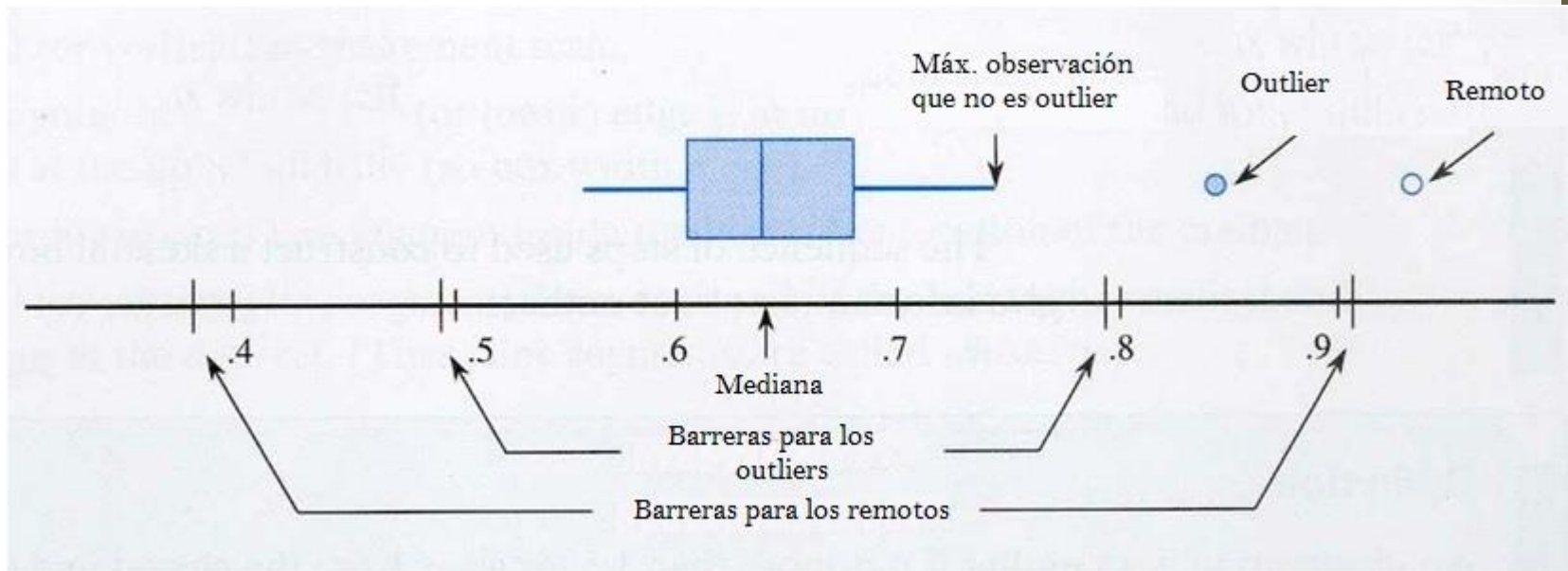
Boxplot modificado

Un boxplot modificado utiliza líneas que se extienden desde los extremos de la caja (cuartiles) hasta las observaciones más extremas que no son outliers (valores adyacentes). Los outliers se identifican con algún símbolo como puntos o asteriscos.

Una observación es un **outlier** si se encuentra a más de 1.5 veces el RI del extremo más cercano de la caja.

Un outlier se considera **remoto** si se encuentra a más de 3 veces el RI del extremo más cercano de la caja en caso contrario

Boxplot modificado

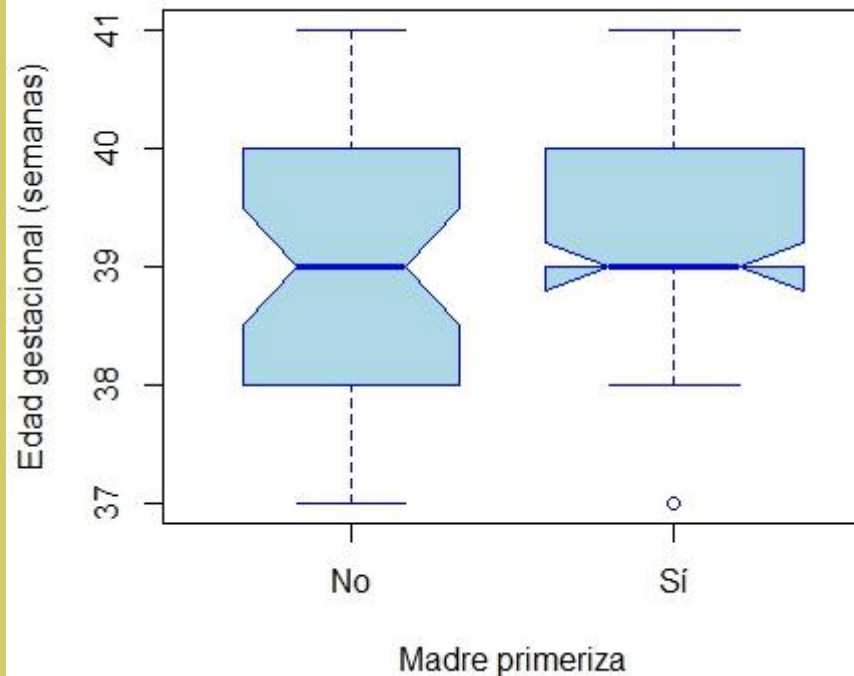


Otros tipos de Boxplot

Edad gestacional según gestas previas

```
boxplot(EDAD_GEST~PRIMERIZA,  
xlab="Madre primeriza",  
ylab="Edad gestacional  
(semanas)",  
names=c("No", "Sí"),  
varwidth=TRUE,  
border="blue", notch=TRUE,  
font=4,  
boxfill="lightblue")
```

En R...



Comentarios acerca del análisis descriptivo

- ✓ Es fundamental definir correctamente la variable que está siendo analizada y determinar cuál es su escala de medida para poder resumirla de forma correcta.
- ✓ El gráfico elegido es apropiado para el tipo de variable analizada?
- ✓ Las medidas de resumen elegidas son las apropiadas?
- ✓ Cómo describiría la forma de la distribución, y qué se podría decir acerca de la variable analizada?
- ✓ Existen *outliers* en el conjunto de datos? Existe una explicación convincente de por qué pueden diferir del resto de los datos?
- ✓ Dónde se ubica la mayoría de los datos?Cuál es un valor típico para la variable?
- ✓ Existe mucha variabilidad en los datos?
- ✓ Tener en cuenta que tanto los gráficos como las tablas deben tener un título que responda las preguntas: qué? Cómo? Dónde? Cuándo?.

Bibliografía

- Agresti A., Franklin C. (2009) . The art and science of learning from data. 2° Ed. New Jersey. Pearson Prentice Hall.
- Ruggieri M. , Arnesi N(2010) . Métodos estadísticos I. Reimpresión. Rosario. UNR editora.