INTRODUCCIÓN

- La Estadística es una ciencia en la cual se hacen inferencias sobre ciertos fenómenos aleatorios en base a material obtenido en muestras.

- El campo de la estadística tiene fundamentalmente dos áreas principales: la matemática y la aplicada.

- La estadística matemática se ocupa de desarrollar nuevos métodos de inferencia estadística y requiere del conocimiento de conceptos matemáticos, en algunos casos complejos, para comprender su implementación.

- La estadística aplicada utiliza los métodos desarrollados por los matemáticos en áreas como la economía, la medicina, la psicología o la biología. En particular la bioestadística es la rama de la estadística aplicada que se ocupa de problemas médicos y biológicos.

Si por ejemplo se quiere controlar el funcionamiento de los aparatos digitales para tomar la presión.

- ¿tienen diferencias con los métodos tradicionales?

- ¿cómo podríamos realizar esta comparación?

- ¿podríamos basarnos en uno solo de estos aparatos?

- ¿a cuántas personas deberíamos tomarle la presión con cada uno de estos aparatos?

- ¿deberíamos medir la presión en forma tradicional antes o después de hacerlo con el aparato digital?

- ¿qué otra información sería importante tener de las personas a quienes les tomamos la presión? ¿cómo chequeamos la precisión de los aparatos digitales?

- Una vez que hayamos decidido de que manera realizar el estudio el primer paso es presentar los datos de forma tal de resumir de alguna forma clara la información recopilada, esto es lo que se llama estadística descriptiva.

- Este material descriptivo puede ser numérico o gráfico.

- Desde el punto de vista numérico se deben dar ciertas medidas resumen que pueden presentarse en forma de tabla, o como una distribución de frecuencias, que presenta cada uno de los valores obtenidos en la muestra y la cantidad de veces que ocurre.

- El tipo de gráfico adecuado depende del tipo de variable que estemos midiendo

ESTADÍSTICA DESCRIPTIVA

Población **---------->** Muestra

<----------
Inferencia

Población: conjunto total de los sujetos o unidades de análisis de interés en el estudio.

Muestra: cualquier subconjunto de sujetos o unidades de análisis de la población en estudio.

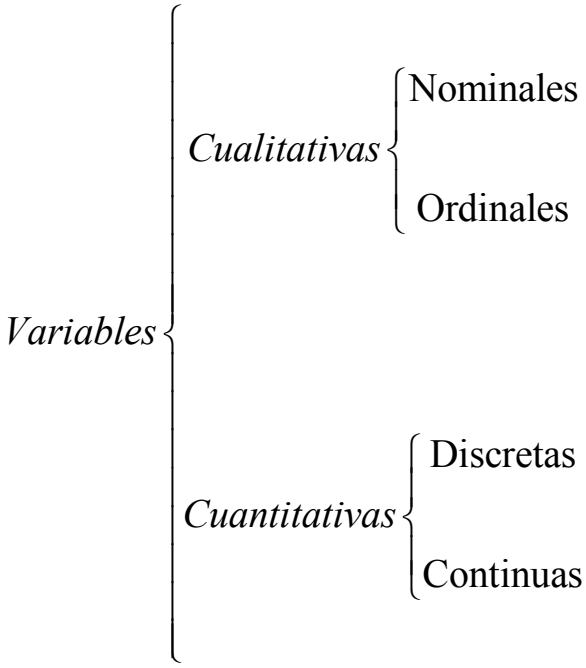
Muestra aleatoria: muestra obtenida a través de un mecanismo tal, que todas las muestras posibles tienen la misma probabilidad de ocurrir. Si se trata de una población de individuos, esto equivale a pedir que todos los individuos tengan igual probabilidad de ser seleccionados.

¿Qué observamos?

Cualquier característica de un individuo u objeto que nos resulte de interés y la expresaremos numéricamente en una variable.

Variable **---------->** Dato
sujeto

VARIABLE: cualquier característica de la unidad de observación que interese registrar y que en el momento de ser registrada puede ser transformada en un número.



Variables Cuantitativas: El resultado de la medición u observación es un número. Se refieren a una cantidad en la que importa el orden y la magnitud.

- **Discretas**: sólo pueden tomar un conjunto finito o infinito numerable de valores.

- **Continuas**: corresponden a una medición que se expresa en unidades y pueden tomar infinitos valores dentro de un intervalo de números reales.

Ejemplos de variables discretas pueden ser: n° de hijos, n° de personas que se contagian de gripe porcina en un país, n° de ratas que responden a un tratamiento, n° de camas disponibles en un hospital.

Ejemplos de variables continuas: altura, peso, volumen de sangre, tiempo hasta que hace efecto un tratamiento.

- En muchos casos si bien la variable de interés es continua, los aparatos de medición no tienen la precisión suficiente o el interés del estudio no necesita de tal precisión y es conveniente discretizarlas.

- Por ejemplo la edad es una variable continua pero en muchos casos sólo nos interesa la edad en años, es decir nos basta con saber que la persona tiene 28 años y no necesitamos saber si son 28 años, 4 meses, 6 días ,3 horas, 2 minutos, etc...

DISTRIBUCIÓN DE FRECUENCIAS

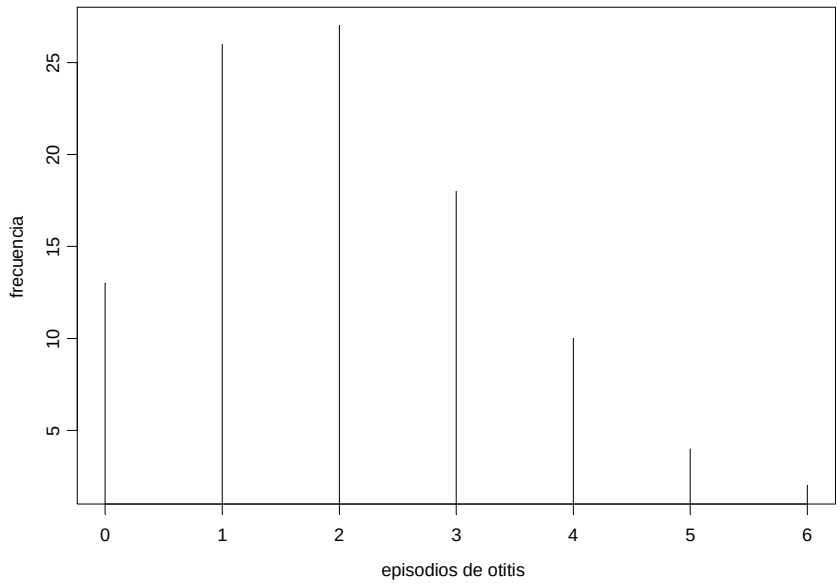
En el caso de una variable discreta para cada valor que toma la variable de interés podemos registrar el número de veces que se repitió dicho valor en la muestra, es decir la frecuencia de dicho valor.

- Supongamos que nos interesa el número de episodios de otitis media que presenta un niño en sus primeros dos años de vida en la Ciudad de Buenos Aires.

- La población estaría integrada por todos los niños de 2 años de la ciudad de Buenos Aires y la variable de interés sería el números de otitis que registró en sus dos años de vida.

- Se tomaron 100 niños al azar dentro de esa población y resumimos la información obtenida en la siguiente tabla:

Nº de otitis	**frecuencia**
0	13
1	26
2	27
3	18
4	10
5	4
6	2



HISTOGRAMAS

- Si la variable es cuantitativa, dividimos el rango donde viven los datos en **intervalos o clases**, que no se superpongan.

- Las clases deben ser **excluyentes** y **exhaustivas,** esto significa que no hay intersección entre dos clases y la unión de todas las clases cubre todo el rango.

- En cada clase medimos la frecuencia.

También podemos usar para cada intervalo la

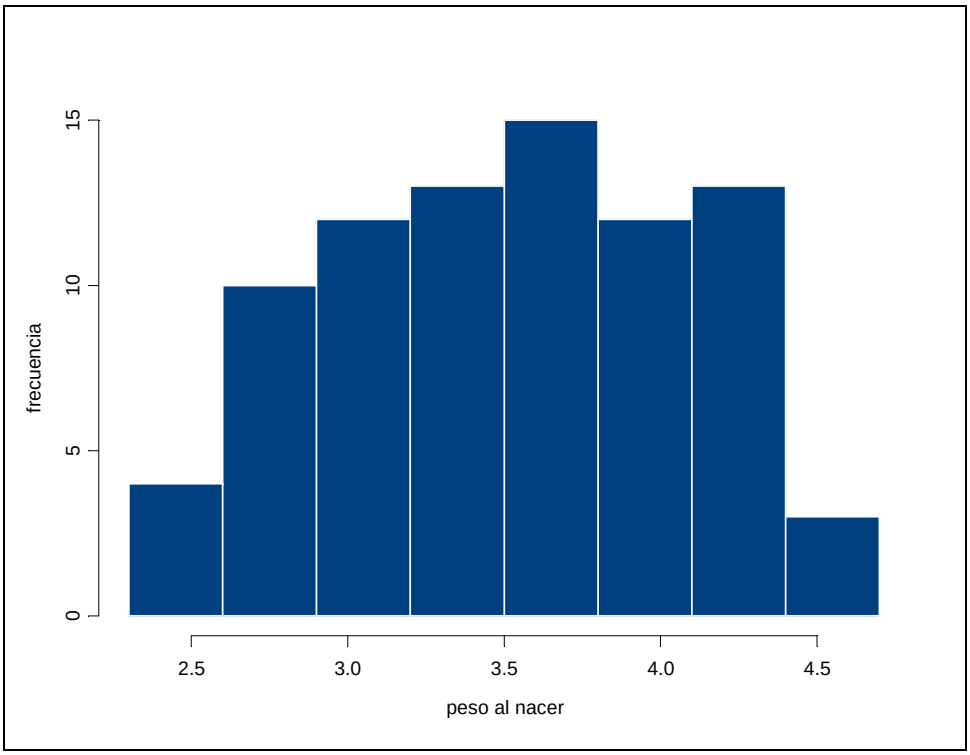
$$\text{frecuencia relativa} = \frac{\text{frecuencia}}{\text{cantidad total de datos}}$$

- Si la población de interés está integrada por todos los niños que nacen en la Argentina y la variable de interés es el peso al nacer, se toma una muestra de 82 de esos niños al azar y se registran los pesos en kilos que tenían al nacer.

- Se consideraron intervalos de longitud 0.3 y se calculó la frecuencia absoluta y relativa de cada intervalo.

Peso al nacer	frecuencia	Frecuencia relativa
[2.3 – 2.6)	4	0.0488
[2.6 – 2.9)	10	0.1220
[2.9 – 3.2)	12	0.1463
[3.2 – 3.5)	13	0.1585
[3.5 – 3.8)	15	0.1829
[3.8 – 4.1)	12	0.1463
[4.1 – 4.4)	13	0.1585
[4.4 – 4.7)	3	0.0367

Se grafican rectángulos cuya altura es la frecuencia o la frecuencia relativa.

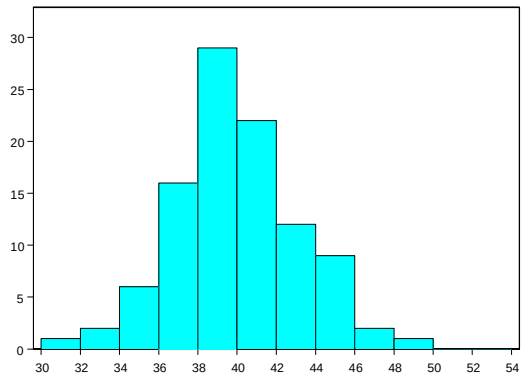


¿Qué forma puede tener un histograma?

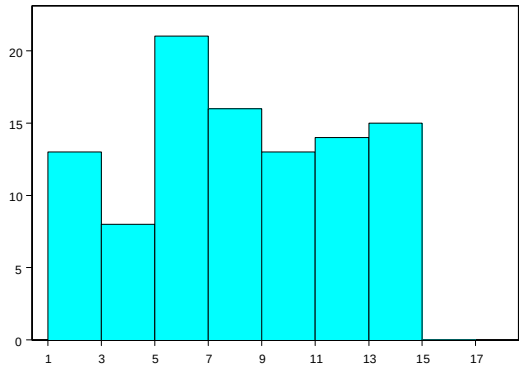
Un aspecto a tener en cuenta en la distribución de los datos es la simetría. Un conjunto de datos que no se distribuye simétricamente, se llama **asimétrico**.

En los siguientes gráficos mostramos algunas de las formas posibles que puede tener un histograma:

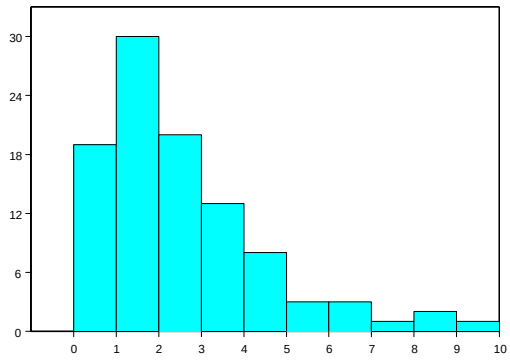
Distribución acampanada



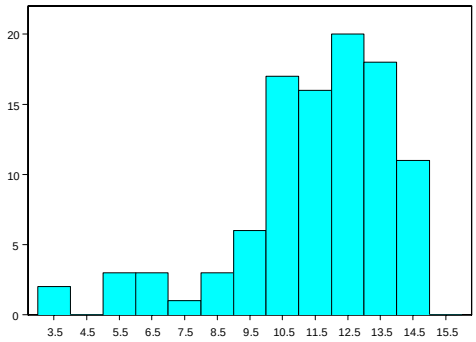
Distribución uniforme



Asimetría derecha



Asimetría izquierda



Intervalos de distinta longitud

- Cuando los intervalos no son de la misma longitud graficar los rectángulos del histograma con la altura de la frecuencia o de la frecuencia relativa puede ser engañoso.

- En ese caso es aconsejable utilizar rectángulos de **área** igual a la frecuencia relativa.

- Es recomendable tomar

$$\text{altura del rectángulo} = \frac{\text{frecuencia relativa}}{\text{longitud del intervalo}}$$

- A esta altura del rectángulo se la llama escala densidad ya que indica el número de datos por unidad de la variable.

- De esta manera el área total de los rectángulos es 1 y dos histogramas son fácilmente comparables independientemente de la cantidad de observaciones en las que se basa cada uno.

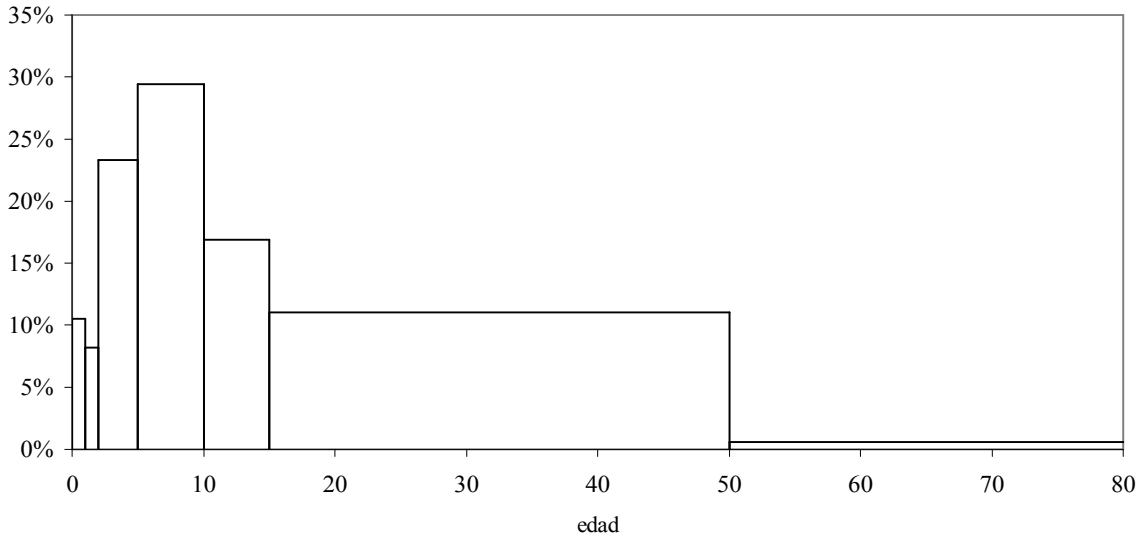
Si la población de interés está integrada por todos los habitantes del país que en el año 2000 tuvieron rubéola y la variable de interés es la edad de esas personas podemos utilizar como muestra los casos de rubéola notificados al SINAVE durante el año 2000 según grupos de edad. Notemos que los intervalos de edad tienen diferente longitud.

Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE

Intervalo (años)	**Frecuencia (f_i)**	**Frecuencia relativa (f_r)**	**Escala densidad**
[0, 1)	497	10.5%	10.53%
[1, 2)	387	8.2%	8.20%
[2, 5)	1100	23.3%	7.77%
[5, 10)	1389	29.4%	5.89%
[10, 15)	798	16.9%	3.38%
[15, 50)	521	11.0%	0.32%
≥ 50	28	0.6%	0.01%
Total	4720	100.00%	-

- Si erróneamente se construye un histograma considerando como altura de la barra la frecuencia relativa se obtiene la gráfica siguiente.

- La última categoría de edad se truncó arbitrariamente en 80 años para poder representarla.

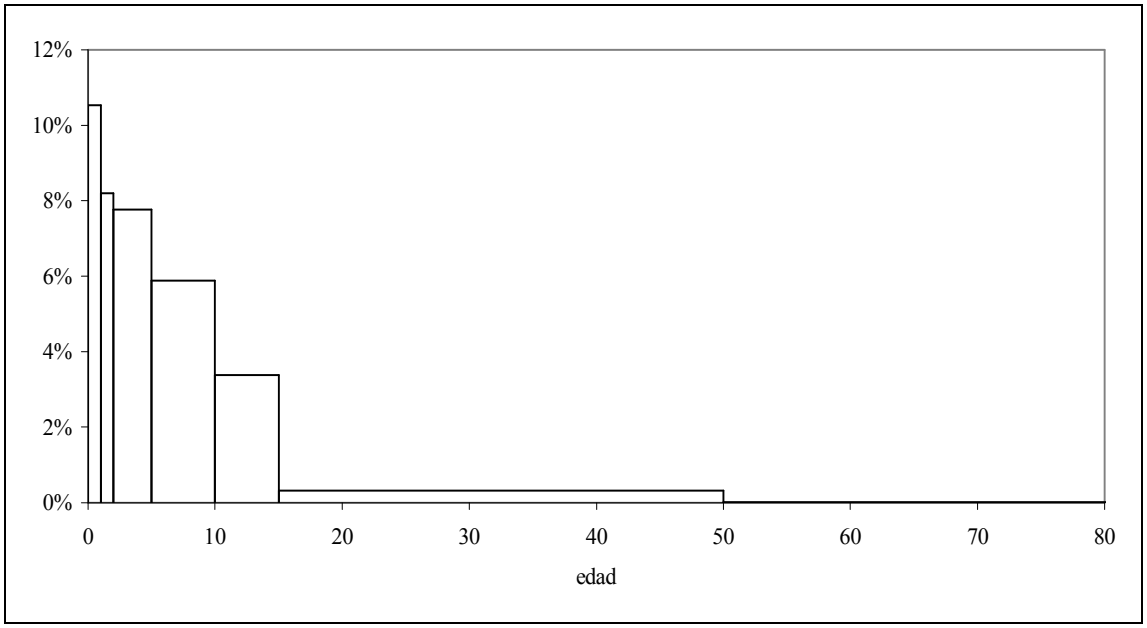


- A partir de este gráfico concluiríamos que la proporción de casos es notablemente mayor en los grupos de 2 a 5 años, de 5 a 10 años o de 10 a 15 años que en los grupos de menores de 1 año o de 1 a 2 años.

- Además, la proporción de casos en el grupo de 15 a 50 años impresiona como notable.

- El problema es que en la imagen visual asociamos la frecuencia de casos con el área de la barra, por ello parece haber más notificaciones de gente de 15 a 50 que de cualquier otro grupo de edad.

Histograma usando escala densidad. Notificaciones de casos de rubéola. Argentina, año 2000. Fuente: SINAVE



- En este gráfico, el porcentaje de casos de rubéola notificados para cada grupo está representado en el área de cada rectángulo.

- Si tuviéramos individuos notificados por rubéola parados en cada grupo etáreo, la altura del histograma representaría el aglutinamiento en cada clase: hay partes del eje de abscisas que están más densamente pobladas que otras.

- El histograma muestra que una gran proporción de casos ocurre en menores de 1 año, y que la proporción desciende a medida que aumenta la edad.

- En este gráfico estamos representando la "densidad de notificaciones" por cada año de edad.

El problema básico en estadística puede verse de la siguiente manera.

- Consideramos una muestra de tamaño *n* de datos $x_1, x_2 \ldots, x_n$, donde x_1 corresponde al primer elemento de la muestra y x_n al último.

- Asumiendo que la muestra fue extraída de una población, ¿qué conclusiones o inferencias podemos hacer con respecto a la población a partir de la muestra?

- Antes de poder responder esta pregunta es importante dar medidas que resuman estos datos.

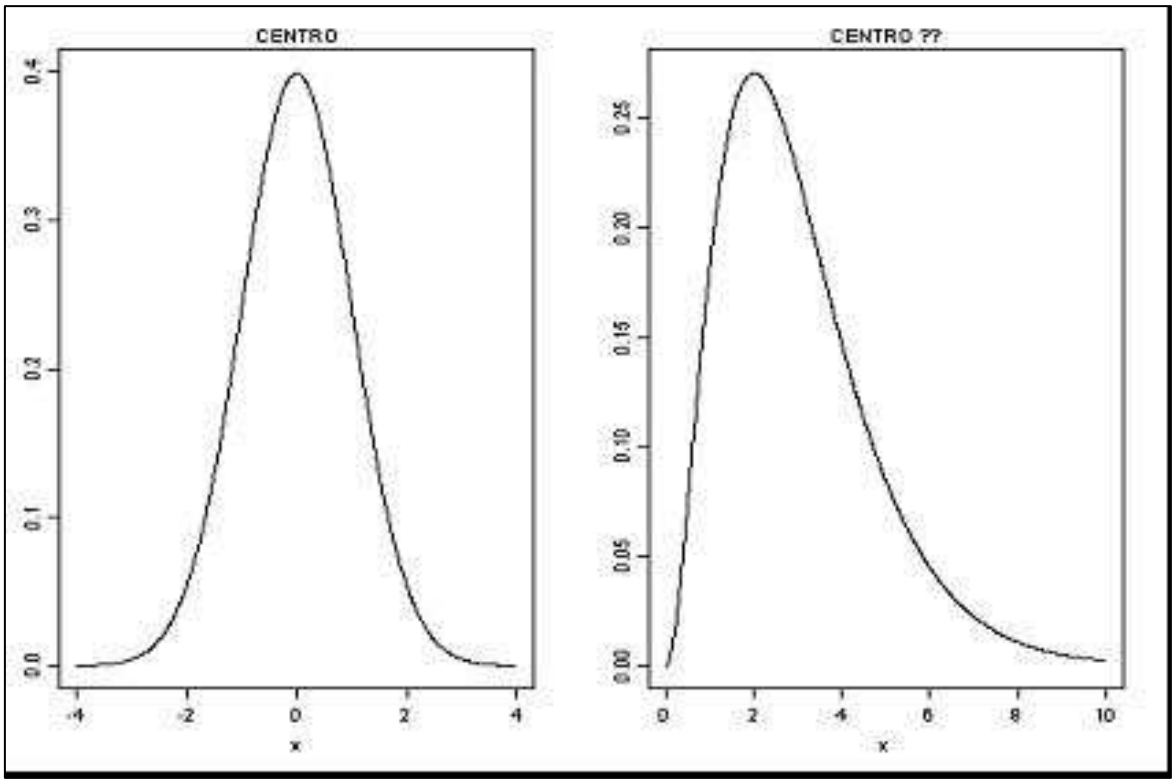
- Un tipo de medidas resumen de interés son las medidas centrales o de posición.

MEDIDAS DE POSICIÓN

- Puede parecer obvio como definir el valor medio de una muestra, sin embargo, cuanto más pensamos en ello menos obvio resulta.

- Una *medida de posición* es un número que pretende indicar dónde se encuentra el *centro* de la distribución de un conjunto de datos.

- Pero, **¿dónde se encuentra el "centro" de una distribución?**



Promedio o media muestral

Sumamos todas las observaciones y dividimos por el número total de datos.

$$\overline{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Representa el *centro de gravedad* o el punto de equilibrio de los datos. Podemos imaginar a los datos como un sistema físico, en el que cada dato tiene una "masa" unitaria y lo ubicamos sobre una barra en la posición correspondiente a su valor. La media representa la posición en que deberíamos ubicar el punto de apoyo para que el sistema esté en equilibrio.

- La suma de las distancias de los datos a la media es cero. Esta propiedad está relacionada con el hecho que la media es el centro de gravedad de los datos.

- La media muestral es en general, una medida natural de posición. Sin embargo, una de sus limitaciones es su sensibilidad frente a valores extremos (outliers). En ese caso

no va a resultar representativa de la mayoría de los datos de la muestra.

El maíz es un alimento importante para los animales pero carece de algunos aminoácidos que son esenciales. Un grupo de científicos desarrolló una nueva variedad que sí contenía niveles apreciables de dichos aminoácidos.

Se llevó a cabo el siguiente experimento: a un grupo de 20 pollos de 1 día se les suministró esta variedad de maíz mejorada y a otro grupo de 20 pollos (grupo de control) se lo alimentó de una forma que sólo se diferenciaba de la anterior en que no contenía harina de la variedad mejorada de maíz.

En este caso tenemos dos poblaciones, una es la de todos los pollos alimentados con la variedad normal de maíz y otra la de todos los pollos alimentados con la variedad mejorada.
En ambos casos la variable de interés es el peso que gana el pollo luego de ser alimentado 21 días con el maíz.

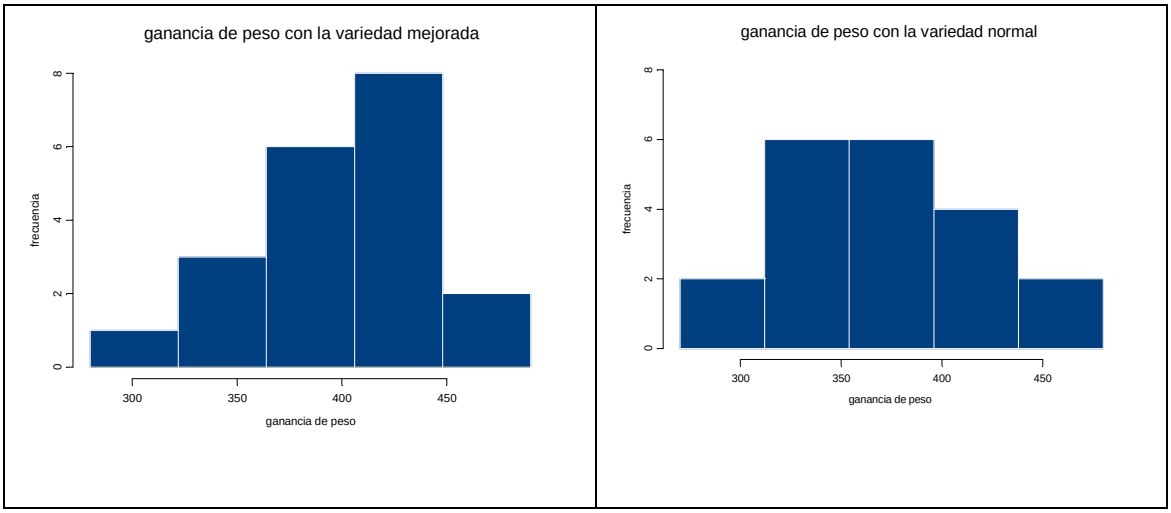
Los resultados que se obtuvieron sobre las ganancias de peso de los pollos (en gramos) al cabo de 21 días de alimentación fueron los siguientes:

• Variedad normal

380 321 366 356 283 349 402 462 356 410 329 399 350 384 316 272 345 455 360 431

• Variedad mejorada

361 447 401 375 434 403 393 426 406 318 467 407 427 420 477 392 430 339 410 326



Media muestral de la ganancia de peso con la variedad mejorada = 402.95
Media muestral de la ganancia de peso con la variedad normal = 366.30

- En el caso de la variedad mejorada, 9 de las ganancias de peso se encuentran antes de la media muestral y 11 después de la misma.

- En cambio en el caso de la variedad normal, 12 se encuentran antes y 8 después.

- Es posible que en casos extremos sólo una observación tome valores menores a la media y todas las restantes valores mayores a la misma.

- Es por eso que en algunos casos será conveniente definir otra medida de posición, la mediana.

Mediana muestral

- Es una medida del centro de los datos en tanto divide a la muestra ordenada en dos partes de igual tamaño.

- Es decir, el 50% de los datos toman valores menores o iguales a la mediana y el 50 % restante es mayor o igual que la mediana.

Para calcularla:

- Ordenamos los datos de menor a mayor.
- La mediana $\tilde{X}$ es el dato que ocupa la posición $\left(\frac{n+1}{2}\right)$ en la muestra ordenada.
- Si el número de datos es impar la mediana es el dato que ocupa la posición central y si el número de datos es par la mediana es el promedio de los dos datos centrales en la muestra ordenada.

- Si la distribución es simétrica la mediana y la media muestral identifican al mismo punto.

- Sin embargo, si la distribución de los datos es asimétrica, la media y la mediana diferirán según el siguiente patrón:

- Asimetría derecha (cola larga hacia la derecha) $\Rightarrow \overline{X} > \tilde{X}$

- Asimetría izquierda (cola larga hacia la izquierda) $\Rightarrow \overline{X} < \tilde{X}$

- La mediana es resistente a la presencia de datos atípicos.

Ordenamos los resultados obtenidos con los pollos para calcular la mediana en cada caso

• Variedad normal

272 283 316 321 329 345 349 350 356 356 360 366 380 384 399 402 410 431 455 462

• Variedad mejorada

318 326 339 361 375 392 393 401 403 406 407 410 420 426 427 430 434 447 467 477

La mediana en ambos casos es el promedio entre $x^{(10)}$ y $x^{(11)}$.

Mediana muestral de la ganancia de peso con la variedad mejorada = 406.5
Mediana muestral de la ganancia de peso con la variedad normal = 358

- En el caso de la variedad mejorada la mediana (406.5) es mayor que la media (402. 95), eso se debe a la asimetría a izquierda de ese conjunto de datos.

- En cambio en la variedad normal la mediana (358) es menor a la media (366.3), en este caso se observa una leve asimetría a derecha.

¿Qué pasaría si por error la observación que ocupa la posición 20 de la variedad mejorada se hubiera registrado como 4770 en vez de 477?

La mediana seguiría siendo la misma, en cambio la media sería 617.6.

Percentiles o cuantiles

El percentil α **%** de la distribución de los datos es el valor por debajo del cual se encuentran el α **%** de los datos en la muestra ordenada.

Para calcularlo:

- Ordenamos la muestra de menor a mayor

- Buscamos el dato que ocupa la posición $\frac{\alpha(n+1)}{100}$ (si este número no es entero se promedian los dos adyacentes o se interpolan los dos adyacentes)

- El percentil 50% coincide con la mediana.

- Llamamos **cuartil inferior** al percentil 25% y **cuartil superior** al percentil 75%.

- Los cuartiles y la mediana dividen a la muestra ordenada en cuatro partes igualmente pobladas.

- Entre los cuartiles se hallan aproximadamente el 50% central de los datos.

	media	mediana	Cuartil inferior	Cuartil superior
Variedad mejorada	402.95	406.50	379.25	429.25
Variedad normal	366.30	358.00	333.00	401.25

- En ambas variedades el cuartil inferior está en la posición 5.25 y el superior en la posición 15.75.

- De acuerdo al software que se utilice puede haber pequeñas diferencias en los cuartiles ya que no todos utilizan el mismo método de cálculo.

MEDIDAS DE DISPERSIÓN

- No sólo es importante conocer alrededor de que valor se encuentran nuestros datos, también es importante saber cuan alejados se encuentran de ese valor.

- Necesitamos cuantificar de alguna manera la dispersión que hay respecto a estas medidas de posición.

Rango Muestral:

Es la diferencia entre el valor más grande y el más pequeño de los datos:

$$\text{Rango} = \text{máx}(X_i) - \text{mín}(X_i)$$

Si bien el rango es una medida muy fácil de calcular, es muy sensible a la presencia de observaciones extremas.

Distancia intercuartil

- Los percentiles pueden utilizarse para medir la variabilidad de un conjunto de datos.

- El rango intercuartil es la diferencia entre el tercer y el primer cuartil y es una medida de variabilidad que no está influenciada por valores extremos.

$$D_q = q_{0.75} - q_{0.25}$$

Varianza y desviación estándar

Si definimos el centro de un conjunto de datos como la media muestral, necesitamos una medida que pueda resumir las diferencias (o desviaciones) entre las observaciones y la media, es decir,

$$x_1 - \overline{x}, x_2 - \overline{x}, ..., x_n - \overline{x} \ .$$

Podríamos promediar estas desviaciones y definir

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

pero no es una buena medida ya que siempre es igual a cero.
Como nos interesa la magnitud de estas desviaciones y no el signo de las mismas, podríamos promediar los módulos o valores absolutos pero habitualmente se promedian los cuadrados de estas desviaciones, de este modo tampoco tenemos en cuenta los signos y además utilizamos una medida con mejores propiedades.

Definimos entonces la varianza muestral como

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

Utilizamos *n-1* en el denominador y no *n* por una propiedad que veremos más adelante.

Una vez obtenida esta medida definimos la desviación estándar como la raíz cuadrada de la misma, es decir que la desviación estándar se define como

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

Desvío Absoluto Mediano (Desviación absoluta respecto de la Mediana)

Es una versión robusta (menos sensible a la presencia de outliers) del desvío estándar, basada en la mediana.

Definimos la MAD como:

$$MAD = mediana\ (|\ x_i - \tilde{x}\ |)$$

¿Cómo calculamos la MAD?

- Ordenamos los datos de menor a mayor.
- Calculamos la mediana.
- Calculamos la distancia de cada dato a la mediana.
- Despreciamos el signo de las distancias y las ordenamos de menor a mayor.
- Buscamos la mediana de las distancias sin signo.

Utilizando todas estas medidas podemos ver que la ganancia en peso de los pollos que fueron alimentados con la variedad de maíz mejorada tiene menor dispersión que la de los pollos alimentados con la variedad normal.

	rango	Distancia intercuartil	Varianza muestral	Desviación estándar	MAD
Variedad mejorada	159	50	1825.73	42.73	22
Variedad normal	190	68.25	2581.17	50.81	33

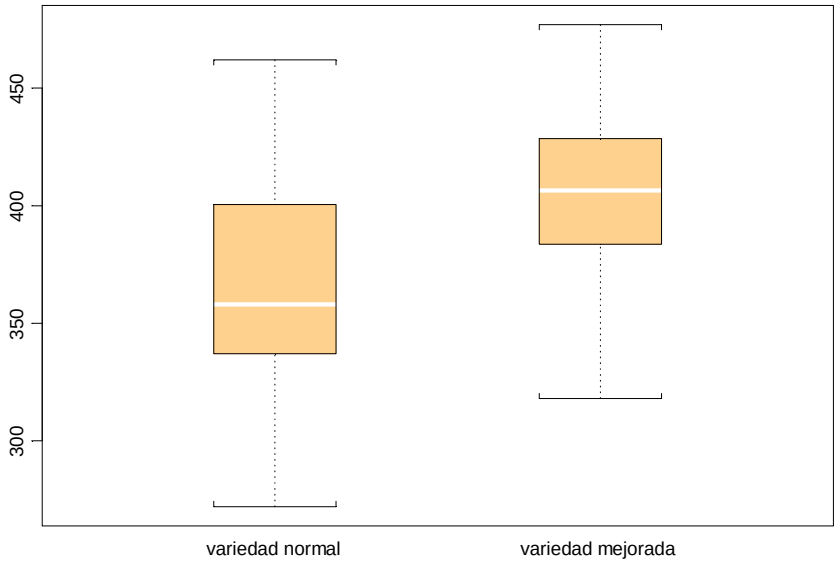
BOXPLOT

El boxplot o diagrama de cajas se construye de la siguiente manera:

- Se determina una escala vertical.
- Se representan en esa escala el primer cuartil, la mediana y el tercer cuartil.
- Se grafica un rectángulo (no importa el ancho) cuyo extremo inferior está a la altura del primer cuartil y el extremo superior a la altura del tercer cuartil, de este modo la altura del rectángulo coincide con la distancia intercuartil.
- Se traza un segmento dentro del rectángulo a la altura de la mediana.
- A partir de cada extremo de la caja dibujamos un segmento hasta el dato más alejado que está a lo sumo 1.5 D_c del extremo de la caja. Estos segmentos se llaman bigotes.
- Marcamos con ***** a aquellos datos que están entre 1.5 D_c y 3 D_c de cada extremo y con **o** a aquellos que están a más de 3 D_c de cada extremo.

El boxplot, entre otras cosas, nos permite evaluar la simetría de la distribución de nuestros datos.

- Si los datos siguen una distribución simétrica en su parte central, el primer y tercer cuartil deberían estar prácticamente a la misma distancia de la mediana.
- Si hay asimetría a derecha, el tercer cuartil debe estar más alejado de la mediana que el primer cuartil.
- Si hay asimetría a izquierda, el primer cuartil debe estar más alejado de la mediana que el tercer cuartil.
- Si el bigote superior es de la misma longitud que el inferior hay simetría en los extremos.
- Si el bigote superior es más largo que el inferior, hay asimetría a derecha en los extremos.
- Si el bigote superior es más corto que el inferior, hay asimetría a izquierda en los extremos.



- En el boxplot de la variedad normal vemos que hay asimetría a derecha en la parte central, mientras que en el boxplot de la variedad mejorada vemos simetría en la parte central.

- La longitud del bigote inferior y superior es muy similar por lo que en los extremos la distribución es bastante simétrica en ambos casos.

- No observamos outliers.

DISTRIBUCIÓN NORMAL

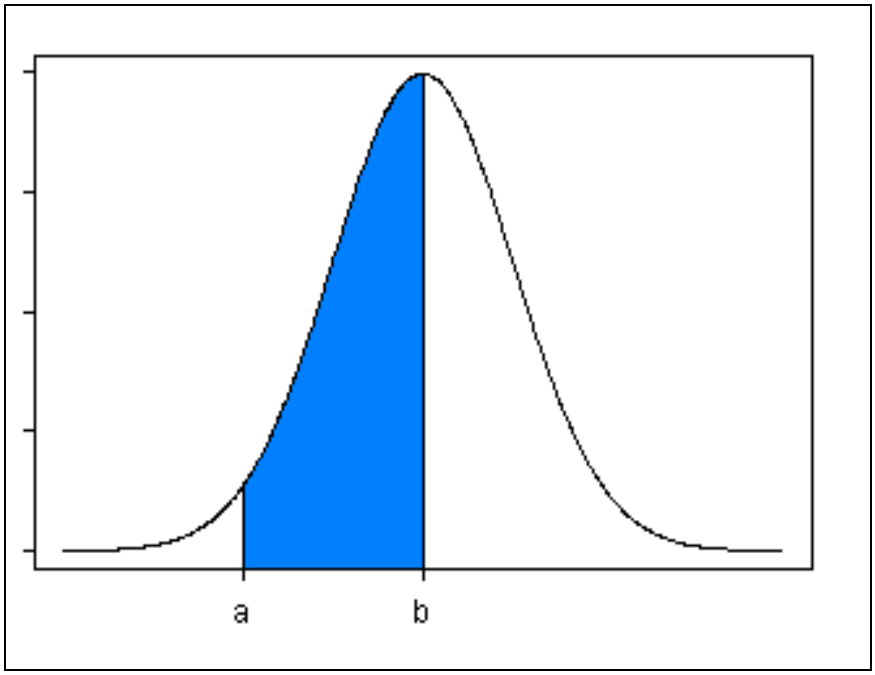
La distribución normal es sumamente usada en estadística. También se la llama distribución Gaussiana o campana de Gauss, su forma depende de dos parámetros μ y σ^2.

Características

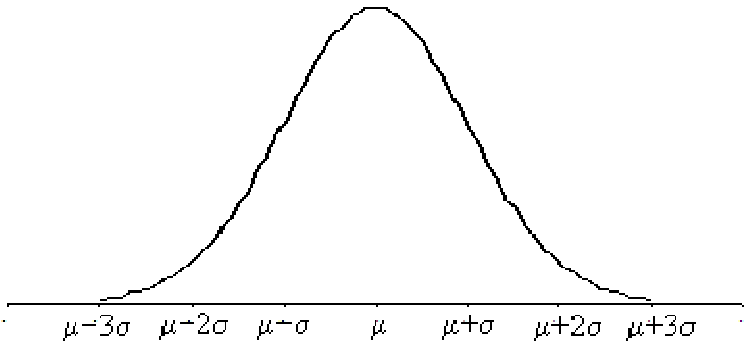
- Curva suave, acampanada y simétrica, con un único pico.
- El punto de simetría corresponde a la media μ de la variable.
- La desviación estándar σ determina el ancho de la campana.
- La curva presenta dos puntos de inflexión (cambios de concavidad) a distancia $-\sigma$ y $+\sigma$ del eje de simetría, μ.
- μ es el parámetro que indica la posición.
- σ es el parámetro que indica la escala o la dispersión de la función.
- El área bajo la curva es 1.

Notación: $X \sim N(\mu , \sigma^2)$ se lee "X tiene distribución Normal con media μ y varianza σ^2".

Si queremos calcular la probabilidad de que X tome valores dentro del intervalo (a,b), debemos calcular el área que queda por debajo de la curva en dicho intervalo.



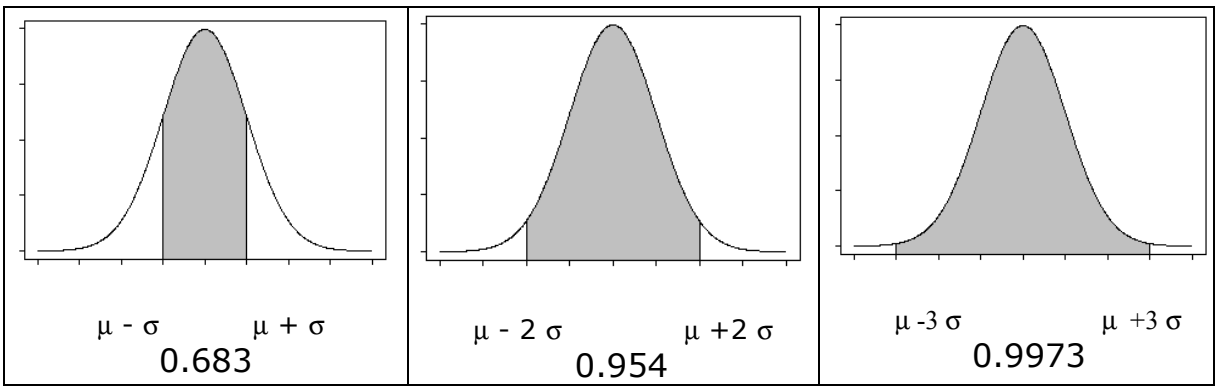
En la distribución normal, prácticamente toda la campana queda comprendida entre la media más menos 3 desviaciones estándar.



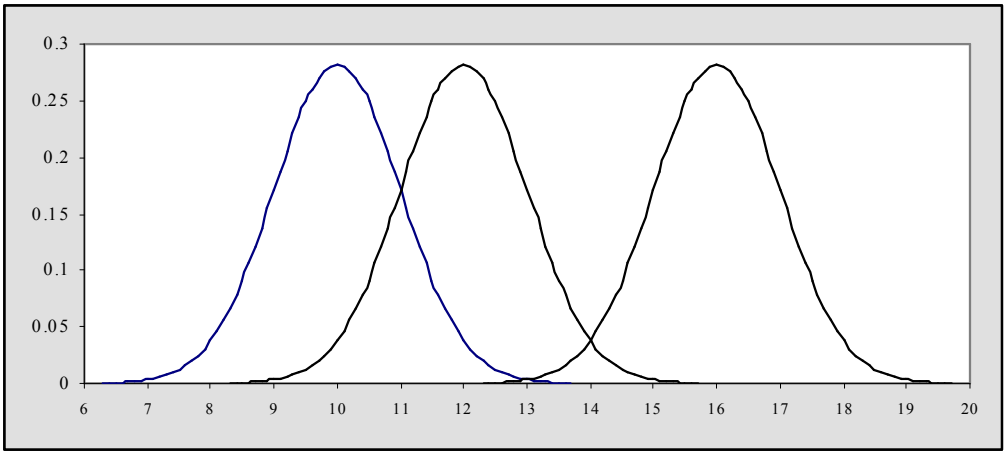
Algunas áreas con características de la curva normal:

Rango	Area (Prob.)
$\mu \pm \sigma$	0.683
$\mu \pm 2\ \sigma$	0.954
$\mu \pm 3\ \sigma$	0.9973

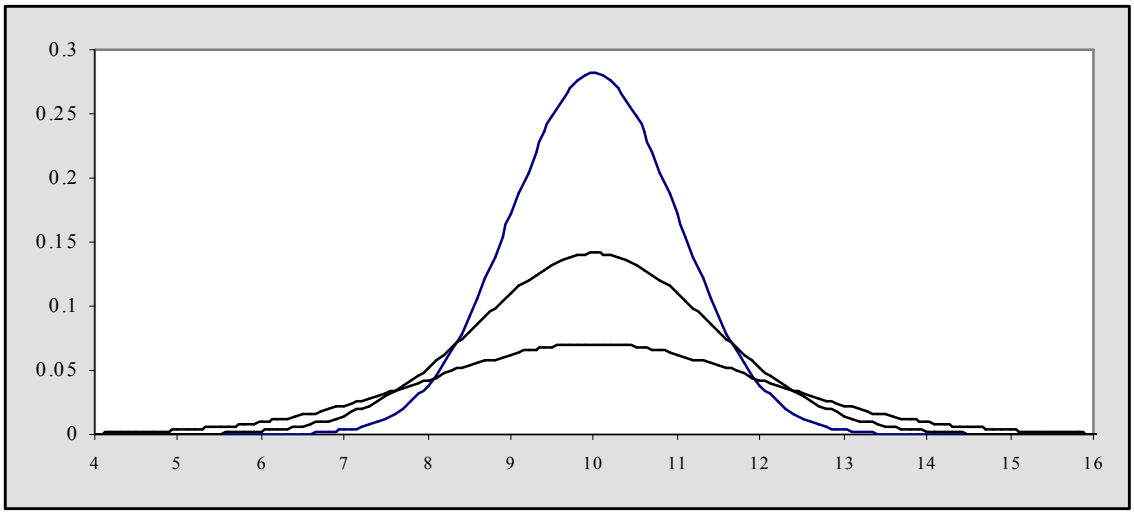
Gráficamente corresponden a:



¿Cómo cambia la normal cuando modificamos el parámetro μ?



¿Cómo cambia la función cuando modificamos el parámetro σ?



Cualquier posición puede expresarse como una distancia a la media medida en desvíos estándar. Es equivalente a considerar una normal con media 0 y desvío 1, que se conoce como **Normal Estándar**.

Variable normal estandarizada

Sea

X ~ N(μ, σ^2), definimos la VARIABLE NORMAL ESTÁNDAR o estandarizada (Z-score) del siguiente modo:

$$Z = \frac{(X - \mu)}{\sigma},$$

Z mide la distancia a la media (al centro) en unidades de desviaciones estándar.

La variable Z es normal con media $\mu_Z = 0$ y desvío estándar $\sigma_Z = 1$, es decir, Z ~ N(0, 1).