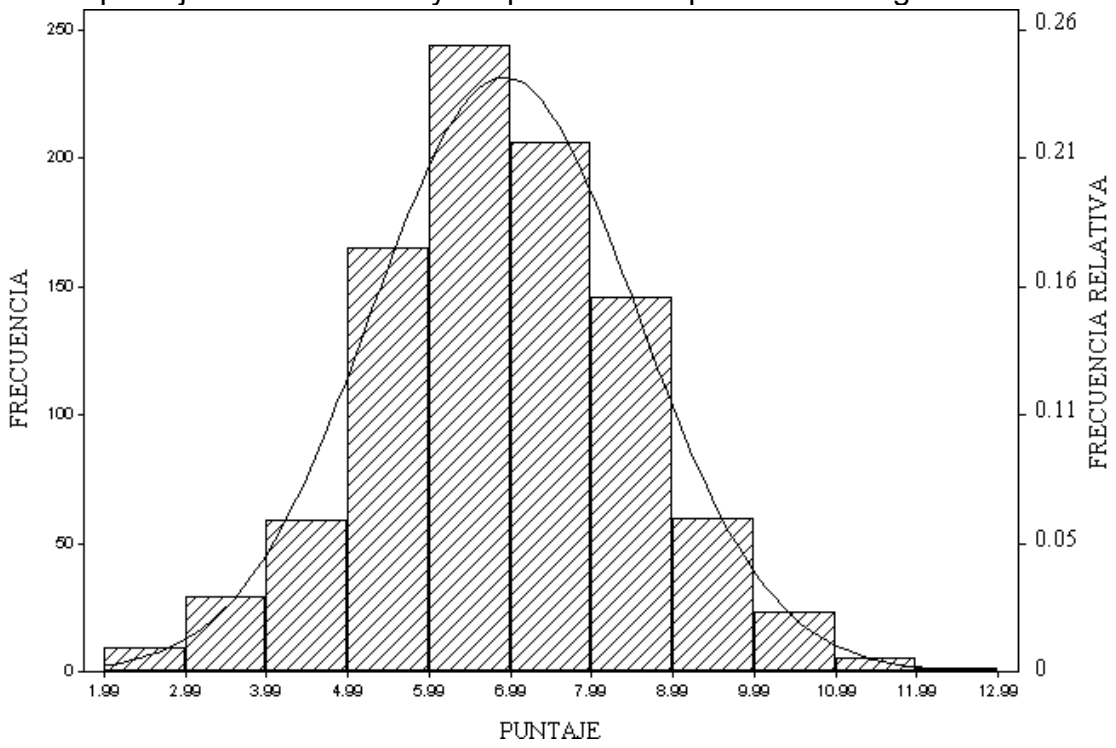
Curvas de densidad

¿Existe alguna manera de describir una distribución completa mediante una única expresión?

- un diagrama tallo-hoja no es práctico pues se trata de demasiados datos
- un histograma elimina los detalles y depende de la elección de las clases
- la mediana y los percentiles registran algunos aspectos específicos de los datos.

Si queremos tener una descripción de la forma global de la distribución, omitiendo valores atípicos y otras desviaciones del patrón general, la respuesta es sí.

Histograma del puntaje de vocabulario y la aproximación por una curva gaussiana.



Aproximamos al histograma por una curva suave que muestre la forma de la distribución sin las irregularidades del histograma.

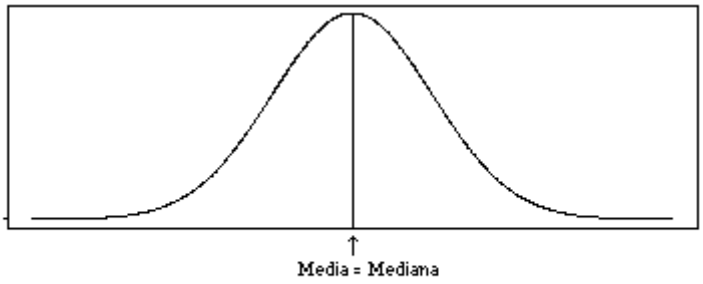
Como la frecuencia relativa de todas las observaciones es 1, requerimos que el *área total bajo la curva* sea 1.

El área bajo la curva y sobre un intervalo, correspondiente a cualquier rango de valores de la variable, es la proporción de observaciones que caen en ese rango. La curva describe la forma de la distribución y el

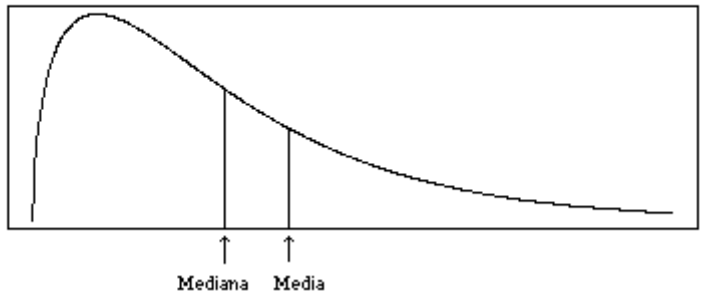
área bajo la curva = frecuencia relativa.

Es llamada *curva de densidad* de la distribución. El eje vertical mide la frecuencia relativa/(longitud del intervalo de clase).

Una curva de densidad con la forma apropiada suele ser una descripción adecuada del patrón global de una distribución. Los datos atípicos, que son desviaciones del patrón global, no están descriptos por la curva.



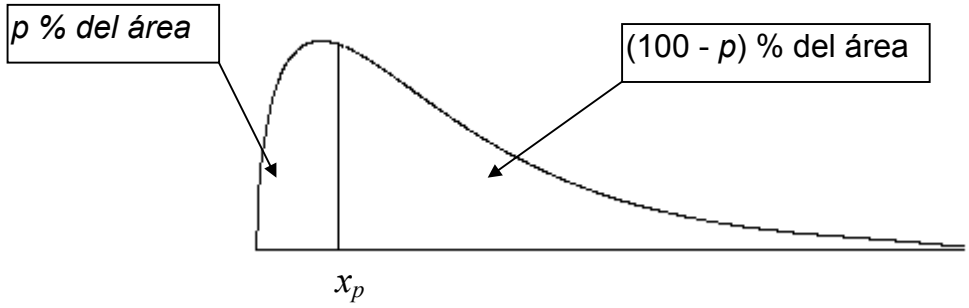
Media y mediana en una curva de densidad simétrica



Media y mediana en una curva de densidad asimétrica a derecha

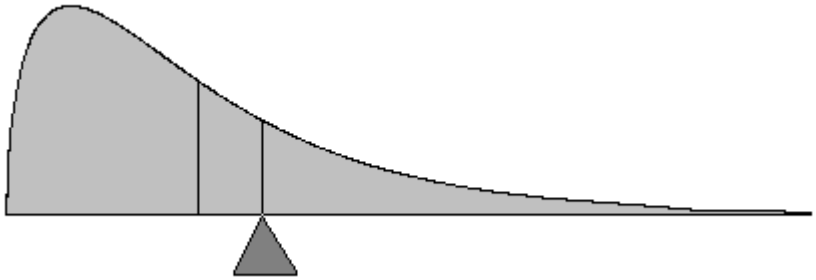
Las medidas de posición y dispersión también se aplican al caso de curvas de densidad.

El p-ésimo percentil, x_p, en una curva de densidad es el punto que *deja a su izquierda un p % del área bajo la curva* y el (100 - *p*) % restante, a la derecha.



En particular la mediana es el punto de áreas iguales, es decir, el punto que deja áreas iguales de cada lado.

Si pensamos a las observaciones como pesos en una vara delgada la media es el punto en que la vara quedaría equilibrada al poner un fiel justo debajo de él. Esta interpretación se extiende a la curva de densidad.



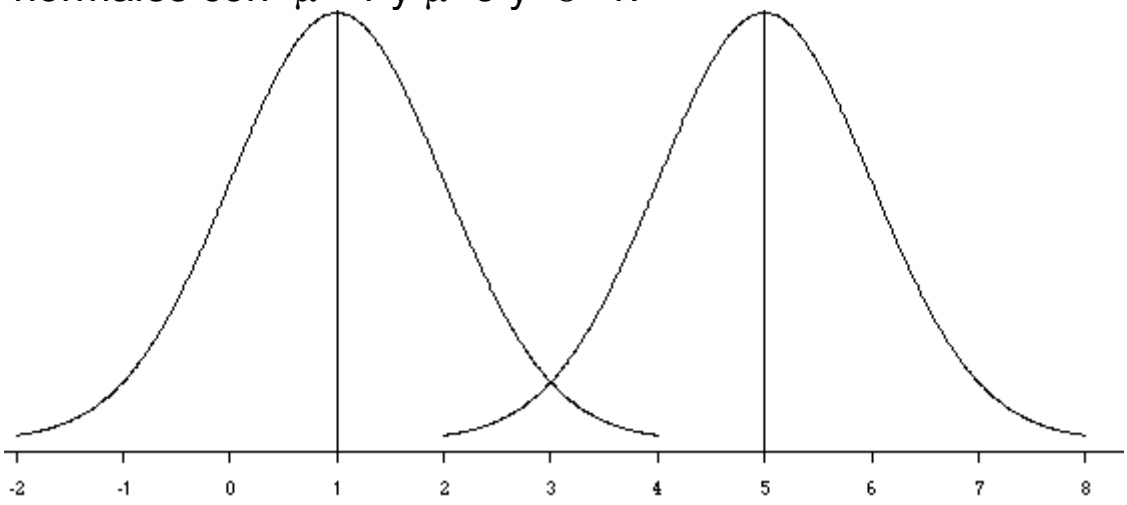
La media es un punto de equilibrio de una curva de densidad.

Las curvas de densidad simétricas son perfectamente simétricas a pesar que los datos reales rara vez mostrarán simetría perfecta.

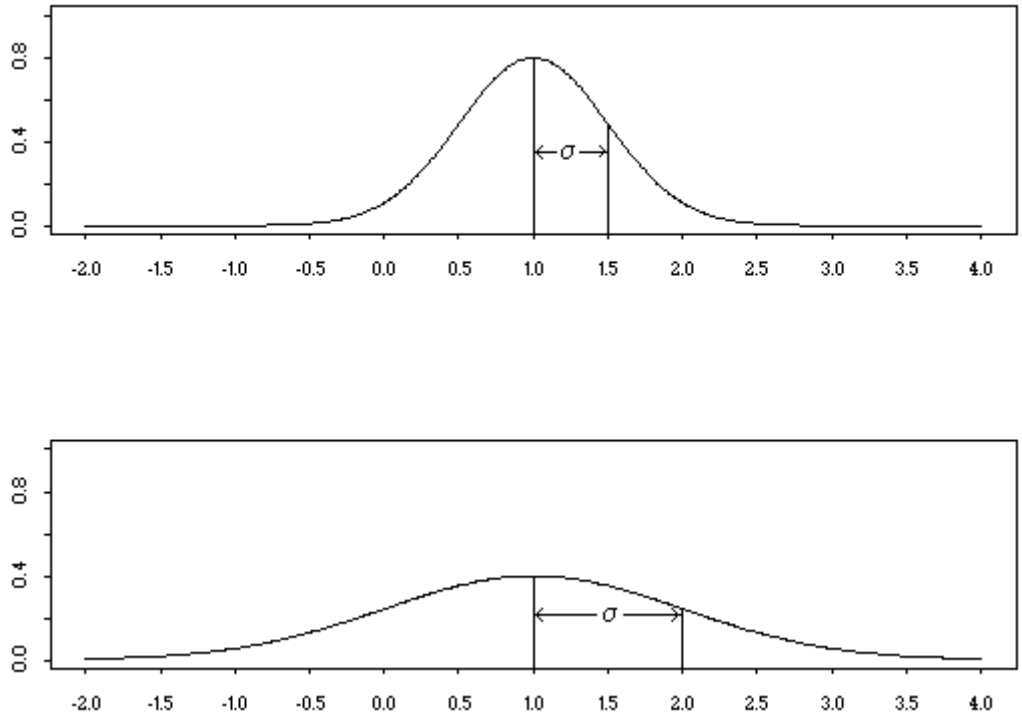
Debemos distinguir los *parámetros poblacionales* , la media = μ y el desvío = σ , de una curva de densidad de los números $\overline{x}$ y DS calculados a partir de las observaciones.

Distribuciones Normales o Gaussianas.

Todas las distribuciones gaussianas tienen la misma forma. Vemos dos curvas normales con μ= 1 y μ=5 y σ=1.



Dos curvas normales con diferente σ.



Podemos localizar σ a ojo en una curva normal. A medida que nos movemos en ambas direcciones desde el centro μ de la curva, ésta aumenta su pendiente

hasta un punto (punto de inflexión) en que la pendiente empieza a disminuir

Los dos puntos en los cuales ocurre este cambio de curvatura están localizados a una distancia σ a cada lado del centro μ.

Recuerde que μ y σ sólos **no** determinan la forma de una distribución en general. Éstas son propiedades de las distribuciones gaussianas.

Existen otras distribuciones no gaussianas con forma de campana.

Las curvas de densidad normal están descriptas por la siguiente ecuación

$$y = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-(\frac{x-\mu}{\sigma})^2} \qquad (2)$$

Observación: la ecuación (2) de la curva queda completamente especificada cuando se conocen los valores de μ y σ.

Las *distribuciones normales* proveen buenos modelos para

- puntajes de pruebas tomadas en poblaciones grandes (pruebas habilidades escolares y muchas pruebas psicológicas),
- mediciones cuidadosamente replicadas y de la misma calidad (datos de Newcomb tabla 2.1 sin outliers),
- caracteristicas de una población biológicamente homogénea (longitudes de las cucarachas, rendimiento de la soja y pérdida de humedad en carne de pollo envasada).

Las distribuciones de las siguientes variables, en cambio, son generalmente *asimétricas*:

- variables económicas (ingreso personal, ventas en firmas comerciales),
- tiempos de sobrevida (de pacientes de cáncer luego de realizado un tratamiento),
- tiempo de vida (de componentes mecánicos o electrónicas).

A pesar que la experiencia puede sugerir si un modelo gaussiano es o no factible en un caso particular, es muy riesgoso suponer la normalidad de los datos sin inspeccionarlos.

Superposición de una curva normal a un histograma:

- grafique una curva simétrica de altura = $\frac{1}{DS\sqrt{2\pi}}$ y puntos de inflexión en $\bar{x} \pm DS$.
- la escala en el eje vertical es la frecuencia relativa, siempre que la longitud de la base de los rectángulos de clase sea 1. En cualquier otro caso, en el eje vertical se grafica (la frecuencia relativa de cada clase) / (longitud de la clase) de manera que el

área de un rectángulo = (longitud de la base)*(altura del rectángulo)= frecuencia relativa

Verifiquemos este procedimiento para la superposición que muestra la figura sabiendo que la media del puntaje es 6.9156, el desvío es 1.6305 y $\frac{1}{DS\sqrt{2\pi}}$ = 0.2447