Estadística Descriptiva

Julio Deride Silva

Área de Matemática
Facultad de Ciencias Químicas y Farmcéuticas
Universidad de Chile

18 de agosto de 2010

Estadística Descriptiva

Julio Deride Silva

Área de Matemática
Facultad de Ciencias Químicas y Farmcéuticas
Universidad de Chile

18 de agosto de 2010

Tabla de Contenidos

Tabla de Contenidos

Tabla de Contenidos

Tabla de Contenidos

Tabla de Contenidos

Outline

Conceptos intuitivos básicos

Definición (Experimento Aleatorio)

Una experimento aleatorio, es un experimento tal que tiene un resultado impredecible antes de realizarlo. Sin embargo, es posible conocer un conjunto Ω en donde están los resultados posibles del experimento.

Definición (Variable Aleatoria)

Una variable aleatoria es una variable cuyo valor depende del resultado de un experimento aleatorio. Es decir, su valor queda determinado sólo después de que el experimento se ha realizado. Los valores que puede tomar esta variable estarán dentro de un conjunto E.

Definiciones

Definición (Realización de una variable aleatoria)

Una realización de una variable aleatoria, es el valor que toma la variable aleatoria después de realizar un experimento. Este valor corresponde a un elemento $e \in E$.

Definición (Muestra de tamaño n.)

Una Muestra de tamaño n para una variable aleatoria X es un vector $M = (x_1, x_2, \ldots, x_n)$, en donde cada componente es una realización de X. La muestra entrega información sobre la variable aleatoria.

Variables Aleatorias Cualitativas

1. **Variables cualitativas:** Se utilizan para describir cualidades, características o modalidades. Si sólo hay dos modalidades o características se denominan *dicotómicas*, de lo contrario diremos que son *politómicas*. Dentro de las variables cualitativas distinguimos 2 subclases: ordinales y nominales.
   1.1. **Variable cualitativa ordinal:** Los valores que toman se encuentran en un conjunto con orden, por ejemplo, $E = \{\text{excelente, bueno, regular, malo, muy malo}\}$.
   1.2. **Variable cualitativa nominal:** No existe ningún orden sobre el conjunto E, por ejemplo, los colores o los grupos sanguíneos.

Variables Aleatorias Cuantitativas

2. **Variables cuantitativas:** Son las que toman valores numéricos. Las variables cuantitativas además pueden ser:
   - 2.1. **Variable discreta:** Es cuando el conjunto E es finito o infinito numerable. Por ejemplo, el número de veces que se debe lanzar una moneda hasta obtener cara.
   - 2.2. **Variable continua:** Es cuando el conjunto E es infinito No-numerable. Por ejemplo el peso o la altura, son variables continuas, pero que no las podemos medir con precisión es otro problema, pues en teoría podrían tomar todos los valores dentro de un intervalo.

Variables Aleatorias Discretas

Definición (Frecuencia Absoluta)

Dada una muestra M de una variable aleatoria, con valores posibles en un conjunto $E = \{e_1, e_2, e_3, \ldots, e_n\}$ finito, se define la frecuencia absoluta de e_i dada la muestra M como:

$$f_i = |\{x \in M \,:\, x = e_i\}|$$

Definición (Frecuencia Relativa)

Dada una muestra M de una variable aleatoria, con valores posibles en un conjunto $E = \{e_1, e_2, e_3, \ldots, e_n\}$ finito, se define la frecuencia relativa de e_i dada la muestra M como:

$$f_i = \frac{|\{x \in M \ : \ x = e_i\}|}{|M|} = f^{rel}(e_i)$$

Observaciones

- $0 \leq f^{rel}(e) \leq 1$
- $\sum_{e \in E} f^{rel}(e) = 1$

Distribución de Frecuencia Empírica - Variables *Continuas*

Procedimiento:

(a) *Ordenar* la muestra de menor a mayor.

(b) Encontrar el *rango* de los datos, que es la diferencia entre el mayor y el menor.

(c) Dividir el rango en un número conveniente de *intervalos*[1]. Cada uno de estos intervalos se denomina *intervalos de clase*. El punto medio de un intervalo de clase se llama *marca de clase*.

(d) Determinar el número de elementos de la muestra en cada intervalo de clase (Frecuencia absoluta del intervalo).

(e) Reunir toda esta información en una tabla.

[1] Hay distintas formas de determinar este número de intervalos

Definiciones

Definición (Frecuencia Absoluta)

La Frecuencia Absoluta para el intervalo I_i es

$$f_i^{abs} = |\{x \in M \ : \ x \in I_i\}|.$$

Definición (Frecuencia Relativa)

La Frecuencia Relativa para el intervalo I_i es

$$f_i^{rel} = \frac{f_i^{abs}}{|M|}.$$

Definiciones

Definición (Frecuencia Acumulada)

La Frecuencia Acumulada, sea relativa o absoluta, hasta el intervalo I_i es:

$$f_i^{Acum} = \sum_{k=1}^{i} f_k^{abs}, \quad f_i^{Rel.Ac} = \sum_{k=1}^{i} f_k^{rel}$$

Tabla de Frecuencias

Consideremos una muestra de tamaño N para la cual en cada intervalo I_i hay una cantidad n_i de realizaciones. Una tabla de frecuencias resume la muestra, indicando la frecuencia absoluta, relativa y acumulada en una tabla, tal como se muestra a continuación:

Intervalo de Clase	Absoluta	Tipo de Relativa	Frecuencia Acumulada	Relativa Acumulada
I_1	n_1	$\frac{n_1}{N}$	n_1	$\frac{n_1}{N}$
I_2	n_2	$\frac{n_2}{N}$	$n_1 + n_2$	$\frac{n_1+n_2}{N}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_i	n_i	$\frac{n_i}{N}$	$\sum_{k=1}^{i} n_k$	$\frac{\sum_{k=1}^{i} n_k}{N}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_{N-1}	n_{N-1}	$\frac{n_{N-1}}{N}$	$\sum_{k=1}^{N-1} n_k$	$\frac{\sum_{k=1}^{N-1} n_k}{N}$
I_N	n_N	$\frac{n_N}{N}$	N	1

Histogramas

Definición (Histograma)

Un $\mathcal{Histograma}$ *consiste un un gráfico de barras con las siguientes características:*

(a) *Cada barra con base sobre el eje* X *tiene su centro en la marca de clase y longitud igual al tama los intervalos de clase.*

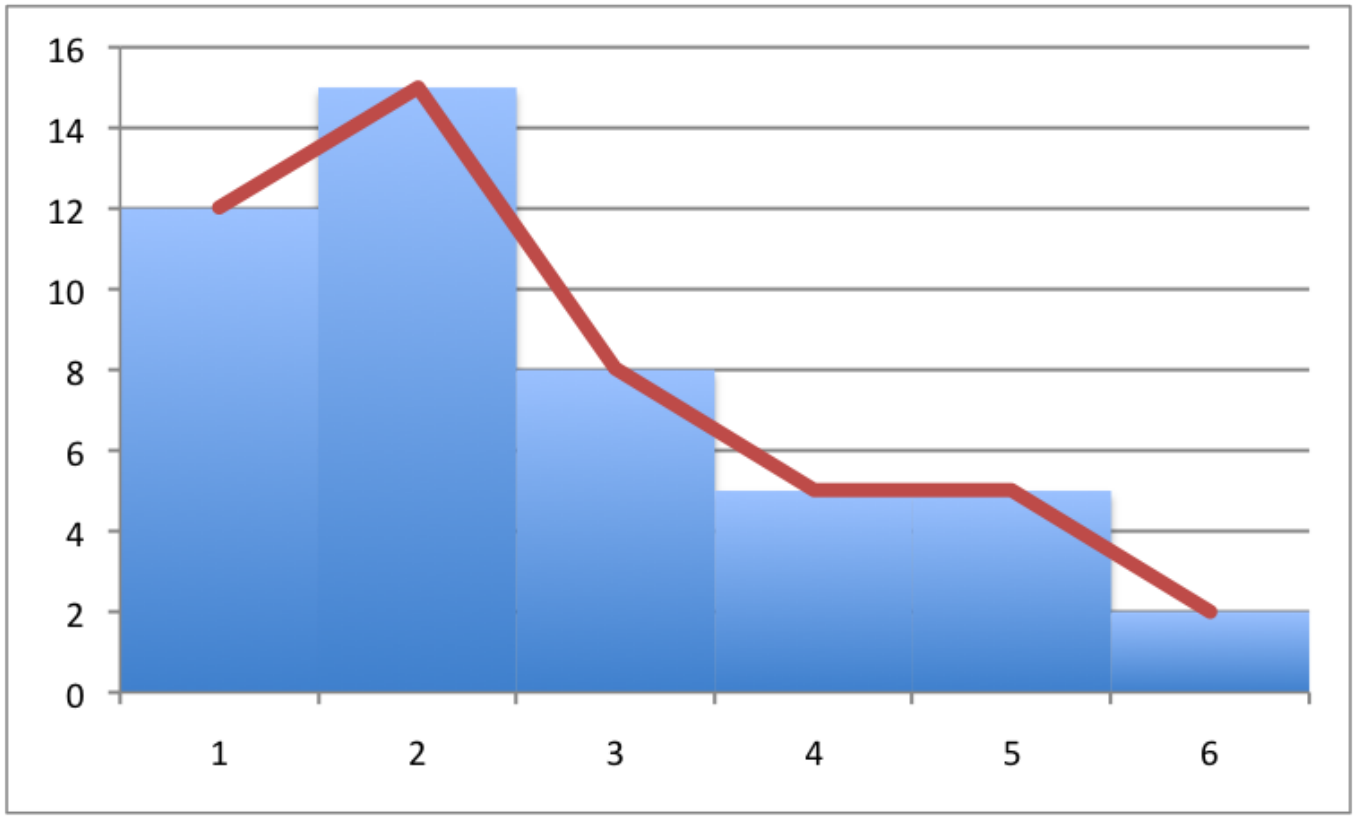
(b) *La superficie de cada barra es proporcional a las frecuencias absolutas.*

Polígono de Frecuencias

Definición (Polígono de frecuencias)

Un Polígono de frecuencias es un gráfico de línea trazado sobre las marcas de clase. Si el histograma ya esta dibujado, puede obtenerse uniendo los puntos medios de los techos de los rectángulos en el histograma.

Ejemplo



Outline

Definición (Media)

Dada una muestra $x_1, \ldots, x_n$ de una variable aleatoria X, se define la media como

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ &= \frac{1}{n}\sum_{i=1}^{K} n_i c_i, \end{aligned}$$

donde c_i es la marca de clase, n_i es la frecuencia absoluta y K es la cantidad de intervalos de clase.

Observación

- *Sensible a valores Extremos.*
- *Entrega poca información en distribuciones asimétricas.*
- *Puede no pertenecer al conjunto de valores posibles de la variable aleatoria.*
- *Tiene la misma unidad que la variable en estudio.*

Definición (Mediana)

Dada una muestra $x_1, \ldots, x_n$ de una variable aleatoria X, se define la mediana como

$$\begin{aligned} \tilde{x} &= x_{\left(\frac{n}{2}\right)}, \\ &= l_i + \frac{\frac{n}{2} - N_{i-1}}{n_i}[u_i - l_i], \end{aligned}$$

donde i es el índice del intervalo de clase $I_i = [l_i, u_i)$ donde se encuentra la mediana, N_{i-1} es la frecuencia absoluta acumulada del intervalo de clase anterior y n_i es la frecuencia absoluta del i-ésimo intervalo.

Observación

- *Estable a valores Extremos.*
- *Fácil de calcular.*
- *Si los datos no están agrupados, entonces siempre pertenece al conjunto de valores que toma la variable.*
- *Tiene la misma unidad que la variable en estudio.*

Definición ($k$ésimo valor de la muestra)

Se define el k-ésimo valor de la muestra $x_{(k)}$, al valor muestral que, luego de haber ordenado la muestra de menor a mayor, ocupa la posición k-ésima.

Definición (Percentil de orden k.)

Se define el percentil de orden k, P_k como el valor que deja por debajo al k % de los valores de la distribución. Dada una muestra $x_1, \ldots, x_n$ agrupada en una tabla de frecuencias, se identifica el intervalo I_i donde pertenece P_k y se calcula como

$$P_k = l_i + \frac{n\frac{k}{100} - N_{i-1}}{n_i}[u_i - l_i],$$

Definición (Cuartil)

$$Q_k = P_{25k}, \quad k = 1, 2, 3$$

Definición (Decil)

$$D_k = P_{10k}, \quad k = 1, \ldots, 9$$

Outline

Introducción.

Para estudiar la composición de una muestra, una de sus características es cuán diferente pueden ser los datos entre sí. Esto es, se pretende construir indicadores capaces de representar la variabilidad o la cercanía de las observaciones y si éstas presentan algún grado de agrupamiento respecto a alguna medida de tendencia central.

Definiciones

Definición (Rango)

Dada una muestra $x_1, \ldots, x_n$ de una variable aleatoria X, se define el rango como

$$\begin{aligned} \mathcal{R} &= x_{max} - x_{min}, \\ &= x_{(n)} - x_{(1)} \end{aligned}$$

esto es, la diferencia entre el mayor y el menor valor de una muestra.

Esta medida recoge el largo del intervalo en el cual toma valores la muestra, sin embargo, presenta deficiencias ante valores extremos.

Definición (Varianza muestral)

Dada una muestra $x_1, \ldots, x_n$ de una variable aleatoria X, se define la varianza muestral como

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

Definición (Desviación estándar o típica)

Llamaremos desviación estándar o típica a la raíz cuadrada de la varianza muestra, esto es

$$S_n = \sqrt{S_n^2}.$$

Observaciones

Observación

Se construye esta medida de dispersión en base a la anterior para hacer consistente el análisis de unidad con respecto a la naturaleza de las observaciones de la muestra.

Observación

Si definimos el intervalo $I = (\overline{x} - 2S_n; \overline{x} + 2S_n)$*, se tiene que en él se encuentra al menos el 75 % de las observaciones de la muestra.*

Definición (Coeficiente de variación)

Se define el coeficiente de variación como

$$\mathcal{CV} = \frac{S_n}{\overline{x}}.$$

Con este indicador, debemos tener las siguientes precauciones:

- Sólo es aplicable a variables estrictamente positivas.
- No es invariante a cambios de origen: Si $Y = X + b \Rightarrow \mathcal{CV}_Y < \mathcal{CV}_X$.
- Es invariante a cambios de escala: $Y = aX \Rightarrow \mathcal{CV}_Y = \mathcal{CV}_X$.

Definición (Dispersión media)

Dada una muestra $x_1, \ldots, x_n$ de una variable aleatoria X, se define la dispersión media como

$$D = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}| \, .$$

Tipificación.

Así como el coeficiente de variación permite comparar variables de distinta naturaleza, si se desea comparar observaciones, se introduce el concepto de *tipificación*, que consiste en definir

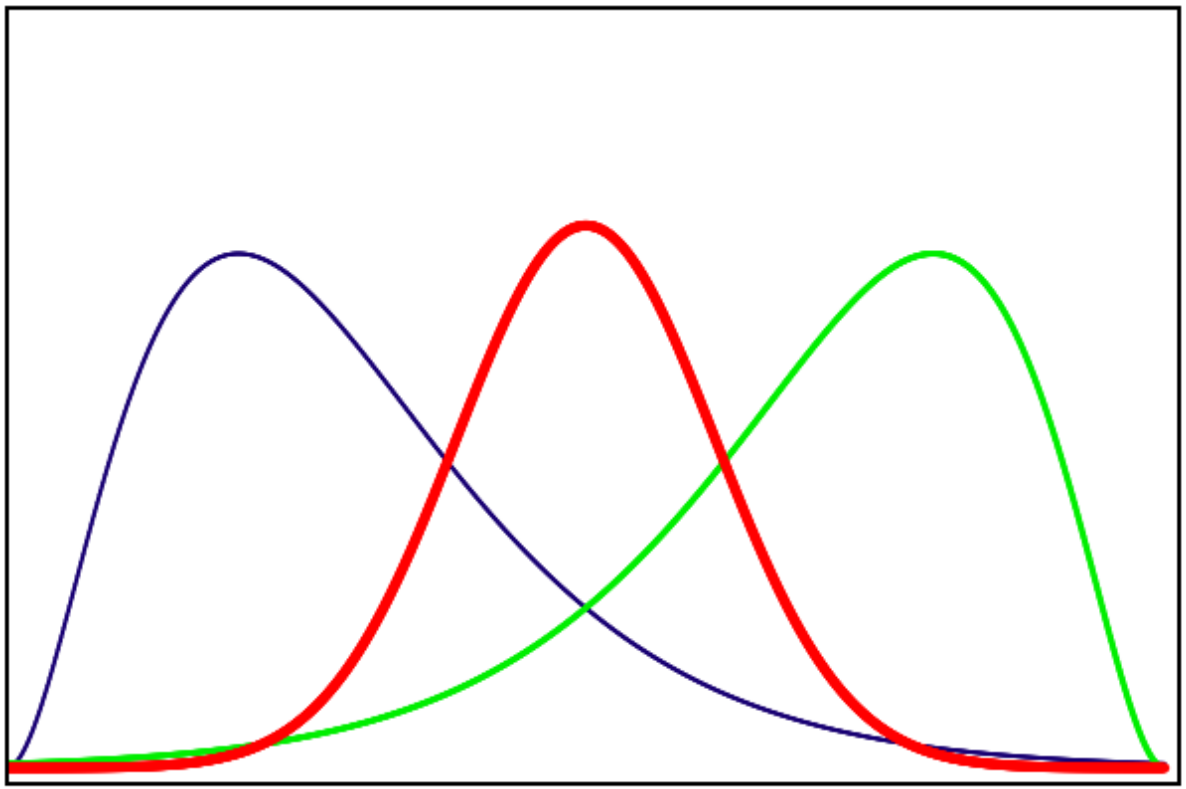
$$z_i = \frac{x_i - \overline{x}}{S_n}.$$

Esta nueva variable tiene la ventaja de ser adimensional y, por lo tanto, hace comparable variables cuya naturaleza puede no tener relación. Se puede demostrar que

- $\overline{z} = 0$.
- $S_n(z) = 1$.

Outline

Introducción.



Medidas de Simetría

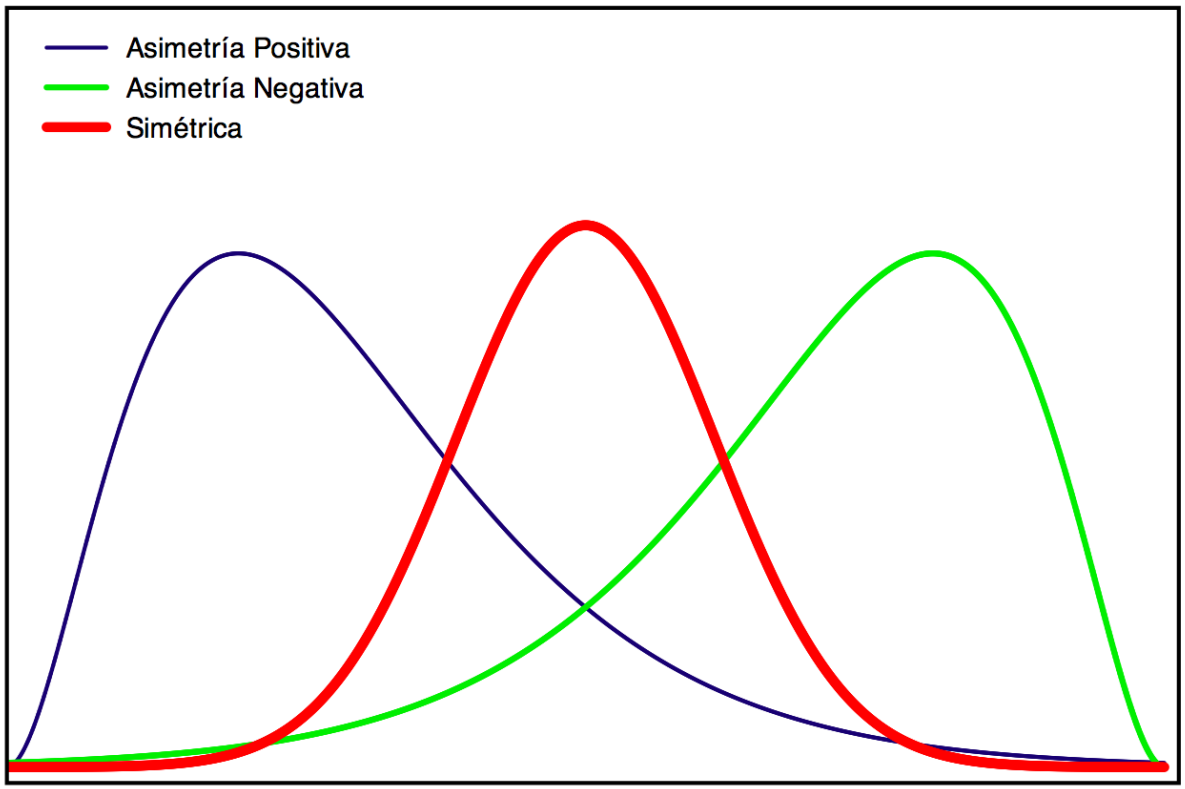
Definición (Simetría)

Una distribución de frecuencia empírica se dirá simétrica si la mitad izquierda de ella es un reflejo de la mitad derecha.

Observación

- *Si la distribución es simétrica y unimodal, entonces* $\overline{x} = \tilde{x} = \hat{x}$.
- *Si se tiene que la mediana y la media no coinciden,* $\tilde{x} \neq \overline{x}$, *necesariamente la muestra presenta asimetría, la cual clasificaremos de la siguiente forma:*
  - *Asimetría positiva: Las frecuencias más altas están al lado izquierdo de la distribución y su cola está a la derecha.*
  - *Asimetría negativa: Las frecuencias más altas están al lado derecho de la distribución y su cola está a la izquierda.*

Gráficamente



Definición (Momento de orden r.)

Dada una muestra $x_1, \ldots, x_n$ de tamaño n, se define el momento de orden r con respecto a A como

$$m_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - A)^r.$$

Observación

- *Si $A = 0$, entonces se dice momento de orden r.*
- *Si $A = \bar{x}$, se dice momento central de orden r.*
- *Es directo de la definición que el momento central de orden 2 corresponde a la varianza muestral. A su vez, el momento de orden 1 es la media muestral.*

Coeficiente de sesgo.

Dado que los momentos centrales de orden impar respetan la posición relativa de las observaciones respecto al promedio a través del signo, utilizaremos, en particular, el momento central de orden 3 para construir un indicador de asimetría.

Definición (Coeficiente de sesgo)

Se define el coeficiente de sesgo como

$$a_3 = \frac{m_3^{central}}{S_n^3}.$$

Observación

Para este indicador, definimos el siguiente criterio de clasificación:

- *Si* $a_3 > 0$, *la distribución presenta asimetría positiva.*
- *Si* $a_3 < 0$, *la distribución presenta asimetría negativa.*
- *Si* $a_3 = 0$, *la distribución es simétrica.*

Definición (Coeficiente de asimetría de Pearson)

$$SK = 3\frac{\bar{x} - \tilde{x}}{S_n},$$

donde $\bar{x}$ es la media y $\tilde{x}$ es la mediana.

Definición (Índice de asimetría Yule Bowley)

$$\mathcal{A}_s = \frac{(\mathcal{Q}_3 - \mathcal{Q}_2) - (\mathcal{Q}_2 - \mathcal{Q}_1)}{(\mathcal{Q}_3 - \mathcal{Q}_1)}.$$

Definición (Índice de asimetría modal)

$$\mathcal{A}_s^{mod} = \frac{\overline{x} - \hat{x}}{S_n},$$

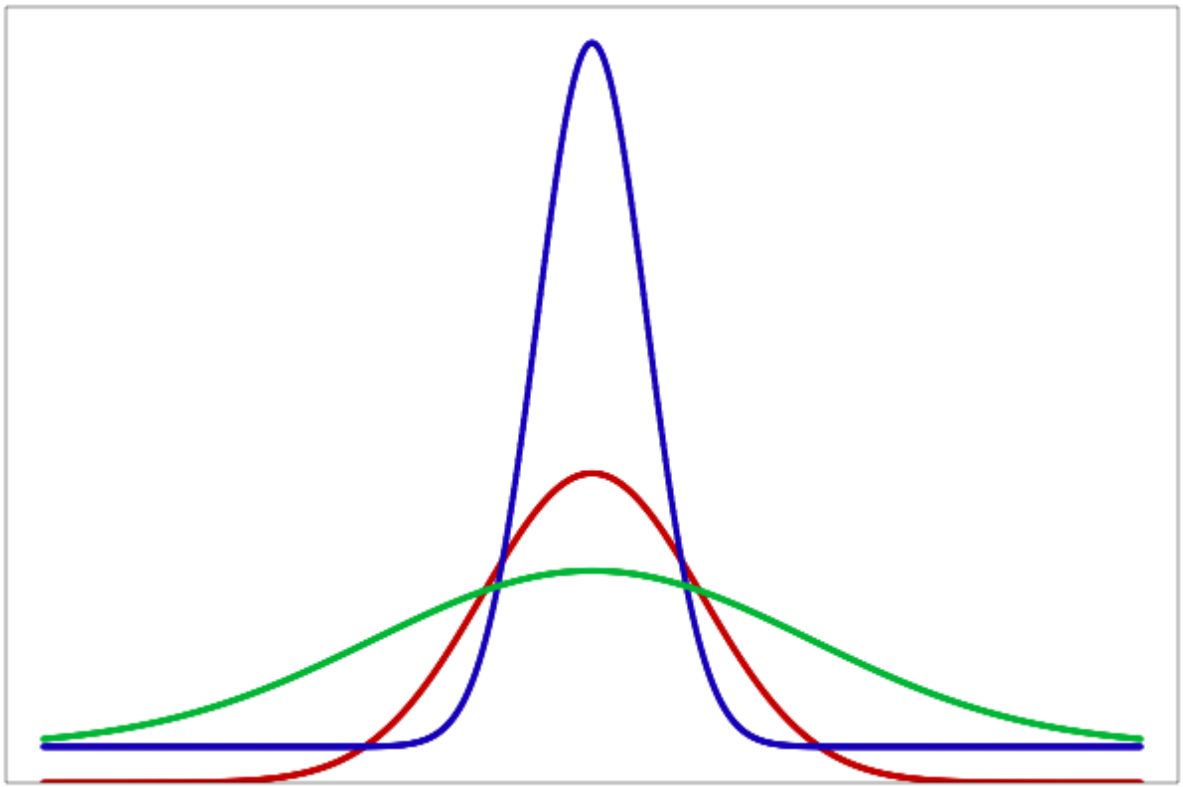
donde $\hat{x}$ corresponde a la moda.

La clasificación se hace de la siguiente forma

Simetría	Sesgo	Pearson	Yule Bowley	Modal
Positiva	$a_3 > 0$	$SK > 0$	$\mathcal{A}_s > 0$	$\mathcal{A}_s^{mod} > 0$
Simétrica	$a_3 = 0$	$SK = 0$	$\mathcal{A}_s = 0$	$\mathcal{A}_s^{mod} = 0$
Negativa	$a_3 < 0$	$SK < 0$	$\mathcal{A}_s < 0$	$\mathcal{A}_s^{mod} < 0$

Outline

Introducción.



Curva Normal

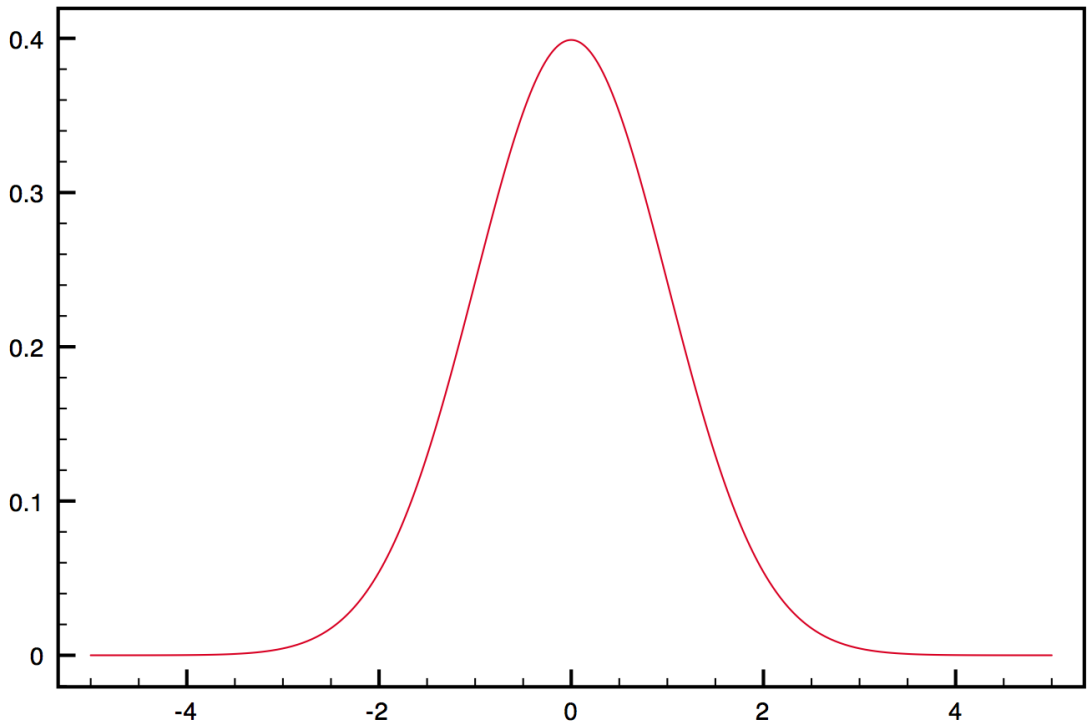


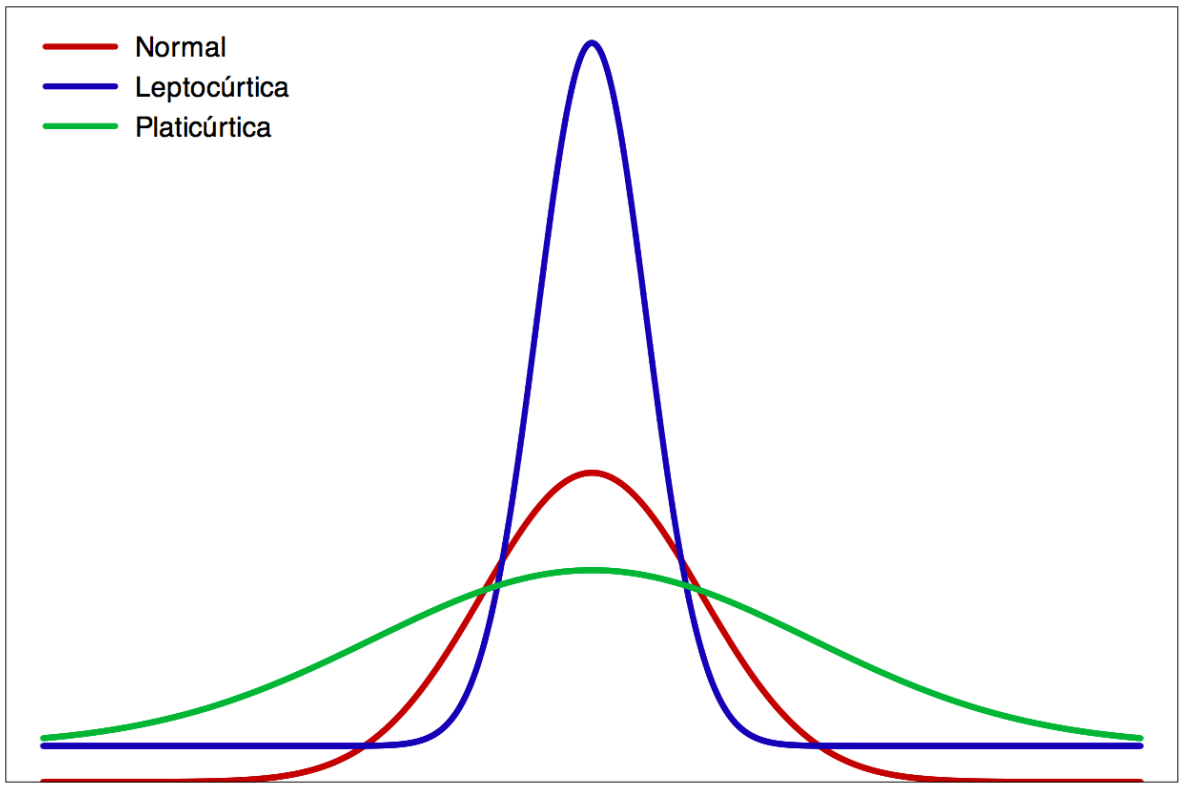
Figura: $f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\frac{1}{2}x^2}$.

Clasificación.

La clasificación de las medidas de apuntamiento, se realizará con las siguientes categorías.

- *Leptocúrtica*: Se dice que una distribución es *Leptocúrtica* si es más apuntada que lo normal.
- *Mesocúrtica*: Se dice que una distribución es *Mesocúrtica* si es tan apuntada como la normal.
- *Platicúrtica*: Se dice que una distribución es *Platicúrtica* si es menos apuntada que lo normal.

Clasificación.



Definimos los siguientes índices:

Definición (Coeficiente de aplastamiento de Fisher)

$$\gamma_2 = \frac{m_4^{central}}{S_n^4} - 3,$$

donde $m_4^{central}$ corresponde al momento central de orden 4.

Definición (Coeficiente de Apuntamiento η)

$$\eta = \frac{\frac{\mathcal{Q}_3 - \mathcal{Q}_1}{2}}{\mathcal{P}_{90} - \mathcal{P}_{10}}.$$

Clasificación.

Finalmente, la clasificación se hace de la siguiente forma

Índice	Leptocúrtica	Mesocúrtica	Platicúrtica
γ_2	$\gamma_2 > 0$	$\gamma_2 = 0$	$\gamma_2 < 0$
η	$\eta > 0,263$	$\eta = 0,263$	$\eta < 0,263$