Distribución Normal

PROGRAMA MATEMÁTICA DUOC UC

Estadística Descriptiva

Medidas de distribución

Las medidas de distribución o de forma permiten comparar **la distribución muestral** con otra teórica que estudiaremos llamada **distribución normal**

Las medidas de forma son:

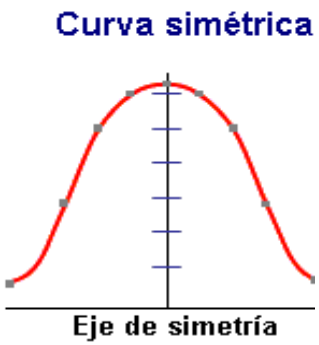
- **Coeficiente de Asimetría**: Tiene relación con la **simetría o asimetría** de la distribución muestral.

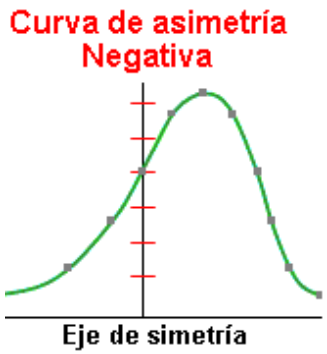
- **Coeficiente de Curtosis** : Tiene relación con la concentración de datos en torno a la media de la distribución muestral.

Coeficiente de Asimetría

Es una medida que permite identificar si los datos se distribuyen de forma simétrica o asimétrica alrededor de la media aritmética, en otras palabras muestra **cómo se distribuyen los datos de una variable con respecto a la media de esa variable**.

El Coeficiente de Asimetría puede indicarnos tres tipos de distribuciones:





El **eje de simetría** se sitúa sobre la **media aritmética** de la variable

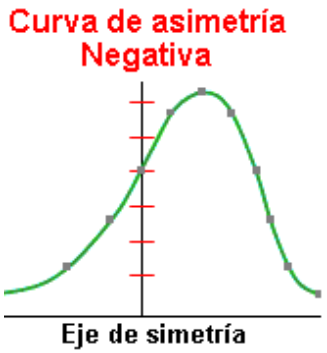
Coeficiente de Asimetría Positivo

Se dice que la curva tiene **asimetría positiva** cuando la mayoría de los datos se encuentran por debajo del valor de la media aritmética.

Debajo de la media se concentra la mayor cantidad de datos

Coeficiente de Asimetría Negativa

Se dice que la curva tiene **asimetría negativa** cuando la mayor cantidad de datos se concentran sobre la media.

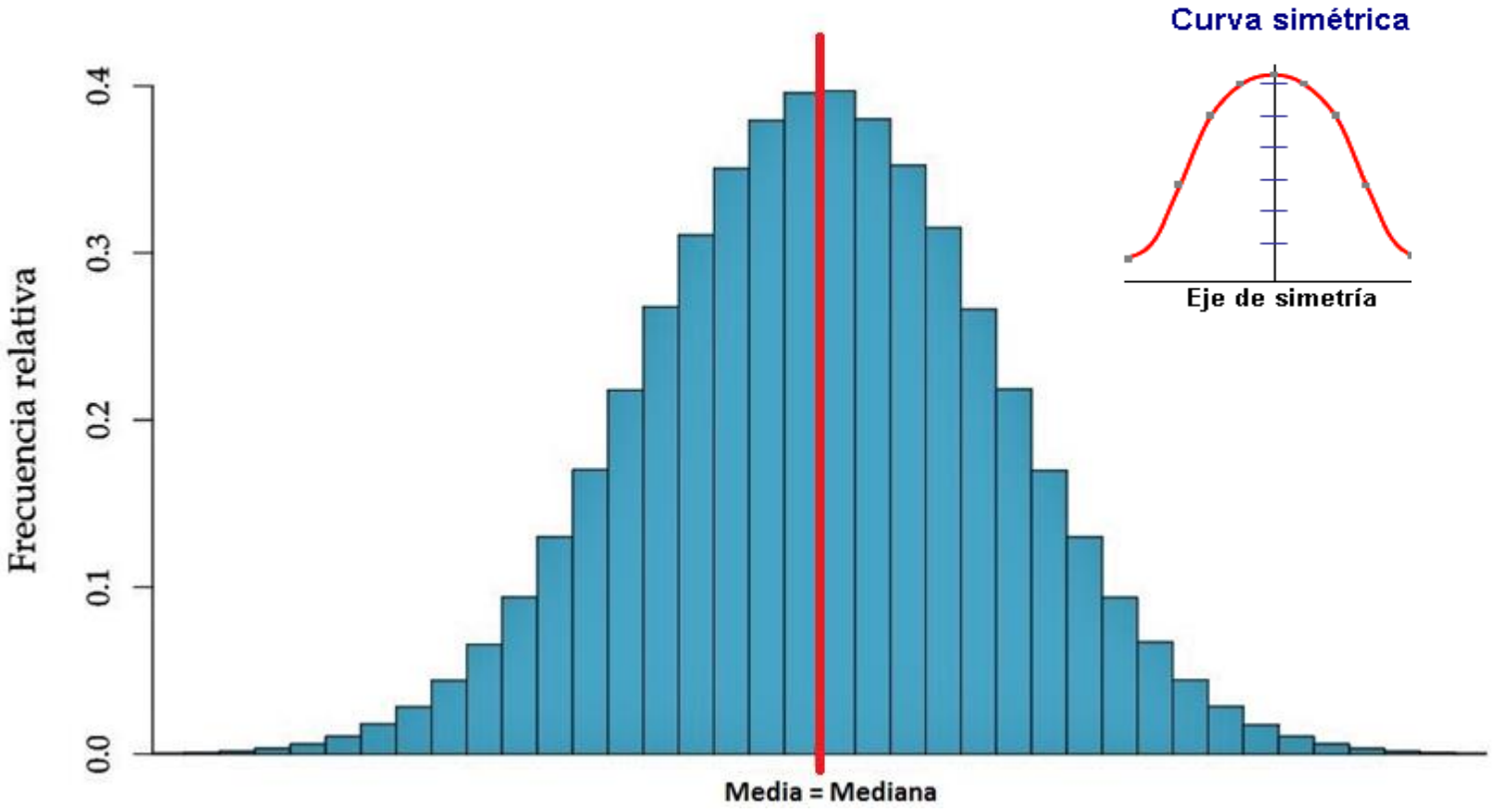


Sobre la media se concentra la mayor cantidad de datos

Asimetría negativa: Cola negativa

Coeficiente de Asimetría próximo a cero

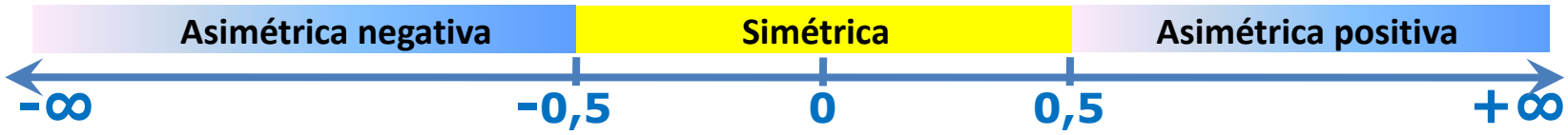
Si la curva es **simétrica,** entonces hay aproximadamente la misma cantidad de datos a ambos lados de la media.



Criterio de simetría

La distribución presenta **asimetría negativa** cuando el coeficiente de asimetría es menor que -0,5

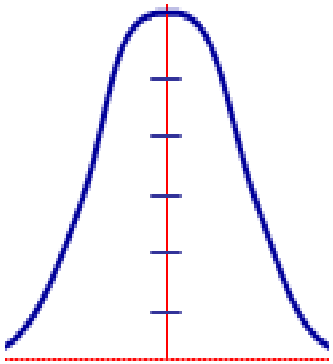
La distribución presenta **asimetría positiva** si el coeficiente de asimetría es mayor a 0,5



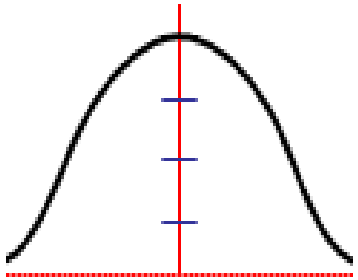
La distribución es **simétrica**, si el coeficiente de asimetría está entre -0,5 y 0,5.

Coeficiente de Curtosis

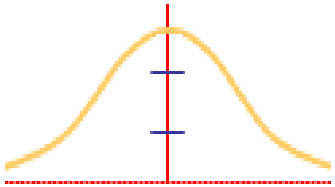
El Coeficiente de Curtosis o Apuntamiento es una medida que permite determinar el grado de concentración de los datos en torno a la media aritmética, en otras palabras mide la mayor o menor cantidad de datos que se agrupan alrededor de la media. Si consideramos que la distribución normal presenta concentración media de observaciones (mesocúrtica), entonces podemos definir otras dos situaciones: distribución leptocúrtica y platicúrtica



Leptocúrtica



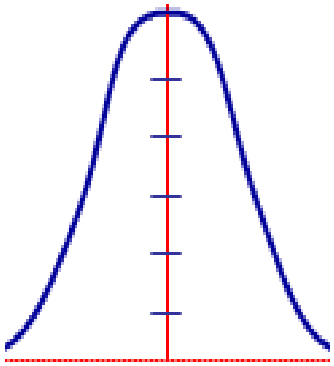
Mesocúrtica



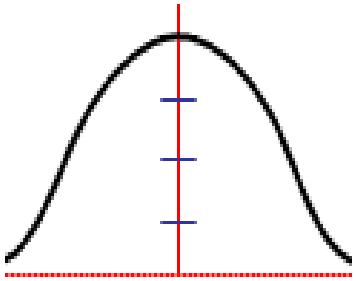
Platicúrtica

Se dice que la distribución es **leptocúrtica, mesocúrtica** y **platicúrtica** cuando presentan alta, normal y baja concentración de observaciones en torno a su media aritmética.

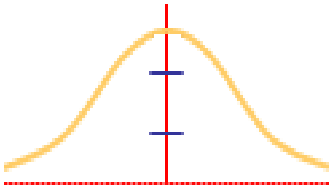
Coeficiente de Curtosis: Forma de la Distribución



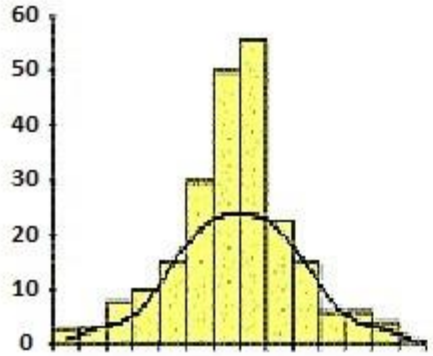
Leptocúrtica



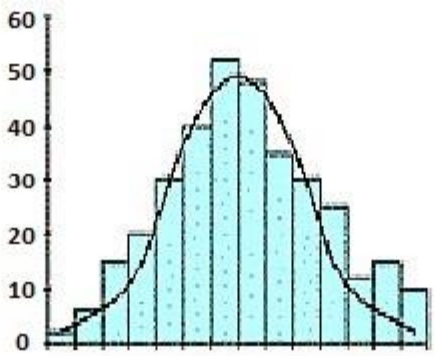
Mesocúrtica



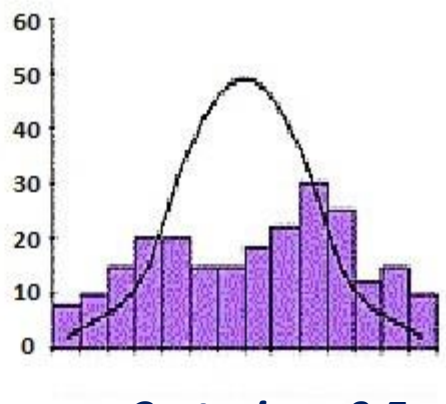
Platicúrtica



Curtosis > 0,5



Curtosis ≈ 0



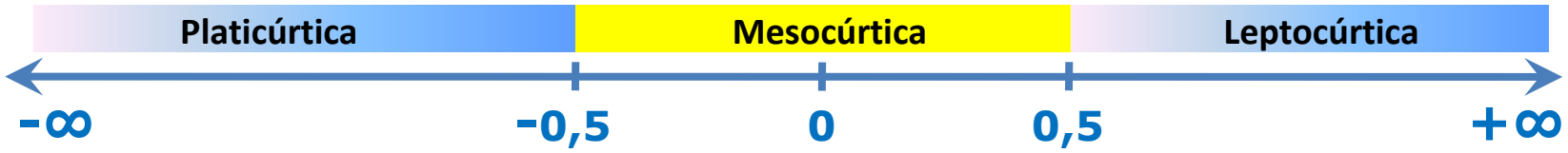
Curtosis < -0,5

Criterio de Curtosis

De la misma manera que en análisis de simetría, algunos autores han determinado un rango de valores para el Coeficiente de Curtosis, tal como se observa en la figura siguiente:

La distribución es platicúrtica, por lo que existe baja concentración de datos u observaciones en torno a la media.

La distribución es leptocúrtica, por lo que existe alta concentración de datos u observaciones en torno a la media.

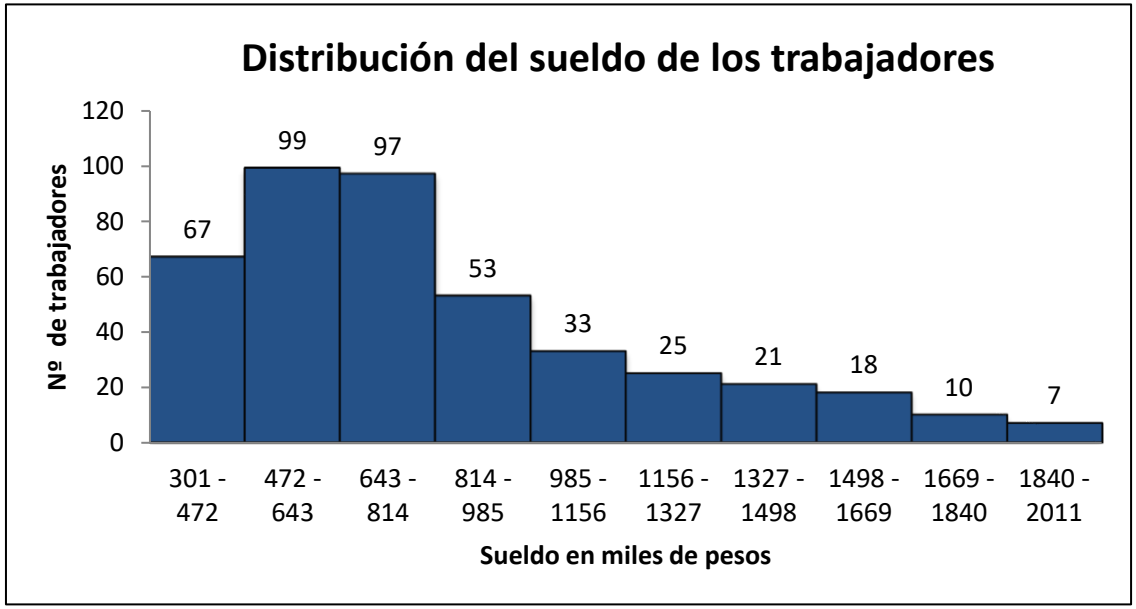


La distribución es mesocúrtica, por lo que existe concentración normal de datos u observaciones en torno a la media.

Ejemplo 1: Se han calculado las medidas de distribución de la variable *presión sistólica medida en mmHg* (milímetros de mercurio) obtenida de una muestra de 300 personas de la tercera edad.

Fórmula Excel	Valor	Interpretación
CURTOSIS(rango de datos)	**-1,097309**	La distribución de la presión sistólica es platicúrtica, por lo que presenta baja concentración de datos en torno a la media.
COEFICIENTE.ASIMETRIA(rango de datos)	**0,176453**	La distribución de la presión sistólica es simétrica por lo que existe aproximadamente la misma cantidad de datos en ambos lados de la media.

Ejemplo 2: El siguiente gráfico muestra la distribución de los *sueldos de los 430 empleados de la empresa Alfa*. Se muestra además el valor del Coeficiente de Asimetría y del Coeficiente Curtosis de esta variable.

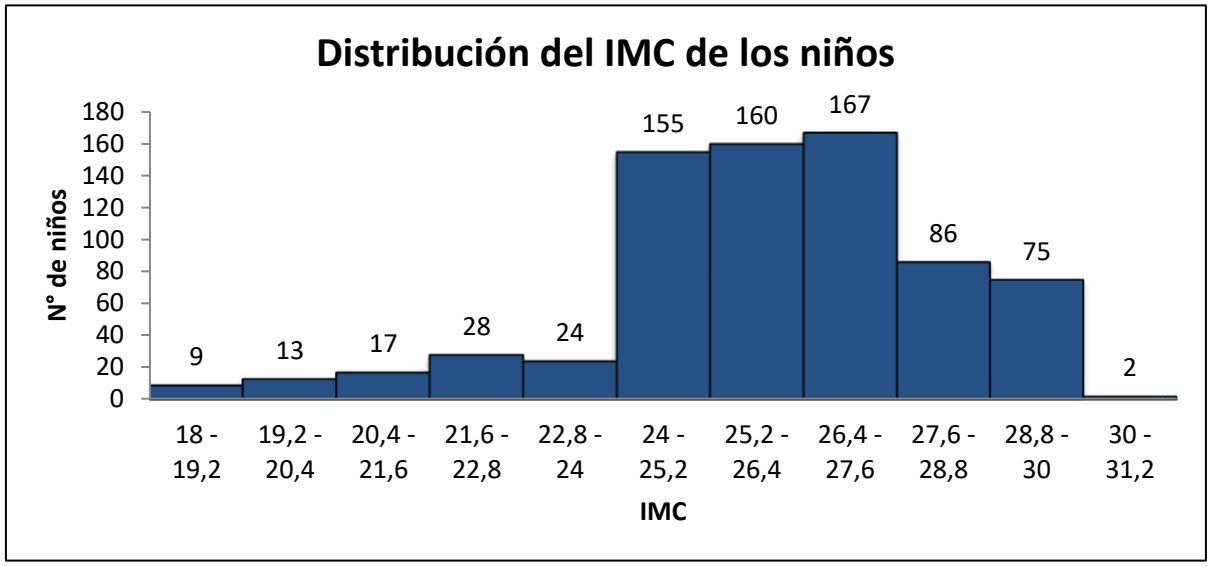


Curtosis	**0,389433707**
Coeficiente de asimetría	**0,996520143**

Interpretación de las medidas de distribución:

La distribución de los sueldos es mesocúrtica, esto significa que los valores de los sueldos se concentran normalmente alrededor de la media. Además la distribución de los sueldos tiene asimetría positiva, es decir, la mayoría de los sueldos se concentran bajo el sueldo promedio.

Ejemplo 3: El siguiente gráfico muestra la distribución del *índice de masa corporal (IMC)* obtenida de una muestra de 736 niños en edad escolar. Se conocen el Coeficiente de Asimetría y del Coeficiente Curtosis.



Curtosis	**0,91321772**
Coeficiente de asimetría	**-0,71358726**

Respuesta:

La distribución del IMC es leptocúrtica, por lo que existe alta concentración de datos en torno a la media. Además la distribución del IMC presenta asimetría negativa lo que significa que la mayoría de los datos se concentran sobre la media.

Variable Aleatoria Continua

Una variable aleatoria X, es una función que asigna un valor numérico al resultado de un experimento aleatorio.
Una variable aleatoria es continua si los valores que puede llegar a tomar pertenecen a un intervalo de números reales.

Ejemplos:

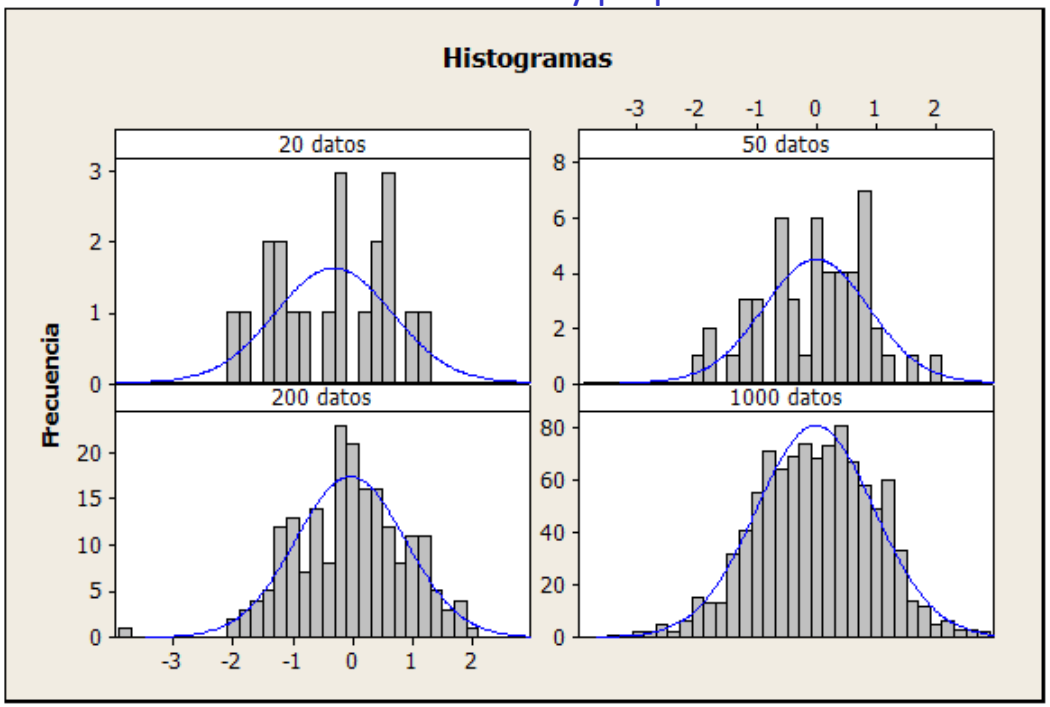
a) X: **Estatura** de un estudiante escogido al azar de la población de estudiantes de un Instituto Profesional.
b) X: **Temperatura media** de un día cualquiera.
c) X: **Sueldo** mensual de un trabajador escogido al azar de la población de trabajadores de una empresa.

Función de densidad de Probabilidad de una V.A. Continua

- Una variable continua se puede representar a través de un histograma de frecuencias relativas, en donde la base de cada barra del histograma es un intervalo de los datos y la altura es la proporción de datos respecto al total, que están contenidos en ese intervalo.

- El histograma muestra gráficamente cómo se distribuyen los datos

- La **función de densidad de probabilidad** de una variable aleatoria continua es aquella cuyo gráfico toma la forma de la distribución de la variable, considerando que se cuenta con muchos datos, clasificados en intervalos muy pequeños.



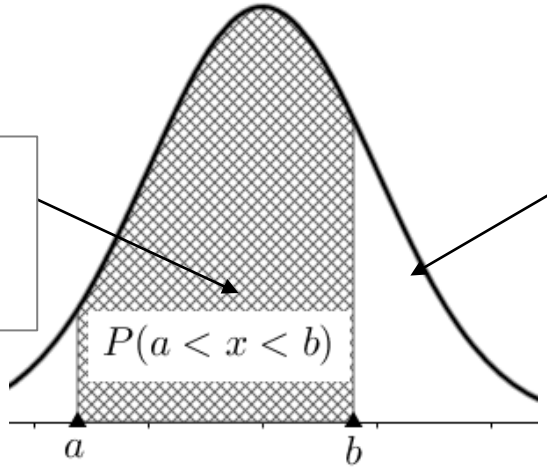
A medida que se aumenta la cantidad de datos y el número de intervalos, la forma del histograma va acercándose al gráfico de la **función de densidad de probabilidad** de la variable.

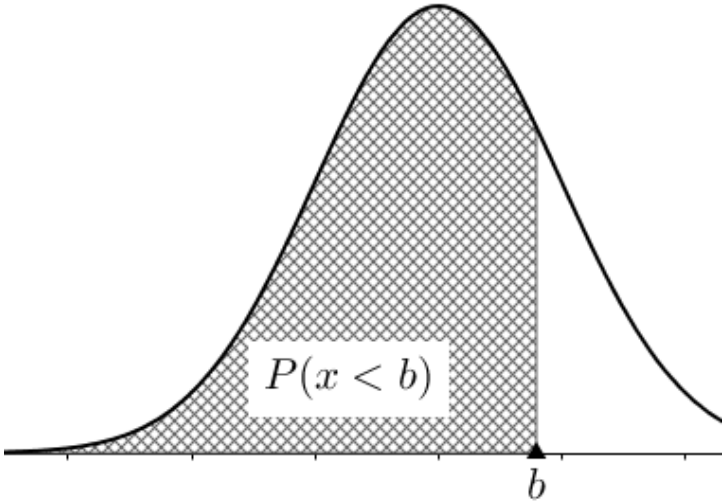
Probabilidad de una variable Aleatoria Continua

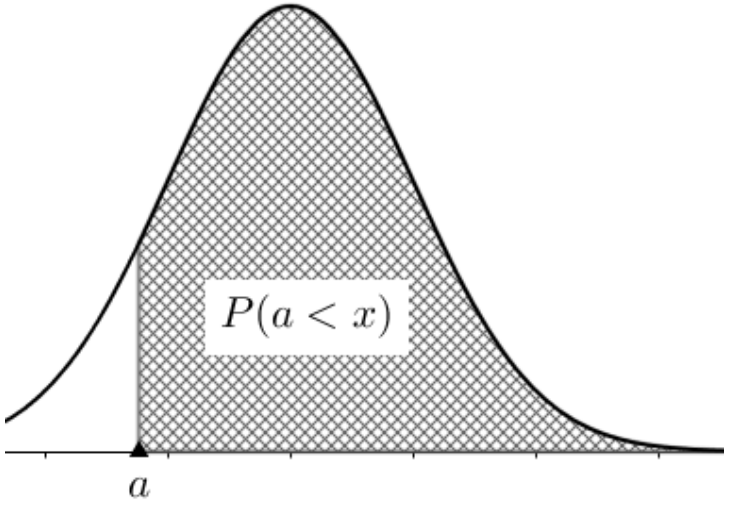
- Sean a y b valores de una variable aleatoria continua. La probabilidad que la variable aleatoria continua se encuentre entre a y b, es igual al área que encierra el gráfico de la función de densidad de probabilidad, delimitada por los extremos a y b.

Área encerrada equivalente a la probabilidad

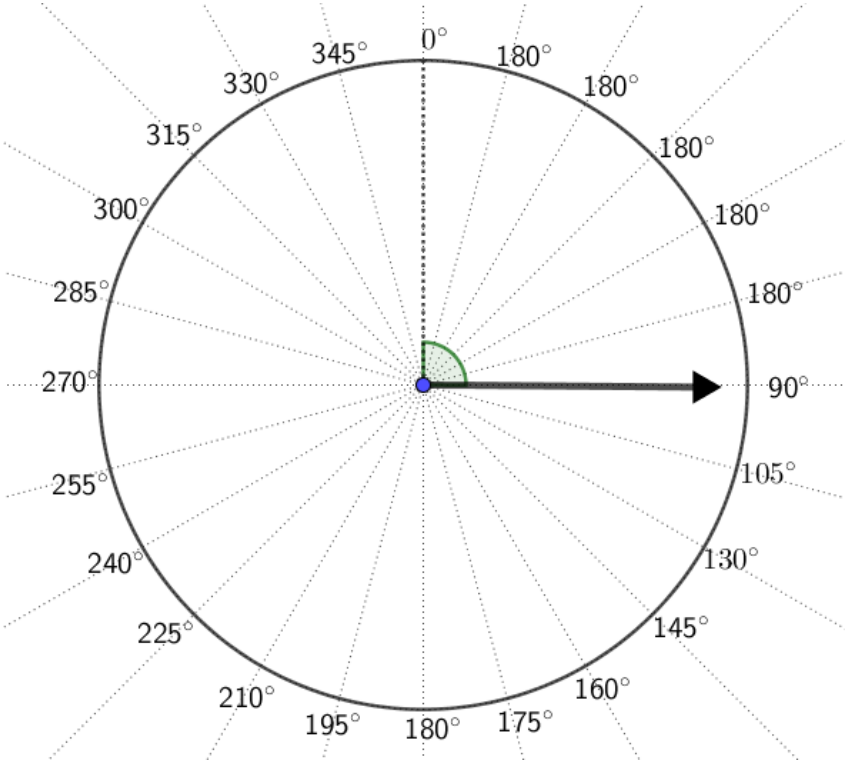
Gráfico de la función de densidad







Ejemplo 4: Suponga que se hace girar la aguja de la figura de modo que se detenga en una posición aleatoria, formando un ángulo con la vertical. En la figura se muestra un ejemplo.



- Note que la variable *ángulo formado con la vertical,* es una **variable aleatoria continua** (llamémosla X) que puede tomar cualquier valor entre 0° y 360°.
- Hay infinitos valores entre 0° y 360°.

Ejemplo 4

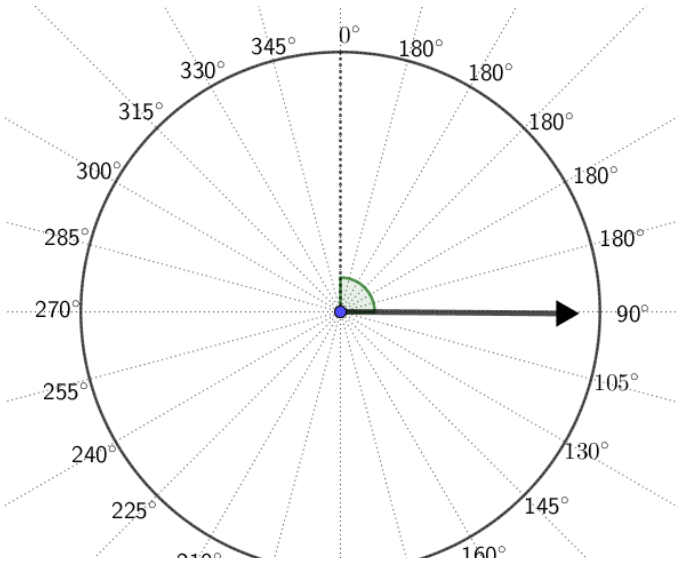
- La función de densidad de probabilidad de esta variable aleatoria es:

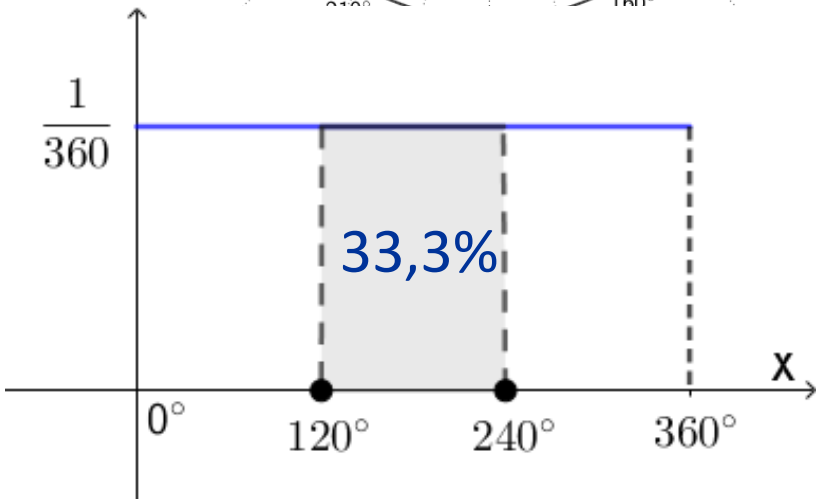
$$f(X) = \frac{1}{360}$$

- Es una función constante. La línea azul corresponde al gráfico de la función de densidad $f(X) = \frac{1}{360}$

- Para conocer, por ejemplo, la probabilidad de que la aguja caiga formando un ángulo que esté entre 120° y 240° con la vertical, calculamos **el área comprendida bajo el gráfico de la función delimitada por 120° y 240° (área sombreada)** que en este caso es la de un rectángulo:

$$P(120 \leq X \leq 240) = (240 - 120) \cdot \frac{1}{360} = 0,\overline{3}$$





$$A_{rectangulo} = base \cdot altura$$

La probabilidad es $0,\overline{3}$ o aprox. 33,3%.

Probabilidad de una variable Aleatoria Continua

- La probabilidad que una variable aleatoria continua tome un valor fijo es cero, esto es: $P(X = a) = 0$

Por ejemplo: Consideremos una v.a. cuyos valores son las estaturas de toda la gente mayor de 18 años de edad. Entre dos alturas cualquiera, por ej. entre 1,63 y 1,65 metros, **hay un número infinito de alturas posibles**, una de las cuales es 1,6435145 metros.

La probabilidad de seleccionar una persona al azar que mida exactamente 1,6435145 metros de altura , es decir, que no sea una persona del conjunto infinitamente grande de alturas cercanas a 1,64 metros es remota, por esto asignamos una probabilidad cero a tal evento.

Distribución Normal

- Una variable aleatoria (V.A.) continua se dice que se distribuye en forma normal cuando su función de densidad de probabilidad está dada por:

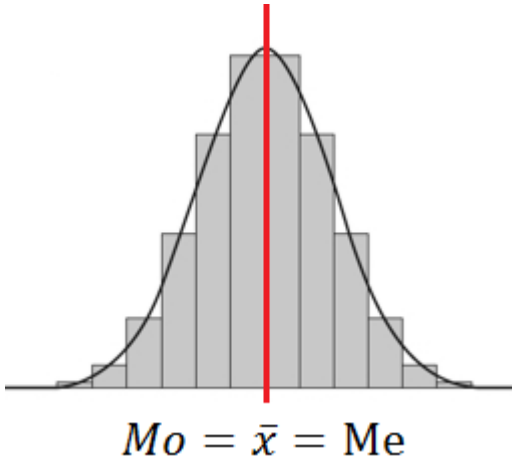
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Es la más importante y la de mayor uso de todas las distribuciones continuas de probabilidad.

Características de la Distribución Normal

Una V.A. continua que se distribuye normalmente cumple con:

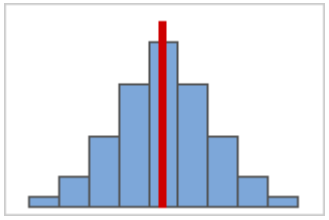
1. La función de densidad de probabilidad tiene **forma de campana**



2. Tiene una única moda, que coincide con su media y su mediana.

Distribución Normal

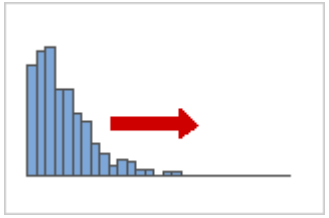
3. Su **Coeficiente de asimetría es muy próximo a cero.**
   - Coeficiente de asimetría: Permite identificar si los datos se distribuyen de forma simétrica o asimétrica alrededor de la media aritmética



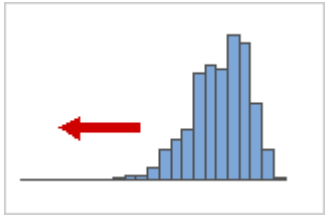
La curva de la distribución normal es simétrica por lo que en teoría este coeficiente debe ser cero, pero en la práctica se admite que esté:

$-0,5 < \text{Coeficiente de asimetría} < 0,5$

Si la distribución **no es simétrica no es normal**. En ese caso podría ser:



Curva asimétrica positiva
Coeficiente de asimetría > 0,5

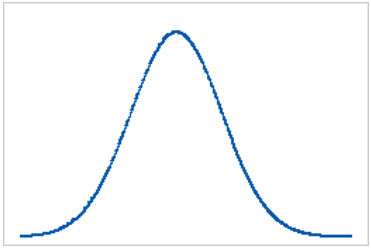


Curva asimétrica negativa
Coeficiente de asimetría < -0,5

Distribución Normal

4. Su curtosis **es muy próxima a cero**

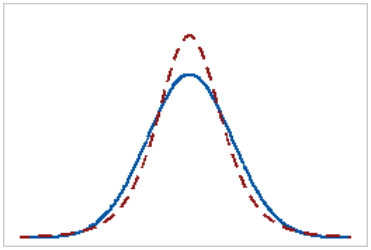
Curtosis: Permite determinar el grado de concentración de los datos en torno a la media aritmética.



Los datos distribuidos normalmente son la línea de base para la curtosis, por lo que tienen curtosis cero, pero en la práctica se admite:

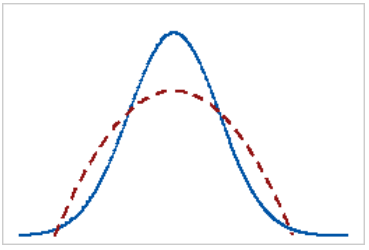
$-0{,}5 < \text{Curtosis} < 0{,}5$

En caso contrario, **no es normal**:



Alta concentración en torno a la media (Leptocúrtica)

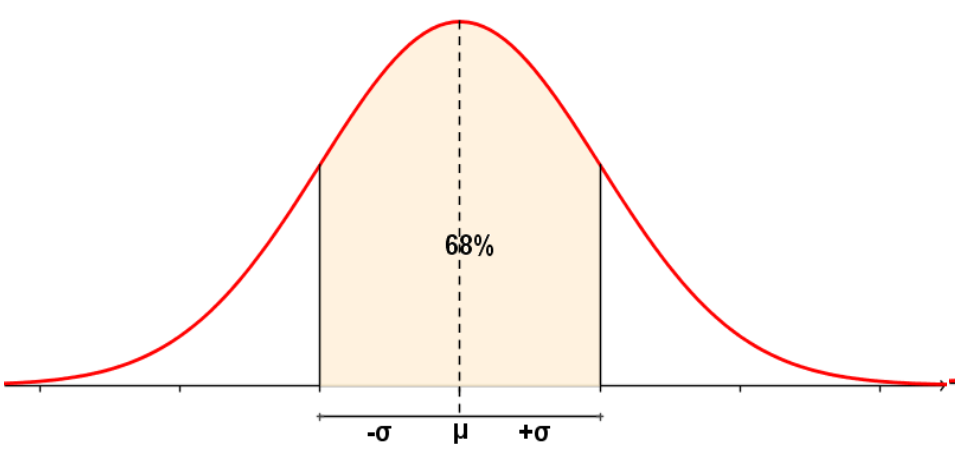
Curtosis > 0,5

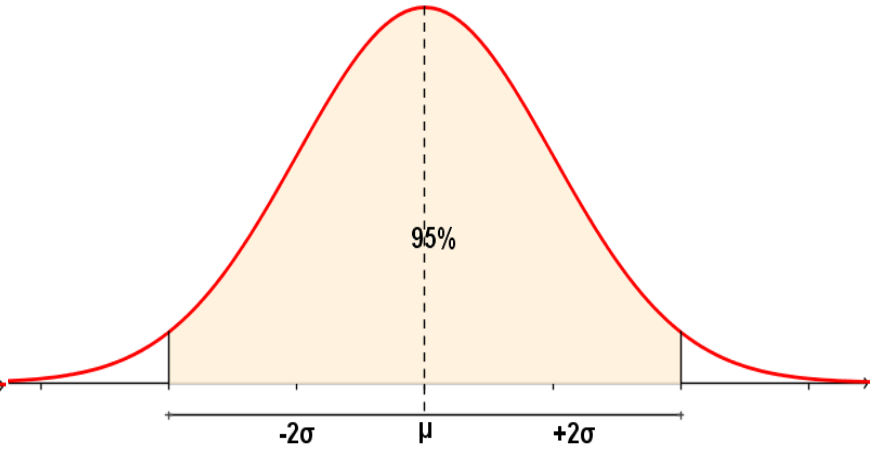


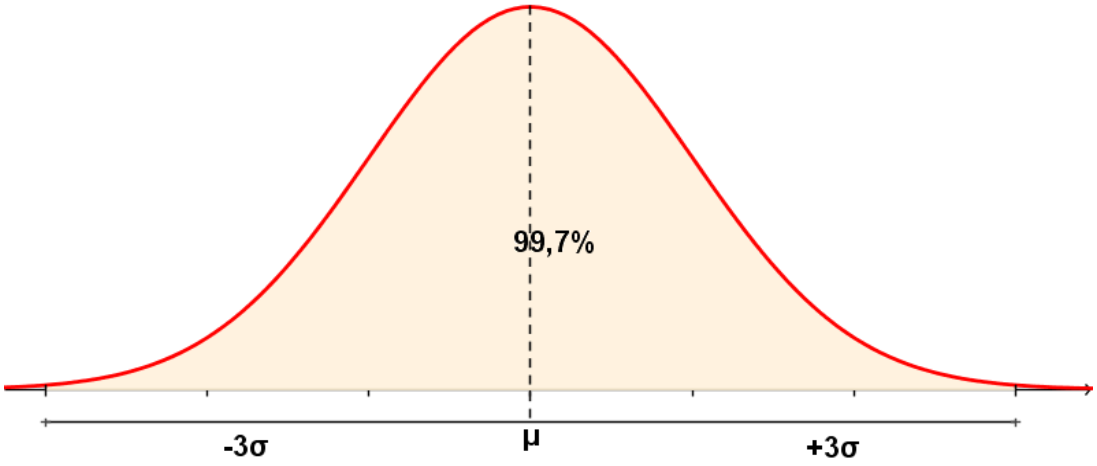
Baja concentración en torno a la media (Platicúrtica)

Curtosis < -0,5

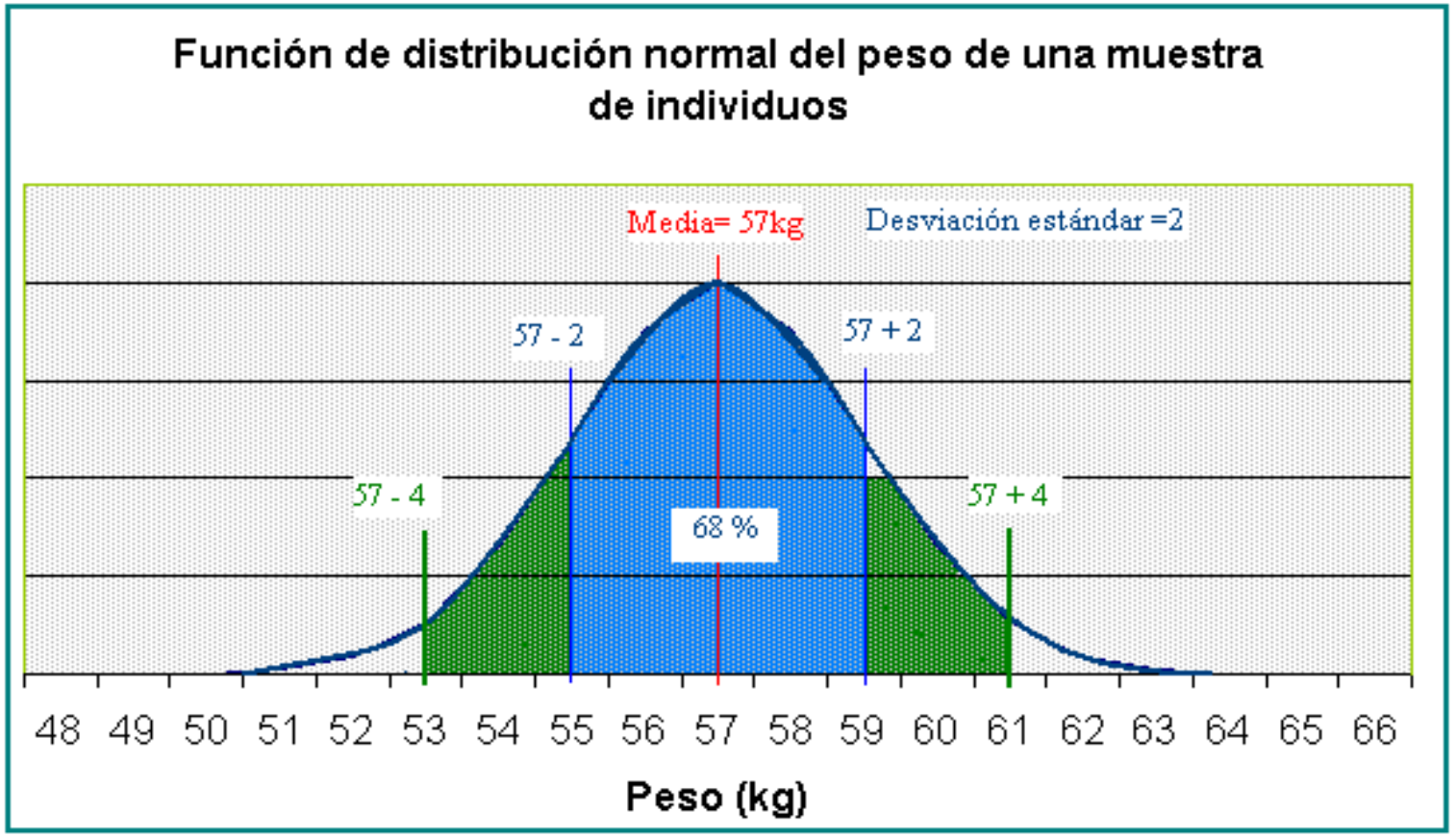
5. El área bajo la curva comprendida entre los valores situados aproximadamente a una desviación estándar de la media es aproximadamente de un 68%, a dos desviaciones estándar de la media es aproximadamente de un 95% y a tres desviaciones estándar de la media es aproximadamente de un 99,7%.







Ejemplo gráfico de una Distribución Normal



Prueba de Kolmogorov - Smirnov

Es una prueba estadística que contrasta si los datos de una variable proceden o no de una distribución especificada. En este caso, **la usaremos para contrastar los datos y confirmar que provienen de una distribución normal**. Esto se realiza comparando las frecuencias relativas acumuladas de una tabla de distribución de frecuencias con los valores obtenidos por la fórmula del modelo normal (=DISTR.NORM(x ; μ ; σ ; 1)).

Para usar esta prueba lo primero que se debe tener es una tabla de distribución de frecuencias, con la frecuencia relativa acumulada en ella. A manera de ejemplo, se adjunta una tabla de distribución de frecuencias para 500 alumnos que rindieron la prueba PSU.

LIM INFERIOR	LIM SUPERIOR
273	331
331	389
389	447
447	505
505	563
563	621
621	679
679	737
737	795
795	850

PSU Grupo 1	***f***	***F***	***H%***
273-331	9	9	0,018
331-389	35	44	0,088
389-447	47	91	0,182
447-505	71	162	0,324
505-563	88	250	0,5
563-621	86	336	0,672
621-679	62	398	0,796
679-737	56	454	0,908
737-795	27	481	0,962
795-850	19	500	1

Prueba de Kolmogorov - Smirnov

Una vez que poseemos la tabla anterior, debemos obtener los valores de las frecuencias relativas, pero ahora usando la fórmula [= DISTR.NORM(x ; μ ; σ ; 1)] , donde "x" será el límite superior del respectivo intervalo, "μ" la media, y "σ" la desviación estándar de todos los datos considerados en el estudio.

= DISTR.NORM(x ; μ ; σ ; 1)

= DISTR.NORM(331 ; 567,41 ; 123,9189 ; 1)

LIM INFERIOR	LIM SUPERIOR
273	331
331	389
389	447
447	505
505	563
563	621
621	679
679	737
737	795
795	850

PSU Grupo 1	*f*	F	H%	F(x)
273-331	9	9	0,018	0,028209828
331-389	35	44	0,088	0,074971624
389-447	47	91	0,182	0,165603914
447-505	71	162	0,324	0,307258649
505-563	88	250	0,5	0,485805521
563-621	86	336	0,672	0,667296537
621-679	62	398	0,796	0,8160751
679-737	56	454	0,908	0,914431006
737-795	27	481	0,962	0,966865844
795-850	19	500	1	0,989406978

$\mu = Media\ aritmética\ o\ Promedio$ = PROMEDIO(rango de datos)
= 567,41

$\sigma = Desviación\ estándar\ poblacional$ = DESVESTP(rango de datos)
= 123,9189

Prueba de Kolmogorov - Smirnov

Luego debemos calcular las diferencias positivas entre las frecuencias relativas acumuladas (H%) y las obtenidas con la fórmula (F(x)).

= ABS(0,018 – 0,028209)

PSU Grupo 1	*f*	F	H%	F(x)	Diferencia +
273-331	9	9	0,018	0,028209828	0,010209828
331-389	35	44	0,088	0,074971624	0,013028376
389-447	47	91	0,182	0,165603914	0,016396086
447-505	71	162	0,324	0,307258649	0,016741351
505-563	88	250	0,5	0,485805521	0,014194479
563-621	86	336	0,672	0,667296537	0,004703463
621-679	62	398	0,796	0,8160751	0,0200751
679-737	56	454	0,908	0,914431006	0,006431006
737-795	27	481	0,962	0,966865844	0,004865844
795-853	19	500	1	0,989406978	0,010593022

= MAX (rango)]
= 0,0200751
Esta es la mayor diferencia

El valor de prueba que usaremos para comparar será el mayor de todas las diferencias obtenidas [= MAX (rango)].

Prueba de Kolmogorov - Smirnov

Para realizar el contraste deberemos comparar el valor de prueba obtenido antes, con el valor crítico de Kolmogorov. Para calcular el valor de Kolmogorov, se realizará la operación:

$$\text{Valor crítico de Kolmogorov} = \frac{0{,}886}{\sqrt{N}} = \frac{0{,}886}{\sqrt{500}} = 0{,}0396\ldots$$

$$donde\ N\ es\ el\ total\ de\ datos$$

En nuestro caso consideramos N = 500, ya que la base de datos contenía 500 puntajes PSU.

Se comparan los valores:

Valor de prueba	**¿Cuál es menor?**	Valor Kolmogorov
0,0201		0,0396

Cuando el valor de prueba sea menor, se podrá concluir que los datos **pertenecen a una distribución normal**, en caso contrario, los datos no pertenecerán a una distribución normal de probabilidades.

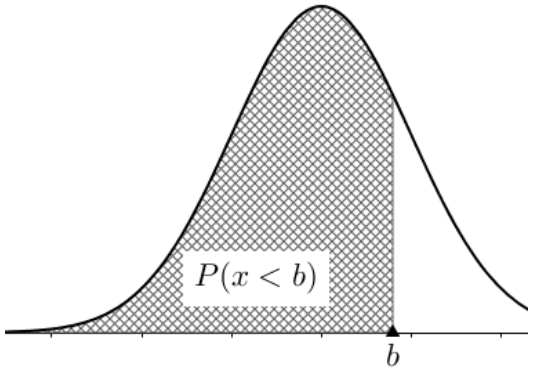
Uso de la función DISTR.NORM en Excel para calcular probabilidades.

Por ejemplo:

Las fórmulas en Excel para calcular probabilidades con el modelo normal, serían las siguientes:

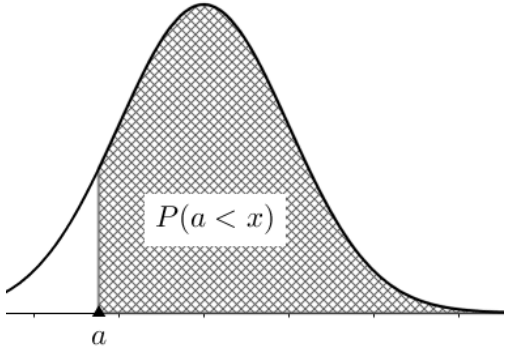
a) La probabilidad de que X sea a lo más ***k***, o menor que ***k***, es:

$$P(X < k) = P(X \leq k) = DISTR.NORM(k; \mu; \sigma; 1)$$



b) La probabilidad de que X sea al menos ***k*** o mayor que ***k***, es:

$$P(X \geq k) = P(X > k) = 1 - DISTR.NORM(k; \mu; \sigma; 1)$$



Observación General:

$\mu = Media\ aritmética\ o\ Promedio$ =PROMEDIO(rango de datos)

$\sigma = Desviación\ estándar\ poblacional$ =DESVESTP(rango de datos)

Ejercicio, utilizando funciones de Excel

El tiempo en horas que una persona dedica a navegar en Internet en una semana se puede considerar como una variable aleatoria, digamos X, la cual tiene un comportamiento aproximadamente normal, con un tiempo medio de navegación de 10 horas a la semana y una desviación estándar del tiempo de navegación en Internet de 2,5 horas. Se selecciona a una persona al azar, determine:

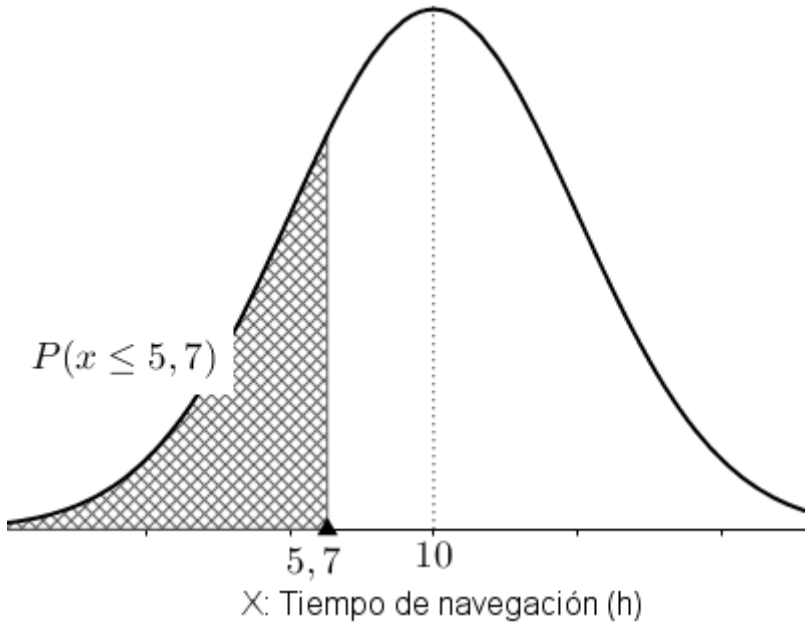
a) ¿Cuál es la probabilidad de que haya navegado en Internet a lo más 5,7 horas en una semana?

b) ¿Cuál es la probabilidad que haya navegado en Internet entre 6,2 horas y 14,8 horas en una semana?

c) ¿Cuál es la probabilidad que haya navegado en Internet por más de 12,5 horas en una semana?

Los resultados se calcularán utilizando la Distribución Normal de Excel.

a) ¿Cuál es la probabilidad de que haya navegado en Internet a los más 5,7 horas en una semana?



Respuesta:

$\mu = Media\ \ aritmética\ \ o\ \ Promedio = 10$

$\sigma = Desviación\ \ estándar\ \ poblacional = 2{,}5$

P(X ≤ 5,7) = DISTR.NORM(5,7;10;2,5;1) = 0,0427

La probabilidad es de un 4,3%.

b) ¿Cuál es la probabilidad que haya navegado en Internet entre 6,2 horas y 14,8 horas en una semana?

Respuesta:

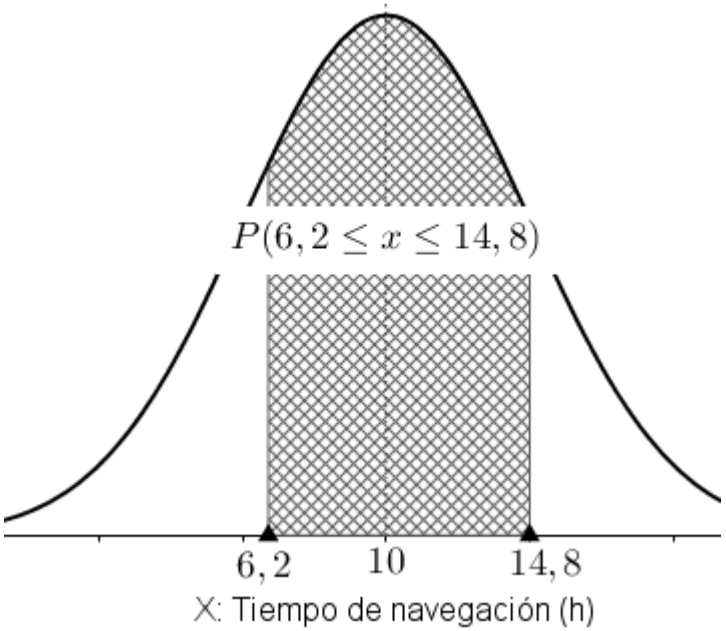
$P(6,2 < X < 14,8) = P(X < 14,8) - P(X < 6,2)$

$= DISTR.NORM(14,8;10;2,5;1) - DISTR.NORM(6,2;10;2,5;1)$

$= 0,97257105 - 0,064255488$

$= 0,9083$

La probabilidad es de un 90,8%.



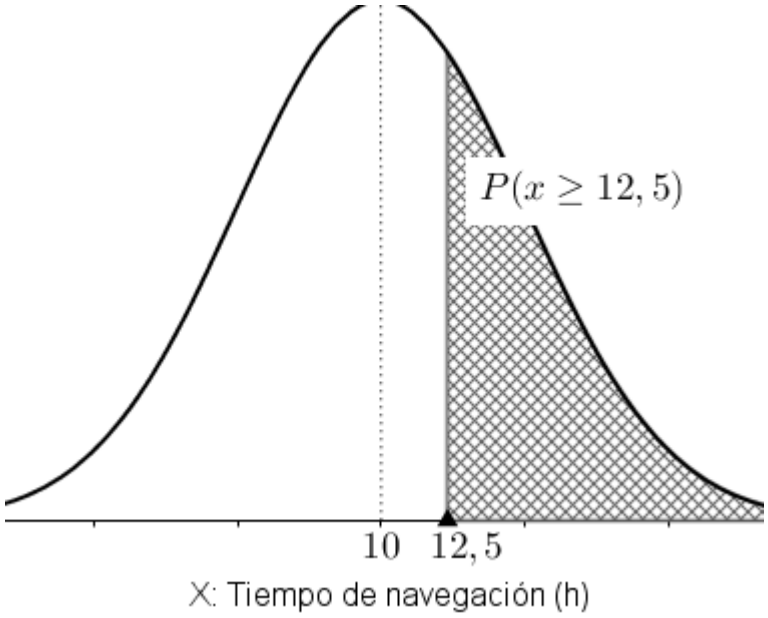
c) ¿Cuál es la probabilidad que haya navegado en Internet por más de 12,5 horas en una semana?

Respuesta:

$P(X > 12,5) = 1 - P(X < 12,5) = 1 - DISTR.NORM(12,5;10;2,5;1)$

$= 1 - 0,84134475 = 0,1587$

La probabilidad es de un 15,9%.



Ejercicios Propuestos

1) Estudios realizados recientemente por una AFP han revelado que el tiempo de duración de sus cuentas de ahorro voluntario de afiliados (lapso en el que las cuentas de ahorro voluntario permanecen abiertas) sigue un comportamiento normal con una media de 24 meses y una desviación estándar de 10,2 meses. Determine:

   a) Si un afiliado abre una cuenta de ahorro voluntario en la AFP, ¿cuál es la probabilidad de que la cuenta tenga dinero después de 32 meses?

   b) Si la AFP tiene 3.460 cuentas, ¿cuántas de ellas tendrán dinero entre 18 y 30 meses?

Ejercicios Propuestos

2) La fábrica de neumáticas GOMALETS produce un tipo de neumático, que tiene vida útil distribuida normalmente con media de 80.000 km. y una desviación estándar de 8.000 km. Se pide:

a) ¿Cuál es la probabilidad de que un neumático dure no más de 89.000 km.?

b) ¿Cuál es la probabilidad de que un neumático tenga una vida útil entre 76.000 y 105.920 km.?

c) Si se fabrican 2.850 neumáticos. ¿Cuántos de ellos tendrían una vida útil de por lo menos 75.800 km.?

Respuestas Ejercicios Propuestos

1) a) La probabilidad es de un 21,6%.

   b) Aproximadamente habrán 1.535 cuentas con dinero entre 18 y 30 meses.

2) a) La probabilidad es de un 87%.

   b) La probabilidad es de un 69,1%.

   c) Aproximadamente habrán 1.996 neumáticos con una vida útil de al menos 75.800 km.

Observación: Para los resultados se utilizó la planilla de cálculo Excel.