ESTADÍSTICA

Introducción

En nuestro contacto con objetos, seres o conceptos, nos damos cuenta que tienen propiedades que se pueden clasificar en cualitativas y en cuantitativas.

El color de las uvas es un ejemplo de propiedad cualitativa y la altura de los árboles es un ejemplo de propiedad cuantitativa.

También se puede hablar de propiedades cuasi cuantitativas como cuando decimos que un programa de televisión es bueno, regular o malo, porque podemos darle un orden a esas clasificaciones.

En la Estadística estas propiedades se toman como variables, y cada valor de ella (repetido o no) se denomina dato.

Las variables cuantitativas se pueden clasificar en continuas y discontinuas. La altura de los árboles es una variable continua, porque toma cualquier valor real, dentro de un intervalo de ellos. En cambio, el número de hijos de un matrimonio toma valores enteros no negativos, solamente.

Estadística

De una forma simple, podemos decir que una estadística es una recolección de datos que se presentan de una manera ordenada y clasificada de acuerdo a un cierto criterio, y de conclusiones a partir de ellos.

De manera más formal, la Estadística como ciencia, trata sobre los métodos y procedimientos para coleccionar, clasificar, presentar y analizar datos obtenidos de la realidad. Además realiza inferencias o conclusiones, lógicamente aceptadas, que sirven para hacer predicciones y tomar decisiones.

En la Estadística tenemos dos ramas bien definidas:

Estadística Descriptiva

En la Estadística Descriptiva se presentan y analizan los datos obtenidos.

Estadística Inferencial

En la Estadística Inferencial se realizan conclusiones a partir de los datos obtenidos.

Observación: también existe la Estadística Matemática que aporta las bases teóricas en que se apoya la Estadística.

Población

Población es el universo en el cual se quiere estudiar, al menos, una propiedad de él.

Muestra

Se denomina muestra al subconjunto de ese universo y del cual se recopilarán los datos. Es necesario que esa muestra sea debidamente representativa.

Supongamos que se requiere conocer el número de hijos por matrimonio de una villa. Para este propósito se elige una muestra representativa de 50 matrimonios de ella, y se obtienen los siguientes datos:

2 , 2 , 4 , 1 , 3 , 5 , 3 , 2 , 1 , 6 , 3 , 4 , 1 , 2 , 0 , 2 , 3 , 1 , 7 , 4 , 2 , 3 , 0 , 5 , 1 , 4 , 3 , 2 , 4 , 1 , 5 , 2 , 1 , 2 , 4 , 0 , 3 , 3 , 2 , 6 , 1 , 5 , 4 , 2 , 0 , 3 , 2 , 4 , 3 , 1 .

El número total de datos se representa con la letra **n**. En este ejemplo **n** = 50.

Junio 2017
http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Frecuencia absoluta (f_i)

La frecuencia absoluta es el número de veces que aparece un valor (x_i) en los datos obtenidos.

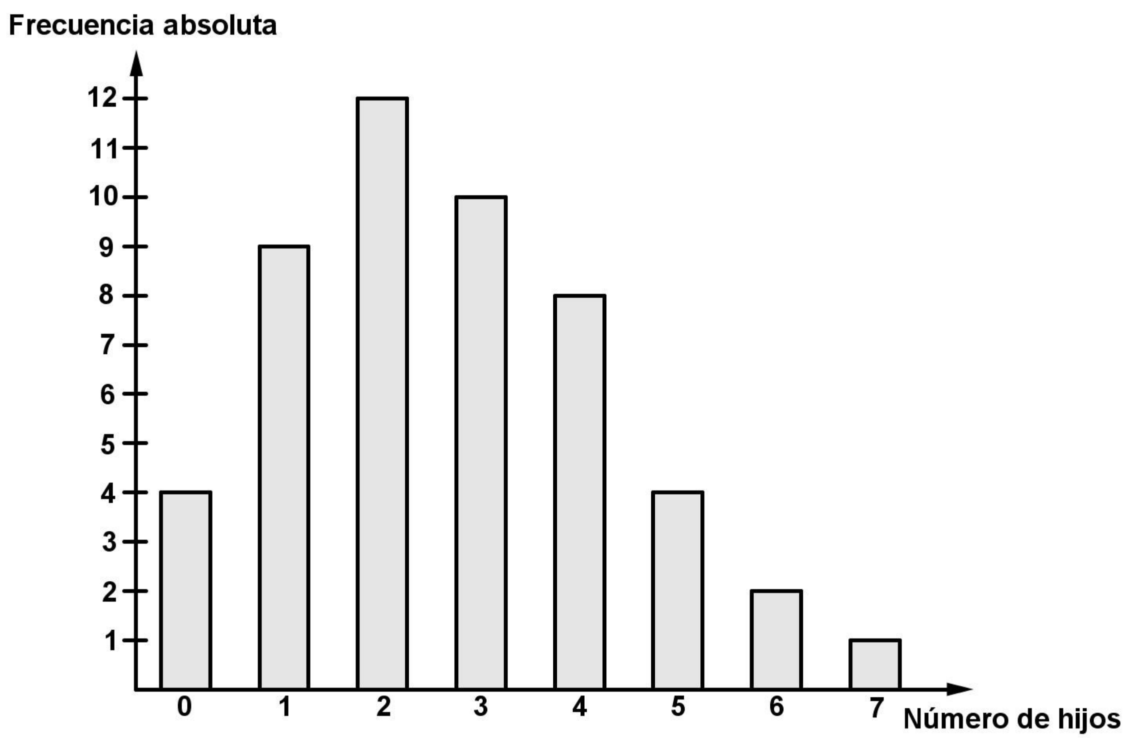
En nuestro ejemplo, la frecuencia absoluta indica el número de familias que tienen una cierta cantidad de hijos:

x_i	f_i
0	4
1	9
2	12
3	10
4	8
5	4
6	2
7	1

Ejemplos 1

Ejercicios 1

Gráfico de barras de frecuencia absoluta



Junio 2017
http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Frecuencia absoluta acumulada (F_i)

La frecuencia absoluta acumulada indica cuantos elementos de la lista de datos son menores o iguales a un valor dado. Es la suma de las frecuencias absolutas desde la primera fila hasta la fila elegida.

Por ejemplo, sabemos que hay 25 matrimonios de la muestra que tienen a lo más 2 hijos:

x_i	f_i	F_i
0	4	4
1	9	13
2	12	25
3	10	35
4	8	43
5	4	47
6	2	49
7	1	50

Ejemplos 2

Ejercicios 2

Frecuencia relativa (h_i)

La frecuencia relativa es el cociente entre la frecuencia absoluta (f_i) y el número total de datos (n) :

x_i	f_i	h_i
0	4	0,08
1	9	0,18
2	12	0,24
3	10	0,20
4	8	0,16
5	4	0,08
6	2	0,04
7	1	0,02

Ejemplos 3

Ejercicios 3

Frecuencia relativa acumulada (H_i)

Análogamente a como se define la frecuencia absoluta acumulada, se define la frecuencia relativa acumulada:

x_i	h_i	H_i
0	0,08	0,08
1	0,18	0,26
2	0,24	0,50
3	0,20	0,70
4	0,16	0,86
5	0,08	0,94
6	0,04	0,98
7	0,02	1,00

http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Frecuencia relativa porcentual (p_i)

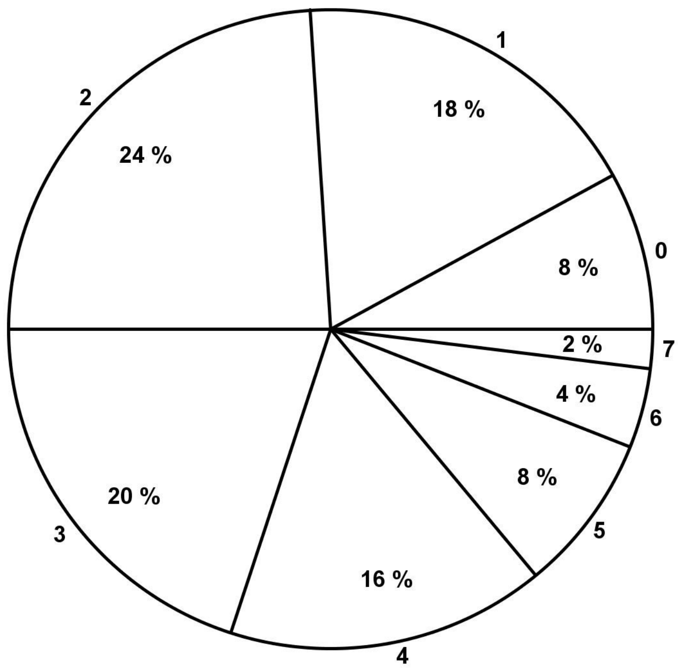
La frecuencia porcentual es la frecuencia relativa (h_i) expresada en forma porcentual. En otras palabras, es la frecuencia relativa (h_i) multiplicada por 100 :

x_i	h_i	p_i (%)
0	0,08	8 %
1	0,18	18 %
2	0,24	24 %
3	0,20	20 %
4	0,16	16 %
5	0,08	8 %
6	0,04	4 %
7	0,02	2 %

Ejemplos 4

Ejercicios 4

Gráfico circular de frecuencia relativa porcentual



Frecuencia relativa porcentual acumulada (P_i)

La frecuencia porcentual acumulada es la frecuencia relativa acumulada (H_i) multiplicada por 100 :

x_i	p_i (%)	P_i (%)
0	8 %	8 %
1	18 %	26 %
2	24 %	50 %
3	20 %	70 %
4	16 %	86 %
5	8 %	94 %
6	4 %	98 %
7	2 %	100 %

Junio 2017
http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Prueba 1
Prueba 2
Prueba 3
Prueba 4
Prueba 5

ESTADÍGRAFOS DE POSICIÓN

Percentil (P_k)

Cuando los datos ordenados de menor a mayor se dividen en 100 grupos iguales, hablamos de percentiles, por ejemplo:

P_{40} : Es el valor menor de la variable que agrupa al menos el 40 % de los datos. En nuestro ejemplo es 2 .

Ejemplos 5

Decil (D_k)

Cuando los datos ordenados de menor a mayor se dividen en 10 grupos iguales, hablamos de deciles, por ejemplo:

D_2 : Es el valor menor de la variable que agrupa al menos el 20 % de los datos. En nuestro ejemplo es 1 .

Quintil (Q_k)

Cuando los datos ordenados de menor a mayor se dividen en 5 grupos iguales, hablamos de quintiles, por ejemplo:

Q_4 : Es el valor menor de la variable que agrupa al menos el 80 % de los datos. En nuestro ejemplo es 4 .

Cuartil (C_k)

Cuando los datos ordenados de menor a mayor se dividen en 4 grupos iguales, hablamos de cuartiles, por ejemplo:

C_3 : Es el valor menor de la variable que agrupa al menos el 75 % de los datos. En nuestro ejemplo es 4 .

Ejemplos 6

ESTADÍGRAFOS DE TENDENCIA CENTRAL

Media aritmética (μ)

La media aritmética es el promedio aritmético de los valores numéricos obtenidos:

$$\mu = \frac{\sum_{i=1}^{i=m} x_i f_i}{n} = \sum_{i=1}^{i=m} x_i h_i$$

En nuestro ejemplo:

$$\mu = \frac{0 \times 4 + 1 \times 9 + 2 \times 12 + 3 \times 10 + 4 \times 8 + 5 \times 4 + 6 \times 2 + 7 \times 1}{50} = 2{,}68$$

O bien:

$$\mu = 0 \times 0{,}08 + 1 \times 0{,}18 + 2 \times 0{,}24 + 3 \times 0{,}20 + 4 \times 0{,}16 + 5 \times 0{,}08 + 6 \times 0{,}04 + 7 \times 0{,}02 = 2{,}68$$

Ejemplos 7
Ejercicios 7

http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Teorema: la media aritmética o promedio de una población finita se calcula promediando todas las medias aritméticas de todas las muestras de igual tamaño que se pueden extraer de esa población.

Ejemplo: una cierta población se compone de 3 grupos de igual tamaño, A , B y C. Para medir una variable aleatoria se tomaron 3 muestras y sus resultados se muestran a continuación:

Muestra	Media aritmética
{ A , B }	18
{ B , C }	12,5
{ C , A }	14,5

La media aritmética de la población es:

$$\mu = \frac{18 + 12{,}5 + 14{,}5}{3} = 15$$

Moda (Mo)

La moda es el valor numérico de mayor frecuencia absoluta (f_i). A veces hay más de un valor numérico que satisface lo anterior. En nuestro ejemplo, la moda es 2.

Ejemplos 8

Ejercicios 8

Mediana (Me)

Al ordenar de mayor a menor (o al revés) los valores numéricos, la mediana es el valor central, si el número de ellos es impar, o promedio aritmético de los valores centrales, si el número de datos es par.

En nuestro ejemplo, los valores centrales son 2 y 3 . Por lo tanto la mediana es 2,5.

P_{50} , D_5 y C_2 son iguales a la mediana.

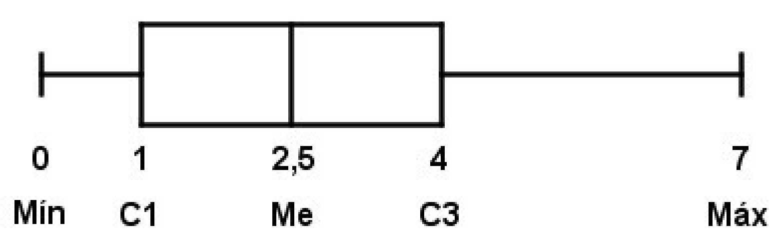
Ejemplos 9

Ejercicios 9

Diagrama de caja y bigotes

Este diagrama se usa para visualizar simetrías y dispersión de los datos. En él aparecen los valores mínimo (Mín) y máximo (Máx), la mediana (Me) y los cuartiles C_1 (25%) y C_3 (75%).

En nuestro ejemplo queda claramente en evidencia que hay mayor dispersión de los datos entre C_3 y Máx que entre C_1 y Mín. Además hay simetría entre la mediana y los cuartiles C_1 y C_3.



Junio 2017
http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Prueba 6

Prueba 7

Prueba 8

Prueba 9

Prueba 10

ESTADÍGRAFOS DE DESVIACIÓN

Recorrido de la variable o cambio de variación (L)

El recorrido de la variable o cambio de variación es la diferencia entre el valor máximo y valor mínimo de la variable.

$$L = x_{máx} - x_{mín}$$

En nuestro ejemplo:

$L = 7 - 0 = 7$

Desviación media (D)

La desviación media es el promedio aritmético de los valores absolutos de las desviaciones con respecto de la media aritmética (μ).

$$D = \frac{\sum_{i=1}^{n} |x_i - \mu|}{n}$$

Varianza (V)

La varianza es el promedio de los cuadrados de las desviaciones con respecto a la media aritmética (μ).

$$V = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

Desviación típica o estándar (σ)

La desviación típica o estándar es la raíz cuadrada de la varianza (V).

$$\sigma = \sqrt{V}$$

En nuestro ejemplo:

x_i	f_i	$\lvert x_i - \mu \rvert$	$(x_i - \mu)^2$
0	4	2,68	7,1824
1	9	1,68	2,8224
2	12	0,68	0,4624
3	10	0,32	0,1024
4	8	1,32	1,7424
5	4	2,32	5,3824
6	2	3,32	11,0224
7	1	4,32	18,6624

$D = 1,36$ $\quad V = 2,7376$ $\quad \sigma = 1,6546$

http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Teorema: Sea $P = \{ x_i : x_i \in R \}$ un conjunto de datos obtenidos de una población; $\mu(x_i)$, la media aritmética de ellos; $Me(x_i)$, su mediana; $Mo(x_i)$, su moda; $V(x_i)$, su varianza y $\sigma(x_i)$, su desviación estándar. Si cada uno de los x_i se transforman en $a x_i + b$, donde a y b son números reales, entonces:

$$\mu(a x_i + b) = a\,\mu(x_i) + b$$

$$Me(a x_i + b) = a\,Me(x_i) + b$$

$$Mo(a x_i + b) = a\,Mo(x_i) + b$$

$$V(a x_i + b) = a^2\,V(x_i)$$

$$\sigma(a x_i + b) = a\,\sigma(x_i)$$

Ejemplo:

x_i	$a x_i + b$	
1,9	$2 \times 1{,}9 + 1$	4,8
1,9	$2 \times 1{,}9 + 1$	4,8
2,1	$2 \times 2{,}1 + 1$	5,2
2,1	$2 \times 2{,}1 + 1$	5,2
Media aritmética		
2	$2 \times 2 + 1$	5
Mediana		
2	$2 \times 2 + 1$	5
Modas		
1,9	$2 \times 1{,}9 + 1$	4,8
2,1	$2 \times 2{,}1 + 1$	5,2
Varianza		
0,01	$2^2 \times 0{,}01$	0,04
Desviación estándar		
0,1	$2 \times 0{,}1$	0,2

Ejemplos 10

http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Distribución de los datos en intervalos

A veces es más conveniente agrupar los datos en intervalos, por ejemplo los sueldos, en miles de pesos, de 40 trabajadores de un colegio están distribuidos de la forma siguiente:

Intervalo	Marca de clase (x_i)	Frecuencia absoluta (f_i)
100 – 200	150	5
200 – 300	250	8
300 – 400	350	11
400 – 500	450	9
500 – 600	550	4
600 – 700	650	3

Observación: a falta de alguna causa especial, se toma como marca de clase al promedio aritmético de los valores extremos de cada intervalo.

De esa manera es posible construir la tabla y gráfico como en el ejemplo anterior.

Intervalo	x_i	f_i	F_i	h_i	H_i	p_i (%)	P_i (%)
100 – 200	150	5	5	0,125	0,125	12,5 %	12,5 %
200 – 300	250	8	13	0,2	0,325	20 %	32,5 %
300 – 400	350	11	24	0,275	0,6	27,5 %	60 %
400 – 500	450	9	33	0,225	0,825	22,5 %	82,5 %
500 – 600	550	4	37	0,1	0,925	10 %	92,5 %
600 – 700	650	3	40	0,075	1	7,5 %	100 %

Cálculo de la media aritmética (μ)

La media aritmética se calcula igual que en el ejemplo anterior:

$$\mu = \frac{150 \times 5 + 250 \times 8 + 350 \times 11 + 450 \times 9 + 550 \times 4 + 650 \times 3}{40} = 370$$

O bien:

$$\mu = 150 \times 0{,}125 + 250 \times 0{,}2 + 350 \times 0{,}275 + 450 \times 0{,}225 + 550 \times 0{,}1 + 650 \times 0{,}075 = 370$$

Cálculo de la moda (Mo)

$$Mo = (x_k)_{inf} + \frac{f_{k+1}}{f_{k+1} + f_{k-1}} \Delta_k$$

Donde:

$(x_k)_{inf}$: valor del extremo inferior del intervalo donde está la moda.

$(x_k)_{sup}$: valor del extremo superior del intervalo donde está la moda.

$\Delta_k = (x_k)_{sup} - (x_k)_{inf}$

f_{k-1} : frecuencia absoluta del intervalo anterior al intervalo donde está la moda.

f_{k+1} : frecuencia absoluta del intervalo posterior al intervalo donde está la moda.

http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

En nuestro ejemplo la moda se encuentra en el tercer intervalo ($k = 3$)

$(x_3)_{inf} = 300$

$(x_3)_{sup} = 400$

$\Delta_3 = 400 - 300 = 100$

$f_2 = 8$

$f_4 = 9$

$$Mo = 300 + \frac{9}{9 + 8} \times 100 \approx 352{,}9$$

Cálculo de la mediana (Me)

$$Me = (x_k)_{inf} + \frac{F - F_{k-1}}{F_k - F_{k-1}} \Delta_k$$

Donde:

$F = n/2$

$(x_k)_{inf}$: valor del extremo inferior del intervalo donde está la mediana.

$(x_k)_{sup}$: valor del extremo superior del intervalo donde está la mediana.

$\Delta_k = (x_k)_{sup} - (x_k)_{inf}$

$F_{k\ 1}$: frecuencia absoluta acumulada del intervalo anterior al intervalo donde está la mediana.

F_k : frecuencia absoluta acumulada del intervalo donde está la mediana.

En nuestro ejemplo la mediana se encuentra en el tercer intervalo ($F = 40/2 = 20 \Rightarrow k = 3$)

$(x_3)_{inf} = 300$

$(x_3)_{sup} = 400$

$\Delta_3 = 400 - 300 = 100$

$F_2 = 13$

$F_4 = 24$

$$Me = 300 + \frac{20 - 13}{24 - 13} \times 100 \approx 363{,}6$$

http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

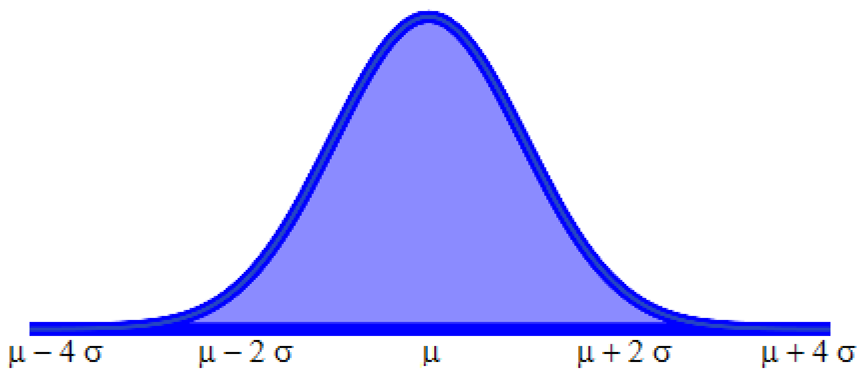
Distribución normal (de Gauss)

Se denomina distribución normal a una distribución (función) de probabilidad de una variable aleatoria continua que modela fenómenos naturales, como por ejemplo la estatura de una población humana. Esta variable recibe el nombre de variable aleatoria normal.

Esta función tiene como dominio a los números reales. Su gráfica tiene una forma acampanada y simétrica con respecto a μ .

Para representar que una variable aleatoria X sigue una distribución normal, de parámetros μ y σ^2 , se emplea la siguiente expresión matemática:

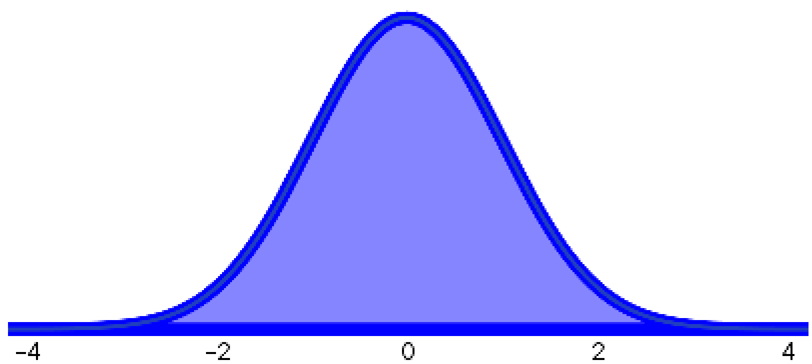
$$X \sim N (\mu , \sigma^2)$$



El área sombreada bajo la curva es igual a 1.

Ejemplos 11

Se habla de distribución normal estándar cuando $X \sim N (0 , 1)$



El área sombreada bajo la curva es igual a 1.

Teorema: si $X \sim N (\mu , \sigma^2)$, y a y b son números reales, entonces $(a X + b) \sim N (a \mu + b , a^2 \sigma^2)$

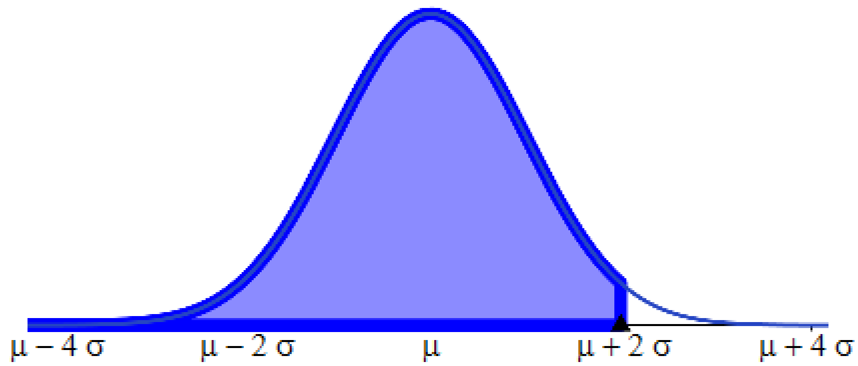
A partir de este teorema se puede relacionar cada variable aleatoria normal con una distribución normal estándar del siguiente modo:

$$Z = \frac{X - \mu}{\sigma} = \frac{1}{\sigma} X - \frac{\mu}{\sigma}$$

Mediante esta transformación se concluye finalmente que si $X \sim N (\mu , \sigma^2)$, entonces $Z \sim N (0 , 1)$.

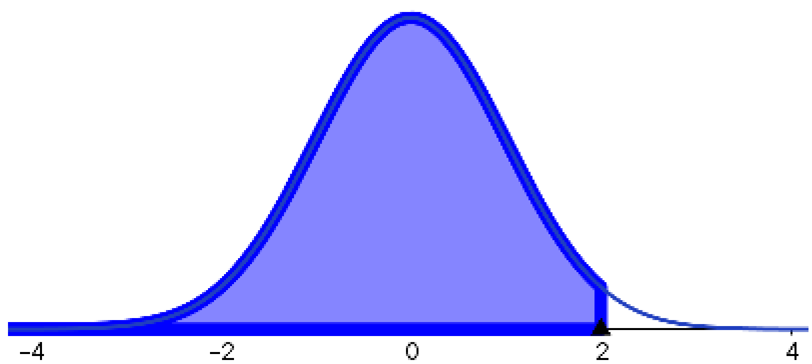
http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Sea X una variable aleatoria continua, tal que $X \sim N(\mu, \sigma^2)$. La parte sombreada de la figura dada a continuación representa a $P(X \leq x)$:



El valor del área no sombreada bajo la curva se designa como $\alpha / 2$.

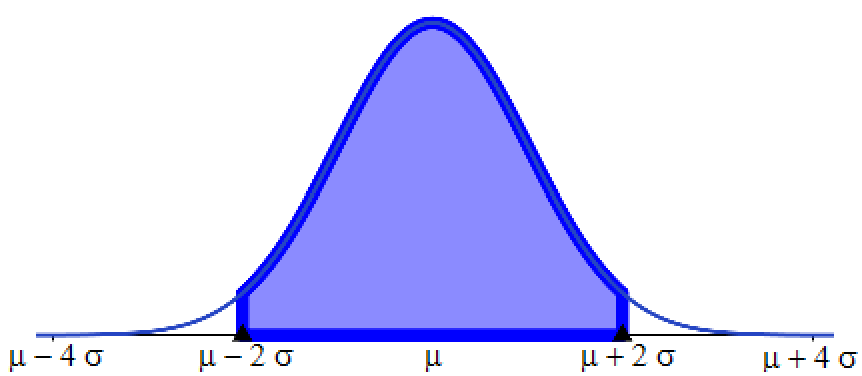
Sea Z una variable aleatoria continua, tal que $Z \sim N(0, 1)$. La parte sombreada de la figura dada a continuación representa a $P(Z \leq z)$:



El valor del área no sombreada bajo la curva se designa como $\alpha / 2$.

Ejemplos 12

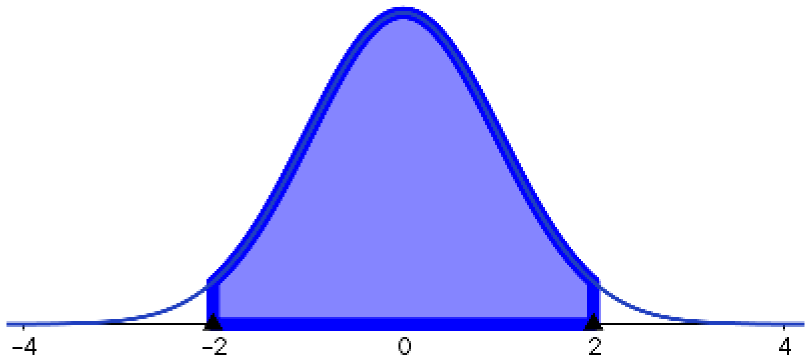
Sea X una variable aleatoria continua, tal que $X \sim N(\mu, \sigma^2)$. La parte sombreada de la figura dada a continuación representa a $P(a \leq X \leq b)$:



El valor del área no sombreada bajo la curva se designa como α.

Junio 2017
http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

Sea Z una variable aleatoria continua, tal que $Z \sim N(0, 1)$. La parte sombreada de la figura dada a continuación representa a $P(-z \le Z \le z)$:



[Ejemplos 13](#)

El valor del área no sombreada bajo la curva se designa como α.

Ejemplo: sea X una variable aleatoria continua, tal que $X \sim N(\mu, \sigma^2)$, donde se sabe que $P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0{,}9544$ y $P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0{,}9974$.

a) ¿Cuál es el valor de $P(\mu + 2\sigma \le X \le \mu + 3\sigma)$?

b) ¿Cuál es el valor de $P(\mu - 2\sigma \le X \le \mu + 3\sigma)$?

c) ¿Cuál es el valor de $P(\mu - 3\sigma \le X \le \mu + 2\sigma)$?

Respuesta: como se puede apreciar en los dos gráficos anteriores, existe una simetría con respecto a μ y 0 . Por lo tanto los valores requeridos son:

a) $\dfrac{0{,}9974 - 0{,}9544}{2} = 0{,}0215$

b) $\dfrac{0{,}9974 + 0{,}9544}{2} = 0{,}9759$

c) $\dfrac{0{,}9974 + 0{,}9544}{2} = 0{,}9759$

Intervalo de confianza para μ conociendo σ

Si M es el valor de la media aritmética de un conjunto de valores que toma una variable aleatoria de una muestra de tamaño n de una población normal con desviación estándar σ conocida, el intervalo de confianza de nivel $1 - \alpha$ $(0 < \alpha < 1)$ para μ es:

$$\left[M - z \cdot \frac{\sigma}{\sqrt{n}} \ , \ M + z \cdot \frac{\sigma}{\sqrt{n}} \right]$$

Esta expresión matemática del intervalo de confianza para μ conociendo σ se puede utilizar también para poblaciones no normales siempre que $n \ge 30$.

[Ejemplos 14](#)

El valor de z se determina usando tablas como la dada a continuación:

TABLA DE DISTRIBUCIÓN NORMAL ESTÁNDAR

z	$\alpha / 2$	$P (Z \leq z)$	$P (-z \leq Z \leq z)$
0,675	0,250	0,750	0,500
0,842	0,200	0,800	0,600
0,935	0,175	0,825	0,650
1,036	0,150	0,850	0,700
1,150	0,125	0,875	0,750
1,175	0,120	0,880	0,760
1,200	0,115	0,885	0,770
1,227	0,110	0,890	0,780
1,254	0,105	0,895	0,790
1,282	0,100	0,900	0,800
1,310	0,095	0,905	0,810
1,340	0,090	0,910	0,820
1,372	0,085	0,915	0,830
1,405	0,080	0,920	0,840
1,440	0,075	0,925	0,850
1,476	0,070	0,930	0,860
1,514	0,065	0,935	0,870
1,555	0,060	0,940	0,880
1,598	0,055	0,945	0,890
1,645	0,050	0,950	0,900
1,695	0,045	0,955	0,910
1,750	0,040	0,960	0,920
1,810	0,035	0,965	0,930
1,880	0,030	0,970	0,940
1,960	0,025	0,975	0,950
2,050	0,020	0,980	0,960
2,170	0,015	0,985	0,970
2,330	0,010	0,990	0,980
2,575	0,005	0,995	0,990

Ejemplo: en una ciudad, el número de hijos que tiene cada mujer fértil, se modela por medio de una distribución normal con media μ y varianza 0,25. Se toma una muestra aleatoria de 100 mujeres fértiles de esa ciudad, obteniéndose un promedio de 1,92 hijos . A partir de los resultados de esa muestra, ¿cuál es el intervalo de confianza de nivel 0,95 para μ?

Respuesta: a partir de la fórmula dada

$$\left[M - z \cdot \frac{\sigma}{\sqrt{n}} \ , \ M + z \cdot \frac{\sigma}{\sqrt{n}} \right]$$

se tiene que

$M = 1,92$

$\sigma = \sqrt{0,25} = 0,5$

$\sqrt{n} = \sqrt{100} = 10$

http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588

$z = 1{,}96$ este valor se obtiene a partir de la cuarta columna de la tabla dada anteriormente. Vemos que a 0,95 le corresponde $z = 1{,}96$. En caso de no contar con esa cuarta columna, tenemos que calcular el promedio entre 1 y 0,95 el cual es 0,975. Luego, usando la tercera columna llegamos al mismo valor de z.

Reemplazando los valores obtenidos, el intervalo pedido es

$$\left[\ 1{,}92 - 1{,}96 \bullet \frac{1}{20}\ ,\ 1{,}92 + 1{,}96 \bullet \frac{1}{20}\ \right]$$

BIBLIOGRAFÍA

Estadística (curso interactivo en línea con examen incluido)

© NELSON LILLO TERÁN
Junio 2017
http://www.eneayudas.cl
matematicayciencias@gmail.com
+56998581588