Módulo 4 – Inferencia Estadística

Intervalos de Confianza

Ciencia de Datos

Objetivos

- Utiliza los conceptos básicos de estadística Inferencial.
- Describir el concepto de intervalo de confianza.
- Realizar estimaciones de la media de una población utilizando intervalos de confianza a partir de una muestra aleatoria.

KIBERNUM

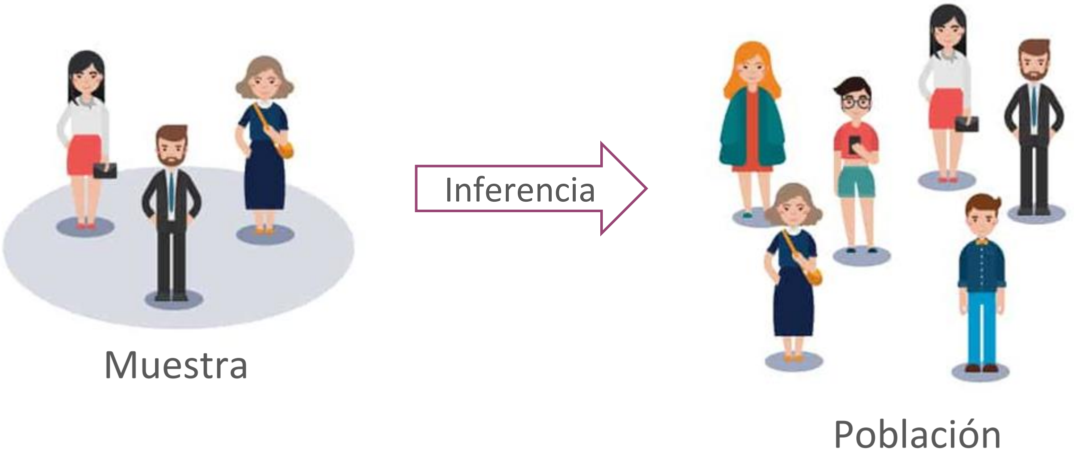
Contenido

- Intervalos de confianza.
- Intervalos de confianza para la media poblacional.

Intervalos de Confianza

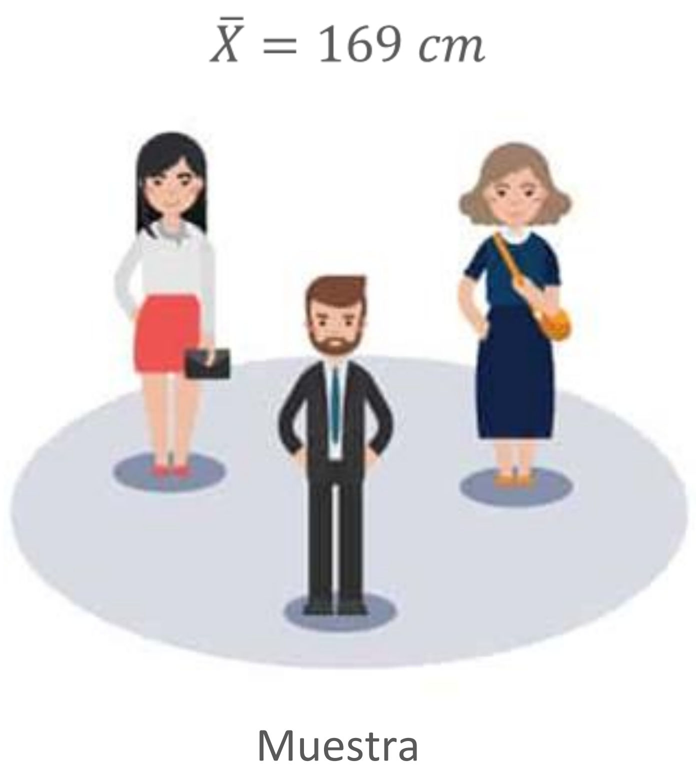
Intuición

Supongamos que se quiere **estimar la media de altura de una población de jóvenes adultos**. Para este efecto, se toma una muestra aleatoria de 50 jóvenes adultos y se mide su altura. Se encuentra que la media de altura de la muestra es de 175 cm y la desviación estándar es de 5 cm.



¿Podemos inferir la media de la población a partir de una muestra?

Intuición



Lo primero que pensaríamos es en utilizar el promedio de las alturas de las muestras, pero no sabemos qué tan alejado está del valor real de la población. La mala noticia es que posiblemente nunca sabremos el valor certero a partir de esta muestra. La buena noticia es que, a pesar de lo anterior, al menos podemos saber entre qué rango de valores podría encontrarse.

Intuición

En este caso, con un 95% de probabilidad podemos inferir que la media de la altura de la población de jóvenes adultos se encuentra en el rango 175 cms ±1.42 cms .

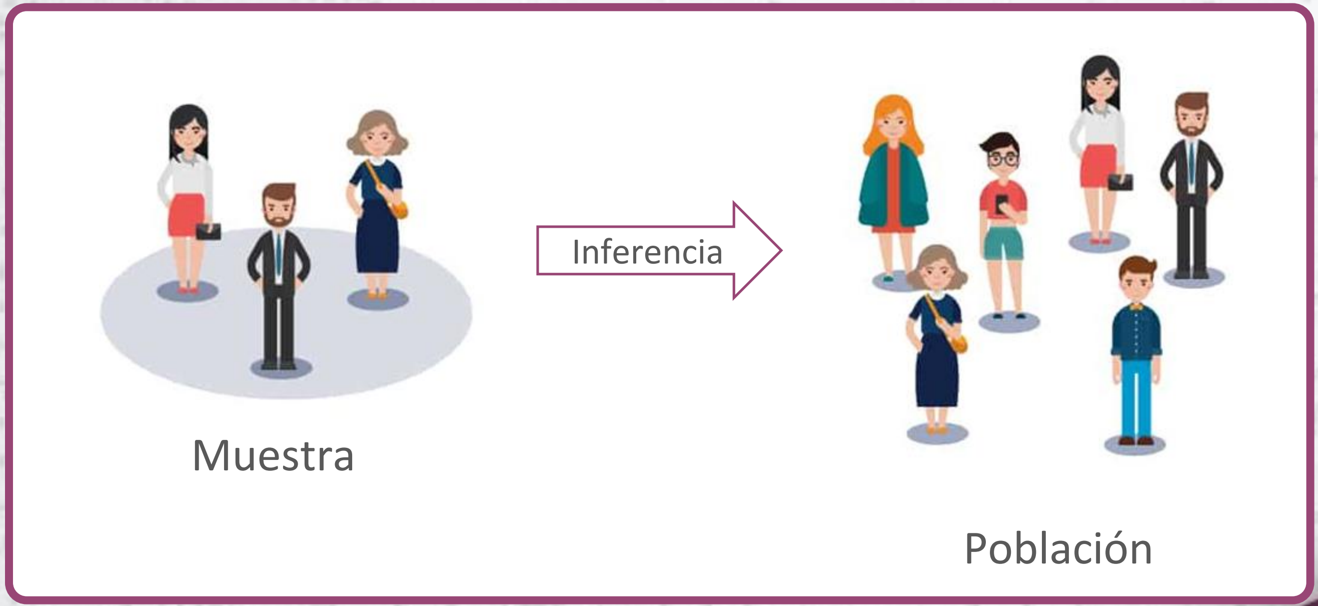
¿Cómo llegar a ese rango?
Lo aprenderemos más adelante.

$$\mu = 175cms \pm 1.42\ cms$$

Población

Definición Intervalo de Confianza

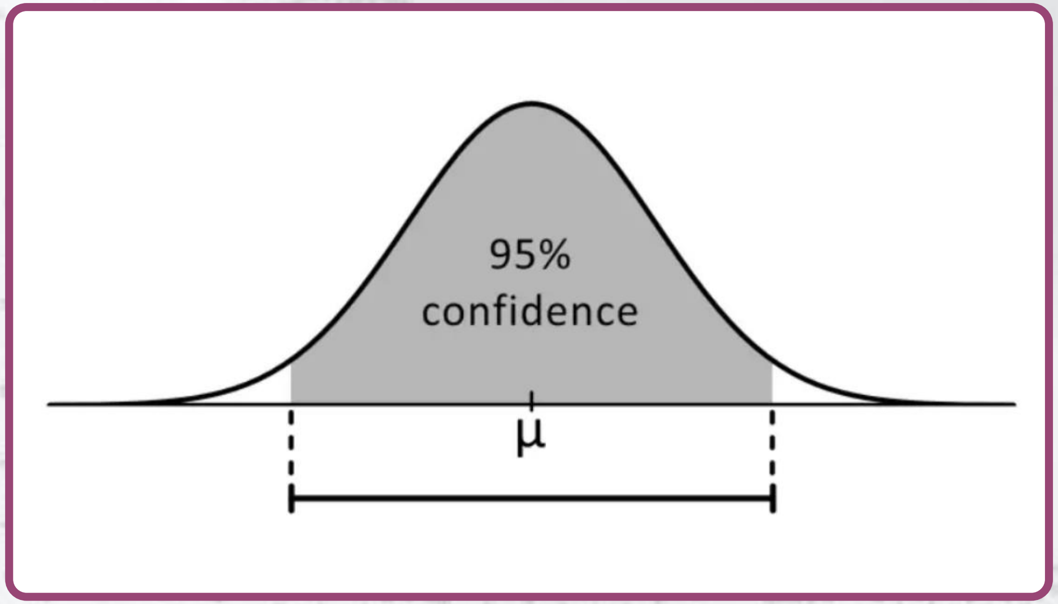
Un intervalo de confianza es un **rango de valores** que se utiliza para **estimar un parámetro desconocido de una población**, como la media o la proporción. Este rango se basa en una **muestra de la población** y se utiliza para proporcionar una **medida de incertidumbre alrededor de la estimación**.



Definición Intervalo de Confianza

En otras palabras, si se tiene una **muestra de una población** y se desea **estimar un parámetro poblacional**, se puede construir un **intervalo de confianza que contenga el valor más probable del parámetro**.

El **nivel de confianza** se refiere a la probabilidad de que el verdadero valor del parámetro poblacional se encuentre dentro de dicho intervalo. Los intervalos de confianza se construyen utilizando técnicas estadísticas y se basan en el tamaño de la muestra y la variabilidad de la muestra.



Algunas definiciones

Estimador Puntual

"Un estimador puntual consiste en un solo valor (punto) deducido de una muestra para estimar el valor de una población" (Lind, Marchal y Whaten, 2008, p.294).

Estimador por intervalo

Un estimador por intervalo es aquel en el que se espera encontrar el parámetro poblacional.

Intervalo de confianza

Intervalo de las estimaciones probables sobre el parámetro.

Límites de los Intervalo de confianza

Son los dos valores extremos del intervalo de confianza.

Nivel de Confianza

Tomamos una muestra de una clase de idiomas y podemos estimar que el promedio general de la clase de idiomas es de 8.35. Al construir un intervalo para esta afirmación, podemos encontrar los dos puntos que van de 7.8 a 8.9.

Ahora, debemos establecer una afirmación que nos garantice que el promedio general del grupo se encuentra justo entre los dos puntos que encontramos. Para ello utilizamos un nivel de confianza que nos permitirá emitir el siguiente enunciado:

“Contamos con el 90% de confianza, o de seguridad, de que el promedio general de la clase de idiomas se encuentra entre los valores de 7.8 y 8.9”

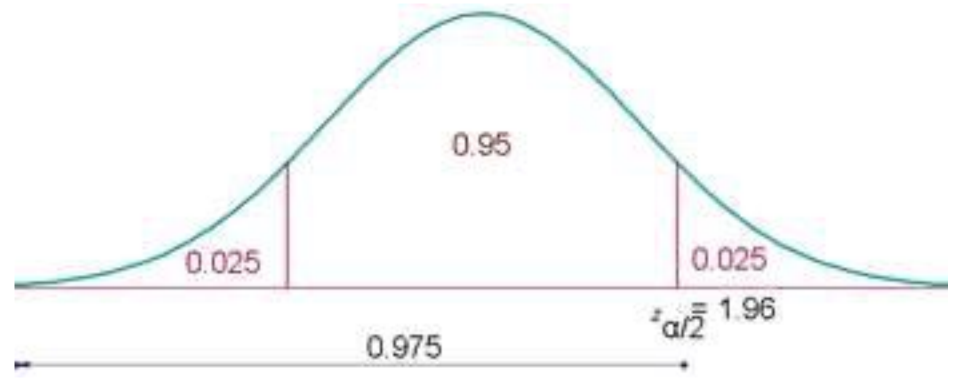
Si quisiéramos tener un nivel de confianza más alto, por ejemplo, de 95%, deberíamos ampliar el rango o intervalo. ¿Cómo hacerlo?

KIBERNUM

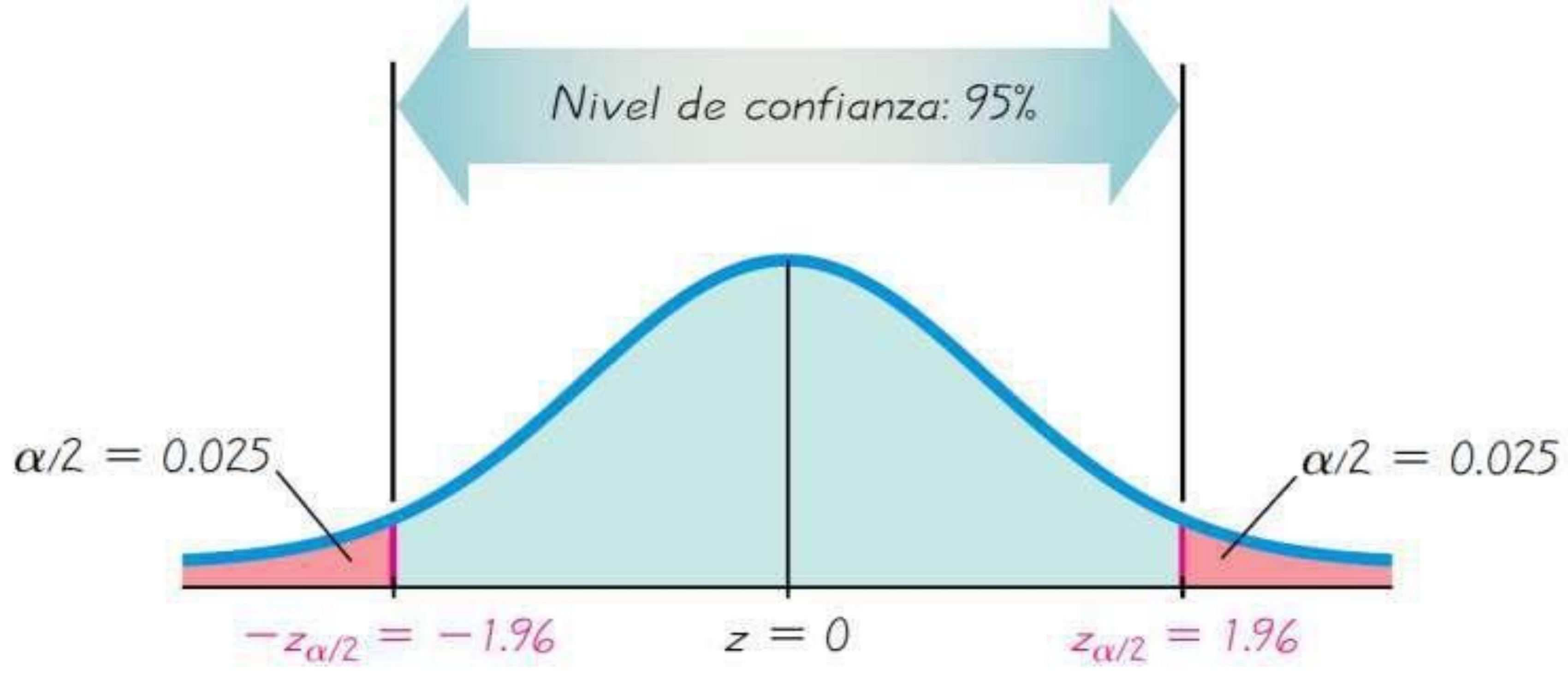
Nivel de Confianza

Ahora debemos asignar los valores al límite inferior y al superior. Sabemos que el área que hay dentro de nuestro intervalo es de 0.95, por lo tanto, el valor de cada cola es de 0.025, al sumar ambas, obtenemos el valor de 0.05.

Nuestro último paso consiste en ubicar dentro de nuestra tabla Z los valores de 0.025 en el caso del límite inferior y 0.975 en el caso del límite inferior. Una vez que encontramos esos valores podemos afirmar que para el nivel de confianza de 95% el valor Z es igual a +1.96 y -1.96 respectivamente.



Nivel de Confianza del 95%



Intervalo de Confianza para la Media Poblacional

Algunas definiciones

Para calcular nuestro intervalo es necesario conocer el valor Z de nuestro nivel de confianza. Una vez que lo descubrimos podemos aplicar la fórmula y de esta manera establecer el valor inferior y superior:

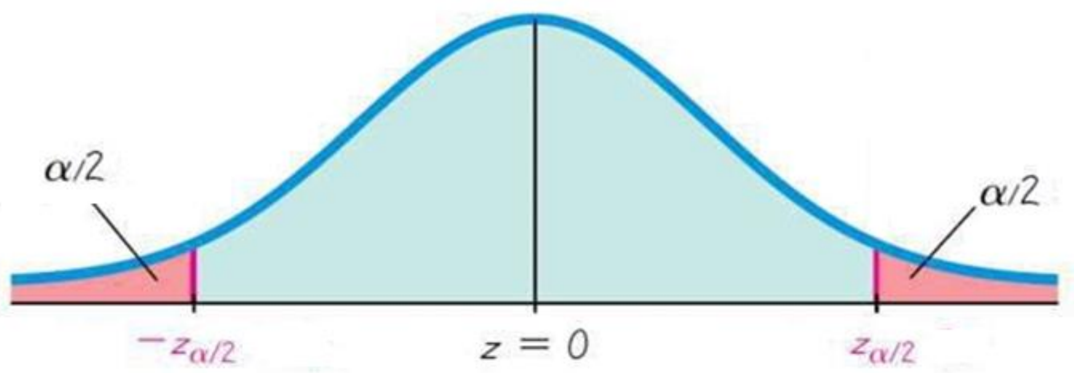
$$\bar{X} \pm z \frac{\sigma}{\sqrt{n}}$$

Un intervalo de confianza nos permite conocer un rango de valores obtenidos a través de una muestra, con el objetivo de conocer un parámetro desconocido, en este caso la media de la población.

IC para la media con varianza conocida (n>=30)

Intervalo de confianza:

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



En donde,

$\boldsymbol{\sigma}$: desviación estándar de la población
$\boldsymbol{n}$: tamaño de la muestra
$\boldsymbol{Z_{\alpha/2}}$: valor crítico de la distribución normal para el nivel de confianza $(1-\alpha)$

IC para la media con varianza conocida (n>=30)

Tamaño de la muestra:

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{e}\right)^2$$

Longitud del Intervalo:

$$L = 2\left(\frac{Z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right)$$

Error estándar:

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

Ejemplo 1

Un fabricante produce focos que tienen un promedio de vida de distribución aproximadamente normal y una desviación estándar de 40 horas. Si una muestra de 30 focos tiene una vida promedio de 780 horas. Encuentre el intervalo de confianza del 95%, 96%, 98% para la media de la población de todos los focos que produce la empresa.

Ejemplo 1

X : variable aleatoria tiempo de vida de focos fabricados.

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

Muestra:
$n = 30$
$\bar{X} = 780$ horas

Población:
$\sigma = 40$
$\mu = ??$

Confianza:
$(1 - \alpha) = 0.95$

$$\alpha = 0.05$$
$$\alpha/2 = 0.025$$

Determinamos valor normal Z

$$Z_{0.025} = -1.96$$

Calculamos intervalo de confianza

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$765.69 < \mu < 794.31$$

IC para la media con varianza desconocida (n < 30)

Cuando la varianza poblacional no es conocida utilizamos la distribución de “t” de “student”, para tamaños de muestra n < 30.

Como σ^2 no se conoce, se estima mediante s^2. La distribución se desvía en forma apreciable cuando los grados de libertad (v = n - 1) son pequeños.

El estadístico t definido resulta de una muestra aleatoria seleccionada de una población normal con varianza σ^2 no conocida.

Intervalo de confianza:

$$\boldsymbol{\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}}$$

En donde,
s : desviación estándar de la muestra
n : tamaño de la muestra
$t_{\alpha/2}$: valor crítico de la distribución t-student para el nivel de confianza $(1-\alpha)$

Ejemplo 2

Una compañía utiliza baterías en sus juegos electrónicos que según ellos duran un promedio de 30 horas, para confirmar esto, se prueba 16 baterías siendo la media muestral de 27.5 horas y su desviación estándar S=5 horas. Encuentre un intervalo de confianza del 95% para la media. Suponga que la distribución de las baterías es aproximadamente normal.

Ejemplo 2

X : variable aleatoria duración de batería
$X \rightarrow N(X, \mu, \sigma^2)$

Muestra:
$n = 16$
$\bar{X} = 27.5$ horas
$S = 5$

Población:
$\mu = 30$ horas
$\sigma = ??$

Confianza:
$(1 - \alpha) = 0.95$
$v = 16 - 1 = 15$

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

Determinamos valor t-Student con v=15:

$$t_{0.025} = 2.131$$

Calculamos intervalo de confianza

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$27.5 - \left(2.131 \cdot {}^{5}/_{4}\right) < \mu < 27.5 + (2.131 \cdot {}^{5}/_{4})$$

$$24.84 < \mu < 30.16$$

Tabla Distribución t-Student

Degrees of freedom (ν)	Amount of area in one tail (α)							
	0.0005	0.001	0.005	0.010	0.025	0.050	0.100	0.200
1	636.6192	318.3088	63.65674	31.82052	12.70620	6.313752	3.077684	1.376382
2	31.59905	22.32712	9.924843	6.964557	4.302653	2.919986	1.885618	1.060660
3	12.92398	10.21453	5.840909	4.540703	3.182446	2.353363	1.637744	0.978472
4	8.610302	7.173182	4.604095	3.746947	2.776445	2.131847	1.533206	0.940965
5	6.868827	5.893430	4.032143	3.364930	2.570582	2.015048	1.475884	0.919544
6	5.958816	5.207626	3.707428	3.142668	2.446912	1.943180	1.439756	0.905703
7	5.407883	4.785290	3.499483	2.997952	2.364624	1.894579	1.414924	0.896030
8	5.041305	4.500791	3.355387	2.896459	2.306004	1.859548	1.396815	0.888890
9	4.780913	4.296806	3.249836	2.821438	2.262157	1.833113	1.383029	0.883404
10	4.586894	4.143700	3.169273	2.763769	2.228139	1.812461	1.372184	0.879058
11	4.436979	4.024701	3.105807	2.718079	2.200985	1.795885	1.363430	0.875530
12	4.317791	3.929633	3.054540	2.680998	2.178813	1.782288	1.356217	0.872609
13	4.220832	3.851982	3.012276	2.650309	2.160369	1.770933	1.350171	0.870152
14	4.140454	3.787390	2.976843	2.624494	2.144787	1.761310	1.345030	0.868055
15	4.072765	3.732834	2.946713	2.602480	2.131450	1.753050	1.340606	0.866245
16	4.014996	3.686155	2.920782	2.583487	2.119905	1.745884	1.336757	0.864667
17	3.965126	3.645767	2.898231	2.566934	2.109816	1.739607	1.333379	0.863279

Ejemplo 3

En una muestra aleatoria de 20 porciones de cereal el contenido Promedio de azúcar fue de 11.3 gramos con desviación estándar de 2.45 gramos. Suponiendo que los contenidos de azúcar están distribuidos normalmente, determine el intervalo de confianza del 95% para el contenido promedio de azúcar en las porciones de dicho cereal.

Ejemplo 3

X : variable aleatoria duración de batería
$X \to N(X, \mu, \sigma^2)$

Muestra:
$n = 20$
$\bar{X} = 11.3$ horas
$S = 2.45$

Población:
$\mu = ??$
$\sigma = ??$

Confianza:
$(1 - \alpha) = 0.95$
$v = 20 - 1 = 19$

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

Determinamos valor t-Student con v=19:

$$t_{0.025} = 2.093$$

Calculamos intervalo de confianza

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$11.3 - \left(2.093 \cdot {}^{2.45}/_{\sqrt{20}}\right) < \mu < 11.3 + \left(2.093 \cdot {}^{2.45}/_{\sqrt{20}}\right)$$

$$10.153 < \mu < 12.446$$

Dudas y consultas

Fin presentación