Matemáticas

3er Semestre

Estadística I

Unidad 2. Estimación

Clave:

05142318/06142318

Universidad Abierta y a Distancia de México

DCEIT

Índice

Estadística I
Unidad 2. Estimación

Presentación de la unidad

Esta segunda unidad entra de lleno en la rama de la estadística inferencial. Uno de los principales objetivos que se persiguen en la estadística inferencial es la posibilidad de hacer generalizaciones y sacar conclusiones basadas en la probabilidad acerca de una población determinada. Todo esto, permitirá la toma de decisiones con base en la información que puede arrogar una muestra dada de esa población y por estas características, es la más usada en los trabajos de investigación.

La estadística inferencial permite inferir a partir de los valores que arrojen las muestras extraídas de una población, como se verá en el desarrollo de esta unidad, el estudio de la estadística inferencial permitirá hacer estimaciones. Estimar no es otra cosa que hacer una buena aproximación de los valores de las características principales de la población de interés. Para fines de este curso se estudiarán las estimaciones, los estimadores y los métodos para construirlos.

Competencia específica

Utilizar diferentes tipos de estimadores para proporcionar un valor aproximado a un parámetro o variable por medio de conjuntos de datos representados en diferentes tamaños de muestras.

Logros

- Utilizar los diferentes tipos de estimación, así como los métodos para generar estimadores, con la finalidad de describir, e interpretar la información obtenida.

2. Estimación

La segunda unidad es bastante extensa. Son muy diversos los temas que abarca. El primer tema, dedicado a las distribuciones de muestreo y teorema central del límite, por si

solo sería una unidad completa. Así las cosas, se hará por lo tanto énfasis en los aspectos más relevantes de cada tema.

La estimación como ya se menciono es uno de los temas de la estadística inferencial. Pero, ¿Qué es la estimación?
En la unidad uno, se vieron conceptos estadísticos tales como la media, la desviación estándar, entre otros. Estos son parámetros de la población, pero muchas veces no es posible trabajar con datos poblacionales, lo que obliga a recurrir a trabajar con muestras representativas de esa población. Pues bien, cuando se calculan parámetros muéstrales tales como la media (promedio) maestral, lo que se está haciendo en si es hacer una estimación del parámetro poblacional. La media muestral es una estimación de la media poblacional. Es solo una aproximación, porque el valor de esta media muestral, varía de muestra a muestra, y se pueden tomar muchas muestras de una población por lo que se tienen diferentes estimaciones de la media poblacional.

De tal forma que la estimación, por tanto, es un proceso en el cual se utilizan parámetros de una muestra para estimar los parámetros de la población. La estimación se realiza utilizando estimadores. La media muestral sería un estimador de la media poblacional.

2.1. Distribuciones de muestreo y teorema central del límite

En la unidad anterior se han abordado temas importantes como la media y la desviación estándar, que nos daban datos sobre la dispersión de la muestra. Sin embargo, también es necesario referirse a la forma de la distribución de muestreo, para ello se requiere un nuevo concepto. Este es el llamado “teorema del límite central”.

El teorema de límite central se abordará en el tema 2.1.1. Pero antes es necesario señalar un tema que resulta útil, esto es las distribuciones de muestreo relacionadas con la normal.

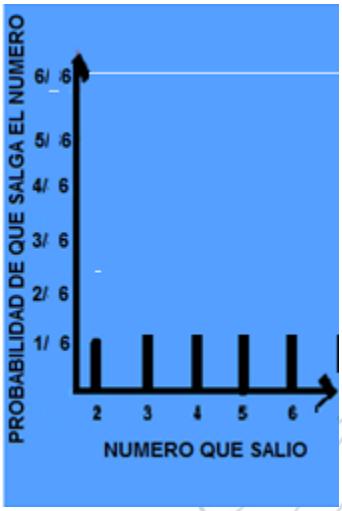
2.1.1. Distribuciones de muestreo relacionadas con la normal

A lo largo de la presente unidad te podrás dar cuenta que los temas tratados en ella están relacionados con la distribución normal. En algunos casos se parte de la hipótesis de que la distribución en cuestión se aproxima a la normal o que bajo ciertas circunstancias se puede considerar como una distribución normal. Así en el punto siguiente 2.1.2. Referido al teorema del límite central, se parte del hecho de que la media maestral se distribuirá de manera cercana a la distribución normal.

Por tanto, resulta indispensable saber lo que es una distribución normal. Para entender mejor este tipo de distribución se aborda primeramente otras distribuciones que son mucho más sencillas y serán de gran ayuda para el entendimiento.

Distribución uniforme.

En estadística para muchos problemas, resulta muy útil graficar el dato contra su probabilidad. Cuando las probabilidades de los datos son las mismas se obtiene una distribución uniforme, un claro ejemplo de esto es el lanzamiento de un dado de seis caras (normal, no truqueado), cuyo espacio muestra es: Ω = {1, 2, 3, 4, 5,6} y la probabilidad para cada uno de los eventos, es idéntica = 1/6. La grafica para este ejemplo es:



En este ejemplo todas las caras del dado tienen la misma probabilidad de salir al ser lanzado. Por lo que la gráfica es uniforme.

Distribución simétrica.

Al igual que en la distribución anterior se trata de graficar el dato contra su probabilidad, en este caso, la principal característica de esta distribución es que se tiene un valor central, las probabilidades de los datos van aumentando hasta llegar a un valor máximo, a partir del cual empiezan a disminuir. Pero este aumento y disminución es simétrico. Cuando se trata de una variable aleatoria continua, una distribución simétrica es en sí una distribución normal.

Ejemplo 1:

Sea el experimento de lanzar dos dados (normales y no truqueados) simultáneamente.

Que llamaremos dado1 

y dado 2, 

Y nos interesa saber la suma de ambos. Es decir, cuantas posibilidades existen que, al lanzar los dos dados, al caer, su suma sea 2, o 3 o 4 o 5 o 6 o 7 o 8 o 9 o 10 o 11 o 12.

Es importante recordar que, al lanzar dos dados, decimos que el número que cae es la suma de los números que aparecen en las caras de ambos dados. Este número es nuestra variable aleatoria “X”. Por ejemplo, si en uno de los dados hay un 1 y en el otro un 3, decimos que cayó un cuatro. Cuando se lanzan dos dados no todos los números tienen la misma probabilidad de caer.

Nos vamos a encontrar entonces con que: para que la suma de ambos sea dos, solo existe una posibilidad y esta es que los dados caigan uno. Ya que, si alguno de los dos cae con un valor superior a 1, sería imposible lograr que la suma de ambos fuese igual a dos. En otras palabras, el 2 sólo puede salir de una manera: que en ambos dados salga un 1 ósea (1+1); pero hay dos maneras de que salga un 3: que en el dado 1 salga 1 y en el dado 2 salga un 2 o que en el dado 1 salga un 2 y en el dado 2 salga un 1,

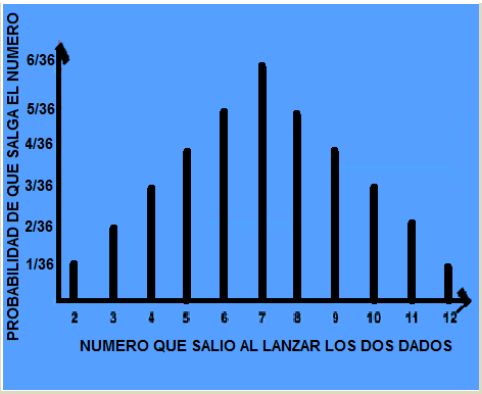
ósea (1 + 2 y 2 + 1). El total de posibilidades para este experimento es de 36. Es decir la cardinalidad de este espacio muestral seria: n (□)=36

Estos resultados los vamos a reportar en una tabla que llamaremos distribución de probabilidad.

Variable aleatoria X (suma de las dos caras del dado)	Probabilidad: p(x)
2	$\frac{1}{36}$
3	$\frac{2}{36}$
4	$\frac{3}{36}$
5	$\frac{4}{36}$
6	$\frac{5}{36}$
7	$\frac{6}{36}$
8	$\frac{5}{36}$
9	$\frac{4}{36}$
10	$\frac{3}{36}$
11	$\frac{2}{36}$
12	$\frac{1}{36}$

Como se puede apreciar claramente en la tabla, las probabilidades se distribuyen simétricamente en torno a un valor central que para este ejemplo es el “7”.

Al graficar los datos de la tabla se obtienen lo siguiente:

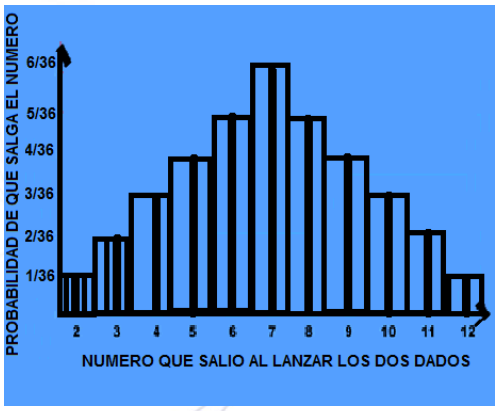


En este ejemplo el valor central es el “7”, a partir de él se forman dos mitades idénticas, por tanto, la figura es simétrica.

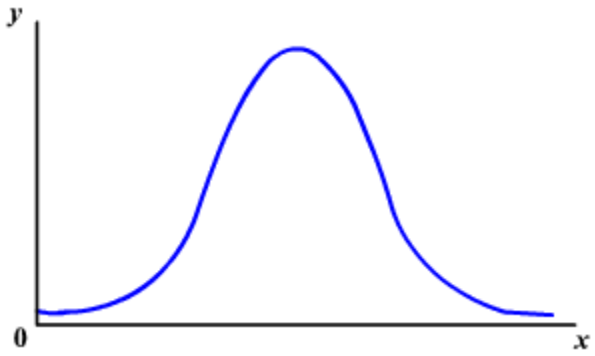
Distribución normal.

Como ya se mencionó, la distribución normal es una distribución de probabilidad simétrica, para variables continuas.

Si se realiza un histograma para muestras suficientemente grandes se asemejan al siguiente gráfico:



Entre más grande sea el tamaño de las muestras, el grafico cada vez más y más se va asemejar a una curva simétrica como la siguiente:



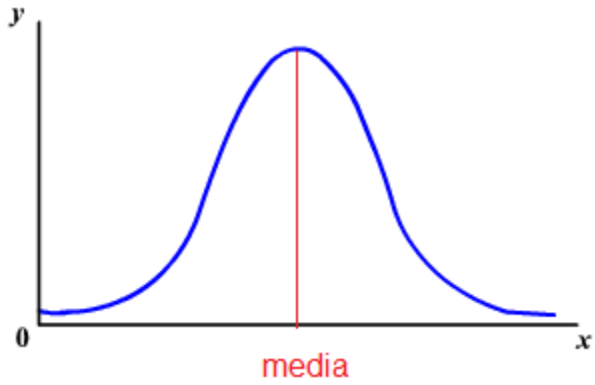
Esta curva es un modelo adecuado para representar la regularidad estadística.

Si se tiene una variable aleatoria continua "x", y se determina que su regularidad estadística queda muy bien representada por esta curva, entonces la distribución de la variable "x", es normal. O dicho en otras palabras la variable "x", se distribuye normalmente.

Por su forma también se le conoce como campana de Gauss.

En torno a la curva es importante señalar lo siguiente:

La curva es simétrica. Es decir que tiene un punto medio que la divide en dos partes iguales.

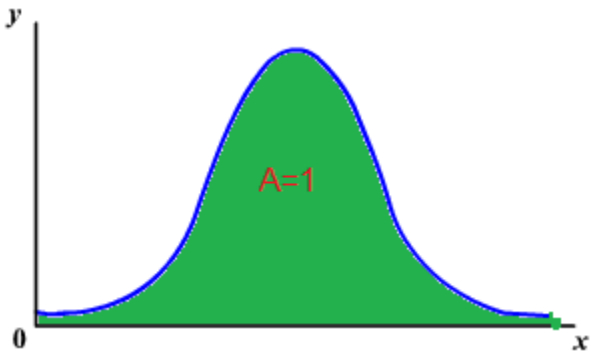


En esta distribución se tiene la peculiaridad de que tanto la media, mediana y moda,

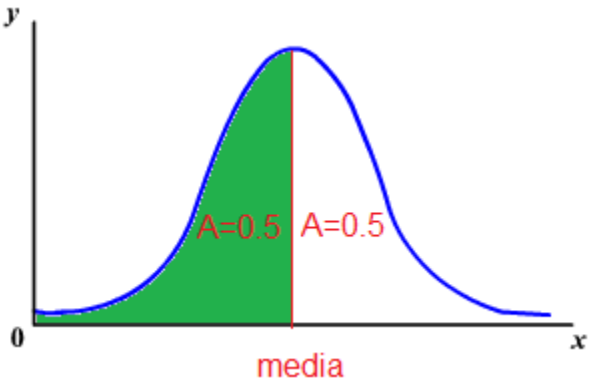
tienen el mismo valor ubicado en el centro de la campana.

El área bajo la curva acotada en dos puntos, representa la probabilidad de que un evento cualquiera tenga un valor entre esos dos puntos.

El área total bajo la curva es igual a la unidad. Recuérdese la definición de integral (como el área bajo la curva) en general en cualquier distribución de probabilidad al calcular el área bajo la curva el valor obtenido será uno.



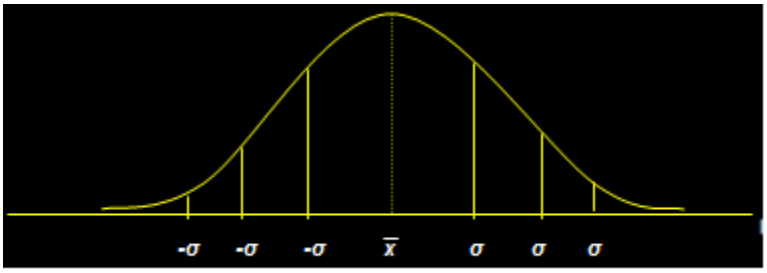
Por tanto, cada mitad de la curva tendrá un valor igual a 0.5. En otras palabras, el área bajo la curva de cada mitad igual a 0.5.



Es cóncava hacia abajo en el centro y cóncava hacia arriba en las colas.

El área bajo la curva se puede subdividir utilizando la media como referente y cualquier otro punto, generalmente se utilizan las desviaciones estándar. Así se dice que un punto

está en función del número de desviaciones estándar que diste de la media.



La expresión matemática que representa la curva normal es la siguiente:

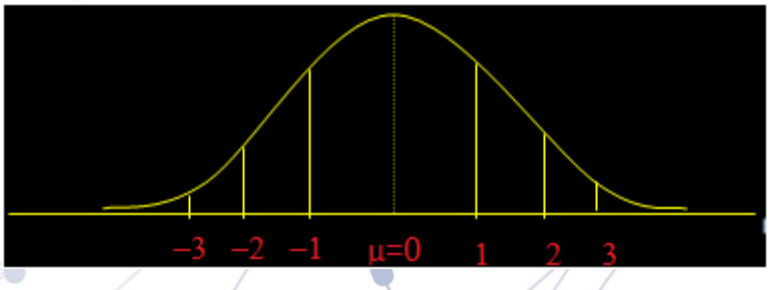
Ecuación de la normal

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

Pero no se trabaja con ella dado que los resultados se encuentran reportados en tablas, así que lo importante es aprender a utilizar adecuadamente las tablas.

Curva normal estándar.

Estas tablas están referidas a una curva normal estandarizada $N\ (\mu, \sigma)$. Una distribución normal estándar, cero, uno, se indica mediante la siguiente notación: N (0,1). En donde el cero corresponde al valor de la media $\mu = 0$ y el uno al valor de la desviación típica σ = 1.



Los valores: 1, 2, 3 hacen referencia a las desviaciones estándar, es decir, si está a una

desviación estándar de la media, a dos, a tres, etc., en la tabla se van a reportar valores de z, que corresponden a los valores antes mencionados, así que, se tendrán valores de $z = 1,2,3$ (y todos los valores intermedios). El valor reportado en la tabla representa el área bajo la curva y esta es la probabilidad de que ocurra el evento. De tal forma que si se busca el valor para $z = 1$, se tiene que es 0.3413, esta es el área bajo la curva que va desde la media hasta $z = 1$ y también indica que la probabilidad de que ocurra un evento entre la media y $z = 1$ es del 34.13% (0.3413).

Como la curva es simétrica los valores serán los mismos si se trata de áreas bajo la curva que van de la media hasta $z = -1$.

La tabla es la siguiente ejemplifica la obtención del dato:

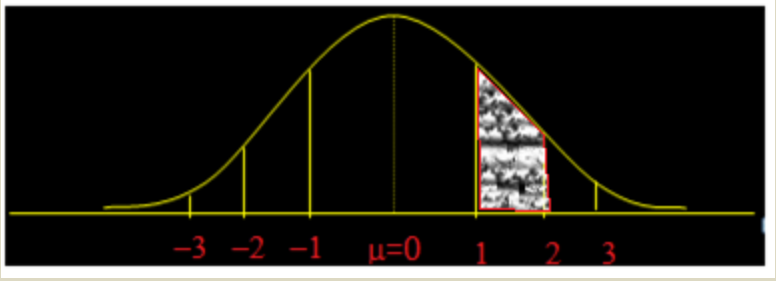
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177

1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Ahora, pasemos a los ejemplos.

Ejemplo 2:

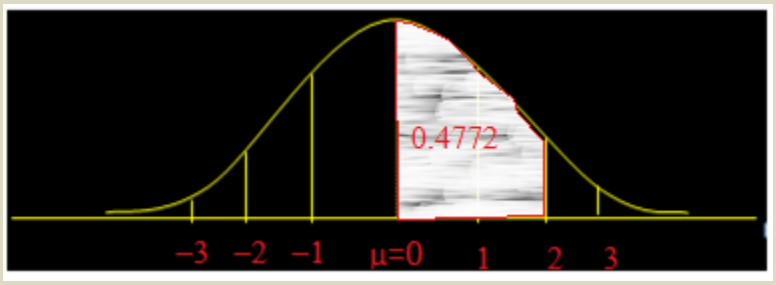
Supóngase que se desea conocer el área bajo la curva acotada entre los valores de $z = 1$ y $z = 2$



Solución:

Se recurre a la tabla "z". Se busca el valor del área bajo al curva descendiendo por la columna de "z". Como ya se vio anteriormente, el valor reportado para $z = 1$ en la tabla es de 0.3413 (34.13%). y para $z = 2$ se tiene un valor de 0.4772. estos valores representan que:

De la media a $z = 2$ el área = 0.4772



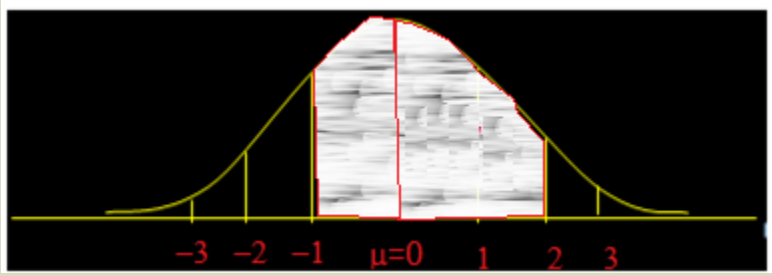
De la media a $z = 1$ el área = 0.3413

Como el área deseada solo abarca de $z = 1$ a $z = 2$, la manera de obtenerla es restándole al área comprendida entre la media y $z = 2$, el área comprendida entre la media y $z = 1$.

Es decir: 0.4772- 0.3413 = 0.1359

Ejemplo 3:

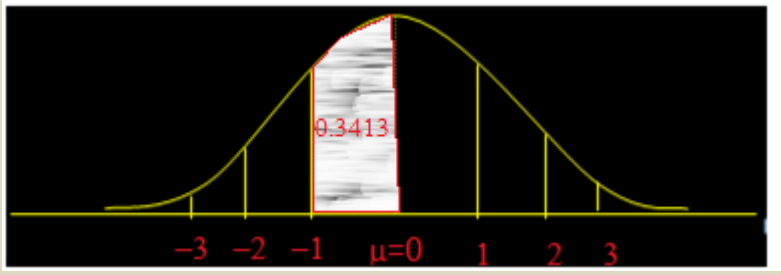
Supóngase que se desea conocer el área bajo la curva acotada entre los valores de $z = -1$ y $z = 2$ como se ilustra en al siguiente figura.



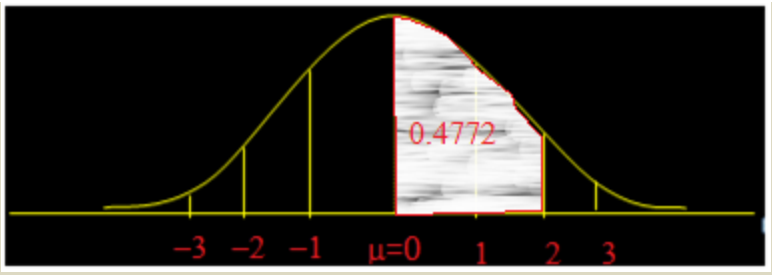
Solución:

Por ser simétrica, la curva del área de la media hasta $z = 1$ tiene el mismo valor que de la media a $z = -1$. Como los valores ya se habían obtenido en el ejemplo anterior se tiene:

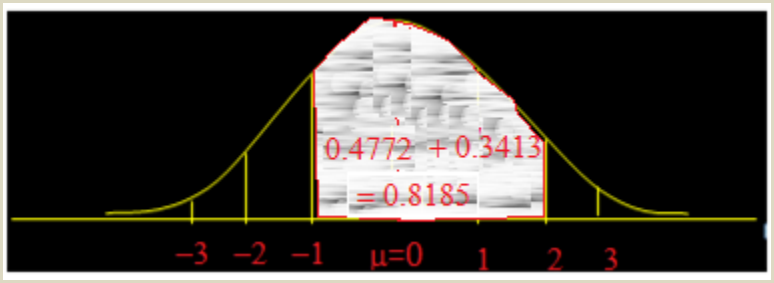
Área de la media a $z1 = 0.3413$



Y área de la media a $z2 = 0.4772$



En este caso, el área buscada se encuentra al sumar las dos áreas encontradas.
De tal forma que se tiene:



Resultado:

El área comprendida entre $z = -1\ y\ z = 2$ es de 0.8185

Puntuaciones tipificadas

Un tema sumamente importante dentro de las distribuciones normales, es el de las puntuaciones "z". Estas servirán para calcular áreas debajo de la curva normal cuando se tengan valores diferentes a $N\ (0,1)$.

Para ello se utilizan las siguientes relaciones matemáticas:

$$z = \frac{xi - \mu}{\sigma}$$

$$z = \frac{xi - \overline{x}}{s}$$

A estos valores se les llamara valores normalizados y a la curva obtenida, curva normal estandarizada.

Ejemplo 4:

$N\ (\mu, \sigma)$ y el área bajo la curva es $P(x)$ $N\ (0,1)$ y el área bajo la curva es $P(z)$

Una puntuación tipificada se puede definir como la expresión de una puntuación origen a la nueva puntuación z. las puntuación tipificada consiste en indicar la cantidad de desviaciones con respecto a la media a la cual se encuentran las puntuaciones originales.

Veamos el siguiente ejemplo:

Ejemplo 5:

Se desea realizar un estudio estadístico sobre el ingreso mensual (en miles de pesos) en una población juvenil. Si se toma una muestra de tamaño n=1000 personas. Esta se distribuye normalmente. La $\mu = 2$ y la $\sigma = 0.5$.

Se desea saber cuál es la probabilidad de que una persona tenga un ingreso mensual superior a 3 (tres mil pesos)

Solución:

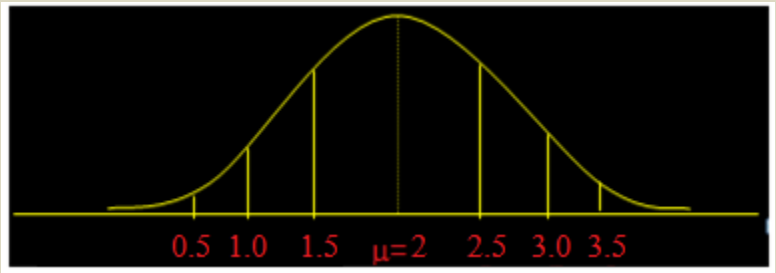
La solución a este problema requiere las formulas vistas anteriormente:

$$z = \frac{xi - \mu}{\sigma}$$

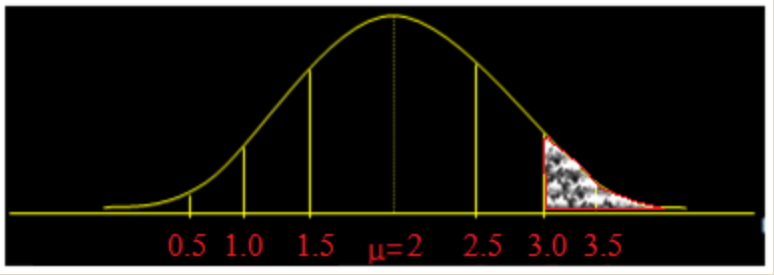
$$z = \frac{xi - \bar{x}}{s}$$

Tenemos la distribución original $N\ (2{,}0.5)$ y se desea la distribución normal estandarizada $N\ (0{,}1)$.

La representación gráfica para la distribución original $N\ (2{,}0.5)$ quedaría de la siguiente forma:



El problema señala que se desea saber la probabilidad de obtener ingresos superiores a los 3 (miles de pesos). Por lo tanto, el área de interés es:



El área en blanco representa la probabilidad buscada, P (3<x)

Ahora se realizará la estandarización con el uso de la ecuación:

$$z = \frac{xi - \mu}{\sigma}$$

xi son los valores: 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5

Sustituyendo los valores en la fórmula:

$$z = \frac{0.5 - 2}{0.5} = \frac{-1.5}{0.5} = -3.0$$

$$z = \frac{1.0 - 2}{0.5} = \frac{-1.0}{0.5} = -2.0$$

$$z = \frac{1.5 - 2}{0.5} = \frac{-0.5}{0.5} = -1.0$$

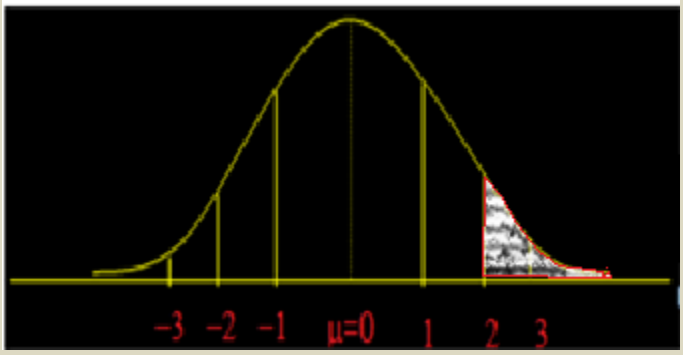
$$z = \frac{2 - 2}{0.5} = \frac{0}{0.5} = 0.0$$

$$z = \frac{2.5 - 2}{0.5} = \frac{0.5}{0.5} = 1.0$$

$$z = \frac{3.0 - 2}{0.5} = \frac{1.0}{0.5} = 2.0$$

$$z = \frac{3.5 - 2}{0.5} = \frac{1.5}{0.5} = 3.0$$

Graficando se obtiene que el área deseada es:



El área en blanco representa la probabilidad buscada, $P\ (3 < z)$

Por lo tanto, hay que buscar en la tabla el valor de $z = 2$ y encontrar el área correspondiente para que por diferencia se encuentre el área deseada.
Para $z = 2$ el área es 0.4772. (Es el área que va de la media hasta $z = 2$)

Como el área total bajo la curva es uno. Y el área total bajo la mitad de ella es 0.5 y como se desea el área posterior a $z = 2$
Se realiza la resta:

0.5-0.4772 = 0.0228, que corresponde al área marcada en la gráfica.
Por lo tanto, la probabilidad buscada es del 2.28%

Resultado:

2.28%

Dado que en el subtema 2.1.3 se verá la aproximación de la normal a la binomial, es necesario abordar la distribución binomial como tal.

Distribución binomial.

La distribución binomial es una distribución para variables discretas y al igual que la distribución normal, tiene una gran variedad de casos en los que es apropiada, quizá es la distribución para variables discretas que tiene más amplia utilización. La distribución binomial se utiliza sobre todo en problemas que involucren variables dicotómicas. Las

variables dicotómicas son variables nominales que tienen dos categorías: falso-verdadero, arriba-abajo, etc.

Existen muchas situaciones donde es propia una distribución binomial. El más claro ejemplo es el experimento de lanzar una moneda al aire, cuyo resultado arroja águila – sol. Tanto en la naturaleza como en la vida cotidiana se encuentran ejemplos de carácter dual: masculino-femenino, día-noche. Entre muchos otros.

La distribución binomial está caracterizada por la probabilidad de éxito y como complemento, la probabilidad de fracaso. La variable de interés "x" es el éxito deseado. Lo contrario es el fracaso.

Ejemplo 6:

Sea el experimento de lanzar dos monedas (normales y no truqueadas) simultáneamente.

Que llamaremos moneda 1  y moneda 2, 

Resolver:

¿Cuál será la probabilidad de que ambas monedas caigan "sol"?

Para este ejemplo se define la variable aleatoria discreta:

$$X = \text{número de soles. Que es el número de éxitos del experimento}$$
$$Y = \text{número de águilas. Que seria el número de fracasos del experimento}$$

En otro experimento el interés podría ser ¿cuántas águilas caen?

Por lo tanto, el éxito sería el número de águilas y no el número de soles como en esta ocasión.

Generalmente la probabilidad de éxito se simboliza con la letra *"p"* y la del fracaso con *"q"*. Como la probabilidad máxima es 1 y siendo complementos entonces se tiene que $q = 1 - p$

La expresión matemática para la distribución binomial es la siguiente:

$$P(x) = \binom{n}{x} p^x q^{n-x} \quad \text{o} \quad P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

Dónde:

P(x)= probabilidad de que sucedan "x" éxitos en "n" intentos.

$\binom{n}{x}$= las combinaciones de n en x.

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$

$x = \text{número de éxitos}$
$n = \text{numero de repeticiones del experimento}$
$p = \text{probabilidad de exito}$
$q = 1 - p = \text{probabilidad de fracaso}$

Ejemplo 7:

Se realiza un experimento que consiste en evaluar un alimento para gatos. Se encontró que 4 de cada 10 gatos prefiere wiskas (el alimento para gatos). Se toma una muestra representativa de tamaño 5 (es decir cinco gatos). ¿Qué probabilidad existe de que dos de esos gatos de la muestra prefieran wiskas?

Solución:

Se usará la siguiente formula:

$$\text{P(x)} = \binom{n}{x} p^x q^{n-x} \text{ o } \quad \text{P(x)} = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

Con la información del problema se tienen los valores de las siguientes variables:

$p = \frac{4}{10} = 0.4$

$q = \frac{6}{10} = 0.6$

$x = 2$

$n = 5$

$n - x = 5 - 2 = 3$

Sustituyendo en la fórmula:

$$\text{P}(2) = \frac{5!}{2!(5-2)!}(0.4)^2(0.6)^{5-2}$$

Resolviendo por partes:

$$5! = 120$$

$$2! = 2$$

$$3! = 6$$

$$(0.4)2 = 0.16$$

$$(0.6)3 = 0.216$$

Sustituyendo se tiene:

$$P(2) = \frac{120}{2 * 6}(0.16)(0.216)$$

P (2)= 0.3456

Que se puede aproximar a:

P (2)= 0.346

¡Bien! ahora se revisarán algunos parámetros para las distribuciones binomiales. Los más sobresalientes son: la media, la varianza, y la desviación estándar.

Dónde:

Media= $\mu = n * p$

Varianza= $\sigma 2 = n * p * q$

Desviación estándar= $\sigma = \sqrt{n * p * q}$

Estas relaciones serán de mucha utilidad cuando se vea el subtema 2.1.3. Referido a la aproximación de la normal a la binomial.

2.1.2. El teorema del límite central

El teorema establece que: cuando se tiene una población cualesquiera y se toman aleatoriamente muestras del mismo tamaño, con una desviación estándar σ y con una media finita μ. La media muestral se distribuirá de manera cercana en forma a la distribución normal siempre y cuando el número de muestras tomado sea lo suficientemente grande, esto es **n** debe ser mayor a 30.

Este teorema es algo parecido a lo visto en la ley de los grandes números, solo que en los grandes números se hablaba de realizar muchas veces el experimento aleatorio y aquí se trata de tomar muchas muestras; entre más muestras se tomen, mucho más cercana será la forma a la de una distribución normal.

Resumiendo, se puede decir que:

La distribución de las medias seguirá una distribución **normal** con la misma media que la distribución original cuando se tomen muestras lo suficientemente grandes.

Es decir, la media muestral $\bar{X}$, se distribuirá en forma semejante a la normal con media μ y desviación estándar:

Desviación estándar muestral:

$$s = \sigma/\sqrt{n}$$

Ahora bien, Si $\bar{X}$ es la media de una muestra aleatoria de tamaño "n" tomada de una población con media μ y varianza finita $\sigma 2$, entonces la forma límite de la distribución queda

$$z = \frac{\bar{X} - \mu}{\sigma}$$

En tanto que, si se va a trabajar con medias de muestras, el valor utilizado para la desviación estándar de las medias muestrales es:

$$s = \sigma/\sqrt{n}$$

Para este caso:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

o

$$z = \frac{\bar{X} - \mu}{s}$$

Ejemplo 8:

Supóngase que se tiene una población a la cual se le ha calculado su media y su desviación estándar, cuyos valores respectivamente son: 100 y 5. Se van a tomar un número suficientemente grande de muestras de tamaño 10 y se calculan sus medias.

La cuestión sería saber el valor de la media de todas las medias muéstrales.

Solución:

Se tienen los siguientes datos:
$\mu = 100$ y $\sigma = 5$,

Para este ejemplo vamos a utilizar el principio del teorema del límite central que establece que la media muestral $\bar{X}$ será la misma que la media de la población μ.

De tal forma que si μ = 100, $\bar{X} = 100$.

Ejemplo 9:

Hagamos otro ejercicio.

Supóngase que se tiene una población estudiantil que desea ingresar a la universidad, para lo cual presentan un examen de admisión. Las calificaciones obtenidas por los aspirantes se distribuyen con una μ =70 y σ= 10. De esta población se toma una muestra de 25 calificaciones y ahora se desea calcular su media muestral. Nos interesa saber la probabilidad de que "x" sea mayor de 80.

Solución:

Utilicemos el teorema del límite central. Así se tiene que:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

y al igual que en el ejemplo anterior vamos a utilizar el principio del teorema del límite central que establece que la media muestral $\bar{X}$ será la misma que la media de la población μ.

De tal forma que si $\mu = \mu_x$

Así que se tiene que buscar:

$$P\ (x > 80)$$

y para ello se buscará:

$$P\ (z =)$$

Se tienen los datos:

$$\bar{X} = 80$$

$$\mu = \mu_x = 70$$

$$\mu_x = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$$

Sustituyendo estos valores en la fórmula:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

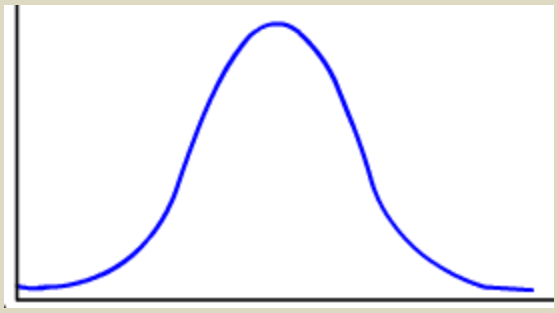
Nos queda:

$$z = \frac{80-70}{2} = \frac{10}{2} = 5$$

Así se tiene que:

P (x>80) = P (z>5)

Recordemos la curva normal:



Como ya se había indicado, se debe buscar en tablas, el área bajo la curva normal.

Ahora bien, es necesario recordar que la curva es simétrica y que el área total bajo de ella es = 1. Así que el área bajo cada mitad de la curva es = 0.5.

En seguida se muestra la tabla que se habrá de usar para resolver este ejemplo:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753

0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952

2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Para este ejemplo $z = 2.0$ implica que busquemos en la columna de la "z" y descendamos hasta localizar el valor de 2. Una vez ahí, se recorre el renglón hacia la derecha en busca de las cifras decimales. En este ejemplo los decimales son = 0. Por lo tanto, el valor buscado de $z = 0.4772$.

Como el área bajo la curva de la mitad de la derecha es = 0.5 y solo nos interesa saber el área que es superior a $z = 2$ se tiene que:

$$P(Z > 2) = 0.5 - 0.4772 = 0.0228$$

Por lo tanto, lo que nos está diciendo el resultado es que la probabilidad de que la muestra tomada promedie una calificación superior a 80 es muy baja, apenas del 2%.

Ejemplo 10

Se obtiene una muestra de alumnos que estudian en la ESAD y cuyo promedio de calificación es de 8.4 con una desviación estándar de 0.9, de una población cuyo promedio es de 8.6, determinar:

a) El porcentaje de alumnos con promedio superior a 8.6
b) Si la población actual de la ESAD es de 11,500 alumnos, encontrar el número de alumnos con este promedio

Solución

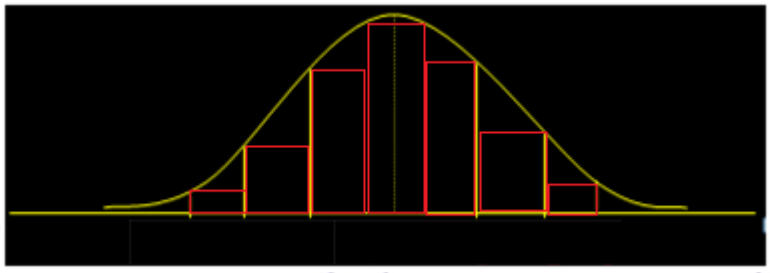
a) Si $\bar{X} = 8.4$, $\mu = 8.6$ y $\sigma = 0.9$ sustituimos en la formula y obtenemos $z = \frac{8.4-8.6}{0.9}$=-0.22 encontramos el valor en tablas y obtenemos 0.0871 de 0 a Z y restamos el valor 0.5 (media campana de Gauss) y obtenemos 0.4129 por cien, nos da el 41.29%

b) Multiplicamos 0.4129 por 11500 y obtenemos 4748.35 que serían los 4748 alumnos con promedio superior a 8.6 en la ESAD

2.1.3. La aproximación de la normal a la distribución binomial

Ya se abordaron las distribuciones, binomial y la normal. Ahora bien, existe la posibilidad de que una distribución binomial se pueda simplificar por medio de la distribución normal, para ello es necesario que se cumpla con varias condiciones.

Ya en el tema anterior donde se vio el teorema central del límite se estableció que conforme va en aumento el tamaño de las muestras las distribuciones tienden hacia una distribución normal.



Cuando la distribución es binomial y se realizan histogramas graficando $\{x, P(x)\}$, y se tiene que: $P(x) = \binom{n}{x} p^x q^{n-x}$ o $P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$

Se obtienen histogramas para la distribución binomial que (al igual que para los anteriores) tienden hacia una distribución normal, a medida que aumentan los valores de "n" (número de ensayos) y "p" ($que\ es\ la\ probabilidad\ de\ éxito$) se aproxima a un valor de ½ = o.5.

En otras palabras la distribución binomial se va a aproximar a la distribución normal, cuando el número de repeticiones del experimento, ósea "n" sea muy grande, o tienda al infinito según se vio en el teorema central del límite. Además se debe cumplir que la tendencia de "p" ósea la probabilidad de éxito tienda a ½.

Ahora bien, cuando se tiene un problema con distribución binomial, si este se puede resolver más fácilmente mediante la aproximación hacia una distribución normal. Se debe tomar en cuenta que la distribución binomial es del tipo discreto y la distribución normal es del tipo continuo. Entonces se tiene que hacer un ajuste llamado ajuste por continuidad. Los valores de este factor de ajuste son ± 0.5.

¿Cuándo se ha de usar un factor de $+0.5$ y cuando uno de -0.5?

Caso 1: se utilizará el valor de +0.5 cuando la probabilidad buscada tenga las expresiones:

$$" \leq " o " > " . P(x \leq\ ), o\ P(x >\ )$$

Caso 2: se utilizara el valor de +0.5 cuando la probabilidad buscada tenga las expresiones: " $\geq$ " o " $<$ "

Y para decidir si el problema es conveniente resolverlo por medio de la distribución normal, se debe verificar que:

$$La\ cantidad\ de\ objetos\ (personas, datos\ etc.) * p \geq 5, y$$
$$La\ cantidad\ de\ objetos\ (personas, datos\ etc.) * q \geq 5.$$

Para encontrar el valor de "z" que se escribirá en las tablas se tendrá que resolver la siguiente ecuación:

$$z = \frac{\mathrm{x} - \mu \pm 0.5}{\sigma}$$

Dónde:

Media= $\mu = n * p$

Varianza= $\sigma 2 = n * p * q$

Desviación estándar= $\sigma = \sqrt{n * p * q}$

Quizá esto se entienda un poco mejor mediante un ejemplo.

Ejemplo 11:

En una escuela se toma una muestra aleatoria de tamaño $n = 100$ alumnos. La escuela tiene una población mixta compuesta de hombres y mujeres. La población de mujeres es del 5%. Se desea saber cuál sería la probabilidad de encontrar 3 o más mujeres en la muestra tomada.

Solución:

Como la población de mujeres es del 5% (0.05), la población de hombres es del 95% (0.95). Dado que la probabilidad total es uno (100%), la probabilidad de encontrar a 3 o más mujeres = 1, la probabilidad de encontrar menos es de 3 mujeres.

Dicho en otras palabras, la probabilidad total es:

$$100\% = P\,(encontrar\ 3\ o\ más\ mujeres) + P\,(encontrar\ menos\ de\ 3\ mujeres).$$

De donde:

$P\,(encontrar\ menos\ de\ 3\ mujeres) = P\,(0) + P\,(1) + P\,(2)$. Significa la probabilidad de

encontrar cero mujeres, más la probabilidad de encontrar una mujer más la probabilidad de encontrar dos mujeres.

$$P\ (encontrar\ 3\ o\ mas\ mujeres)\ =\ P\ (3) + P\ (4) + \cdots + P\ (100)$$

¡Bien! ahora se checa que sea conveniente utilizar la distribución normal.

$$La\ cantidad\ de\ objetos\ (personas, datos\ etc.)\ *\ p\ \geq\ 5,$$

y

$$La\ cantidad\ de\ objetos\ (personas, datos\ etc.)\ *\ q\ \geq\ 5.$$

Sustituyendo datos:

$$100 * 0.05 = 5, entonces\ cumple\ con\ ser\ \geq 5\ y$$

$$100 * 0.95 = 95\ cumple\ con\ ser\ \geq 5.$$

Cumple con ambas condiciones, por lo tanto, si es conveniente resolver mediante la distribución normal.

Ahora, se necesitan calcular los valores de la media y la desviación estándar.

Media= $\mu\ = n * p$

Sustituyendo:

$$\mu\ = n * p = 100 * 0.05 = 5$$

Desviación estándar= $\sigma\ =\ \sqrt{n * p * q}$

Sustituyendo:

$$\sigma = \sqrt{100 * 0.05 * 0.95} = \sqrt{4.75} = 2.18$$

Estas son las ecuaciones que se utilizaran para encontrar los valores que se sustituyan en la ecuación que nos permita calcular el valor de "z".

$$z = \frac{\mathrm{x} - \mu \pm 0.5}{\sigma}$$

Se desea encontrar: $P\ (encontrar\ 3\ o\ más\ mujeres) = \ P\ (x \geq 3)$, eso significa que se utilizará un factor de ajuste de -0.5.

Por lo tanto sustituyendo estos valores en la ecuación de z, se tiene:

$$z = \frac{3 - 5 - 0.5}{2.18}$$

$$z = \frac{-2.5}{2.18} = -1.15$$

Buscando en la tabla de z, el área encontrada es: 0.3749

Este valor corresponde al área que va de 3 hasta la μ=5, por lo tanto, hay que sumar el resto del área que correspondería a los valores superiores a 5.

Se tiene, por lo tanto:

$$0.3749 + 0.5 = 0.8749$$

Que corresponde a una probabilidad del 87.49%

$$P\ (x \geq 3) = 87.49\%$$

2.2. Métodos de construcción de estimadores

Existen dos tipos de estimación: la estimación puntual y la estimación por intervalos que se verán más adelante en el tema 2.4 de esta unidad. Respecto a la estimación puntual o por punto, existen algunos métodos que se utilizan para realizar dicha estimación: el método de estimación por momentos y el método de estimación máximo verosímil. Que se verán en seguida.

2.2.1. Momentos

Como ya se ha mencionado existen dos métodos para la construcción de estimadores, el de momento y el de máxima verosimilitud, aunque hay algunos autores que mencionan un tercero (por mínimos cuadrados). De estos, el más antiguo, es el de momentos y también es el más sencillo. El método de los momentos está basado en un principio estadístico que sugiere que se pueden obtener buenas estimaciones de los momentos poblacionales a partir de sus respectivos momentos muéstrales. Consiste entonces en establecer una igualdad entre los momentos (poblacionales y muéstrales).

2.2.2. Máxima verisimilitud

El método de máxima verosimilitud es un poco más complejo comparado con el método anterior (método por momentos). Este método se basa en la selección del valor del parámetro que hace máxima la función de verosimilitud. Esto es la función de probabilidad conjunta o la función de densidad conjunta. En otras palabras, de lo que se trata es de considerar como valor del parámetro poblacional aquel valor que hace máxima la probabilidad. De hecho, es en parte algo intuitivo, ya que si se tiene una serie de sucesos ¿Cuál va a aparecer?, la lógica y la estadística dirán que aquel cuya probabilidad se a mayor. Por ejemplo, al lanzar dos dados de seis caras (normales no truqueados) la intuición dirá que el número que aparecerá será el 7, porque es el suceso más probable.

2.3. Propiedades de los estimadores

Como ya se ha mencionado, existen parámetros muéstrales (variables aleatorias, como la media muestral, entre otras), que son utilizadas para aproximarse a los valores de los parámetros poblacionales, a estas variables se les conoce como estimadores. A los valores que toman estos estimadores se les llama estimaciones de los parámetros poblacionales. En la unidad uno, se revisaron los conceptos de media muestral representada por "$\bar{x}$" y la media poblacional, simbolizada por "μ". Así se tiene que "$\bar{x}$" es un estimador de "μ".

Estos estimadores tienen ciertas propiedades, entre las que se destacan: el sesgo, la eficiencia, la consistencia y la suficiencia. A continuación, se pasará a revisar dichas propiedades.

2.3.1. Sesgo

Existen varias maneras de clasificar a los estimadores, una de ellas está basada en el sesgo o su parte contraria el insesgo de un estimador. Veamos que significa que un estimador este sesgado o insesgado.

A grosso modo se puede decir que un estimador es insesgado cuando coincide con los parámetros poblacionales, trátese de la media o cualquier otro. Por el contrario, un estimador se considera sesgado o con sesgo, cuando no coincide con los parámetros poblacionales, ya sea que se trate de la media, la desviación estándar o cualquier otro.

2.3.2 Eficiencia relativa

El termino eficiencia es de cierta manera de mucho uso cotidiano, pongamos por ejemplo a un trabajador de una empresa. ¿Qué significa que este trabajador sea eficiente? Una posible respuesta es, que el trabajador, cumpla con las tareas que se le han asignado en los tiempos estipulados. Es decir, el trabajador debe cumplir con ciertas características y de manera análoga, un estimador, será eficiente si cumple con ciertas características

estadísticas.

Las características estadísticas a evaluar en un estimador, para saber si es eficiente, son ante todo la varianza y el sesgo. Un estimador será más eficiente que otro si cumple en ser insesgado y con una varianza pequeña. Por el contrario, un estimador será de menor eficiencia si presenta un alto valor de varianza y además es sesgado.

En estadística se dice que un estimador es eficiente cuando existe precisión del estadístico muestral como estimador del parámetro poblacional. Como es una cuestión de precisión eso implica que un estimador será más eficiente tanto más próximo se encuentre del valor del parámetro poblacional.

2.3.3. Consistencia

Para revisar el concepto de consistencia es necesario recordar los conceptos de muestra y población. Si se toman muestras cada vez más grandes, es obvio que los estimadores muéstrales se acerquen cada vez más a los parámetros poblacionales. En otras palabras, el estimador se va igualando al parámetro poblacional. Como para cada muestra se tiene un valor diferente del estimador, se puede tener una sucesión de estos valores del estimador. La consistencia radica en que esta sucesión sea convergente en probabilidad hacia el parámetro poblacional. En otras palabras, que coincida con el parámetro poblacional.

2.3.4. Suficiencia. Teorema de Rao Blackwell

El concepto de suficiencia es de uso cotidiano en las calificaciones, al menos en ciertos sistemas educativos. El alumno obtiene la calificación de seis, "s" o suficiente. ¿Que significa? que, con los conocimientos adquiridos y demostrados en las pruebas y tareas, se considera suficiente para que el chico apruebe su curso.

En estadística un estimador muestral siempre estará referido a un parámetro poblacional. La suficiencia tiene que ver con que el estimador muestral, reúna la mayor cantidad de

características o información relevante, para poder ser utilizado como estimador del parámetro poblacional.

2.4. Estadística y estimadores

En la unidad uno de este curso de estadística se vieron los diferentes tipos de medidas: medidas de posición, de asociación y de tendencia central. En todas ellas se trabajó con varios parámetros estadísticos, cabe recordar algunos: la media, la desviación estándar entre otros. Ahora bien, como se señaló, estos parámetros pueden referirse a la población o aun muestra representativa (aleatoria) de dicha población.

μ representa a la media poblacional, pero por diferentes motivos no siempre se puede trabajar con la poblacional. Así se recurre al parámetro muestral $\bar{x}$, que es una buena aproximación al valor de μ.

Cuando se utiliza el parámetro $\bar{x}$, se está haciendo una estimación del parámetro μ. Por tal motivo algunos autores señalan que: $\bar{x}$ estima a μ.

En tanto que el parámetro poblacional "σ" (desviación estandar) se estima mediante el parámetro muestral "s".

Los estimadores son los agentes para realizar una estimación y como ya se ha indicado, en estadística se tienen dos tipos de estimaciones

En seguida se verá en que consiste cada una de ellas.

2.4.1. Estimación puntual

En estadística se utilizan dos tipos de estimación. Estimación puntual y estimación por intervalos. Una estimación puntual es aquella en donde con un solo valor numérico basta para estimar el parámetro. Es decir que con un solo valor que se obtenga será posible hacer una conclusión acerca de la población que es objeto de estudio.

Un claro ejemplo de este tipo de estimación la encontramos en las medidas de tendencia central. En la unidad uno de esta materia, se vio el tema y se señaló la importancia que tienen este tipo de medidas. Entre los conceptos vistos se revisó el concepto de la media poblacional "μ", y el de la media muestral "$\bar{x}$". En esta unidad ahora se expone que la media poblacional "μ", se puede estimar mediante un estimador, en este caso la media muestral "$\bar{x}$".

Algunos ejemplos de estimación puntual: El promedio de vida para un humano, el promedio de solicitado par ingreso a universidad.

La estimación puntual presenta una clara desventaja en comparación con la estimación por intervalos que se verá en seguida. Al ser un solo dato no es posible expresar de alguna manera el grado de incertidumbre de una estimación. Por ello la estimación por intervalos cobra una mayor importancia.

2.4.2. Estimación por intervalos

La estimación por intervalos como su nombre lo indica, se trata de un rango de valores. En este intervalo de valores, con cierto nivel de probabilidad, se encuentra el parámetro poblacional. Como ya se vio en la unidad uno de esta materia, el concepto de rango consta de dos valores, un valor máximo y un valor mínimo. En estimación, se les conoce como límite superior y límite inferior respectivamente. Así se tiene que dentro de estos valores (máximo-mínimo) se puede concluir que se encuentra el parámetro de interés.

Ya en el párrafo anterior se vio lo que es una estimación por intervalos, se vio la definición y se trató de entender su diferencia con la estimación puntual.

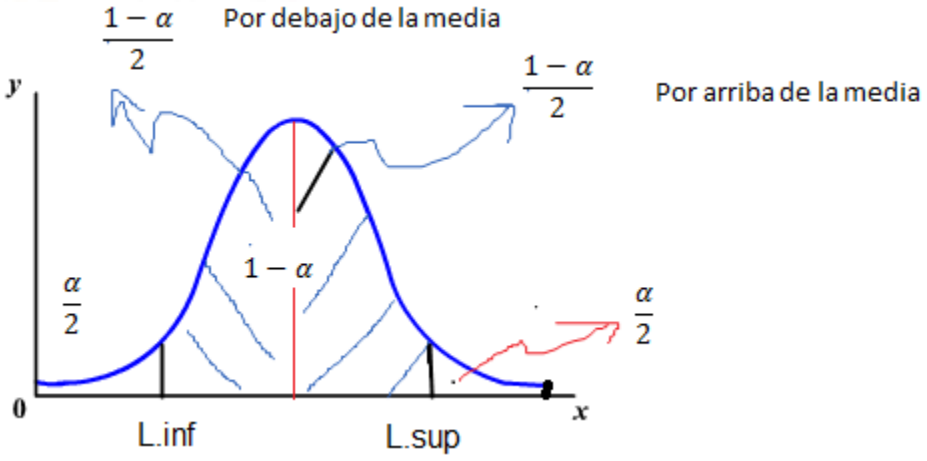
2.5. Estimación por intervalo

En este tema ahora se desarrollará la estimación por intervalos dada la importancia estadística que tiene. Así se abordarán diferentes subtemas entre los que destacan el estudio de los intervalos de confianza. El intervalo de confianza tiene mucho uso en los estudios estadísticos como se podrá constatar más adelante.

2.5.1. Intervalo aleatorio

Como ya se vio se tienen básicamente dos tipos de estimación, estas son la puntual y la estimación por intervalos. Un intervalo como se vio en el subtema anterior se puede asemejar al concepto de rango (rango=valor máximo – valor mínimo). Estos valores máximo y mínimo, en intervalos se les llaman límite superior y límite inferior. Ahora bien, se le llama intervalo aleatorio, al intervalo que se construye y al menos uno de sus límites es una variable aleatoria.

Veámoslo en el esquema siguiente:



Dónde:

α = significancia o nivel de significancia o nivel de significación

$1 - \alpha$ = nivel de confianza. Es el área bajo la curva comprendida entre los límites superior e inferior.

L.inf = Límite inferior de confianza
L.sup = Límite superior de confianza

Estos límites de confianza entran justamente en nuestro siguiente tema.

2.5.2. Intervalo de confianza

Se han revisado los dos tipos de estimación, se vio que una estimación por intervalos consta de dos valores, un valor máximo y un valor mínimo y que en estimación se les conoce como límite superior y límite inferior respectivamente. Cuando se trata de un intervalo de confianza ese valor máximo se le llama límite superior de confianza y al valor mínimo se le llama límite inferior de confianza.

Los intervalos de confianza están asociados a cierto nivel de confianza, por ejemplo: un 90%, o expresado en forma decimal un 0.90 de confianza.

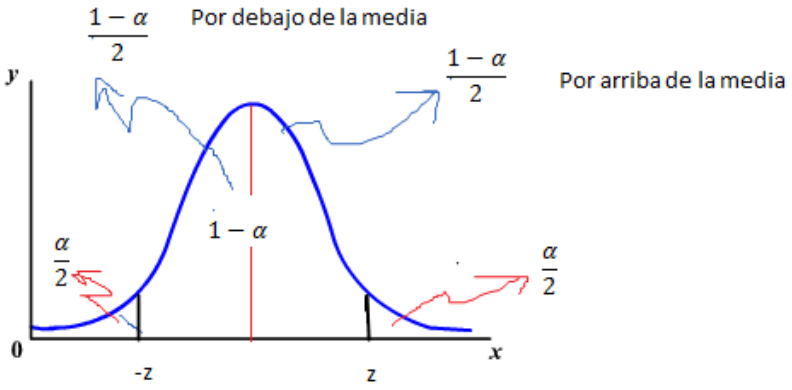
El intervalo de confianza por tanto expresa un cierto nivel de probabilidad, esto implica que existe un riesgo de error conocido. Por otra parte, el intervalo de confianza proporciona valores enfocados en el valor estadístico muestral, pero el parámetro poblacional también se ubica dentro del intervalo.

El intervalo de confianza es importante porque permite responder a ciertas cuestiones como conocer qué tan grande es el error de muestreo, esto es, la toma de la muestra respecto a la población. Por ejemplo, se tiene una media muestral "$\bar{x}$" y es importante saber cuánto se aproxima a la media poblacional "μ" y cuál es su error de muestreo.

En esta unidad se abordará el intervalo de confianza para la media poblacional siendo conocida su desviación típica. Es importante recordar que en la distribución de muestreo de las medias los promedios que se obtengan de diferentes muestras se distribuyen de

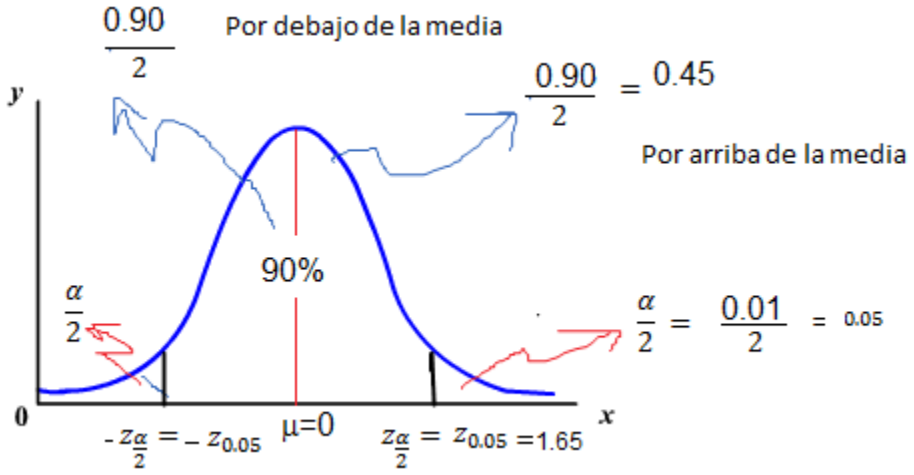
forma normal.

Para construir los intervalos de confianza es necesario saber qué tipo de distribución siguen los datos que se están trabajando. Si se tiene que son del tipo de una distribución normal, entonces se procede de acuerdo al siguiente esquema, en donde se tienen varios componentes.



Como ya se había visto en el subtema 2.1.1 la curva es simétrica y el área total bajo de ella es igual a 1. Representa el 100% de probabilidad. Donde, 1- α es el nivel de confianza. Está acotado por los valores de $-z\ y\ z$. Las zonas antes de $-z$ y después de z. son las colas de la distribución y representan el riesgo de error. Sumar las dos colas da como resultado "α" ósea: $\frac{\propto}{2}+\frac{\propto}{2}=\propto$. Así que entre más pequeño sea α, el área demarcada por $1-\alpha$ será máxima, y esto implica una probabilidad máxima para la muestra seleccionada. Es decir 1- α es el área bajo la curva acotada por $(-z, z)$ y es la probabilidad que se tiene de encontrarse ahí el parámetro poblacional.

Por ejemplo si se tiene que 1- α es del 90% (0.9), esto implica que α es del 10% (0.01). Es decir se tiene el 10% de probabilidad de que los valores medios de la muestra estén a -1.65 desviaciones estándar (el 5% de la cola izquierda) y a $+1.65$ (el otro 5% que corresponde a la cola derecha) desviaciones estándar, de la media. Esto se puede comprender mucho mejor con la siguiente imagen:



La expresión matemática para el cálculo del intervalo de confianza es:

$$IC = \bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right) \qquad IC = \bar{x} \pm z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

O bien

$$\bar{x} \pm z\left(\frac{s}{\sqrt{n}}\right)$$

Dónde:

$\bar{X}$ = media o promedio muestral

z = este valor es el reportado en las tablas de distribución normal. La obtención del valor de z depende del intervalo de confianza deseado. Para el caso de los intervalos se utiliza z para la mitad de la significancia y se simboliza como: $z_{\frac{\alpha}{2}}$

$$z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = \text{Margen de error E}$$

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

σ =Desviación estándar

n= tamaño de la muestra

Generalmente se trabajan los problemas con un nivel de confianza dado, por ejemplo, del: 90%, 95% y 99%. Que equivalen respectivamente a: 0.9, 0.95, 0.99.

Para buscar el valor de z, es necesario usar el nivel de confianza entre 2, dado que se tiene una curva normal simétrica que tiene dos mitades. El nivel de confianza entre 2, arroja un área bajo la curva, este es el dato con el que se ha de buscar el valor de z en las tablas. Por ejemplo, si se trabaja con un nivel de confianza del 99% o su equivalente 0.99, no es este valor el que hay que buscar en tablas. Se busca la mitad, es decir $0.99/2 = 0.495$, ósea

$$\frac{1-\alpha}{2} = 0.495$$

Ejemplo 12:

Se desea hacer un estudio con la población estudiantil de una escuela para lo cual se toma una muestra aleatoria. Obteniéndose la siguiente información:

El tamaño de la muestra n= 49 con un promedio aritmético de 55 y cuya desviación estándar fue de 10.

Se desea calcular el intervalo de confianza del 99% para la media poblacional μ.

Solución:

Para resolver el problema se recurrirá a la expresión matemática señalada anteriormente:

$$\bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$$

De los datos del problema el único que no se tiene es el valor de z.
Este se obtiene de las tablas. Para ello se trabaja con el nivel de confianza del 99% que equivale a tener 0.99 de probabilidad, lo que nos da un valor para "α" del 1% o bien del 0.01.

Así:
$\alpha = 0.01$ y $1 - \alpha = 0.99$ y por tanto, $\frac{\alpha}{2}$=0.005

Como se tiene que la distribución es de tipo normal, y esta es simétrica, entonces para cada mitad se tiene:

$\frac{\alpha}{2}$=0.005

$$1 - \alpha = 0.99$$
$$\frac{1-\alpha}{2} = 0.495$$

Este es el valor que se utilizara para encontrar el valor de z.

En seguida se muestra la tabla que se habrá de usar para encontrar el valor de z de este ejemplo:

Es la misma tabla utilizada en el cálculo del teorema central del límite, solo que ahora se comienza buscando el área bajo la mitad de la curva, que para este ejemplo es del 0.495. Una vez localizado este valor se obtiene el valor de z correspondiente. Como se puede observar en la tabla se marca el área =0.495, siguiendo la fila se ve que corresponde a un valor de $z = 2.5$ pero como el valor de 0.495, está en la columna que corresponde a 0.08 decimales, el valor total de z=2.58

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753

0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952

2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Para este ejemplo $z = 2.58$ corresponde al valor de

$$z_{0.05}$$

Ahora se sustituyen los datos en la fórmula:

$$\text{IC}= \bar{x} \pm z\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\text{IC} = 55 \pm 2.58\left(\frac{10}{\sqrt{49}}\right) = 55 \pm 2.58\left(\frac{10}{7}\right)$$

$$\text{IC} = 55 \pm 3.68$$

Otra forma de reportarlo es:

$$51.31 < IC < 58.68$$

Algunos autores prefieren la notación:

$$IC = (51.31, \quad 58.68)$$

No hay mayor problema en utilizar una u otra notación.

Del resultado obtenido se desprende por lo tanto que existe la probabilidad del 99% de que el valor de μ este en el intervalo comprendido entre: $51.31\ y\ 58.68$.

En general se puede decir que:

Cuando se trabaja con muestras de tamaño grande, disminuye el error de muestreo y por lo mismo el intervalo será más preciso. También disminuirá el error típico y el intervalo será menos ancho.

A mayor nivel de confianza el ancho del intervalo será mayor.

Cabe mencionar que también se pueden construir intervalos de confianza usando otras distribuciones diferentes a la distribución normal. Pero ese tema será parte de otro curso.

2.5.3. Método pivotal de intervalos de confianza

Ya que se está abordando el intervalo de confianza, resulta pertinente decir que este se puede construir mediante el uso del método del pivote (método pivotal). Este método está basado en la determinación de una expresión pivote. Esta expresión pivote es función de las mediciones de la muestra y del parámetro poblacional desconocido. El parámetro poblacional desconocido es el único valor desconocido. En tanto el pivote presenta una distribución que no depende del parámetro poblacional.

Supóngase que se tiene un estadístico "r" que presenta una distribución normal. Se puede construir un intervalo para "r" $P(z1 \leq r \leq z2) = 1 - \alpha$.

2.5.4. Intervalos con muestras grandes

Este tema de los intervalos con muestras grandes ya se ha venido utilizando en los temas anteriores, sobre todo en la construcción de intervalos de confianza. Es importante, sin

embargo, dejar en claro que en la estadística inferencial se tienen básicamente dos categorías para los tamaños de muestras, las categorías tamaño grande y tamaño pequeño.

Una muestra de tamaño grande se considera a partir de $n > 30$.
Una muestra de tamaño pequeña por lo tanto será aquella que este por debajo de este valor es decir $n \leq 30$.es importante hacer notar que esto es solo una convención, y que por lo tanto hay autores que manejan el valor de 25 elementos para considerar a una muestra grande.

En general para la construcción de intervalos de confianza se tiene que:

Cuando se tiene el primer caso es decir que el tamaño de las muestras es grande, se trabaja con la distribución normal. (Esto ya se hizo en el subtema anterior)
Cuando se trata de muestras pequeñas se utiliza otro tipo de distribución, sobre todo la distribución t de Student (siempre y cuando la población se distribuya normalmente).

Willian Gosset desarrollo la distribución "t" para muestras pequeñas.

Al trabajar con la distribución normal se observa que sobre todo para muestras pequeñas, la aproximación de la desviación estándar "s" de la muestra (o muestras) a la desviación estándar de la población, es deficiente. Por ello la distribución "t" de Student es una distribución que tenderá a una mayor dispersión en comparación con la distribución normal.

2.5.5. Método general de intervalos de confianza

Recién se vio el método del pivote, el cual consistía en localizar una cantidad pivotal. Este método a veces no es adecuado para la determinación de los intervalos de confianza, por tal motivo se es necesario otro método, al cual se le conoce como método general. Este método alternativo al pivotal va a generar los mismos resultados, para las condiciones adecuadas. La diferencia está en que este método no requiere de un pivote. De tal

manera que en este método la función muestral que se utilice, podrá tener una distribución de probabilidad que dependa del parámetro poblacional.

Cierre de la unidad

Las medidas estadísticas aprendidas en esta unidad son de suma importancia y representan la base para prácticamente cualquier estudio estadístico. Con estos conceptos bien aprendidos será mucho más fácil abordar la siguiente unidad, por lo que se recomienda al estudiante, revisar los temas de la unidad cuantas veces sea necesario hasta tener cierto dominio sobre ellos, dado que como ya se comentó son los pilares para cualquier estudio posterior.

Fuentes de consulta

Básica

- Huntsberger, D. (1983). *Elementos de Estadística inferencial*. España: Continental.
- Kuby, J. (2012). *Estadística elemental.* México: Cengage.
- Ojer, L. (1990*). Estadística básica.* Madrid: Dossat.