Dra. Teresa Peralta

LA DISTRIBUCIÓN NORMAL Y LOS PUNTAJES Z

La distribución normal es muy útil porque nos permite entender algunas de sus características y cómo éstas se reflejan en distintas variables. Por ejemplo, de la distribución normal podemos decir:

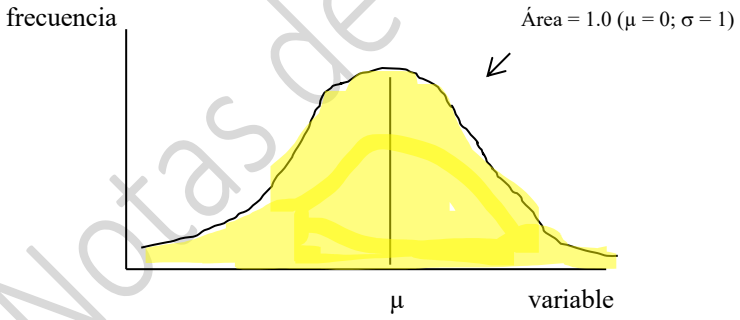
1. La media se ubica en la mitad y la curva es simétrica alrededor de la media. La media y mediana son las mismas.
2. La mayoría de las observaciones están cercanas a la media, de tal suerte que la frecuencia es más alta alrededor de la media. Hay muy pocas observaciones que son mucho más grandes o mucho más pequeñas que la media.
3. La curva nunca toca el eje X en ninguno de sus lados, sólo se acerca a éste.
4. Una curva normal se define por su media y su desviación estándar y se construye a partir de estos dos datos.
5. Todas las curvas normales tienen la misma forma, pero pueden ser altas y delgadas o bajitas y ensanchadas. La diferencia es que la segunda tiene una desviación estándar más grande.

Ahora bien, debido a que su forma básica es fija, la proporción de área debajo de la curva puede ser calculada en varios puntos. Por ejemplo, si la variable que estamos observando fuese la altura de la población y si supiéramos la media y la desviación estándar correspondientes, podríamos saber que proporción de gente estaría por arriba de los 2 metros de altura o entre 1.60 mts. y 1.80 mts.

Áreas por debajo de la curva normal

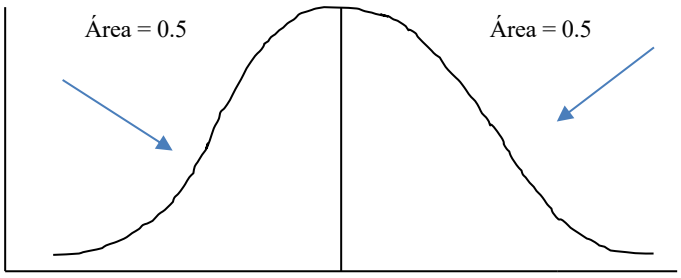
Al área total por debajo de la curva se le denomina 1.0, como se muestra en la gráfica 1, así que cualquier proporción de la curva será un número entre 0 y 1.

Gráfica 1.



La mitad del área se ubica por arriba de la media y la otra mitad por debajo de la media, así que podemos decir que la proporción del área más grande que la media es de 0.5 y la proporción por debajo de la media es también de 0.5 (gráfica 2). También podemos decir que la probabilidad de que una observación sea más grande que la media es de 0.5, porque la mitad de las observaciones se ubican por arriba de la media.

Gráfica 2

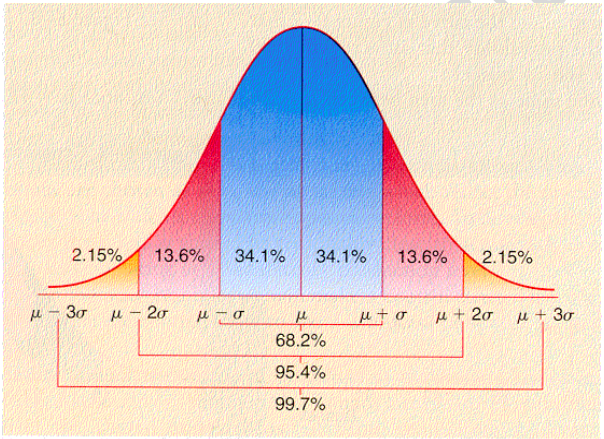


Otras áreas por debajo de la curva pueden describirse en términos de las desviaciones estándar con respecto de la media. Estas áreas son las mismas independientemente de si la curva es alta y delgada o baja y ancha. Por ejemplo, en la gráfica 3:

-0.682 del área se ubica entre $\mu - \sigma$ y $\mu + \sigma$. Esto significa que 68.2% de las observaciones se ubican dentro de una desviación estándar en cualquier lado de la media.

-0.954 del área se ubica entre $\mu - 2\sigma$ y $\mu + 2\sigma$. Esto significa que 95.4% de las observaciones se ubica dentro de dos desviaciones estándar en cualquier lado de la media.

Gráfica 3.



Preguntas:

a. ¿Qué proporción del total del área se ubica entre la media y una desviación estándar por encima de la media?
b. ¿Qué proporción del total del área se ubica entre la media y 1 desviación estándar a la izquierda de la media?
c. ¿Qué proporción del total del área se ubica hacia la izquierda de 1 desviación estándar por debajo de la media?

d. ¿Qué proporción del total del área se ubica fuera de la región entre la media – 2DE y la media + 2DE?

Ejemplo

Suponga que se afirma que los salarios para un tiempo completo de recién egresados universitarios tendrían una distribución normal con una media de $15,000 y una desviación estándar de $1,500. ¿Qué podemos saber acerca de los salarios de recién egresados si observamos la distribución normal?

Sabemos que el 95.4% de las observaciones, en una distribución normal, se ubican entre 2DE en ambos lados de la media. Podemos calcular los salarios de 2DE a ambos lados de la media:

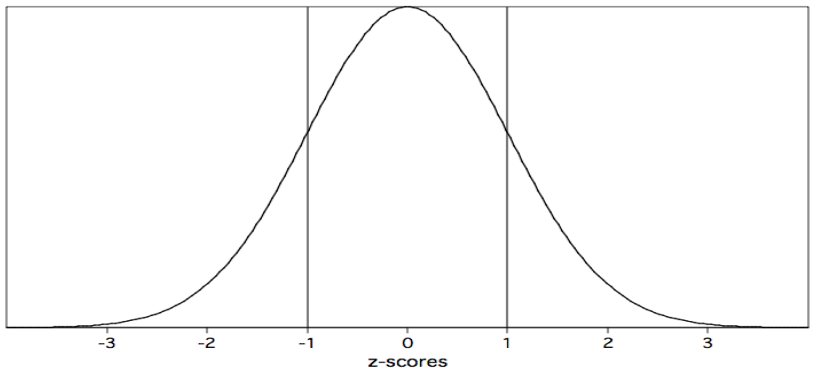
$2\sigma = 2 \times 1500 = 3000$
$\mu + 2\sigma = 15\,000 + 3000 = 18\,000$
$\mu - 2\sigma = 15\,000 - 3000 = 12\,000$

Ahora, grafique la información:
Gráfica 4.

PUNTAJES Z Y LA DISTRIBUCIÓN NORMAL

Sabemos que cuando estandarizamos una variable al convertir los valores en puntajes Z, la media de los puntajes Z es de 0 y la desviación estándar es de 1. La distribución de una variable estandarizada es conocida como "**una distribución normal estándar**" (gráfica 5).

Gráfica 5.



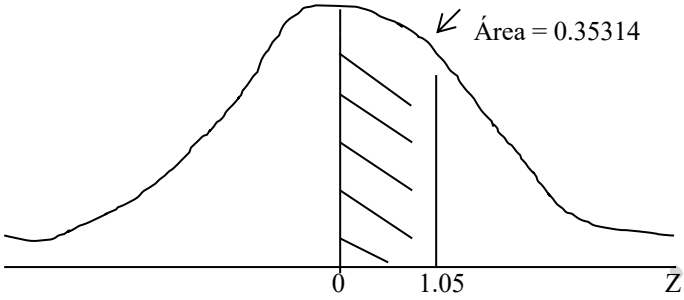
Un puntaje Z de +1 indica que estamos a 1DE por arriba de la media, así que el área entre la media y Z = +1 es de 0.341 o, en otras palabras, el área entre μ y μ + σ es de 0.341.
Ahora bien, como los puntajes Z no siempre contienen valores redondeados (como +1), para encontrar el área entre la media y cualquier puntaje Z debemos **consultar una tabla de distribución normal**. Una tabla de distribución normal proporciona la proporción del área de la curva entre la media y un cierto valor de Z. Esto es lo mismo que la proporción entre la media y el mismo valor negativo de Z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.4440
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.4990

Áreas

Si se observa la tabla, la primera columna proporciona el puntaje Z hasta un decimal. La primera fila proporciona el segundo decimal del puntaje Z que uno quiere ubicar. Por ejemplo, si quisiéramos buscar el puntaje Z de 1.05 la respuesta sería 0.35314. Esto significa que la proporción del área que se ubica entre la media y Z es igual a 1.05 (1.05 desviaciones estándar por arriba de la media. Gráficamente se vería así:

Gráfica 6.



Preguntas:

a. ¿Cuál es la proporción entre la media y Z = 2.00? _____
b. ¿Cuál es la proporción entre la media y Z = 0.59? _____
c. ¿Cuál es la proporción entre la media y Z = 1.66? _____
d. ¿Cuál es la proporción entre la media y Z = -0.03? _____

Áreas por fuera de los puntajes Z

a. ¿Cuál es el área a la derecho de Z = 1.00? _____
b. ¿Cuál es el área a la derecha de Z = 2.33? _____
c. ¿Cuál es el área a la izquierda de Z = -0.86? _____
d. ¿Cuál es el área a la izquierda de Z = -1.17? _____

Ahora veamos un ejemplo.

Suponga que usted está muy contento porque obtiene un puntaje de 80% en su examen de Estadística. Para saber cómo se compara su resultado con respecto del de sus compañeros, usted descubre que el puntaje medio para la clase fue de 74 con una desviación estándar de 4. Usted asume que los puntajes están distribuidos de forma normal. ¿Cuál es su puntaje Z?

$$Z = \frac{80 - 74}{4} = 1.5$$

Su puntaje es de 1.5 desviaciones estándar por arriba de la media. Sin embargo, usted tiene curiosidad de saber a cuántos estudiantes pasó. Lo que necesitamos saber es la proporción de puntajes que se ubican por debajo de Z = 1.5 y toda el área por debajo de la media.
A partir de las tablas normales, el área entre la media y Z = 1.5 es 0.43319. A esto le debemos sumar el área por debajo de la media, que es la mitad de las observaciones o 0.5. (0.43319 + 0.5 = 0.93319). Esto significa que el 93.32% de la clase obtuvo un puntaje más bajo que el suyo (o sea menos de 80 puntos).

La muestra y la población

Con frecuencia queremos averiguar algo acerca de un grupo determinado de individuos y no nos es posible preguntarles a todos y por eso seleccionamos una muestra de ese grupo de personas. Para poder hacer nuestra selección es importante considerar los siguientes aspectos:

1. La población debe estar claramente definida.
2. La propiedad o característica que nos interesa también debe estar definida.
3. La muestra debe ser representativa de la población.

Con respecto del tercer punto, necesitamos escoger nuestra muestra con mucho cuidado para minimizar:

a. El error en la muestra
b. La variabilidad de la muestra

Tipos de errores en las muestras

Hay tres tipos de errores que podemos cometer y que tienen como consecuencia que nuestra muestra no tenga las mismas características que la población de la cual fueron seleccionadas.

1. Variabilidad de la muestra

Si nosotros seleccionamos varias muestras de la misma población, las medias y las desviaciones estándar no serán las mismas.

2. El error en la muestra

Un estimado de una muestra (por ejemplo, la media) no será exactamente el mismo que el valor estimado de la población. En otras palabras, el estimado será "el error". Esto no significa falta de cuidado, sino un artefacto del método usado para seleccionar la muestra. Por ejemplo, si en la muestra se escoge individuos a través del directorio telefónico, se estaría excluyendo a aquellos que no cuentan con una línea telefónica. El error en la muestra se puede reducir utilizando un método óptimo para la selección de la muestra.

3. Errores no provenientes de la selección de la muestra

Aquí se pueden ubicar cuestiones tales como un diseño pobre en las preguntas de un cuestionario; la manera de codificar las respuestas, entre otros.

Muestras representativas

Hay dos estrategias básicas en la selección:

a. Muestreo aleatorio simple. Aquí cada miembro de una población tiene la misma oportunidad de ser seleccionado.
b. Muestreo estratificado. Se usa cuando sabemos que hay grupos específicos en una población que sean diferentes entre sí. Queremos que la muestra represente a todos los estratos de la población. La estratificación más común en muchas muestras es por sexo y por edad. La pregunta que sigue es: ¿cuántos miembros de cada estrato deben estar en la muestra? El número debería ser proporcional al número de la población

(no siempre) ya que sería recomendable incluir un número mayor de aquellos estratos con mayor variabilidad y por lo tanto menos predecibles.

¿Qué tan grande debe ser la muestra?
El tamaño de la muestra debe reflejar el tamaño de la población, pero no siempre será proporcional al tamaño de la población. El tamaño de la muestra depende de:

a. La variabilidad entre los miembros de la población. Una población con mayor variabilidad (con una desviación estándar grande) requerirá una muestra grande. Si la población es uniforme, una pequeña muestra será más que adecuada.
b. Nivel de precisión requerido para calcular un estimado. Entre más precisión necesitemos, más grande tendrá que ser la muestra.
c. Costo de la muestra. No es necesario hacer la muestra innecesariamente grande ya que esto consume mayor tiempo y es más costoso.

LA DISTRIBUCIÓN DE LA MUESTRA PARA LA MEDIA

Suponga que tomamos 5 muestras, cada una de 5 mujeres, y les preguntamos cuántos hijos tienen. Los datos los concentramos en el cuadro siguiente:

Cuadro 1. Datos de 5 muestras de 5 mujeres que se les preguntó el # de hijos que tienen

Muestra	Número de hijos				
A	0	2	2	1	1
B	3	0	1	0	2
C	2	1	2	2	0
D	2	0	4	1	0
E	1	3	0	2	2

Podríamos calcular la media del número de hijos para cada muestra:

Cuadro 2. Media de # de hijos para las 5 muestras de mujeres[1]

Muestra	Media de # de hijos
A	1.2
B	1.2
C	1.4
D	1.4
E	1.6

Si las muestras son representativas, la mayoría de las medias deberían estar muy cercanas a la media real de la población en cuestión. Sin embargo, un pequeño número de muestras tendrá una media que se aleje de la media de la población. Esto nos llevaría a decir que la *distribución de una media muestral seguirá aproximadamente la distribución normal* (hecho 1)

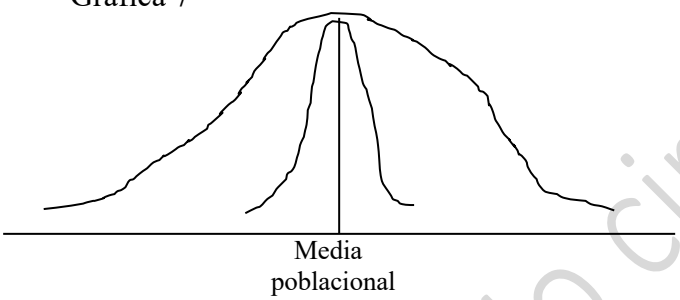
Si hemos tomados un número razonable de muestras, entonces *la media de la media muestral será aproximadamente igual a la media real de la población* (hecho 2). En nuestro ejemplo la media de las medias muestrales es de 1.36

[1] Las medias de diferentes muestras se conocen con el nombre de *media muestral*.

$$\text{Media de las medias muestrales} = \frac{1.2 + 1.2 + 1.4 + 1.4 + 1.6}{5} = 1.36$$

Suponiendo que el número de muestras para el ejemplo es razonable, podríamos decir que la media poblacional también estaría en 1.36. Ahora, si consideramos las medias muestrales, ¿variarán tanto como los valores de la población? En la siguiente gráfica, ¿cuál línea muestra la distribución de todos los valores de la población y cuál la distribución de las medias muestrales?

Gráfica 7



La distribución con una variación mayor es la distribución de toda la población y la que tiene un rango de distribución más angosto es la media muestral. ¿Por qué? Porque de los valores de toda la población, algunos valores extremos la harán más amplia, pero las medias muestrales son medias y por lo tanto se han eliminado los valores extremos. Entonces las medias muestrales son menos variables que los valores reales de la población. Esto significa que la desviación estándar de una media muestral es más pequeña que la desviación estándar de toda la población. De hecho, existe una fórmula para calcular la desviación estándar de una media muestral. La desviación estándar de una media muestral se conoce con el nombre de *error estándar* (hecho 3):

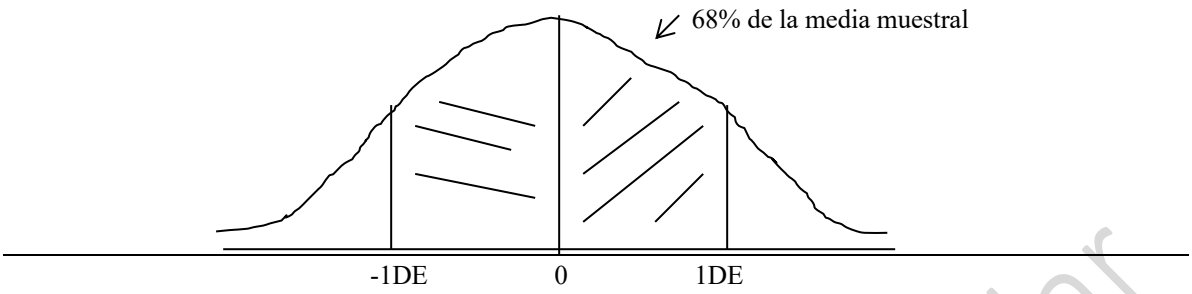
$$\text{Error estándar} = \frac{DE_{pob}}{\sqrt{n}}$$

El error estándar es igual a la desviación estándar de la población dividido por la raíz cuadrada del tamaño de la muestra.

A la desviación estándar de una media muestral se le llama error estándar para distinguirlo de otras desviaciones estándar comunes de muestras o poblaciones.

¿Para qué sirve saber el error estándar? Por ejemplo, si sabemos que la media muestral sigue una distribución normal, podemos decir que el 68% de todas las medias muestrales se ubicarán entre la media y una desviación estándar en cada lado (gráfica 8).

Gráfica 8



Esto nos permite estimar las posibilidades de que una media muestral sea mucho más grande o más pequeña que la media de la población. Podemos calcular un puntaje Z para una media muestral para ver que tanto más alta o baja es con respecto de la media de la población.

Pero antes vamos a ver que los tres hechos antes mencionados constituyen lo que se conoce con el nombre de teorema del límite central:

Teorema del límite central

Si unas muestras de tamaño *n* son seleccionadas de una población, las medias muestrales están aproximadamente distribuidas de forma normal con:

Media = media de la población

y

Error estándar = $\frac{DE_{pob}}{\sqrt{n}}$

¿Qué tan grande debe ser una muestra para que esto funcione?

- Si la distribución de la población es normal, una muestra de un tamaño de 10 o más es necesaria.
- Si la distribución no es normal (si es asimétrica), el tamaño de la muestra debe de ser de por lo menos de 25.

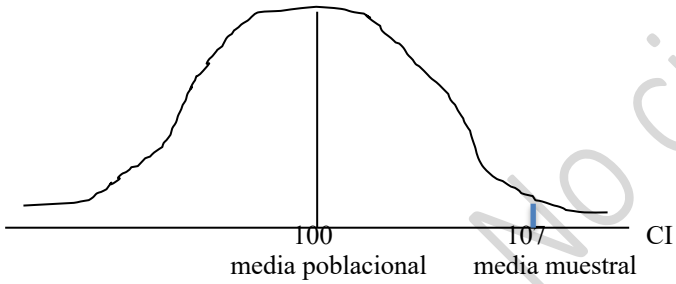
¿Para que nos sirve esto? Nos sirve para saber si una media muestral que hemos encontrado es diferente de la media poblacional debido al azar (la variabilidad de la muestra) o si la muestra es genuinamente diferente de la población. También sirve para estimar la precisión de cualquier muestra que tomemos.

Ejemplo. El coeficiente intelectual de los estudiantes

El coeficiente intelectual de la población general está distribuido normalmente con una media de 100 y una desviación estándar de 15. Imagine que selecciona aleatoriamente una muestra de 40 estudiantes y encuentra que su CI medio es de 107. Suponiendo que aceptamos que el CI realmente mide inteligencia, ¿son estos estudiantes realmente más inteligentes que el promedio de la población?

Lo que estamos tratando de averiguar es la posibilidad de que de una muestra de 40 estudiantes que tienen un CI medio de 107 sea debido al azar o si realmente son más inteligentes. Esto lo podemos representar así:

Gráfica 9



Lo primero que debemos hacer es averiguar la distribución muestral para las medias de una muestra de 40. De acuerdo con el Teorema de límite central, la media muestral estará distribuida normalmente cuando:

$$\text{Media} = \text{media poblacional} = 100$$

$$\text{Error estándar} = \frac{DE_{pob}}{\sqrt{n}} = \frac{15}{\sqrt{40}} = 2.37$$

Posteriormente, calculamos el puntaje Z utilizando la formula:

$$Z = \frac{\text{media muestral} - \text{media poblacional}}{\text{error estándar}}$$

$$Z = \frac{107 - 100}{2.37} = 2.95$$

Un puntaje Z de 2.95 es muy alto. Si buscamos Z = 2.95 en la tabla normal nos da un área de 0. 49841 entre la media y Z = 2.95. Así que el área por arriba de Z = 2.95 será de (0.5 – 0.49841) = 0.00016

Esto nos dice que la probabilidad de obtener una muestra de 40 estudiantes con un CI de 115 o más es extremadamente baja ya que los individuos varían mucho más que las medias muestrales.

El puntaje Z para un individuo con un puntaje de 107 sería de:

$$Z = \frac{\text{observación – media poblacional}}{\text{DE poblacional}} = \frac{107 - 100}{15} = 0.47$$

Si buscamos un puntaje Z de 0.47 en la tabla normal tendremos un área de 0.1808. Por lo tanto, el área por arriba de Z = 0.47 será de 0.5 – 0.1808 = 0.3192. Así que la probabilidad de que un individuo tenga un CI de 107 o más es de 0.3192 (alrededor de 32%), lo cual es bastante razonable.

Ejemplo 2.

A trabajadores eventuales de una ocupación cualquiera se les paga un salario promedio de $4.60 la hora con una desviación estándar de $0.40. Los salarios están distribuidos de forma normal.

1. ¿Cuál es la probabilidad de que un individuo gane $4.50 o menos por hora?
2. ¿Cuál es la probabilidad de obtener una muestra de 20 trabajadores con un salario promedio de $4.50 o menos?
3. ¿Cuál es la probabilidad de obtener una muestra de 50 trabajadores con un salario medio de $4.50 o menos?
4. ¿Por qué sus respuestas a 1, 2 y 3 son diferentes?