Clase 4

26/8/2019

Variable Aleatoria Continua

A diferencia de una variable discreta que puede asumir una cantidad finita de valores en un intervalo determinado, la variable aleatoria continua puede asumir infinitos valores dentro de un intervalo.
Existen diversas distribuciones de probabilidad continuas que son aplicables como modelos a una amplia gama de variables aleatorias de este tipo. Entre estos modelos uno de los más importantes es el que veremos a continuación

Distribución Normal

La distribución normal es una de las más importantes en estadística por sus numerosas aplicaciones teóricas, también es llamada de Gauss o de Laplace-Gauss, en virtud de quienes la formulan en forma independiente cada uno de ellos.
De la observación de los distintos fenómenos estadísticos se concluye que muchas variables aleatorias continuas tienen un comportamiento como el descripto por esta distribución.

Ejemplos de aplicaciones

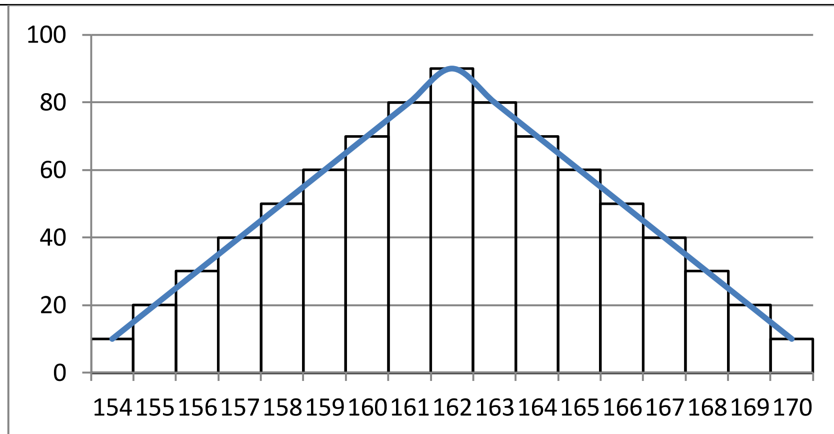
- Características morfológicas de personas, animales o plantas, como ser altura, peso, diámetros, perímetros, etc
- Características Fisiológicas, como efectos de determinadas dosis de un fármaco, o resultado de aplicar determinada cantidad de fertilizante sobre las plantas o suelos.
- Características sociológicas: consumo de determinados productos por algunos grupos de personas, puntuaciones de examen.
- Características Psicológicas: coeficiente intelectual, grado de adaptación a un medio.
- Comportamiento de valores estadísticos muestrales, como la media.

Función de Densidad de la Normal

$$P(X) = (1 / \delta *\sqrt{(2\prod)}) * e^{-((X - \mu)2)/2\delta}$$

Ejemplo de trabajo

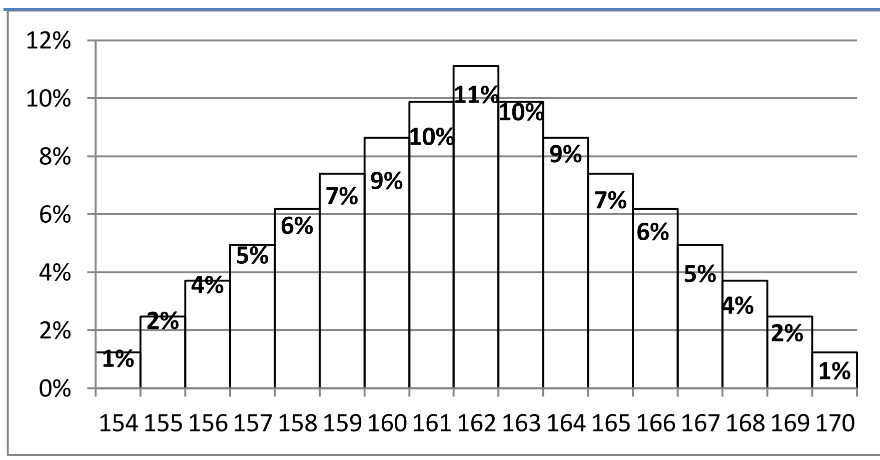
Altura	FA	FR	FR (%)
154	10	0,01	1%
155	20	0,02	2%
156	30	0,04	4%
157	40	0,05	5%
158	50	0,06	6%
159	60	0,07	7%
160	70	0,09	9%
161	80	0,10	10%
162	90	0,11	11%
163	80	0,10	10%
164	70	0,09	9%
165	60	0,07	7%
166	50	0,06	6%
167	40	0,05	5%
168	30	0,04	4%
169	20	0,02	2%
170	10	0,01	1%
	810	1,00	100%



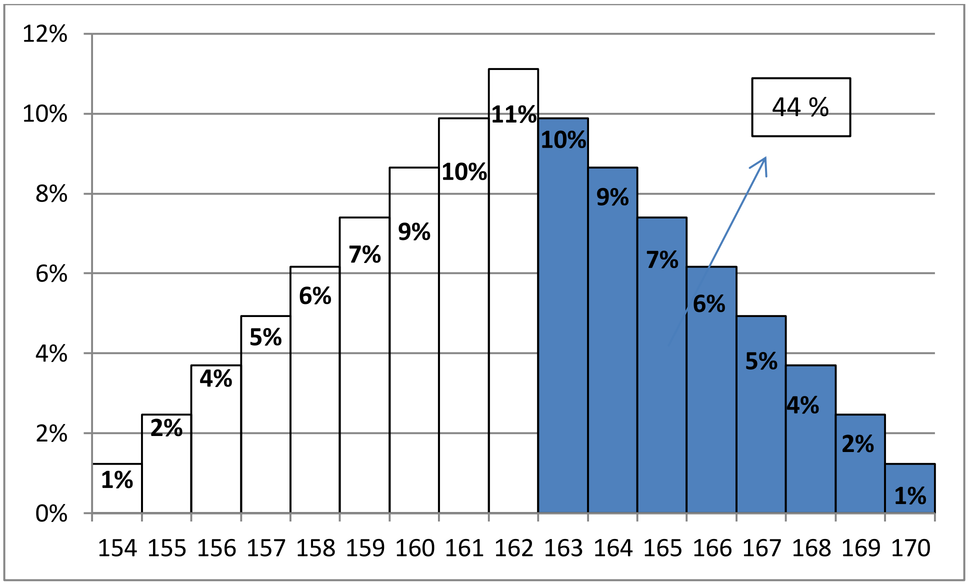
Características de Distribución Normal

- Es simétrica, lo cual implica que la M (X) = Mdna (X) = Mo
- Tiene forma de campana
- Depende de dos parámetros: la media (ư) y la desviación estándar (δ), estos definirán la forma de la distribución
- Es asintótica horizontalmente
- La suma del área bajo la curva es igual a 1

Algunos análisis posibles



Probabilidades asociadas a cada intervalo de altura de esa población



Probabilidad de que mida 163 cm o más, es 44 %

Relación entre la media y la desviación estándar en la distribución normal

<table>
<tr><td>Ư + δ</td><td rowspan="2">68 % de los datos</td></tr>
<tr><td>Ư - δ</td></tr>
<tr><td>Ư + 2δ</td><td rowspan="2">95 % de los datos</td></tr>
<tr><td>Ư - 2δ</td></tr>
<tr><td>Ư + 3δ</td><td rowspan="2">99 % de los datos</td></tr>
<tr><td>Ư - 3δ</td></tr>
</table>

Distribución Normal estandarizada

Es el procedimiento aplicable a cualquier variable con distribución normal que genera otra variable que vamos a llamar Z y que tiene la particularidad de que su media es igual a 0 y su desviación estándar es 1, simbólicamente

Z (ư=0; δ=1) = (X - ư) / δ

Clase 5: Ejercicios prácticos

30/08/2019

Distribución Poisson

1. En promedio 6 personas utilizan un cajero automático en una sucursal del banco. Cuál es la probabilidad de:
   a. Exactamente 6 personas utilicen el cajero durante una 1 hora seleccionada al azar?
   b. Menos de 5 personas lo utilicen durante el mismo periodo?
   c. Nadie utilice la caja durante un intervalo de 10 minutos
   d. Nadie utilice la caja durante un intervalo de 5 minutos?

2. Un libro de texto tiene 50 errores de mecanografía en el total de 500 páginas del mismo. Cuál es la probabilidad de que:
   a. Un capitulo de 30 páginas tenga 2 o más errores
   b. Un capítulo de 50 páginas tenga 2 o más errores
   c. Una página elegida al azar no tenga error

3. Una empresa electrónica observa que el número de componentes que fallan antes de cumplir 100 horas de funcionamiento es una variable aleatoria de Poisson. Si el número promedio de estos fallos es ocho. Cuál es la probabilidad de que:
   a. Falle un componente en 25 horas?
   b. Fallen no más de dos componentes en 50 horas?
   c. Fallen por lo menos diez en 125 horas?

Distribución Normal

Ejercicios

1. Se sabe que el tiempo útil de un componente eléctrico tiene una distribución normal con media $\mu = 2000$ y desviación estándar $\sigma = 200$ horas.
   a. Determine la probabilidad de que un componente elegido al azar dure entre 2000 y 2400 horas
   b. Realice el mismo cálculo mediante la distribución normal estandarizada

2. Se ha ajustado el proceso de fabricación de un tornillo de precisión de manera que la longitud exacta es de 13 cm. No todos los tornillos tienen esa longitud exacta debido a distintas fuentes de variabilidad existentes en el proceso. La desviación estándar de la longitud de los tornillos es σ = 0.1 cm y se conoce que la distribución de las longitudes es normal. Determine la probabilidad de que un tornillo elegido al azar tenga una longitud de entre 13 y 13.2 cm
3. En el caso del ejercicio anterior. Cuál es la probabilidad de que la longitud del tornillo exceda de 13.25 cm?. Grafique
4. Cuál es la probabilidad de que la longitud del tornillo este entre 12.9 y 13.1 cm?. Y la probabilidad de que no esté entre el intervalo mencionado?. Grafique

5. El promedio de vida útil de ciertos tipos de llantas tiene una distribución normal, con media μ = 38000 km y σ = 3000 km. Se pide
   a. Cuál es la probabilidad de que una llanta elegida al azar tenga una vida útil de cuando menos 35000 km?. Grafique
   b. Cuál es la probabilidad de que dure más de 45000 km?. Grafique

6. Un distribuidor hace un pedido de 500 llantas de las descriptas en el ejercicio anterior
   a. Aproximadamente cuantas llantas durarán entre 40000 y 45000 km?
   b. 45000 km o más?

Clase 6: Aproximación a distribuciones Binomial y Poisson

6/9/2019

Aproximación normal a la probabilidad binomial

Cuando el número de casos observados en una variable con distribución binomial es relativamente grande se puede utilizar la distribución normal de probabilidades para aproximar las probabilidades binomiales. Una regla generalmente utilizada plantea que cuando el número de casos es mayor o igual 30 y np y nq son mayores o iguales a 5, esto puede darse.

- $\mu = np$
- $\sigma = \sqrt{(npq)}$

Ejercicios

1. De un grupo se ventas se observa que el 20% de los que un vendedor visita en forma personal realizan la compra. Si un representante de ventas visita a 30 posibles clientes , cual es la probabilidad de que se realicen 10 o más ventas?.
    a. Calcule la probabilidad utilizando la distribución binomial
    b. Calcule utilizando la aproximación según la distribución normal

Aproximación normal a la probabilidad de poisson

Cuando la media (λ) de una distribución poisson es relativamente grande, se puede recurrir a la distribución normal para aproximar los respectivos valores de probabilidad. Por lo general esto sucede cuando λ es mayor o igual a 10. Además tener en cuenta que:

- $\mu = \lambda$
- $\sigma = \sqrt{(\lambda)}$

Ejercicios

El número promedio de solicitudes de servicio que reciben en un departamento de reparación de maquinaria por cada turno de 8 horas es 10. Puede determinarse la probabilidad de que se reciban más de 15 solicitudes en un turno de 8 horas?

a. Calcule la probabilidad utilizando la distribución Poisson
b. Calcule utilizando la aproximación según la distribución normal

Clase 8: Distribución en el Muestreo

27/09/2019

La estadística inferencial trata de la posibilidad de estimar valores poblacionales a partir de valores muestrales. Por lo tanto lo que vamos a empezar a ver ahora, es un conjunto de procedimientos que nos permiten “inferir” conclusiones sobre un universo de información a partir de una parte no muy grande de estos.
En general en diversas investigaciones es habitual analizar un subconjunto de elementos y generalizar las conclusiones a un universo mayor, por cuestiones de costo y de tiempo.
Estos procedimientos nos permiten que los errores propios de inferir conclusiones a partir de un subconjunto de datos, puedan ser conocidos y controlados.

Definiciones Preliminares

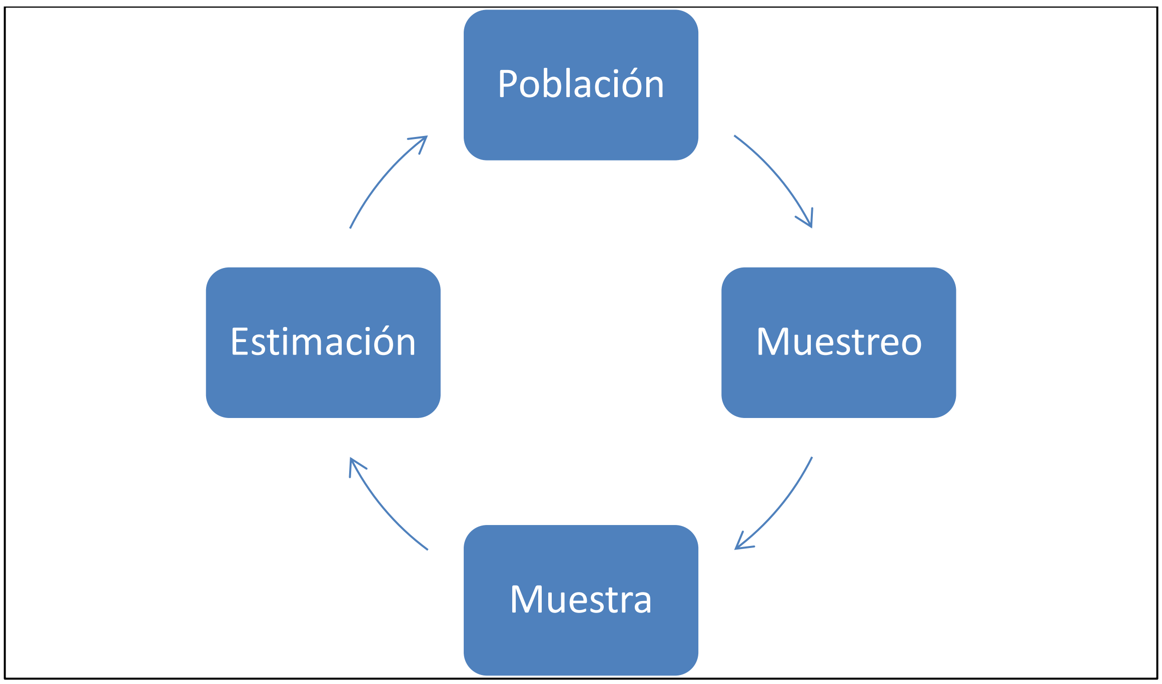
Población:
Es el nombre genérico que le damos a un conjunto de unidades de análisis que son objeto de un estudio o una investigación particular.

- Ejemplos:
- Los pacientes con trastornos severos del hospital San Roque de la ciudad de Córdoba durante el periodo 2015/2019
- Los Estudiantes de IES21 durante el año 2018
- Los votantes de la República Argentina durante las elecciones “pasos” del 12 de mayo de 2019.
- La incorporación de las nuevas tecnologías en las empresas de la provincia de Córdoba

Muestra:
Se llama así a un subconjunto de la población que comparte sus características en aquellos aspectos que son de interés para la investigación. El concepto de muestra va asociado siempre a la idea de “Representatividad”, es decir, la muestra seleccionada debe tener características similares a la población bajo estudio en los aspectos

que nos interesan. Por ejemplo si en una muestra de votantes nos interesa la percepción diferente que puede haber entre los distintos niveles socioeconómicos de los mismos y sabemos por estudios anteriores que la población total está compuesta por 30% de niveles bajos, un 50% de niveles medios y un 20% de niveles altos, luego estos estratos deberán tener una proporción similar en la muestra.

Esquema de relación entre la población y la muestra



El objetivo de una investigación por muestreo es obtener información acerca de las características de la población a partir de datos provenientes de la muestra. La característica poblacional que pretende conocerse la llamaremos **Parámetro.** Los valores calculados de esta manera se llamarán **Estimadores**.

Clase 9 y 10: Estimación Puntual y por intervalos

4, 7 y 11/10/2019

En lo visto sobre muestreo indicamos la importancia del carácter aleatorio en la elección de las unidades que constituirían la muestra. Esta exigencia se funda en la necesidad de generar variables aleatorias cuya distribución de probabilidad conozcamos, o bien podamos suponer. Si podemos hacer eso, conoceremos las probabilidades asociadas a determinados resultados muestrales y podremos, entonces, usarlos para inferir sobre la población, que es, no lo olvidemos, nuestro objetivo.
Las medidas que calculemos a partir de datos muestrales dependerán del azar. La media muestral, por ser una medida calculada sobre los valores obtenidos en una muestra aleatoria, es una variable aleatoria, su valor depende de cuáles sean los casos que constituyen la muestra y esto depende del azar, que es lo que hemos pedido como requisito al procedimiento de muestreo. Estamos entonces refiriéndonos a entidades diferentes cuando hablamos de la media poblacional y la media muestral, aunque el procedimiento de cálculo sea el mismo. La media poblacional es un valor que puede ser conocido (si hemos observado a toda la población) o desconocido, pero en todo caso es fijo. Cuando hacemos estimaciones no sabemos cuánto vale la media de la población, pero sí sabemos que es un número estable, fijo. Por el contrario, la media muestral depende del azar, luego **es una variable aleatoria**.
Las medidas que se refieren a la población se denominan parámetros. Si se trabaja con variables cualitativas, se tratará de la **proporción** de alguna categoría, se tratará allí de **la proporción paramétrica o proporción poblacional**. O bien de la varianza, la desviación estándar o también de coeficientes de correlación. Todas las medidas que hemos mencionado hasta aquí pueden calcularse sobre la población completa si se hace un censo, o bien sobre una muestra que debe ser aleatoria si se quiere luego hacer inferencias. En todos los casos que se trate de medidas que se refieran a la población completa, las llamaremos paramétricas. Solo podrán ser conocidas en los casos en que la población completa sea observada, cuando se haga un relevamiento exhaustivo, un censo.

Para distinguir entre las medidas de la población y las de la muestra usaremos diferentes símbolos. De manera general, las letras latinas se usarán para identificar medidas descriptivas obtenidas sobre datos muestrales, mientras que usaremos letras griegas para referirnos a los valores de la población.
La siguiente es la notación que usamos y la correspondencia entre valores poblacionales y muestrales.

Convenciones Población-Muestra

Variables	En la Población	En la muestra
Cantidad de casos	N	n
Media	μ	$\overline{x}$
Proporción	P	ṕ
Varianza	σ^2	S^2
Desviación Estandard	σ	S

Ejemplo para estudiar algunos conceptos de interés

Paciente	Meses Internado
A	3
B	4
C	5

Calculemos los parámetros Media Poblacional y Varianza Poblacional:

$\mu = (3 + 4 + 5)/3 = 4$

Si calculamos la Varianza poblacional obtenemos:

$\delta = ((3 - 4)^2 + (4 - 4)^2 + (5 - 4)^2) / 3 = 0.66$

Trabajemos ahora con todas las muestras posibles de tamaño 2 de esta población total de tamaño 3 (este es solo un ejercicio teórico que pretende analizar las posibilidades reales que suceden al estimar parámetros poblacionales a partir de una muestra, sólo tiene ese fin) Supuesto de origen: Se seleccionan muestra de tamaño 2, con reposición

Muestra	MI	$\bar{x}$
AA	3-3	3
AB	3-4	3.5
AC	3-5	4
BA	4-3	3.5
BB	4-4	4
BC	4-5	4.5
CA	5-3	4
CB	5-4	4.5
CC	5-5	5

Función de distribución de la media muestral

Media Muestral

$\bar{x}$	FA	FR
3	1	0.11
3.5	2	0.22
4	3	0.33
4.5	2	0.22
5	1	0.11
Total	9	1

Comentarios

La media muestral es una variable aleatoria. Diferentes muestras ofrecen valores diferentes para la misma.

Al hacer un histograma de frecuencias de la media muestral y analizar las probabilidades de estar cerca del verdadero valor de la media poblacional, observamos que esta se distribuye en forma normal, independientemente de la distribución de la variable original (días de internación)

Al Calcular la E ($\bar{x}$) y analizar su resultado, tenemos que:

$E(\bar{x}) = (3*0.11)+(3.5*0.22)+(4*0.33)+(4.5*0.22)+(5*0.11) = 4$, es decir la esperanza de la media muestral es igual a la media poblacional (μ), por lo que concluimos que:

La media muestral es un estimador **insesgado** de la media poblacional. Por esta razón

Ya que justamente esta es la definición de estimadores insesgados, cuando la esperanza matemática del estimador es igual al valor poblacional.

Del mismo modo si calculamos la varianza de la media muestral obtenemos el siguiente resultado:

$V(\bar{X}) = (3-4)^2 * 0.11 + (3.5-4)^2 * 0.22 + (4-4)^2 * 0.33 + (4.5-4)^2 * 0.22 + (5-4)^2 * 0.11 = 0.33$, aquí observamos que la varianza de la media muestral es exactamente la mitad de la media poblacional, y eso sucede porque el tamaño de la muestra es 2, es decir, el valor de la varianza poblacional dividido el tamaño de la muestra, por lo que concluimos que $S^2 = \delta^2 / n$ y por lo tanto $S = \delta / \sqrt{n}$

Aprovechamos este resultado para definir otra propiedad de los estimadores que es la consistencia, la misma dice que un estimador es consistente si al aumentar el tamaño muestral disminuye la dispersión resultante, es decir los datos están más cercanos a la media estimada. En este caso esto se verifica ya que n está en el denominador por lo que el resultado de S, disminuye al aumentar n. Con lo que concluimos que:

La media muestral es un estimador **Consistente** de la media poblacional, ya que su varianza disminuye a medida que aumenta el tamaño de la muestra

Estimación puntual

La media muestral es un estimador de la media poblacional, por lo que ya tenemos una primera estimación de ese parámetro. Se llaman así porque ofrecen un único valor como estimación del parámetro de interés. Por ejemplo si en una muestra de 50 médicos que egresaron en los últimos diez años hallamos que han terminado la carrera con una nota promedio de 6,50, disponemos de una media muestral; si ahora preguntamos por el promedio con que terminaron la carrera todos los médicos que egresaron en los últimos diez años, la respuesta es tentativa, diremos que "debe ser cercano a 6,50". Con esta expresión imprecisa, hacemos una estimación de la media poblacional. Así hacemos una estimación deficiente, imprecisa, ya que no sabemos cuán cerca puede estar la verdadera nota promedio de 6,50. Estas son las que se denominan estimaciones puntuales.

Estimación por intervalo

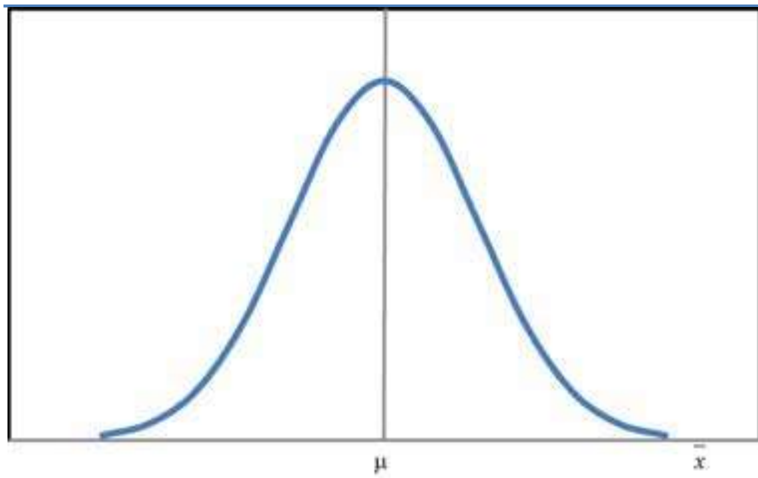
Una estimación más completa de los parámetros que nos interesan, se denomina estimación por intervalo. Ella consiste en ofrecer no ya un número como en la estimación puntual, sino un intervalo, acerca del cual tendremos cierta certeza (o confianza) que contenga al parámetro. Así, en lugar de decir que el promedio con que egresa el total de médicos de la facultad "debe ser cercano a 6,50", construiremos un intervalo, que dirá, por ejemplo, "tenemos una certeza del 95% que el intervalo 6,10; 6,90 contiene al promedio con que egresan los médicos de la facultad de Medicina". Veamos a continuación cómo construir estos intervalos de confianza para estimar la media poblacional.

Estimación de la media

Vamos a hacer uso de lo que sabemos hasta el momento sobre las distribuciones en el muestreo para mejorar la calidad de las estimaciones puntuales y construir los intervalos de confianza.

Para ello, empezaremos con la media, la muestra ha sido sacada de manera aleatoria, la media muestral es una variable aleatoria, cuya distribución tiene media y desviación estándar. Además, a medida que aumenta el tamaño de la muestra, esa distribución tiende a ser normal, es decir que será tanto más cercana a una distribución normal cuanto más grande sea n. A los fines prácticos, una muestra de 30 casos se considera suficientemente grande como para usar la distribución normal en la estimación de la media poblacional. Si la muestra es más pequeña que ese tamaño, no podemos usar inmediatamente la distribución normal, sino que deberemos apelar a la distribución t de Student. Trabajaremos primero suponiendo que se trata de muestras lo suficientemente grandes y usaremos la distribución normal.

Distribución de la media Muestral



Con esta información y recurriendo a la distribución normal estandarizada podemos estimar un intervalo de confianza para la misma donde se concentran el 95 % de los datos de esta, calculando los valores de Z, entre los cuales se encuentran el 95% de los datos en la mencionada distribución.

Para ello podemos obtener los valores de Z entre los cuales se concentra el 95% de los valores en la normal estandarizada, recurriendo a la Inversa normal estandarizada y obteniendo los valores de Z, con los valores acumulados de probabilidad hasta ellos. En el primer caso hasta 0.025 (2,5%, 5%/2), nos da el valor de Z_1 = -1.96 y de la probabilidad acumulada hasta 0.975 (97.5%, 95% + 2.5%), el que nos da Z_2 = 1.96

Si queremos obtener el intervalo en la variable media muestral, debemos recordar, primero la definición de la variable Z

$Z = (x - \mu) / \delta$

Pero en este caso la variable X es en realidad $\bar{X}$, por lo que nos queda $Z = (\bar{X} - \mu) / (\delta /\sqrt{n})$, recordando lo obtenido anteriormente respecto a que la desviación estándar de la media muestral es igual a lo que figura en el denominador de esta última formula.
Despejando de esta fórmula a μ que es lo que deseamos estimar, nos queda

$\mu = \bar{X} + Z * (\delta /\sqrt{n})$ y reemplazando Z, por sus respectivos valores correspondientes al intervalo de confianza del 95%, de Z1 = -1.96 y Z2 = 1.96, Obtenemos el intervalo buscado.

Lo Hagamos en función del ejemplo desarrollado al comienzo. Aquí deberíamos tomar algunas de las muestras posibles de la tabla anterior, por ejemplo la que corresponde a los valores AB, de días de internación que eran 3 y 4 y su media es 3.5 y reemplazar en la anterior formula, obtenemos
$\mu = 3.5 - 1.96 * (0.66 / \sqrt{2}) = 2.59$ el Límite inferior del correspondiente al valor de Z_1
$\mu = 3.5 + 1.96 * (0.66 / \sqrt{2}) = 4.41$ El Límite superior para el intervalo de la media poblacional

Es decir, podemos concluir que con un 95% de confianza el verdadero valor de la media poblacional esta entre el valor 2.59 y 4.41, a partir de la estimación puntual estimada con la muestra seleccionada.

Ejercicio de Práctica: realice este mismo cálculo para todas las muestras posibles y verifique en cuantas de estas estimaciones está comprendido el verdadero valor de la media poblacional. Extraiga conclusiones

Clase 11 y 12: Prueba de Hipótesis

1 y 4/11/2019

Prueba de hipótesis

La prueba de hipótesis tiene como objetivo el darnos argumentos para decidir en contextos de incertidumbre. A partir de lo que sabemos sobre las distribuciones en el muestreo, ése es el caso cuando necesitamos concluir acerca de una población, a partir de información que tenemos disponible en una muestra aleatoria. El resultado de la prueba permitirá decidir si lo que se observa en la muestra es compatible con un planteo hipotético sobre la población. Nunca será posible decidir de manera taxativa que la hipótesis es verdadera, eso es algo que no podemos saber; por el contrario, podemos ver hasta qué punto lo que observamos en la muestra contradice o no lo que se afirma a escala poblacional. Es decir que podremos descartar una hipótesis por no ser compatible con lo que se observa, pero no a la inversa: no será posible "confirmar" una hipótesis, solo podremos concluir que la evidencia no la contradice, lo que también se expresa diciendo que no hay evidencia para rechazarla. Empezaremos con ejemplos no muy cercanos a la estadística, a fin de ver que esta forma de razonar no es para nada ajena a lo cotidiano. Una prueba de hipótesis puede compararse con un juicio: el acusado no es condenado hasta que no hay evidencia suficiente para hacerlo. La evidencia (las pruebas, en el lenguaje de la justicia) rara vez son completas, se trata de información fragmentada, sujeta a interpretaciones diferentes. En el inicio del juicio, "el acusado es inocente", en nuestra notación llamaremos a esa afirmación, hipótesis nula, y la indicaremos Ho Esta expresión indica que se trata de un estado inicial: todos son inocentes hasta que se prueba lo contrario, por lo que la hipótesis nula señala que esta persona en particular (el acusado), no es diferente de cualquier ciudadano que no ha cometido delito. Mientras no haya pruebas suficientes, la hipótesis nula se considerará aceptada. En el juicio, el fiscal aportará pruebas en dirección contraria a esta hipótesis. Buscará información para probar que debe rechazarse la hipótesis nula y condenar al acusado. Difícilmente estarán a la vista todos los datos necesarios para reconstruir la situación y dar una respuesta absolutamente inequívoca, pero si hay suficiente evidencia, se dará la hipótesis nula por rechazada. La decisión de condenar al acusado solo se tomará cuando haya muy poco riesgo de equivocarse, cuando la probabilidad de decidir de manera errada sea muy pequeña.

En este ejemplo, la población es el conjunto completo de información necesaria para tomar la decisión de manera certera. Se trataría de un conjunto de datos muy amplio, que no está disponible, por lo que la decisión debe tomarse a partir de un fragmento de información, que son las pruebas que han podido reunirse, en la analogía que hacemos con nuestros procedimientos, esto constituye la muestra a partir de la que se tomará la decisión sobre la hipótesis nula: aceptarla o rechazarla.

En investigación, la prueba de hipótesis suele formularse de tal modo que rechazar H0 implica aportar un nuevo hallazgo, por el contrario, aceptar H0 equivale a que no hay cambios respecto de la situación inicial.

Algunos ejemplos de hipótesis nulas:

- Esta droga no produce ningún efecto sobre la memoria.
- La técnica terapéutica A es igualmente eficaz que la B.
- Los métodos A y B para enseñar a leer a los niños producen iguales resultados.

En casi todos los casos, la expectativa del investigador está en rechazar la H0, porque eso significa que ha hallado algo de interés: que la droga produce efectos, que hay técnicas terapéuticas mejores que otras y por tanto recomendables, que se pueden elegir mejores métodos para enseñar a leer, que el favor del electorado hacia un político es más o menos extendido.

Al realizar pruebas de hipótesis lo que haremos será evaluar cuál sería la probabilidad de hallarlo si fuera cierta la hipótesis nula. Cuando esta probabilidad sea grande no habrá evidencia para rechazarla, cuando sea pequeña decidiremos rechazarla.

Prueba de Hipótesis para la media poblacional

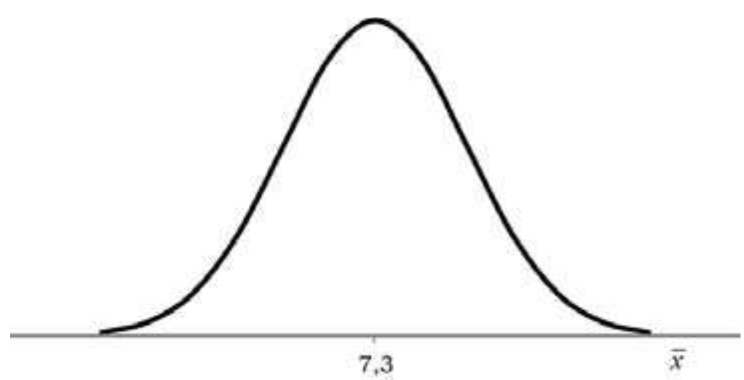
Supongamos que para una determinada carrera universitaria, históricamente los alumnos han tardado para recibirse un promedio de 7,30 años. Decimos históricamente para indicar que son datos acumulados por largo tiempo y que provienen de los registros de la facultad de años atrás. Se ha introducido un cambio en el plan de estudios de la carrera y puede creerse que con ese cambio los alumnos tardarán un tiempo distinto en recibirse. Tenemos entonces un promedio de la población de quienes se recibieron en las anteriores condiciones (una media poblacional histórica), y queremos hacer inferencia sobre la media poblacional de los alumnos que cursan con el nuevo plan. Estos últimos no están todos accesibles, porque hay alumnos que están cursando y otros que lo harán en el futuro, por lo que de esa población solo puedo conocer a una muestra de los que ya han egresado y ver cuánto tiempo han tardado ellos en recibirse.

Expresamos las hipótesis de este modo:

H0 : $\mu = 7.30$

H1 : $\mu \neq 7.30$

La hipótesis nula indica que la media poblacional de los alumnos que cursan con el nuevo plan es la misma que antes, que no hay diferencia, que no hay cambios. La hipótesis alternativa afirma lo contrario: que el tiempo promedio que tardan los alumnos en terminar la carrera con el nuevo plan es diferente a los 7,30 años históricos. Ambas son afirmaciones sobre la población (sobre el parámetro media poblacional), por eso son hipótesis. Si la hipótesis nula fuera cierta, por lo que sabemos sobre las distribuciones en el muestreo, la siguiente sería la distribución de las medias muestrales:

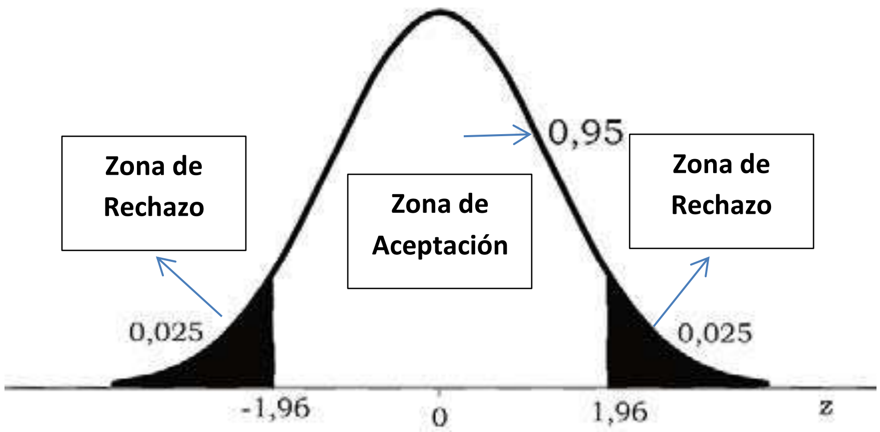


Es decir tiene una distribución normal con media 7.3 (recordemos lo que vimos en el teorema central del límite sobre la distribución de la media muestral)

A fin de realizar la prueba de hipótesis debemos obtener una muestra. Supongamos que seleccionamos 100 egresados y que encontramos un tiempo promedio para terminar la carrera de 7,50 años con una desviación estándar de 1,30 años. Debemos tener un criterio para decidir si este valor observado es compatible con la hipótesis nula ($\mu = 7{,}30$) o si constituye evidencia suficiente para rechazarla a favor de la hipótesis alternativa ($\mu \neq 7{,}30$). El criterio es el de ver cuán probable sería este valor observado si la hipótesis nula fuera cierta. En consecuencia, debemos calcular la probabilidad que tiene la media muestral de asumir el valor observado, por lo que buscaremos la probabilidad de hallar valores como el observado (7,50) o más extremos que él. Esto significa que nos preguntamos por la probabilidad que tiene la variable media muestral de asumir el valor 7,50 o uno más extremo, es decir un valor que se aleje más de la media hipotética. Hemos dicho que "alejado" equivale a "poco probable si H0 fuera cierta", por lo que los valores alejados se encuentran en los extremos de la distribución de la variable analizada, bajo el supuesto de H0 verdadera (es decir, centrada en la media hipotética). Para decidir si la evidencia hallada en la muestra es suficiente para rechazar la hipótesis nula, vamos a establecer a priori un valor máximo para la probabilidad de ocurrencia del valor muestral, o lo que es lo mismo, un valor máximo para el área extrema donde consideraremos que se encuentran los valores "alejados".

Criterio de toma de decisión

La H0 se rechazará si hay poca probabilidad de hallar un valor como el observado o uno más extremo que él. Lo que llamamos poca probabilidad, puede establecerse a priori, por ejemplo en 0,05. Eso indica que consideraremos a los resultados con probabilidad menor a 0,05 como muy improbables de hallar si H0 fuera cierta y nos conducirán a rechazarla. Por el contrario, si encontramos valores cuya probabilidad de ocurrencia es superior a 0,05, los trataremos como valores esperables y nos conducirán a aceptar la H0. Como sabemos de la distribución normal, los valores de $Z = \pm 1{,}96$ delimitan un área central de 95%, es decir que dejan fuera un área de 5%. Los valores de Z superiores a 1,96 ó inferiores a -1,96 tienen una probabilidad de ocurrencia de 0,05, repartida en las dos "colas" de la distribución normal.



Los valores de la media muestral que correspondan a puntajes Z que superen a 1,96 ó sean inferiores a -1,96 serán valores con probabilidad menor a 0,05, por lo que serán considerados como poco probables y conducirán a rechazar H0. Por el contrario, los valores que tengan Z comprendido entre -1,96 y 1,96 serán probables y nos llevarán a que aceptemos H0. Estos dos puntos (-1,96 y 1,96) se denominan valores críticos de Z y podemos indicarlo con un subíndice: Zc. En nuestro ejemplo, el valor observado es $\bar{X}$ = 7,50, lo podemos llamar $\bar{X}obs$. El puntaje Z equivalente a ese $\bar{X}obs$ se llama Z observado (Zobs) y vale:

$Zobs = (\bar{X}obs - \mu) /(S/\sqrt{n}) = (7{,}50 – 7{,}30) / (1{,}3/\sqrt{100}) = 1{,}54$

Al comparar este valor con los valores extremos considerados (-1.96, 1.96), observamos que este valor positivo es inferior a 1.96 y por lo tanto cae en la zona de aceptación y decimos que en base al valor obtenido de la muestra seleccionada no hay evidencia suficiente para rechazar la Hipótesis de que la media poblacional es de 7.30 para este grupo de alumnos. Estamos diciendo que dado el valor obtenido en la muestra y como este no está en los valores extremos, es posible que esta muestra provenga de una distribución como la que propone la hipótesis Nula

Se llama nivel de significación a la probabilidad de hallar al valor muestral en la zona de rechazo de H0, si H0 es verdadera. Se indica como α, y es elegido por el investigador

Tipos de Errores en Pruebas de Hipótesis

Dado que la decisión de aceptar o rechazar la H0 se toma de manera probabilística, siempre existe la posibilidad de tomar una decisión incorrecta. Esto sucede porque las muestras son tomadas al azar y puede suceder que la que usamos para tomar la decisión sea una muestra extrema. Aunque es un resultado poco probable, no es imposible.

Como hemos visto, el nivel de significación mide la probabilidad de hallar un determinado resultado muestral si la H0 fuera cierta, es una probabilidad pequeña, que habitualmente fijamos en 0,05 ó 0,01. Si la H0 es cierta y la muestra sobre la que basamos la decisión es extrema, es decir, tiene un valor ubicado en alguna de las colas de la distribución, nuestra decisión será la de rechazar H0 y esa decisión será errónea. Al momento de decidir, no podemos saber si H0 es verdadera y obtuvimos una de esas muestras muy poco probables, o si efectivamente H0 es falsa. Por esta razón el nivel de significación mide la probabilidad de errar en la decisión de esta manera: rechazando una H0 que es verdadera. Éste se conoce como Error de Tipo I (ETI).

En consecuencia, establecer α es afirmar que se está dispuesto a correr ese riesgo de cometer el ETI. En un experimento que consiste en decidir si una droga produce efectos sobre un determinada patología, la Ho dirá que no hay efecto, por lo que cometer el ETI será creer que hay efecto (rechazar H0) cuando en realidad no lo haya (H0 verdadera). Como no sabemos si H0 es verdadera o falsa, cada vez que rechacemos H0 debemos recordar que hay una probabilidad α de haber tomado una decisión incorrecta.

Existe otro tipo de error, al que llamaremos Error de Tipo II (ETII) y sucederá cuando aceptemos H0 siendo falsa. Es decir a partir de la muestra obtenida no rechazamos la hipótesis nula, pero bien podría suceder que esa muestra obtenida perteneciera a otra distribución y no a la que estamos "suponiendo" en la hipótesis nula (la media poblacional bien podría ser otra distinta de 7.30)

Cuadro resumen

	Decisión

Alternativas	Acepta H0	Rechazar H0
H0 Verdadera	Decisión Correcta. Probabilidad (1-α)	Error de Tipo I. Probabilidad α
H0 Falsa	Error de tipo II. Probabilidad ß	Decisión Correcta: Probabilidad (1 – ß). Potencia de la Dócima

Ejercitación

1. Se ha obtenido una muestra de 25 alumnos de una Facultad para estimar la calificación media de los expedientes de los alumnos en la Facultad. Se sabe por otros cursos que la desviación típica de las puntuaciones en dicha Facultad es de 2.01 puntos. La media de la muestra fue de 4.9. Se pide
    a. Intervalo de confianza al 90 %.
    b. Intervalo de confianza al 99 %.
2. El peso (en gramos) de las cajas de cereales de una determinada marca sigue una distribución). N(μ, 5). Se han tomado los pesos de 16 cajas seleccionadas aleatoriamente, y los resultados obtenidos han sido: 506, 508, 499, 503, 504, 510, 497, 512, 514, 505, 493, 496, 506, 502, 509, 496.
    a. Obtener los intervalos de confianza del 90%, 95% y 99% para la media poblacional.
    b. Suponiendo que la varianza es desconocida, calcule los mismos intervalos
3. Una muestra aleatoria extraída de una población normal de varianza 100, presenta una media muestral x = 160. Con una muestra de tamaño 144, se pide:
    a. Calcular un intervalo de confianza del 95 por ciento para la media poblacional.

b. Calcular un intervalo de confianza del 90 por ciento para la media poblacional.
c. Comparar ambos intervalos, desde el punto de vista de la información que generan

4. Los visitantes del parque nacional en un mes, medida a través de una muestra aleatoria durante 10 días elegidos al azar, han sido los siguientes: 682, 553, 555, 666, 657, 649, 522, 568, 700, 552
Suponiendo que los niveles de afluencia siguen una distribución normal.

   a. Construya un intervalo con el 95% de confianza
   b. Cuál sería el intervalo para un nivel de confianza del 99%?

5. Se desea estimar con un nivel de confianza del 95 % la altura media de los hombres de 18 o más años de un país. Suponiendo que la desviación típica de estas, en la población vale 4,

   a. Obtenga un intervalo de confianza del 95 % con una muestra de n=15 hombres seleccionados al azar, cuyas alturas son: 167 167 168 168 168 169 171 172 173 175 175 175 177 182 195

6. Y si no conociéramos la desviación típica poblacional?
Una empresa está implementado un programa de capacitación para

sus empleados. Se decide capacitar a 15 empleados. En la siguiente tabla se observan los tiempos de capacitación de cada uno

Tiempos de capacitación					
Empleado	Tiempo	Empleado	Tiempo	Empleado	Tiempo
1	52	6	59	11	54
2	44	7	50	12	58
3	55	8	54	13	60
4	44	9	62	14	62
5	45	10	46	15	63

a. Calcular el intervalo de confianza al 95% para la media Poblacional

7. Se realiza una encuesta sobre el nivel de conocimientos generales de los estudiantes de secundario de la ciudad de Córdoba, se eligen al azar a 9 estudiantes con las siguientes calificaciones

7.8	6.5	5.4	7.1	5	8.3	5.6	6.6	6.2

a. Estime un intervalo de confianza del 98 % para la media poblacional

Ejercicios de prueba de Hipótesis

1. Se sabe que la desviación típica de las notas de cierto examen de Matemáticas es 2,4. Para una muestra de 36 estudiantes se obtuvo una nota media de 5,6. ¿Sirven estos datos para confirmar la hipótesis de que la nota media del examen fue de 6, con un nivel de confianza del 95%?

2. La duración de las lámparas de 100 watt que fabrica una empresa sigue una distribución normal con una desviación de 120 horas. Su vida media está garantizada durante un mínimo de 800 horas. Se escoge al azar una muestra de 50 lámparas de un lote de producción y, después de comprobarlas, se obtiene una vida media de 750 horas. Con un nivel de significación de 0,01, ¿habría que rechazar el lote por no cumplir la garantía?

3. Un fabricante de lámparas eléctricas está ensayando un nuevo método de producción que se considerará aceptable si las lámparas obtenidas por este método dan lugar a una población normal de duración media 2400 horas, con una desviación típica igual a 300. Se toma una muestra de 100 lámparas producidas por este método y esta muestra tiene una duración media de 2320 horas. ¿Se puede aceptarr la hipótesis de validez del nuevo proceso de fabricación con un riesgo igual o menor al 5%?

4. Ante reiterados reclamos de algunos clientes de un supermercado se decide realizar una muestra aleatoria de 12 sobres de café de los que están a la venta. Se encuentra que el peso promedio de estos sobres es de 15.97 grs y una desviación estándar de esa muestra de 0.15 grs. Los empacadores afirman que el peso neto promedio del café es de 16 grs por sobre. ¿Puede rechazarse esta afirmación con un nivel de significancia de 10%?

5. Una cadena de restaurant afirma que el tiempo promedio de espera de clientes es de 3 minutos en promedio y con una desviación estándar de 1 minuto. El departamento de calidad obtuvo una muestra de 50 clientes y obtuvo que el tiempo medio de espera de cada cliente es de 2.75 minutos. Con un nivel de significación de

0.05, se puede concluir que el tiempo promedio de espera es menor a 3 minutos?