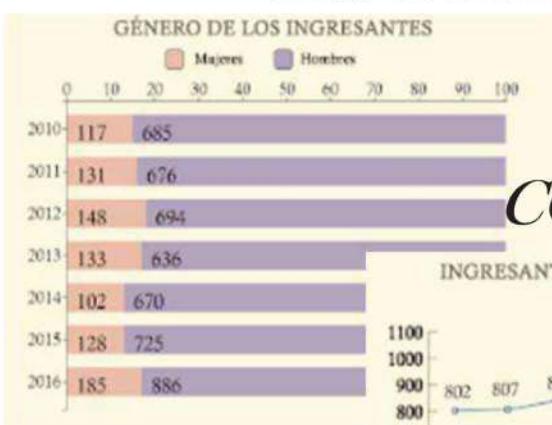


# *PROBABILIDADES Y ESTADÍSTICAS*



## *CONCEPTOS BÁSICOS*



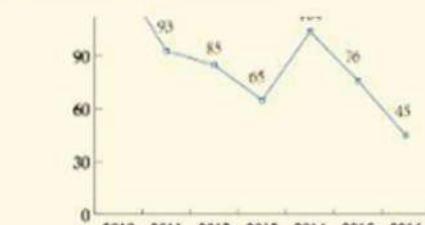
**26% MÁS INGRESANTES  
 EN 2016 QUE EN 2015.**

Es la mayor cantidad en los últimos 18 años.

**CASI UN QUINTO DE LOS  
 INGRESANTES ES MUJER.**



Sin embargo hasta ahora el 14% de los ingresantes 2017 son mujeres.



**LA MITAD DE LOS  
 PRIMERIZOS REGRESA.**

El 47% de los ingresantes 2015 volvieron este año.

**AÑO 2015**

*Prof. Cdra. Gladys M. Rouadi*



*La Estadística constituye una disciplina científica que trata de la selección, análisis y uso de datos con el fin de resolver problemas. A toda persona, tanto en su ejercicio profesional como en su actividad diaria en contacto con diferentes medios, se le ofrece información en forma de datos. Consecuentemente, algunos conocimientos de Estadística le serán de utilidad a la población en general, pero en particular en conocimiento estadístico será de vital importancia para ingenieros de todas las especialidades, científicos y administradores, debido a que manejan y analizan datos cotidianamente. En consecuencia las herramientas básicas de la Estadística les resultan de gran importancia a la hora del ejercicio profesional.*

*Las aplicaciones de la Probabilidad y la Estadística son numerosas en todos los casos de la ciencia aplicada en donde existan variaciones y donde las conclusiones acerca de un sistema están basadas en datos observados.*

*Por Estadística y Probabilidad se entiende los métodos para describir y modelar la variabilidad, además de permitir la toma de decisiones cuando la variabilidad está presente.*

*Del diseño a la producción, los procesos tienen que ser permanentemente mejorados. Con sus conocimientos técnicos y dotados de habilidades estadísticas básicas para la recolección y representación gráfica de datos, ingenieros y científicos podrán desenvolverse eficientemente. Agradezco a los integrantes de la Cátedra y a los alumnos que colaboraron en la detección de errores que permiten año tras año mejorar al presente material.*

---

*Probabilidades y Estadística: conceptos básicos. 1º ed.-Córdoba. ROUADI, Gladys Margarita. Eudecor. 2013. ISBN 978-987-1536-38-2. Fecha de catalogación: 12/04/2013*



## Unidad N° 8: Teoría del Muestreo

### Objetivos Específicos

Que el estudiante:

**Visualice la necesidad y ventajas del muestreo.**

**Identifique los distintos tipos de muestreo.**

**Conozca cómo se selecciona una muestra.**

**Interprete a través de las distribuciones por muestreo los errores, la probabilidad de cometerlos y la importancia de la variabilidad de los estadísticos.**

**Conozca y aplique la Ley de los Grandes Números y el Teorema Central del Límite.**

### Contenidos

**1. Generalidades.**

**2. Razones para el muestreo.**

**3. Base teórica del muestreo.**

**4. Procedimientos para la selección de muestras.**

**4.1. Generalidades.**

**4.2. Muestreo no probabilístico.**

**4.2.1. Características.**

**4.2.2. Muestreo de criterio.**

**4.2.3. Muestreo de la muestra disponible.**

**4.2.4. Muestreo por cuotas.**

**4.3. Muestreo probabilístico.**

**4.3.1. Características.**

**4.3.2. Muestreo aleatorio simple.**

**4.3.3. Muestreo aleatorio estratificado.**

**4.3.4. Muestreo sistemático.**

**4.3.5. Muestreo por conglomerados.**

**5. Distribuciones en el muestreo.**

**5.1. Distribución por muestreo de la media muestral.**

**5.1.1. Muestreo con reposición.**

**5.1.2. Muestreo sin reposición.**

**5.2. Distribución por muestreo de la proporción muestral.**

**5.2.1. Muestreo con reposición.**

**5.2.2. Muestreo sin reposición.**

**5.3. Distribución por muestreo de la varianza muestral corregida.**

**5.3.1. Muestreo con reposición.**

**5.3.2. Muestreo sin reposición.**

**6. Ley de los Grandes Números.**

**7. Teorema Central del Límite.**

**8. Parámetros y estadísticas para variables y parámetros para variables aleatorias.**



## **1. GENERALIDADES**

Hemos mencionado con anterioridad el concepto de POBLACIÓN y MUESTRA. Pero recordemos nuevamente sus conceptos:

### ***Población, Universo o Colectivo***

Es el conjunto de individuos, sean personas o cosas, sobre las cuales se desea información, o de las cuales se estudia alguna característica. Son las unidades estadísticas, poseedoras de la característica bajo estudio.

Así, los alumnos de un curso, constituyen un conjunto de individuos, pero esto solo, no configura una población en sentido estadístico, pues es necesario definir, además, alguna o algunas características que nos interesan analizar de ese conjunto. Por ejemplo, el peso, la estatura, la edad, etc.; o bien medir en forma simultánea peso y altura; peso y edad; peso, altura y edad; etc.

Entonces, una población en sentido estadístico queda configurada por el conjunto de individuos, sean personas o cosas, acompañadas por una o más características que se miden o cuentan en cada una de ellas.

### ***Muestra***

Es una parte de la Población bajo estudio, convenientemente seleccionada, con el objetivo de obtener conclusiones, tomar decisiones o realizar predicciones válidas sobre el comportamiento de la totalidad de la Población de la cual fue extraída, en relación a la característica o a las características estudiadas.

Debe quedar claro que nuestro propósito es siempre el conocimiento del comportamiento de la característica o de las características estudiadas en la Población.

Pero si ésta no puede examinarse en su totalidad, recurrimos a una parte de ella, llamada Muestra, arribando de esta manera a la *Teoría Del Muestreo*. Esta Teoría nos brindará y fijará pautas y enseñanzas necesarias para un correcto uso de las muestras y los procedimientos para su selección, a los fines de que las mismas sean representativas y en consecuencia válidas para lograr el objetivo planteado para con la población.

Las técnicas y procedimientos que se utilizan para inferir en forma válida desde la muestra a toda la población corresponden a la llamada *Inferencia Estadística*.

Dos aspectos importantes de la Inferencia Estadística son:

- 1- *Estimación de Parámetros* de Población, partiendo de estimadores muestrales (o estadísticos) y calculando la precisión de la estimación (error de estimación).
- 2- *Docimásia o Prueba de Hipótesis*, que consiste en la utilización de las muestras, para verificar si un determinado supuesto sobre la Población es verdadero o falso, midiendo los riesgos de cometer un error.

Recuérdese que toda medida calculada en base a datos poblacionales, se llama *Parámetro*, y toda medida calculada en base a datos de muestras, se llama *Estadístico*,



*Estadística O Estadígrafo.* Entonces, los datos de muestra pueden usarse para *Estimar Características De La Población Con Determinados Niveles De Confianza y Error (Estimación Estadística)*, o para *Tomar Decisiones, Tales Como, Aceptar o Rechazar Valores Poblacionales Supuestos, Midiendo Los Riesgos De Cometer Un Error De Aceptarlo Siendo Falsos o De Rechazarlos Siendo Ciertos (Docimásia De Hipótesis)*.

## **2. RAZONES PARA EL MUESTREO.**

Las poblaciones que se analizan pueden ser infinitas, y en tales situaciones el muestreo es el único procedimiento posible.

Hay casos en los que la captación del dato requiere de la destrucción de los elementos, por lo cual un censo implicaría la destrucción de la población.

Por ejemplo, si un fabricante desea probar la calidad de sus lámparas sujetándolas a pruebas de duración, esto es, encendiéndolas, deberá recurrir indefectiblemente a un muestreo, pues de lo contrario destruiría su producción.

En general, para muchos tipos de datos, la población no es accesible.

Si la población es finita, enfrentamos la alternativa de obtener información completa, mediante un examen del 100% de todos los elementos de la población, es decir, realizar un censo (con excepción de poblaciones que siendo finitas, constan de millones de elementos que hacen imposible su enumeración completa, por lo que el único procedimiento práctico es el muestreo, pues, de realizarse un censo, los costos de localizar, visitar y entrevistar serían prohibitivos) o recurrir al muestreo, que es lo aconsejable.

Cabe preguntar: ¿Por qué es aconsejable utilizar una muestra que proporciona sólo información incompleta acerca de la población, cuando un censo proporcionaría información completa?

Para responder analizaremos las siguientes razones:

### ***Mayor Exactitud.***

Aunque financiera, práctica y físicamente sea posible observar a toda la población, el muestreo es más eficiente debido a una mayor exactitud. Dicho de otra manera, los resultados obtenidos a través del muestreo, son casi tan precisos o en ciertos casos, más precisos, que los obtenidos mediante un censo.

Cualquier encuesta estadística, utilizada para la captación del dato (muestreo o censo) siempre contiene cierto error. Los errores estadísticos son de dos clases:

a- *Errores de observación, o de no muestreo*, es decir errores ajenos al muestreo:

- Falta de respuesta de algunas unidades seleccionadas en la muestra. Esto puede suceder por omisión, por fracaso en la localización de algunas unidades, o por renuencia de algunos individuos a contestar las preguntas de la encuesta.



- Errores de medición en alguna unidad. Puede existir inexactitud en algún aparato de medición, la persona entrevistada puede no conocer la respuesta o falsearla, etc.
- Errores introducidos en la toma del dato, en la codificación, tabulación y análisis de los resultados.

b- *Errores de estimación ó de muestreo:*

- Son el resultado de la elección casual de unidades de muestreo y ocurren cuando trabajamos con una parte del conjunto. En un censo desaparecen.

Un error de estimación, es la diferencia entre el resultado de la muestra y el del censo, cuando ambos resultados se obtienen usando los mismos procedimientos.

En otras palabras, al calcular los parámetros de la población en base a estimaciones muestrales, se comete un error de estimación.

Lo importante aquí, a diferencia de lo que sucede con los errores de observación, es que la Estadística proporciona las técnicas necesarias para medir estos errores.

Entonces, no sólo puede esperarse que el error total sea menor en un estudio de muestreo, sino que los resultados de ella, también pueden ser usados con un mayor grado de confianza, por nuestro conocimiento del tamaño probable del error.

En resumen, el resultado final, es más exacto cuando proviene de una muestra, que cuando proviene de un censo, pues son mayores los errores de observación en el censo, que los errores de estimación y observación juntos, en el muestreo.

Por otro lado, los errores de observación no pueden medirse con la precisión con que se miden los errores de estimación.

**Costo.**

Una muestra, generalmente, es menos costosa que un censo. El costo como argumento a favor del muestreo está basado en que puede proporcionar datos con la suficiente precisión y a un costo mucho más bajo que un censo. El costo puede reducirse mientras no se perjudique la precisión que se desea. Es decir, no solo debemos elegir el muestreo considerando su menor costo, si con esto sacrificamos la calidad (precisión) de los resultados.

**Tiempo.**

Otra ventaja de una muestra sobre un censo, es que la primera produce en general, información con mucha más rapidez, fundamentalmente en problemas en los cuales éste es esencial.

La rapidez se debe a dos razones principales:

- a- Extraer una muestra, requiere menor tiempo que levantar un censo, ya que es una tarea a menor escala.



b- La corrección, codificación y tabulación de los resultados insume menos tiempo y cualquier encuesta proporciona información útil, sólo después que los datos han sido recopilados y tabulados.

La información debe llegar en tiempo oportuno, pasado el cual no es de utilidad.

### **3. BASE TEÓRICA DEL MUESTREO.**

Los datos estadísticos poseen dos importantes características:

- a- Diversidad.
- b- Regularidad o uniformidad.

Analizaremos a cada una de ellas y veremos que son fundamento para inferir sobre la población en base a datos muestrales.

#### ***Diversidad***

Las unidades elementales de cualquier población son afectadas por una multiplicidad de fuerzas que, aunque relacionadas, actúan sobre los elementos individuales con un considerable grado de independencia.

Estas causas explican las variaciones de una unidad a otra en la población.

Así, naranjas del mismo árbol pueden diferir en tamaño, color, peso, o dulzura.

Aunque la diversidad es una cualidad universal de los datos, no hay ninguna población estadística cuyos elementos varíen entre sí, sin límite. Así, siguiendo con las naranjas, las mismas varían en un grado limitado en tamaño, color, peso o dulzura, pero siempre serán identificadas como naranjas.

El hecho de que cualquier población tiene propiedades características y que las variaciones en sus elementos son limitadas, hacen posible que elijamos una muestra relativamente pequeña y al azar, que puede reflejar bastante bien las características de la población.

#### ***Regularidad o Uniformidad***

Las fuerzas relacionadas, pero independientes, que producen variabilidad en una población, están a menudo tan equilibradas y concentradas que tienden a generar iguales valores por arriba y por debajo de cierto valor central, alrededor del cual tienden a agruparse la mayor parte de los valores.

Por ejemplo, los saldos de las cuentas de caja de ahorro en cualquier banco, pueden variar desde \$1.000 hasta más de \$1.000.000 y la mayoría se hallarán en un lugar intermedio entre esos dos valores.

Así, los elementos individuales de una población tienden a variar entre sí y al mismo tiempo, a adaptarse a ciertas normas. Por ello, tenemos diversidad y uniformidad en los datos.



Debido a la *Uniformidad Estadística*, si se escoge una muestra grande al azar, las características de esta muestra diferirán muy poco de las de la población.

Por la *Diversidad*, si se toman algunas muestras al azar, aunque muy similares en muchos aspectos, las muestras nunca coincidirán completamente unas con otras.

La *Uniformidad o Regularidad* es la tendencia de las características mensurables a concentrarse alrededor de una medida de tendencia central (promedio), del que las observaciones individuales divergen en cierta forma definida.

Los promedios son más estables que los valores individuales, y resultan más estables a medida que se incrementa el número de observaciones (tamaño de la muestra).

Esto se debe a que en una muestra grande, unas pocas observaciones extraordinarias, escasamente afectan a la media, porque hay muchas otras observaciones más típicas.

Pero, si la muestra es pequeña, no se presenta la oportunidad de que muchas observaciones típicas inmovilicen una observación extrema, por lo que, los promedios en muestras pequeñas, exhiben mayor variabilidad que en muestras grandes.

En la práctica, para hacer inferencias, generalmente tomamos una sola muestra. Sea ésta grande o pequeña, estamos casi seguros de que sus características no son exactamente las de la población. ¿Cómo podemos estar seguros entonces, sobre el grado de confianza de nuestras conclusiones? La respuesta es *Aleatoriedad*, o sea, se requiere que la muestra sea al azar.

En base a lo analizado y recordando lo establecido en el punto 1.3. Etapas de la Investigación Científica, como etapas 1 y 2, es decir:

- 1- Formulación o definición del problema.
- 2- Diseño del experimento.

Donde mencionamos que las investigaciones pueden consistir en Experimentos, Estudios Muestrales ó Estudios Observacionales, y que nuestro análisis se centrará en los Estudios Muestrales, debemos entonces realizar el *Diseño de Muestras* que comprende:

- 1- El Plan de Muestreo
- 2- Elección del estimador a utilizar.

Es decir, deberá definirse la población objeto de estudio, tanto en sus elementos componentes como en su probable distribución (muestreo de poblaciones finitas), así como la extensión (lugar físico donde se llevará a cabo la investigación) y tiempo (período para la recolección de la información), lo que dijimos de acotar en tiempo y espacio.

Luego, deberá identificarse el Marco Muestral, es decir el listado de todas las unidades de muestreo que pueden ser seleccionadas en alguna etapa del proceso de muestreo (Ejemplo: Guía de Teléfono, Padrón Electoral, etc.), el que deberá ser cuidadosamente analizado a fin de evitar duplicaciones o ausencias, es decir, deberá verificarse si contiene todos los elementos necesarios para evitar estimaciones sesgadas.



Inmediatamente, deberá determinarse el tamaño de la muestra, procedimiento de estimación, obtención de las estimaciones y cálculo de su precisión, que varían según el procedimiento de muestreo utilizado, para lo cual seguidamente los desarrollaremos.

#### **4. PROCEDIMIENTOS PAR LA SELECCIÓN DE MUESTRAS**

##### **4.1. Generalidades.**

Para que la Inferencia Estadística sea válida es necesario que la muestra sea representativa de la población.

Para seleccionar una muestra representativa deben considerarse dos criterios:

###### *Fiabilidad*

Es de esperar que en el muestreo haya errores, un error de muestreo, como ya dijimos, es la diferencia entre el valor de un estadígrafo, obtenido mediante una muestra aleatoria, y el valor del correspondiente parámetro de población, debido a variaciones fortuitas en la selección de las unidades elementales.

Se mide, por lo que se llama *fiabilidad o precisión* del muestreo, que está relacionada con la varianza del estadígrafo. Cuanto mayor la varianza, menor la fiabilidad del resultado de la muestra, según demostraremos en desarrollos posteriores.

###### *Efectividad*

El criterio de efectividad está asociado al costo del muestreo. Un diseño de muestreo se considera efectivo, si se obtiene el mismo grado de fiabilidad al menor costo posible. Un diseño de muestreo, se considera más efectivo que otro, si el primero tiene menor costo que el segundo, dentro del mismo grado de fiabilidad.

Veamos ahora los tipos de procedimientos que pueden utilizarse:

##### ***4.2. Muestreo No Probabilístico.***

###### ***4.2.1. Características***

Las unidades de la población que integrarán la muestra se eligen según el criterio o juicio del investigador, por lo que no permite conocer:

- La probabilidad que tiene la muestra de ser seleccionada.
- El error de muestreo, ni su evaluación en términos de probabilidad (confianza o riesgo).
- Precisión del estimador.

Es decir, que la elección de los elementos de la muestra se realiza de manera casual. No hay aleatoriedad y no pueden realizarse Generalizaciones o Inferencias en relación a la población de la cual la muestra fue extraída.



#### **4.2.2. Muestreo de Criterio**

También llamado Muestreo Intencional, Muestreo por Juicio o Purposive Sampling.

Es una muestra en la que el criterio o juicio experto del investigador desempeña un papel fundamental en la selección de objetos que se han de incluir en la muestra y/o al tomar decisiones respecto a las partes de la población para las que la muestra nos proporciona información. Si la muestra con base a un criterio determinado puede proporcionar resultados útiles, no existen métodos disponibles para estimar el error de muestreo y por consiguiente la precisión, la que dependerá de la solidez del criterio ejercido, pero no podrá evaluarse sobre la base de la muestra misma.

Ejemplo: Para estudiar la preferencia por algún producto alimenticio a base de pescado, se elige una región donde el consumo de pescado está muy arraigado.

#### **4.2.2. Muestreo de la Muestra Disponible**

También llamado Muestreo a la Mano o Muestreo por Conveniencia.

La muestra queda constituida por una parte de la población que se encuentra convenientemente disponible. Pueden ser útiles para propósitos limitados, pero no pueden proporcionar la seguridad de que los resultados obtenidos sean indicativos de las características de la población completa bajo estudio. Las conclusiones pueden contener un error considerable.

Ejemplo: Seleccionar personas en la caja de un supermercado y entrevistarlas para conocer su opinión sobre un producto determinado.

#### **4.2.3. Muestreo por Cuotas**

(Tomado del material de Estudio del Curso de Postgrado “Estadística Aplicada a la Investigación”. U.N.C. Facultad de Ciencias Económicas. Est. Nidia Blanch-Est. Silvia Joeckes. Modulo X- Pág. 31/35.)

Es un caso especial del Muestreo por Juicio y se utiliza en encuestas de opinión, investigaciones de mercado, etc.

El investigador establece pasos explícitos para obtener una muestra que sea similar a la población objetivo, ejerciendo ciertos “controles” sobre algunas características de sus elementos. Se estiman los tamaños de subconjuntos de la población en base a datos de un listado, un censo, etc., a partir de lo cual se calculan proporcionalmente “cuotas” ó número deseado de observaciones muestrales con respecto a los subconjuntos de la población. Los encuestadores hacen lo que pueden por encontrar personas que satisfagan las restricciones de sus controles de cuotas.

#### Ejemplo

Supongamos, por ejemplo, que un investigador desea realizar una encuesta para medir rating televisivos. Le interesa en particular evaluar la respuesta de acuerdo a subconjuntos de la población determinados por edad, educación e ingreso de los encuestados.



Luego, las características que desea controlar son:

Edad. 4 categorías: Menos de 20    21-30    31-50    Más de 50

Nivel Educativo Alcanzado. 3 categorías: Primario    Secundario    Universitario

Ingresos. 3 categorías: Menos de 500    500-1000    Más de 1000

Estas características pueden resumirse en la siguiente tabla:

Edad	Nivel Educativo	Ingresos		
		< 500	500 - 1000	> 1000
< 20	Primario			
	Secundario			
	Universitario			
21 - 30	Primario			
	Secundario			
	Universitario			
31 - 50	Primario			
	Secundario			
	Universitario			
> 50	Primario			
	Secundario			
	Universitario			

La muestra, entonces, tendrá que estar conformada por personas menores de 20 años con nivel de estudios primarios y que ganan menos de \$500; menores de 20 años con estudios secundarios que ganen menos de \$500, etc.

En total, deberá tener  $4 \times 3 \times 3 = 36$  grupos de personas bien diferenciados, para controlar las tres características que ha establecido como relevantes por su influencia sobre el rating televisivo.

Para obtener adecuadamente su muestra, el investigador deberá contar con información previa respecto a la proporción de individuos de la población que componen cada uno de estos grupos. Ello es posible, por ejemplo, a partir de datos provenientes de un Censo Poblacional.

Pero, en otros casos, tal descripción de la población puede ser extremadamente difícil, o a veces, imposible de encontrar.

Si el investigador posee la información, puede determinar el tamaño que deberá tener la muestra en cada uno de los grupos efectuando el siguiente cálculo:

Tamaño total de la muestra x Proporción del grupo.

Luego, si deseamos seleccionar una muestra de 1000 personas y se conoce por algún censo de población que la proporción de menores de 20 años con estudios primarios que ganan menos de \$500 es del 10%, tendríamos:

$$1000 \times 0,10 = 100$$



El investigador debe, entonces, buscar 100 personas que reúnan las 3 características mencionadas. Este mismo procedimiento tiene luego que ser repetido en cada uno de los grupos preestablecidos. La manera de encontrar a estas personas queda ligada, generalmente, al juicio del investigador.

Los problemas que presenta el Muestreo por Cuotas son varios, a saber:

- Antes de comenzar el procedimiento de selección hay que contar con una proporción bastante aproximada de individuos que componen cada grupo, lo cual resulta a veces imposible.

- Se deben seleccionar todas las características que estén relacionadas con la información que queremos obtener a partir de la muestra. Por ejemplo, si queremos averiguar la actitud de la gente hacia el uso del pelo largo en los varones, evidentemente la edad deberá ser una característica a tener en cuenta.

Si se deja de lado alguna característica importante, los resultados muestrales pueden ser completamente erróneos.

- Cuando se establecen características de control múltiples se crean numerosos grupos y, a veces, resulta imposible ubicar una cierta cantidad de personas que reúnan todas las características.

### 4.3. Muestreo Probabilístico

#### 4.3.1. Características

La selección de unidades que integrarán la muestra se realiza utilizando las propiedades proporcionadas por la Teoría de Probabilidades, por lo que permite conocer:

- La probabilidad de la muestra seleccionada.
- El error de muestreo y su evaluación en términos de probabilidad (Confianza ó Riesgo).
- La precisión del estimador.

Permite realizar inferencias sobre la población de la cual fue extraída la muestra por medio de métodos estadísticos al introducir la aleatorización en el procedimiento de selección.

Nótese que:

- No permite discreción acerca de los elementos que, de la población, deberán incluirse en la muestra.
- Una vez que un elemento ha sido seleccionado, requiere que deba incluirse en la muestra, sin permitir sustitución.

#### 4.3.2. Muestreo Aleatorio Simple.

Es el procedimiento por el cual se selecciona una muestra ( $n$  unidades) de una población ( $N$  elementos), de manera tal que cada una de las muestras posibles del mismo tamaño tengan la misma probabilidad de ser seleccionadas, es decir, que las probabilidades de todas las muestras posibles del mismo tamaño ( $n$ ) que se pueden obtener de una población de  $N$  elementos, son iguales. Esto equivale a decir, que cada



uno de los elementos de la población tiene la misma probabilidad de ser incluido en la muestra, lo cual implica la aplicación de técnicas de azar.

#### Ejemplo:

Si se desea conocer las opiniones de los estudiantes de una gran escuela, no se puede tener una muestra aleatoria con solo entrevistar a los estudiantes a la entrada de la cafetería hasta obtener el número de entrevistas que se deseé.

Muchos estudiantes pueden no entrar jamás en la cafetería y así entren, la probabilidad de entrevistarlos es nula.

Una muestra así obtenida no es aleatoria.

Uno de los métodos comúnmente utilizados para lograr una muestra aleatoria es enumerar todos los elementos de la población definida, escribir los números en tarjetas, fichas o bolillas, ponerlos en una bolsa, mezclarlos y extraer los elementos según el tamaño de muestra determinado.

Si se toma con reemplazo, una vez analizado el elemento, vuelve a la población, se mezclan nuevamente los objetos y se realiza la extracción siguiente. Esto implica que la probabilidad permanezca invariable para cada objeto y la probabilidad de cada muestra es igual a  $\frac{1}{N^n}$ .

Si se toma sin reemplazo, la probabilidad de cada objeto se hace mayor en cada nueva extracción, pues el número disponible para elegir, se reduce en el número de objetos ya sacados, pero los objetos restantes tienen idéntica probabilidad de ser seleccionados. La probabilidad para cada muestra se obtiene como  $\frac{1}{C_N^n}$ .

Este procedimiento se simplifica haciendo uso de la Tabla de Números Aleatorios.

El Muestreo Aleatorio Simple es de aplicación en poblaciones pequeñas y homogéneas (los elementos se comportan de manera homogénea en relación a la característica bajo estudio) y requiere la identificación de todos los elementos de la población.

A continuación ejemplificaremos el uso de la Tabla de Números Aleatorios y realizaremos la estimación de  $\mu$ , a través de su mejor estimador  $\bar{x}$  considerando tres muestras de entre el total de muestras posibles.

#### Ejemplo del uso de la Tabla de Números Aleatorios

Dada la siguiente información, correspondiente al total de fichas con saldo deudor existentes en una empresa y clasificadas en 8 intervalos:



	Saldo deudor					
	$y'_{i+1} - y'_i$	$y_i$	$n_i$	$y_i n_i$	$N_i$	
1	100 - 150	125	322	40.250	322	(322 deudores con saldo menor a 150)
2	150 - 200	175	145	25.375	467	(467 deudores con saldo menor a 200, de los cuales 322 tiene saldo menor a 150)
3	200 - 250	225	115	25.875	582	
4	250 - 300	275	114	31.350	696	
5	300 - 350	325	104	33.800	800	
6	350 - 400	375	91	34.125	891	
7	400 - 450	425	85	36.125	976	
8	450 - 500	475	24	11.400	1.000	
			1.000	238.300		

Se pide realizar la estimación de  $\mu$ , para lo cual se deberá:

1- Seleccionar tres muestras de tamaño cinco con reposición, considerando:

$$m: \text{cantidad de muestras} = 3$$

$$n: \text{tamaño de cada muestra} = 5$$

2- Comparar  $\bar{x}$  con  $\mu$ .

### Resolución:

En primer lugar, calculamos

$$\mu = \frac{\sum_{i=1}^N y_i n_i}{N} = \frac{238.300}{1.000} = 238,30$$

Recuérdese que cuando tenemos un intervalo, tomamos como valor de la variable el punto medio o marca de clase, que es la media aritmética de los extremos de cada intervalo, es decir:

$$\frac{y'_{i-1} + y'_i}{2}$$

Luego, construimos las tres muestras y calculamos sus medias. Para ello necesitamos elegir cinco números aleatorios de los dígitos de la tabla para cada muestra, por lo que en total necesitamos 15 números aleatorios.

La regla de decisión queda planteada en función de las frecuencias absolutas acumuladas:



Ó

1	001 - 322	000 - 321	<i>Tomamos un individuo del 1º intervalo.</i>
2	323 - 467	322 - 466	<i>Tomamos un individuo del 2º intervalo.</i>
3	468 - 582	467 - 581	<i>Tomamos un individuo del 3º intervalo.</i>
4	583 - 696	582 - 695	<i>Tomamos un individuo del 4º intervalo.</i>
5	697 - 800	696 - 799	<i>Tomamos un individuo del 5º intervalo.</i>
6	801 - 891	800 - 890	<i>Tomamos un individuo del 6º intervalo.</i>
7	892 - 976	891 - 975	<i>Tomamos un individuo del 7º intervalo.</i>
8	977 - 000	976 - 999	<i>Tomamos un individuo del 8º intervalo.</i>

Además, elegimos al azar los números aleatorios (que en este caso representan el número de fichas de deudores varios). Una vez elegido el punto de comienzo en la tabla, extraemos los dígitos siguiendo una columna hasta llegar al final de la página, luego comenzamos con la columna siguiente siempre de arriba abajo, y así hasta completar el número de dígitos a elegir.

En nuestro caso, elegimos quince números de tres dígitos cada uno, todos seguidamente.

Los cinco primeros, constituirán la primera muestra, los cinco segundos la segunda y así sucesivamente:

FICHA Nº	$y_i$	FICHA Nº	$y_i$	FICHA Nº	$y_i$
100	125	660	275	985	475
375	175	310	125	118	125
084	125	852	375	834	375
990	475	635	275	886	375
128	125	737	325	995	475
$\Sigma$	1.025		1.375		1.825

Estos dígitos representan la “ficha nº”, pero, qué saldo tiene esa ficha?

Puesto que nuestra media es promedio de saldos, no de fichas, nos fijamos a qué intervalo corresponde ese número de ficha y vemos cuál es el saldo representativo para ese intervalo, dado por  $Y$ .

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$\bar{y}_1 = \frac{1025}{5} = 205 \quad \bar{y}_2 = \frac{1375}{5} = 275 \quad \bar{y}_3 = \frac{1825}{5} = 365$$

La diferencia con  $\mu$ , se debe a que mientras la media poblacional es un parámetro cuyo valor es constante en una población determinada, la media muestral es una variable



aleatoria, cuyo valor puede variar dependiendo de las observaciones muestrales que forman parte de la muestra elegida.

En el caso planteado, se obtuvieron muestras a los fines de estimar el parámetro Media Poblacional ( $\mu$ ), a través de su mejor estimador, la Media Muestral ( $\bar{x}$ ).

Podemos interesarnos en estimar el parámetro Proporción Poblacional ( $P$ ), a través de su estimador Proporción Muestral ( $\hat{P}$ ), o bien el parámetro Varianza Poblacional ( $\sigma^2$ ) ó Varianza Poblacional Corregida ( $s^2$ ), ó sus correspondientes desviaciones estándar ( $\sigma$  ó  $S$  respectivamente), recurriendo al valor muestral correspondiente ( $s^2$  ó  $\hat{s}$ ).

En relación al análisis de la precisión de los estimadores requerimos calcular las desviaciones estándares. Así,  $\sigma_{\bar{x}}$  es el error estándar de la media muestral,  $\sigma_{\hat{P}}$  es el error de la proporción muestral, etc., y nos permitirán:

- Estimar la precisión de la estimación.
- Estimar el tamaño de la muestra.
- Comparar la precisión de los estimadores, obtenida por medio de distintos métodos de muestreo.

Para determinar el tamaño de la muestra, se hace necesario especificar el error de estimación que se desea y el nivel de confianza asociado, es decir, la evaluación del error de muestreo en términos de probabilidad.

En el siguiente tema a tratar: *Distribuciones En El Muestreo*, aprenderemos a calcular errores estándares, es decir, las desviaciones de los estimadores. En la Bolilla sobre Estimación Estadística, aprenderemos la determinación del tamaño de la muestra.

#### **4.3.2. Muestreo Aleatorio Estratificado**

Se utiliza especialmente:

- 1- Cuando la información se desea con distinta precisión para algunas subdivisiones de la población.
- 2- Cuando los elementos difieren sensiblemente en los distintos estratos de la población, o sea cuando la población no es homogénea, es decir, sus elementos se comportan de manera heterogénea en relación a la característica bajo estudio.
- 3- Cuando los problemas de muestreo que se presentan en cada parte son distintos y necesitan un tratamiento diferencial.

#### **Procedimiento**

La población se divide en cierto número de grupos, llamados estratos ( $r$  grupos de  $N_i$  individuos cada uno), mutuamente excluyentes y colectivamente exhaustivos, es decir, no pueden haber sobre posiciones, ni omisiones.

O sea que: 
$$N = N_1 + N_2 + N_3 + \dots + N_r$$



Una vez establecidos los estratos, se extrae de cada uno de manera independiente una muestra de tamaño  $n_i$ , siendo la muestra total  $n$ , igual a la suma de los  $n_i$ .

O sea que:  $n = n_1 + n_2 + n_3 + \dots + n_r$

Debe tratarse de que los estratos tengan mayor homogeneidad que en la población total, lo que permitirá una mejor estimación (más precisa) de los parámetros, utilizando muestras relativamente más pequeñas.

En la muestra obtenida en cada estrato se calcula el estadístico de interés. Estas medidas por estrato se ponderan adecuadamente para formar una estimación combinada de la población completa.

### Ejemplo

Supongamos que con una muestra de 5 alumnos queremos estimar el promedio de edad de un grupo de 50 alumnos, de 1º a 5º año, estratificados en la siguiente forma:

$$\begin{aligned} N_1: & \text{alumnos de 1º, 2º y 3º año} = 30 \\ N_2: & \text{alumnos de 4º y 5º año} = 20 \end{aligned}$$

Se toma una muestra de  $n=5$ , donde  $n_1=3$  y  $n_2=2$ .

Así, los resultados obtenidos en la muestra son :

Estrato 1		Estrato 2	
Alumno	Edad	Alumno	Edad
32	22	05	25
37	21	17	26
41	19		

$$\begin{aligned} 1^{\circ}) \quad \bar{x} &= \frac{\sum_{i=1}^n x_i}{n} = \frac{22 + 21 + 19 + 25 + 26}{5} = \underline{22,6} \\ \hat{s}^2 &= \frac{\sum_{i=1}^n x_i^2 - \bar{x}^2}{n-1} = \frac{(484 + 441 + 361 + 625 + 676) - 5(22,6)^2}{4} = \frac{2587 - 2553,8}{4} = \underline{8,3} \end{aligned}$$

$$\hat{s} = \sqrt{\hat{s}^2} = \underline{2,88}$$

2º)

### Estrato 1

$$\begin{aligned} \bar{x}_1 &= \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1} = \frac{22 + 21 + 19}{3} = \frac{62}{3} = \underline{20,7} \\ \hat{s}_1^2 &= \frac{\sum_{i=1}^{n_1} x_{1i}^2 - n_1 \bar{x}_1^2}{n_1 - 1} = \frac{(484 + 441 + 361) - 3(20,7)^2}{2} = \frac{1286 - 1285,47}{2} = \underline{0,27} \end{aligned}$$



### Estrato 2

$$\bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2} = \frac{25+26}{2} = \underline{25,5}$$

$$\hat{s}_2^2 = \frac{\sum_{i=1}^{n_2} x_{2i}^2 - n_2 \bar{x}_2^2}{n_2 - 1} = \frac{(625+676) - 2(25,5)^2}{1} = \frac{1301 - 1300,5}{1} = \underline{0,50}$$

Resumiendo:

	Media	Varianza	Desviación
Sin estratificación	22,6	8,30	2,88
Dentro del Estrato 1	20,7	0,27	0,52
Dentro del Estrato 2	25,5	0,50	0,71

La media poblacional a estimar será

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}$$

Como la información está separada en estratos se requiere calcular un promedio ponderado, luego el estimador será el promedio ponderado de las estimaciones de los estratos individuales, donde las ponderaciones son los tamaños de cada estrato.

$$\bar{x} = \frac{\sum_{i=1}^r N_i \bar{x}_i}{\sum_{i=1}^r N_i} = \frac{(30 \times 20,7) + (20 \times 25,5)}{50} = \underline{22,62}$$

Siguiendo nuestro análisis, diremos que el Muestreo Estratificado puede ser más eficiente que el Muestreo Aleatorio Simple.

Para demostrarlo, plantearemos el siguiente ejemplo:

Pretendemos estimar la antigüedad promedio de 2000 empleados de una compañía.

Si fijamos una muestra de 50 empleados, existirán  $C_{2000}^{50}$  maneras de combinar 50 entre 2000. Así, existirán casos en los cuales todos los empleados de la muestra provengan de tal planta, o todos provengan de la oficina, o todos sean mujeres, o todos sean hombres. Tales muestras podrían considerarse “no representativas” y el hecho de que tales muestras no representativas sean posibles usando el Muestreo Aleatorio Simple, puede aumentar el error de muestreo y, por consiguiente, disminuir la precisión de la estimación a partir de la muestra.



¿Por qué estas muestras “no representativas” pueden aumentar el error de muestreo?

El punto significativo es considerar en qué aspectos estas posibles combinaciones de las muestras son no representativas.

Si son no representativas de características que no están relacionadas con el elemento bajo estudio (en este caso tiempo de servicio), el error de muestreo no incrementa.

Ejemplo: una combinación puede estar formada por 50 empleados, todos ellos fumadores.

Tal grupo no sería representativo si estuviésemos interesados en saber la proporción de empleados que fuman; sin embargo, en lo que respecta a la duración de sus servicios, el hecho de fumar no quita representación y, por lo tanto, no incrementa el error de muestreo para la estimación del tiempo de servicio promedio.

Por otra parte, la falta de representación de las muestras con respecto al lugar en que trabajan dentro de la compañía, incrementarían el error de muestreo de la estimación puesto que existe una relación definida en la compañía XX, entre el tiempo de servicios y el lugar en que trabajan los empleados (o sea en la planta, en la oficina, o en otra parte).

En consecuencia, si la muestra está formada sólo de empleados de planta o solo de oficina, puede llevar a una estimación que difiera considerablemente de la media de la población.

Tales combinaciones de muestras, que producen medias de la muestra alejadas de la media de la población, tienden a hacer más variable la distribución de muestreo de  $\bar{x}$ , incrementando así el error de muestreo.

El muestreo aleatorio estratificado puede reducir tales causas del incremento del error de muestreo, haciendo imposible tomar algunas de estas muestras “no representativas” y, en esta forma, haciendo que la distribución de muestreo de  $\bar{x}$  sea menos variable, lo que produce una mayor precisión.

Para definir los estratos se pueden emplear datos anteriores, intuición o bien resultados preliminares procedentes de otros estudios.

Es entonces, una combinación de submuestras de los estratos, que son muestras aleatorias simples o sistemáticas. En cuanto tales, todo elemento disponible de cada estrato tiene igual probabilidad de ser seleccionado.

### *Estratificación doble*

Cuando de ser posible, los estratos son subdivididos en subestratos.

### *Selección de los estratos*

Los estratos deben establecerse de forma tal que los elementos en cada estrato difieren tanto como sea posible, respecto a la característica bajo investigación de los elementos en los otros estratos, si bien dentro de cada estrato deben ser tan homogéneos



como sea posible. Entonces, debe buscarse heterogeneidad entre los estratos y homogeneidad dentro del estrato.

### Planeación del tamaño de las muestras

Es conveniente que se planee el tamaño necesario de la muestra para minimizar los costos y para maximizar la precisión.

La magnitud de la muestra se llama Afijación, y tenemos:

#### 1- Afijación Igual

Donde todos los  $n_i$  son iguales, o sea:

$$n_i = \frac{n}{r} \quad \text{Donde : } r: \text{cantidad de estratos}$$

**n:** tamaño de la muestra

#### 2- Afijación Proporcional

Cada  $n_i$  posee en la muestra la misma proporción o participación que cada  $N_i$  posee en la población, entonces:

$$n_i = \left( \frac{N_i}{N} \right) n$$

#### 3- Afijación Óptima o de Neyman

$$n_i = \left( \frac{N_i \sigma_i}{\sum N_i \sigma_i} \right) n$$

Donde:  $\sigma_i$  es la desviación estándar de los elementos en el estrato i.

Esta fórmula maximiza la precisión del estimador de la muestra.

Entonces, la razón de muestreo se hace en cada estrato proporcional a la desviación típica de ese estrato. Cuanto más homogéneo es el estrato, menor su desviación típica y menor su proporción en la muestra.

#### Ejemplo

Afijaciones					
Estrato	$N_i$	$\sigma_i$	Igual	Proporcional	Óptima
1	1.800	4,2	62	90	64
2	1.500	5,5	62	75	70
3	1.200	7,1	63	60	72
4	500	10,3	63	25	44
$\Sigma$	5.000		250	250	250



Cálculos:  
 $n = 250$

Estrato	$N_i$	$N_i \sigma_i$
1	1.800	7.560
2	1.500	8.250
3	1.200	8.520
4	500	5.150
$\Sigma$		<b>29.480</b>

1) Igual:

$$n_i = \frac{n}{r} = \frac{250}{4} = 62,5$$

2) Proporcional:

$$n_i = \left( \frac{N_i}{N} \right) n$$

$$n_1 = \left( \frac{1800}{5000} \right) 250 = 0,36 \times 250 = 90$$

$$n_2 = \left( \frac{1500}{5000} \right) 250 = 0,30 \times 250 = 75$$

$$n_3 = \left( \frac{1200}{5000} \right) 250 = 0,24 \times 250 = 60$$

$$n_4 = \left( \frac{500}{5000} \right) 250 = 0,10 \times 250 = 25$$

3) Óptima:

$$n_i = \left( \frac{N_i \delta_i}{\sum N_i \delta_i} \right) n$$

$$n_1 = \left( \frac{1.800 \times 4,2}{29.480} \right) 250 = 64$$

$$n_2 = \left( \frac{1.500 \times 5,5}{29.480} \right) 250 = 70$$

$$n_3 = \left( \frac{1.200 \times 7,1}{29.480} \right) 250 = 72$$

$$n_4 = \left( \frac{500 \times 10,3}{29.480} \right) 250 = 44$$



*Obsérvese:*

- 1) Si los estratos tienen idéntica participación en el total y las desviaciones son parecidas, la afijación igual puede aplicarse.
- 2) Si los estratos tienen diferente participación en el total y las desviaciones son parecidas, la afijación proporcional es aplicable.
- 3) Si los estratos tienen igual ó diferente participación en el total y la desviaciones son disímiles, la afijación óptima es la aplicable.

#### **4.3.3. Muestreo Sistemático**

Se selecciona un elemento de la población, cada  $k$  elementos, después de haberlos colocado en un cierto orden especificado.

El punto de partida, arranque o raíz, se selecciona al azar entre los  $k$  primeros elementos.

##### Ejemplo

Deseamos seleccionar una muestra sistemática de 50 cuadras en una comunidad que tiene 500 cuadras en la ciudad.

$$k = \frac{N}{n} = \frac{500}{50} = 10 \quad \text{Donde } k: \text{razón de muestreo}$$

Después seleccionamos un número al azar entre 1 y 10, a partir de una tabla de dígitos al azar.

Supongamos que ese número sea 3, entonces la tercera cuadra será la primera en la muestra, la segunda cuadra, será la decimotercera, la siguiente el número 23 y así sucesivamente.

Esto asegurará que una muestra sistemática contenga cuadras de todas las partes de la ciudad (ventaja sobre el muestreo aleatorio).

Ocasionalmente puede ser menos eficiente que el Muestreo Aleatorio Simple, esto es cuando la población se ordena en cierto orden periódico. Como ejemplo, supongamos que todas las cuadras de la ciudad contienen 8 casas. Un muestreo sistemático de cada octava casa podría contener solamente casas en las esquinas y éstas pueden tener características diferentes respecto a las familias de la cuadra.

Se puede emplear fácilmente cuando se dispone de una lista de las unidades de la población, como por ejemplo, una guía telefónica.

Selecciona una muestra más representativa que el Muestreo Aleatorio Simple, si los elementos cercanos de la población se asemejan más entre sí, de los que se parecen a los que quedan distantes.



Tiene la desventaja de requerir la numeración u ordenamiento de los elementos de la población, lo cual podría ser físicamente imposible si la población abarca todo un país o una zona geográfica considerable, es decir, si se trata de una gran población.

Debido a que el método nos asegura una muestra regularmente espaciada, nos asegura una representación uniforme de los elementos de la población, y permite una estimación más precisa de la media de la población, que una tomada por el Muestreo Aleatorio Simple, salvo que las unidades  $k$ -ésimas que constituyen la muestra resulten análogas o estén correlacionadas.

#### 4.3.4. Muestreo por Conglomerados

Diametralmente opuesto al muestreo por estratos está el muestreo por conglomerados, que consiste en seleccionar primero, al azar, grupos llamados conglomerados, de elementos individuales de la población, y en tomar luego todos los elementos o una submuestra de ellos, dentro de cada conglomerado, para constituir así la muestra global.

Se hacen tan pequeñas como se puedan las diferencias entre conglomerados, en tanto que las diferencias entre los elementos individuales dentro de cada conglomerado se hacen tan grandes como sea posible.

Lo ideal sería que cada conglomerado sea una miniatura de toda la población y así un solo conglomerado se llama *Unidad De Muestreo Primaria*.

Si todos los elementos o unidades elementales de cada conglomerado seleccionados se incluyen en la muestra, se llama *Muestreo De Una Etapa*. Si se saca una submuestra aleatoria de elementos de cada conglomerado seleccionado, se llama *Muestreo En Dos Etapas*. Si intervienen más de dos etapas en la obtención de la muestra global, es un *Muestreo De Etapas Múltiples*. Desde luego, los métodos aleatorios se emplean en cada etapa.

El objetivo es el estudio de las características de los elementos individuales o unidades elementales, si bien se eligen inicialmente las unidades de muestreo primarias.

*Ventajas:* reducción de costos para un grado de fiabilidad dado.

*Desventaja:* ausencia de fiabilidad para un tamaño dado de muestra. La varianza tiende a ser mayor. Pero si el costo de seleccionar una unidad elemental se reduce mucho, la misma cuantía de gastos permitirá seleccionar una muestra más grande.

Ejemplo:

Seleccionar una muestra para entrevistar a la población de Córdoba.

No existe una lista al día de los habitantes, y construirla resultaría muy costoso, además encontraríamos que las personas de la muestra estarían esparcidas por toda la provincia y el costo de enviar entrevistadores hasta dichos lugares sería alto.

Podemos utilizar el muestreo por conglomerados (áreas).



Las áreas (barrios), por ejemplo, reciben el nombre de *Unidades De Muestreo*, debido a que son las unidades que se muestrearán.

Las unidades de las cuales se obtendrá información son *Unidades Elementales*, por ejemplo los individuos.

En conclusión:

- El muestreo por áreas elimina la necesidad de formar una lista de todas las unidades elementales, pues utiliza una unidad de muestreo para la cual ya existe una lista completa o puede obtenerse con facilidad.

- Se requieren muestras más grandes que el Muestreo Aleatorio Simple, para obtener la misma precisión, debido a que las personas que viven dentro de cualquier área a menudo tienden a ser más parecidas en sus características de opiniones, que las personas que viven en áreas distintas.

#### *Selección entre muestras de probabilidad y muestras de criterio*

Si la muestra debe ser extremadamente pequeña debido a razones de costo (financieras) o de otra índole, debe preferirse una muestra de criterio a un muestra de probabilidad.

Las muestras de probabilidad deben ser consideradas cuando se requieren resultados de alta precisión, o cuando se requieren resultados objetivos y desprovistos de error sistemático, debido a que se determinarán decisiones o cursos de acción importantes sobre la base de los resultados de la muestra.

Consideraremos cualquier tipo de procedimiento de muestreo de probabilidad, que proporcione en el mismo nivel de confianza, la misma precisión en la estimación, a menor costo, o más precisión al mismo costo, como una técnica más eficiente que el Muestreo Aleatorio Simple.



## **5. DISTRIBUCIONES EN EL MUESTREO**

Hemos señalado que las estadísticas de muestra raramente se obtienen por sí mismas, sino que sirven de base para generalizar acerca de parámetros de población desconocidos.

La *Inferencia Estadística* supone métodos que nos permiten inferir de datos limitados (muestras) lo que es cierto de mayores conjuntos de datos (poblaciones).

Las conclusiones inductivas (de lo particular a lo general) pueden ser erróneas debido a la variabilidad casual en el muestreo al azar, por lo tanto, se espera generalmente que una estadística de muestra, sea diferente de su correspondiente parámetro, que es un valor desconocido, pero *Constante*; es decir, medir sólo una parte, en vez de toda la población, da por resultado cierto margen de error llamado error de muestreo (diferencia entre el valor de una estadística y el de un parámetro).

Estos errores se originan en las variaciones al azar en el valor de una estadística de muestra de una muestra a otra. Como tales, pueden ser evaluados sólo en términos de distribuciones de probabilidades de estadísticas de muestra.

Una estadística de muestra, que es calculada de una muestra al azar, es una *Variable Aleatoria* y por consiguiente tiene distribución de probabilidad propia (Recuérdese que una distribución de probabilidades se muestra en una tabla indicando todos los posibles valores para la variable y su respectiva probabilidad).

Tal distribución, se conoce como *Distribución Por Muestreo De Una Estadística*, y tiene propiedades bien definidas.

Un valor importante de estas distribuciones por muestreo, es la ayuda que nos prestan para revelar los tipos de errores de muestreo y sus magnitudes.

Veremos a continuación y a través de un ejemplo sencillo la Distribución por Muestreo de algunas estadísticas, tales como: Media Muestral, Proporción Muestral y Varianza Muestral Corregida.

Lo primero que debemos hacer es definir la población a ser muestreada.

Para ello, supongamos que nos interesa estudiar el número de materias aprobadas durante un curso lectivo determinado de los estudiantes de tercer año de la carrera de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional - Facultad Regional Córdoba.

Para simplificar el problema consideremos que hay tan sólo seis estudiantes, o sea que el tamaño poblacional ( $N$ ) es igual a 6. (Lógicamente que no usaríamos una muestra en este caso, sino que directamente recurriríamos a un censo, pero a los fines del desarrollo del tema es conveniente esta simplificación).

Entonces, la población queda definida de la siguiente manera:



Estudiantes	Nº de materias aprobadas
A	1
B	2
C	3
D	3
E	4
F	5

Calcularemos ahora, los siguientes parámetros, pues serán de utilidad para demostraciones posteriores:

- 1- Media Poblacional ( $\mu$ )
- 2- Varianza Poblacional ( $\sigma^2$ ) y Desviación Estándar Poblacional ( $\sigma$ )
- 3- Varianza Poblacional Corregida ( $s^2$ ) y Desviación Estándar Poblacional Corregida ( $s$ )
- 4- Proporción Poblacional ( $P$ ), considerando que nos interesa la proporción de estudiantes con 2 o menos materias aprobadas.

Nuestra variable, es el número de materias aprobadas, que representaremos por  $X$ . Luego:

$X$	$X^2$
1	1
2	4
3	9
3	9
4	16
5	25
<b>18</b>	<b>64</b>

Entonces:

$$1) \quad \mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{18}{6} = 3$$

$$2) \quad \sigma_X^2 = \frac{\sum_{i=1}^N X_i^2}{N} - \mu^2 = \frac{64}{6} - 3^2 = 1,666 \quad \Longrightarrow \quad \sigma_X = \sqrt{1,666} = 1,29$$

$$3) \quad S_X^2 = \frac{\sum_{i=1}^N X_i^2}{N-1} - \frac{N\mu^2}{N-1} = \frac{64}{5} - \frac{6(3^2)}{5} = 2 \quad \Longrightarrow \quad S_X = \sqrt{2} = 1,4142$$

Podríamos haber calculado a  $S_X^2$ , en función de su relación con  $\sigma_X^2$ , de la siguiente forma:

$$S_X^2 = \sigma_X^2 \cdot \frac{N}{N-1} = 1,666 \times \frac{6}{5} = 2$$



Y a  $S_x$ , en función de su relación con  $\sigma_x$ , de la siguiente manera:

$$S_x = \sigma_x \sqrt{\frac{N}{N-1}} = 1,29 \sqrt{\frac{6}{5}} = 1,4142$$

$$4) P = \frac{X}{N} = \frac{2}{6} = 0,33$$

Supongamos que se ha de extraer una muestra de 2 estudiantes ( $n = 2$ ) mediante un muestreo aleatorio simple.

A esta altura podremos preguntarnos cuántas son las muestras posibles de tamaño 2 que podemos extraer de una población de 6 individuos. La respuesta diferirá según el muestreo sea con o sin reposición.

Analizaremos en primer término el **Muestreo Con Reposición**.

Entonces, en general, la cantidad de muestras con reemplazo de tamaño  $n$  que pueden extraerse de una población de tamaño  $N$ , viene dada por los arreglos con reposición de  $N$  (tamaño poblacional) elementos tomados de  $n$  en  $n$  (tamaño muestral). Estos arreglos se simbolizan y definen como:

$${}^r A_N^n = N^n$$

Y para nuestro ejemplo:

$${}^r A_6^2 = 6^2 = 36$$

Luego podemos extraer 36 muestras.

Nos interesa ahora conocer cuáles son cada una de estas 36 muestras posibles.

Para ello es conveniente construir una tabla de doble entrada, donde representamos al fenómeno aleatorio por el par de variables aleatorias  $(x_1, x_2)$ , donde  $x_1$  es la variable aleatoria que representa todos los resultados posibles que se pueden presentar en la primera extracción y  $x_2$  es la variable aleatoria que representa todos los resultados posibles que se pueden presentar en la segunda extracción.

Lógicamente que cada par de variables aleatorias, constituye una muestra posible. Así, las 36 muestras posibles, quedan indicadas en la siguiente tabla:

$x_2$ $x_1$	1	2	3	3	4	5
1	(1,1)	(1,2)	(1,3)	(1,3)	(1,4)	(1,5)
2	(2,1)	(2,2)	(2,3)	(2,3)	(2,4)	(2,5)
3	(3,1)	(3,2)	(3,3)	(3,3)	(3,4)	(3,5)
3	(3,1)	(3,2)	(3,3)	(3,3)	(3,4)	(3,5)
4	(4,1)	(4,2)	(4,3)	(4,3)	(4,4)	(4,5)
5	(5,1)	(5,2)	(5,3)	(5,3)	(5,4)	(5,5)

TABLA A



El conjunto de toda las muestras posibles que se han confeccionado, constituyen el espacio muestral, y cada una de las muestras es un evento elemental.

Ciertamente, la probabilidad de que el par ordenado, es decir, cada uno de los eventos elementales, asuma un determinado valor es igual a:

$\frac{1}{N^2}$ , o sea que:  $P_r(x_1, x_2) = \frac{1}{36}$  para nuestro caso, siendo  $\frac{1}{N^n}$  para el caso de  $n$  extracciones, o sea,  $P_r(x_1, x_2, \dots, x_n) = \frac{1}{N^n}$ , es decir que cada una de las muestras tiene igual probabilidad de aparecer.

Ahora bien, si el muestreo fuera **Sin Reposición**, la cantidad de muestras sin reemplazo de tamaño  $n$  que pueden extraerse de una población de tamaño  $N$ , viene dada por el combinatorio de  $N$  (tamaño poblacional) elementos tomados de  $n$  en  $n$  (tamaño muestral).

Las combinaciones se simbolizan y definen como:

$$C_N^n = \frac{N!}{n!(N-n)!}$$

**NOTA:** algunos autores reemplazan  $C_N^n$  por  $\binom{N}{n}$ .

Y para nuestro ejemplo:

$$C_6^2 = \frac{6!}{2!(6-2)!} = \frac{6!}{2!4!} = \frac{6 * 5 * 4 * 3 * 2 * 1}{(2 * 1)(4 * 3 * 2 * 1)} = \frac{30}{2} = 15$$

Luego, podemos extraer 15 muestras.

Al igual que antes, nos interesa conocer cuáles son cada una de estas muestras posibles.

Para determinarlas, construimos idéntica tabla que la anterior, a la que haremos algunas salvedades.

$x_2$ $x_1$	1	2	3	$3'$	4	5
1	(1,1)	(1,2)	(1,3)	(1,3')	(1,4)	(1,5)
2	(2,1)	(2,2)	(2,3)	(2,3')	(2,4)	(2,5)
3	(3,1)	(3,2)	(3,3)	(3,3')	(3,4)	(3,5)
$3'$	(3',1)	(3',2)	(3',3)	(3',3')	(3',4)	(3',5)
4	(4,1)	(4,2)	(4,3)	(4,3')	(4,4)	(4,5)
5	(5,1)	(5,2)	(5,3)	(5,3')	(5,4)	(5,5)

TABLA B

En primer término hemos diferenciado a 3 de  $3'$ , puesto que estos valores son elementos distintos, y al ser el muestreo sin reposición se puede cometer el error de eliminar a uno de ellos, y decimos error, puesto que de eliminar a uno de esos



elementos, estamos cambiando la población, pues pasaríamos a tener 5 elementos en vez de los 6 que constituyen nuestra población bajo estudio.

Por otro lado, todos los elementos de la diagonal principal, o sea aquellas muestras en que se repite el valor, no pueden constituir ahora el espacio probabilístico, pues el muestreo es sin reposición, o sea que si en la primera extracción se ha presentado el 1, no puede volver a presentarse, por lo tanto todos los elementos de la diagonal principal no deben ser considerados. Nos quedan así dos bloques, uno arriba de la diagonal principal y otro debajo. Cualquiera de ellos, pero solo uno, constituirá nuestro espacio probabilístico con 15 resultados posibles. (Obsérvese que en cada bloque, cada par es idéntico sólo que cambia el orden de presentación de sus elementos). La probabilidad de cada evento elemental es ahora 1/15, para nuestro caso, siendo en

general de:

$$\frac{1}{C_N^n}$$

Una vez planteada la población objeto de análisis y definido el tamaño muestral y el espacio probabilístico, deduciremos la *Distribución Por Muestreo*, de algunas estadísticas.

### 5.1. Distribución Por Muestreo De La Media Muestral

La media muestral es, como toda media, un promedio de las observaciones.

Entonces, se calcula y se simboliza como:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Cada muestra posible de tamaño n, extraída de una población de N elementos, sea el muestreo, con o sin reposición, dará lugar a un valor para  $\bar{x}$ . Así tendremos, que la cantidad de medias muestrales en el muestreo con reposición es  $N^n$  y el muestreo sin reposición  $C_N^n$ .

En el cálculo de la media muestral, intervienen las observaciones que son variables aleatorias, como consecuencia de ello  $\bar{x}$  es también una variable aleatoria.

Ahora bien, dentro de una muestra,  $\bar{x}$  es una constante, pues en esa muestra es la única media que puede calcularse, pero en el conjunto de muestras posibles,  $\bar{x}$  es una variable aleatoria, pues no sabemos cual muestra ha de presentarse.

Entonces, como a toda variable aleatoria, es posible calcularle:

- a- su función de probabilidad
- b- su esperanza
- c- su varianza y desviación

A tal fin calcularemos en primer término, los valores que  $\bar{x}$  puede asumir.



### 5.1.1. Muestreo Con Reposición.

En la **TABLA A**, hemos definido las 36 muestras posibles, en base a ellas calculamos entonces las 36 medias muestrales.

Así, si la muestra es:

$$(1, 1); \bar{x}_1 = \frac{1+1}{2} = 1$$

$$(1, 2); \bar{x}_2 = \frac{1+2}{2} = 1,5$$

$$(1, 3); \bar{x}_3 = \frac{1+3}{2} = 2$$

y así sucesivamente.

Construiremos ahora una tabla idéntica a la **TABLA A**, donde en lugar de cada muestra escribiremos su media muestral.

$\frac{x_1}{x_2}$	1	2	3	3	4	5
1	1	1,5	2	2	2,5	3
2	1,5	2	2,5	2,5	3	3,5
3	2	2,5	3	3	3,5	4
3	2	2,5	3	3	3,5	4
4	2,5	3	3,5	3,5	4	4,5
5	3	3,5	4	4	4,5	5

**TABLA C**

### Función de Probabilidad

Obsérvese que hay valores iguales que se presentan para más de una muestra, en tal caso, podemos sistematizar estos valores en una tabla, tomando los valores distintos que puede asumir  $\bar{x}$  y calcular su probabilidad, teniendo en cuenta que cada descripción representa un caso igualmente posible y mutuamente excluyente (no se pueden presentar juntos), de modo que las probabilidades se computan sumando el número de descripciones que pertenecen a cada valor en relación al total de valores.

$\bar{x}_i$	$P(\bar{x}_i)$
1,0	1/36
1,5	2/36
2,0	5/36
2,5	6/36
3,0	8/36
3,5	6/36
4,0	5/36
4,5	2/36
5,0	1/36
	36/36

**TABLA D**

Hemos arribado de esta manera a la función de probabilidades de la media muestral que constituye la *Distribución Por Muestreo de  $\bar{x}$* .



## Esperanza

La media de la distribución de muestreo de  $\bar{x}$  se calcula de la misma forma que la media de cualquier otra distribución de probabilidades, sólo que ahora reemplazamos a la variable  $x$  por  $\bar{x}$ .

$$E(\bar{x}) = \bar{x} = \sum_{i=1}^m \bar{x}_i P(\bar{x}_i) = \frac{\sum_{i=1}^m \bar{x}_i n_i}{N^n} \quad \text{Considerando que:}$$

$$P(\bar{x}_i) = \frac{n_i}{N^n} \quad \text{Y } m \text{ es la cantidad de valores distintos de } \bar{x}.$$

Para encontrar su valor construiremos la siguiente tabla de cálculos, donde además de  $P(\bar{x}_i)$ , consideraremos a  $n_i$  (frecuencia absoluta).

$\bar{x}_i$	$P(\bar{x}_i)$	$\bar{x}_i P(\bar{x}_i)$	$\bar{x}_i^2 P(\bar{x}_i) = \bar{x}_i [\bar{x}_i P(\bar{x}_i)]$	$n_i$	$\bar{x}_i n_i$	$\bar{x}_i^2 n_i$
1,0	1/36	1/36	1/36	1	1,0	1,0
1,5	2/36	3/36	4,5/36	2	3,0	4,5
2,0	5/36	10/36	20/36	5	10,0	20,0
2,5	6/36	15/36	37,5/36	6	15,0	37,5
3,0	8/36	24/36	72/36	8	24,0	72,0
3,5	6/36	21/36	73,5/36	6	21,0	73,5
4,0	5/36	20/36	80/36	5	20,0	80,0
4,5	2/36	9/36	40,5/36	2	9,0	40,5
5,0	1/36	5/36	25/36	1	5,0	25,0
	<b>36/36</b>	<b>108/36</b>	<b>354/36</b>	<b>36</b>	<b>108,0</b>	<b>354,0</b>

Entonces:

$$E(\bar{x}) = \bar{x} = \sum_{i=1}^m \bar{x}_i P(\bar{x}_i) = \frac{108}{36} = 3$$

O bien:

$$E(\bar{x}) = \bar{x} = \frac{\sum_{i=1}^m \bar{x}_i n_i}{N^n} = \frac{108}{36} = 3$$

Entonces, los valores posibles de  $\bar{x}$ , varían de 1 a 5, estando la distribución centrada en 3. Este valor es el mismo que la *Media Poblacional*, es decir que  $E(\bar{x}) = \mu$ , ya sea que la población se finita o infinita.

## Varianza y Desviación Estándar

La varianza y desviación estándar se calculan de la misma forma que la varianza y desviación estándar de cualquier distribución de probabilidades, sólo que reemplazamos a  $x$  por  $\bar{x}$ . Así, la varianza de la media muestral o varianza de  $\bar{x}$ , es :



$$\sigma_x^2 = \sum_{i=1}^m \bar{x}_i^2 P(\bar{x}_i) - [E(\bar{x})]^2 = \frac{\sum_{i=1}^m \bar{x}_i^2 n_i}{N^n} - [E(\bar{x})]^2$$

Reemplazando según datos:

$$\sigma_x^2 = \frac{354}{36} - 3^2 = 0,8333$$

Y la desviación estándar de la media muestral o desviación estándar de  $\bar{x}$ , es:

$$\sigma_{\bar{x}} = \sqrt{\sigma_x^2} = \sqrt{\sum_{i=1}^m \bar{x}_i^2 P(\bar{x}_i) - [E(\bar{x})]^2} = \sqrt{\frac{\sum_{i=1}^m \bar{x}_i^2 n_i}{N^n} - [E(\bar{x})]^2}$$

Luego para el ejemplo:

$$\sigma_{\bar{x}} = \sqrt{0,8333} = 0,9129$$

Nótese que la distribución de muestreo de  $\bar{x}$  no es tan variable como la de la población muestreada, así  $\sigma_{\bar{x}} = 0,9129$ , mientras que  $\sigma_X = 1,29$ .

La distribución de muestreo de  $\bar{x}$  es siempre menos variable que la de la población muestreada.

$\sigma_{\bar{x}}$  Mide la variabilidad de las medias muestrales posibles  $\bar{x}$  en la distribución de muestreo de  $\bar{x}$ , y  $\sigma_X$  mide la variación de los valores de X en la población.

La relación exacta entre  $\sigma_{\bar{x}}$  y  $\sigma_X$  para muestras aleatorias simples, si el muestreo es con reemplazo (poblaciones finitas ó infinitas) es:

$$\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$$

Y en consecuencia:

$$\sigma_{\bar{x}}^2 = \frac{\sigma_X^2}{n}$$

Comprobemos varias relaciones:

$$\sigma_{\bar{x}}^2 = 0,8333 \quad \text{y} \quad \sigma_X^2 = 1,666$$

$$\sigma_{\bar{x}}^2 = \frac{1,666}{2} = 0,833$$

$$\sigma_{\bar{x}} = 0,9129 \quad \text{y} \quad \sigma_X = 1,29, \text{ luego:} \quad \sigma_{\bar{x}} = \frac{1,29}{\sqrt{2}} = 0,9129$$



Obsérvese que la varianza de la media muestral es directamente proporcional a la varianza poblacional e inversamente proporcional al tamaño de la muestra. Esto quiere decir que a medida que aumenta el tamaño de muestra, menor es la variabilidad de la media muestral, mientras que, cuanto más variable es la característica en estudio en la población ( $\sigma^2$ ), mayor será también la varianza de la media muestral.

Ahora bien, la distribución de muestreo de  $\bar{x}$  está basada en todas las muestras aleatorias posibles de un tamaño dado que pueden ser seleccionadas de una población. No obstante en la práctica solamente se toma una muestra aleatoria. En este caso ¿es la distribución de muestreo de  $\bar{x}$ , un concepto útil? La respuesta es sí.

Sabemos ahora, que cuando tomamos una muestra aleatoria de un tamaño dado de una población, y calculamos la media de la muestra realmente obtenida, éste es tan solo uno de los muchos valores posibles que podría asumir la media de la muestra. La pregunta, entonces es, si la media de la muestra particular obtenida está cercana a la media de la población.

¿Cómo nos ayuda la distribución de muestreo de  $\bar{x}$  a responder a esta pregunta?

La distribución de muestreo de  $\bar{x}$  en la **TABLA D** revela que ninguna media de la muestra difiere de la media de la población en más de  $\pm 2$  [(5-3) y (1-3)] materias, lo cual significa que el error posible debido al muestreo, es cuando mucho del 66%[(2/3)x100].

Además, existe una alta probabilidad (56%) de que la media de la muestra no difiera de la media poblacional en más de 0,5 (alrededor de un 17% de error). Así pues, podemos confiar mucho (o no) en que una muestra aleatoria simple de dos estudiantes indicará el número de materias aprobadas promedio de la población con un error de no más el 17%. Es, en esta forma, que la distribución del muestreo de  $\bar{x}$ , conteniendo los resultados de todas las muestras posibles aleatorias, nos permite conocer la precisión con la cual estimar la media de la población en base de cualquier muestra aleatoria simple.

### 5.1.2. Muestreo Sin Reposición

En la **TABLA B**, hemos definido las 15 muestras posibles. En base a ellas calcularemos las 15 medias muestrales, siguiendo idéntico procedimiento que en el caso anterior.

$x_2$ $x_1$	1	2	3	3	4	5
1		1,5	2	2	2,5	3
2			2,5	2,5	3	3,5
3				3	3,5	4
3					3,5	4
4						4,5
5						

**TABLA E**



## Función de Probabilidad

$\bar{x}_i$	$P(\bar{x}_i)$
1,5	1/15
2,0	2/15
2,5	3/15
3,0	3/15
3,5	3/15
4,0	2/15
4,5	1/15
	<b>15/15</b>

Y ésta es, al igual que antes, es la *Distribución Por Muestreo De  $\bar{x}$* .

## Esperanza

$\bar{x}_i$	$P(\bar{x}_i)$	$\bar{x}_i P(\bar{x}_i)$	$\bar{x}_i^2 P(\bar{x}_i)$	$n_i$	$\bar{x}_i n_i$	$\bar{x}_i^2 n_i$
1,5	1/15	1,5/15	2,25/15	1	1,5	2,25
2,0	2/15	4,0/15	8,00/15	2	4,0	8,00
2,5	3/15	7,5/15	18,75/15	3	7,5	18,75
3,0	3/15	9,0/15	27,00/15	3	9,0	27,00
3,5	3/15	10,5/15	36,75/15	3	10,5	36,75
4,0	2/15	8,0/15	32,00/15	2	8,0	32,00
4,5	1/15	4,5/15	20,25/15	1	4,5	20,35
	<b>15/15</b>	<b>45/15</b>	<b>145/15</b>	<b>15</b>	<b>45,0</b>	<b>145,00</b>

Entonces:

$$E(\bar{x}) = \bar{x} = \sum_{i=1}^m \bar{x}_i P(\bar{x}_i) = \frac{\sum_{i=1}^m \bar{x}_i n_i}{C_N^n}$$

Considerando que en M.S.R.  $P(\bar{x}) = \frac{n_i}{C_N^n}$

Luego:  $E(\bar{x}) = \frac{45}{15} = 3$

Así, hemos arribado nuevamente a que  $E(\bar{x}) = \mu$

## Varianza y Desviación Stándar

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^m \bar{x}_i^2 P(\bar{x}_i) - [E(\bar{x})]^2 = \frac{\sum_{i=1}^m \bar{x}_i^2 n_i}{C_N^n} - [E(\bar{x})]^2$$

Remplazando según los datos:

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^m \bar{x}_i^2 P(\bar{x}_i) - [E(\bar{x})]^2 = \frac{145}{15} - 3^2 = 0,666$$

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{0,666} = 0,8165$$



Si bien la  $E(\bar{x})$  es igual a  $\mu$ , tanto si el muestreo es con reemplazo o sin reemplazo, no sucede lo mismo con las relaciones entre:

$$\sigma_{\bar{x}}^2 \text{ y } \sigma_X^2 \quad \text{o} \quad \sigma_{\bar{x}} \text{ y } \sigma_x$$

Así en el muestreo sin reemplazo:

$$\sigma_{\bar{x}}^2 = \frac{\sigma_X^2}{n} \cdot \frac{N-n}{N-1} \quad \text{Y} \quad \sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

El factor  $\frac{N-n}{N-1}$  recibe el nombre de factor de corrección finita, y se aproxima a 1 si el tamaño de la población es relativamente grande, comparado con el tamaño de la muestra, por lo cual, con el muestreo aleatorio simple de una población infinita, la fórmula se convierte en:

$$\sigma_{\bar{x}}^2 = \frac{\sigma_X^2}{n} \quad \text{o} \quad \sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$$

Por lo tanto, las últimas fórmulas, sólo aplicables a poblaciones infinitas, con frecuencia se usan para poblaciones finitas, siempre y cuando el tamaño de la muestra no exceda al 5% del tamaño de la población.

### 5.2. Distribución Por Muestreo De La Proporción Muestral.

Recordemos que la proporción muestral que simbolizamos por  $p$  ó  $\hat{P}$ , se define como el cociente entre el número de éxitos en  $n$  pruebas, y el tamaño de la muestra. En símbolos:

$$p = \hat{P} = \frac{x}{n}$$

Supongamos ahora, que en el ejemplo dado, nos interesa la proporción de estudiantes con dos ó menos materias aprobadas.

También la proporción muestral es una variable aleatoria, por lo tanto es posible calcularle:

- su función de probabilidad
- su esperanza
- su varianza y desviación

Para ello, calcularemos en primer término los valores que puede asumir:

#### 5.2.1. Muestreo Con Reposición

En la **TABLA A**, se calcularon 36 muestras posibles y cada una de ellas dará lugar a una proporción muestral, que para determinarla, razonamos así:

$p = \hat{P} = \frac{x}{n}$ , donde  $x$  es el número de éxitos, que para nuestro caso, será la cantidad de veces que en la muestra se presenten dos ó menos materias aprobadas.



Entonces, si la muestra es :

$$(1,1), \quad x = 2 \quad y \quad \hat{P} = \frac{2}{2} = 1$$

$$(1,2), \quad x = 2 \quad y \quad \hat{P} = \frac{2}{2} = 1$$

$$(1,3), \quad x = 1 \quad y \quad \hat{P} = \frac{1}{2} = 0,50$$

y así sucesivamente, hasta calcular las 36 proporciones.

Las mismas se exponen en la tabla siguiente:

$\frac{x_2}{x_1}$	1	2	3	3	4	5
1	1	1	0,5	0,5	0,5	0,5
2	1	1	0,5	0,5	0,5	0,5
3	0,5	0,5	0	0	0	0
3	0,5	0,5	0	0	0	0
4	0,5	0,5	0	0	0	0
5	0,5	0,5	0	0	0	0

TABLA F

### Función de Probabilidad

Al igual que para la media muestral, hay valores que se presentan para más de una muestra, por lo que podemos sistematizar estos valores en una tabla como sigue:

$\hat{P}_i$	$P(\hat{P}_i)$
0	16/36
0,5	16/36
1	4/36
	36/36

TABLA G

Que constituye la *Distribución Por Muestreo De La Proporción Muestral*.

### Esperanza

Se simboliza y define:

$$E(\hat{P}) = \sum_{i=1}^m \hat{P}_i P(\hat{P}_i) = \frac{\sum_{i=1}^m \hat{P}_i n_i}{N^n}$$



Entonces:

$\hat{P}_i$	$P(\hat{P}_i)$	$\hat{P}_i P(\hat{P}_i)$	$\hat{P}_i^2 P(\hat{P}_i)$	$n_i$	$\hat{P}_i n_i$	$\hat{P}_i^2 n_i$
0	16/36	0	0	16	0	0
0,5	16/36	8/36	4/36	16	8	4
1	4/36	4/36	4/36	4	4	4
	<b>36/36</b>	<b>12/36</b>	<b>8/36</b>	<b>36</b>	<b>12</b>	<b>8</b>

$$E(\hat{P}) = \sum_{i=1}^m \hat{P}_i P(\hat{P}_i) = \frac{\sum_{i=1}^m \hat{P}_i n_i}{N^n} = \frac{12}{36} = 0,33$$

Si recordamos el valor obtenido para  $P$ , proporción poblacional, vemos que la distribución por muestreo de  $\hat{P}$  también está centrada en la correspondiente característica de la población, o sea, alrededor de la proporción poblacional 0,33.

Luego, ya sea que la población muestreada sea finita o infinita, la media de la distribución por muestreo de  $\hat{P}$  es siempre igual a la proporción de la población  $P$ , es decir:

$$E(\hat{P}) = P$$

### Varianza y Desviación Stándar

La varianza, se simboliza y define:

$$\sigma_p^2 = \sum_{i=1}^m \hat{P}_i^2 P(\hat{P}_i) - [E(\hat{P}_i)]^2 = \frac{\sum_{i=1}^m \hat{P}_i^2 n_i}{N^n} - [E(\hat{P})]^2$$

Recurriendo a la tabla de cálculos anterior, obtenemos:

$$\sigma_p^2 = \frac{8}{36} - \left( \frac{12}{36} \right)^2 = 0,11$$

La desviación, se simboliza y define:

$$\sigma_{\hat{p}} = \sqrt{\sum_{i=1}^m \hat{P}_i^2 P(\hat{P}_i) - [E(\hat{P}_i)]^2} = \sqrt{\frac{\sum_{i=1}^m \hat{P}_i^2 n_i}{N^n} - [E(\hat{P})]^2}$$

Nótese que la variabilidad relativa de la distribución por muestreo de  $\hat{P}$  es 100%  $[(0,33/0,33) \times 100]$ , la cual es mucho mayor que la de la distribución de  $\bar{x}$  de 30,43%  $[(0,9129/3) \times 100]$ . Esta mayor variabilidad puede asegurarse al hecho de que  $\hat{P}$  ignora el número de materias reales, aprobadas por los estudiantes de la muestra y considera sólo si son menores o iguales a dos.



Ahora bien, no necesitamos desarrollar la distribución de  $\hat{P}$  para calcular  $\sigma_{\hat{P}}$  (variación de  $\hat{P}$  de una muestra a la otra), pues la teoría estadística indica que, si el muestreo es aleatorio simple y con reposición (poblaciones infinitas o finitas), entonces:

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}} \quad \text{Donde } P \text{ es la proporción poblacional.}$$

Así, reemplazando a  $P$  y  $n$  por su igual, obtenemos que:

$$\sigma_{\hat{P}} = \sqrt{\frac{0,33(0,67)}{2}} = \underline{0,33}$$

Análogamente:

$$\sigma_{\hat{P}}^2 = \frac{P(1-P)}{n} = \frac{0,33(0,67)}{2} = \underline{0,11}$$

### 5.2.2. Muestreo Sin Reposición

Siguiendo los mismos pasos anteriores, plantearemos los valores que  $\hat{P}$  puede asumir, así como su función de probabilidad, esperanza, varianza y desviación estándar.

Entonces:

$x_2$ $x_1$	1	2	3	3	4	5
1	1	1	0,5	0,5	0,5	0,5
2	1	1	0,5	0,5	0,5	0,5
3	0,5	0,5	0	0	0	0
3	0,5	0,5	0	0	0	0
4	0,5	0,5	0	0	0	0
5	0,5	0,5	0	0	0	0

TABLA H

Al igual que para la media muestral, en este caso los valores que  $\hat{P}$  puede asumir, son los correspondientes al bloque superior ó inferior a la diagonal principal, entonces:

### Función de Probabilidad

$\hat{P}_i$	$P(\hat{P}_i)$
0	6/15
0,5	8/15
1	1/15
	<b>15/15</b>

Así, esta tabla, constituye la *Distribución Por Muestreo De  $\hat{P}$* .



### Esperanza

Al igual que antes:

$$E(\hat{P}) = \sum_{i=1}^m \hat{P}_i P(\hat{P}_i) = \frac{\sum_{i=1}^m \hat{P}_i n_i}{C_N^n}$$

Luego:

$\hat{P}_i$	$P(\hat{P}_i)$	$\hat{P}_i P(\hat{P}_i)$	$\hat{P}_i^2 P(\hat{P}_i)$	$n_i$	$\hat{P}_i n_i$	$\hat{P}_i^2 n_i$
0	6/15	0	0	6	0	0
0,5	8/15	4/15	2/15	8	4	2
1	1/15	1/15	1/15	1	1	1
	<b>15/15</b>	<b>5/15</b>	<b>3/15</b>	<b>15</b>	<b>5</b>	<b>3</b>

$$E(\hat{P}) = \frac{5}{15} = \underline{0,33}$$

Entonces, en el muestreo sin reemplazo (poblaciones finitas), también se verifica que:

$$E(\hat{P}) = P$$

### Varianza y Desviación Estándar

$$\sigma_{\hat{p}}^2 = \sum_{i=1}^m \hat{P}_i^2 P(\hat{P}_i) - [E(\hat{P}_i)]^2 = \frac{\sum_{i=1}^m \hat{P}_i^2 n_i}{C_N^n} - [E(\hat{P})]^2$$

Reemplazando:

$$\sigma_{\hat{p}}^2 = \frac{3}{5} - \left( \frac{5}{15} \right)^2 = \underline{0,089}$$

y

$$\sigma_{\hat{p}} = \sqrt{\sum_{i=1}^m \hat{P}_i^2 P(\hat{P}_i) - [E(\hat{P}_i)]^2} = \sqrt{\frac{\sum_{i=1}^m \hat{P}_i^2 n_i}{C_N^n} - [E(\hat{P})]^2}$$

Reemplazando:

$$\sigma_{\hat{p}} = \sqrt{0,089} = \underline{0,298}$$

En el caso del muestreo sin reemplazo (poblaciones finitas), se verifica que:

$$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}} \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0,33(0,67)}{2}} \sqrt{\frac{6-2}{6-1}} = \underline{0,298}$$



Análogamente:

$$\sigma_{\hat{P}}^2 = \frac{P(1-P)}{n} \frac{N-n}{N-1} = \frac{0,33(0,67)}{2} \frac{6-2}{6-1} = 0,089$$

El factor  $\frac{N-n}{N-1}$  recibe igual tratamiento que el ya analizado para  $\bar{x}$  en el muestreo sin reemplazo.

### 5.3. Distribución Por Muestreo De La Varianza Muestral Corregida.

Así como a cada muestra le hemos calculado su media, podemos determinarle su varianza, que indicará como toda varianza, el grado de dispersión o concentración de las observaciones muestrales respecto a su valor central.

Esta varianza, recibe el nombre de *Varianza Muestral* y se simboliza y calcula como:

#### 1- Serie Simple

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2$$

La primera expresión corresponde a la fórmula definicional (recuérdese que la varianza se definía como un promedio del cuadrado de las desviaciones con respecto a la media), mientras que la segunda expresión, constituye una fórmula de cálculo rápido y es la que generalmente utilizaremos.

#### 2- Datos Agrupados

$$s^2 = \frac{\sum_{i=1}^m (x_i - \bar{x})^2 n_i}{n} = \sum_{i=1}^m (x_i - \bar{x})^2 P(x_i)$$

Considerando que  $\frac{n_i}{n} = P(x_i)$ , luego la fórmula de cálculo rápido se expresa como sigue:

$$s^2 = \frac{\sum_{i=1}^m x_i^2 n_i - n \bar{x}^2}{n} = \frac{\sum_{i=1}^m x_i^2 n_i}{n} - \bar{x}^2 = \sum_{i=1}^m x_i^2 P(x_i) - \bar{x}^2$$

Definiremos a continuación a la *Varianza Muestral Corregida*, concepto que será de utilidad en los próximos capítulos, según veremos.

Entonces la varianza muestral corregida, se simboliza y calcula:



### 1- Series Simples

$$\hat{s}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}$$

### 2- Datos Agrupados

$$\hat{s}^2 = \frac{\sum_{i=1}^m (x_i - \bar{x})^2 n_i}{n-1} = \frac{\sum_{i=1}^m x_i^2 n_i - n\bar{x}^2}{n-1}$$

Obsérvese que los numeradores para el cálculo de  $s^2$  y  $\hat{s}^2$  son iguales, y la diferencia radica en el denominador de las fórmulas,  $n$  para  $s^2$  y  $n-1$  para  $\hat{s}^2$ .

Luego las relaciones entre ambas son:

$$\hat{s}^2 = s^2 \frac{n}{n-1} \quad \text{y} \quad s^2 = \hat{s}^2 \frac{n-1}{n}$$

Por las mismas razones dadas en el capítulo I, al estudiar las medidas de dispersión, se puede definir la desviación estándar muestral y la desviación estándar muestral corregida.

Entonces la desviación estándar muestral, se simboliza y calcula:

$$s = \sqrt{s^2}$$

Mientras que la desviación estándar muestral corregida sería:

$$\hat{s} = \sqrt{\hat{s}^2}$$

Por otro lado, las relaciones entre ambas son:

$$\hat{s} = s \sqrt{\frac{n}{n-1}} \quad s = \hat{s} \sqrt{\frac{n-1}{n}}$$

La varianza muestral corregida, por idénticas razones que las dadas para la media muestral, es también una variable aleatoria. Por lo tanto, calcularemos:

- a- su función de probabilidad.
- b- su esperanza.

#### 5.3.1. Muestreo Con Reposición

En base a la **TABLA A**, donde definimos las muestras posibles para nuestro ejemplo, calcularemos las 36 varianzas corregidas. Así, si la muestra es  $(1,1)$ , siendo  $\bar{x} = 1$ ,



$$\hat{s}^2 = \frac{\sum_{i=1}^2 (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2}{n-1} + \frac{(x_2 - \bar{x})^2}{n-1} = \frac{(1-1)^2}{2-1} + \frac{(1-1)^2}{2-1} = 0 + 0 = 0$$

Si la muestra es (1,2), siendo  $\bar{x} = 1,5$ :

$$\begin{aligned}\hat{s}^2 &= \frac{(x_1 - \bar{x})^2}{n-1} + \frac{(x_2 - \bar{x})^2}{n-1} = \frac{(1-1,5)^2}{2-1} + \frac{(2-1,5)^2}{2-1} = \\ &= (-0,5)^2 + (0,5)^2 = 0,25 + 0,25 = \underline{0,50}\end{aligned}$$

O bien:

$$\begin{aligned}\hat{s}^2 &= \frac{(x_1^2 + x_2^2) - n\bar{x}^2}{n-1} = \frac{(1^2 + 2^2) - 2(1,5)^2}{2-1} = \\ &= \frac{(1+4)-(2 \cdot 2,25)}{1} = 5 - 4,5 = \underline{0,50}\end{aligned}$$

Y así sucesivamente.

Luego, las varianzas muestrales corregidas posibles son:

$x_1$	1	2	3	3	4	5
1	0	0,5	2	2	4,5	8
2	0,5	0	0,5	0,5	2	4,5
3	2	0,5	0	0	0,5	2
3	2	0,5	0	0	0,5	2
4	4,5	2	0,5	0,5	0	0,5
5	8	4,5	2	2	0,5	0

TABLA I

### Función de Probabilidad

Con igual criterio que el utilizado en los casos anteriores definimos la función de probabilidad que representa al igual que antes la Distribución por muestreo de  $\hat{s}^2$ .

Así:

$\hat{s}_i^2$	$P(\hat{s}_i^2)$
0,0	8/36
0,5	12/36
2,0	10/36
4,5	4/36
8,0	2/36
	36/36

Y ésta es la *Distribución Por Muestreo De La Varianza Muestral Corregida*.



## Esperanza

Con igual concepto, se simboliza y calcula como:

$$E(\hat{s}^2) = \sum_{i=1}^m \hat{s}_i^2 P(\hat{s}_i^2) = \frac{\sum_{i=1}^m \hat{s}_i^2 n_i}{N^n}$$

Entonces:

$\hat{s}_i^2$	$P(\hat{s}_i^2)$	$\hat{s}_i^2 P(\hat{s}_i^2)$	$n_i$	$\hat{s}_i^2 n_i$
0,0	8/36	0	8	0
0,5	12/36	6/36	12	6
2,0	10/36	20/36	10	20
4,5	4/36	18/36	4	18
8,0	2/36	16/36	2	16
	<b>36/36</b>	<b>60/36</b>	<b>36</b>	<b>60</b>

Reemplazando:

$$E(\hat{s}^2) = 60/36 = 1,6666.$$

Pero 1,66 es el valor de la varianza poblacional.

Luego cuando el muestreo es con reposición:

$$E(\hat{s}^2) = \sigma_x^2$$

### 5.3.2. Muestreo Sin Reposición

Al igual que en los casos vistos para el muestreo sin reposición, los valores posibles para  $\hat{s}^2$  corresponden al bloque superior o inferior de la diagonal principal, luego podemos definir:

#### Función de Probabilidad

$\hat{s}_i^2$	$P(\hat{s}_i^2)$
0	1/15
0,5	6/15
2	5/15
4,5	2/15
8,0	1/15
	<b>15/15</b>

Y ésta es la *Distribución Por Muestreo De La Varianza Muestral Corregida*.



### Esperanza

Lo simbolizamos y calculamos como:

$$E(\hat{s}^2) = \sum_{i=1}^m \hat{s}_i^2 P(\hat{s}_i^2) = \frac{\sum_{i=1}^m \hat{s}_i^2 n_i}{C_N^n}$$

Entonces:

$\hat{s}_i^2$	$P(\hat{s}_i^2)$	$\hat{s}_i^2 P(\hat{s}_i^2)$	$n_i$	$\hat{s}_i^2 n_i$
0,0	1/15	0/15	1	0
0,5	6/15	3/15	6	3
2,0	5/15	10/15	5	10
4,5	2/15	9/15	2	9
8,0	1/15	8/15	1	8
	<b>15/15</b>	<b>30/15</b>	<b>15</b>	<b>30</b>

Reemplazando:

$$E(\hat{s}^2) = 30/15 = 2$$

Pero 2 es el valor de la varianza poblacional corregida.

Luego, cuando el muestreo es sin reposición:  $E(\hat{s}^2) = S^2$

En resumen, hemos mostrado hasta ahora, que para cualquier estadística de la muestra basada en un muestreo aleatorio simple de un tamaño dado y de una población especificada, existe una distribución de muestreo de esa estadística que indica:

- 1- Todos los diferentes valores posibles de la estadística de muestra que pueden ser obtenidos a partir de todas las diferentes muestras aleatorias posibles de un tamaño dado, de la población.
- 2- Las probabilidades de que se presenten estos valores de la estadística de muestra.

Cualquier distribución de muestreo siempre se refiere a:

- 1- La población específica que está siendo muestreada.
- 2- Un tamaño específico de muestra aleatoria simple.

Si cambia la población o el tamaño de muestra aleatoria simple, obtenemos una nueva distribución de muestreo.



## 6. LEY DE GRANDES NÚMEROS

Expresa que “La probabilidad de que la diferencia entre el estadístico y el parámetro sea superior a un número  $d$ , arbitrariamente elegido, tiende a 0, a medida que  $n$  tiende a  $\infty$ ”.

Si simbolizamos por  $\theta$  al parámetro y por  $\hat{\theta}$  al estadístico, simbólicamente la ley de grandes números queda indicada por:

$$\lim_{n \rightarrow \infty} \Pr(|\hat{\theta} - \theta| > d) = 0$$

O bien, podemos expresarla como

$$\lim_{n \rightarrow \infty} \Pr(|\hat{\theta} - \theta| \leq d) = 1$$

En otras palabras, cuando el tamaño de la muestra es grande, hay una probabilidad cercana a 1 de que el valor muestral (estadístico) esté cerca del valor poblacional (parámetro).

Si consideramos

$$\begin{aligned}\theta &= \mu && \text{Media poblacional} \\ \hat{\theta} &= \bar{x} && \text{Media muestral}\end{aligned}$$

Diremos

$$\lim_{n \rightarrow \infty} \Pr(|\bar{x} - \mu| \leq d) = 1$$

Si consideramos

$$\begin{aligned}\theta &= P && \text{Proporción poblacional} \\ \hat{\theta} &= \hat{P} && \text{Proporción muestral}\end{aligned}$$

Diremos

$$\lim_{n \rightarrow \infty} \Pr(|\hat{P} - P| \leq d) = 1$$

## 7. TEOREMA CENTRAL DE LÍMITE

No realizamos ninguna demostración de este teorema, nos limitamos a enunciarlo.

Cualquiera sea la distribución de la población en la medida que posea varianza finita, la variable aleatoria  $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$  tenderá a distribuirse con media 0 y varianza 1, a



medida que  $n$  crece indefinidamente, donde  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ , es la medida de una muestra al azar de tamaño  $n$ .

En general, la normalidad de una distribución de probabilidad para la media muestral, es el llamado T.C.L. y puede ser establecido como sigue:

- 1) Cuando la población es bastante grande y está normalmente distribuida, la distribución de probabilidad de las medias muestrales será normal.
- 2) Cuando la población no es normal, la distribución de probabilidad de las medidas muestrales se aproximará a una distribución normal si el tamaño de la muestra es suficientemente grande, usualmente 30 o más.
- 3) La distribución normal de las medias muestrales tiene la media igual al valor esperado de la muestra  $E(\bar{x})$  y el error estándar  $\sigma_{\bar{x}}$ . Los valores de  $E(\bar{x})$  y  $\sigma_{\bar{x}}$  teóricamente pueden ser calculados a partir de la media  $\mu$  y la desviación estándar  $\sigma$  de la población, respectivamente.

También es a partir de ese teorema que aproximamos cualquier distribución de probabilidad discreta a la distribución normal para tamaños de muestra grandes, según lo ya analizado.

## **8. PARÁMETROS Y ESTADÍSTICAS PARA VARIABLES Y PARÁMETROS PARA VARIABLES ALEATORIAS.**

Según vimos, cuando se cuenta con los valores observados de la Variable bajo estudio, sea en base a una población o en base a una muestra, se generan medidas descriptivas llamadas parámetros o estadísticos.

Por otro lado, si contamos con los valores posibles que una variable puede asumir, se genera una variable aleatoria, la descripción se realiza a través de medidas que llamamos parámetros.

Exponemos seguidamente una *Síntesis De Fórmulas*, para cada una de ellos y las relaciones existentes.



## Parámetros

<b>1- MEDIA POBLACIONAL</b> $E(x) = \mu$	1.1 Series Simples $\frac{\sum_{i=1}^N x_i}{N}$		1.2 Datos Agrupados $\frac{\sum_{i=1}^m X_i n_i}{N} = \sum_{i=1}^m X_i P(X_i)$
<b>2- VARIANZA</b>	2.1 Poblacional $\sigma_x^2$	Series Simples $\frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^N x_i^2}{N} - \mu^2$	Datos Agrupados $\frac{\sum_{i=1}^m (x_i - \mu)^2 n_i}{N} = \frac{\sum_{i=1}^m x_i^2 n_i}{N} - \mu^2 = \sum_{i=1}^m x_i^2 P(x_i) - \mu^2$
	2.2 Poblacional Corregida $S_x^2$	Series Simples $\frac{\sum_{i=1}^N (x_i - \mu)^2}{N-1} = \frac{\sum_{i=1}^N x_i^2}{N-1} - \frac{N \mu^2}{N-1}$	Datos Agrupados $\frac{\sum_{i=1}^m (x_i - \mu)^2 n_i}{N-1} = \frac{\sum_{i=1}^m x_i^2 n_i}{N-1} - \frac{N \mu^2}{N-1}$
	2.3 Relaciones	$\sigma_x^2 = S_x^2 \frac{N-1}{N}$	$S_x^2 = \sigma_x^2 \frac{N}{N-1}$
<b>3- DESVIACIÓN TÍPICA</b>	3.1 Poblacional	$\sigma_x = \sqrt{\sigma_x^2}$	
	3.2 Poblacional Corregida	$s_x = \sqrt{s_x^2}$	
	3.3 Relaciones	$\sigma_x = s_x \sqrt{\frac{N-1}{N}}$	$s_x = \sigma_x \sqrt{\frac{N}{N-1}}$
<b>4- PROPORCIÓN POBLACIONAL</b>	$P = \frac{X}{N}$		



## Estadísticos

		1.1 Series Simples		1.2 Datos Agrupados	
<b>1- MEDIA MUESTRAL</b>  $\bar{x}$		$\frac{\sum_{i=1}^n x_i}{n}$		$\frac{\sum_{i=1}^m x_i n_i}{n}$	
<b>2- VARIANZA</b>	<b>2.1 Muestral</b>  $s^2$	<i>Series Simples</i> $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2$		<i>Datos Agrupados</i> $\frac{\sum_{i=1}^m (x_i - \bar{x})^2 n_i}{n} = \frac{\sum_{i=1}^m x_i^2 n_i}{n} - \bar{x}^2 = \sum_{i=1}^m x_i^2 P(x_i) - \bar{x}^2$	
	<b>2.2 Muestral Corregida</b>  $\hat{s}^2$	<i>Series Simples</i> $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2}{n-1} - \frac{n \bar{x}^2}{n-1}$		<i>Datos Agrupados</i> $\frac{\sum_{i=1}^m (x_i - \bar{x})^2 n_i}{n-1} = \frac{\sum_{i=1}^m x_i^2 n_i}{n-1} - \frac{n \bar{x}^2}{n-1}$	
	<b>2.3 Relaciones</b>	$s^2 = \hat{s}^2 \frac{n-1}{n}$		$\hat{s}^2 = s^2 \frac{n}{n-1}$	
<b>3- DESVIACIÓN TÍPICA</b>	<b>3.1 Muestral</b>	$s = \sqrt{s^2}$			
	<b>3.2 Muestral Corregida</b>	$\hat{s} = \sqrt{\hat{s}^2}$			
	<b>3.3 Relaciones</b>	$s = \hat{s} \sqrt{\frac{n-1}{n}}$		$\hat{s} = s \sqrt{\frac{n}{n-1}}$	
<b>4- PROPORCIÓN MUESTRAL</b>		$\hat{P} = p = \frac{x}{n}$			



## Parámetros Para Variables Aleatorias

<b>5- MEDIA DE LAS MEDIAS MUESTRALES</b> $E(\bar{x}) = \bar{x}$	<i>M.C.R.</i>	$\frac{\sum_{i=1}^m \bar{x}_i n_i}{N^n} = \sum_{i=1}^m \bar{x}_i P(\bar{x}_i)$
	<i>M.S.R.</i>	$\frac{\sum_{i=1}^m \bar{x}_i n_i}{C_N^n} = \sum_{i=1}^m \bar{x}_i P(\bar{x}_i)$
<b>6- VARIANZA DE LAS MEDIAS MUESTRALES</b> $\sigma_{\bar{x}}^2$	<i>M.C.R.</i>	$\frac{\sum_{i=1}^m \bar{x}_i^2 n_i}{N^n} - [E(\bar{x})]^2 = \sum_{i=1}^m \bar{x}_i^2 P(\bar{x}_i) - [E(\bar{x})]^2$
	<i>M.S.R.</i>	$\frac{\sum_{i=1}^m \bar{x}_i^2 n_i}{C_N^n} - [E(\bar{x})]^2 = \sum_{i=1}^m \bar{x}_i^2 P(\bar{x}_i) - [E(\bar{x})]^2$
<b>7- DESVIACIÓN DE LAS MEDIAS MUESTRALES</b> $\sigma_{\bar{x}}$	<i>M.C.R.</i>	
	<i>M.S.R.</i>	$\sqrt{\sigma_{\bar{x}}^2}$
<b>8- ESPERANZA DE LA PROPORCIÓN MUESTRAL</b> $E(\hat{P})$	<i>M.C.R.</i>	$\frac{\sum_{i=1}^m \hat{P}_i n_i}{N^n}$
	<i>M.S.R.</i>	$\frac{\sum_{i=1}^m \hat{P}_i n_i}{C_N^n}$
<b>9- VARIANZA DE LA PROPORCIÓN MUESTRAL</b> $\sigma_{\hat{P}}^2$	<i>M.C.R.</i>	$\frac{\sum_{i=1}^m \hat{P}_i^2 n_i}{N^n} - [E(\hat{P})]^2$
	<i>M.S.R.</i>	$\frac{\sum_{i=1}^m \hat{P}_i^2 n_i}{C_N^n} - [E(\hat{P})]^2$



## Parámetros Para Variables Aleatorias

<b>10- DESVIACIÓN DE LA PROPORCIÓN MUESTRAL</b> $\sigma_{\hat{P}}$	$\sqrt{\sigma_{\hat{P}}^2}$					
<b>11- ESPERANZA DE LA VARIANZA MUESTRAL CORREGIDA</b> $E(\hat{s}^2)$	<i>M.C.R.</i>			<i>M.S.R.</i>		
	$\frac{\sum_{i=1}^m \hat{s}_i^2 n_i}{N^n} = \sum_{i=1}^m \hat{s}_i^2 P(\hat{s}_i^2)$			$\frac{\sum_{i=1}^m \hat{s}_i^2 n_i}{C_N^n} = \sum_{i=1}^m \hat{s}_i^2 P(\hat{s}_i^2)$		
<b>12- RELACIONES</b>	M.C.R.	$E(\bar{x}) = \mu$	$E(\hat{s}^2) = \sigma^2$	$E(\hat{P}) = P$	$\sigma_x^2 = \frac{\sigma_x^2}{n} \rightarrow \sigma_x = \frac{\sigma_x}{\sqrt{n}}$	$\sigma_{\hat{P}}^2 = \frac{P(1-P)}{n} \rightarrow \sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}}$
	M.S.R.	$E(\bar{x}) = \mu$	$E(\hat{s}^2) = s^2$	$E(\hat{P}) = P$	$\sigma_x^2 = \frac{\sigma_x^2}{n} \cdot \frac{N-n}{N-1} \rightarrow$ $\sigma_x = \frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$	$\sigma_{\hat{P}}^2 = \frac{P(1-P)}{n} \cdot \frac{N-n}{N-1} \rightarrow$ $\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$



## Unidad N° 9: Estimación Estadística

---

### Objetivos Específicos

*Que el estudiante*

**Comprenda** los fundamentos teóricos y la lógica subyacente de la Inferencia Estadística en una de sus dos grandes ramas: Estimación de Parámetros.

**Diferencie** las formas entre estimación puntual y por intervalos, teniendo en cuenta las condiciones de los buenos estimadores.

**Reconozca** las particularidades de cálculo de intervalos en distintos casos.

**Determine** los tamaños de muestra necesarios, para el caso de un Muestreo aleatorio simple.

**Conozca** el concepto, alcance e interpretación del error de estimación, el riesgo, la confianza y las relaciones entre ellos y el tamaño de la muestra.

### Contenidos

*Estimación estadística.*

**1. Generalidades.**

**2. Propiedades de los buenos estimadores.**

    2.1. Inseguridad

    2.2. Eficiencia

    2.3. Consistencia

    2.4. Suficiencia

**3. Estimación puntual. Limitaciones.**

**4. Error, riesgo y tamaño de la muestra.**

**5. Estimación por intervalos.**

    5.1. Ejemplo especial.

    5.2. Elementos y terminología.

    5.3. Nivel de confianza. Significado y selección.

    5.4. Intervalo de confianza para estimar  $\mu$ . Uso de la distribución Normal y "t" de Student.

    5.5. Determinación del tamaño de la muestra en la estimación de la  $\mu$ .

    5.6. Intervalo de confianza para estimar  $p$ . Uso de la distribución Normal.

    5.7. Determinación del tamaño de la muestra en la estimación de  $p$ .

    5.8. Intervalo de confianza para estimar la varianza de una población normal. Uso de la distribución  $\chi^2$  (Chi cuadrado).



## I. GENERALIDADES

Según lo visto hasta ahora, existen situaciones en las cuales no podemos realizar un estudio poblacional, ya sea porque la población que desea estudiarse es, o muy grande, o infinita, o resulta difícil acceder a su conocimiento total por diversos motivos, vinculados al costo o no.

Ello nos imposibilita el conocimiento de los parámetros. Es, entonces, que se recurre a la idea de estimar esos valores. ¿Y cómo puede estimarse el valor de un parámetro?

Es necesario poseer alguna información sobre la población objetivo, ya que sería imposible intentar alguna estimación sin ninguna información de ella. Esta información está provista por la muestra, a partir de la cual, se calculan medidas que proporcionan una idea de los valores posibles de esos parámetros poblacionales desconocidos, a través de las estimaciones realizadas.

Entonces, diremos que los Parámetros se estiman utilizando Estadísticas Muestrales.

Según hemos analizado, existen distintos Tipos De Muestreo, pero para aplicar la Teoría Estadística y extender las conclusiones de la muestra hacia la población, todos ellos deben garantizar de que cada elemento de la población tiene igual probabilidad de ser elegido en la muestra que los demás.

El método para hacer la estimación recibe el nombre de *Estimador*, mientras que el resultado real obtenido de la muestra seleccionada se llama *Estimación* del parámetro.

Por ejemplo, para estimar la media poblacional ( $\mu$ ), seleccionamos como estimador a la media muestral ( $\bar{X}$ ). El valor numérico de la media muestral en una muestra determinada es la estimación de la media poblacional.

Por otro lado, un Estimador, es un *Estimador Puntual* de una característica de población, si proporciona un solo número como estimación.

En correspondencia, la estimación se llama *Estimación Puntual* porque la característica de población está estimada con un solo número basado en la muestra.

La estimación puntual se contrasta con la *Estimación Por Intervalos*, que veremos más adelante, donde el parámetro se estima situado entre dos límites para una confianza dada.

A partir de estos conceptos convenimos en simbolizar como:

$\theta$  Al parámetro a estimar

$\hat{\theta}$  Al estimador, que es una función de las observaciones muestrales, es decir,  $\hat{\theta} = f(x_1, x_2, \dots, x_n)$ , luego, por ser función de las observaciones muestrales es una variable aleatoria.



Para poder elegir un estimador, de entre varios propuestos para estimar un parámetro, existen algunos criterios que fijan ciertas propiedades deseables para los estimadores. Y así, será utilizado el estimador de un parámetro que cumpla con todas o la mayoría de esas propiedades. Existen también, métodos de estimación que proporcionan los mejores estimadores, como por ejemplo el llamado *Método De Máxima Verosimilitud o De La Mayor Probabilidad*.

## **2. PROPIEDADES DE LOS BUENOS ESTIMADORES**

Las propiedades que debe poseer un buen estimador puntual se relacionan con la *Distribución Por Muestreo* del estimador, ya que como Distribución De Probabilidades, indica qué tan lejos tiende a encontrarse la estimación proporcionada por el estimador, de la característica de la población que debe ser estimada.

Las propiedades de los buenos estimadores son:

- 2.1. Insesgabilidad**
- 2.2. Eficiencia**
- 2.3. Consistencia**
- 2.4. Suficiencia**

### ***2.1. Insesgabilidad***

Un estimador es insesgado, cuando su valor esperado, es decir su esperanza matemática, es igual al parámetro de la población que se estima. Es decir, es una propiedad del estimador, pero en relación al parámetro que estima.

Entonces, si  $\theta$ , es el parámetro a estimar y  $\hat{\theta}$ , un estimador propuesto,  $\hat{\theta}$  es un estimador insesgado de  $\theta$ , si  $E(\hat{\theta}) = \theta$ , cualquiera sea su distribución.

Si existiera una diferencia entre la esperanza del estimador y el parámetro a estimar, esa diferencia se denomina sesgo.

### **Ejemplos**

Supongamos que la media poblacional  $\mu$  se estima con el estimador media muestral ( $\bar{x}$ ). Entonces:

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

Hemos demostrado en Distribuciones de Muestreo que la esperanza matemática de  $\bar{x}$ , es igual a la media poblacional  $\mu$ , tanto en el muestreo con reemplazo, como en el muestreo sin reemplazo. Es decir que la media muestral ( $\bar{x}$ ), se distribuye con esperanza igual a la media poblacional  $\mu$ :

$$E(\bar{x}) = \mu$$

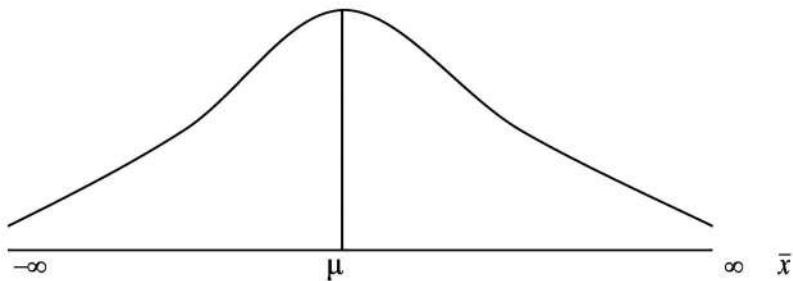


Por lo tanto la media muestral  $\bar{x}$ , es un estimador insesgado de la media poblacional  $\mu$ .

Si este estimador insesgado tiene distribución normal o asintótica normal, se distribuye:

$$\bar{x} \sim N(\mu, \sigma_{\bar{x}})$$

Y su gráfica es:



Supongamos ahora, que la media poblacional,  $\mu$ , se estima con el estimador  $\bar{x} + c$ , donde  $c$  es una constante. Entonces:

$$\begin{aligned}\theta &= \mu \\ \hat{\theta} &= \bar{x} + c\end{aligned}$$

Para determinar si el estimador propuesto es insesgado, debemos calcular su esperanza matemática y verificar si arribamos o no al parámetro a estimar.

Así, considerando las propiedades de la esperanza:

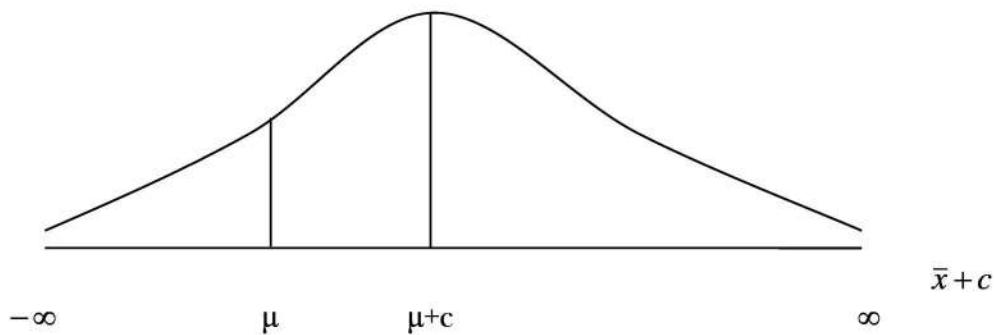
$$E(\hat{\theta}) = E(\bar{x} + c) = E(\bar{x}) + E(c) = \mu + c$$

Este estimador de  $\mu$  se distribuye con esperanza matemática igual a  $\mu + c$ , luego no es insesgado, sino que es sesgado.

Si se distribuye normal, entonces:

$$\bar{x} + c \sim N(\mu + c, \sigma_{\bar{x}+c})$$

La gráfica es:





La esperanza del estimador se sitúa a la derecha del parámetro a estimar, entonces tiene sesgo positivo.

En general, para un estimador  $\hat{\theta}$  del parámetro  $\theta$ , tendremos:

Cuando  $E(\hat{\theta}) = \theta$ ,  $\hat{\theta}$  es un estimador insesgado de  $\theta$ .

Cuando  $E(\hat{\theta}) > \theta$ ,  $\hat{\theta}$  es un estimador sesgado, con sesgo positivo.

Cuando  $E(\hat{\theta}) < \theta$ ,  $\hat{\theta}$  es un estimador sesgado, con sesgo negativo.

El sesgo relacionado con estimadores, no es el sesgo debido a errores en los registros, ni a las preguntas dirigidas usadas en un cuestionario, ni a otros factores similares que llevan a errores que no pertenecen al muestreo.

Veamos algunas otras aplicaciones:

1-

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

Luego:

$E(\bar{x}) = \mu$ , cualquiera sea la distribución de  $\bar{x}$  y  $\bar{x}$  es estimador insesgado de  $\mu$ , igualdad demostrada prácticamente al tratar la Distribución por Muestreo de  $\bar{x}$ .

Su demostración teórica puede realizarse de la siguiente manera:

Dada una muestra de tamaño  $n$ , donde las observaciones son  $x_1, x_2, \dots, x_n$

Se define: 
$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Tomando Esperanza en ambos miembros:

$$E(\bar{x}) = E\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right)$$

Aplicando propiedades de esperanza:

$$E(\bar{x}) = \frac{1}{n} E(x_1 + x_2 + \dots + x_n) = \frac{1}{n} [E(x_1) + E(x_2) + \dots + E(x_n)]$$

Considerando que las observaciones corresponden a la misma población y  $\mu$  es una constante:

$$E(\bar{x}) = \frac{1}{n} (\mu + \mu + \dots + \mu) = \frac{1}{n} n \mu = \mu$$



2-

$$\theta = M_e$$

$$\hat{\theta} = m_e$$

Consideramos en este caso que el parámetro es la mediana poblacional y el estimador la mediana muestral. Se verifica que  $E(m_e) = M_e$ , cualquiera sea la distribución.

Luego, la mediana muestral es un estimador insesgado de la mediana poblacional.

3-

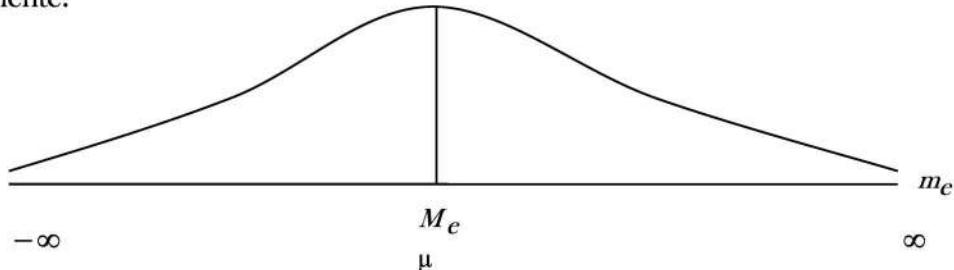
$$\theta = \mu$$

$$\hat{\theta} = m_e$$

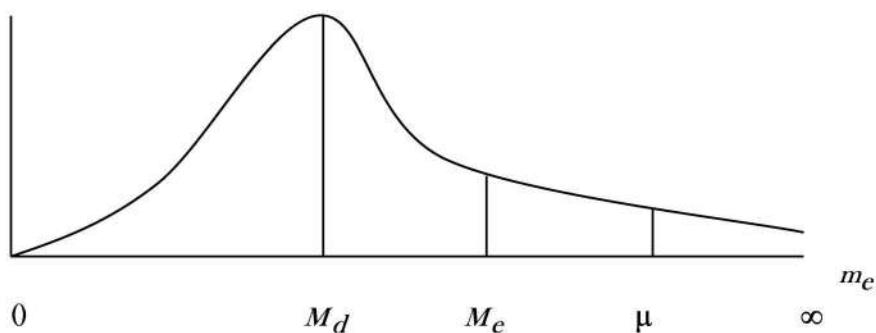
Consideramos en este ejemplo a la mediana de la muestra como estimador de la media poblacional  $\mu$ . Por lo mencionado en 2, sabemos que  $E(m_e) = M_e$ , luego, se pueden presentar tres situaciones:

a- Cuando la distribución de la población sea normal, y en general simétrica, la mediana muestral será un estimador insesgado de la media poblacional  $\mu$ , debido a que en las distribuciones simétricas la media y la mediana son iguales. Es decir:  $M_e = \mu$

Gráficamente:



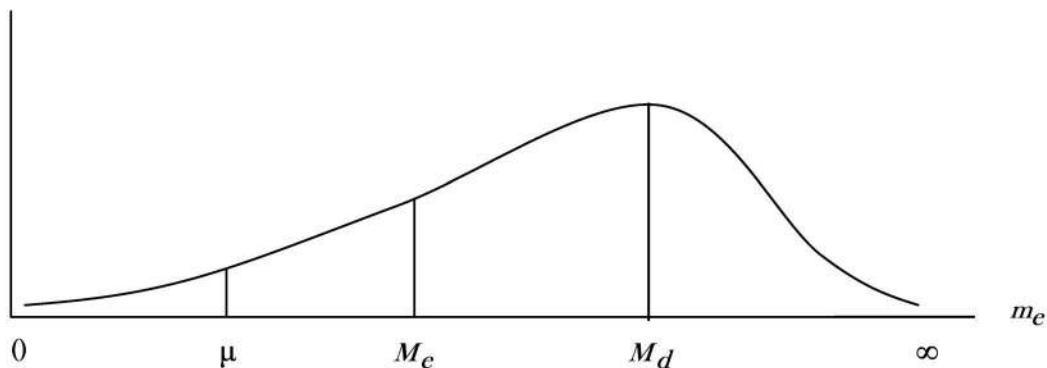
b- Cuando la distribución de la población sea asimétrica derecha, el estimador  $m_e$ , del parámetro  $\mu$ , será sesgado y tendrá un sesgo negativo.



$$E(m_e) = M_e \quad ; \quad \theta = \mu \quad ; \quad \text{luego} \quad E(m_e) = M_e \quad \theta = \mu$$



c- Cuando la distribución de la población sea asimétrica izquierda, el estimador  $m_e$  del parámetro  $\mu$ , será sesgado y tendrá sesgo positivo.



$$E(m_e) = M_e \quad ; \quad \theta = \mu \quad ; \quad \text{luego: } E(m_e) = M_e > \theta = \mu$$

4-

$$\begin{aligned} \theta &= \sigma^2 \\ \hat{\theta}_1 &= s^2 & \hat{\theta}_2 &= \hat{s}^2 \end{aligned}$$

Se desea estimar la varianza poblacional, para lo cual se proponen 2 estimadores:

La varianza muestral ( $s^2$ ) y la varianza muestral corregida ( $\hat{s}^2$ ).

Según demostramos prácticamente en la Distribución Por Muestreo de  $\hat{s}^2$ :

*MCR*                           *MSR*

$$E(\hat{s}^2) = \sigma_X^2 \quad E(\hat{s}^2) = S_X^2$$

Luego, la varianza muestral corregida es un estimador insesgado y la varianza muestral  $s^2$ , es un estimador sesgado de la varianza poblacional  $\sigma^2$ . Además:

$$E(s^2) = \frac{n-1}{n} \sigma_X^2 \quad \text{Siendo } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

El sesgo es negativo.

Demostración

$$E(\hat{s}^2) = \sigma_X^2, \quad \text{pero} \quad \hat{s}^2 = s^2 \cdot \frac{n}{n-1}$$



$$E(s^2 \frac{n}{n-1}) = \sigma_x^2$$

La esperanza de una variable por una constante, es igual a la constante por la esperanza de la variable.

$$E(s^2) \left( \frac{n}{n-1} \right) = \sigma_x^2$$

Luego:

$$E(s^2) = \left( \frac{n-1}{n} \right) \sigma_x^2$$

La demostración teórica de que  $E(\hat{s}^2) = \sigma_x^2$ , puede realizarse considerando:

$$E(s^2) = \left( \frac{n-1}{n} \right) \sigma_x^2$$

Si multiplicamos ambos miembros por el factor  $\frac{n}{n-1}$  se obtiene:

$$\frac{n}{n-1} E(s^2) = \sigma_x^2$$

Haciendo uso de la propiedad de la Esperanza de una constante por una variable:

$$E\left(\frac{n}{n-1}s^2\right) = \sigma_x^2 \quad \text{Y siendo } \hat{s}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = s^2 \frac{n}{n-1}$$

$$E(\hat{s}^2) = \sigma_x^2$$

5-

$$\theta = P$$

$$\hat{\theta} = \hat{P}$$

Se demostró al tratar prácticamente la Distribución Por Muestreo De  $\hat{P}$  que  $E(\hat{P}) = P$ , luego  $\hat{P}$  (Proporción Muestral) es un estimador insesgado de  $P$ , Proporción Poblacional.

En resumen, son estimadores insesgados para los correspondientes parámetros, cualquiera sea la distribución:

$\theta$	$\hat{\theta}$
$\mu$ , media poblacional	$\bar{x}$ , media muestral
$P$ , proporción poblacional	$\hat{P}$ , proporción muestral
$\sigma_x^2$ , varianza poblacional	$\hat{s}^2$ , varianza muestral corregida



Ya se explicó que los estimadores son variables aleatorias, pues para cada una de las muestras posibles, pueden asumir algunos valores diferentes, y cada valor o intervalo de valores posibles tiene asociada una probabilidad, conformando en conjunto la Distribución por muestreo, y por lo tanto puede calcularse su Esperanza y su Desviación.

El hecho de que la Esperanza del Estimador sea igual al parámetro a estimar, por supuesto no asegura, ni aproximadamente, que de la estimación a realizar surgirá un resultado igual al parámetro, sólo se trata de una propiedad teórica que afirma que al tomar todas las muestras posibles de un mismo tamaño, de una población determinada, el promedio de los valores del estimador será igual al valor del parámetro.

En realidad, lo que interesa de un estimador, no es tanto que su valor esperado sea igual al parámetro, porque en cada realización puede resultar muy alejado de ese valor, sino de su *Variabilidad*, es decir el desvío estándar (o error estándar) del estimador.

En efecto siempre que el valor del estimador no esté muy lejos del parámetro (aunque exista alguna diferencia), será preferible entre dos estimadores posibles, aquel que tenga menor variabilidad, puesto que en este caso las probabilidades de que en cada muestra el estimador esté más cercano al parámetro serán mayores. En efecto parece razonable intuitivamente desear un estimador que proporcione estimaciones que se acumulen o concentren alrededor de la característica de la población que debe ser estimada (parámetro).

## 2.2. Eficiencia

Esta propiedad es muy importante porque se refiere precisamente a la variabilidad de los estimadores a la que venimos haciendo referencia.

La varianza de un estimador proporciona una idea del grado de confianza que se puede tener en el mismo.

Si dos estimadores son insesgados, o si uno involucra un pequeño sesgo, preferimos el estimador que tenga menor variabilidad.

Esto nos lleva a definir la *Eficiencia Relativa*.

Dados dos estimadores de un mismo parámetro, insesgados, se dice que es relativamente más eficiente el que tenga menor dispersión, o sea, aquel cuya distribución por muestreo tenga la varianza menor, pues está más concentrado alrededor de la media.

De esta manera los errores de estimación (o de muestreo) son menos probables, pues los valores para la estimación se encuentran más cerca del parámetro que se estima.

Sean dos estimadores  $\hat{\theta}_1$  y  $\hat{\theta}_2$ , insesgados, para un mismo parámetro  $\theta$ , la eficiencia relativa se define como el cociente entre las varianzas de estos estimadores:



$$E_f = \frac{\sigma_{\hat{\theta}_1}^2}{\sigma_{\hat{\theta}_2}^2}$$

Conduciendo a tres situaciones posibles:

$$E_f = \begin{cases} = 1 & \text{Ambos son igualmente eficientes} \\ < 1 & \hat{\theta}_1 \text{ Es relativamente más eficiente que } \hat{\theta}_2 \\ > 1 & \hat{\theta}_2 \text{ Es relativamente más eficiente que } \hat{\theta}_1 \end{cases}$$

### Ejemplo

$$\theta = \mu$$

Para una distribución Normal y en general simétrica

$$\hat{\theta}_1 = \bar{x} \quad \hat{\theta}_2 = m_e$$

Por otro lado:

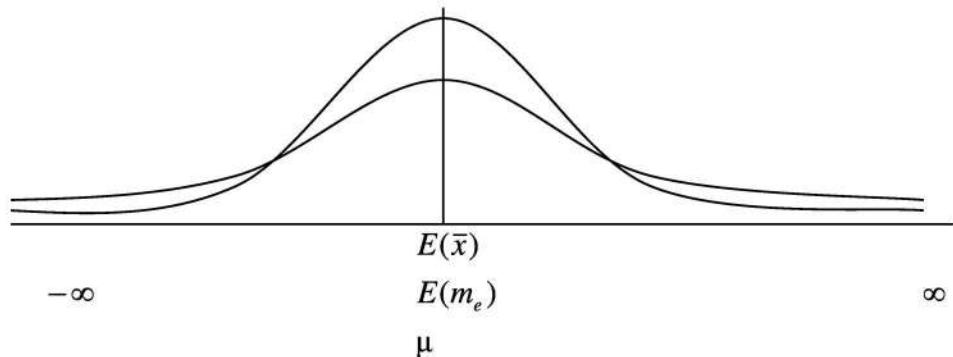
$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} \quad \text{y} \quad \sigma_{m_e}^2 = \frac{\sigma_x^2}{n} \frac{\pi}{2}$$

Luego:

$$Ef = \frac{\sigma_{\bar{x}}^2}{\sigma_{m_e}^2} = \frac{\frac{\sigma_x^2}{n}}{\frac{\sigma_x^2}{n} \frac{\pi}{2}} = \frac{2}{\pi} \cong 0,64$$

Este resultado indica que la varianza de la media muestral es el 64% de la varianza de la mediana de la muestra, luego la distribución muestral de la media está más concentrada alrededor de  $\mu$  que la distribución muestral de la mediana.

Gráficamente:





### Varianza Mínima

Si en lugar de comparar dos estimadores, consideramos en general, el estimador que tenga varianza mínima, éste puede utilizarse para medir la eficiencia, y se dice que este estimador de varianza mínima es un estimador eficiente.

Así por ejemplo, mediante la aplicación del *Teorema De Rao-Cramer* se demuestra que si  $\hat{\mu}$  es un estimador de la media de la población  $\mu$ , la varianza de  $\hat{\mu}$  no puede ser inferior a  $\sigma^2/n$ , que es la varianza de la media muestral:

$$\sigma_{\hat{\mu}}^2 \geq \frac{\sigma_x^2}{n}$$

Pero dado que:

$$\frac{\sigma_x^2}{n} = \sigma_x^2$$

Resulta que entre los estimadores de la media de la población,  $\bar{x}$  es el que tiene menos varianza. Podemos decir, en consecuencia, que  $\bar{x}$  es un estimador insesgado y eficiente, o de varianza mínima de  $\mu$ .

### 2.3. Consistencia

Un estimador es consistente si para muestras grandes hay una probabilidad cercana a uno, de que la estimación esté cerca de la *Característica De La Población* (Parámetro), que se desea estimar.

Es decir, el estimador  $\hat{\theta}$  del parámetro  $\theta$  es consistente si para  $n \rightarrow N$  (o a  $\infty$ ), se verifica que:  $\hat{\theta} \rightarrow \theta$

Es decir, si el estimador se aproxima al parámetro, a medida que el tamaño de la muestra se aproxima al tamaño de la población.

En otras palabras, se dice que un estimador  $\hat{\theta}$  es consistente, si el límite de la probabilidad de que  $|\hat{\theta} - \theta| < \delta$  es uno cuando la muestra tiende a  $\infty$ .

$$\lim_{n \rightarrow \infty} \Pr \{ |\hat{\theta} - \theta| < \delta \} = 1$$

La ley de los grandes números proporciona una evidencia de la consistencia de los estimadores media y proporción muestral, ya que establece:

$$\lim_{n \rightarrow \infty} \Pr \{ |\bar{x} - \mu| < \delta \} = 1$$

Y



$$\lim_{n \rightarrow \infty} \Pr\{ |\hat{P} - P| < \delta \} = 1$$

No obstante, intuitivamente vemos que a medida que el tamaño de la muestra se aproxima al tamaño de la población, la media de la muestra tiende a aproximarse a la media de la población, de tal manera que cuando  $n \rightarrow N$  ambas medidas son iguales y por consiguiente  $\bar{x}$  es un estimador consistente de  $\mu$ . Idéntico razonamiento se hace para  $\hat{p}$ .

Lo mismo sucede con la mediana de la muestra,  $m_e$ , como estimador de la mediana poblacional,  $M_e$ , a medida que el tamaño de la muestra crece, la mediana de la muestra tiende a la mediana de la población.

Dada la relación que existe entre  $\mu$  y  $M_e$ , el estimador  $m_e$ , será un estimador consistente de  $\mu$ , cuando la distribución de la población sea simétrica, en caso contrario el estimador citado no será consistente o bien será inconsistente.

Con respecto a la varianza muestral,  $s^2$ , si bien no es insesgado, sí es consistente como estimador de la  $\sigma_x^2$ , ya que intuitivamente vemos que a medida que se tiende a  $N$ , la varianza muestral ( $s^2$ ), tiende a la varianza poblacional.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

En cuanto al estimador insesgado  $\hat{s}^2$ , es también un estimador consistente, puesto que cuando  $n \rightarrow N$ ,  $n$  es suficientemente grande como para el cociente  $\frac{n}{n-1}$  sea prácticamente igual a la unidad.

$$\hat{s}^2 = \frac{n}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

## 2.4. Suficiencia

Un estimador suficiente, es un estimador que utiliza toda la información contenida en la muestra, respecto al parámetro a estimar.

La  $\bar{x}$ ,  $\hat{s}^2$  y  $\hat{P}$ , son estimadores suficientes de los parámetros  $\mu$ ,  $\sigma^2$  y  $P$ , respectivamente, pues en su cálculo intervienen todas las observaciones muestrales.

Por último, diremos que un estimador es asintóticamente normal, si además de ser insesgado y eficiente, cumple con la propiedad de tener distribución normal cuando el tamaño de la muestra se incrementa.



### **3. ESTIMACIÓN PUNTUAL. LIMITACIONES**

Cualquier estimador puntual, aún cuando tenga todas las propiedades deseadas, tiene la limitación de que no proporciona información acerca de la precisión de la estimación obtenida, esto es, de la magnitud del error debido al muestreo.

Cuando analizamos la distribución por muestreo de  $\bar{x}$ , por ejemplo, se observó que el promedio basado en una muestra aleatoria simple, puede tomar muchos valores diferentes, indicados en dicha distribución de  $\bar{x}$ . Además, estos valores posibles de  $\bar{x}$  están agrupados en la distribución por muestreo de  $\bar{x}$ , alrededor de la verdadera media de población,  $\mu$ , y la mayoría de las  $\bar{x}$  caen a la derecha o a la izquierda de la media de población. Por lo tanto, es bastante seguro que cualquier  $\bar{x}$ , es decir, cualquier estimación puntual, generalmente no proporcionará el valor real de la media poblacional.

¿Qué valor, entonces, tiene la estimación puntual? Realmente tiene poco uso, a menos, que tengamos alguna indicación sobre la precisión de la estimación, la que puede evaluarse por medio de la estimación de un intervalo, que toma como punto de partida a la estimación puntual.

### **4. ERROR, RIESGO Y TAMAÑO DE LA MUESTRA**

Definiremos ahora, el error de estimación y el riesgo de cometer un error superior a un determinado valor, para luego analizar la relación entre ellos y el tamaño de la muestra.

Si para estimar a  $\theta$ , utilizamos a  $\hat{\theta}$ , que es insesgado y tiene distribución normal, es decir que:

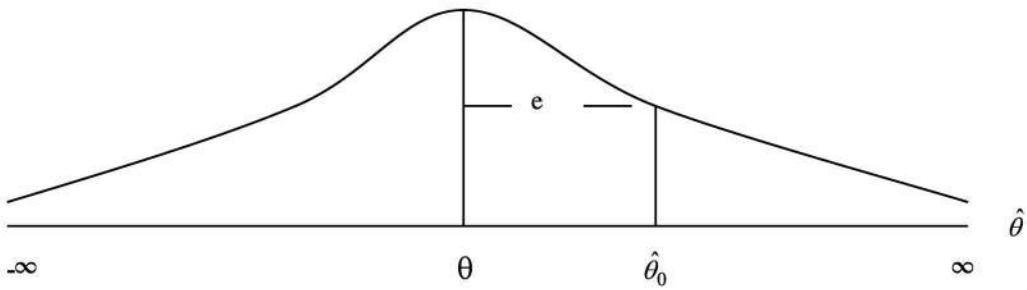
$$E(\hat{\theta}) = \theta \quad \text{por ser insesgado} \quad y \quad \hat{\theta} \sim N(\theta, \sigma_{\hat{\theta}})$$

$\hat{\theta}$  Es una variable aleatoria, según ya se explicó y puede asumir un conjunto de valores posibles según las muestras que puedan tomarse de la población bajo estudio. Cuando se toma una muestra,  $\hat{\theta}$  asume un valor particular, que es la estimación puntual, por ejemplo  $\hat{\theta}_0$ , que se utiliza para estimar a  $\theta$ . Entonces, la diferencia entre esta estimación puntual y el parámetro, se llama *Error De Muestreo*.

Simbólicamente:

$$e = \hat{\theta}_0 - \theta$$

Gráficamente:



Ahora bien, si el error máximo aceptable es, en valor absoluto,  $|e|$ , que expresamos por:

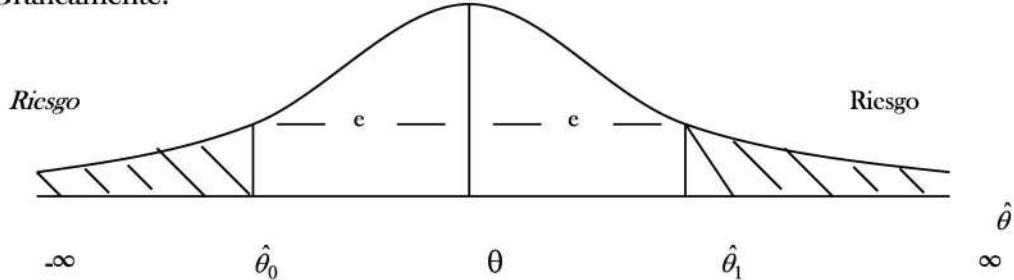
$$|\hat{\theta}_0 - \theta| \leq e$$

Habrá un riesgo de cometer un error superior a él, que se mide por la probabilidad de que se produzca un error superior a  $e (|\hat{\theta}_0 - \theta| > e)$ .

En símbolos:

$$\Pr \{ |\hat{\theta}_0 - \theta| > e \} = \text{Riesgo de cometer un error mayor a } e.$$

Gráficamente:



Donde  $\hat{\theta}_0$  y  $\hat{\theta}_1$  son los máximos valores particulares del estimador que se está dispuesto a aceptar, pues producen un error dentro del nivel que se admite. Pero, ¿Qué ocurre si los estimadores asumen valores superiores a estos límites? Se producirá un nivel de error que no estamos dispuestos a aceptar. Luego, la probabilidad de que aparezcan estos valores para el estimador, es nuestro riesgo de cometer un error superior al aceptable.

Para calcular la probabilidad mencionada debemos previamente estandarizar la variable, a través de:

$$z = \frac{\hat{\theta}_0 - \theta}{\sigma_{\hat{\theta}}} = \frac{e}{\sigma_{\hat{\theta}}}$$



Identificaremos estos conceptos a partir de la Distribución por Muestreo de  $\bar{x}$  suponiendo  $MCR$ , que construyéramos en la Unidad 8, donde  $\mu = 3 = E(\bar{x})$ .

	$\bar{x}_i$	$P(\bar{x}_i)$
	1,0	1/36
	1,5	2/36
	2,0	5/36
	2,5	6/36
$\mu = E(\bar{x})$	3,0	8/36
	3,5	6/36
	4,0	5/36
	4,5	2/36
	5,0	1/36
		<b>36/36</b>

Se observa, en este caso particular, que el mínimo valor posible para  $\bar{x}$  es 1,0 y el máximo valor posible es 5,0; produciendo un máximo nivel de error de  $\pm 2$  [(1-3);(5-3)].

Supongamos ahora, que fijamos un máximo nivel de error de  $\pm 0,5$ , ¿cuál es la probabilidad de cometerlo?

$$\text{Simbólicamente: } P\left\{ |\bar{x} - \mu| \leq 0,5 \right\}$$

Para encontrar la respuesta determinemos cuáles valores de  $\bar{x}$  producen una diferencia  $\leq 0,5$  con  $\mu$ , y qué probabilidad tienen.

Así:

$\bar{x}_i$	$P(\bar{x}_i)$
2,5	6/36
3,0	8/36
3,5	6/36
	<b>20/36</b>

$$\text{Luego: } P\left\{ |\bar{x} - \mu| \leq 0,5 \right\} = \frac{20}{36} = \underline{\underline{0,56}}$$

Esta probabilidad indica la confianza de no superar el nivel de error máximo aceptable. Por otro lado, podremos preguntarnos, cuál es la probabilidad de superar este error, es decir:

$$P\left\{ |\bar{x} - \mu| \geq 0,5 \right\} = \frac{16}{36} = \underline{\underline{0,44}}$$

Esta probabilidad mide el riesgo de cometer un error superior al máximo aceptable.



En este caso, contamos con la Distribución de Probabilidades, pero en la práctica sólo tendremos una muestra, entonces fijaremos el máximo nivel de error aceptado y las probabilidades se calcularán según corresponda a través de los modelos especiales de Probabilidad.

En general, para el estimador media muestral la teoría estadística establece dos situaciones:

1- Cuando la muestra se selecciona aleatoriamente de una población con distribución normal, la variable aleatoria media muestral responde a las características de la distribución normal de la población de la que fue extraída.

2- Cuando la muestra se selecciona aleatoriamente de una población sin distribución normal, la distribución de la variable aleatoria media muestral se aproxima a la distribución normal a medida que el tamaño de la muestra es aumentado.

### Ejemplo

Supongamos una población que se distribuye normal, con media igual a 500 y desviación típica igual a 30. Tomamos una muestra de 100 unidades y queremos conocer cuál es el riesgo de cometer un error de estimación de  $\mu$ , superior a 6.

Tenemos los siguientes valores:

$$\text{Población} \sim N(500; 30)$$

$$n = 100$$

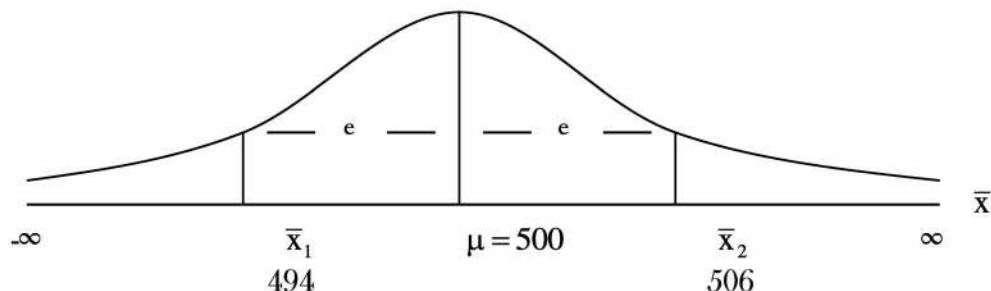
$$|e| > 6$$

$$\hat{\mu} = \bar{x}$$

Hay dos valores particulares de  $\bar{x}$  que nos interesan:

$$\bar{x}_0 = 500 - 6 = 494 \quad y \quad \bar{x}_1 = 500 + 6 = 506$$

Gráficamente:



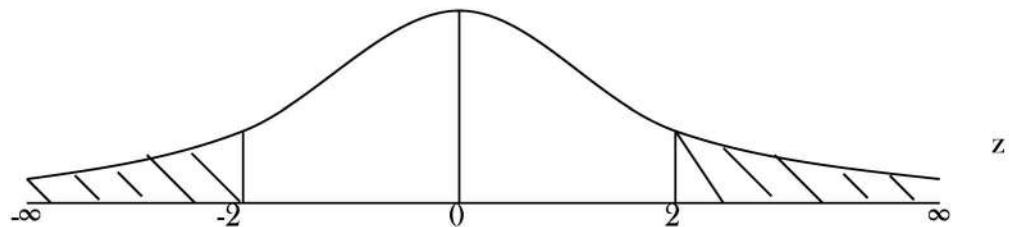
La variable normal tipificada es:



$$z_0 = \frac{\bar{x}_0 - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{494 - 500}{\frac{30}{\sqrt{100}}} = -2$$

y

$$z_1 = \frac{\bar{x}_1 - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{506 - 500}{\frac{30}{\sqrt{100}}} = 2$$



La probabilidad, (o sea el riesgo) de que al efectuar la estimación puntual se cometa un error, en valor absoluto, superior a 6, es decir:  $\Pr\{|\bar{x} - \mu| > 6\}$

Es igual a:

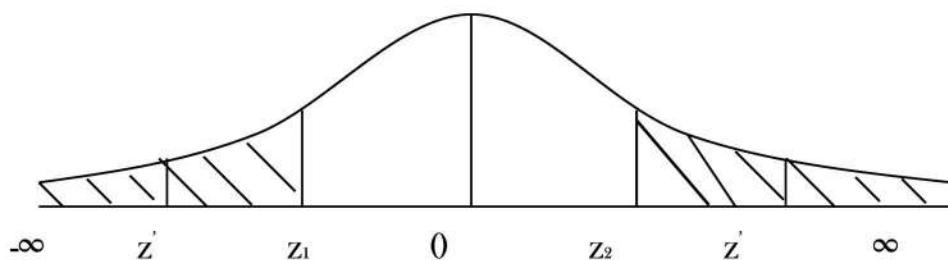
$$\Pr\{|z| > 2\} = \Pr\{z < -2\} + \Pr\{z > 2\} = 2[1 - \Pr\{z < 2\}] = 2(1 - 0.9772) = \underline{0.0456}$$

### Relación entre error, riesgo y tamaño de la muestra

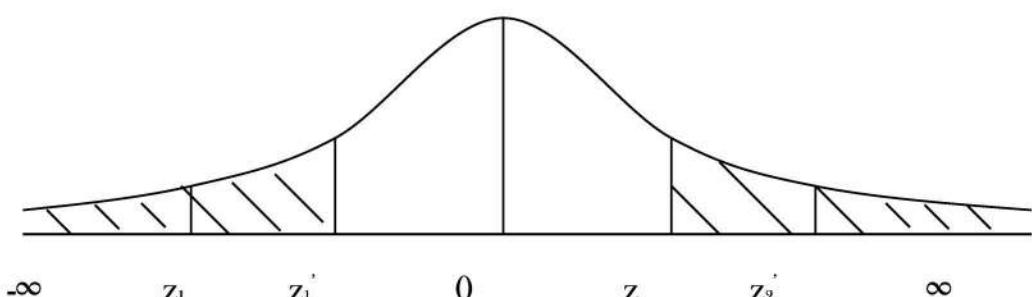
Antes de realizar el análisis, diremos que:

(1) *El riesgo es una función inversa de z:*

Cuando z aumenta, el riesgo disminuye.



Cuando z disminuye, el riesgo aumenta.





(2) Si  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$  depende de  $\begin{cases} \bar{x} - \mu & = \text{error} \\ \sigma \\ n \end{cases}$  El riesgo también depende de ellos

Si  $z$  está en función de los tres elementos mencionados y a su vez  $z$  mantiene relación con el riesgo, entonces el riesgo también depende de esos elementos.

Analizaremos ahora las relaciones, teniendo en cuenta que cuando correspondemos a dos de ellos, el resto permanece constante, de lo contrario no podríamos asegurar las conclusiones.

a) Riesgo

Partiendo de:

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

Deducimos que:

*Si incrementamos el error, incrementa  $z$  y disminuye el riesgo.*

*Si incrementamos el tamaño de la muestra, incrementa  $z$  y disminuye el riesgo.*

*Si incrementamos la desviación, disminuye  $z$  e incrementa el riesgo.*

b) Error

Despejando de:

$$z = \frac{e\sqrt{n}}{\sigma} \quad \text{Obtendremos que: } e = \frac{z\sigma}{\sqrt{n}}$$

Deducimos que:

*Si incrementamos el riesgo, disminuye  $z$  y disminuye el error.*

*Cuanto mayor es la desviación, mayor será el error.*

*Si incrementamos el tamaño de la muestra, disminuye el error.*

c) Tamaño de la muestra

Partiendo de:

$$z = \frac{e\sqrt{n}}{\sigma}$$

Y despejando  $n$ , obtendremos:

$$n = \frac{z^2 \sigma^2}{e^2}$$

Deducimos que:

Si incrementamos el riesgo, disminuye  $z$  y disminuye  $n$ , es decir, requerimos de menor información.

A mayor varianza, mayor tamaño de la muestra.

A mayor error, menor tamaño de la muestra.

En todos los casos las conclusiones serían a la inversa, si partimos de disminuciones.

## ***5. ESTIMACIÓN POR INTERVALOS***

La estimación por intervalos, consiste en obtener un cierto intervalo aleatorio  $[L_l ; L_s]$ , a partir de la estimación puntual, considerando un cierto error de estimación, y un determinado grado de confianza de que el intervalo construido contiene al parámetro que queremos estimar.

### ***5.1. Ejemplo Especial***

Supongamos que obtenemos una estimación por intervalo de la media de la población en la siguiente forma: seleccionamos una muestra aleatoria simple de la población y calculamos la media muestral  $\bar{x}$ .



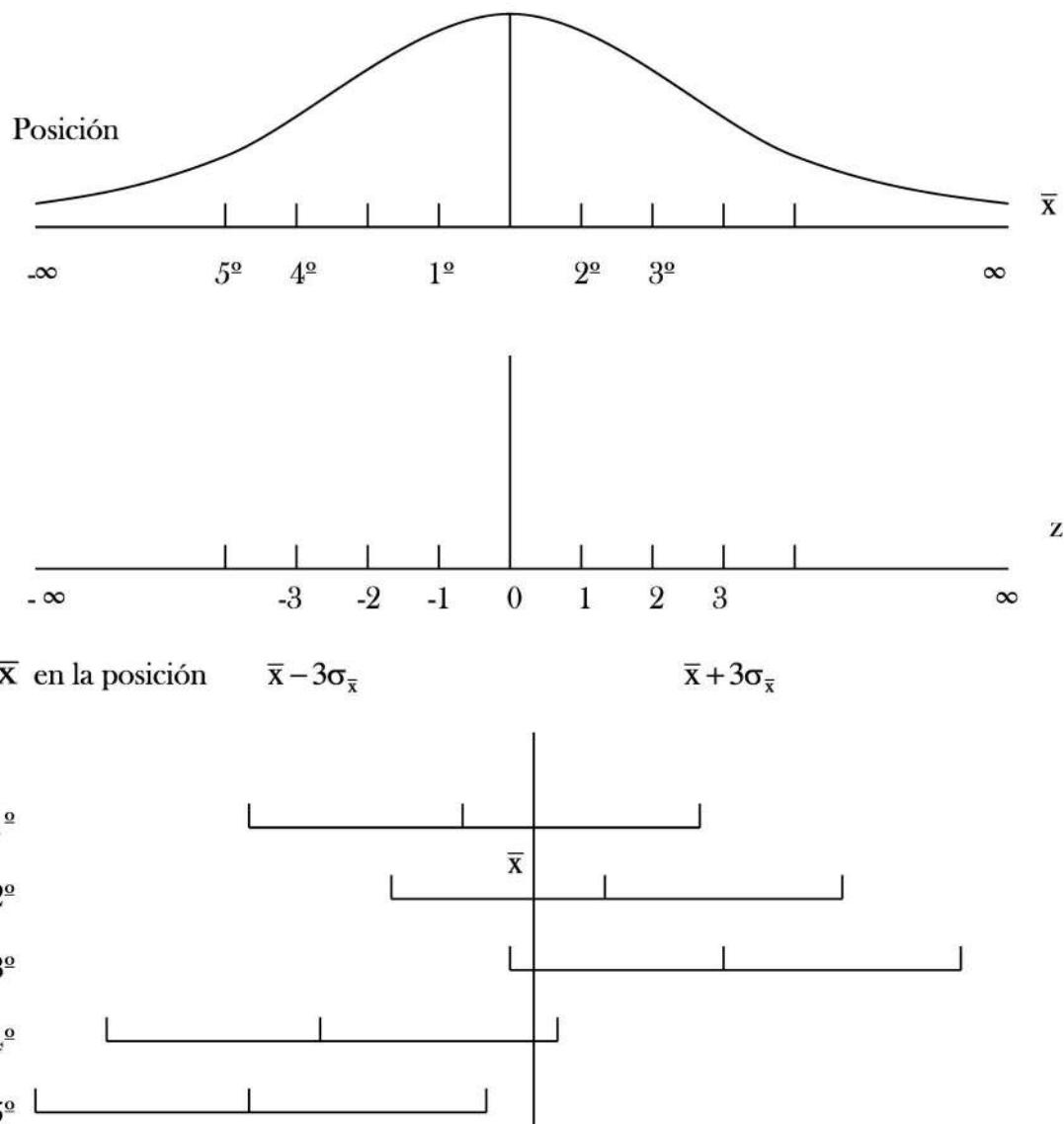
Después, afirmamos que la media de la población se encuentra entre  $\bar{x} - 3\sigma_{\bar{x}}$  y  $\bar{x} + 3\sigma_{\bar{x}}$ , siendo el primero, el límite inferior del intervalo y el segundo, el límite superior del intervalo.

¿Cuáles son las implicancias de esta estimación por intervalos?

Para responder, debemos referirnos a la distribución por muestreo de  $\bar{x}$ , puesto que estamos interesados en su comportamiento.

Suponiendo que el tamaño de muestra es suficiente grande, la distribución de  $\bar{x}$  es aproximadamente normal con  $(\mu, \sigma_{\bar{x}})$ .

Veamos la siguiente gráfica:



(Pag. 277 Neter)

Distribución Por Muestreo De  $\bar{x}$  y Diferentes Intervalos De Confianza Posibles.



Obsérvese que la escala de  $\bar{x}$  se encuentra en blanco, pues sólo conocemos a partir de la teoría estadística, que la  $E(\bar{x})=\mu$ , pero no conocemos el valor de  $\mu$ .

Debajo de la escala  $\bar{x}$  se encuentra la escala z, que indica la distancia desde la media, en unidades de desviación estándar de  $\bar{x}$  ( $\sigma_{\bar{x}}$ ).

Recordemos que:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \sim N(0,1)$$

Así, +1 en la escala z, significa que este punto se encuentra  $1\sigma_{\bar{x}}$  arriba de la media de la distribución de muestreo de  $\bar{x}$ , que es  $\mu$ ; +2 en la escala z, indica que este punto se encuentra  $2\sigma_{\bar{x}}$  arriba de  $\mu$ , y así sucesivamente.

Si usamos la estimación por intervalos  $\bar{x} \pm 3\sigma_{\bar{x}}$  para estimar  $\mu$ , la estimación particular por intervalos obtenida sobre la base de una muestra aleatoria simple, depende de la localización de  $\bar{x}$  en la distribución de todas las  $\bar{x}$  posibles. La gráfica indica cinco posibles posiciones en las cuales  $\bar{x}$  obtenida de una muestra particular, podría caer en la distribución por muestreo de  $\bar{x}$ .

Supongamos que el valor de  $\bar{x}$  corresponde a la posición 1º. Aquí, la estimación del intervalo estaría indicada por el segmento centrado en  $\bar{x}$ , en la posición 1. Los límites del intervalo se obtienen agregando a  $\bar{x}$  la longitud que representa  $3\sigma_{\bar{x}}$ , obtenida de la escala z y restando de  $\bar{x}$  la longitud que representa  $3\sigma_{\bar{x}}$ . Los límites obtenidos en esta forma corresponden entonces a  $\bar{x} + 3\sigma_{\bar{x}}$  y  $\bar{x} - 3\sigma_{\bar{x}}$ , que es el intervalo estimado que queríamos encontrar.

Nótese que este intervalo cubre la media de la población  $\mu$ . Luego, si la media de la muestra cae en la posición 1º, el intervalo  $\bar{x} \pm 3\sigma_{\bar{x}}$  incluirá a  $\mu$ , y la aseveración de que  $\mu$  se encuentra dentro del intervalo será correcta.

Supongamos que la media de la muestra cae en la posición 2º, luego los límites  $\bar{x} + 3\sigma_{\bar{x}}$  son construidos alrededor de  $\bar{x}$  localizada en la posición 2. Nuevamente  $\mu$  está dentro del intervalo. Lo mismo ocurre para el caso en que  $\bar{x}$  se localiza en la posición 3º y 4º.

¿Deja alguna vez el intervalo  $\bar{x} \pm 3\sigma_{\bar{x}}$  de incluir a  $\mu$ ?

Sí. Puesto que los límites se obtienen sumando y restando  $3\sigma_{\bar{x}}$  a  $\bar{x}$ , el intervalo no incluirá a  $\mu$ , siempre que  $\bar{x}$  caiga más allá de tres desviaciones estándar ( $3\sigma_{\bar{x}}$ ) de la media  $\mu$ , es decir, siempre que  $\bar{x}$  caiga arriba de +3 o debajo de -3 en la escala z, situación que ocurre cuando  $\bar{x}$  cae en la posición 5º.



Entonces, la declaración, “el valor de la media se encuentra en algún punto entre  $\bar{x} \pm 3\sigma_{\bar{x}}$ ”, será correcta si la media de la muestra  $\bar{x}$  cae a una distancia menor que  $3\sigma_{\bar{x}}$  de la media de la población.

¿Cuál es la probabilidad de que  $\bar{x}$  caiga dentro de ese intervalo?

La distribución de  $\bar{x}$  es aproximadamente normal, así que sabemos que la probabilidad de que  $\bar{x}$  caiga a menos de  $3\sigma_{\bar{x}}$ , de la media de la distribución por muestreo de  $\bar{x}$  que es igual a  $\mu$ , es 0,997. Así pues, si tomamos un gran número de muestras aleatorias simples, suficientemente grandes, calculamos el intervalo  $\bar{x} \pm 3\sigma_{\bar{x}}$  para cada una y en cada caso declaramos que la media de la población se encuentra dentro del intervalo, alrededor del 99,7 % de estas aseveraciones serán correctas.

La estimación del intervalo para la media  $\mu$  puede escribirse más formalmente en la siguiente forma:  $\bar{x} - 3\sigma_{\bar{x}} \leq \mu \leq \bar{x} + 3\sigma_{\bar{x}}$

La cual se lee: la media de la población  $\mu$  se encuentra en algún punto entre  $\bar{x} - 3\sigma_{\bar{x}}$  y  $\bar{x} + 3\sigma_{\bar{x}}$ . Este intervalo recibe el nombre de: *Intervalo De Confianza Para La Media De La Población*, y la probabilidad de una aseveración correcta, en este caso 0,997 recibe el nombre de Coeficiente De Confianza o Nivel De Confianza, si se expresara en términos de porcentaje, es decir, 99,7%.

## 5.2. Elementos que Intervienen y Terminología a Utilizar

Simbología	Descripción
$\theta$	Parámetro A Estimar
$x_1, x_2, \dots, x_n$	n Observaciones Muestrales De Una Población
$\hat{\theta} = f(x_1, x_2, \dots, x_n)$	Estimador De $\theta$ y Como Tal Una Variable Aleatoria.
$K = K(\hat{\theta}, \theta)$	Estadístico, Que Por Ser Función De $\hat{\theta}$ Es Una Variable Aleatoria. Depende Matemáticamente Del Parámetro $\theta$ , El Cual Está Indicado Implicítamente En Él; $K(\hat{\theta}, \theta)$ Se Elije De Manera Que Tenga Una Función De Probabilidad Conocida Y Tabulada.
$1 - \alpha = \Pr\{K_1 \leq K(\hat{\theta}, \theta) \leq K_2\}$	Nivel De Confianza O Coeficiente De Confianza.
$K_1, K_2$	Coeficientes De Confianza (Si $1 - \alpha$ Es Llamado Nivel De Confianza) o Puntos Críticos (Si $1 - \alpha$ Es Llamado Coeficiente De Confianza). Se Obtienen De La Tabla De La Distribución De K. Son Función De $1 - \alpha$ . Al Fijar $1 - \alpha$ Quedan Determinados $K_1$ Y $K_2$ .
$L_1, L_2$	Límites Del Intervalo De Confianza Para Estimar $\theta$ . Se Obtienen Al Despejar $\theta$ En La Desigualdad.



De tal manera que:

$$L_i \leq \theta \leq L_s$$

$L_i$  y  $L_s$  son variables aleatorias y sus valores dependen de las observaciones muestrales, de los coeficientes de confianza, del nivel de confianza y del tamaño de la muestra.

Cabe aclarar, que la mayoría de los textos llama a  $1 - \alpha$  coeficiente de confianza y a  $(1 - \alpha) 100$ , nivel de confianza. En tales casos se suele designar a  $K_1$  y  $K_2$  como "puntos críticos" o "multiplicador de confianza".

Veremos la forma de construcción de intervalos de confianza para los parámetros media poblacional, proporción poblacional y varianza poblacional.

También pueden determinarse intervalos para otros parámetros, tales como, Diferencia de Medias, Diferencia de Proporciones y Cociente de Varianzas, que no trataremos en este curso.

### 5.3. Nivel De Confianza. Significado Y Selección

#### Significado

Si consideramos una muestra aleatoria de 50 empleados, para estudiar la antigüedad media de los 2000 empleados de una compañía, y se encuentra que con un nivel de confianza del 99,7%, el intervalo es  $4,1 \leq \mu \leq 7,9$  es decir, que la antigüedad media tenía un valor que se encontraba entre 4,1 y 7,9 años.

Sin embargo, esta aseveración con un nivel de confianza de 99,7%, no significa que exista una probabilidad de 0,997 de que la media poblacional tenga un valor que se encuentre entre 4,1 y 7,9 años. El tiempo de servicio promedio de todos los empleados es algún número definido y aún desconocido. Puede ser o no un valor que se encuentre entre 4,1 y 7,9 años; aquí no hay ninguna probabilidad involucrada, pues no existe variable aleatoria en la expresión, ya que en el centro del intervalo está un parámetro, y en los extremos se tienen números obtenidos, sumando y restando el error (medido en términos de desviación estándar) al valor puntual.

Es por ello, que la expresión final es:

$$L_i \leq \theta \leq L_s, \text{ con un } 99,7\%$$

En lugar de  $P\{L_i \leq \theta \leq L_s\} = 0,95$ .

Que la declaración de que el tiempo promedio de servicios se encuentre entre 4,1 y 7,9 años sea correcta o equivocada, dependerá, de si la media de la muestra cae dentro de  $\pm 3\sigma_{\bar{x}}$ , a un lado y otro de  $\mu$ ; esto no lo sabemos.

Sabemos, sin embargo, que tomando un gran número de muestras aleatorias simples de esta población y calculando un intervalo de confianza con un nivel del 99,7% para cada una, alrededor del 99,7% de estas declaraciones serán correctas.



Es decir, sobre 100 muestras aleatorias de un cierto tamaño  $n$  de una población, si en cada una se calcula la media muestral y a partir de ella, se construyen 100 intervalos de confianza para el parámetro que se desea estimar, 99,7 contendrán al verdadero valor del parámetro poblacional, mientras que 0,3 no lo contendrán.

### Selección

En cualquier situación dada, mientras menor sea el nivel de confianza más estrecho será el intervalo de confianza; mientras mayor sea el nivel de confianza, más amplio será el intervalo de confianza.

La generalización establecida indica que niveles de confianza más pequeños nos llevarán a intervalos más estrechos, lo cual es conveniente, pero también significa una mayor probabilidad de que un intervalo sea incorrecto, lo cual, por supuesto, es inconveniente.

De hecho, un intervalo de confianza basado en un tamaño de muestra, solo puede ser estrechado aumentando el riesgo de una estimación incorrecta.

Por otra parte, el riesgo de una declaración incorrecta puede hacerse más pequeño aumentando el nivel de confianza. Tal incremento, sin embargo, tiene el efecto de ampliar el intervalo de confianza en cualquier situación dada, y en esta forma puede hacer que el intervalo sea menos útil.

Puesto que la amplitud del intervalo se relaciona con el riesgo de que la estimación del intervalo sea incorrecta, los usuarios de las estimaciones de muestras, deberán especificar este riesgo previamente, sobre la base de la situación del problema para el cual se requiere la estimación.

### 5.4. Intervalo De Confianza Para Estimar $\mu$ . Uso de la Distribución Normal y "t" de Student.

Seguiremos la siguiente regla para determinar la distribución aplicable:

1) Poblaciones Normales		2) Poblaciones No Normales	
$\sigma^2$ Conocida $n$ Cualquiera	Distribución Normal	$\sigma^2$ Conocida $n \geq 30$	Por TCL Normal
$\sigma^2$ Desconocida		$\sigma^2$ Desconocida	
$n < 30$	"t" De Student	$n < 30$	—
$n \geq 30$	Por TCL Normal	$n \geq 30$	Por TCL Normal



Es frecuente que para estimar un parámetro se utilice el estimador análogo. Así, para estimar la media poblacional  $\mu$ , utilizaremos la media muestral  $\bar{x}$ .

Cuando la población es normal  $(\mu, \sigma)$ , la media muestral  $\bar{x}$ , tiene una distribución normal  $\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$  y lo mismo sucede en las poblaciones no normales cuando el tamaño de la muestra  $n$ , es suficientemente grande.

El estadístico  $K(\hat{\theta}, \theta)$ , con distribución conocida, generalmente tabulada, que se utiliza para estimar la media poblacional es:

$$K(\bar{x}, \mu) = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Si la varianza poblacional  $\sigma^2$  es conocida, la expresión anterior es una Normal  $(0,1)$ , pero si  $\sigma^2$  es desconocida hay que estimarla con los datos de muestra, mediante:

$$\hat{s}^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

En tal caso el estadístico:

$$K(\bar{x}, \mu) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

Es el cociente de dos variables aleatorias, y en consecuencia no tiene distribución normal, salvo que  $n$  sea grande, porque entonces adquiere valores próximos a  $\sigma$ .

Cuando  $n$  sea pequeña el estadístico  $\frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$  tendrá una distribución  $t_{n-1}$ .

### Caso 1: Poblaciones Normales

#### a) Varianza Poblacional Conocida. Muestras de cualquier tamaño

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$



$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Luego, planteamos que el nivel de confianza  $1 - \alpha$ , es:

$$1 - \alpha = \Pr\{z_1 \leq K(\bar{x}, \mu) \leq z_2\}, \text{ es decir}$$

$$1 - \alpha = \Pr\left\{z_1 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_2\right\}$$

A partir de esta expresión despejamos al parámetro  $\mu$ :

$$1 - \alpha = \Pr\left\{z_1 \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq z_2 \frac{\sigma}{\sqrt{n}}\right\}$$

$$1 - \alpha = \Pr\left\{z_1 \frac{\sigma}{\sqrt{n}} - \bar{x} \leq -\mu \leq z_2 \frac{\sigma}{\sqrt{n}} - \bar{x}\right\}$$

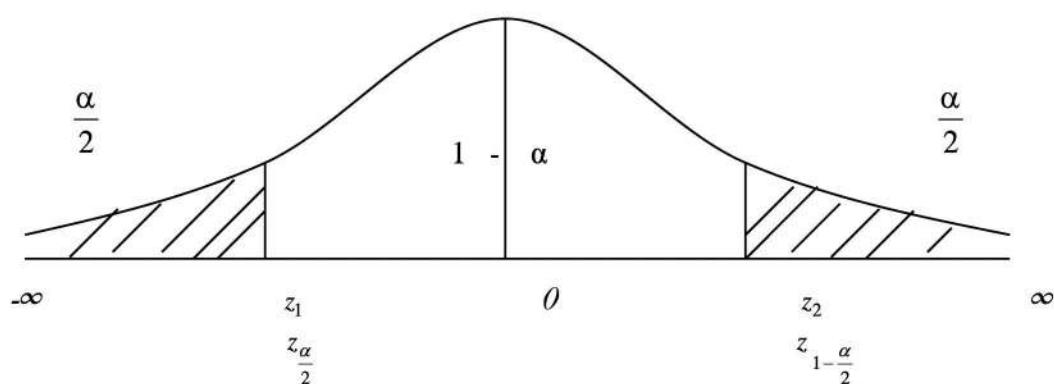
Como el parámetro a estimar es  $\mu$ , multiplicamos por (-1), invirtiendo el signo de la desigualdad:

$$1 - \alpha = \Pr\left\{\bar{x} - z_1 \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x} - z_2 \frac{\sigma}{\sqrt{n}}\right\}$$

Colocamos el término menor a la izquierda y el mayor a la derecha:

$$1 - \alpha = \Pr\left\{\bar{x} - z_2 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} - z_1 \frac{\sigma}{\sqrt{n}}\right\}$$

Gráficamente:





Por simetría  $-z_{1-\frac{\alpha}{2}}$

Donde el subíndice de z indica la probabilidad acumulada hasta ese punto.

$$\frac{\alpha}{2} = \Pr\left\{z \leq z_{\frac{\alpha}{2}}\right\} \quad \text{y} \quad 1 - \frac{\alpha}{2} = \Pr\left\{z \leq z_{1-\frac{\alpha}{2}}\right\}$$

Reemplazando:

$$1 - \alpha = \Pr\left\{\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right\}$$

$$1 - \alpha = \Pr\left\{\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} - (-z_{1-\frac{\alpha}{2}}) \frac{\sigma}{\sqrt{n}}\right\}$$

Llegamos así al intervalo buscado

$$1 - \alpha = \Pr\left\{\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right\}$$

Este enunciado de probabilidad declara que si se toman de la población muchas muestras aleatorias de tamaño n, y para cada una se hace la aseveración:

$$\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

(1 - α) % de estas aseveraciones serán correctas.

Luego, los límites de la confianza obtenidos son:

$$L_i = \bar{x} - z_2 \frac{\sigma}{\sqrt{n}} \quad \text{y} \quad L_s = \bar{x} - z_1 \frac{\sigma}{\sqrt{n}}$$

Recordemos que:  $z \frac{\sigma}{\sqrt{n}} = e$

Luego:

$$L_i = \bar{x} - e \quad L_s = \bar{x} + e$$



La amplitud del respectivo intervalo aleatorio  $\{L, L\}$  depende del nivel de confianza  $1-\alpha$ , de la varianza poblacional conocida y del tamaño de la muestra. El intervalo tiene siempre la misma amplitud, mientras no se modifique el nivel de confianza, ni el tamaño de la muestra.

Su diferencia nos dará una idea sobre la precisión de la estimación.

La precisión es directamente proporcional al tamaño de la muestra  $n$ , e inversamente proporcional a la desviación estándar ( $\sigma$ ).

Como síntesis, diremos, que para reducir la amplitud de un intervalo de confianza y en consecuencia, aumentar su precisión, debemos reducir el error estándar de la media muestral que es  $\frac{\sigma}{\sqrt{n}}$ . Esto puede lograrse, solamente, disminuyendo la variabilidad de los datos, ya sea homogeneizando el material ó, si esto no puede llevarse a cabo, aumentando el tamaño de la muestra.

El intervalo planteado supone muestreo con reposición.

### Ejemplo

El dueño de una estación de servicio, desea saber la cantidad de nafta diaria que vende, en promedio, por cliente.

Toma una muestra al azar de 36 clientes, y encuentra que, en promedio vendió 15 litros de nafta. Si sabe por estudios anteriores, que la población se distribuye aproximadamente normal, con una desviación estándar de 2 litros, se pide encontrar:

- La estimación puntual de la media poblacional.
- Un intervalo de confianza del 95% para la media de la venta diaria de combustible, en dicha estación de servicio.
- El nivel de error, e interpretar  $1-\alpha$ , relacionándolo con el riesgo.

### Resolución

- La estimación puntual, es el valor que asume el estimador en una muestra dada. En el ejemplo, el parámetro a estimar es el promedio de nafta diario que vende, es decir  $\mu$ , por lo que el estimador a utilizar es  $\bar{x}$ , y su estimación puntual es de 15 litros, valor obtenido para  $\bar{x}$ , en la muestra de 36 clientes.

- Dijimos en a, que :

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$



La distribución a aplicar es la normal, puesto que la población es normal y la varianza poblacional es conocida, sin interesar el tamaño de la muestra.

Luego, el intervalo es:

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\} = 0,95$$

Los datos son:

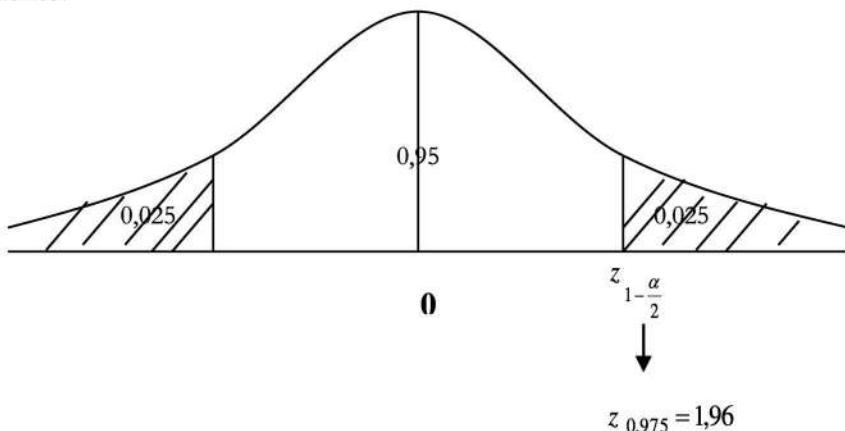
$n = 36$  clientes

$\bar{x} = 15$  litros

$\sigma = 2$  litros

$1 - \alpha = 0,95$ , a partir del cual determinamos el valor de  $z_{1-\frac{\alpha}{2}}$ .

Gráficamente:



Reemplazamos:

$$15 - 1,96 \frac{2}{\sqrt{36}} \leq \mu \leq 15 + 1,96 \frac{2}{\sqrt{36}}$$

$$14,35 \leq \mu \leq 15,65$$

c- Nivel de error:  $z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \pm = \pm 1,96 \frac{2}{\sqrt{36}} = \pm 0,65$

*Interpretación: Por cada 100 intervalos construidos, 95 contendrán al verdadero valor del parámetro, y 5 no lo contendrán. Confiamos en que el intervalo construido, sea uno de los 95, pero corremos el riesgo de que sea uno de los 5 que no contiene al verdadero valor del parámetro. Es decir, existe un riesgo del 5%.*

En caso de muestreo sin reposición, las desviaciones llevarán el factor de corrección de poblaciones finitas, planteándose:

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$



$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}} \sim N(0,1) \quad \text{pues} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Y el intervalo:

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$$

### b) Varianza Poblacional Desconocida. $n < 30$

Como la varianza poblacional es desconocida, debe ser estimada con datos de la muestra. Un estimador insesgado de  $\sigma^2$  es:

$$\hat{s}^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Entonces planteamos:

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \approx t_{n-1}$$

A partir de lo cual, definimos:

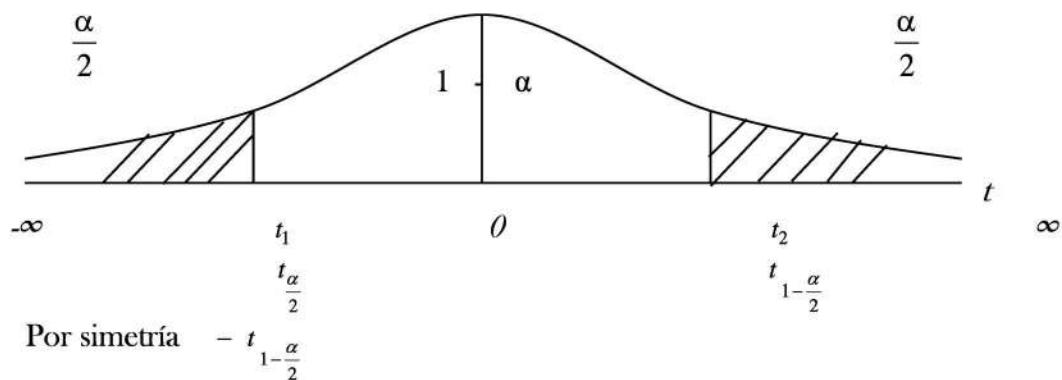
$$1 - \alpha = \Pr \left\{ t_1 \leq \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \leq t_2 \right\}$$

Realizando igual procedimiento que en a, despejamos  $\mu$ , llegando a la siguiente expresión:

$$1 - \alpha = \Pr \left\{ \bar{x} - t_2 \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} - t_1 \frac{\hat{\sigma}}{\sqrt{n}} \right\}$$

La distribución "t" de Student tiene idéntico comportamiento que la distribución normal, es decir, es unimodal, en forma de campana y simétrica, sólo que la curva es más achatada, puesto que se distribuye con media 0 y varianza ligeramente superior a 1.

Gráficamente:



Las entradas a la tabla se realizan considerando los grados de libertad, que se calculan como  $n-1$ .

Entonces:

$$1 - \alpha = \Pr \left\{ \bar{x} - t_{(n-1); 1 - \frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(n-1); \frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\}$$

$$1 - \alpha = \Pr \left\{ \bar{x} - t_{(n-1); 1 - \frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} - (-t_{(n-1); 1 - \frac{\alpha}{2}}) \frac{\hat{\sigma}}{\sqrt{n}} \right\}$$

Llegamos así al intervalo buscado

$$1 - \alpha = \Pr \left\{ \bar{x} - t_{(n-1); 1 - \frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(n-1); \frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\}$$

Son igualmente válidas las consideraciones realizadas en el punto **a**, luego:

$$e = t_{(n-1); 1 - \frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

Y los límites de confianza son:

$$L_i = \bar{x} - t_{(n-1); 1 - \frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \quad \text{y} \quad L_s = \bar{x} + t_{(n-1); \frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

$$L_i = \bar{x} - e$$

$$L_s = \bar{x} + e$$

El muestreo supuesto es con reposición.



### Ejemplo

Se seleccionaron al azar, de un archivo de la UTN, los pesos de 20 empleados varones. La media de la muestra fue de 75 Kg. Se conoce que la población se distribuye aproximadamente normal. La varianza muestral corregida, fue de 16 Kg<sup>2</sup>. Calcular para la media poblacional, un intervalo de confianza del 90%.

### Resolución

$$\theta = \mu$$
$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\hat{\sigma}} \approx t_{n-1}$$
 Pues, la población es normal, la varianza poblacional desconocida y  $n < 30$

$$1 - \alpha = \Pr \left\{ \bar{x} - t_{(n-1); 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(n-1); 1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right\} = 0,90$$

Los datos son:

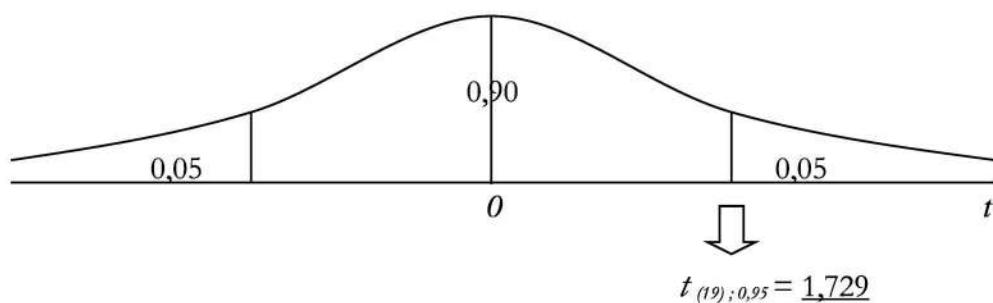
$$\bar{x} = 75 \text{ Kg.}$$

$$n = 20 \text{ empleados}$$

$$\hat{\sigma}^2 = 16 \text{ Kgs}^2 \rightarrow \hat{\sigma} = 4 \text{ Kgs}$$

$$1 - \alpha = 0,90$$

Gráficamente:



Reemplazamos:

$$75 - 1,729 \frac{4}{\sqrt{20}} \leq \mu \leq 75 + 1,729 \frac{4}{\sqrt{20}}$$

$$73,45 \leq \mu \leq 76,55$$



En caso de muestreo Sin Reposición, se plantea:

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}} \approx t_{n-1}$$

$$1 - \alpha = \Pr \left\{ \bar{x} - t_{(n-1); 1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + t_{(n-1); 1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$$

c) Varianza Poblacional Desconocida.  $n \geq 30$

Según vimos en el punto anterior, cuando  $\sigma^2$  es desconocida, la distribución que corresponde es “t” de Student, pero por aplicación del Teorema Central de Límite, cuando n es grande, todas las distribuciones tienden a la Normal. Luego, planteamos:

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim N(0, 1)$$

El Intervalo quedará:

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\}$$

Entonces:

$$e = z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

Y

$$L_I = \bar{x} - e$$
$$L_S = \bar{x} + e$$

El muestreo supuesto es con reposición.



### Ejemplo

Para determinar el rendimiento anual de ciertas acciones, un grupo de inversores, tomó una muestra de 50. La media y la desviación resultaron:  $\bar{x} = 8,71\%$  y  $\hat{\sigma} = 2,1\%$ .

Suponiendo que el rendimiento de esta clase de acciones se distribuye en forma aproximadamente Normal, estimar su verdadero rendimiento anual promedio usando un intervalo de confianza del 90 %.

### Resolución

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim N(0,1)$$

Pues, la Población es Normal, Varianza Poblacional Desconocida y  $n \geq 30$ , luego por aplicación del Teorema Central del Límite aproximamos a la Normal.

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{\frac{1-\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{1-\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\} = 0,90$$

Los datos son:

$$\bar{x} = 8,71\%$$

$$n = 50$$

$$\hat{\sigma} = 2,1\%$$

$$1 - \alpha = 0,90, \text{ de donde } z_{0,95} = 1,645$$

Reemplazando:

$$8,71 - 1,645 \frac{2,1}{\sqrt{50}} \leq \mu \leq 8,71 + 1,645 \frac{2,1}{\sqrt{50}}$$

$$8,22 \leq \mu \leq 9,20$$

Si el muestreo fuera sin reposición se plantea:

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}} \sim N(0,1)$$



Y el intervalo:

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$$

### Caso 2: Poblaciones No Normales

a) Varianza Poblacional Conocida.  $n \geq 30$

$$\theta = \mu$$
$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \quad \text{pues } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\}$$

Supuesto Muestreo con Reposición.

#### Ejemplo

Para llegar a una negociación salarial adecuada, en un determinado sindicato, se requiere una estimación precisa del salario actual de los empleados sindicalizados. Un agente laboral, tomó una muestra de  $n = 60$  empleados sindicalizados, y en ella encontró una media del salario quincenal de \$ 247,45.

Se sabe, de estudios anteriores, que la desviación estándar poblacional es de \$ 21,60. Determinar un intervalo de confianza del 95%, para un salario quincenal promedio de todos los empleados sindicalizados.

#### Resolución

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

No indica población normal, varianza poblacional conocida y  $n \geq 30$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\} = 0,95$$

Los datos son:



$$\bar{x} = 247,45$$

$$n = 60$$

$$\sigma = 21,6$$

$$1-\alpha = 0,95, \text{ de donde } z_{0,975} = 1,96$$

Reemplazando:

$$247,45 - 1,96 \frac{21,6}{\sqrt{60}} \leq \mu \leq 247,45 + 1,96 \frac{21,6}{\sqrt{60}}$$

$$241,98 \leq \mu \leq 252,92$$

Si el muestreo fuera sin reposición, se considera, como en los casos anteriores, el factor de corrección de poblaciones finitas:

$$\theta = \mu$$
$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}} \sim N(0,1)$$

Y el intervalo:

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$$

b) Varianza Poblacional Desconocida.  $n \geq 30$

$$\theta = \mu$$
$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim N(0,1)$$

El Intervalo quedará:

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\}$$

Luego:



Y

$$e = z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

$$L_l = \bar{x} - e \quad L_s = \bar{x} + e$$

Las demostraciones y planteos se realizaron bajo el supuesto de Muestreo Con Reposición

### Ejemplo

En una semana de trabajo determinada se elige al azar una muestra de 300 empleados de una empresa manufacturera. Los trabajadores realizan una labor a destajo y se encuentra que el promedio de pago por pieza trabajada es de \$ 18 con una desviación estándar corregida de \$ 1,4.

Estimar con un nivel del 95% un intervalo de confianza para el pago promedio a destajo, de todos los empleados de la empresa.

### Resolución:

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim N(0, 1) \quad \text{Pues, no indica Población Normal, Varianza Poblacional desconocida}$$

y n grande.

$$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\} = 0,95$$

Los datos son:

$$\bar{x} = 18$$

$$n = 300$$

$$\hat{\sigma} = 1,4$$

$$1 - \alpha = 0,95, \text{ de donde } z_{0,975} = 1,96$$

Reemplazando:

$$18 - 1,96 \frac{1,4}{\sqrt{300}} \leq \mu \leq 18 + 1,96 \frac{1,4}{\sqrt{300}}$$

$$17,84 \leq \mu \leq 18,16$$

Si el muestreo fuera sin reposición se plantea:

$$\theta = \mu$$

$$\hat{\theta} = \bar{x}$$



$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}} \sim N(0,1)$$

Y el intervalo:

$$1-\alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$$

### 5.5. Determinación Del Tamaño De La Muestra En La Estimación De $\mu$

En base a:

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

Y considerando que:  $\bar{x} - \mu = e$ , obtenemos:  $z = \frac{e}{\frac{\sigma}{\sqrt{n}}}$

Luego de esta fórmula despejamos n:

$$z = \frac{e \sqrt{n}}{\sigma} \rightarrow z\sigma = e \sqrt{n} \rightarrow \frac{z\sigma}{e} = \sqrt{n}$$

$$n \geq \frac{z^2 \sigma^2}{e^2}$$

Por lo tanto, para determinar el tamaño de la muestra, se deben conocer tres factores:

- 1) El nivel de confianza deseado, a partir del cual se determina z.
- 2) El error muestral permitido,  $e$ .
- 3) La desviación estándar.

En primer lugar, el valor de z se puede determinar una vez que se conoce el nivel de confianza deseado.

En segundo lugar, el error muestral  $e$ , es la cantidad de error que se está dispuesto a aceptar en el uso del estadístico muestral para estimar el parámetro.

En tercer lugar, se debe tener disponible una estimación de la desviación estándar a fin de determinar el tamaño requerido de la muestra. En algunos casos, se conoce la desviación estándar de la variable. En otros casos, pueden estar disponibles datos pasados (históricos) que se pueden extrapolar para determinar la desviación estándar actual. Si la desviación estándar no se puede determinar con los datos pasados, se puede efectuar un



estudio piloto empleando los resultados, para obtener una estimación de la desviación estándar.

### Ejemplo

La división de investigación de la Dirección Nacional de Vialidad, desea conocer el promedio de Kilómetros, recorridos durante una semana, por sus camiones. El jefe de la división indicó que:

- El máximo error muestral no deberá ser mayor a 15 Km por arriba o por debajo de la media verdadera.
- El nivel de confianza será 95,45%.
- La desviación estándar de la población, según estudios previos es de 120 Km.

¿Cuál es el tamaño de muestra adecuado a estos requerimientos?

### Resolución

Datos:

$$\begin{aligned} |\bar{x} - \mu| &= |e| \leq 15 \\ 1 - \alpha &= 0,9545 \quad \alpha = 0,0455 \quad \alpha/2 = 0,0227 \\ \sigma &= 120 \end{aligned}$$

Luego:

$$n \geq \frac{z^2 \sigma^2}{e^2}$$

Siendo:

$$1 - \alpha = 0,9545 \rightarrow z_{0,9772} = 2,00$$

Reemplazando:

$$n \geq \frac{(2)^2 (120)^2}{(15)^2} = \frac{57.600}{225} = 256$$

$$n \geq 256$$

### 5.6. Intervalo De Confianza Para Estimar P. Uso De La Distribución Normal.

El parámetro a estimar es  $P$ , proporción poblacional y el estimador es  $\hat{P} = \frac{x}{n}$ , proporción muestral. El estimador tiene distribución binomial.

En este caso no se puede determinar un estadístico  $K(\hat{P}, P)$  que se distribuya independientemente del parámetro  $P$ , pero si consideramos muestras grandes, por aplicación del Teorema Central del Límite,  $\hat{P}$  tendrá Distribución Normal.



Entonces,  $\hat{P}$  es un estimador sin sesgo de la proporción poblacional  $P$ , y puesto que la distribución por muestreo de  $\hat{P}$  es aproximadamente normal para muestras grandes, el intervalo de confianza para  $P$  será de la misma forma que el intervalo de confianza para la media  $\mu$ , dado anteriormente.

El estadístico:

$$K(\hat{P}; P) = \frac{\hat{P} - P}{\sigma_{\hat{P}}} = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}} \sim N(0,1)$$

Obsérvese, que  $\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}}$ , depende del parámetro  $P$ , desconocido; entonces debemos utilizar su estimación:

$$\hat{\sigma}_{\hat{P}} = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Luego:

$$1 - \alpha = \Pr \left\{ z_1 \leq \frac{\hat{P} - P}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \leq z_2 \right\}$$

El intervalo de confianza es:

$$\boxed{\Pr \left\{ \hat{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq P \leq \hat{P} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right\} = 1 - \alpha}$$

Los límites son:

$$L_1 = \hat{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad L_2 = \hat{P} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

También se pueden extender estas conclusiones para estimar la media  $\mu = nP$ , en una población binomial, multiplicando los límites señalados anteriormente, por  $n$ , en tal caso:

$$\boxed{\Pr \left\{ x - z_{1-\frac{\alpha}{2}} \sqrt{n\hat{P}(1-\hat{P})} \leq \mu \leq x + z_{1-\frac{\alpha}{2}} \sqrt{n\hat{P}(1-\hat{P})} \right\} = 1 - \alpha}$$



Donde:

$$L_1 = x - z_{\frac{1-\alpha}{2}} \sqrt{n\hat{P}(1-\hat{P})}$$

$$L_2 = x + z_{\frac{1-\alpha}{2}} \sqrt{n\hat{P}(1-\hat{P})}$$

En resumen:

$$\theta = P$$

$$\theta = nP = \mu$$

$$\hat{\theta} = \hat{P}$$

$$\hat{\theta} = n\hat{P} = x$$

$$K(\hat{P}; P) = \frac{\hat{P} - P}{\sigma_{\hat{P}}} = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}} \sim N(0,1)$$

$$K(x, \mu) = \frac{n\hat{P} - nP}{\sigma_x} = \frac{x - \mu}{\sigma_x} \sim N(0,1)$$

Donde:

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}}$$

$$\sigma_x = \sqrt{nP(1-P)}$$

Al desconocer P, utilizamos su estimador  $\hat{P}$ , luego:

$$\hat{\sigma}_{\hat{P}} = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Desviación  
Estándar estimada  
de la Proporción  
Muestral

$$\hat{\sigma}_x = \sqrt{n\hat{P}(1-\hat{P})}$$

Desviación  
Estándar estimada  
del número de

Por lo último, los límites de confianza definidos requieren un gran tamaño de muestra. Cuando más se desvía la proporción de 0.5, tanto mayor debe ser n.

*W. G. Cochran, da reglas prácticas sobre este punto:*

Si $\hat{P}$ es igual a:	Use aproximación normal cuando construya un intervalo de confianza del 95%, solo si n es por lo menos = a:
0.5	30
0.4 ó 0.6	50
0.3 ó 0.7	80
0.2 ó 0.8	200
0.1 ó 0.9	600
0.5 ó 0.95	1400

Fuente: Análisis Estadístico - Ya - Lun Chou - Segunda Edición. Interamericana.

Según este cuadro, por ejemplo, si  $\hat{P} = 0.1$ , el tamaño de la muestra debe ser por lo menos 600 para usar la aproximación normal y construir un intervalo de confianza del 95%. Si se desea un grado mayor de confianza de 95% debe incrementarse n por encima de los valores mostrados en el cuadro.

Todos los casos planteados han supuesto muestreo con reposición, si el muestreo fuera sin reposición, tendríamos:



$$K(\hat{P}, P) = \frac{\hat{P} - P}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{\frac{N-n}{N-1}}} \quad \text{Pues} \quad \hat{\sigma}_{\hat{P}} = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\Pr \left\{ \hat{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{\frac{N-n}{N-1}} < P < \hat{P} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{\frac{N-n}{N-1}} \right\} = 1 - \alpha$$

### Ejemplo

Tomada una muestra al azar de 500 directores de empresas, se encuentra que 100 de ellos habían pasado sus vacaciones en el exterior. Estimar: mediante intervalos:

- La proporción poblacional de directores de empresa que vacacionan en el extranjero mediante un intervalo de confianza del 99,73%.
- El número de directores que en la población vacacionan en el exterior.

### Resolución

$$n = 500$$

$$x = 100$$

$$1 - \alpha = 0,9973 \rightarrow z_{0,9987} = 3 \quad \alpha = 0,0027 \quad \frac{\alpha}{2} = 0,0014$$

Luego:

a-

$$\theta = P$$

$$\hat{\theta} = \hat{P}$$

$$K(\hat{P}, P) = \frac{\hat{P} - P}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \sim N(0,1)$$

$$\Pr \left\{ \hat{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq P \leq \hat{P} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right\} = 1 - \alpha = 0,9973$$

$$\hat{P} = \frac{x}{n} = \frac{100}{500} = 0,20$$

Reemplazando:



$$0,20 - 3 \sqrt{\frac{0,20 \times 0,80}{500}} < P < 0,20 + 3 \sqrt{\frac{0,20 \times 0,80}{500}}$$

$$\underline{0,146 < P < 0,254}$$

b- El intervalo para el número de directores que vacacionan en el exterior, se obtiene:

1) Teniendo el intervalo para la proporción, se multiplica por el tamaño de la muestra, así:

$$(0,146)(500) < nP < (0,254)(500)$$

$$\underline{73 < nP < 127}$$

2) Se encuentra el intervalo haciendo:

$$\theta = nP$$

$$\hat{\theta} = n\hat{P} = x$$

$$K(x, nP) = \frac{n\hat{P} - nP}{\sigma_{n\hat{P}}} = \frac{x - \mu}{\sigma_x} \sim N(0, 1)$$

$$\Pr\left\{x - z_{\frac{1-\alpha}{2}} \sqrt{n\hat{P}(1-\hat{P})} \leq nP \leq x + z_{\frac{1-\alpha}{2}} \sqrt{n\hat{P}(1-\hat{P})}\right\} = 1 - \alpha = 0,9973$$

Reemplazando:

$$100 - 3 \sqrt{500(0,2)(0,8)} < nP < 100 + 3 \sqrt{500(0,2)(0,8)}$$

$$\underline{73 < nP < 127}$$

Considerando:  $1 - \alpha = 0,9973 \rightarrow z_{0,9987} = 3$  y  $\hat{P} = \frac{x}{n} \rightarrow n\hat{P} = x$

### 5.7. Determinación Del Tamaño De La Muestra En La Estimación De P

Partiendo de: 
$$z = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

Despejamos n:



$$\hat{P} - P = z \sqrt{\frac{P(1-P)}{n}} \rightarrow (\hat{P} - P)^2 = z^2 \frac{P(1-P)}{n} \rightarrow \frac{(\hat{P} - P)^2}{z^2 P(1-P)} = \frac{1}{n}$$

$$n \geq \frac{z^2 P(1-P)}{(\hat{P} - P)^2}$$

De esta manera, se determina un tamaño mínimo de muestra para satisfacer las condiciones impuestas.

En la determinación del tamaño de la muestra para estimar la proporción  $P$ , se necesitan tres factores:

- 1) El nivel de confianza deseado, a partir del cual se determina  $z$ .
- 2) El error muestral permitido  $e$ .
- 3) La proporción real estimada.

El nivel de confianza deseado en la estimación del valor real de la proporción, permitirá obtener el valor de  $z$  apropiado en la distribución normal.

El error muestral es la cantidad de error que se está dispuesto a aceptar al estimar la proporción real.

La proporción real (verdadera) de éxitos en la población,  $P$ , es la cantidad que se querría estimar al tomar la muestra. En este caso hay dos caminos alternos posibles.

Si la proporción real de éxitos se puede estimar con base en los datos o experiencia pasados, esta estimación se puede utilizar para  $P$ . Pero, ¿y si no hay información disponible? ¿Qué se puede utilizar para estimar  $P$ ?

En este caso, se trataría de ser lo más conservador posible al estimar  $P$ .

En la fórmula de  $n$  se desearía usar el valor de  $P$  que hace a la cantidad  $P(1-P)$  lo más grande posible. Se puede demostrar empíricamente que cuando  $P=0,50$ , entonces  $P(1-P)$  está a su valor máximo.

Si:

$P=0,5$	$P(1-P)=0,5(0,5)=0,25$
$P=0,4$	$P(1-P)=0,4(0,6)=0,24$
$P=0,7$	$P(1-P)=0,7(0,3)=0,21$

Por lo tanto, cuando no se tiene conocimiento o estimación previos de la proporción  $P$  verdadera, se debería usar  $P=0,50$ , como el medio más conservador, para determinar el tamaño de la muestra.

Al no conocer  $P$ , asignándole el valor 0,50 aseguro el tamaño de muestra más grande, cubriendo de este modo, todas las posibilidades, pues  $P(1-P)$  es máximo, y  $z$  y  $e$ , están prefijados.



Pero, por otro lado, el uso de  $P = 0,50$  puede dar por resultado una sobreestimación del tamaño de la muestra. Como la proporción de la muestra se utiliza en el intervalo de confianza, si difiere mucho de 0,50, entonces, el ancho del intervalo de confianza puede ser bastante menor que el pretendido originalmente.

### Ejemplo

El dueño de una radio quiere conocer la proporción de gente que gusta de los programás deportivos. A cuántas personas se deberá encuestar si:

- a) - El error muestral no debe ser mayor al 2%.
  - El nivel de confianza es de 95%.
  - La proporción de gente que gusta de estos programás es aproximadamente de 0,60.
- b) No se conoce nada a cerca de la proporción de gente que gusta de este tipo de programás.

### Resolución

- a) Datos:

$$|\hat{P} - P| = |e| \leq 0,02$$

$$1 - \alpha = 0,95$$

$$P = 0,60$$

Luego: 
$$n \geq \frac{z^2 P(1-P)}{(\hat{P} - P)^2}$$

Siendo:

$$1 - \alpha = 0,95 \rightarrow z_{0,975} = 1,96$$

Reemplazamos:

$$n \geq \frac{(1,96)^2 0,60(0,40)}{(0,02)^2} = \frac{0,9221984}{0,0004} = 2.305$$

$$n \geq 2.305$$

b)

Datos

$$|e| \leq 0,02$$

$$1 - \alpha = 0,95 \rightarrow z = 1,96$$

$$n \geq \frac{z^2 P(1-P)}{e^2}$$

Al no conocer nada sobre la verdadera proporción consideramos a  $P = 1 - P = 0,50$

Reemplazando:



$$n \geq \frac{(1,96)^2 (0,5)(0,5)}{(0,02)^2} = \frac{0,9604}{0,0004} = 2.401$$

$$n \geq 2.401$$

**5.8. Intervalo De Confianza Para Estimar La Varianza De Una Población Normal. Uso de la Distribución  $\chi^2$  (Chi Cuadrado).**

$$\theta = \sigma^2$$

$$\hat{\theta} = \hat{\sigma}^2 = \hat{s}^2$$

$$K(\hat{\theta}; \theta) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} = \frac{\hat{\sigma}^2(n-1)}{\sigma^2} \approx \chi^2_{(n-1)}$$

Luego, planteamos que el nivel de confianza  $1-\alpha$ , es:

$$1 - \alpha = \Pr \left\{ k_1 \leq \frac{\hat{\sigma}^2(n-1)}{\sigma^2} \leq k_2 \right\}$$

Ahora, procederemos a despejar al parámetro  $\sigma^2$ :

$$1 - \alpha = \Pr \left\{ \frac{k_1}{\hat{\sigma}^2(n-1)} \leq \frac{1}{\sigma^2} \leq \frac{k_2}{\hat{\sigma}^2(n-1)} \right\}$$

Tomando recíproca, se invierte el sentido de la desigualdad:

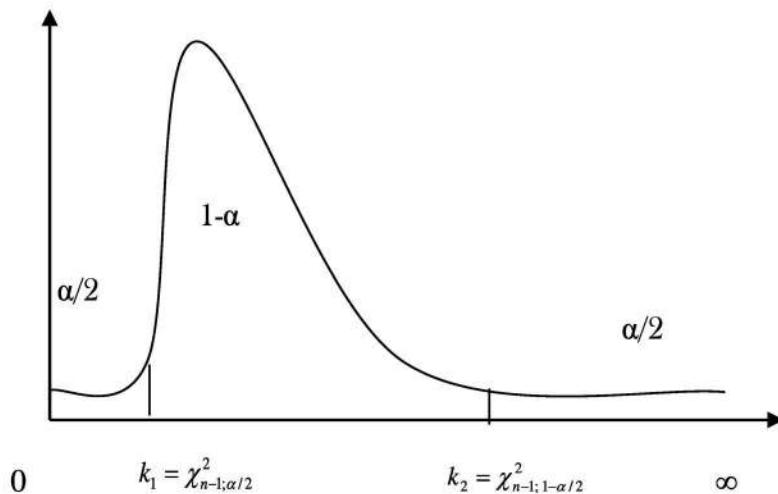
$$1 - \alpha = \Pr \left\{ \frac{\hat{\sigma}^2(n-1)}{k_1} \geq \sigma^2 \geq \frac{\hat{\sigma}^2(n-1)}{k_2} \right\}$$

Luego:

$$1 - \alpha = \Pr \left\{ \frac{\hat{\sigma}^2(n-1)}{k_2} \leq \sigma^2 \leq \frac{\hat{\sigma}^2(n-1)}{k_1} \right\}$$

$k_2$  y  $k_1$  son valores particulares de  $\chi^2$  que se encuentran en la respectiva tabla, así:

$$k_1 = \chi^2_{n-1; \alpha/2} \quad k_2 = \chi^2_{n-1; 1-\alpha/2}$$



Reemplazando, llegamos al intervalo buscado:

$$1 - \alpha = \Pr \left\{ \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; 1-\alpha/2}} \leq \sigma^2 \leq \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; \alpha/2}} \right\}$$

Este enunciado de probabilidad declara que si se toman de la población muchas muestras aleatorias de tamaño  $n$ , y para cada una se hace la aseveración:

$$\frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; 1-\alpha/2}} \leq \sigma^2 \leq \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; \alpha/2}}$$

(1- $\alpha$ ) % de estas aseveraciones serán correctas.

Luego, los límites de la confianza obtenidos son:

$$L_i = \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; 1-\alpha/2}} \quad L_s = \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; \alpha/2}}$$

#### Ejemplo:

Para un determinado electrodoméstico el promedio de ventas por comercio durante el año pasado, y de acuerdo a una muestra aleatoria de  $n = 10$  negocios, fue de  $\bar{x} = \$ 3.425$  con una  $\hat{\sigma} = \$ 200$ . Se supone que las ventas por comercio tienen una distribución aproximadamente normal.

Estimar la varianza de las ventas de ese electrodoméstico en todos los comercios, para el año anterior, utilizando un intervalo de confianza del 90%.

Resolución:



Datos:

$$\begin{aligned} n &= 10 \\ \bar{x} &= \$ 3.425 \\ \hat{\sigma} &= \$ 200. \\ 1 - \alpha &= 0,90 \end{aligned}$$

Luego:

$$\theta = \sigma^2$$

$$\hat{\theta} = \hat{\sigma}^2 = \hat{s}^2$$

$$K(\hat{\theta}; \theta) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} = \frac{\hat{\sigma}^2(n-1)}{\sigma^2} \approx \chi^2_{(n-1)}$$

$$\Pr \left\{ \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; 1-\alpha/2}} \leq \sigma^2 \leq \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; \alpha/2}} \right\} = 0,90$$

Siendo  $1 - \alpha = 0,90$ 

$$k_1 = \chi^2_{n-1; \alpha/2} = \chi^2_{9; 0,05} = 3,3 \quad k_2 = \chi^2_{n-1; 1-\alpha/2} = \chi^2_{9; 0,95} = 16,9$$

Reemplazando:

$$\frac{9(200)^2}{16,9} \leq \sigma^2 \leq \frac{9(200)^2}{3,3}$$

$$\underline{21,302 \leq \sigma^2 \leq 109,09}$$



Síntesis de Estimación por Intervalos - Muestreo Con Reposición					
Parámetro	Estimador	Estadístico	Distribución Del Estadístico	Intervalo	Supuesto
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$N(0,1)$	$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\}$	Población Normal $\sigma^2$ Conocida n Cualquiera
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$	$t_{n-1}$	$1 - \alpha = \Pr \left\{ \bar{x} - t_{(n-1), 1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(n-1), 1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\}$	Población Normal $\sigma^2$ Desconocida n < 30
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$	$N(0,1)$	$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\}$	Población Normal $\sigma^2$ Desconocida n ≥ 30
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$N(0,1)$	$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\}$	Población No Normal $\sigma^2$ Conocida n ≥ 30
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$	$N(0,1)$	$1 - \alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right\}$	Población No Normal $\sigma^2$ Desconocida n ≥ 30
$P$	$\hat{P}$	$\frac{\hat{P} - P}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}}$	$N(0,1)$	$1 - \alpha = \Pr \left\{ \hat{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq P \leq \hat{P} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right\}$	Según Regla De Cochran
$\sigma^2$	$\hat{\sigma}^2 = \hat{s}^2$	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} = \frac{\hat{\sigma}^2(n-1)}{\sigma^2}$	$\chi^2_{(n-1)}$	$1 - \alpha = \Pr \left\{ \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; 1-\alpha/2}} \leq \sigma^2 \leq \frac{\hat{\sigma}^2(n-1)}{\chi^2_{n-1; \alpha/2}} \right\}$	Población Normal



Síntesis de Estimación por Intervalos – Muestreo Sin Reposición					
Parámetro	Estimador	Estadístico	Distribución Del Estadístico	Intervalo	Supuesto
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$	$N(0,1)$	$1-\alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$	Población Normal $\sigma^2$ Conocida n Cualquiera
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$	$t_{n-1}$	$1-\alpha = \Pr \left\{ \bar{x} - t_{(n-1), 1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + t_{(n-1), 1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$	Población Normal $\sigma^2$ Desconocida n < 30
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$	$N(0,1)$	$1-\alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$	Población Normal $\sigma^2$ Desconocida n ≥ 30
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$	$N(0,1)$	$1-\alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$	Población No Normal $\sigma^2$ Conocida n ≥ 30
$\mu$	$\bar{x}$	$\frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$	$N(0,1)$	$1-\alpha = \Pr \left\{ \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right\}$	Población No Normal $\sigma^2$ Desconocida n ≥ 30
$P$	$\hat{P}$	$\frac{\hat{P} - P}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{\frac{N-n}{N-1}}}$	$N(0,1)$	$1-\alpha = \Pr \left\{ \hat{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{\frac{N-n}{N-1}} \leq P \leq \hat{P} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \sqrt{\frac{N-n}{N-1}} \right\}$	Según Regla De Cochran



## Unidad N° 10: Contraste, Docima ó Verificación De Hipótesis

### **Objetivos Específicos**

*Que el estudiante:*

**Comprenda** los fundamentos teóricos y la lógica subyacente de la Inferencia Estadística en una de sus dos grandes ramas: *Docimásia de Hipótesis*.

**Analice** el proceso de prueba de hipótesis estadística para diferentes casos, teniendo en cuenta reglas de decisión adecuadas, errores que se pueden cometer en dicho proceso y el cálculo de la función de potencia del test elegido y su correspondiente curva OC.

### **Contenidos**

1. Decisión estadística.
2. Hipótesis estadísticas.
3. Concepto de docima.
4. Errores y sus probabilidades.
5. Distintos tipos de docimás.
6. Docima para  $\mu$ . Uso de la distribución Normal y "t" de Student.
7. Docima para  $p$ . Uso de la distribución Normal.
8. Curva operatoria característica (OC) y curva de potencia.
9. Docima para la varianza. Uso de la  $\chi^2$  (Chi cuadrado).
10. Docima e intervalos de confianza.



## CONTRASTE, DOCIMA Ó VERIFICACIÓN DE HIPÓTESIS

También llamada Prueba ó Test de Hipótesis. Es un procedimiento de la Inferencia Estadística para la toma de decisiones.

### 1. DECISIÓN ESTADÍSTICA

Se llama decisión estadística, a una decisión que se toma con respecto a algún aspecto de la población, en base a evidencias proporcionadas por las muestras.

Para tomar una decisión estadística se formulan conjeturas sobre las características de una distribución poblacional. Estas conjeturas se llaman *Hipótesis Estadísticas*.

### 2. HIPÓTESIS ESTADÍSTICAS

Consisten en suposiciones ó conjeturas que se establecen acerca del valor de un parámetro (característica de una distribución poblacional).

Las hipótesis pueden ser:

#### ***Hipótesis nula o Hipótesis de Nulidad***

Es un supuesto acerca de uno o más parámetros de la población que debe ser rechazado o no en base a la evidencia muestral.

Indica que no se han producido en la población, efectos o cambios.

Se denomina nula, en el sentido de que no existe diferencia real entre el verdadero valor del parámetro de la población de la que hemos obtenido la muestra y el valor hipotetizado. Se simboliza por  $H_0$

#### ***Hipótesis Alternativa o Hipótesis Alterna***

Si la hipótesis nula es falsa, deberá existir otra hipótesis que sea verdadera. Esta hipótesis recibe el nombre de hipótesis alternativa. Indica la presencia de cambios ó efectos en la población. Se simboliza por  $H_a$ .

### 3. CONCEPTO DE DOCIMA

A partir de los conceptos dados, definiremos a la *Docimásia de Hipótesis* como un *Experimento Aleatorio* que se realiza para decidir sobre la veracidad o falsedad de una hipótesis, llamada hipótesis nula, la que, por lo general, indica la ausencia de cambios, o efectos en la población, en base a evidencia proporcionada por la muestra, midiendo los riesgos de cometer un error, aceptando (no rechazando) la hipótesis nula siendo falsa, o rechazándola siendo cierta.

Antes de ver cómo opera en términos estadísticos, analicemos el siguiente ejemplo, que nos permitirá mostrar el razonamiento para efectuar una prueba de hipótesis.

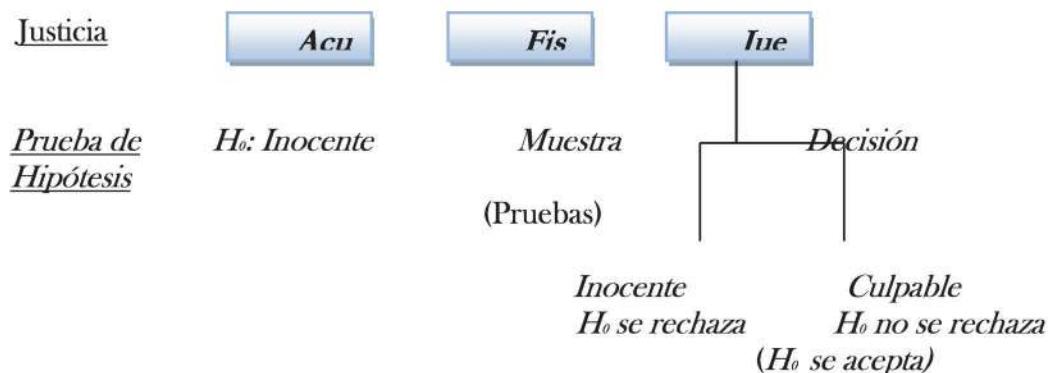


Así, el razonamiento necesario, es muy similar al que se utiliza en una Corte de Justicia cuando se debe tomar la decisión de declarar o no culpable a una persona acusada de cometer un delito.

Los actores de un juicio pueden resumirse en un acusado, un fiscal y un juez. Es, éste último, quien debe tomar la decisión de declarar o no culpable al acusado.

El acusado será considerado inocente hasta tanto las pruebas presentadas por el fiscal demuestren lo contrario.

Así:



Si la evidencia presentada al juez no es contundente, éste decidirá por la inocencia del acusado (no rechazo de  $H_0$ ). En caso contrario, cuando la evidencia condene al acusado, el juez tiene a mano una alternativa, la acusación de culpabilidad (lo que equivale al rechazo de la hipótesis nula planteada).

En cualquiera de los dos casos, el juez puede cometer un error:

- 1- Decidir por la inocencia del acusado (No rechazo o aceptación de  $H_0$ ), y en realidad, éste sea culpable. (No rechazar o aceptar una hipótesis falsa).
- 2- Decidir por la culpabilidad del acusado (Rechazo de  $H_0$ ) y en realidad, éste sea inocente. (Rechazar una hipótesis cierta).

Plantearemos ahora, a través de un ejemplo, cómo opera la docimásia de hipótesis:

Dadas:

$$H_0 : \theta = \theta_0 \quad (\text{El parámetro asume un valor particular } \theta_0)$$

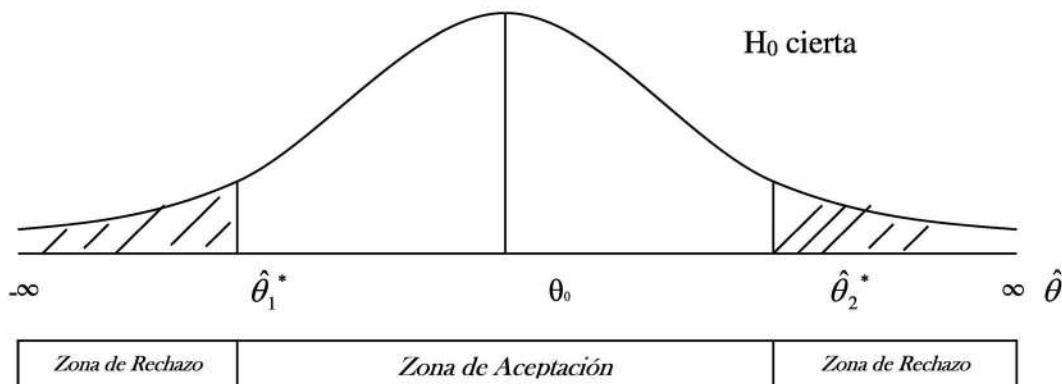
$$H_1 : \theta \neq \theta_0 \quad (\text{El parámetro asume un valor distinto de } \theta_0, \text{ que puede ser mayor o menor})$$

Recordemos que los estimadores muestrales ( $\hat{\theta}$ ) son variables aleatorias, y además, entre otras propiedades, deben ser insesgados, lo que implica que  $E(\hat{\theta}) = \theta$ .



Luego, si  $H_0$  es cierta, indica que  $\theta = \theta_0$ , es decir que la muestra seleccionada proviene de una población centrada en  $\theta_0$ , entonces el estimador  $\hat{\theta}$  se distribuirá con parámetro  $\theta_0$  [ $E(\hat{\theta}) = \theta_0$  ].

Considerando que  $\hat{\theta} \sim N(\theta_0, \sigma_{\hat{\theta}})$ , podemos graficar:



Luego, se puede determinar la probabilidad de que una estimación puntual (valor particular para  $\hat{\theta}$  en la muestra seleccionada) se aleje del parámetro  $\theta_0$  más allá de ciertos límites, llamados puntos críticos ( $\hat{\theta}_1^*$  Y  $\hat{\theta}_2^*$ ), y asuma valores en una zona llamada Región Crítica o Zona de Rechazo. (Valores de la variable entre  $-\infty$  y  $\hat{\theta}_1^*$  ó entre  $\hat{\theta}_2^*$  y  $\infty$ ). Si esta posibilidad se fija a niveles muy bajos (0,05; 0,01 ó menores), serán muy pocas las veces que la estimación puntual caiga en la región crítica cuando la hipótesis nula sea verdadera.

Cuando el punto estimado caiga en región crítica puede ser que la hipótesis nula sea cierta y se halla presentado un evento de baja probabilidad de ocurrencia; pero es más lógico pensar que este resultado significativo se debe a que la hipótesis nula es falsa (es decir  $\theta \neq \theta_0$ ) y se ha producido, en la población, el efecto o cambio que señala la hipótesis alternativa, por lo tanto se rechaza la hipótesis nula como falsa.

Es por ello, que la región crítica es, en realidad una *Zona De Rechazo* de la hipótesis nula.

Reiteramos que cuando  $H_0$  es cierta, el valor del parámetro es el indicado en ella. Cuando  $H_0$  es falsa, el valor del parámetro no es el indicado en ella, sino, él o uno de los indicados en  $H_1$ , de acuerdo al tipo de docima, según veremos más adelante.

Existe, en consecuencia, una probabilidad  $\alpha$ , de que la hipótesis nula sea cierta y se la rechace como falsa (área sombreada). A esta probabilidad la fija el investigador a niveles muy bajos, y se la llama *Nivel De Significación*.

$$\text{Nivel de significación} = \alpha = \Pr\{\text{Rechazar } H_0 / H_0 \text{ Cierta}\}$$



#### 4. ERRORES Y SUS PROBABILIDADES

Cuando la estimación puntual cae en zona de rechazo, se decide que  $H_0$  es falsa y se la rechaza, pero se puede cometer un error, llamado ERROR TIPO I, de rechazarla siendo cierta.

Es decir que:

$$\text{Error Tipo I} = \text{Rechazar } H_0 / H_0 \text{ Cierta}$$

Y la probabilidad de cometerlo es igual a  $\alpha$ , luego:

$$\alpha = \text{Nivel de Significación} = \Pr\{\text{Error Tipo I}\} = \Pr\{\text{Rechazar } H_0 / H_0 \text{ Cierta}\}$$

Por otra parte, cuando la estimación puntual cae en zona de aceptación, nos lleva a aceptar  $H_0$  como verdadera, pudiendo cometer otro error, aceptar  $H_0$  como verdadera cuando en realidad es falsa. Éste es el ERROR TIPO II.

Es decir que:

$$\text{Error Tipo II} = \text{Aceptar } H_0 / H_0 \text{ falsa}$$

Y la probabilidad de cometerlo es igual a  $\beta$ , luego:

$$\beta = \Pr\{\text{Error tipo II}\} = \Pr\{\text{Aceptar } H_0 / H_0 \text{ Falsa}\}$$

Entonces, las decisiones y sus probabilidades, se resumen en la siguiente tabla:

Decisiones	$H_0$ Cierta $\theta = \theta_0$	$H_0$ Falsa $\theta > \theta_0; \theta < \theta_0; \theta \neq \theta_0$
Aceptar $H_0$ $\hat{\theta} \notin ZR$	Decisión Correcta Probabilidad: $1-\alpha$	Error Tipo II Probabilidad: $\beta$
Rechazar $H_0$ $\hat{\theta} \in ZR$	Error Tipo I Probabilidad: $\alpha$	Decisión Correcta Probabilidad: $1-\beta$

Luego:

$$1 - \alpha = \Pr\{\text{Aceptar } H_0 / H_0 \text{ cierta}\}$$

$$\alpha = \Pr\{\text{Rechazar } H_0 / H_0 \text{ cierta}\} = \Pr\{\text{error tipo I}\} = \text{Nivel de Significación}$$

$$\beta = \Pr\{\text{Aceptar } H_0 / H_0 \text{ falsa}\} = \Pr\{\text{error tipo II}\}$$

$$1 - \beta = \Pr\{\text{Rechazar } H_0 / H_0 \text{ falsa}\} = \text{Potencia de la docima}$$

Gráficamente: consideraremos la docima planteada al inicio de este tema, es decir:



$$H_0 : \theta = \theta_0$$

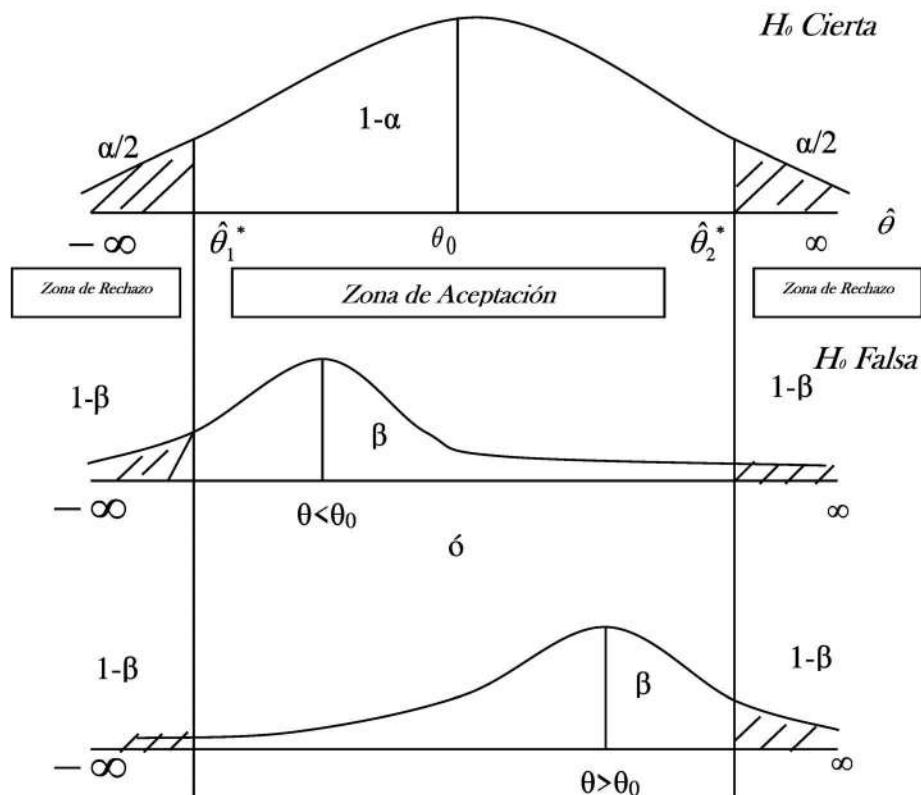
$$H_1 : \theta \neq \theta_0$$

Aclaremos que las conclusiones a las que arribaremos son válidas para todos los tipos de docima.

Si  $H_0$  es cierta, la distribución del estimador tendrá como parámetro a  $\theta_0$ .

Si  $H_0$  es falsa, habrá un conjunto de distribuciones para el estimador, cada una concentrada en un valor distinto, según lo supuesto en  $H_1$ .

Lógicamente, sólo será válida una de ellas, pues el parámetro es una constante para una población dada, pero al desconocerlo, podemos suponer valores que el mismo puede asumir. Entonces, según  $H_1 : \theta \neq \theta_0$ , supondremos, considerando a  $H_0$  falsa, valores menores y mayores a  $\theta_0$ , graficadas, para simplificar, en dos distribuciones, identificando las distintas probabilidades.



Obsérvese:

1- La ubicación de las probabilidades de ambos errores:  $\alpha$ , definida en la zona de rechazo en la distribución del estadístico bajo  $H_0$  cierta y  $\beta$ , en la zona de aceptación en la distribución del estadístico cuando  $H_0$  es falsa.



2- Las probabilidades  $\alpha$  y  $\beta$ , no son complementarias porque se refieren a distintas distribuciones. Cuando  $H_0$  es cierta ( $\theta = \theta_0$ ) no hay valores para  $\beta$ , ya que en este caso sólo hay dos eventos posibles: rechazar  $H_0$  cierta y aceptar  $H_0$  cierta, cuyas probabilidades son  $\alpha$  y  $1-\alpha$  respectivamente.

Cuando  $H_0$  es falsa los eventos son: aceptar  $H_0$  y rechazar  $H_0$  falsa, con probabilidades  $\beta$  y  $1-\beta$  respectivamente.

3.  $\alpha$  y  $1-\beta$  son probabilidades de rechazar  $H_0$  mientras que  $\beta$  y  $1-\alpha$ , son probabilidades de aceptar  $H_0$ , en distintas circunstancias.

4- Los valores de  $\beta$  corresponden a la zona de aceptación, la que está determinada por  $\alpha$ , que corresponde a la zona de rechazo. Mientras mayor sea  $\alpha$ , menor será  $\beta$ , y viceversa.

Es decir,  $\alpha$  y  $\beta$  varían en sentido inverso y no puede disminuirse una sin que aumente la otra. Por esta razón, solo se fija  $\alpha$ , la probabilidad  $\beta$  depende de ella.

La única manera de hacer disminuir las probabilidades de ambos tipos de errores simultáneamente, es aumentando el tamaño de la muestra.

Algunas veces, en ciertos problemas, se fijan  $\alpha$  y  $\beta$  y posteriormente se calcula el tamaño de la muestra para satisfacer las probabilidades prefijadas.

En particular, consideremos el siguiente ejemplo:

Decisiones	$H_0 : Llueve$	$H_0 : No Llueve$
Llevo Paraguas (Acepto $H_0$ )	<i>DECISIÓN CORRECTA</i> Aceptar $H_0 / H_0$ Cierta Probabilidad = $1-\alpha$	<i>ERROR TIPO II</i> Aceptar $H_0 / H_0$ Falsa Probabilidad = $\beta$
No Llevo Paraguas (Rechazo $H_0$ )	<i>ERROR TIPO I</i> Rechazar $H_0 / H_0$ cierta Probabilidad = $\alpha$	<i>DECISIÓN CORRECTA</i> Rechazar $H_0 / H_0$ falsa Probabilidad = $1-\beta$

## **5. DISTINTOS TIPOS DE DOCIMA**

Dado un parámetro  $\theta$ , objeto de la docima y un valor particular de este parámetro,  $\theta_0$ , para el cual se verifica la hipótesis nula, se pueden plantear:

### **A) Docimas de Hipótesis Compuestas ó Inexactas**

- 1- Docimas Bilaterales
- 2- Docimas Laterales
  - 2.1. Derechas
  - 2.2. Izquierdas

**B) Docimas de Hipótesis Simples ó Exactas**

- 1- Docimas Laterales Derechas
- 2- Docimas Laterales Izquierdas

**A) Docimas de Hipótesis Compuestas ó Inexactas**

Se contrasta un valor contra un conjunto de valores. ( $\neq ; > ; <$ )

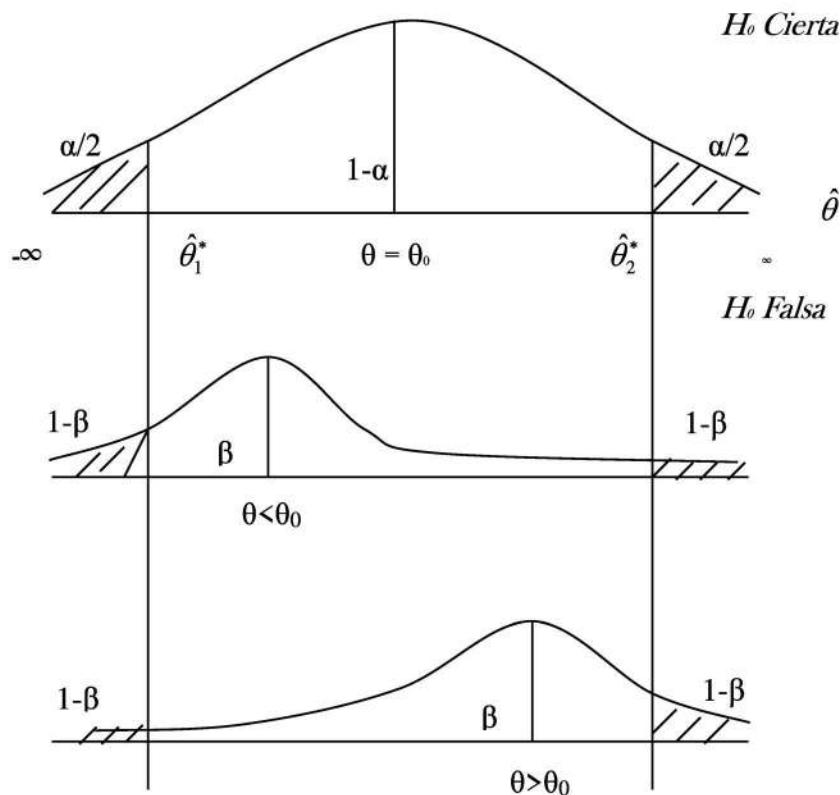
**1) Docimas Bilaterales**

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

La hipótesis nula dice que el parámetro es igual a  $\theta_0$ , y la hipótesis alternativa dice que es distinto.

Considerando un estimador insesgado con distribución normal:  $\hat{\theta} \sim N(\theta_0, \sigma_{\hat{\theta}})$ , y bajo el supuesto de la  $H_0$  cierta, la gráfica es:



La zona de rechazo de  $H_0$ , está ubicada en ambos extremos de la distribución, según  $H_1$ , bajo la hipótesis nula cierta.

Cuando la hipótesis nula es falsa, el estimador se distribuye alrededor de otro valor del parámetro distinto  $\theta_0$ .



Obsérvese, además, que los puntos críticos calculados en base a  $\alpha$ , en el supuesto de la  $H_0$  cierta, son quienes determinan las zonas de rechazo en los supuestos de  $H_0$  falsa, la cual no permite en estas distribuciones, hablar de simetría, por lo cual  $1 - \beta$ , es suma de las dos áreas sombreadas, pero no podemos considerar iguales a las mismas.

Luego:

$$1 - \alpha = \Pr\{\hat{\theta}_1^* \leq \hat{\theta} \leq \hat{\theta}_2^* / H_0 \text{ cierta}\}$$

$$\alpha = \Pr\{\hat{\theta} < \hat{\theta}_1^* / H_0 \text{ cierta}\} + \Pr\{\hat{\theta} > \hat{\theta}_2^* / H_0 \text{ cierta}\}$$

$$\beta = \Pr\{\hat{\theta}_1^* \leq \hat{\theta} \leq \hat{\theta}_2^* / H_0 \text{ falsa}\}$$

$$1 - \beta = \Pr\{\hat{\theta} < \hat{\theta}_1^* / H_0 \text{ falsa}\} + \Pr\{\hat{\theta} > \hat{\theta}_2^* / H_0 \text{ falsa}\}$$

## 2) Docimas Laterales

En ambas, la  $H_0$  es cierta para un conjunto de valores, con lo cual se cubre el espacio paramétrico, así:

Derecha	Izquierda
$H_0 : \theta \leq \theta_0$	$H_0 : \theta \geq \theta_0$
$H_1 : \theta > \theta_0$	$H_1 : \theta < \theta_0$

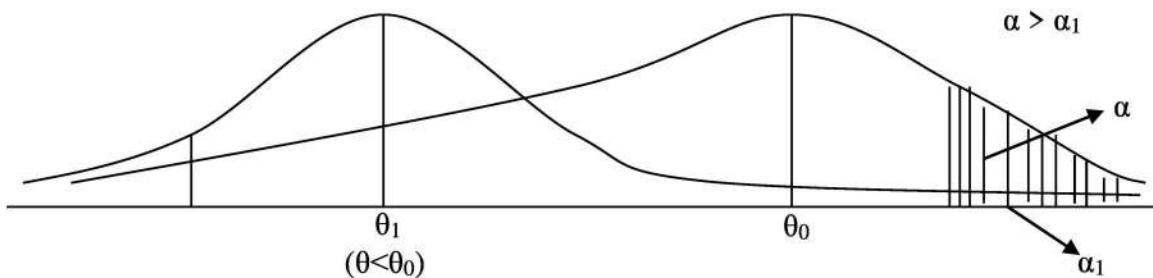
Pero, para determinar  $\alpha$  y la zona de rechazo sólo se tiene que encontrar el valor  $\theta = \theta_0$ , es decir, el MAYOR de los posibles valores del parámetro en la hipótesis nula de la docima lateral derecha y el MENOR en la misma hipótesis de la docima lateral izquierda.

Una vez establecida la zona de rechazo, hay para cada uno de los distintos valores posibles de  $\theta$ , una probabilidad diferente de rechazar  $H_0$  cierta, pero estas probabilidades son siempre menores a  $\alpha$ .

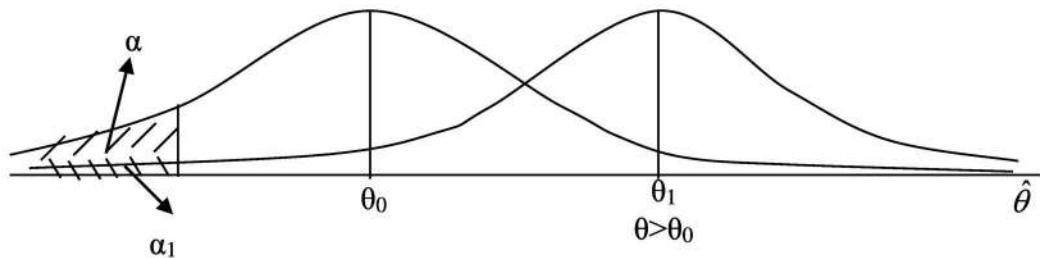
En efecto, en una docima lateral derecha ( $H_0 : \theta \leq \theta_0$ ) la hipótesis nula se verifica para todo valor del parámetro menor o igual a  $\theta_0$ , y dado que el estimador se distribuye alrededor de  $\theta_0$  que sirvió de base para calcular la zona de rechazo, calculamos el mayor de los riesgos que se corre en este supuesto.



Gráficamente:



Con un razonamiento análogo deducimos que en la docima lateral izquierda, la probabilidad  $\alpha_1$  de rechazar la  $H_0$  cierta, para  $\theta$  mayor a  $\theta_0$ , también es menor a  $\alpha$ .



Utilizamos a partir de esta explicación:

*Docima Lateral Derecha*

$$\begin{aligned} H_0 : \theta &= \theta_0 \\ H_1 : \theta &> \theta_0 \end{aligned}$$

*Docima Lateral Izquierda*

$$\begin{aligned} H_0 : \theta &= \theta_0 \\ H_1 : \theta &< \theta_0 \end{aligned}$$

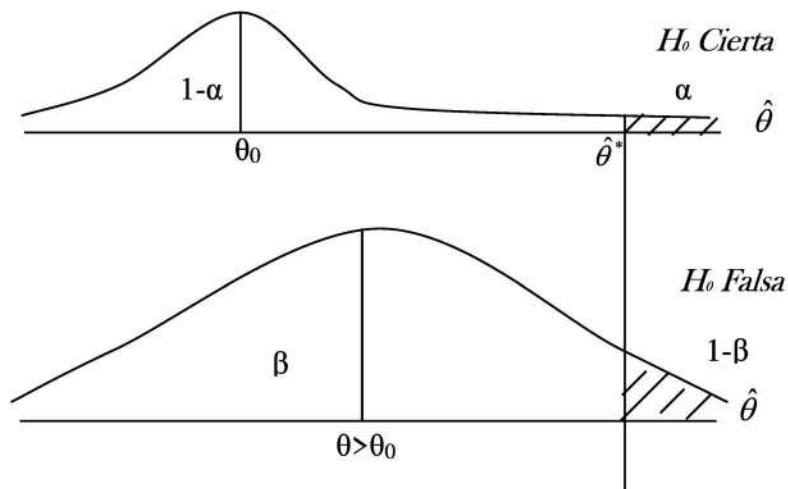
### 2.1. Docimas Laterales Derechas

$$\begin{aligned} H_0 : \theta &= \theta_0 \\ H_1 : \theta &> \theta_0 \end{aligned}$$

La hipótesis nula dice que el parámetro es igual  $\theta_0$ , y la hipótesis alternativa dice que es mayor.



Gráficamente:



La  $H_0$  se rechazará cuando la evidencia proporcionada por la muestra nos haga pensar que el parámetro es superior a  $\theta_0$ .

La zona de rechazo está ubicada en el extremo derecho de la distribución, según  $H_0$ , bajo la hipótesis nula cierta.

Luego:

$$1 - \alpha = \Pr\{\hat{\theta} \leq \hat{\theta}^* / H_0 \text{ cierta}\}$$

$$\alpha = \Pr\{\hat{\theta} > \hat{\theta}^* / H_0 \text{ cierta}\}$$

$$\beta = \Pr\{\hat{\theta} \leq \hat{\theta}^* / H_0 \text{ falsa}\}$$

$$1 - \beta = \Pr\{\hat{\theta} > \hat{\theta}^* / H_0 \text{ falsa}\}$$

## 2.2. Docimas Laterales Izquierdas

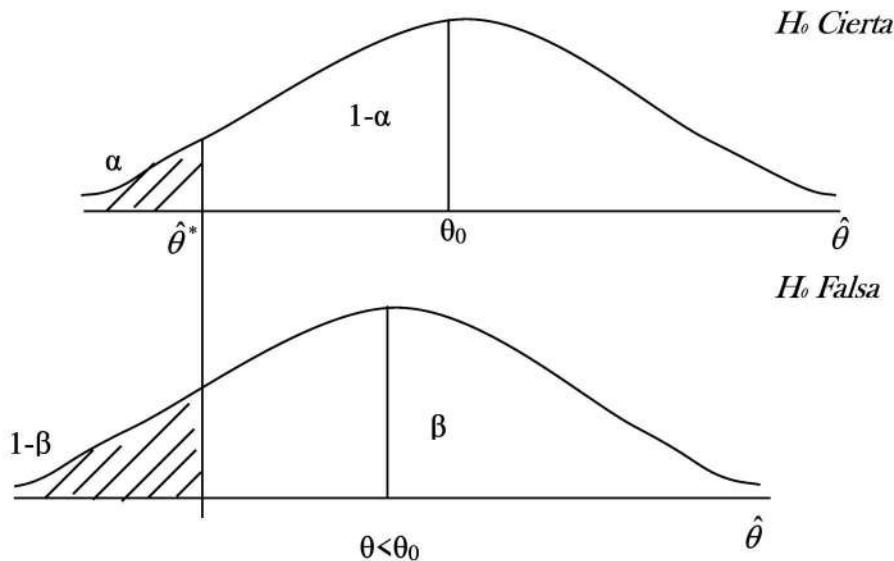
$$H_0 : \theta = \theta_0$$

$$H_1 : \theta < \theta_0$$

La hipótesis nula plantea que el parámetro es igual a  $\theta_0$  y la alternativa que es menor a  $\theta_0$ .



Gráficamente:



La  $H_0$ , se rechazará si la evidencia proporcionada por la muestra nos hace suponer que el parámetro es inferior a  $\theta_0$ .

La zona de rechazo está en el extremo izquierdo de la distribución del estimador, según  $H_0$ , supuesto que  $H_0$  es cierta.

Luego:

$$1 - \alpha = \Pr\{\hat{\theta} \geq \hat{\theta}^* / H_0 \text{ cierta}\}$$

$$\alpha = \Pr\{\hat{\theta} < \hat{\theta}^* / H_0 \text{ cierta}\}$$

$$\beta = \Pr\{\hat{\theta} \geq \hat{\theta}^* / H_0 \text{ falsa}\}$$

$$1 - \beta = \Pr\{\hat{\theta} < \hat{\theta}^* / H_0 \text{ falsa}\}$$

### Nota

Obsérvese que la zona de rechazo se determina con la distribución del estimador bajo la hipótesis nula cierta, pero la ubicación de esta zona está indicada por la hipótesis alternativa, señalando cuál es el tipo de docima.

Además, decir que:  $H_0$  cierta, implica  $\theta = \theta_0$

$H_0$  falsa, implica  $\theta \neq \theta_0$ ;  $\theta > \theta_0$ ;  $\theta < \theta_0$  es decir  $H_1$ , según la docima.

El cálculo de los puntos críticos lo realizaremos cuando tratemos la docima de los parámetros  $\mu$  y  $P$ .

**B) Docimas de Hipótesis Simples ó Exactas**

Cuando se contrasta un valor para  $H_0$  contra un valor para  $H_1$ .

Así:

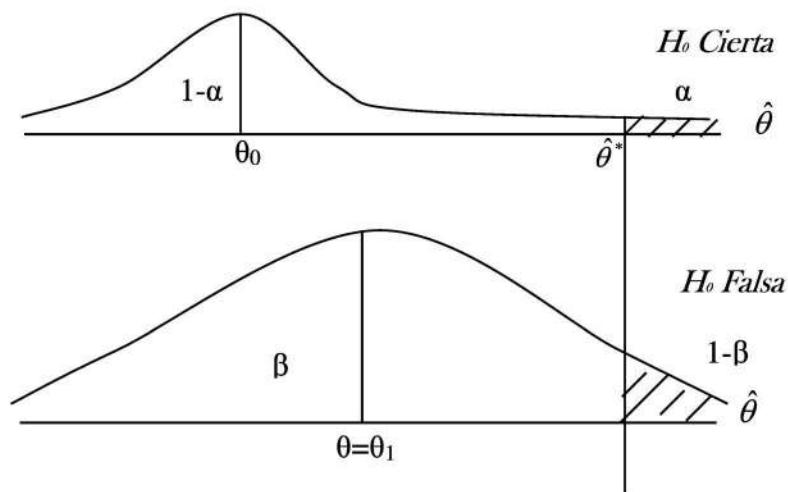
$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

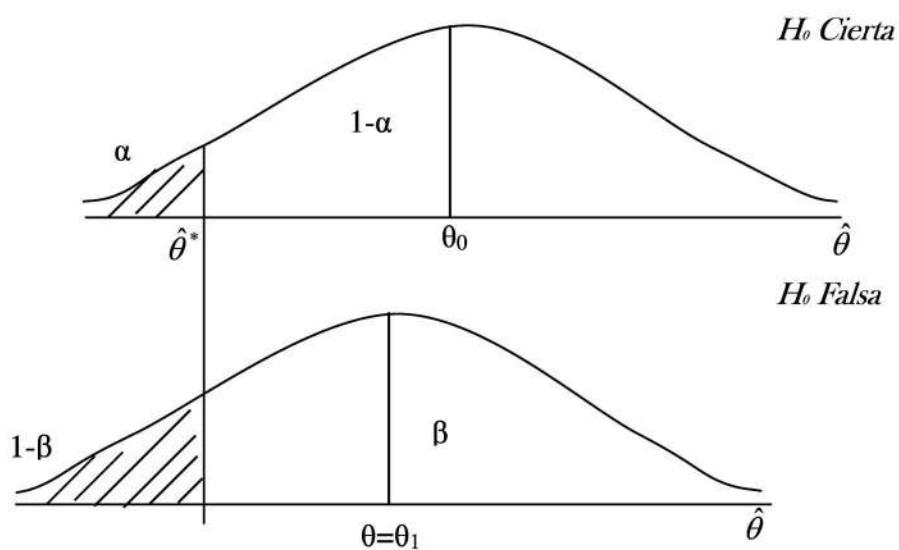
Las docimas exactas son laterales derechas o izquierda, según sea  $\theta_1$  mayor o menor a  $\theta_0$ .

Gráficamente:

*Docima Lateral Derecha*



*Docima Lateral Izquierda*





## 6. DOCIMA PARA $\mu$ . Uso de la Distribución Normal y "t" de Student.

Consideraremos para el desarrollo de este tema, los casos vistos en Estimación Estadística, a saber:

1) Poblaciones Normales		2) Poblaciones No Normales	
$\sigma^2$ Conocida $n$ Cualquiera	Distribución Normal	$\sigma^2$ Conocida $n \geq 30$	Por TCL Normal
$\sigma^2$ Desconocida		$\sigma^2$ Desconocida	
$n < 30$	"t" De Student	$n < 30$	—
$n \geq 30$	Por TCL Normal	$n \geq 30$	Por TCL Normal

Plantearemos para el caso de Poblaciones Normales,  $\sigma^2$  conocida,  $n$  cualquiera, el siguiente esquema de trabajo:

1. Identificación del Parámetro a docimar  $\theta$
2. Selección del Estimador  $\hat{\theta}$
3. Determinación del Estadístico  $K(\hat{\theta}, \theta)$
4. Cálculo de los Puntos críticos  $\hat{\theta}^*$
5. Regla de decisión.
6. Cálculo de probabilidades involucradas

El desarrollo se realizará teniendo en cuenta:

- 1- Docima de Hipótesis Simples: Lateral Derecha y Lateral Izquierda.
- 2- Docima de Hipótesis Compuestas: Lateral derecha, Lateral Izquierda y Bilateral.

En el resto de los casos, se indicarán las modificaciones a realizar, en función de la distribución que corresponda, pues los procedimientos y razonamientos son idénticos.

### Caso 1: Poblaciones Normales.

#### a) Varianza Poblacional Conocida. Muestras de cualquier tamaño.

$$\theta = \mu$$

$$\hat{\theta} = \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$K(\hat{\theta}, \theta) = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$



## Docima de Hipótesis Simples

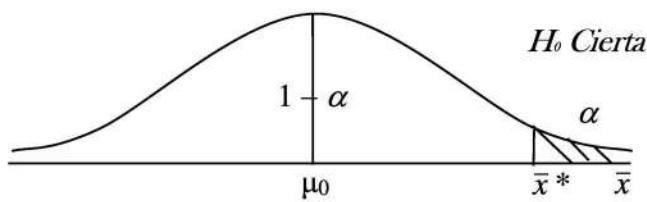
### Lateral Derecha

#### Hipótesis

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0 \quad \text{donde } \mu_a > \mu_0$$

#### Gráfica



#### Punto Crítico

$$\bar{x}^* = \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{o} \quad z^* = z_{1-\alpha}$$

#### Regla de Decisión

$$\begin{array}{lll} \text{Si:} & \bar{x} \leq \bar{x}^* & \text{Aceptar } H_0 \text{ (No rechazar } H_0) \\ & \bar{x} > \bar{x}^* & \text{Rechazar } H_0 \end{array}$$

Si se ha determinado  $z$ :

$$\begin{array}{lll} \text{Si:} & z \leq z^* & \text{Aceptar } H_0 \text{ (No rechazar } H_0) \\ & z > z^* & \text{Rechazar } H_0 \end{array}$$

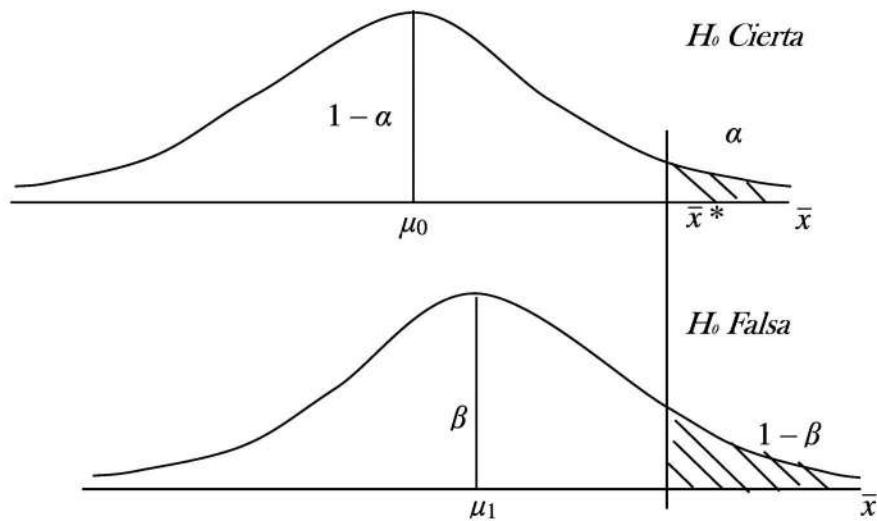
$$\text{Donde } z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

#### Probabilidades involucradas.

Si  $H_0$  no es cierta, es decir,  $\mu$  no es igual a  $\mu_0$ , entonces existe otra distribución, centrada en el valor mayor supuesto en la hipótesis alternativa, es decir  $\mu_a$ , generando las probabilidades  $\beta$  y  $1 - \beta$ .



Gráficamente:



Luego:

$$\alpha = \Pr\{\bar{x} > \bar{x}^* / H_0\} = \Pr\left\{z > \frac{\bar{x}^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$1 - \alpha = \Pr\{\bar{x} \leq \bar{x}^* / H_0\} = \Pr\left\{z \leq \frac{\bar{x}^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$\beta = \Pr\{\bar{x} \leq \bar{x}^* / H_1\} = \Pr\left\{z \leq \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

$$1 - \beta = \Pr\{\bar{x} > \bar{x}^* / H_1\} = \Pr\left\{z > \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

### ***Lateral Izquierda***

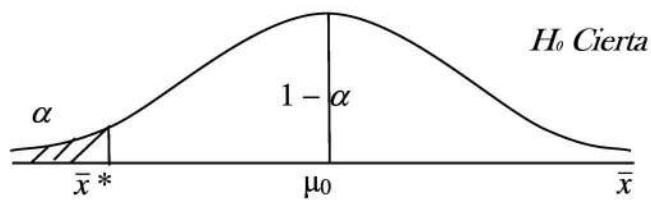
#### Hipótesis

$$H_0: \mu = \mu_0$$

$$H_1: \mu = \mu_1 \quad \text{donde } \mu_1 < \mu_0$$



## Gráfica



## Punto Crítico

$$\bar{x}^* = \mu_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{ó} \quad z^* = -z_{1-\alpha}$$

## Regla de Decisión

Si:  $\bar{x} \geq \bar{x}^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
 $\bar{x} < \bar{x}^*$       Rechazar  $H_0$

Si se ha determinado  $z^*$ :

Si:  $z \geq z^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
 $z < z^*$       Rechazar  $H_0$

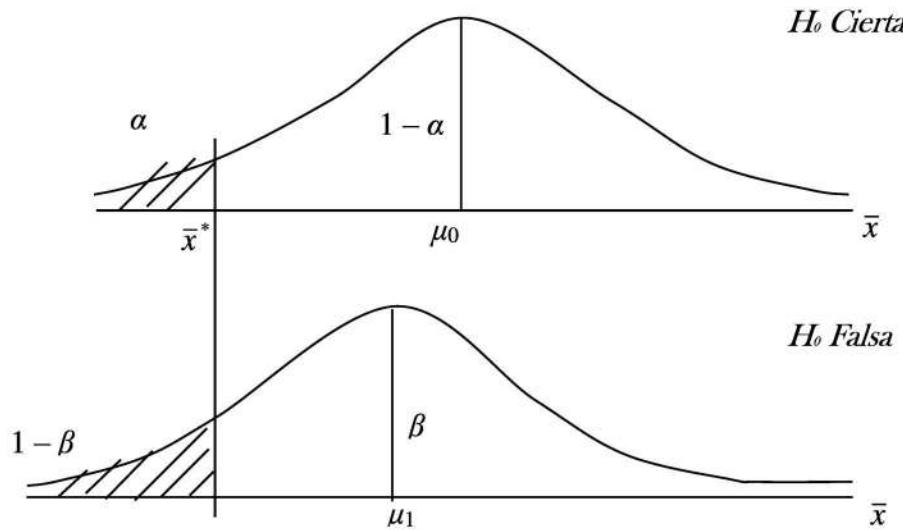
Donde  $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

## Probabilidades involucradas.

Si  $H_0$  es cierta, es decir  $\mu$  no es igual a  $\mu_0$ , entonces existe otra distribución, centrada en el valor de la hipótesis alternativa, es decir  $\mu_1$ , generando las probabilidades  $\beta$  y  $1 - \beta$ .



Gráficamente:



Luego:

$$\alpha = \Pr\{\bar{x} < \bar{x}^* / H_0\} = \Pr\left\{z < \frac{\bar{x}^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$1 - \alpha = \Pr\{\bar{x} \geq \bar{x}^* / H_0\} = \Pr\left\{z \geq \frac{\bar{x}^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$\beta = \Pr\{\bar{x} \geq \bar{x}^* / H_1\} = \Pr\left\{z \geq \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

$$1 - \beta = \Pr\{\bar{x} < \bar{x}^* / H_1\} = \Pr\left\{z < \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

**Docima de hipótesis compuestas:**

**Lateral derecha:**

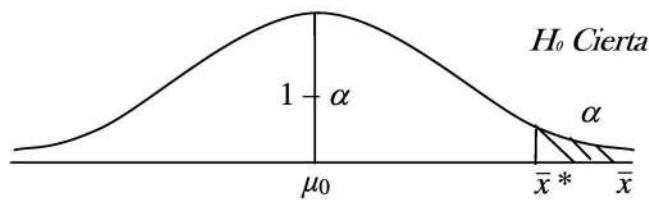
Hipótesis

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$



## Gráfica



## Punto Crítico

$$\bar{x}^* = \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{ó} \quad z^* = z_{1-\alpha}$$

## Regla de Decisión

Si:  $\bar{x} \leq \bar{x}^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
 $\bar{x} > \bar{x}^*$       Rechazar  $H_0$

Si se ha determinado  $z^*$ :

Si:  $z \leq z^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
 $z > z^*$       Rechazar  $H_0$

Donde  $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

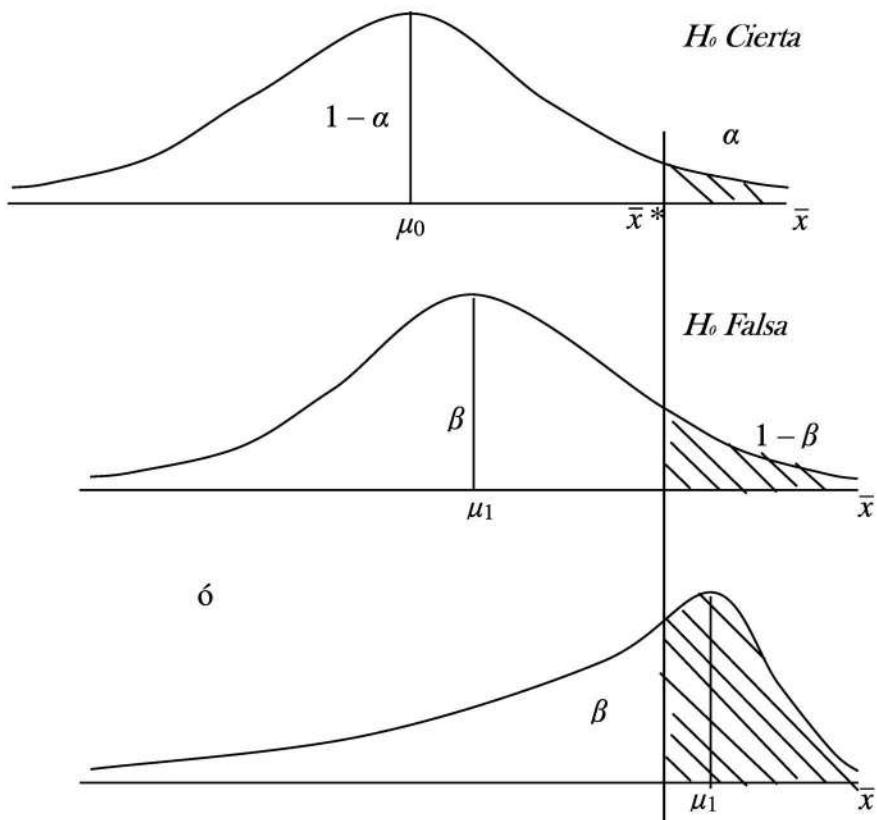
## Probabilidades involucradas.

Si  $H_0$  no es cierta, es decir  $\mu$  no es igual a  $\mu_0$ , entonces existen un conjunto de distribuciones posibles, centradas en los valores de  $\mu$  mayores a  $\mu_0$ , según indica la hipótesis alternativa, generando las probabilidades  $\beta$  y  $1 - \beta$ .

Recuérdese: el valor del parámetro es único, el hecho de no conocerlo nos permite suponer valores para él, dentro de un conjunto de posibles, llamado espacio paramétrico.



Gráficamente:



Y así podríamos suponer más distribuciones.

Luego:

$$\alpha = \Pr\{\bar{x} > \bar{x}^* / H_0\} = \Pr\left\{z > \frac{\bar{x}^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$1 - \alpha = \Pr\{\bar{x} \leq \bar{x}^* / H_0\} = \Pr\left\{z \leq \frac{\bar{x}^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$\beta = \Pr\{\bar{x} \leq \bar{x}^* / H_1\} = \Pr\left\{z \leq \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

$$1 - \beta = \Pr\{\bar{x} > \bar{x}^* / H_1\} = \Pr\left\{z > \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$



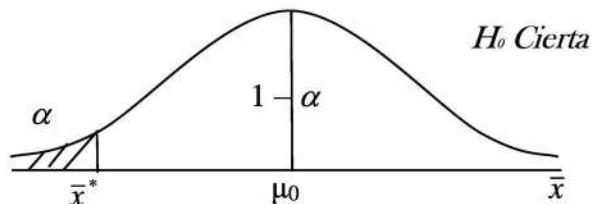
## Lateral Izquierda

### Hipótesis

$$H_0: \mu = \mu_0$$

$$H_a: \mu < \mu_1$$

### Gráfica



### Punto Crítico

$$\bar{x}^* = \mu_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \quad \text{ó} \quad z^* = -z_{1-\alpha}$$

### Regla de Decisión

Si:  $\bar{x} \geq \bar{x}^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
 $\bar{x} < \bar{x}^*$       Rechazar  $H_0$

Si se ha determinado  $z^*$ :

Si:  $z \geq z^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
 $z < z^*$       Rechazar  $H_0$

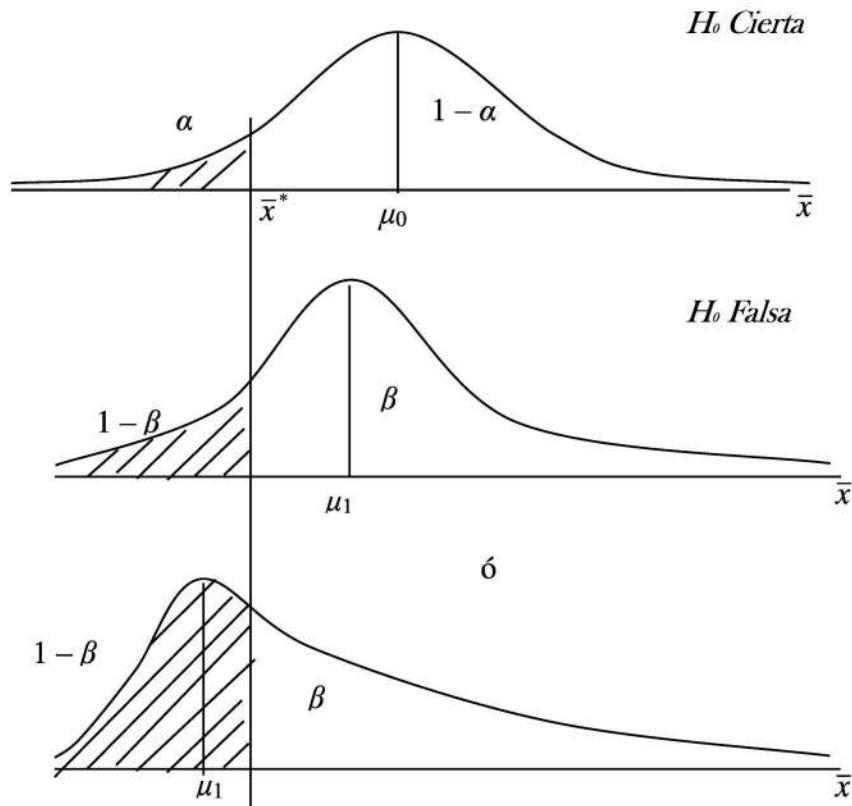
Donde  $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

### Probabilidades involucradas.

Si  $H_0$  no es cierta, es decir  $\mu$  no es igual a  $\mu_0$ , entonces existe un conjunto de distribuciones posibles centradas en valores menores a  $\mu_0$ , según indica la hipótesis alternativa, generando las probabilidades  $\beta$  y  $1 - \beta$ .



Gráficamente:



Y así podríamos suponer más distribuciones.

Luego:

$$\alpha = \Pr\{\bar{x} < \bar{x}^* / H_0\} = \Pr\left\{z < \frac{\bar{x}^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$1 - \alpha = \Pr\{\bar{x} \geq \bar{x}^* / H_0\} = \Pr\left\{z \geq \frac{\bar{x}^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

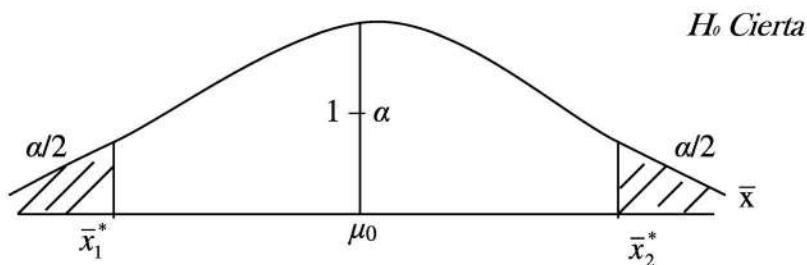
$$\beta = \Pr\{\bar{x} \geq \bar{x}^* / H_1\} = \Pr\left\{z \geq \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

$$1 - \beta = \Pr\{\bar{x} < \bar{x}^* / H_1\} = \Pr\left\{z < \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

**Bilateral**Hipótesis

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

Gráficamente**Puntos Críticos**

$$\bar{x}_1^* = \mu_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{ó} \quad z_1^* = -z_{1-\frac{\alpha}{2}}$$

$$\bar{x}_2^* = \mu_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{ó} \quad z_2^* = z_{1-\frac{\alpha}{2}}$$

Regla de Decisión

Si:  $\bar{x}_1^* \leq \bar{x} \leq \bar{x}_2^*$  Aceptar  $H_0$  (No rechazar  $H_0$ )

$\bar{x} < \bar{x}_1^*$  Rechazar  $H_0$

$\bar{x} > \bar{x}_2^*$  Rechazar  $H_0$

Si se ha determinado  $z$ :

Si:  $z_1^* \leq z \leq z_2^*$  Aceptar  $H_0$  (No rechazar  $H_0$ )

$z < z_1^*$  Rechazar  $H_0$

$z > z_2^*$  Rechazar  $H_0$

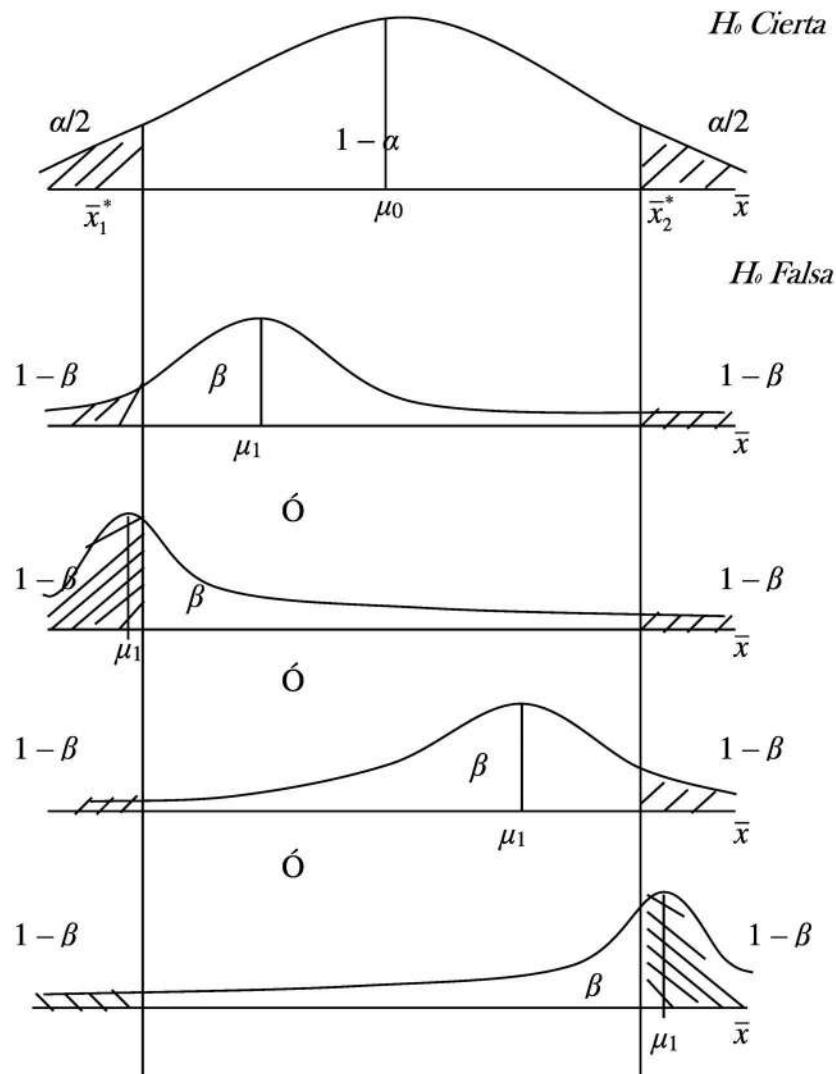
Donde  $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$



### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $\mu$  no es igual a  $\mu_0$ , entonces existe un conjunto de distribuciones posibles, centradas en valores de  $\mu$  menores o mayores a  $\mu_0$ , según indica la hipótesis alternativa, generando las probabilidades  $\beta$  y  $1 - \beta$ .

Gráficamente:



### Nota:

La distribución bajo la  $H_0$  cierta es simétrica respecto a  $\mu_0$ . Es, en base a esta distribución, que se calculan los puntos críticos, que determinan bajo la  $H_0$  falsa  $\beta$  y  $1 - \beta$ . Por lo tanto, las distribuciones bajo el supuesto de la  $H_0$  falsa, son simétricas en relación a sus respectivos valores centrales  $\mu_1$ , pero no en relación a los puntos críticos. Esto hace que  $1 - \beta$  sea la suma del área sombreada, áreas, que insistimos, no son simétricas.



Luego:

$$\alpha = \Pr\{\bar{x} < \bar{x}_1^* / H_0\} + \Pr\{\bar{x} > \bar{x}_2^* / H_0\} = \Pr\left\{z < \frac{\bar{x}_1^* - \mu_0}{\sigma/\sqrt{n}}\right\} + \Pr\left\{z > \frac{\bar{x}_2^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$1 - \alpha = \Pr\{\bar{x}_1^* \leq \bar{x} \leq \bar{x}_2^* / H_0\} = \Pr\left\{\frac{\bar{x}_1^* - \mu_0}{\sigma/\sqrt{n}} \leq z \leq \frac{\bar{x}_2^* - \mu_0}{\sigma/\sqrt{n}}\right\}$$

$$\beta = \Pr\{\bar{x}_1^* \leq \bar{x} \leq \bar{x}_2^* / H_1\} = \Pr\left\{\frac{\bar{x}_1^* - \mu_1}{\sigma/\sqrt{n}} \leq z \leq \frac{\bar{x}_2^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

$$1 - \beta = \Pr\{\bar{x} < \bar{x}_1^* / H_1\} + \Pr\{\bar{x} > \bar{x}_2^* / H_1\} = \Pr\left\{z < \frac{\bar{x}_1^* - \mu_1}{\sigma/\sqrt{n}}\right\} + \Pr\left\{z > \frac{\bar{x}_2^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

### b) Varianza Poblacional Desconocida.

$n < 30$

Corresponde al caso de la distribución "t" de Student, es decir que el estadístico seleccionado deberá ser:

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

$n \geq 30$

Por aplicación al Teorema Central del Límite se aproxima la distribución "t" a la distribución normal, definiendo el estadístico.

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} \sim N(0, 1)$$

### Caso 2: Poblaciones No Normales

*Varianza poblacional Conocida y desconocida con  $n \geq 30$ .*

Por aplicación del T.C.L. se aproxima a la distribución normal, definiendo el estadístico como:

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \text{ Si } \sigma^2 \text{ es conocida.}$$

ó

$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} \sim N(0, 1) \text{ Si } \sigma^2 \text{ es desconocida.}$$



## 7. DOCIMA PARA P. USO DE LA DISTRIBUCIÓN NORMAL

Aplicaremos en este caso lo explicado en estimación de P.

Luego:

$$\theta = P$$

$$\hat{\theta} = P \sim N\left(P, \sqrt{\frac{PQ}{n}}\right)$$

$$K(\hat{\theta}, \theta) = \frac{\hat{P} - P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$$

Nota:

En este caso de Docima utilizaremos  $\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}}$ , pues el valor para P será el supuesto de  $H_0$  ó  $H_1$  según corresponda, de acuerdo a lo que desarrollaremos seguidamente.

### ***Docima de Hipótesis Simples***

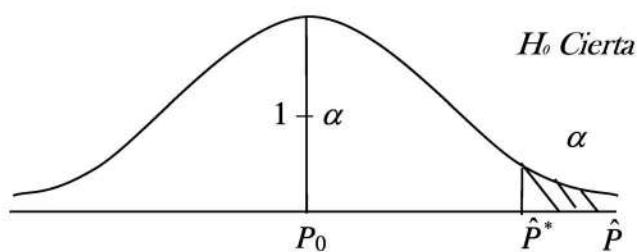
#### ***Docima Lateral Derecha***

##### Hipótesis

$$H_0: P = P_0$$

$$H_1: P = P_1 \quad \text{donde} \quad P_1 > P_0$$

##### Gráfica



##### Punto Crítico

$$\hat{P}^* = P_0 + z_{1-\alpha} \sqrt{\frac{P_0(1-P_0)}{n}} \quad \text{Ó} \quad z^* = z_{1-\alpha}$$



### Regla de decisión

Si:       $\hat{P} \leq \hat{P}^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
               $\hat{P} > \hat{P}^*$       Rechazar  $H_0$

Si se ha determinado  $z^*$

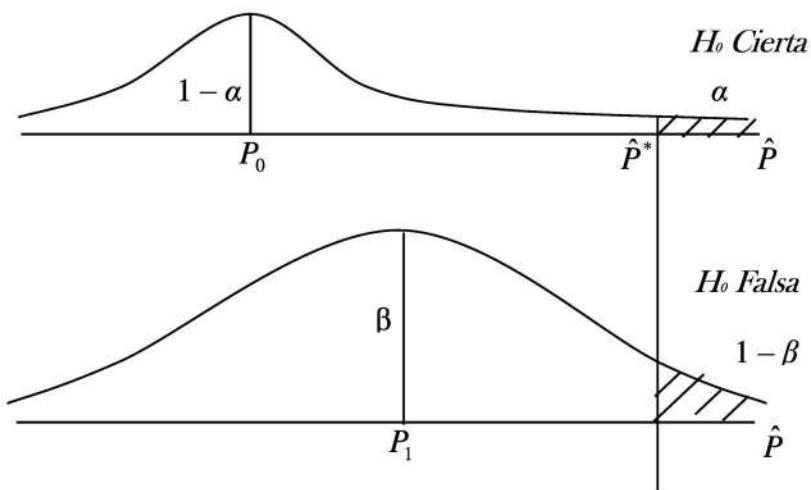
Si       $z \leq z^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
               $z > z^*$       Rechazar  $H_0$

Donde:  $z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$

### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $P_{no}$  es igual a  $P_1$ , entonces existe otra distribución centrada en el valor mayor supuesto en la hipótesis alternativa, es decir  $P_1$ , generando las probabilidades  $\beta$  y  $1 - \beta$ .

Gráficamente





Luego

$$\alpha = \Pr\{\hat{P} > \hat{P}^* / H_0\} = \Pr\left\{z > \frac{\hat{P}^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$
$$1 - \alpha = \Pr\{\hat{P} \leq \hat{P}^* / H_0\} = \Pr\left\{z \leq \frac{\hat{P}^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$
$$\beta = \Pr\{\hat{P} \leq \hat{P}^* / H_1\} = \Pr\left\{z \leq \frac{\hat{P}^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$
$$1 - \beta = \Pr\{\hat{P} > \hat{P}^* / H_1\} = \Pr\left\{z > \frac{\hat{P}^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$

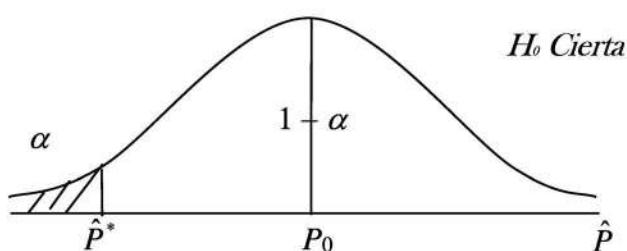
### Decima Lateral Izquierda

#### Hipótesis

$$H_0: P = P_0$$

$$H_1: P = P_1 \quad \text{donde} \quad P_1 < P_0$$

#### Gráfica



#### Punto Crítico

$$\hat{P}^* = P_0 - z_{1-\alpha} \sqrt{\frac{P_0(1-P_0)}{n}} \quad \text{Ó} \quad z^* = -z_{1-\alpha}$$



### Regla de decisión

Si:       $\hat{P} \geq \hat{P}^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
               $\hat{P} < \hat{P}^*$       Rechazar  $H_0$

Si se ha determinado  $z^*$

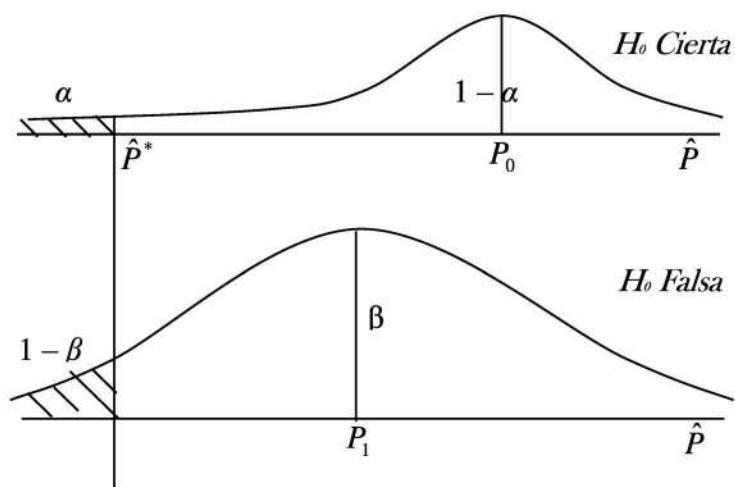
Si       $z \geq z^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
               $z < z^*$       Rechazar  $H_0$

Donde:  $z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$

### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $P$  no es igual a  $P_0$ , entonces existe otra distribución centrada en el valor mayor supuesto en la hipótesis alternativa, es decir  $P_1$ , generando las probabilidades  $\beta$  y  $1 - \beta$ .

Gráficamente





Luego:

$$\alpha = \Pr\{\hat{P} < \hat{P}^* / H_0\} = \Pr\left\{z < \frac{\hat{P}^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$
$$1 - \alpha = \Pr\{\hat{P} \geq \hat{P}^* / H_0\} = \Pr\left\{z \geq \frac{\hat{P}^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$
$$\beta = \Pr\{\hat{P} \geq \hat{P}^* / H_1\} = \Pr\left\{z \geq \frac{\hat{P}^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$
$$1 - \beta = \Pr\{\hat{P} < \hat{P}^* / H_1\} = \Pr\left\{z < \frac{\hat{P}^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$

### Docima de Hipótesis Compuestas

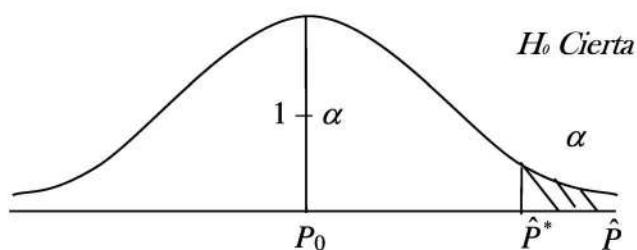
#### Docima Lateral Derecha

##### Hipótesis

$$H_0: P = P_0$$

$$H_1: P > P_0$$

##### Gráficamente



##### Punto Crítico

$$\hat{P}^* = P_0 + z_{1-\alpha} \sqrt{\frac{P_0(1-P_0)}{n}} \quad \text{Ó} \quad z^* = z_{1-\alpha}$$



### Regla de decisión

Si:       $\hat{P} \leq \hat{P}^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
               $\hat{P} > \hat{P}^*$       Rechazar  $H_0$

Si se ha determinado  $z^*$

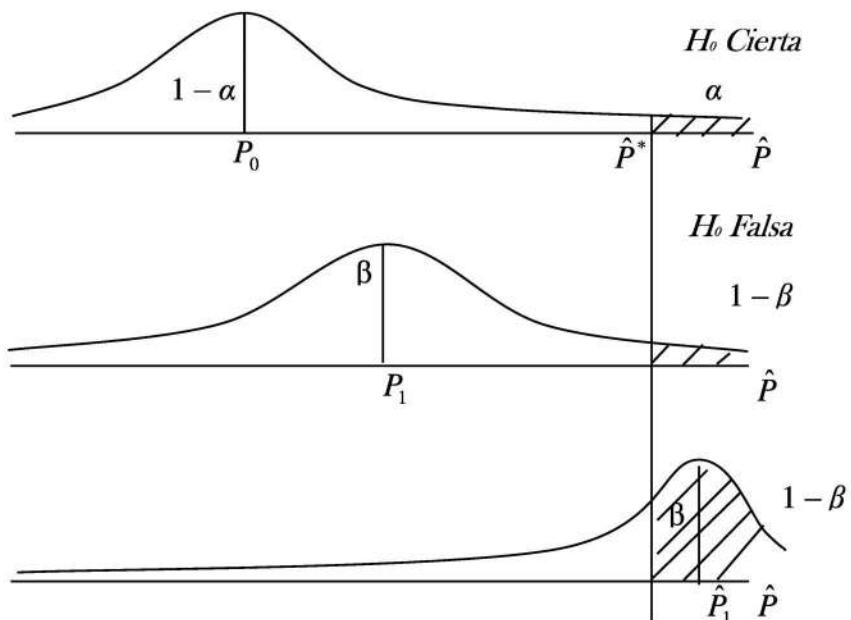
Si       $z \leq z^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
               $z > z^*$       Rechazar  $H_0$

Donde:  $z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$

### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $P$  no es igual a  $P_0$ , entonces existe otro conjunto de distribuciones posibles, centradas en valores de  $P$  mayores a  $P_0$ , según indica la hipótesis alternativa, generando las probabilidades  $\beta$  y  $1 - \beta$ .

Gráficamente



Y así, como antes, podríamos suponer más distribuciones.



Luego:

$$\alpha = \Pr\{\hat{P} > \hat{P}^* / H_0\} = \Pr\left\{z > \frac{\hat{P}^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$
$$1 - \alpha = \Pr\{\hat{P} \leq \hat{P}^* / H_0\} = \Pr\left\{z \leq \frac{\hat{P}^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$
$$\beta = \Pr\{\hat{P} \leq \hat{P}^* / H_1\} = \Pr\left\{z \leq \frac{\hat{P}^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$
$$1 - \beta = \Pr\{\hat{P} > \hat{P}^* / H_1\} = \Pr\left\{z > \frac{\hat{P}^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$

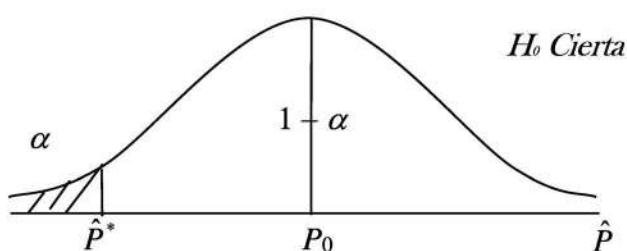
### Decima Lateral Izquierda

#### Hipótesis

$$H_0: P = P_0$$

$$H_1: P < P_0$$

#### Gráficamente



#### Punto Crítico

$$\hat{P}^* = P_0 - z_{1-\alpha} \sqrt{\frac{P_0(1-P_0)}{n}}$$
      ó       $z^* = -z_{1-\alpha}$



### Regla de decisión

Si:       $\hat{P} \geq \hat{P}^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
               $\hat{P} < \hat{P}^*$       Rechazar  $H_0$

Si se ha determinado  $z^*$

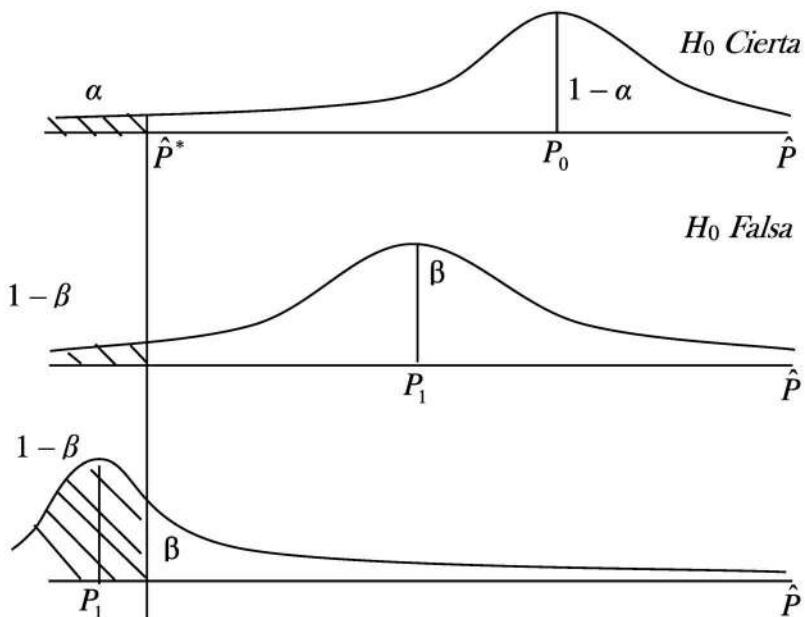
Si       $z \geq z^*$       Aceptar  $H_0$  (No rechazar  $H_0$ )  
               $z < z^*$       Rechazar  $H_0$

Donde:  $z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}}$

### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $P$  no es igual a  $P_0$ , entonces existe otro conjunto de distribuciones posibles, centradas en valores de  $P$  menores a  $P_0$ , según indica la hipótesis alternativa, generando las probabilidades  $\beta$  y  $1 - \beta$ .

Gráficamente





Luego:

$$\alpha = \Pr\{\hat{P} < \hat{P}^*/H_0\} = \Pr\left\{z < \frac{\hat{P}^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$

$$1 - \alpha = \Pr\{\hat{P} \geq \hat{P}^*/H_0\} = \Pr\left\{z \geq \frac{\hat{P}^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$

$$\beta = \Pr\{\hat{P} \geq \hat{P}^*/H_1\} = \Pr\left\{z \geq \frac{\hat{P}^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$

$$1 - \beta = \Pr\{\hat{P} < \hat{P}^*/H_1\} = \Pr\left\{z < \frac{\hat{P}^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$

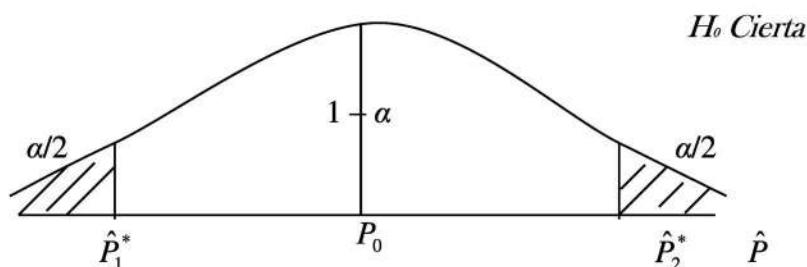
### **Docima Bilateral**

#### Hipótesis

$$H_0: P = P_0$$

$$H_1: P \neq P_0$$

#### Gráficamente



#### Punto Crítico

$$\hat{P}_1^* = P_0 - z_{\frac{\alpha}{2}} \sqrt{\frac{P_0(1-P_0)}{n}} \quad \text{Ó} \quad z_1^* = -z_{1-\alpha}$$

$$\hat{P}_2^* = P_0 + z_{\frac{\alpha}{2}} \sqrt{\frac{P_0(1-P_0)}{n}} \quad \text{Ó} \quad z_2^* = z_{1-\alpha}$$

#### Regla de decisión



Si:  $\hat{P}_1^* \leq \hat{P} \leq \hat{P}_2^*$  Aceptar  $H_0$  (No rechazar  $H_0$ )

$\hat{P} < \hat{P}_1^*$  Rechazar  $H_0$

$\overset{\circ}{\hat{P}} > \hat{P}_2^*$  Rechazar  $H_0$

Si se ha determinado  $z^*$

Si  $z_1^* \leq z \leq z_2^*$  Aceptar  $H_0$  (No rechazar  $H_0$ )

$z < z_1^*$  Rechazar  $H_0$

$\overset{\circ}{z} > z_2^*$  Rechazar  $H_0$

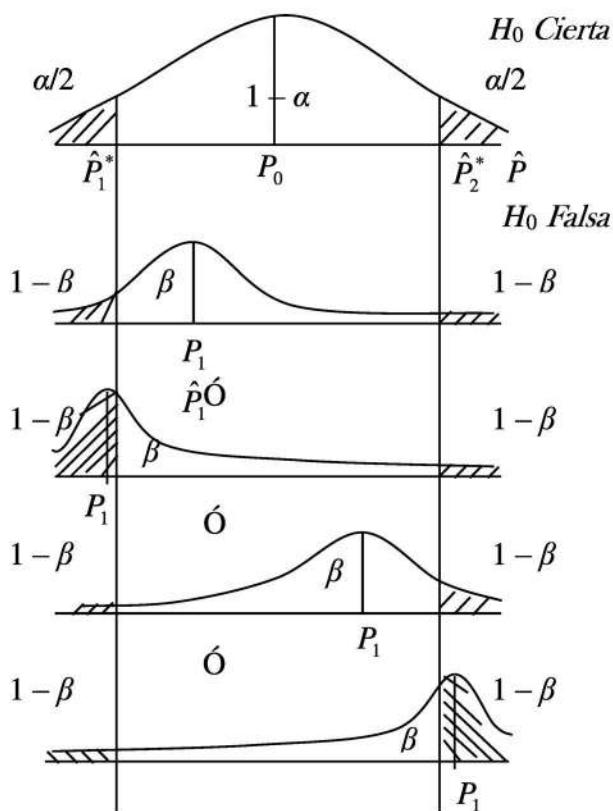
Donde:  $z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$

### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $P$  no es igual a  $P_0$ , entonces existe otro conjunto de distribuciones posibles, centradas en valores de  $P$  menores ó mayores a  $P_0$ , según indica la hipótesis alternativa, generando las probabilidades  $\beta$  y  $1 - \beta$ .



Gráficamente



**Nota:** Remitimos a nota en Probabilidades involucradas de la docima bilateral para  $\mu$ , en relación a  $1 - \beta$ .

Luego:

$$\alpha = \Pr\{\hat{P} < \hat{P}_1^* / H_0\} + \Pr\{\hat{P} > \hat{P}_2^* / H_0\} = \Pr\left\{z < \frac{\hat{P}_1^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\} + \Pr\left\{z > \frac{\hat{P}_2^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$

$$1 - \alpha = \Pr\{\hat{P}_1^* \leq \hat{P} \leq \hat{P}_2^* / H_0\} = \Pr\left\{\frac{\hat{P}_1^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \leq z \leq \frac{\hat{P}_2^* - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}\right\}$$

$$\beta = \Pr\{\hat{P}_1^* \leq \hat{P} \leq \hat{P}_2^* / H_1\} = \Pr\left\{\frac{\hat{P}_1^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}} \leq z \leq \frac{\hat{P}_2^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$

$$1 - \beta = \Pr\{\hat{P} < \hat{P}_1^* / H_1\} + \Pr\{\hat{P} > \hat{P}_2^* / H_1\} = \Pr\left\{z < \frac{\hat{P}_1^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\} + \Pr\left\{z > \frac{\hat{P}_2^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}\right\}$$



## 8. CURVA OPERATORIA CARACTERÍSTICA (OC) Y CURVA DE POTENCIA

$1 - \beta = \Pr \{ \text{Rechazar } H_0 / H_0 \text{ Falsa} \} = \text{Potencia de la docima.}$

Esta probabilidad mide la potencia que tiene la docima para rechazar la hipótesis nula cuando es falsa.

Además, dijimos que  $\alpha$  y  $\beta$  varían en sentido inverso, en consecuencia  $\alpha$  y  $1 - \beta$  varían en el mismo sentido.

Cuando la hipótesis alternativa es compuesta, según ya mencionamos, hay un conjunto de valores posibles para el parámetro (espacio paramétrico) y en consecuencia un conjunto de distribuciones posibles para el estimador. Así, en cada una de estas distribuciones se genera una probabilidad distinta ( $\beta$ ) de aceptar  $H_0$  cuando es falsa, y por lo tanto, hay un valor diferente para  $1 - \beta$ .

En base a ello se definen:

**Curva de Potencia**, a aquella que relaciona todos los valores posibles del parámetro con la probabilidad de rechazar  $H_0$  para un determinado  $\alpha$ . Si la  $H_0$  es cierta,  $1 - \beta$  será distinto de  $\alpha$  cuando  $H_0$  sea falsa.

**Curva Operatoria Característica o Curva OC**, es aquella que relaciona los posibles valores del parámetro con la probabilidad de aceptar  $H_0$ , para un determinado  $\alpha$ . Es el complemento de la función de potencia. Cuando  $H_0$  es cierta,  $\beta = 1 - \alpha$ , sólo será distinta ( $\beta \neq 1 - \alpha$ ) si  $H_0$  es falsa.

### Ejemplificando

Gráficamente para una docima lateral derecha

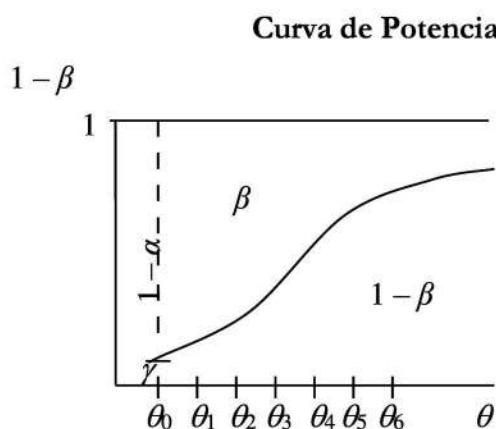


Fig. 1

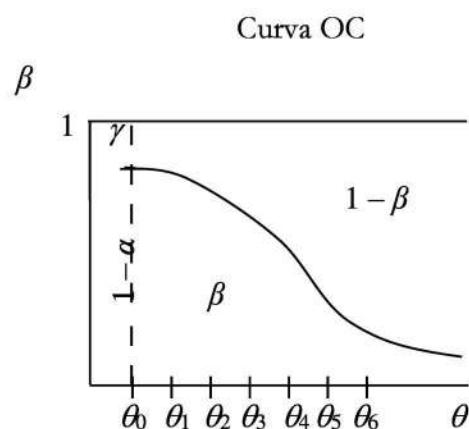


Fig. 2

Fig. 1

Hemos considerado, en la abscisa a los diferentes valores del parámetro  $\theta$  y en la ordenada sus correspondientes probabilidades  $1 - \beta$ . Debajo de la curva encontramos la potencia de la docima para cada valor supuesto de  $\theta$ , y su diferencia con 1 (máxima probabilidad) indica  $\beta$ , probabilidad del Error Tipo II. En el punto en que  $\theta = \theta_0$ ,  $1 - \beta = \alpha$  y  $\beta = 1 - \alpha$ .

Fig. 2

Realizaremos un razonamiento idéntico, pero ahora los valores por debajo de la curva corresponden a  $\beta$  (ya que la ordenada se rotula con las correspondientes Probabilidades  $\beta$ ) y sus diferencias con 1 (probabilidad máxima) corresponden a  $1 - \beta$ .

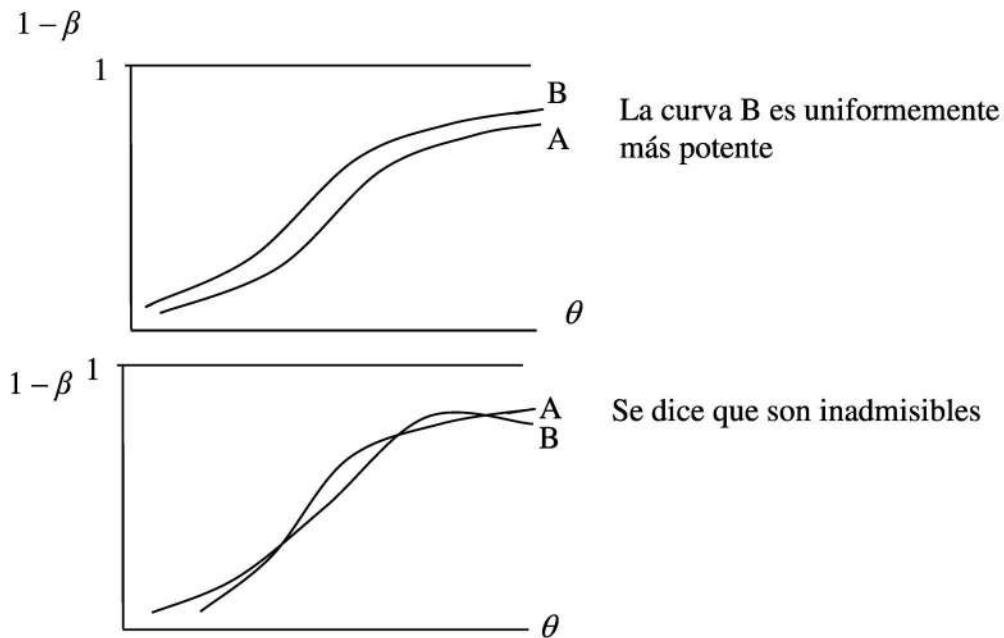
Por otro lado, consideraremos la influencia que  $\alpha$ ,  $n$  y  $\sigma^2$  ejercen sobre la potencia de la docima, así diremos:

1) Aumentando o disminuyendo  $\alpha$ , se puede aumentar o disminuir  $1 - \beta$ , pues varían en idéntico sentido.

2) Mientras mayor sea  $n$  (tamaño de la muestra), menor será la variabilidad del estimador de  $\theta$ , o sea, menor será  $\sigma_{\hat{\theta}}$  y por lo tanto más concentrado alrededor de  $\theta$ , lo que hará disminuir a  $\beta$ , e incrementar  $1 - \beta$ , lo cual hará más potente a la docima. Este efecto también se logra a través de una menor varianza poblacional, es decir, reduciendo  $\sigma^2$ .

Las curvas de potencia son útiles para comparar dos o más docima y determinar cuál de ellas es más potente, o sea cuál es más eficaz.

La docima más conveniente será aquella que sea uniformemente más potente, es decir que para todos los posibles valores del parámetro (en todo el espacio paramétrico) la probabilidad  $1 - \beta$  sea superior, lo cual asegura para un  $\alpha$  dado, una menor probabilidad para el Error Tipo II. Se dice que las otras son inadmisibles, pues según el valor del parámetro pueden tener un  $1 - \beta$  mayor o menor, y al no conocer el verdadero valor del parámetro, no sabemos a qué situación vamos a enfrentarnos.



Entonces, la docima uniformemente más potente, minimiza  $\beta$ , manteniendo  $\alpha$ . Como no conocemos el verdadero valor de  $\theta$ , la única forma de minimizar, es utilizando una docima uniformemente más potente, porque así tenemos la seguridad de que cualquiera sea el valor de  $\theta$ ,  $\beta$  será siempre menor que en las otras docima.

### Ejemplo

Vamos a plantear un caso base, a partir del cual produciremos tres tipos de modificaciones, con el fin de ejemplificar los distintos tipos de docima. En cada una de ellas, construiremos las Curvas de Potencia y OC.

Supongamos que según un proceso de fabricación se produce una pieza que tiene una característica variable con distribución normal, cuya media es de 100 y desviación típica de 12.

1) Se propone un nuevo método de fabricación, que se asegura, aumenta la media, lo cual significa una ventaja para la fábrica. A los fines de verificarlo se fabrican 36 piezas con el nuevo proceso, y se establece un nivel de significación del 0,0228. La muestra dio como resultado una media de 103.

Se pregunta, ¿Es aconsejable cambiar el proceso?

Nos enfrentaremos a un problema de toma de decisiones: continuar con el actual proceso ó cambiarlo. Entonces una docima permitirá esta decisión.

En primer término el parámetro objetivo es la Media Poblacional. La población de la cual se extrae la muestra es Normal, y la varianza poblacional es conocida. Luego será de aplicación de Distribución Normal.

Entonces:

$$\theta = \mu$$



$$\hat{\theta} = \bar{x} \sim N(\mu; \sigma/\sqrt{n})$$
$$K(\hat{\theta}; \theta) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Antes de plantear la hipótesis, extraigamos los datos:

Proceso Actual:

$$\begin{aligned}\mu &= 100 \\ \sigma &= 12\end{aligned}$$

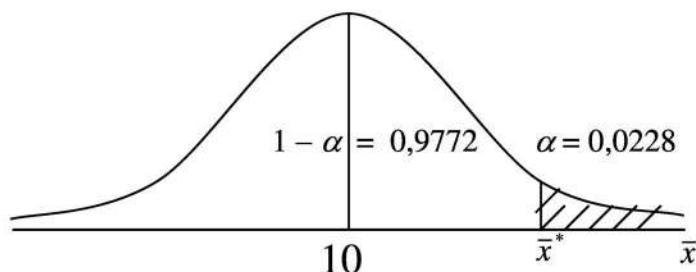
Se asegura que el cambio aumentará la media.

$$\begin{aligned}n &= 36 \\ \bar{x} &= 103 \\ \alpha &= 0,0228\end{aligned}$$

Luego:

$$\begin{aligned}H_0 : \mu &= 100 && \text{(Media sin modificación)} \\ H_1 : \mu &> 100 && \text{(Modificación propuesta)}\end{aligned}$$

La docima es lateral derecha, y su gráfica bajo el supuesto de  $H_0$  cierta será:



Punto Crítico:

$$\begin{aligned}x^* &= \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} = \mu_0 + z_{0,9772} \frac{\sigma}{\sqrt{n}} = \\ &= 100 + 2 \times \frac{12}{\sqrt{36}} = \underline{\underline{104}}\end{aligned}$$

Regla de decisión:

$$\begin{array}{lll} \text{Si} & \bar{x} \leq \bar{x}^* & \text{Aceptar } H_0 \quad \bar{x} \leq 104 \\ & \bar{x} > \bar{x}^* & \text{Rechazar } H_0 \quad \bar{x} > 104 \end{array}$$

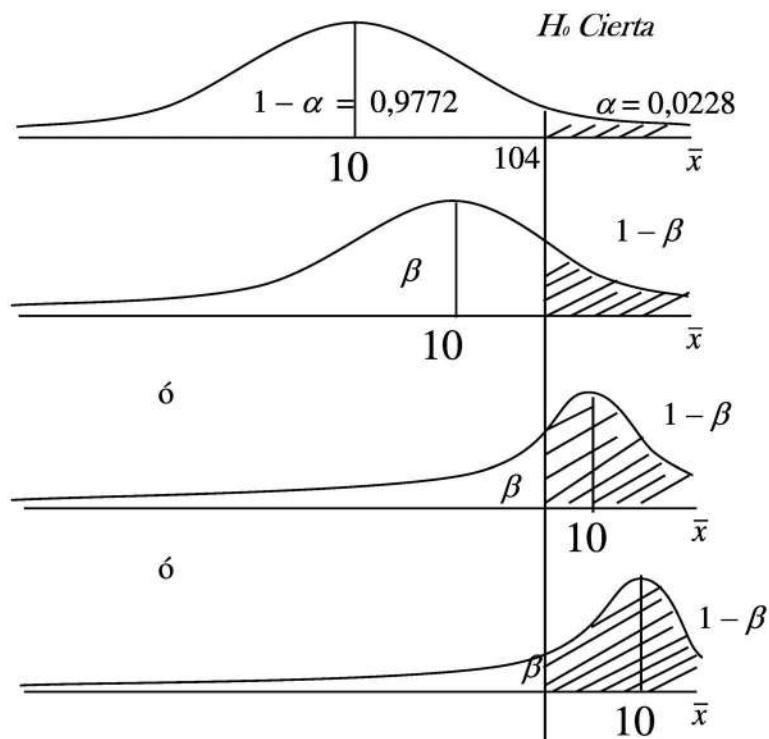


Siendo  $\bar{x} = 103 \leq \bar{x}^* = 104$ , aceptamos, o bien no rechazamos  $H_0$ . Esto indicará que el nuevo proceso no aumenta la media, pues la muestra extraída sigue teniendo media igual a 100 (Cuando aceptamos  $H_0$ , aceptamos el valor supuesto en ella). Por lo tanto se aconseja no modificar el proceso actual.

Además consideremos que el hecho de aceptar, no significa que  $H_0$  sea cierta. Entonces, si fuera falsa, ¿Cuál es el verdadero valor de entre todos los posibles? ¿Cuáles son las probabilidades  $\beta$  y  $1 - \beta$  a las que nos podemos enfrentar?

Para calcularlas, supondremos otros valores para  $\mu$ , que en el caso de una docima lateral derecha, serán mayores que 100. Consideraremos por ejemplo 102, 105 y 107.

Gráficamente:



Entonces:

$$\begin{aligned}\beta &= \Pr\{\bar{x} \leq \bar{x}^* / H_1\} = \Pr\left\{z \leq \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\} \\ 1 - \beta &= \Pr\{\bar{x} > \bar{x}^* / H_1\}\end{aligned}$$

Calculemos  $\beta$  y  $1 - \beta$  surgirá como su complemento.

$$\begin{aligned}1) \quad \beta &= \Pr\{\bar{x} \leq \bar{x}^* / \mu_1 = 102\} = \Pr\left\{z \leq \frac{104 - 102}{12/\sqrt{36}}\right\} = \Pr\{z \leq 1\} = 0,8413 \\ 1 - \beta &= 1 - 0,8413 = \underline{\underline{0,1583}}\end{aligned}$$



$$2) \quad \beta = \Pr\{\bar{x} \leq \bar{x}^* / \mu_1 = 105\} = \Pr\left\{z \leq \frac{104 - 105}{12/\sqrt{36}}\right\} = \Pr\{z \leq -0,5\} = 1 - \Pr\{z < 0,5\} = \\ = 1 - 0,6915 = 0,3085$$

$$1 - \beta = 1 - 0,3085 = \underline{\underline{0,6915}}$$

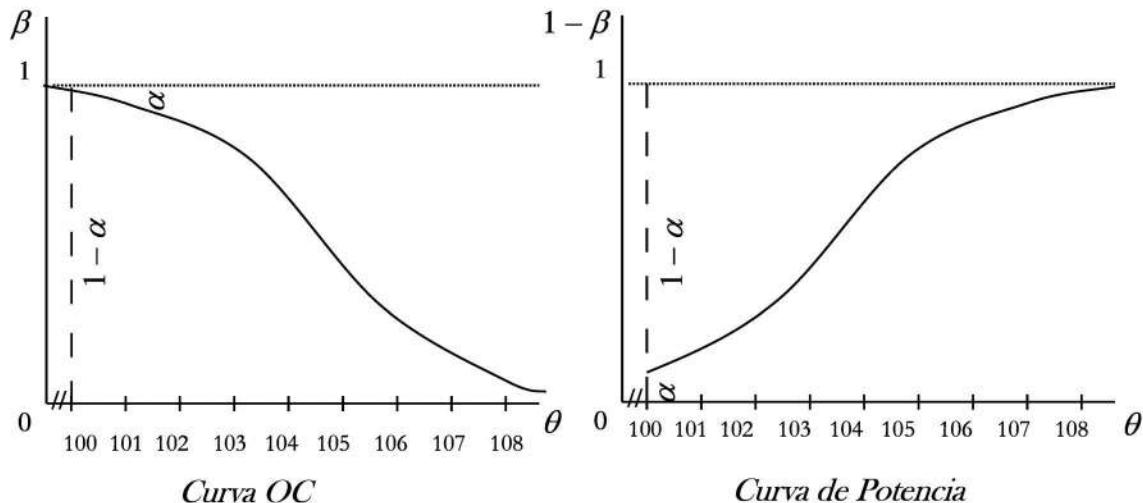
$$3) \quad \beta = \Pr\{\bar{x} \leq \bar{x}^* / \mu_1 = 107\} = \Pr\left\{z \leq \frac{104 - 107}{12/\sqrt{36}}\right\} = \Pr\{z \leq -1,5\} = 1 - \Pr\{z < 1,5\} = \\ = 1 - 0,9332 = 0,0668$$

$$1 - \beta = 1 - 0,0668 = \underline{\underline{0,9332}}$$

Así, podremos construir la siguiente tabla:

$\mu$	$z^* = \frac{104 - \mu}{2}$	$\beta$	$1 - \beta$
100	2	0.9772 = $1 - \alpha$	0.0228 = $\alpha$
101	1.5	0.9332	0.0668
102	1	0.8413	0.1587
103	0.5	0.6915	0.3085
104	0	0.50	0.50
105	-0.5	0.3085	0.6915
106	-1	0.1587	0.8413
107	-1.5	0.0668	0.9321
108	-2	0.0228	0.9772

A partir de esta información graficamos:



2) Supongamos ahora que la ventaja para la fábrica está en disminuir el valor de la media y que se espera que el nuevo método propuesto produzca ese efecto. En la muestra de 36 piezas que se toma, la media es ahora de 92.

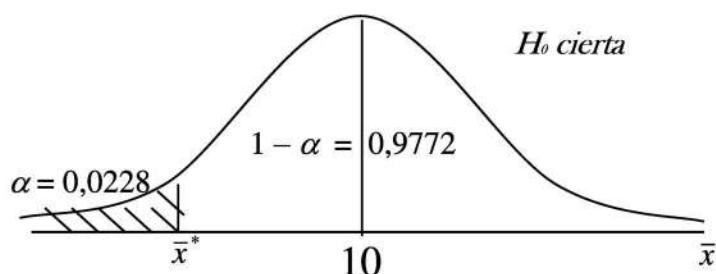
Esto implica plantear la hipótesis de la siguiente manera:

$$H_0: \mu = 100$$

Entonces ahora es izquierda.

$$H_1: \mu < 100$$

Gráficamente:



Punto Crítico:

$$\begin{aligned} x^* &= \mu_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}} = \mu_0 - z_{0,9772} \frac{\sigma}{\sqrt{n}} = \\ &= 100 - 2 \times \frac{12}{\sqrt{36}} = 96 \end{aligned}$$

Regla de decisión:

$$\begin{array}{ll} \text{Si} & \bar{x} < 96 \text{ Rechazar } H_0 \\ & \bar{x} \geq 96 \text{ Aceptar } H_0 \quad (\text{No rechazar } H_0) \end{array}$$



Siendo  $\bar{x} = 92 < \bar{x}^* = 96$ , rechazamos  $H_0$ , entonces la muestra tomada evidencia que ha sido extraída de una población que tiene una media inferior a 100, lo cual manifiesta que el nuevo proceso reduce la media. La acción a aconsejar es implementar el nuevo proceso.

Aclaramos aquí que el hecho de rechazar no significa que sea totalmente falsa.

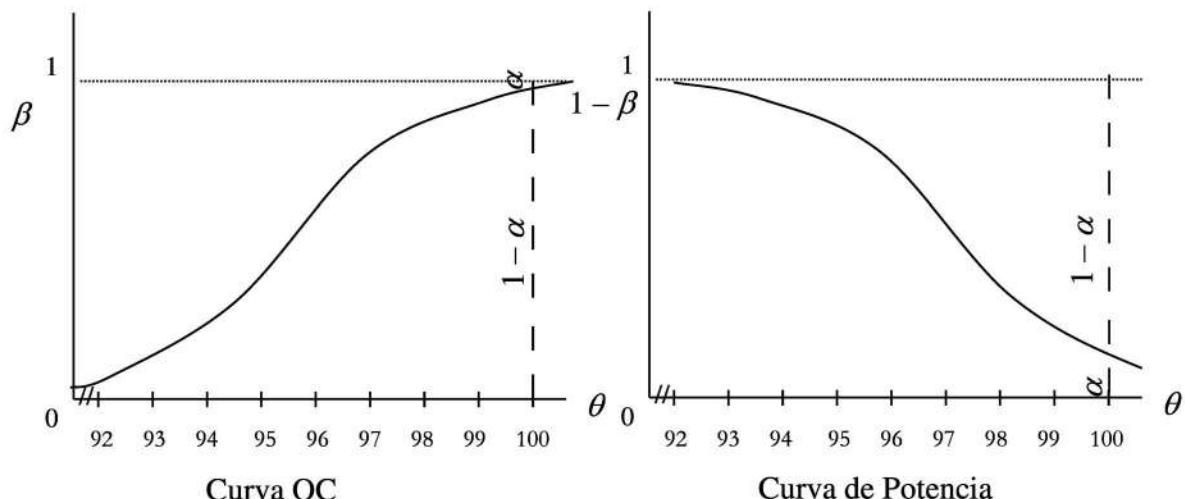
Calculamos ahora las probabilidades  $\beta$  y  $1 - \beta$ , considerando en este caso valores inferiores a 100.

Además:

$$\beta = \Pr\{\bar{x} \geq \bar{x}^* / \mu_1\} = \Pr\left\{z \geq \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$
$$1 - \beta = \Pr\{\bar{x} < \bar{x}^* / \mu_1\} = \Pr\left\{z < \frac{\bar{x}^* - \mu_1}{\sigma/\sqrt{n}}\right\}$$

$\mu$	$z^* = \frac{96 - \mu}{2}$	$\beta$	$1 - \beta$
92	2	0.0228	0.9772
93	1.5	0.0668	0.9332
94	1	0.1587	0.8413
95	0.5	0.3085	0.6915
96	0	0.50	0.50
97	-0.5	0.6915	0.3085
98	-1	0.8413	0.1587
99	-1.5	0.9321	0.0668
100	-2	0.9772 = 1 - $\alpha$	0.0228 = $\alpha$

A partir de esta información graficamos:





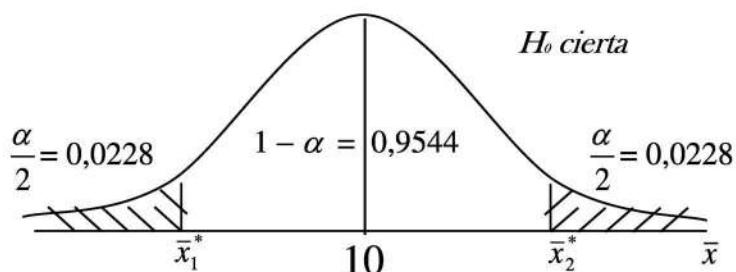
3) Si ahora consideramos que la transformación hace temer que se modifique sustancialmente la media, lo que perjudicaría a la fábrica, definiendo  $\alpha = 0,0456$  y encontrando una media de 95, ¿qué se le aconsejará a la fábrica?

El planteo la hipótesis será:

$$H_0 : \mu = 100$$

$$H_1 : \mu \neq 100$$

Gráficamente:



Punto Crítico:

$$x_1^* = \mu_0 - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 100 - 2 \times \frac{12}{\sqrt{36}} = \underline{\underline{96}}$$

$$x_2^* = \mu_0 + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 100 + 2 \times \frac{12}{\sqrt{36}} = \underline{\underline{104}}$$

Regla de decisión:

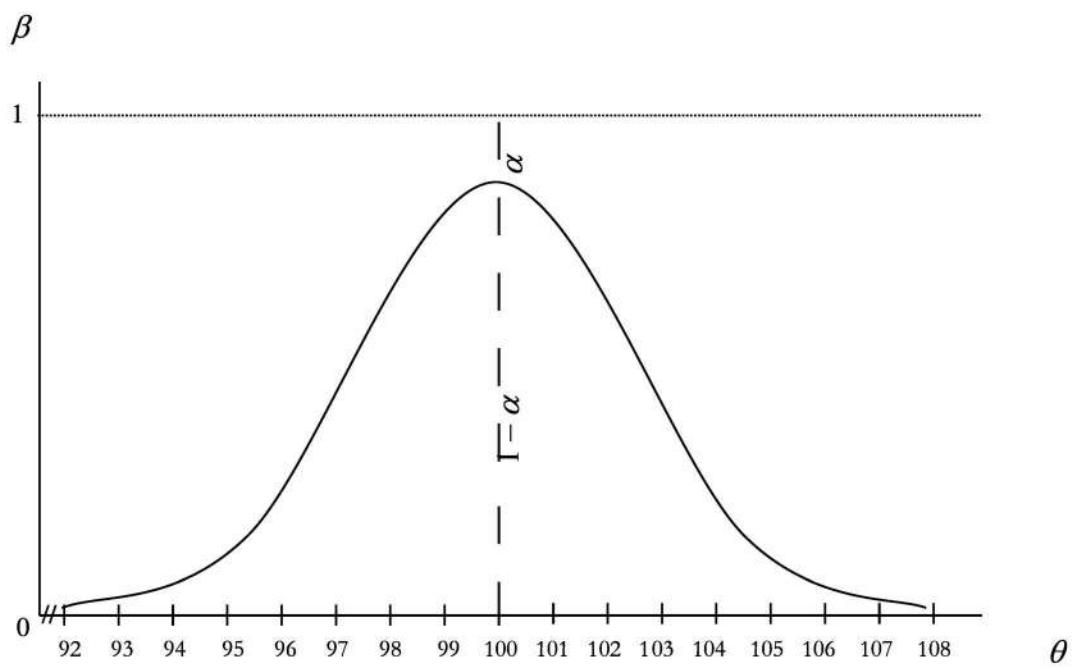
Si	$96 \leq \bar{x} \leq 104$	Aceptar $H_0$	(No rechazar $H_0$ )
	$\bar{x} < 96$	Rechazar $H_0$	
	$\bar{x} > 104$	Rechazar $H_0$	

Siendo  $\bar{x} = 95 < \bar{x}^* = 96$ , rechazamos  $H_0$ , por lo que aconsejaremos no efectuar la modificación propuesta.

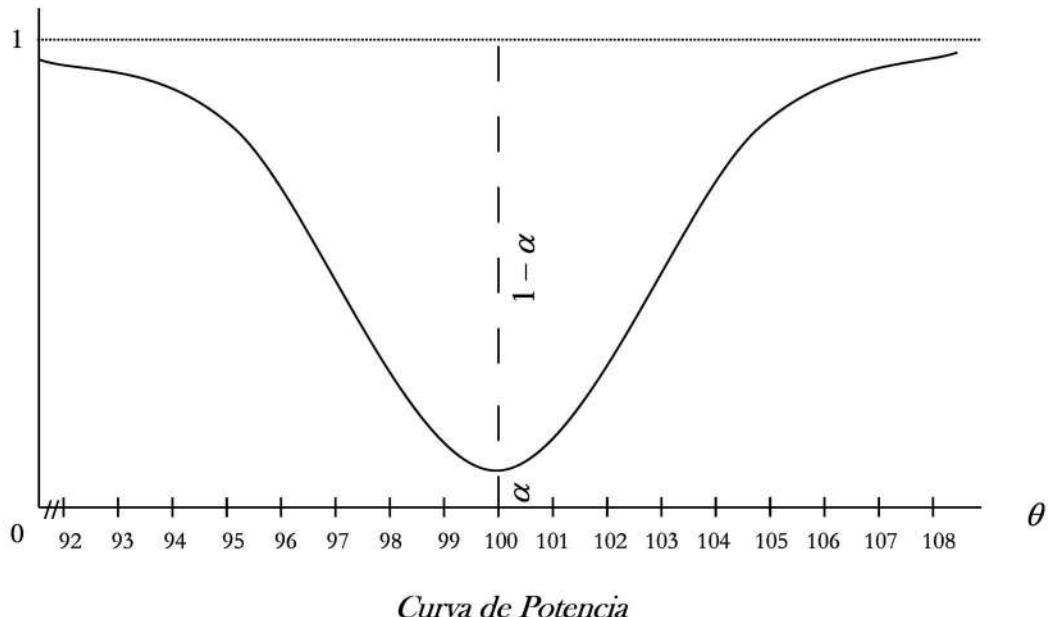
El cálculo de las probabilidades  $\beta$  y  $1 - \beta$  se sintetiza en la siguiente tabla, considerando valores menores y mayores a 100.



$\mu$	$z_2^*$	$z_1^*$	$\beta$	$1 - \beta$
92	6	2	0.0228	0.9772
93	5.5	1.5	0.07	0.93
94	5	1	0.16	0.84
95	4.5	0.5	0.31	0.09
96	4	0	0.50	0.50
97	3.5	-0.5	0.69	0.31
98	3	-1	0.840	0.16
99	2.5	-1.5	0.927	0.073
100	2	-2	0.9544	0.0456( $\alpha$ )
101	1.5	-2.5	0.927	0.073
102	1	-3	0.84	0.16
103	0.5	-3.5	0.69	0.31
104	0	-4	0.50	0.50
105	-0.5	-4.5	0.31	0.69
106	-1	-5	0.16	0.84
107	-1.5	-5.5	0.07	0.93
108	-2	-6	0.0228	0.9772



Curva OC

 $I - \beta$ 

Curva de Potencia

## 9. DOCIMA PARA LA VARIANZA

Planteamos en el parámetro a docimar, el estimador y el estadístico

$$\theta = \sigma^2$$

$$\hat{\theta} = \hat{\sigma}^2 = \hat{s}^2$$

$$K(\hat{\theta}; \theta) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} = \frac{\hat{\sigma}^2(n-1)}{\sigma^2} \approx \chi^2_{(n-1)}$$

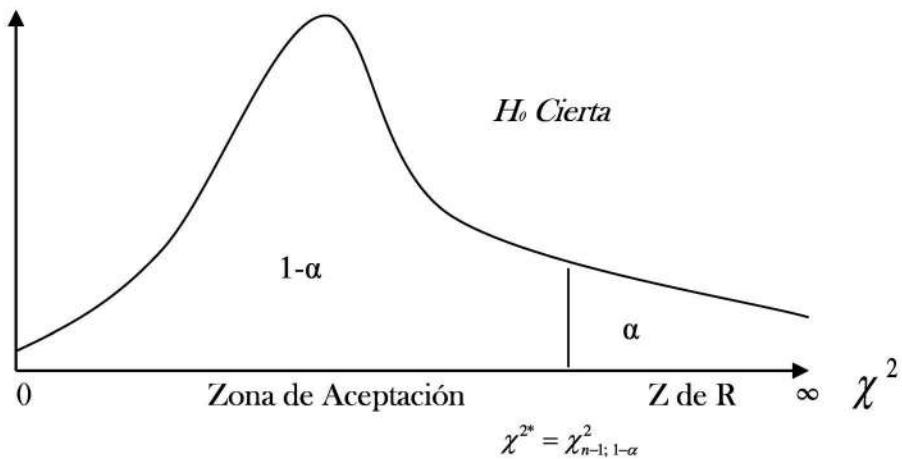
Docima de Hipótesis Simples

Lateral Derecha

Hipótesis

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 = \sigma_1^2 \quad \text{donde } \sigma_1^2 > \sigma_0^2$$



Punto Crítico:  $\chi^{2*} = \chi_{n-1; 1-\alpha}^2$

#### Regla de decisión

Si  $\chi^2 \leq \chi^{2*}$  Acepto  $H_0$   
 $\chi^2 > \chi^{2*}$  Rechazo  $H_0$       Donde  $\chi^2 = \frac{\hat{\sigma}^2(n-1)}{\sigma_0^2}$

#### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $\sigma^2$  no es igual a  $\sigma_0^2$ , entonces existe otra distribución, con una mayor dispersión, generando las probabilidades  $\beta$  y  $1-\beta$ , luego:

$$\alpha = \Pr\{\chi^2 > \chi^{2*} / H_0\}$$

$$1-\alpha = \Pr\{\chi^2 \leq \chi^{2*} / H_0\}$$

$$\beta = \Pr\{\chi^2 \leq \chi^{2*} / H_1\}$$

$$1-\beta = \Pr\{\chi^2 > \chi^{2*} / H_1\}$$

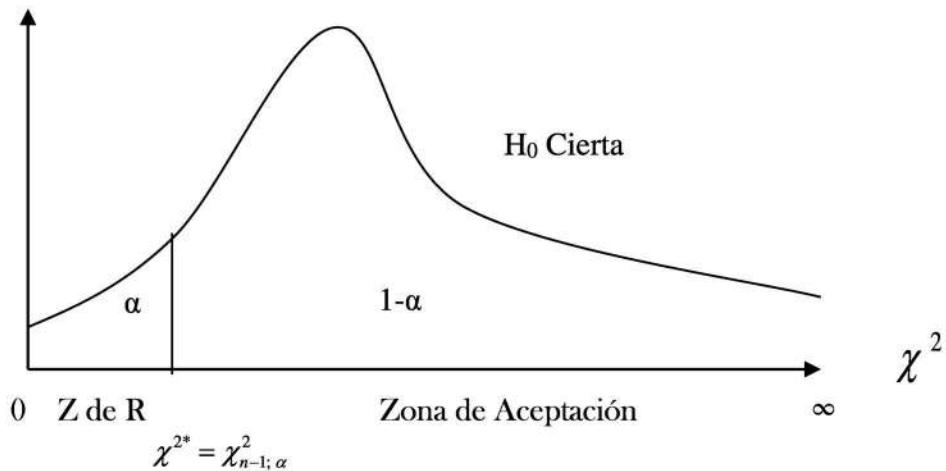


### Lateral Izquierda

#### Hipótesis

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2 \quad \text{donde } \sigma^2 < \sigma_0^2$$



#### Punto Crítico:

$$\chi^{2*} = \chi_{n-1; \alpha}^2$$

#### Regla de decisión

Si  $\chi^2 \geq \chi^{2*}$  Acepto  $H_0$

$\chi^2 < \chi^{2*}$  Rechazo  $H_0$       Donde  $\chi^2 = \frac{\hat{\sigma}^2(n-1)}{\sigma_0^2}$

#### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $\sigma^2$  no es igual a  $\sigma_0^2$ , entonces existe otra distribución, con una menor dispersión, generando las probabilidades  $\beta$  y  $1-\beta$ , luego:



$$\alpha = \Pr\{\chi^2 < \chi^{2*} / H_0\}$$

$$1 - \alpha = \Pr\{\chi^2 \geq \chi^{2*} / H_0\}$$

$$\beta = \Pr\{\chi^2 \geq \chi^{2*} / H_1\}$$

$$1 - \beta = \Pr\{\chi^2 < \chi^{2*} / H_1\}$$

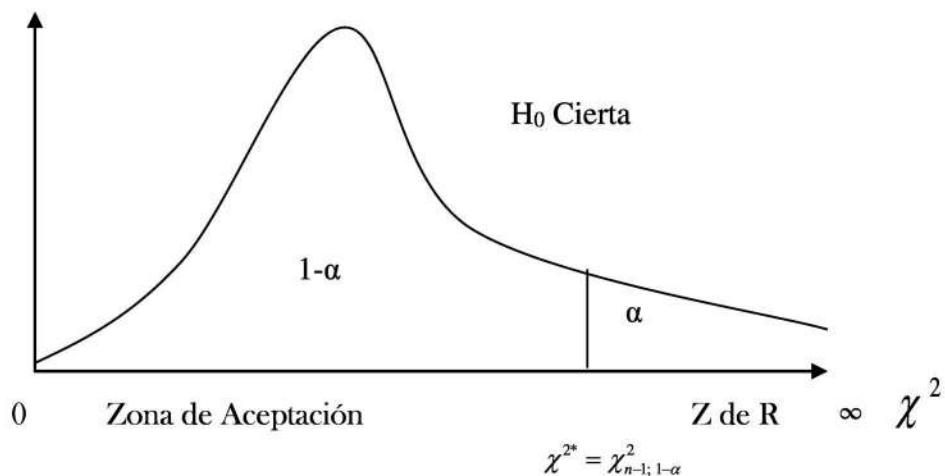
### Decima de Hipótesis Compuestas

#### Lateral Derecha

##### Hipótesis

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 > \sigma_0^2$$



Punto Crítico:  $\chi^{2*} = \chi_{n-1; 1-\alpha}^2$

##### Regla de decisión

Si  $\chi^2 \leq \chi^{2*}$  Acepto  $H_0$

$\chi^2 > \chi^{2*}$  Rechazo  $H_0$

Donde  $\chi^2 = \frac{\hat{\sigma}^2(n-1)}{\sigma_0^2}$

### Probabilidades involucradas



Si  $H_0$  no es cierta, es decir  $\sigma^2 \neq \sigma_0^2$ , entonces existen otras distribuciones, con una mayor dispersión, generando las probabilidades  $\beta$  y  $1-\beta$ , luego:

$$\alpha = \Pr\{\chi^2 > \chi^{2*} / H_0\}$$

$$1 - \alpha = \Pr\{\chi^2 \leq \chi^{2*} / H_0\}$$

$$\beta = \Pr\{\chi^2 \leq \chi^{2*} / H_1\}$$

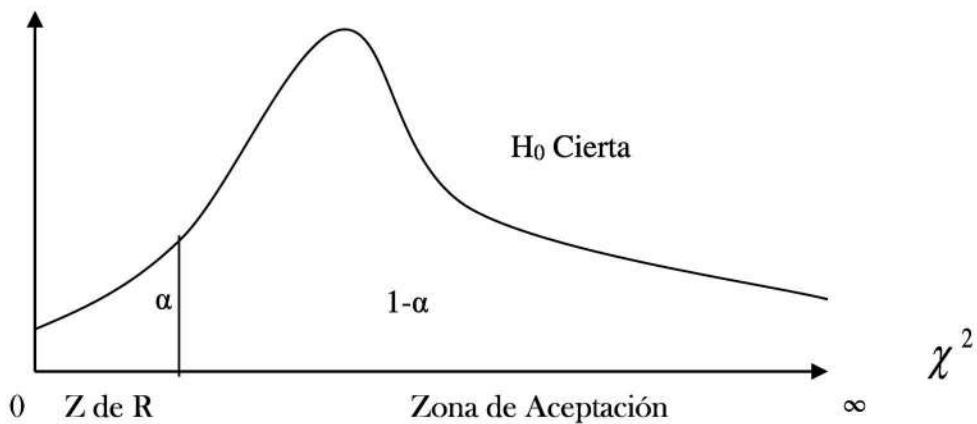
$$1 - \beta = \Pr\{\chi^2 > \chi^{2*} / H_1\}$$

### Lateral Izquierda

#### Hipótesis

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$



Punto Crítico:  $\chi^{2*} = \chi^2_{n-1; \alpha}$

#### Regla de decisión

Si  $\chi^2 \geq \chi^{2*}$  Acepto  $H_0$

$$\chi^2 < \chi^{2*} \text{ Rechazo } H_0 \quad \text{Donde } \chi^2 = \frac{\hat{\sigma}^2(n-1)}{\sigma_0^2}$$

#### Probabilidades involucradas



Si  $H_0$  no es cierta, es decir  $\sigma^2 \text{ no es igual a } \sigma_0^2$ , entonces existen otras distribuciones, con menores o mayores dispersiones, generando las probabilidades  $\beta$  y  $1-\beta$ , luego:

$$\alpha = \Pr\{\chi^2 < \chi^{2*} / H_0\}$$

$$1 - \alpha = \Pr\{\chi^2 \geq \chi^{2*} / H_0\}$$

$$\beta = \Pr\{\chi^2 \geq \chi^{2*} / H_1\}$$

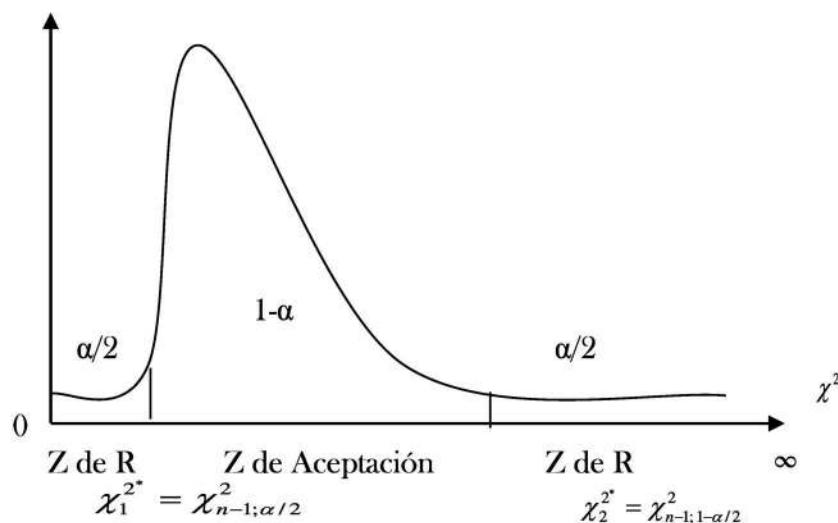
$$1 - \beta = \Pr\{\chi^2 < \chi^{2*} / H_1\}$$

### Bilateral

#### Hipótesis

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2$$



Punto Crítico:  $\chi_1^{2*} = \chi_{n-1; \frac{\alpha}{2}}^2$  y  $\chi_2^{2*} = \chi_{n-1; 1-\frac{\alpha}{2}}^2$

#### Regla de decisión

Si  $\chi^2 \leq \chi_1^{2*} \leq \chi^2 \leq \chi_2^{2*}$  Acepto  $H_0$

$\chi^2 < \chi_1^{2*}$  ó  $\chi^2 > \chi_2^{2*}$  Rechazo  $H_0$  Donde  $\chi^2 = \frac{\hat{\sigma}^2(n-1)}{\sigma_0^2}$



### Probabilidades involucradas

Si  $H_0$  no es cierta, es decir  $\sigma^2 \neq \sigma_0^2$ , entonces existen otras distribuciones, con una menor ó mayor dispersión, generando las probabilidades  $\beta$  y  $1-\beta$ , luego:

$$\alpha = \Pr\{\chi^2 < \chi_1^{2*} / H_0\} + \Pr\{\chi^2 > \chi_2^{2*} / H_0\}$$

$$1 - \alpha = \Pr\{\chi_1^{2*} \leq \chi^2 \leq \chi_2^{2*} / H_0\}$$

$$\beta = \Pr\{\chi_1^{2*} \leq \chi^2 \leq \chi_2^{2*} / H_1\}$$

$$1 - \beta = \Pr\{\chi^2 < \chi_1^{2*} / H_1\} + \Pr\{\chi^2 > \chi_2^{2*} / H_1\}$$

### Ejemplos

1- Supongamos que un operador de bolsa, al aconsejar a un cliente con respecto a la inversión en una acción particular, destaca la poca variabilidad de su cotización. De acuerdo a lo estipulado por el operador económico, esta acción presentaría una varianza de las cotizaciones diarias  $\sigma^2 = 0,2$ .

El cliente, quien debe realizar una fuerte inversión, decide poner a prueba la hipótesis del operador, estableciendo las siguientes hipótesis estadísticas:

$$H_0: \sigma^2 = 0,2$$

$$H_1: \sigma^2 > 0,2$$

Para probar estas hipótesis, se selecciona una muestra de 15 días donde se registra la cotización diaria, arrojando  $\hat{\sigma}^2 = 0,4$ , y se pretende trabajar con un nivel de significación del 0,05.

Entonces estamos frente a un contraste para la varianza, por lo cual plantearemos:

$$\theta = \sigma^2$$

$$\hat{\theta} = \hat{\sigma}^2 = \hat{s}^2$$

$$K(\hat{\theta}; \theta) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} = \frac{\hat{\sigma}^2(n-1)}{\sigma^2} \approx \chi_{(n-1)}^2$$

Las hipótesis a probar son, según el enunciado:

$$H_0: \sigma^2 = 0,2$$

$$H_1: \sigma^2 > 0,2$$



Este planteo corresponde a una docima lateral derecha, por lo que:

$$\chi^2* = \chi^2_{n-1; 1-\alpha} = \chi^2_{15-1; 1-0,05} = \chi^2_{14; 0,95} = 23,7$$

Estableciendo la siguiente regla de decisión:

$$\begin{aligned}\chi^2 \leq \chi^2* &\quad \text{Acepto } H_0 \\ \chi^2 > \chi^2* &\quad \text{Rechazo } H_0\end{aligned}$$

Donde  $\chi^2 = \frac{\hat{\sigma}^2(n-1)}{\sigma_0^2} = \frac{0,4(14)}{0,2} = 28$

Este resultado nos lleva al rechazo de la Hipótesis Nula, lo que indicaría que la evidencia proporcionada por la muestra indica que el error del operador, ya que la cotización diaria de la acción es más variable de lo que aseguraba.

2- El departamento de Control de Calidad de una empresa que fabrica computadoras electrónicas estima que si la longitud de una determinada pieza presenta una varianza mayor a  $4 \text{ mm}^2$ , irremediablemente se producirá la inutilización de una placa en el término de 6 meses de uso.

Como la empresa no tiene la intención de perder la porción de mercado lograda hasta el presente, sino que en sus planes está el de incrementarla, decide tomar muestras aleatorias durante el proceso de fabricación de la pieza en cuestión a los fines de controlar su variabilidad.

Una muestra de 15 piezas arrojó una longitud media de 5 mm con una desviación estándar de 1,2 mm.

¿Qué conclusiones puede obtener el Departamento de Control de Calidad de la empresa en cuanto a la calidad de la pieza analizada a un nivel de significación del 0,05?

Seguiremos la misma estructura de planteo:

$$\theta = \sigma^2$$

$$\hat{\theta} = \hat{\sigma}^2 = \hat{s}^2$$

$$K(\hat{\theta}; \theta) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} = \frac{\hat{\sigma}^2(n-1)}{\sigma^2} \approx \chi^2_{(n-1)}$$

Datos:

$$n = 15$$

$$\bar{x} = 5 \text{ mm}$$

$$s = 1,2 \text{ mm} \rightarrow \hat{s} = s \sqrt{\frac{n}{n-1}} = 1,2 \sqrt{\frac{15}{14}} = 1,24 \quad \text{y} \quad \hat{s}^2 = 1,53$$



Las hipótesis a probar son, según el enunciado:

$$H_0: \sigma^2 = 4$$

$$H_1: \sigma^2 < 4$$

Este planteo corresponde a una docima lateral izquierda, por lo que:

$$\chi^{2*} = \chi_{n-1; \alpha}^2 = \chi_{14-0,05}^2 = 6,57$$

Estableciendo la siguiente regla de decisión:

$$\begin{array}{ll} \chi^2 \geq \chi^{2*} & \text{Acepto } H_0 \\ \chi^2 < \chi^{2*} & \text{Rechazo } H_0 \end{array}$$

Donde  $\chi^2 = \frac{\hat{\sigma}^2(n-1)}{\sigma_0^2} = \frac{1,53(14)}{4} = 5,36$

Entonces, de acuerdo con este resultado, se rechaza la hipótesis nula, indicando que la variabilidad de las piezas, de acuerdo a evidencia proporcionada por la muestra, es menor a 4.

## **10. DOCIMA E INTERVALO DE CONFIANZA**

Por haber visto solamente intervalos bilaterales, diremos: Si el resultado de una docima bilateral lleva a aceptar  $H_0$ , entonces el correspondiente intervalo que se construya contendrá al verdadero valor poblacional, es decir será uno de los  $1 - \alpha$  que contienen al verdadero parámetro.

Si por el contrario, el resultado indica el rechazo de  $H_0$ , el intervalo correspondiente no contendrá al verdadero valor del parámetro, es decir será uno de los  $\alpha$  intervalos que no lo contienen.



## BIBLIOGRAFÍA

- **DORFLINGER, José H.** "Notas de Estadística y Probabilidad". Tomo I. Facultad de Ciencias Económicas. U.N.C 1978
- **DORFLINGER, José H.** "Notas de Estadística y Probabilidad". Tomo II. Facultad de Ciencias Económicas. U.N.C 1977
- **CARRIZO, José F. - CARRIZO, José F. (H).** "Nociones de Inferencia Estadística". Facultad de Ciencias Económicas. U.N.C. 1974.
- **CARRIZO, José F.** "Distribución en el muestreo". Facultad de Ciencias Económicas. U.N.C. 1974.
- **FERRERO, Fernando** "Muestreo". Facultad de Ciencias Económicas. U.N.C. 1973
- **SHAO, Stephen P.** "Estadística para Economistas y Administradores de empresas". Herrero Hnos. Sues. S.A. 1979.
- **CHOU, Ya-Lun.** "Análisis estadístico". Segunda Edición Interamericana.. 1977.
- **CHAO, Lincoln L.** "Estadística para las Ciencias Administrativas". Segunda Edición. Mc Graw Hill. 1975.
- **YAMANE, Taro** "Estadística". Tercera Edición. Harla, S.A. de C.V. 1977.
- **NETER-WASSERMAN-WHITMORE** "Fundamentos de Estadística para Negocios y Economía". Nueva Edición. C.E.C.S.A. 1978.
- **GILBERT, Norma.** "Estadística". Interamericana 1981
- **BERENSON, M. L. - LEVINE, D. M.** "Estadística para Administración y Economía" Conceptos y aplicaciones. Primera Edición. Interamericana. 1982.
- **HANKE, John E - REITSCH, Arthur G.** "Estadística para Negocios"
- **SÁNCHEZ CARRIÓN, Juan Javier.** "Manual de Análisis de datos". Alianza Editorial. S.A. Madrid. 1995.
- **NIDIA BLANCH- SILVIA JOECKES.** "Estadística Aplicada a la Investigación". Modulo IX: Inferencia Estadística. Estimación de Parámetros y Prueba de Hipótesis. Departamento de Educación a Distancia. U.N.C. Facultad de Ciencias Económicas. Marzo 1995.
- **NIDIA BLANCH- SILVIA JOECKES.** "Estadística Aplicada a la Investigación". Modulo X: Diseño de experimentos y selección de muestras aleatorias de aplicaciones finitas. Departamento de Educación a Distancia. U.N.C. Facultad de Ciencias Económicas. Marzo 1995.
- **PAULO AFONSO LOPES.** "Probabilidad y Estadística", Conceptos, Modelos, Aplicaciones en Excel. 1<sup>a</sup> Edición. Pearson Educación. Colombia. 2000.