ESTADÍSTICA INFERENCIAL

1. Introducción

El muestreo es uno de los métodos estadísticos que hoy en día tiene mayor utilidad e importancia. Su técnica consiste, a grandes rasgos, en obtener una información deseada sobre una población a partir de los datos extraídos de una parte de la misma.

Los problemas de muestreo se plantean en multitud de ocasiones en la vida cotidiana, tal es el caso, por ejemplo, de una entidad bancaria que piensa en la posibilidad de abrir una sucursal en un barrio de una ciudad. Para saber si esa sucursal será rentable se deben conocer los ingresos de la población del barrio. Lógicamente el método perfecto para conocer este dato sería visitar a cada uno de los vecinos y obtener, directamente de ellos, los datos necesarios. Es evidente que este método, además de inusual, resultaría costosísimo en tiempo y dinero. En este caso, el banco recurrirá a la utilización de un método de muestreo, que pasará por los siguientes puntos:

- Selección de una parte (***muestra***) de la población total.
- Estudio en la muestra de la variable buscada (en el caso de la entidad bancaria ***los ingresos***).
- Extensión de los resultados obtenidos para la muestra a la población total.

Estos tres pasos dan origen a un conjunto de problemas que deben resolverse para una correcta aplicación del método:

1. Cómo debe llevarse a cabo la selección de la muestra.
2. Cuál debe ser el tamaño adecuado de la muestra elegida.
3. Cuál es el grado de fiabilidad de la extensión de los resultados obtenidos desde la muestra a la población total

En la actualidad, la técnica de muestreo se aplica a los más variados problemas: detección de la intención de voto a un determinado partido político, control de calidad de un producto, obtención de las medidas antropomórficas de una población para confeccionar prendas de vestir, etc.

2. Estadística descriptiva e inferencial

Estadística es la ciencia que tiene por objeto el estudio de los colectivos. Basándose en el cálculo de probabilidades y en el análisis matemático, la estadística se centra en la toma, organización, recopilación, presentación y análisis de datos tanto para la obtención de conclusiones como para la toma de decisiones razonables de acuerdo con tales análisis.

En una colección de datos que atañen a las características de un colectivo de individuos u objetos (por ejemplo, las alturas y pesos de los estudiantes de un Instituto de Educación Secundaria, o las piezas defectuosas y no defectuosas producidas por una fábrica), es a menudo imposible o poco práctico observar la totalidad de los individuos u objetos, sobre todo si éstos son muchos. En lugar de observar el colectivo entero, llamado ***población*** *o* ***universo****,* se observa una pequeña parte del mismo, llamada ***muestra.***

Es conveniente considerar una muestra:

- Cuando el número de elementos de la población es muy grande, puesto que el estudio completo de la misma exigiría demasiado tiempo y muchos recursos.

- Cuando sea muy costosa la observación estadística, aunque el número de elementos de la población no sea muy grande.

Es necesario considerar una muestra:

- Cuando la observación estadística exige destruir los elementos observados (por ejemplo, algunos ensayos fisicoquímicos de control de calidad).

- Cuando la población es infinita.

La parte de la estadística que trata solamente de describir y analizar un colectivo dado sin *sacar conclusiones e inferencias* de un colectivo mayor se llama ***estadística descriptiva.***

Si una muestra es representativa de una población, se pueden deducir importantes conclusiones acerca de ésta, a partir del análisis de la muestra. La ***estadística inductiva*** *o* ***estadística inferencial*** es la parta de la estadística cuya finalidad es obtener conocimientos de una población a partir de las observaciones relativas a una muestra de la misma. Al no poder estar totalmente seguros de la veracidad de tales *inferencias,* se utiliza la *teoría de probabilidades* para dar una medida de la certidumbre de las mismas.

➔ 2.1 Consideraciones sobre la toma de muestras

En la elección de una muestra deben tenerse en cuenta dos cuestiones fundamentales:

1. La representatividad de la muestra elegida.

2. Tamaño de la muestra.

➢ ***La representatividad*** se conseguirá siempre que los individuos que forman la muestra constituya una reproducción, a pequeña escala, de la totalidad de la población. Esto es, que la muestra sea una reproducción sin sesgo de la población.

Esta cuestión es de una importancia transcendental, ya que si la muestra no es representativa, a la hora de extender los resultados obtenidos a través de la muestra a la población se producirán unos desfases con la realidad, ocasionando grandes errores de predicción.

➢ La otra cuestión que debe tenerse en cuenta es el ***tamaño de la muestra***, que está relacionado íntimamente con el grado de fiabilidad que se desea obtener cuando se extiendan los resultados obtenidos a partir de ella a la población total.

Lógicamente, cuanto mayor sea el tamaño de la muestra mayor será la fiabilidad de los resultados obtenidos. Por ejemplo, si se quiere establecer la media de la estatura de una población de 1.000 individuos y se consideran dos muestras, una de 10 y otra de 900 personas, cabe esperar que el resultado obtenido a partir de la segunda muestra sea más fiable que el de la primera.

La determinación del tamaño de la muestra es un problema que se estudiará con detalle más adelante.

3.- Técnicas de muestreo

Se han desarrollado diversas técnicas o diseños para la elección de muestras y lograr una buena representatividad. Cada diseño es adecuado para determinados tipos de problemas.

Los diseños para muestras se clasifican en ***probabilísticos*** y ***no probabilísticos***.

- **Definición:**

*Una **muestra es probabilística** si cada unidad de la población tiene asignada **alguna probabilidad** de ser seleccionada en la muestra, probabilidad que debe ser conocida. Al contrario de lo que generalmente se cree, no es necesario que esta probabilidad sea la misma para todas las unidades de la población; en todo caso cada unidad debe tener una probabilidad de ser seleccionada y esta probabilidad debe ser conocida por el investigador.*

Dado que la probabilidad de seleccionar cada unidad de la población es conocida, un estadístico puede utilizar las diversas reglas y leyes sobre probabilidad para evaluar la confianza de las conclusiones que se obtengan a partir de muestra probabilísticas. En otras palabras, cuando una muestra es probabilística, el riesgo de decisiones y conclusiones incorrectas se puede medir utilizando la teoría de la probabilidad.

- **Definición:**

*Una **muestra es no probabilística** cuando algunas unidades no tienen posibilidad de ser seleccionadas o si la probabilidad de seleccionar una unidad cualquiera no es conocida o no puede determinarse.*

Una vez más, cuando se toma una decisión en base a la información de una muestra, existe siempre un riesgo de error. Así, si 1.000 artículos seleccionados de las producciones de dos máquinas A y B dan 20 % y 24 % de defectuosos, respectivamente, existe un riesgo de error al concluir que la máquina A es mejor que la B. Existe la posibilidad de que ambas máquinas tengan la misma eficiencia y que la diferencia observada se deba a la casualidad o a fluctuaciones muestrales. El riesgo de llegar a una conclusión errónea puede ser medido sólo si la muestra es una muestra probabilística; no puede ser determinado en el caso de una muestra no probabilística. Por esta razón, sólo las muestras probabilísticas pueden ser objeto de un análisis y un tratamiento estadísticos. En este tema, estudiaremos en detalle las muestras aleatorias simples, que constituyen el tipo más común de muestras probabilísticas. Además examinaremos brevemente otro tipo de muestras.

4.- Muestreo probabilístico

➔ 4.1 Muestreo aleatorio simple

- **Ejemplo:**

Para ilustrar el procedimiento utilizado en la selección de una muestra aleatoria simple, consideremos una población compuesta de cinco personas: Álvarez, Bravo, Cascos, Díaz y Espinete. Supongamos que deseamos formar una comisión (una muestra) de tres miembros de esta población, utilizando una selección aleatoria simple.

¿Cuántas muestras de ***tres*** elementos podemos seleccionar de esta población?

Representando cada nombre por su letra inicial, tenemos las $\binom{5}{3} = \frac{5!}{3!\cdot 2!} = \frac{5\cdot 4\cdot 3\cdot 2}{3\cdot 2\cdot 2} = 10$ muestras posibles siguientes:

ABC, ABD, ABE, ACD, ACE

ADE, CDE, BCD, BCE, BDE

Nuestro método de selección nos proporcionaría una muestra aleatoria simple si cada una de estas 10 muestras tiene la misma probabilidad de ser seleccionada. Una forma de hacerlo sería escribir cada una de estas 10 combinaciones en una hoja de papel, echar las 10 hojas en una caja, revolver bien y extraer una hoja. Si la combinación extraída es ***CDE***, nuestra muestra estará formada por Cascos, Díaz y Espinete. Si bien este método de obtener una muestra aleatoria simple es directo, los estadísticos prefieren un procedimiento alternativo que es igualmente bueno pero más fácil de aplicar.

Se ha demostrado que cada una de las 10 muestras tiene la misma probabilidad de ser seleccionada si se escribe cada uno de los cinco nombres en una hoja de papel, se echan las cinco hojas en una caja, se revuelve bien y se extraen sucesivamente tres hojas. Este método alternativo elimina la tarea de enumerar todas las muestras del tamaño deseado que pueden seleccionarse de una población, tarea que es enorme cuando el número de unidades en la población es grande.

Puede ser de interés demostrar que ambos métodos para seleccionar una muestra aleatoria simple dan a cada muestra posible igual oportunidad de ser seleccionada. Con el primer método, cada una de las 10 muestras posibles tiene la misma oportunidad de ser seleccionada. La probabilidad de seleccionar cualquiera de estas muestras, digamos ***CDE*** (Cascos, Díaz y Espinete) es $\frac{1}{10}$. Para demostrar nuestra proposición, por lo tanto, debemos probar que la probabilidad de seleccionar una comisión formada por Cascos, Díaz y Espinete con el segundo método es también $\frac{1}{10}$.

Utilizando el segundo método de selección de la muestra, hay seis maneras con las cuales se puede obtener una muestra formada por Cascos, Díaz y Espinete (no necesariamente en ese orden). La muestra resultante es Cascos, Díaz y Espinete si se elige a Cascos en la primera extracción, a Díaz en la segunda y a Espinete en la tercera. La muestra resultante es también Cascos, Díaz y Espinete si se elige a Cascos en la primera extracción, a Espinete en la segunda y a Díaz en la tercera. A continuación se enumeran las seis maneras posibles de seleccionar una comisión formada por Cascos, Díaz y Espinete:

1. Cascos, después Díaz, después Espinete (CDE);
2. Cascos, después Espinete, después Díaz (CED);
3. Díaz, después Cascos, después Espinete (DCE);
4. Díaz, después Espinete, después Cascos (DEC);
5. Espinete, después Cascos, después Díaz (ECD);

6. Espinete, después Díaz, después Cascos (EDC).

Cada uno de los seis sucesos anteriores representa la selección de una comisión formada por Cascos, Díaz y Espinete. Como los seis sucesos son mutuamente incompatibles, la probabilidad de seleccionar una comisión formada por Cascos, Díaz y Espinete es igual a la suma de las probabilidades de estos seis sucesos. Estas probabilidades son:

$$P(C \cap D \cap E) = \tfrac{1}{5} \cdot \tfrac{1}{4} \cdot \tfrac{1}{3} = \tfrac{1}{60}\,, \quad P(C \cap E \cap D) = \tfrac{1}{5} \cdot \tfrac{1}{4} \cdot \tfrac{1}{3} = \tfrac{1}{60}$$

$$P(D \cap C \cap E) = \tfrac{1}{5} \cdot \tfrac{1}{4} \cdot \tfrac{1}{3} = \tfrac{1}{60}\,, \quad P(D \cap E \cap C) = \tfrac{1}{5} \cdot \tfrac{1}{4} \cdot \tfrac{1}{3} = \tfrac{1}{60}$$

$$P(E \cap C \cap D) = \tfrac{1}{5} \cdot \tfrac{1}{4} \cdot \tfrac{1}{3} = \tfrac{1}{60}\,, \quad P(E \cap D \cap C) = \tfrac{1}{5} \cdot \tfrac{1}{4} \cdot \tfrac{1}{3} = \tfrac{1}{60}$$

De dónde deducimos que la probabilidad de seleccionar la muestra Cascos, Díaz y Espinete es: $\tfrac{1}{60} + \tfrac{1}{60} + \tfrac{1}{60} + \tfrac{1}{60} + \tfrac{1}{60} + \tfrac{1}{60} = \tfrac{6}{60} = \tfrac{1}{10}$.

- **Definición:**

  El muestreo aleatorio simple consiste en seleccionar ***n*** *elementos sin reemplazamiento de entre los* ***N*** *que componen la población, de tal modo que todas las muestras de tamaño* ***n*** *que se pueden formar* $C_{N,n} = \binom{N}{n}$ *tengan la misma probabilidad de salir elegidas.*

  La probabilidad de elegir una muestra es : $P = \dfrac{1}{\binom{N}{n}}$.

  La probabilidad de que un elemento determinado de la población forme parte de la muestra viene dada por: $P = \dfrac{\binom{N-1}{n-1}}{\binom{N}{n}} = \tfrac{n}{N}$

➔ 4.2 Uso de números aleatorios

La utilización de hojas de papel y de una caja para seleccionar una muestra aleatoria simple es un procedimiento bastante elemental, que se hace aún más engorroso e impráctico a medida que el número de unidades elementales de la población se hace mayor. En lugar de este procedimiento, resulta más práctico utilizar una tabla de números aleatorios.

- ***¿Qué son los números aleatorios?*** Son las cifras 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 dispuestas de una manera particular. La ***tabla 1*** es un ejemplo de una tabla de números aleatorios. Las diferentes cifras están distribuidas de una forma que todas aparecen aproximadamente con la misma frecuencia en la tabla; además, una cifra determinada cualquiera no tiene

ninguna relación con las que la rodean. En otras palabras, las diversas cifras están distribuidas al azar; sin embargo, por razones de conveniencia, en cada fila se han colocado 40 cifras en grupos de cinco.

- **Ejemplo:**

Para ilustrar el uso de una tabla de números aleatorios, utilizaremos el problema práctico de seleccionar una muestra aleatoria simple, consistente en 10 cuentas pendientes elegidas de una población de 6532 cuentas.

Como primer paso, asignamos un número de serie a cada cuenta, en nuestro caso, un número de cuatro cifras. (si el número de cuentas en la población hubiera sido 623, habríamos asignado a cada una un número de tres cifras; así la cantidad de cifras de cada número de serie depende del total de cuentas en la población.) Asignamos el número 0001 a la primera cuenta, 0002 a la segunda, 0123 a la que ocupa el lugar 123º y 6532 a la última. Una vez que todas las cuentas han sido numeradas, podemos comenzar a seleccionar nuestra muestra leyendo las primeras cuatro cifras de la primera línea de la tabla de números aleatorios; el número es 1507, al cual le corresponde la cuenta que ocupa el lugar 1507 en la lista de cuentas de la población, cuenta que pasa a ser entonces la primera cuenta de la muestra. Continuando de esta manera obtenemos las otras nueve cuentas:

1920, 5385, 1259, 5183, 6093, 6262, 2737, 4042, 2110

(*Nótese que cuando un número de la serie no corresponde a una cuenta, pasamos a la siguiente línea y que si un número de la serie aparece dos veces, debe ignorarse la segunda vez.*)

La tabla de números aleatorios podría haberse utilizado de alguna otra manera; por ejemplo, podríamos haber leído las primeras cuatro cifras a partir de cualquier lugar de la tabla; además, para seleccionar las segundas cuatro cifras podríamos habernos desplazado en cualquier sentido (hacia los lados, en diagonal) con tal que ese desplazamiento se mantuviera sistemáticamente.

➔ 4.3 Muestreo aleatorio sistemático

Se empieza numerando todos los elementos de la población desde 1 a N. Para seleccionar los n elementos que constituyen la muestra, es preciso obtener el **coeficiente de elevación** $h = \frac{N}{n}$. Después se elige al azar un número i, llamado **origen**, comprendido entre 1 y h ($1 \leq i \leq h$), que nos indica el punto de arranque de la selección.

La muestra está formada por los elementos:

$$i,\ i+h,\ i+2h,\ ....,\ i+(n-1)h$$

Este procedimiento exige, que para que se pueda aplicar correctamente, la población no presente ninguna ordenación por la variable objeto de estudio y, si la hay, previamente habrá que desordenarla.

- **Ejemplo:**

  Siguiendo con el ejemplo de las 6532 cuentas, para elegir las 10 cuentas de la muestra a través del muestreo aleatorio sistemático obtenemos el coeficiente de elevación *h*:

$$h = \frac{N}{n} = \frac{6532}{10} = 653'2 \approx 653$$

  Elegimos en las tablas de números aleatorios un número al azar comprendido entre 1 y 653. Para ello seleccionamos una columna de números, por ejemplo la primera, y en ella, por ejemplo, los tres últimos dígitos, hasta encontrar un número *i* entre 1 y 653, que resulta ser el 70.

  La muestra de las 10 cuentas la componen: 70, 70+653, 70+1306, 70+1959, 70+2612, 70+3265, 70+3918, 70+4571, 70+5224, 70+5877

  es decir, las cuentas de número:

  70, 723, 1376, 2029, 2682, 3335, 3988, 4641, 5294, 5947.

- Antes de concluir este apartado, debemos señalar que para seleccionar una muestra aleatoria simple es necesario disponer de una lista de todas las unidades elementales de la población. Esta lista se denomina ***marco poblacional***. En forma ideal, un marco poblacional debe contener cada unidad elemental de la población y excluir los duplicados.

- En muchas investigaciones de mercado es difícil disponer de un buen marco poblacional. Los investigadores utilizan "marcos de trabajo" tales como guías telefónicas, listas de contribuyentes o de electores, u otro tipo de listas preparadas con fines específicos. Si bien algunos de estos marcos son adecuados en muchos casos, todos ellos tienen diversas imperfecciones que introducen algunos riesgos y limitan su utilización.

- Hasta ahora hemos considerado solamente la selección de una muestra aleatoria simple de una población finita para la cual era posible preparar una lista de todas las unidades elementales. Ahora, ***¿cómo podemos seleccionar una muestra aleatoria simple de una población infinita para la cual no podemos obtener una tal lista?*** Por ejemplo, ¿cómo podemos seleccionar una muestra aleatoria simple de 100 bombillas eléctricas de un proceso productivo en el que se fabrican miles de bombillas por día? La respuesta a esta pregunta es simple: un proceso productivo es un proceso aleatorio y 100 bombillas cualesquiera constituyen una muestra aleatoria simple de la producción total. Felizmente, la mayoría de las poblaciones infinitas son generadas por procesos que son al menos aproximadamente aleatorios y cualquier número de unidades elementales constituyen una muestra aleatoria simple de la población formada por todas las unidades elementales.

➔ 4.4 Muestreo estratificado

En una muestra estratificada, la población se divide en un determinado número de grupos o estratos. El objeto de esta estratificación es obtener un grupo relativamente más homogéneo respecto de la característica en estudio. De cada grupo o estrato, se selecciona una muestra aleatoria simple y estas submuestras se combinan en una gran muestra.

Así, por ejemplo, si se desea estimar el gasto anual en arriendo por familia en una ciudad o región geográfica puede ser conveniente dividir las familias de esa ciudad en diversos estratos, de acuerdo a cuales sean sus ingresos. Como ilustración podrían dividirse en tres grupos: aquellos con ingresos anuales inferiores a 1.000.000 PTA.; aquellos cuyos ingresos anuales están entre 1.000.000 y 5.000.000 PTA. y, por último, quienes tienen ingresos superiores a 5.000.000 PTA. Una vez que la población se ha dividido en estos tres estratos, se selecciona una muestra aleatoria simple de cada estrato y estas tres submuestras se combinan en una gran muestra.

A condición de que nuestra estratificación sea efectivamente adecuada para dividir la población en grupos más homogéneos con respecto a los gastos anuales en arriendo, obtendremos una estimación más precisa del gasto anual medio en arriendo utilizando una muestra estratificada que con una muestra aleatoria simple seleccionada de la población total, en el supuesto que ambas muestras sean del mismo tamaño. La estratificación tiene una ventaja adicional: proporciona información tanto para cada estrato como para toda la población. Debe mencionarse, sin embargo, que el costo de obtención de una muestra estratificada es mayor que el de la obtención de una muestra aleatoria simple del mismo tamaño ya que la estratificación involucra el trabajo adicional de clasificar la población en los diversos estratos.

5.- Muestreo no probabilístico

Las muestras no probabilísticas se caracterizan por el hecho de que no es posible determinar la probabilidad de inclusión de cada unidad elemental de la población en la muestra. Por esta razón no hay forma de medir el riesgo de llegar a conclusiones erróneas a partir de muestras no probabilísticas. Dado que la confiabilidad de los resultados de estas muestras no puede medirse, las muestras no probabilísticas no se prestan para tratamiento y análisis estadísticos. Los tipos más comunes de esta clase de muestra son las muestras de conveniencia y las muestras basadas en la opinión.

➔ 5.1 Muestreo de conveniencia

Las unidades elementales que se incluyen en una muestra de conveniencia se eligen por su facilidad de acceso y su conveniencia. Un fabricante de un cierto artículo alimenticio que desea tener información acerca de la preferencia que el consumidor dispensa a su producto puede, por ejemplo, incluir un cupón en algunos paquetes de su producto. El cliente debe llenar el cupón contestando las preguntas ahí formuladas. Los cupones recibidos por el fabricante constituyen la muestra.

➔ 5.2 Muestreo basado en la opinión

Las unidades elementales que se incluyen en una muestra basada en la opinión son seleccionadas por un experto en base a que son representativas de la mayoría de las unidades elementales de la población. Estas muestras se utilizan en los casos en que el costo o el tiempo disponible hacen necesario que la muestra sea de tamaño muy pequeño.

6.- Parámetros poblacionales y estadísticos muestrales

- **Definición:**

Un estadístico es un número que describe un determinado aspecto de una muestra. *La media aritmética, la mediana, la moda, la fracción o proporción y la desviación típica son todos estadísticos si se calculan para una muestra (parte de la población).*

Debe hacerse una clara distinción entre un parámetro y un estadístico. En tanto que ***un parámetro describe un determinado aspecto de la población***, un estadístico describe un determinado aspecto de una muestra. La media aritmética es un parámetro si se calcula para toda la población en tanto que si se calcula para una muestra, es un estadístico.

Dado que un parámetro se calcula para toda la población, su valor es constante; en cambio, el valor de un estadístico varía de una muestra a otra; luego, ***un estadístico es una variable.*** Para ilustrar esto, consideremos el siguiente ejemplo:

:

Supongamos que nuestra población consiste en las "edades" de las siguientes cinco casas:

$$x_1 = 2 \quad años; \quad x_2 = 2 \quad años; \quad x_3 = 4 \quad años; \quad x_4 = 5 \quad años; \quad x_5 = 2 \quad años$$

La media aritmética de la población es constante y vale 3:

$$\mu = \frac{2+2+4+5+2}{5} = 3 \quad años.$$

Si seleccionamos una muestra aleatoria simple de dos casas y estas resultan ser la primera y la segunda, la media aritmética de esta muestra es 2 años; en cambio, si la muestra consiste en la tercera y cuarta casas, la media aritmética es ahora 4'5 años. Se ve así que la media aritmética puede cambiar de una muestra a otra.

Los estadísticos más comunes son la media aritmética, la proporción y la desviación típica. La media aritmética de una *muestra aleatoria simple*, que se designa por $\bar{x}$, se calcula de la misma manera que la media aritmética de la población. En símbolos,

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

donde x_i indica el i-ésimo elemento de la muestra y n el tamaño de la muestra.

La proporción de una *muestra aleatoria simple*, que se designa por p, también se calcula de la misma manera que la proporción de la población.

El principal objetivo de muchos problemas de Estadística es determinar el valor de algún parámetro de una población, por ejemplo, la media aritmética μ. Para lograr este objetivo, los estadísticos rara vez calculan el valor del parámetro utilizando los datos de la población completa, sino que tratan de estimar su valor a partir de una muestra cuidadosamente seleccionada.

Así, los estadísticos utilizan la media muestral $\bar{x}$ como un estimador de la media poblacional μ; también utilizan la proporción muestral μ_p como estimador de la proporción poblacional p y la desviación típica muestral $\sigma_{\bar{x}}$ como estimador de la desviación típica poblacional σ.

Ahora, ¿cuán confiable es $\bar{x}$ como estimador de μ? El valor de μ es una constante en tanto que el valor de $\bar{x}$ varía de una muestra a otra. La confiabilidad de $\bar{x}$ como estimador de μ no puede determinarse a menos que podamos describir en forma precisa el comportamiento de $\bar{x}$ o bien el modo de su variabilidad de una muestra a otra. Si μ es conocido, ¿cuáles son los posibles valores de $\bar{x}$ en las diferentes muestras? Si esta pregunta se contesta satisfactoriamente, entonces se podrá establecer fácilmente la confiabilidad de $\bar{x}$ como estimador de μ.

- **Definición:**

  Se llama ***error muestral de un estadístico*** al valor absoluto de la diferencia entre dicho estadístico y su parámetro correspondiente.

  Los errores muéstrales se clasifican en:

  - Errores de sesgo: Son debidos a que la muestra no representa a la población sino que ha sido elegida con cierta parcialidad o sesgo. Esta clase de errores provienen del método aplicado para elegir la muestra y, como consecuencia de ello, el valor del estadístico o estimador obtenido no tiende a aproximarse al valor del parámetro respectivo de la población.

  - Errores aleatorios: Se deben al azar y no a la parcialidad o sesgo. Su cuantía puede ser estimada aplicando la teoría de probabilidades.

Decimos que un estadístico es preciso cuando el error muestral de dicho estadístico es pequeño; es decir, cuando su valor se acerca mucho al de su parámetro correspondiente.

7. Distribución de un estadístico

➔ 7.1 Distribución muestral de medias

La relación entre la media poblacional y las medias de las diversas muestras que pueden seleccionarse de una población se puede ilustrar muy bien mediante una operación de muestreo a partir de una población conocida.

☐ Ejemplo 1:

Consideremos que la población consiste en las alturas de tres plantas diferentes cultivadas en un laboratorio de botánica y son las siguientes:

$$x_1 = 1cm \quad x_2 = 3cm \quad x_3 = 5cm$$

La media y la desviación típica de la población se calculan a continuación:

$$\mu = \tfrac{1+3+5}{3} = 3cm$$

$$\sigma = \sqrt{\tfrac{(1-3)^2+(3-3)^2+(5-3)^2}{3}} = \sqrt{\tfrac{8}{3}} = \tfrac{2\sqrt{2}}{\sqrt{3}}cm$$

- En la población anterior vamos a formar todas las muestras posibles ***con reemplazamiento*** de tamaño 2, y a determinar la media aritmética $\bar{x}_i$ en cada una de ellas.

Hay $CR_{3,2} = C_{4,2} = \begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$ muestras posibles con reemplazamiento de tamaño dos que pueden seleccionarse de nuestra población de tres plantas; en consecuencia, hay 6 medias muestrales posibles. ¿Cuál es la representatividad de estas medias respecto de la media poblacional?

Tabla 1

Pares	***Muestras***	$\bar{x}_i$	$\bar{x}_i^2$
(1,1)	1,1	1	1
(1,3)	1,3	2	4
(1,5)	1,5	3	9
(3,1)	1,3	2	4
(3,3)	3,3	3	9
(3,5)	3,5	4	16
(5,1)	1,5	3	9
(5,3)	3,5	4	16
(5,5)	5,5	5	25
Totales		27	93

Tabla 2

$\bar{X} = \bar{x}_i$	$p_i = P(\bar{X} = \bar{x}_i)$	$\bar{x}_i \cdot p_i$	$\bar{x}_i^2 \cdot p_i$
1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
2	$\frac{2}{9}$	$\frac{4}{9}$	$\frac{8}{9}$
3	$\frac{3}{9}$	$\frac{9}{9}$	$\frac{27}{9}$
4	$\frac{2}{9}$	$\frac{8}{9}$	$\frac{32}{9}$
5	$\frac{1}{9}$	$\frac{5}{9}$	$\frac{25}{9}$
	Totales	$\frac{27}{9}$	$\frac{93}{9}$

Para examinar estas 6 muestras con mayor detalle, observemos la *tabla 1* y calculemos primero su media aritmética o esperanza matemática, designada por $\mu_{\bar{x}}$:

$$\mu_{\bar{x}} = \tfrac{1+2+3+2+3+4+3+4+5}{9} = \tfrac{27}{9} = 3cm = \mu$$

Podemos entonces decir que el promedio de todas las medias muestrales posibles es igual a la media poblacional, o $\mu_{\bar{x}} = \mu$. Sin embargo, una media muestral particular puede ser mayor o menor que la media poblacional pero, en todo caso, su promedio es igual a la media poblacional.

Continuando nuestro examen de las propiedades de las medias muestrales, calculemos la desviación típica de las 6 medias muestrales posibles. Se designará por $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \sqrt{\frac{93}{9} - 3^2} = \sqrt{\frac{93-81}{9}} = \sqrt{\frac{12}{9}} = \sqrt{\frac{4}{3}} = \frac{2}{\sqrt{3}} cm$$

La distribución anterior de las medias muestrales equivale a definir una variable aleatoria $\bar{X}$ de ley de probabilidad representada en la *tabla 2*.

La esperanza matemática $E[\bar{X}]$ y varianza $V[\bar{X}]$ de esta variable aleatoria es:

$$E[\bar{X}] = \Sigma\, p_i \cdot \bar{x}_i = \frac{27}{9} = 3$$

$$V[\bar{X}] = \Sigma\, p_i \cdot \bar{x}_i^2 - (E[\bar{X}])^2 = \frac{93}{9} - 3^2 = \frac{93-81}{9} = \frac{12}{9} = \frac{4}{3}$$

Se cumple: $E[\bar{X}] = \mu_{\bar{x}} \quad y \quad V[\bar{X}] = \sigma_{\bar{x}}^2$

Observa que se han obtenido:

$\mu =$ media de la población = 3 cm. ; $\bar{x}_i =$ media de cada muestra de tamaño 2 (*con reemplazamiento*) ; $\mu_{\bar{x}} =$ media de todas las medias de las muestras de tamaño 2 = 3 cm. ; $E[\bar{X}] =$ esperanza matemática de la variable aleatoria $\bar{X}$ = 3

y que se cumple: $\mu = \mu_{\bar{x}} = E[\bar{X}] = 3$

De igual forma se han obtenido:

$\sigma =$ desviación típica de la población $= \frac{2\sqrt{2}}{\sqrt{3}} cm$; $\sigma_{\bar{x}} =$ desviación típica de la distribución formada por todas las medias de las muestras de tamaño 2 = $= \frac{2}{\sqrt{3}} = \frac{\frac{2\sqrt{2}}{\sqrt{3}}}{\sqrt{2}} = \frac{\sigma}{\sqrt{n}}$; $V[\bar{X}] =$ varianza de la variable aleatoria $\bar{X} = \frac{4}{3}$

y se cumple: $V[\bar{X}] = \sigma_{\bar{x}}^2 = \frac{4}{3} = \frac{\sigma^2}{2}$

☐ **Ejemplo 2:**

Consideremos la misma población del ***ejemplo 1*** formada por las alturas de tres plantas diferentes: $x_1 = 1cm, \quad x_2 = 3cm \quad y \quad x_3 = 5cm$. La media y desviación típica poblacionales eran: $\mu = 3cm \quad y \quad \sigma = \frac{2\sqrt{2}}{\sqrt{3}} = \frac{2\sqrt{6}}{3} cm.$

- En la población anterior vamos a formar todas las muestras posibles ***sin reemplazamiento*** de tamaño 2, y a determinar la media aritmética $\bar{x}_i$ en cada una de ellas.

Tabla 3

Pares	***Muestras***	$\bar{x}_i$	$\bar{x}_i^2$
(1,3)	1,3	2	4
(1,5)	1,5	3	9
(3,1)	1,3	2	4
(3,5)	3,5	4	16
(5,1)	1,5	3	9
(5,3)	3,5	4	16
Totales		18	58

Tabla 4

$\bar{X} = \bar{x}_i$	$p_i = P(\bar{X} = \bar{x}_i)$	$\bar{x}_i \cdot p_i$	$\bar{x}_i^2 \cdot p_i$
2	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{4}{3}$
3	$\frac{1}{3}$	$\frac{3}{3}$	$\frac{9}{3}$
4	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{16}{3}$
	Totales	$\frac{9}{3}$	$\frac{29}{3}$

Hay $C_{3,2} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} = 3$ muestras posibles, sin reposición, de tamaño dos que pueden seleccionarse de nuestra población de 3 plantas; en consecuencia hay tres medias muestrales posibles. Para examinar estas tres medias muestrales con mayor detalle, calculemos primero su media aritmética, observando los datos de la *tabla 3:*

$$\mu_{\bar{x}} = \frac{18}{6} = 3cm$$

Podemos entonces decir que el promedio de todas las medias muestrales posibles es igual a la media poblacional, $\mu_{\bar{x}} = \mu$. Continuando con nuestro examen de las propiedades de las medias muestrales, calculemos la desviación típica:

$$\sigma_{\bar{x}} = \sqrt{\frac{58}{6} - 3^2} = \sqrt{\frac{29}{3} - 9} = \sqrt{\frac{2}{3}} = \frac{\sqrt{2}}{\sqrt{3}} = \frac{\sqrt{6}}{3} cm$$

La desviación típica $\sigma_{\bar{x}}$(o ***error estándar***) de la distribución de medias muestrales indica la diferencia *"promedio"* entre los diversos valores de $\bar{x}_i$ y $\mu_{\bar{x}}$. En promedio, cada media muestral difiere de la media poblacional en $\sqrt{\frac{2}{3}}$ cm. Un valor pequeño de $\sigma_{\bar{x}}$indica dos hechos:

primero, los diversos valores de $\bar{x}_i$ son cercanos entre si;

segundo, la diferencia promedio entre estos valores y $\mu_{\bar{x}} = \mu$ es pequeña.

En consecuencia cualquier valor $\bar{x}_i$ es una buena estimación para μ.

Si bien el error estándar $(\sigma_{\bar{x}})$ de la media mide la diferencia promedio entre las medias muestrales y la media poblacional, no es necesario considerar todas la

muestras para determinar este error. Felizmente, es posible calcular $\sigma_{\bar{x}}$si se conoce la desviación típica de la población σ. Se cumple:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

donde N es el tamaño de la población y n el tamaño de la muestra.

Refiriéndonos a nuestra población de tres plantas con $\mu = 3cm, \quad \sigma = \frac{2\sqrt{6}}{3}$, N = 3 y n = 2, *entonces:* $\sigma_{\bar{x}} = \frac{\frac{2\sqrt{6}}{3}}{\sqrt{2}} \cdot \sqrt{\frac{3-2}{3-1}} = \frac{2\sqrt{3}}{3} \cdot \frac{1}{\sqrt{2}} = \frac{2\sqrt{3}}{3\sqrt{2}} = \frac{2\sqrt{6}}{6} = \frac{\sqrt{6}}{3}$

Este valor de$\sigma_{\bar{x}}$es idéntico al obtenido anteriormente.

La distribución anterior de las medias muestrales equivale a definir una variable aleatoria $\bar{X}$ de ley de probabilidad representada en la *tabla 4*.

La esperanza matemática $E[\bar{X}]$ y varianza $V[\bar{X}]$de esta variable aleatoria es:

$$E[\bar{X}] = \sum_{i=1}^{3} \bar{x}_i \cdot p_i = \frac{9}{3} = 3$$

$$V[\bar{X}] = \sum_{i=1}^{3} \bar{x}_i^2 \cdot p_i - (E[\bar{X}])^2 = \frac{29}{3} - 9 = \frac{2}{3}$$

- Podemos decir entonces que si se seleccionan todas las muestras posibles, sin reposición, de tamaño *n* de una población dada de tamaño *N*, se tiene:

$$\mu = \mu_{\bar{x}} = E[\bar{X}] \quad y \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = + \sqrt{V[\bar{X}]}$$

- El factor $\sqrt{\frac{N-n}{N-1}}$ se denomina ***corrección por población finita*** y es aproximadamente 1 si se muestra una población ***infinita.***

- Si la población es ***finita*** pero el tamaño de la muestra es pequeño en relación al de la población (*tamaño muestra no superior al 10 % de la población*), el factor de corrección por población finita también será aproximadamente 1. Por ejemplo, si en un problema N = 1001, n = 49, entonces

$$\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1001-49}{1001-1}} = \sqrt{\frac{952}{1000}} = \sqrt{0'952} = 0'9757$$

- Es de gran interés observar que el error estándar $(\sigma_{\bar{x}})$ se hace más pequeño cuanto mayor sea el tamaño de la muestra. A medida que el tamaño de la muestra aumenta, las diversas medias muestrales se hacen más uniformes en su valor y, en consecuencia, cualquier media muestral es una buena estimación de la media poblacíonal. En otras palabras, una muestra grande es más confiable que una pequeña.

- Concluiremos este examen de las propiedades de la media muestral, con una última e importante propiedad: ***las medias muestrales están distribuidas en forma aproximadamente normal cuando el tamaño de la muestra es 30 o más.*** Debe destacarse que esto es válido aun cuando las muestras se obtengan a partir de una población origen que no esté distribuida normalmente. Por otra parte, cuando la población origen tiene distribución normal, las medias muestrales tienen distribución normal *por pequeño que sea* el tamaño de la muestra.

- Se toma, normalmente, como criterio que en todas aquellas muestras que cumplan que $5 \cdot n > N$ será necesario distinguir los dos tipos de muestreo, mientras que en caso contrario, que será el más habitual, se usará únicamente el muestreo con reposición.

Resumen

Consideremos todas las posibles muestras aleatorias (con o sin reposición), de tamaño n, de una población de tamaño $N > n$, en la que se estudia la variable estadística X.

Calculemos las medias, $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ de las diferentes muestras, y construyamos, a partir de ellas, la ***distribución muestral de medias.***

Se verifican las siguientes propiedades:

Población / **Extracciones**	**Infinita**	**Finita**
Con reposición	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$n < \frac{10}{100}N, \quad \sqrt{\frac{N-n}{N-1}} \approx 1, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Sin reposición	$\sqrt{\frac{N-n}{N-1}} \approx 1, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

- Si $n \geq 30$, entonces la distribución muestral de medias es normal $\bar{X} = N(\mu, \sigma_{\bar{x}})$, tanto si la población de origen es normal como si no lo es.

- Si la población de origen es normal, entonces la distribución muestral de medias es normal.

- En muestras cuyo tamaño verifica $5 \cdot n > N$ es necesario distinguir los dos tipos de muestreo.

☐ Ejemplo 1:

El cociente de inteligencia medio de la población estudiantil de Zaragoza es $\mu = 95$ *y la desviación típica es* $\sigma = 14$. *Se extrae una muestra aleatoria (sin reposición) de 49 estudiantes de esa población. ¿Qué probabilidad hay de que resulte una media igual o inferior a 92 ?*

❑ Solución:

La distribución muestral de medias se aproxima a una normal (por ser n = 49 > 30) con:

$$\mu_{\bar{x}} = \mu = 95 \quad y \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{14}{\sqrt{49}} = 2$$

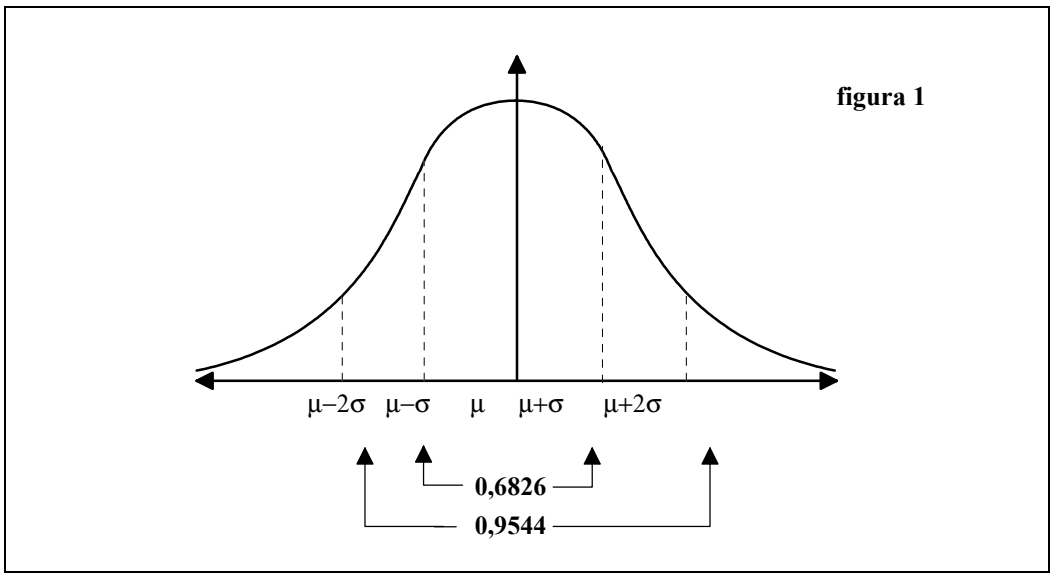
ya que al ser N mucho mayor que n, se verifica: $\frac{\sqrt{N-n}}{\sqrt{N-1}} \approx 1$.

Tipificando el valor $\bar{x}_1 = 92$ de la distribución muestras de medias , podremos acudir a las tablas de la distribución normal N(O, 1) y averiguar la probabilidad correspondiente al valor tipificado z_1:

$$P(\bar{X} \leq 92) \overset{Tipificando}{=} P\left(\frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} \leq \frac{92-95}{2}\right) = P(Z \leq -1'50) = 0'0668$$

☐ Ejemplo 2:

Dos mil muestras aleatorias diferentes de 100 alumnos son seleccionadas en una gran universidad en la que la edad promedio de los alumnos es $\mu = 20$*años y la desviación típica es* $\sigma = 2$ *años. Habrá 2000 medias muestrales. La media de la primera muestra puede ser 22 años; la de la segunda, 19; la de la tercera, 20, etc.; Sin embargo, las*

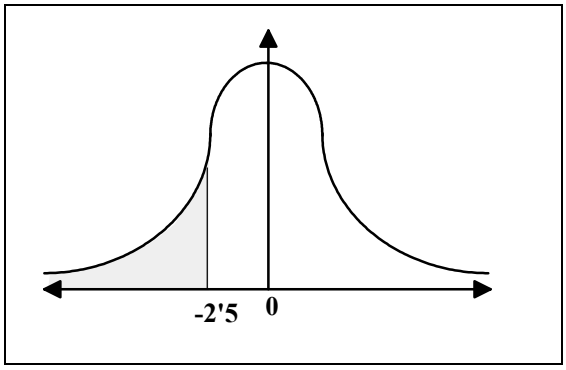


2000 medias muestrales tendrán una distribución aproximadamente normal con una media de 20 años y una desviación típica $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}} = 0'2$ *años.*

La media de cada una de las 2000 muestras puede ser o no 20 años; sin embargo, al ser $\bar{X} = N(20, 0'2)$*, el 68 % de las medias muestrales estará en el intervalo* $(19'8, 20'2) = (\mu - \sigma_{\bar{x}}, \mu + \sigma_{\bar{x}})$*. El 95 % de las medias muestrales estará en el intervalo* $(19'6, 20'4) = (\mu - 2\sigma_{\bar{x}}, \mu + 2\sigma_{\bar{x}})$

- **Ejemplo 3:**

El valor promedio de las cuentas pendientes que figuran en los archivos de una firma comercial es de 125 dólares con una desviación típica de 24 dólares. ¿Cuál es la probabilidad de que una muestra aleatoria simple de 36 cuentas seleccionadas del archivo tenga un promedio igual o menor a 115 dólares?



- ***Solución:***

Si se seleccionan todas las muestras posibles de tamaño 36 del archivo de la firma comercial, entonces los valores de $\bar{X}$ están distribuidos normalmente con una media $\mu_{\bar{x}} = \mu = 125$ dólares y una desviación típica $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{24}{\sqrt{36}} = 4$ dólares. La probabilidad de que la media de la muestra de tamaño 36 sea igual o menor que 115 dólares se puede calcular como sigue:

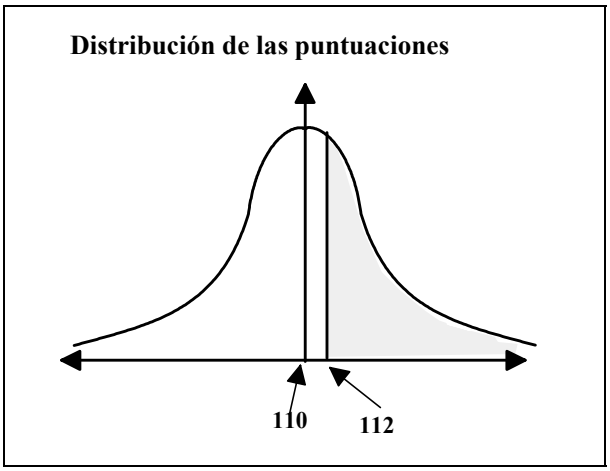
$$P(\bar{X} \leq 115) \overset{Tipificando}{=} P(Z \leq \tfrac{115-125}{4}) = P(Z \leq -2'5) = 0'0062$$

- **Ejemplo 4:**

La media de las puntuaciones del C.I. (coeficiente de inteligencia) de los alumnos de una universidad es 110 y la desviación típica 10. ***(a)*** *Si las puntuaciones del C.I. están distribuidas normalmente, ¿cuál es la probabilidad de que la puntuación de un alumno cualquiera sea mayor que 112?* ***(b)*** *¿Cuál es la probabilidad de que la puntuación media de una muestra de 36 alumnos sea mayor que 112?* ***(c)*** *¿Cuál es la probabilidad de que la puntuación media de una muestra de 100 alumnos sea mayor que 112?*

- ***Solución:***

(a) La probabilidad de que la puntuación del C.I. de un alumno cualquiera sea mayor que 112 se calcula como sigue:



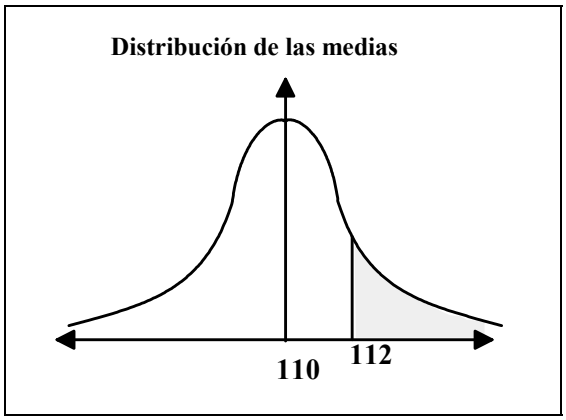
puntuación del C.I.= $= X = N(110, 10)$

$$P(X > 112) = P(Z > \tfrac{112-110}{10})$$

$$= P(Z > 0'2) = 0'4207$$

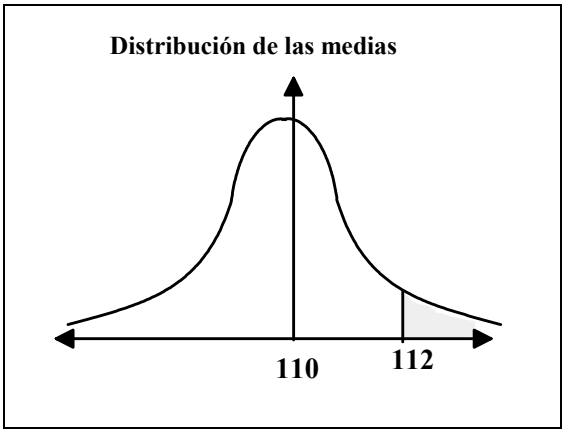
(b) Las medias de las diversas muestras de 36 alumnos tienen una distribución normal $\bar{X} = N(\mu_{\bar{x}} = 110, \sigma_{\bar{x}} = \frac{10}{\sqrt{36}}) = N(110, 1'67)$. Por lo tanto:

$$P(\bar{X} > 112) = P\left(Z > \tfrac{112-110}{1'67}\right) = P(Z > 1'2) = 0'1151$$



Puede ser de utilidad comparar las dos distribuciones normales utilizadas en las partes (a) y (b) del ejemplo. La distribución normal en (a) describe una población real, lo cual consiste en las puntuaciones del C.I. de todos los alumnos. La media poblacional es 110 y la desviación típica 10. Por otra parte, la distribución normal en (b) describe una ***población teórica,*** la cual consiste en las medias de todas las muestras posibles de 36 alumnos que pueden seleccionarse en dicha universidad. La media es también 110, pero su desviación típica, llamada error estándar, sólo 1'67.

(c) Las medias de las diversas muestras de 100 alumnos tienen una distribución normal, $N\left(\mu_{\bar{x}} = 112, \sigma_{\bar{x}} = \frac{10}{\sqrt{100}} = 1\right)$. La probabilidad de que la puntuación media del C.I. en una muestra aleatoria de 100 alumnos sea mayor que 112 puede determinarse del modo que sigue:



$$P(\bar{X} > 112) = P(Z > \tfrac{112-110}{1})=$$

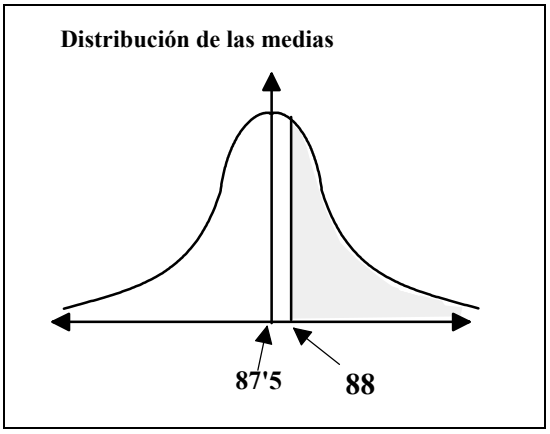
$$= P(Z > 2) = 0'0228$$

Observemos finalmente que la distribución normal de la figura de la izquierda también describe una población teórica. Esta

población, sin embargo, está formada por la medias de todas las muestras aleatorias posibles de 100 alumnos que pueden seleccionarse de esa universidad.

☐ Ejemplo 5:

Se sabe que la media de ingresos, en el barrio donde la entidad bancaria está estudiando abrir una sucursal, es de 87.500 PTA. con una desviación típica de 8.500 PTA. Se desea saber cuál es la probabilidad de elegir una muestra de 60 personas donde la media de ingresos sea mayor que 88.000 PTA.



❑ *Solución:*

Las medias muestrales se distribuirán según una normal $N(87'5, 1'097)$ ya que:

$\mu_{\bar{x}} = \mu = 87'5$ miles de PTA.,
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8'5}{\sqrt{60}} = 1'097$ miles de PTA. y además $\sqrt{\frac{N-n}{N-1}} \approx 1$. La gráfica de la izquierda nos define la distribución de las medias y la probabilidad buscada es el área de la parte rayada. Se calcula así:

$$P(\bar{X} > 88) = P\left(Z > \frac{0'5}{1'097}\right) =$$

$$= P(Z > 0'46) = 0'3228$$

➔ 4.2 Distribución muestral de proporciones

En el presente apartado examinaremos problemas que tratan acerca de la *proporción poblacional*. Describiremos la relación entre la proporción muestras, μ_p, y la proporción poblacional, p**.** La relación entre la proporción poblacional y las proporciones de las diversas muestras que pueden seleccionarse de esa población puede ilustrarse mediante la descripción de la operación de muestreo a partir de una población conocida.

☐ Ejemplo 1:

Supongamos que esta población consta de cinco bolitas: una de ellas es verde y las cuatro restantes son blancas. Designando la bolita verde por V y las blancas por B, la población puede representarse por:

Población de cinco bolitas

Bolita	Color
1	V
2	B
3	B
4	B
5	B

La proporción de bolitas verdes, designada por p , en esta población, es 0'20. Supongamos ahora que seleccionamos *todas* las muestras posibles de cuatro bolitas de esta población y calculamos la proporción de bolitas verdes, p_i, para cada muestra,

Muestras posibles n = 4	**Proporción muestral p_i**
1,2,3,4 (V,B,B,B)	0'25
1,2,3,5 (V,B,B,B)	0'25
1,2,4,5 (V,B,B,B)	0'25
1,3,4,5 (V,B,B,B)	0'25
2,3,4,5 (B,B,B,B)	0'00

Así, mientras la proporción poblacional p de bolitas verdes es 0'20, la proporción muestral p_i es 0,25 para las cuatro primeras muestras y es cero para la última. Para examinar con más detalle estas cinco proporciones muestrales posibles, calculemos su media aritmética o promedio. El promedio de todas las proporciones muéstrales, designado por $\mu_p = E(p)$, es

$$\mu_p = \frac{p_1+p_2+p_3+p_4+p_5}{5} = \frac{0'25+0'25+0'25+0'25+0}{5} = \frac{1}{5} = 0'20 = p$$

Podemos por lo tanto decir que si bien las proporciones muéstrales individuales pueden sobrestimar o subestimar la proporción poblacional, su promedio es igual a esta proporción poblacional.

Calculemos ahora la desviación típica de estas cinco proporciones muestrales posibles. Su desviación típica, llamada ***error estándar de la proporción*** y designada por $\sigma_p$se calcula a continuación.

p_i	$(p_i - p)$	$(p_i - p)^2$
0'25	0'05	0'0025
0'25	0'05	0'0025
0'25	0'05	0'0025
0'25	0'05	0'0025
0'00	0'20	0'04

$$\sigma_p = \sqrt{\frac{\sum(p_i-p)^2}{n^o \ \ muestras}} = \sqrt{\frac{0'05}{5}} = \sqrt{0'01} = 0'1$$

El error estándar de la proporción indica la disparidad «promedio» entre las diversas p_i y p. Las primeras cuatro proporciones muestrales difieren de la proporción poblacional en 0'05 y la quinta en 0'20. Sin embargo, en promedio, cada proporción muestras difiere de la proporción poblacional en 0'10.

Un valor pequeño de σ_p, indica dos cosas: ***(1)*** Los diversos valores de $p_i$son cercanos entre si; ***(2)*** la diferencia promedio entre estos valores de p_iy $p$es pequeña; en consecuencia, cualquier valor de $p_i$es una buena estimación de p.

Si bien el error estándar de la proporción mide la diferencia «promedio» entre todas las proporciones muéstrales posibles y la proporción poblacional, no es necesario en la práctica seleccionar todas las muestras posibles para determinar su valor. Existe un modo alternativo de calcular este error estándar de la proporción. Se ha encontrado que

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}},$$

donde $p$es la proporción poblacional, N es el tamaño de la población y n el de la muestra.

Para nuestra población de cinco bolitas, donde $p = 0'20$, el error estándar de las proporciones de todas las muestras posibles de tamaño 4 es

$$\sigma_p = \sqrt{\frac{0'20 \cdot 0'80}{4}} \sqrt{\frac{5-4}{5-1}} = \sqrt{\frac{0'16}{4}} \cdot \sqrt{\frac{1}{4}} = \sqrt{0'01} = 0'1$$

Este valor es idéntico al obtenido anteriormente.

Supongamos una población de tamaño N, en la que se estudia el atributo o la variable X. Sea $x_i$un valor de la variable X y n_i su frecuencia absoluta (número de elementos que presentan dicho valor). Definimos la ***proporción***, p, de los elementos de la población que presentan el valor $x_i$como el cociente: $p = \frac{n_i}{N}$. En consecuencia, la proporción de elementos de la población que no presentan el valor x_i será $1 - p$.

Por ejemplo, en las elecciones municipales de un municipio de 50.000 habitantes censados, el partido A obtuvo 15.200 votos. La proporción, p, de los que votaron al partido A fue, por tanto, $p = \frac{15200}{50000} = 0'304$, y la de los que no votaron fue: $1 - 0'304 = 0'696$.

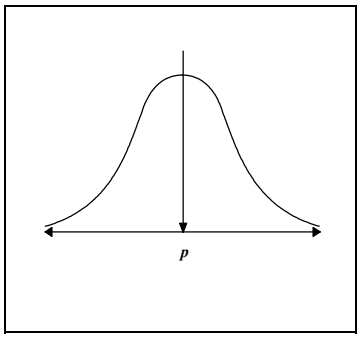
Así, podemos decir que si se seleccionan todas las muestras posibles de tamaño n de una población dada, y calculamos en cada una de ellas, la proporción, p_i, de elementos que presentan el valor x_i, obtenemos la ***distribución muestral de proporciones***, cuya media designaremos por μ_p y su desviación típica por σ_p.

Se verifican las siguientes propiedades:

(a) $\mu_p = p$ (en todos los casos)

(b) $\sigma_p = \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ (si la población es ***finita*** y las muestras son ***sin reposición***)

(c) El factor de ***corrección por población finita***, $\sqrt{\frac{N-n}{N-1}}$, se aproximará a 1 cuando la población sea de tamaño ***infinito,*** cuando el tamaño de la muestra sea inferior al 10% del de la población o si las muestras son ***con reposición.*** Bajo cualquiera de estas condiciones, el error estándar de la proporción estará dado por $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$.

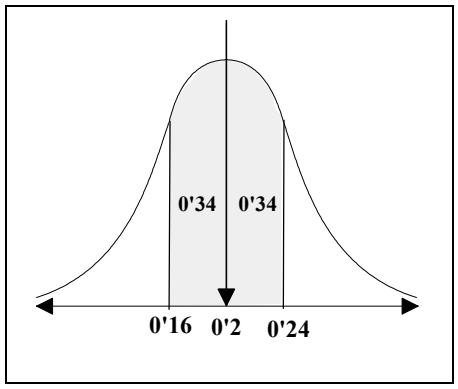


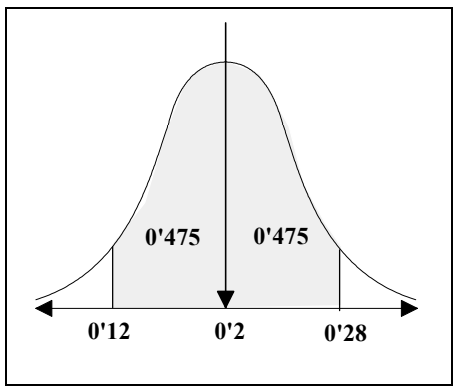
- Concluyendo nuestro examen de las propiedades de las diversas proporciones muestrales, agregaremos una última e importante propiedad: *Las proporciones muéstrales tienen una distribución aproximadamente normal cuando el tamaño de la muestra es suficientemente grande* $(n \geq 30)$.
- En resumen, podemos enunciar este **teorema general:** *Si se seleccionan todas las muestras posibles de tamaño n de una población con proporción* p*, entonces las proporciones muéstrales resultantes tienen distribución aproximadamente normal, tienen promedio igual a* $\mu_p = p$ *y desviación estándar o error estándar igual a* σ_p.

Para aclarar las aplicaciones del teorema anterior, supongamos que se seleccionan todas las muestras posibles de 100 alumnos de una universidad en la cual la proporción de alumnos del último curso es del 20 %. De acuerdo con nuestro teorema, las diversas proporciones muéstrales estarán distribuidas en forma aproximadamente normal, con una media de $0'2$ y un error estándar o desviación típica de $0'04$.

Como la distribución de las proporciones muestrales es aproximadamente normal, el 68% de dichas proporciones estará en el intervalo $(p-\sigma_p, p+\sigma_p) = (0'16, 0'24)$,o sea, entre el 16 % y 24 %.

Del mismo modo, el 95% de estas proporciones muestrales estará en el intervalo $(p-2\sigma_p, p+2\sigma_p) = (0'12, 0'28)$, o, aproximadamente entre 12% y 28%.





❑ Ejemplo 1:

En el "referéndum" sobre la permanencia de España en la OTAN, celebrado el 12 de marzo de 1986, participaron 17217290 votantes de un censo cifrado en 28828434. La proporción, p, de los que votaron fue:

$$p = \frac{17217290}{28828434} = 0'5972$$

y, por tanto, la de los que no votaron 1 - p = 1 - 0'5972 = 0'4028. En una muestra aleatoria (sin reposición) de 100 personas censadas, ¿cuál es la probabilidad de que la proporción de votantes sea mayor que 0'7 ?

❑ Solución:

La distribución muestras de proporciones se aproxima a una normal con:

$$\mu_p = p = 0'5972 \quad y \quad \sigma_p = \sqrt{\frac{p\cdot(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Por ser N mucho mayor que n, se verifica: $\frac{N-n}{N-1} \approx 1$

Por lo que: $\sigma_p = \sqrt{\frac{0'5972\cdot 0'4028}{100}} = 0'0490$

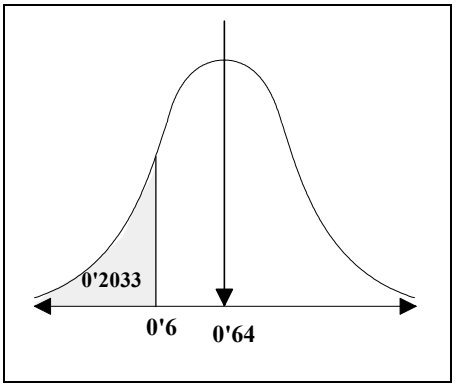
Tipificando el valor $p_1 = 0'7$ de la distribución muestra de proporciones , podremos acudir a las tablas de la distribución normal N(O,1) y averiguar la probabilidad correspondiente al valor tipificado z_1:

Por tanto:

$$P(P > 0'7) = P\left(\frac{P-\mu_p}{\sigma_p} > \frac{0'7-0'5972}{0'0490}\right) = P(Z > 2'10) = 0'0179$$

❑ Ejemplo 2:

Es sabido que el 64 % de los votantes de un cierto distrito electoral apoyan al partido A. ¿Cuál es la probabilidad de que una muestra aleatoria simple de 100 votantes de ese distrito de una proporción de simpatizantes de A del 60 % o menos?

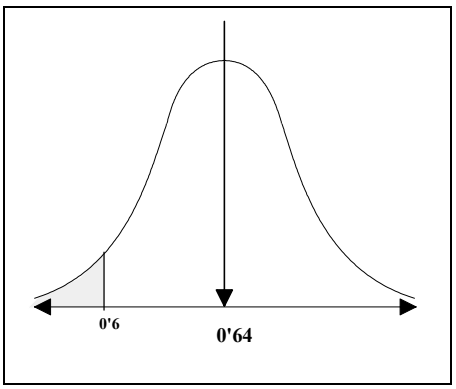


❑ *Solución:*

Por lo que ya hemos visto, sabemos que si se seleccionan todas las muestras posibles de 100 votantes de ese distrito, las proporciones

muestrales estarán distribuidas en forma aproximadamente normal y tendrán promedio de $\mu_p = p = 0'64$ y error estándar de $\sigma_p = \sqrt{\frac{0'64 \cdot 0'36}{100}} = 0'048$

La probabilidad de que la proporción de simpatizantes de A en una muestra de 100 votantes del distrito sea del 60 % o menos se calcula así: $P(p \leq 0'6) = P\left(Z \leq \frac{0'6 - 0'64}{0'048}\right) = P(Z \leq -0'83) = 0'2033$



Supongamos ahora que se toma una muestra de 400 votantes del distrito. ¿Cuál es la probabilidad de que la proporción muestral sea del 60 % o menos?

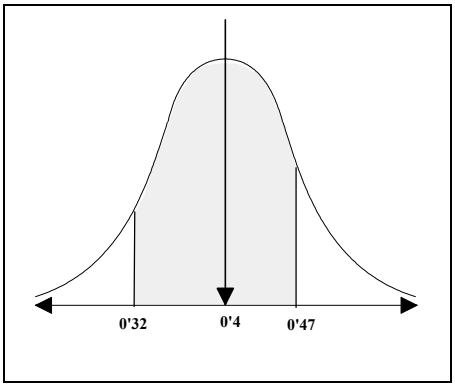
Si se toman muchas muestras de votantes de tamaño 400, entonces las diversas proporciones muestrales están distribuidas de forma aproximadamente normal con una media de $\mu_p = 0'64$ y un error estándar de $\sigma_p = \sqrt{\frac{0'64 \cdot 0'36}{400}} = 0'024$, La probabilidad de que la proporción muestral de simpatizantes del partido A en una muestra de 400 votantes sea de 60 % o menos se calcula de la siguiente forma:

$$P(P \leq 0'6) = P\left(Z \leq \frac{0'6 - 0'64}{0'024}\right) = P(Z \leq -1'67) = 0'0475.$$

Así, en tanto que la probabilidad de que una muestra de 100 votantes de una proporción de simpatizantes de A del 60 % o menos es de 0'2033, esta probabilidad se reduce a 0'0475 cuando el tamaño de la muestra se aumenta.

☐ Ejemplo 3:

El 40 % de los estudiantes de postgrado de una universidad son casados. Si se seleccionan al azar 100 alumnos de postgrado, ¿cuál es la probabilidad de que la proporción de alumnos casados de esta muestra esté entre el 32 % y el 47 % ?



☐ *Solución:*

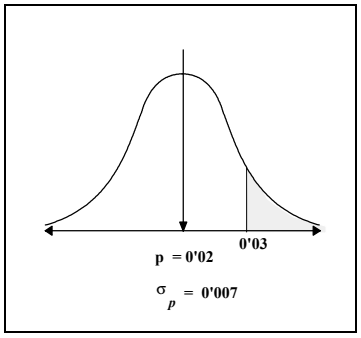
Si se toma una gran cantidad de muestras de 100 alumnos de postgrado, la distribución de las diversas proporciones muestrales es normal: $N(\mu_p = 0'4, \sigma_p = 0'049)$. La probabilidad de que la proporción de alumnos casados en una muestra de tamaño 100 esté entre el 32 % y el 42 % se determina a continuación:

$$P(0'32 \le P \le 0'47) = P(-1'63 \le Z \le 1'43) = 08720$$

- **Ejemplo 4:**

El departamento de adquisiciones de una gran ensambladora de televisores utiliza la regla siguiente para decidir si acepta o rechaza una partida de 10.000 partes enviadas semanalmente por un proveedor: Seleccionar una muestra de 400 partes de cada partida recibida. Si el 3 % o más de las partes seleccionadas son defectuosas, rechazar toda la partida: si la proporción de piezas defectuosas es inferior al 3 %, aceptar la partida. ¿Cuál es la probabilidad de rechazar una partida que contiene realmente 2 % de partes defectuosas?

- ***Solución:***



Si bien la partida contiene un 2 % de partes defectuosas, o sea que $p = 0'02$, es posible que una muestra de 400 partes tenga una proporción del 3 % o más partes defectuosas y, en consecuencia, la partida sea rechazada. Para determinar la probabilidad de rechazar esta partida, veamos en primer lugar la distribución de las proporciones de las diversas muestras de tamaño 400 que pueden seleccionarse de esta partida. Su media es 0'02 y la desviación típica es 0'007. La probabilidad de obtener una proporción muestral del 3 % o más y en consecuencia rechazar la partida se calcula a continuación:

$$P(p \ge 0'03) = P(Z \ge 1'43) = 0'0764$$

8. Estimación de parámetros

La técnica de muestreo se utiliza, normalmente, para determinar un dato estadístico (***parámetro***) de una variable en una población. Por ejemplo, se puede desear hallar la media de la altura de los hombres de una ciudad. La variable que se ha de estudiar es la altura de los ciudadanos y el parámetro deseado es la media de todas las alturas.

El proceso que se debe seguir consiste en elegir una muestra aleatoria y representativa de los individuos de la ciudad y a partir de ella obtener un valor (***estadístico a***) que permita concluir que la media de la variable altura, dentro de la población total, es ***a*****,** con un cierto margen de error que será necesario cuantificar.

Parece claro que el valor más probable de la media de alturas de la población coincide con el valor de la media de la muestra, pero esta suposición es bastante arriesgada, ya que no existe ningún inconveniente inicial para suponer que el valor más aproximado a la media de la población sea la media de la variable en la muestra, la moda, o cualquier otro valor.

Por ejemplo, si la media de la población total es 172 cm. puede darse el caso de seleccionar una muestra cuya media sea 165 cm. y la mediana 169 cm., con lo que en esta muestra sería mejor estimación la mediana que la media.

Sin embargo cuando se habla de estimar un parámetro no se considera una sola muestra, sino el conjunto de todas las posibles muestras. Es un hecho estadístico probado por métodos que superan el alcance de este curso (***método de máxima verosimilitud y método de mínimos cuadrados)*** que las medias de todas las posibles muestras se concentran más próximamente a la media de la población que el conjunto de las medianas de dichas muestras. Por esta razón se determina que la media de la muestra es mejor estimación de la media de la población que la mediana.

Hay dos formas de expresar la estimación de parámetros:

- **Estimación puntual**.
- **Estimación por intervalos.**

En la primera, a partir de la muestra, se decide que el valor del parámetro que se desea estimar es un número real con un cierto margen de error, mientras que en el segundo se establece un intervalo al que el parámetro en cuestión debe pertenecer con un determinado nivel de confianza.

Así, en el cálculo de la media de las alturas de la población podría establecerse a partir de la muestra, resultados como los siguientes:

- *La media de la población es de 173 cm. con un error de +4, con un riesgo del 95 %* (***estimación puntual***) o,
- *la media de la población está comprendida entre 170 y 175 cm. con un nivel de confianza de 99 %* (***estimación por intervalos***).

➔ 8.1 Estimación de la media de una variable de población

El problema que se resuelve en este apartado es la estimación de una variable en una población a partir de la media de una muestra.

Estimación puntual

Utilizando el método de estimación de parámetros de máxima verosimilitud se llega al resultado siguiente:

Si μ y σ son, respectivamente, la media y la desviación típica de la variable en la población y $\bar{x}_i$ y σ_i la media y la desviación típica de la variable en la muestra de tamaño n, se cumple que las estimaciones más probables de μ y σ son, respectivamente, $\bar{x}_i$ y $\sigma_i \cdot \sqrt{\frac{n}{n-1}}$ expresado habitualmente del modo siguiente:

$$\mathrm{E}(\mu) = \bar{x}_i \qquad \mathrm{E}(\sigma) = \sigma_i \cdot \sqrt{\frac{n}{n-1}}$$

Sin embargo decir que $\bar{x}_i$es el valor más probable de la media de la población es decir muy poco, ya que es necesario cuantificar el riesgo que se asume al considerar como μ

el valor de $\bar{x}_i$. Esta cuantificación se realiza a partir del hecho, ya conocido, de que la distribución muestral de las medias se ajusta, siempre que $n \geq 30$, a una distribución normal

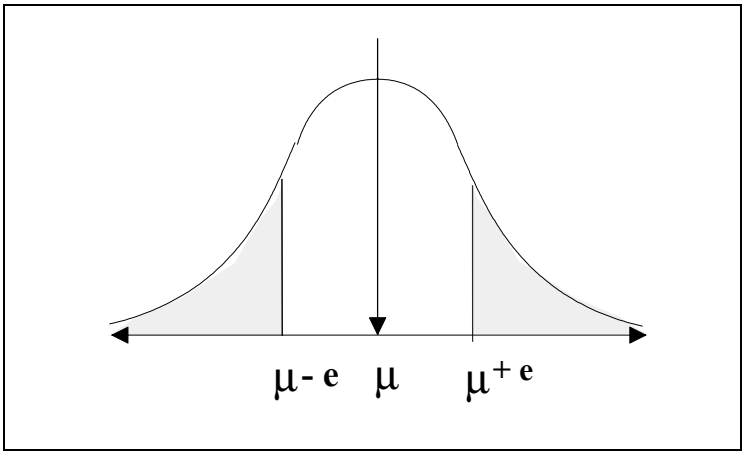
$$N(\mu, \tfrac{\sigma}{\sqrt{n}})$$

Así, cuando se toma como media de la población la media de la muestra $\bar{x}_i$, el riesgo de que el valor $\bar{x}_i$se aleje más de e unidades de μ viene dado por

$$R = P(\bar{X} > \mu + e) + P(\bar{X} < \mu - e) \overset{Tipificando}{=}$$

$$= P(Z > \tfrac{\mu+e-\mu}{\sigma_{\bar{x}}}) + P(Z < \tfrac{\mu-e-\mu}{\sigma_{\bar{x}}}) = P(Z > \tfrac{e}{\sigma_{\bar{x}}}) + P(Z < \tfrac{-e}{\sigma_{\bar{x}}}) =$$

$$= 2 \cdot P(Z > \tfrac{e}{\sigma_{\bar{x}}})$$



Por tanto, a la hora de realizar una estimación puntual aparecen tres valores que están relacionados entre sí: el riesgo (R), el error (e) y el tamaño de la muestra (n). Siempre que se conozcan dos de ellos será posible determinar el otro.

➔ Caso 1: Se conocen *e* y *n* y se quiere conocer *R*.

Bastará con utilizar la expresión:

$$R = 2 \cdot P(Z > \tfrac{e}{\sigma_{\bar{X}}})$$

para resolver el problema.

Hay que tener en cuenta que para calcular $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ es necesario conocer σ, dato que en muchas ocasiones será desconocido. Si se da esta situación bastará con utilizar como σ su estimación puntual:

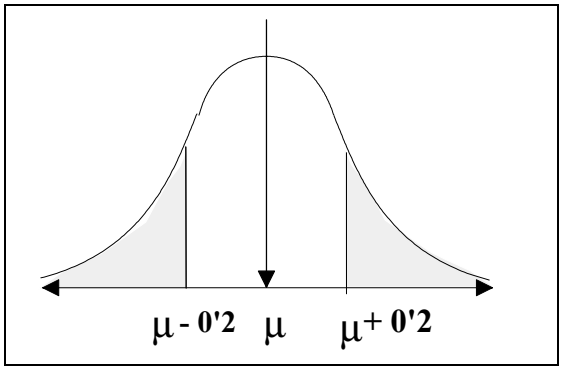
$$\sigma_i \cdot \sqrt{\tfrac{n}{n-1}}$$

☐ **Ejemplo 1:**

Una máquina fabrica clavos y se sabe por experimentación que la desviación típica de la variable (longitud de los clavos) es 1 mm. Se quiere calcular la probabilidad (riesgo) de que al elegir una muestra de 125 clavos su media se aleje de la media real más de $0'2$ mm.

La media de la muestra $\bar{x}_i$ tiene que cumplir:

$$\bar{x}_i > \mu + 0'2 \quad o \quad bien \quad \bar{x}_i < \mu - 0'2$$



por lo que la probabilidad buscada será:

$$R = 2 \cdot P\left(Z > \frac{0'2}{\sigma_{\bar{x}}}\right)$$

y como $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{125}} = 0'089$

$$R = 2 \cdot P(Z > 2'247) = 0'0246$$

➔ Caso 2: Se conocen *n* y *R* y se quiere calcular *e.*

También se solucionará este problema a partir de la fórmula:

$$R = 2 \cdot P(Z > \frac{e}{\sigma_{\bar{x}}})$$

aunque ahora se realizará una búsqueda inversa en las tablas de la normal.

☐ **Ejemplo 2:**

En el caso del ejemplo anterior se conoce ahora $n = 125$, $\sigma = 1$mm. *y* $R = 0'0802$ *y se quiere calcular* e.

De acuerdo con la fórmula anterior:

$$0'0802 = 2 \cdot P(Z > \frac{e}{\sigma_{\bar{x}}})$$

y por tanto: $P(Z > \frac{e}{\sigma_{\bar{x}}}) = 0'0401$

Igualdad que, gracias a la tablas, da lugar a:

$$\frac{e}{\sigma_{\bar{x}}} = 1'75 \Rightarrow e = 1'75 \cdot \frac{1}{\sqrt{125}} = 0'156$$

En este ejemplo se observa claramente que, a medida que el tamaño de la muestra aumenta, el error disminuye.

➔ Caso 3: Se conocen *R* y *e* y se quiere calcular *n*.

Quizás éste sea uno de los problemas más interesantes que pueden plantearse ya que, en la práctica, determinar el tamaño de la muestra que se selecciona es una de las primeras dificultades que hay que solventar.

También en esta situación, la fórmula básica es: $R = 2 \cdot P(Z > \frac{e}{\sigma_{\bar{x}}})$

☐ Ejemplo 3:

Ahora se busca el tamaño que debe tener la muestra para que la probabilidad de que la media de la muestra se aleje más de 0'18 sea igual a 0'1286, sabiendo que $\sigma = 1$.

Así, $R = 0'1286$, $e = 0'18$ y $\sigma = 1$. Por tanto:

$$0'1286 = 2 \cdot P\left(Z > \frac{0'18}{\sigma_{\bar{x}}}\right) \Rightarrow \quad 0'0643 = P\left(Z > \frac{0'18}{\sigma_{\bar{x}}}\right)$$

$$\frac{0'18}{\sigma_{\bar{x}}} = 1'52 \Rightarrow \quad \sigma_{\bar{x}} = 0'1184$$

Como $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ se tiene $0'1184 = \frac{1}{\sqrt{n}}$ de donde $n = 71'308$.

Por consiguiente, el tamaño de la muestra tendrá que ser de, al menos, 72 clavos.

En los tres casos vistos anteriormente se ha tenido en cuenta que la distribución muestral de las medias se aproximaba a la normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. En caso de muestras pequeñas, estos cálculos no son válidos a no ser que la distribución de variable dentro de la población total sea normal. En la estimación puntual no se estudiará el caso $n < 30$, pero si se hará en la estimación por intervalos que es la que habitualmente se realiza en los casos prácticos.

Estimación por intervalos

El problema consiste ahora en determinar dos valores ***a*** y ***b*** entre los que la media estará situada con un cierto ***nivel de confianza.*** Se habla de nivel de confianza y no de probabilidad de que se verifique

$$a < \mu < b$$

ya que el concepto de probabilidad en esta situación no tiene sentido pues, una vez fijado el intervalo (a, b), se cumple una de estas dos situaciones:

$\mu \in (a, b)$ lo que da lugar a $P(\mu \in (a, b)) = 1$

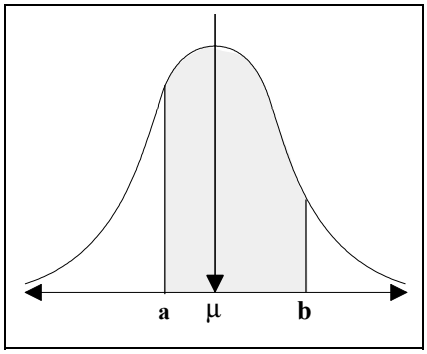
$\mu \notin (a,b)$ lo que da lugar a $P(\mu \in (a,b)) = 0$

Por esta razón, J. Neyman acuñó el nuevo término: nivel de confianza. ¿Qué significado tiene decir que **(*a, b*)** es el intervalo de estimación de la media μ al nivel de confianza $1-\alpha$?. La respuesta a esta pregunta es la siguiente:

Si se eligieran 100 muestras y a partir de cada una de ellas se obtuviera un intervalo, de la misma forma que se ha hecho con el **(*a, b*)**, cabría esperar que en $100 \cdot (1-\alpha)$ de estos intervalos la media estaría contenida en ellos.

Sabiendo que la distribución muestras se aproxima a una normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$la respuesta en sentido gráfico sería la que se observa en la figura de la página siguiente.

El área rayada es igual a $1-\alpha$. Al valor de α también se le denomina ***nivel de significación.***



Para estimar μ mediante un intervalo se distinguirán dos casos:

*** estimación de μ mediante muestras de tamaño** ≥ 30

*** estimación de μ mediante muestras de tamaño < 30**

Tanto en un caso como en otro, aparecen dos nuevos casos según sea conocida o no la desviación típica de la variable dentro de la población total (σ). Sin embargo, esto no representa ningún problema ya que en caso de desconocerla, se tomará su estimación puntual $\sigma_i \cdot \sqrt{\frac{n}{n-1}}$ y se trabajará del mismo modo que si σ fuese conocida.

➔ Muestras de tamaño mayor o igual a 30.

Si se tipifica la variable "media de la muestra", se obtiene que la nueva variable:

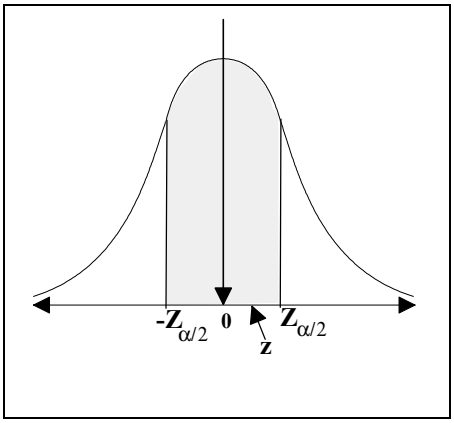
$$Z = \frac{\bar{x}_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

se distribuye según una normal $N(0, 1)$.

El nivel de confianza $1-\alpha$ con el que se desea realizar la estimación de la media da lugar a dos valores opuestos.

$$-Z_{\frac{\alpha}{2}} \qquad Z_{\frac{\alpha}{2}}$$

de tal modo que el área determinada por la función densidad de $N(0,\ 1)$, el eje de abscisas y las coordenadas en $-Z_{\frac{\alpha}{2}}$ y $Z_{\frac{\alpha}{2}}$ sea $1-\alpha$.



De acuerdo con el concepto de nivel de confianza, si $\bar{x}_i$ es la media de la muestra elegida y z es la variable tipificada, se cumple:

$$P\left(-Z_{\frac{\alpha}{2}} < z < Z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

y por tanto, para un nivel de confianza $1-\alpha$ se cumple:

$$-Z_{\frac{\alpha}{2}} < \frac{\bar{x}_i-\mu}{\sigma_{\bar{x}}} < Z_{\frac{\alpha}{2}} \quad \Rightarrow \quad -\sigma_{\bar{x}}.Z_{\frac{\alpha}{2}} < \bar{x}_i-\mu < \sigma_{\bar{x}}.Z_{\frac{\alpha}{2}} \quad \Rightarrow$$

$$\Rightarrow \quad \bar{x}_i-\sigma_{\bar{x}}.Z_{\frac{\alpha}{2}} < \mu < \bar{x}_i+\sigma_{\bar{x}}.Z_{\frac{\alpha}{2}}$$

Por lo que el intervalo buscado es:

$$\left(\bar{x}_i-\sigma_{\bar{x}}.Z_{\frac{\alpha}{2}}, \bar{x}_i+\sigma_{\bar{x}}.Z_{\frac{\alpha}{2}}\right)$$

El único problema que falta por resolver es el cálculo de $Z_{\frac{\alpha}{2}}$, pero esta cuestión se salva fácilmente sin más que imponer la condición

$$2\cdot P\left(Z > Z_{\frac{\alpha}{2}}\right) = \alpha$$

y buscar en las tablas de la normal tipificada los valores de $Z_{\frac{\alpha}{2}}$.

Precisamente el valor de $Z_{\frac{\alpha}{2}}$ recibe el nombre de **valor crítico** correspondiente al nivel de confianza $1-\alpha$. Así, si se quiere conocer el valor crítico asociado a un nivel de confianza del 95%, bastaría con realizar las siguientes operaciones:

$$2\cdot P\left(Z > Z_{\frac{\alpha}{2}}\right) = 0'05 \Rightarrow P\left(Z > Z_{\frac{\alpha}{2}}\right) = 0'025$$

y, consultando las tablas de la normal tipificada, se obtiene:

$$Z_{\frac{\alpha}{2}} = 1'96$$

Del mismo modo, podría obtenerse el valor crítico correspondiente a cualquier nivel de confianza. La siguiente tabla muestra los valores críticos más usuales:

nivel de confianza	90 %	95 %	96 %	98 %	99 %
valor crítico	1'645	1'96	2'055	2'325	2'575

- **Ejemplo 4:**

Una fábrica conservera desea conocer el tiempo que tarda en estropearse un producto que tiene almacenado. Elige una muestra de 200 unidades, resultando que el tiempo medio de descomposición de estos productos es de 172 horas. Por experiencias anteriores se conoce que la desviación típica de la variable ***"tiempo de descomposición"*** *es de 3'5 horas. Para el nivel de confianza del 95 %, ¿entre qué valores podrá estimarse el tiempo medio de descomposición para la totalidad del producto almacenados.*

- **Solución:**

A partir de lo visto anteriormente

$$1 - \alpha = 0'95 \Rightarrow Z_{\frac{\alpha}{2}} = 1'96$$

y, por tanto, el intervalo buscado es:

$$\mu \in (171'515, 172'485)$$

Por consiguiente, cabría esperar que si se erigiesen 100 muestras de tamaño 200 y se hallara para una de ellas el correspondiente intervalo, en 95 de estos intervalos estaría la media buscada.

Un problema que se suele plantear a menudo es el de calcular el tamaño de la muestra para que la anchura del intervalo al cual pertenece la media con un nivel de confianza $1 - \alpha$, sea una cantidad prefijada de antemano.

- **Ejemplo 5:**

Se desea saber que tamaño debe tener la muestra elegida al azar entre todos los productos de la fábrica del ejemplo anterior para que el intervalo al que pertenece la media del tiempo de descomposición de la totalidad de los productos de la fábrica, con un nivel de confianza del 95 %, sea igual a 2 horas.

- **Solución:**

Ahora es conocido $\sigma_{\bar{x}} \cdot Z_{\frac{\alpha}{2}} = 1$ y se pide hallar n.

Como $Z_{\frac{\alpha}{2}} = 1'96 \Rightarrow \sigma_{\bar{x}} = \frac{1}{1'96} = 0'5102$, y a partir de $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ se tiene

$$n = \frac{\sigma^2}{\sigma_{\bar{x}}^2} = \frac{3'5^2}{0'5102^2} = 47'059$$

Es decir, será necesaria una muestra de al menos 48 unidades.

- ***De una forma general este problema puede resolverse del siguiente modo:***

Si la anchura del intervalo es 2L se cumple:

$\sigma_{\bar{x}} \cdot Z_{\frac{\alpha}{2}} = L \Rightarrow \frac{\sigma}{\sqrt{n}} \cdot Z_{\frac{\alpha}{2}} = L$, y por tanto, $n = \frac{\sigma^2 \cdot Z_{\frac{\alpha}{2}}^2}{L^2}$

➔ El tamaño de la muestra es menor de 30

La distribución de la variable ***"media de las muestras"*** ya no se aproxima con garantías a la normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ y todos los procedimientos que se han explicado hasta ahora ya no son válidos. Se puede demostrar que la variable

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

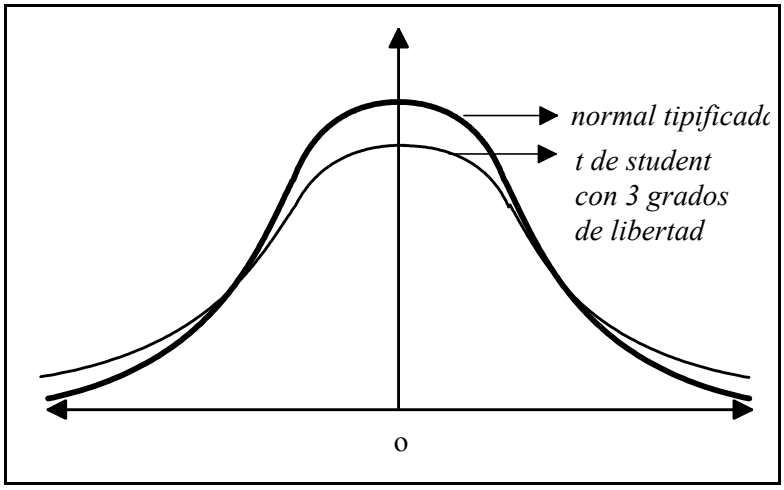
responde a una distribución que recibe el nombre de **distribución t de Student,** con $n - 1$ grados de libertad.

Esta distribución de Student se caracteriza por:

- t toma valores entre $-\infty \quad y \quad +\infty$
- La distribución es simétrica respecto al eje de ordenadas y posee un máximo en $t = 0$.
- La gráfica de la función densidad es semejante a la de la normal, aunque algo más aplanada.
- Para conocer la probabilidad de que t sea mayor que una determinada cantidad, por ejemplo

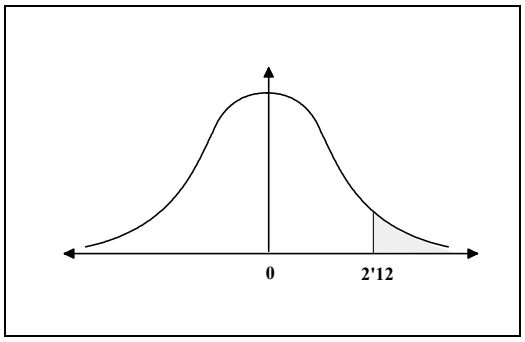
$$P(\text{t} > 2{,}120)$$

es necesario fijar una nueva variable que recibe el nombre de ***"grado de libertad",*** variable que condiciona la gráfica de la función densidad.



- Si la muestra crece de tamaño, la distribución *t* de Student se aproxima a la normal.

El cálculo de probabilidades con esta nueva distribución es semejante al que se realiza a partir de la normal, con la diferencia de que, en este caso, es necesario indicar el número de grados de libertad. La ***Tabla 4*** nos facilita este cálculo, fijando en la filas los grados de libertad y en las columnas las correspondientes probabilidades. Así, con 16 grados de libertad se cumple:



$$P(t > 2'120) = 0'025$$

lo que gráficamente indicaría que el área rayada en la figura de la izquierda es igual a 0,025.

De acuerdo con esto y siguiendo un proceso análogo al empleado con la normal para muestras de tamaño grande, si se desea estimar la media μ de una variable dentro de una población con la media $\bar{x}$ de una muestra de tamaño n (n < 30) y con un nivel de confianza $1-\alpha$se cumplirá:

$$P\left(-t_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

dónde $t_{\frac{\alpha}{2}}$ indica el valor crítico dentro de una distribución *t* de Student con n-1 grados de libertad para un nivel de confianza $1-\alpha$; en este caso, los valores críticos dependen, además del valor de $1-\alpha$, del número de grados de libertad. Precisamente la tabla de la distribución *t* es una tabla de valores críticos pues permite calcularlos de una forma directa, no como sucedía en la normal.

Es decir, con un nivel de confianza $1-\alpha$ se verifica: :

$$-t_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < t_{\frac{\alpha}{2}} \Rightarrow -t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \bar{x}-\mu < t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow$$

$$\Rightarrow \bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

lo cual da el intervalo estimado para

$$\left(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

- **Ejemplo 6:**

Se desea estimar mediante un intervalo la media de la velocidad punta de un modelo de automóvil con un nivel de confianza del 95 % sabiendo que la desviación típica es de 2 Km/h. Para resolver este problema se eligen al azar 15 vehículos y se observa que la velocidad punta media que alcanzan los vehículos de la muestra es de 192 Km/h.

- **Solución:**

Con estos datos, el cálculo de $t_{\frac{\alpha}{2}}$ con $\alpha = 0'05$ y 14 grados de libertad, buscado en las tablas es:

$$t_{0'025,14} = 2'145$$

y por consiguiente el intervalo es:

$$\left(192 - 2'145 \cdot \frac{2}{\sqrt{15}}, 192 + 2'145 \cdot \frac{2}{\sqrt{15}}\right) = (190'8923, 193'1077)$$

➔ 8.2 Estimación a partir de una muestra de la proporción de un suceso en una población

El proceso que se seguirá es idéntico al realizado para la estimación de la media, variando solamente la normal, ya que, en este caso, se trabajará con:

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

y, por tanto, todas las fórmulas vistas anteriormente son válidas para este caso, sustituyendo μ, por p y $\sigma_{\bar{x}}$ por $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$.

Ejemplo 7:

Una cadena de televisión, para decidir sobre la continuidad en antena de un programa, realiza una encuesta de aceptación de dicho programa. Se encuesta a mil personas y se obtiene que el 54 % es partidario de la continuidad de dicho programa. Se pide determinar un intervalo del porcentaje de la población con un nivel de confianza del 98 %.

Solución:

Si se llama p al porcentaje de toda la población, se obtiene:

$$p \in \left(0'54 - \sqrt{\tfrac{0'54 \cdot 0'46}{1000}} \cdot 2'325, 0'54 + \sqrt{\tfrac{0'54 \cdot 0'46}{1000}} \cdot 2'235\right)$$

$$p \in (0'54034, 0'5766)$$

y por lo tanto con un nivel de confianza del 98 % puede asegurarse que el porcentaje de personas favorables a la continuidad del programa estará comprendido entre el 50'34 % y el 57'66 %.

Ejemplo 8:

En una encuesta realizada a 80 de las 2000 personas activas con que cuenta un municipio, se ha comprobado que 36 trabajan en el sector primario. Calcula los límites de confianza del 95 % de la proporción exacta del total de empleados en este sector.

Solución:

El intervalo de confianza correspondiente a este nivel de confianza es:

$$(p_i - 1'96 \cdot \sigma_p, p_i + 1'96 \cdot \sigma_p), \text{ siendo } p_i = \tfrac{36}{80} = 0'45$$

Aun cuando el enunciado del problema no lo especifique, parece razonable que la muestra ha sido obtenida sin reposición, por lo que el error típico deberá estimarse a partir de :

$$\sigma_p = \sqrt{\tfrac{p_i(1-p_i)}{n}} \cdot \sqrt{\tfrac{N-n}{N-1}} = \sqrt{\tfrac{0'45 \cdot 0'55}{80}} \cdot \sqrt{\tfrac{2000-80}{2000-1}} = 0'0545$$

Por lo que el intervalo de confianza será: $(0'3432, 0'5568)$

En conclusión, la proporción poblacional, de la que la muestra extraída constituye la mejor estimación, se halla comprendida entre 0'3432 y 0'5568 con una probabilidad o nivel de confianza de 0'95.

9. Hipótesis estadísticas

En el trabajo científico nos vemos obligados con frecuencia a tomar decisiones relativas a la población sobre la base de la información obtenida de muestras de la misma. Tales decisiones se llaman ***decisiones estadísticas***. Por ejemplo, podemos querer decidir, basados en datos muéstrales, si un método pedagógico es mejor que otro, o si una moneda está trucada o no.

Al intentar tomar una decisión, es útil hacer hipótesis o conjeturas sobre la población implicada. Tales hipótesis, que pueden ser ciertas o no, se llaman ***hipótesis estadísticas.*** Son, en general, enunciados sobre las distribuciones de probabilidad de las poblaciones.

HIPÓTESIS NULA: En muchos casos, formulamos una hipótesis estadística con el único propósito de rechazarla o invalidarla. Así, si queremos decidir si una moneda está trucada, formulamos la hipótesis de que la moneda es buena (o sea, que la probabilidad de obtener cara es p=O'5). Análogamente, si deseamos decidir si un procedimiento es mejor que otro, formulamos la hipótesis de que no hay diferencia entre ellos (o sea, que cualquier diferencia observada se debe simplemente a fluctuaciones en el muestreo de la misma población). Una hipótesis de este tipo suele llamar ***hipótesis nula*** y se denota por Ho.

HIPÓTESIS ALTERNATIVA: Toda hipótesis que difiera de una hipótesis nula dada se llama ***hipótesis alternativa***. Por ejemplo, si la hipótesis nula es p=O'5, hipótesis alternativas podrían ser: p≠0'5, p>O'5 ó p<O'5. Una hipótesis alternativa de la hipótesis nula se denota por H_1 .

➔ 9.1 Contraste de hipótesis

Si suponemos que una hipótesis es cierta pero vemos que los resultados hallados en una muestra aleatoria difieren notablemente de los esperados bajo la hipótesis (o sea, esperados sobre la base del puro azar), entonces diremos que las diferencias observadas son ***significativas,*** y nos veremos inclinados a rechazar la hipótesis (o al menos a no aceptarla ante la evidencia obtenida). Por ejemplo, si en 40 lanzamientos de una moneda salen 35 caras, estaríamos inclinados a rechazar la hipótesis de que la moneda es buena, aunque cabe la posibilidad de equivocarnos.

Los procedimientos que permiten determinar si las muestras observadas difieren *significativamente* de los resultados esperados y, por tanto, nos ayudan a decidir si aceptamos o rechazamos hipótesis, se llaman ***pruebas de decisión o contrastes (o test) de hipótesis o reglas de decisión.***

El contraste de hipótesis más simple se plantea cuando se dispone de un estadístico observado en una muestra aleatoria de tamaño n obtenida de una población formada por N elementos y se desea saber si el valor desconocido del respectivo parámetro poblacional es o no igual a un determinado parámetro teórico dado. La formulación de las correspondientes hipótesis nula, H_0, y alternativa, H_1, podría ser así:

- Hipótesis nula, H_0: La muestra en la que ha sido observado un cierto estadístico $(\bar{x}_i, p_i, \sigma_i, ...)$ procede de una población cuyo respectivo parámetro es igual a un determinado parámetro teórico dado $(\mu, p, \sigma, ...)$

- Hipótesis alternativa, H_1: La muestra en la que ha sido observado un cierto estadístico $(\bar{x}_i, p_i, \sigma_i, ...)$procede de una población cuyo respectivo parámetro difiere de un determinado parámetro teórico dado $(\mu, p, \sigma, ...)$

Cualquiera que sea la solución que tomemos, en favor o en contra de la hipótesis, no estará exenta del riesgo de cometer errores, cuya naturaleza es necesario conocer antes de iniciar cualquier contraste de hipótesis.

➔ 9.2 Errores

Las pruebas de decisión estadística no garantizan la certeza absoluta de la decisión, tan sólo una probabilidad, más o menos alta, de acertar al rechazar o no una hipótesis. Consideraremos dos tipos de errores de decisión:

ERROR DE TIPO I: Se comete al rechazar la hipótesis nula, H_0, cuando es verdadera.

ERROR DE TIPO II: Se comete al no rechazar la hipótesis nula, H_0, cuando es falsa.

Para que las reglas de decisión (o contrastes de hipótesis) sean buenas, deben diseñarse de modo que minimicen los errores de decisión. Y no es una cuestión sencilla porque, para cualquier tamaño de la muestra, un intento de disminuir un tipo de error suele ir acompañado de un crecimiento del otro tipo. En la práctica, un tipo de error puede ser más grave que el otro, y debe alcanzarse un compromiso que disminuya el error más grave. La única forma de disminuir ambos a la vez es aumentar el tamaño de la muestra, que no siempre es posible.

➔ 9.3 Nivel de significación y de confianza

Al contrastar una hipótesis, la probabilidad de cometer un error de tipo I se llama ***nivel de significación*** (o ***riesgo)*** del contraste. Esta probabilidad, denotada a menudo por α, se suele especificar antes de tomar la muestra, de manera que los resultados obtenidos no influyan en la decisión.

Se llama ***nivel de confianza (***o ***fiabilidad)*** de un contraste hipótesis al número $1-\alpha$.

Si, por ejemplo, se escoge el nivel de significación 0'05 (5%) al diseñar una regla de decisión, entonces hay una probabilidad de 0'05 (5%) de rechazar la hipótesis nula cuando sea verdadera; es decir, tenemos un nivel de confianza de 0'95 (95%) de tomar la decisión correcta al rechazar la hipótesis nula. En tal caso decimos que *la hipótesis nula ha sido rechazada a un nivel de significación 0'05 (o a un nivel de confianza 0* 95), lo cual quiere decir que la hipótesis nula tiene una probabilidad de 0'05 de ser verdadera.

El nivel de significación α se fija para cada contraste de hipótesis y, en consecuencia, es siempre conocido.

Razones de orden práctico han consagrado el uso de los niveles de significación de 0'05 y 0'01, sobre todo el primero, si bien se pueden usar otros valores.

La probabilidad de cometer un error de tipo II, denotada por β*, no puede fijarse y, aunque en general disminuye a medida que aumenta la probabilidad de cometer un error de tipo I, la relación entre ambas no es lineal.*

Puesto que β es desconocida, ***la conclusión de un contraste de hipótesis nunca será la aceptación de la hipótesis nula*** H_0. A lo sumo podrá anunciarse en términos de no oposición a su aceptación o de inexistencia de diferencias estadísticas significativas, mediante fórmulas similares a las siguientes: "nada se opone a aceptar la hipótesis nula ..." o bien "la diferencia encontrada no es estadísticamente significativa

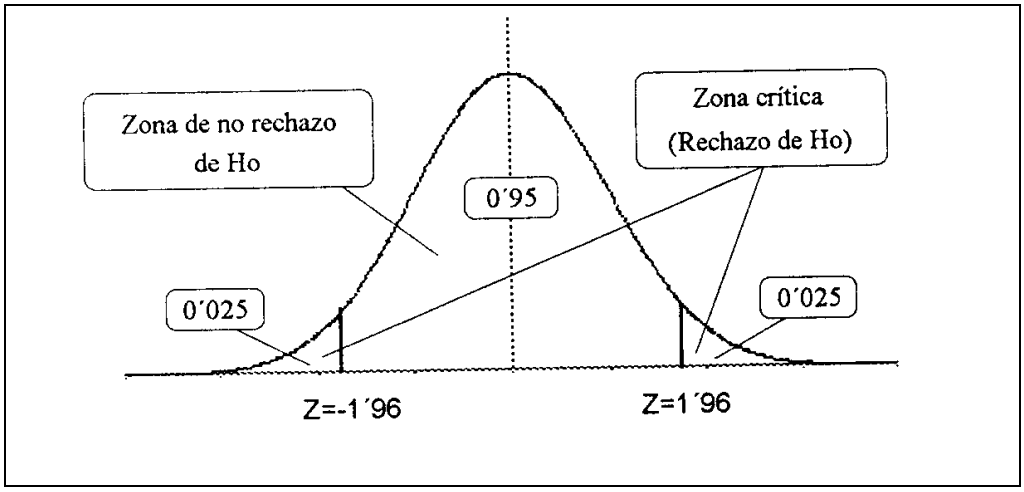
En cambio, ***un contraste de hipótesis sí puede concluirse con un rechazo de*** **H0 ,*o una aceptación de*** $\mathbf{H_1}$ ***con un determinado riesgo*** α, porque la única probabilidad de error al rechazar la hipótesis nula es de tipo I, que ha sido fijada previamente.

➔ 9.4 Contrastes de hipótesis basados en la distribución normal

Supongamos que, bajo cierta hipótesis nula H_0, la distribución muestral de un estadístico E es normal con media μ_e y desviación típica σ_e. Así, pues, la distribución de la variable tipificado Z, dada por:

$$Z = \frac{E-\mu_e}{\sigma_e}$$

es la distribución normal standard N(O, 1)

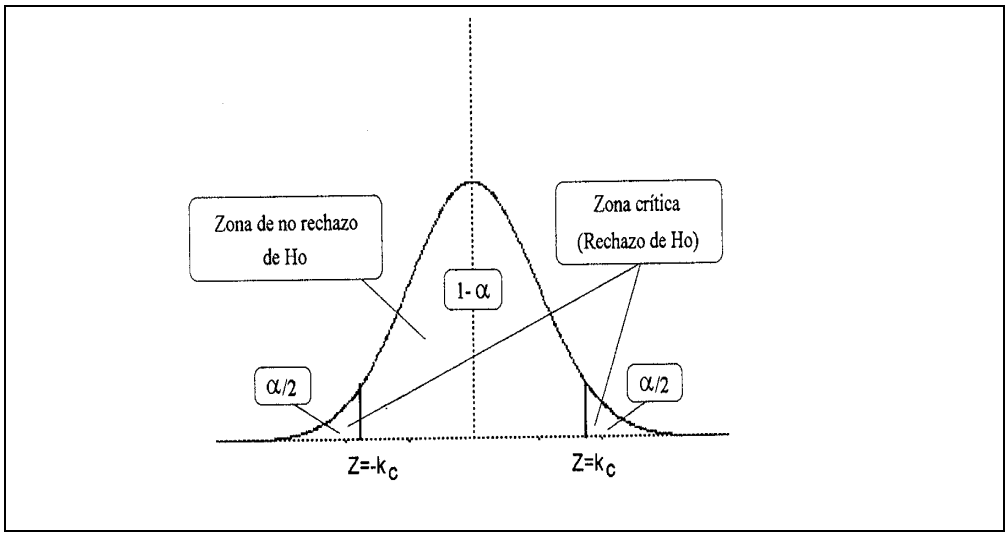


Como se ve en la figura, si al escoger una sola muestra aleatoria hallamos que el valor de Z de su estadístico, E, está fuera del intervalo [-1'96,1'96], debemos concluir que tal suceso podría ocurrir con una probabilidad de sólo 0'05 (la zona crítica señalada en la figura), si la hipótesis nula Ho dada fuera cierta. Diremos, entonces, que *este valor de Z difiere de forma significativa* de lo que cabría esperar bajo la hipótesis nula, y nos vemos obligados a rechazar la hipótesis nula Ho.

El área señalada, 0'05, es el nivel de significación (o riesgo) del contraste. Representa la probabilidad de equivocamos al rechazar la hipótesis nula (o sea, la probabilidad de cometer un error de tipo I). Así, pues, decimos que la *hipótesis nula se rechaza al nivel de significación 0'05,* o que *el valor de Z del estadístico muestral dado es significativo al nivel 0'05.* También podemos decir que *la hipótesis nula se rechaza a un nivel de confianza 0'95.*

En general, si al escoger una sola muestra aleatoria encontramos que el valor de Z de su estadístico está fuera del intervalo $[-k_c, , k_c]$, debemos concluir que tal suceso podría ocurrir

con una probabilidad α, si la hipótesis nula H_0 dada fuera cierta. En estas condiciones, se *rechaza la hipótesis nula a un nivel de significación* α.



El conjunto de valores de Z fuera del intervalo $[-k_c , k_c]$ se llama ***zona crítica*** de la hipótesis (o ***zona de rechazo*** de la hipótesis o ***zona de significación)***.

El conjunto de valores de Z en el intervalo $[-k_c , k_c]$ se conoce como ***zona de no rechazo*** de la hipótesis (o ***zona de no significación)***.

kc recibe el nombre de ***valor crítico*** de Z.

Basándonos en lo que acabamos de exponer, podemos formular la siguiente ***regla de decisión:***

- Rechazar la hipótesis nula al nivel de significación α, si el valor de Z para el estadístico E está fuera del intervalo $[-k_c , k_c]$ (o sea, si $Z > k_c$ ó $Z < -k_c$). Esto equivale a decir que el estadístico muestral observado es significativo al nivel α.

- No rechazar la hipótesis nula en caso contrario (o sea, si $-k_c < Z < k_c$).

El valor de Z juega un papel muy importante en el contraste de hipótesis. Se le llama ***estadístico de contraste o razón crítica,*** y lo denotaremos por Zc .

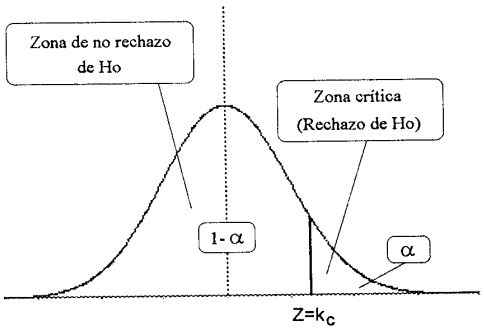
➔ 9.5 Contrastes unilaterales y bilaterales

En el contraste o test precedente estábamos interesados en los valores extremos del estadístico E o en su correspondiente valor de Z a *ambos lados de la media* (o sea, en las dos colas de la distribución normal). Tales contrastes se llaman contrastes ***bilaterales*** o contrastes ***de dos colas***.

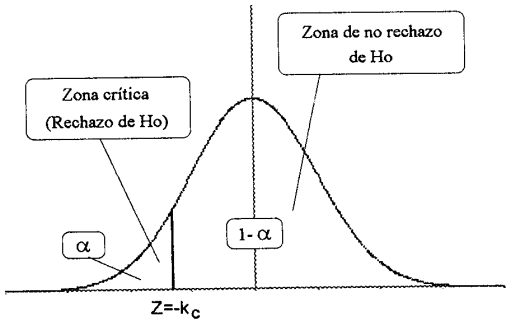
No obstante, podemos estar interesados tan sólo en valores extremos a un lado de la media (o sea, en una de las colas de la distribución normal), tal como sucede cuando se contrasta la hipótesis de que un proceso es mejor que otro (lo cual no es lo mismo que

contrastar si un proceso es mejor o peor que otro). Tales contrastes se llaman ***unilaterales*** o ***de una cola***.

En tales situaciones, *la zona crítica es una región situada a un lado de la distribución* (a la derecha o a la izquierda), con área igual al nivel de significación α, tal como se ve en la figura.



ZONA CRÍTICA A LA DERECHA



ZONA CRÍTICA A LA IZQUIERDA

La tabla siguiente muestra los valores críticos de Z (para contrastes unilaterales y bilaterales) correspondiente a los niveles de significación más importantes. Los valores críticos de Z para otros niveles de significación se hallan utilizando la tabla de la distribución normal (Tabla 3).

Nivel de significación: α	0'10 (10 %)	0'05 (5 %)	0'01 (1 %)	0'005 (0'5 %)
Valor crítico de Z para tests unilaterales: k_c	1'28	1'645	2'33	2'58
Valor crítico de Z para tests bilaterales: k_c	1'645	1'96	2'59	2'81

☐ Ejemplo 1:

Se pretende saber, mediante un contraste de hipótesis, si una muestra formada por 1000 españoles con derecho a voto en marzo de 1986, de los que 640 participaron en el referéndum sobre la OTAN celebrado el día 12 de dicho mes, procedía de una extracción aleatoria realizada entre las personas censadas en la Comunidad Valenciana, donde el porcentaje de participantes en la consulta fue del 66'5%.

❑ *Solución:*

Para ello, se procede a formular las hipótesis nula y alternativa siguientes:

Hipótesis nula H_0: La muestra en la que ha sido observada la proporción de participantes, $p_i = \frac{640}{1000} = 0'64$, procede de una población cuya proporción es igual a la proporción, p = 0'665, registrada en la Comunidad Valenciana.

Hipótesis alternativa H_1: La muestra procede de una población con una proporción de participantes diferente a la proporción, p = 0'665, registrada en la Comunidad Valenciana.

Se trata de un contraste bilateral.

Comenzaremos por tipificar el valor de $p_i = 0'64$

$$Z = \frac{p_i - \mu_p}{\sigma_p} = \frac{p_i - \mu_p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0'64 - 0'665}{\sqrt{\frac{0'665 \cdot 0'335}{1000}}} = -1'67$$

Si fijamos un nivel de significación $\alpha = 0'05$, el valor crítico de Z (Ver la tabla anterior) correspondiente es kc = 1'96.

Observamos que -1'67 < Z < 1'67; por lo que la diferencia entre la proporción observada, $p_i = 0'64$, y la proporción poblacional, $p = 0'665$, puede ser debida a las variaciones inherentes a la aleatoriedad del muestreo; por cuyo motivo nada se opondría a aceptar H_0, es decir, no rechazará esta hipótesis, lo cual, sin embargo, no significa que se haya demostrado su cumplimiento.

➔ 9.6 Contraste entre un estadístico y su parámetro

Los contrastes de hipótesis más sencillos tienen como objetivo decidir si una muestra aleatoria, en la que se ha observado un determinado estadístico, puede proceder de una población cuyo respectivo parámetro es conocido.

Para grandes muestras, las distribuciones muéstrales de algunos estadísticos son normales (o casi normales), por lo que en estos casos podremos aplicar los contrastes de hipótesis basados en la distribución normal que hemos explicado en el apartado 9.1.

Un procedimiento práctico sistemático para realizar un contraste de hipótesis entre un estadístico y su parámetro consiste en:

1. Establecer las hipótesis:

- Hipótesis nula H_0

- Hipótesis alternativa H_1

2. Hallar el valor crítico kc

3. Hallar la razón crítica $Z_c = \frac{E - \mu_e}{\sigma_e}$

siendo:

E : Valor del estadístico obtenido de la muestra.

μ_e: Media de la distribución muestral del estadístico.

σ_e: Error típico del estadístico.

4. Tomar la decisión estadística:

- Si $|Z_c| \leq k_c$: Nada se opone a aceptar H_0.

- Si $|Z_c| > k_c$: Se rechaza H_0 y se acepta H_1 con riesgo α.

► Contraste de la media

En el caso de la media, la razón crítica, Z_c, es: $Z_c = \frac{\bar{x}_i - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$

donde:

$\bar{x}_i$: Media obtenida de la muestra aleatoria.

$\mu_{\bar{x}}$: Media de la distribución muestral de medias.

$\sigma_{\bar{x}}$: Error típico de la distribución muestral de medias.

μ : Media teórica de la población en la que se toma la muestra aleatoria.

☐ Ejemplo 2:

La superficie media de las explotaciones agrarias españolas, según el censo de 1982, es de 18'90 Ha y su desviación típica de 116'52.

(a) ¿Puede afirmarse que difiere de la superficie media de las explotaciones censadas, la de una comarca en la que la media de una muestra aleatoria de 500 explotaciones ha resultado ser de 14'5 0 Ha, con una desviación típica de 110'30, si se acepta un riesgo de 0'05 ?

(b) ¿Sería correcto afirmar, con un riesgo de 0'01, que la superficie media de las explotaciones de una comarca, en la que se supone una mayor abundancia de las unidades de explotación grandes, sea realmente superior a la del conjunto de España, si la observada en una muestra de 2000 explotaciones pertenecientes a dicha comarca, tomadas al azar, es 25'90 Ha?

☐ *Solución:*

(a) Debe decidirse si la superficie media desconocida, μ, de las explotaciones de la comarca en la que se ha tomado la muestra, es igual o no a la media poblacional conocida, 18'90 Ha; por lo que plantearemos las dos hipótesis:

- Hipótesis nula H_0: μ= 18'90

- Hipótesis alternativa H_1: $\mu \neq 18'90$

En consecuencia, optaremos por un *contraste bilateral* con los datos siguientes:

Tamaño de la muestra: n = 500

Riesgo aceptado: α = 0'05

Valor crítico: k_c = 1'96

Media aritmética de la muestra: $\bar{x}_i$ = 14'50

Media teórica poblacional: μ = 18'90

Desviación típica poblacional: σ = 116'52

No será preciso utilizar la desviación típica de la muestra dada en el enunciado del problema porque se conoce la desviación típica poblacional.

El error típico de la media, $\sigma_{\bar{x}}$, será: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{116'52}{\sqrt{500}} \approx 5'21$

La razón crítica, Zc, valdrá:

$$Z_c = \frac{\bar{x}_i - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{14'50 - 18'90}{5'21} \approx -0'84$$

Puesto que $|Z_c| = 0'84 < 1'96$*, nada se opone a aceptar la hipótesis nula* H0. Es decir, que la superficie media de las explotaciones de la comarca aludida puede ser igual a la media de las explotaciones censadas y no sería correcto afirmar, con los datos conocidos, que sea diferente a ésta.

(b) Se trata de comprobar si la superficie media desconocida, μ, de las explotaciones de la comarca origen de la muestra, es superior a la media poblacional conocida, 18'90 Ha. La decisión consistirá en optar por una de las dos hipótesis siguientes:

- Hipótesis nula H_0: μ = 18'90

- Hipótesis alternativa H_1: μ > 18'90

En consecuencia, aplicaremos un *contraste unilateral (con la zona crítica a la derecha)* con los datos siguientes:

Tamaño de la muestra: n = 2000

Riesgo aceptado: α = 0'01

Valor crítico: k_c = 2'33

Media aritmética de la muestra: $\bar{x}_i$ = 25'90

Media teórica poblacional: μ = 18'90

Desviación típica poblacional: σ = 116'52

El error típico de la media, $\sigma_{\bar{x}}$, será: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{116'52}{\sqrt{2000}} \approx 2'61$

La razón crítica, Zc, valdrá: $Z_c = \frac{\bar{x}_i - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{25'90 - 18'90}{2'61} \approx 2'68$

Como $|Z_c| = 2'68 > 2'33$, *se rechaza la hipótesis nula Ho y se acepta la hipótesis alternativa H_1 con un riesgo* α = 0'01. Es altamente probable que la superficie media de las explotaciones agrarias de la comarca sea superior a la media de la totalidad de las explotaciones censadas en España.

► Contraste de una proporción

En el caso de una proporción, la razón crítica, Zc , es:

$$Z_c = \frac{p_i - \mu_p}{\sigma_p}$$

donde:

p_i: Proporción obtenida de la muestra aleatoria.

μ_p: Media de la distribución muestral de proporciones.

σ_p: Error típico de una proporción.

p : Proporción teórica de la población en la que se toma la muestra aleatoria.

☐ Ejemplo 3:

La proporción de población ocupada en la agricultura entre el total de personas con trabajo durante el cuarto trimestre de 1985 en Extremadura era equivalente a 0'332. Comprueba:

(a) Si una muestra de 500 personas con trabajo, de las que 142 estaban ocupadas en la agricultura, podía haber sido extraída al azar entre las personas que trabajaban en la Comunidad de Extremadura a finales de 1985.

(b) Si era superior a la aludida proporción de población agraria calculada para el conjunto de esa comunidad, la de una comarca en la que, según los resultados de una encuesta realizada a 1000 personas con trabajo tomadas al azar, 360 estaban ocupadas en la agricultura.

En ambos casos se acepta un riesgo del 5%.

❑ *Solución:*

(a) Se trata de decidir si la proporción desconocida, p, de la comarca donde se ha extraído la muestra difiere o no de la proporción poblacional 0'332; es decir, elegir entre las hipótesis:

- Hipótesis nula H_0: p = 0'332
- Hipótesis alternativa H_1: $p \neq$ 0'332

Por tanto, aplicamos un *contraste bilateral* con los datos:

Tamaño de la muestra: n = 500
Nivel de significación: $\alpha = 0'05$
Proporción teórica poblacional: $p = 0'332$
Proporción de la muestra: $p_i = \frac{142}{500} = 0'284$
Valor crítico: kc = 1'96

El error típico, σ_p, será: $\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0'332 \cdot 0'668}{500}} \approx 0'021$

La razón crítica, Z_c, es: $Z_c = \frac{p_i - \mu_p}{\sigma_p} = \frac{0'284 - 0'332}{0'021} \approx -2'29$

Puesto que $|Z_c| = 2'29 > 1'96$, *se rechaza la hipótesis nula* H0 *y se acepta la hipótesis alternativa* H1*, con un riesgo* $\alpha = 0'05$. Por tanto, es poco probable que la muestra haya sido obtenida al azar de entre la totalidad de las personas que ocupaban un puesto de trabajo en la Comunidad de Extremadura en el tercer trimestre de 1985.

(b) Se trata de comprobar si la proporción desconocida, p, de la comarca, de la que únicamente se conoce la proporción p_i observada en una muestra aleatoria, es superior a la proporción poblacional conocida, 0'332. Equivale a elegir entre las dos hipótesis siguientes:

- Hipótesis nula H_0: $p = 0'332$
- Hipótesis alternativa H_1: $p > 0'332$

Puesto que sólo interesa decidir si la proporción desconocida, p, es superior o no a la proporción poblacional dada, 0'332, optaremos por la aplicación de un *contraste unilateral (con zona crítica a la derecha)* con los datos:

Tamaño de la muestra: n = 1 000
Nivel de significación: $\alpha = 0'05$
Proporción teórica poblacional: $p = 0'332$
Proporción de la muestra: $p_i = \frac{360}{1000} = 0'360$
Valor crítico: kc = 1'645

El error típico, σ_p, será: $\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0'332 \cdot 0'668}{1000}} \approx 0'0149$

La razón crítica, Zc, es: $Z_c = \frac{p_i - \mu_p}{\sigma_p} = \frac{0'360 - 0'332}{0'0149} \approx 1'88$

Puesto que $|Z_c| = 1'88 > 1'645$, *se rechaza la hipótesis nula* H_0 *y se acepta la hipótesis alternativa* H_1 *con un riesgo* $\alpha = 0'05$. Por consiguiente, la proporción de personas ocupadas en la agricultura entre los que contaban con un puesto de trabajo el tercer trimestre de 1985 era superior en la comarca de referencia que en el conjunto de Extremadura o, al menos, es altamente probable que así sucediera.

Ejercicios de contraste de hipótesis

1. Para justificar su petición de aumento de salarios, los empleados del departamento de despachos de una firma de venta por correo sostienen que en promedio el departamento

completa una orden en 13 minutos. Si usted es el gerente general de la firma, ¿qué conclusión obtiene si una muestra de 400 órdenes da un tiempo medio de terminación de 14 minutos y una desviación típica de 10 minutos? Utiliza un nivel de significación 0'05.

2. En la investigación de varias denuncias respecto del rótulo "PESO NETO 300 g." que aparece en los frascos de una marca de pasta de cacahuetes tostados de la ciudad, la Comisión de Control del Comercio y la Industria seleccionó una muestra de 36 frascos. La muestra dio un peso neto medio de 298 g. y una desviación típica de 7'5 g. Utilizando el nivel de significación 0'01, ¿qué conclusión debe sacar la Comisión acerca de la operación de la compañía fabricante?

3. Un cierto tipo de fusible está diseñado para fundirse cuando la intensidad de la corriente llega a 20 amperios. De un lote de 100.000 fusibles se seleccionan 36, los cuales son probados en relación con su punto de saturación. ¿Qué conclusión se obtiene acerca de la especificación del amperaje del lote si la muestra da una media de 20'9 amperios y una desviación típica de 1'5 amperios? Utiliza un nivel de significación 0'01.

4. En un informe preparado por el Departamento de Investigación Económica de un gran banco comercial, se expresa que el ingreso familiar anual medio en la zona de Valparaiso, Chile, es de $25.296. ¿Qué conclusión se saca acerca de la validez del informe si una muestra aleatoria simple de 400 familias de la zona da un ingreso medio de $25.722 con una desviación típica de $6.000. Utiliza un nivel de significación 0'05.

5. La Asociación de dueños de establecimientos comerciales detallistas sostiene que el salario medio por hora de sus empleados es de $22,00. El Sindicato de empleados sospecha que la Asociación exagera el valor del salario medio por hora. En una muestra aleatoria de 400 empleados, el Sindicato encuentra que el salario medio por hora es de $21'00 con una desviación típica de $8'00. Si el Sindicato desea rechazar una afirmación verdadera no más de una vez en 100, ¿rechazará el Sindicato la afirmación de la Asociación?

6. Un organismo gubernativo de control analiza una muestra de 36 paquetes de carne picada que vende el supermercado "El Económico". El rótulo en cada paquete dice *"contiene no más de 25 % de grasa"*. ¿Puede el organismo gubernativo concluir que la carne picada que vende el supermercado tiene más del 25 % de grasa si la muestra da un contenido medio de grasa de 0'265 y una desviación estándar de 0'030? Utiliza un nivel de significación de 0'05.

7. El Departamento de Salud y Bienestar piensa que sólo el 10 % de las personas de más de 65 años de edad dispone de un seguro de salud adecuado. En una muestra aleatoria de 900 personas mayores de 65 años, 99 tienen un seguro de salud adecuado. ¿Qué condición puede obtenerse? Utiliza un nivel de significación de 0'05.

8. El presentador de un programa semanal de TV desearía que la asistencia al estudio de grabación se distribuyera en igual proporción entre hombres y mujeres. De 400 personas que asisten al programa en una noche determinada, 220 son hombres. Utilizando un nivel de significación de 0'01, ¿puede el presentador concluir que la proporción por sexo de la concurrencia no es la deseada?

9. Un laboratorio farmacéutico ha elaborado un medicamento para tratar la presión sanguínea alta. El laboratorio afirma que el medicamento efectivamente baja la presión en el 80 % de los casos. Si 175 de 225 pacientes tratados con el medicamento experimentaron una disminución substancial de la presión sanguínea, ¿concluiría usted

que el laboratorio ha exagerado la efectividad del medicamento? Utiliza un nivel de significación de 0'01.

10. En una conferencia de prensa, una alta autoridad anuncia que el 90 % de los habitantes adultos del país están a favor de cierto proyecto económico del Gobierno. Una muestra aleatoria de 625 adultos indica que 550 están en favor del proyecto. Si usted desea rechazar una hipótesis verdadera no más de una vez en 100, ¿concluiría que la popularidad del proyecto ha sido exagerada por la autoridad?

11. La duración de las bombillas de 100 vatios que fabrica una empresa sigue una distribución normal con una desviación típica de 120 horas. Su vida media está garantizada durante un mínimo de 800 horas. Se escoge al azar una muestra de 50 bombillas de un lote y, después de comprobarlas, se obtiene una vida media de 750 horas. Con un nivel de significación de 0'01, ¿habría que rechazar el lote por no cumplir la garantía? (***Selectividad Madrid)***

12. Supongamos una población $N(\mu, \sigma = 8)$. Se extrae de ella una muestra aleatoria simple. Si se sabe que la probabilidad de cometer un error de 3'92 o más al estimar la media poblacional μ mediante la media muestral es de 0'05, ¿qué tamaño ha de tener la muestra? (***Selectividad Castilla y León***)

13. El nivel medio de protombina en una población normal es de 20 mg/100 ml de plasma de plasma con una desviación típica de 4 mg/100 ml. Se toma una muestra de 40 individuos en los que la media es de 18'5 mg/100 ml. ¿Es la muestra comparable con la población, con un nivel de significación del 0'05? (***Selectividad Castilla y León***)

14. Un fabricante de lámparas eléctricas está ensayando un nuevo método de producción que se considerará aceptable si las lámparas obtenidas por este método dan lugar a una población normal de duración media de 2.400 horas, con una desviación típica igual a 300 horas. Se toma una muestra de 100 lámparas producidas por este método y esta muestra da una duración media de 2.320 horas. ¿Se puede aceptar la hipótesis de validez del nuevo proceso de fabricación con un riesgo igual o menor al 5 %?

Formula la hipótesis alternativa. Define error de I y error de tipo II. (***Selectividad Castilla - La Mancha***)

15. Se realizó una encuesta a 350 familias preguntando si poseían ordenador en casa, encontrándose que 75 de ellas lo poseían. Estima la proporción real de familias que disponen de ordenador con un nivel de confianza del 95 %.

16. Al medir el tiempo de reacción, un psicólogo estima que la desviación típica del mismo es de 0'5 segundos. ¿Cuál será el número de medidas que deberá hacer para que sea del 99 % la confianza de que el error de su estimación no excederá de 0'1 segundos?

17. La desviación típica de la altura de los habitantes de un país es de 10 centímetros. Calcular el tamaño mínimo que ha de tener una muestra de habitantes de dicho país para que el error cometido al estimar la altura media sea inferior a 1 centímetro con un nivel de confianza del 99 %. ¿Y si el nivel de confianza es del 95 %? Explicar los pasos seguidos para obtener las respuestas. (***Selectividad - Junio 98***)