UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

DIVISIÓN SISTEMA UNIVERSIDAD ABIERTA Y
EDUCACIÓN A DISTANCIA

LICENCIATURA
en ADMINISTRACIÓN

APUNTES DIGITALES
PLAN 2012

SUAyED UNA OPCIÓN PARA TI

ESTADÍSTICA INFERENCIAL

Plan 2012

Clave:		**Créditos: 8**
Licenciatura: ADMINISTRACIÓN		**Semestre: 2º**
Área: Matemáticas		**Horas asesoría:**
Requisitos:		**Horas por semana: 4**
Tipo de asignatura:	**Obligatoria (X)**	**Optativa ( )**

AUTOR

ELISEO FLORES ALAMILLA

ADAPTACIÓN EN LÍNEA

ELISEO FLORES ALAMILLA

ACTUALIZACIÓN AL PLAN DE ESTUDIOS 2012

LUIS FELIPE ZÚÑIGA

INTRODUCCIÓN AL MATERIAL DE ESTUDIO

Las modalidades abierta y a distancia (SUAYED) son alternativas que pretenden responder a la demanda creciente de educación superior, sobre todo, de quienes no pueden estudiar en un sistema presencial. Actualmente, señala Sandra Rocha (2006):

> Con la incorporación de las nuevas tecnologías de información y comunicación a los sistemas abierto y a distancia, se empieza a fortalecer y consolidar el paradigma educativo de éstas, centrado en el estudiante y su aprendizaje autónomo, para que tenga lugar el diálogo educativo que establece de manera semipresencial (modalidad abierta) o vía Internet (modalidad a distancia) con su asesor y condiscípulos, apoyándose en materiales preparados ex profeso.

Un rasgo fundamental de la educación abierta y a distancia es que no exige presencia diaria. El estudiante SUAYED aprende y organiza sus actividades escolares de acuerdo con su ritmo y necesidades; y suele hacerlo en momentos adicionales a su jornada laboral, por lo que requiere flexibilidad de espacios y tiempos. En consecuencia, debe contar con las habilidades siguientes.

- Saber estudiar, organizando sus metas educativas de manera realista según su disponibilidad de tiempo, y estableciendo una secuencia de objetivos parciales a corto, mediano y largo plazos.

- Mantener la motivación y superar las dificultades inherentes a la licenciatura.
- Asumir su nuevo papel de estudiante y compaginarlo con otros roles familiares o laborales.
- Afrontar los cambios que puedan producirse como consecuencia de las modificaciones de sus actitudes y valores, en la medida que se adentre en las situaciones y oportunidades propias de su nueva situación de estudiante.
- Desarrollar estrategias de aprendizaje independientes para que pueda controlar sus avances.
- Ser autodidacta. Aunque apoyado en asesorías, su aprendizaje es individual y requiere dedicación y estudio. Acompañado en todo momento por su asesor, debe organizar y construir su aprendizaje.
- Administrar el tiempo y distribuirlo adecuadamente entre las tareas cotidianas y el estudio.
- Tener disciplina, perseverancia y orden.
- Ser capaz de tomar decisiones y establecer metas y objetivos.
- Mostrar interés real por la disciplina que se estudia, estar motivado para alcanzar las metas y mantener una actitud dinámica y crítica, pero abierta y flexible.
- Aplicar diversas técnicas de estudio. Atender la retroalimentación del asesor; cultivar al máximo el hábito de lectura; elaborar resúmenes, mapas conceptuales, cuestionarios, cuadros sinópticos, etcétera; presentar trabajos escritos de calidad en contenido, análisis y reflexión; hacer guías de estudio; preparar exámenes; y aprovechar los diversos recursos de la modalidad.

Además de lo anterior, un estudiante de la modalidad a distancia debe dominar las herramientas tecnológicas. Conocer sus bases y metodología; tener habilidad en la búsqueda de información en bibliotecas virtuales; y manejar el sistema operativo Windows, paquetería, correo electrónico, foros de discusión, chats, blogs, wikis, etcétera.

También se cuenta con materiales didácticos como éste elaborados para el SUAYED, que son la base del estudio independiente. En específico, este documento electrónico ha sido preparado por docentes de la Facultad para cada una de las asignaturas, con bibliografía adicional que te permitirá consultar las fuentes de información originales. El recurso comprende referencias básicas sobre los temas y subtemas de cada unidad de la materia, y te introduce en su aprendizaje, de lo concreto a lo abstracto y de lo sencillo a lo complejo, por medio de ejemplos, ejercicios y casos, u otras actividades que te posibilitarán aplicarlos y vincularlos con la realidad laboral. Es decir, te induce al “saber teórico” y al “saber hacer” de la asignatura, y te encauza a encontrar respuestas a preguntas reflexivas que te formules acerca de los contenidos, su relación con otras disciplinas, utilidad y aplicación en el trabajo. Finalmente, el material te da información suficiente para autoevaluarte sobre el conocimiento básico de la asignatura, motivarte a profundizarlo, ampliarlo con otras fuentes bibliográficas y prepararte adecuadamente para tus exámenes. Su estructura presenta los siguientes apartados.

1. *Información general de la asignatura.* Incluye elementos introductorios como portada, identificación del material, colaboradores, datos oficiales de la asignatura, orientaciones para el estudio, contenido y programa oficial de la asignatura, esquema general de contenido, introducción general a la asignatura y objetivo general.
2. *Desarrollo de cada unidad didáctica.* Cada unidad está conformada por los siguientes elementos:

   Introducción a la unidad.

   Objetivo particular de la unidad.

   Contenidos.

   Actividades de aprendizaje y/o evaluación. Tienen como propósito contribuir en el proceso enseñanza-aprendizaje facilitando el afianzamiento de los contenidos esenciales. Una función importante de estas actividades es la retroalimentación: el asesor no se limita a valorar el trabajo realizado, sino que además añade comentarios, explicaciones y orientación.

   Ejercicios y cuestionarios complementarios o de reforzamiento. Su finalidad es consolidar el aprendizaje del estudiante.

   Ejercicios de autoevaluación. Al término de cada unidad hay ejercicios de autoevaluación cuya utilidad, al igual que las actividades de aprendizaje, es afianzar los contenidos principales. También le permiten al estudiante calificarse él mismo cotejando su resultado con las respuestas que vienen al final, y así podrá valorar si ya aprendió lo suficiente para presentar el examen correspondiente.

Para que la autoevaluación cumpla su objeto, es importante no adelantarse a revisar las respuestas antes de realizar la autoevaluación; y no reducir su resolución a una mera actividad mental, sino que debe registrarse por escrito, labor que facilita aún más el aprendizaje. Por último, la diferencia entre las actividades de autoevaluación y las de aprendizaje es que éstas, como son corregidas por el asesor, fomentan la creatividad, reflexión y valoración crítica, ya que suponen mayor elaboración y conllevan respuestas abiertas.

3. *Resumen* por unidad.
4. *Glosario* de términos.
5. *Fuentes* de consulta básica y complementaria. Mesografía, bibliografía, hemerografía, sitios web, entre otros, considerados tanto en el programa oficial de la asignatura como los sugeridos por los profesores.

Esperamos que este material cumpla con su cometido, te apoye y oriente en el avance de tu aprendizaje.

Recomendaciones (orientación para el estudio independiente):

- Lee cuidadosamente la introducción a la asignatura, en ella se explica la importancia del curso.
- Revisa detenidamente los objetivos de aprendizaje (general y específico por unidad), en donde se te indican los conocimientos y habilidades que deberás adquirir al finalizar el curso.
- Estudia cada tema siguiendo los contenidos y lecturas sugeridos por tu asesor, y desarrolla las actividades de aprendizaje. Así podrás aplicar la teoría y ejercitarás tu capacidad crítica, reflexiva y analítica.
- Al iniciar la lectura de los temas, identifica las ideas, conceptos, argumentos, hechos y conclusiones, esto facilitará la comprensión de los contenidos y la realización de las actividades de aprendizaje.
- Lee de manera atenta los textos y mantén una actitud activa y de diálogo respecto a su contenido. Elabora una síntesis que te ayude a fijar los conceptos esenciales de lo que vas aprendiendo.
- Debido a que la educación abierta y a distancia está sustentada en un principio de autoenseñanza (autodisciplina), es recomendable diseñar desde el inicio un plan de trabajo para puntualizar tiempos, ritmos, horarios, alcance y avance de cada asignatura, y recursos.
- Escribe tus dudas, comentarios u observaciones para aclararlas en la asesoría presencial o a distancia (foro, chat, correo electrónico, etcétera).

- Consulta al asesor sobre cualquier interrogante por mínima que sea.
- Revisa detenidamente el plan de trabajo elaborado por tu asesor y sigue las indicaciones del mismo.

Otras sugerencias de apoyo

- Trata de compartir tus experiencias y comentarios sobre la asignatura con tus compañeros, a fin de formar grupos de estudio presenciales o a distancia (comunidades virtuales de aprendizaje, a través de foros de discusión y correo electrónico, etcétera), y puedan apoyarse entre sí.
- Programa un horario propicio para estudiar, en el que te encuentres menos cansado, ello facilitará tu aprendizaje.
- Dispón de periodos extensos para al estudio, con tiempos breves de descanso por lo menos entre cada hora si lo consideras necesario.
- Busca espacios adecuados donde puedas concentrarte y aprovechar al máximo el tiempo de estudio.

TEMARIO DETALLADO

	Horas
1. Introducción al muestreo	4
2. Distribuciones muestrales	8
3. Estimación de parámetros	10
4. Pruebas de hipótesis	10
5. Pruebas de hipótesis con la distribución ji cuadrada	8
6. Análisis de regresión lineal simple	10
7. Análisis de series de tiempo	8
8. Pruebas estadísticas no paramétricas	6
TOTAL	**96**

INTRODUCCIÓN

En esta asignatura el estudiante dará continuación al curso previo de Estadística I. Observando la importancia que tiene el aprenderla, así:

En la unidad 1 investigará y aplicará la teoría del muestreo a diferentes tipos de problemas y, en consecuencia, diferentes tipos de muestras. Observará los retos que implica la correcta selección de una muestra con el objetivo de que su estudio tenga la validez científica y la exactitud de la matemática.

En la unidad 2 estudiará las distribuciones muestrales y el teorema central del límite, los cuales pueden ayudar para la posterior elaboración de los intervalos de confianza.

En la unidad 3 estimará los parámetros principales con el fin de tomar decisiones en un entorno de incertidumbre.

En la unidad 4 aplicará las pruebas de hipótesis en el ambiente administrativo y contable para poder decidir continuar o desechar alguna forma de actuar de la compañía donde se encuentre laborando, basado en hechos científicos.

En la unidad 5 se analizarán las pruebas de hipótesis con la distribución ji cuadrada y su aplicación.

En la unidad 6 investigará el análisis de regresión lineal simple para averiguar el comportamiento de las variables y sus diferentes relaciones.

En la unidad 7 se analizarán las series de tiempo para observar su aplicación a diferentes problemas de la vida diaria de las empresas.

En la unidad 8 analizará las pruebas estadísticas no paramétrica para poder racionalizar fenómenos que no son cuantificables, pero que por su importancia merecen ser estudiados.

OBJETIVO GENERAL

Al finalizar el curso, el alumno será capaz de inferir las características de una población con base en la información contenida, así como de contrastar diversas pruebas para la toma de decisiones.

ESTRUCTURA CONCEPTUAL

UNIDAD 1

INTRODUCCIÓN AL MUESTREO

OBJETIVO ESPECÍFICO

Al terminar la unidad el alumno reconocerá los diferentes tipos de muestreo y sus características.

INTRODUCCIÓN

La teoría del muestreo es útil en numerosas ocasiones y en diferentes campos de la ciencia, sobre todo cuando no se cuenta con los recursos necesarios para hacer un censo (tiempo y dinero) o cuando no es necesario o recomendable hacer un estudio completo de toda la población de interés. Sin embargo, el no hacer el estudio completo, no significa de ninguna manera que el estudio no sea importante, pues extraer una muestra que sea representativa de una población y hacer inferencias que sean correctas de la población basándose en los datos arrojados por la muestra, es todo un proceso que debe ser cuidadosamente diseñado y elaborado; desde el objetivo del muestreo, tamaño de la muestra, técnica de muestreo a emplear, homogeneidad de la población, hasta las inferencias obtenidas al termino del estudio apoyadas en la teoría de la estimación.

Cabe aclarar que es imposible que una sola persona logre tal estudio completo y que una gran cantidad de expertos en diferentes campos se ve involucrada en tales estudios. Tales expertos incluyen no solo a los expertos en estadística, en mercados, en el giro mismo al que se esté dirigiendo el estudio, etc.

Todo esto hace que sea necesario poseer un conocimiento claro de lo que es la teoría del muestreo y la teoría de la estimación que estudiaremos en la presente unidad.

LO QUE SÉ

Selecciona si las siguientes aseveraciones son verdaderas (V) o falsas (F).

	Verdadera	**Falsa**
1. El siguiente es un axioma de probabilidad, **"**La probabilidad de un hecho existe y es restringida a la amplitud de cero a uno, inclusive. Es decir, si designamos la probabilidad de un hecho E como P (E), entonces: $0 \leq P(E) \leq 1$".	( )	( )
2. La siguiente es una propiedad de los logaritmos: $\log_a u^n = n \log_a u$	( )	( )

3. La siguiente expresión no es una propiedad de los logaritmos: $\log_a uv = \log_a u + \log_a v$	( )	( )
4. La teoría de conjuntos es un instrumento matemático muy útil para analizar un problema, permitiéndonos enfocar en él lo que es fundamental de lo que no lo es.	( )	( )
5. El sentido de una desigualdad debe ser invertido al multiplicar o dividir toda la desigualdad por un número negativo.	( )	( )
6. La derivada de una función es el límite del incremento de la función al incremento de la variable independiente cuando este último tiende a cero.	( )	( )
7. Una función matemática es una regla que asigna a cada elemento de un conjunto “A” uno y solo un elemento de un conjunto “B”.	( )	( )

TEMARIO DETALLADO

(4 horas)

1.1. Parámetros estadísticos y estimadores
1.2. Estimación de parámetros y pruebas de hipótesis
1.3. Muestreo aleatorio y muestreo de juicio
1.4. Muestras únicas y muestras múltiples
1.5. Muestras independientes y muestras relacionadas
1.6. Tipos de muestreo aleatorio

1.1. Parámetros, estadísticos y estimadores

La teoría del muestreo estudia la relación entre una población y las muestras tomadas de ella; es decir, se utiliza para estimar magnitudes desconocidas de una población —tales como valores promedio y de dispersión, llamadas a menudo parámetros de la población o simplemente parámetros— a partir del conocimiento de esas magnitudes sobre muestras, que se llaman estadísticos de la muestra o simplemente estadísticos.

1.2. Estimación de parámetros y pruebas de hipótesis

Desde un punto de vista práctico, es muy importante ser capaz de inferir información sobre una población a partir de muestras suyas. Con tal situación se enfrenta la inferencia estadística, que usa los principios de la teoría del muestreo.

Un problema importante de la inferencia estadística es la estimación de **parámetros** de la población, o brevemente parámetros (tales como la media o la varianza de la población), de los correspondientes estadísticos muestrales, o simplemente estadísticos (tales como la media y la varianza de la muestra).

- Método de máxima verosimilitud

En cualquier situación de muestreo es posible encontrar un estimador de un parámetro, utilizando el método de máxima verosimilitud de R. A. Fisher, el cual es un procedimiento general para la selección de estimadores.

Hay varias razones por las que se quiere utilizar un estimador de máxima verosimilitud para un parámetro; aunque dichos estimadores no siempre son eficientes e insesgados, por lo general son la mejor opción que se tiene debido a las siguientes propiedades:

- A medida que se incrementa el tamaño muestral, el sesgo del estimador de máxima verosimilitud tiende a cero.
- Su error estándar se aproxima al mínimo error estándar posible.
- Su distribución muestral se aproxima a la normal.

Debido a estas propiedades, muchos investigadores están a favor del uso de los estimadores de máxima verosimilitud en gran cantidad de situaciones de muestreo.

Pero veamos con más detalle cómo podemos encontrar un estimador de máxima verosimilitud; por lo tanto, empecemos por entender qué es la función de verosimilitud.

- Función de verosimilitud

Si denotamos a la función de verosimilitud con la letra “L” y la definimos como la probabilidad de observar los datos tomados de manera independiente de una variable aleatoria cualquiera, entonces dicha función de verosimilitud tendrá la forma siguiente:

$$L(y_1,y_2,\ldots,y_n, a) = P(y_1)P(y_2)\ldots P(y_n)$$

En el caso discreto y la siguiente forma en el caso continuo:

$$L(y_1,y_2,\ldots,y_n, a) = f(y_1)f(y_2)\ldots f(y_n)$$

Como podemos observar, independientemente de cual fuere el caso (variable aleatoria discreta o variable aleatoria continua), la función de verosimilitud se obtiene simplemente sustituyendo en la función original cada uno de los datos y multiplicando la función por sí misma para cada uno de los casos.

Por ejemplo supóngase que independientemente de lo que sucede el resto de los días, el número de trabajos que llegan en un día a un despacho contable tiene una distribución de Poisson con media desconocida. Supóngase además que el primer día de la muestra llega sólo un trabajo y que el segundo (y último) día llegan cuatro. Escribe la función de verosimilitud.

Para resolver este problema, la metodología es la siguiente:

- Primer paso

Debemos escribir la fórmula básica de la cual se parte y debemos identificar exhaustivamente todas sus variables; en este caso, la fórmula corresponde a una distribución de Poisson; por lo tanto, recordando que la distribución de Poisson es discreta con:

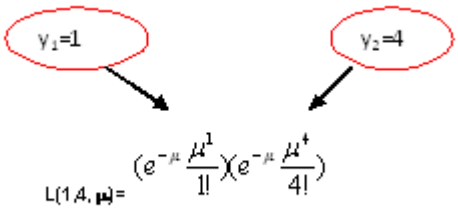
$$P(y) = e^{-\mu}\frac{\mu^{y}}{y!}$$

Donde: μ es el número esperado de eventos que suceden en un periodo y e= 2.71828.

- Segundo paso

Sustituir los valores o datos dados por el problema en la fórmula original, considerando la teoría de la función de verosimilitud. Los valores observados son y_1=1 e y_2=4; por lo tanto, la función de verosimilitud estará formada por el producto para cada uno de los datos de la fórmula misma.

Es decir:



- Tercer paso

Realizar las operaciones algebraicas correspondientes a la reducción de la fórmula, lo cual quiere decir que finalmente la fórmula anterior se puede reducir a:

$$L(1,4,\mu) = \frac{e^{-2\mu}\,\mu^{5}}{(1!)(4!)}$$

Éste es el último resultado de la función de verosimilitud solicitada en el problema.

A continuación, es necesario entender qué es una estimación de máxima verosimilitud.

Estimación máxima verosímil

Para valore observados en una muestra y_1, y_2,...y_n, la estimación máximo verosímil de un parámetro e es el valor ê que maximiza la función de verosimilitud L (y_1..y_2, e).

En un principio siempre es posible encontrar estimadores de máxima verosimilitud calculando numéricamente la función de verosimilitud. No obstante, utilizar el cálculo diferencial simplifica el trabajo de encontrar tales estimadores.

La idea básica (Kreyszig, 1990 [2], p. 959) del método de máxima verosimilitud es muy sencilla.

Se elige aquella aproximación para el valor desconocido que en este caso y para efectos de explicación llamaremos a de manera que “L” sea tan grande como sea posible.

Si “L” es una función diferenciable de a, una condición necesaria para que “L” tenga un máximo (no en la frontera) es:

Se escribe una derivada parcial debido a que “L” también depende de: y_1, y_2,...,y_n y una estimación de esta ecuación: $\frac{\partial L}{\partial A}=0$ que depende de y_1, y_2,...,y_n, se llama estimación de máxima verosimilitud para “a”.

Recordemos que para determinar el máximo de una función se iguala a cero la primera derivada y se resuelve la ecuación que de ello resulta.

En los problemas de máxima verosimilitud con frecuencia es más conveniente trabajar con el logaritmo natural de la verosimilitud que con la verosimilitud misma. Por lo tanto, podemos reemplazar la ecuación: $\frac{\partial L}{\partial A}=0$ por:

$$\frac{\partial \ln(L)}{\partial A}=0$$

Debido a que $f \geq 0$; un máximo de “f” en general es positivo y “ln (L)” es una función monótona creciente[1] de “L”. Esto a menudo simplifica los cálculos.

En principio se debería utilizar el criterio de la segunda derivada para asegurarse de que lo que se obtiene es un máximo y no un mínimo. No obstante, es muy claro que la solución de la ecuación correspondiente a la primera derivada produce un estimador de máxima verosimilitud y no un mínimo.

[1] En virtud de que el logaritmo natural es una función creciente, a medida que la verosimilitud se incrementa hacia su máximo, también lo hace su logaritmo.

Finalmente, si la distribución de "Y" contiene "r" parámetros: a_1, a_2,...,a_r, entonces en lugar de $\frac{\partial L}{\partial A} = 0$ se tiene las "r" condiciones:

$$\frac{\partial L}{\partial A_1} = 0 \quad , \quad \frac{\partial L}{\partial A_2} = 0 \quad ,\ldots, \quad \frac{\partial L}{\partial A_r} = 0$$

y en lugar de $\frac{\partial \ln(L)}{\partial A} = 0$ tenemos:

$$\frac{\partial \ln(L)}{\partial A_1} = 0 \quad , \quad \frac{\partial \ln(L)}{\partial A_2} = 0 \quad ,\ldots, \quad \frac{\partial \ln(L)}{\partial A_r} = 0$$

Por lo tanto, continuando con el ejemplo anterior, tenemos que la función de verosimilitud era:

$$L(1,4,\ \mu) = e^{-2\mu}\frac{\mu^5}{(1!)(4!)}$$

En donde el valor desconocido es en este caso μ

De modo que continuando con el proceso, el logaritmo natural de la verosimilitud es:

$$l(1,4,\ \mu) = \ln e^{-2\mu} + \ln\frac{\mu^5}{(1!)(4!)}$$

en donde por leyes de los logaritmos esta ecuación queda de la siguiente manera:

$$l(1,4,\mu) = -2\mu\,(\ln e) + \ln \mu^5 - \ln[(1!)(4!)]$$

Continuando con las leyes de los logaritmos, la expresión toma la forma siguiente:

$$l(1,4,\mu) = -2\mu + 5 \ln \mu - \ln [(1!)(4!)[$$

Posteriormente, al obtener la primera derivada a esta ecuación, esta cobra la siguiente forma:

$$\frac{dl(1,4,\mu)}{d\mu} = \frac{d}{d\mu}(-2\mu) + \frac{d}{d\mu}(5\ln \mu) - \frac{d}{d\mu}[\ln(1!)(4!)]$$

Si a la ecuación anterior le aplicamos las leyes de la derivación matemática, tenemos que esta expresión se convierte en:

$$\frac{dl(1,4,\mu)}{d\mu} = -2 + \frac{5}{\mu}$$

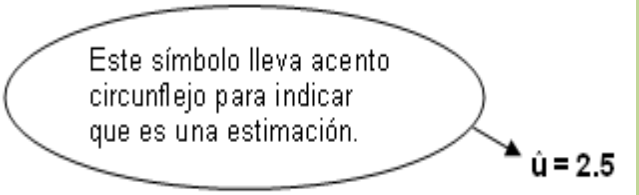
Continuando con el proceso, igualamos a “cero” esta primera derivada, por lo que la expresión resultante se indica a continuación:

$$\frac{dl(1,4,\mu)}{d\mu} = -2 + \frac{5}{\mu} = 0$$

que es lo mismo que:

$$-2 + \frac{5}{\mu} = 0$$

Resolviendo la última ecuación de primer grado con una incógnita tenemos que:



De modo que la estimación de máximo verosímil o de máxima verosimilitud de μ es $\hat{u}=2.5$.

En resumen, la metodología para encontrar una estimación de máximo verosímil es la siguiente:

Primer paso	Identificar la fórmula básica a que se refiere el problema junto con todas sus variables de manera exhaustiva.
Segundo paso	Encontrar la función de verosimilitud correspondiente (sustituyendo los datos dados en la fórmula original y considerando la teoría de la función de verosimilitud).
Tercer paso	Aplicar la función del logaritmo natural a la función de verosimilitud.
Cuarto paso	Realizar las operaciones propias de los logaritmos para desglosar la función en sumas y restas, dentro de las cuales es común que queden comprendidas multiplicaciones y divisiones.
Quinto paso	Aplicar la primera derivada a la función logaritmo natural.

Sexto paso	Realizar operaciones correspondientes a la teoría de derivación.
Séptimo paso	Igualar el resultado reducido de la primera derivada a cero.
Octavo paso	Resolver la ecuación de primer grado resultante, con lo cual obtenemos el resultado del estimador de máxima verosimilitud.

- Estimación por el método de momentos

Otra forma de hacer una estimación puntual de un parámetro es a través del llamado método de los momentos, el cual es otra metodología utilizada, en la cual, se igualan los momentos muestrales con los momentos poblacionales.

Si consideramos que el primer momento poblacional es E(X) (valor esperado de X), el segundo momento poblacional es $E(X^2)$ y así sucesivamente. Mientras que el primer momento muestral es $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ (el promedio de la muestra), el segundo momento muestral es $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ y así sucesivamente.

Considere el caso de una población cuya función densidad de probabilidad es $f_x(x)$ y parámetro desconocidoθ, como sigue:

$$f(x) = \begin{cases} (\theta+1)x^{\theta} & 0 \le x \le 1 \\ 0 & 0.C. \end{cases}$$

Si quisiéramos estimar el parámetro θ, entonces debemos calcular el primer momento poblacional e igualarlo con el primer momento muestral, a saber:

Estimar θ por el metodo de momentos.

$$E(x) = \int x f_x(x)\,dx$$

$$E(x) = \int_0^1 (\theta+1)x^{\theta}\,dx = \int_0^1 (\theta+1)x^{\theta+1}\,dx = \frac{(\theta+1)}{(\theta+2)} x^{\theta+2}\Big|_0^1 = \frac{\theta+1}{\theta+2}$$

Igualando el primer momento poblacional con el primer momento muestral, tenemos:

$$\frac{\theta+1}{\theta+2} = \frac{\sum X_1}{n} = \bar{x}$$

Y despejando θ, tenemos :

$$\hat{\theta}+1 = \bar{x}(\hat{\theta}+2)$$

es decir :

$$\hat{\theta}(1-\bar{x}) = 2\bar{x}-1$$

$$\hat{\theta} = \frac{2\bar{x}-1}{1-\bar{x}}$$ estimando puntual por momentos.

Así, si la variable estudiada X es el porcentaje de agrado de un producto y dicho porcentaje (de 0 a 100) se distribuye de acuerdo con la función de densidad $f_x(x)$ (que para asumir cierto modelo se puede utilizar una prueba de bondad de ajuste), entonces para estimar θ se determina una muestra aleatoria en la cual consideramos que arroja un promedio $\bar{x}=0.39$ (es decir 39% de satisfacción). Por lo cual en este caso el estimador de θ es: $\hat{\theta}=\frac{2\bar{x}-1}{1-\bar{x}}=\frac{2(0.39)-1}{1-0.39}=-0.36$, valor que no tiene significado práctico, pero que a partir del cual se describe el comportamiento de la población y en la cual el promedio es $E(X)=\frac{\theta+1}{\theta+2}=\frac{-0.36+1}{-0.36+2}=0.39$; asimismo se puede calcular la mediana, moda, varianza, entre otras características.

Resulta claro que siendo un estimador puntual, un estadístico tomado de una muestra que es utilizado para estimar un parámetro, dicho estimador es tan bueno como lo sea la muestra de la cual proviene, sin embargo, para diferentes muestras representativas de la misma población, se tendrán diferentes estimaciones puntuarles. Así las cosas, estimar un parámetro utilizando una estimación de intervalo (que veremos en el tema 3) resulta muchas veces preferible a utilizar una estimación puntual.

La teoría del muestreo es útil también para determinar si las diferencias observadas entre dos muestras son debidas a variaciones fortuitas o si son realmente significativas. Tales cuestiones aparecen, por ejemplo, al probar un nuevo suero como tratamiento de una enfermedad o al decidir si un proceso de producción es mejor que otro.

Las respuestas implican el uso de los llamados contrastes (o tests) de hipótesis y de significación, que son importantes en la teoría de las decisiones.

En general, un estudio de las inferencias hechas sobre una población a partir del análisis de diferentes muestras obtenidas de ésta, con indicación de la precisión de tales inferencias, se llama inferencia estadística.

La teoría de las probabilidades es el fundamento de los métodos de muestreo; para usarla hay que poseer un buen nivel de conocimiento, desde el punto de vista de la matemática, de álgebra, cálculo y probabilidades, así como de los métodos generales de estadística y de la teoría básica de las estimaciones, desde el punto de vista estadístico; todo ello es esencial para un entendimiento adecuado del desarrollo riguroso de la teoría del muestreo.

Así pues, “muestreo” es el proceso para obtener información acerca del conjunto de una población o universo examinando sólo una parte del mismo.

1.3. Muestreo aleatorio y muestreo de juicio

Existen básicamente dos métodos para seleccionar una muestra. Si cada elemento de una población tiene la misma posibilidad de ser seleccionado para integrar la muestra, el método se denomina **muestreo aleatorio**; por el contrario, si los elementos tienen diferentes posibilidades de ser elegidos, el método se denomina muestreo no aleatorio.

Cuando un muestreo se realiza devolviendo al conjunto el elemento una vez analizado se dice que el muestreo se realizó con reemplazo; si el elemento seleccionado no se regresa al conjunto, el muestreo es sin reemplazo. Esta condición resulta muy importante cuando se desea asignar un valor de probabilidad a la selección.

Su ventaja es que todos los datos tienen la misma posibilidad de ser seleccionados y en consecuencia podemos obtener información importante de la población de la cual fue extraída la muestra y, su desventaja es que si la población es heterogénea o que se encuentre agrupada en segmentos de diferentes tamaños, entonces la muestra puede no ser representativa de la población.

Debido a que si uno de los segmentos de la población es muy pequeño entonces cabe la posibilidad de que ninguno de sus elementos pueda ser incluido en la muestra y en consecuencia no ser tomado en cuenta.

Muestreo de juicio o no probabilístico. Una muestra es llamada muestra de juicio cuando sus elementos son seleccionados mediante juicio personal. La persona que selecciona los elementos de la muestra, usualmente es un experto en la medida dada, es decir el investigador con su experiencia designa cuáles elementos forman parte de la muestra, sin embargo, debe evitarse, ya que no puede hacerse ninguna afirmación probabilística o inferencia válida si la muestra se eligió usando este tipo de muestreo.

1.4. Muestras únicas y muestras múltiples

En el muestreo a estadios múltiples se subdivide la población en varios niveles ordenados que se extraen sucesivamente por medio de un procedimiento de embudo. El muestreo se desarrolla en varias fases o extracciones sucesivas para cada nivel.

Por ejemplo, si tenemos que construir una muestra de profesores de primaria en un país determinado, estos pueden subdividirse en unidades primarias representadas por circunscripciones didácticas y unidades secundarias que serían los propios profesores.

En primer lugar extraemos una muestra de las unidades primarias (para lo cual debemos tener la lista completa de estas unidades) y en segundo lugar extraemos aleatoriamente una muestra de unidades secundarias de cada una de las primarias seleccionadas en la primera extracción.

1.5. Muestras independientes y muestras relacionadas

Los contrastes permiten comprobar si hay diferencias entre las distribuciones de dos poblaciones a partir de dos muestras dependientes o relacionadas; es decir, tales que cada elemento de una muestra está emparejado con un elemento de la otra, de tal forma que los componentes de cada pareja se parezcan entre sí lo más posible por lo que hace referencia a un conjunto de características que se consideran relevantes. También es posible que cada elemento de una muestra actúe como su propio control.

Algunas de las pruebas que pueden realizarse con el programa SPSS son: la prueba de Wilcoxon, la de signos y la de McNemar. (Alea, Guillén, Muñoz, Torrelles y Viladomiu, 2000, p. 117)

1.6. Tipos de muestreo aleatorio

Muestreo aleatorio sistemático

Aclaremos esto observando que el procedimiento en este tipo de muestreo: se acomodan los elementos o personas de la población de forma ascendente de preferencia y se selecciona un punto de partida aleatorio y luego se toma cada k-esimo miembro para formar la muestra.

Del muestreo aleatorio simple puede ser difícil en ciertos casos. Por ejemplo, suponga que la población que nos interesa consiste de 2000 facturas que se localizan en cajones. Tomar una muestra aleatoria sencilla requeriría primero numerar las facturas, del 0001 al 1999; posteriormente, se seleccionaría luego una muestra de, por ejemplo, 100 números utilizando una tabla de números aleatorios; luego, en los cajones deberá localizarse una factura que concuerde con cada uno de estos 100 números; en fin, esta tarea puede requerir mucho tiempo. En lugar de ello, se podría seleccionar una muestra aleatoria sistemática utilizando el siguiente método: se recorren simplemente los cajones y se cuentan las facturas; finalmente, se toman las que coincidan con el número 20 para su estudio. Así, la primera factura debería elegirse utilizando un proceso aleatorio, por ejemplo, una tabla de números aleatorios. Si se eligió la décima factura como punto de partida, la muestra consistiría en las facturas décima, trigésima, quincuagésima, septuagésima, etcétera.

Debido a que el primer número se elige al azar, todos tienen la misma probabilidad de seleccionarse para la muestra. Por lo tanto, se trata de un muestreo cuasi-aleatorio. La ventaja para este tipo de muestreo sería que es más rápido que un muestreo aleatorio formal y su desventaja es que puede no reflejar información importante contenida en el conjunto de datos debido a que no todos los elementos estrictamente hablados, tienen la misma oportunidad de ser seleccionados.

Muestreo aleatorio estratificado

Otro tipo de muestreo es el aleatorio estratificado (Lind, Marchal & Mason, 2004, p. 226): divide una población en subgrupos llamados estratos y se selecciona una muestra de cada uno de ellos con lo cual se garantiza la representación de cada subgrupo o estrato.

Una vez que la población se divide en estratos, es posible seleccionar una muestra proporcional o no proporcional. Como el nombre señala, un procedimiento de muestreo proporcional requiere que el número de artículos de cada estrato esté en la misma proporción que en la población.

Ejemplo

Los gastos en mercadotecnia de las 352 empresas mexicanas más grandes seleccionadas por la revista *Fortune*. Supóngase que el objetivo de estudio consiste en determinar si las empresas con altos rendimientos sobre su inversión (una medición de la rentabilidad) han gastado una mayor proporción de su presupuesto de ventas en mercadotecnia que las empresas que tienen un menor rendimiento o incluso un déficit.

Supóngase que las 352 empresas se dividieron en cinco estratos; si seleccionamos una muestra de 50 empresas, entonces de acuerdo con el muestreo aleatorio estratificado se deberían incluir:

Estrato	Rentabilidad	# empresas	# muestreado	?
1	30% y más	8	1	(8/352)(50)
2	De 20 a 30%	35	5	(35/352)(50)
3	De 10 a 20%	189	27	(189/352)(50)
4	De 0 a 10%	115	16	(115/352)(50)
5	Déficit	5	1	(5/352)(50)
	Total	352	50	

En la quinta columna de la tabla anterior, podemos observar los cálculos realizados para determinar el número de elementos muestreados por estrato, garantizando con este procedimiento, que cada uno de los estratos de interés se encuentra representado en la muestra por estudiar.

Una muestra estratificada no proporcional es aquella en la cual, la cantidad de elementos que se seleccionan en cada estrato no guarda proporción con la cantidad de elementos respectivos en la población.

En algunos casos, el muestreo estratificado tiene la ventaja de poder reflejar con mayor precisión las características de la población que un muestreo aleatorio simple o sistemático, dado que puede darse el caso en ambos muestreos (aleatorio simple o sistemático), de que alguno de los estratos de interés no quede considerado en la muestra al no ser elegido al menos alguno de sus elementos y la desventaja para este tipo de muestreo estratificado es que puede caerse en el exceso de estratos haciendo el proceso de muestreo más difícil y tardado que si aplicamos un muestreo aleatorio simple.

Muestreo por conglomerados

Otro tipo de muestreo que es común es el muestreo por conglomerados. Se entiende como conglomerado de elementos de una población, a cualquier subconjunto de la misma, que se defina como tal, es decir, como un conglomerado. (Lind, Marchal & Mason, 2004, p. 227)

La definición de un conglomerado, así como su tamaño, se definen y dependen de los objetivos del estudio que se esté realizando, y en general, los conglomerados definidos en un estudio pueden o no tener el mismo tamaño (véase, Flores, 1998, p. 225).

Muchas veces se le emplea para reducir el costo de realizar un muestreo de una población dispersa en una gran área geográfica. Supóngase que se desea determinar el punto de vista de los industriales de toda la República Mexicana con respecto a las reformas fiscales del año 2004. La selección de una muestra aleatoria de los industriales de toda la República Mexicana y el contacto personal con cada uno de ellos serían muy onerosos en cuanto a tiempo y dinero.

En lugar de ello, se podría emplear un muestreo por conglomerados subdividiendo la República Mexicana en unidades pequeñas, ya fueran estados o regiones.

Muchas veces, éstas se conocen como unidades primarias. Supóngase que se subdividió a la República Mexicana en 12 unidades primarias y luego se escogió a cuatro de ellas; de esta forma, los esfuerzos se concentran en estas cuatro unidades, tomando una muestra aleatoria de los industriales de cada una de estas regiones y entrevistarlos (obsérvese que se trata de una combinación del muestreo por conglomerados y el muestreo aleatorio simple).

Tamaño de la muestra

Para la determinación del tamaño de la muestra se requiere tomar en consideración la mayor cantidad posible de los siguientes elementos.

1. Tamaño del universo.
2. Tasa de error esperada.
3. Homogeneidad-heterogeneidad del fenómeno.
4. Precisión o margen de error.
5. Exactitud o nivel de confianza.
6. Número de estratos.
7. Etapas de muestreo.
8. Conglomeración de unidades.
9. Estado del marco muestral.
10. Efectividad de la muestra.
11. Técnica de recolección de datos.
12. Recursos disponibles (véase, Galindo, 1998, pp. 49-62)

Fórmula genérica

Dependiendo del problema mismo, no todos los problemas incluyen la totalidad de los elementos mencionados.

Como es de observarse, dentro de las teorías del muestreo y probabilidad existen diversos procedimientos para el cálculo de los tamaños de la muestra; todos ellos consideran a la mayoría de los elementos que hemos enumerado.

La fórmula utilizada es la siguiente:

$$n = \frac{NPQ}{\left[\frac{Me^2}{Nc^2}(N-1)\right] + PQ}$$

Variables

Las variables que considera la fórmula son los siguientes:

Variable	Descripción
N	Tamaño de la muestra
N	Tamaño del universo
P	Probabilidad de ocurrencia (homogeneidad del fenómeno)
Q	Probabilidad de no ocurrencia (1-p)
Me	Margen de error o precisión. Expresado como probabilidad.
Nc	Nivel de confianza o exactitud. Expresado como valor z que determina el área de probabilidad buscada.

Ejemplo

Se requiere calcular el tamaño de una muestra para el siguiente caso:

Variable	Descripción
N	?
N	3,000,000
P	Desconocemos la probabilidad de ocurrencia. Por esta razón asumimos el mayor punto de incertidumbre, que es de 50%, que al ser expresada como probabilidad queda como: 0.5
Q	1 – 0.5 = 0.5
Me	+/- 5% de margen de error. Que expresado como probabilidad queda como: 0.05
Nc	95% de nivel de confianza o exactitud. Que expresado como valor “z” que determina el área de probabilidad buscada queda como: 1.96

Al sustituir estos valores en la fórmula, tenemos:

$$n = \frac{(3{,}000{,}000)(0.5)(0.5)}{\left[\frac{(0.05)^2}{(1.96)^2}(3{,}000{,}000-1)\right]+(0.5)(0.5)}$$

De donde, al realizar las operaciones indicadas, el valor de “n” obtenido es de 384.1. Finalmente, haciendo un redondeo, el tamaño de la muestra será de 384 elementos.

El valor de “z” se busca en las tablas de distribución normal estándar y la forma de encontrarlo es la siguiente:

1. El porcentaje deseado entre 2 (debido a la simetría de la curva de distribución normal), en este caso el resultado sería:

$$\frac{95}{2} = 47.5$$

2. Este resultado (47.5) se divide entre 100 para convertirlo de porcentaje a decimal, es decir:

$$\frac{47.5}{100} = 0.475$$

3. Este valor de 0.475 se busca en el cuerpo de la tabla de la curva de distribución normal estándar (La mayoría de los textos de probabilidad y estadística contienen esta tabla), donde encontramos el valor correspondiente de z = 1.96.

RESUMEN DE LA UNIDAD

Como pudimos observar, las técnicas de muestreo son variadas y su aplicación depende del estado de la población (homogeneidad-heterogeneidad), sin embargo la metodología de aplicación del proceso de muestrear es mucho más completa, pues tiene que cuidar de numerosos detalles tales como el objetivo mismo del muestreo, el tamaño de la muestra, el nivel de confianza, etc.

El apoyo que brinda la teoría de la estimación es muy importante para poder obtener inferencias correctas de la población y en consecuencia, las personas que deban tomar las decisiones correspondientes puedan hacer su trabajo de manera eficiente teniendo como sustento de tales decisiones herramientas estadísticas poderosas tales como la Teoría del muestreo y la Teoría de la estimación.

GLOSARIO DE LA UNIDAD

Aleatorio

Suceso incierto que tiene algún grado de inseguridad de ocurrir (también es llamado estocástico).

Censo

Es el estudio en el que se incluye a toda la población.

Cuestionario

Instrumento recolector autoadministrable. En él, el cuestionado lee y contesta por sí mismo las preguntas.

Desviación estándar

Raíz cuadrada de la suma de los cuadrados de las desviaciones de cada valor que asume la variable en relación a la media. Raíz cuadrada de la varianza para la muestra “s” para la población (sigma).

Distribución normal

Estudia la concentración de probabilidad en un intervalo cualquiera, que está contenido en el área bajo la curva de una función de probabilidades en forma de campana.

Distribución normal estandarizada

Estandariza las probabilidades de la distribución normal.

Entrevista

Instrumento recolector empleado en una conversación a niveles profundos o específicos. Puede ser libre o estructurada.

Error sistemático

Error de respuesta o de encuesta que se produce constantemente a lo largo de la investigación.

Estadística

Es una ciencia relativamente nueva que tiene por objeto la colección e interpretación de datos.

Estadística inferencial

Estimación de las características de una población, validación de distribuciones o la toma de decisiones sobre algún factor de la población, sin conocerla enteramente y basándose en los resultados de un muestreo, que se manifiestan en la estadística descriptiva de ese conjunto de datos.

Muestra

Es un conjunto de “n” observaciones extraídas de entre los “N” elementos de la población.

Muestreo a juicio

Es la selección de “n” elementos de entre los “N” de una población elegida según el criterio del sujeto que los elige. Se basa en suposiciones muy amplias acerca de las variables que se van a estudiar en la población. Generalmente lo realizan expertos en la materia.

Muestreo aleatorio simple

Requiere de un marco muestral aleatorizado o no, en el que estén contenidos sin repetición todas las unidades de la población.

Parámetro

Medida que caracteriza a una población.

ACTIVIDADES DE APRENDIZAJE

ACTIVIDAD 1

Elabora un cuadro comparativo del muestreo por conglomerados y del muestreo estratificado.

ACTIVIDAD 2

Forma un equipo de cuatro integrantes y consulten la página de *Food and Agriculture Organization of the United Nations* www.fao.org escribe en el buscador "muestreo" y revisa cada uno de los apartados desarrollados en los artículos.

Comenta con tu equipo tus hallazgos.

CUESTIONARIO DE REFORZAMIENTO

1. ¿Qué es la teoría del muestreo?
2. ¿En qué situaciones es conveniente recurrir al muestreo?
3. ¿Cuáles son los aportes de la teoría del muestreo?
4. ¿Qué es un muestreo aleatorio simple?
5. ¿Para qué se utiliza la teoría del muestreo?
6. ¿Qué es un muestreo aleatorio sistemático?
7. ¿Qué es un muestreo aleatorio estratificado?
8. ¿Qué es un muestreo por conglomerados?
9. ¿Qué es el nivel de confianza?
10. ¿Qué es el error de muestreo?

EXAMEN DE AUTOEVALUACIÓN 1

Elige la respuesta correcta a las siguientes preguntas, una vez que concluyas, obtendrás de manera automática tu calificación.

1. A los valores numéricos obtenidos del análisis estadístico descriptivo de una muestra se les denomina:
    a) población
    b) parámetros
    c) estadísticos
    d) sesgo
    e) desviación estándar

2. Cuando se selecciona una muestra con el fin de realizar un análisis estadístico debe cuidarse que los elementos:
    a) tengan características similares entre sí
    b) se encuentren dentro del mismo lote
    c) sean seleccionados de manera aleatoria
    d) sean lo más parecidos a la población
    e) estén lo más alejados del centro de la población

3. Al proceso mediante el cual se obtienen los elementos de una muestra representativa de la población se le denomina:
    a) proceso estadístico
    b) procedimiento de muestreo
    c) proceso de selección
    d) muestreo aleatorio
    e) seccionamiento

4. Al obtener una muestra se debe asegurar que durante el proceso todos los elementos:
    a) resulten del mismo tipo
    b) resulten como deseamos
    c) se encuentren del intervalo seleccionado
    d) resulten sin defectos
    e) tengan la misma probabilidad de ser escogidos

5. Una técnica para muestrear, en la cual se asegura la no intervención de la mano del hombre, es:
    a) el uso de un dado
    b) una moneda
    c) una tabla de números aleatorios
    d) el criterio del analista a cargo
    e) el criterio del cliente

6. Una población finita en la que se realiza un muestreo con reemplazamiento puede ser considerada como:
    a) modelo
    b) infinita
    c) muestra

d) acotada
e) estratificada

7. El muestreo realizado mediante la aplicación de un criterio personal de preferencia o aversión hacia determinados elementos constituye un método:
   a) probabilístico
   b) aleatorio simple
   c) aleatorio directo
   d) de conglomerados
   e) no probabilístico

8. Supóngase que hay un inventario con 15 diferentes líneas de producto. Si para efectuar un muestreo tomamos una sola línea de producto se dice que el muestreo fue:
   a) probabilístico
   b) por conglomerados
   c) aleatorio simple
   d) aleatorio sistemático
   e) aleatorio subjetivo

9. Se denomina así a la diferencia entre un estadístico y su parámetro poblacional correspondiente:
   a) media poblacional
   b) proporción
   c) error de muestreo
   d) parámetro poblacional
   e) sesgo

10. Un auditor va a realizar una prueba donde espera una tasa de error no mayor al 5%. Si fija una precisión de ± 3% y un nivel de confianza de 95% en una población de 15 000 facturas, si la prueba se realizara en el mes de marzo y si la última factura del mes de febrero es la No. 28 974, el tamaño de la muestra es de:
    a) 15 000
    b) 375
    c) 7 500
    d) 28 974
    e) 1 500

EXAMEN DE AUTOEVALUACIÓN 2

Selecciona si las siguientes aseveraciones son verdaderas (V) o falsas (F).

	Verdadera	Falsa
1. En un muestro aleatorio cada elemento de una población tiene la misma posibilidad de ser seleccionado para integrar la muestra.	( )	( )
2. En un muestreo no aleatorio los elementos tienen diferentes posibilidades de ser elegidos para integrar la muestra.	( )	( )

3. El muestreo por conglomerados consiste en dividir una población en subgrupos llamados estratos y se selecciona una muestra de cada uno de ellos con lo cual se garantiza la representación de cada subgrupo o estrato en la muestra final.	( )	( )
4. El muestreo estratificado muchas veces se emplea para reducir el costo de realizar un muestreo de una población dispersa en una gran área geográfica.	( )	( )
5. El error de muestreo es la diferencia que se presenta entre los resultados obtenidos en el análisis de las muestras respecto de los que en realidad corresponden a la población.	( )	( )
6. El error de muestreo se presenta con mayor intensidad cuando las muestras no son representativas de la población de la cual fueron extraídas.	( )	( )
7. El error de muestreo se presenta de forma azarosa y no hay forma de evitarlo, calcularlo o minimizarlo.	( )	( )

LO QUE APRENDÍ

Considera una distribución binomial con n=5, y y=2. Encuentra la estimación de máxima verosimilitud correspondiente.

MESOGRAFÍA

Bibliografía sugerida

Autor	Capítulo	Páginas
Berenson y otros (2001)	7	319-342
Levin (1996)	6	236-246
Lind y otros (2004)	8	261-270
Christensen (1990)	1-7	26-316

Bibliografía básica

Berenson L. Mark; Levine M. David; Krehbiel C. Timothy. (2001). *Estadística para Administración*. (2ª ed.) México: Prentice Hall.

Kreyszig, Erwin. (1990). *Matemáticas avanzadas para ingeniería*. (vol. 2) México: Limusa.

Levin, Richard I. y Rubin, David S. (1996). *Estadística para administradores*. México: Alfaomega.

Lind A. Douglas, Marchal G. William, Mason D. Robert. (2004). *Estadística para Administración y Economía*. (11ª ed.) Madrid: Alfaomega.

Bibliografía complementaria

Alea Riera, Ma. Victòria; Guillén, Montserrat; Muñoz, Carmen; Torrells, Elizabeth; Viladomiu, Núria. (2000). *Estadística con SPSS v.10.0.* Barcelona: Universitat de Barcelona (Textos Docents 226) [vista previa]

Ato, Manuel y López, Juan J. (1996). *Fundamentos de estadística con SYSTAT*. México: Addison Wesley Iberoamericana.

Christensen, H. (1990). *Estadística paso a paso*. (2ª ed.) México: Trillas.

Flores García, Rosalía. (1998). *Estadística aplicada para administración*. México: Iberoamericana.

Galindo Cáceres, Luis Jesús. (1998). *Técnicas de investigación en sociedad, cultura y comunicación*. México: Pearson.

Garza, Tomás. (1996). *Probabilidad y estadística*. México: Iberoamericana.

Hanke, Jonh E. y Reitsch, Arthur G. (1997). *Estadística para Negocios.* México: Prentince Hall.

Sitios de Internet

Sitio	Descripción
http://ocw.upm.es/estadistica-e-investigacion-operativa/matematicas-y-estadistica-aplicada/contenidos/OCW/Tecnicas-de-muestreo/Mat_Clase/tec_muestreo.pdf	Martín Fernández, Susana y Ayuga Téllez, Esperanza. (2008). Introducción al muestreo. Ciencias Ambientales, UPM
http://aulasvirtuales.wordpress.com/2010/04/30/introduccion-al-muestreo	Rodríguez, Manuel Luis. (2010). "Introducción al muestreo", (30/04/10), Aulas Virtuales [blog]
http://www.itch.edu.mx/academic/industrial/estadistica1/cap01.html	Torre, Leticia de la. (2003). "Teoría del Muestreo", Estadística I, Instituto Tecnológico de Chihuahua
http://www.eumed.net/libros/2006c/203/2l.htm	Ávila Baray, Héctor Luis. (2006). "Introducción a la Teoría del Muestreo", *Introducción a la metodología de la investigación*.
http://www.ub.edu/aplica_infor/spss/cap6-3.htm	Alea, V. "Pruebas para dos muestras relacionadas", *SPSS Análisis de datos,* Estadística, Universidad de Barcelona

UNIDAD 2

DISTRIBUCIONES MUESTRALES

OBJETIVO ESPECÍFICO

Al terminar la unidad el alumno identificará e interpretará los diferentes tipos de distribuciones muestrales.

INTRODUCCIÓN

La distribución de la población de la cual extraemos la muestra con la que trabajamos en estadística es importante para saber qué tipo de distribución debemos aplicar en cada una de las situaciones que se nos presenten en la práctica; en esta unidad veremos algunas de estas distribuciones que se encuentran relacionadas con la distribución normal, además de observar la distribución muestral para la media y para la proporción y su relación con el teorema central del límite.

LO QUE SÉ

Elige la respuesta correcta a las siguientes preguntas:

1. La distribución chi-cuadrada χ^2 es útil para analizar la relación:

   a) entre la varianza de la muestra y la varianza de la población

   b) entre la media de la muestra y la media de la población

   c) entre una muestra y otra

2. La fórmula para calcular la media aritmética de una muestra es:

   a) $\chi^2 = \frac{s^2(gl)}{\sigma^2}$

   b) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

   c) $\frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}$

3. La fórmula para calcular la varianza de una muestra es:

   a) $\frac{s^2(n-1)}{\chi^2_{\alpha/2}}$

   b) $\frac{s^2(n-1)}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}$

   c) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

4. La distribución “t” de Student se utiliza cuando:

a) El investigador lo decide
b) cuando la desviación estándar de la población es desconocida
c) cuando no hay otra alternativa

5. La distribución “F” se utiliza para:

a) analizar la relación entre las varianzas de dos muestras extraídas de la misma población.
b) Analizar la relación entre la varianza de la muestra y la varianza de la población
c) Calcular la desviación estándar

6. La fórmula para calcular la desviación estándar de una población es:

a) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

b) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

c) $\sigma = \sqrt{\frac{1}{N}\sum_{1}^{N}(x_i - \mu)^2}$

7. La fórmula correcta para el cálculo de combinaciones es:

a) ${}_nP_r = \frac{n!}{(n-r)!}$

b) ${}_nC_r = \frac{n!}{r!(n-r)!}$

c) $F_{(X)} = \binom{n}{x} P^x (1-P)^{n-x}$

8. Las combinaciones se utilizan cuando:

a) no importa el orden

b) si importa el orden

c) no hay otra opción

9. La simetría es una característica de la distribución:

a) chi-cuadrada χ^2

b) F

c) Normal

TEMARIO DETALLADO

(8 horas)

2.1. La distribución muestral de la media

2.2. El teorema central del límite

2.3. La distribución muestral de la proporción

2.4. La distribución muestral de la varianza

2.1. La distribución muestral de la media

El estudio de determinadas características de una población se efectúa a través de diversas muestras que pueden extraerse de ella.

El muestreo puede hacerse con o sin reposición, y la población de partida puede ser infinita o finita. Una población finita en la que se efectúa muestreo con reposición puede considerarse infinita teóricamente. También, a efectos prácticos, una población muy grande puede considerarse como infinita. En todo nuestro estudio vamos a limitarnos a una población de partida infinita o a muestreo con reposición.

Consideremos todas las posibles muestras de tamaño n en una población. Para cada muestra podemos calcular un estadístico (media, desviación típica, proporción,...) que variará de una a otra. Así obtenemos una distribución del estadístico que se llama distribución muestral.

Las dos medidas fundamentales de esta distribución son la media y la desviación típica, también denominada error típico.

Hay que hacer notar que si el tamaño de la muestra es lo suficientemente grande, las distribuciones muestrales son normales y en esto se basarán todos los resultados que alcancemos.

Distribución muestral de medias

Cada muestra de tamaño n que podemos extraer de una población proporciona una media. Si consideramos cada una de estas medias como valores de una variable aleatoria podemos estudiar su distribución que llamaremos distribución muestral de medias.

Si tenemos una población normal N (μ,n) y extraemos de ella muestras de tamaño n, la distribución muestral de medias sigue también una distribución normal

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

Si la población no sigue una distribución normal pero n>30, aplicando el llamado Teorema central del límite la distribución muestral de medias se aproxima también a la normal anterior.

2.2. El teorema central del límite

El enunciado formal del teorema del límite central es el siguiente: si en cualquier población se seleccionan muestras de un tamaño específico, la distribución muestral de las medias de muestras es aproximadamente una distribución normal. Esta aproximación mejora con muestras de mayor tamaño.

Ésta es una de las conclusiones más útiles en estadística pues nos permite razonar sobre la distribución muestral de las medias de muestras sin contar con información alguna sobre la forma de la distribución original de la que se toma la muestra. En otras palabras, de acuerdo con el teorema del límite central, es válido aproximar la distribución de probabilidad normal a cualquier distribución de valores medios muestrales, siempre y cuando se trate de una muestra suficientemente grande.

El teorema central del límite o teorema del límite central se aplica a la distribución muestral de las medias de muestras que veremos a continuación y permite utilizar la distribución de probabilidad normal para crear intervalos de confianza para la media de la población.

2.3. La distribución muestral de la proporción

Hoy es bien sabido que si la investigación produce datos mensurables tales como el peso, distancia, tiempo e ingreso, la media muestral es en ocasiones el estadístico más utilizado, pero, si la investigación resulta en artículos “contables” como por ejemplo: cuántas personas de una muestra escogen la marca “Peñafiel” como su refresco, o cuántas personas de una muestra tienen un horario flexible de trabajo, utilizar la proporción muestral es generalmente lo mejor.

Mientras que la media se calcula al promediar un conjunto de valores, la “proporción muestral” se calcula al dividir la frecuencia con la cual una característica dada se presenta en una muestra entre el número de elementos de la muestra. Es decir:

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

Donde: x = número de elementos de una muestra que tienen la característica.

n = número de elementos de la muestra.

Ejemplo; supóngase que una comercializadora pretende establecer un nuevo centro y desea saber la proporción del consumidor potencial que compraría el principal producto que vende para lo cual realiza un estudio de mercado mediante una encuesta a 30 participantes, lo cual permitirá saber quiénes lo comprarían y quiénes no; se obtuvieron los siguientes resultados:

*x*1=1	*x*7=1	*x*13=0	*x*19=1	*x*25=0
*x*2=0	*x*8=0	*x*14=1	*x*20=0	*x*26=0
*x*3=0	*x*9=0	*x*15=1	*x*21=1	*x*27=0
*x*4=0	*x*10=0	*x*16=0	*x*22=1	*x*28=1
*x*5=0	*x*11=0	*x*17=0	*x*23=1	*x*29=0
*x*6=1	*x*12=0	*x*18=1	*x*24=0	*x*30=1

Donde “1” significa que está dispuesto a comprar el producto y “0” no está dispuesto a comprarlo.

En este caso, la proporción de la población (P) que compraría el producto, se puede estimar con $\bar{p}$ (proporción de la muestra que lo compraría), cuyo valor esperado sería $E(\bar{p}) = P$, y el error de $\bar{p}$ al estimar P es:

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}}$$

Si la población es finita, y si la población es infinita o si el muestreo es con reposición, los resultados anteriores se reducen a:

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{P(1-P)}{n}}$$

Es decir, de acuerdo con el teorema del límite central, $\bar{p}$ muestral se comportará como una normal con media P (la verdadera proporción poblacional) y desviación estándar $\sigma_{\bar{p}}$.

En el ejemplo de la comercializadora se tiene que $\bar{p} = \frac{12}{30} = 0.40$.

Pero suponiendo que el verdadero parámetro de la población es P=0.30; es decir, sólo el 30% de la población lo compraría, entonces el promedio $\bar{p}$ estimará a P poblacional pero con un error igual a $\sigma_{\bar{p}}$ que en este caso es:

$$\sigma_{\bar{p}} = \sqrt{\frac{0.30(0.70)}{30}} = 0.1195$$

En este caso $\bar{p}$ muestral tendrá distribución normal con media P=0.30 y desviación estándar $\sigma_{\bar{p}} = 0.1195$.

Dado que todas las muestras aleatorias que sean tomadas de una misma población en general serán distintas y tendrán por ende diferentes valores para sus estadísticos tales como la media aritmética o la desviación estándar, entonces resulta importante estudiar la distribución de todos los valores posibles de un estadístico, lo cual significa estudiar las distribuciones muestrales para diferentes estadísticos (véase, Weimer, 1996, p. 353). La importancia de éstas distribuciones muestrales radica en el hecho de que en estadística inferencial, las inferencias sobre poblaciones se hacen utilizando estadísticas muestrales pues con el análisis de las distribuciones asociadas con éstos estadísticos se da la confiabilidad del estadístico muestral como instrumento para hacer inferencias sobre un parámetro poblacional desconocido.

2.4. La distribución muestral de la varianza

La varianza de las muestras sigue un proceso distinto a los de la media y proporción. La causa es que el promedio de todas las varianzas de las muestras no coincide con la varianza de la población s2. Se queda un poco por debajo.

Comúnmente se utiliza el subíndice $_n$ para recordar que en la varianza se divide entre n. Si deseamos que la media de la varianza coincida con la varianza de la población, tenemos que acudir a la cuasivarianza o varianza insesgada, que es similar a la varianza, pero dividiendo las sumas de cuadrados entre n-1.

Su raíz cuadrada es la cuasidesviación típica o desviación estándar. Si se usa esta varianza, si coinciden su media y la varianza de la población lo que nos indica que la cuasivarianza es un estimador insesgado, y la varianza lo es sesgado.

RESUMEN DE LA UNIDAD

El teorema central del límite es útil para entender que la distribución de las medias de muestras tomadas de una misma población y del mismo tamaño es aproximadamente normal y que esta aproximación mejora a medida que se incrementa el tamaño de la muestra; dando pie al estudio de la distribución muestral para la media y para la proporción y a la elaboración de “intervalos de confianza” que se analizarán en el apartado 3.4., la proporción muestral es el mejor estadístico por utilizar cuando en la investigación se trata de averiguar cuestiones tales como: ¿Cuántos integrantes de la población tienen una característica en particular o una tendencia similar?

Con todo lo analizado hasta aquí, podemos ir observando que la estadística nos ofrece la oportunidad de analizar el comportamiento de una población utilizando diferentes herramientas tales como las distribuciones relacionadas con la normal entre otras, a demás de diferentes teorías tales como la del muestreo y la de la estimación estadística, con lo cual, los tomadores de decisiones pueden aunar estos conocimientos a su experiencia en el medio en el que se estén desenvolviendo y en consecuencia tomar decisiones más certeras que cada vez más necesarias en un mundo globalizado como el nuestro.

GLOSARIO DE LA UNIDAD

Distribución muestral

Es una distribución de probabilidades que consta de todos los valores posibles de un estadístico de muestra.

Error estándar

Es la desviación estándar de un estimador puntual.

Factor de corrección para población finita

El término $\sqrt{\frac{N-n}{N-1}}$ que se usa en las fórmulas de $\sigma_{\bar{x}}$ y $\sigma_{\bar{p}}$ cuando se selecciona una muestra de una población finita, no de una población infinita. La regla fácil que generalmente se acepta es no tomar en cuenta el factor de corrección para población finita siempre que $\frac{n}{N} \leq 0.05$

Muestras pareadas

Muestras en las que con cada dato de una muestra se forman parejas con el dato correspondiente.

Parámetro

Es una característica numérica de una población, tal como la media aritmética poblacional, la desviación estándar poblacional o la proporción poblacional.

Teorema del límite central

También conocido como teorema central del límite, es un teorema que permite usar la distribución de probabilidad normal para aproximar la distribución de muestra de $\bar{x}$ y $\bar{p}$ cuando el tamaño de la muestra es grande.

ACTIVIDADES DE APRENDIZAJE

ACTIVIDAD 1

Para una proporción poblacional de 0.25 ¿Cuál es la probabilidad de obtener una proporción muestral menor o igual a 0.21 para n = 120?

ACTIVIDAD 2

Supóngase una proporción poblacional de 0.58 y que una muestra aleatoria de 410 artículos se muestrea al azar. ¿Cuál será la probabilidad de que la proporción muestral sea mayor a 0.70?

CUESTIONARIO DE REFORZAMIENTO

1. ¿Qué es una distribución de muestreo?
2. Si el estadístico utilizado es la media muestral, ¿qué nombre recibe la distribución de este estadístico?
3. ¿Qué es la distribución muestral de las medias de las muestras?
4. ¿Qué relación existe entre la media de las medias de la muestra y la media de la población?
5. ¿Cómo es la dispersión de las medias de la muestra en comparación con la de los valores de la población?
6. ¿Cómo es la forma de la distribución muestral de las medias de muestras y la forma de la distribución de frecuencia de los valores de la población?
7. ¿Cómo es la desviación estándar de las medias de las muestras comparada con la desviación estándar de la población?
8. Para una población infinita ¿qué implicación tiene el hecho de que la distribución de muestreo sea asintóticamente normal?
9. ¿Cómo es la distribución de muestreo de medias cuando la población de origen está normalmente distribuida?
10. En una empresa se tienen 4 puestos de gerente nivel C disponibles y 7 candidatos que pueden ocupar esos puestos, ¿de cuántas formas podemos tomar la decisión correspondiente?

EXAMEN DE AUTOEVALUACIÓN 1

Lee las siguientes afirmaciones y marca Verdadera o Falsa, según corresponda.

	Verdadera	Falsa
1. El enunciado formal del teorema central del límite dice que si en cualquier población se seleccionan muestras de un tamaño específico, la distribución muestral de las medias de muestras es aproximadamente una distribución normal y que esta aproximación mejora con muestras de mayor tamaño.	( )	( )
2. La conclusión del teorema central del límite es una de las conclusiones menos útiles en estadística pues no permite razonar sobre la distribución muestral de las medias de muestras sin contar con información alguna sobre la forma de la distribución original de la que se toma la muestra.	( )	( )

3. El teorema central del límite permite aproximar la distribución de probabilidad normal a cualquier distribución de valores medios muestrales, siempre y cuando se trate de una muestra suficientemente grande.	( )	( )
4. El teorema central del límite se aplica a la distribución muestral de las medias de muestras y permite utilizar la distribución de probabilidad normal para crear intervalos de confianza.	( )	( )
5. La media muestral es uno de los estadísticos más utilizados en estadística inferencial.	( )	( )
6. Para que un investigador pueda asignar un valor probabilístico a una media muestral, es necesario que conozca la distribución muestral de las medias.	( )	( )
7. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ es la fórmula para calcular la desviación estándar de las medias de las muestras cuando la población es finita.	( )	( )
8. $\mu_{\bar{x}} = \mu\sqrt{\frac{N-n}{N-1}}$ es la fórmula para calcular la media de las medias para una población finita.	( )	( )
9. La media de las medias siempre es igual a la media de la población, independientemente de si la población es finita o infinita.	( )	( )

EXAMEN DE AUTOEVALUACIÓN 2

Elige la respuesta correcta a las siguientes preguntas.

1. Al considerar todas las muestras de tamaño “n” que pueden extraerse de una población, si se calcula el valor medio para cada una de ellas y se integran estos valores en un solo conjunto de datos es posible obtener una:
    a) campana de Gauss
    b) tendencia paramétrica
    c) curva de ajuste
    d) distribución muestral
    e) parámetro muestral

2. En el proceso de inferencia estadística paramétrica existen dos maneras de estimar los parámetros de una población, una de ellas es la:
    a) estadística descriptiva
    b) estimación puntual
    c) prueba de significancia
    d) medida de sesgo
    e) medida de tendencia central

3. Calcular el factor de corrección para la población finita de un inventario que consta de 250 productos y a la cual se le efectuará un muestreo de 40%:
    a) 0.881
    b) 0.918
    c) 0.819
    d) 0.991
    e) 0.989

4. Qué concepto establece que si se selecciona una muestra aleatoria suficientemente grande de n observaciones, la distribución muestral de las medias de las muestras se aproxima a una distribución normal.
    a) Definición de distribución muestral
    b) Proceso aleatorio
    c) Proceso de muestreo
    d) Teorema del límite central
    e) Distribución de probabilidad

5. Si una población se distribuye normalmente (con media y desviación estándarσ), la distribución muestral de las medias construida a partir de la misma población también se distribuye normalmente. Esta definición corresponde a la (el):
    a) teorema de Bayes
    b) ley de las probabilidades
    c) teorema del límite central
    d) ley de la distribución normal
    e) teorema de Markov

6. Una población se compone de los siguientes cinco números 2, 3, 6, 8, y 11. Calcula la media de la distribución muestral para tamaños de muestra 2 con reemplazamiento:
   a) 6.2
   b) 5.7
   c) 6.0
   d) 6.1
   e) 5.8

7. Cuando se lleva a cabo un estudio estadístico paramétrico se requiere una muestra suficientemente grande, lo cual significa que debe tener un tamaño igual o mayor a:
   a) 64
   b) 50
   c) 40
   d) 30
   e) 20

8. Si las distribuciones muestrales tienen la misma media, la elección de una de ellas deberá entonces basarse en la que tenga el menor valor del estadístico. Esta definición corresponde a:
   a) rango
   b) varianza
   c) sesgo
   d) mediana
   e) moda

9. Se tiene una lista de 120 estudiantes, 60 de ellos son de Contaduría y el resto de Administración. Si se toma una muestra al azar, halla la probabilidad de que se escojan entre el 40% y el 60% de contadores del tamaño de la muestra:
   a) 98.5%
   b) 96.7%
   c) 95.8%
   d) 97.7%
   e) 99.1%

10. De un lote muy grande (población infinita) de facturas, la desviación estándar es $10. Se extraen diversas muestras; cada una de ellas es de 200 facturas y se calculan las desviaciones estándar de cada muestra. Halla la media de la distribución muestral de desviaciones estándar:
    a) 0.30
    b) 0.50
    c) 2.77
    d) 7.41
    e) 10.0

LO QUE APRENDÍ

Preocupado por la variabilidad aparente de dos máquinas exactamente iguales y que fabrican el mismo tipo de botella para agua “ciel”, el dueño de la fábrica solicita un estudio en el que se muestreen al azar 10 botellas para cada máquina, obteniendo los siguientes resultados:

Máquina no. 1	Máquina no. 2
5.3	5.9
5.5	5.7
5.9	5.8
5.8	5.7
4.7	5.5
4.5	5.4
4.4	5.3
4.2	5.1
4.7	5.5
5.1	5.9

Si el diámetro de la botella debe ser de 5 cm. Y los valores de la tabla están dados en la misma escala, determina si las varianzas de ambas máquinas son diferentes.

MESOGRAFÍA

Bibliografía sugerida

Autor	Capítulo	Páginas
Berenson y otros (2001)	7	205-217
Levin y otros (1996)	6	247 -261
Christensen (1990)	5	235 - 250
Lind y otros (2004)	8	270 - 281

Bibliografía básica

Berenson, L. Mark; Levine, M. David; Krehbiel, C. Timothy. (2001). *Estadística para Administración*. (2ª ed.) México: Prentice Hall.

Levin, Richard I. y Rubin, David S. (1996). *Estadística para administradores*. México: Alfaomega.

Lind, A. Douglas; Marchal, G. William; Mason, D. Robert. (2004). *Estadística para Administración y Economía*. (11ª ed.) México: Alfaomega.

Bibliografía complementaria

Ato, Manuel y López, Juan J. (1996). *Fundamentos de estadística con SYSTAT*. México: Addison/Wesley.

Christensen, H. (1990). *Estadística paso a paso* (2ª ed.) México: Trillas.

Garza, Tomás. (1996). *Probabilidad y estadística*. México: Iberoamericana.

Hanke, John E. y Reitsch, Arthur G. (1997). *Estadística para Negocios.* México: Prentice Hall.

Weimer, Richard C. (1996). *Estadística*. México, CECSA.

Sitios de Internet

Sitio	Descripción
http://recursostic.educacion.es/descartes/web/materiales_didacticos/inferencia_estadistica/distrib_muestrales.htm	García Cebrian, María José. (2001). "Distribuciones muestrales", Estadística, Descartes 2D, Matemáticas interactivas.

http://www.ugr.es/~ramongs/laborales/tema6.pdf	Gutiérrez Sánchez, Ramón. (2007). "Distribuciones muestrales", Curso de Estadística, Diplomatura en Laborales, Universidad de Granada.
http://www.uoc.edu/in3/emath/docs/Distrib_Muestrales.pdf	Juan, Ángel A.; Sedano, Máximo, Vila, Alicia. (2002). "Distribuciones muestrales", Proyecto e-Math, UOC.
http://www.itch.edu.mx/academic/industrial/estadistica1/cap01.html	Torre, Leticia de la. (2003). "Teoría del Muestreo", Estadística I, Instituto Tecnológico de Chihuahua

UNIDAD 3

ESTIMACIÓN DE PARÁMETROS

OBJETIVO ESPECÍFICO

Al terminar la unidad el alumno aprenderá los métodos de estimación de parámetros y su interpretación.

INTRODUCCIÓN

En el momento de tomar decisiones el conocimiento de los parámetros de población es de vital importancia, tal conocimiento generalmente solo se puede tener al estimar el valor de dichos parámetros, sin embargo, la estimación es mejor cuando se da un margen de confianza y uno de error, siendo importante la correcta estimación de dichos parámetros a través de la construcción de intervalos de confianza que puedan sustentar la toma de decisiones de manera eficiente.

En el momento de tomar decisiones, es de vital importancia tener el conocimiento de los parámetros de población aunque estos solo pueden ser estimados de sus valores, sin embargo, la estimación es mejor cuando se tiene un margen de confianza y uno de error. Para ello se debe contar con una correcta estimación de los parámetros por medio de la construcción de intervalos de confianza que puedan sustentar la toma de decisiones de manera más eficiente.

LO QUE SÉ

Elige la respuesta correcta a las siguientes preguntas.

1. La media de la distribución de las medias de las muestras $\mu_{\bar{x}}$ siempre es

a) mayor que la de la población

b) menor que la de la población

c) igual a la de la población

2. La fórmula para la desviación estándar de la distribución de las medias de las muestras para una población suficientemente grande es:

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

b) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

c) $\sigma_{\bar{x}} = \sigma$

TEMARIO DETALLADO

(10 horas)

3.1. Estimaciones por punto y estimaciones por intervalo

3.2. Error de muestreo y errores que no son de muestreo

3.3. Propiedades de los estimadores

3.4. Estimación de una media con muestras grandes

3.5. Estimación de una media con muestras pequeñas

3.6. Estimación de una proporción

3.7. Otros intervalos de confianza

3.1. Estimaciones por punto y estimaciones por intervalo

Una estimación de un parámetro de la población dada por un solo número se llama una estimación de punto del parámetro. No obstante, un estimador puntual sólo refiere una parte de la historia. Si bien se espera que el estimador puntual esté próximo al parámetro de la población, se desearía expresar qué tan cerca está. Un intervalo de confianza sirve a este propósito.

Para realizar un análisis requerimos de una definición técnica. Utilicemos “a” como un símbolo genérico de un parámetro poblacional y, “â” para indicar una estimación de “a” basada en datos de la muestra. Una vez acordado esto podemos decir que un estimador “â” de un parámetro “a” es una función de los valores muestrales aleatorios, que proporciona una estimación puntual de “a”. Un estimador es en sí una variable aleatoria y por consiguiente tiene una distribución muestral teórica.

Se llama estimador puntual (Kreyszig, 2000[2], p.958) al número (punto sobre la recta real o recta de los números reales), que se calcula a partir de una muestra dada y que sirve como una aproximación (estimación) del valor exacto desconocido del parámetro de la población; es decir, es un valor que se calcula a partir de la información de la muestra, y que se usa para estimar el parámetro de la población.

Existe una distinción técnica entre un estimador como una función de variables aleatorias y una estimación como un único número. Tal distinción se refiere al proceso en sí (estimador) y el resultado de dicho proceso (la estimación.) Lo que en realidad importa de esta definición es que nosotros sólo podemos definir buenos procesos (estimadores), mas no garantizar buenos resultados (estimaciones).

Por ejemplo, la media muestral es el mejor estimador de una población normal (μ); sin embargo, no podemos garantizar que el resultado sea óptimo todas las veces. Es decir, no podemos garantizar que para cada muestra la media muestral esté siempre más cerca de la media poblacional, que, digamos, la mediana muestral (es decir, puede darse el caso en el que la mediana muestral esté más próxima a la media poblacional que la media muestral). Así, lo más que podemos hacer es encontrar estimadores que den buenos resultados en el límite.

Como una aproximación de la media de una población (Kreyszig, 2000[2]) puede tomarse la media $\bar{x}$ de una muestra correspondiente, lo cual da la estimación: $\hat{u}=\bar{x}$, para μ, es decir:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i \quad \text{----------(1)}$$

Donde n= tamaño de la muestra.

Del mismo modo, una estimación para la varianza de una población es la varianza de una muestra correspondiente; es decir:

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

Evidentemente, (1) y (2) son estimaciones de los parámetros para distribuciones en las que tanto la media como la varianza aparecen explícitamente como parámetros, tales como las distribuciones normal y de Poisson. Aquí, podemos mencionar que (1) es un caso muy especial del llamado método de los momentos, en la que los parámetros que van a estimarse se expresan en términos de los momentos de la distribución en las fórmulas resultantes (véase, Kreyszig, 2000[2], § 19.8); esos momentos se reemplazan por los momentos correspondientes de la muestra, lo cual proporciona las estimaciones deseadas.

Aquí, el k-ésimo momento de una muestra $x_1, x_2,...x_n$, es:

$$m_k = \frac{1}{n}\sum_{i=1}^{i=n}(x_i)^k$$

3.2. Error de muestreo y errores que no son de muestreo

La desviación estándar de una distribución, en el muestreo de un estadístico, es frecuentemente llamada el error estándar del estadístico. Por ejemplo, la desviación estándar de las medias de todas la muestras posibles del mismo tamaño, extraídas de una población, es llamada el error estándar de la media. De la misma manera, la desviación estándar de las proporciones de todas las muestras posibles del mismo tamaño, extraídas de una población, es llamada el error estándar de la proporción.

La diferencia entre los términos “desviación estándar” y “error de estándar” es que la primera se refiere a los valores originales, mientras que la última está relacionada con valores calculados. Un estadístico es un valor calculado, obtenido con los elementos incluidos en una muestra.

Error muestral o error de muestreo

La diferencia entre el resultado obtenido de una muestra (un estadístico) y el resultado el cual deberíamos haber obtenido de la población (el parámetro correspondiente) se llama el error muestral o error de muestreo. Un error de muestreo usualmente ocurre cuando no se lleva a cabo la encuesta completa de la población, sino que se toma una muestra para estimar las características de la población. El error muestral es medido por el error estadístico, en términos de probabilidad, bajo la curva normal. El resultado de la media indica la precisión de la estimación de la población basada en el estudio de la muestra. Mientras más pequeño es el error de muestras, mayor es la precisión de la estimación. Deberá hacerse notar que los errores cometidos en una encuesta por muestreo, tales como respuestas inconsistentes, incompletas o no determinadas, no son considerados como errores muéstrales. Los errores no muéstrales pueden también ocurrir en una encuesta completa de la población.

3.3. Propiedades de los estimadores

Insesgadez. Un estimador es insesgado o centrado cuando verifica que E($\hat{\theta}$) = θ . (Obsérvese que deberíamos usar $\hat{\theta}$ **(x)** y no $\hat{\theta}$, pues hablamos de estimadores y no de estimaciones pero como no cabe la confusión, para simplificar, aquí, y en lo sucesivo usaremos $\hat{\theta}$). En caso contrario se dice que el estimador es sesgado. Se llama sesgo a

$$B(\hat{\theta}) = \theta - E(\hat{\theta})$$

[Se designa con B de BIAS, sesgo en inglés]

Como ejemplo podemos decir que: la media muestral es un estimador insesgado de la media de la población (y lo es sea cual fuere la distribución de la población) ya que: si el parámetro a estimar es

$$\theta = \mu$$

y establecemos como estimador de $\mu = \hat{\mu} = \hat{\theta}(x) = \bar{x};$

Tendremos que $E\left[\hat{\theta}(x)\right] = E[\bar{x}] = \mu$ luego la media muestral es un estimador insegado de la media poblacional.

En cambio la varianza muestral es un estimador sesgado de la varianza de la población, ya que: si utilizamos como estimador de σ^2 la varianza muestral s^2 es decir: $\hat{\theta}(x) = s^2$ tendremos que $E\left[\hat{\theta}(x)\right] = E\left[s^2\right] = \sigma^2(n-1)/n = \sigma^2(1-\frac{1}{n}) \neq \sigma^2$ es el parámetro a estimar. Existe pues un sesgo que será

$$B\left[\hat{\theta}(x)\right] = B\left[s^2\right] = \sigma^2 - E\left[s^2\right] = \sigma^2 - (\sigma^2(1-1/n)) = \frac{\sigma^2}{n}$$

Dado que la varianza muestral no es un estimador de la varianza poblacional con propiedades de insesgadez, conviene establecer uno que si las tenga; este estimador no es otro que la cuasivarianza muestral, de ahí su importancia; así la cuasivarianza es en función de la varianza $s^{*2} = \frac{n s^2}{n-1}$ y tomada como estimador tendríamos que

$$E\left[s^{*2}\right] = E\left[\frac{n s^2}{n-1}\right] = \frac{n}{n-1} E\left[s^2\right] = \left(\frac{n}{n-1}\right)\left(\frac{n-1}{n}\right) \sigma^2 = \sigma^2$$

Dado que la esperanza del estimador coincide con el parámetro a estimar podemos decir que la cuasivarianza muestral es un estimador insesgado de la varianza de la población.

No obstante, y dado que, cuando el tamaño de la muestra tiende a infinito el sesgo tiende a cero, se dice que el estimador es asintóticamente insesgado o asintóticamente centrado: podemos establecer que:

$$\lim_{n\to\infty} B\left(s^2\right) = \lim_{n\to\infty} \sigma^2/n = 0$$

Por tanto la varianza muestral es un estimador sesgado pero asintóticamente insesgado de la varianza de la población.

Consistencia. Un estimador es consistente si converge en probabilidad al parámetro a estimar. Esto es: si

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right|<\varepsilon\right)=1$$

Linealidad. Un estimador es lineal si se obtiene por combinación lineal de los elementos de la muestra; así tendríamos que un estimador lineal sería:

$$\hat{\theta}=\hat{\theta}(x)=\sum_{i=1}^{n}k_i x_i$$

Eficiencia. Un estimador es eficiente u óptimo cuando posee varianza mínima o bien en términos relativos cuando presenta menor varianza que otro. Quedando claro que el hecho puede plantearse también en términos más coherentes de Error Cuadrático Medio (ECM). Tendríamos que:

$$\text{ECM}(\hat{\theta}) = E\left[\left(\hat{\theta}-\theta\right)^2\right] = D^2\left[\hat{\theta}\right]+B^2\left[\hat{\theta}\right]$$

por lo expresado podemos aventurar que un estimador insegado, luego

$$B^2\left[\hat{\theta}\right]=0$$

es el único capaz de generar eficiencia.

Suficiencia. Un estimador es suficiente cuando

$f(x_1, x_2, \ldots x_n / \hat{\theta})$ no depende del parámetro a estimar θ. En términos más simples: cuando se aprovecha toda la información muestral. [CEACES, aquí]

3.4. Estimación de una media con muestras grandes

El cálculo de Z (Unidades estandarizadas) para la distribución muestral de la media.

El valor de z estandarizada es igual a la diferencia que existe entre la media muestral $\bar{X}$ y la media poblacional μ, dividida por el error estándar de la media $\sigma_{\bar{x}}$.

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

Despejando el valor de $\bar{X}$, obtenemos:

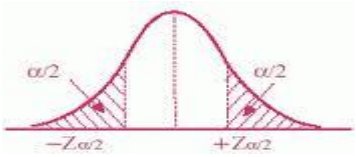
$$\bar{X} = \mu + z\frac{\sigma}{\sqrt{n}}$$

Pero como para el intervalo se debe encontrar un intervalo que contenga la media poblacional, entonces reemplazamos a μ por $\bar{X}$ y cada uno de los límites estará dado por:

$$\bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

$$\bar{X} - Z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z$$

Donde Z = valor correspondiente a una área acumulada 1 - $\frac{\alpha}{2}$ de la distribución normal estandarizada, esto es, una probabilidad de la cola superior de $\frac{\alpha}{2}$



3.4.1 Determinación del tamaño de muestra necesario para estimar una media

Si se desea estimar una media habrá que conocer:

(a) El nivel de confianza o seguridad (1−α). El nivel de confianza prefijado da lugar a un coeficiente (*z* α). Para un nivel de seguridad del 95 %, α=1.96, para un nivel de seguridad del 99 % α= 2.58;

(b) La precisión con que se desea estimar el parámetro (2×*d* es la amplitud del intervalo de confianza);

(c) Una idea de la varianza s^2 de la distribución de la variable cuantitativa que se supone existe en la

Población

$$n = Z^2 \alpha\ S^2 / d^2$$

Por ejemplo, si se desea conocer la media de la glucemia basal de una población, con una seguridad del 95 % y una precisión de ± 3 mg/dl y se tiene información a través un estudio piloto o de una revisión bibliográfica que la varianza es de 250 mg/dl:

$$n= 1.96^2 \times 250 / 3^2 = 106.7$$

3.5. Estimación de una media con muestras pequeñas

Si $\bar{x}$=85. σ=8, n=64, y suponiendo que la población se distribuye normalmente, construya una estimación del intervalo de confianza del 95% de la media poblacional μ.

La desviación estándar para la media

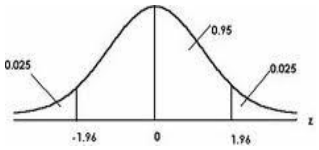
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{64}} = 1$$

Para intervalo del 95%

Cola 100%-95% = 5%.

Cada extremo tiene área de 2.5%.

Representa una área de 0.025.



Hallamos los límites del intervalo de confianza.

$$\bar{X}_i = \bar{x} \pm Z_i \frac{\sigma}{\sqrt{n}}$$

Para Z_1= -1.96, Tenemos $\bar{X}_1 = 85 - (1.96)(1)$

85 – 1.96

83.04

Para Z_1= 1.96, Tenemos $\bar{X}_1 = 85 + (1.96)(1)$

85 + 1.96

86.96

El intervalo de confianza es : 83.04 < μ < 86.96

Nos indica con el 95% de seguridad, que el promedio de las medias muéstrales de las cuentas está entre 83.04 y 86.96.

Por otra parte,

> Se podrán utilizar muestras pequeñas siempre y cuando la distribución de donde proviene la muestra tenga un comportamiento normal. Esta es una condición para utilizar las tres distribuciones que se manejarán en esta unidad; t de Student, X^2 ji-cuadrada y Fisher.
>
> A la teoría de pequeñas muestras también se le llama teoría exacta del muestreo, ya que también la podemos utilizar con muestras aleatorias de tamaño grande.
>
> En esta unidad se verá un nuevo concepto necesario para poder utilizar a las tres distribuciones mencionadas. Este concepto es "*grados de libertad*".
>
> Para definir grados de libertad se hará referencia a la varianza muestral:
>
> $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$
>
> Esta fórmula está basada en *n-1 grados de libertad* (*degrees of freedom*). Esta terminología resulta del hecho de que si bien s^2 está basada en *n* cantidades $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$, éstas suman cero, así que especificar los valores de cualquier *n-1* de las cantidades determina el valor restante.

Por ejemplo, si *n=4 y* $x_1 - \bar{x} = 8$; $x_2 - \bar{x} = -6$ y $x_4 - \bar{x} = -4$, entonces automáticamente tenemos $x_3 - \bar{x} = 2$, así que sólo tres de los cuatro valores de $x_i - \bar{x}$ están libremen [sic.] te determinamos 3 grados de libertad. [Torre, 2003]

3.6. Estimación de una proporción

Un estimador puntual de la proporción P en un experimento binomial está dado por la estadística P=X/N, donde x representa el número de éxitos en n pruebas. Por tanto, la proporción de la muestra p =x/n se utilizará como estimador puntual del parámetro P.

Si no se espera que la proporción P desconocida esté demasiado cerca de 0 ó de 1, se puede establecer un intervalo de confianza para P al considerar la distribución muestral de proporciones.

En este despeje podemos observar que se necesita el valor del parámetro P y es precisamente lo que queremos estimar, por lo que lo sustituiremos por la proporción de la muestra p siempre y cuando el tamaño de muestra no sea pequeño. [Torre, 2003]

Intervalo para estimar la proporción

En el caso de la proporción, el estadístico por utilizar es:

$$\frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p} - P}{\sqrt{P(1-P)/n}}$$

El cual, de acuerdo con el teorema del límite central, tendrá distribución normal estándar. En este caso, P es la proporción de la población con una característica dada y que se puede estimar por medio de $\overline{p}$, que es la proporción de la muestra con la característica.

Ejemplo:

Considera el caso de la Bolsa Mexicana de Valores; se desea estimar la proporción de las 250 acciones que tendrán una baja en precio al cierre del día. Para ello se observa una muestra de las primeras 4 horas sobre 50 acciones operadas y se observó que la proporción que bajó de precio es el 0.10 (10%). En el día se estima que no se presenten turbulencias por información importante o privilegiada. Se pide determinar el intervalo de confianza para la proporción total de acciones a la baja con un nivel de confianza del 90%.

De acuerdo con la metodología indicada el intervalo estará determinado por:

$$Z_{\alpha/2} < \frac{\overline{p} - P}{\sqrt{\overline{p}(1-\overline{p})/n}} < Z_{1-\alpha/2}$$

Pero de acuerdo con tablas normal estándar $Z_{\alpha/2} = Z_{0.05} = -1.64$ y $Z_{0.95} = 1.64$ y como $\overline{p} = 0.10$ entonces el intervalo se deduce de:

$$-1.64 < \frac{0.10 - P}{\sqrt{0.10(1-0.10)/50}} < 1.64$$

que equivale a:

$$-1.64(0.0424264) < 0.10 - P < 1.64(0.0424264)$$

y despejando P se tiene:

$$-1.64(0.04242064) - 0.10 < -P < 1.64(0.0424264) - 0.10$$

iguala:

$$1.64(0.0424264) + 0.10 > P > -1.64(0.0424264) + 0.10$$

Por lo cual el intervalo es:

$$0.169 > P > 0.0304$$

Es decir aproximadamente entre el 3% y 17%.

3.6.1 Determinación del tamaño de muestra para estimar una proporción

Si deseamos estimar una proporción, debemos saber:

a) El nivel de confianza o seguridad (1-α). El nivel de confianza prefijado da lugar a un coeficiente (Z_α). Para una seguridad del 95% = 1.96, para una seguridad del 99% = 2.58.

b) La precisión que deseamos para nuestro estudio.

c) Una idea del valor aproximado del parámetro que queremos medir (en este caso una proporción). Esta idea se puede obtener revisando la literatura, por estudio pilotos previos. En caso de no tener dicha información utilizaremos el valor p = 0.5 (50%).

Ejemplo: ¿A cuántas personas tendríamos que estudiar para conocer la prevalencia de diabetes?

Seguridad = 95%; Precisión = 3%: Proporción esperada = asumamos que puede ser próxima al 5%; si no tuviésemos ninguna idea de dicha proporción utilizaríamos el valor p = 0,5 (50%) que maximiza el tamaño muestral:

$$n = \frac{Z_\alpha^2 * p * q}{d^2}$$

Donde:

$Z_\alpha^2 = 1.96^2$ (ya que la seguridad es del 95%)
p = proporción esperada (en este caso 5% = 0.05)
q = 1 – p (en este caso 1 – 0.05 = 0.95)
d = precisión (en este caso deseamos un 3%)

$$n = \frac{1.96^2 * 0.05 * 0.95}{0.03^2} = 203$$

Si la población es finita, es decir conocemos el total de la población y deseásemos saber cuántos del total tendremos que estudiar la respuesta seria:

$$n = \frac{N * Z_\alpha^2 p * q}{d^2 * (N-1) + Z_\alpha^2 * p * q}$$

Donde:

N = Total de la población
$Z_\alpha^2 = 1.96^2$ (si la seguridad es del 95%)
p = proporción esperada (en este caso 5% = 0.05)
q = 1 – p (en este caso 1-0.05 = 0.95)
d = precisión (en este caso deseamos un 3%).

¿A cuántas personas tendría que estudiar de una población de 15.000 habitantes para conocer la prevalencia de diabetes?

Seguridad = 95%; Precisión = 3%; proporción esperada = asumamos que puede ser próxima al 5%; si no tuviese ninguna idea de dicha proporción utilizaríamos el valor p = 0.5 (50%) que maximiza el tamaño muestral.

$$n = \frac{15.000 * 1.96^2 * 0.05 * 0.95}{0.03^2 (15.000 - 1) + 1.96^2 * 0.05 * 0.95} = 200$$

Según diferentes seguridades el coeficiente de Zα varía, así:

Si la seguridad Zα fuese del 90% el coeficiente sería 1.645

Si la seguridad Zα fuese del 95% el coeficiente sería 1.96
Si la seguridad Zα fuese del 97.5% el coeficiente sería 2.24
Si la seguridad Zα fuese del 99% el coeficiente sería 2.576.

(Fernández, 1996, pp. 1-2)

3.7. Otros intervalos de confianza

Intervalo de confianza

Un rango de valores que se construye a partir de datos de la muestra de modo que el parámetro ocurre dentro de dicho rango con una probabilidad específica se conoce como nivel de confianza.

Es decir, una estimación de un parámetro de la población dada por dos números, entre los cuales se puede considerar encajado al parámetro, se llama una estimación de intervalo del parámetro.

Las estimaciones de intervalo indican la precisión de una estimación y son, por tanto, preferibles a las estimaciones puntuales.

Por ejemplo: si decimos que el porcentaje de productos defectuosos que produce una máquina es del 6%, entonces el nivel se ha medido en 0.06 y estamos dando una estimación de punto. Por otra parte, si decimos que el porcentaje es 0.05±0.03 m (o sea, que está entre 2% y 8%), estamos dando una estimación de intervalo).

El margen de error (o la precisión) de una estimación nos informa de su fiabilidad.

Intervalo para estimar la media

De acuerdo con tablas de la distribución normal estándar el área bajo la curva entre z=-1 y z=+1 es 0.6826; por consiguiente, y de acuerdo con la definición de la función normal estándar de probabilidad, las desigualdades siguientes se cumplen con probabilidad de 0.6826.

$$-1 < z < +1$$

Como la distribución de las medias de las muestras (con media $\mu_{\bar{x}}$ y desviación estándar $\sigma_{\bar{x}}$) es normal, entonces:

Si reemplazamos z por $\frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$ en las desigualdades anteriores,

Se deberá cumplir:

$$-1 < \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} < +1$$

Con probabilidad 0.6826. Esto es equivalente a que las desigualdades:

$$\bar{X} - \sigma_x < \mu_x < \bar{X} + \sigma_x$$

Se cumplan también con probabilidad 0.6826; sustituyendo ahora:

$\sigma_{\bar{x}}$ Por $\frac{s}{\sqrt{n}}$ y $\mu_{\bar{x}}$ por μ_x

Se tiene que:

$$\bar{X} - \frac{s}{\sqrt{n}} < \mu_x < \bar{X} + \frac{s}{\sqrt{n}}$$

Se cumple con la misma probabilidad.

Podemos esperar entonces que con una probabilidad de 0.68 que μ_x se encuentre dentro del intervalo:

(69– 0.58, 69+0.58)

Es decir: $68.42 < \hat{\mu}_x < 69.58$ aquí, la media aritmética de la población lleva un acento circunflejo debido a que se trata de una estimación.

Se dice que éste es un intervalo de confianza de 0.68 o 68%, ya que se tiene una confianza de 68% de que el intervalo contenga la media de la población.

Si una confianza de 68% fuese insuficiente se pueden construir otros intervalos con porcentajes de confianza que sean más útiles.

Por ejemplo, si se deseara encontrar un intervalo de confianza de 0.95 para se requeriría determinar “k” de tal manera que las desigualdades siguientes se cumplieran con probabilidad de 0.95.

$$-k < z < +k \quad \text{----------1}$$

En términos generales, para encontrar un intervalo de cualquier porcentaje de confianza, se hace lo siguiente:

1º Se divide el porcentaje de confianza requerido entre 100.
2º El resultado del punto anterior se divide entre 2.
3º El valor así obtenido se busca en las tablas de la curva de distribución normal.
4º El valor encontrado en las tablas se sustituye en 1 y comenzamos el proceso nuevamente.

Es decir, en nuestro caso el valor resultante es de 0.475; por lo tanto, el valor en las tablas que se encuentra junto a éste último es “1.96”. Es decir, el área bajo la curva normal estándar entre –1.96 y +1.96 es 0.9544, o sea, aproximadamente 0.95. Así, la probabilidad de que z se encuentre dentro del intervalo:

(-1.96, +1.96)

Es, aproximadamente 0.95 o, en otra forma, las desigualdades:

-1.96 <z<+1.96

Se cumplen con probabilidad 0.95;

Y puesto que se sabe que la distribución de las medias de las muestras es normal,

Se puede reemplazar z por $\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{x}}}$

Expresión que aproximada a:

$$\frac{\bar{X}-\mu_x}{\frac{s}{\sqrt{n}}}$$

En las desigualdades anteriores, se obtiene:

$$-1.96 < \frac{\bar{X}-\mu_x}{\frac{s}{\sqrt{n}}} < +1.96$$

Resolviendo estas desigualdades para μ , se tiene que:

$$\bar{X} - \frac{1.96s}{\sqrt{n}} < \mu_x < \bar{X} + \frac{1.96s}{\sqrt{n}} \quad \text{----------2}$$

Como un intervalo con 0.95 de confianza para μ . Por lo tanto, se puede afirmar con 95% de confianza que μ se encuentra dentro del intervalo:

$$\bar{X} - \frac{1.96s}{\sqrt{n}} \quad \text{y} \quad \bar{X} + \frac{1.96s}{\sqrt{n}}$$

Por lo tanto, sustituyendo los valores de la media y de la desviación estándar, así como del tamaño de la muestra para el ejercicio anterior (media 69, desviación estándar 3.5 y tamaño de muestra 36) en 2 se tiene que el intervalo con 95% de confianza es:

$$69 - \frac{1.96 \ \ 3.5}{\sqrt{36}} < \mu_x < 69 + \frac{1.96 \ \ 3.5}{\sqrt{36}}$$

$$67.8 < \mu_x < 70.1$$

Intervalo para estimar la varianza

Propiedades de los estimadores, sabemos que el estimador para varianza poblacional (σ^2) es S^2; sin embargo, para estimar un intervalo de confianza para σ^2 es necesario conocer la distribución del estadístico; más aún, la metodología implica que es necesario tener un estadístico que involucre el parámetro desconocido y que además tenga distribución perfectamente conocida. Por lo cual, en este caso el estadístico es:

$$\frac{(n-1)S^2}{\sigma^2}$$

Que de acuerdo con lo estudiado en el tema 2 tiene una distribución Chi-cuadrada con n-1 grados de libertad. Así que para una muestra particular, dicho estadístico tiene una probabilidad de estar en un rango dado.

Ejemplo:

Considera el caso de estimar si no hay deficiencias en una máquina que llena envases con capacidad de 500 ml.; para ello, se extrae una muestra periódicamente; si la muestra indica que hay una variación de ±5 ml. alrededor de los 500 y con un nivel de confianza del 95%, entonces se puede decir que el proceso está bajo control.

En este caso lo que importa es la variación en el llenado, pues el nivel promedio de llenado se puede controlar programando la máquina. Por ello, si la muestra arroja una variación arriba de 5 unidades, entonces el proceso no estará bajo control.

Suponga que la muestra de tamaño 41 arroja una varianza de 13 unidades (desviación estándar de 3.60 ml). Entonces, de acuerdo con la estimación por intervalos de confianza, se tendrá que:

$$X^2_{0.025} < \frac{(n-1)S^2}{\sigma^2} < X^2_{0.975}$$

El resultado anterior de acuerdo con tablas de Chi-cuadrada con 40 grados de libertad $X^2_{0.025}$=24.433 y $X^2_{0.09750}$ = 59.342.
(Recuerda que el uso de las tablas y de los grados de libertad se encuentra en el apartado 3.2)

Entonces el intervalo es:

$$24.433 < \frac{(n-1)S^2}{\sigma^2} < 59.342$$

Sustituyendo los resultados de la muestra se tiene:

$$24.433 < \frac{(40-1)(13)}{\sigma^2} < 59.342$$

Al obtener inversos multiplicativos tenemos:

$$\frac{1}{24.433} > \frac{\sigma^2}{(40-1)(13)} > \frac{1}{59.342}$$

Despejando todas las constantes y dejar solo σ^2 se tiene el intervalo:

$$\frac{1}{24.433} > \frac{\sigma^2}{(40-1)(13)} > \frac{1}{59.342}$$

$$20.75 > \sigma^2 > 8.54$$

Obteniendo raíz cuadrada, se tiene:

$$4.555 > \sigma > 2.92$$

Por lo cual se puede decir que el proceso está bajo control.

RESUMEN DE LA UNIDAD

Las inferencias acerca de una población que se obtienen del estudio de una muestra pueden ser tan buenas como lo sean las estimaciones obtenidas, aquí, el cuidado va evidentemente sobre la recolección de los datos, pues existe una gran variedad de estimadores que pueden ser utilizados dependiendo del contexto pero el éxito de la aplicación de un estimador (estimación) dependerá necesariamente de la calidad de los datos mismos, resulta evidente que esto es extensible a los intervalos de confianza tanto para la media como para proporciones.

En el presente análisis únicamente nos restringimos a la aplicación de estimadores, considerando que los datos son de una calidad suficientemente buena para obtener buenas estimaciones de los diferentes parámetros de interés así como también del tamaño de la muestra necesario tanto para la media como para la proporción.

GLOSARIO DE LA UNIDAD

Distribución t

Es en realidad una familia de distribuciones de probabilidad que se emplea para construir un intervalo de confianza para la media poblacional, siempre que la desviación estándar σ se estime mediante la desviación estándar muestral "s" y la población tenga una distribución de probabilidad normal o casi normal.

Error muestral

Es el valor absoluto de la diferencia entre el valor de un estimador puntual insesgado, tal como la media de la muestra $\bar{x}$ y el valor del parámetro poblacional que estima, (en este caso, la media de la población μ); es decir, en este caso el error muestral es: $\left|\bar{x}-\mu\right|$

Estimación de intervalo

Estimación de un parámetro de la población que define un intervalo dentro del que se cree está contenido el valor del parámetro. Tiene la forma de: Estimación puntual $\pm$ margen de error.

Grados de libertad

Es el número de observaciones independientes para una fuente de variación menos el número de parámetros independientes estimado al calcular la variación.

Margen de error

Es el valor $\pm$ sumado y restado a una estimación puntual a fin de determinar un intervalo de confianza de un parámetro poblacional.

Nivel de confianza

Es la confianza asociada con una estimación de intervalo. Por ejemplo si en un proceso de estimación de intervalo, el 90% de los intervalos formados con este procedimiento contienen el valor del parámetro buscado, se dice que éste es un intervalo de 90% de confianza.

ACTIVIDADES DE APRENDIZAJE

ACTIVIDAD 1

Completa el siguiente cuadro sobre los tipos de estimadores.

	Ventajas	Desventajas
Estimadores sesgados		
Estimadores insesgados		
Estimadores consistentes		
Estimadores inconsistentes		

ACTIVIDAD 2

Construye un intervalo de confianza para la varianza de forma general.

ACTIVIDAD 3

Resuelve los siguientes problemas, escribe tu respuesta.

1. Considera una empresa que comercializa productos genéricos de limpieza y deseamos estimar el consumo promedio anual de una población potencial. Si obtenemos una muestra piloto de 15 personas en donde S=28.9 l y queremos un nivel de confianza de 95% con un error en la estimación de B=2 l. Determina el tamaño de la muestra que debe evaluarse.

2. El día de hoy la bolsa mexicana de valores estimó con una muestra de 20 acciones que el promedio de las que estuvieron a la alza fue de $\overline{p}=0.10$; con un nivel de confianza de 90% y un error a lo más de 5%. Determina el tamaño de la muestra que debe ser estudiada.

CUESTIONARIO DE REFORZAMIENTO

1. ¿Cuál será la probabilidad de que un auditor tenga cuatro éxitos si va a realizar cinco auditorías, suponiendo que las probabilidades de éxito y por ende las de fracaso son independientes de una auditoría a otra?
2. Supón que la llegada de trabajos a un despacho contable obedece a una distribución de Poisson y que en dicho despacho realizamos un muestreo; durante el primer día llegaron dos trabajos; en el segundo, cuatro; en el tercero, tres; en el cuarto, cinco; y en el quinto, dos. Encuentra el estimador de máxima verosimilitud correspondiente.
3. Si realizamos un muestreo en un autolavado donde durante la primera hora llegaron dos automóviles; en la segunda, cuatro; en la tercera, tres; en la cuarta, cinco; y en la quinta, dos; encuentra el estimador de máxima verosimilitud correspondiente.
4. Una muestra aleatoria de tamaño 49 tiene una media de 157 y una desviación estándar de 14.7. Determina un intervalo con 95% de confianza para la media verdadera de la muestra.
5. Suponiendo que 64 mediciones de la densidad del cobre dieron por resultado una media de 8.81 y una desviación estándar de 0.24, calcula un intervalo con 99% de confianza para la densidad verdadera.

6. En una muestra aleatoria de 125 llantas para automóvil, se encontró que la vida media fue de 35,000 km. y la desviación estándar de 4,000. Determina un intervalo con 68% de confianza para la vida media.
7. Un estudio sobre ciertas acciones comunes permitió conocer que en una muestra aleatoria de 100 acciones la rentabilidad anual promedio fue de 4.2%, mientras que su desviación estándar es de 0.6%. Determina un intervalo, con 95% de confianza, para la rentabilidad promedio.
8. ¿Cuál es la diferencia entre una estimación y un estimador?
9. ¿Qué es un intervalo de confianza?
10. Señala, ¿por qué son preferibles las estimaciones de intervalo a las estimaciones puntuales?

EXAMEN DE AUTOEVALUACIÓN

Elige la respuesta correcta a las siguientes preguntas.

1. En este estimador su esperanza matemática es igual a parámetro en cuestión:
    a) robusto
    b) insesgado
    c) sesgado

2. Es un estimador para un problema particular y tiene el error estándar más pequeño de todos los estimadores insesgados posibles:
    a) el más eficiente
    b) sesgado
    c) ineficiente

3. Este tipo de estimaciones se usan con frecuencia a causa de la relativa sencillez con que se obtienen algunas de ellas
    a) consistentes
    b) robustas
    c) ineficientes

4. Este tipo de estimadores son estadísticos casi insesgados y casi eficientes para una gran variedad de distribuciones poblacionales:
    a) consistentes
    b) robustos
    c) eficientes

5. Este tipo de estimador se aproxima al parámetro poblacional con probabilidad uno a medida que el tamaño de la muestra tiende a infinito:
    a) consistente
    b) robusto
    c) inconsistente

LO QUE APRENDÍ

Construye un intervalo de confianza de 95% para la vida media de los neumáticos muestreados en la tabla mostrada a continuación. (Nota. Los datos están dados en miles de kilómetros.)

85,000	90,000	100,000	105,000
90,000	95,000	92,300	97,200
91,000	98,000	97,000	97,500
88,000	89,900	99,600	99,500
97,890	99,870	95,490	94,789
90,890	99,810	98,900	
97,870	97,980	99,950	
96,190	96,710	95,498	
98,990	97,900	95,267	
96,876	96,930	99,900	

MESOGRAFÍA

Bibliografía sugerida

Autor	Capitulo	Páginas
Levin y otros (1996)	7	273-313
Berenson y otros (2001)	10	344-382
Christensen (1990)	7	311-356
Lind y otros (2004)	9	294-309

Bibliografía básica

Berenson, L. Mark; Levine, M. David; Krehbiel, C. Timothy. (2001). *Estadística para Administración*. (2ª ed.) México: Prentice Hall.

Levin, Richard I. y Rubin, David S. (1996). *Estadística para administradores*. México: Alfaomega.

Lind, A. Douglas; Marchal, G. William; Mason, D. Robert. (2004). *Estadística para Administración y Economía*. (11ª ed.) México: Alfaomega.

Bibliografía complementaria

Ato, Manuel y López, Juan J. (1996). *Fundamentos de estadística con SYSTAT*. México: Addison/Wesley.

Christensen, H. (1990). *Estadística paso a paso* (2ª ed.) México: Trillas.

Garza, Tomás. (1996). *Probabilidad y estadística*. México: Iberoamericana.

Hanke, Jonh E. y Reitsch, Arthur G. (1997). *Estadística para Negocios.* México: Prentice Hall.

Kreyszig, Erwin. (2000). *Matemáticas avanzadas para ingeniería, vol. 2,* (3ª ed.) México: Limusa.

Sitios de Internet

Sitio	Descripción
http://www.itescam.edu.mx/principal/sylabus/fpdb/recursos/r53794.PDF	Fernández, Pita. (1996). "Determinación del tamaño muestral", Cad Aten Primaria 1996; 3: 138-14, actualizado 06/03/01
http://www.uv.es/ceaces/tex1t/4%20estimacion/estimacion.html#2.Propiedades%20de%20los%20Estimadores	Martínez de Lejarza Esparducer, Juan y otros. (2011). "Inferencia

	estadística / Estimación puntual / propiedades de los estimadores", Contenedor Hipermedia de Estadística Aplicada a las Ciencias Económicas y sociales", (Proyecto CEACES), Universidad de Valencia.
http://www.itch.edu.mx/academic/industrial/estadistica1/cap01.html	Torre, Leticia de la. (2003). "Teoría del Muestreo", Estadística I, Instituto Tecnológico de Chihuahua

UNIDAD 4

PRUEBAS DE HIPÓTESIS

OBJETIVO ESPECÍFICO

Al terminar la unidad el alumno conocerá las pruebas de hipótesis y su aplicación.

INTRODUCCIÓN

En esta unidad, el alumno investigará y analizará el concepto de prueba de hipótesis y lo aplicará sobre varianzas, medias, etc.; ello le permitirá percatarse de la importancia que tienen las pruebas de hipótesis para la toma de decisiones dentro de las empresas.

Actualmente, sabemos que la matemática es una herramienta importante en la toma de decisiones, y la estadística junto con todos sus procesos no es la excepción; así, es importante que el alumno desarrolle todos los conceptos y ejercicios aquí planteados, enriqueciendo su cultura matemática para su futuro desempeño profesional.

Sabemos que cuando las personas toman decisiones, inevitablemente lo hacen con base en las creencias que tienen en relación con el mundo que los rodea; llevan en la mente una cierta imagen de la realidad, piensan que algunas cosas son verdaderas y otras falsas y actúan en consecuencia, así, los ejecutivos de empresas toman todos los días decisiones de importancia crucial porque tienen ciertas creencias tales como:

- De que un tipo de máquina llenadora pone al menos un kilogramo de detergente en una bolsa.
- De que cierto cable de acero tiene una resistencia de 100 kg o más a la rotura.
- De que la duración promedio de una batería es igual a 500 horas.
- De que en un proceso de elaboración de cápsulas éstas contengan precisamente 250 miligramos de un medicamento.
- Que la empresa de transportes de nuestra competencia tiene tiempos de entrega más rápidos que la nuestra.
- De que la producción de las plantas de oriente contiene menos unidades defectuosas que las de occidente.

Incluso los estadistas basan su trabajo en creencias tentativas:

- Que dos poblaciones tienen varianzas iguales.
- Que esta población está normalmente distribuida.
- Que estos datos muestrales se derivan de una población uniformemente distribuida.

En todos estos casos, y en muchos más, las personas actúan con base en alguna creencia sobre la realidad, la cual quizá llegó al mundo como una simple conjetura, como un poco más que una suposición informada; una proposición adelantada tentativamente como una verdad posible es llamada hipótesis.

Sin embargo, tarde o temprano, toda hipótesis se enfrenta a la evidencia que la comprueba o la rechaza y, en esta forma, la imagen de la realidad cambia de mucha a poca incertidumbre.

Por lo tanto, de una manera sencilla podemos decir que una prueba de hipótesis es un método sistemático de evaluar creencias tentativas sobre la realidad, dicho método requiere de la confrontación de tales creencias con evidencia real y decidir, en vista de esta evidencia, si dichas creencias se pueden conservar como razonables o deben desecharse por insostenibles.

A continuación estudiaremos la forma en que las creencias de las personas pueden ser probadas de manera sistemática.

LO QUE SÉ

Elige la respuesta correcta a las siguientes preguntas.

1. La fórmula del estadístico “z” es:

a) $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

b) $\frac{\alpha}{2}$

c) $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

2. La fórmula correcta para calcular un intervalo de confianza es:

a) $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

b) $\frac{\alpha}{2}$

c) $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

TEMARIO DETALLADO

(10 horas)

4.1. Planteamiento de las hipótesis
4.2. Errores tipo I y tipo II
4.3. Pruebas de uno y de dos extremos, y regiones de aceptación y de rechazo
4.4. Pruebas de hipótesis para una media poblacional
4.5. Tres métodos para realizar pruebas de hipótesis
4.6. Prueba de hipótesis sobre una proporción poblacional
4.7. Pruebas de hipótesis sobre la diferencia entre dos medias
4.8. Pruebas de hipótesis sobre la diferencia entre dos proporciones
4.9. Prueba para la diferencia entre dos varianzas

4.1. Planteamiento de las hipótesis

1. Formulación de dos hipótesis opuestas

El primer paso para probar una hipótesis es siempre formular dos hipótesis opuestas, que sean mutuamente excluyentes y, también colectivamente exhaustivas, del experimento que estemos evaluando. Cada una de estas hipótesis complementarias es una proposición sobre un parámetro de la población tal que la verdad de una implique la falsedad de la otra. La primera hipótesis del conjunto, simbolizada por H_0, se denomina hipótesis nula; la segunda, simbolizada por H_1 o bien por Ha, es la hipótesis alternativa.

2. Selección de un estadístico de prueba

El segundo paso para probar una hipótesis es la selección de un estadístico de prueba. Un estadístico de prueba es aquel calculado con base en una sola muestra aleatoria simple tomada de la población de interés; en una prueba de hipótesis sirve para establecer la verdad o falsedad de la hipótesis nula.

3. Derivación de una regla de decisión

Una vez que hemos formulado de manera apropiada las dos hipótesis opuestas y seleccionado el tipo de estadístico con que probarlas, el paso siguiente en la prueba de hipótesis es la derivación de una regla de decisión:

Una regla de decisión es una regla para prueba de hipótesis que nos permite determinar si la hipótesis nula debe ser aceptada o si debe ser rechazada a favor de la alternativa.

Se dice que los valores numéricos del estadístico de prueba para los que H_0 es aceptada están en la región de aceptación y son considerados no significativos estadísticamente.

Por el contrario, si el valor numérico del estadístico de prueba se encuentra en la región de rechazo, esto aconseja que la hipótesis alternativa sustituya a la desacreditada hipótesis nula; entonces este valor es considerado estadísticamente significativo.

Es importante notar que la aceptación o rechazo se refiere a la hipótesis nula H_0.

4. Toma de una muestra, cálculo del estadístico de prueba y confrontación con la regla de decisión (véase, Kohler, 1996, p.384).

El paso final en la prueba de hipótesis requiere:

- Seleccionar una muestra aleatoria simple de tamaño n, de la población de interés,
- Calcular el valor real (opuesto al crítico) del estadístico de prueba (seleccionado en el paso 2).
- Confrontar con la regla de decisión (derivada en el paso 3).

4.2. Errores tipo I y tipo II

Error tipo I

En una prueba estadística, rechazar la hipótesis nula cuando es verdadera se denomina error tipo I. Y a la probabilidad de cometer un error tipo I se le asigna el símbolo α (letra griega alfa).

La probabilidad de α aumenta o disminuye a medida que aumenta o disminuye el tamaño de la región de rechazo. Entonces, ¿por qué no se disminuye el tamaño de la región de rechazo para hacer α tan pequeña como sea posible?

Desgraciadamente, al disminuir el valor de α aumenta la probabilidad de no rechazar la hipótesis nula cuando ésta es falsa y alguna hipótesis alternativa es verdadera. Aumenta entonces la probabilidad de cometer el llamado error de tipo II para una prueba estadística.

Ejemplo

Incurrir en un riesgo α

Un fabricante de varillas de acero especial que son utilizadas en la construcción de edificios muy altos ha contratado a un estadista para que pruebe si sus varillas ciertamente tienen un promedio de resistencia a la tensión de al menos 2000 libras ¿Cuáles son las implicaciones si el nivel de significancia de la prueba de hipótesis se fija en: α = 0.08?

Solución:

Dadas las hipótesis: $H_0 : \mu_0 \geq 2000$ y $H_1 : \mu_0 < 2000$

El procedimiento asegura lo siguiente: aun cuando las varillas tengan un promedio de resistencia a la tensión de 2000 libras o más, en el 8% de todas las pruebas la conclusión será lo contrario.

Error Tipo II

En una prueba estadística, aceptar la hipótesis nula cuando es falsa se denomina error tipo II. A la probabilidad de cometer un error de tipo II se le asigna el símbolo β (letra griega beta)

Para un tamaño de muestra fijo, α y β están inversamente relacionados; al aumentar uno, el otro disminuye. El aumento del tamaño de muestra produce mayor información sobre la cual puede basarse la decisión y, por lo tanto, reduce tanto a α como a β. En una situación experimental, las probabilidades de los errores de tipo I y II para una prueba miden el riesgo de tomar una decisión incorrecta. El experimentador selecciona los valores de estas probabilidades y la región de rechazo y el tamaño de muestra se escogen de acuerdo con ellas.

Ejemplo

Incurrir en un riesgo β

El fabricante de computadoras ha contratado a un estadista para probar si el ensamble de una computadora toma un promedio de al menos 50 minutos. ¿Cuáles son las implicaciones si el riesgo β de la prueba es igual a 0.2?

Solución

Dadas las hipótesis: $H_0 : \mu_0 \geq 50$ y $H_1 : \mu_0 < 50$

El procedimiento asegura lo siguiente: incluso si el tiempo de ensamble en efecto promedia más de 50 minutos, en el 20% de todas las pruebas la conclusión será lo contrario. Sin embargo, en el 80% de dichas pruebas este tipo de error se evita, lo que indica la potencia de la prueba.

Nivel de significancia

El nivel de significancia o significación es la probabilidad de cometer un error tipo I, es decir, el valor que se le asigna a α.

Potencia de la prueba

Es posible determinar (Weimer, 1996, p. 461) la probabilidad asociada con tomar una decisión correcta: no rechazar H_0 cuando es verdadera o rechazarla cuando es falsa. La probabilidad de no rechazar H_0 cuando es verdadera es igual a 1-α.

Esto se puede demostrar notando que:

P(rechazar Ho cuando es verdadera) + P(no rechazar Ho cuando es verdadera) = 1

Como

P (rechazar Ho cuando es verdadera) = α

Tenemos:

$$P \text{ (no rechazar Ho cuando es verdadera)} = 1 - \alpha$$

Nota que la probabilidad de no rechazar Ho cuando es verdadera es el nivel de confianza $1-\alpha$

La probabilidad de rechazar Ho cuando es falsa es igual a $1-\beta$. Esto se puede demostrar notando que:

$$P(\text{rechazar Ho cuando es falsa}) + P(\text{no rechazar Ho cuando es falsa}) = 1$$

Pero como:

$$P \text{ (no rechazar Ho cuando es falsa)} = \beta$$

Tenemos:

$$P \text{ (rechazar Ho cuando es falsa)} = 1-\beta$$

La probabilidad de rechazar la hipótesis nula H_0 cuando es falsa se llama potencia de la prueba.

Probabilidades asociadas con los cuatro resultados posibles de una prueba de hipótesis.

Símbolo de la probabilidad	**Definición**
α	Nivel de significancia. Probabilidad de un error tipo I
β	Probabilidad de un error tipo II

1-α	Nivel de confianza. Probabilidad de no rechazar H_0 cuando es verdadera
1-β	Potencia de la prueba. Probabilidad de rechazar H_0 cuando es falsa.

4.3. Pruebas de uno y de dos extremos y regiones de aceptación y de rechazo

a) Prueba bilateral o de dos extremos: la hipótesis planteada se formula con la igualdad.

Ejemplo
H0 : μ = 200
H1 : μ ≠ 200

b) Pruebas unilateral o de un extremo: la hipótesis planteada se formula con ≥ o ≤
H0 : μ ≥ 200 H0 : μ ≤ 200
H1 : μ < 200 H1 : μ > 200

En las pruebas de hipótesis para la media (μ), cuando se conoce la desviación estándar (σ) poblacional, o cuando el valor de la muestra es grande (30 o más), el valor estadístico de prueba es z y se determina a partir de:

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

El valor estadístico z, para muestra grande y desviación estándar poblacional desconocida se determina por la ecuación:

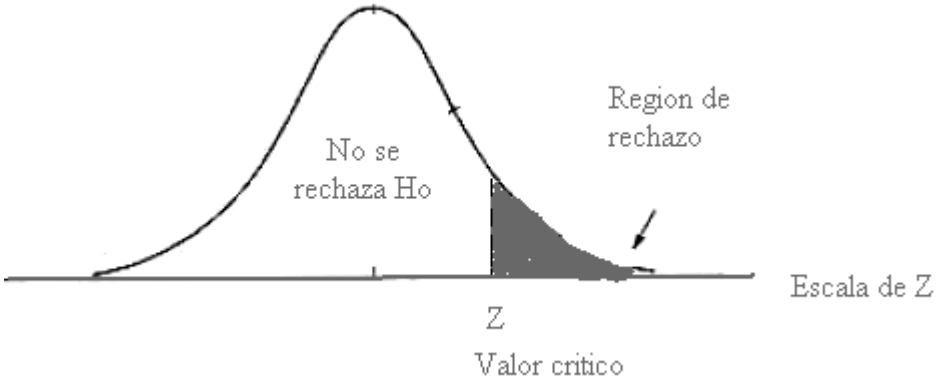
$$Z = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

En la prueba para una media poblacional con muestra pequeña y desviación estándar poblacional desconocida se utiliza el valor estadístico t.

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

[Paso 4] Formular la regla de decisión

Se establece las condiciones específicas en la que se rechaza la hipótesis nula y las condiciones en que no se rechaza la hipótesis nula. La región de rechazo define la ubicación de todos los valores que son tan grandes o tan pequeños, que la probabilidad de que se presenten bajo la suposición de que la hipótesis nula es verdadera, es muy remota



Distribución muestral del valor estadístico z, con prueba de una cola a la derecha

Valor crítico: Es el punto de división entre la región en la que se rechaza la hipótesis nula y la región en la que no se rechaza la hipótesis nula.

[Paso 5] Tomar una decisión

En este último paso de la prueba de hipótesis, se calcula el estadístico de prueba, se compara con el valor crítico y se toma la decisión de rechazar o no la hipótesis nula. Tenga presente que en una prueba de hipótesis solo se puede tomar una de dos decisiones: aceptar o rechazar la hipótesis nula. Debe subrayarse que siempre existe la posibilidad de rechazar la hipótesis nula cuando no debería haberse rechazado (error tipo I). También existe la posibilidad de que la hipótesis nula se acepte cuando debería haberse rechazado (error de tipo II). (Cruz, 2009)

4.4. Pruebas de hipótesis para una media poblacional

Dentro de la inferencia estadística, un contraste de hipótesis (también denominado test de hipótesis o prueba de significación) es un procedimiento para juzgar si una propiedad que se supone cumple una población estadística es compatible con lo observado en una muestra de dicha población. Fue iniciada por Ronald Fisher y fundamentada posteriormente por Jerzy Neyman y Karl Pearson.

Mediante esta teoría, se aborda el problema estadístico considerando una hipótesis determinada H_0 y una hipótesis alternativa H_1, y se intenta dirimir cuál de las dos es la hipótesis verdadera, tras aplicar el problema estadístico a un cierto número de experimentos.

Está fuertemente asociada a los considerados errores de tipo I y II en estadística, que definen respectivamente, la posibilidad de tomar un suceso verdadero como falso, o uno falso como verdadero.

Existen diversos métodos para desarrollar dicho test, minimizando los errores de tipo I y II, y hallando por tanto con una determinada potencia, la hipótesis con mayor probabilidad de ser correcta. Los tipos más importantes son los test centrados, de hipótesis y alternativa simple, aleatorizados, etc. Dentro de los tests no paramétricos, el más extendido es probablemente el test de la U de Mann-Whitney. (Wikipedia: Contraste de hipótesis)

4.5. Tres métodos para realizar pruebas de hipótesis

Las pruebas de hipótesis se clasifican como direccionales o no direccionales, dependiendo de cuando la hipótesis nula involucra o no el signo de igualdad (=).

Si la afirmación de H_0 contiene el signo de igualdad, entonces la prueba se llama no direccional, mientras que si tal afirmación no contiene el signo de igualdad (esto es, si involucra los signos menor o mayor que), entonces la prueba se llama direccional. Las pruebas no direccionales se llaman también pruebas de dos colas y las direccionales se nombran pruebas de una cola.

Así, si la afirmación de "H_0" contiene el símbolo ">", entonces la prueba se llama prueba direccional de cola izquierda; por el contrario, Si la afirmación de H_0 tiene el símbolo "<", entonces la prueba se denomina prueba direccional de cola derecha.

Quienes investigan el mercado de consumo tienen una hipótesis alternativa o de investigación: el nuevo producto es superior al anterior. Formalmente, una hipótesis alternativa, denotada con H_1, es un enunciado acerca de la población. La hipótesis nula, denotada con H_0, es la negación de la hipótesis alternativa H_1.

La estrategia básica en las pruebas de hipótesis es tratar de apoyar la hipótesis alternativa “contradiciendo” la hipótesis nula.

4.5.1. El método del intervalo

En contrastes en los que la hipótesis nula es simple y la alternativa es bilateral, se puede utilizar el intervalo de confianza sobre el parámetro para obtener un contraste de nivel, siendo 1- el nivel de confianza del intervalo. Esta práctica es usual en este tipo de contrastes, en los que no existe uno uniformemente más potente.

Por tanto, en estos casos, donde la hipótesis nula es simple y la alternativa bilateral, utilizaremos el intervalo de confianza para determinar el contraste asociado.

Un concepto importante en el planteamiento de la inferencia estadística es la de función de distribución de la muestra, definida en una muestra de tamaño *n* como:
Siendo Ni el número de observaciones muestrales o iguales que xi, es decir, la frecuencia acumulada. Esta función presenta tantos saltos como valores muestrales haya, siendo la cuantía del salto cuando no se repite el valor xi, y cuando xi se repite ni veces, lo que indica que la función de distribución empírica es siempre discreta. (Muñoz, s/f)

4.5.2. El método estadístico de prueba

Los datos se deben sintetizar en un estadístico de la prueba. Dicho estadístico se calcula para ver si es razonablemente compatible con la hipótesis nula. Cuando se prueba una proporción el estadístico de la prueba es muy simple: se cuenta el número de éxitos en la muestra para encontrar el estadístico.

En las pruebas de hipótesis es necesario trazar una línea entre los valores del estadístico de la prueba que son relativamente probables dada la hipótesis nula y los valores que no lo son. ¿En qué valor del estadístico de la prueba comenzamos a decir que los datos apoyan a la hipótesis alternativa?

Para contestar a esta pregunta se requiere conocer la distribución muestral del estadístico de la prueba. Los valores del estadístico de la prueba que son sumamente improbables bajo la hipótesis nula (tal como los determina la distribución muestral) forman una región de rechazo para la prueba estadística.

La región de rechazo es la región asociada al contraste de hipótesis, al conjunto de valores muestrales bajo los cuales se rechaza la hipótesis nula.
Fijada la región de rechazo automáticamente se tiene la regla de decisión. Si nuestra muestra pertenece a la región de rechazo rechazamos H0 y si no, la aceptamos.
Precisamente el objetivo de la teoría de los contrastes o prueba de hipótesis es determinar para cada contraste cuál es la región de rechazo óptima en base a criterios que se especificarán. (Muñoz, s/f)

4.5.3. El método del valor de la P

Al probar hipótesis en las que la estadística de prueba es discreta, la región crítica se puede elegir de forma arbitraria y determinar su tamaño. Si es demasiado grande, se puede reducir al hacer un ajuste en el valor crítico. Puede ser necesario aumentar el tamaño de la muestra para compensar la disminución que ocurre de manera automática en la potencia de la prueba (probabilidad de rechazar Ho dado que una alternativa específica es verdadera).

Por generaciones enteras de análisis estadístico, se ha hecho costumbre elegir un nivel de significancia de 0.05 ó 0.01 y seleccionar la región crítica en consecuencia. Entonces, por supuesto, el rechazo o no rechazo estricto de Ho dependerá de esa región crítica. En la estadística aplicada los usuarios han adoptado de forma extensa la aproximación del valor P. La aproximación se diseña para dar al usuario una alternativa a la simple conclusión de “rechazo” o “no rechazo”.

La aproximación del valor P como ayuda en la toma de decisiones es bastante natural pues casi todos los paquetes de computadora que proporcionan el cálculo de prueba de hipótesis entregan valores de P junto con valores de la estadística de la prueba apropiada.

* Un valor P es el nivel (de significancia) más bajo en el que el valor observado de la estadística de prueba es significativo.

* El valor P es el nivel de significancia más pequeño que conduce al rechazo de la hipótesis nula Ho.

* El valor P es el mínimo nivel de significancia en el cual Ho sería rechazada cuando se utiliza un procedimiento de prueba especificado con un conjunto dado de información. Una vez que el valor de P se haya determinado, la conclusión en cualquier nivel particular resulta de comparar el valor P con α. (Torre, 2003b)

4.6. Prueba de hipótesis sobre una proporción poblacional

Las pruebas de hipótesis se realizan sobre los parámetros poblacionales desconocidos, es decir, sólo tiene sentido realizarlas cuando se estudia una muestra de la población objeto y deseamos hacer inferencias hacia el total poblacional. Si estudiaste al total de los elementos de tú población objeto (definida de acuerdo a los objetivos de tú [sic.] investigación), no tiene sentido realizar PH ni otro tipo de inferencia.

Antes de realizar una prueba de hipótesis, debes revisar cuidadosamente las características de los datos (naturaleza de las variables), la forma de selección de la muestra y su tamaño, en fin, valorar el cumplimiento de los supuestos necesarios para aplicar la prueba adecuada a cada caso. Fijando el nivel de significación antes de realizar la prueba y no después de obtener el resultado, al igual que debes valorar seriamente si debes enunciar el problema de forma bilateral o unilateral antes de realizar la prueba. Violar el cumplimiento de los supuestos implica que la prueba pierda potencia, pudiendo no encontrarse diferencias cuando realmente las hay o lo contrario. (Mitecnológico, Prueba de hipótesis para proporción)

4.7. Pruebas de hipótesis sobre la diferencia entre dos medias

Puesto que deseamos estudiar dos poblaciones, la distribución de muestreo que nos interesa es la distribución de muestreo de la diferencia entre medias muestrales.

Conceptos básicos de las distribuciones de población, distribuciones de muestreo de la media y distribuciones de muestreo de diferencias entre medias muestrales.

Ambas tienen medias y desviaciones estándar, respectivamente, debajo de cada población se muestra distribución de muestreo de la media para esa población. Las dos distribuciones teóricas de muestreo de la media están integradas todas las muestras posibles de determinado tamaño que pueden extraerse de la correspondiente distribución de la población 2, si después restamos las dos medias muestrales, obtenemos la diferencia entre medias muestrales. Esta diferencia será positiva si X1 es mayor que X2 y negativa si X3 es mayor que X1. Al construir esta distribución de todas las diferencias posibles de muestreo de X1 – X2, terminamos teniendo la distribución de muestreo entre las medias muestrales. (Mitecnológico, Prueba de hipótesis para diferencias de medias)

4.8. Pruebas de hipótesis sobre la diferencia entre dos poblaciones

Las pruebas de hipótesis a partir de proporciones se realizan casi en la misma forma utilizada cuando nos referimos a las medias, cuando se cumplen las suposiciones necesarias para cada caso. Pueden utilizarse pruebas unilaterales o bilaterales dependiendo de la situación particular.
La proporción de una población
Las hipótesis se enuncian de manera similar al caso de la media.
Ho: p = p0
H1: p ¹ p0

Regla de decisión: se determina de acuerdo a la hipótesis alternativa, si es bilateral o unilateral

En el caso de muestras pequeñas se utiliza la distribución Binomial. La situación más frecuente es suponer que existen diferencias entre las proporciones de dos poblaciones, para ello suelen enunciarse las hipótesis de forma similar al caso de las medias:
Ho: p1 = p2 Þ p1 - p2 = 0
H1: p1 ¹ p2

Puede la hipótesis alternativa enunciarse unilateralmente.

Siendo a1 y a2, el número de sujetos con la característica objeto de estudio en las muestras 1 y 2 respectivamente, es decir, en vez de calcular la varianza para cada muestra, se calcula una p conjunta para ambas muestras bajo el supuesto que no hay diferencias entre ambas proporciones y así se obtiene la varianza conjunta. Recuerda que q = 1-p.

La regla de decisión se determina de manera similar a los casos ya vistos anteriormente.

El objetivo de la prueba es comparar estas dos proporciones, como estimadores
H1: p1 ¹ p2

Recuerda que la H1 también puede plantearse de forma unilateral. (Mitecnológico, Prueba hipótesis para proporción y diferencia de proporciones)

4.9. Prueba para la diferencia entre dos varianzas

En ocasiones es importante comparar dos poblaciones para ver si una es más variable que la otra en alguna medida específica. La hipótesis nula es que las dos poblaciones tienen la misma varianza, y la hipótesis alternativa es que una tiene mayor varianza que la otra. Se obtienen muestras aleatorias de cada población y se calculan las varianzas muestrales. Estos valores se usan entonces en la ecuación siguiente para calcular el estadístico de la muestra:

Cociente F
S12
F = ---------
S22
Donde:

- S12 = Varianza de la muestra 1
- S22 = Varianza de la muestra 2

Nota: Por convivencia, para encontrar los valores de F, por lo general se pone en el numerador la varianza muestral más grande.

El estadístico de prueba dado por la ecuación anteriormente nombrado, es el cociente *F*. Si la hipótesis nula de varianzas poblacionales iguales es cierta, la razón de las varianzas muestrales se obtiene de la distribución *F* teórica. Al consultar la tabla *F* se puede evaluar la probabilidad de este suceso.

Si parece probable que el cociente *F* pueda haberse obtenido de la distribución muestral supuesta, la hipótesis nula no se rechaza. Si es poco probable que el cociente *F* se haya obtenido de la distribución supuesta, la hipótesis nula se rechaza.

La distribución *F* específica que se aplica a una prueba en particular queda determinada por dos parámetros: los grados de libertad para el numerador y los grados de libertad para el denominador. Cada uno de estos valores es n-1. Si se conocen estos valores y se elige un valor alfa, al valor crítico de *F* se puede encontrar en la tabla *F*. (Hereas, s/f)

RESUMEN DE LA UNIDAD

Las pruebas de hipótesis, como herramienta estadística, son importantes porque nos indican el camino, al aceptar o desechar un hipótesis de manera tentativa a favor de otra, sin embargo no aportan mayor información; pero si apoyamos nuestra decisión con un intervalo de confianza apropiado, podemos obtener datos que pueden ser transformados en información y utilizarlos como sustento de una decisión que generalmente en cualquier ámbito representa dinero. Evidentemente se debe de tomar en consideración todos los errores posibles que se puedan cometer durante el proceso, de donde nacen los errores tipo i y II para las pruebas de hipótesis, además de la potencia de una prueba de hipótesis para que nuestra opinión sea lo más certera posible.

GLOSARIO DE LA UNIDAD

Curva de la potencia de la prueba

Es la gráfica de la probabilidad de rechazar H_0 para todos los valores posibles del parámetro poblacional que no satisfacen la hipótesis nula.

Error tipo I

Es el error que se comete al rechazar H_0 cuando ésta es verdadera.

Error tipo II

Es el error que se comete al aceptar H_0 cuando ésta es falsa.

Estadístico de prueba

Es el estadístico cuyo valor se utiliza para determinar si se rechaza una hipótesis nula.

Hipótesis nula (H_0)

Es la hipótesis que tentativamente se supone que es verdadera en una prueba de hipótesis.

Hipótesis alternativa (H_1) o hipótesis de estudio

Es la hipótesis que se desea comprobar y que se concluye como verdadera cuando se rechaza la hipótesis nula.

Nivel de significancia

Es la probabilidad máxima de cometer un error tipo I.

Potencia de la prueba

Es la probabilidad de rechazar correctamente H_0 cuando es falsa.

Prueba direccional o de una cola

Prueba de hipótesis en la que la región de rechazo se tiene en un extremo de la distribución muestral.

Prueba no direccional o de dos colas

Prueba de hipótesis en la que la región de rechazo se ubica en ambos extremos de la distribución muestral.

Región de rechazo

Es la zona de valores en la cual se rechaza la hipótesis H_0.

Valor crítico

Es un valor contra el cual se compara el obtenido en el estadístico de prueba para determinar si se debe rechazar o no la hipótesis nula.

Valor p

Es la probabilidad de que, cuando la hipótesis nula sea verdadera, se obtenga un resultado de una muestra que sea al menos tan improbable como el que se observa. También se le conoce como nivel observado de significancia.

ACTIVIDADES DE APRENDIZAJE

ACTIVIDAD 1

Explica lo que entiendes por hipótesis nula e hipótesis alternativa.

ACTIVIDAD 2

Considerando únicamente la hipótesis nula y la hipótesis alternativa. Cuántos tipos de hipótesis hay y explícalas.

CUESTIONARIO DE REFORZAMIENTO

1. Una hipótesis nula se establece como:
2. El nivel de significancia en una prueba de hipótesis es:
3. Un estadístico de prueba en una prueba de hipótesis es:
4. ¿Cuáles son las etapas básicas en pruebas de hipótesis?
5. En una prueba de una cola el signo de la hipótesis nula puede ser:
6. El nivel de significancia en una prueba de hipótesis corresponde a:

7. Un artículo de prensa señaló que la edad promedio de los accionistas de empresas está decreciendo. El gerente de una de ellas decide realizar una prueba de hipótesis para verificar si este señalamiento aplica a su empresa. Se considera una desviación estándar de 12 años y una muestra de tamaño 250, cuya media muestral es de 53 años. Para un nivel de significancia del 5%, ¿cuál es el valor crítico para la prueba?
8. La Ingeniería de Control de Calidad probó un lote de tubos fluorescentes y encontró una vida promedio de 1,570 horas con desviación estándar de 120 horas. Con un nivel de significación del 5%, determinar la regla de decisión.
9. Se prueba un lote de un nuevo modelo experimental de 100 lámparas de vapor de sodio; su vida es de 43,000 horas y su desviación estándar, de 2,000 horas. Si la vida normal de las lámparas es de 40,000 horas. Probar con un nivel de significación del 10%
10. En una planta embotelladora de leche se toma una muestra de 500 botellas; 40 de ellas se obtienen con impurezas. Si se supone que el límite máximo de impurezas es 7%. Establece la regla de decisión para un nivel de significancia del 4%

EXAMEN DE AUTOEVALUACIÓN

Elige la respuesta correcta a las siguientes preguntas.

1. Supón que formas parte de un grupo de protección al consumidor, y estás interesado en determinar si el peso promedio de cierta marca de arroz, empacado en paquetes de 1 kg, es menor que el peso anunciado; para ello, eliges una muestra aleatoria de 50 bolsas, de las cuales obtienes una media de 980 gr. y una desviación estándar de 70 gr. Para un nivel de significancia es del 5%, la hipótesis nula se:
   a) acepta
   b) es indiferente
   c) rechaza
   d) debe replantear

2. Se supone que un medicamento que sirve como antibiótico contiene 1000 unidades de penicilina. Una muestra aleatoria de 100 de estos antibióticos produjo una media de 1020 gramos y una desviación estándar de 140 gramos. Para un nivel de significancia del 5%, la hipótesis nula se:
   a) acepta
   b) rechaza
   c) es indiferente
   d) replantea

3. Se sabe que los voltajes de una marca de pilas “AAA” para calculadora se distribuyen normalmente con un promedio de 1.5 volts; se probó una muestra aleatoria de 15 y se encontró que la media fue de 1.3 volts y que la desviación estándar fue de 0.25 volts. Para un nivel de significancia del 5%, la hipótesis nula se:
   a) acepta
   b) rechaza
   c) es indiferente
   d) replantea

LO QUE APRENDÍ

Elige un tipo de empresa comercial. Elabora una propuesta del procedimiento general que se deberá realizarse para el desarrollo de un software que lleve el control de sus ventas.

MESOGRAFÍA

Bibliografía sugerida

Autor	Capitulo	Páginas
Levin y otros (1996)	9	359-390
Berenson y otros (2001)	11, 12	384-460
Christensen (1990)	8	378-458
Lind y otros (2004)	10	331-353

Bibliografía básica

Bruegge, Bernd. (2001). *Ingeniería de software orientada a objetos*. México: Prentice Hall.

Joyanes, Luis. (2003). *Fundamentos de programación Algoritmos Estructuras de datos y objetos*. (3ª ed.) Madrid: McGraw-Hill.

Kohler, Heinz. (1996). *Estadística para negocios y economía*. México: CECSA.

Pfleeger, Shari Lawrence. (2002). *Ingeniería de software, Teoría y práctica*. México: Prentice Hall.

Piattini, Mario y Félix García (coords.) (2003). *Calidad en el desarrollo y mantenimiento de software*. México: Alfa omega / Ra-Ma.

Piattini, M., *et al*. (2003). *Análisis y diseño de aplicaciones informáticas de gestión*. México: Alfa omega / Ra-Ma.

Pressman, Roger S. (2002). *Ingeniería de software.* (5ª. ed.) México: McGraw-Hill.

Sommerville, Ian (2001). *Ingeniería de software.* (6ª ed.) México: Addison Wesley.

Weitzenfield, Alfredo. (2003). *Ingeniería de software orientada a objetos con UML, Java e Internet*. México: Thomson.

Bibliografía complementaria

Brown, David. (1997). *Object-Oriented Analysis.* USA: John Wiley & Sons.

Dennis, Alan (2000). *Systems Analysis and Design and applied approach.* USA: John Wiley & Sons.

Ince, Darrel (1993). *Ingeniería de Software.* México: Addison-Wesley.

Kendall, Kenneth (1990). *Análisis de diseño de sistemas.* México: Prentice Hall.

Larman Craig (1999). *UML y patrones.* México: Prentice-Hall.

Márquez Vite, Juan Manuel (2002). *Sistemas de información por computadora, Metodología de desarrollo.* México: Trillas.

Meyer, Bertrand (1999). *Construcción de Software Orientado a Objetos*. Madrid: Prentice-Hall.

Sitios de Internet

Sitio	Descripción
http://www.monografias.com/trabajos30/prueba-de-hipotesis/prueba-de-hipotesis.shtml	Cruz Ramírez, Armando Pedro. (2009). "Pruebas de hipótesis para una muestra". Monografías
http://html.rincondelvago.com/analisis-de-la-varianza_1.html	Hereas, "Análisis de la varianza", Rincón del vago
http://www.mitecnologico.com/Main/PruebaHipotesisParaProporcionYDiferenciaDeProporciones	Mitecnológico, "Prueba hipótesis para proporción y diferencia de proporciones"
http://www.mitecnologico.com/Main/PruebaDeHipotesisParaDiferenciasDeMedias	Mitecnológico (4.3.2) Prueba de hipótesis para diferencias de medias
http://html.rincondelvago.com/contraste-de-hipotesis_1.html	Muñoz, Gonzalo. (s/f). Contraste de hipótesis. Rincón del vago.

http://www.itch.edu.mx/academic/industrial/estadistica1/cap02c.html#u02usovaloresp	Torre, Leticia de la. (2003b). "Uso de valores P para la toma de decisiones", Estadística I, Instituto Tecnológico de Chihuahua
http://es.wikipedia.org/wiki/Contraste_de_hip%C3%B3tesis	Wikipedia: "Contraste de hipótesis", actualizado el 13/10/11
lc.fie.umich.mx/~jrincon/pruebas%20de%20hipotesis.ppt	Rincón Pasaye, José Juan. (2008) "Pruebas de hipótesis", Probabilidad y estadística, [diapositivas] UMICH
www.cyta.com.ar/biblioteca/bddoc/bdlibros/guia_estadistica/modulo_9.htm	Ciencia y Técnica Administrativa. (2005). "Módulo 9. Pruebas de hipótesis, muestras grandes", Guía de Estadísticas
http://www.geociencias.unam.mx/~ramon/Estadistica/Clase5b.pdf	Zúñiga, F. Ramón. (2008). "Clase 5. Pruebas de hipótesis", Estadística, Querétaro: Geociencias, UNAM
http://uvigen.fcien.edu.uy/utem/genmen/06chi2.htm	"La prueba de Chi-cuadrado", Genética Mendeliana, UVIGEN, Universidad de la República, Montevideo. (Traducción de McClean, Phillip, 2000 *)

UNIDAD 5

PRUEBAS DE HIPÓTESIS CON LA DISTRIBUCIÓN JI CUADRADA

OBJETIVO ESPECÍFICO

Al terminar la unidad el alumno relacionará los conceptos de prueba de hipótesis con la distribución ji cuadrada.

INTRODUCCIÓN

En esta unidad, el alumno investigará y analizará el concepto de prueba de hipótesis y lo aplicará sobre varianzas, medias, etc.; ello le permitirá percatarse de la importancia que tienen las pruebas de hipótesis para la toma de decisiones dentro de las empresas.

Actualmente, sabemos que la matemática es una herramienta importante en la toma de decisiones, y la estadística junto con todos sus procesos no es la excepción; así, es importante que el alumno desarrolle todos los conceptos y ejercicios planteados en la presente unidad, enriqueciendo su cultura para su futuro desempeño profesional.

Sabemos que cuando las personas toman decisiones, inevitablemente lo hacen con base en las creencias que tienen en relación con el mundo que los rodea; llevan en la mente una cierta imagen de la realidad, piensan que algunas cosas son verdaderas y otras falsas y actúan en consecuencia, así, los ejecutivos de empresas toman todos los días decisiones de importancia crucial porque tienen ciertas creencias tales como:

- De que un tipo de máquina llenadora pone al menos un kilogramo de detergente en una bolsa.
- De que cierto cable de acero tiene una resistencia de 100 kg. o más a la rotura.
- De que la duración promedio de una batería es igual a 500 horas.
- De que en un proceso de elaboración de cápsulas éstas contengan precisamente 250 miligramos de un medicamento.
- Que la empresa de transportes de nuestra competencia tiene tiempos de entrega más rápidos que la nuestra.
- De que la producción de las plantas de oriente contiene menos unidades defectuosas que las de occidente.

Incluso los estadistas basan su trabajo en creencias tentativas:

- Que dos poblaciones tienen varianzas iguales.
- Que esta población está normalmente distribuida.
- Que estos datos muestrales se derivan de una población uniformemente distribuida.

En todos estos casos y en muchos más, las personas actúan con base en alguna creencia sobre la realidad, la cual quizá llegó al mundo como una simple conjetura, como un poco más que una suposición informada; una proposición adelantada tentativamente como una verdad posible es llamada hipótesis.

Sin embargo, tarde o temprano, toda hipótesis se enfrenta a la evidencia que la comprueba o la rechaza y, en esta forma, la imagen de la realidad cambia de mucha a poca incertidumbre.

Por lo tanto, de una manera sencilla podemos decir que una prueba de hipótesis es un método sistemático de evaluar creencias tentativas sobre la realidad, dicho método requiere de la confrontación de tales creencias con evidencia real y decidir, en vista de esta evidencia, si dichas creencias se pueden conservar como razonables o deben desecharse por insostenibles.

A continuación estudiaremos la forma en que las creencias de las personas pueden probarse de manera sistemática.

LO QUE SÉ

Elige la respuesta correcta a las siguientes preguntas.

1. La distribución chi-cuadrada χ^2 es útil para analizar la relación:

   a) entre la varianza de la muestra y la varianza de la población

   b) entre la media de la muestra y la media de la población

   c) entre una muestra y otra

2. La fórmula para calcular la media aritmética de una muestra es:

   a) $\chi^2 = \frac{s^2(gl)}{\sigma^2}$

   b) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

   c) $\frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}$

3. La fórmula para calcular la varianza de una muestra es:

   a) $\frac{s^2(n-1)}{\chi^2_{\alpha/2}}$

   b) $\frac{s^2(n-1)}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}$

   c) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

TEMARIO DETALLADO

(8 horas)

5.1. La distribución ji cuadrada, χ^2

5.2. Pruebas de hipótesis para la varianza de una población

5.3. Prueba para la diferencia entre n proporciones

5.4. Pruebas de bondad de ajuste a distribuciones teóricas

5.5. Pruebas sobre la independencia entre dos variables

5.6. Pruebas de homogeneidad

5.1. La distribución ji cuadrada, χ^2

En ocasiones los investigadores muestran más interés en la varianza poblacional que en la proporción o media poblacionales y las razones llegan desde el campo de la calidad total, por ejemplo, donde la importancia en demostrar una disminución continua en la variabilidad de las piezas que la industria de la aviación llega a solicitar es de vital importancia. Por ejemplo, el aterrizaje de un avión depende de una gran cantidad de variables, entre las que encontramos la velocidad y dirección del aire, el peso del avión, la pericia del piloto, la altitud, etc.; si en el caso de la altitud, los altímetros del avión tienen variaciones considerables, entonces podemos esperar con cierta probabilidad un aterrizaje algo abrupto, por lo tanto la variabilidad de estos altímetros debe mostrar un disminución continua; y qué decir de los motores que impulsan al avión mismo, si las piezas que los conforman son demasiado grandes, el motor puede incluso no poder armarse y si son demasiado pequeñas, entonces los motores tendrán demasiada vibración y en ambos casos las pérdidas de la industria son cuantiosas.

Así, la relación entre la varianza de la muestra y la varianza de la población está determinada por la distribución Chi-cuadrada (χ^2) siempre y cuando la población de la cual se toman los valores de la muestra se encuentre normalmente distribuida.

Y aquí debemos tener especial cuidado, pues la distribución Chi-cuadrada es sumamente sensible a la suposición de que la población está normalmente distribuida y por ejemplo construir intervalos de confianza para estimar una varianza poblacional, puede que los resultados no sean correctos dependiendo de si la población no está normalmente distribuida.

La distribución Chi-cuadrada (χ^2) es la razón que existe entre la varianza de la muestra (s^2) multiplicada por los grados de libertad y la varianza de la población. Es decir:

$$\chi^2 = \frac{s^2(gl)}{\sigma^2}$$

El término 'grados de libertad' (Black, 2005, p. 264) se refiere al número de observaciones independientes para una fuente de variación menos el número de parámetros independientes estimado al calcular la variación.

Para la distribución Chi-cuadrada (χ^2), los grados de libertad vienen dados por (n – 1), por lo tanto, la fórmula anterior quedaría expresada como:

$$\chi^2 = \frac{s^2(n-1)}{\sigma^2}$$

Donde podemos observar que la variación de la distribución Chi-cuadrada (χ^2) depende del tamaño de la muestra y de los grados de libertad que posea.

En general y debido a que la distribución Chi-cuadrada (χ^2) no es simétrica a medida que se incrementa el número de grados de libertad, la curva característica de la distribución se vuelve menos sesgada.

La distribución Chi-cuadrada (χ^2) es en sí toda una familia de distribuciones por lo que, existe una distribución Chi-cuadrado para cada grado de libertad.

Algebraicamente podemos manipular la fórmula anterior $\chi^2 = \frac{s^2(n-1)}{\sigma^2}$ con el objetivo de que nos sea de utilidad para construir intervalos de confianza para varianzas poblacionales, quedando de la siguiente manera:

$$\frac{s^2(n-1)}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}$$

Ejemplo

Supóngase que una muestra de 7 pernos especiales utilizados en el ensamblado de computadoras portátiles arrojó los siguientes resultados:

2.10 mm; 2.00 mm, 1.90 mm, 1.97 mm, 1.98 mm, 2.01 mm, 2.05 mm

Si quisiéramos una estimación puntual de la varianza de la población, sería suficiente con calcular la varianza de la muestra, de la siguiente manera:

Primero calculamos la media aritmética de los datos utilizando la siguiente fórmula:

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

por lo tanto sustituyendo datos tenemos que:

$$\bar{X} = \frac{2.10 + 1.90 + 1.98 + 2.05 + 2.00 + 1.97 + 2.01}{7}$$

y al efectuar cálculos, el resultado de la media aritmética (redondeado a 2 decimales) es de:

$$\bar{X} = 2.00$$

a continuación elaboramos una tabla para facilitar el cálculo de la varianza de los datos:

I-dato	DATOS	Dato-media	(Dato - media) elevado al cuadrado
I	x_i	$(x_i - \mu)$	$(x_i - \mu)^2$
1	2,10	0,10	0,00972

2	1,90	-0,10	0,01029
3	1,98	-0,02	0,00046
4	2,05	0,05	0,00236
5	2,00	0,00	0,00000
6	1,97	-0,03	0,00099
7	2,01	0,01	0,00007
	14,01	0,01	0,02389

Recordando ahora la fórmula correspondiente a la varianza de una muestra:

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

y sustituyendo datos en esta fórmula, podemos ver que el valor obtenido en la esquina inferior derecha de la tabla anterior corresponde a:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2$$

por lo tanto:

$$s^2 = \frac{1}{7-1}(0.02389)$$

de donde al efectuar cálculos vemos que:

$$s^2 = 0.003981$$

Es decir, la varianza de la muestra tiene un valor de: 0.003981, pero si consideramos que el valor de la estimación puntual puede cambiar de una muestra a otra, entonces será mejor construir un intervalo de confianza, para lo cual debemos suponer que la población de los diámetros de los pernos está normalmente distribuida, y como vemos que n=7 entonces los

grados de libertad serán: gl=7-1=6, si queremos que el intervalo sea del 90% de confianza, entonces el nivel de significancia α será de 0.10 siendo esta la parte del área bajo la curva de la distribución Chi-cuadrada que está fuera del intervalo de confianza, esta área es importante porque los valores de la tabla de distribución Chi-cuadrada están dados de acuerdo con el área de la cola derecha de la distribución. Además en nuestro caso $\alpha/2$ = 0.05 es decir, 0.05 del área está en la cola derecha y 0.05 está en la cola izquierda de la distribución.

Es importante hacer notar que debido a la forma de curva de la distribución Chi-cuadrada, el valor para ambas colas será diferente, así, el primer valor que se debe obtener es el de la cola derecha, mismo que se obtiene al ubicar en el primer renglón de la tabla el valor correspondiente al nivel de significancia, que en este caso es de 0.05 y, posteriormente se ubica en el lugar de las columnas los correspondientes grados de libertad ya calculado, que en este caso es de 6 grados de libertad, por lo tanto el valor de Chi-cuadrada obtenido es de:

$$\chi^2_{0.05,6} = 12.5916$$

Obsérvese que en la nomenclatura se escribe la denotación de Chi-cuadrada teniendo como subíndice el nivel de significancia y los grados de libertad y, a continuación se escribe el valor correspondiente (véase, Black, 2005, p. 779).

El valor de Chi-cuadrada para la cola izquierda se obtiene al calcular el área que se encuentra a la derecha de la cola izquierda, entonces:

A a la derecha de la cola izquierda = 1 – 0.05

A a la derecha de la cola izquierda = 0.95

por lo tanto, el valor de Chi-cuadrada para la cola izquierda será, utilizando

el mismo procedimiento anterior para un área de 0.95 y 6 grados de libertad, de:

$$\chi^2_{0.95,6} = 1.63538$$

incorporando estos valores a la fórmula, tenemos que el intervalo de 90% de confianza para los 7 pernos utilizados en el ensamblado de computadoras portátiles tendrá la forma mostrada a continuación:

$$\frac{s^2(n-1)}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}$$

$$\frac{0.0034122(7-1)}{12.5916} \leq \sigma^2 \leq \frac{0.0034122(7-1)}{1.63538}$$

$$0.0001625 \leq \sigma^2 \leq 0.0125189$$

Este intervalo de confianza nos dice que con 90% de confianza, la varianza de la población está entre 0.0001625 y 0.0125189.

La prueba estadística de X^2 para una muestra se emplea frecuentemente como prueba de bondad de ajuste, sin embargo, en un plan experimental en el que se cuenta con un grupo muestral, con diversas subclases y las mediciones están en escala nominal, resulta muy útil este procedimiento.

La eficacia de la prueba está de acuerdo con el tamaño de la muestra, pues con un grado de libertad, si hay dos subclases, algunos autores consideran que la prueba es insensible, no obstante la información que aporta más de dos categorías es satisfactoria en función de la fórmula:

$$X^2 = \Sigma^H_{N=1} \frac{(fo - fe)^2}{fe}$$

Donde:

χ^2 = valor estadístico de ji cuadrada.

fo = frecuencia observada.

fe = frecuencia esperada.

La ji cuadrada se utiliza cuando:

- Cuando los datos puntualizan a las escalas nominal u ordinal.
- Se utiliza solo la frecuencia.
- Poblaciones pequeñas.
- Cuando se desconocen los parámetros media, moda, etc.
- Cuando los datos son independientes.
- Cuando se quiere contrastar o comparar hipótesis.
- Investigaciones de tipo social -muestras pequeñas no representativas >5.
- Cuando se requiere establecer el nivel de confianza o significatividad en las diferencias.
- Cuando la muestra es seleccionada no probabilísticamente.
- X^2 permite establecer diferencias entre f y se utiliza solo en escala nominal.
- Población > a 5 y < a 20.

Pasos

1. Arreglar las categorías y las frecuencias observadas.
2. Calcular los valores teóricos esperados para el modelo experimental o tipo de distribución muestral: normal, binomial y de Poisson.
3. Calcular las diferencias de las frecuencias observadas en el experimento con respecto a las frecuencias esperadas.
4. Elevar al cuadrado las diferencias y dividirlas entre los valores esperados de cada categoría.
5. Efectuar la sumatoria de los valores calculados.

6. Calcular los grados de libertad (gl) en función de número de categorías [K]: gl = K - 1.
7. Comparar el estadístico X^2 con los valores de la distribución de ji cuadrada en la tabla.
8. Decidir si se acepta o rechaza la hipótesis X^2c ³ X2t se rechaza Ho.

5.2. Pruebas de hipótesis para la varianza de una población

En ocasiones analistas investigan la variabilidad de una población, en lugar de su media o proporción.

Esto es debido a que la uniformidad de la producción muchas veces es crítica en la práctica industrial.

La variabilidad excesiva es el peor enemigo de la alta calidad y la prueba de hipótesis está diseñada para determinar si la varianza de una población es igual a algún valor predeterminado.

La desviación estándar de una colección de datos se usa para describir la variabilidad en esa colección y se puede definir como la diferencia estándar entre los elementos de una colección de datos y su media.

La varianza de un conjunto de datos se define como el cuadrado de su desviación estándar; y la varianza muestral se utiliza para probar la hipótesis nula que se refiere a la variabilidad y es útil para entender el procedimiento de análisis de la varianza.

La hipótesis nula; para la prueba de la varianza, es que la varianza poblacional es igual a algún valor previamente especificado. Como el aspecto de interés, por lo general es si la varianza de la población es mayor que este valor, siempre se aplica una de una cola.

Para probar la hipótesis nula, se toma una muestra aleatoria de elementos de una población que se investiga; y a partir de esos datos, se calcula el estadístico de prueba.

Por ejemplo si se desea averiguar si la variabilidad de edades en una comunidad local es la misma o mayor que la de todo el Estado. La desviación estándar de las edades del Estado, conocida por un estudio reciente es de 12 años. Tomamos una muestra aleatoria de 25 personas de la comunidad y determinamos sus edades. Calcular la varianza de la muestra y usar la ecuación anteriormente explicada para obtener el estadístico muestral.

5.3. Prueba para la diferencia entre n proporciones

Las pruebas de hipótesis a partir de proporciones se realizan casi en la misma forma utilizada cuando nos referimos a las medias, cuando se cumplen las suposiciones necesarias para cada caso. Pueden utilizarse pruebas unilaterales o bilaterales dependiendo de la situación particular.

La proporción de una población
Las hipótesis se enuncian de manera similar al caso de la media.
Ho: p = p0
H1: p ¹ p0

Regla de decisión: se determina de acuerdo a la hipótesis alternativa (si es bilateral o unilateral), lo cual puedes fácilmente hacerlo auxiliándote de la tabla 4.4.1.

En el caso de muestras pequeñas se utiliza la distribución Binomial. La situación más frecuente es suponer que existen diferencias entre las proporciones de dos poblaciones, para ello suelen enunciarse las hipótesis de forma similar al caso de las medias:
Ho: p1 = p2 Þ p1 - p2 = 0
H1: p1 ¹ p2

Puede la hipótesis alternativa enunciarse unilateralmente. (Mitecnológico, Prueba de hipótesis para proporción).

5.4. Pruebas de bondad de ajuste a distribuciones teóricas

Una hipótesis estadística se definió como una afirmación o conjetura acerca de la distribución f(x,q) de una o más variables aleatorias. Igualmente se planteó que la distribución podía tener uno o más parámetros desconocidos, que denotamos por q y que la hipótesis se relaciona con este parámetro o conjunto de parámetros En otros casos, se desconoce por completo la forma de la distribución y la hipótesis entonces se relaciona con una distribución específica f(x,q) que podamos asignarle al conjunto de datos de la muestra. El primer problema, relacionado con los parámetros de una distribución conocida o supuesta es el problema que hemos analizado en los párrafos anteriores. Ahora examinaremos el problema de verificar si el conjunto de datos se puede ajustar o afirmar que proviene de una determinada distribución. Las pruebas estadísticas que tratan este problema reciben el nombre general de “Pruebas de Bondad de Ajuste”.

Se analizarán dos pruebas básicas que pueden aplicarse: La prueba Chi-Cuadrado y la prueba de Smirnov-Kolmogorov. Ambas pruebas caen en la categoría de lo que en estadística se denominan pruebas de “Bondad de Ajuste” y miden, como el nombre lo indica, el grado de ajuste que existe entre la distribución obtenida a partir de la muestra y la distribución teórica que se supone debe seguir esa muestra.

Ambas pruebas están basadas en la hipótesis nula de que no hay diferencias significativas entre la distribución muestral y la teórica. Ambas pruebas están basadas en las siguientes hipótesis:

H0: $f(x,q) = f_0(x,q)$

H1: $f(x,q) \neq f_0(x,q)$

Donde $f_0(x, q)$ es la distribución que se supone sigue la muestra aleatoria. La hipótesis alternativa siempre se enuncia como que los datos no siguen la distribución supuesta. Si se desea examinar otra distribución específica, deberá realizarse de nuevo la otra prueba suponiendo que la hipótesis nula es esta nueva distribución. Al especificar la hipótesis nula, el conjunto de parámetros definidos por q puede ser conocido o desconocido. En caso de que los parámetros sean desconocidos, es necesario estimarlos mediante alguno de los métodos de estimación analizados con anterioridad.

Para formular la hipótesis nula deberán tenerse en cuenta los siguientes aspectos o criterios:

a) La naturaleza de los datos a analizar. Por ejemplo, si tratamos de investigar la distribución que siguen los tiempos de falla de unos componentes, podríamos pensar en una distribución exponencial, o una distribución gama o una distribución Weibull, pero en principio no consideraríamos una distribución normal. Si estamos analizando los caudales de un río en un determinado sitio, podríamos pensar en una distribución logarítmica normal, pero no en una distribución normal.

b) Histograma. La forma que tome el histograma de frecuencia es quizás la mejor indicación del tipo de distribución a considerar. (Mitecnológico, Prueba de bondad de ajuste)

5.4.1 Ajuste a una distribución normal

Con frecuencia conviene saber si puede suponerse que una serie de datos obtenidos experimentalmente proceden de una población distribuida normalmente

Recordemos que en una distribución normal:

- el 68% de los datos está en el intervalo (x-s, x+s)
- el 95% de los datos está en el intervalo (x-2s, x+2s)
- el 99% de los datos está en el intervalo (x-3s, x+3s)

Si calculadas la media x y la desviación típica s de nuestros datos, se cumplen aproximadamente estos porcentajes podemos considerar que la población de partida es normal.

Veamos un ejemplo en el que seguimos un proceso un poco más elaborado: (García Cebrian, 2001b, Ajuste de un conjunto de datos a una normal)

5.4.2 Ajuste a una distribución Poisson

Este proceso se puede utilizar con cualquier tipo de variable en escala nominal u ordinal y sirve para cualquier tipo de distribución.

El Prueba está basada en la distribución chi cuadrado (χ^2) y fue creada por uno de los más reputados estadísticos de los últimos tiempos, Karl Pearson. Su base, como en todos los test de hipótesis, consiste en establecer dos hipótesis, la hipótesis nula que considera que los datos que tenemos se ajustan a una determinada distribución y la hipótesis alternativa que es la negación de la nula, es decir, nuestros datos no se ajustan a la distribución. Dicho así no parece muy claro, pero es como se suele explicar la teoría.

Tenemos unos datos que *'parece'* que siguen una determinada distribución, pero hay unas diferencias entre los datos que tenemos (observados) y los que deberían de ser (esperados). ¿Son esas diferencias lo suficientemente grandes para que sean provocadas por el azar. La respuesta a esta pregunta la obtendremos con la prueba de bondad de ajuste.

Si nuestros datos no siguen la distribución de Poisson, todas las predicciones que hagamos utilizando las fórmulas para esta distribución serán erróneas y si nos basamos en ellos, tenemos muchas posibilidades de error.

Ejemplo, si tomamos los datos del total de goles marcados por partido en la primera división durante la temporada 2010-2011, de estas estadísticas tendremos el resumen de los datos que necesitamos. Estos serían nuestros valores 'Observados'. El siguiente paso que debemos hacer es calcular la media de los goles totales marcados por partido. Al tener los datos resumidos no podemos utilizar la función *promedio ()* si no que debemos hacer una especie de 'desagrupamiento'.

La primera es crear una nueva columna en la que multiplicaremos el número de goles por la cantidad de partidos (Columna C). Sumaremos todos esos productos y dividiremos este valor por el total de partidos jugados.

Una vez calculada la media, lo que hacemos es determinar los valores 'Esperados' según una distribución de Poisson con esa media. Esto lo calculamos multiplicando la probabilidad de Poisson para cada resultado, por el total de partidos.

Para calcular el estadístico χ^2 con la siguiente fórmula:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

Nos proporciona información de donde se producen las mayores discrepancias. Cuanto mayor sea el valor que obtengamos, mayor es la discrepancia entre el valor observado y el esperado. Más alejado está ese punto de su lugar teórico predicho por la curva de Poisson y más probabilidad tenemos que el resultado de la prueba nos diga que nuestros datos no se ajustan bien a la curva.

Finalmente se suman todos estos valores y se busca dentro de la función χ^2 y comprobar si las diferencias que hemos encontrado son lo suficientemente grandes o no para rechazar o no rechazar la hipótesis nula. Este es un error muy común en la interpretación de los resultados de prueba de este tipo. La función χ^2 tiene dos parámetros, el primero de ellos es el valor de nuestra suma, y el segundo son los grados de libertad.

Los grados de libertad se obtienen con la siguiente fórmula: GL = Nc - Np – 1

Siendo Nc = al número de categorías que tenemos y Np = número de parámetros que se está estimando. Para nuestro caso tenemos 10 categorías y se va a estimar un solo parámetro que es la media: GL = 10 - 1 - 1 = 8

El valor que nos devuelve es lo que en estadística se llama P-Value, y corresponde a la probabilidad de equivocarnos si rechazamos la hipótesis nula. Como norma general se suele tomar como valores de corte el 5% ó el 1% dependiendo de lo restrictivos que seamos. Este valor se debe de tomar antes de la realización de la prueba y es el límite para rechazar o no rechazar la hipótesis nula.

En el ejemplo se tiene un P-Value de 0.54 con lo que se debe decir que las diferencias que se han encontrado no son lo suficientemente grandes como para decir que nuestros datos no siguen una distribución de Poisson (Buzjss, 2008)

5.4.3 Ajuste a una distribución binomial

Ajuste de una serie de datos a una distribución binomial:
Disponemos de una serie de k datos que toman los valores
0, 1, ... ,n.

Para saber si estos datos siguen pueden aproximarse por una distribución binomial:

1. Calculamos la media de los k datos y la igualamos a la esperanza teórica de la Binomial (n ·p). Despejamos de aquí el valor de p.

2. Calculamos los valores teóricos de p (X = r), multiplicándolos por k para obtener los valores teóricos de cada posible valor de la variable aleatoria en series de k datos.
3. Si la diferencia es "suficientemente pequeña" aceptamos como buena la aproximación Binomial, si no, la rechazamos.

Nota: La fundamentación estadística que nos permitiría decidir de manera objetiva si la diferencia entre los datos teóricos y los reales es "suficientemente pequeña" escapa de los objetivos de esta unidad didáctica, con lo cual la decisión se deberá tomar de manera subjetiva. (Martín, 2001)

5.5. Pruebas sobre la independencia entre dos variables

Cuando cada individuo de la población a estudio se puede clasificar según dos criterios A y B, admitiendo el primero a posibilidades diferentes y b el segundo, la representación de las frecuencias observadas en forma de una matriz a x b recibe el nombre de Tabla de contingencia.

La hipótesis nula a contrastar admite que ambos caracteres, A y B, se presentan de forma independiente en los individuos de la población de la cual se extrae la muestra; siendo la alternativa la dependencia estocástica entre ambos caracteres. La realización de esta prueba requiere el cálculo del estadístico.

El estadístico L se distribuye como una con (a - 1)(b - 1) grados de libertad. El contraste se realiza con un nivel de significación del 5%.

Para estudiar la dependencia entre la práctica de algún deporte y la depresión, se seleccionó una muestra aleatoria simple de 100 jóvenes, con los siguientes resultados:

	Sin depresión	Con depresión	total
Deportista	38	9	47
No. deportista	31	22	53
	69	31	100

L = (38 – 32,43)2/32,43 + (31 – 36,57)2/36,57 + (9 – 14,57)2/14,57 + (22 – 16,43)2/16,43
= 0,9567 + 0,8484 + 2,1293 + 1,8883 = 5,8227

El valor que alcanza el estadístico L es 5,8227. Buscando en la tabla teórica de Chi Cuadrado para 1 grado de libertad se aprecia Lt = 3,84146 < 5,8227 lo que permite rechazar la hipótesis de independencia de caracteres con un nivel de significación del 5%, admitiendo por tanto que la práctica deportiva disminuye el riesgo de depresión. (Mitecnológico, Prueba de hipótesis para Proporción)

5.6. Pruebas de homogeneidad

Se plantea el problema de la existencia de homogeneidad entre r poblaciones, para lo cual se realizan muestras independientes en cada una de ellas. Los datos muestrales vienen clasificados en s clases y sus frecuencias absolutas se presentan en forma de una matriz r x s.

El estadístico L se distribuye como una con (r - 1)(s - 1) grados de libertad. El contraste se realiza con un nivel de significación del 5%.

Un estudio sobre caries dental en **niños** de seis ciudades con diferentes cantidades de flúor en el suministro de **agua**, ha proporcionado los resultados siguientes:

Comunidad	Nº niños sin caries	Nº niños con caries	
A	38	87	125
B	8	117	125
C	30	95	125
D	44	81	125
E	64	61	125
F	32	93	125
	216	534	750

L = (38 – 36)2/36 + (8 – 36)2/36 + (30 – 36)2/36 + (44 – 36)2/36 + (64 – 36)2/36 + (32 – 36)2/36 + (87 – 89)2/89 (117 – 89)2/89 + (95 – 89)2/89 + (81 – 89)2/89 + (61 – 89)2/89 + (93 – 89)2/89
L = 0,1111 + 21,7778 + 1,0000 + 1,7778 + 21,7778 + 0,4444 + 0,0449 + 8,8089 + 0,4045 + 0,7191 + 8,8089 + 0,1797
L = 65,85
Se quiere saber si la incidencia de caries infantil es igual en las seis poblaciones.

La propia tabla hace pensar que la incidencia de la enfermedad no es igual en todas las poblaciones; basta observar los datos correspondientes a las comunidades B y E. El contraste arroja un valor del estadístico L de 65,85, lo que lleva a rechazar la hipótesis de homogeneidad y aceptar que el diferente contenido de flúor en el suministro del agua puede ser la causa de la disparidad en el número de niños con caries. El Lt esperado según la tabla de la distribución Chi Cuadrado es 11,0705 que es menor 65,85. (Pérez, 2006)

RESUMEN DE LA UNIDAD

En esta unidad, se revisó el concepto de prueba de hipótesis aplicado sobre varianzas, medias, etc.; lo que nos conlleva a hacer conciencia de la relevancia de las pruebas de hipótesis en la toma de decisiones de las empresas.

Como se analizó desde el comienzo de la unidad, los seres humanos actuamos con base en alguna creencia sobre la realidad, basadas en muchos casos en conjeturas o en proposiciones adelantadas, en otras palabras en hipótesis, las cuales se comprueban o se rechazan dando certidumbre o incertidumbre a la realidad.

En el desarrollo de la unidad vimos cómo una prueba de hipótesis es un método sistemático de evaluar creencias tentativas sobre la realidad, que las confronta con evidencia real que nos ayudan a determinar si son razonables o deben desecharse.

GLOSARIO DE LA UNIDAD

Curva de la potencia de la prueba

Es la gráfica de la probabilidad de rechazar H_0 para todos los valores posibles del parámetro poblacional que no satisfacen la hipótesis nula.

Error tipo I

Es el error que se comete al rechazar H0 cuando ésta es verdadera.

Error tipo II

Es el error que se comete al aceptar H0 cuando ésta es falsa.

Estadístico de prueba

Es el estadístico cuyo valor se utiliza para determinar si se rechaza una hipótesis nula.

Nivel de significancia

Es la probabilidad máxima de cometer un error tipo I.

Potencia de la prueba

Es la probabilidad de rechazar correctamente H0 cuando es falsa.

Prueba direccional o de una cola

Prueba de hipótesis en la que la región de rechazo se tiene en un extremo de la distribución muestral.

Prueba no direccional o de dos colas

Prueba de hipótesis en la que la región de rechazo se ubica en ambos extremos de la distribución muestral.

Región de rechazo

Es la zona de valores en la cual se rechaza la hipótesis H0.

Valor crítico

Es un valor contra el cual se compara el obtenido en el estadístico de prueba para determinar si se debe rechazar o no la hipótesis nula.

Valor p

Es la probabilidad de que, cuando la hipótesis nula sea verdadera, se obtenga un resultado de una muestra que sea al menos tan improbable como el que se observa. También se le conoce como nivel observado de significancia.

ACTIVIDADES DE APRENDIZAJE

ACTIVIDAD 1

Revisa los diferentes tipos de pruebas de hipótesis desarrolladas en esta unidad y compáralas, escribe tus conclusiones.

CUESTIONARIO DE REFORZAMIENTO

1. ¿Por qué los investigadores muestran más interés en la varianza poblacional que en la proporción o media poblacionales?
2. ¿A qué se refiere el término grados de libertad?
3. ¿Qué es una prueba de hipótesis?
4. ¿Qué es la distribución Chi-cuadrada (χ^2)?
5. ¿Qué determina la relación entre la varianza de la muestra y la varianza de la población?
6. ¿La prueba estadística de X^2 para una muestra se emplea frecuentemente?
7. ¿Por qué la variabilidad excesiva es el peor enemigo de la alta calidad?
8. ¿Cómo se define una hipótesis estadística?
9. ¿En qué consisten la pruebas de Bondad de Ajuste?
10. ¿Cuáles son los aspectos que deben considerarse para formular la hipótesis nula?

EXAMEN DE AUTOEVALUACIÓN

Elige la respuesta correcta a las siguientes preguntas, una vez que concluyas, obtendrás de manera automática tu calificación.

1. La relación entre la varianza de la muestra y la varianza de la población está determinada por la distribución Chi-cuadrada (χ^2) siempre y cuando la población de la cual se toman los valores de la muestra se encuentre:
    a) normalmente distribuida
    b) indiferente
    c) rechazada
    d) replanteada

2. La desviación estándar de una colección de datos se usa para describir la variabilidad en esa colección y se puede definir como la diferencia estándar entre los elementos de una colección de:
    a) datos y su media
    b) información indiferente
    c) datos aleatorios
    d) datos y su variabilidad

3. Es el error que se comete al aceptar H_0 cuando ésta es falsa:
    a) Tipo I
    b) Tipo II
    c) Tipo III
    d) Estándar

4. La prueba estadística de X^2 para una muestra se emplea frecuentemente como prueba:
    a) bondad de ajuste
    b) Tipo II
    c) Tipo III
    d) Estándar

LO QUE APRENDÍ

Elabora un mapa conceptual sobre los tipos de pruebas desarrolladas en esta unidad.

MESOGRAFÍA

Bibliografía sugerida

Autor	Capítulo	Páginas
Levin y otros (1996)	11	447-501
Lind y otros (2004)	11	369-392
Christensen (1990)	9	459-498
Hanke y otros (1997)	9	275-297

Bibliografía básica

Levin, Richard I. y Rubin, David S. (1996). *Estadística para administradores*. México: Alfaomega.

Lind, A. Douglas; Marchal, G. William; Mason, D. Robert. (2004). *Estadística para Administración y Economía*. (11ª ed.) México: Alfaomega.

Bibliografía complementaria

Black, Ken. (2005). *Estadística en los negocios para la toma de decisiones*. (4ª ed.) México: CECSA.

Christensen, H. (1990). *Estadística paso a paso* (2ª ed.) México: Trillas.

Garza, Tomás. (1996). *Probabilidad y estadística*. México: Iberoamericana.

Hanke, John E. y Reitsch, Arthur G. (1997). *Estadística para Negocios.* México: Prentice Hall.

Sitios de Internet

Sitio	Descripción
http://buzjss.blogspot.com/2008/10/la-distribucin-de-poisson-test-de.html	Buzjss, "Estadística y apuestas deportivas", 16/10/08, [blog]
http://recursostic.educacion.es/descartes/web/materiales_didacticos/distribuciones_probabilidad/aplic_normal.htm	García Cebrian, María José. (2001). "Distribuciones muestrales", Estadística y Probabilidad, Descartes 2D, Matemáticas interactivas.

http://recursostic.educacion.es/descartes/web/materiales_didacticos/Distribucion_binomial/binomial.htm	Martín Álvarez, Pablo Antonio. (2001). "Ajuste de una serie de datos a una distribución binomia", La distribución nominal B (n, p), Descartes 2D, Matemáticas interactivas
http://www.mitecnologico.com/Main/PruebaHipotesisParaProporcion	Mitecnológico, "Prueba de hipótesis para proporción"
http://www.mitecnologico.com/Main/PruebaDeBondadDeAjuste	Mitecnológico, "Prueba de bondad de ajuste"
http://www.mitecnologico.com/Main/PruebaDeIndependencia	Mitecnológico, "Prueba de independencia"
http://www.monografias.com/trabajos15/prueba-de-independencia/prueba-de-independencia.shtml	Pérez Leal, José. (2006). "Prueba de homogeneidad: Prueba de independencia", Monografías
http://html.rincondelvago.com/analisis-de-la-varianza_1.html	"Prueba de la varianza con una población", Rincón del vago

UNIDAD 6

ANÁLISIS DE REGRESIÓN LINEAL SIMPLE

OBJETIVO ESPECÍFICO

Al terminar la unidad el alumno conocerá el método de regresión lineal simple así como su aplicación e interpretación.

INTRODUCCIÓN

El uso de la regresión lineal simple es muy utilizado para observar el tipo de relación que existe entre dos variables y poder llevar a cabo la toma de decisiones correspondiente dependiendo de la relación entre dichas variables, así por ejemplo, pudiera darse el caso en el que después de aplicar la regresión lineal no exista relación entre las variables involucradas y en consecuencia la decisión podría ser buscar cuál es la variable independiente que tiene influencia sobre la dependiente y volver a realizar el estudio completo; pero si fuera el caso en el cual si existiera una relación positiva entre las variables involucradas, la obtención del coeficiente de correlación nos daría más información sobre el porcentaje de relación existente y pudiendo determinar si es necesario la inclusión de otra variable independiente en el problema mismo, para lo cual el análisis de regresión ya sería del tipo múltiple.

LO QUE SÉ

Elige la respuesta correcta a las siguientes preguntas.

1. Es una condición para determinar la ecuación de una recta:
   a) conocer la pendiente de la ordenada al origen
   b) conocer la pendiente y la ordenada al origen de la recta misma
   c) conocer dos ordenadas al origen de la recta misma

2. La pendiente de una recta nos indica:
   a) si la recta pasa por el origen
   b) si la recta se encuentra en un cuadrante en particular
   c) la inclinación de la recta

3. En la ecuación de una recta, la ordenada al origen nos indica:
   a) el punto donde la recta intersecta al eje “x”
   b) un punto fuera del plano
   c) el punto donde la recta intersecta al eje “y”

4. Cuando se dice que la relación entre dos variables es de tipo lineal, sabemos que la gráfica de su relación es:
   a) una línea recta
   b) una parábola
   c) una circunferencia

5. De las siguientes ecuaciones, cuál representa una línea recta:
   a) $x^2 + y^2 = 1$
   b) $y = mx + b$
   c) $y = mx^2 + b$

TEMARIO DETALLADO

(10 horas)

6.1. Ecuación y recta de regresión

6.2. El método de mínimos cuadrados

6.3. Determinación de la ecuación de regresión

6.4. El modelo de regresión y sus supuestos

6.5. Inferencias estadísticas sobre la pendiente de la recta de regresión

6.6. Análisis de correlación

6.1. Ecuación y recta de regresión

Observando el diagrama de dispersión, podemos obtener una primera idea de si existe relación o no entre las variables estadísticas. Con el coeficiente de correlación podemos medir la correlación lineal, en caso de existir. Vamos ahora a calcular las líneas que mejor se aproximen a la nube de puntos. A estas líneas se les llama líneas de regresión.

La función que mejor se aproxima a la nube de puntos puede ser lineal, de segundo grado, exponencial, logarítmica,... En este tema vamos a calcular únicamente funciones lineales, que vamos a llamar rectas de regresión.

La forma de obtener estas rectas es por el procedimiento conocido como el método de los mínimos cuadrados. Buscamos una recta de ecuación $y=mx+n$ que sea la mejor aproximación. Cada punto x_i de la primera variable tendrá, por una parte, el valor correspondiente a la segunda variable y_i, y por otra, su imagen por la recta de regresión $y=mx_i+n$. Entre estos dos valores existirá una diferencia $d_i=mx_i+n-y_i$. Vamos a calcular la recta con la condición de que la suma de los cuadrados de todas estas diferencias $\Sigma(mx_i+n-y_i)^2$ sea mínima. Derivando respecto de m y de n y realizando los cálculos matemáticos necesarios, llegamos a la recta de regresión de Y sobre X, que tiene por ecuación en la forma punto-pendiente.

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2} \cdot (x - \bar{x})$$

(Barrios, 2005)

6.2. El método de mínimos cuadrados

Cualquier método estadístico que busque establecer una ecuación que permita estimar el valor desconocido de una variable, a partir del valor conocido de una o más variables, se denomina análisis de regresión.

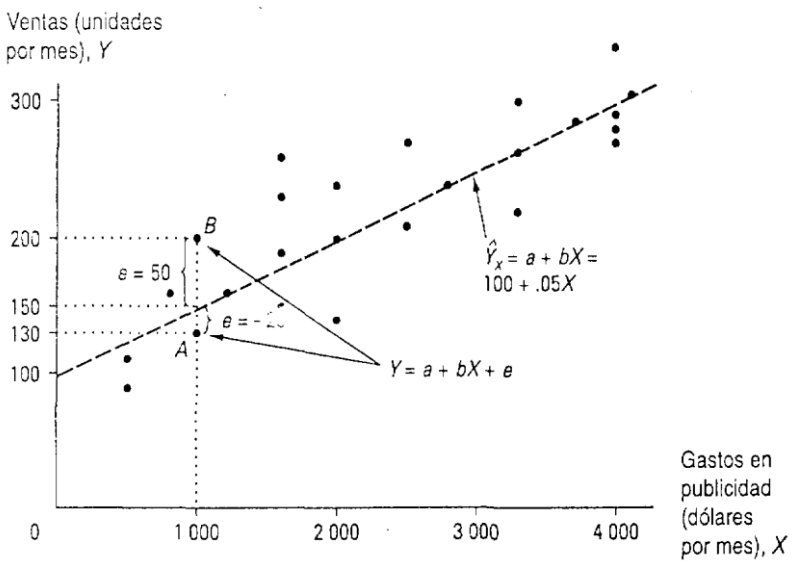
El método de mínimos cuadrados es un procedimiento para encontrar la ecuación de regresión que se origina al estudiar la relación estocástica que existe entre dos variables. Fue Karl Friedrich Gauss (1777-1855) quien propuso el método de los mínimos cuadrados y fue el primero en demostrar que la ecuación estimada de regresión minimiza la suma de cuadrados de errores.

En el análisis de regresión, (Kohler, 1996, pp. 528-529), una variable cuyo valor se suponga conocido y que se utilice para explicar o predecir el valor de otra variable de interés se llama variable independiente y se simboliza por “X”. Por el contrario, una variable cuyo valor se suponga desconocido y que se explique o prediga con ayuda de otra se llama variable dependiente y se simboliza por “Y”.

Una relación estocástica, (Kohler, 1996, p. 530), entre dos variables cualesquiera, *x* y *y*, es imprecisa en el sentido de que muchos valores posibles de “y” se pueden asociar con cualquier valor de “*x”*.

Sin embargo, un resumen gráfico de la relación estocástica entre la variable independiente "*x*" y la variable dependiente "*y*" estará dado por una línea de regresión, misma que reduce al mínimo los errores cometidos cuando la ecuación de esa línea se utilice para estimar y a partir de *x*.

Gráfica que muestra la relación existente entre los gastos de publicidad y las ventas.



De esta gráfica podemos ver claramente que las ventas dadas en unidades por mes (variable dependiente) en este caso, si guardan relación con los gastos en publicidad y, que dicha relación puede ser denotada por la "recta de regresión".

De este análisis de relación estocástica, que se da entre dos variables, surgen las ecuaciones que nos provee el método de mínimos cuadrados, que a saber son:

Ecuación de la recta de regresión: $\hat{y}_i = b_0 + b_1 x_i$

En la que:

- x_i = es un valor dado de la variable independiente para el cual se quiere estimar el valor correspondiente de la variable dependiente
- b_0 = ordenada al origen de la línea estimada de regresión,
- b_1 = pendiente de la línea estimada de regresión,
- $\hat{Y}_i$ = valor estimado de la variable dependiente, para el i-ésimo valor de la variable independiente

Resulta claro que para poder determinar la recta de regresión, es necesario que antes sean calculados los valores correspondientes a la pendiente de la recta y a la ordenada al origen.

La pendiente de la recta de regresión se calcula mediante la siguiente fórmula:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

y la ordenada al origen se calcula mediante la fórmula:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Antes de continuar, es necesario advertir que el análisis de regresión no se puede interpretar como un procedimiento para establecer una relación de causa a efecto entre variables. Sólo puede indicar cómo o hasta qué grado las variables están asociadas entre sí. Cualquier conclusión acerca de causa y efecto se debe basar en el juicio del o los individuos con más conocimientos sobre la aplicación. Por ejemplo, un estadista puede llegar a determinar que la relación entre las ventas y el presupuesto asignado a mercadotecnia es positiva y que se tiene un coeficiente de correlación de 0.96, lo cual prácticamente nos indica que es recomendable incrementar el presupuesto al departamento de mercadotecnia para obtener mejores ingresos dentro de la compañía, sin embargo el director de operaciones puede llegar a determinar que debido a condiciones internas del país en el que se encuentre la empresa, o bien la aparición de una nueva ley que regule los medios utilizados por el mencionado departamento de mercadotecnia, pueden llegar a frenar o incluso generar conflictos dentro de la empresa si incrementamos el presupuesto al departamento correspondiente.

6.3. Determinación de la ecuación de regresión

En estadística la **regresión lineal** o **ajuste lineal** es un método matemático que modeliza la relación entre una variable dependiente *Y*, las variables independientes X_i y un término aleatorio ε. Este modelo puede expresarse como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Donde β_0 es la intersección o término "constante", las $\beta_i \ (i > 0)$ son los parámetros respectivos a cada variable independiente, y *p* es el número de parámetros independientes para tener en cuenta en la regresión. La regresión lineal puede ser contrastada con la regresión no lineal.

6.4. El modelo de regresión y sus supuestos

Con frecuencia, nos encontramos en economía con modelos en los que el comportamiento de una variable, *Y,* se puede explicar a través de una variable *X;* lo que representamos mediante

$$Y = f(X) \quad (1)$$

Si consideramos que la relación *f,* que liga *Y* con *X*, es lineal, entonces (1) se puede escribir así:

$$t\ 1\ 2\ t\ Y = \beta + \beta\ X \quad (2)$$

Como quiera que las relaciones del tipo anterior raramente son exactas, sino que más bien son aproximaciones en las que se han omitido muchas variables de importancia secundaria, debemos incluir un término de perturbación aleatoria, *t u* , que refleja todos los factores – distintos de *X* -que influyen sobre la variable endógena, pero que ninguno de ellos es relevante individualmente. Con ello, la relación quedaría de la siguiente forma:

Modelo de regresión simple

$$Yt = \beta + \beta\ Xt + ut \quad (3)$$

(Uriel, 2004, p. 1)

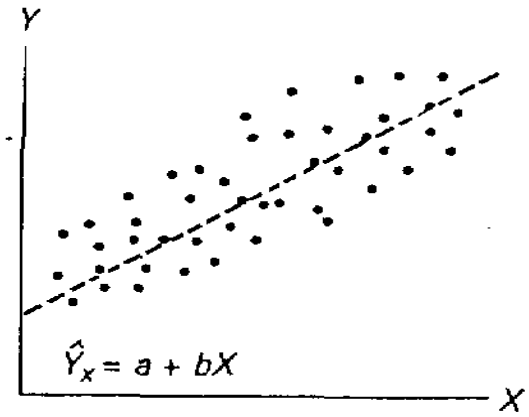
6.5. Inferencias estadísticas sobre la pendiente de la recta de regresión

Las inferencias acerca de la pendiente de la recta de regresión son importantes dado que la relación entre las dos variables en cuestión depende de ella precisamente, es decir, si la pendiente de la recta de regresión es positiva, entonces la naturaleza de la relación entre ambas variables será positiva, y la pendiente de la recta es negativa, entonces la relación entre las variables será negativa también, con lo cual podemos iniciar la toma de decisiones dependiendo del contexto del problema mismo. Como se mencionó anteriormente la ecuación de la recta de regresión:

$$\hat{y}_i = b_0 + b_1 X_i$$

b_0 representa la ordenada al origen de la línea estimada de regresión, y b_1 es la pendiente de la línea estimada de regresión.

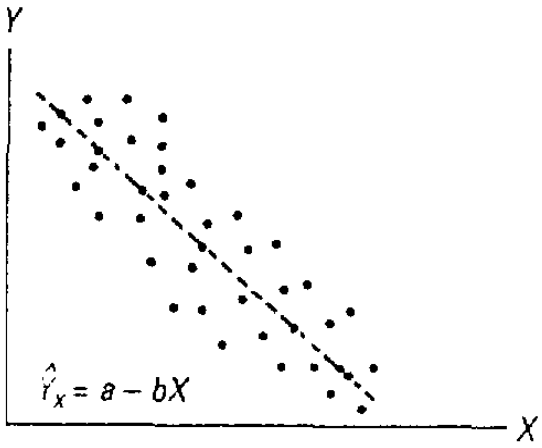
Donde b_0 es en sí, el punto donde la recta corta al eje de las “x” y b_1 nos da el grado de inclinación de la recta, de tal forma que cuando la pendiente de la recta es positiva, se dice que la relación que existe entre las dos variables dependiente e independiente es de naturaleza positiva, es decir, que posee una gráfica como la indicada a continuación:



Relación positiva entre dos variables en regresión lineal

En este tipo de relación, los incrementos en los valores de la variable independiente traen como consecuencia un incremento en los valores correspondientes de la variable dependiente y la gráfica tiene como podemos apreciar una forma ascendente.

Pero cuando la pendiente de la recta de regresión es negativa, es decir, que dicha ecuación tuviera la forma $\hat{y}_i = b_0 - b_1 X_i$ entonces la relación existente entre las variables es de tipo negativa, lo cual quiere decir, que a incrementos en los valores de la variable independiente, la variable dependiente responde con decrementos; la gráfica resultante tendría la forma siguiente:



Relación negativa entre dos variables en regresión lineal

En esta gráfica podemos observar que la tendencia de la recta de regresión es descendente, lo cual implica como ya habíamos mencionado, que la relación entre ambas variables es negativa.

6.6. Análisis de correlación

Cuando es necesario resumir aún más los datos (de una gráfica por ejemplo) se utiliza un solo número, que de alguna forma mide la fuerza de asociación entre dos variables como son el ingreso real y el nivel de educación escolar en nuestro caso. El análisis de correlación nos ayuda a obtener dicho número que se conoce como: coeficiente de correlación. Los valores de coeficiente de correlación siempre están entre –1 y +1 un valor de +1 indica que las dos variables tienen una relación lineal positiva perfecta.

Esto es, todos los puntos de datos están en una línea recta con pendiente positiva. Un valor de –1 indica que las variables tienen una relación lineal negativa perfecta, y que todos los puntos de datos están en una recta con pendiente negativa. Los valores del coeficiente de correlación cercanos a cero indican que las variables no tienen relación línea, (véase, Anderson, Sweeney & Willimas, 1999, p. 555).

A continuación presentamos la ecuación para calcular el coeficiente de correlación de la muestra. Si ya se ha hecho un análisis de regresión y se ha calculado el coeficiente de determinación, entonces, el coeficiente de correlación se puede calcular como sigue:

$$r = (signo de b_1)\sqrt{r^2}$$

Donde b1 es la pendiente de la ecuación de regresión.

De esta fórmula, resulta claro que el signo del coeficiente de correlación es positivo si la ecuación de regresión tiene pendiente positiva (b1 > 0), y negativo si la ecuación de regresión tiene pendiente negativa (b1 < 0).

RESUMEN DE LA UNIDAD

En esta unidad se revisó el método de regresión lineal simple así como su aplicación e interpretación, la importancia de este método radica en que se utiliza para observar el tipo de relación que existe entre dos variables y poder llevar a cabo la toma de decisiones correspondiente dependiendo de la relación entre dichas variables. Si fuera el caso en el cual existiera una relación positiva entre las variables involucradas, la obtención del coeficiente de correlación nos daría más información sobre el porcentaje de relación existente y con esto determinar si es necesario incluir otra variable independiente en el problema mismo.

GLOSARIO DE LA UNIDAD

Análisis de residuales

Análisis que se aplica para determinar si los supuestos acerca del modelo de regresión parecen válidos. También se usa para determinar observaciones extraordinarias o influyentes.

Coeficiente de correlación

Medida de la intensidad de la relación lineal entre dos variables.

Coeficiente de determinación

Medida de la bondad del ajuste de la recta de regresión. Se interpreta como la parte de la variación de la variable dependiente “y” que explica la recta de regresión.

Diagrama de dispersión

Gráfica de datos de dos variables en la que la variable independiente está en el eje horizontal y la variable dependiente en el eje vertical.

Método de mínimos cuadrados

Procedimiento que se usa para determinar la recta de regresión. Su objeto es minimizar $\sum\left(y_i - \hat{y}_i\right)^2$

Observación influyente

Observación que tiene una fuerte influencia sobre el efecto de los resultados de la regresión.

Puntos de gran influencia

Observaciones con valores extremos de la variable independiente.

Recta de regresión

Estimación hecha a partir de datos de una muestra aplicando el método de mínimos cuadrados para la regresión lineal simple, la ecuación de regresión estimada es: $\hat{y}_i = b_0 + b_1 X_i$

Regresión lineal simple

Análisis de regresión donde intervienen una variable independiente y una variable dependiente; en ella, la relación entre las variables se aproxima mediante una recta.

Residual i-ésimo

Diferencia entre el valor observado de la variable dependiente y el valor predicho usando la recta de regresión; para la i-ésima observación, el residual es: $y_i - \hat{y}_i$

Variable dependiente

Es la variable que se predice o se explica. Se representa matemáticamente por “y”.

Variable independiente

Es la variable que sirve para predecir o explicar. Se representa matemáticamente por “x”.

ACTIVIDADES DE APRENDIZAJE

ACTIVIDAD 1

Explica las implicaciones del signo y valor del coeficiente de determinación del problema resuelto en la autoevaluación.

ACTIVIDAD 2

Explica las implicaciones del signo y valor del coeficiente de correlación del problema resuelto en la autoevaluación.

CUESTIONARIO DE REFORZAMIENTO

1. ¿Qué es el análisis de regresión lineal o bivariada?
2. ¿Cuándo se aplica la regresión múltiple?
3. ¿Qué es el método de los mínimos cuadrados?
4. ¿Quién propuso el método de los mínimos cuadrados?
5. ¿Qué es el coeficiente de determinación?
6. ¿Cuál es el rango del coeficiente de determinación?
7. ¿Qué es el coeficiente de correlación?
8. ¿Cuál es el rango del coeficiente de correlación?
9. ¿Quién desarrolló por primera vez los métodos estadísticos para el estudio de la relación entre dos variables?
10. ¿Es el análisis de regresión un procedimiento para establecer una relación de causa y efecto?

EXAMEN DE AUTOEVALUACIÓN

Elige la respuesta correcta a las siguientes preguntas, una vez que concluyas, obtendrás de manera automática tu calificación.

1. ¿Por qué son importantes las inferencias acerca de la pendiente de la recta de regresión?
    a) porque de ella depende la relación entre las variables en cuestión
    b) porque matemáticamente es obligatorio calcularla.
    c) porque nos indica el punto donde la recta de regresión intersecta al eje de las “y”.

2. ¿Cuándo la pendiente de la recta de regresión es positiva, la relación entre las variables es?
    a) negativa
    b) cero
    c) positiva

3. Cuándo la pendiente de la recta de regresión es negativa, la relación entre las variables, ¿cómo es?
    a) cero
    b) negativa
    c) positiva

4. ¿Es el símbolo comúnmente utilizado para denotar a la pendiente de la recta de regresión?:

a) b_0

b) b_1

c) b_2

Un economista del Departamento del Distrito Federal está preparando un estudio sobre el comportamiento del consumidor. Los datos que obtuvo los plasmó en la siguiente tabla.

Consumidor	1	2	3	4	5	6	7	8	9	10	11	12
Ingreso	24.3	12.5	31.2	28	35.1	10.5	23.2	10	8.5	15.9	14.7	15
Consumo	16.2	8.5	15	17	24.2	11.2	15	7.1	3.5	11.5	10.7	9.2

5. Considerando el consumo como variable dependiente el coeficiente de determinación es:

a) $r^2 =$ 0.844740208

b) $r^2 =$ -0.844740208

c) $r^2 =$ 1.844740208

6. Para el problema anterior, el coeficiente de correlación es:

a) r =1.919097496

b) r =-0.919097496

a) r = 0.919097496

LO QUE APRENDÍ

Una tienda departamental está considerando otorgar tarjetas de crédito a sus clientes, para lo cual realiza un estudio con el fin de observar el comportamiento de sus gastos en función de su salario. Los datos obtenidos en una muestra aleatoria de tamaño 11 se encuentran en la siguiente tabla.

Sueldo del cliente	18.0	15.0	19.0	9.2	8.6	12.0	10.7	14.3	17.8	16.0	15.0
Gastos del cliente	14.8	10.4	15.7	7.1	5.3	8.0	8.5	10.2	13.0	14.0	11.3

Nota: tanto el sueldo como los gastos del cliente son mensuales y están dados en miles de pesos.

Haz un análisis de regresión, define las variables involucradas y determina:

a) la pendiente de la recta de regresión
b) la ordenada al origen de la recta de regresión
c) la recta de regresión lineal resultante.
d) el coeficiente de determinación
e) el coeficiente de correlación
f) el pronóstico de gasto para un cliente que gana $21,000.00

En conclusión, para este problema, entre más ganan los empleados, más gastan.

MESOGRAFÍA

Bibliografía sugerida

Autor	Capítulo	Páginas
Levin y otros (1996)	12 y13	509-612
Lind y otros (2004)	13	458 -489
Christensen (1990)	10	557 - 609
Hanke (1997)	14	522 - 561

Bibliografía básica

Levin, Richard I. y Rubin, David S. (1996). *Estadística para administradores*. México: Alfaomega.

Lind, A. Douglas; Marchal, G. William; Mason, D. Robert. (2004). *Estadística para Administración y Economía*. (11ª ed.) México: Alfaomega.

Bibliografía complementaria

Anderson, David R.; Sweeney, Dennis J.; Williams. Thomas A. (1999). *Estadística para administración y economía*. México: Thomson.

Christensen, H. (1990). *Estadística paso a paso* (2ª ed.) México: Trillas.

Garza, Tomás. (1996). *Probabilidad y estadística*. México: Iberoamericana.

Hanke, John E. y Reitsch, Arthur G. (1997). *Estadística para Negocios.* México: Prentice Hall.

Sitios de Internet

Sitio	Descripción
http://recursostic.educacion.es/descartes/web/materiales_didacticos/bidimensional_lbarrios/regresion_est.htm	Barrios Calmaestra, Luis. (2005). "Regresión lineal", Estadísticas II, Distribuciones bidimensionales. Descartes 2D Matemáticas interactivas.
http://www.uv.es/uriel/material/Morelisi.pdf	Uriel Jiménez, Ezequiel. (2004). *Modelos de regresión lineal simple*, UV.

UNIDAD 7

ANÁLISIS DE SERIES DE TIEMPO

OBJETIVO ESPECÍFICO

Al terminar la unidad el alumno conocerá los métodos para el análisis de series de tiempo, así como su aplicación e interpretación.

INTRODUCCIÓN

Una serie de tiempo es el conjunto de datos que se registran a través del tiempo sobre el comportamiento de una variable de interés, generalmente los registros se realizan en periodos iguales de tiempo.

Las series de tiempo resultan especialmente útiles cuando se requiere realizar un pronóstico sobre el comportamiento futuro que puede tener una variable determinada, imaginemos por ejemplo la necesidad de tomar una decisión sobre el comportamiento a futuro de la demanda, el precio y las ventas de un producto, los ingresos en el próximo año, los precios de bienes y servicios, los valores de los energéticos, etc. En todas estas situaciones resulta útil el análisis de las series de tiempo que los representan, bajo la hipótesis de que los factores que han influenciado su comportamiento en el pasado estarán presentes de manera similar en el futuro.

De esta manera, el objetivo principal del conocimiento de las series de tiempo es la identificación de los factores que intervienen y la separación de cada uno de ellos, con el fin de pronosticar cuál será el comportamiento en el futuro.

LO QUE SÉ

Elige la respuesta correcta a las siguientes preguntas, una vez que concluyas, obtendrás de manera automática tu calificación.

1. La fórmula que caracteriza la recta de regresión es:

a) $\hat{y}_i = b_0 + b_1 X^2{}_i$

b) $\hat{y}_i = b_0 + b_1 X_i$

c) $\bar{X} = \frac{\sum_{i=1}^{i=n} x_i}{n}$

2. La fórmula para determinar la pendiente de la recta de regresión es:

a) $b_0 = \bar{Y} - b_1 \bar{X}$

b) $b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$

c) $\hat{y}_i = b_0 + b_1 X_i$

3. La fórmula para determinar la ordenada al origen de la recta de regresión es:

a) $b_o = \bar{Y} - b_1 \bar{X}$

b) $$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

c) $\hat{y}_i = b_o + b_1 X_i$

4. La fórmula para calcular el coeficiente de determinación es:

a) $$r = \sqrt{\frac{\sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2}{\sum_{i=1}^{n} (Y - \bar{Y}_i)^2}}$$

b) $$r^2 = signo \quad de \quad b_1 \left(\sqrt{\frac{\sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2}{\sum_{i=1}^{n} (Y - \bar{Y}_i)^2}} \right)$$

c) $$r^2 = \frac{\sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2}{\sum_{i=1}^{n} (Y - \bar{Y}_i)^2}$$

5. La fórmula para calcular el coeficiente de correlación es:

a) $r = (signo \quad de \quad b_1)\sqrt{r^2}$

b) $r = (signo \quad de \quad b_o)\sqrt{r^2}$

c) $r^2 = signo \quad de \quad b_0 \left(\sqrt{\frac{\sum_{i=1}^{n} (\hat{Y} - \bar{Y})^2}{\sum_{i=1}^{n} (Y - \bar{Y}_i)^2}} \right)$

6. ¿Cuál es el rango de los valores que puede tomar el coeficiente de determinación?

a) $-\infty, +\infty$

b) $-1, +1$

c) $0, +1$

7. ¿Cuál es el rango de los valores que puede tomar el coeficiente de correlación?

a) $-\infty, +\infty$

b) $-1, +1$

c) $0, +1$

TEMARIO DETALLADO

(8 horas)

7.1. Los cuatro componentes de una serie de tiempo

7.2. Análisis gráfico de la tendencia

7.3. Tendencia secular

7.4. Variaciones estacionales

7.5. Variaciones cíclicas

7.6. Fluctuaciones irregulares

7.7. Modelos autoregresivos de promedios móviles

7.1. Los cuatro componentes de una serie de tiempo

La componente cíclica es la fluctuación que puede observarse que ocurre alrededor de la tendencia. Cualquier patrón regular de variaciones arriba o debajo de la recta que representa a la tendencia puede atribuirse a la componente cíclica.

Estacionalidad (E)

La componente estacional muestra un comportamiento regular en los mismos periodos de tiempo, reflejando costumbres o modas que se repiten regularmente dentro del periodo de observación. En la gráfica la estacionalidad quedaría representada por ejemplo por las variaciones semanales en los rendimientos, no visibles por el periodo de información que se está manejando.

Componente irregular (I)

Es la componente que queda después de separar a las otras componentes, es el resultado de factores no explicables que siguen un comportamiento aleatorio, siendo por ello una parte no previsible de la serie.

Ejemplo

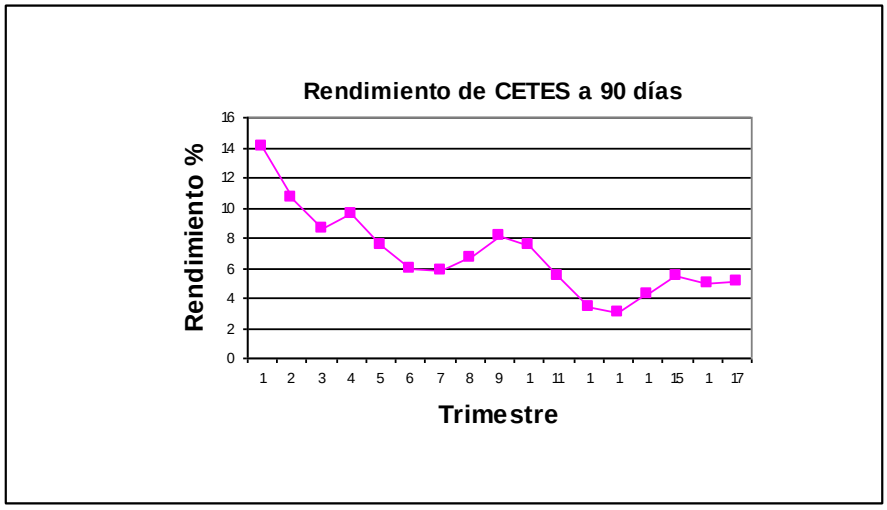
Supongamos que tenemos la información siguiente, correspondiente al comportamiento del rendimiento de los Certificados de la Tesorería, denominados CETES a 90 días, el tiempo está expresado en trimestres y el valor de la variable en valores de la tasa de interés que ganan en cada trimestre.

Trimestre	%
1	14.03
2	10.69
3	8.63
4	9.58
5	7.48
6	5.98
7	5.82
8	6.69
9	8.12
10	7.51
11	5.42
12	3.45
13	3.02
14	4.29
15	5.51
16	5.02
17	5.07

Rendimiento de CETES a 90 días

El registro de rendimientos trimestrales de los CETES representa una serie de tiempo, ya que se han obtenido en periodos sucesivos.

Si se analiza el registro podemos observar que hay una disminución en los valores de rendimiento, de mayor a menor, pero nos resulta difícil afirmar en qué proporción ha ocurrido y de cuánto han sido las variaciones. Si este registro lo analizamos como una serie tendremos la gráfica siguiente:



Rendimiento de los certificados de la tesorería a 90 días

Utilizando el ejemplo anterior procederemos a descomponer la serie de tiempo en cada uno de sus componentes, lo cual haremos en los siguientes incisos.

La separación de la tendencia utiliza la metodología de la línea de regresión, hemos mencionado que esta línea puede ser una recta o una curva, en este curso únicamente analizaremos el modelo lineal, por su simpleza y facilidad de cálculo, de esta manera podemos representar a la tendencia por medio de la expresión matemática siguiente:

$$Yt = b_o + b_1X$$

En donde:

- Yt tasa de rendimiento calculada
- X tiempo, en este caso expresado en trimestres
- b_o valor de Y cuando el valor del tiempo es cero
- b_1 pendiente de la recta de tendencia

Una vez definido el modelo, se procede a la determinación de los valores de los coeficientes b_o y b_1 de la recta de regresión. En nuestro problema en particular, la ecuación de regresión, que representa a la tendencia del comportamiento de la tasa de rendimiento de los CETES a 90 días aplicando las fórmulas correspondientes para el cálculo primero de "b_1"

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}$$

y posteriormente para el cálculo de "b_0"

$$b_o = \bar{Y} - b_1 \bar{X}$$

es:

$$Yt = 10.8553676 - 0.44595588\ X$$

Además, aplicando las fórmulas correspondientes primero al cálculo del coeficiente de determinación:

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

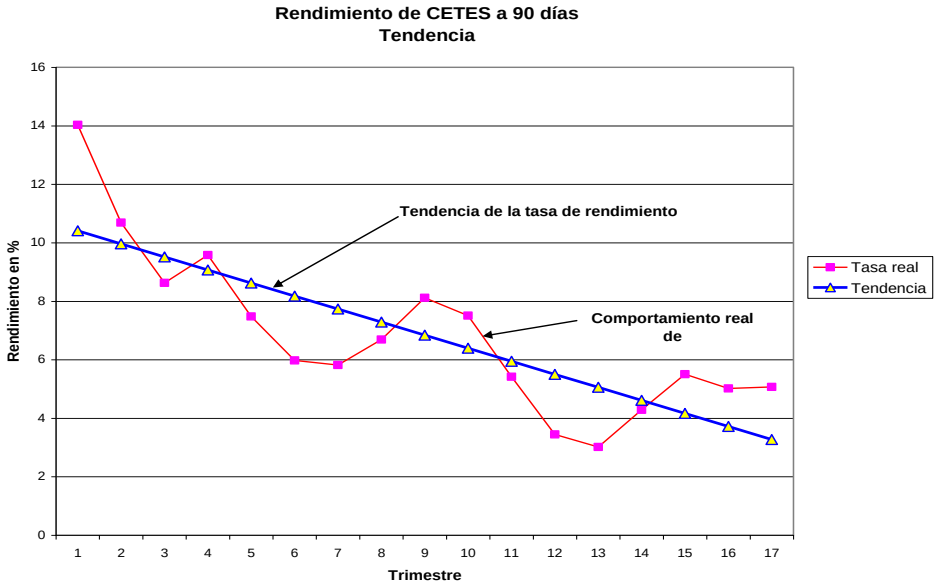
y finalmente al cálculo del coeficiente de correlación:

$$r = (signo de b_1)\sqrt{r^2}$$

Tenemos que el valor del coeficiente de correlación es de r = -0.8078, lo que nos indica que el ajuste logrado con la recta de regresión es adecuado, recordemos que el coeficiente de correlación es una medida de la precisión lograda en el ajuste, valores del coeficiente de correlación iguales a +1 ó -1 son la indicación de un ajuste perfecto, un valor igual a cero nos dirá que este no existe. (nota: se deja al estudiante corroborar los valores obtenidos de "b_1", "b_0" y "r")

7.2. Análisis gráfico de la tendencia

Una vez definida la ecuación de la recta de tendencia es posible compararla gráficamente con los valores de la serie, como se muestra en la gráfica siguiente (Gráfica de comparación de la recta de tendencia contra el comportamiento real de los CETES a 90 días.), en ella podemos observar que la tendencia de las tasas de rendimiento es descendente, el signo del coeficiente b1, que representa la pendiente de la recta, ya nos lo había indicado. También podemos observar que son evidentes valores por arriba y por debajo de esta línea, estos representan a los valores cíclicos de la serie.



Gráfica de comparación de la recta de tendencia contra el comportamiento real de los CETES a 90 días

En el análisis de tendencias podemos ver clara y rápidamente mediante el cálculo de la pendiente de la recta de regresión (b_1) si la tendencia de la variable de medición (en nuestro caso en particular “el rendimiento de los CETES a 90 días”) es a la baja (pendiente negativa), a la alza (pendiente positiva) o a mantenerse sin variación (pendiente cero); lo cual dentro del análisis de la serie de tiempo, es muy importante.

7.3. Tendencia secular

Se denomina tendencia secular o simplemente tendencia a la trayectoria temporal de crecimiento, decrecimiento o estabilidad que sigue una serie cronológica a largo plazo. Movimiento unidireccional y persistente que describe la evolución temporal de una determinada variable, una vez depurada de sus variaciones estacionales, cíclicas y accidentales. Para obtener la tendencia secular de una serie temporal se pueden emplear diferentes métodos, como por ejemplo el de las medias móviles o el de los mínimos cuadrados.

7.4. Variaciones estacionales

Método de la razón a la media móvil para determinar la componente estacional en una serie temporal

1º Se determina la tendencia por el método de las medias centradas en los períodos (Yt) (estamos aplicando cuatro observaciones para el cálculo de la media aritmética)

2º Como este método se basa en la hipótesis multiplicativa, si dividimos la serie observada Yt, por su correspondiente media móvil centrada, eliminamos de forma conjunta las componentes del largo plazo (tendencia y ciclo), pero la serie seguirá manteniendo el efecto de la componente estacional.

3º Para eliminar el efecto de la componente estacional, calcularemos las medias aritméticas a nivel de cada estación (cuatrimestre). Estas medias representan de forma aislada la importancia de la componente estacional.

4º. Calcularemos los índices de variación estacional, para lo que previamente calcularemos la media aritmética anual de las medias estacionales (M1, M2, M3, M4), que será la base de los índices de variación estacional. Existirán tantos índices como estaciones o medias estacionales tengan las observaciones

5º Una vez obtenidos los índices de variación estacional puede desestacionalizarse la serie observada, dividiendo cada valor de la correspondiente estación por su correspondiente índice.

Método de la Tendencia por Ajuste Mínimo-Cuadrático El objetivo sigue siendo aislar la componente estacional de la serie por eliminación sucesiva de todos los demás. La diferencia con el método anterior es que, en este caso, las componentes a l/p (tendencia-ciclo) las obtenemos mediante un ajuste mínimo-cuadrático de las medias aritméticas anuales yt calculándose bajo la hipótesis aditiva.

Sigue los siguientes pasos:
• Se calculan las medias anuales de los datos observados y :

i las observaciones son trimestrales estas medias se obtienen con 4 datos, si son mensuales con 12 datos, etc. para el caso de que el periodo de repetición sea el año

• Se ajusta una recta por mínimos cuadrados y a b t t = + que nos representa, como sabemos, la tendencia, siendo el coeficiente angular de la recta el incremento medio anual de la tendencia, que influirá de forma distinta al pasar de una estación a otra
• Se calculan, con los datos observados, las medias estacionales (M1, M2, M3, ...) con objeto de eliminar la componente accidental. Estas medias son brutas pues siguen incluyendo los componentes a l/p (tendencia-ciclo) que deben someterse a una corrección.

• Empleando el incremento medio anual dado por el coeficiente, se obtienen las medias estacionales corregidas de las componentes a largo plazo (M'1, M'2, M'3, ...) bajo el esquema aditivo:

• Los índices de variación estacional se obtienen con la misma sistemática del método anterior: con las medias estacionales corregidas se obtiene la media aritmética anual M'A que sirve de base para calcular los índices:

• Obtenidos estos índices, podemos desestacionalizar la serie como en el método anterior. (Ruíz, 2004, §5.4)

7.5. Variaciones cíclicas

Las fluctuaciones de los valores de rendimientos alrededor de la línea de tendencia constituyen la componente cíclica, estas son el resultado de la ocurrencia de fenómenos que pueden tener origen social, económico, político, costumbres locales, etc., pero que pueden afectar el comportamiento de la variable, de ahí que su separación resulte importante.

Supongamos ahora que nos interesa conocer la variación que han tenido los rendimientos respecto de la tendencia, es decir la componente cíclica, la cual queda representada en la gráfica (Gráfica de apreciación de la componente cíclica de los CETES a 90 días) por los valores mayores y menores respecto de la tendencia. Si deseamos conocer el valor numérico de este comportamiento debemos proceder como sigue:

Calcular para cada trimestre el valor del rendimiento de acuerdo con la ecuación de la tendencia (Yt) y compararlo con el correspondiente del registro, estableciendo una proporción entre estos dos valores de la manera siguiente:

$$c = \frac{Y}{Yt} 100$$

En donde:

Y representa el rendimiento registrado.

Yt representa el rendimiento calculado con la ecuación de tendencia.

Los valores así calculados se muestran en la tabla siguiente, expresados en porcentaje respecto del valor de la tendencia, los valores que estén por encima de la recta de tendencia alcanzarán un porcentaje superior a cien, mientras que los que se encuentren por debajo de ella tendrán valores inferiores a cien.

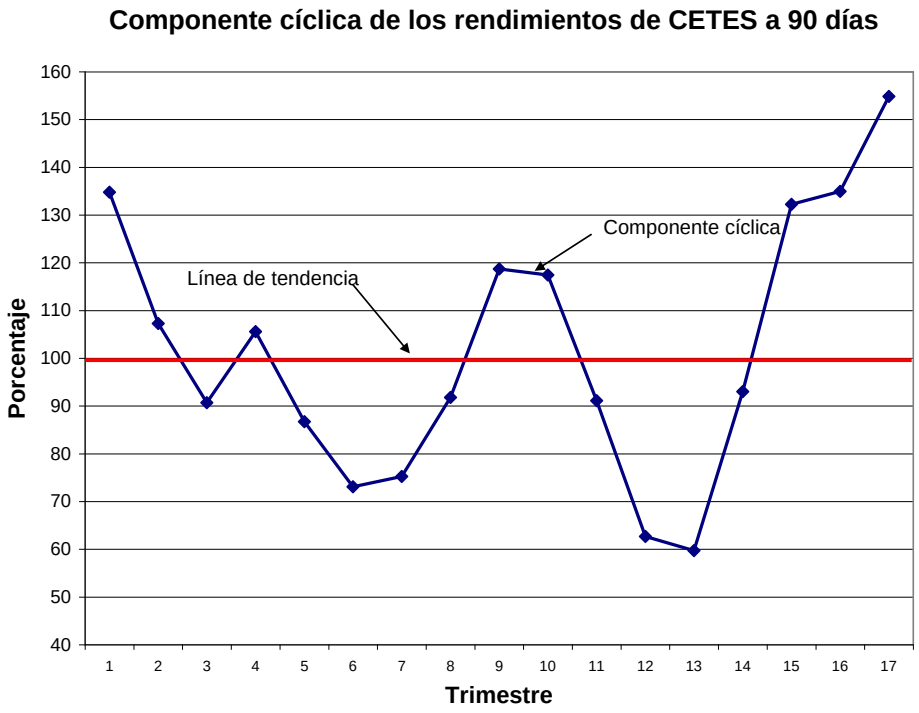
Valores de la componente cíclica

	Rendimiento		Componente
Trimestre	Real	Tendencia	cíclica
	Y	Yc	%
1	14.03	10.41	134.78
2	10.69	9.96	107.29
3	8.63	9.52	90.68
4	9.58	9.07	105.60
5	7.48	8.63	86.72
6	5.98	8.18	73.11
7	5.82	7.73	75.26
8	6.69	7.29	91.80
9	8.12	6.84	118.68
10	7.51	6.40	117.42
11	5.42	5.95	91.09
12	3.45	5.50	62.68
13	3.02	5.06	59.71
14	4.29	4.61	93.02
15	5.51	4.17	132.26

16	5.02	3.72	134.94
17	5.07	3.27	154.85

Las componentes cíclicas pueden ser graficadas para observar los posibles patrones que se presentan, la línea de la tendencia corresponde en la gráfica a la línea del 100%, observemos que la variación cíclica se presenta hacia arriba y hacia abajo de la recta de tendencia.

Gráfica de apreciación de la componente cíclica de los CETES a 90 días.



Es posible ver con mucha claridad cuál ha sido el comportamiento de los rendimientos respecto de la tendencia. Podemos observar que las fluctuaciones a la baja han sido más importantes que las correspondientes a la alza.

Esto es muy importante, pues si alguna persona compró CETES a 90 días durante el primer trimestre, podemos observar que el rendimiento de estos bajó a continuación y apenas pudieron igualarse los rendimientos alrededor del trimestre 16, presentando una alza alrededor del trimestre 17, lo cual puede representar una pérdida de tiempo y dinero para la persona que bien pudo invertir algunos otros instrumentos que tuvieran mejores rendimientos.

7.6. Fluctuaciones irregulares

Finalmente, una vez separada la componente estacional, procedemos a calcular la componente irregular, lo cual se realiza utilizando nuevamente la ecuación del modelo multiplicativo, relacionándola con el producto de las componentes conocidas hasta ahora, es decir obteniendo la relación:

$$\frac{(T)(C)(E)(I)}{(T)(C)(E)} = I$$

Los valores obtenidos se expresan en porcentaje, el cálculo de esta componente se muestra en la tabla siguiente:

	Rendimiento	Componentes			
Trimestre	Real	tendencia	cíclica	temporal	Irregular
		Yc	C	E	I
1	14.03	10.41	134.78	96.52	103.61

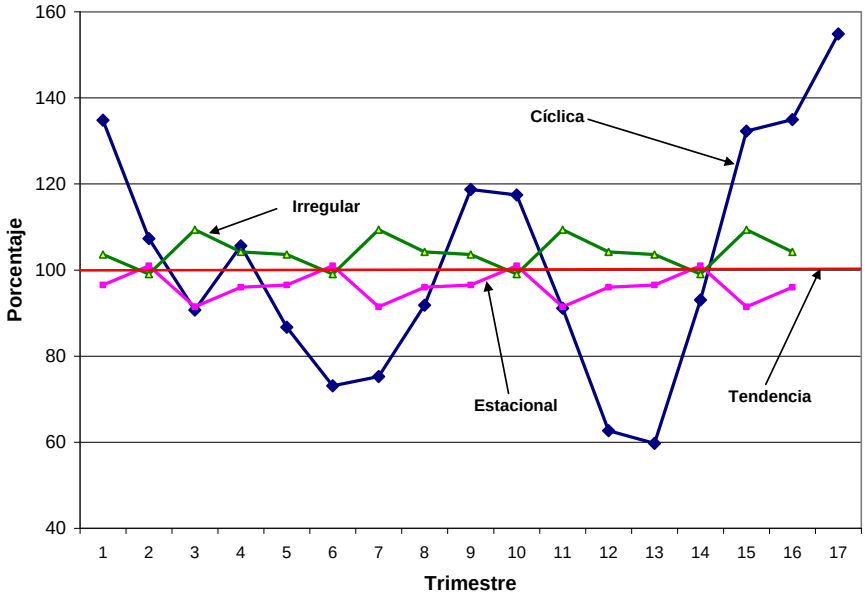
2	10.69	9.96	107.29	100.96	99.05
3	8.63	9.52	90.68	91.46	109.34
4	9.58	9.07	105.60	95.98	104.19
5	7.48	8.63	86.72	96.52	103.61
6	5.98	8.18	73.11	100.96	99.05
7	5.82	7.73	75.26	91.46	109.34
8	6.69	7.29	91.80	95.98	104.19
9	8.12	6.84	118.68	96.52	103.61
10	7.51	6.40	117.42	100.96	99.05
11	5.42	5.95	91.09	91.46	109.34
12	3.45	5.50	62.68	95.98	104.19
13	3.02	5.06	59.71	96.52	103.61
14	4.29	4.61	93.02	100.96	99.05
15	5.51	4.17	132.26	91.46	109.34
16	5.02	3.72	134.94	95.98	104.19
17	5.07	3.27	154.85		

Cálculo de la componente irregular

En la tabla se presentan los valores de cada una de las componentes, los correspondientes a la cíclica, estacional e irregular se expresan como un porcentaje del valor de la tendencia, la gráfica (Gráfica de los componentes de la serie de tiempo para nuestro ejemplo del rendimiento de los CETES a 90 días) que relaciona todos los valores se presenta enseguida.

Gráfica de los componentes de la serie de tiempo para nuestro ejemplo del rendimiento de los CETES a 90 días.

Rendimientos de CETES a 90 días
Componentes de la serie de tiempo



Una vez separadas cada una de los componentes es posible conocer la influencia que cada una de ellas tiene sobre el valor del rendimiento, y tomar una decisión sobre las consideraciones que deban realizarse para llevar a cabo una predicción, en este caso deberá analizarse con mucha atención la relación que cada una de ellas haya tenido con los fenómenos económicos y hacer la consideración de las probabilidades que tiene de ocurrir de la misma manera, para considerar o no su participación en la predicción sobre el comportamiento del rendimiento de los CETES.

7.7. Modelos autoregresivos de promedios móviles

Un proceso estocástico { zt } con índice temporal discreto se dice estacionario si las distribuciones conjuntas de probabilidad asociadas con un vector *(zt^,z^2,...,z,k)* son idénticas a las asociadas con el vector *(zl1+h,z ►Z+h,...,z^,^+h)* obtenido por una traslación temporal, y esto para todo conjunto *(tl,t,,...,t^)* de índices, para todo *k* y para todo *h.* Un proceso estacionario tiene todos sus momentos invariantes a cambios en el tiempo. Un proceso se dice "estacionario débil" si sus momentos de primer y segundo orden (esperanzas matemáticas, varianzas, covarianzas) son invariantes a cambios en el tiempo.

RESUMEN DE LA UNIDAD

Esta unidad es una introducción básica a los métodos elementales de análisis y pronóstico de series de tiempo; primero se mostró, que para explicar el comportamiento de una serie de tiempo es conveniente suponer que la serie está formada por sus cuatro componentes básicos: tendencia, cíclico, estacional e irregular. Posteriormente separamos cada uno de estos componentes para medir su efecto, con lo cual logramos pronosticar valores futuros de la serie de tiempo.

También se mencionaron los métodos de suavizamiento como medio para pronosticar una serie de tiempo que no presenta algunos de sus componentes de manera apreciable. Además se ejemplificó el uso del análisis de regresión lineal en series de tiempo que solo tengan una tendencia a largo plazo.

Finalmente es fácil observar que las series de tiempo son métodos cualitativos de pronóstico que se utilizan cuando se tienen pocos datos históricos o carecemos de ellos. Las series de tiempo también se utilizan cuando se espera que su comportamiento continúe en el futuro.

GLOSARIO DE LA UNIDAD

Componente cíclico

Componente del modelo de la serie de tiempo que causa una variación periódica sobre y debajo de la tendencia, y la variación dura más de un año.

Componente estacional

Componente del modelo de una serie de tiempo que muestra un patrón periódico de un año o menos.

Componente irregular

Componente del modelo de una serie de tiempo que refleja la variación aleatoria de los valores de la serie de tiempo, adicionales a los que se pueden explicar con los componentes de tendencia, cíclico y estacional.

Constante de suavizamiento

Parámetro del modelo de suavizamiento exponencial, con el que se calcula el factor de ponderación asignado al valor más reciente de la serie de tiempo en el cálculo del valor del pronóstico.

Elaboración de escenarios

Método cualitativo de pronóstico que consiste en formar un escenario conceptual del futuro, basado en un conjunto bien definido de supuestos.

Error cuadrático medio

Es un método con el que se mide la precisión de un modelo de pronóstico. Es el promedio de la suma de las diferencias entre los valores pronosticados y los valores reales de la serie de tiempo estando elevadas al cuadrado esas diferencias.

Modelo auto-regresivos

Modelo de serie de tiempo donde se usa una relación de regresión basada en valores anteriores de la serie para predecir valores futuros de la misma.

Modelos causales de pronóstico

Métodos de pronóstico que relacionan una serie de tiempo con otras variables que se cree explican o causan su comportamiento.

Modelo multiplicativo de serie de tiempo

Modelo en el cual se multiplican los componentes de la serie de tiempo, entre sí, para identificar el valor real de dicha serie. Cuando se suponen presentes los cuatro componentes de tendencia, cíclico, estacional e irregular, se obtiene: $Y_t = (T_t)(C_t)(E_t)(I_t)$. Cuando se modela el componente cíclico se obtiene: $Y_t = (T_t)(E_t)(I_t)$.

Promedios móviles

Método de pronóstico o suavizamiento de una serie de tiempo, en el que se promedia cada grupo sucesivo de puntos de datos.

Promedios móviles ponderados

Método de pronóstico o suavizamiento de una serie de tiempo con el que se calcula un promedio ponderado de los valores de datos en el pasado. La suma de los factores de ponderación debe ser igual a uno.

Pronóstico

Proyección o predicción de valores futuros de una serie de tiempo.

Serie de tiempo

Es un conjunto de observaciones medidas en puntos sucesivos en el tiempo, o durante periodos sucesivos en el tiempo.

Serie de tiempo des-estacionalizada

Serie de tiempo en la que se ha eliminado el efecto estacional, dividiendo cada observación original de la serie entre el correspondiente índice estacional.

Suavizamiento exponencial

Técnica de pronóstico que emplea un promedio ponderado de una serie de tiempo en el pasado para determinar valores de una serie de tiempo suavizada, que se pueden usar para elaborar pronósticos.

Tendencia

Desplazamiento o movimiento de la serie de tiempo, a largo plazo, observable a través de varios periodos.

ACTIVIDADES DE APRENDIZAJE

ACTIVIDAD 1

Elabora un cuadro comparativo de lo que representa cada una de las cuatro componentes de una serie de tiempo.

	Representa
Componente de tendencia	
Componente cíclica	
Componente de estacionalidad	
Componente irregular	

ACTIVIDAD 2

Elabora un resumen de la forma en que se separa la componente de tendencia en una serie de tiempo.

CUESTIONARIO DE REFORZAMIENTO

1. ¿Qué es una serie de tiempo?
2. ¿Cuáles son los elementos de una serie de tiempo?
3. ¿Cuál es el modelo más utilizado para descomponer una serie de tiempo?
4. Explica qué es la tendencia en una serie de tiempo.
5. ¿Cómo se produce la tendencia de una serie de tiempo?
6. Explica qué es la componente cíclica en una serie de tiempo.
7. Explica qué es la componente estacional en una serie de tiempo.
8. Explica qué es la componente irregular en una serie de tiempo.
9. ¿Cómo se produce la componente irregular de una serie de tiempo?
10. ¿Cuál es el objetivo del responsable del pronóstico en el análisis de predicciones?

EXAMEN DE AUTOEVALUACIÓN

Elige la respuesta correcta a las siguientes preguntas, una vez que concluyas, obtendrás de manera automática tu calificación.

1. En una serie de tiempo ¿Qué es la variación cíclica?
   a) Son las fluctuaciones de los valores alrededor de la línea de tendencia.
   b) Son fluctuaciones u oscilaciones ocasionadas por movimientos telúricos
   c) Es la oscilación armónica del modelo multiplicativo de la serie de tiempo.

2. ¿Qué fenómenos dan origen a la componente cíclica?
   a) Naturales como la lluvia y el viento
   b) Geológicos tales como los terremotos, temblores, etc.
   c) Sociales, económicos, políticos, costumbres locales, etc.,

3. La componente cíclica se calcula para cada valor real obtenido mediante la fórmula:
   a) $\hat{y}_i = b_0 + b_1 X_i$
   b) $c = \frac{Y}{Yt} 100$
   c) $C = \frac{Y}{T \quad E \quad I}$

4. En el cálculo de la componente cíclica para cada valor real, debemos auxiliarnos con la ecuación:
   a) de la recta de regresión
   b) del modelo multiplicativo de una serie de tiempo
   c) de tendencia de la serie de tiempo.

5. Cuando la serie de tiempo contiene datos diarios, semanales o mensuales, la primera componente que debe ser aislada es la:
   a) tendencia
   b) componente temporal.
   c) componente irregular.

6. En la expresión $\frac{(T)(C)(E)(I)}{(T)(C)} = (E)(I)$ obtenida a partir del modelo multiplicativo de una serie de tiempo, el resultado contiene:
   a) los efectos estacionales, junto con las fluctuaciones irregulares.
   b) la tendencia, junto con las fluctuaciones irregulares.
   c) solo las fluctuaciones irregulares.

7. Para separar la componente temporal es necesario tener:
   a) muy pocos datos
   b) una fuerte cantidad de datos
   c) ningún dato

LO QUE APRENDÍ

Los siguientes valores corresponden al tipo de cambio del dólar para 17 días consecutivos. Con estos datos pronostique usted mediante una serie de tiempo el tipo de cambio correspondiente para el día numero 18.

Día	($)
1	
2	13.9058
3	13.9777
4	13.9382
5	13.9145
6	13.9325
7	14.0950
8	13.9342
9	14.1675
10	14.1513
11	14.1975
12	14.3097
13	14.5404
14	14.4667
15	14.2945
16	14.1778
17	14.1392

MESOGRAFÍA

Bibliografía sugerida

Autor	Capitulo	Páginas
Levin y otros (1996)	15	673-712
Christensen (1990)	12	625 - 643
Lind y otros (2004)	12	602 - 624
Hanke y otros (1997)	16	668 - 691

Bibliografía básica

Levin, Richard I. y Rubin, David S. (1996). *Estadística para administradores*. México: Alfaomega.

Lind, A. Douglas; Marchal, G. William; Mason, D. Robert. (2004). *Estadística para Administración y Economía*. (11ª ed.) México: Alfaomega.

Bibliografía complementaria

Christensen, H. (1990). *Estadística paso a paso* (2ª ed.) México: Trillas.

Garza, Tomás. (1996). *Probabilidad y estadística*. México: Iberoamericana.

Hanke, John E. y Reitsch, Arthur G. (1997). *Estadística para Negocios.* México: Prentice Hall.

Sitios de Internet

Sitio	Descripción
http://ciberconta.unizar.es/LECCION/seriest/100.HTM	Arellano, M. (2001): "Introducción al Análisis Clásico de Series de Tiempo", [en línea] 5campus.com, Estadística
http://maxsilva.bligoo.com/content/view/186499/Series-de-Tiempo.html	Silva Quiroz, Maximiliano. (2008). “Series de tiempo”, Estadística y empresa (13/05/08)
http://ciberconta.unizar.es/LECCION/seriest/inicio.html	Arellano, Mireya. (2001). “Introducción al análisis clásico de series de tiempo”, 5campus,com. Estadística
http://www.eumed.net/cursecon/libreria/drm/1n.htm	Ruíz Muñoz, David. (2004). “Series temporales: Determinación de las variaciones estacionales”. *Manual de estadística*. EUMED.

UNIDAD 8

PRUEBAS ESTADÍSTICAS NO PARAMÉTRICAS

OBJETIVO ESPECÍFICO

Al terminar la unidad el alumno identificará las pruebas no paramétricas más utilizadas.

INTRODUCCIÓN

En esta unidad se revisarán las pruebas no paramétricas y su utilidad sobre todo cuando no se conoce la distribución de la cual provienen los datos, lo cual impide hacer una estimación por intervalos de confianza o una prueba de hipótesis.

Como se verá, las pruebas no paramétricas resultan más accesibles de realizar y comprender ya que no requieren cálculos laboriosos ni el ordenamiento o clasificación formal de datos o mediciones más exactas de parámetros poblacionales.

También se analizarán las pruebas como la prueba de rachas, definida como una secuencia de uno o más símbolos similares que se expresa como una serie continua de uno o más símbolos. La prueba del signo para probar la hipótesis de que “no hay diferencia en las medianas entre las distribuciones continuas de dos variables aleatorias X y Y, en la situación en la que podemos extraer muestras de X y Y”.

La prueba de signos y rangos de Wilcoxon utilizada como una alternativa no paramétrica cuando se trata de comparar los datos de 2 poblaciones o de una misma población mediante una muestra apareada. Y la prueba de los rangos con signo que usa los rangos de los valores absolutos de las diferencias pareadas.

LO QUE SÉ

Elige la respuesta correcta a la siguiente pregunta:

La fórmula del estadístico “z” es:

a) $z = \dfrac{\sum_{i=1}^{k} f_o^2}{f_e} - n$

b) $z = \dfrac{x - \mu}{\sigma}$

c) $z = \sum_{i=1}^{k} \dfrac{(f_o - f_e)^2}{f_e}$

TEMARIO DETALLADO

(6 horas)

8.1. Diferencias entre los métodos estadísticos paramétricos y no paramétricos
8.2. La prueba de rachas para aleatoriedad
8.3. La prueba del signo
8.4. La prueba de signos y rangos de Wilcoxon

8.1. Diferencias entre los métodos estadísticos paramétricos y no paramétricos

Las pruebas no paramétricas son útiles sobre todo cuando no se conoce la distribución del cual provienen los datos y, por tanto, no se conoce la distribución del estadístico para hacer una estimación por intervalos de confianza o una prueba de hipótesis. Estas pruebas son útiles por ejemplo cuando el tipo de datos es nominal u ordinal.

Generalmente son más fáciles de realizar y comprender ya que no requieren cálculos laboriosos ni el ordenamiento o clasificación formal de datos o mediciones más exactas de parámetros poblacionales.

Tipo de pruebas no paramétricas

Paso 1. Establecer la hipótesis nula (H_o) y la hipótesis alternativa (H_1).

La H_o indica que no hay diferencias significativas entre las frecuencias observadas y las frecuencias esperadas. Cualquier diferencia puede atribuirse al muestreo o a la casualidad. La H_i indica por lo tanto que si hay diferencias significativas entre una distribución esperada y la estimada para la población.

Paso 2. Elegir un nivel de significación (α).

Paso 3. Elegir y calcular el estadístico de prueba χ_e^2

Paso 4. Establecer la regla de decisión.

Paso 5. Calcular el valor de Chi-cuadrada crítica (χ_c^2) y tomar la decisión.

8.2. La prueba de rachas para aleatoriedad

Es una prueba que se utiliza para comprobar la aleatoriedad de muestras. Es muy importante demostrar la aleatoriedad de las muestras en los estudios estadísticos. Si no es así, se crea una gran desconfianza en los procesos de muestreo.

En una prueba de rachas, se asigna a todas las observaciones de la muestra uno o dos símbolos. Una racha se designa como una secuencia de uno o más símbolos similares y también se expresa como una serie continua de uno o más símbolos. Si el número de rachas es menor de 20, se utilizan tablas específicas en donde se muestran valores críticos mínimos y máximos por lo que si el número de rachas (r) es menor o excede de esos valores críticos, se indica una ausencia de aleatoriedad.

Si se tienen 2 categorías y los datos muestrales no caen en alguna de ellas, se puede utilizar la mediana como valor de referencia.

Una importante aplicación de la prueba de rachas se encuentra en el método de mínimos cuadrados en el análisis de regresión. Una propiedad básica en estos modelos de regresión es que los errores son aleatorios.

Las hipótesis para probar son:

H_o : Existe aleatoriedad en las muestras.

H_1 : No existe aleatoriedad en las muestras.

Si el número de datos en 2 categorías n_1 y n_2 son mayores que 20, la distribución de muestreo para "*r*" se aproxima a una distribución normal.

Las fórmulas son:

Media de la distribución muestral del número de rachas:

$$\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1$$

Desviación estándar:

$$\sigma_r = \sqrt{\frac{2n_1n_2 \;\; 2n_1n_2 - n_1 - n_2}{n_1+n_2 \;^2 \;\; n_1+n_2-1}}$$

Estadístico de prueba:

$$z = \frac{r-\mu_r}{\sigma_r}$$

Ejemplo de aplicación; en una campaña a 100 posibles compradores de un producto especializado, se realizaron 52 ventas, 48 no ventas y 40 rachas. A un nivel de significación del 1% probar la hipótesis que la muestra es aleatoria.

Las hipótesis son:

H_o : La muestra es aleatoria.

H_1 : La muestra no es aleatoria.

Estadístico de prueba: $z = \dfrac{r - \mu_r}{\sigma_r}$

La media es: $\mu_r = \dfrac{2n_1n_2}{n_1 + n_2} + 1 = \dfrac{2 \cdot 52 \cdot 48}{52 + 48} + 1 = 50.92$

La desviación estándar:

$$\sigma_r = \sqrt{\frac{2n_1n_2 \ \ 2n_1n_2 - n_1 - n_2}{n_1 + n_2 \ ^2 \ \ n_1 + n_2 - 1}} = \sqrt{\frac{2 \cdot 52 \cdot 48 \ \ 2 \cdot 52 \cdot 48 - 52 - 48}{52 + 48 \ ^2 \ \ 52 + 48 - 1}} = \sqrt{24.67} = 4.97$$

Por lo tanto: $z = \dfrac{r - \mu_r}{\sigma_r} = \dfrac{40 - 50.92}{4.97} = -2.20$

Nivel de significación: $\alpha = 0.01$ por lo que $z_c = \pm 2.58$ ya que es una prueba de 2 colas. Como $z < z_c$ cae en la zona de aceptación se puede concluir que no hay evidencia suficiente para rechazar la hipótesis nula, por lo que se puede indicar que la muestra es aleatoria.

8.3. La prueba del signo

En las estadísticas, la prueba de los signos se utiliza para probar la hipótesis de que “no hay diferencia en las medianas entre las distribuciones continuas de dos variables aleatorias X y Y, en la situación en la que podemos extraer muestras de X y Y”.

Se trata de una prueba no paramétrica que hace unos pocos supuestos muy cerca de la naturaleza de las distribuciones bajo prueba -esto significa que tiene una aplicación muy general, pero pueden carecer de la potencia estadística de otras pruebas como el dos a dos muestras de T (test).

Formalmente, sea p = Pr $(X> Y)$, y luego probar la hipótesis nula $H_{0:}\ p$ = 0.50. En otras palabras, bajo la hipótesis nula de que dado un azar par de medidas $(x_{i,}\ y_{i),}$ entonces x_i e y_i son las mismas probabilidades de ser más grande que el otro. (Wikipedia, 2001, “Sign test”)

Puesto que la distribución normal es simétrica, la media de una distribución normal es igual a la mediana. Por consiguiente, la prueba del signo puede emplearse para probar hipótesis sobre la media de una población normal.

8.4. La prueba de signos y rangos de Wilcoxon

Se utiliza como una alternativa no paramétrica cuando se trata de comparar los datos de 2 poblaciones o de una misma población mediante una muestra apareada en la que cada unidad experimental genera 2 observaciones pareadas o ajustadas, una de la población 1 y una de la población 2. Las diferencias entre las observaciones pareadas permiten tener una buena perspectiva respecto de la diferencia entre las 2 poblaciones.

La metodología del análisis paramétrico de una muestra pareada requiere de datos de intervalo y de la suposición de que la población de las diferencias entre los pares de observaciones tenga una distribución normal. Con este supuesto se puede usar la distribución “t” para probar la hipótesis nula es decir que no hay diferencias entre las medias poblacionales. Si no es así, se debe utilizar la prueba de rango con signo de Wilcoxon.

La prueba de los rangos con signo usa los rangos de los valores absolutos de las diferencias pareadas, asignando el rango 1 a la diferencia con valor absoluto mínimo, el rango 2 a la siguiente diferencia con menor valor absoluto y así se procede sucesivamente. Se deben descartar los rangos con diferencias de cero y en caso de valores absolutos repetidos, a cada uno de ellos se les otorga el valor promedio de los rangos ocupados por los valores repetidos. A cada uno de los rangos positivos o negativos, se les asocia el signo correspondiente.

La suma de los rangos positivos se indica por T^+, la suma de los rangos negativos se denota por T^- y el máximo valor entre estos 2 valores se escribe solamente *"T"* y se utiliza generalmente como estadístico de prueba. Si el número de diferencias es igual o mayor que 15 entonces la distribución muestral de *"T"* es aproximadamente normal por lo que se utilizará la variable parametrizada "z". Si es menor se deberán utilizar tablas especiales que proporcionan los valores críticos para la prueba de rangos con signo.

La suma de los rangos es: $S = \frac{n\ n+1}{2}$ y deberá ser igual a $T^+ + T^-$

Las fórmulas de la media y desviación estándar de la distribución muestral *"T"* son las siguientes:

Media: $\mu_T = \frac{n\ n+1}{4}$

Desviación estándar: $\sigma_T = \sqrt{\frac{n\ n+1\ \ddot{}2n+1}{24}}$

y el estadístico de prueba es: $z = \frac{T - \mu_T}{\sigma_T}$.

Ejemplo de aplicación; se desea saber si un programa de capacitación en cómputo en una empresa especializada, mejoró las habilidades de los empleados en dicha materia. Por ello se observa el nivel de habilidades antes del programa y después del programa en una muestra de 22 empleados, obteniéndose los siguientes resultados y probar la hipótesis a un nivel de significación del 1%.

					Diferencias	Rango	Rangos
Número	Puntaje		Diferencia		absolutas		con signos
Empleado	Antes (a)	Después (b)	b-a		ordenadas		correctos
1	18	15	-3		2	1	1
2	60	70	10		3	2	-2
3	81	75	-6		4	3	-3
4	15	20	5		5	4	4.5
5	20	50	30		5	5	4.5
6	17	40	23		6	6	-6
7	26	50	24		8	7	-7.5
8	11	30	19		8	8	7.5
9	20	40	20		9	9	-9
10	38	30	-8		10	10	10.5
11	80	85	5		10	11	10.5
12	59	86	27		11	12	12
13	12	72	60		19	13	13
14	87	98	11		20	14	15
15	88	79	-9		20	15	15
16	64	88	24		20	16	15
17	88	90	2		23	17	17
18	76	96	20		24	18	18.5
19	43	39	-4		24	19	18.5
20	90	98	8		27	20	20
21	40	60	20		30	21	21
22	50	60	10		60	22	22

Se obtienen las diferencias de los puntajes antes y después, sus diferencias, las diferencias absolutas ordenadas, sus rangos y los rangos con signos correctos.

La suma de rangos positivos es: $T^{+} = 225.5$

La suma de rangos negativos es: $T^{+} = 27.5$

Comprobación: $$S = T^{+} + T^{-} = \frac{n\ \ n+1}{2} = \frac{22\ \ 22+1}{2} = 253.0$$

Por lo tanto $T = 225.5$

La hipótesis por probar son:

Ho: No hay diferencia significativa debido al tratamiento.

Ha: Hay diferencia significativa por el tratamiento

La columna de rangos con signos correctos se determinó mediante el promedio de rangos, si la diferencia absoluta se repite y los rangos son signos correctos se preserva el signo de la diferencia que le dio origen. Por ejemplo, para el rango 4 y 5 se promedió (4+5)/2=4.5 y como el rango 4 corresponde a una diferencia 5 positiva entonces se le asigna 4.5 positivo, lo mismo para el rango 5. En el caso de los rangos 7 y 8 (correspondientes a una diferencia de 8), el promedio es 7.5 y como la diferencia de 8 corresponde a un valor negativo y otro positivo, entonces se le asigna un rango con signo correcto de -7.5 y 7.5.

Estadístico de prueba: $$z = \frac{T - \mu_T}{\sigma_T}$$

La media es: $$\mu_T = \frac{n\ \ n+1}{4} = \frac{22 \cdot 23}{4} = 126.5$$

La desviación estándar: $\sigma_T = \sqrt{\frac{n\ \ n+1\ \ 2n+1}{24}} = \sqrt{\frac{22 \cdot 23 \cdot 43}{24}} = 30.1$

Por lo tanto: $z = \frac{T - \mu_T}{\sigma_T} = \frac{225.5 - 126.5}{30.1} = 3.29$

Nivel de significación: $\alpha = 0.01$ por lo que $z_c = 2.33$

Como $z > z_c$ cae en la zona de rechazo, se puede concluir que el programa de capacitación de cómputo en esta empresa sí mejoró las habilidades del personal.

RESUMEN DE LA UNIDAD

En esta unidad se revisaron las pruebas no paramétricas más utilizadas, cuando no se conoce la distribución de la cual provienen los datos, como se pudo observar, las pruebas no paramétricas resultan más accesibles de realizar y comprender ya que no requieren mediciones más exactas de parámetros poblacionales.

También se analizó las pruebas como la prueba de rachas y la prueba del signo para probar la hipótesis de que “no hay diferencia en las medianas entre las distribuciones continuas de dos variables aleatorias X y Y”, y la prueba de los rangos con signo que usa los rangos de los valores absolutos de las diferencias pareadas.

GLOSARIO DE LA UNIDAD

Métodos no paramétricos

Métodos estadísticos que requieren muy pocos o ningún supuesto acerca de las distribuciones de probabilidad de la población, y acerca del nivel de medición. Estos métodos se pueden aplicar cuando se dispone de datos nominales u ordinales.

Métodos sin distribución

Es otro nombre que se da a los métodos estadísticos no paramétricos, que indica la carencia de supuestos sobre la distribución de probabilidad de la población.

Prueba de signo

Prueba estadística no paramétrica que permite identificar diferencias entre dos poblaciones basándose en el análisis de datos nominales.

Prueba de rango con signo de Wilcoxon

Prueba estadística no paramétrica con la cual se identifican diferencias entre dos poblaciones, basada en el análisis de dos muestras pareadas o ajustadas.

ACTIVIDADES DE APRENDIZAJE

ACTIVIDAD 1

1. Una manufacturera automotriz desea conocer la preferencia de los clientes por los colores ocre o índigo del modelo de lujo, pues sólo uno saldrá al mercado. Se invitó a los 20 mejores vendedores para que opinaran y se encontró que doce prefirieron el color ocre, siete el índigo y uno indeciso. En un nivel del 10% probar si:

H_0: Cualquier color gustará por igual a los clientes

H_1: Hay preferencia por alguno de los colores de los clientes

2. Para el aniversario de la empresa se organizó una convención y se dio a escoger entre el menú tradicional o uno especial. La muestra fue de 81 clientes de los cuales 42 prefirieron el especial. Utilizando la prueba del signo y un nivel de 0.02, pruebe si a los clientes les gustó más el menú especial que el tradicional:

H_0: Ambos menús gustaron por igual ($p=0.50$)

H_1: Gustó más el menú especial ($p>0.50$)

CUESTIONARIO DE REFORZAMIENTO

1. ¿Qué pruebas son más efectivas: las paramétricas o las no paramétricas?

2. El entrenador de un equipo de ciclismo determina al azar la presión de las llantas de las bicicletas antes de la carrera. Si la presión no es correcta la registra como muy baja (B) o muy alta (A). A continuación se dan los datos. Utiliza la prueba correcta para determinar, con un nivel de significancia de 0.10, si la presión de las llantas tiende a ser muy alta o muy baja, o si la ocurrencia de alta y baja se puede considerar igual.

   A A B B B A A B B B B A A B B B

3. El número de juegos de video vendidos por semana durante varias semanas se organiza en una secuencia de baja a alta; se designan por A o B, que representan a los dos vendedores clave de la compañía. El distribuidor de videos está interesado en analizar su volumen de ventas.

4. ¿Pueden los vendedores considerarse igualmente efectivos? Prueba con un nivel de significancia de 0.05.

   A,A,B,A,A,B,B,A,A,A,A,B,B,A,A,B

   A,B,A,B,B,B,A,B,A,B,B,B,A,B,B,B

5. ¿Cuál es la diferencia esencial entre los métodos estadísticos paramétricos y los no paramétricos?

6. Enumera las razones por las que elegirías un método no paramétrico para analizar datos muestrales.

7. ¿Qué prueba no paramétrica es similar a la prueba del signo de una muestra?

8. Un generador de números aleatorios genera números positivos y negativos en forma aleatoria. Después de verificar la primera serie de números, el analista piensa que la serie parece aleatoria, pero decide que debe realizarse una prueba estadística antes de usar el programa en toda la empresa. Se presenta la serie de números observada, donde P representa a un número positivo y N a uno negativo. ¿Parece ser aleatorio el programa?

   a) Prueba con un nivel de significancia de 0.5

   b) Prueba con un nivel de significancia de 0.10

EXAMEN DE AUTOEVALUACIÓN

Elige la respuesta correcta a las siguientes preguntas, una vez que concluyas, obtendrás de manera automática tu calificación.

1. Se utiliza para probar la hipótesis de que “no hay diferencia en las medianas entre las distribuciones continuas de dos variables aleatorias X y Y, en la situación en la que podemos extraer muestras de X y Y”.
    a) la prueba de los signos
    b) prueba de Mann-Whitney-Wilcoxon
    c) coeficiente de correlación de rango de Spearman

2. Son útiles sobre todo cuando no se conoce la distribución del cual provienen los datos
    a) la prueba de los signos
    b) prueba de Mann-Whitney-Wilcoxon
    c) las pruebas no paramétricas

3. Estas pruebas son útiles por ejemplo cuando el tipo de datos es nominal u ordinal.
    a) la prueba de los signos
    b) las pruebas no paramétricas
    c) prueba de Mann-Whitney-Wilcoxon

4. Se utiliza como una alternativa no paramétrica cuando se trata de comparar los datos de 2 poblaciones o de una misma población mediante una muestra apareada
    a) la prueba de signos y rangos de Wilcoxon
    b) las pruebas no paramétricas
    c) prueba de Mann-Whitney-Wilcoxon

LO QUE APRENDÍ

Explica la diferencia entre una prueba estadística paramétrica y una prueba estadística no paramétrica.

MESOGRAFÍA

Bibliografía sugerida

Autor	Capítulo	Páginas
Berenson y otros (2001)	13	461-522
Levin y otros (1996)	14	621-663
Christensen (1990)	12	623 - 643
Lind y otros (2004)	18	671 - 693

Bibliografía básica

Berenson, L. Mark; Levine, M. David; Krehbiel, C. Timothy. (2001). *Estadística para Administración*. (2ª ed.) México: Prentice Hall.

Levin, Richard I. y Rubin, David S. (1996). *Estadística para administradores*. México: Alfaomega.

Lind, A. Douglas; Marchal, G. William; Mason, D. Robert. (2004). *Estadística para Administración y Economía*. (11ª ed.) México: Alfaomega.

Bibliografía complementaria

Ato, Manuel y López, Juan J. (1996). *Fundamentos de estadística con SYSTAT*. México: Addison/Wesley.

Christensen, H. (1990). *Estadística paso a paso* (2ª ed.) México: Trillas.

Garza, Tomás. (1996). *Probabilidad y estadística*. México: Iberoamericana.

Hanke, John E. y Reitsch, Arthur G. (1997). *Estadística para Negocios.* México: Prentice Hall

----------. (1996). *Pronósticos en los Negocios*, México; Prentice Hall.

Hildebran, David y Lyman, Ott. (1998). *Estadística aplicada a la administración y a la economía*. (3ª ed.) México: Addison Wesley

Kazmier L. y Díaz Mata, A. (1998). *Estadística aplicada a la administración y economía*, México; McGraw-Hill.

Mendenhall W. y Sheaffer, R.L. (1986). *Estadística matemática con aplicaciones*, México; Iberoamérica.

----------. (1987). *Elementos de Muestreo,* México; Iberoamericana.

Meyer, Paul L. (2002). *Probabilidad y aplicaciones estadísticas,* México; Addison Wesley Iberoamericana.

Weimer, Richard C. (1996). *Estadística*. México, CECSA.

Sitios de Internet

Sitio	Descripción
http://www.itch.edu.mx/academic/industrial/estadistica1/cap04.html	Torre, Leticia, de la. (2003) "Pruebas chi-cuadrada y estadística no paramétrica", Curso de Estadística I, Instituto Tecnológico de Chihuahua.
http://scientific-european-federation-osteopaths.org/es/prueba-estadisticas	Scientific European Federation of Osteopaths. (2012). "Las pruebas estadísticas" *Metodología de la investigación científica*.
http://www.uclm.es/actividades0708/cursos/estadistica/pdf/descargas/SPSS_PruebasNoParametricas.pdf	Sánchez Sánchez, Fco. (2008). "SPSS Pruebas no paramétricas", Curso de Estadística avanzada, UCLM,
http://scientific-european-federation-osteopaths.org/es/test-estadisticos	Scientific European Federation of Osteopaths. (2012). "Los test estadísticos" *Metodología de la investigación científica*.