ESTADÍSTICA APLICADA

A LAS CIENCIAS I

CÓDIGO 3125

GUÍA DE ESTUDIO PARA EL LIBRO

Bioestadística. Principios y Procedimientos.

STEEL y TORRIE

Preparada por

Ana Lorena Solís

UNIVERSIDAD ESTATAL A DISTANCIA

VICERRECTORÍA ACADÉMICA

ESCUELA DE CIENCIAS EXACTAS Y NATURALES

Edición académica

Virginia Ramírez

Encargado de cátedra

Carmen Andrés Jiménez

Revisión filológica

Armando Ríos

Esta guía de estudio ha sido confeccionada para ser utilizada en los programas de Ingeniería Agronómica, Ingeniería Agroindustrial, Manejo y Conservación de Recursos Naturales y Administración de Empresas Agropecuarias que imparte la UNED.

PRESENTACIÓN

Desde la década de los setentas hasta la fecha, Costa Rica ha experimentado un auge en el uso de la estadística en sus actividades cotidianas, es común ver como empresas privadas e instituciones públicas ponen a disposición de los gestores datos que apoyarán la solución de problemas y la forma de analizarlos, esta situación genera una cultura de toma de decisiones basada en datos estadísticos.

En el escenario mundial, se exponen los problemas alimentarios debido a la velocidad de crecimiento de la población en relación con la capacidad para disponer de alimentos para la humanidad, esto ha obligado a que especialistas de diferentes áreas se interesen en mejorar las condiciones de producción, evaluar alternativas de solución a problemas y diseñar sistemas agroproductivos consecuentes con la conservación del ambiente.

Esta preocupación se traduce en investigación aplicada que intenta llevar a la práctica el conocimiento para descubrir los elementos que permiten mejorar tanto la calidad y cantidad como la seguridad de los productos agrícolas.

Sin embargo, aunque la cultura de toma de decisiones, basada en datos, se encuentra en proceso, el mayor problema que enfrentan algunos profesionales es la falta de formación estadística para el uso adecuado de las herramientas técnicas que les permitan el logro de objetivos de investigación, de mejoramiento continuo y de análisis en la búsqueda de soluciones de forma exitosa.

En este contexto, formarse en estadística es una necesidad urgente, un profesional que no conozca sobre el manejo de datos se encuentra en desventaja para el desempeño de sus funciones de una forma científica.

La presente guía de estudio tiene como propósito apoyar al estudiante en el aprendizaje de la Estadística aplicada a los campos de la Agroindustria, Administración de Empresas Agropecuarias, Ingeniería Agronómica y Manejo de Recursos Naturales, en particular, con el uso del libro de *"Bioestadística: Principios y Procedimientos"* que estimula el pensamiento analítico que requieren estos profesionales en la ejecución de sus funciones.

CONTENIDO

ÍNDICE DE CUADROS

DESCRIPCIÓN DEL CURSO

1. PROPÓSITO DE LA ASIGNATURA

Lograr que el estudiante de ciencias exactas domine las técnicas y métodos de análisis estadístico básico en el nivel descriptivo e inferencial que le permitan desarrollar de forma exitosa trabajos de investigación, una adecuada presentación de resultados y obtener conclusiones válidas para la toma de decisiones en el desempeño profesional.

El enfoque del curso está orientado a la aplicación de las técnicas y su correcta interpretación, de forma que el futuro profesional cuente con los elementos necesarios para garantizar un uso correcto de la estadística en el ejercicio de sus funciones.

2. OBJETIVOS DE LA ASIGNATURA

Al finalizar el curso de Estadística Aplicada a las Ciencias I usted será capaz de:

1. Adquirir conocimientos sobre los elementos básicos de la estadística y su aplicación para la comprensión de los fenómenos propios del campo.
2. Aplicar conocimientos sobre los elementos básicos acerca de la presentación de datos estadísticos para su uso e interpretación.
3. Aplicar conocimientos sobre el manejo de medidas de posición y tendencia central y de su variabilidad para el análisis estadístico de datos.
4. Adquirir conocimientos generales sobre la técnica de asociación entre dos características cuantitativas o dos cualitativas para la generación de conocimiento predictivo en el campo afín.
5. Adquirir conocimientos generales del proceso de inferencia estadística como herramienta de análisis para la generación de conocimiento en el campo de desempeño profesional.

6. Adquirir conocimientos sobre los elementos básicos de la recolección de datos estadísticos a partir de fuentes de información existentes o en su defecto los conocimientos para crear los instrumentos necesarios para obtener la información que se requiere en el proceso de investigación.

3. PROPÓSITO DE LA GUÍA DE ESTUDIO

La guía de estudio orienta al estudiante en las lecturas propias de cada tema y en los aprendizajes mínimos esperados, los cuales podrá alcanzar con ayuda de los ejercicios resueltos y sugeridos para su autoevaluación.

Se ofrece un glosario de términos, un resumen de cada tema y ayudas esquemáticas que facilitarán su aprendizaje. Adicionalmente en algunos temas se ofrece un apoyo extra mediante presentaciones con Power Point ® que resumen fórmulas, interpretaciones y ejemplos aplicados.

Además se induce al estudiante en el manejo de las herramientas de internet, las cuales reforzarán los conocimientos adquiridos mediante el uso de los materiales básicos.

4. MATERIAL DEL CURSO

El material básico del curso es el siguiente:

STEEL y TORRIE. (1992). *Bioestadística. Principios y Procedimientos*. México: Editorial Graf América (622 p.)

METODOLOGÍA

De forma sistemática encontrará, en el abordaje de cada tema, lo siguiente:

1. El capítulo o los capítulos que cubre el tema en el libro de texto indicado en los materiales básicos del curso.
2. Información de material adicional en formatos de documentos de Word ®, presentaciones en Power Point o enlaces digitales en la web.
3. Ejercicios que refuerzan los conceptos o técnicas estadísticas tratados en el tema, incluye algunas bases de datos en formato Excel ® como apoyo.
4. Indicaciones generales sobre aspectos de precaución, importancia y advertencia acerca de aspectos técnicos que deben considerarse para asegurar el adecuado uso de la estadística.
5. Elementos claves presentados en forma de resumen que debe conocer para concluir de manera exitosa el estudio.
6. Una guía de ejercicios de autoevaluación que le permiten valorar el conocimiento alcanzado o identificar, si requiere reforzar, algunos conceptos.

1. RECOMENDACIONES PARA EL APRENDIZAJE

Recomendaciones para la elaboración de mapas conceptuales

Antes de dar alguna recomendación al respecto, es necesario definirlo y comprender las razones para su uso. Un mapa conceptual es una herramienta que permite representar el conocimiento en forma sencilla, práctica y esquemática; a la vez, facilita el aprendizaje de lo inductivo a lo deductivo y de lo general a lo específico con lo cual se establece un orden jerárquico en su construcción.

Se ha adoptado una forma estándar de presentación de este tipo de mapas, donde los conceptos son encerrados en círculos o cajas y las relaciones entre ellos se muestran mediante el uso de líneas, se acostumbra incluir descripciones sencillas.

Para ilustrar se presenta un mapa conceptual de la estadística, en la figura 1, su relación con la generación de conocimiento y la interrelación con el método científico. Incluye el concepto de sus dos grandes ramas: descriptiva e inferencial.

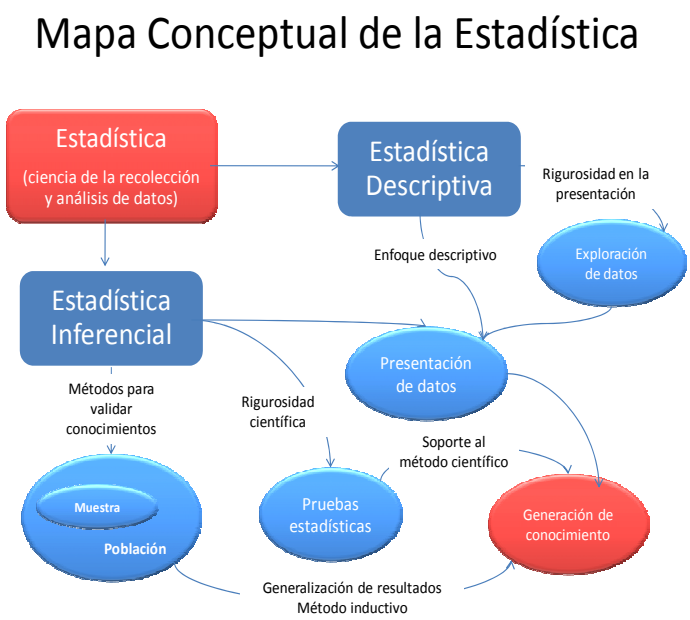


Figura 1. Mapa conceptual de la Estadística

Recomendaciones para la elaboración de un esquema resumen

En términos generales se puede definir un esquema como una síntesis descriptiva que representa los conceptos e ideas más relevantes de un tema, con frecuencia se utilizan formas gráficas que ayudan en la exposición con el fin de simplificarlo.

Existe variedad de esquemas, los más conocidos son: los de llaves, los de barras y los de resumen. En este último, se busca sintetizar la relación entre conceptos y deducciones logradas, así como la representación gráfica que facilite su comprensión y memorización.

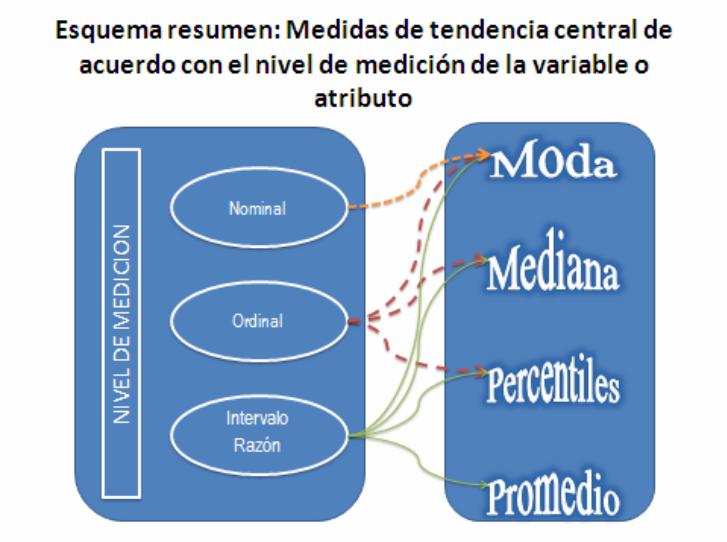


Figura 2. Esquema resumen. Medidas de tendencia central de acuerdo con el nivel de medición de la variable o atributo

Recomendaciones para la elaboración de un cuadro comparativo

Un cuadro comparativo es una herramienta que permite determinar similitudes y diferencias entre dos o más temas u objetos mediante el uso de una tabla, se coloca en las columnas los temas u objetos y en las filas se establecen las categorías de comparación, de esta forma se facilita la interpretación. A manera de ejemplo, se presentan, en el cuadro 1, algunas diferencias entre los conceptos de estimador y parámetro, se identifican tres categorías básicas en las que se pueden establecer diferencias.

Cuadro 1. Cuadro comparativo entre estimador y parámetro

Categorías	Estimador	Parámetro
Es función de	Una muestra	Una población
Se conoce su valor	Por medio de muestreo	Se supone conocido y se aproxima por medio de un estimador
Tiene un valor único	Varía de muestra a muestra	Tiene un único valor en la población y puede ser estimado por varios estimadores

Recomendaciones para la elaboración de una ficha técnica

Una ficha técnica es un resumen de las condiciones y características de un estimador como la que se muestra en el cuadro 2, por lo general, se utilizan tarjetas de forma rectangular, en las cuales se incluye la definición de la medida estadística, su alcance, su fórmula de cálculo y su interpretación.

Cuadro 2: Ficha técnica para la desviación estándar

Ficha técnica para la desviación estándar	
Definición	Mide la dispersión o variabilidad del conjunto de datos sobre la característica de interés. Tiene las mismas unidades de medida que la variable en estudio. Su cuadrado es la varianza. No se puede obtener la desviación sin antes calcular la varianza
Alcance	Aplica solamente en características de tipo cuantitativo
Interpretación	Representa la desviación promedio entre la media de la muestra y cualquier otro valor en el conjunto de datos, está expresada en unidades absolutas. Por ejemplo si se trata de la característica edad medida en años, su desviación estándar estará medida en años
Fórmula	Se calcula la varianza y se obtiene la raíz cuadrada: $s^2 = \frac{1}{n-1}\sum_{1}^{n}\left(x_i - \bar{x}\right)^2$

ENLACES DIGITALES

Se presentan a continuación una lista de enlaces sugeridos, donde se pueden encontrar definiciones, explicaciones, ejercicios resueltos y algunas recomendaciones para el tratamiento de la estadística y su aplicación.

Cuadro 3. Enlaces recomendados

Nombre del sitio	Enlace
Sistema de información del sector agrario costarricense	http://www.infoagro.go.cr/
Open Course Ware Universidad de Sevilla	http://ocwus.us.es/metodos-de-investigacion-y-diagnostico-en-educacion/analisisde-datos-en-la-investigacion-educativa/Bloque_II/page_26.htm
Depósito de documentos de la FAO	http://www.fao.org/docrep/p3350S/p3350s05.htm
Estadística para Ciencias Ambientales	http://www.uam.es/personal_pdi/ciencias/ajustel/docencia/estadisccaa/estadisccaa.html
Estadística para Agronomía	http://www.unlu.edu.ar/~estadistica/agronomia/
Manual de Estadística elaborado por David Ruiz Muñoz de la Universidad Pablo de Olavide	http://www.eumed.net/cursecon/libreria/drm/ped-drm-est.htm
Universidad Nacional de Luján	http://www.unlu.edu.ar/~estadistica/agronomia/
Ciencia y Técnica Administrativa: instrumento destinado a transferir el conocimiento generado por la actividad científica en Argentina	http://cyta.com.ar/biblioteca/bddoc/bdlibros/guia_estadistica/

TEMA 1: GENERALIDADES DE LA ESTADÍSTICA

1. OBJETIVO

Adquirir conocimientos sobre los elementos básicos de la estadística y su aplicación para la comprensión de los fenómenos propios del campo.

Cuadro 4. Generalidades de la Estadística

Tema	Capítulo del libro	Páginas
Fases de la investigación científica	Capítulo 1	4 y 5
Estadística y método científico: generalidades de la investigación, fuentes de información, elaboración de encuestas, tipos de preguntas	Capítulo 1	5
Clasificación de la estadística: descriptiva e inferencial, paramétrica-no paramétrica	Capítulo 24 Texto en la guía	520 y 521
Conceptos básicos: universo, población, muestra, unidad, observación, atributo, variable continua, variable discreta	Capítulo 2	7 a 10
Elementos de muestreo: tipos, sesgo y error de muestreo	Capítulo 4 Capítulo 25	541 a 543
Tamaño y selección de la muestra	Presentación de Power Point: Elementos de muestreo	
	Hoja de *Excel:* Datos sobre muestreo de árboles	
Tipos de muestreo: aleatorio y no aleatorio	Capítulo 4	65 a 68 incluye sección 4.3
	Capítulo 25	543 a 547
Muestreo estratificado, polietápico y de conglomerados	Capítulo 25	547 a 557
Recolección de datos: fuentes de información, instrumentos	Capítulo 2	10 a 11

2. DEFINICIÓN DE ESTADÍSTICA Y SU CLASIFICACIÓN

Estadística descriptiva e inferencial

La palabra "estadística" tiene dos acepciones:

- Como sinónimo de dato: este se refiere al concepto popular de la estadística, es equivalente a datos con algún ordenamiento coherente y sistemático.
- Como ciencia: en la cual la estadística estudia el comportamiento de los fenómenos naturales y sociales en todas las ciencias. Esto se debe a la particularidad de ofrecer técnicas y métodos precisos para obtener información y analizarla.

La Estadística, dependiendo del alcance de sus técnicas y métodos, puede ser clasificada en dos grandes tipos:

1. Estadística descriptiva o deductiva: se utiliza para resumir de forma numérica o gráfica un conjunto de datos, su análisis está limitado a lo que estos describen.
2. Estadística inferencial o inductiva: permite realizar conclusiones o generalizaciones, basándose en los datos resumidos y analizados de una muestra hacia la población o universo del cual proviene la muestra.

Estadística paramétrica versus no paramétrica

La estadística inferencial es comúnmente llamada Estadística paramétrica, debido a que se basa en el muestreo de una población a partir de la cual se estimará un parámetro como, por ejemplo la media (μ), la desviación estándar (σ) o la proporción (P). Estos métodos se ajustan a condiciones estrictas, así como el requisito de que los datos de la muestra provengan de una población normalmente distribuida.

Cuando el supuesto de normalidad no se cumple, surge lo que se conoce como la Estadística no paramétrica, la cual puede ser definida como un conjunto de técnicas que no requieren de supuestos en torno a la población de la cual

provienen, es decir, adopta una distribución "libre", el objetivo es completar el proceso de inferencia mediante la aplicación de técnicas menos robustas.

En aplicaciones que involucran datos que no poseen un alto nivel de medición, es decir nominal u ordinal, se dispone de muestras pequeñas o el supuesto de normalidad no se cumple, el uso de técnicas no paramétricas es recomendado.

La principal desventaja de estos métodos es que pierden información, en general, las pruebas de esta rama de la Estadística son menos eficientes que sus contrapartes paramétricas.

3. EJERCICIOS

Con el objetivo de reforzar los conocimientos adquiridos en este tema se recomienda que desarrolle los siguientes ejercicios:

1. Desarrolle los ejercicios 2.2.1, 2.4.2, 2.5.2, 2.5.4 y 2.5.7.

2. Utilice la lista de árboles (archivo Excel "Datos sobre Muestreo de Árboles") y la tabla A.1 Diez mil dígitos aleatorizados del libro para seleccionar una muestra aleatoria simple de tamaño 10, seleccione como punto de arranque la columna 30-31-32 en la fila 00, siguiendo de forma vertical, luego salte a las columnas 35-36-37 para completar la selección.

3. Seleccione una muestra de 10 observaciones de la lista de árboles (archivo Excel "Datos sobre Muestreo de Árboles") utilizando el método de selección forma sistemática. Inicie en la posición 5.

4. OBSERVACIONES FINALES

Se debe prestar atención al tipo de muestra cuando se trabaja en el campo de la inferencia estadística, es común enfocar la investigación al tamaño o número de casos de estudio, siendo esta interpretación incorrecta, pues, igual importancia, tiene el tipo de muestreo y la selección misma de los elementos que la conforman; en general, estos conceptos definen el diseño del muestreo que va más allá del número de casos. Cuidar de estos aspectos contribuye a elaborar muestras

estadísticas representativas y aleatorias, elementos básicos para la validez de las conclusiones en cualquier investigación.

El error más común en estos estudios se concentra en la medición del "error de muestreo" asociado a las estimaciones, principalmente, cuando se trata de selecciones no aleatorias. Se debe tener claridad de que el objetivo de calcular el error es conocer de forma aproximada las diferencias que se observan en la población que se atribuyen a comportamientos aleatorios y no a efectos controlados. Por ejemplo, la valoración del impacto de un tratamiento médico específico debe realizarse con diseños experimentales y no con estudios por muestreo tradicionales, debido a que el interés está en conocer como este puede afectar la respuesta del individuo.

5. RESUMEN DEL TEMA

Al finalizar el tema se espera que usted sea capaz de:

- Definir el concepto de Estadística descriptiva e inferencial, sus alcances y limitaciones.
- Definir el concepto de Estadística paramétrica y no paramétrica, sus alcances y limitaciones.
- Definir características de orden estadístico con interés observacional e investigativo.
- Definir los conceptos básicos del muestreo: unidad estadística, muestra, universo y tipos de muestreo.
- Describir la relación entre estadística e investigación científica.
- Diferenciar e identificar variables y atributos.
- Elaborar una muestra simple al azar para elaborar una muestra estadística.
- Elaborar un instrumento (cuestionario) para la adecuada recolección de datos en un trabajo de investigación de campo.
- Valorar la estadística como medio de representación de la realidad.

6. SOLUCIÓN A LOS EJERCICIOS

1. Desarrolle los ejercicios 2.2.1, 2.4.2, 2.5.2, 2.5.4 y 2.5.7.

Ejercicio 2.2.1

Clasificar las siguientes variables como cuantitativas, cualitativas, continuas o discretas según el caso como se muestra en el cuadro 5.

Cuadro 5. Clasificación de variables

Característica	Tipo	Continua o discreta
Color de los ojos	Cualitativa	Continua
Número de errores por estudiante en un examen de deletreo	Cuantitativa	Discreta
Tiempo para recargar de tinta un estilógrafo que se usa normalmente	Cuantitativa	Continua
Número de niños en el hospital más cercano en el día de año nuevo	Cuantitativa	Discreta
Número de peces en un estanque	Cuantitativa	Discreta
Recuento de insectos	Cuantitativa	Discreta
Kilómetros recorridos por una llanta hasta el primer pinchazo	Cuantitativa	Continua
Posibles rendimientos de maíz en un campo determinado	Cuantitativa	Continua
Posibles resultados al lanzar 50 monedas	Cuantitativa	Discreta

Ejercicio 2.4.2

Son muestras aleatorias las que aparecen en el cuadro 6.

Cuadro 6. Muestras aleatorias

Muestras aleatorias	
Las truchas pescadas en un día en un lago de tamaño moderado	Puede ser considerada una muestra aleatoria, bajo el supuesto de que no existen concentraciones en partes específicas del lago y que las truchas están en constante movimiento
Las ardillas capturadas en un día en una trampa	No se puede considerar aleatorio en el sentido de que las ardillas que se encuentran más cerca de la trampa tienen mayor probabilidad de caer en ella
Las respuestas escritas a un pronunciado político solicitadas en un anuncio de televisión	No puede ser considerado aleatorio, se van a recibir respuestas solamente de personas con acceso al anuncio por TV, aquellos que no tienen oportunidad de ver el anuncio no podrán participar
Una muestra autorizada de un botánico de la vegetación de un campo	No puede ser considerado aleatorio, puesto que el botánico experto seleccionará la vegetación bajo su criterio, el cual está afectado por la especialización, esto puede producir el sesgo de exclusión por omisión del especialista. La muestra final puede ser representativa pero puede excluir unidades de la muestra por juicio

Ejercicio 2.5.4

Supóngase que una población tiene 40 elementos y que se desea extraer una muestra de 5 observaciones sin reemplazo. ¿Cómo se haría?

Una opción es numerar cada elemento, iniciando en 01 y terminando en 40, poner la lista de números en una bolsa, luego pedirle a una persona que seleccione sin ver un número dentro de la bolsa, anotar el número y depositarlo de nuevo en la bolsa, de manera que cada elemento puede ser seleccionado cada vez que un nuevo número se extrae. Este procedimiento permite asignar la misma probabilidad de selección a cada unidad.

Ejercicio 2.5.7

1. ¿Cómo se extraería, por ejemplo, una muestra completamente aleatoria de 100 números telefónicos de un directorio telefónico? ¿Podría usarse un plan de dos etapas que implicara menor esfuerzo?

Una opción es definir cuántos números por página se van a obtener, por ejemplo 5 números telefónicos, esto significa que se requieren 20 páginas. Si se asume que el directorio telefónico tiene 600 páginas con 100 números cada una, entonces se podrá seleccionar una de cada 30 páginas de forma sistemática y dentro de cada una uno de cada 20 números, iniciando en los primeros 20.

2. Utilice la lista de árboles (archivo Excel "Datos sobre Muestreo de Árboles") y la tabla A.1 Diez mil dígitos aleatorizados del libro para seleccionar una muestra aleatoria simple de tamaño 10, seleccione como punto de arranque la columna 30-31-32 en la fila 00, siguiendo de forma vertical, luego salte a las columnas 35-36-37 para completar la selección.

La muestra estará conformada por las observaciones en las siguientes posiciones: 095, 084, 025, 093, 012, 060, 030, 036, 053, 038.

3. Seleccione una muestra de 10 observaciones de la lista de árboles (archivo Excel "Datos sobre Muestreo de Árboles") utilizando el método de selección en forma sistemática. Inicie en la posición 5.

Dado que se dispone de una lista de 100 árboles y se requiere de una muestra de 10, se obtiene el intervalo de selección dividiendo 100/10=10, lo que indica que debe seleccionarse uno de cada 10 árboles. Si se inicia en 5, entonces la muestra estará conformada por los árboles en las posiciones: 5, 15,25, 35, 45, 55, 65,75, 85, 95.

Esta selección es equivalente a numerar los árboles de 1 a 10 en forma repetida, como se muestra en la columna Muestreo sistemático dentro del archivo en Excel *y seleccionar los árboles que corresponden a la numeración 5.*

TEMA 2. ESTADÍSTICA DESCRIPTIVA

1. OBJETIVO

Aplicar conocimientos sobre los elementos básicos acerca de la presentación de datos estadísticos para su uso e interpretación.

Cuadro 7. Estadística descriptiva

Tema	Capítulo del libro	Páginas
Procesamiento y presentación de datos	Capítulo 2	12 y 13, 14 y 15
Codificación y tabulación	Capítulo 2	32-33 y 35 Sec 2.19
Presentación de datos: textual, semitextual, cuadros estadísticos, gráficos (lineal, barras, columnas, circular, dispersión, pictogramas, series de tiempo, etc.)	Presentación de Power Point: Cuadros y gráficos estadísticos	
Elaboración de cuadros estadísticos simples y de doble entrada mediante distribuciones de frecuencias. Análisis gráfico de series estadísticas simples y de doble entrada	Hoja de Excel: "Datos sobre Muestreo de Arboles"	

2. EJERCICIOS

Con el objetivo de reforzar los conocimientos adquiridos en este tema se recomienda que desarrolle los siguientes ejercicios:

1. Ejercicios 2.6.1 (ítemes 1,7 y 8).
2. Utilice los datos del archivo Excel "Datos sobre Muestreo de Árboles" y elabore un cuadro para resumir los datos Diámetro a la Altura de Pecho 2009, asegúrese de contemplar todos los aspectos formales de un cuadro. Utilice rangos de tamaño 0,5 cm., iniciando en 6 cm.
3. Elabore un polígono de frecuencias para los datos presentados en el punto anterior.
4. Describa los datos más relevantes del polígono de frecuencias.

5. Para las características que se muestran en el cuadro 6, señale el tipo y el de gráfico más adecuado.

Cuadro 8. Características según tipo y gráfico recomendado

Característica	Tipo	Gráfico
Diámetro a la altura de pecho (cm) 2007		
Altura (m)		
Posición (cuadrante de ubicación del árbol)		

3. OBSERVACIONES FINALES

Al elaborar cuadros estadísticos se deben considerar las preguntas básicas en la presentación de datos, esto permitirá que el lector ubique fácilmente lo que el autor pretende al exponer los resultados. Es recomendable que considere la unidad de estudio para facilitar la elaboración del título y sus componentes. Seguir las recomendaciones contribuye en la estandarización de la exhibición de información y da formalidad al estudio. El uso de gráficos ayuda a organizar los datos, facilita la detección de patrones, permite observar agrupamientos, relaciones y la forma de la distribución de los mismos, esto se logra fácilmente siguiendo los lineamientos dados para su construcción. Se inicia con la identificación del tipo de característica y la escala de medición utilizada, aspectos que definen el gráfico ideal que se ajusta al requerimiento. Los aspectos estéticos deben tratarse con discreción, pues el exceso puede distorsionar el objetivo básico de comunicación de un resultado.

La omisión de alguno de los componentes del cuadro estadístico o de respuesta a las preguntas básicas afectan la credibilidad y formalidad de la investigación, al mismo tiempo, esto puede llevar a interpretaciones incorrectas de los datos y a fallas en la toma de decisiones. Algunas veces los investigadores omiten explorar la forma de distribución de las características de estudio, porque les parece irrelevante; sin embargo, de la representación gráfica, se derivan conclusiones básicas que pueden variar desde la forma misma hasta conceptos de simetría y

asimetría que ayudarán al análisis descriptivo de resumen mediante la correcta utilización de alguna de las medidas de posición y tendencia central.

El error más común en el uso de estas técnicas de presentación de resultados es la elaboración del título, pocas investigaciones presentan la rigurosidad de contemplar la respuesta a las preguntas básicas y, en consecuencia, la interpretación es parcial, incluso, algunas veces la información que se puede derivar no es utilizada. En la representación gráfica, las fallas se refieren al abuso que existe de la gama de posibilidades que ofrecen los programas computacionales en la actualidad, es así como se encuentran gráficos que representan variables continuas, cuando su utilización es recomendada en características cualitativas, de manera que no todas las opciones que ofrece el computador son aplicables técnicamente.

Otro de los errores importantes es la presentación de datos sin ser ordenados previamente, de manera que la idea de magnitud no llega a partir de la primera mirada al gráfico, sino después de que el lector hace el ejercicio mental para obtener la información relevante.

4. RESUMEN DEL TEMA

Al finalizar el tema se espera que usted sea capaz de:

- Identificar las formas de presentación de datos estadísticos.
- Construir cuadros estadísticos en forma adecuada.
- Interpretar adecuadamente cuadros y gráficos estadísticos.
- Valorar la importancia de conocer las técnicas estadísticas para la presentación de datos.
- Mantener una actitud crítica ante las formas de presentación de datos estadísticos.

5. SOLUCIÓN A LOS EJERCICIOS

1. Desarrolle los ejercicios 2.6.1 (ítems 1,7 y 8), 2.6.5

Ejercicio 2.6.1

En la "Encuesta, Nacional de Caza, Pesca y Vida Silvestre en 1975" realizada por el Servicio de Pesca y Vida Silvestre de los Estados Unidos, se encuestaron más de 2 000 familias por teléfono y se enviaron cuestionarios por lo menos a 1 000 cazadores y pescadores en cada estado. Toda persona de 9 o más años que cazó o pescó, al menos en un día, en 1975 era aceptable para participar en la encuesta postal. De esas personas de 9 años o más, 95,9 millones cooperaron en alguna actividad relacionada con la vida silvestre. Entre los datos reportados estuvieron los siguientes (a menudo aproximados a partir de una gráfica):

Ítem 1: 53 millones participaron en pesca; 50 millones en observación de la vida silvestre; 26 millones en recolección de calamares, cangrejos y otros; 21 millones en caza, 17 millones en tiro recreacional; 15 millones en fotografía de vida salvaje, y 5 millones en disparo con arco.

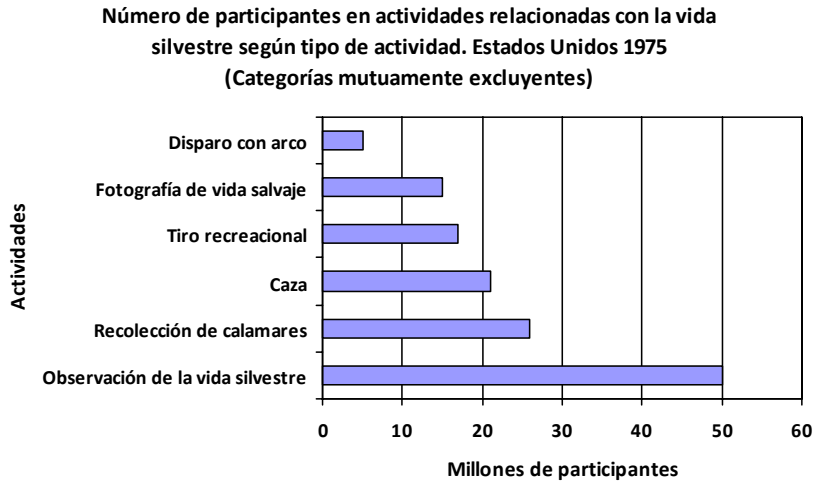
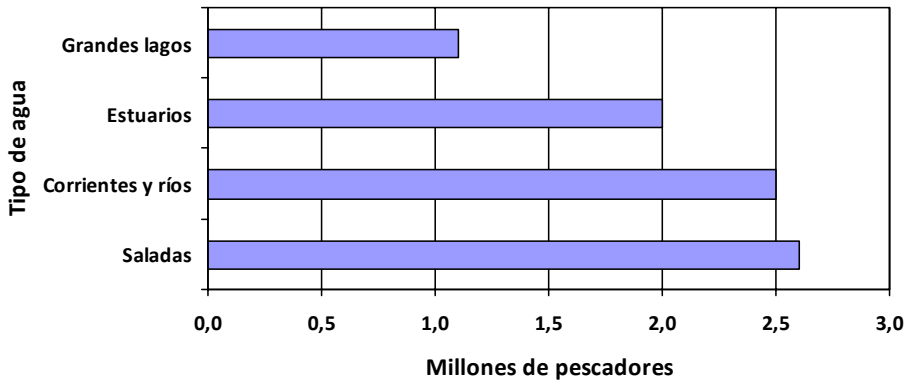


Figura 3

Ítem 2: El número, en millones, de pescadores en ríos por tipo de agua fue: 2,6 en aguas saladas, 2,5 en corrientes y ríos, 2,0 en estuarios y 1,1 en los grandes lagos.

Número de pescadores según tipo de agua. Estados Unidos 1975
(Categorías mutuamente excluyentes)

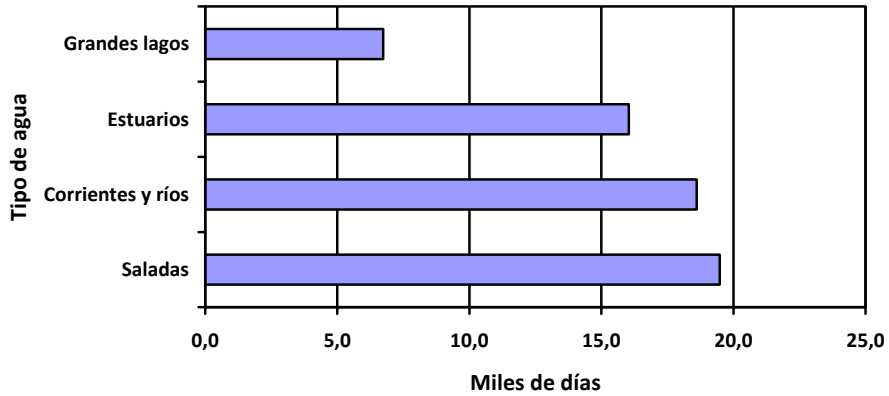


FUENTE: Servicio de Pesca y Vida Silvestre de los Estados Unidos

Figura 4

Ítem 3. El número en miles de días de participación en pesca en ríos por tipo de agua especificada en el numeral 7 fue, respectivamente: 19 478, 18 606, 16 039 y 6 739.

Miles de días de participación en pesca en ríos según tipo de agua. Estados Unidos 1975
(Categorías mutuamente excluyentes)



FUENTE: Servicio de Pesca y Vida Silvestre de los Estados Unidos

Figura 5

2. Utilice los datos del archivo Excel "Datos sobre Muestreo de Árboles" y elabore un cuadro para resumir los datos diámetro a la altura de pecho 2009, asegúrese de contemplar todos los aspectos formales de un cuadro. Utilice rangos de tamaño 0,5 cm, inicie en 6 cm.

Cuadro 9. Distribución del número de árboles de acuerdo con el diámetro a la altura de pecho 2009

Límites de clase			Frecuencia
6,0	a menos de	6,5	5
6,5	a menos de	7,0	20
7,0	a menos de	7,5	25
7,5	a menos de	8,0	26
8,0	a menos de	8,5	20
8,5	a menos de	9,0	4
Total			100

3. Elabore un polígono de frecuencias para los datos presentados en el punto anterior.

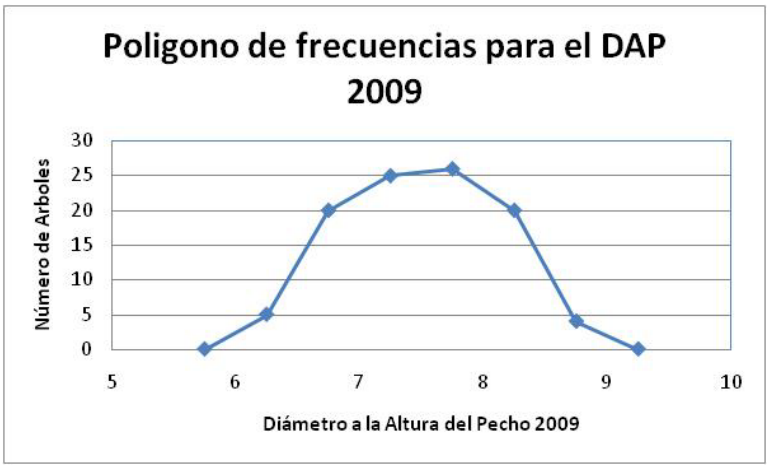


Figura 6

4. Describa los datos más relevantes del polígono de frecuencias.

Se observa que el DAP es una variable continua con un alto grado de simetría, es unimodal con una alta concentración de casos en las clases del medio. La moda se presenta entre 7 y 8, lo mismo que la mediana.

5. Para las siguientes características señale el tipo y el de gráfico más adecuado.

Cuadro 10. Características según tipo y gráfico recomendado

Característica	Tipo	Gráfico
Diámetro a la altura de pecho (cm) 2007	Cuantitativa	Barras verticales Lineal
Altura (m)	Cuantitativa	Barras verticales Lineal
Posición (cuadrante de ubicación del árbol)	Cualitativa	Barras horizontales Circular

TEMA 3. MEDIDAS DE POSICIÓN Y DE VARIABILIDAD

1. OBJETIVO

Aplicar conocimientos sobre uso de medidas de posición y tendencia central, así como la variabilidad para su uso en el análisis estadístico de datos.

Cuadro 11. Medidas de posición y de variabilidad

Tema	Capítulo del libro	Páginas
Concepto de medidas de posición y de variabilidad	Capítulo 2	15 y 19
Medidas de posición y tendencia central en datos simples	Capítulo 2	15 a 18
Medidas de variabilidad en datos simples	Capítulo 2	20 a 27, excluye sección 2.12
Medidas de posición y tendencia central en datos agrupados	Capítulo 2	32 a 35
Medidas de variabilidad en datos agrupados	Capítulo 2	
Cálculo de percentiles en datos simples y agrupados	Texto en la guía	
Interpretación de las medidas de tendencia central y de variabilidad	Presentación de Power Point: Medidas de tendencia central y variabilidad Distribuciones de frecuencias	

2. CÁLCULO DE PERCENTILES EN DATOS SIMPLES Y AGRUPADOS

Los percentiles o cuantiles son valores que dividen al conjunto de datos en 100 partes proporcionalmente complementarias. Su aplicación tiene sentido cuando la característica bajo estudio es ordinal, de intervalo o razón. Su uso requiere que el conjunto de datos esté ordenado cuando el procesamiento es manual; los sistemas automatizados, estadísticos o matemáticas, no requieren de esta condición.

Casos particulares son los cuartiles que dividen el conjunto en cuatro partes iguales:

- Primer cuartil (Q_1)= Percentil 25 (P_{25}) = cuantil 0,25: representa el valor por debajo del cual se encuentra el 25% de las observaciones y complementariamente por encima de este se encuentra el 75% restante.

- Segundo cuartil (Q_2)= Percentil 50 (P_{50}) = cuantil 0,5 = mediana: representa el valor por debajo del cual se encuentra el 50% de los datos y complementariamente sobre él está el 50% restante.

- Tercer cuartil (Q_3)= Percentil 75 (P_{75}) = cuantil 0,75: representa el valor por debajo del cual se encuentra el 75% de los datos y complementariamente sobre él está el 25% restante.

Datos no agrupados

En el caso de datos sin agrupar (datos simples), el percentil se expresa matemáticamente como:

$$P_m = \frac{m}{100} * (n+1)$$

Donde ***m*** representa el percentil de interés, puede tomar valores entre 1% y 99%, ejemplo 20%, 45%, entre otros.

P_m: indica la posición en la que se encuentra el percentil dentro del conjunto.

Si la posición del percentil (P_m) no es un valor entero, se toma el inmediato superior, esto funcionaría como una aproximación al real en la práctica; desde el punto de vista teórico, se sugiere realizar una interpolación para obtener una mejor estimación.

Ejemplo 1. Sobre cálculo de percentiles

Para el siguiente conjunto de datos, de tamaño 15, calcule el percentil 25 y el 60:

7, 19 ,8 ,1, 7, 17 ,4, 7,1, 7, 3, 7, 3, 13, 3

- Primero, ordene el conjunto de datos: 1, 1, 3, 3, 3, 4, 7, 7, 7, 7, 7, 8, 13, 17, 19

- Segundo, calcule la posición del percentil 25 siguiendo la fórmula anterior: P_{25}=25/100*(15+1)=4. Esto es el percentil 25 ocupa la posición 4 que equivale a 3.

- Tercero, interprete este valor: El 25% de las observaciones son menores o iguales a 3 y el 75% restante es mayor que 3.

Para el percentil 60, repita el segundo y tercer paso:

- Calcule la posición del percentil 60 siguiendo la fórmula anterior: P_{25}=60/100*(15+1)=9,6. El percentil 60 ocupa la posición entre 9 y 10, en términos prácticos, se utiliza la posición 10, es decir, el P_{60} se ubica en la posición 10 que equivale a 7.

- Interprete este valor: El 60% de las observaciones son menores o iguales a 7 y el 40% restante es mayor que 7.

Datos agrupados

Para el cálculo de percentiles en datos agrupados, es importante introducir el concepto de clase percentil, la cual se define como aquella donde se acumula el tanto por ciento de interés o más de los casos. Para el cálculo manual es importante obtener previamente las frecuencias acumuladas.

Matemáticamente se utiliza la siguiente fórmula: $P_m = L_i + c * \left(\frac{\frac{n*m}{100} - F_{i-1}}{f_i} \right)$

Donde:

- ***m*** representa el percentil de interés, puede tomar valores entre 1% y 99%, ejemplo 20%, 45%, entre otros.

- ***Pm*** indica la posición en la que se encuentra el percentil dentro del conjunto.

- L_i representa el límite inferior de la clase percentil.

- ***C*** representa el intervalo real de la clase percentil

- F_{i-1} representa la frecuencia acumulada de la clase percentil

- F_i representa la frecuencia simple de la clase percentil.

Ejemplo 2. Sobre cálculo de percentiles

Utilice la distribución de frecuencias del cuadro 10 y calcule el percentil 60.

Cuadro 12. Distribución de frecuencias para ilustrar el cálculo de percentiles

Límites de clase	Frecuencia simple (f)	Frecuencia absoluta acumulada (F)
De 40 a menos de 70	6	6
De 70 a menos de 100	8	14
De 100 a menos de 130	12	26
De 130 a menos de 160	18	44
De 160 a menos de 190	15	59
De 190 a menos de 210	9	68
De 210 a menos de 240	2	70
Total	70	

Pasos:

1. Ubicar la clase que alcanza el 60% o más de los casos: 70*60/100=42. El percentil 60 se ubica donde se acumulan 42 o más valores, es la clase de "De 130 a menos de 160".

2. Identificar los distintos valores requeridos:

- $L_i = 130$.
- $f_i = 18$
- $C = 30$
- $n = 70$
- $F_{i-1} = 26$
- $m = 60$

3. Realizar el cálculo de acuerdo con la fórmula anterior, esto equivale a 156,67.

4. Interpretar este valor: El 60% de los casos u observaciones se encuentran por debajo de 156,67 y el 40% restante está sobre este valor.

3. EJERCICIOS

Con el objetivo de reforzar los conocimientos adquiridos en este tema, se recomienda que desarrolle los siguientes ejercicios:

1. Ejercicios 2.7.3 y 2.7.8 del libro de texto.

2. Se dispone de información referente a los niveles de precipitación de algunos días del año organizados en la distribución de frecuencias del cuadro 11, donde f, F y f_r representan respectivamente, la frecuencia absoluta, acumulada y relativa:

a. Complete el cuadro.

a. Calcule la media, la moda y mediana.

a. Calcule la desviación estándar.

a. Interprete los datos obtenidos en los puntos b y c en términos del problema.

Cuadro 13. Distribución de frecuencias sobre los niveles de precipitación

x	F	F	fr
236	4		0,08
301	4		
405		16	0,16
442	7		0,14
521	5	28	
562		38	
620	7	45	
631			
Total			

3. ¿Cuál gráfico sería el más adecuado para representar estos datos?

4. Calcule el coeficiente de variación y determine si existe simetría en esta distribución.

5. Dos muestras aleatorias presentan la información en el cuadro 12:

Cuadro 14. Promedio y desviación estándar en dos muestras aleatorias

Muestra	Promedio	Desviación estándar
1	45	7
2	75	8

a. ¿Cuál de las muestras es más dispersa?

4. OBSERVACIONES FINALES

La adecuada aplicación de las medidas de posición o tendencia central y de variabilidad está determinada por tres factores básicos: el tipo de características, su nivel de medición y la forma de la distribución. Las dos primeras indican cuáles medidas de posición y de variabilidad son posibles de calcular, al mismo tiempo, la forma de la condición de simetría permite decidir sobre el estimador más apropiado.

En ocasiones, se presentan resultados de análisis estadísticos inadecuados debido a un mal manejo de la codificación de las variables, así, por ejemplo, si no se tiene en cuenta que se trata de una característica de tipo cualitativo y se codificó como numérica, al momento del procesamiento de datos, se calcula la media y su desviación estándar (por su condición numérica), aun cuando esto no tiene sentido. Lo correcto en estos casos es utilizar la moda únicamente para describir los resultados.

El error más común en la aplicación de medidas de posición o tendencia central se presenta en no saber escoger cuál medida describe mejor el conjunto de datos, esto ocurre cuando no se considera la condición de simetría del mismo. Cada estimador tiene sus ventajas y desventajas. La media, por ejemplo, se ve afectada por valores extremos, razón por la cual no siempre representa de la mejor manera al conjunto de datos en presencia de asimetría.

Otro de los errores es utilizar únicamente una medida de posición sin considerar la dispersión de la variable, con excepción de los atributos, esta es una práctica

incorrecta, pues con la variabilidad es posible tener una idea de cuán preciso es el estimador.

Es frecuente pensar que un valor de tendencia central describe bien un conjunto de datos, sin embargo, en la práctica, es mucho más valioso asociar un rango de variación, por ejemplo, es mejor saber que la temperatura varía entre 20° y 25° que pensar que la temperatura promedio es 22°, el rango describe de mejor forma en contraposición a un valor puntual.

5. RESUMEN DEL TEMA

Al finalizar el tema se espera que usted sea capaz de:

- Calcular e interpretar las medidas de tendencia central en datos simples y agrupados.
- Calcular e interpretar las medidas de variabilidad en datos simples y agrupados.
- Identificar los errores frecuentes en el uso de medidas de posición.
- Valorar las medidas de posición y de variabilidad como herramientas para describir una realidad.

6. SOLUCIÓN A LOS EJERCICIOS

1. Desarrolle los ejercicios 2.7.3 y 2.7.8

Ejercicio 2.7.3
¿La media aritmética puede considerarse como una media ponderada con pesos iguales? ¿Cuáles son estos pesos?

La media aritmética puede ser considerada como media ponderada con pesos iguales a 1/n.

Ejercicio 2.7.8
Comentar la afirmación: "50% de los estadounidenses tienen una inteligencia por debajo del promedio".

Esta afirmación supone que la distribución del coeficiente de inteligencia es simétrica, puesto que la media y la mediana coinciden.

2. Se dispone de información referente a los niveles de precipitación de algunos días del año organizados en la siguiente distribución de frecuencias, donde f, F y f_r representan, respectivamente, la frecuencia absoluta, acumulada y relativa:

    a.. Complete el cuadro.

Cuadro 15. Distribución de frecuencias sobre los niveles de precipitación (cuadro completo)

x	f	F	fr
236	4	4	0,08
301	4	8	0,08
405	8	16	0,16
442	7	23	0,14
521	5	28	0,10
562	10	38	0,20
620	7	45	0,14
631	5	50	0,10
Total		50	1,00

b. Calcule la media, la moda y mediana.

A efecto de calcular las medidas de tendencia central se ha confeccionado el cuadro 14:

Cuadro 16. Distribución de frecuencias sobre los niveles de precipitación, cálculo de medidas de posición

Valores	Frecuencia	Frecuencia Acumulada	(x*f)	(x^2*f)
236	4	4	944	222 784
301	4	8	1 204	362 404
405	8	16	3 240	1 312200
442	7	23	3 094	1 367 548
521	5	28	2 605	1 357 205
562	10	38	5 620	3 158 440
620	7	45	4 340	2 690 800
631	5	50	3 155	1 990 805
	50		24 202	12 462 186

Con base en los cálculos anteriores se tiene lo siguiente:

Moda *562,00*

Mediana *521,00*

Promedio *484,04*

c. Calcule la desviación estándar.

Con base en los cálculos anteriores se tiene lo siguiente:

Desviación *123,51*

d. Interprete los datos obtenidos en los puntos b y c en términos del problema.

El nivel de precipitación que se ha observado con más frecuencia es de 562.

El 50% de los días observados se presentó un nivel de precipitación menor o igual a 521.

De acuerdo con lo observado puede esperarse que se alcance un nivel de precipitación de 484,04 por día.

En promedio, el nivel de precipitación se aleja de su media en 123,51.

3. Cuál gráfico sería el más adecuado para representar estos datos?

El gráfico lineal es el más adecuado para representar estos datos.

4. Calcule el coeficiente de variación y determine si existe simetría en esta distribución.

El coeficiente de variación (desviación /media) es de 26%, superior a 20%, con lo cual esta muestra puede ser considerada asimétrica.

5. Dos muestras aleatorias presentan la información del cuadro 15:

Cuadro 17. Promedio y desviación estándar en dos muestras aleatorias, cálculo del coeficiente de variación

Muestra	Promedio	Desviación Estándar	Coeficiente de variación
1	45	7	15,6%
2	75	8	10,7%

¿Cuál de las muestras es más dispersa?

La primera muestra es más dispersa pues presenta un coeficiente de variación más alto, en promedio las observaciones se alejan en un 15,6% de la media; la segunda muestra presenta un desvío relativo menor de 10,7%.

TEMA 4: VARIABLES BIDIMENSIONALES

1. OBJETIVO

Adquirir conocimientos generales sobre la técnica de asociación entre dos características cuantitativas o dos cualitativas para la generación de conocimiento predictivo en el campo afín.

Cuadro 18. Variables bidimensionales

Tema	Capítulo del libro	Páginas
Dependencia funcional e independencia	Texto en la Guía	
Concepto de covarianza	Texto en la Guía	
Asociación o relación	Capítulo 10	Páginas 231 a 248, se excluye la sección 10.7
Regresión lineal simple	Capítulo 10	
Coeficiente de correlación de Pearson	Capítulo 11	Páginas 263 a 269, excluye 11.4 Presentación de Power Point: Regresión y correlación en dos variables

2. DEPENDENCIA FUNCIONAL E INDEPENDENCIA

El estudio de las variables bidimensionales se enfoca en describir la relación estadística o matemática entre ellas; se busca definir cuál es el comportamiento de una conforme la otra varía y viceversa.

En algunas ocasiones, las variaciones en una variable se describen mediante una función matemática, de manera que, para cualquier valor de la variable X, siempre existe un único valor en la variable Y. Esto es lo que se conoce como dependencia funcional; matemáticamente, se expresa como $y_j = f(x_i)$, donde los subíndices j e i representan un valor de la variable Y y X respectivamente. Si Y depende

funcionalmente de X entonces también habrá una dependencia funcional de X sobre Y.

En Estadística no es posible definir una relación funcional uno a uno, es decir, a un mismo valor de la variable X puede corresponder más de una observación en Y, en este caso, lo que interesa es establecer la correspondencia entre ambas variables de la manera más precisa posible; intuitivamente, puede verse esta como una expresión funcional más un error, positivo o negativo, que hace las veces de ajuste para describir de una mejor manera lo observado.

Por ejemplo, para un tiempo de crecimiento (X) de una planta pueden existir diferentes alturas (Y) y, aunque es posible comprender que el tamaño está afectado por el tiempo de crecimiento, no es factible describir una relación matemática única; sin embrago, se puede predecir la elevación de la planta en función del tiempo de crecimiento con algún error, lo cual puede llevar a pensar que no existe un valor único sino un rango de variación.

La relación de dependencia entre dos variables estadísticas se presenta cuando los cambios en una afectan o condicionan el comportamiento de la otra y viceversa.

La asociación descrita entre el tiempo de crecimiento y estatura de una planta es lo que se conoce como relación estadística o dependencia, esta técnica busca definir si entre dos variables existe una relación y la fuerza con que esta se presenta o, por el contrario, determinar si el comportamiento observado es más aleatorio.

La condición de independientes se presenta entre dos variables cuando el cambio en una de ellas no afecta o condiciona los cambios en la otra variable, por ejemplo, experiencia laboral de un trabajador y la distancia recorrida entre su casa y su lugar de trabajo.

Como ejemplos de relaciones estadísticas pueden citarse: consumo de gasolina respecto a los kilómetros recorridos por un vehículo; nivel de gasto en función de ingreso familiar, estatura *versus* peso del individuo; consumo de litros cúbicos de agua potable en una familia de acuerdo con el número de miembros.

3. CONCEPTO DE COVARIANZA

La covarianza es una medida de la variación conjunta de dos variables, su valor describe la relación, positiva o negativa, que existe entre ellas y el grado o fuerza de la asociación.

- σ_{xy} es positiva si los valores altos de *X* están asociados a los valores altos de *Y* y viceversa.
- σ_{xy} es negativa si los valores altos de *X* están asociados a los valores bajos de *Y* y viceversa.
- Si *X* e *Y* son variables aleatorias independientes $cov(x,y) = 0$.
- La independencia es condición suficiente, pero no necesaria para que la $cov(x,y)$ sea nula.

4. ASOCIACIÓN O RELACIÓN

El coeficiente de correlación mide la fuerza de la asociación entre pares de variables y la dirección de la relación, de manera que puede predecirse o esperarse los cambios en una de ellas cuando la otra presenta variaciones.

La magnitud y fuerza de la asociación entre variables, descrita por medio del coeficiente de correlación, varía entre -1 y 1, como se muestra en el cuadro 17.

Cuadro 19. Rango de variación del coeficiente de correlación

-1	**0**	**1**
Asociación **negativa** y alta	No existe asociación	Asociación **positiva** y alta

En este conjunto de técnicas el más frecuentemente utilizado es el de Pearson, sin embargo, se debe tener cuidado porque este se aplica solamente cuando ambas variables son cuantitativas, si, al menos, una variable presenta un nivel de medición diferente al cuantitativo, este ya no es aplicable.

5. EJERCICIOS

Con el objetivo de reforzar los conocimientos adquiridos en este tema se recomienda que desarrolle los siguientes ejercicios:

Suponga que tiene una parcela de árboles de una especie, dividida en cuadrantes, la cual ha sido evaluada en dos años diferentes (2007 y 2009). Utilice el archivo Excel "Datos sobre Muestreo de Árboles" para lo siguiente:

1. Determine si existe relación entre el diámetro a la altura del pecho y la altura del árbol al inicio del periodo de estudio.
2. Haga un diagrama de dispersión y descríbalo.
3. Ajuste un modelo de regresión lineal simple para predecir la altura del árbol de acuerdo con el diámetro a la altura del pecho. Valore la bondad del ajuste.
4. Calcule el coeficiente de correlación de Pearson usando el coeficiente de determinación.
5. Interprete los coeficientes del modelo, la correlación y la bondad del ajuste.
6. ¿Cuál es la altura esperada de un árbol si el diámetro a la altura del pecho es 4,3 cm?
7. ¿Cuál es la altura esperada de un árbol si el diámetro a la altura del pecho es 6 cm?
8. ¿Cuál de las estimaciones anteriores es más confiable?

6. OBSERVACIONES FINALES

El estudio de variables bidimensionales permite conocer de una forma más cercana a la realidad el comportamiento de características cualitativas y cuantitativas de forma conjunta. Sin embargo, debe tenerse especial cuidado en considerar los niveles de medición de estas, debido a que esto condiciona el uso de técnicas específicas.

La exploración de los datos permite evaluar la práctica estadística más apropiada para estudiar las relaciones entre atributos y variables de una manera más intuitiva, de tal forma que pueda confirmarse con el análisis estadístico aquello que se percibe como cierto.

Este tipo de razonamiento puede ser altamente complejo o relativamente simple, en este capítulo, se establece el nivel más básico dentro de este conjunto de técnicas que se refiere a la relación lineal entre dos variables cuantitativas, existen otras para construir modelos más elaborados que permiten establecer relaciones de mayor orden matemático, pues consideran diferentes niveles de medición en las variables o ajustes lineales.

El coeficiente de correlación de Pearson se utiliza estrictamente para cuantificar la relación existente entre dos variables, cuando se trate de atributos debe buscarse el coeficiente más adecuado.

De igual manera, cuando la relación entre dos variables no es lineal, debe buscarse el modelo que describa mejor la dependencia existente entre ellas.

El error más común cuando se realiza el análisis de variables bidimensionales es el uso de las técnicas inapropiadas, como basarse en el coeficiente de correlación de Pearson para explicar la relación entre una variable cuantitativa y un atributo, por ejemplo, describir la asociación entre el nivel de ingreso y el sexo de las personas.

Otro de los errores comunes es ajustar modelos de regresión lineal cuando se trata de describir la dependencia funcional entre dos variables y no evaluar el ajuste o los efectos en los supuestos. Cuando esta técnica, se utiliza se debe tener la precaución de valorar el coeficiente de determinación para tener una idea de que tan bueno es el ajuste entre ellas.

Es recomendable corroborar la consistencia entre la correlación, el coeficiente de regresión y la bondad de ajuste para garantizar la adecuada aplicación del modelo, es fácil cometer errores en la interpretación y en las conclusiones cuando no se vigila esta consistencia.

7. RESUMEN DEL TEMA

Al finalizar el tema se espera que usted sea capaz de:

- Definir el concepto de asociación estadística de dos variables cuantitativas y dos cualitativas.
- Explicar el concepto de regresión lineal simple.
- Ajustar e interpretar un modelo de regresión mediante el apoyo de un *software* estadístico.
- Calcular e interpretar adecuadamente el coeficiente de correlación de Pearson.
- Examinar un conjunto de datos y seleccionar la técnica estadística adecuada para realizar un análisis de asociación y correlación.

8. SOLUCIÓN A LOS EJERCICIOS

Suponga que tenemos una parcela de árboles, de una especie, dividida en cuadrantes, que ha sido evaluado en dos años diferentes (2007 y 2009). Utilice el archivo hoja de Excel: Datos sobre Muestreo de Árboles para lo siguiente:

1. Determine si existe relación entre el diámetro a la altura del pecho y la altura del árbol al inicio del periodo de estudio. Haga un diagrama de dispersión y descríbalo.

El diagrama de dispersión muestra que existe una relación lineal positiva entre las variables DAP y altura del árbol, es posible observar el efecto en la altura cuando el DAP aumenta.

2. A partir de este gráfico se sabe que la pendiente de la línea de ajuste debe ser mayor que cero y, consecuentemente, el coeficiente de correlación también.

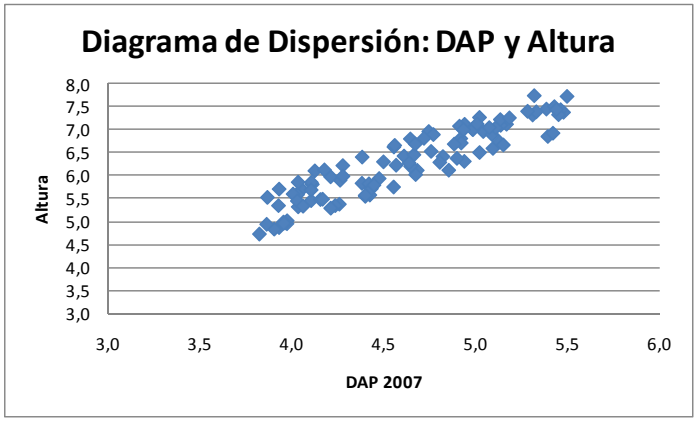


Figura 7

3. Ajuste un modelo de regresión lineal simple para predecir la altura del árbol de acuerdo con el diámetro a la altura del pecho. Valore la bondad del ajuste.

   El ajuste del modelo es el siguiente: y = 1,4729x - 0,5397

   R^2 = 0,8546

4. Calcule el coeficiente de correlación de Pearson usando el coeficiente de determinación.

   El coeficiente de correlación es equivalente a la raíz cuadrada del coeficiente de determinación:

   $$r = \sqrt{R^2} = \sqrt{0{,}8546} = 0{,}924$$

5. Interprete los coeficientes del modelo, la correlación y la bondad del ajuste.

   *Los coeficientes del modelo de regresión son dos: la intersección, que en este caso equivale a -0,5397 y la pendiente que es 1,4729, este es un modelo donde el intercepto no tiene sentido, pues se considera la altura esperada cuando el diámetro a la altura del pecho es cero, es de decir se espera una altura negativa cuando el DAP es cero, esto no tiene sentido y por tanto se considera como un componente necesario para ajustar las estimaciones. El coeficiente de regresión b = 1,4729 se interpreta como el cambio esperado en la altura del árbol cuando el DAP cambia en 1 cm, es decir que si el DAP cambia en 2 cm se espera que la altura del árbol se incremente en 1,4729*2 =2,9458.*

El coeficiente de correlación r = 0,924, es positiva y fuerte, cercana a 1, se interpreta como: existe una fuerte asociación lineal entre las variables DAP y altura del árbol, el grado de la asociación es fuerte, se espera que incrementos en una variable impacte de manera positiva la otra.

El valor del R^2 o bondad de ajuste es alto, cercano a 100%, con lo cual es posible deducir que el modelo que describe la relación lineal entre DAP y altura del árbol es bueno estadísticamente.

6. ¿Cuál es la altura esperada de un árbol si el diámetro a la altura del pecho es 4,3 cm?

   y = 1,4729(4,3) - 0,5397 = 5,79377*

7. ¿Cuál es la altura esperada de un árbol si el diámetro a la altura del pecho es 6 cm.?

   y = 1,4729(6) - 0,5397= 8,2977*

8. ¿Cuál de las estimaciones anteriores es más confiable?

   La estimación más confiable es la primera, esta se conoce como interpolación, debido a que el valor de x se encuentra dentro del rango de variación del modelo (3,5-4,5). La banda de confianza del modelo se ensancha conforme el valor de x se aleja a los extremos del rango.

TEMA 5: INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

1. OBJETIVO

Adquirir los conocimientos generales del proceso de inferencia estadística como herramienta de análisis para la generación de conocimiento en el campo.

Cuadro 20. Introducción a la inferencia estadística

Tema	Capítulo del libro	Páginas
Concepto de inferencia estadística	Texto en la Guía	
Parámetros y estimadores	Texto en la Guía	
Elementos de probabilidad	Capítulo 3 páginas 37 a 40 se excluye la sección 3.3	Presentación en Power Point: Introducción a la inferencia estadística
Distribución de probabilidades (gaussiana, binomial, poisson distribución F y chi cuadrado).	Capítulo 3 Capítulo 23, páginas 510 a 519 Incluye las secciones 3.3, 3.5, 3.6, 3.7, 3.9	Texto sobre distribución *F* en la Guía
Estimación puntual y por intervalos		Presentación en Power Point: Introducción a la inferencia estadística
Cálculo de intervalos de confianza		Presentación en Power Point: Introducción a la inferencia estadística

2. CONCEPTO DE INFERENCIA ESTADÍSTICA

La Estadística inferencial o inductiva permite realizar conclusiones o generalizaciones, con base en los datos resumidos y analizados de una muestra hacia la población o universo de la cual provienen.

El objetivo fundamental de esta rama de la estadística es generalizar las observaciones de una parte hacia un todo, en este proceso, toman relevancia los conceptos de representatividad y aleatoriedad de la muestra, debido a que, si estas características no se cumplen, el resultado de la inferencia será débil y, en la mayor parte de los casos, erróneo.

El proceso de inferencia debe basarse en métodos estadísticos apropiados que permitan generalizar de forma sólida lo que se observa en la muestra, entre estos, se pueden citar: el muestreo aleatorio, las técnicas de estimación, las pruebas de hipótesis, el diseño experimental, la teoría bayesiana y algunos métodos no paramétricos.

3. PARÁMETROS Y ESTIMADORES

Se conoce como estimadores estadísticos a la medida que se obtiene respecto a la distribución de probabilidades poblacional, ejemplos de estos son: la media, la varianza, la proporción, entre otros, que consideran todas las unidades que la conforman la población, de manera que, desde el punto de vista práctico, esta información es imposible de conocer; a partir de ello, la teoría de la estimación de parámetros adquiere importancia, pues, en el mundo real, normalmente, esta información se obtiene a partir del estudio por muestreo, donde estos métodos se aplican con base en el principio de inferencia.

Las características deseables de un estimador estadístico para ser considerado de valor son las siguientes:

- Consistencia: se refiere al efecto sobre la estimación al incrementarse el tamaño de la muestra, pues se espera que el valor estimado se aproxime al parámetro desconocido, por cuanto, entre más observaciones, más

conocimiento se tiene sobre la variabilidad de la población y, en consecuencia, del verdadero valor.

- Insesgado: se refiere a la expectativa que se tiene del estimador en muestras repetidas, donde se espera que no presente diferencias entre una medición y otra o, en caso contrario, que estas sean despreciables.
- Eficiencia: esta característica se refiere a la medición de las diferencias esperadas en el estimador, de manera que la variabilidad asociada sea mínima.
- Suficiencia: se refiere a la importancia de que el estimador utilice toda la información de la muestra y no que se base en unas pocas observaciones.

4. DISTRIBUCIÓN *F* DE FISHER-SNEDECOR

Esta distribución de probabilidad se usa para evaluar varias hipótesis estadísticas, como por ejemplo:

1. Igualdad de varianzas entre dos o más poblaciones o muestras.
2. Comparación de dos o más medias poblacionales, esta prueba es conocida como análisis de varianza (ANDEVA o en inglés ANOVA).

Matemáticamente, sean U y V dos variables aleatorias independientes con distribución Chi-cuadrado con m y n grados de libertad respectivamente.

Sea X una variable definida como el cociente de U y V divididas por los grados de libertad: $X = \frac{U/m}{V/n}$. La variable X definida de esta manera sigue una distribución *F Fisher-Snedecor* con m y n grados de libertad, que se representa como: $X \to F_{m,n}$.

Dentro de la familia de distribuciones *F*, se utilizan los grados de libertad en el numerador y en el denominador como parámetros para definirlas como únicas, de manera que, para cada par de m y n grados de libertad, se define una función.

Esta es una distribución asimétrica positiva, de manera que no existen valores negativos de *F*, acercándose de manera asintótica al eje x.

La forma de la representación gráfica depende de los valores *m* y *n*, de tal forma que si *m* y *n* tienden a infinito, la distribución tiende a parecerse a la distribución normal.

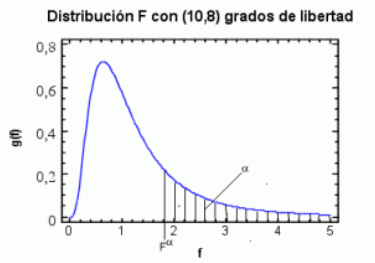


Figura 8

La media existe si "*n*" es mayor o igual que 3 y la varianza existe si "*n*" es mayor o igual que 5 y sus valores son:

$$\mu = \frac{n}{(n-2)} \quad y \quad \sigma^2 = \frac{2\,n^2\,(m+n-2)}{m\,(n-2)^2\,(n-4)}$$

En la tabla A.6, en el apéndice de tablas del libro de texto, pueden verse los valores de *F* en función de *m* y *n* sus respectivos grados de libertad.

5. EJERCICIOS

Con el objetivo de reforzar los conocimientos adquiridos en este tema se recomienda que desarrolle los siguientes ejercicios:

1. Resuelva los ejercicio 3.7.1 y 3.7.2
2. Una variable aleatoria sigue una distribución χ^2 de Pearson. Se pide calcular:

   (a) Los puntos críticos: $\chi^2_{0,90;5}$, $\chi^2_{0,01;26}$, $\chi^2_{0.025;8}$, $\chi^2_{0,08;10}$.

(b) Las probabilidades:

i. $P(\chi^2_8 \geq 3.49)=$

ii. $P(\chi^2_8 \leq 15.51)=$

iii. $P(\chi^2_{10} \geq 4)=$

iv. $P(\chi^2_{20} \leq 29)=$

3. Una variable aleatoria sigue una distribución *t* de Student. Calcule:

(a) Los puntos críticos: $t_{0,20;20}$, $t_{0,99;10}$, $t_{0,25;10}$

(b) Las probabilidades:

i. $P(t_{10} \geq 1.372)$

ii. $P(t_8 \leq 1.2)$

ii. $P(-0.5 \leq t_6 \leq 0.6)$;

4. Una variable aleatoria sigue una distribución F de Fisher-Snedecor. Calcule:

(a) Los puntos críticos: $F_{0.1;10,12}$, $F_{0.05;5,24}$, $F_{0.90;28,30}$

(b) Las probabilidades: $P(2 \leq F_{10;20} \leq 2.25)$

6. OBSERVACIONES FINALES

La necesidad de contar con información veraz, oportuna y representativa es un requisito en la vida diaria, decidir sobre el mejor tratamiento para un cultivo, el medicamento más adecuado para una enfermedad específica, un cambio en la macroeconomía, entre otros, requiere un análisis del contexto de forma objetiva.

La inferencia estadística es la base del método científico para generar nuevo conocimiento y valorar otro ya existente, por lo tanto, es oportuno conocer sobre las técnicas básicas y sus alcances, en particular, lo que corresponde al muestreo estadístico, sus características y validez.

La muestra está conformada por unidades estadísticas definidas en tiempo y espacio que dan un alcance y validez al estudio. Así, por ejemplo, el no tener presente estas características, en relación con las conclusiones derivadas puede llevar a conclusiones equivocadas. Este pareciera ser el error persistente en la aplicación de la estadística.

Las muestras elaboradas sin representatividad y aleatoriedad presentan limitaciones en la medición del error de muestreo, lo cual no permite la aplicación de la inferencia estadística, de manera que conservar las condiciones deseables de una muestra es básico para la adecuada utilización de la inferencia.

7. RESUMEN DEL TEMA

Al finalizar el tema se espera que usted sea capaz de:

- Explicar el concepto de inferencia estadística.
- Identificar las diferencias entre un parámetro y un estimador.
- Calcular estimaciones puntuales y de intervalo para diferentes parámetros.
- Definir el concepto de hipótesis estadística.
- Utilizar el concepto de significancia estadística para generar conocimiento.
- Utilizar el lenguaje propio de las pruebas de hipótesis estadísticas.
- Realizar pruebas de hipótesis estadísticas simples.

8. SOLUCIÓN A LOS EJERCICIOS

1. Resuelva los ejercicios 3.7.1 y 3.7.2.

Ejercicio 3.7.1

Dada una distribución normal de Y con media 5 y varianza 16, encuentre:

$$P(Y \leq 10) = P(Y-\mu)/\sigma \leq (10-5)/4) = P(Z \leq 1,25) = 0,8944$$

$$P(Y \leq 0) = P((Y-\mu)/\sigma \leq (0-5)/4) = P(Z \leq -1,25) = 0,1056$$

$$P(0 \leq Y \leq 15) = P((0-5)/4 \leq (Y-\mu)/\sigma \leq (15-5)/4) = (-1,25 \leq Z \leq 2,5)$$

$$P(Z \leq 2,5) - P(Z \leq -1,25) = 0,9938 - 0,1056 = 0,8881$$

$$P(Y \geq 5) = 1 - P(Y \leq 5) = 1 - P(z \leq 0) = 1 - 0,50 = 0,50$$

$$P(Y \geq 15) = 1 - P(Y \leq 15) = 1 - P(z \leq 2,5) = 1 - 0,9938 = 0,0062$$

Ejercicio 3.7.2

Dada una distribución normal de Y con media 20 y varianza 16, encuentre Y_0, tal que:

$$P(Y \leq Y_0) = P((Y-\mu)/\sigma \leq (Y_0 - 20)/4) = P(Z \leq Z_0) = 0,025$$

$$\rightarrow Z_0 = -1,960 \rightarrow -1,960 = (Y_0 - 20)/4 \rightarrow Y_0 = 12,1601$$

$$P(Y \leq Y_0) = P((Y-\mu)/\sigma \leq (Y_0 - 20)/4 = P(Z \leq Z_0)) = 0,01$$

$$\rightarrow Z_0 = -2,3263 \rightarrow -2,3263 = (Y_0 - 20)/4 \rightarrow Y_0 = 10,6946$$

$$P(Y \leq Y_0) = 0,95$$

$$\rightarrow Z_0 = 1,6449 \rightarrow 1,6449 = (Y0 - 20)/4 \rightarrow Y_0 = 26,5794$$

$$P(Y \geq Y_0) = 0,90 \rightarrow 1 - P(Y \leq Y_0) = 0,90 \rightarrow P(Y \leq Y_0) = 0,1$$

$$\rightarrow Z_0 = -1,2816 \rightarrow -1,2816 = (Y0 - 20)/4 \rightarrow Y_0 = 14,8738$$

2. Una variable aleatoria sigue una distribución x^2 de Pearson.

Es importante recordar que la tabla de la distribución Chi-cuadrado calcula la probabilidad de un valor más extremo que un $x_0 \rightarrow P\left(X > x_0\right)$, donde X es una variable aleatoria de x^2.

b. Los puntos críticos:

i. $x^2_{0,90,5} = 1,61$

ii. $x^2_{0,01,26} = 45,6$

iii. $x^2_{0,025,8} = 17,5$

c. Las probabilidades:

i. $P\left(x^2_8 \geq 3,49\right) = 0,9000$

ii. $P\left(x^2_8 \leq 15,5\right) = 0,9500$

iii. $P\left(x^2_{10} \geq 4\right) = 0,0404$

iv. $P\left(x^2_{20} \leq 29\right) = 0,9122$

3. Una variable aleatoria sigue una distribución t de Student. Se pide calcular:

Recuerde que la tabla de la distribución T-Student de una cola calcula la probabilidad de un valor más extremo que un $x_0 \rightarrow P\left(X > x_0\right)$ donde X es una variable aleatoria que sigue la distribución *t*; si el interés está en una probabilidad de dos colas, entonces la probabilidad es $P\left(\left|X\right| > x_0\right) = P\left(X > x_0 \text{ ó } X < -x_0\right)$.

b. Los puntos críticos:

i. $t_{0,20,20} = 1,3253$

ii. $t_{0,99,10} = 0,0129$

iii. $t_{0,25,10} = 1,2213$

c. Las probabilidades:

i. $P\left(t_{10} \geq 1{,}372\right) = 0{,}1000$

ii. $P\left(t_{8} \leq 1{,}2\right) = 0{,}8678$

ii. $P\left(-0{,}5 \leq t_{6} \leq 0{,}6\right) = 0{,}3973$

v. $P\left(\left|t_{24}\right| > 2\right) = 0{,}8861$

4. Una variable aleatoria sigue una distribución F de Fisher-Snedecor. Se pide calcular:

Recuerde que la tabla de la distribución F de Fisher-Snedecor calcula la probabilidad de un valor más extremo que un $x_0 \rightarrow P\left(X > x_0\right)$, donde X es una variable aleatoria con una distribución F con n grados de libertad en el numerador y m en el denominador.

a. Los puntos críticos:

i. $F_{0,1,10,12} = 4{,}71$

ii. $F_{0,05,5,24} = 2{,}62$

iii. $F_{0,90,28,30} = 0{,}62$

b. Las probabilidades: $P\left(2 \leq F_{10;20} \leq 2{,}25\right) = 0{,}0309$

TEMA 6. CONTRASTACIÓN DE HIPÓTESIS

1. OBJETIVO

Aplicar los conocimientos generales del proceso de inferencia estadística, a través de las pruebas de significancia para formular hipótesis en la generación de conocimiento.

Cuadro 21. Contrastación de hipótesis

Tema	Capítulo del libro	Páginas
Definición de hipótesis estadísticas	Capítulo 5	Presentación de Power Point: Pruebas de hipótesis estadísticas
Significancia estadística	Capítulo 5	
Pruebas de significancia para el promedio en una y dos muestras	Capítulo 5	
Pruebas de significancia para proporciones en una y dos poblaciones	Capítulo 5	
Pruebas de significancia para la varianza en una y dos poblaciones	Capítulo 5	
Pruebas de bondad de ajuste. Pruebas de Kolmogorov-Smirnov, chi-cuadrado	Capítulo 20	
Estadística no paramétrica: pruebas U Mann-Whitney y Kruskal-Wallis	Capítulo 24, sección 24.9	

2. PRUEBA DE KOLMOGOROV-SMIRNOV

La prueba de Kolmogorov-Smirnov es considerada no paramétrica, se utiliza para evaluar si una muestra proviene de una población específica; en esencia, esta valora las diferencias que se presentan entre las frecuencias observadas y las que, teóricamente, se esperan, bajo una distribución de probabilidad determinada con el fin de valorarlas mediante una hipótesis estadística y concluir acerca de su aleatoriedad. La distribución supuesta puede ser Binomial, Poisson, Normal, entre otras, aunque se prefiere que sea continua, debido a que las frecuencias esperadas

deben calcularse a partir del supuesto bajo hipótesis y, en consecuencia, se pueden presentar inconvenientes para calcular las probabilidades exactas.

Esta prueba también es conocida como "prueba de bondad de ajuste" y como "prueba de concordancia".

El estadístico de prueba, dentro del proceso de hipótesis, debe contrastarse con los valores tabulares definidos, puede revisarse en el apéndice del libro de texto las tablas A.22 y A.23ª según se requiera.

Algunas ventajas de la prueba de Kolmogorov-Smirnov son las siguientes:

- Se puede aplicar para tamaños de muestra pequeños.
- Es una prueba que se considera robusta, aun cuando la variable no sea continua, esto se refleja en la confianza de rechazar o no la hipótesis de nulidad, es decir que se puede tener confianza en la decisión pese a que la probabilidad no sea exacta.
- Adicionalmente, esta se considerada mejor que la "prueba de bondad de ajuste" basada en la distribución de chi-cuadrado.

Los pasos para aplicar la prueba de Kolmogorov-Smirnov, una vez planteada la hipótesis, son los siguientes:

1. Ordenar los datos de la muestra observada o empírica en clases, calcular la frecuencia simple y acumulada en términos relativos.
2. Obtener las frecuencias esperadas, simples y acumuladas, de acuerdo con la distribución teórica específica para cada uno de los valores o rangos definidos según corresponda.
3. Calcular las diferencias entre lo observado y lo esperado e identificar la máxima, este valor estadístico normalmente se identifica con la letra *D*.
4. Utilizar la tabla de valores críticos de *D* para comparar el valor estadístico *D* de Kolmogorov-Smirnov y tomar la decisión respecto a la hipótesis nula.

Ejemplo de aplicación de la prueba de Kolmogorov-Smirnov

Se tienen datos relacionados con el contenido de proteínas totales del plasma en prematuros normales de 15 días. Los datos fueron redondeados a la unidad más próxima, como se muestra en el cuadro 20.

Utilice un nivel de significancia del 5% y pruebe la hipótesis de que la variable aleatoria "proteínas totales en plasma" sigue una distribución normal.

Cuadro 22. Distribución de frecuencias de prematuros normales de 15 días según contenido de proteínas totales del plasma

Gramos de proteína por litro	Número de casos
De 40 a 44	2
De 45 a 49	6
De 50 a 54	12
De 55 a 59	13
De 60 a 64	5
De 65 a 69	2

Planteamiento de la hipótesis de nulidad (Ho) y la alterna (Ha)

Ho: la variable aleatoria "proteínas totales en plasma" sigue una distribución normal.

Ha: la variable aleatoria "proteínas totales en plasma" no sigue una distribución normal.

De forma equivalente puede plantearse:

Ho: las diferencias entre los valores teóricos de la distribución normal y los observados son aleatorios.

Ha: las diferencias entre los valores teóricos de la distribución normal y los observados no son aleatorios.

Aplicación de la prueba

- Ordenar los datos de la muestra observada o empírica en clases, calcular la frecuencia simple y acumulada en términos relativos.

- Obtener las frecuencias esperadas, simples y acumuladas, de acuerdo con la distribución teórica específica para cada uno de los valores o rangos definidos según corresponda.

Cuadro 23. Distribución de frecuencias relativas simples y acumuladas de prematuros normales de 15 días según contenido de proteínas totales del plasma

Gramos de proteína por litro	Número de casos	Frecuencia relativa	
		Simple	Acumulada
De 40 a 44	2	0,05	0,05
De 45 a 49	6	0,15	0,20
De 50 a 54	12	0,30	0,50
De 55 a 59	13	0,33	0,83
De 60 a 64	5	0,13	0,95
De 65 a 69	2	0,05	1,00
Total	40	1,00	

- Calcular las diferencias entre lo observado y lo esperado e identificar la diferencia máxima, este estadístico normalmente se identifica con la letra *D*.
    - Primero, calcule los límites reales para obtener el punto medio de la clase, luego obtenga la media y la varianza:

Cuadro 24. Distribución de frecuencias de prematuros normales de 15 días, cálculo del punto medio, la media y la desviación estándar

Gramos de proteína por litro	Número de casos	Punto medio	Suma Producto	Desviaciones
De 40 a 44	2	42,0	84	306,28
De 45 a 49	6	47,0	282	326,34
De 50 a 54	12	52,0	624	67,69
De 55 a 59	13	57,0	741	89,58
De 60 a 64	5	62,0	310	290,70
De 65 a 69	2	67,0	134	318,78
Total	40		2 175	1 399,38

Con base en los datos anteriores se obtiene:

Media	54,38
Desviación estándar	5,99

- Utilice la información anterior para obtener los límites estandarizados:

Cuadro 25. Distribución de frecuencias de prematuros normales de 15 días, cálculo de límites estandarizados

Límites reales			Límites estandarizados		
De 39,5	a	44,5	De 2,48	a	-1,65
De 44,5	a	49,5	De -1,65	a	-0,81
De 49,5	a	54,5	De -0,81	a	0,02
De 54,5	a	59,5	De 0,02	a	0,86
De 59,5	a	64,5	De 0,86	a	1,69
De 64,5	a	69,5	De 1,69	a	2,52

- Utilice la tabla A.4 de la curva normal estándar del apéndice del libro o utilice algún *software* para obtener las probabilidades acumuladas esperadas en cada clase:

Cuadro 26. Distribución de frecuencias de prematuros normales de 15 días, calculo de frecuencias esperadas acumuladas

Límites estandarizados			Frecuencia esperada acumulada
De -2,48	a	-1,65	0,049619841
De -1,65	a	-0,81	0,207867859
De -0,81	a	0,02	0,508324417
De 0,02	a	0,86	0,803884111
De 0,86	a	1,69	0,954513406
De 1,69	a	2,52	0,994215014

- Calcule las diferencias absolutas entre lo que, teóricamente, se espera y lo observado, luego identifique la diferencia máxima, este es el valor estadístico de Kolmogorov-Smirnov.

Cuadro 27. Distribución de frecuencias de prematuros normales de 15 días. Calculo de diferencias absolutas entre las frecuencias observadas y las teóricas

Límites estandarizados			**Frecuencias relativas**		**Diferencias absolutas**
			Esperada acumulada	**Observada acumulada**	
De -2,48	a	-1,65	0,050	0,050	0,0004
De -1,65	a	-0,81	0,208	0,200	0,0079
De -0,81	a	0,02	0,508	0,500	0,0083
De 0,02	a	0,86	0,804	0,825	0,0211
De 0,86	a	1,69	0,955	0,950	0,0045
De 1,69	a	2,52	0,994	1,000	0,0058

- Como se observa en el cuadro 27 el valor estadístico *D* de Kolmogorov-Smirnov es:

D máximo es 0,0211

- Utilice la tabla de valores críticos de *D* para comparar el valor estadístico *D* de Kolmogorov-Smirnov y tome la decisión respecto a la hipótesis nula. Utilice la tabla de K-S para un nivel de significancia de 5%, se obtiene que

$$D_t = \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{40}} = 0.22$$

- Decisión: como la diferencia máxima observada es menor que el valor crítico establecido por el estadístico *D* de Kolmogorov Smirnov, se concluye que hay evidencia estadística para pensar que la variable aleatoria "proteínas totales en plasma" sigue una distribución normal. Las diferencias observadas tienen un comportamiento aleatorio.

3. EJERCICIOS

Con el objetivo de reforzar los conocimientos adquiridos en este tema se recomienda que desarrolle los siguientes ejercicios:

1. ¿Qué es la hipótesis nula?
2. ¿Qué es la hipótesis alternativa?
3. ¿Qué tan útil es el valor *p* en la prueba de hipótesis?
4. ¿Cuál es la diferencia entre un error tipo I y un tipo II?
5. ¿Qué tan recomendable es probar una hipótesis por medio de intervalos de confianza?
6. Un organismo de control farmacéutico investiga una muestra de 35 frascos de cierto medicamento para controlar el contenido de una droga específica que afecta el ritmo cardíaco. Se pretende determinar si se están cumpliendo las especificaciones del caso, las cuales establecen que ese contenido no debe

diferir de $0,12gr/100ml$ Al evaluar la muestra, se encontró que el contenido medio es de $0,10gr/100ml$, con una desviación estándar de $0.02gr/100ml$. ¿Se estarán o no infringiendo las especificaciones? (Use $\alpha = 0,01$)

7. El desempleo en la zona urbana, durante varios años, se ha mantenido alrededor de 6,7 personas de cada 100. En el 2009, se realizó un estudio en diferentes épocas del año con base en 1 500 individuos y se estimó que 5,3 de cada 100 individuos estaban desempleados. ¿Se podrá decir, con el 5% de significancia que el porcentaje de desempleo ha variado en el 2009?

8. Resuelva el ejercicio 5.3.3 del libro.

4. OBSERVACIONES FINALES

La técnica estadística sobre contraste de prueba de hipótesis ha sido utilizada por científicos en todos los campos como un instrumento para la valoración de teorías y generación de nuevo conocimiento.

Pese a ser una técnica de amplio uso, es posible notar en trabajos de investigación formal errores metodológicos y de aplicación que deben llamar la atención sobre como la estadística es aplicada. Errores en la definición de las hipótesis, en la conceptualización del nivel de significancia, entre otros, llevan a concluir de manera equivocada lo que la evidencia estadística conlleva. La selección del valor estadístico de prueba es causa de errores de aplicación, puesto que se trabaja con condiciones que violentan los supuestos teóricos y no se realiza el ajuste a la técnica.

Los errores comunes en la aplicación de técnicas de inferencia estadística son varios, empezando con la condición de la muestra sobre la cual se basa la generalización, hasta errores propios de las pruebas de hipótesis; por ejemplo, es común leer la expresión “se acepta la hipótesis nula”, “existe evidencia estadística para aceptar la hipótesis nula”.La existencia de los errores tipo I y tipo II impiden aceptar una hipótesis estadística, sin embargo, permiten no rechazarla, lo correcto es concluir acerca de la posibilidad o no de rechazarla conforme la evidencia lo sugiera.

5. RESUMEN DEL TEMA

Al finalizar el tema se espera que usted sea capaz de:

- Formular hipótesis estadísticas para crear o validar conocimiento.
- Utilizar el lenguaje propio de las pruebas de hipótesis estadísticas.
- Identificar y aplicar el procedimiento para la prueba de hipótesis estadísticas.
- Valorar la importancia de conocer la técnica sobre pruebas de hipótesis y su impacto en la generación de conocimiento.

6. SOLUCIÓN A LOS EJERCICIOS

1. ¿Qué es la hipótesis nula?

   Es una afirmación de igualdad respecto a un parámetro estadístico.

2. ¿Qué es la hipótesis alternativa?

   Es una afirmación respecto al parámetro poblacional cuando Ho se rechaza.

3. ¿Qué tan útil es el valor *p* en la prueba de hipótesis?

   El valor de p en la prueba de hipótesis es fundamental para tomar la decisión respecto a Ho.

4. ¿Cuál es la diferencia entre un error tipo I y un tipo II?

   El error tipo I se comete cuando Ho se rechaza siendo cierta y el error tipo II es no rechazar Ho siendo falsa.

5. ¿Qué tan recomendable es probar una hipótesis por medio de intervalos de confianza?

   La prueba mediante intervalos de confianza es equivalente a calcular el valor estadístico de prueba o calcular el valor crítico de p. Los tres procedimientos son matemáticamente equivalentes.

6. Un organismo de control farmacéutico investiga una muestra de 35 frascos de cierto medicamento para controlar el contenido de cierta droga que afecta el ritmo cardíaco. Se pretende determinar si se están cumpliendo las especificaciones del caso, las cuales establecen que ese contenido no debe diferir de $0{,}12 gr/100 ml$ Al evaluar la muestra, se encontró que el contenido medio es de $0{,}10 gr/100 ml$, con una desviación estándar de $0.02 gr/100 ml$. ¿Se estarán o no infringiendo las especificaciones? (Use $\alpha = 0.01$)

H_0: $\mu = 0{,}12$ gr/100 ml

H_1: $\mu \neq 0{,}12$ gr/100 ml

μ_0	0,12		
Media	0,10		
n	35,00		
Desviación	0,02		
Número de colas	2,00	z_t	
Nivel de significancia	0,01	-2,58	
Zcal=	-5,92	*p*-calculado	0,00

El valor de la prueba de hipótesis muestra que hay evidencia estadística para rechazar la hipótesis de nulidad (|Zcal|>Zt o bien p<α), lo cual quiere decir que no se está cumpliendo con las especificaciones técnicas respecto al contenido de la droga en el medicamento.

En esta prueba debe utilizarse el estadístico Z debido a que el tamaño de muestra es mayor de 30 observaciones.

7. El desempleo en la zona urbana, durante varios años, se ha mantenido alrededor de 6,7 personas de cada 100. En 2009, se realizó un estudio en diferentes épocas del año con base en 1 500 individuos y se estimó que 5,3 de cada 100 individuos estaban desempleados. ¿Se podrá decir, con el 5% de significancia que el porcentaje de desempleo ha variado en 2009?

H_0: P_0= 6,7%

H_1: P_0 ≠6,7%

P_0	6,70%		
p	5,30%		
n	1500		
Desviación	0,00645554		
Zcal=	-2,17	p-value	0,02

El valor de la prueba de hipótesis muestra que hay evidencia estadística para rechazar la hipótesis de nulidad (|Zcal|>Zt o bien p<α), lo cual quiere decir que la tasa de desempleo ha variado en el 2009. En esta prueba, debe utilizarse el valor estadístico Z, debido a que el parámetro bajo análisis es una proporción.

Ejercicio 5.3.3

Los rendimientos de 10 plantas de fresas en un ensayo de uniformidad los presenta Baker y Baker (5,1) así: 239, 176, 235, 217, 234, 216, 318, 190, 181 y 225 g. Al 95% y 99%, calcule los intervalos de confianza para la media poblacional. Probar la hipótesis μ =205 (escogida arbitrariamente) con la alternativa μ≠ 205 al 5% de nivel de significancia.

H_0: μ = 205,00 g

H_1: μ ≠205,00 g

μ_0	205,00 g		
Media	223,10		
n	10,00		
Desviación	40,41		
Número de colas	2,00	t_t	
Nivel de significancia	0,05	2,69	
Número de colas	2,00	t_t	
Nivel de significancia	0,01	3,69	
t_{cal}=	1,42	p-calculado	0,19

El valor de la prueba de hipótesis muestra que hay evidencia estadística para no rechazar la hipótesis de nulidad ($|tcal|>t_t$ o bien $p>\alpha$), lo cual quiere decir que el rendimiento esperado en cada planta de fresa es de 250 g. En esta prueba, debe utilizarse el valor estadístico t, debido a que el tamaño de muestra es menor que 30 observaciones.

GLOSARIO

aleatorio. Describe eventos, cuya ocurrencia no es posible predecir, es decir, no responde a patrones sistemáticos o controlados.

asociación estadística. Busca definir si entre dos variables existe una relación o, por el contrario, el comportamiento observado responde a la aleatoriedad.

categoría. Representa las opciones posibles en que puede ser ordenado un atributo o característica cualitativa para construir una distribución de frecuencias. Además, los rangos en que puede ser ordenada una variable cuantitativa para comprender la forma en que los datos se agrupan para confeccionar una distribución de frecuencias.

coeficiente de correlación. Mide la fuerza de la asociación entre pares de variables y la dirección de la asociación, de manera que puede predecirse o esperarse, lo que puede ocurrirle a una cuando la otra presenta variaciones. La magnitud y fuerza de la asociación entre variables, descrita por medio de este, varía entre -1 y 1. Si la relación es fuerte, el resultado del coeficiente se acerca a los extremos -1 o 1 e indica una relación inversa o directa respectivamente.

coeficiente de regresión. Representa una medida de la dependencia funcional que existe entre dos variables, indica de forma cuantitativa cuál es el cambio esperado en la variable dependiente (Y) ante cambios unitarios en la variable independiente (X).

coeficiente de determinación. Es una medida de la bondad de ajuste de un modelo de regresión, el coeficiente de determinación (r^2) representa la proporción de la variación total en la variable dependiente (Y) que puede ser explicada por la variación en la variable independiente (X). Este equivale al cuadrado del coeficiente de correlación toma valores de 0 a 1 conforme su resultado se acerca a 1, indica que el modelo ajustado es bueno para describir funcionalmente la relación entre ambas variables.

covarianza. La covarianza (σxy) es una medida de la variación conjunta de dos variables, su valor describe la relación, positiva o negativa, que existe entre ellas y el grado o fuerza de su asociación.

cuartiles. Existen tres cuartiles: Q_1, Q_2 y Q_3. Estos números dividen las observaciones muestrales una vez ordenados en cuatro partes iguales. Así Q_1 determina el valor que divide al conjunto en un 25% de observaciones por debajo de él y el restante 75% está por encima del mismo valor. El segundo cuartil Q_2 corresponde a la mediana.

dependencia estadística. Término utilizado para indicar que dos o más variables presentan una relación, es decir, que el comportamiento de una afecta o condiciona el comportamiento de las otras.

desviación estándar (típica). Mide la dispersión o variabilidad presente en la característica de estudio. Tiene las mismas unidades de medida que la variable. Su cuadrado es la varianza.

diagrama de puntos o de dispersión. Es un gráfico bidimensional o tridimensional que permite observar los valores muestrales de dos o tres variables de forma conjunta, este es útil para valorar relaciones entre dos o tres variables.

diagrama de barras verticales. Representación gráfica que se utiliza para representar variables discretas.

diagrama de barras horizontales. Figura gráfica que se utiliza para representar atributos o características de tipo cualitativo.

error de estimación. Es un término que se utiliza en estadística para describir la diferencia que se espera entre una estimación basada en una muestra y el respectivo valor poblacional.

escala estadística. Se refiere al nivel de medición de las variables: nominal, dicotómica, discreta o continua, la cual es básica para aplicar medidas estadísticas resumen.

frecuencias absolutas. Representan el número de casos que se localizan en una determinada clase o categoría.

frecuencias relativas. Representan el tanto por uno o tanto por cien de los casos que se ubican en una clase o categoría. Se obtienen dividiendo la frecuencia absoluta por el tamaño muestral; estas suman 1 ó 100 según se expresen en tanto por uno o en tanto por ciento, respectivamente.

independencia estadística. Dos variables son consideradas independientes cuando la ocurrencia de una de ellas no influye en la ocurrencia de la otra; se espera que el comportamiento en ambas variables responda a la aleatoriedad.

máximo. Es el valor más alto observado en la muestra.

media. Es una medida de tendencia central que se utiliza para describir variables cuantitativas. Su cálculo se obtiene sumado las observaciones y dividiendo por el número de observaciones en la muestra.

mediana. Corresponde al percentil 50%. Es el valor que divide al conjunto de datos (muestras) en dos partes proporcionalmente iguales. Es el valor sobre el cual se encuentran el 50% de los valores y por debajo de ella el 50% restante. Es también una medida de tendencia central utilizada en variables cuantitativas.

mínimo. Es el menor valor observado en la muestra.

moda. Es el valor que se repite con más frecuencia dentro de la muestra.

muestra. Subconjunto de unidades estadísticas.

muestreo aleatorio simple. Es aquel en el que cada elemento de la población tiene la misma probabilidad de ser seleccionado.

muestreo con reemplazo. Es aquel en el que un elemento puede ser seleccionado más de una vez en la muestra, para ello, se extrae uno de ellos, se observa y se devuelve a la población, de esta forma, es posible hacer infinitas extracciones del conjunto aun siendo este finito.

muestreo sin reemplazo. Los elementos extraídos son observados y se mantienen fuera de la población.

parámetro. Es la medida que se obtienen sobre la distribución de probabilidades de la población, son medidas que se obtienen como la media, la varianza, la proporción, etc.

percentiles. Son valores que dividen al conjunto de datos en 100 partes proporcionalmente complementarias. Su aplicación tiene sentido cuando la característica bajo estudio es ordinal, de intervalo o razón. Su aplicación requiere que el conjunto de datos esté ordenado cuando el procesamiento es manual.

población. Es un conjunto de personas, entidades u objetos del cual se quiere saber algo que interesa para tomar una decisión acertada.

proporción. Número de elementos, personas u objetos que presentan la condición de interés dentro de la muestra respecto al total de observaciones. Se puede expresar sobre la base de tanto por uno o en tanto por cien.

rango. Diferencia entre el valor máximo y mínimo de un conjunto de datos (muestra o población). Se aplica solamente en variables cuantitativas.

rango intercuartílico. La diferencia entre los valores de la característica de interés que representan el percentil 75% y el percentil 25%, respectivamente.

sesgo. Representa un error que se debe a factores distintos del azar, es difícil de medir, pero es posible de ser controlado. Estos afectan la representatividad de un resultado obtenido por muestreo.

simetría. Es un término utilizado para describir la forma de una distribución de probabilidad o de una variable aleatoria; se considera simétrica si existe el mismo número de valores a la derecha que a la izquierda de la media, por tanto, se espera el mismo número de desviaciones con signo positivo que con signo negativo.

tablas de contingencia. Tablas de dos o más variables, en cada celda se representa el número de observaciones que cumplen con la combinación de los niveles de estas características.

unidad estadística. Aquella persona, elemento u objeto sobre el cual recae la observación, esta debe definirse en tiempo y espacio.

variable. Es una función que asocia a cada elemento de la población una medición en la característica que se desea observar.

variable bidimensional. Una variable bidimensional es aquella en la que cada elemento está definido por un par de características (X, Y) entre las que puede existir o no relación.

variables categóricas. Son aquellas que asumen valores cualitativos, es decir, que indican cualidades, etiquetas alfanuméricas o "nombres".

variables métricas. Las variables métricas asumen valores numéricos, a estas les corresponde las escalas de medición de intervalo o razón.

FUENTES BIBLIOGRÁFICAS

Impresas

Gómez Barrantes, Miguel (1998). *Elementos de Estadística Descriptiva* (Tercera Edición). Editorial EUNED.. San José.

Moya Meoño, Ligia (1986). *Introducción a la Estadística de la Salud* (Primera Edición). Editorial Universidad de Costa Rica. San José.

STEEL, Robert G. D. y TORRIE, James H (1992). *Bioestadística. Principios y Procedimientos*. Editorial Graf América. México

Digitales

Departamento de Ciencia de la Computación e Inteligencia Artificial (2005). *El problema de la dependencia entre variables medibles*. Recuperado en enero, 2010 de <http://www.dccia.ua.es/dccia/inf/asignaturas/EST/PRACTICA5E05.pdf>.

Galbiati, Jorge M (Sin año). *Distribucion f de Snedecor*. Recuperado en enero, 2010 de <http://www.jorgegalbiati.cl/nuevo_06/Fsned.pdf>.

García Mancilla, Hugo (Sin año). *Estadística Descriptiva e Inferencial I.* Recuperado en enero, 2010 de <http://www.conevyt.org.mx/bachileres/material_bachilleres/cb6/5sempdf/edin1/edin1_f1.pdf>.

Justel, Ana (2005-2006). *Aulas de Informática de la U.A.M.*. Recuperado en enero, 2010 de <http://www.uam.es/personal_pdi/ciencias/ajustel/docencia.html>.

Robledo Martín, Juana (Febrero, 2005). *Diseños de muestreo.* Recuperado en enero, 2010 de

<http://www.nureinvestigacion.es/FICHEROS_ADMINISTRADOR/F_METODOLOGICA/FMetod_12.pdf>.

Roldán Martínez, Antonio (Sin año). *Temas de Estadística Práctica*. Recuperado en enero, 2010 de <http://www.hojamat.es/estadistica/iniestad.htm>.

Ruiz Muñoz, David (Sin año). *Apuntes de Estadística*. Recuperado en Enero, 2010 de <http://www.eumed.net/cursecon/libreria/drm/ped-drm-est.htm>.

Saravia Gallardo, Marcelo Andrés (2001-2004). *Metodología de Investigación Científica.* Recuperado en enero, 2009 de <http://www.ciencia y tecnología.gob.bo/convocatorias/publicaciones/Metodologia.pdf>.