



UTPL
La Universidad Católica de Loja

Modalidad Abierta y a Distancia

Estadística

Guía didáctica



Índice

Primer
bimestre

Segundo
bimestre

Solucionario

Referencias
bibliográficas



Departamento de Ciencias Biológicas

Sección departamental de Ecología y Sistemática

Estadística

Guía didáctica

Autor:

Pablo Ancelmo Ramón Contento



Asesoría virtual
www.utpl.edu.ec

Índice

Primer
bimestre

Segundo
bimestre

Solucionario

Referencias
bibliográficas

Índice

Primer bimestre

Segundo bimestre

Solucionario

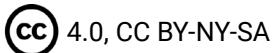
Referencias bibliográficas

Estadística

Guía didáctica

Pablo Ancelmo Ramón Contento

Universidad Técnica Particular de Loja



Diagramación y diseño digital:

Ediloja Cía. Ltda.

Telefax: 593-7-2611418.

San Cayetano Alto s/n.

www.ediloja.com.ec

edilojainfo@ediloja.com.ec

Loja-Ecuador

ISBN digital - 978-9942-25-746-8



La versión digital ha sido acreditada bajo la licencia Creative Commons 4.0, CC BY-NY-SA: Reconocimiento-No comercial-Compartir igual; la cual permite: copiar, distribuir y comunicar públicamente la obra, mientras se reconozca la autoría original, no se utilice con fines comerciales y se permiten obras derivadas, siempre que mantenga la misma licencia al ser divulgada. <https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es>

27 de abril, 2020

Índice

1. Datos de información.....	8
1.1. Presentación-Orientaciones de la asignatura.....	8
1.2. Competencias genéricas de la UTPL.....	8
1.3. Competencias específicas de la carrera.....	9
1.4. Problemática que aborda la asignatura.....	9
2. Metodología de aprendizaje.....	10
3. Orientaciones didácticas por resultados de aprendizaje	11
Primer bimestre.....	11
Resultado de aprendizaje 1	11
Contenidos, recursos y actividades de aprendizaje	11
Semana 1	11
Unidad 1. Introducción a la estadística	12
1.1. Nocións básicas	12
Actividad de aprendizaje recomendada	14
Semana 2	15
1.2. Variables estadísticas y escalas de medición.....	15
Actividad de aprendizaje recomendada	18
Autoevaluación 1	20
Semana 3	22
Unidad 2. Exploración de datos.....	22
2.1. Tablas de frecuencias	25
2.2. Gráficas estadísticas para variables categóricas	33
Actividad de aprendizaje recomendada	40
2.3. Gráficas estadísticas para variables numéricas	41
Actividad de aprendizaje recomendada	52

Índice	
Primer bimestre	
Segundo bimestre	
Solucionario	
Referencias bibliográficas	
Semana 4	53
2.4. Gráficas relacionales (bi-variadas).....	53
Actividad de aprendizaje recomendada	65
Autoevaluación 2	67
Semana 5	69
Unidad 3. Estadísticos descriptivos	69
3.1. Medidas de centralización	70
Actividad de aprendizaje recomendada	75
3.2. Medidas de variación	76
Actividad de aprendizaje recomendada	82
Semana 6	83
3.3. Medidas de posición (posición relativa)	83
Actividad de aprendizaje recomendada	87
Autoevaluación 3	88
Actividades finales del bimestre.....	90
Semana 7	90
Actividad de aprendizaje recomendada	90
Semana 8	91
Actividad de aprendizaje recomendada	91
Segundo bimestre	92
Resultado de aprendizaje 1	92
Contenidos, recursos y actividades de aprendizaje	92
Semana 9	92
Unidad 4. Probabilidad	92
4.1. Noción es básicas de probabilidades	94
4.2. Propiedades operacionales	99
Actividad de aprendizaje recomendada	101

Semana 10	102
4.3. Técnicas de conteo	102
4.4. Teoremas básicos de la probabilidad	104
Actividad de aprendizaje recomendada	109
Autoevaluación 4	110
Semana 11	112
Unidad 5. Distribuciones de variables aleatorias (discretas y continuas).....	112
5.1. Variables aleatorias y distribuciones de probabilidad	113
5.2. Distribución Binomial	119
Actividad de aprendizaje recomendada	124
Semana 12	125
5.3. Distribución de Poisson	126
5.4. Distribución normal	131
Actividad de aprendizaje recomendada	140
Autoevaluación 5	143
Semana 13	146
Unidad 6. Estimación estadística - Intervalos de confianza.....	146
6.1. Tipos de estimadores.....	149
6.2. Intervalo de confianza para la media	151
Actividad de aprendizaje recomendada	156
Semana 14	156
6.3. Intervalo de confianza para la proporción	157
Actividad de aprendizaje recomendada	161
Autoevaluación 6	162
Actividades finales del bimestre.....	165

Índice

Primer
bimestreSegundo
bimestre

Solucionario

Referencias
bibliográficas

Semana 15	165
Actividad de aprendizaje recomendada	165
Semana 16	166
4. Solucionario	167
5. Referencias bibliográficas	174

Índice

Primer
bimestre

Segundo
bimestre

Solucionario

Referencias
bibliográficas

Índice

Primer
bimestre

Segundo
bimestre

Solucionario

Referencias
bibliográficas



1. Datos de información

1.1. Presentación-Orientaciones de la asignatura



1.2. Competencias genéricas de la UTPL

- Pensamiento crítico y reflexivo
- Trabajo en equipo
- Comportamiento ético
- Organización y planificación del tiempo.

Índice

Primer
bimestre

Segundo
bimestre

Solucionario

Referencias
bibliográficas

1.3. Competencias específicas de la carrera

- Aplicar conocimientos de física, matemática, química y biología.

1.4. Problemática que aborda la asignatura

Debilidad en al análisis de información para desarrollar procesos óptimos de gestión en cuanto a la seguridad y salud ocupacional.



2. Metodología de aprendizaje

Para garantizar un proceso de aprendizaje significativo y el desarrollo de las competencias propuestas, las metodologías que se aplicarán en el desarrollo de la asignatura son:

- Aprendizaje por indagación, a través de esta metodología de aprendizaje se induce al estudiante a construir el conocimiento derivando en un entendimiento profundo, se provee de una diversidad de maneras flexibles para aproximarse a las preguntas de investigación motivo del análisis estadístico. Esta metodología fomenta en los estudiantes ciertos hábitos mentales que los estimula a plantearse preguntas sobre procesos de la vida real, puntos de vista, establecer relaciones y supuestos o hipótesis.
- Aprendizaje basado en análisis del estudio de caso, con esta técnica se desarrollan habilidades como el análisis, la síntesis y la evaluación de la información; así como, el pensamiento crítico que facilita no solo la integración de los conocimientos de la materia, sino que también, ayuda al alumno a generar y fomentar el trabajo en equipo, y la toma de decisiones, además de otras actitudes como la innovación y la creatividad.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas



3. Orientaciones didácticas por resultados de aprendizaje



Primer bimestre

Resultado de aprendizaje 1

Es capaz de aplicar los principios de la estadística y análisis de probabilidades.

Contenidos, recursos y actividades de aprendizaje



Semana 1

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas



Unidad 1. Introducción a la estadística

Estimado estudiante, en esta unidad nos vamos a centrar en conocer los aspectos básicos de la estadística, para ello es necesario que siga atentamente las instrucciones que se presentan en esta guía.

1.1. Nociones básicas

El autor del texto básico inicia definiendo a la estadística como una “disciplina que nos enseña a realizar juicios y tomar decisiones en presencia de incertidumbre”. Tabak (2011), propone dos tipos de incertidumbre, una de ellas es típica y se presenta cuando las cosas cambian naturalmente (ejemplo la temperatura durante el día), y otra llamada epistémica cuando se dispone de conocimiento insuficiente acerca de un proceso que se está analizando. Por ahora la que nos compete tener presente es la primera. La presencia de incertidumbre y variabilidad en la información estadística hace que sea necesario el uso de metodologías adecuadas para el tratamiento de los datos y poder extraer conclusiones confiables. Justamente de esto se encarga la estadística, nos provee de una amplia colección de métodos o técnicas para procesar la información. Su aplicabilidad se da en todas las esferas de la vida cotidiana, pero fundamentalmente en el campo de las ciencias; por ejemplo, en el ámbito de las ciencias biológicas y en general las ciencias de la vida, asume el nombre de bioestadística.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Así, la estadística ha llegado a ocupar un amplio escenario en el desarrollo de la ciencia y la tecnología, por lo que podemos decir que esta disciplina llegó para expandirse e incorporarse en la sociedad del conocimiento y la información. Tratar de definir la estadística mediante una sola expresión, sería limitar el amplio contexto de su aplicación.

Si quisieramos hablar de una clasificación general de la estadística podríamos decir que se divide en **descriptiva** e **inferencial**. La primera se encarga de presentar la información en forma resumida mediante valores numéricos, tablas o gráficas; esta etapa incluye lo que se denomina exploración de datos. La segunda se emplea para extraer conclusiones de una **población** a partir de un segmento representativo llamado **muestra**.

La mayoría (por no decir todos) de estudios o investigaciones trabajan con información extraída solamente de una parte de la población, precisamente porque resulta muy complicado o prácticamente imposible recopilar información de toda la población. Así, para referirnos a los valores que resumen los elementos de una población los llamaremos **parámetros**, mientras que, en el caso de la muestra, **estadísticos**.

La materia prima de la estadística, podría decirse que está compuesta por datos; cuyas técnicas de recolección también son desarrolladas por una rama de la estadística denominada **técnicas de muestreo**. A manera de resumen podemos describir las cuatro técnicas de muestreo más utilizadas: muestreo aleatorio simple (MAS), muestro sistemático (MS), muestreo estratificado (ME) y muestreo por conglomerados (MC). El MAS consiste en extraer una muestra aleatoria (al azar) de la población en estudio, donde cada unidad muestral tiene la misma posibilidad de ser elegida. El MS parte de una unidad muestral escogida al azar en la población y el resto de las unidades serán elegidas de manera uniforme considerando cierta distancia entre una y otra (por ejemplo, todas

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

las unidades que están en una posición múltiplo de cinco). El ME tiene por objetivo dividir la población de estudio en varios grupos o estratos homogéneos, y de cada estrato se procede a extraer una submuestra aleatoria que generalmente será proporcional al tamaño del estrato. Finalmente, para aplicar el MC también se procede a seccionar la población en varios grupos, pero sin importar que los grupos sean homogéneos, de los cuales se escogerán solamente algunos para constituir la muestra final. Más detalles sobre este tema encontrarán en el apartado introductorio y en las secciones 1.1 y 1.2 del texto básico.



Actividad de aprendizaje recomendada

Los recursos de aprendizaje que se recomienda revisar, para el estudio de los temas propuestos son:

Mendenhall, W., Beaver, R., & Beaver, B.. (2016). Introducción a la Probabilidad y Estadística. 13^a edición. México: CENGAGE Learning.

Este texto aborda los temas de forma clara y simplificada con lenguaje y estilo “amigable”, sin sacrificar la integridad estadística de la presentación. Trata de enseñar cómo aplicar los procedimientos estadísticos, al igual que para explicar: cómo describir de modo significativo conjuntos de datos reales, qué significan los resultados de las pruebas estadísticas en términos de sus aplicaciones prácticas; cómo evaluar la validez de los supuestos detrás de las pruebas estadísticas; y qué hacer cuando se han violado los supuestos estadísticos.



Semana 2

Para identificar los diferentes tipos de variables que podrían presentarse en un estudio o investigación sea experimental u observacional, le sugiero revisar los ejemplos de variables que se presentan en la sección 1.2 (Tipos de variables), del texto básico.

1.2. Variables estadísticas y escalas de medición

Luego de la lectura del texto usted pudo identificar que cualquier característica de una población que puede cambiar entre los individuos o elementos de la población, se denomina **variable**. Una clasificación muy general de las variables las presenta como **cualitativas** y **cuantitativas**. Las primeras son aquellas que están conformadas por dos o más categorías, sobre las cuales no se pueden efectuar ningún tipo de operaciones algebraicas; cuando se constituyen por dos categorías toman el nombre de dicotómicas (o binarias). Las cuantitativas en cambio representan magnitudes numéricas que pueden ser **discretas** (cuando se expresa mediante números enteros incluido el cero) y **continuas** (números enteros y fraccionarios, lo que comúnmente llamamos decimales).

Para complementar la explicación del texto, vamos a decir que el proceso de asignar números, letras, palabras o símbolos a una variable se le llama **medición** y la forma de asignación determina el tipo de **escala** de medición. La selección adecuada del método que nos ayudará a describir y analizar los datos dependerá del conocimiento de la escala a la que pertenece una medición. Según Stevens (1946, 1957) las escalas de medición se clasifican en: nominal, ordinal, intervalo y escala de razón. Cada escala tiene sus

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

propiedades matemáticas que determinarán el análisis estadístico adecuado en cada caso.

Escala nominal, las unidades observacionales (individuos o elementos de la muestra) se agrupan de acuerdo a una determinada propiedad o categoría de la variable. Por ejemplo, la variable sexo podría codificarse como M (masculino) y F (femenino), o también mediante números 1=masculino, 2= femenino. El hecho de emplear número para codificar una variable no quiere decir que la variable sea numérica; son simplemente identificadores arbitrarios, donde la operación matemática es el conteo (frecuencia).

Escala ordinal, aquí se emplea la relación de orden para la asignación de categorías o grupos, no se puede determinar distancias entre categorías. Por ejemplo, nivel socioeconómico, nivel de dolor, grado de intervención de un bosque, etc.

Escala de intervalo, puede establecerse orden entre sus valores, comparaciones de igualdad y medir distancias entre cada valor de la escala. El valor cero de la escala no es absoluto sino arbitrario, es decir no representa ausencia de la variable, por ejemplo, una temperatura de 0°C no implica ausencia de temperatura, así mismo la distancia numérica entre **25°C y 30°C** es la misma entre **35°C y 40°C** , pero esto no implica que una temperatura de **40°C** sea el doble de intensa que una temperatura de **20°C** .

Escala de razón, es la forma más completa de medición, posee las propiedades de la escala de intervalo, pero además incluye el cero absoluto, es decir el cero representa ausencia total de la variable que se está midiendo. Por ejemplo, la longitud, el peso, el tiempo, etc.

Una vez que es capaz de diferenciar entre poblaciones y muestras, y además distinguir los tipos de variables y sus escalas de medición, le propongo el siguiente ejemplo de un estudio observacional de campo donde se describen algunas variables cuya relación interesa analizar.

Ejemplo 1.1: Supongamos que usted está interesado en estudiar la relación entre el tiempo de servicio de los empleados, la edad y el área funcional de la empresa donde labora, en tres tipos de empresas.

Las variables que se observan en el estudio son: tiempo de servicio (variable numérica continua que puede expresarse en años y meses), edad (variable numérica discreta expresada en años), especie (variable categórica), área funcional (puede considerarse como nominal si se expresa en niveles como: producción, finanzas, marketing, ventas, etc.) y tipo de empresa también sería nominal. La población estadística estaría conformada por las edades o tiempos de servicio de todos los empleados en la empresa. En la práctica es muy complejo levantar toda la información de la población, en caso de que la empresa esté conformada por cientos o hasta miles de empleados, resultando más efectivo realizar las medidas solamente en una fracción de la población. En este caso lo más adecuado sería ubicar submuestras en las diferentes áreas y en cada área seleccionar una muestra aleatoria de empleados y realizar las mediciones de las variables de interés sólo de las personas dentro del área. Estas personas constituyen un subconjunto de la población.

En cuanto al tipo de empresa, está conformado por un rango discreto de clases que puede ser de acuerdo con la actividad económica, por ejemplo: industrial, comercial, salud, educación, etc. En un muestreo, podrían ser considerados como estratos o conglomerados.

Espero que el ejemplo de estudio observacional descrito en el párrafo anterior le haya servido para identificar los diferentes tipos de variables estadísticas, así como las respectivas escalas de medición. Podemos también hacer uso del software estadístico R para la identificación de variables (ver documento de introducción al R).



Actividad de aprendizaje recomendada

Luego de haber revisado los fundamentos de la estadística, así como los diferentes tipos de variables que podrían presentarse en un estudio o investigación, con la ayuda del texto básico le recomiendo dar respuesta a los enunciados siguientes:

En la sección de Introducción del texto básico, así como en las secciones 1.1 y 1.2 del texto básico:

1. El autor hace una breve descripción de la estadística descriptiva e inferencial, lea estos conceptos, y mediante ejemplos construya las similitudes y diferencias.
2. El autor comenta sobre los pasos para realizar una inferencia, lea el apartado y diferencie cada paso.
3. Luego, en la sección 1.2 (variables y datos), se hace una descripción detallada sobre los tipos de variables, lea y realice un cuadro resumen, proponiendo un ejemplo para cada tipo de variable.

Retroalimentación: Como su nombre lo indica, el término univariado hace referencia a una sola variable, mientras que multivariado se refiere a dos o más variables. La estadística se clasifica en dos ramas generales: descriptiva e inferencial. Por otro lado, la

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

estadística moderna se involucra en muchas áreas del conocimiento, para ejemplificar considere solamente tres áreas: Biología, Ecología y Ciencias Ambientales. Un aspecto básico a tener en cuenta al momento de recopilar información es la aplicación de muestreos aleatorios, aunque hay situaciones particulares donde el muestreo puede ser orientado.

Además, se sugiere desarrollar la **autoevaluación 1**. Para contestar las preguntas de la autoevaluación 1, es necesario que haya revisado los temas de las semanas 1 y 2, tanto en el texto básico como en los otros recursos recomendados. Esta temática está relacionada con la introducción a la estadística, variables y escalas de medición, lo cual le permitirá fortalecer su conocimiento en cuanto a la clasificación de variables estadísticas que se pueden identificar en un estudio, proyecto o investigación.



Autoevaluación 1

Lea con atención los enunciados del 1 al 5 y encierre en un círculo el literal que corresponda a la opción correcta.

1. Una de las ramas de la estadística es:
 - a. Una variable.
 - b. La estadística inferencial.
 - c. La escala de medición.

2. Una característica que puede cambiar entre los elementos de una población o de una muestra se denomina:
 - a. Una variable.
 - b. Una muestra.
 - c. Estadística descriptiva.

3. Realizar un censo equivale a:
 - a. Tomar datos de algunos elementos de la población.
 - b. Extraer varias muestras de la población.
 - c. Extraer información de todos los elementos de la población.

4. Los parámetros se relacionan con las:
 - a. Características de la muestra.
 - b. Características de la población.
 - c. Variables numéricas.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

5. La variable “área basal” de un conjunto de árboles, es de tipo:

- a. Numérica discreta.
- b. Numérica continua.
- c. Categórica.

En los ítems del 6 al 10, dentro del paréntesis, escriba V si la afirmación es correcta y F si es falsa.

- 6. () Un ejemplo de escala nominal sería: “La cantidad de hectáreas de bosque, deforestadas anualmente en el Ecuador”.
- 7. () La etapa de exploración de los datos está vinculada exclusivamente con las variables categóricas.
- 8. () La estadística inferencial busca extraer conclusiones para la población a partir de la muestra.
- 9. () El muestreo estratificado se caracteriza por dividir la población en grupos, todos de igual tamaño.
- 10. () Una muestra se denomina de conveniencia cuando se eligen los individuos u objetos que van a conformar la muestra, sin aleatorizar.

[Ir al solucionario](#)

Luego de haber respondido la autoevaluación, compare con el solucionario que se encuentra al final de la guía didáctica, y realice los correctivos si es necesario.



Semana 3



Unidad 2. Exploración de datos

El análisis exploratorio de datos es considerado un enfoque o filosofía para analizar datos que emplea variedad de técnicas con el propósito de: maximizar la comprensión de un conjunto de datos, descubrir la estructura subyacente de los datos, detectar anomalías (*outliers* o valores atípicos), probar suposiciones, desarrollar modelos, etc. Análisis exploratorio no equivale estrictamente a gráficas estadísticas, aunque a veces se los emplea indistintamente; aunque el análisis exploratorio utiliza en gran medida gráficas estadísticas. En esta etapa exploratoria, podemos refinar la pregunta de investigación o recolectar nuevos datos en caso de ser necesario.

En la presente guía, para una efectiva exploración de datos, se abordará las técnicas básicas de agrupamiento y resumen de datos mediante tablas y gráficas estadísticas.

¿Por qué son importantes las gráficas estadísticas?

Partiendo de la conocida frase “una imagen vale más que mil palabras”, se puede decir que esto se cumple efectivamente en el campo de la estadística. Las gráficas estadísticas constituyen una de

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

las más potentes herramientas disponibles para describir y apoyar en un proceso de análisis de datos. La fortaleza radica en que mediante las gráficas se puede transmitir gran cantidad de información de forma rápida y eficiente. Su representación comporta un lenguaje universal, a pesar de las diferencias culturales o de idioma, su interpretación es posible.

Las gráficas permiten almacenar y resumir grandes conjuntos de datos, y pueden integrarse estrechamente dentro de metodologías estadísticas más complejas y formales como técnicas de modelamiento, análisis multivariados entre otros. Es por ello que previo a la realización de un análisis estadístico, es recomendable y necesario conocer la estructura o naturaleza de los datos, características que se pueden revelar mediante las gráficas. Es así que las gráficas son a la vez una herramienta que facilita el razonamiento crítico a cerca de los datos. En la presente guía, se revisará solamente las gráficas más básicas y a la vez más usuales a la hora de representar información estadística. En la figura 1 se presenta un resumen de las gráficas uni y bi-variadas.

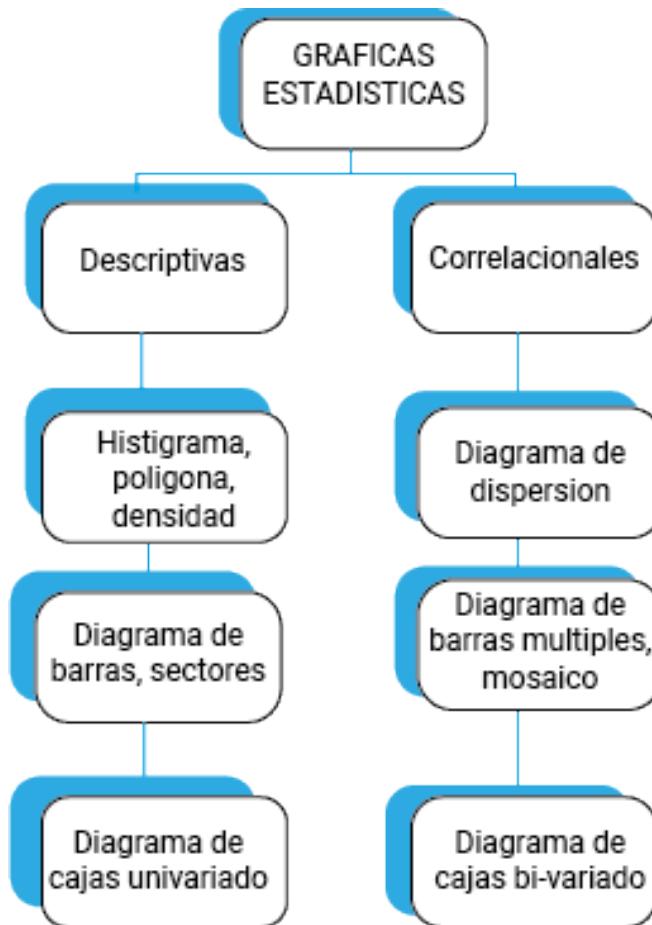


Figura 1. Principales gráficas estadísticas univariadas y bivariadas.

Elaborar una gráfica estadística implica tener en cuenta aspectos de diseño como tamaño, escalas de las variables entre otros, sin embargo, es más importante considerar el objetivo o mensaje que se desea trasmitir. En la medida que las gráficas transmiten la información para la cual han sido construidas, se puede hablar de eficiencia o falta de pertinencia en la gráfica. Excederse en detalles estéticos innecesarios puede distraer la atención e interferir en una clara interpretación de la gráfica, se recomienda buscar un equilibrio entre simplicidad del diseño y simplicidad en la interpretación.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

En la actualidad, la estadística moderna con el apoyo de la tecnología (software estadístico comercial o libre) puede hacer mayor uso de las técnicas gráficas como parte de una rutina de análisis exploratorio. De la flexibilidad del software estadístico, dependerá el proceso de refinar una gráfica, el cual conlleva manipulación de ejes, extracción de subconjuntos de datos, etiquetado, colores, leyendas entre otros. Precisamente el entorno R que hemos elegido para ejecutar los análisis estadísticos en esta guía, tiene la ventaja de ser altamente flexible y permitir elaborar gráficas de alta calidad. Al final de la guía didáctica, usted cuenta con una introducción al aprendizaje de R, diseñado fundamentalmente para principiantes. Además, para fortalecer el aprendizaje de R, puede acceder a los tutoriales y cursos en línea descritos en el plan docente (REA 1), así como al documento de introducción al programa R (Anexo de la guía).

Antes de dar algunos lineamientos acerca del agrupamiento de datos, es necesario tratar de conceptualizar el término “datos”. El término **datos** puede ser definido como información que representa atributos cualitativos o cuantitativos de una variable o un conjunto de variables. Estadísticamente los datos pueden ser clasificados en agrupados y no-agrupados. Cualquier dato que se recopila en primer lugar es un dato no agrupado; por ejemplo, el índice de masa corporal de una persona, el tiempo que se tarda en desarrollar un determinado proceso de producción, el número de accidentes laborales en una industria etc.

2.1. Tablas de frecuencias

El objetivo de agrupar grandes cantidades de datos es facilitar el cálculo de varias medidas descriptivas como porcentajes, promedios, varianzas, etc. Hoy en día, gracias al desarrollo de software estadístico (libre y comercial), el proceso de agrupamiento de datos es sencillo, facilitando en gran medida el resumen de la información.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Una tabla de frecuencias es el resultado de la tabulación de datos, en la que aparecen de forma bien organizada los valores (frecuencias) de las variables que se están estudiando.

Exclusivamente cuando la variable de análisis es cuantitativa (de preferencia continua), y el número de valores de la variable es grande, los valores de la variable se agrupan en **intervalos de clase**, que generalmente poseen la misma amplitud y son mutuamente excluyentes (no se traslapan). Entonces surge la pregunta ¿Cuántos intervalos se deben considerar?, no hay una receta estricta sino, la mejor guía en estos casos es conocer la naturaleza de los datos. Una regla empírica establece que deben ser entre 6 y 15 intervalos para obtener un resumen adecuado. Otra pregunta que debe responderse se refiere a la amplitud que debe tener cada intervalo. En la mayoría de los procesos de construcción de una distribución de frecuencias, se opta por intervalos de igual amplitud. En este caso la amplitud (A) se calcula mediante la relación:

$$A = \frac{R}{k}$$

Donde, R es el rango de los datos ($R = \max(x) - \min(x)$) y k representa el número de clases que se desea construir. Sin embargo, con esta relación se podría obtener valores de amplitud poco convenientes debido a que la división es inexacta y en ese caso habrá que considerar paralelamente el sentido común. Una regla sencilla que da buenos resultados es generar amplitudes de 5 o 10 unidades (o si la escala de la variable está entre 0 y 1, puede considerarse amplitudes de 0.05, 0.10, ...), en estos casos el límite inferior del primer intervalo debe ser menor o igual al valor mínimo de la variable, y análogamente el límite superior del último intervalo debe ser mayor o igual al máximo valor de la variable.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Para realizar los ejemplos ilustrativos, vamos a considerar algunos conjuntos de datos que el programa R trae incorporados. Para acceder a los datos del programa escribimos en la consola del programa (R console) la siguiente instrucción: `data()` y ejecutamos con <<enter>>. Luego aparecerá un listado de nombres, donde cada uno corresponde a un set de datos (llamados también objetos), conforme se observa en la figura 2.

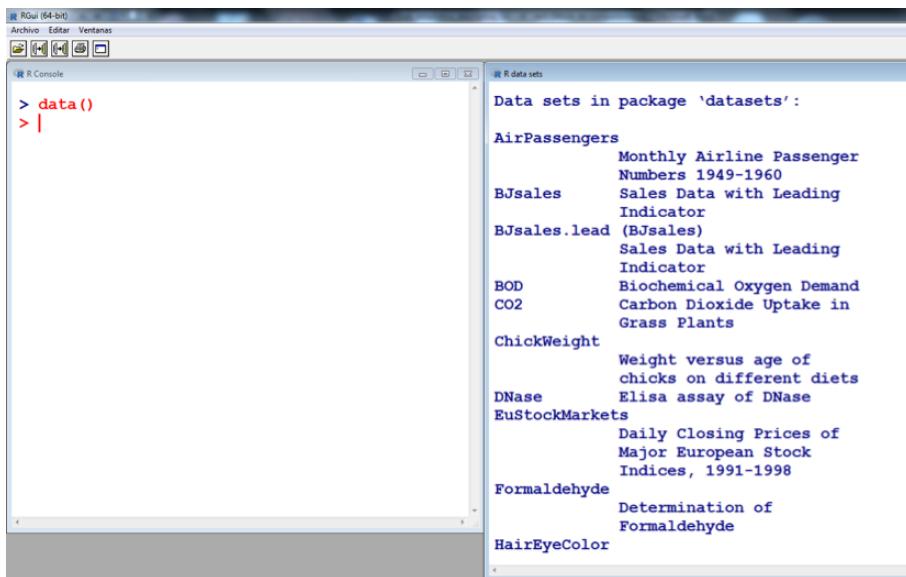


Figura 2. Listado de datos disponibles en el programa R.

Fuente: R Core Team (2017)

Ejemplo 2.1.1

Utilizando la tabla de datos “airquality” (Chambers, Cleveland, Kleiner, y Tukey 1983) que se encuentra incorporada en el programa R, se desea construir una distribución de frecuencias de 10 clases (intervalos) para la variable “Temp” (temperatura en grados Fahrenheit °F).

El ejemplo consiste en construir una tabla con 5 columnas. Las columnas están conformadas por: intervalos de clases (o

simplemente clases), frecuencias absolutas, frecuencias relativas, y las respectivas frecuencias acumuladas. Todas estas operaciones las realizaremos con ayuda del programa estadístico. En la consola del programa R, vamos a seguir el siguiente procedimiento:

1. Para ver la estructura de la tabla se escribe la función `str(airquality)`. Esta función nos permite saber cuántas y de qué tipo son las variables contenidas en la tabla. En este caso, se obtienen los siguientes resultados:

```
'data.frame': 153 obs. of 6 variables:  
  
 $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA  
 ...  
 $ Solar.R: int 190 118 149 313 NA NA 299 99  
 19 194 ...  
  
 $ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6  
 13.8 20.1 8.6 ...  
  
 $ Temp : int 67 72 74 62 56 66 65 59 61 69  
 ...  
 $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
 $ Day : int 1 2 3 4 5 6 7 8 9 10 ...
```

El término “`data.frame`” en el programa se interpreta como conjunto de datos, o simplemente una tabla. Observamos que se trata de una tabla de 153 filas (renglones) y 6 variables (columnas). Todas las variables son numéricas, cinco variables discretas y una continua, así lo indica la descripción abreviada “`int`” y “`num`”, respectivamente.

Por ahora nuestro interés es resumir en clases la variable Temp (que se ubica en la cuarta columna).

2. Para efectuar más fácilmente las operaciones, vamos a crear una copia de la variable Temp, con el nombre de “temperatura”, para ello escribimos:

```
temperatura=airquality$Temp
```

3. Ahora, es necesario conocer el rango o amplitud de la variable “temperatura”, para ello escribimos: range(temperatura)

Observamos que el rango de valores de la temperatura va desde 56 hasta 97°F.

4. Creamos un vector llamado clases, mediante una secuencia numérica. Puesto que deseamos 10 clases, escribimos:

```
clases=seq(56, 97, 4.1)
```

donde 56 y 97 son los valores obtenidos en el rango, y 4.1 sería el ancho de clase que se obtiene así: $(97 - 56)/10$

5. Utilizamos la función “hist” para crear la distribución de frecuencias, y almacenamos los resultados con el nombre H:

```
H=hist(temperatura,breaks=clases)
```

Finalmente, obtenemos los resultados digitando H:

```
$breaks
```

```
60.1 64.2 68.3 72.4 76.5 80.6 84.7 88.8 92.9
97.0
```

```
$counts
```

```
8 8 13 10 22 24 29 20 12 7
```

\$density

```
0.01275307 0.02072374 0.01594134 0.03507094
0.03825921 0.04622987 0.03188267 0.01912960
0.01115894
```

\$mids

```
62.15 66.25 70.35 74.45 78.55 82.65 86.75 90.85
94.95
```

Donde:

H\$breaks: son los límites de los intervalos de clase

H\$counts: son las frecuencias de cada clase (F)

H\$mids: representa los puntos medios de cada clase o intervalo.

Calculamos también las frecuencias relativas (si es de interés) de la siguiente manera:

$Fr = H\$counts / 153$, donde 153 es el número total de individuos (o número de renglones en la tabla). En resumen, la frecuencia relativa es el cociente entre la frecuencia absoluta y el total de individuos n.

En este caso, para obtener el tamaño de muestra ($n=153$) basta con digitar: `length(temperatura)`. Las frecuencias acumuladas (Fa, Fra) absoluta y relativa, respectivamente, se calculan utilizando la función `cumsum()`, esta función realiza la operación suma sucesiva desde el primer elemento hasta el último, así

$Fa = cumsum(H\$counts)$

$Fra = cumsum(H\$counts) / 153$

Para controlar o fijar el número de dígitos decimales, podemos hacerlo de la siguiente forma:

```
Fra= cumsum(round(H$counts/153, digits=3))
```

Donde la instrucción “round”, permite redondear las frecuencias a tres cifras decimales. Los resultados del agrupamiento de la variable “temperatura” se observan en la Tabla 1.

Tabla 1. *Agrupamiento de los datos de la variable “Temp” de la tabla “airquality” en 10 clases de igual amplitud. F: frecuencia absoluta, Fa: frecuencia absoluta acumulada, Fr: frecuencia relativa, Fra: frecuencia relativa acumulada.*

Clase	F	Fa	Fr	Fra
[56.0 60.1)	8	8	0.052	0.052
[60.1 64.2)	8	16	0.052	0.104
[64.2 68.3)	13	29	0.085	0.189
[68.3 72.4)	10	39	0.065	0.254
[72.4 76.5)	22	61	0.144	0.398
[76.5 80.6)	24	85	0.157	0.555
[80.6 84.7)	29	114	0.190	0.745
[84.7 88.8)	20	134	0.131	0.876
[88.8 92.9)	12	146	0.078	0.954
[92.9 96.0)	7	153	0.046	1

Fuente: Chambers et al. 1983

Nótese que cada una de las clases está descrita por un intervalo semi-aberto, es decir, abierto en el extremo derecho y cerrado en el extremo izquierdo, lo que asegura que cada observación pertenezca exclusivamente a una sola clase.

A continuación, le propongo una función que le permitirá calcular de forma rápida y sencilla la tabla de frecuencias.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

```
### Función para construir una tabla de frecuencias

distrib.frec<-function(datos,n.clases){

  datos<-na.omit(datos)

  ac=(range(datos)[2]-range(datos)[1])/n.clases

  clases=seq(range(datos)[1],range(datos)[2],ac)

  H=hist(datos,breaks=clases)

  F=H$counts

  Fa=cumsum(H$counts)

  n=length(datos)

  Fr= H$counts/n

  Fra= cumsum(H$counts/n)

  C=1:n.clases

  tabla.frec=cbind(C,F,Fa,Fr,Fra)

  print(tabla.frec)

}
```

Para utilizar esta función, debe copiar todo el texto que está en el recuadro y pegarlo en la consola de R, y ejecutar con “enter”. Luego, para utilizar la función con los datos de temperatura, escribimos de la siguiente forma:

```
distrib.frec(temperatura,10)
```

Y obtendrá la tabla de frecuencias en la consola del programa (Figura 3). Además de la tabla, se obtiene una gráfica de la distribución de frecuencias. Este tipo de gráficas y otras más, se discutirán en la siguiente unidad de la guía.

Como se puede observar el uso de esta función “distrib.freq” es bastante sencillo, solo requiere ingresar los datos (en este caso temperatura) y el número de clases (en este caso 10).

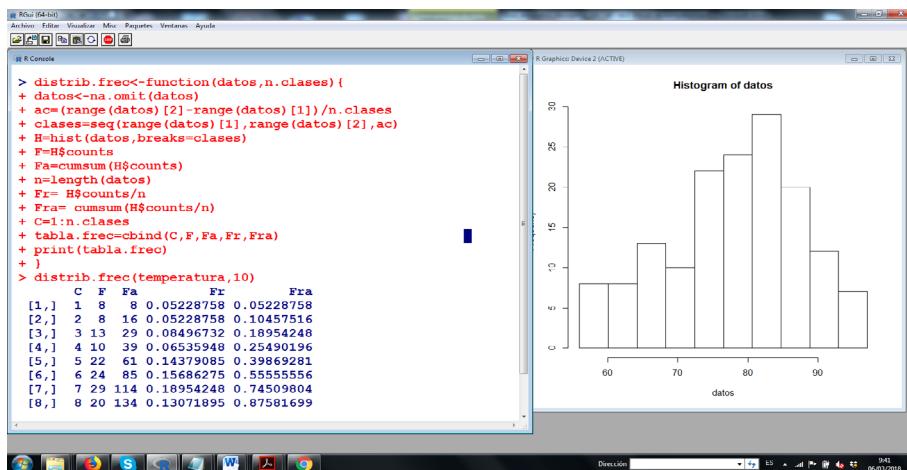


Figura 3. Tabla de frecuencias para los datos de la variable Temperatura.

Fuente: R Core Team (2017)

2.2. Gráficas estadísticas para variables categóricas

Usualmente iniciamos el trabajo con datos categóricos resumiendo la información mediante tablas de frecuencia (ver Secc. 2.1 del texto básico); posteriormente necesitamos una forma más intuitiva de presentar la información, es ahí donde hacemos uso de las gráficas. En el caso de variables categóricas (o nominales), las gráficas más comunes son el diagrama circular de sectores (pie chart) y el diagrama de barras (bar plot).

Diagrama circular

El gráfico circular (o sectores), es una representación con regiones o cortes de un círculo con diferentes colores, el área de cada región corresponde a las frecuencias absolutas o relativas de las categorías de la variable nominal.

En el programa estadístico R, un diagrama circular se construye utilizando la función `pie(x, labels)`, donde `x` toma valores positivos (frecuencias) y `labels` un vector de caracteres o nombres para cada región.

Ejemplo 2.2.1

Crear una nueva variable llamada `wind.cat` a partir de categorizar la variable velocidad del viento (`Wind`) de la tabla “airquality” del programa R, en tres categorías: ≤ 5 ; $(5, 15]$ y > 15 millas por hora. Luego, representar mediante un diagrama circular la nueva variable `wind.cat`.

En el programa R, para crear una variable categórica a partir de otra numérica empleamos la función `cut()` que significa cortar. A continuación, el proceso completo.

1. Creamos una variable numérica en este caso:

`viento=airquality$Wind`. Asignamos un nombre a la nueva variable categórica que vamos a construir, en este caso le llamaremos `wind.cat`, y escribimos la siguiente línea:

```
wind.cat=cut(viento, breaks=c(0,5,15,Inf),  
labels=c("baja","media","alta"))
```

Donde la condición “breaks” indica los límites de las clases, la primera clase inicia en 0 y termina en 5, la segunda clase de 5.1 a 15, y la tercera clase considera todos los valores superiores a 15, y la condición “labels” asigna nombres a las categorías (esto es opcional).

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

2. Ahora la variable está creada, podemos observar lo que hemos creado, digitando `wind.cat` y ejecutando con “enter”. Una forma sencilla de resumir una variable categórica es generar la frecuencia absoluta de cada categoría, esto se denomina tabulación de la variable y lo hacemos con ayuda de la función “`table`”, así:

```
t=table(wind.cat)
```

3. (Opcional) Definimos los colores para el diagrama, caso contrario, el programa asignará colores por defecto, en este caso vamos a definir tres colores con ayuda de la función `heat.colors`:

```
colores=heat.colors(3)
```

4. Calculamos los porcentajes de cada sector, estos porcentajes corresponden a la frecuencia relativa porcentual de cada categoría, para ello digitamos lo siguiente:

```
p=round(t/sum(t) * 100, 1)
```

5. Agregamos el signo “%” a cada valor

```
p=paste(p, "%", sep="")
```

6. En resumen, generamos el diagrama circular con ayuda de la función `pie`:

```
pie(t, col=colores, labels=p, cex=1.5)
```

7. Finalmente, creamos una leyenda para indicar lo que representa cada color de la gráfica, digitando:

```
legend("topright", names(t), cex=1.2,  
fill=colores, title="Velocidad-viento", bty="n")
```

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Donde “`topright`” (superior derecho) determina la ubicación de la leyenda, la función “`names`” permite que se impriman en la leyenda los nombres de las categorías de la variable que constan en la tabla `t`, “`fill`” imprime los colores fijados, y “`title`” permite incluir un título a la leyenda.

Y ahora el diagrama de sectores para la variable “`wind.cat`” está listo (Figura 4A). Usted puede hacer cambios adicionales a la gráfica, por ejemplo, modificar los colores, tipos de letra, etc. Para más información al respecto, consulte la ayuda de la función pie digitando `help(pie)`.

Una alternativa al diagrama de sectores bi-dimensional es el tri-dimensional, aunque no muy utilizado en publicaciones científicas, pero es posible también construirlo en este programa:

Para construir el diagrama circular 3D (comúnmente llamado diagrama de pastel), se requiere instalar una librería adicional en el programa R, llamada “`plotrix`”. Para ello deberá seguir los siguientes pasos:

Paso 1: Instalar la librería digitando:

```
install.packages("plotrix")
```

Paso 2: Activar (habilitar) la librería escribiendo:

```
library(plotrix)
```

Paso 3: Finalmente, construir el diagrama en 3D digitando lo siguiente:

```
pie3D(t, labels=p, explode=0, main="Velocidad del viento", theta=pi/4, col=c(2,3,4))

legend("top", names(t), cex=1.2,
fill=c(2,3,4), bty="n", horiz=T)
```



En la Figura 4 A y B, se puede observar dos formatos del diagrama circular correspondientes a la misma variable categórica, velocidad del viento, denotada por wind.cat. Una interpretación sería que la categoría predominante es la velocidad media con una frecuencia relativa porcentual de 86.9%, en otras palabras, de todos los 154 días observados, en el 86.9% (133 días) de las observaciones se registró una velocidad del viento entre 5 y 15 millas por hora.

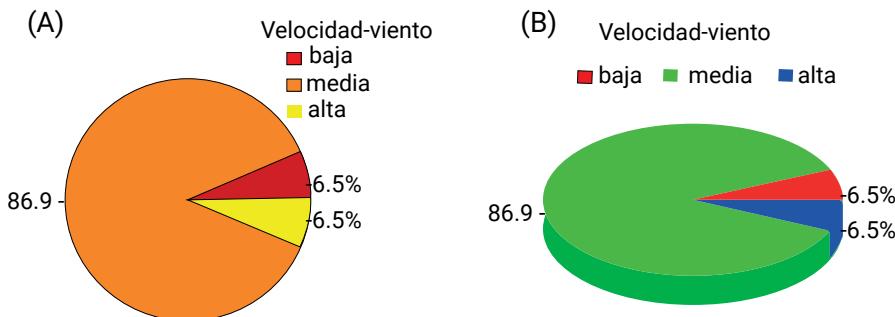


Figura 4. Diagrama circular de la velocidad del viento. (A) Diagrama bi-dimensional, (B) Diagrama tri-dimensional.

Fuente: Chambers et al. 1983

Diagrama de barras

Es otra alternativa de representación gráfica para variables categóricas o cualitativas. Para cada categoría o nivel de la variable le corresponde una barra vertical (u horizontal si lo prefiere) cuya altura o longitud estará fijada proporcionalmente por la frecuencia absoluta o relativa de cada categoría, respectivamente. Esta forma de representación es particularmente útil cuando se dispone de dos columnas de datos, una categórica por ejemplo la especie vegetal y otra numérica por ejemplo la abundancia de cada especie. Hay literatura estadística que incluso sugiere emplear diagramas de barras antes que diagramas circulares ya que las personas somos capaces de juzgar una longitud más adecuadamente que el volumen o el área de una región.

Para ilustrar el diagrama de barras vamos a utilizar los datos del ejemplo 2.2.1. Las funciones que se utilizan en el programa R para esta gráfica se muestran a continuación.

Podemos generar un diagrama de barras con las frecuencias absolutas (Figura 5A); para ello hacemos uso de la función “barplot”, es decir, digitamos o copiamos en la consola lo siguiente:

```
barplot(t, xlab="Velocidad del viento", ylab="Frecuencia absoluta", cex.lab=1.5, cex.names=1.2, cex.axis=1.2)
```

O también podemos emplear las frecuencias relativas (Figura 5 - B):

```
barplot(t/sum(t), xlab="Velocidad del viento", ylab="Frecuencia relativa",
cex.lab=1.5, cex.names=1.2, cex.axis=1.2, las=1)
```

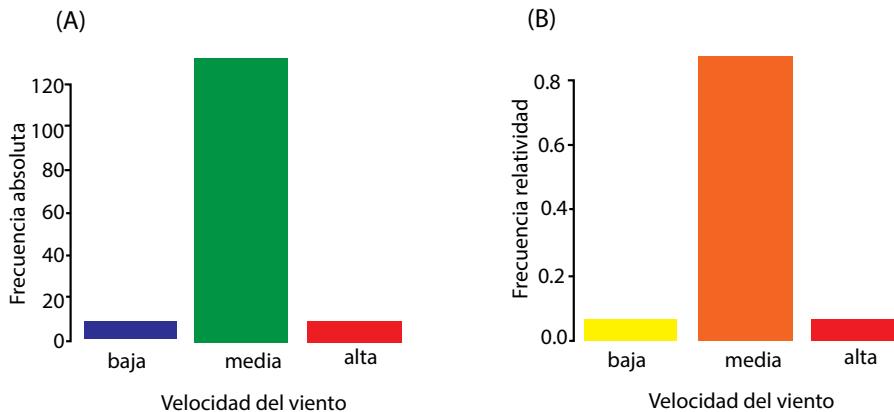


Figura 5. Diagrama de barras para la velocidad del viento categórica. (A) Utilizando frecuencia absoluta, (B) Utilizando frecuencia relativa.

Fuente: Chambers et al. 1983

La interpretación de este diagrama es similar a la del diagrama circular, donde el área de la región en el diagrama circular se corresponde con la altura de la barra de cada categoría respectivamente.

Ejemplo 2.2.2

Otra aplicación del diagrama de barras es la representación de series de tiempo mediante barras; por ejemplo, la velocidad máxima del viento por cada mes observado.

Entonces, para cumplir con nuestro objetivo, es necesario obtener el valor máximo de la velocidad del viento en cada uno de los meses observados.

Obtenemos un vector de velocidades máximas para cada mes, esta operación la realizamos en el programa con ayuda de la función “*tapply*”, digitando lo siguiente:

```
m=tapply(viento, airquality$Month, max)
```

Luego, para ver los resultados digitamos m

Tenga presente que para utilizar la función *tapply*, es necesario ingresar tres elementos: primero una variable numérica, en segundo lugar, una categórica y en tercer lugar la función estadística que desea calcular. En este caso la función estadística “max” nos devuelve el valor máximo.

Luego graficamos los valores utilizando la función *barplot*, así:

```
bp=barplot(m, xlab="Mes", ylab="Velocidad máxima  
(m/h)",
```

```
cex.lab=1.5,cex.names=1.2,cex.axis=1.2)
```

Para facilitar la lectura de la gráfica, se puede incluir los valores de velocidad máxima en cada barra, digitando lo siguiente:

```
text(bp, m, labels = m, pos = 1,cex=1.2)
```

En el diagrama (Figura 6) se observa que los meses 5 y 6 presentaron mayor velocidad del viento alcanzando las 20 millas/hora, mientras que para los tres meses subsiguientes la velocidad máxima desciende hasta 15 millas/hora.

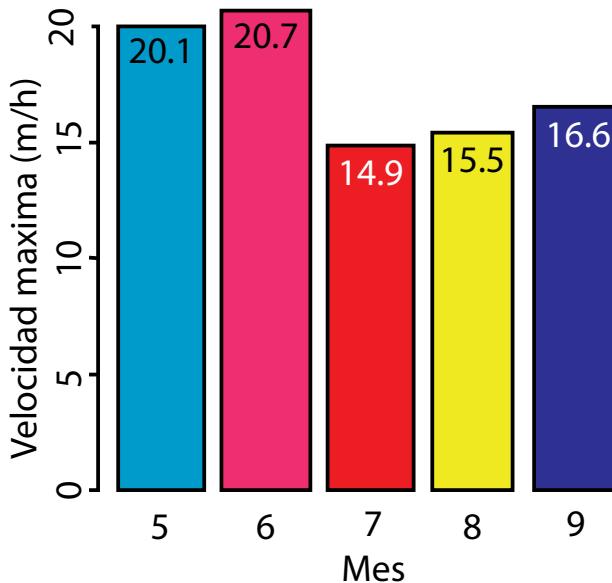


Figura 6. Diagrama de barras para la velocidad máxima del viento por cada mes.

Fuente: Chambers et al. (1983)

Así, el diagrama de barras puede ser utilizado también para representar otros valores estadísticos como el promedio, la mediana, la amplitud, etc. Estas funciones estadísticas las revisaremos en la Unidad 3.



Actividad de aprendizaje recomendada

Estimado estudiante, luego de la revisión de las gráficas de variables categóricas, le sugiero el desarrollo de la siguiente actividad:

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Ejercicio práctico: Agrupar los datos de la variable ozono (Ozone) de la tabla “airquality” que consta en el programa R, en siete clases y reportar los resultados conforme la Tabla 1.

Retroalimentación: mayor información acerca de los datos se puede obtener digitando `help(airquality)` en la consola del programa. Ahí encontrará una descripción de cada variable, se observa que la variable Ozone está dada en partes por billón (ppb). Similar al ejemplo de la temperatura, ahora va a crear una nueva variable llamada ozono, tenga en cuenta que esta nueva variable tiene datos incompletos o faltantes, que aparecen como “NA”, antes de avanzar con los cálculos, deberá omitir los datos faltantes, así:

```
ozono=airquality$Ozone  
ozono=na.omit(ozono)
```

O también, puede utilizar la función “`distrib.frec`” que le permitirá construir de forma rápida y sencilla la tabla de frecuencias.

2.3. Gráficas estadísticas para variables numéricas

Cuando disponemos de variables numéricas el interés fundamental es conocer la *distribución* de los datos. Conocer la distribución nos ayudará a responder preguntas como ¿cuál es la amplitud de los datos?, ¿cuál es la tendencia central?, ¿qué tan dispersos están los valores? En esta sección intentaremos dar una respuesta gráfica a estas interrogantes. Cuando disponemos de un conjunto grande de datos, podemos aprovechar esta información y resumirla de algunas maneras, una de ellas es la representación gráfica que permitirá identificar la forma de distribución de los datos caracterizada por el sesgo o la simetría.

Histograma

Es una representación visual de la distribución de un conjunto de datos, permitiendo al usuario identificar dónde se concentra la

mayor (o menor) cantidad de datos. Consiste de barras verticales paralelas y adyacentes que gráficamente muestran la distribución de frecuencias de una variable cuantitativa. Son útiles básicamente para muestras grandes ($n>30$).

En R la función que permite construir un histograma es `hist(x, breaks=...)`, donde `x` es el vector numérico y `breaks` una secuencia que representa los límites de las clases . Esta función ya la utilizamos en el ejemplo 2.1.1, para construir la tabla de frecuencias. A continuación, un ejemplo detallado sobre la construcción de un histograma.

Ejemplo 2.3.1

Representar gráficamente la distribución de los datos de la variable temperatura (tabla “airquality” en R).

Primero construiremos un histograma sin controlar el número de clases, por defecto el programa va a generar las clases. En R utilizamos la siguiente función (puede copiar y pegar en la consola las dos líneas siguientes):

Para una histograma simple digitamos y ejecutamos:
`hist(temperatura)`

Para un histograma personalizado podemos definir el color (`col`), el tamaño de los número en los ejes (`cex.axis`), una etiqueta para los ejes X e Y (`xlab`, `ylab`), girar 90° los números del eje Y (`las`), controlamos el título de la gráfica (`main`), en este caso no interesa el título. Ejecutamos:

```
hist(temperatura, col="gray", cex.  
axis=1.2, xlab="Temperatura  
(°F)", main="", ylab="Frecuencia", cex.lab=1.5, las=1)
```

Esta función genera el histograma de la figura 7-A.

Luego controlando el número de clases, para obtener una distribución de cinco clases seguimos los siguientes pasos:

Paso 1: Creamos un vector que permita definir las clases.

Necesitamos tres valores: inicial, final y un incremento. Como inicial podemos escoger el mínimo o un valor anterior al mínimo, en este caso 55, como valor final podemos seleccionar uno mayor al máximo, en este caso 100, y finalmente el incremento será el cociente entre la diferencia (final - inicial) y el número de clases $(100-55)/5=9$.

Entonces el vector será:

```
clases=seq(55,100,9)
```

Paso 2: Graficamos el histograma digitando la función siguiente:

```
hist(temperatura,breaks=clases,col="gray",cex.axis=1.2, xlab="Temperatura  
(°F)",main="",ylab="Frecuencia",cex.lab=1.5,las=1)
```

Con esta función obtenemos el histograma de la figura 7-B.

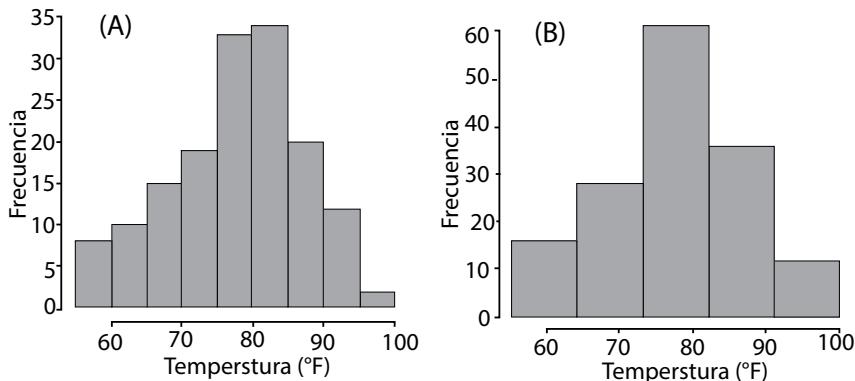


Figura 7. Distribución de la variable temperatura. (A) Histograma con nueve clases, (B) Histograma con cinco clases definidas por el usuario.

Fuente: Chambers et al. 1983

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

La figura 7-A muestra un histograma con nueve clases generadas por defecto en el programa R (algoritmo de Sturges); sin embargo, el usuario puede cambiar el número de clases. Al modificar el número de clases, consecuentemente se modifica la frecuencia de cada clase. Al reducir el número de clases se incrementa el ancho de clase y el valor de la frecuencia (Figura 7-B). Respecto a la forma de la distribución podemos decir que es bastante simétrica, es decir las frecuencias más altas se ubican en la parte central de la distribución y decrecen “simétricamente” en ambos lados de la distribución (Figura 7-B); no obstante, con mayor número de clases el decrecimiento de las frecuencias no es muy simétrico (Figura 7-A) puesto que las frecuencias de las clases de la izquierda decrecen más lentamente que aquellas que se ubican a la derecha de la gráfica.

A manera de observación se puede resaltar que una distribución con mayor número de clases es más susceptible de identificar sesgos o asimetrías. La pregunta que suele presentarse en esta situación es ¿Cuál es el número óptimo de clases?, no hay una regla exacta, sino que la literatura sugiere un número entre 5 y 12 clases, dependiendo del conjunto de datos.

Polígono de frecuencias

En muchos textos de estadística, un polígono de frecuencias se muestra como complemento a un histograma. Para su construcción extraemos el punto medio y la frecuencia de cada clase (altura de cada barra) respectivamente, luego la conexión de estos puntos (vértices) mediante segmentos dará lugar al polígono de frecuencias.

A continuación pongo a su disposición la función `poli.frec`, que le servirá para construir el polígono de frecuencias. Para ello deberá copiar y pegar todo el bloque de la función en la consola del programa R, ejecutar con “enter” y estará lista para utilizarla.

```
### Función poli.freq para graficar
poli.freq<-function(datos, n.clases,
titulo.x){

  datos<-na.omit(datos)

  ac=(max(datos)-min(datos))/n.clases

  clases=seq(min(datos), max(datos),
  ac)

  H=hist(datos, breaks=clases)

  x=H$mids

  y=H$counts

  plot(x, y, type="o", lwd=3,cex.
axis=1.3,las=1,xlab=titulo.x,
ylab="Frecuencia",xlim=c(H$mids[1]
-ac,H$mids[n.clases] + ac),cex.
lab=1.5)

  segments(H$mids[n.clases],H$counts[n.
clases],H$mids[n.clases] + ac,0,
lwd=3)

  segments(H$mids[1]
-ac,0,H$mids[1],H$counts[1], lwd=3)
```

Ejemplo 2.3.2

Construir el polígono de frecuencias para la variable temperatura (tabla “airquality” en el programa R), utilizando 10 clases.

Para construir el polígono, vamos a utilizar la función poli.freq, no olvide que previamente debe copiar y pegar el código de la función poli.freq del recuadro anterior, en la consola de R. Para que la

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

función genere la gráfica, se debe ingresar tres elementos: el primero corresponde al vector de datos (en este caso temperatura), en segundo lugar, el número de clases (en este ejemplo ingresamos 10), y finalmente entre comillas el título que desea que aparezca en el eje X de la gráfica, en este ejemplo escribimos la expresión “Temperatura ($^{\circ}\text{F}$)”.

De esta forma, la función que debe escribir sería:

```
poli.freq(temperatura,10,"Temperatura ( $^{\circ}\text{F}$ )")
```

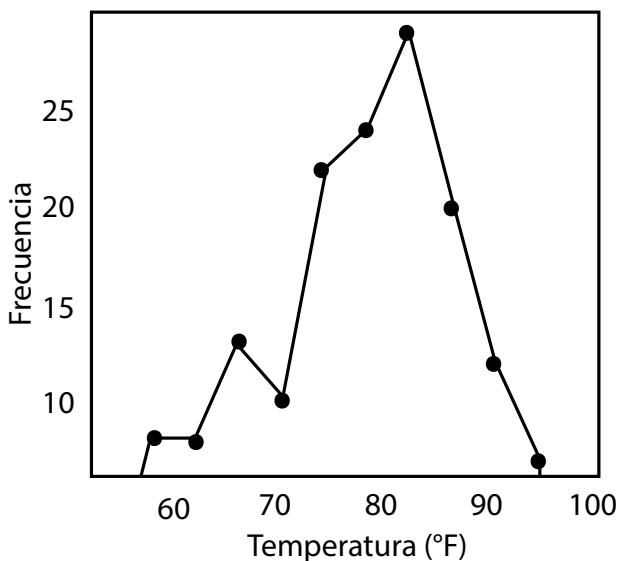


Figura 8. Polígono de frecuencias de la variable temperatura (tabla airquality) con 10 clases.

Fuente: Chambers et al. 1983

El polígono de frecuencias revela con más detalle la simetría (o ausencia de la misma) en una gráfica de distribución de frecuencias; se observa un ligero sesgo negativo (alargamiento a la izquierda) en la distribución de la temperatura (Figura 8).

Sin embargo, hay situaciones en las cuales sería más deseable estimar directamente la densidad, ya que tanto el histograma como el polígono son dependientes del número de clases que se consideren.

Curva de densidad

Un diagrama de densidad, también conocido como “kernel density plot”, se construye a partir de una variable numérica y muestra la distribución suavizada de los puntos a lo largo del eje numérico constituido por la variable de interés. Graficar una curva de densidad implica construir un estimador de la función de densidad de un conjunto de datos observados (Silverman, 1998). Los picos de la curva de densidad son las ubicaciones donde existe la mayor concentración de puntos. Esta gráfica es una variación del histograma y utiliza una técnica estadística llamada “kernel smoothing” para estimar una función de valor real. Una ventaja de esta gráfica sobre el histograma, es que determina con mayor precisión la forma de la distribución, ya que no está afectada por el número de clases. La curva de densidad proporciona información valiosa de características como son el sesgo y multi-modalidad en los datos. A continuación, un ejemplo ilustrativo sobre el proceso para graficar una curva de densidad en el programa R.

Ejemplo 2.3.3

Construir la curva de densidad para la variable temperatura (tabla airquality) sin considerar el mes. Las funciones que se deben escribir y ejecutar en R son:

```
plot(density(temperatura), type="n", main="", xlab=
"Temperatura (°F)",
ylab="Densidad", cex.axis=1.2, cex.lab=1.5)

polygon(density(temperatura, col="gray60", border=NA
```

Luego de ejecutar las funciones anteriores, se obtendrá la figura 9-A.

En el caso que interese hacer la representación mes por mes, podemos construir la densidad de la temperatura por cada mes observado y comparar entre los 5 meses registrados. Las funciones que deberá escribir y ejecutar en R son:

```
plot(density(airquality[airquality$Month==5, 4]),  
col=1, ylim=c(0, 0.12), xlim=c(40, 120),  
xlab="Temperatura  
(°F)", ylab="Densidad", main="", lwd=3, cex.axis=1.2,  
cex.lab=1.5)  
  
lines(density(airquality[airquality$Month==6, 4]),  
col=2, lwd=3)  
  
lines(density(airquality[airquality$Month==7, 4]),  
col=3, lwd=3)  
  
lines(density(airquality[airquality$Month==8, 4]),  
col=4, lwd=3)  
  
lines(density(airquality[airquality$Month==9, 4]),  
col=5, lwd=3)  
  
legend("topleft", legend=c(5:9), lty=1, col=c(1:5),  
lwd=3, bty="n", title="Mes", cex=1.5)
```

Finalmente obtendrá la figura 9-B.

En la densidad de la temperatura global (Figura 9-A) se observa una relativa simetría respecto al punto más alto de la curva, con ligero sesgo (alargamiento) a la izquierda; sin embargo, si se representa la misma variable por mes (Figura 9-B), las densidades muestran formas diferentes, con sesgos más pronunciados en algunos meses (ejemplo meses 7 y 9).

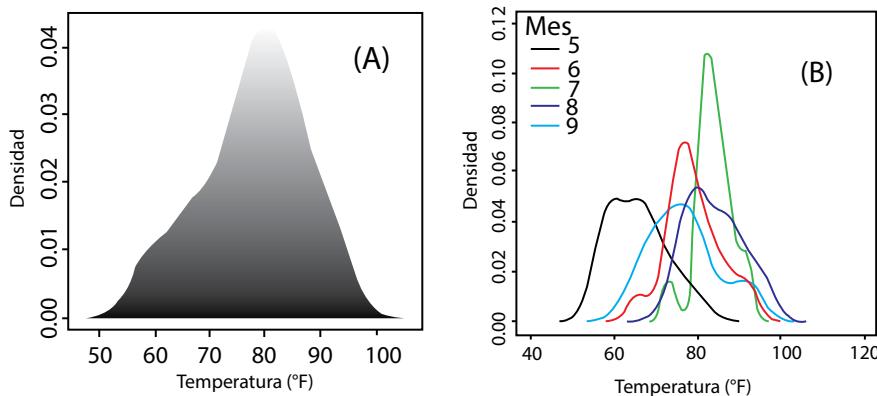


Figura 9. Curva de densidad de la variable temperatura. (A) Sin considerar el mes, (B) Densidades por cada mes observado.

Fuente: Chambers et al. 1983

En general esta forma de representación, es más ajustada a la distribución de los datos y se compara generalmente con una distribución normal (que la revisaremos en las siguientes secciones).

Diagrama de cajas

Previamente hemos revisado técnicas elementales de representación de la distribución de datos (histograma, polígono de frecuencias). En este apartado, presentamos otra importante gráfica estadística llamada diagrama de cajas. Los diagramas de cajas son útiles para identificar valores anómalos (“outliers” en inglés) y para comparar distribuciones. La distribución de los datos se realiza a través de cinco estadísticos resumen: el mínimo, el primer cuartil (Q1), la mediana (Q2), el tercer cuartil (Q3) y el máximo.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Por ahora revisaremos la forma más simple de construir un diagrama de cajas, conformado por un rectángulo central cuya altura denota el rango inter-cuartil (IQR), el segmento dentro del rectángulo muestra la mediana y las líneas (bigotes) arriba y abajo del rectángulo muestran los límites superior e inferior, respectivamente.

Para una adecuada interpretación del diagrama de cajas, es necesario tener algunas consideraciones, por ejemplo, una distribución son sesgo positivo tendría un “bigote” más largo en la dirección positiva que en la negativa. Cuando el valor medio es mayor que la mediana, también habría indicios de sesgo positivo. La presencia de valores atípicos no necesariamente indica que se son “malos” datos u observaciones, de hecho, son muy importantes porque poseen información valiosa del conjunto de datos; no deben ser removidos directamente, sino merecen consideración especial, puesto que podrían contener información clave del fenómeno de estudio.

Ejercicio 2.3.4

Representar la distribución de la velocidad del viento (numérica) mediante diagrama de cajas, utilizando el programa R.

La función que se emplea en R es “boxplot”, y dentro de ella se incluyen algunas instrucciones que permiten dar formato al diagrama. Para construir la gráfica digite y ejecute lo siguiente:

```
boxplot(airquality$Wind, pch=19, cex=2, lwd=2, las=1,  
       ylab="Velocidad del viento (mi/h)", cex.axis=1.3,  
       cex.lab=1.5)
```

Si previamente no ha creado una copia de la variable Wind (conforme se hizo para la temperatura), se puede referir a la variable como airquality\$Wind, esta notación tiene la estructura: nombre-de-la-tabla\$nombre-de-la-variable. El signo de dólar (\$) especifica una variable de entre todas las que conforman la tabla.

Luego de ejecutar la función `boxplot(...)`, se obtendrá la Figura 10, donde se muestran todos los componentes del diagrama de cajas.

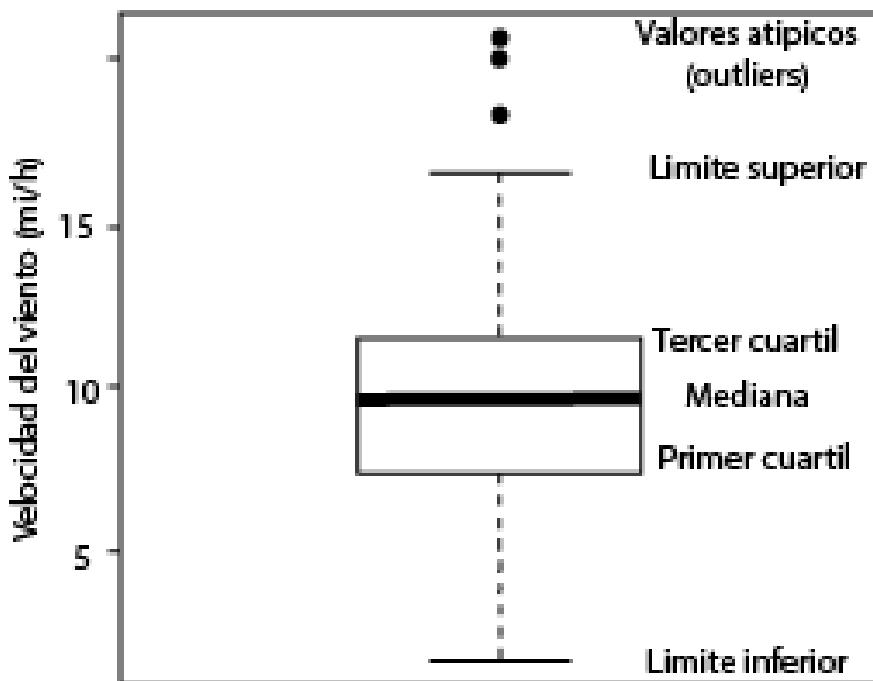


Figura 10. Distribución de la velocidad del viento mediante diagrama de cajas.

Fuente: Chambers et al. 1983

La caja que se ubica aproximadamente en el centro de la gráfica, abarca el 50% de las observaciones, puesto que su amplitud está dada por la distancia entre el primer y tercer cuartil se denomina rango inter-cuartil (IQR).

Por ahora nos limitaremos a decir que el valor central de la velocidad está alrededor de 10 mi/h, y que hay presencia de valores atípicos de la velocidad, ubicados por arriba del límite superior. Estos atípicos corresponden a valores muy altos de la velocidad (días con viento

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

muy fuerte). Otro aspecto importante es que los límites (inferior y superior) no siempre van a coincidir con los valores mínimo y máximo de la variable. El cálculo de los límites se realiza mediante las siguientes relaciones:

$$\text{Límite inferior} = Q1 - (1.5 * \text{IQR})$$

$$\text{Límite superior} = Q3 + (1.5 * \text{IQR})$$

Más adelante, en la sección de estadísticos descriptivos, se detallará la forma de calcular cada uno de los elementos del boxplot, para facilitar la interpretación de la gráfica.



Actividad de aprendizaje recomendada

Para complementar la comprensión de las gráficas numéricas, se recomienda realizar la siguiente actividad:

- Analizar la distribución de datos de la variable ozono (columna Ozone de la tabla “airquality” del programa R) mediante: un histograma de siete clases, un polígono de frecuencias y una curva de densidad.

Retroalimentación: Las tres formas de representación le permitirán identificar la distribución de los datos de ozono, y responder cuestiones como: ¿es simétrica la distribución?, es decir, ¿la moda se ubica en el centro?, si no es simétrica, ¿cuál es la dirección del sesgo? Responder esto le permitirá asimilar la idea intuitiva de la propiedad de normalidad, muy importante en la estadística.



Semana 4

2.4. Gráficas relacionales (bi-variadas)

En la semana 4 continuamos revisando las gráficas estadísticas, ahora nos referimos a las gráficas que incluyen dos variables.

Las gráficas correlacionales se emplean en situaciones donde los datos representan observaciones correspondientes a dos variables o caracteres, cuyas observaciones se han efectuado en los individuos de cierta población (o una parte de la población denominada muestra). La representación conjunta de dos variables nos va a permitir identificar relaciones o asociaciones entre ellas. Si las variables son categóricas o cualitativas, lo más común es emplear diagramas de barras, pero además hay otras opciones como el diagrama de mosaico.

Relación de variables categórica vs categórica

Para ello es necesario partir de una tabla resumen (Tabla 2) de dos variables cualitativas (X, Y), denominada tabla de doble entrada (o tabla de contingencia).

Tabla 2. *Esquema de tabla de doble entrada (tabla de contingencia)*

$X \setminus Y$	B1	B2	...	Bk	Total fila
A_1	n_{11}	n_{12}	...	n_{1k}	TF_1
A_2	n_{21}	n_{22}	...	n_{2k}	TF_2
...
A_l	n_{l1}	n_{l2}	...	n_{lk}	TF_l
Total columna	TC_1	TC_2	...	TC_k	N

Donde A_1, \dots, A_l y B_1, \dots, B_k son las categorías de X e Y respectivamente, N el número total de individuos observados, n_{ik} es la frecuencia absoluta del par (A_i, B_k) de entre los N individuos que poseen la categoría A_i de X y la categoría B_k de Y a la vez.

Diagrama de barras múltiples

Se emplea para representar la distribución cuando ambas variables tienen pocas categorías. Consiste en dibujar para cada par (A_i, B_k) una barra de longitud proporcional a la frecuencia absoluta (o relativa). Las barras se pueden disponer en forma horizontal o vertical. Las categorías de X generalmente se representan en el eje horizontal, y las categorías de Y mediante colores diferentes en las barras, por ello es necesario incluir una leyenda que indique las categorías que representan los distintos colores.

Diagrama de mosaico

Este diagrama es una representación gráfica de una tabla de contingencia o tabla de doble entrada, donde las filas y las columnas representan las dos variables categóricas respectivamente. Esta gráfica permite examinar la relación entre dichas variables. Sobre el eje Y se presentan las categorías (modalidades) de una de las variables, y sobre cada una se levanta un rectángulo con área proporcional a la frecuencia marginal de la categoría. A la vez, cada rectángulo se subdivide en sub-rectángulos de base proporcional a la frecuencia condicionada de las categorías de la otra variable, respectivamente. Por ejemplo, esta gráfica revela independencia de las variables cuando las cajas o rectángulos en todas las categorías tienen áreas similares.

Ejemplo 2.4.1

Representar mediante diagrama de barras y diagrama de mosaico, la relación entre la velocidad del viento (baja, media, alta) y el mes observado.

Para el diagrama de barras, en R seguimos los siguientes pasos:

Paso 1: Definimos los colores, sino, el programa asignará por defecto varias tonalidades de gris, escribimos:

```
colores=heat.colors(3)
```

Paso 2: Construimos la tabla de doble entrada digitando los siguiente:

```
t=table(wind.cat,airquality$Month)
```

Paso 3: Dibujamos el diagrama de barras ejecutando lo siguiente:

```
barplot(t,beside=T,xlab="Mes",ylab="Frecuencia",
        ylim=c(0,40),col=colores,cex.axis=1.3,cex.names=1.3
        ,cex.lab=1.5,las=1)
```

Paso 4: Insertamos una leyenda para indicar lo que representa cada color, escribiendo lo siguiente:

```
legend("top",rownames(t),horiz=T,fill=colores,
      bty="n",cex=1.2,title="Velocidad del viento")
```

Para el diagrama de mosaico, el proceso es más abreviado en R, seguimos los siguientes pasos:

Paso 1: Dibujamos el diagrama escribiendo lo siguiente:

```
mosaicplot(t,cex.axis=1.3,xlab="",main="")
```

Paso 2: Insertamos título a los ejes horizontal y vertical respectivamente. Para ello ejecutamos lo siguiente:

```
mtext("Velocidad del viento",1, line=2,cex=1.5)
```

```
mtext("Mes",2,line=2,cex=1.5)
```

Para conocer más detalles acerca de la función mtext, se le sugiere revisar la ayuda: `help(mtext)` .

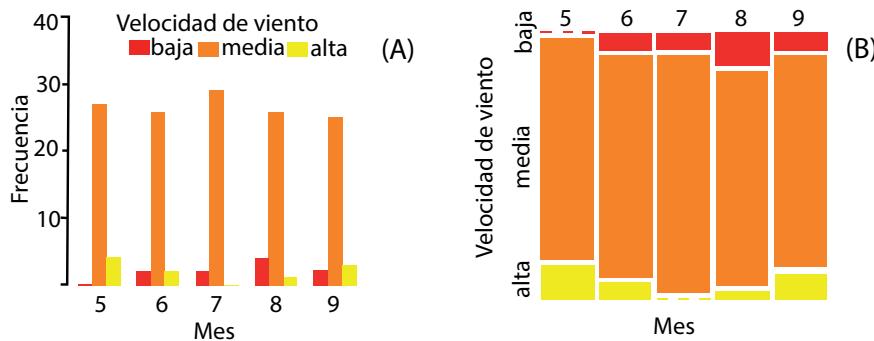


Figura 11. Relación entre la velocidad del viento (baja: $\leq 5\text{m/h}$, media: $5-15\text{m/h}$, alta: $>15\text{m/h}$) y el mes observado. (A) Diagrama de barras. (B) Diagrama de mosaico.

Fuente: Chambers et al. 1983

En la Figura 11 A y B, se observa que prevalece la velocidad media en todos los meses, además en los meses que presentaron mayor frecuencia de velocidad alta (meses 5 y 9) hay poca frecuencia (o ausencia) de velocidad baja y viceversa. Sólo hay tres meses donde la velocidad fue alta, media y baja (meses 6, 8 y 9).

Relación de variables categórica vs numérica

Diagrama de cajas (boxplot)

En la subsección 2.3.4 utilizamos el diagrama de cajas para representar la distribución de una sola variable numérica; ahora emplearemos el mismo diagrama para identificar la relación entre una variable numérica (respuesta) y una variable categórica (explicativa). Hay situaciones donde se habla de relación por ejemplo el cambio de la temperatura entre un mes y otro, durante todo el año; por otro lado, esa relación puede establecer un grado de efecto de la variable explicativa sobre la variable respuesta. Ejemplos: el cambio del área foliar de cierta especie vegetal por efecto de la

altitud; la variación de la presión arterial por efecto del medicamento; la variación de la concentración de metales pesados en un río por efecto de la contaminación de la minería; etc. Para establecer la relación o el efecto, es necesario que la variable explicativa esté conformada por dos o más grupos (categorías o tratamientos). A continuación, se presenta un ejemplo al respecto.

Ejemplo 2.4.2

- A. Con los datos de la tabla “airquality” del programa R, se desea saber cómo cambia la temperatura en cada mes observado.
- B. Con los datos de la “iris” del programa R, se quiere conocer gráficamente la relación entre la longitud del sépalos y la especie.

Visualmente estas relaciones se pueden identificar mediante un diagrama de cajas bi-variado.

Para el literal (A), construimos el diagrama en R utilizando las siguientes funciones:

Paso 1: Creamos una copia de la variable Month y le damos el nombre de mes:

```
mes=airquality$Month
```

Paso 2: Construimos el diagrama de cajas, observe que dentro de la función boxplot(), escribimos primero la variable numérica (en este caso la temperatura) y en segundo lugar la categórica (mes), enlazando las variables con el símbolo “~”, que en el programa R se lee “en función de”, es decir, “temperatura en función del mes”. Escriba y ejecute:

```
boxplot(temperatura~mes, xlab="Mes", ylab="Velocidad  
del viento (mi/h)", cex.axis=1.3, cex.lab=1.5,  
pch=19, cex=2, lwd=2, las=1)
```

Con esto se obtiene la Figura 12–A de la que podemos puntualizar algunas observaciones.

- El mes 5 presentó temperatura más baja, y los meses 7 y 8 la temperatura más alta.
- Los meses 6 y 7 presentan valores anómalos, registros muy bajos de la temperatura. Consecuentemente distribución sesgada de la temperatura.
- El mes 6 muestra mayor variación en la temperatura, esto se puede deducir por la longitud de los bigotes, y la presencia de atípicos.
- La relación no es constante, tampoco directa (lineal creciente) a lo largo de los meses observados, pues la trayectoria que forman las cajas es parabólica.

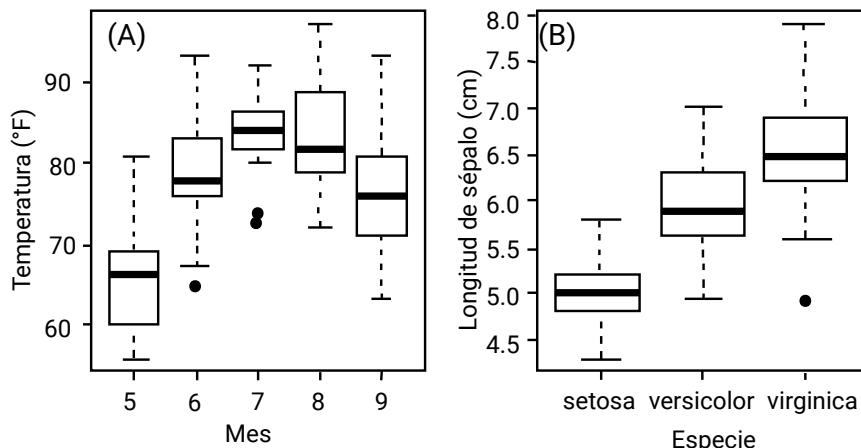


Figura 12. (A) Relación de la temperatura en Nueva York y el mes observado.
 (B) Relación entre la longitud del sépalo y la especie (datos tabla “iris” en R).

Fuente: Chambers et al. (1983)

Para el literal (B), seguimos los siguientes pasos en el programa R.

Paso 1: Creamos una copia de las variables que se requieren para la gráfica, así:

```
SL=iris$Sepal.Length
```

```
especie=iris$Species
```

Paso 2: Construimos el diagrama de cajas, definiendo "LS en función de especie". Escriba y ejecute:

```
boxplot(SL~espceie,xlab="Especie",ylab="Longitud de  
sépalo (cm)",  
cex.axis=1.3, cex.lab=1.5, pch=19, cex=2,  
lwd=2,las=1)
```

En la Figura 12-B, las flores de la especie *virginica* presentaron mayor longitud de sépalo respecto a las especies restantes. También se identificó un valor atípico (valor extremo muy bajo en el grupo) de la longitud en la especie *virginica*. Finalmente, a partir del ancho de caja en cada grupo, la variación de la longitud de sépalo es mayor en las especies *versicolor* y *virginica* respecto de la especie *setosa*.

Relación de variables numérica vs numérica

Diagrama de dispersión (scatter plot)

Es un conjunto de puntos graficados en un plano de coordenadas, donde cada eje del plano representa una variable numérica, por ejemplo, la relación altura vs área de dosel en los árboles de cierta especie. Los diagramas de dispersión son útiles como herramientas de visualización de datos para ilustrar tendencias o identificar posibles asociaciones entre dos variables, donde una de ellas podría ser considerada explicativa (ejemplo la altitud) y otra puede ser considerada variable de respuesta (ejemplo la altura o longitud de las plantas). Cuando la tendencia es creciente en ambos ejes, se habla

de asociación positiva, si es creciente en un eje y decreciente en el otro, entonces la asociación será negativa. En el caso de no existir una tendencia (puntos dispersos aleatoriamente en el plano), las variables no están correlacionadas.

Algunas consideraciones acerca del diagrama de dispersión:

1. Incluso si el diagrama muestra una relación, no se puede asumir que una variable causa a la otra, ambas pueden estar influenciadas por una tercera variable.
2. Cuanto más la formación de los puntos en el diagrama se asemeja a una recta diagonal, más fuerte es la relación.
3. La fuerza de una relación se determina mediante estadísticos de prueba (estas técnicas se revisarán en estadística inferencial).
4. Si el diagrama no muestra relación, puede deberse a que la variable explicativa (X) no cubre un rango suficientemente amplio, incluso considerar si los datos pueden estratificarse.
5. Dibujar un diagrama de dispersión es el primer paso en el análisis relacional entre variables numéricas.

Ejemplo 2.4.3

Mediante diagramas de dispersión, analizar la relación entre la radiación solar y la temperatura, también la relación entre la temperatura y la velocidad del viento. Los datos de estas variables se encuentran en la tabla “airquality” de R.

A. Gráfica de la relación *radiación solar vs temperatura*.

Para construir la gráfica en el programa R, seguimos los siguientes pasos.

Paso 1: creamos una copia de cada variable numérica, así:

```
temperatura=airquality$Temp  
radiacion=airquality$Solar.R  
velocidad=airquality$Wind
```

Paso 2: Dibujamos el diagrama con la función plot, así:

```
plot(radiacion,temperatura,pch=20,cex=2,cex.  
axis=1.3,  
xlab="Radiación (A)",ylab="Temperatura (°F)",cex.  
lab=1.5)
```

Obteniéndose así la Figura 13-A.

B. Gráfica de la relación *velocidad del viento vs temperatura*.

Para construir este diagrama (Figura 13-B), ejecutamos las siguientes líneas:

```
plot(velocidad, temperatura,pch=20,cex=2,cex.  
axis=1.3,  
xlab="Velocidad (mi/h)",ylab="Temperatura  
(°F)",cex.lab=1.5)
```

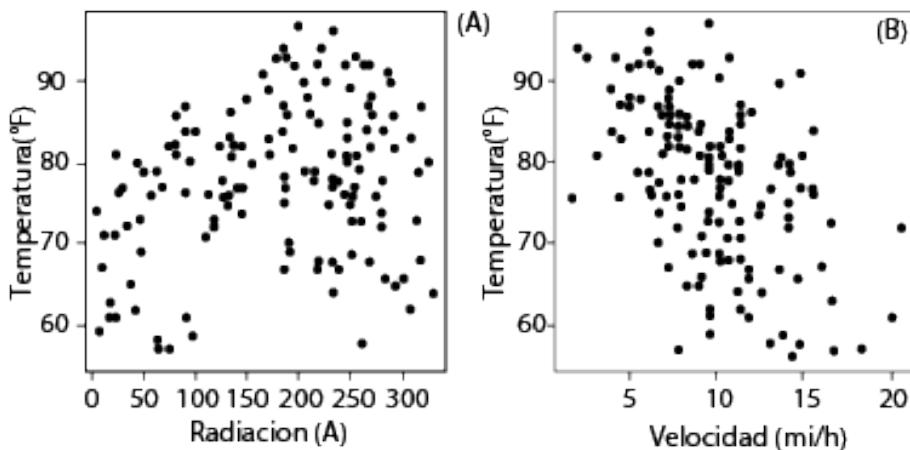


Figura 13. Relaciones bivariadas. (A) Temperatura vs radiación solar. (B) Temperatura vs velocidad del viento.

Fuente: Chambers et al. (1983)

Hay una ligera (leve) relación positiva entre la temperatura del ambiente y la radiación solar (Figura 13-A), sin embargo, si se observa con detalle, la tendencia parece ser cuadrática (forma parabólica); por otro lado, se evidencia una relación negativa (inversa) mediana entre la temperatura y la velocidad del viento (Figura 13-B). Hay una clara diferencia entre los dos diagramas, el segundo (Figura 13-B) muestra una relación más definida que el primero. Esta fuerza de relación puede ser cuantificada mediante un coeficiente de correlación (esta metodología forma parte de la estadística inferencial).

A continuación, les propongo un ejemplo donde el diagrama de dispersión inicial no muestra asociación entre las variables, sin embargo, estratificando las variables se puede identificar relaciones significativas. Para este ejemplo vamos a considerar las variables Longitud de sépalo y ancho de sépalo (Sepal.Length y Sepal.Width) de la tabla iris, y analizar mediante gráfica si presentan indicios de correlación.

Graficamos el diagrama de dispersión:

Paso 1: creamos una copia de las variables para facilitar su manipulación, así:

```
SL=iris$Sepal.Length
```

```
SW=iris$Sepal.Width
```

```
especie=iris$Species
```

Paso 2: construimos los diagramas utilizando la función plot en R:

```
plot(SL,SW, pch=19, cex=1.5, xlab="Longitud de  
sépalo (cm)",  
ylab="Ancho de sépalo (cm)", cex.axis=1.3, cex.  
lab=1.5)
```

También existe la posibilidad de estratificar el diagrama de dispersión en base a la especie, diferenciando con símbolos diferentes las especies. En el programa R se lo realiza así:

Paso 1: Construye el diagrama digitando:

```
plot(SL,SW, pch=as.numeric(especie), cex=1.7,  
xlab="Longitud de sépalo (cm)", ylab="Ancho de  
sépalo (cm)", cex.axis=1.3, cex.lab=1.5)
```

Paso 2: Incluye una leyenda en el diagrama indicando lo que representa cada símbolo:

```
legend("topright",c("setosa","versicolor",  
"virginica"),pch=c(1:3),  
pt.cex=1.3,bty="n",text.font=3, cex=1.2)
```

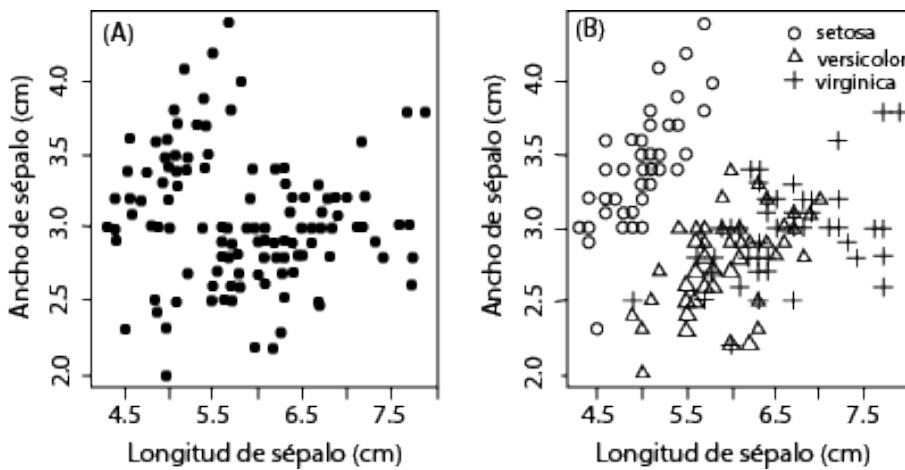


Figura 14. Relación entre la longitud y ancho del sépalo. (A) Diagrama de dispersión inicial global. (B) Diagrama estratificado por la especie.
Fuente: R core team (2017)

En la Figura 14-A no se observa relación bien definida entre la longitud y el ancho del sépalo, los puntos están dispersos en el plano sin evidenciar relación lineal; sin embargo si se considera la especie en la misma gráfica discriminando los puntos con diferente símbolo por cada especie (Figura 14-B), se revela relación directa para las especies *setosa* y *versicolor*, mientras que para la especie *virginica* la relación no es tan lineal puesto que forma una trayectoria ligeramente parabólica con puntos más dispersos.

A partir del último ejemplo podemos subrayar que hay situaciones donde la relación es evidente sin necesidad de estratificar las variables, y en otros casos será conveniente incluir una tercera variable (categórica) para aclarar la relación.



Actividad de aprendizaje recomendada

Luego de haber revisado las tres formas de gráficas correlacionales, se sugiere realizar las siguientes actividades:

- Utilizando la tabla “airquality” (del programa R), crear una variable llamada “temp.cat” a partir de categorizar la variable numérica “temperatura” en tres clases: (0-70], (70-90] y (90, Inf). Luego tabular conjuntamente las variables tem.cat y Month. Con esta tabla realizar diagrama de barras y mosaico. Escribir una interpretación de la gráfica.

Retroalimentación: Este ejercicio busca fortalecer dos aspectos ya revisados: la creación de variables categóricas a partir de numéricas, y luego identificar relaciones entre dos variables numéricas a partir de la gráfica. Recuerde si tiene dos variables categóricas A y B, y quiere generar la tabla (t) de frecuencias cruzadas de ambas variables simultáneamente, debe hacerlo con la función table, así: `t=table(A,B)`.

- Analizar la relación entre la variable Ozono (tabla airquality) y el mes (Month). Realice un diagrama de cajas ¿Qué puede concluir a partir de la gráfica?

Retroalimentación: Tenga presente que para este tipo de gráficas bi-variadas, una variable debe ser numérica y la otra categórica, la numérica es la dependiente (eje Y) y la categórica es la independiente (eje X). Luego, en la gráfica analice el valor central de cada grupo y la variación en cada grupo, reflejada en los anchos de caja, y en base a las diferencias que observe, escriba una interpretación.

- Utilizando los datos de la tabla “trees” del programa R, mediante diagramas de dispersión, analizar la relación entre las variables: Volumen de madera (Volume) vs altura del árbol (Height), y Diámetro del árbol (Girth) vs altura del árbol (Height).

Retroalimentación: Para ver las variables de la tabla “trees” puede digitar y ejecutar la función: str(trees), ahí observará que hay tres variables numéricas, las cuales le servirán para construir los diagramas de dispersión.

Adicionalmente se recomienda responder la autoevaluación de fin de unidad, la misma que le ayudará a sustentar la temática revisada.



Autoevaluación 2

Como una forma de cuantificar el aprendizaje de lo revisado en la Unidad 2, acerca de formas básicas de representación gráfica de datos estadísticos, le propongo las siguientes preguntas.

A. Seleccione y marque el literal que corresponde a la opción correcta.

1. La frecuencia relativa es el cociente entre:
 - a. El total de los datos y la frecuencia absoluta.
 - b. La frecuencia acumulada y la absoluta.
 - c. La frecuencia absoluta y el número total de datos.
2. Un diagrama de barras sirve para representar:
 - a. Una variable categórica.
 - b. Dos variables numéricas.
 - c. Una variable numérica y una categórica.
3. En base a sus características podemos decir que son equivalentes los diagramas:
 - a. Diagrama de barras y diagrama de cajas.
 - b. Diagrama de barras y diagrama circular.
 - c. Diagrama de cajas y diagrama dispersión.
4. La gráfica que permite identificar la variación y el valor central es:
 - a. Histograma.
 - b. Nube de puntos.
 - c. Diagrama de cajas.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

5. Se quiere analizar la relación entre el tipo de bosque y la riqueza de especies presentes, sería adecuado utilizar:
- Diagrama de cajas.
 - Diagrama de dispersión.
 - Diagrama de densidad.
- B. Complete con el término adecuado en cada afirmación de manera que sea correcta.
6. Se denomina _____ a la gráfica que está formada por barras adyacentes y que sirve para representar la distribución de un conjunto de datos numéricos.
7. La altura de las barras en un histograma, representan las _____ de cada clase.
8. Una distribución se dice sesgada a la _____ cuando la gráfica presenta un alargamiento a la derecha y acumulación de datos a la izquierda.
9. La figura que se forma al unir los puntos medios de cada clase en la parte superior de las barras del histograma se denomina _____.
10. Si en el diagrama de dispersión de dos variables numéricas (X, Y), Y aumenta mientras X disminuye, se dice que la relación es _____.

[Ir al solucionario](#)

Luego de haber respondido la autoevaluación, compare con el solucionario que se encuentra al final de la guía didáctica, y realice los correctivos si es necesario.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas



Semana 5



Unidad 3. Estadísticos descriptivos

En la semana cinco se abordan temas relacionados con el cálculo de estadísticos descriptivos tales como medidas centrales, de dispersión y de posición. Para profundizar acerca de conceptos, fórmulas y ejemplos de los estadísticos descriptivos, lea el tema: "Descripción de datos con medidas numéricas", en el texto básico.

En la unidad anterior revisamos el uso de algunas gráficas para resumir y presentar datos estadísticos; sin embargo, en muchas ocasiones resulta muy eficaz condensar dicha información y expresarla mediante indicadores numéricos que también son de fácil interpretación. Estos indicadores comúnmente se los conoce como estadísticos descriptivos.

Analizar datos en el campo de las ciencias biológicas o ambientales, conlleva tratar con información muy variable, por ello es necesario definir tipos de medidas (estadísticos) que sinteticen la información. En la presente unidad revisaremos las medidas más utilizadas en la práctica para resumir datos: medidas de centralización, medidas de dispersión o variación y medidas de posición.

[Índice](#)[Primer bimestre](#)[Segundo bimestre](#)[Solucionario](#)[Referencias bibliográficas](#)

Estimado estudiante, esta información es clave para avanzar en las técnicas de análisis estadístico, por ello le invito a seguir con atención el desarrollo de la unidad tanto en la caracterización teórica como en el proceso de cálculo.

3.1. Medidas de centralización

Estas medidas sirven para resumir un conjunto de datos (o una distribución) y presentar un solo valor que “represente” a dicho conjunto. Las tres medidas de centralización (o tendencia central) más usuales son: la media aritmética, la mediana y la moda (o modo).

Media aritmética

La media aritmética es la medida de tendencia central más usada para resumir una variable numérica. Se obtiene mediante la sumatoria de todos los valores de la variable, y dividiendo por el tamaño de la muestra. La fórmula general para su cálculo es:

Datos no agrupados:

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

Datos agrupados:

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

Donde x_i representa el i-ésimo valor de la variable X y n el total de valores de la variable. Para el caso de datos agrupados, x_i representa el valor medio de la clase (intervalo) y f_i la frecuencia absoluta respectiva.

Como podemos darnos cuenta, al incluir en el cálculo todos los valores de la variable, si tales valores son homogéneos entonces la

media aritmética será un buen resumen de grupo, caso contrario se tornará inestable. Por ello se recomienda tener en cuenta:

- La media aritmética es sensible ante valores extremos o atípicos, básicamente cuando la muestra es pequeña.
- No es recomendable usar la media aritmética como medida central en distribuciones muy asimétricas (sesgadas).
- Cuando la variable es numérica discreta (por ejemplo, número de árboles en un cuadrante), la media aritmética no refleja en valor exacto de la variable, porque se podría obtener valores fraccionarios (por ejemplo, media= 2.5); en tales casos se recomienda redondear al entero inmediato.

Nota: Formalmente hablando, podemos usar el término “promedio muestral”, y el término “media poblacional” (Madsen 2011).

Hay otras medias generalizadas como la media geométrica, geométrica ponderada, la media armónica, la media cuadrática (Manikandan 2011), que serían de utilidad en situaciones donde la media aritmética es inestable. Sin embargo, también podría decirse que las medias generalizadas mencionadas son todas “medias aritméticas en disfraz”, es decir lo único que cambia entre ellas es la transformación de los valores de la variable; así para la media geométrica se emplea el logaritmo de los valores de la variable original, para la media armónica se emplea los valores recíprocos. Por lo mencionado, no siempre será factible el cálculo de las medias generalizadas, por ejemplo, el logaritmo o el recíproco de cero no es posible calcular. La media geométrica es apropiada cuando los valores cambian exponencialmente y en casos donde la distribución original es asimétrica y esa falta de simetría puede corregirse mediante el logaritmo. Estos casos comúnmente ocurren en datos microbiológicos, serológicos o geoquímicos donde es imposible obtener concentraciones negativas de algún elemento químico

(Reimann et al. 2008). O también donde cada observación está representada por un porcentaje creciente respecto de la observación previa. Por otro lado, ante presencia de grandes (o muy pequeños) valores atípicos que causen sesgo en el promedio, es preferible emplear la media armónica.

La mediana

El propósito de la mediana de la muestra (o de la población si es posible) es reflejar la tendencia central de la muestra de manera que no esté afectada por los valores atípicos (outliers).

Para su cálculo partimos de una serie ordenada de datos (x_1, x_2, \dots, x_n) en forma creciente, entonces la mediana muestral será:

Si n es impar

$$Md = x_{\frac{n+1}{2}}$$

Si n es par

$$Md = \frac{1}{2} \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$$

Algunos investigadores sugieren que la mediana es útil por ejemplo cuando se quiere evaluar los recursos completos de un país.

La ventaja fundamental de la mediana es que será menos afectada que la media aritmética en distribuciones sesgadas, es decir es invariante ante la presencia de valores atípicos (extremos).

La moda

Es el valor más probable en una muestra, es decir aquel que ocurre con mayor frecuencia. Corresponde al punto más alto en la gráfica de densidad. Cuando la curva de densidad presenta más de un pico (máximos) entonces se dice que la distribución es polimodal.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

No existe una fórmula simple para estimar el valor de la moda, lo único requerido es una frecuencia de conteo para cada valor del conjunto de datos. Sin embargo, cuando el conjunto de datos está conformado por valores diferentes donde a menudo solo hay una ocurrencia de cada valor, o por casualidad hay dos o tres ocurrencias de los valores, esto puede ser solo una coincidencia estadística, en este caso la moda no es significativa. Por tanto, la moda no es muy utilizada en la práctica (Madsen 2011). La moda, a menudo es estimada a partir de un histograma o una curva de densidad, identificando el valor donde la gráfica presenta un máximo.

Ejemplo 3.1 Utilizando los datos de la tabla “airquality” del programa R, calcular la media aritmética, la mediana y la moda de la variable ozono (Ozone). Representar gráficamente.

Paso 1: creamos una copia de la variable Ozone con el nombre de ozono

```
ozono<-airquality$Ozone
```

Paso 2: calculamos la media aritmética utilizando la función “mean”. Note que además se agregó la instrucción “na.rm=TRUE” con la finalidad que, en caso de existir datos perdidos en la variable, esto sea omitidos para el cálculo de la media.

```
mean(ozono, na.rm=TRUE)
```

Paso 3: calculamos la mediana, digitando:

```
median(ozono, na.rm=TRUE)
```

Paso 4: Calculamos la moda. Una forma de estimar la moda es identificar el valor de la variable con la mayor frecuencia. Entonces, calculamos las frecuencias así:

```
table(ozono)
```

Otra forma es identificar el valor máximo de la curva de densidad:

```
d=density(ozono,na.rm=TRUE)
```

```
d$x [which.max(d$y) ]
```

Los resultados se reportan en la tabla 3.

Tabla 3. *Estadísticos centrales de la variable ozono (ppb)*

Variable	Media aritmética	Mediana	Moda estimada
Ozono	42.12	31.50	20.58

Fuente: Chambers et al. (1983)

Ahora graficamos la variable ozono mediante una curva de densidad y representamos las medidas centrales.

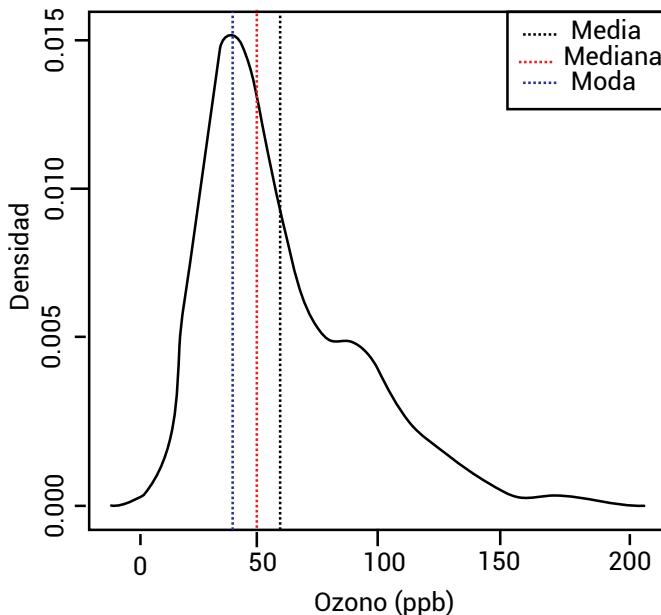


Figura 15. Distribución de la variable ozono. Identificación de los tres estadísticos muestrales: media, mediana y moda.

Fuente: Chambers et al. (1983)

La distribución de la variable ozono es claramente sesgada a la derecha (Figura 14), por ello la moda se ubica a la izquierda y la media aritmética a la derecha. Se observa que ante la falta de simetría los valores de las medidas centrales son diferentes.



Actividad de aprendizaje recomendada

Una vez que ha revisado lo concerniente a las medidas de centralización, se sugiere realizar la siguiente actividad:

- Utilizando los datos de la tabla “mdeaths” del programa R correspondiente al número de muertes por enfermedades pulmonares en UK, calcular las medidas centrales para cada año, además representar gráficamente la serie completa con los seis promedios. Realizar una pequeña interpretación de la gráfica.

Retroalimentación: Si está trabajando en el programa R, puede observar la tabla digitando y ejecutando mdeaths. Si observa, las filas de la tabla corresponden a los años empezando en 1974 hasta 1979. Primero debe crear una copia de la tabla, así: m= mdeaths, luego si quiere extraer los datos del año 1974, lo puede hacer digitando y ejecutando m[1:12], para 1975 sería m[13:24], así sucesivamente para cada año. Para la representación gráfica, seleccione cualquiera de las gráficas descritas en la Unidad 2, que sean relacionadas con variables numéricas. Si quiere identificar valores atípicos, lo más adecuado sería emplear un diagrama de cajas.

3.2. Medidas de variación

En la sección anterior hemos visto algunas medidas centrales para comparar grupos de datos; no obstante, puede haber situaciones donde a pesar de presentar los mismos valores centrales, los datos muestren distribuciones diferentes (Reimann et al. 2008). Una ilustración gráfica de esto se observa en la Figura 15, donde se compara la longitud de los árboles en dos zonas diferentes (tres árboles por zona), la longitud media en ambas zonas es la misma (2.1m), los árboles de la zona 1 parecen tener longitudes similares, mientras que los árboles de la zona 2 presentan longitud muy variable.

Entonces, para mejorar la descripción de un conjunto de datos debemos apoyarnos en medidas de variabilidad. En la literatura estadística podemos encontrar variedad de ellas, pero las más usuales son: el rango (o amplitud), el rango inter-cuartil, la varianza, la desviación estándar, el coeficiente de variación, entre otras; cada una de ellas posee un concepto diferente y aplicación en situaciones puntuales. A continuación, una breve descripción de las medidas de dispersión más utilizadas en la práctica.

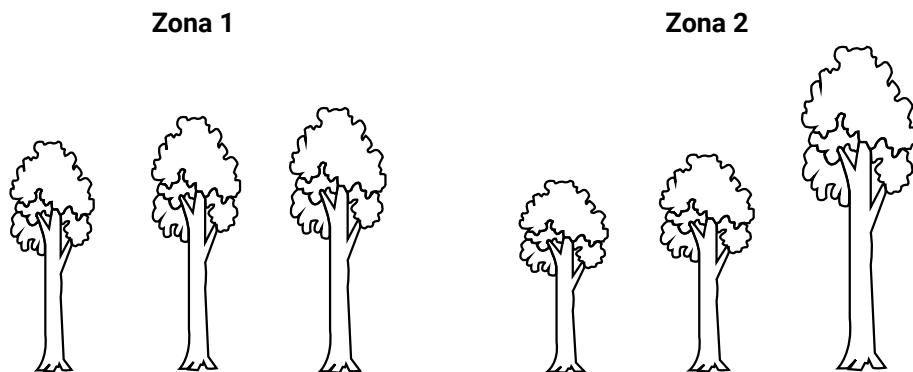


Figura 16. Comparación de la longitud de árboles de cierta especie en dos zonas diferentes.

Fuente: Ramón, P. (2018)

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

El rango (o amplitud)

Es la forma más sencilla de cuantificar la variación de un conjunto de datos numéricos. Se define como la diferencia entre el valor máximo y el valor mínimo, expresando así la amplitud del intervalo de los datos; es fácil de calcular y de comprender. Para muestras pequeñas y tamaño constante, el rango es un buen indicador de la variación (DeCoursey 2003), sin embargo, cuando el tamaño de muestra es diferente y aumenta, también el rango tiende a crecer, en tal caso es recomendable usar otra medida de dispersión. La fórmula de cálculo es la siguiente:

$$R = X_{\max} - X_{\min}$$

Otro inconveniente del rango es que su valor depende únicamente de dos valores muestrales (el máximo y el mínimo) y no hace uso del resto de valores de la muestra. Una aplicación común del rango es en control estadístico de la calidad, para la construcción de diagramas de control con la finalidad de detectar rápidamente cualquier cambio en un proceso de producción (Madsen 2011).

Rango inter-cuartil (IQR)

Una medida más robusta que el rango es el rango inter-cuartil, para su cálculo se determina la diferencia entre el primer y el tercer cuartil. De esta forma podemos decir que el IQR mide el rango del 50% central de los datos. Su cálculo se efectúa mediante la siguiente relación:

$$IQR = Q3 - Q1$$

donde Q1 representa el primer cuartil y Q3 el tercer cuartil. Q1 es un valor que divide los datos en dos partes: 25% de los datos a la izquierda de Q1 y 75% a la derecha; de forma análoga el valor del tercer cuartil Q3.

[Índice](#)[Primer bimestre](#)[Segundo bimestre](#)[Solucionario](#)[Referencias bibliográficas](#)

La fortaleza del IQR está en que es insensible ante valores extremos, ya que considera solamente al 50% de las observaciones centrales, descartando el 25% de los valores superiores e inferiores de la distribución. La representación del IQR se puede ver en un diagrama de cajas (Figura 10 de la sección anterior, exploración de datos) y está dado por el ancho de la caja.

La varianza

Es una de las más importantes medidas de variabilidad y puede referirse a ella como "la media de los cuadrados de las desviaciones de cada observación respecto de la media poblacional (o muestral)". Su fórmula está definida como:

Varianza poblacional:

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

Varianza muestral:

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

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

Nótese que las dos fórmulas presentan dos pequeñas diferencias: (1) para denotar la varianza poblacional (y la media poblacional) se emplea símbolos griegos, (2) el denominador en la varianza muestral es n-1. Para mayores detalles sobre las diferencias entre estas fórmulas se le sugiere realizar la lectura que se indica en las actividades recomendadas, al final de la unidad.

Según las fórmulas anteriores, la varianza tendría unidades cuadradas (metros cuadrados, minutos cuadrados, kilogramos cuadrados, etc.), lo que complica su interpretación y representación, por ello se recomienda utilizar la desviación estándar.

La desviación estándar

Podemos arriesgarnos a decir que es la más importante de las medidas de variabilidad, está definida como la raíz cuadrada de la varianza. Posee las mismas unidades de medida de la variable original y es un valor representativo de las desviaciones respecto de la media aritmética.

Puesto que para su cálculo intervienen todos los valores del conjunto de datos, el problema nuevamente (al igual que la media) es que cada observación tiene el mismo peso. Si hay presencia de valores extremos (atípicos), la desviación estándar será aún más sesgada que la media aritmética (Reimann et al. 2008). Por ello se recomienda que antes de calcular o reportar la desviación estándar como medida de dispersión, revisar la distribución de los datos.

El coeficiente de variación (CV)

El coeficiente de variación, también conocido como desviación estándar relativa o dispersión relativa, es independiente de la magnitud y de la medida de los datos. Usualmente se expresa en porcentaje y es muy adecuado para comparar la variación de los datos expresados en diferentes medidas o en diferentes grupos. Para su cálculo es necesario disponer de la desviación estándar (S) y de la media aritmética () de los datos, conforme se expresa en la ecuación:

$$CV = \left(\frac{S}{\bar{X}} \right) * 100\%$$

Para que el CV tenga significado, su límite inferior debe ser cero; es decir valores negativos no deberían ocurrir (Madsen 2011).

Ejemplo 3.2.1 Para la variable temperatura de la tabla “airquality” del programa R, calcular: la amplitud, el rango inter-cuartil, la varianza y la desviación estándar.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Paso 1: Creamos una copia de la variable temperatura (si antes ya creó esta copia, no es necesario hacerlo de nuevo):

```
temperatura=airquality$Temp
```

Paso 2: Se calcula la amplitud restando del valor máximo, el valor mínimo, así:

```
max(temperatura) - min(temperatura)
```

Amplitud= 41

Paso 3: En el programa R ya existe una función que calcula el rango intercuartil, así:

```
IQR(temperatura)
```

Rango inter-cuartil= 13

Paso 4: Calculamos la varianza digitando:

```
var(temperatura)
```

Varianza= 89.59

Paso 5: Calculamos la desviación estándar como la raíz cuadrada de la varianza. En R se utiliza la función “sqrt” para obtener la raíz cuadrada de un número, así:

```
sqrt(var(temperatura))
```

Otra forma es utilizar directamente la función “sd” definida en R, así:

```
sd(temperatura)
```

Desviación estándar= 9.47

Se observa que la variación máxima de la temperatura dada por la amplitud es de **41°F**, con una desviación estándar de **9.5°F**.

Ejemplo 3.2.2 Para la variable temperatura de la tabla “airquality” (en R) calcular el coeficiente de variación (dispersión relativa) de la temperatura por cada mes observado, e identificar en qué mes varió más la temperatura.

No siempre es conveniente construir medidas de variación para todos los datos de la variable, en algunas ocasiones puede ser más adecuado generar esas medidas condicionando a las categorías de otra variable (categórica), en nuestro caso el mes. Esta forma de análisis servirá para identificar en qué mes cambió más la temperatura.

Para hallar el coeficiente de variación debemos contar con la temperatura media y la desviación estándar por cada mes, y luego efectuar el cociente.

Paso 1: creamos una copia de la variable Month, así:

```
meses=airquality$Month
```

Paso 2: calculamos la temperatura media por cada mes, usando la función “*tapply*” y guardamos los resultados con el nombre m, así:

```
m=tapply(temperatura, meses, mean)
```

Es importante el orden de los argumentos que se escriben dentro de la función *tapply*, así: primero debe ir la variable numérica, luego la variable categórica, y finalmente la operación estadística que se quiere efectuar.

Paso 3: calculamos la desviación estándar por cada mes, de forma similar a la media, dentro de la función *tapply* se reemplaza *mean* por *sd*, y guardamos los resultados con el nombre s, así:

```
s=tapply(temperatura, meses, sd)
```

Paso 4: con los resultados del paso 2 y paso 3, calculamos el coeficiente de variación por cada mes, dividiendo los promedios para las desviaciones estándar, así:

$$(s/m) * 100$$

Los resultados se observan en la tabla 4, la cual muestra que la variación máxima (10.9%) de la temperatura se presentó en el mes de septiembre (9).

Tabla 4. Variación de la temperatura por cada mes observado.

Estadístico	Mes				
	5	6	7	8	9
Media (oF)	65.5	79.1	83.9	83.9	76.9
Desviación estándar (oF)	6.9	6.6	4.3	6.6	8.4
Coef. Var. (%)	10.5	8.3	5.1	7.8	10.9

Fuente: Chambers et al. (1983)



Actividad de aprendizaje recomendada

Lectura: Para aclarar las diferencias entre varianza poblacional y varianza muestral, lea el tema "Medidas de variabilidad", en el texto básico.

Ejercicio práctico: Utilizando los datos de la tabla "iris", determine los estadísticos de variación: rango, IQR, varianza, desviación estándar y coeficiente de variación de las variables longitud de sépalo (Sepal.Length) y ancho de sépalo (Sepal.Width) por separado para cada especie; e identificar en qué especie hay mayor variación.

Retroalimentación: El rango puede ir expresado como un intervalo con los valores mínimo y máximo, o un solo valor como la amplitud (ver ejemplo 3.2.2). Otra forma de calcular el rango inter-cuartil (IQR) es calcular por separado los cuartiles 1 (Q1) y 3 (Q3) y luego establecer la diferencia: $IQR = Q3 - Q1$.



Semana 6

3.3. Medidas de posición (posición relativa)

En secciones anteriores hablamos de la mediana como una medida central y del rango inter-cuartil como medida de variación. Ambas se expresan en términos de medidas de posición denominadas *percentiles*. Es así que la mediana divide los datos en dos partes iguales, es decir la mediana está dada por aquel valor que deja tanto a su derecha como a su izquierda el 50% de las observaciones. Esta propiedad de la mediana nos lleva a generalizar el concepto de medida de posición.

Las medidas de posición relativa indican la ubicación de una observación en comparación con los valores de otras observaciones. Para su descripción nos ayudaremos de los *cuantiles*. Un cuantil de orden P divide los datos en 100P% a su izquierda y 100(1-P)% a su derecha. Aquí el valor de P oscila entre 0 y 1. Así entonces la mediana es el cuantil 0.50.

Para entender mejor este concepto nos referiremos a los *percentiles* que son un caso particular de cuantiles. Un percentil de orden K deja a su izquierda K% de las observaciones y a su derecha el complemento (1-k)%. La escala de un percentil está entre 0 y 100. Un

caso particular de los percentiles son los cuartiles, quintiles y deciles. Estas medidas dividen al conjunto de datos (distribución) en cuatro, cinco y diez partes iguales, respectivamente. De esto podemos concluir que la mediana entonces equivale al percentil de orden 50, al segundo cuartil, al quinto decil.

Para el cálculo de estas medidas de posición (como anteriormente lo hicimos para la mediana) los datos deben estar ordenados en forma ascendente. Goos & Meintrup (2015) proponen los siguientes pasos para el cálculo de los percentiles.

- Ordenar las n observaciones del vector de datos en orden ascendente.
- Calcular la posición del $(100 \cdot P)$ -ésimo percentil como $Q = P \cdot (n+1)$
- Si Q es un entero, entonces el valor $X[Q]$ será el percentil muestral $(100 \cdot P)$.
 - Se determina el valor entero anterior a Q (aquí denominado como A)
 - Determinar la diferencia $F = Q - A$
 - El percentil deseado será:

Para entender mejor, vamos a ejemplificar este procedimiento.

Ejemplo 3.3 Utilizando los datos de la altura (Height) de los árboles de la tabla “trees” del programa R, calcular el percentil 45.

Creamos una copia del vector de alturas

```
alturas=trees$Height  
n=length(alturas)
```

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Calculamos la posición del percentil 45

$$P=0.45$$

$$Q=p \cdot (n+1)$$

$$Q$$

Puesto que Q no es entero, determinamos el entero anterior A y la diferencia F:

$$A=14$$

$$F=0.40$$

Calculamos el percentil 45, para ello ordenamos los datos usando la función sort()

$$C.45 = (1-F) * \text{sort}(alturas)[14] + F * \text{sort}(alturas)[15]$$

$$\text{Respuesta } C.45 = 75.4$$

Este valor nos indica que el 45% de los árboles poseen altura inferior o igual a 75.4 pies, consecuentemente, el 55% posee altura mayor que 75.4 pies.

En el programa R está definida una función para el cálculo de percentiles, denominada quantile(); así para nuestro ejemplo sería:

```
quantile(alturas, probs=c(0.50, 0.75))
```

75.5

El valor obtenido por la función quantile() es muy próximo al valor obtenido por la fórmula.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Con los datos de la altura podemos también calcular los cuartiles 2 y 3, mismos que equivalen a los percentiles 50 y 75 respectivamente. Utilizando la función quantile() sería:

```
quantile(alturas, probs=c(0.50, 0.75))
```

50% 75%

76 80

Cuartil 2 (la mediana) = 76, el 50% de los árboles presentaron altura inferior o igual a 76 pies.

Cuartil 3 = 80, el 75% de los árboles presentaron altura inferior o igual a 80 pies.

Adicionalmente, en R contamos con una función denominada summary(), que nos permite obtener un resumen de los estadísticos descriptivos básicos incluidos los tres cuartiles.

```
summary(alturas)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
63	72	76	76	80	87

La altura mínima es 63 y la máxima 87pies. El 25% de las alturas son inferiores o iguales a 72 pies (cuartil 1), el 50% de las alturas son inferiores o iguales a 76 pies (cuartil 2 o mediana), el 75% de las alturas son inferiores o iguales a 80 pies, y la altura media es 76 pies.



Actividad de aprendizaje recomendada

Ejercicio práctico: Utilizando los datos de la tabla “trees” del programa R, para la variable diámetro (Girth) determine las siguientes medidas de posición: percentil 10, percentil 90, cuartiles 1, 2 y 3. Verifique que efectivamente el cuartil 2 coincide con la mediana.

Retroalimentación: Tenga presente que los cuartiles son casos particulares de percentiles, así, el cuartil 1 será el percentil 25, el cuartil 2 corresponde al percentil 50 y el cuartil 3 será el percentil 75. Si representa por P10 al percentil 10 y así análogamente para el resto de percentiles, usted debe obtener los siguientes valores:

$$P10=10.50, P25=11.05, P50=12.90, P75=15.25, P90=17.90$$

Para completar el ejercicio se recomienda que escriba una corta interpretación.



Autoevaluación 3

Una vez que ha culminado la revisión de los principales estadísticos descriptivos, le recomiendo responder la siguiente autoevaluación conforme se indique en cada enunciado.

A. Escriba en el paréntesis, V si el enunciado es correcto y F si es falso:

1. () La amplitud es medida de centralización.
2. () Se conoce como rango inter-cuartil a la diferencia entre el valor máximo y el mínimo.
3. () La medida central que es insensible ante valores extremos se denomina mediana.
4. () El valor más alto en la curva de densidad corresponde a la varianza.
5. () El coeficiente de variación es una medida de dispersión que se obtiene dividiendo la desviación estándar para la media aritmética.

B. En cada uno de los numerales, seleccione el literal que corresponde a la respuesta correcta.

6. La desviación estándar es:
 - a. El cuadrado de la varianza.
 - b. La raíz cuadrada de la varianza.
 - c. El cuadrado de la amplitud.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

7. Gráficamente en un diagrama de cajas se puede identificar:

- a. La Mediana.
- b. El percentil 10.
- c. La varianza.

8. Cuando la distribución de una variable numérica es simétrica, entonces:

- a. La media y la varianza coinciden.
- b. La media, mediana y moda coinciden.
- c. La mediana es menor que la media aritmética.

9. El percentil 50 también se denomina:

- a. Rango inter-cuartil.
- b. Mediana.
- c. Primer cuartil.

10. El percentil 90:

- a. Deja a su izquierda el 90% de las observaciones.
- b. Está por debajo el 90% de las observaciones.
- c. Deja 10 observaciones a su derecha.

[Ir al solucionario](#)

Luego de haber respondido la autoevaluación, compare con el solucionario que se encuentra al final de la guía didáctica, y realice los correctivos si es necesario.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Con esto hemos concluido la tercera unidad, recuerde que hemos abordado el uso y cálculo de los estadísticos descriptivos básicos, clasificados como: medidas centrales, de variación y de posición. Ahora le invito a que pase a la siguiente unidad, a la vez que le animo a realizar las consultas que considere necesarias ya sea por correo electrónico, por vía telefónica o mediante el Entorno Virtual de Aprendizaje.



Actividades finales del bimestre



Semana 7

Con el propósito de prepararse para el examen presencial, se recomienda que revise los diferentes recursos educativos relacionados con las temáticas de las unidades 1, 2 y 3.

Para aquellos estudiantes que no participaron en la actividad síncrona, evalúe su aprendizaje participando en la actividad suplementaria.



Actividad de aprendizaje recomendada

Luego de haber revisado las unidades 4, 5 y 6 de la guía didáctica, se le propone realizar los siguientes ejercicios prácticos.

- Del texto básico, capítulo 2, resolver los ejercicios suplementarios: 2.54 y 2.65. Para el ejercicio 2.54 intente

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas



Semana 8



Actividad de aprendizaje recomendada

Actividad 2:

Se recomienda que revise los diferentes recursos educativos relacionados con las temáticas de las unidades 1, 2 y 3, texto básico, anexos, REAs y links.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas



Segundo bimestre

Resultado de aprendizaje 1

Es capaz de aplicar los principios de la estadística y análisis de probabilidades.

Contenidos, recursos y actividades de aprendizaje



Semana 9



Unidad 4. Probabilidad

En la semana nueve del segundo bimestre empieza el estudio sobre las probabilidades, un tema muy importante porque permite entender el concepto de probabilidad, sus propiedades y básicamente su relación con las técnicas de estimación de parámetros (intervalos de confianza) que se verá más adelante.

“La probabilidad es lo que usualmente ocurre”. Aristóteles

"Probabilidad es la verdadera guía de la vida". Cicerón

"Todo lo que existe en el universo es fruto del azar". Demócrito

He querido iniciar este nuevo tema citando algunas frases que hacen relación a la probabilidad, pues la probabilidad es un marco referencial (entorno) que nos permite construir enunciados estadísticos y analizar datos (Seefeld & Linder 2007). Por ejemplo, si sembramos diez lotes con la misma cantidad de árboles en cada lote, ¿por qué nunca (o casi nunca) obtenemos la misma tasa de mortalidad en todos los lotes?, más allá de los factores (climáticos, edáficos, etc.) que puedan determinar las causas de la mortalidad de las plantas, es el *azar* el que determina tales variaciones. Entonces, la estadística nos ayudará a determinar cuál es el rango de valores que probablemente se obtengan por azar al medir la ocurrencia de un suceso.

Si los sucesos no cambiaron al azar, serían siempre predecibles y entonces no tendríamos que hacer uso de la estadística. Aquí intervienen las probabilidades como un elemento básico y a la vez fundamental para el desarrollo de las metodologías de análisis estadístico y fundamento de la inferencia estadística. A través del cálculo de las probabilidades se puede determinar la probabilidad que tiene un suceso de ocurrir bajo determinadas condiciones, y su variación debida al azar. A continuación, algunos ejemplos sencillos.

1. La probabilidad de obtener cara en el lanzamiento de una moneda
2. La probabilidad de obtener 6 en el lanzamiento de un dado
3. La probabilidad de obtener número par en el lanzamiento de un dado
4. La probabilidad de encontrar una especie vegetal en peligro de extinción en una reserva natural, etc.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

La condición básica en estos ejemplos está dada por el número de resultados posibles, así en el último ejemplo, la probabilidad estará condicionada a la riqueza de especies presentes en dicha reserva.

La interpretación de la probabilidad se dará en base de su valor numérico, por ejemplo, si buscamos una especie en peligro de extinción, su probabilidad de existencia será bastante baja (quizá 1 en 1000) de que ocurra, entonces es muy improbable que ocurra solo por azar; a diferencia de una especie dominante con alta probabilidad de ocurrencia (quizá mayor a 0.90). Si quisieramos expresar el concepto de probabilidad en una frase, podemos decir que "*la probabilidad es la medida de la incertidumbre*".

El enfoque de la probabilidad que se pretende dar en este capítulo, no es tanto matemático, pues sólo se estudiarán los conceptos más importantes como base para posteriormente abordar los temas de inferencia estadística

4.1. Nociones básicas de probabilidades

Hasta hace pocos años, los textos de estadística presentaban la probabilidad como un fenómeno objetivo que se derivaba de procesos objetivos. Así, la probabilidad objetiva puede dividirse en clásica (a priori) y frecuentista (frecuencia relativa, a posteriori) (Wayne 2002). Las características fundamentales de los diferentes tipos de probabilidad se describen en la tabla 5.

Tabla 5. *Tipos de probabilidad y sus características principales*

Tipos de la probabilidad	Características
Probabilidad clásica	Data del siglo XVII (época de Pascal y Fermat), teoría creada para resolver problemas de juegos de azar (ejemplo el lanzamiento de los dados, las cartas, etc.), no es necesario efectuar el experimento para su cálculo. Así Wayne (2002) propone la definición "Si un evento puede ocurrir de N formas, las cuales se excluyen mutuamente y son igualmente probables, y si m de esos eventos poseen una característica E, entonces la probabilidad de ocurrencia de E se expresa como m/N ".
Probabilidad frecuentista	Depende de la repetición del proceso o experimento, así "si algún proceso se repite un gran número de veces (n), y si algún evento resultante con la característica E, ocurre m veces, la frecuencia relativa de la ocurrencia de E es aproximadamente igual a la probabilidad de E dada por m/n " (Wayne 2002). Al calcular probabilidades con el método de frecuencias relativas, se obtiene un valor estimado en vez de uno exacto. Conforme el número de observaciones se incrementa, el valor aproximado tiende a acercarse al valor real; esta propiedad se conoce como " <i>Ley de los grandes números</i> " (Yañez Canal & Jaimes 2013).
Probabilidad subjetiva (personalista)	Surge a inicios de la década de 1950, y sostiene que la probabilidad mide la confianza que un individuo tiene en la certeza de una determinada proposición. Con base en toda la evidencia que tiene disponible, fundamentado en la intuición, opiniones, creencias y demás información indirecta. Se aplican al cálculo de la probabilidad de sucesos únicos. Hoy en día la teoría bayesiana plantea la solución a un problema estadístico desde el enfoque subjetivo de la probabilidad. El propósito de este curso no incluye al enfoque subjetivo, pues amerita de un alto nivel estadístico para su comprensión y aplicación.

Fuente: Wayne (2002)

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Ejemplo 4.1.1 Para ilustrar la aplicación de la probabilidad frecuentista, y a la vez la propiedad de los grandes números, vamos a simular el lanzamiento de la moneda varias veces, en repetidas ocasiones aumentando el número de lanzamientos cada vez. Luego calculamos la frecuencia relativa (probabilidad) del suceso A = "obtener cara". Finalmente graficamos el experimento.

Este procedimiento podemos realizarlo en el programa R siguiendo los siguientes pasos.

Paso 1: creamos un vector que será el espacio muestral, así:

```
omega=c ("C", "S")
```

Paso 2: creamos un vector donde se almacenarán las frecuencias relativas, para este ejemplo lo llamamos caras:

```
caras=vector()
```

Paso 3: definimos un vector para almacenar el número de lanzamientos:

```
n2=vector()
```

Paso 4: ejecutamos el experimento con 20 ensayos. Cada ensayo tiene diferente número de lanzamientos, por ejemplo, para el ensayo $i=5$, habrá $i^2=5^2=25$ lanzamientos.

```
for(i in 1:20){  
  
  caras[i]<-  
  table(sample(omega,i*i,replace=TRUE)) / (i*i)  
  [1]  
  
  n2[i]<-i*i  
  
}
```

Paso 5: graficamos el experimento, así:

```
plot(n2,caras,type='o',ylim=c(0,1),lwd=2,xaxt="n",
cex.lab=1.5,
cex.axis=1.2,xlab="Nro.
lanzamientos",ylab="Frecuencia relativa
(caras)")

axis(1,at=n2, las=1, cex=.4)

abline(h=0.5,col="red",lty=2,lwd=2)
```

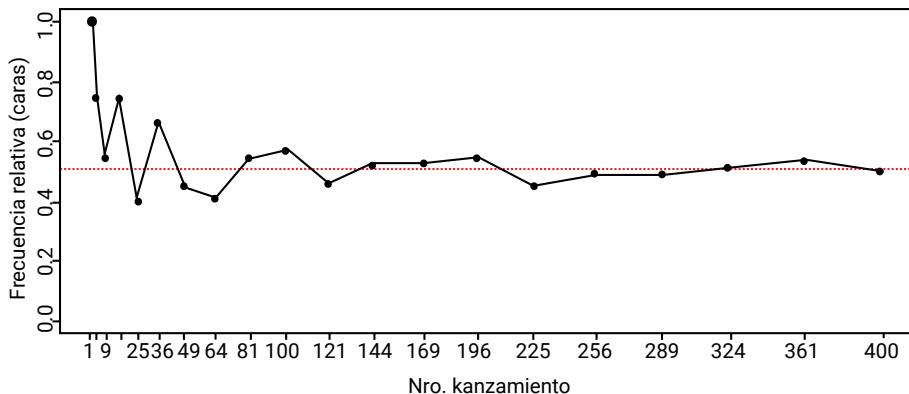


Figura 17. Simulación del experimento “lanzamiento de la moneda”.

El experimento se realizó 20 veces, incrementando el número de lanzamientos en cada experimento. La línea negra representa la frecuencia relativa (probabilidad) observada y la línea roja discontinua representa la probabilidad teórica (0.5) de obtener cara.

Se observa que conforme aumenta el número de lanzamientos, la probabilidad observada tiende hacia la probabilidad teórica (ley de los grandes números. Figura 17).

Para que resulte más sencillo entender este tema relativo a la probabilidad, a continuación, les propongo algunos (de entre muchos)

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

términos y conceptos básicos, necesarios para familiarizarse con el lenguaje probabilístico.

Conceptos previos

Fenómenos aleatorios, características o procesos de interés de los individuos de una población, aquellos cuyos resultados son impredecibles de antemano (Moncho Vasallo 2015). Por ejemplo, el área de dosel de los árboles, la edad, la especie, la respuesta a un tratamiento específico (Ejm: efecto del nitrógeno sobre el incremento diamétrico), etc. Es importante distinguir entre fenómenos aleatorios a partir de estudios observacionales (no experimentales) y experimentales. Los primeros son aquellos que se efectúan sin modificar las condiciones donde se da el fenómeno, y los experimentales son sujeto de cambios, de influencia, de modificaciones por parte del investigador.

Suceso (o evento), es cada uno de los resultados posibles de un fenómeno aleatorio. Por ejemplo, si se trata del fenómeno aleatorio sexo de las personas, los sucesos simples son {masculino, femenino}. Generalmente a los sucesos se los denota por letras mayúsculas.

Suceso A= “obtener número impar en el lanzamiento del dado”

Espacio muestral, es el conjunto de los sucesos simples o elementales. Por ejemplo si se trata del fenómeno (o experimento) aleatorio “lanzamiento del dado”, el espacio muestral es:
 $E=\{1,2,3,4,5,6\}$.

Suceso contrario (complementario) de A, formado por todos los elementos que no están en A, se denota por A' . Por ejemplo:

$A' =$ “Obtener número par”

Suceso unión ($A \cup B$), está formado por todos los elementos que están en A más los elementos que están en B. Por ejemplo:

A= “Especies arbustivas en un ecosistema seco”

B= “Especies herbáceas en un ecosistema seco”

$A \cup B = \{\text{especies arbustivas o herbáceas de un ecosistema seco}\}$

Suceso intersección ($A \cap B$), formado por los elementos que están tanto en A como B. Es decir, si hablamos de un experimento, serían los resultados experimentales que están simultáneamente en A y B. Ejemplo:

A= “Plantas forestales”

B= “Plantas caducifolias”

$A \cap B = \{\text{Plantas forestales y caducifolias}\}$

Lectura recomendada: Para profundizar sobre esta terminología y relaciones sobre la teoría de conjuntos, se recomienda leer el tema relacionado con “Espacios muestrales y eventos”, en el texto básico

4.2. Propiedades operacionales

Para asegurar una noción consistente de que una probabilidad representa el azar de sucesos relacionados a experimentos aleatorios, se emplean reglas (o axiomas).

Sea A cualquier suceso del espacio muestral Ω , entonces decimos que P representa una probabilidad si verifica lo siguiente:

1. Toda probabilidad es no negativa: $P(A) \geq 0$

[Índice](#)[Primer bimestre](#)[Segundo bimestre](#)[Solucionario](#)[Referencias bibliográficas](#)

2. Toda probabilidad se define en el intervalo $[0,1]$: $0 \leq P(A) \leq 1$
3. La probabilidad del espacio muestral: $P(\Omega) = 1$, siempre que Ω esté compuesto de sucesos excluyentes.
4. P es aditiva, es decir, si $\{A_i\}_{i=1,\dots,n}$ son sucesos disjuntos o excluyentes entonces:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

5. $P(\emptyset) = 0$
6. $P(A^c) = 1 - P(A)$
7. Si $A \subset B$ entonces $P(A) \leq P(B)$
8. (**Regla formal de la suma**) Si A y B no son sucesos disjuntos entonces:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Ejercicio 4.2.1 Utilizar los datos de la tabla 6 para el cálculo de probabilidades de un suceso.

Tabla 6. *Número estimado de especies animales endémicas y amenazadas del Ecuador.*

Grupo	Especies Endémicas (E)	Especies Amenazadas (A)	Total
Mamíferos (M)	24	36	60
Aves (V)	38	92	130
Reptiles (R)	121	12	133
Anfibios (N)	163	45	208
Total	346	185	531

Fuente: Sierra (1999)



Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Con estos datos, si se escoge un individuo al azar, calcular las siguientes probabilidades:

- a. (Suceso simple) ¿Cuál es la probabilidad de que sea una especie endémica?

$$P(E) = 346/531 = 0.652$$

- b. (Suceso compuesto: intersección) ¿Cuál es la probabilidad de que sea especie amenazada y anfibio?

$$P(A \text{ y } N) = 45 / 531 = 0.085$$

- c. (Suceso compuesto: unión) ¿Cuál es la probabilidad de que sea reptil o especie endémica?

$$P(R \text{ o } E) = (163+121+12)/531 = 296/531 = 0.557$$



Actividad de aprendizaje recomendada

Para apoyar la adquisición de conocimientos de la temática relacionada a la probabilidad, se recomienda realizar las siguientes actividades:

Lectura recomendada: Todo lo relacionado a las reglas básicas de la probabilidad y sus propiedades operacionales, usted puede encontrar en el apartado “Axiomas, interpretaciones y propiedades de la probabilidad”, en el texto básico.

La idea de realizar esta actividad es que pueda tener una visión más amplia de las probabilidades, sus reglas operacionales, y su relación con la teoría de conjuntos.



Semana 10

4.3. Técnicas de conteo

En la semana diez se amplía el estudio de las probabilidades, que viene a ser un resumen de las operaciones relacionadas con las probabilidades como son las técnicas para realizar permutaciones o combinaciones de eventos o sucesos.

Partiendo del concepto frecuentista de probabilidad, sabemos que una probabilidad se define como:

$P = (\text{número de eventos simples favorables}) / (\text{número total de eventos simples})$

Sin embargo, en la práctica puede resultar tedioso o muy complejo contar el número de eventos simples; para facilitar este cálculo podemos ayudarnos con reglas de conteo.

Regla multiplicativa

Para ilustrar esta regla consideremos k conjuntos de tamaños: n_1, n_2, \dots, n_k . Si un elemento es elegido al azar de cada conjunto, entonces el número total de diferentes resultados es: $(n_1)(n_2)\dots(n_k)$ (Kaps & Lamberson 2004).

Permutaciones

Supongamos que disponemos de un conjunto de n elementos, el número de formas que esos n elementos pueden arreglarse (en diferentes órdenes), se denomina permutación de los n elementos y

[Índice](#)[Primer bimestre](#)[Segundo bimestre](#)[Solucionario](#)[Referencias bibliográficas](#)

se define por: . El símbolo $n!$ se lee como el factorial de n y representa el producto de todos los número naturales de 1 hasta n .

Por ejemplo: A partir de un conjunto de 3 animales $\{x,y,z\}$, de cuántas formas pueden arreglarse en tripletas?

$$P(3) = 3! = 3 \cdot 2 \cdot 1 = 6.$$

Puede formarse tripletas de seis formas. En el entorno R esta operación la realizamos ejecutando:

```
factorial(3)
```

```
[1] 6
```

La respuesta es 6, el corchete [1] indica que el resultado está compuesto de un solo término.

En general, podemos definir permutaciones de un conjunto de n elementos, tomados k a la vez. En este caso el número de permutaciones está dado por la relación:

$$P_{n,k} = \frac{n!}{(n - k)!}$$

Nótese que en este caso, el orden de los subconjuntos es importante. Por ejemplo, de un conjunto de 10 especies arbustivas perennes, cuántos pares de especies se pueden formar considerando que el orden de los pares es importante.

Reemplazando en la fórmula sería:

$$P_{10,2} = \frac{10!}{(10 - 2)!} = \frac{10 \cdot 9 \cdot 8}{8!} = 90$$

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Esto nos dice que podemos formar 90 pares de especies, considerando el orden.

Combinaciones

De un conjunto de n elementos, el número de formas diferentes que los n elementos pueden ser tomados k a la vez, sin importar el orden ($xy = yx$) sería:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\dots(n-k+1)}{k!}$$

Considerando el mismo conjunto anterior, de 10 especies de arbustos, se quiere combinaciones de especies de dos en dos, donde el orden no importa, entonces:

$$\binom{10}{2} = \frac{10!}{2!(8)!} = \frac{90}{2!} = 45$$

En R podemos hacer esta operación utilizando la siguiente función:

```
choose(10,2)
```

```
[1] 45
```

Hay 45 formas diferentes de pares de especies que se pueden combinar.

4.4. Teoremas básicos de la probabilidad

Probabilidad condicional

Intervienen dos sucesos (A,B), donde la probabilidad de ocurrencia del primer suceso (A) está condicionada a la ocurrencia del segundo suceso (B). La fórmula la puede observar en la sección 2.4 (Probabilidad condicional) del texto básico.

Representación gráfica de probabilidad condicional

Para entender mejor el concepto de probabilidad condicional, ilustramos la definición en relación a un diagrama de conjuntos.

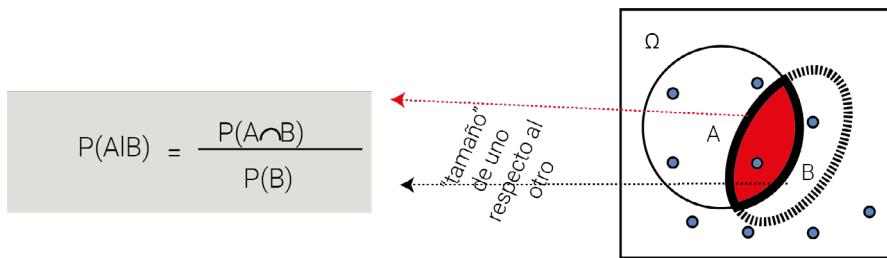


Figura 18. Relación entre la probabilidad condicional y la teoría de conjuntos.
Fuente: Barón López F. J. (2013).

A partir de la Figura 18 podemos destacar algunas observaciones importantes:

1. Para que la probabilidad condicional exista, $P(B)$ debe ser mayor que cero
2. La probabilidad condicional será diferente de cero siempre que los sucesos A y B no sean excluyentes. Es decir, exista al menos un elemento en la intersección.
3. La probabilidad condicional será mayor a medida en que A y B tengan más elementos en común.

Ejemplo 4.4.1 Utilizando los datos de la tabla 4.2.1 de la guía didáctica, calcule las siguientes probabilidades condicionadas:

- ¿Cuál es la probabilidad de que una especie sea endémica (E), dado que pertenece al grupo de las aves (V)?
$$P(E|V) = P(E \cap V) / P(V) = 38/130 = 0.292$$
- ¿Cuál es la probabilidad de que una especie escogida al azar sea anfibio (N) dado que es amenazada (A)?
$$P(N|A) = P(N \cap A) / P(A) = 45/185 = 0.243$$

A partir de la probabilidad condicional, si los sucesos no son independientes, la probabilidad de la intersección de los sucesos, denominada *regla de la multiplicación*, se expresa como:

$$P(A \cap B) = P(B) + P(A|B) = P(A) + P(B|A)$$

esto porque $P(A \cap B) = P(B \cap A)$.

Independencia de sucesos

Se dice que dos sucesos son independientes cuando la ocurrencia de uno no afecta la probabilidad de ocurrencia del otro suceso (Triola 2009). Por ejemplo, consideremos los sucesos A= "Obtener cara en el lanzamiento de una moneda" y B="Obtener número par en el lanzamiento de un dado". Decimos, que A y B son independientes si cumplen:

$$P(A|B) = P(A)$$

De la misma forma, se ha definido la regla de la multiplicación cuando los sucesos son independientes, y se expresa como:

$$P(A \cap B) = P(A) * P(B)$$

Teorema de la probabilidad total

Este teorema parte de la expresión "divide y vencerás". Supongamos que tenemos dos conjuntos: A= "especies forestales", B= "arbustos", mutuamente excluyentes. Un tercer conjunto C= "especies caducifolias".

En este caso, el teorema de la probabilidad total nos servirá para determinar la probabilidad de encontrar especies caducifolias tanto en A como en B.

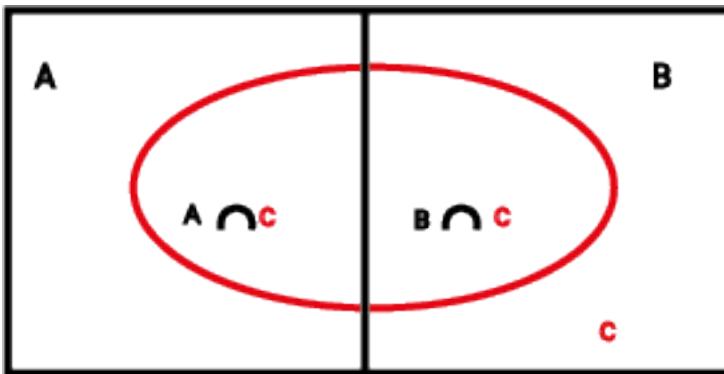


Figura 19. Representación gráfica el teorema de la probabilidad total.

El objetivo es encontrar la probabilidad del suceso C, este suceso se divide en dos sub-regiones:

$$C = (A \cap C) \cup (B \cap C), \text{ entonces:}$$

$$P(C) = P(C \cap A) + P(C \cap B)$$

Luego, aplicando la regla de la multiplicación tenemos:

$$P(C) = P(C|A)P(A) + P(C|B)P(B)$$

Ejemplo 4.4 .1 Utilizando el esquema de la figura 18, supongamos que, en cierto ecosistema, el 70% de las plantas son forestales y de ellas el 20% son caducifolias, mientras que del total de arbustos, el 10% son caducifolias. ¿Cuál es el porcentaje total de caducifolias?

Utilizando el teorema de la probabilidad total tenemos:

$$P(C) = P(C|A)P(A) + P(C|B)P(B)$$

$$P(C) = (0.70)*(0.20) + (0.30)*(0.10)$$

$$P(C) = 0.17$$

De esto se deduce que el 17% de las plantas son caducifolias.

Otra forma de representación es mediante un diagrama de árbol. Para el ejercicio 4.4.1 tenemos el siguiente diagrama:

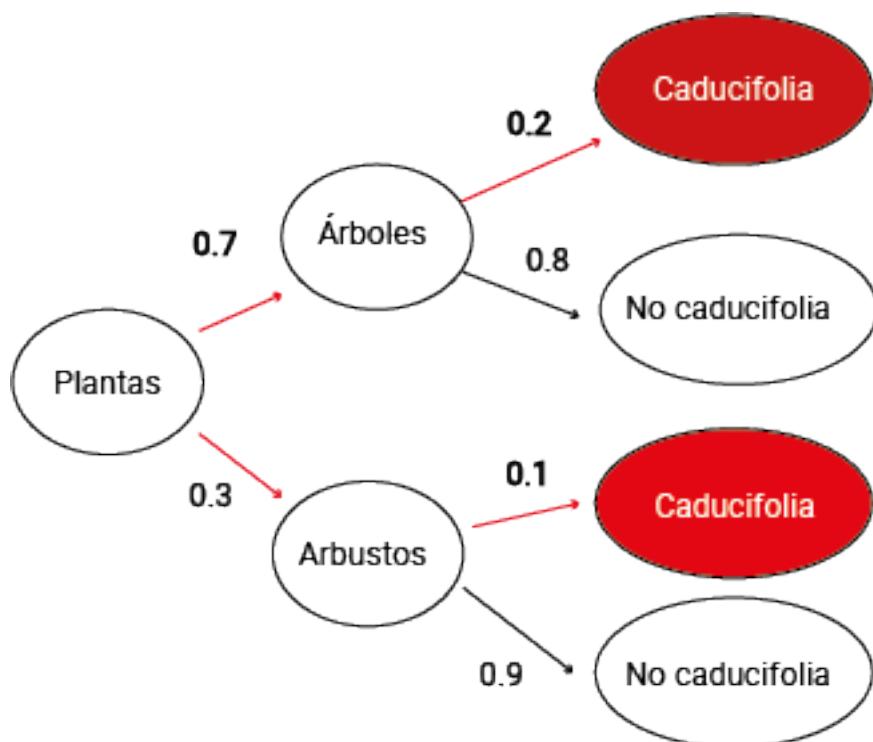


Figura 20. Diagrama de árbol para los datos del ejemplo 4.4.1.

Los caminos a través de los nodos representan intersecciones, y las bifurcaciones representan uniones disjuntas. En la figura 19 observamos dos rutas (ramas del árbol) o dos intersecciones señaladas con flechas de color rojo que conducen al suceso objetivo (C).

Ambas formas funcionan, usted seleccione la que considere más entendible o funcional.



Actividad de aprendizaje recomendada

Para profundizar los conceptos y propiedades de la probabilidad, se sugiere realizar las siguientes revisiones de bibliografía:

Lectura recomendada: Para mayores detalles, se le recomienda leer el tema relativo a “Técnicas de muestreo”, en el texto básico.

Lectura recomendada: Para profundizar sobre este tema de la probabilidad condicional, le sugiero leer el tema Probabilidad condicional, en el texto básico.

Lectura recomendada: Para conocer más acerca de sucesos independientes, le recomiendo leer el tema de Independencia en el texto básico.

Complementariamente se recomienda responder la autoevaluación correspondiente a la unidad 4.



Autoevaluación 4

Una vez que ha culminado la revisión de los fundamentos teóricos y ejemplos de aplicación de las probabilidades, le recomiendo responder la siguiente autoevaluación conforme se indica en cada enunciado.

En cada uno de los enunciados siguientes, complete con el término adecuado de manera que la afirmación sea verdadera.

1. La probabilidad _____, depende de la repetición del experimento.
2. Los estudios que se efectúan sin modificar las condiciones del entorno se denominan _____.
3. La probabilidad del espacio muestral es igual a la unidad siempre que los sucesos sean _____.
4. La probabilidad de un suceso A condicionada a un suceso B, será nula cuando _____.
5. Dos sucesos son independientes cuando la probabilidad de la intersección se expresa como _____ de las probabilidades de cada suceso.

En los siguientes ítems, seleccione y encierre el literal que corresponde a la respuesta correcta.

6. La expresión divide y vencerás se relaciona con:
 - a. La probabilidad de la unión de sucesos.
 - b. El valor de una probabilidad.
 - c. El teorema de la probabilidad total.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

7. Por sus propiedades, las probabilidades se relacionan con:
 - a. La teoría de conjuntos.
 - b. La Física.
 - c. La Geometría.
8. Si la probabilidad de identificar una especie arbórea introducida en un bosque protector es igual a X, entonces la probabilidad de no encontrar dicha especie será igual a:
 - a. 1.
 - b. $1-X$.
 - c. $X-1$.
9. ¿Cuándo dos sucesos M y N son mutuamente excluyentes?
 - a. $P(M \cup N) = P(M) - P(N)$
 - b. $P(M \cup N) = P(M) + P(N)$
 - c. $P(M \cup N) = 1 - (P(M) + P(N))$
10. Un conjunto está formado de 4 elementos, ¿cuántos arreglos de dos en dos, sin importar el orden, serían?:
 - a. 6
 - b. 8
 - c. 12

[Ir al solucionario](#)

Luego de haber respondido la autoevaluación, compare con el solucionario que se encuentra al final de la guía didáctica, y realice los correctivos si es necesario.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas



Semana 11



Unidad 5. Distribuciones de variables aleatorias (discretas y continuas)

El tema a estudiar en la semana 11 es muy importante porque trata sobre los dos tipos fundamentales de variables aleatorias (discretas y continuas) y cómo estas intervienen en problemas cotidianos.

Esta temática permite al lector, formular y resolver modelos probabilísticos sencillos, ya sea de forma analítica o con la ayuda de un software estadístico.

Los recursos de aprendizaje para el estudio del tema son:

Devoré, J. (2016). Probabilidad y estadística para ingeniería y ciencias. 9^a edición. México: CENGAGE Learning.

Para establecer con claridad la diferencia entre los dos tipos de variables aleatorias, revisar el tema Variables aleatorias y Distribuciones de probabilidad para variables aleatorias discretas.

Para conocer mayores detalles sobre el concepto y propiedades del valor esperado y la varianza de una variable aleatoria discreta, se le recomienda revisar el tema Valores esperados, en la sección 3.3 del texto básico.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Una vez que haya revisado el contenido indicado en el texto básico, se espera que el estudiante identifique la función de probabilidad de masa y la función de distribución acumulada, además conozca en detalle los parámetros de la distribución binomial.

En esta unidad revisaremos aspectos relacionados con las variables aleatorias (v.a.), muy útiles para cuantificar la incertidumbre y resumir resultados de experimentos o fenómenos. Un fenómeno puede ser determinista o aleatorio, será determinista cuando podemos predecir sus resultados, mientras que es aleatorio cuando no existe una certeza a priori sobre las observaciones que ocurrirán (Barragués et al. 2014). En este curso nos concierne el análisis de información procedente de fenómenos aleatorios, para lo cual haremos uso de modelos probabilísticos que permitirán describir la regularidad de las variables a ser analizadas.

5.1. Variables aleatorias y distribuciones de probabilidad

Con frecuencia los modelos utilizados en estadística incorporan al menos un elemento de tipo probabilístico, por tanto, para su formulación es necesario tener presente los conceptos de probabilidad y variable aleatoria. Los fundamentos básicos de probabilidad ya los revisamos en la sección precedente, entonces ahora nos concentraremos en lo que se refiere a las variables aleatorias. Según su naturaleza, las variables aleatorias pueden ser discretas o continuas. Las discretas toman valores enteros incluido el cero, mientras que las continuas pueden tomar asumir cualquier valor real. Para facilitar la comprensión de las variables aleatorias, les propongo algunos ejemplos en la Tabla 7.

Tabla 7. *Ejemplos de experimentos, variables aleatorias (discreta y continua) y valores de la variable aleatoria.*

Variables aleatorias discretas		
Experimento	Variable aleatoria (v.a.)	Valores posibles de la v.a.
Seleccionar 5 empleados al azar en una fábrica	Nro. de empleados satisfechos con el ambiente laboral	0,1,2,3,4,5
Inspeccionar un curso de 40 estudiantes	Cantidad de estudiantes que han sufrido accidentes el último mes	0,1,2,...,40
Visita a un parque recreacional un fin de semana	Cantidad de personas	0,1,2,3.....
Analizar la calidad del agua mediante pruebas de laboratorio	Resultado de la prueba	0: no-contaminada 1: contaminada
Variables aleatorias continuas		
Atención en oficina turística	Tiempo en minutos, entre las llegadas de turistas	$X \geq 0$
Llenar una lata de bebida (máx =12.1 onzas)	Cantidad de onzas de la bebida	$0 \leq x \leq 12.1$
Proyecto para construir un centro comercial	Porcentaje de avance del proyecto	$0 \leq x \leq 100$
Ensayar un nuevo proceso químico	Temperatura cuando se lleva a cabo la reacción deseada (min 150° F; máx 212° F)	$150 \leq x \leq 212$

Para resolver problemas relacionados con las variables aleatorias, hay tres elementos que es necesario tener en cuenta: el experimento, la variable y los resultados de la variable (Tabla 7). La diferencia fundamental entre las variables discreta y continua radica en el “dominio de la variable”, aquellos valores posibles que puede asumir la variable aleatoria. Para el caso discreto, hay situaciones en que

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

la variable puede tomar únicamente dos valores (0,1) ausencia/presencia, y en la mayoría de los casos el dominio está conformado por secuencias de valores enteros que inician en cero (ausencia de la variable); mientras que, en el caso continuo, los posibles valores de la variable están constituidos por intervalos en los números reales.

En términos un poco formales podemos decir que una variable aleatoria X es una función que asocia a cada elemento (suceso) del espacio muestral Ω de un experimento aleatorio, un valor numérico real.

Función de probabilidad (F), es generada a partir de la variable aleatoria que puede ser discreta o continua. Es una relación entre el conjunto de los números reales y el intervalo unitario $[0,1]$; es decir, toma un número real (un valor que puede asumir la v.a.) y lo convierte en un valor numérico en el intervalo $0 \leq x \leq 1$, verificando la relación: $f(x) = P(X = x)$, para cualquier número real.

Interpretación: F es la probabilidad de que la v.a. X tome un valor exactamente igual a x.

Función de distribución (F) de una variable aleatoria X, es una función F que va del conjunto de los reales al conjunto unitario $[0,1]$, similar a la función de probabilidad, pero con la diferencia que ahora debe cumplirse la relación: $F(x) = P(X \leq x)$, para cualquier número real.

Interpretación: F es la probabilidad de que la v.a. X tome valores menores o iguales a x.

Se puede observar la diferencia que para la función de probabilidad (f) hay una relación de igualdad, mientras que para la función de distribución (F) hay una relación de orden, por ello a esta última también se la denomina función de distribución acumulada.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Ejemplo 5.1.1: Construir las funciones de probabilidad y de distribución acumulada para el experimento de lanzar la moneda dos veces, donde la v.a. está dada por X = “número de caras observadas”.

Empezamos definiendo el espacio muestral: $\Omega = \{SS, CS, SC, CC\}$

Luego, a partir del espacio muestral cuantificamos los valores (frecuencias) que la variable aleatoria X aparece en cada uno de los cuatro resultados de Ω . Estos valores son $\{0,1,1,2\}$. Entonces el conjunto de valores posibles de X es $\{0, 1, 2\}$.

Posteriormente, a partir de los valores de X , generamos las probabilidades:

$P(X=0)=1/4$, $P(X=1)=2/4$, $P(X=2)=1/4$, donde 4 es el número de resultados en Ω .

Finalmente, con la ayuda del programa R construimos la distribución de probabilidades:

Paso 1. Definimos un vector de valores de la variable aleatoria

```
x=c(0,1,2)
```

Paso 2. Definimos un vector de probabilidades

```
f=c(0.25,0.5,0.25)
```

Paso 3. Graficamos el diagrama de probabilidades así:

```
plot(x,f,ylab="Probabilidad",xlab="X", type='h',
xaxt="n", ylim=c(0,1), cex.lab=1.5, cex.
axis=1.3, lwd=3)

axis(1, at=c(0,1,2),cex.axis=1.3)
```

Paso 4. Definimos un vector de probabilidades acumuladas y graficamos el respectivo diagrama:

```
fa=c(0.25, 0.75, 1)
```

```
plot(x, fa, ylab="Probabilidad acumulada", xlab="X",
type='s', xaxt="n", ylim=c(0,1), cex.lab=1.5, cex.
axis=1.3, lwd=3)

axis(1, at=c(0,1,2),cex.axis=1.3)
```

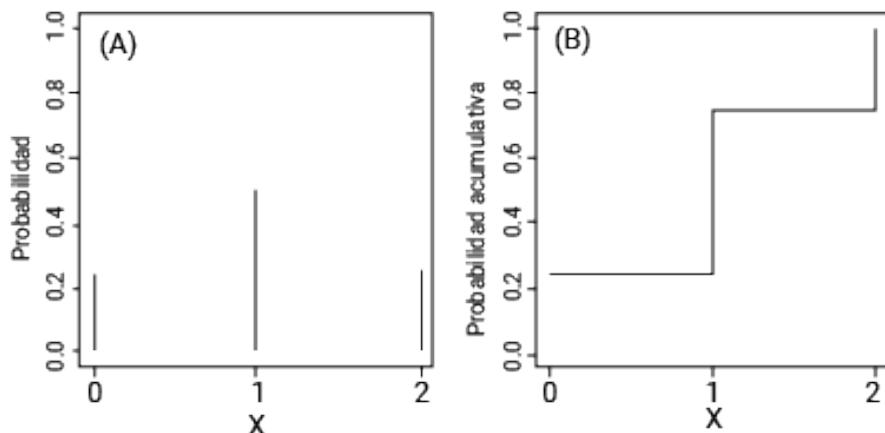


Figura 21. (A) Función de probabilidad, (B) Función de distribución acumulada

La recta vertical con mayor altura (Figura 21-A) representa el valor esperado de la variable aleatoria X, es decir aquel valor de la variable con mayor probabilidad de ocurrencia: $P(X=1)=0.5$. La figura 20-B representa la distribución de probabilidad acumulada, con tendencia siempre creciente y alcanza el máximo valor de la probabilidad ($P=1$). Por ejemplo si nos fijamos en el valor máximo que alcanza la gráfica para $x=1$ (Figura 20-B) es aproximadamente 0.80 en el eje Y, este valor representa la probabilidad $P(X \leq 1)$, es decir la suma de las probabilidades: $P(X=0) + P(X=1)$.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Complementariamente a la distribución de probabilidades de una variable aleatoria, podemos calcular el valor esperado (aquel valor de la v.a. donde está centrada la distribución de probabilidad) y la desviación estándar como una medida de dispersión de las probabilidades que servirá para cuantificar la variabilidad en X.

Ejemplo 5.1.2: Con los datos del ejemplo 5.1.1 determinar el valor esperado y la desviación estándar de la variable aleatoria, y realizar la respectiva interpretación.

Para calcular el valor esperado y la desviación estándar de una v.a. discreta en el programa R, digitamos y ejecutamos los siguiente:

```
E=sum (x*f)
```

```
V=sum ( (x-E)^2*f)
```

```
SD=sqrt (V)
```

Luego, para ver los resultados digitamos: E;V;SD

Los resultados que se obtienen respectivamente son:

```
[1] 1
```

```
[1] 0.5
```

```
[1] 0.7071068
```

El valor esperado es igual a $x=1$, este valor representa el número de caras con más posibilidad de ocurrencia; por otro lado, la desviación estándar es 0.707, pero al tratarse de una v.a. discreta los valores deben ser enteros, entonces para facilitar la interpretación de la desviación estándar podemos redondear a 1, esto nos dice que en promedio los valores de la variable aleatoria distan del valor esperado en una unidad.

5.2. Distribución Binomial

Como dato histórico, el cálculo de las probabilidades tuvo notable desarrollo con el trabajo del matemático suizo Jacob Bernoulli (1654-1705), el cual definió el proceso conocido por su nombre “*ensayo de Bernoulli*”, estableciendo las bases para el desarrollo y utilización de la distribución binomial. Una variable aleatoria de Bernoulli surge de un experimento donde hay únicamente dos alternativas de respuesta, generalmente denotados como “éxito” y “fracaso”. Para los resultados de éxito, la variable aleatoria asume el valor de 1, y para resultados de fracaso, el valor de 0. La probabilidad de éxito se denota por “p” y la de fracaso “q”, donde $q = 1 - p$. La distribución de una v.a. de Bernoulli puede ser descrita como:

$$p(x) = p^x (1-p)^{1-x}$$

Algunos ejemplos que hacen referencia a esta distribución: El nacimiento de un bebé, el resultado puede ser niño/niña; durante una epidemia una persona puede ser catalogada como enferma/sana, una pregunta dicotómica en un examen objetivo puede ser Verdadera/Falsa, al experimentar un nuevo tratamiento se obtiene un resultado que puede ser éxito/fracaso, una muestra de agua puede estar contaminada/no-contaminada, etc.

Si se repite un experimento de Bernoulli n veces, esto dará lugar a la distribución binomial, donde los ensayos de Bernoulli son mutuamente excluyentes. Existen dos parámetros que caracterizan a la distribución binomial, la cantidad de pruebas o ensayos (n) y la probabilidad de éxito (p). Por esta razón una v.a. X de tipo binomial se denota como $X \sim B(n,p)$. Una forma de escribir la ecuación matemática de la distribución binomial es:

$$P(x = K) = \binom{n}{k} p^k (1-p)^{n-k}$$

Donde k es el número de aciertos, n el número de ensayos o resultados del experimento, p la probabilidad de éxito y 1-p la probabilidad de fracaso.

Como se mencionó en la sección anterior, complementariamente a los valores de probabilidad, es necesario conocer el valor esperado de una variable aleatoria binomial X (número promedio de 'éxitos' en muestras repetidas de tamaño n), cuyo resultado se obtiene por: $E[X] = np$, y la varianza de X se determina por: $\text{Var}[X] = npq$. De esto se puede deducir que la varianza asume el mayor valor cuando la probabilidad de éxito es $p=1/2$.

Ejemplo 5.2.1: Consideremos el caso de tener un frasco con plántulas in vitro, donde la variable de Bernoulli representa si el frasco está contaminado. Asumamos que la probabilidad de contaminación de un frasco en cultivo de plántulas in vitro es del 8% (valor que dependerá de la especie y las condiciones de laboratorio), este valor será la probabilidad de éxito.

Sea X= "Obtener frasco contaminado" (éxito)

Tabla 8. *Probabilidades de éxito/fracaso en el experimento de "cultivo de plántulas in vitro"*

Resultado	X = x	P(X = x)
Contaminado	1	p = 0.08
No-contaminado	0	q = 0.92

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

A partir de este problema podemos plantear varias interrogantes como:

- a. ¿Cuál es la probabilidad que, en un lote de 20 frascos, se encuentren exactamente 5 frascos contaminados?
- b. ¿Cuál es la probabilidad que menos de 5 frascos escogidos al azar estén contaminados?
- c. ¿Cuál es la probabilidad que más de 5 frascos escogidos al azar estén contaminados?
- d. ¿Cuál es la probabilidad que a lo mucho 5 frascos escogidos al azar estén contaminados?
- e. ¿Cuál es la probabilidad que entre 4 y 6 frascos (inclusive) escogidos al azar estén contaminados?

Cada ensayo es independiente, el resultado de haber obtenido un frasco contaminado en el primer ensayo, no tiene influencia estadística sobre el segundo frasco contaminado.

A continuación, vamos a dar respuesta a las interrogantes planteadas.

- a. Para responder la primera interrogante lo único que hacemos es reemplazar los valores de $n=20$, $k=5$ y $p=0.08$ en la ecuación de la distribución binomial:

$$P(x=5) = \binom{20}{5} (0.008)^5 (1 - 0.08)^{20-5} = \frac{20!}{5!15!} 9.381393e-7 = 0.0145$$

Este proceso lo podemos abreviar utilizando la función “dbinom” del programa R, digitando lo siguiente:

```
dbinom(5, 20, 0.08)
```

Y obtenemos el resultado:

```
[1] 0.01454491
```

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Por tanto, existe una probabilidad de 0.0145 de encontrar 5 frascos contaminados en un lote de 20 frascos.

b. $P(x < 5) = P(x=0)+P(x=1)+ P(x=2)+P(x=3)+ P(x=4)$

Resolver esta ecuación implica reemplazar cada uno de los cinco términos en la ecuación de la ley binomial, y finalmente sumar cada respuesta. Esta operación la realizamos en el programa R utilizando la función “`pbinom`”, digitando así:

```
pbinom(4, 20, 0.08)
```

```
[1] 0.9816556
```

De esta forma se observa que hay alta probabilidad (0.98) de que menos de 5 frascos estén contaminados.

c. $P(X > 5) = 1 - P(X \leq 5)$

En R digitamos y ejecutamos:

```
1- pbinom(5, 20, 0.08)
```

```
[1] 0.003799487
```

La probabilidad de identificar más de 5 frascos contaminados es de 0.0038

d. $P(x \leq 5) = P(x=0)+P(x=1)+ P(x=2)+P(x=3)+ P(x=4) + P(x=5)$

Similar a la pregunta (b), pero en este caso se incluye el valor $x=5$ (a lo mucho 5 frascos, es decir hasta cinco frascos o menos).

```
pbinom(5, 20, 0.08)
```

```
[1] 0.9962005
```

e. $P(4 \leq x \leq 6) = P(X=4)+P(X=5)+P(X=6)$

La condición inclusive quiere decir que los extremos 4 y 6 sí se incluye en el intervalo. Usando el programa R, esta operación se realiza de la siguiente forma:

```
sum(dbinom(4:6, 20, 0.08))
```

```
[1] 0.06997763
```

De esta forma se obtiene una probabilidad de 0.07 aproximadamente, de identificar entre 4 y 6 frascos contaminados en el lote de 20.

Complementariamente a las preguntas que acabamos de responder, podemos graficar la distribución de probabilidades de la v.a. X, asociada al experimento completo para un lote de 20 frascos observados. Esto lo realizamos de la siguiente forma:

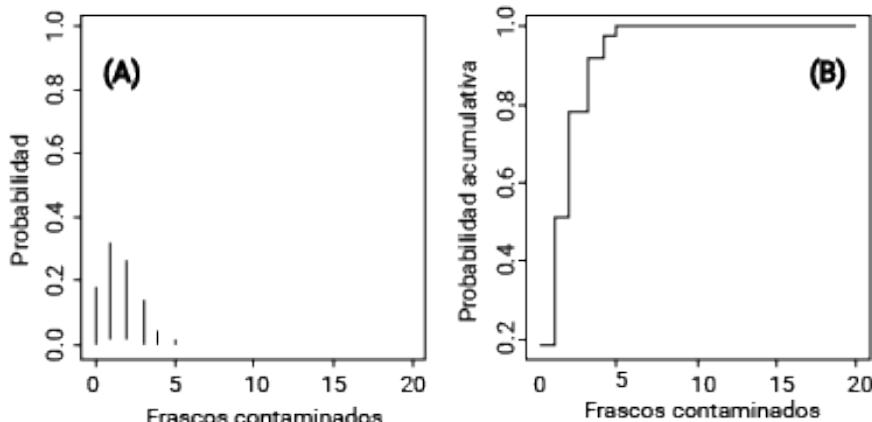


Figura 22. Distribución de probabilidades de la v.a. binomial X= “identificar frascos contaminados”, con probabilidad de éxito p= 0.08. (A) Función de probabilidad, (B) Función de distribución.

En la Figura 22-A se puede observar el valor esperado en $x=1$, es decir hay mayor probabilidad de detectar 1 frasco contaminado en un lote de 20 frascos, mientras que a partir de 6 frascos contaminados es

prácticamente imposible identificar contaminación en el lote. Por otro lado, la probabilidad acumulada hasta un valor de $x = 5$, se alcanza el valor máximo de probabilidad de 1 (Figura 22-B). Dicho en términos del problema, la probabilidad de identificar no más de 5 frascos contaminados es aproximadamente de 1.



Actividad de aprendizaje recomendada

Se recomienda realizar las siguientes actividades para reforzar los conocimientos de esta semana:

- *Actividad recomendada:* Se inspeccionan tres muestras de agua de un río para identificar la presencia de metales pesados. Si denotamos por "P" cuando la muestra da positivo, y "N" cuando da negativo, el espacio muestral será:

$$\Omega = \{NNN, PNN, NPN, NNP, PPN, PNP, NPP, PPP\}$$

Si la variable aleatoria es $X = \text{"Obtener muestras positivas"}$, construya gráficamente las funciones de probabilidad y distribución acumulada, y estime el valor esperado y la desviación estándar de X .

Retroalimentación: Para llevar a cabo este ejercicio, deberá definir primero dos vectores: el vector de valores de X y el vector de probabilidades asociado a X . Luego de forma similar al ejemplo 5.1.1 obtenga la gráfica. Ya en la gráfica puede fijarse en la línea más alta, y esta corresponderá al valor esperado de X .

- *Actividad recomendada:* Un reporte de prensa afirma que el 45% de los ciudadanos de cierta población se oponen a la construcción de un centro comercial en un área designada

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

como reserva ecológica. Si se encuesta a un grupo de 30 personas de forma aleatoria. Cuál es la probabilidad de que:

- Menos de la mitad se opongan a la construcción
- Más de 20 se opongan a la construcción
- Exactamente la tercera parte se opongan a la construcción
- Entre 10 y 20 personas inclusive, se opongan.
- Graficar la distribución de probabilidades usando el programa R.

Retroalimentación: Recuerde que, para calcular probabilidades con la ley binomial, usando el programa R, debemos disponer de 3 argumentos: el valor de la variable aleatoria x (en cada literal se conoce dicho valor, por ejemplo en (a) el valor sería 14, en (c) sería 10), el tamaño de muestra $n=30$ para este ejercicio, y la probabilidad de éxito en este caso $p=0.45$, con estos datos puede guiarse de acuerdo al ejercicio 5.2.1



Semana 12

En la semana doce revisamos dos distribuciones de probabilidad muy importantes como son la distribución de Poisson (discreta) y la distribución normal (continua). La primera muy útil para analizar datos que se expresan en forma de conteos y cuya probabilidad de ocurrencia es baja, y la segunda cuando se trata con variables continuas.

Para una mejor comprensión de los temas indicados, se recomienda revisar los **recursos de aprendizaje**:

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Mendenhall, W., Beaver, R., & Beaver, B.. (2016). Introducción a la Probabilidad y Estadística. 13^a edición. México: CENGAGE Learning.

De este texto, en el capítulo 3, sección 3.6, revise la distribución de probabilidad de Poisson, su ecuación y propiedades.

Para mayores detalles sobre las funciones de densidad y distribución de variables aleatorias continuas, y otras características complementarias, se le recomienda revisar los temas: Funciones de densidad de probabilidad, Funciones de distribución acumulada y valores esperados, en las secciones 4.1 y 4.2 del texto básico.

Más detalles sobre la distribución normal y normal estándar, respecto a las fórmulas, fundamentos matemáticos y propiedades, podrá encontrar al revisar el tema Distribución Normal, en la sección 4.3 del texto básico

5.3. Distribución de Poisson

La otra distribución más comúnmente utilizada en la vida real (conjuntamente con la Binomial) es la distribución de Poisson. Llamada así en honor a Simeon D. Poisson (1781-1840) francés que desarrolló esta distribución basándose en estudios realizados en la última etapa de su vida. Esta distribución expresa a partir de una frecuencia de ocurrencia media, la probabilidad que ocurra un determinado número de eventos (llamados también *sucesos raros*) durante un período de tiempo o espacio (Gutiérrez 2012). Los siguientes son ejemplos de variables aleatorias que se distribuyen de acuerdo a una ley de Poisson:

- Número de accidentes laborales durante un mes
- Número de bacterias en una muestra de agua



- Número de trabajadores reportados por mal comportamiento en un año
- Número de fallas por m^2 de construcción
- Número de mutaciones en una secuencia genética
- Número de especies herbáceas en un cuadrante de $1m^2$, etc.

Todas las v.a. de tipo Poisson tiene las características: (1) son discretas, (2) el evento ocurre en un período específico de tiempo, espacio, volumen, etc., (3) es considerada como el límite al que tiende la distribución binomial, cuando n tiende a infinito y la probabilidad de éxito (p) tiende a cero. De esta forma, se emplea la distribución de Poisson como aproximación de experimentos binomiales, mediante el siguiente criterio: $n > 50$, $p < 0.1$, $np < 5$.

En la ecuación de la distribución de Poisson se identifica el parámetro μ (en algunos textos se emplea en su lugar el parámetro lambda λ), que representa el valor medio de la v.a. X y es prácticamente el único parámetro de la ley de Poisson. Este parámetro puede ser obtenido a partir del valor esperado de una variable aleatoria binomial: $\mu = np$.

Una característica propia de la distribución de Poisson es que coincide el valor esperado con la varianza de la v.a. X, es decir: $E[X] = V[X] = \mu$

Además, la distribución de Poisson conlleva un conjunto de supuestos fundamentales como (Pagano 2001):

1. La probabilidad que acontezca un suceso en un intervalo es proporcional a la amplitud del intervalo.
2. Teóricamente es posible que suceda un número infinito de sucesos en un intervalo dado. No hay límite de ensayos.
3. Los sucesos ocurren independientemente tanto en el mismo intervalo como entre intervalos.

Ejemplo 5.3.1 (ejercicio 86, sección 3.6 del texto básico) En el agua de lastre que es descargada de un barco hay organismos con una concentración de 10 organismos/m³, de acuerdo con un proceso de Poisson:

- ¿Cuál es la probabilidad de que 1 m³ de descarga tenga al menos 8 organismos?
- ¿Cuál es la probabilidad de que el número de organismos en 1.5 m³ de agua de descarga exceda su valor medio por más de una desviación estándar?
- ¿Para qué cantidad de descarga la probabilidad de que haya menos de un organismo sería igual a 0.999?
- Graficar la distribución de probabilidades de la variable aleatoria.

Definimos la variable aleatoria X= “Número de organismos”, y el valor medio $\mu = 10$ por cada metro cúbico de volumen.

Para calcular la distribución de Poisson en el programa R utilizamos la función ‘dpois’, y para la probabilidad acumulada, la función ‘ppois’. A continuación, damos respuesta a las interrogantes.

- Es importante simbolizar cada una de las interrogantes, antes de ejecutar las funciones en el programa. Por ejemplo, en el literal (a) el término “al menos 8” es equivalente a decir “8 o más”, o también “mayor o igual a 8”. Así:

$$P(X \geq 8) = 1 - P(X < 8)$$

Esta probabilidad en el programa R, se calcula ejecutando la función:

$$1 - ppois(7, 10)$$

Obteniéndose:

[1] 0.7797794

Note que la función “`ppois`”, a diferencia de la binomial “`pbinom`”, requiere solamente dos argumentos, el primero corresponde al valor de la variable X (en este caso 7), y el segundo al valor medio para el intervalo respectivo, en este caso 10 para el intervalo (unidad de volumen) 1 metro cúbico.

- b. Se busca la probabilidad para un volumen de 1.5m^3 , esto implica que debemos ajustar el valor del parámetro valor medio. Para 1m^3 , $\mu = 10$, entonces para 1.5m^3 , $\mu = 15$.

El valor medio es 15 y una desviación estándar es 3.872983, podemos tomar el valor redondeado de 4. Así, simbólicamente lo que se quiere determinar es:

$$P(X > 15+4) = P(X > 19) = 1 - P(X \leq 19)$$

En el programa R, esta probabilidad se calcula ejecutando la función:

```
1- ppois(19, 15)
```

```
[1] 0.1247812
```

- c. En este caso, el proceso es inverso a los literales anteriores, puesto que conociendo el valor de probabilidad (0.999) se pide determinar la descarga para la cual prácticamente no haya organismos

$P(X < 1) = 0.999$, para ello usamos la función “`qpois`” que permite calcular el cuantil conociendo la probabilidad y un valor medio, en este caso por inducción el valor medio es 0.001 organismos por volumen de descarga. Así:

```
qpois(0.999, 0.001)
```

```
[1] 0
```

Cero organismos ocurren cuando tenemos 0.001 organismos en promedio, y esto se cumple para una descarga= 0.001/10 = 0.0001 m³.

- d. Finalmente graficamos la distribución de probabilidades para un valor medio de 10 organismos por m³.

Definimos un vector de valores de la v.a. X, puesto que no conocemos en tamaño de n una guía puede ser duplicar el valor esperado.

```
x=0:20
```

```
plot(x,dpois(x,10),type='h',xlab="organismos/m^3",ylab="Probabilidad",cex.axis=1.3, cex.lab=1.5,lwd=3,ylim=c(0,1))
```

```
plot(x,ppois(x,10),type='s',xlab="organismos/m^3", ylab="Probabilidad acumulada", cex.axis=1.3,cex.lab=1.5,lwd=3)
```

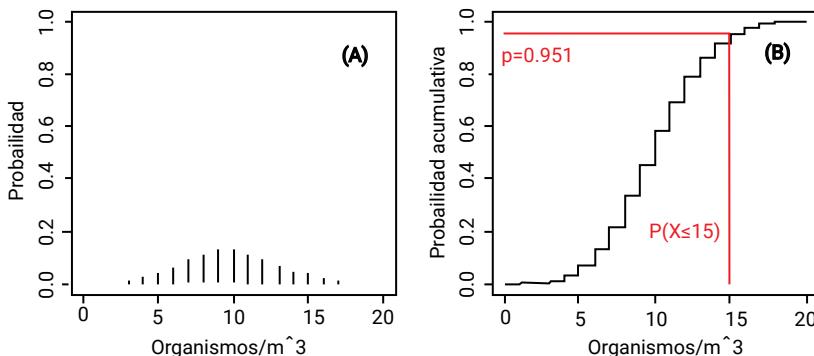


Figura 23. (A) Función de probabilidad de Poisson de la v.a. X="Nro. De organismos por m³ de descarga", (B) Función de distribución acumulada de X.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

En la gráfica de la probabilidad absoluta (Figura 23-A), se observa que la probabilidad de encontrar organismos en 1m³ de descarga no excede al 20% (todas las líneas están por debajo de 0.20). Por otro lado, la probabilidad de que hasta 15 organismos (o menos) estén presentes en 1m³ de descarga es mayor al 95% (Figura 22-B).

Hasta ahora hemos tratado con variables aleatorias discretas cuyos posibles valores pueden ser escritos como sucesiones o listas de números enteros incluido el cero. En la presente unidad hablaremos de las variables aleatorias continuas. Para iniciar con este tema se le recomienda revisar el desarrollo de los contenidos en el texto básico.

Una gran cantidad de variables aleatorias utilizadas en aplicaciones científicas y de ingeniería son descritas mediante variables continuas, estas variables pueden asumir cualquier valor en un intervalo dado de los números reales. La representación gráfica de la distribución de una v.a. continua se denomina *función de densidad* que a menudo se interpreta como el límite de un histograma cuando el número de observaciones crece hacia el infinito. Con frecuencia este tipo de distribuciones presentan forma acampanada, pero de todas ellas la que más se asemeja a una campana es la distribución normal.

5.4. Distribución normal

Considerada como la más importante y útil en el campo de la probabilidad y la estadística, su uso frecuente se debe a que hay muchas variables asociadas a fenómenos naturales que siguen el modelo de la normal. Variables tales como: caracteres morfológicos de individuos cierta etnia o especie (talla, peso, diámetros, área de dosel, etc.), caracteres fisiológicos, caracteres sociológicos, caracteres psicológicos, errores cometidos a realizar mediciones, distribuciones de estadísticos muestrales como la media, la varianza, etc. Su gráfica se denomina “curva normal”, tiene forma acampanada y por esto también se le denomina “campana de Gauss” (Figura 24-A).

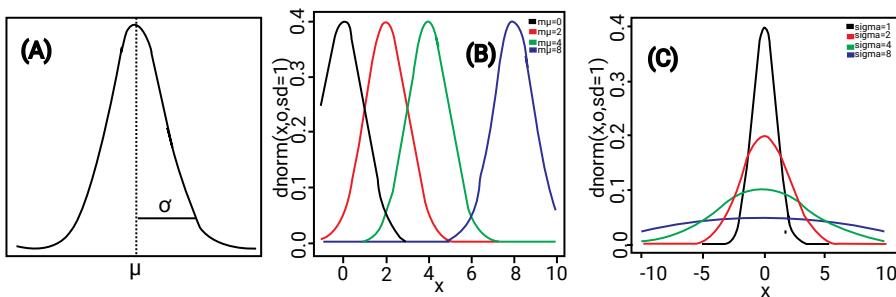


Figura 24. (A) Curva normal $N(\mu, \sigma)$. (B) Efecto del parámetro de centralización (μ). (C) Efecto del parámetro de escala (σ).

La curva normal se caracteriza por sus parámetros poblacionales media μ y varianza σ^2 (o desviación estándar σ), presentando simetría respecto a la media. Una interpretación física relaciona a la media con el centro de gravedad, aquel punto de equilibrio de la distribución, y la varianza como la inercia o resistencia en hacer girar la distribución alrededor de la media (Cobo et al. 2007). Convencionalmente, para indicar que una v.a. X sigue una distribución normal, se emplea la expresión: $X \sim N(\mu, \sigma)$. Esta expresión, algunos autores la consideran como abuso de lenguaje, por esta razón sería más adecuado decir que mediante el modelo normal se consigue representar en buena forma el comportamiento empírico de dicha variable. Así por ejemplo, la altura de las plantas de *Croton sp.* en un matorral seco es $N(60\text{cm}, 8\text{cm})$, equivale a decir que la altura de las plantas de *Crotno sp.* se comporta de forma normal con media 60cm y desviación estándar 8cm.

Características de la distribución normal

La distribución normal tiene propiedades muy particulares que la hacen importante al momento de realizar análisis estadísticos, entre ellas se destacan:

- Es simétrica respecto de la media aritmética, mediana y moda.
- Es asintótica respecto del eje X (o Z si se trata de la normal estándar)

- Puede tomar cualquier valor en los números reales ($-\infty$, $+\infty$).
- Hay más probabilidad de ocurrencia para los valores próximos a la media aritmética.
- Conforme nos alejamos de la media, la probabilidad decrece dependiendo de la desviación estándar
- La forma de la campana depende de los dos parámetros μ y σ . μ se denomina parámetro de centralización y σ parámetro de escala (Figura 24-B, 24-C).
- Aunque la forma de la curva normal siempre es simétrica, sin embargo, no siempre tiene la misma dispersión (Figura 25).

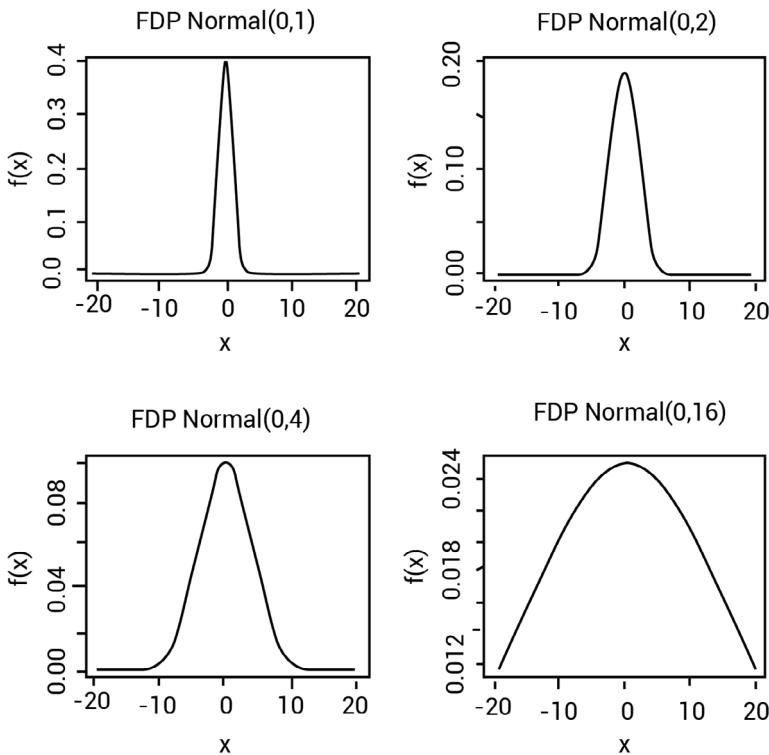


Figura 25. Curvas normales con igual media y dispersión variante.

Función de densidad de probabilidad (FDP)

Para representar probabilidad mediante la curva normal, empleamos la noción de la operación “integral” que en cálculo sirve para determinar áreas de regiones irregulares. Así, el área entre dos números reales (a,b) bajo la curva normal, representa la probabilidad de que una v.a. X tome un determinado valor en dicho intervalo. Esto lo detallaremos más adelante en el tema probabilidad como área.

La distribución normal estándar

La distribución de probabilidad normal no es única, de hecho, es toda una familia ilimitada, por esto resulta imposible definir una tabla de probabilidades para cada una de ellas. Para superar este problema, se emplea una sola distribución de entre todas las que podrían existir, se trata de la *distribución normal estándar*. Ahora la pregunta que nos planteamos es: ¿Cómo podemos obtener una distribución normal estándar?

Partimos de una distribución continua X (por ejemplo el diámetro de los árboles de un bosque), convertimos esta variable X en una nueva variable Z mediante un proceso de centrado y reducción (algunos autores lo definen como estandarización o tipificación). Este proceso consiste en tomar los valores de x , restar a cada uno la media aritmética (μ) y dividir para la desviación estándar (σ).

La distribución normal estándar cumple con las características básicas de la normal mencionadas anteriormente, pero adicionalmente posee otras propiedades particulares como:

- Es la única distribución normal donde los parámetros son conocidos: $\mu=0$ y $\sigma=1$. Por ello generalmente se denota por $N(0,1)$.
- Las unidades de medida de la variable normal estándar (Z) están dadas en términos de desviaciones estándar.

- El área total bajo la curva es igual a 1 (esto similar a otras distribuciones).
- Esta distribución sirve para el cálculo de probabilidades (Figura 6.1.1-B).

Probabilidad como área

La función de densidad de toda distribución continua de probabilidad se construye de tal forma que el área bajo la curva limitada por las ordenadas “a” y “b” sea igual a la probabilidad de que la variable aleatoria X tome cualquier valor en el intervalo [a,b] (Figura 25).

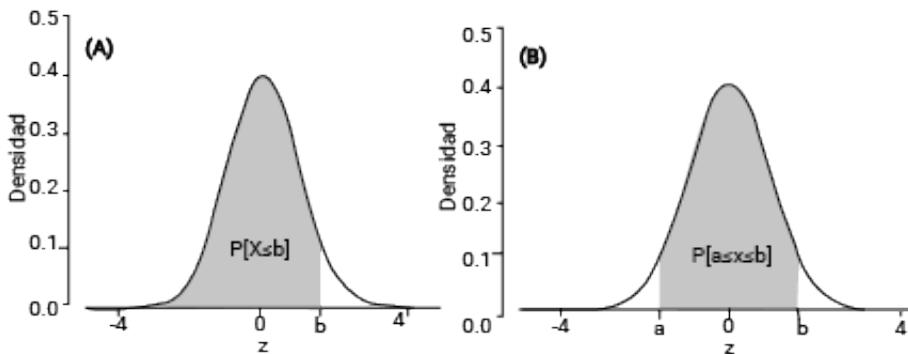


Figura 26. Representación de la probabilidad para la distribución normal.

- (A) Probabilidad de que la variable X tome valores a la izquierda de b. (B) Probabilidad de que la variable X presente valores comprendidos en el intervalo [a,b].

En las figuras 26-A y 26-B, el eje de las abscisas está definido por la variable normal estándar Z, un valor de Z mide la distancia entre un valor específico de X y la media aritmética en unidades de la desviación estándar. Entonces, al determinar el valor de Z mediante estandarización es posible hallar el área bajo cualquier curva normal con base en la normal estándar.

Para facilitar el cálculo de probabilidades utilizando la curva normal, se sugiere seguir los siguientes pasos:

1. Interpretar gráficamente (haces un bosquejo) el área de interés.
2. Calcular el valor de Z asociado a la variable aleatoria X.
3. Buscar el valor del área en una tabla de probabilidades de la normal estándar.
4. Realizar operaciones elementales (suma o resta si es necesario) para encontrar la probabilidad deseada.

En la sección 4.3 (Distribución normal) del texto básico puede encontrar ejercicios resueltos a cerca del cálculo de probabilidades con base en la normal y haciendo uso de las tablas estadísticas (ver ejemplo 4.13)

NOTA: Los pasos mencionados anteriormente se pueden resumir a uno solo mediante el uso del software estadístico R. A continuación, un ejemplo ilustrativo.

Ejemplo 5.4.1 (Datos hipotéticos) Asumamos que la temperatura durante el mes de septiembre en cierta localidad se distribuye normalmente con media 18.7°C y desviación estándar 5°C . Con base en esta información, calcule las siguientes probabilidades:

- a. La probabilidad de que la temperatura durante septiembre sea inferior o igual a 15°C .
- b. La probabilidad de que la temperatura en septiembre exceda los 21°C .
- c. La probabilidad de que la temperatura en septiembre esté comprendida entre 14 y 22°C .
- d. Halle el valor de la temperatura tal que sólo el 20% de las observaciones excedan dicho valor.

Solución:

Definimos la v.a. X = "Temperatura del mes de septiembre expresada en °C". O simplemente "Temperatura".

Hacemos un bosquejo (si es posible) del área correspondiente a la probabilidad que se desea calcular en cada literal (Figura 26).

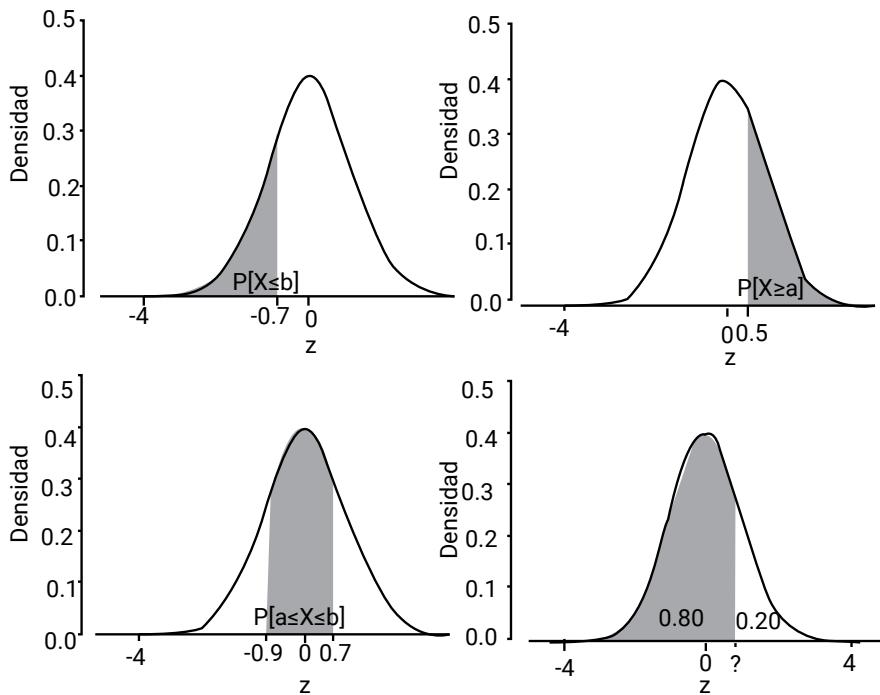


Figura 27. Representación gráfica de las probabilidades que se busca calcular en el ejemplo 6.2.1.

- Siempre es conveniente simbolizar la probabilidad que se va a calcular para identificar correctamente la relación de orden (< o >) y escribir adecuadamente la función en el programa. En este caso se desea determinar: $P(X \leq 15)$

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Usando el programa R, el proceso lo realizamos con la función pnorm():

`pnorm(15, 18.5, 5)`

Por tanto, hay una probabilidad de 0.24 de que se presenten temperaturas menores a 15°C durante septiembre.

Importante: Se debe tener en cuenta el orden de los elementos dentro de la función pnorm(), primero el valor de la variable ($X=15$), segundo el valor de la media aritmética (18.5) y tercero el valor de la desviación estándar (5). Si alteramos el orden de estos elementos, se altera el resultado.

El valor -0.74 en la abscisa de la Figura 26-A se obtiene estandarizando el valor de $X=15$: $Z=(15-18.5)/5$. Este paso es necesario si hacemos los cálculos con ayuda de la tabla estadística de la distribución normal; sin embargo cuando usamos el programa R, este proceso ya se realiza al ejecutar la función pnorm().

- b. Se busca determinar: $P(X > 21)$

Podría también simbolizarse $P(X > 21)$ (Figura 26-B), por tratarse de una variable continua, prácticamente no habría diferencia en el resultado. En el programa este cálculo podemos realizarlo al menos de dos formas:

Primera forma: Transformando la desigualdad $P(X > 21)$ a su complemento $1-(X \leq 21)$

`1-pnorm(21, 18.5, 5)`

0.3085375

Segunda forma: Definiendo la instrucción “lower.tail=FALSE”, así:

```
pnorm(21, 18.5, 5, lower.tail=FALSE)
```

0.3085375

Entonces, hay una probabilidad de 0.31 (redondeado a 2 decimales) de que se presenten temperaturas superiores a 21°C durante septiembre.

- c. Equivale a determinar el área comprendida entre los valores a=14 y b=22.

Esta operación la representamos como una diferencia de áreas (Figura 26-C):

```
pnorm(22, 18.5, 5)-pnorm(14, 18.5, 5)
```

0.5739762

Deducimos que es bastante probable (0.57) que se registren temperaturas entre 14°C y 22°C durante el mes de septiembre.

- d. A diferencia de los literales anteriores (a, b y c) donde conociendo el valor de X se buscaba la probabilidad, ahora conociendo la probabilidad (o área) debemos hallar el valor de X.

Conforme se puede observar en la figura 26-D, hallar el valor de la variable X que deja a su derecha el 20% de las observaciones, es equivalente a encontrar el valor de X que deja a su izquierda el 80% de área. Esta operación la realizamos en el programa R usando la función qnorm() ingresando el valor de área de la izquierda, así:

```
qnorm(0.80, 18.5, 5)
```

```
[1] 22.70811
```

Por tanto, podemos decir que solamente el 20% de todas las temperaturas registradas excederán a 22.7°C. Consecuentemente, el 80% de los registros de temperatura serán inferiores a 22.7°C.

Importante: Al hacer uso de la función qnorm(), tener en cuenta que el primer valor que se ingresa corresponde al rango percentil del valor de la variable que se busca; es decir el porcentaje de área que se encuentra a la izquierda.



Actividad de aprendizaje recomendada

Para fortalecer la comprensión sobre el uso de las leyes Poisson y Normal, en la solución de casos reales, se sugiere resolver los siguientes problemas:

- **Problema 1 (ejercicio tomado de Zar (2010)):** Supongamos que se conoce que la longitud de pétalo de una población de plantas de cierta especie X es normalmente distribuida con media $\mu=3.2\text{cm}$ y desviación estándar $\sigma=1.8\text{cm}$. Qué proporción de la población se esperaría que tenga longitud de pétalo:
 - a. Mayor a 4.5cm?
 - b. Superior a 1.78cm?
 - c. Entre 2.9 y 3.6cm?
 - d. Menor a 2cm?
 - e. ¿Cuál es la longitud de pétalo tal que el 90% de la población excede dicho valor?

Retroalimentación: Note que se trata de una variable continua y además se conocen los parámetros media y desviación estándar, por tanto, la función adecuada es la normal. Para resolver este ejercicio utilizando el programa R, deberá emplear la función *pnorm* en los literales a, b, c y d, y para el literal e, la función adecuada será *qnorm*. No olvide que si busca áreas mayores o superiores a cierto valor de X, debe simbolizar la probabilidad mediante complemento: $1 - \text{pnorm}(\dots)$. Lea y analice con atención el ejemplo 6.2.1 que le servirá de guía.

- **Problema 2:** El recuento de glóbulos blancos de un individuo sano puede presentar un promedio en valor mínimo de hasta 6000 por cada mm^3 de sangre. Para detectar una deficiencia de glóbulos blancos se determina su número en una gota de sangre de 0.001mm^3 :
 - a. Defina la variable aleatoria
 - b. ¿Cuánto de raro sería encontrar un máximo de 2 glóbulos blancos en una gota de sangre?
 - c. ¿Cuán probable sería encontrar menos de 5 glóbulos blancos en dos gotas de sangre?
 - d. ¿Cuál es la probabilidad de que el número de glóbulos blancos esté entre 10 y 15 inclusive, en dos gotas de sangre?
 - e. Graficar la distribución de probabilidad para el número de glóbulos blancos en una gota de sangre.

Retroalimentación: Tenga en cuenta que antes de utilizar las funciones del programa R, debe simbolizar las probabilidades en cada literal. Además, las funciones “*dpois*” y “*ppois*” requieren de dos argumentos para su cálculo, el primer argumento se refiere al valor de la variable X, y el segundo corresponde al valor medio por unidad de

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

volúmen. Por ejemplo, si el valor medio es 6 glóbulos por cada gota de sangre, consecuentemente, en dos gotas de sangre el valor medio será 12 glóbulos, y así de forma análoga para cualquier cantidad de gotas de sangre.

Para reforzar el fundamento teórico y de análisis relacionado con las distribuciones discretas, se sugiere responder a la autoevaluación 5 de la unidad



Autoevaluación 5

Finalmente, para “cuantificar” el aprendizaje en el tema de las distribuciones probabilísticas discretas, le propongo responder la autoevaluación que se presenta a continuación, y luego de haber completado, compare sus respuestas con el solucionario que se encuentra al final de la guía didáctica.

Lea con atención los enunciados del 1 al 5 y marque la opción correcta:

1. Un fenómeno se dice aleatorio cuando:
 - a. Es posible predecir sus resultados.
 - b. No tiene posibilidad de ocurrencia.
 - c. No existe certeza de los resultados que ocurrirán.
2. Identifique la variable aleatoria continua:
 - a. El diámetro del tronco de un árbol.
 - b. El número de especies animales en un área protegida.
 - c. La cantidad de árboles que se talan diariamente.
3. Entre los siguientes ejemplos, identifique aquel que corresponde a un ensayo de Bernoulli:
 - a. Encuestar a un grupo de personas para identificar si conocen o no la normativa ambiental.
 - b. Seleccionar una persona que puede conocer o no las formas de reciclar los residuos sólidos.
 - c. Muestrear 10 árboles para identificar si están o no afectados por una plaga.

4. Se quiere cuantificar el número de hojas de un árbol afectadas por un patógeno, entonces la v.a. puede tomar los siguientes valores:
- $X \geq 0$
 - $X = 0,1,2,\dots$
 - $X = 1,2,3,\dots$
5. Una variable aleatoria establece una relación entre:
- Dos conjuntos cualesquiera de los números reales.
 - Dos sucesos del espacio muestral.
 - Elementos del espacio muestral con números reales.

En los enunciados del 6 al 10, escriba dentro del paréntesis V si la afirmación es verdadera, o F si es falsa.

6. () Simbólicamente la función de distribución acumulada de probabilidad se representa por: $P(X \geq x)$.
7. () Los parámetros que definen la distribución binomial son: la probabilidad de éxito y el número de ensayos n .
8. () El valor esperado de una variable aleatoria discreta se define como el valor que puede asumir la variable con mayor probabilidad de ocurrencia.
9. () Si la variable aleatoria (X) consiste en cuantificar la precipitación diaria en una región árida, los valores pueden ser: $X \geq 0 \text{ mm}$.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

10. () Se conoce como ley de los sucesos raros porque la probabilidad de ocurrencia de la variable aleatoria es próxima a 1.

[Ir al solucionario](#)

Luego de haber respondido la autoevaluación, compare con el solucionario que se encuentra al final de la guía didáctica, y realice los correctivos si es necesario.



Semana 13



Unidad 6. Estimación estadística - Intervalos de confianza

Con esta unidad iniciamos el estudio de la inferencia estadística, nos proponemos conocer las técnicas básicas para estimar parámetros (como la media y la proporción de la población) disponiendo de datos solamente de una muestra. Les animo a que trabajen resolviendo ejercicios y compartiendo entre compañeros sus ideas y dudas a través del EVA.

Los **recursos de aprendizaje** que le permitirán reforzar sus conocimientos en esta temática son:

Mendenhall, W., Beaver, R., & Beaver, B. (2016). Introducción a la Probabilidad y Estadística. 13^a edición. México: CENGAGE Learning.

Para conocer más acerca de la estimación estadística, conceptos generales y características de los estimadores, le recomiendo leer el tema “Algunos conceptos generales de la estimación puntual” del texto básico.

Durante el proceso de estimación surgen algunas interrogantes, por ejemplo: ¿cómo se calculan los límites de intervalo?,

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

¿existe algún valor definido para la confianza del intervalo?, ¿un intervalo confiable quiere decir que es preciso a la vez?, etc.

Para responder las preguntas formuladas anteriormente, les recomiendo leer los contenidos del tema Propiedades básicas de los intervalos de confianza, en el texto básico. Ahí podrán encontrar una descripción de las características fundamentales de los intervalos de confianza, así como las fórmulas de cálculo y lineamientos para la respectiva interpretación.

Para entender de dónde se obtienen las fórmulas para el cálculo del intervalo de confianza, le sugiero observar la figura 7.4 del texto básico. En esa figura se identifican dos regiones, la región sombreada denominada región crítica, y la región en blanco denominada región de confianza.

Para conocer más acerca de la distribución t-Student, le sugiero revisar el tema relacionado con Intervalos basados en una distribución de población normal, en el texto básico. Ahí usted puede establecer diferencias entre la distribución de Z y la distribución de t.

Apreciados estudiantes, todos los contenidos que hemos visto en las unidades anteriores, conforman el material necesario para entender la inferencia estadística, es decir realizar inferencias a cerca de una población a partir de una muestra. Para ello los estadísticos estudiados como la media aritmética, la desviación estándar, la proporción, además las distribuciones muestrales, las probabilidades y otros elementos más, nos servirán para realizar procesos de inferencia estadística.

La inferencia estadística intenta dar respuesta a dos problemas concretos: *la estimación y el contraste de hipótesis*. Entendemos por estimación el proceso de encontrar una aproximación del valor del parámetro con información basada en la muestra (Figura 28), mientras que el contraste de hipótesis implica tomar una decisión en base a la prueba de un supuesto o afirmación a cerca del parámetro.

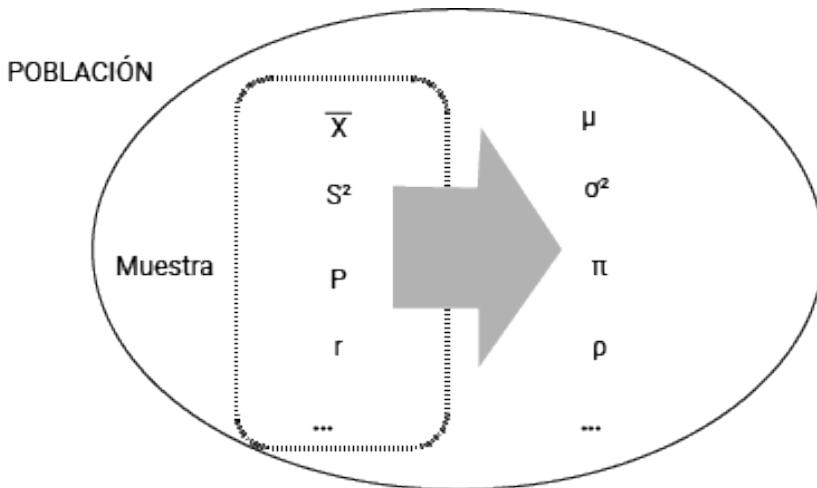


Figura 28. Ilustración de uno de los objetivos de la inferencia estadística.
Estimación de parámetros poblacionales a partir de estadísticos muestrales

En esta unidad revisaremos aspectos útiles a tener en cuenta para obtener estimaciones. A continuación, algunos ejemplos donde es adecuado emplear la inferencia estadística.

- Un médico desea conocer si un fármaco es más efectivo para el tratamiento de una infección.
- El director de personal ensaya dos métodos de entrenamiento de los empleados, y quiere conocer si producen resultados diferentes.
- Un gestor ambiental quiere saber si hay efecto (y nocivo) de los desechos industriales en la calidad del agua de los ríos.
- Un epidemiólogo investiga la relación entre el tipo de actividad de los trabajadores en una industria y las enfermedades laborales más prevalentes, etc.

Así, podemos describir innumerables ejemplos de aplicación de la inferencia estadística. Ahora la interrogante que surge es

¿Cuál método de inferencia debe usarse?, ¿estimar un parámetro o probar una hipótesis respecto al parámetro? No obstante, independientemente del método que escojamos, es conveniente una “*medida de bondad*” de la inferencia.

6.1. Tipos de estimadores

En la literatura estadística se definen básicamente dos formas diferentes de calcular estimadores, estas son: **estimación puntual y estimación de intervalo**.

Les comento que todos los estadísticos que hemos generado en las unidades anteriores se tratan de estimaciones puntuales (calcular la media, la varianza, la proporción muestrales, etc.), cuyos valores resultantes los denominamos estimadores puntuales. Por otro lado, también con base en los datos de la muestra, calculamos un rango de valores (o intervalo) definido por dos límites (inferior y superior) dentro de los cuales se espera con cierta confianza que esté contenido el valor del parámetro, es decir construimos los llamados *intervalos de confianza*.

A continuación, les comparto una figura ilustrativa que les ayudará en entender el proceso de estimación puntual.

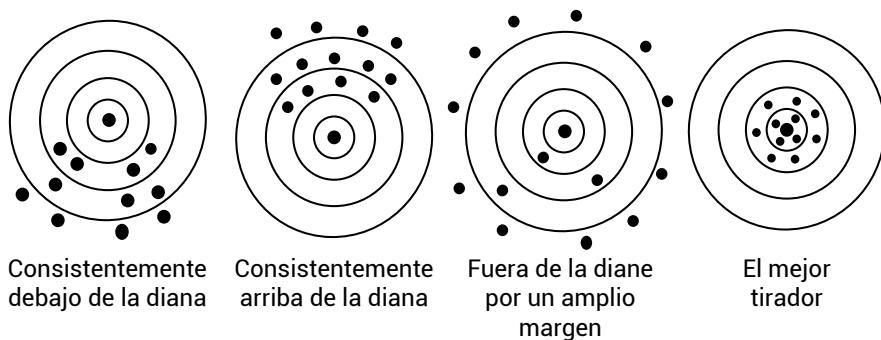


Figura 29. Semejanza entre la estimación puntual y el tiro al blanco.

Fuente: Mendenhall et al. (2010)

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

El punto central de los círculos correspondería a un parámetro, por ejemplo, la media poblacional, y cada punto en negrita representa un estimador de la media obtenido a partir de una muestra. De los cuatro escenarios presentes, el último es el mejor por dos razones: los puntos están más próximos al centro y además están más próximos entre sí, esto estadísticamente implica sesgo y variación pequeños.

Sin embargo, la realidad respecto a esta figura ilustrativa tiene dos diferencias básicas: (1) Generalmente solo tenemos la oportunidad de un disparo (un solo punto en negrita), y (2) el disparo prácticamente se realiza con los ojos vendados; es decir no sabemos con certeza dónde está ubicado el parámetro. Por ello podemos decir que cuando realizamos estimaciones de intervalo básicamente hay dos noticias, una buena y otra mala; la buena es que estas técnicas son altamente confiables, y la mala es que no sabemos si hemos acertado en nuestro caso.

A continuación, haremos una revisión de la forma de estimación más empleada en la práctica, denominada estimación por intervalos de confianza.

Estimación de intervalo (intervalos de confianza IC)

Supongamos que conocemos que la captura de carbono (en toneladas) en hojarasca por hectárea de bosque es aproximadamente 3.51t en promedio. Ciertamente esta estimación puntual nos da una referencia sobre la variable de interés, pero debemos estar conscientes que es el resultado solamente de una muestra. Entonces, para que los resultados sean más confiables, sería mejor reportar un intervalo antes que un único valor. Por ejemplo, el intervalo [2.53 – 4.49] nos indicaría que se espera que la cantidad media de carbono capturado por hojarasca esté contenida en dicho intervalo, con cierto grado de confianza.

¿Cómo definimos a un intervalo de confianza?



Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Podemos definirlo como un “*rango de valores dentro del cual se espera se encuentre el parámetro poblacional con cierto grado de confianza*”. Es decir, solamente es probable, más no es seguro, que ahí esté realmente el parámetro (por ejemplo, la media, la proporción, etc.).

Dentro del tema Propiedades básicas de los intervalos de confianza, el autor del texto básico propone una de las interpretaciones más aceptadas de intervalo de confianza, y la relaciona con el concepto de frecuencia relativa. Así, si extrajéramos 100 muestras de la población, se esperaría que 95 de ellas arrojen valores del estadístico dentro del IC.

Ahora, para la construcción de un intervalo de confianza debemos valernos de una distribución estadística, por sus propiedades, la más idónea es la distribución normal estándar. Recuerda usted que la curva normal estándar es aquella que tiene como media 0 y como desviación estándar 1, y además en el centro se ubican la media, la mediana y la moda.

6.2. Intervalo de confianza para la media

De esta forma si queremos calcular un intervalo de confianza para la media empleando la distribución normal estándar Z, debemos utilizar la relación: , donde el error estándar de la media está dado por: . Por supuesto, en la mayoría de los casos no conoceremos el valor de la desviación estándar poblacional (σ), en tales situaciones, empleamos el estimador que está dado por la desviación estándar muestral (S) y la puntuación Z se sustituye por la puntuación t, que tiene distribución t-Student con $n-1$ grados de libertad. Entonces el intervalo de confianza quedará definido como , $x \pm (t^*EE)$, donde $EE=S/\sqrt{n}$.

Ahora ¿Cómo seleccionamos Z?, la puntuación Z dependerá del nivel de confianza que fijemos para el IC, así los valores de Z más

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

típicos son: 2.58, 1.96 y 1.64 para los niveles de confianza: 0.99, 0.95 y 0.90 respectivamente. Mientras que los valores de t no dependen únicamente del nivel de confianza, sino también del tamaño muestra con el que se trabaja. La diferencia entre Z y t se evidencian básicamente para muestras pequeñas (observar la Figura 7.7, del texto básico).

Ejemplo 6.2.1 En el programa R disponemos de una tabla denominada “iris”, en ella se muestran cinco variables, cuatro numéricas y una categórica, entre las numéricas hay una que se denomina “Sepal.Length” (que en español se traduce como longitud del sépalo). Utilizando esta variable, estimar la longitud media del sépalo mediante intervalos de confianza al 95% y 99% de confianza.

Les propongo dos versiones de la solución, una mediante el uso de fórmulas y otra utilizando las funciones del programa R.

Utilizando las fórmulas, al 95% de confianza, calculamos los límites inferior y superior (LI, LS), y utilizamos la distribución t, puesto que no se conoce la varianza poblacional de la longitud de sépalo. Tome en cuenta que los 150 datos con que vamos a trabajar, constituyen una muestra de la población total que es desconocida.

Datos:

$$x = 5.84 \text{ cm}, S = 0.828, n = 150, t = 1.976, EE = 0.0676$$

$$LI = 5.84 - (1.976 * 0.0676) = 5.71 \text{ cm}$$

$$LS = 5.84 + (1.976 * 0.0676) = 5.98 \text{ cm}$$

Entonces el 0.95IC= [5.71 – 5.98]. Donde 0.95IC es la nomenclatura de IC al 95%.

De esto se puede concluir que al 95% de confianza se espera que la longitud media del sépalo esté entre 5.71cm y 5.98cm. Conscientes que hay un 5% de posibilidad que la longitud media poblacional esté fuera de esos límites.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Ingresando las fórmulas en el programa tenemos:

Paso 1: Creamos una copia del vector Sepal.Length, denotada por SL, para abreviar la escritura:

`SL=iris$Sepal.Length`

Paso 2: Calculamos los valores que se requieren para construir el IC y les asignamos un nombre a cada uno, me refiero a: la media (m), la desviación estándar (s), el tamaño de muestra (n), el error estándar de la media (EE), y definimos el nivel de error alfa (complemento del nivel de confianza).

`m=mean(SL)`

`s=sd(SL)`

`n=length(SL)`

`EE=s/sqrt(n)`

`alfa=0.05`

Paso 3: Calculamos la puntuación Z usando la función qnorm y obtenemos 1.959964 (redondeado 1.96)

`z=qnorm(1-(alfa/2),0,1)`

Paso 4: Finalmente, ingresamos las fórmulas de los límites del IC, LI para el límite inferior y LS para el superior, así:

`LI= m - (z*EE)`

`LS= m + (z*EE)`

`LI; LS`

5.709732

5.976934

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Obtenemos los mismos límites del IC, y en consecuencia aplica la misma interpretación.

Otra forma más simplificada utilizando el programa R, sería:

1. Crear una copia de la variable de interés (los datos para el intervalo):

```
SL=iris$Sepal.Length
```

2. Utilizar la función “t.test” que ya está definida en el programa y que realiza dos operaciones simultáneas: probar una hipótesis y crear un intervalo a cualquier nivel de confianza, utilizando el estadístico t-Student. En este caso queremos solamente el cálculo del intervalo, por ello al final escribimos \$conf.int, así:

```
t.test(SL,conf.level=0.95)$conf.int
```

La respuesta que se obtiene es:

```
[1] 5.709732 5.976934
```

El primer valor corresponde al límite inferior y el segundo al límite superior. En caso de que interese un intervalo al 99% de confianza sería:

```
t.test(SL,conf.level=0.99)$conf.int
```

```
[1] 5.666920 6.019747
```

Nótese que el intervalo al 99% de confianza es más amplio que el intervalo al 95%.

Es importante destacar que, si empleamos la puntuación Z en lugar de t, obtenemos el mismo intervalo, esto se debe a que el tamaño de muestra con que trabajamos es grande ($n=150$).

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

NOTA: para tamaños de muestra grandes ($n > 100$), la puntuación t converge a la puntuación Z.

En resumen: podemos emplear la puntuación Z cuando conocemos la varianza poblacional (es decir en pocas ocasiones), al contrario, emplearemos la puntuación t cuando se desconoce la varianza poblacional y ésta se estima por la varianza muestral, y además con mayor razón cuando se trabaja con muestras pequeñas.

Además, podemos tener una idea de la precisión del IC, calculando la longitud media del IC, así podemos tener una confianza de $(1-\alpha)100\%$ de que el error no excederá al producto Z^*EE (t^*EE en el caso de usar t).

¿Qué tan grande debe ser la muestra para obtener buenas estimaciones de la media?

Si se emplea como estimación de μ , podemos tener $(1-\alpha)100\%$ de confianza que el error no excederá una cierta cantidad E, cuando el tamaño de muestra sea al menos:

$$n = \left(\frac{Z * \sigma}{E} \right)^2$$

E: margen de error muestral.

Ejemplo 6.2.2 Para los datos del ejemplo 8.2.1 supongamos que la desviación estándar poblacional es de 0.8cm y queremos tener estimaciones de la media al 95% que no excedan E=0.25cm. ¿Cuál deberá ser el tamaño de muestra?

$$n = \left(\frac{1.96 * 0.8^2}{0.25} \right) = 39.3 \approx 39$$

Por lo tanto, se requiere al menos una muestra de tamaño 39 para una precisión máxima de 0.25cm.





Actividad de aprendizaje recomendada

El ejercicio práctico que se proponen a continuación le permitirá complementar los conocimientos de aplicación de los intervalos de confianza para la media poblacional.

Ejercicio (tomado de Devoré (2016)): Una muestra aleatoria de $n=15$ bombas térmicas de cierto tipo produjeron las siguientes observaciones de vida útil (en años): 2, 1.3, 6, 1.9, 5.1, 0.4, 1, 5.3, 15.7, 0.7, 4.8, 0.9, 12.2, 5.3, 0.6

- a. Obtenga un IC de 95% para la vida útil promedio
- b. Obtenga un IC al 99%.

Retroalimentación: Si va utilizar el programa R, primero debe ingresar el vector de valores, así:

```
X=c(2, 1.3, 6, 1.9, 5.1, 0.4, 1, 5.3, 15.7, 0.7, 4.8, 0.9, 12.2, 5.3, 0.6)
```

Copiamos y pegamos este vector en la consola de R, y está listo para utilizar. En las fórmulas del ejemplo 8.2.1, reemplace SL por X.

Respuesta al 95%: [1.92 – 6.50]



Semana 14

En esta semana se estudia un nuevo tema relacionado con la determinación de costos de servicios, el mismo que no se encuentra muy desarrollado por la razón que todas las metodologías, sistemas

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

o procedimientos de cálculo de costos de productos, tienen igual aplicación al momento de costear servicios. Es decir, no hay diferencias, excepto por aquella que tiene relación con el hecho de que, en algunas empresas que prestan servicios no se incurre en costos por materias primas o materiales; un ejemplo es una oficina de asesoría contable, o una entidad educativa.

Estratégicamente se planteó en la tarea del segundo bimestre, un ejercicio de ABC con aplicación a una entidad educativa con la finalidad de que comprenda en la práctica, lo indicado en el párrafo anterior.

Los **recursos de aprendizaje** que le permitirán tener una amplia comprensión del tema objeto de estudio son:

Devoré, J. (2016). Probabilidad y estadística para ingeniería y ciencias. 9^a edición. México: CENGAGE Learning.

Para conocer más acerca de la estimación de la proporción, conceptos generales y características, le recomiendo leer el tema Intervalos de confianza de muestra grande para una media y para una proporción de población.

6.3. Intervalo de confianza para la proporción

En esta sección revisaremos aspectos del estimador relacionado con las variables categóricas. Con seguridad esto no les resultará novedoso, ya que anteriormente tratamos con ese tipo de variables. Se debe destacar una característica básica de la variable binomial como aproximación a la normal.

El parámetro al que nos referimos, y no menos importante que la media poblacional, es la *proporción poblacional*. En situaciones donde la variable de estudio es cualitativa (categórica, nominal o dicotómica) y queremos aproximar la proporción de ocurrencia de una determinada categoría.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Si queremos calcular un IC para la proporción poblacional empleando la distribución normal estándar Z, debemos utilizar la relación: $p \hat{=} (Z^*EE_p)$, donde el error estándar de la proporción está dado por la relación: $EE_p = \sqrt{((p \hat{q})/n)}$.

La puntuación Z, al igual que para el caso de la media, dependerá del nivel de confianza que fijemos para el IC, así los valores de Z más típicos son: 2.58, 1.96 y 1.64 para los niveles de confianza: 0.99, 0.95 y 0.90 respectivamente.

Ejemplo 6.3.1 (Tomado de Walpole et al. 2007) Un genetista se interesa en la proporción de hombres africanos que tienen cierto trastorno sanguíneo menor. En una muestra aleatoria de 100 hombres africanos, se encontró que 24 lo padecen. Calcule un IC del 99% para la proporción de hombres africanos que tienen este trastorno sanguíneo.

Les propongo dos versiones de la solución, una mediante el uso de fórmulas y otra utilizando las funciones del programa R.

Utilizando las fórmulas, al 99% de confianza, calculamos los límites inferior y superior (LI, LS) respectivamente.

Datos:

$$\hat{p} = \frac{24}{100} = 0.24, n = 100, Z = 2.58, EE_p = 0.0427$$

$$LI = 0.24 - (1.96 * 0.0427) = 0.13$$

$$LS = 0.24 + (1.96 * 0.0427) = 0.35$$

Entonces el 0.99IC= [13% – 35%]. Donde 0.99IC es la nomenclatura de IC al 99%.

De esto se puede concluir que al 99% de confianza se espera que la proporción de hombres africanos con trastorno sanguíneo esté entre

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

13% y 35%. Con el 1% de posibilidad que dicho intervalo no contenga al valor de la proporción real.

Ingresando las fórmulas en el programa tenemos:

$$p=24/100$$

$$n=100$$

$$EE=\sqrt{(p*(1-p))/n}$$

$$\alpha=0.01$$

$$z=qnorm(1-(\alpha/2), 0, 1)$$

$$LI= p - (z*EE)$$

$$LS= p + (z*EE)$$

$$LI; LS$$

0.1299907

0.3500093

Como podrán darse cuenta, la estimación para la proporción consiste de un proceso similar a la media poblacional, con una ligera variación en el cálculo del error estándar.

¿Qué tan grande debe ser la muestra para obtener buenas estimaciones de la proporción?

Si \hat{p} se emplea como estimación de p , podemos tener $(1-\alpha)100\%$ de confianza que el error no excederá una cierta cantidad E , cuando el tamaño de muestra sea al menos:

$$n = \frac{Z^2 * \hat{p}\hat{q}}{E^2}$$

E: margen de error muestral

Z: puntuación normal estándar al nivel 1-(α/2).

Ejemplo 8.2.3 Para los datos del ejemplo 8.2.3 queremos tener estimaciones de la proporción poblacional al 95% que no excedan los dos puntos porcentuales ($E=0.02$). Suponer que una estimación preliminar de hombres africanos con el trastorno es del 30%. ¿Cuál deberá ser el tamaño de muestra?

Utilizando la ecuación anterior, obtenemos el tamaño de muestra deseado:

$$n = \frac{Z^2 * \hat{p}\hat{q}}{E^2} = \frac{1.96 * (0.30 * 0.70)}{(0.02)^2} = 1029$$

De esta manera, podemos ver que se requiere una muestra mínima de 1029 personas para obtener estimaciones de la proporción tan ajustadas como del 2%.

¿Qué pasará con el tamaño de muestra si el margen de error sube al 5%?

Le invito a dar respuesta a la interrogante, considerando la misma relación anterior para el cálculo del tamaño mínimo de la muestra.

Hemos terminado de realizar la revisión correspondiente a la temática del segundo bimestre. A continuación, le propongo una autoevaluación, para que la responda y a la vez le sirva como un indicador del nivel de aprendizaje y comprensión de lo que hemos visto.



Actividad de aprendizaje recomendada

Para adquirir mayor destreza en la construcción de intervalos de confianza para la proporción poblacional, se sugiere resolver el siguiente ejercicio práctico.

Ejercicio práctico: La unidad de transporte municipal está interesada en verificar si el transporte público cumple con los requerimientos para el cuidado del ambiente en lo que se refiere a contaminación mínima. Se realiza un muestreo aleatorio de 60 vehículos y se observa que solamente 50 cumplieron con el requerimiento. Estime la proporción de unidades que cumplen con el requerimiento, al 95% y 99% de confianza.

Retroalimentación: Sea que resuelva mediante fórmulas o mediante el programa R, primero debe calcular los valores necesarios: proporción muestral (p), tamaño de muestra (n), error estándar de la proporción (EE), nivel de error alfa, puntuación Z; con estos valores se procede a calcular los límites inferior (LI) y superior (LS).

Complementariamente, es importante que responda la Autoevaluación 6, la misma que le permitirá cuantificar la asimilación de la temática de la Unidad.



Autoevaluación 6

Con el propósito de fortalecer el fundamento teórico de los temas relacionados con estimación estadística mediante intervalos de confianza, se proponen la siguiente autoevaluación.

Lea con atención los enunciados siguientes (1 al 5) y marque la opción correcta:

1. La estadística inferencial busca dar respuesta a dos problemas básicos:
 - a. Calcular tamaños de muestras, y estimar la media poblacional.
 - b. Estimar parámetros y probar hipótesis.
 - c. Plantear hipótesis y hacer gráficas.
2. El punto central de un intervalo de confianza corresponde a:
 - a. El estadístico muestral.
 - b. La media poblacional.
 - c. Al tamaño de muestra.
3. Suponga que, al estimar la proporción poblacional, se obtiene un intervalo dado por: [0.10 – 0.25] al 90% de confianza. Esto nos dice que:
 - a. Se espera que el 90% de las medias muestrales estén contenidas en el intervalo.
 - b. El intervalo es 100% que contenga a la proporción buscada.
 - c. Se espera que el 90% de las proporciones muestrales estimadas estén contenidas en el intervalo.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

4. Escoger un nivel de confianza alto para construir un intervalo, implica:

- a. Obtener un intervalo estrecho.
- b. Mayor amplitud en el intervalo.
- c. Que el intervalo sea más preciso.

5. Para obtener un intervalo de confianza estrecho, es necesario:

- a. Un tamaño de muestra pequeño.
- b. Con cualquier tamaño de muestra siempre será estrecho.
- c. Un tamaño de muestra grande.

En los literales del 6 al 10, escriba entre paréntesis V si el enunciado es verdadero o F si es falso.

6. () El error estándar de la media muestral se relaciona inversamente con la desviación estándar muestral.

7. () El error estándar de la media muestral se define como el cociente entre la varianza y el tamaño de la muestra.

8. () El intervalo de confianza para la proporción poblacional se emplea cuando se trata con variables binomiales o cualitativas.

9. () Buscando estimar la edad media de los estudiantes que ingresan a la universidad, al 99% se obtuvo el rango [17.5 – 19.5], entonces el estimador puntual de la media fue 18.5

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

10. () Cuando no se conoce la varianza poblacional, para estimar la media utilizamos la puntuación normal estándar Z.

[Ir al solucionario](#)

Luego de haber respondido la autoevaluación, compare con el solucionario que se encuentra al final de la guía didáctica, y realice los correctivos si es necesario.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas



Actividades finales del bimestre



Semana 15

Con el propósito de prepararse para el examen presencial, se recomienda que revise los diferentes recursos educativos relacionados con las temáticas de las unidades 4, 5 y 6.

Para aquellos estudiantes que no participaron en la actividad síncrona, evalúe su aprendizaje participando en la actividad suplementaria.



Actividad de aprendizaje recomendada

Compilando todo lo revisado en las unidades 4, 5 y 6, se le propone realizar los siguientes ejercicios prácticos.

Ejercicio práctico 1 (probabilidad): Una fábrica recibe lotes de material de tres proveedores en proporciones de 50%, 30% y 20%. Se sabe que el 0.1% de los lotes del primer proveedor son rechazados, el 0.5% de los del segundo y el 1% de los del tercero es rechazado en el control de calidad que realiza la fábrica a la recepción del material. ¿Cuál es la probabilidad de que un lote escogido al azar sea rechazado?

Retroalimentación: Utilice el teorema de probabilidad total, para ello puede apoyarse en un diagrama de Ven o diagrama de árbol.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Ejercicio práctico 2 (distribuciones de probabilidad): Supongamos que el número de inasistencias de un grupo de trabajadores de cierta empresa sigue una distribución Poisson con una media de 4 faltas por semana.

- a. Determine la probabilidad de 2 faltas en una semana.
- b. Determine la probabilidad de 10 faltas en 2 semanas.
- c. Determine la probabilidad de menos de 3 faltas en una semana.

Retroalimentación: Utilice la ley de Poisson, puede apoyarse en el programa estadístico R, y utilizar las funciones dpois() para los literales a y b, ppois() para el literal c. Revise el ejercicio resuelto en la guía didáctica.

Ejercicio práctico 3 (intervalos de confianza): En un reporte de prensa (El Comercio, 01 de mayo de 2015) se afirma que 42 de cada 1000 trabajadores sufren accidentes laborales, con esta información, ¿es posible asegurar que en Ecuador la accidentabilidad laboral es inferior al 5%? Pruebe la hipótesis construyendo un intervalo de confianza para la proporción al 95%.

Retroalimentación: Para construir el intervalo, puede utilizar la distribución normal estándar Z o la distribución Chi-cuadrado. Esta última la puede ejecutar con la función prop.test() del programa R. Ver ejercicio resuelto en la unidad 6 de la guía didáctica.



Semana 16

Se recomienda que revise los diferentes recursos educativos relacionados con las temáticas de las unidades 4, 5 y 6, de manera especial realice un repaso en el texto básico, la guía didáctica virtualizada y las autoevaluaciones.



4. Solucionario

Autoevaluación 1		
Pregunta	Respuesta	Retroalimentación
1	b	Generalmente se describen dos ramas de la estadística: estadística descriptiva e inferencial.
2	a	Variable por definición es aquella que cambia entre individuos (o elementos) de un conjunto, muestra o población de estudio.
3	c	Censar significa levantar información de todos los elementos que conforman la población.
4	b	Los parámetros tienen relación con la población, y los estadísticos con la muestra.
5	b	El área basal de un árbol puede expresarse también en decimales (fracciones), por tanto es continua.
6	F	Al hablar de cantidad de hectáreas, se hace referencia a una variable numérica.
7	F	Cuando se realiza exploración de datos, se toma en cuenta los diversos tipos de variables, no exclusivamente las categóricas.
8	V	Inferir se relaciona con el método inductivo, es decir, partir de lo particular hacia lo general. En el caso de la estadística sería, partir de la muestra hacia la población.
9	F	Estratificar significa separar la población en varios estratos o grupos de acuerdo a cierto criterio, eso implica que puede haber a la vez grupos grandes y grupos pequeños.
10	V	Seleccionar la muestra sin criterio aleatorio, sino de forma direccional, se denomina muestra de conveniencia.

Ir a la
autoevaluación

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Autoevaluación 2		
Pregunta	Respuesta	Retroalimentación
1	c	Obtener la frecuencia relativa, implica dividir la absoluta para el número total de individuos.
2	a	Las variables categóricas generalmente se representan mediante diagramas de barras, aunque también mediante diagrama de sectores.
3	b	Conforme lo indicado en el numeral (2), las dos gráficas tienen el mismo fin.
4	c	El diagrama de cajas nos proporciona dos tipos básicos de información visual: variación y tendencia central de los datos.
5	a	El tipo de bosque es variable categórica, y la riqueza de especies es numérica, entonces, para relacionarlas es adecuado un diagrama de cajas.
6	Histograma	La forma más típica para representar la distribución de datos numéricos es mediante un histograma.
7	Frecuencias	El histograma resulta de formar barras donde el ancho de cada barra representa la amplitud de clase, y su altura es una frecuencia (absoluta o relativa).
8	Derecha	La dirección del alargamiento en una distribución gráfica, señala hacia donde está el sesgo de los datos.
9	Polígono de frecuencias	El polígono es otra forma de representar la distribución de datos numéricos, y se genera al unir los puntos medios de las clases (barras) en un histograma.
10	Inversa	Cuando dos variables se incrementan al mismo tiempo hablamos de relación directa, mientras que si una incrementa y la otra disminuye, se trata de una relación inversa.

Ir a la
autoevaluación

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Autoevaluación 3		
Pregunta	Respuesta	Retroalimentación
1	F	La amplitud sirve para cuantificar la variación de un conjunto de datos.
2	F	El rango inter-cuartil se obtiene de la diferencia entre el primer y tercer cuartil.
3	V	A la mediana no le afecta si en un conjunto de datos se incluye un valor extremadamente alto o bajo, porque para su cálculo sólo toma en cuenta los datos centrales.
4	F	Por definición, el valor más alto en una gráfica de distribución se corresponde con la moda.
5	V	Coeficiente de variación o dispersión relativa se obtiene de dividir la desviación estándar para la media aritmética, y ese resultado puede expresarse en porcentaje.
6	b	La varianza es el cuadrado de la desviación estándar.
7	a	La línea que resalta dentro de la caja corresponde a la mediana, no la media aritmética.
8	b	Las medidas de tendencia central coinciden en distribuciones simétricas.
9	b	La mediana separa a una serie de datos en dos partes iguales, es decir deja tanto a la izquierda como a la derecha el 50% de las observaciones.
10	a	Un valor percentil de orden K, deja tras de sí (a la izquierda) el K% de las observaciones.

Ir a la
autoevaluación



[Índice](#)[Primer bimestre](#)[Segundo bimestre](#)[Solucionario](#)[Referencias bibliográficas](#)

Autoevaluación 4

Pregunta	Respuesta	Retroalimentación
1	Frecuentista	Efectuar un experimento en repetidas ocasiones y a partir de ello estimar la probabilidad de un suceso específico, se denomina proceso frecuentista.
2	Observacionales	Generalmente los estudios de investigación son de tipo experimentales y observacionales, en los primeros no se cambia ninguna condición, y en los segundos se modifica o controla uno o varios factores.
3	Excluyentes	Sucesos excluyentes son aquellos que no comparten elementos o valores de la variable aleatoria, por ejemplo, en el lanzamiento de la moneda los dos sucesos (cara y sello) son excluyentes porque no hay un resultado que sea cara y sello a la vez, y la probabilidad del espacio muestral es la suma de las probabilidades. En este caso: $0.50+0.50=1$.
4	Los sucesos son disjuntos	La probabilidad condicionada es el cociente entre la probabilidad de la intersección y la probabilidad del suceso condicionante, por tanto, un cociente será cero solo cuando el denominador sea cero, es decir cuando los sucesos no comparten elementos.
5	Producto	Sucesos independientes es equivalente a sucesos excluyentes, y su intersección no involucra elementos. Por la regla del producto: $P(A \text{ y } B) = P(A)*P(B)$.
6	c	El teorema de la probabilidad total tiene como objetivo dividir la región (o probabilidad) objetivo en varias sub-regiones.
7	a	Las leyes de la teoría de conjuntos son equivalentes a las leyes de las probabilidades.
8	b	Por ley del complemento, dado un suceso A, entonces: $P(A)=1 - P(A_c)$.

Índice

Primer bimestre

Segundo bimestre

Solucionario

Referencias bibliográficas

Autoevaluación 4

Pregunta	Respuesta	Retroalimentación
9	b	Por la regla de la suma habría que sumar las dos probabilidades y restar la probabilidad de la intersección, pero al ser excluyentes, dicha probabilidad de intersección es nula.
10	a	El número de formas sería: $4! / (2! * (4-2)!)$, donde el símbolo “!” representa al factorial.

Ir a la
autoevaluación

[Índice](#)[Primer bimestre](#)[Segundo bimestre](#)[Solucionario](#)[Referencias bibliográficas](#)

Autoevaluación 5		
Pregunta	Respuesta	Retroalimentación
1	c	La aleatoriedad se relaciona con la incertidumbre, que es un factor que está presente en los fenómenos que se estudian con la estadística.
2	a	El diámetro de un árbol (abreviado como DAP) se puede expresar en fracción.
3	b	Efectuar un experimento binomial una sola vez, corresponde a un ensayo de Bernoulli. Por ejemplo, lanzar la moneda una vez, el resultado solo puede ser cara o sello.
4	b	El número de hojas infectadas se describe con el conjunto de los números enteros positivos incluido el cero.
5	c	Una variable aleatoria es también una función que toma un elemento del espacio muestral y lo expresa como un número.
6	F	La distribución de probabilidad acumulada hace referencia al concepto de percentil, aquella región que está a la izquierda de cierto valor ($\leq x$).
7	V	La ley binomial se caracteriza por dos parámetros: el número de ensayos y la probabilidad de éxito del suceso en estudio.
8	V	Gráficamente el valor esperado se relaciona con la barra (o línea) más alta, es decir, que representa la mayor probabilidad.
9	V	La variable aleatoria “precipitación pluvial” es numérica que además puede expresarse en forma continua dependiendo de la precisión del equipo de medición. Por ejemplo: precipitación = 38.5mm
10	F	Suceso raro es aquel que tiene pocas posibilidades de ocurrencia.

[Ir a la autoevaluación](#)

[Índice](#)[Primer bimestre](#)[Segundo bimestre](#)[Solucionario](#)[Referencias bibliográficas](#)

Autoevaluación 6

Pregunta	Respuesta	Retroalimentación
1	b	La estimación de parámetros y la prueba de hipótesis, son dos problemas básicos que se abordan con las técnicas de estadística inferencial.
2	a	Para construir un intervalo de confianza para la media poblacional, se debe disponer de la media muestral, que se ubica en el centro del intervalo.
3	c	Un nivel de confianza (por ejemplo, el K%) indica que, de 100 muestras aleatorias, se espera que K arrojen un valor estimado del parámetro, dentro de dicho intervalo.
4	b	La relación entre el nivel de confianza del intervalo y el ancho del intervalo es directa.
5	c	La relación entre el tamaño de muestra y la amplitud del intervalo de confianza es inversa.
6	F	El error estándar se obtiene del cociente entre la desviación estándar y la raíz cuadrada del tamaño de muestra, por tanto, la relación con la desviación estándar es directa.
7	F	El cociente entre la desviación estándar, no la varianza.
8	V	A partir de variables cualitativas, es posible estimar proporciones.
9	V	El estimador puntual se ubica en el centro del intervalo de confianza.
10	F	Se emplea la distribución normal estándar (Z) para realizar estimaciones de la media, cuando se conoce la varianza poblacional; caso contrario se emplea la distribución t-Student.

[Ir a la autoevaluación](#)



5. Referencias bibliográficas

- Barragués, J. I., Morais, A. y Guisasola J. (2014). *Probability and Statistics: A didactic introduction*. Boca Raton: CRC Press. Taylor & Francis Group.
- Chambers, J. M., Cleveland, W. S., Kleiner, B. y Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.
- Cobo, E. et al. (2007) *Bioestadística para no estadísticos*. Barcelona: Elsevier Doyma, S.L.
- DeCoursey W. J. (2003). *Statistics and Probability for Engineering Applications with Microsoft Excel*. Boston, USA: Newnes.
- Goos, P. y Meintrup, D. (2015). *Statistics with JMP: graphs, Descriptive Statistics, and Probability*. Chichester: John Wiley & Sons Ltd.
- Gutiérrez Banegas, A. (2012). *Probabilidad y estadística. Enfoque por competencias*. México: McGraw-Hill/Interamericana editores S.A.
- Kaps, M. y Lamberson, W. R. (2004). *Biostatistics for Animal Science*. Massachusetts: CABI Publishing.
- Madsen B. (2011). *Statistics for Non-Statisticians*. Berlin: Springer-Verlag.
- Manikandan, S. (2011). Measures of central tendency: The mean. *J. Pharmacol Pharmacoter*, 2(2), 140-142.

Índice

Primer
bimestre

Segundo
bimestre

Solucionario

Referencias
bibliográficas

Mendenhall, W., Beaver, R. J. y Beaver, B. M. (2010). *Introducción a la probabilidad y estadística*. 13a edición. México: CENGAGE Learning.

Moncho Vasallo, J. (2015). *Estadística aplicada a las ciencias de la salud*. Barcelona: Elsevier España, S.L.

Pagano, M. & Gauvreau, K. (2001). *Fundamentos de bioestadística*. Buenos Aires: Thomson Learning.

R Core Team (2017). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria.

Reimann, C., Filzmoser, P., Garrett, R. G. y Dutter, R. (2008). *Statistical Data Analysis Explained. Applied Environmental Statistics with R*. England: John Wiley & Sons, Ltd.

Seefeld, K. & Linder, E. (2007). *Statistics using R with biological examples*. University of New Hampshire, Durham, NH.
Recuperado de: <http://cran.espol.edu.ec/>

Sierra, R. (1999). Propuesta Preliminar de un Sistema de Clasificación de Vegetación para el Ecuador Continental. Loja, Ecuador: Editorial Universitaria de la Universidad Técnica Particular de Loja.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677-680.

Tabak, J. (2011). *Probability and Statistics. The science of uncertainty*. Revised edition. New York: Facts On File.

Triola, M. F. (2009). *Estadística*. México: Pearson Educación.

Índice

Primer
bimestre

Segundo
bimestre

Solucionario

Referencias
bibliográficas

Walpole, R. E., Myers, R. H., Myers, S. L. & Ye, K. (2007). *Probabilidad y estadística para ingeniería y ciencias*. México: Pearson Educación.

Wayne, D. (2002). *Bioestadística: base para el análisis de las ciencias de la salud*. México: Limusa Wiley S.A.

Zar, J. H. (2010). *Biostatistical Analysis*. New Jersey: Pearson Prentice Hall.