UTPL
La Universidad Católica de Loja

Vicerrectorado de Modalidad Abierta y a Distancia

Estadística

Guía didáctica

UTPL

Facultad Ciencias Exactas y Naturales

Estadística

Guía didáctica

Carrera	PAO Nivel
Seguridad y Salud Ocupacional	I

Autor:

Pablo Ancelmo Ramón Contento

ESTA_2025

Universidad Técnica Particular de Loja

Estadística

Guía didáctica

Pablo Ancelmo Ramón Contento

Diagramación y diseño digital

Ediloja Cía. Ltda.
Marcelino Champagnat s/n y París
edilojacialtda@ediloja.com.ec
www.ediloja.com.ec

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

Año de edición: abril, 2020

Edición: primera edición reestructurada en enero 2025 (con un cambio del 30%)

Loja-Ecuador

Índice

1. Datos de información

1.1 Presentación de la asignatura

1.2 Competencias genéricas de la UTPL

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

1.3 Competencias del perfil profesional

Contar con una formación ética, científica y tecnológica de calidad, con pensamiento crítico y reflexivo orientado a la innovación y a la investigación para la generación de nuevos modelos de gestión en Seguridad y Salud Ocupacional.

1.4 Problemática que aborda la asignatura

Debilidad en el 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 estimulan 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.

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.

Identifica gráficas estadísticas para variables categóricas y numéricas. Calcula estadísticos descriptivos de centralización, variación y posición haciendo uso de programas informáticos en un contexto de problemas reales.

Contenidos, recursos y actividades de aprendizaje recomendadas

Recuerde revisar de manera paralela los contenidos con las actividades de aprendizaje recomendadas y actividades de aprendizaje evaluadas.

Semana 1

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.

1.1 Nociones básicas

Mendenhall et al. (2015) 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 (por 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; así, por ejemplo, en el ámbito de la salud y en general las ciencias de la vida, asume el nombre de bioestadística**.**

Es por ello que, 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 quisiéramos hablar de una clasificación general de la estadística, podríamos decir que se divide en dos ramas: **descriptiva** e **inferencial**. La primera se encarga de presentar la información de 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 acerca 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 un segmento 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), muestreo 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 las unidades que están en una posición múltiplo de cinco). El muestreo estratificado 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 Muestreo por Conglomerados también se procede a dividir 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.

Actividad de aprendizaje recomendada

Estimado estudiante, con el propósito de reforzar su conocimiento, realice la actividad que se describe a continuación:

Revise para el estudio de los temas propuestos el siguiente documento: Mendenhall, W., Beaver, R., & Beaver, B. (2015). Introducción a la Probabilidad y Estadística. 14.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.

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 2

Unidad 1. Introducción a la estadística

Mendenhall et al. (2015) propone la clasificación de variables estadísticas en dos grupos: cualitativas y cuantitativas.

1.2 Variables estadísticas y escalas de medición

Luego de realizar la lectura de la referencia indicada en el párrafo anterior, 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 puede 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 únicamente mediante números enteros, incluido el cero) y **continuas** (cuando se expresa mediante números reales).

Para complementar la explicación del tema, 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 la 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) las escalas de medición se clasifican en: nominal, ordinal, intervalo y escala de razón. Cada escala tiene sus

propiedades matemáticas que determinarán el análisis estadístico adecuado en cada caso. A continuación, en la siguiente infografía se presentan las propiedades más relevantes de cada escala.

Propiedades de las escalas de medición

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), 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 solo 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 parágrafo anterior le haya servido para identificar los diferentes tipos de variables estadísticas, así como las respectivas escalas de medición.

Actividades de aprendizaje recomendadas

Estimado estudiante, luego de haber revisado los fundamentos de la estadística, así como los diferentes tipos de variables que podrían presentarse en un estudio observacional o de investigación, con la ayuda de la bibliografía citada en la presente guía le recomiendo realizar las siguientes actividades:

Actividad 1:

1. Lea los conceptos, y mediante ejemplos construya las similitudes y diferencias entre estadística descriptiva e inferencial.
2. Describa y diferencie los pasos para realizar una inferencia estadística.
3. Realice un cuadro resumen sobre los tipos de variables estadísticas, proponiendo un ejemplo para cada 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 estadística moderna se involucra en muchas áreas del conocimiento; para ejemplificar, considere solamente tres áreas: Biología, Ciencias de la Salud y Seguridad Ocupacional.

Un aspecto básico a tener en cuenta al momento de recopilar información es la aplicación del muestreo aleatorio, aunque hay situaciones particulares donde el muestreo puede ser orientado.

Actividad 2.

Se sugiere desarrollar la autoevaluación 1 y para ello, es necesario que haya revisado los temas de las semanas 1 y 2, tanto en la presente guía, así 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 el conocimiento en cuanto a la clasificación de las 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.

5. La variable “tiempo de servicio en días de un empleado” de cierta dependencia, 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 empleados que pierden su trabajo anualmente en Ecuador”.

7. ( ) La etapa de exploración de los datos está vinculada exclusivamente con las variables categóricas.

8. ( ) La estadística inferencial busca extraer conclusiones hacia 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

Contenidos, recursos y actividades de aprendizaje recomendadas

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. El análisis exploratorio no siempre equivale estrictamente a gráficas estadísticas, no obstante, que a veces se las emplea indistintamente; aunque el análisis exploratorio utiliza en gran medida las gráficas estadísticas. En esta etapa exploratoria, podemos refinar la pregunta de investigación o recolectar nuevos datos en caso de ser necesario.

Para una efectiva exploración de datos, se abordarán las técnicas básicas de agrupamiento y resumen de datos mediante tablas y gráficas estadísticas.

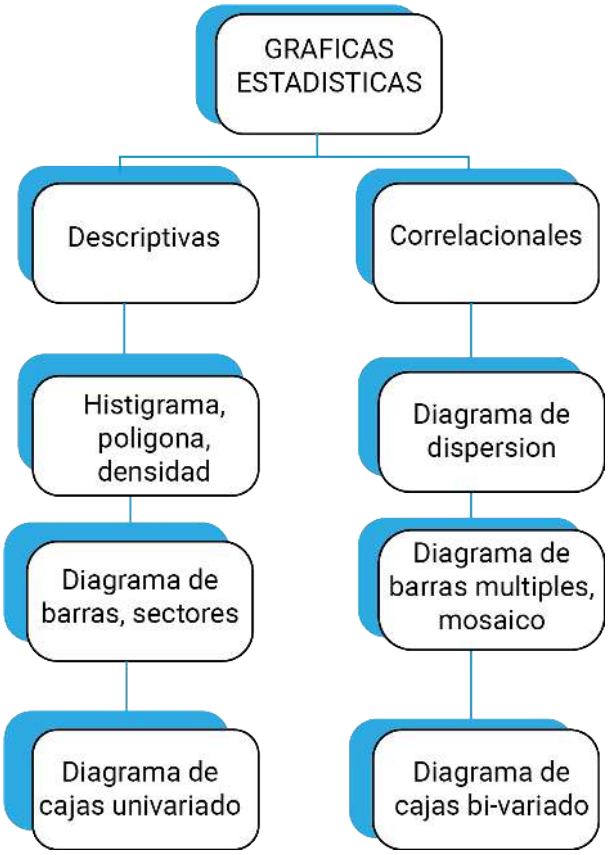
¿Por qué son importantes las gráficas estadísticas?

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 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 lo 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 como las gráficas son a la vez una herramienta que facilita el razonamiento crítico acerca de los datos. Se revisarán 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



Nota. Ramón, P., 2020.

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 trasmiten 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 el equilibrio entre simplicidad del diseño y simplicidad en la interpretación.

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 estas representaciones gráficas las puede realizar con MS Excel que permite visualización efectiva y dispone de una plataforma que transforma datos crudos en información valiosa.

Al final de la guía didáctica, usted cuenta con una introducción a la estadística usando Excel, diseñado fundamentalmente para principiantes. Además, para fortalecer el aprendizaje de esta herramienta, puede acceder a los tutoriales en línea descritos en el plan docente (REA 4). Por otro lado, también el uso de MSExcel en la obtención de resultados estadísticos por ejemplo las gráficas.

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 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 las 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 manera el resumen de la información.

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 (> 100), estos valores de se agrupan en **intervalos de clase**, que generalmente poseen la misma amplitud y son mutuamente excluyentes (no se traslapan, no se sobreponen). 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.

Ejemplo 2.1.1

Revise el siguiente ejemplo donde hacemos uso de la base denominada "airquality" (Chambers et al. 1983) que se refiere a variables de calidad del aire en la ciudad de Nueva York.

Ejemplo 2.1.1. Datos “airquality” en programa R

2.2 Gráficas estadísticas para variables categóricas

Usualmente, iniciamos el trabajo con datos categóricos resumiendo la información mediante tablas de frecuencia (Mendenhall et al (2015), Sección 2.1); posteriormente necesitamos una forma más intuitiva y eficaz de presentar la información, es ahí donde hacemos uso de las gráficas. En el caso de las 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 la frecuencia absoluta o relativa de cada categoría de la variable nominal.

Estimado estudiante, le invito a revisar el Anexo 1. Estadística usando Excel, donde se indica la construcción de este diagrama con el uso de Excel.

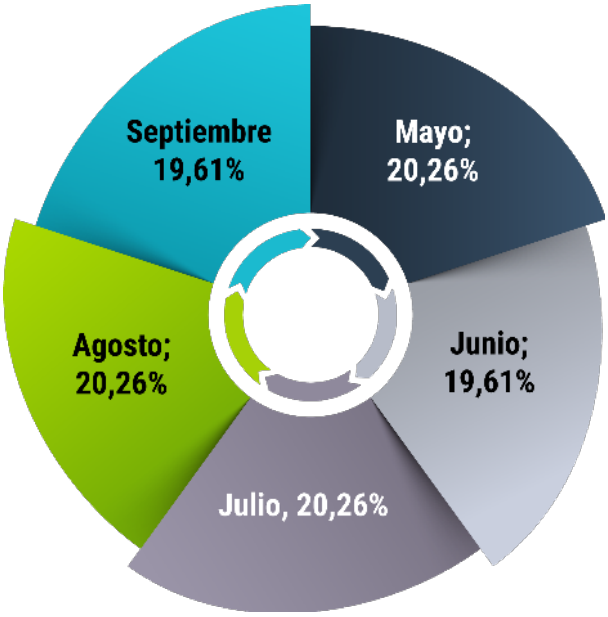
Ejemplo 2.2.1

Para ilustrar el diagrama circular, utilizamos la tabla de calidad de aire (airquality) específicamente la variable mes (Month) para representar el porcentaje de observaciones de cada mes (Figura 2).

En la figura 2, se puede observar un formato bidimensional del diagrama circular correspondiente al porcentaje de observaciones o registros por mes. Siendo los meses junio y septiembre aquellos con menor número de registros (19.6%), mientras que los meses restantes presentaron mayor cantidad de observaciones (20.26%)

Figura 2

Diagrama circular del porcentaje de observaciones por mes (Tabla "airquality")



Nota. Ramón, P., 2020.

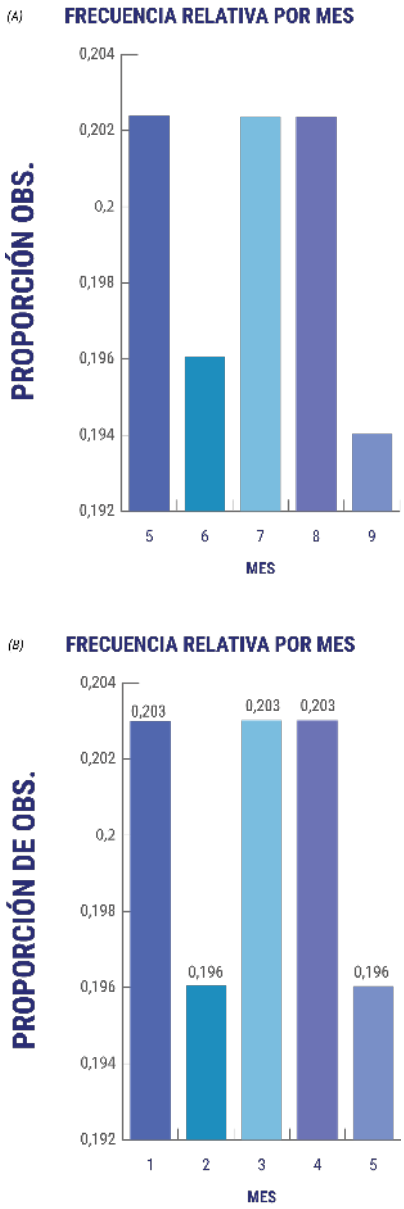
Diagrama de barras

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 frecuencia relativa de cada categoría, respectivamente. Esta forma de representación es particularmente útil cuando se dispone de dos columnas de datos, una categórica, por ejemplo el mes y otra numérica por ejemplo el porcentaje de observaciones. 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.

Figura 3

Diagrama de barras para la velocidad del viento categórica: (A) utilizando frecuencia relativa, (B) incluyendo la frecuencia en la gráfica



Nota. Ramón, P., 2025.

Para incluir los valores de la frecuencia, activamos la opción etiqueta de datos en Excel.

La interpretación de este diagrama es similar al 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

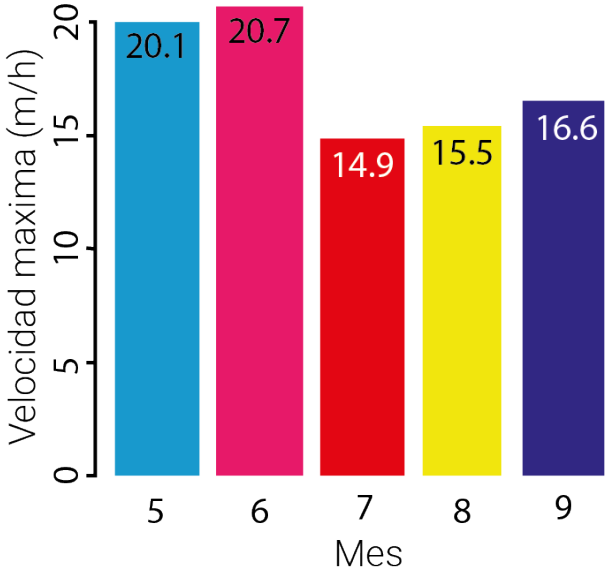
Otra aplicación del diagrama de barras es la representación de series de tiempo mediante barras; por ejemplo, la velocidad máxima del viento para cada mes observado.

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

En la Figura 4 se observa que los meses 5 y 6 presentaron mayor velocidad del viento superando las 20 millas/hora, mientras que para los tres meses subsiguientes la velocidad máxima desciende hasta 15 millas/ hora.

Figura 4

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



Nota. Tomado de Graphical Methods for Data Analysis (p. 347), por Chambers et al., 1983, Belmont, CA: Wadsworth. CC BY 4.0.

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.

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. Conformado por 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 el anexo al final de la guía un ejemplo de construcción de histogramas usando Excel. A continuación, un ejemplo detallado sobre la construcción de un histograma.

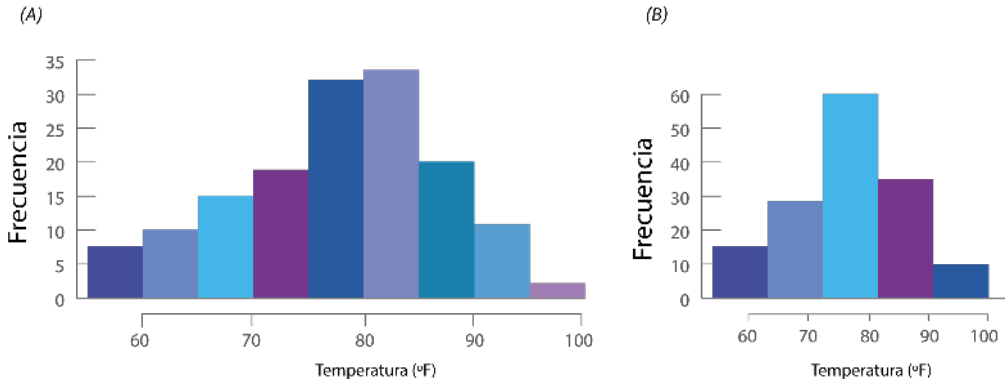
Ejemplo 2.3.1

Representar gráficamente la distribución de los datos de la variable temperatura (tabla "*airquality*").

Primero construiremos un histograma sin controlar el número de clases, por defecto el programa va a generar las clases. Al final de la guía el proceso completo para crear histogramas con Excel.

Figura 5

Distribución de la variable temperatura: (A) Histograma con nueve clases, (B) Histograma con cinco clases



Nota. Ramón, P., 2025.

Se observa mejor la distribución de los datos con mayor cantidad de clases en el histograma (Figura 5-A). El número de clases se ha calculado con el algoritmo de Sturges, este algoritmo se basa en la siguiente ecuación: $N=1 + \log_2(n)$.

Donde N: número de clases, n: número de datos.

Sin embargo, el usuario puede ajustar el número de clases según el tamaño de la serie de datos, 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 5-B). Respecto a la forma de la distribución podemos decir que es bastante simétrica, es decir las frecuencias más altas se ubican en el centro de la distribución y decrecen "simétricamente" en ambos lados de la gráfica (Figura 5-B); no obstante, con mayor número de clases el decrecimiento de las frecuencias no es muy simétrico (Figura 5-A) puesto que las frecuencias de las clases de la izquierda decrecen más lentamente que aquellas que se ubican en la cola derecha de la gráfica.

A manera de observación se puede resaltar que una distribución con mayor número de clases permite más fácilmente identificar sesgos o asimetrías. La pregunta que suele presentarse en esta situación es ¿Cuál es el número óptimo de clases?, para responder esta pregunta la regla de Sturges, mencionada arriba es la más común.

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.

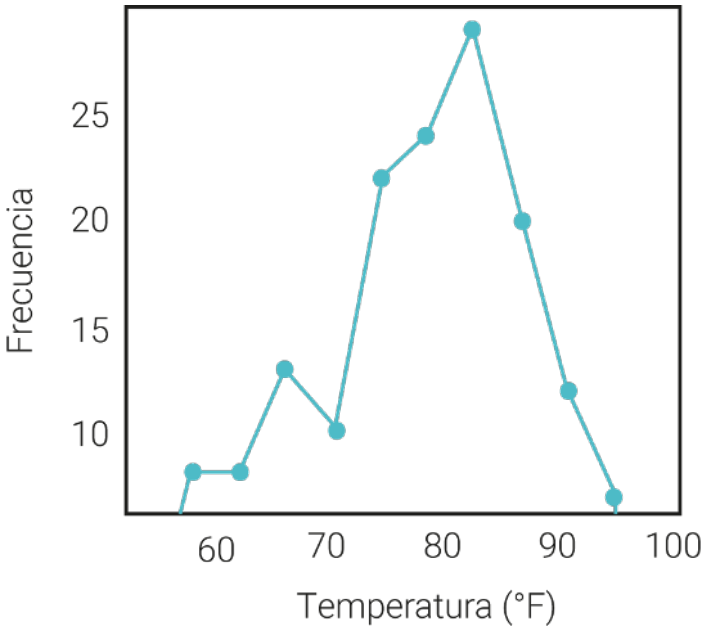
A continuación pongo a su disposición la función poli.frec, que le servirá para construir el polígono de frecuencias. Para ello deberá copiar y pegar todo el bloque de la función en la consola del programa R, ejecutar con "*enter*" y estará lista para utilizarla.

Ejemplo 2.3.2

Construir el polígono de frecuencias para la variable temperatura.

Figura 6

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



Nota. Tomado de Graphical Methods for Data Analysis (p. 147), por Chambers et al., 1983, Belmont, CA: Wadsworth. CC BY 4.0.

El polígono de frecuencias revela con más detalle la simetría (o ausencia de esta) en una gráfica de distribución de frecuencias; se observa un ligero sesgo negativo (alargamiento hacia la izquierda) en la distribución de la temperatura (Figura 6).

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 consideran.

Curva de densidad

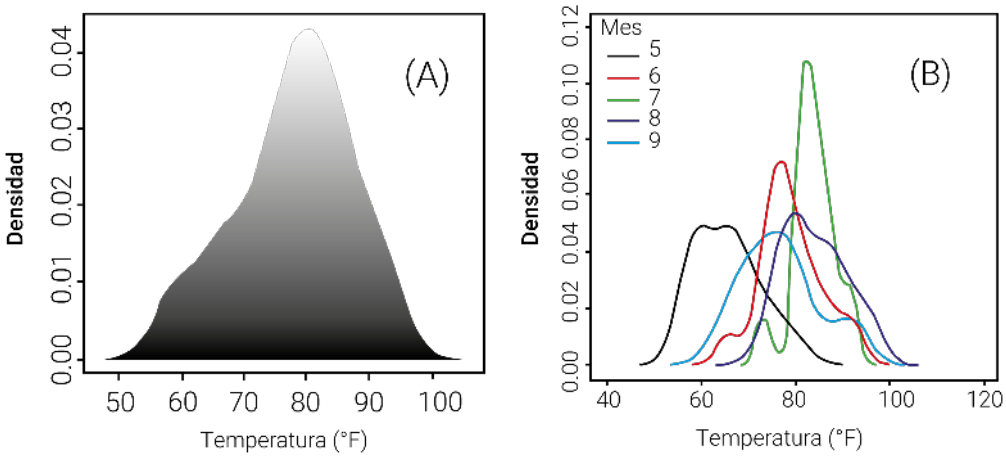
Un diagrama de densidad, también conocido como "*kernel density plot*", se construye a partir de una variable numérica y muestra la distribución suavizada de los puntos a lo largo del eje numérico constituido por la variable de interés. Graficar una curva de densidad implica construir un estimador de 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 para su construcción 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.

Ejemplo 2.3.3

En la densidad de la temperatura global (Figura 7-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 7-B), las densidades muestran formas diferentes, con sesgos más pronunciados en algunos meses (ejemplo meses 7 y 9).

Figura 7

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



Nota. Ramón, P., 2020.

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 más adelante).

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. El diagrama de cajas es útil para identificar valores anómalos o extremos ("*outliers*" en inglés) y para comparar distribuciones. La distribución de los datos se representa y resume a través de cinco estadísticos: el mínimo, el primer cuartil (Q1), la mediana (cuartil 2, Q2), el tercer cuartil (Q3) y el máximo.

Por ahora revisaremos la forma más sencilla de construir un diagrama de cajas, conformado por un rectángulo central cuya altura define el rango inter-cuartil (IQR), el segmento dentro del rectángulo representa 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 con sesgo positivo tendría un "*bigote*" más largo en la dirección positiva (hacia arriba) que en la negativa.

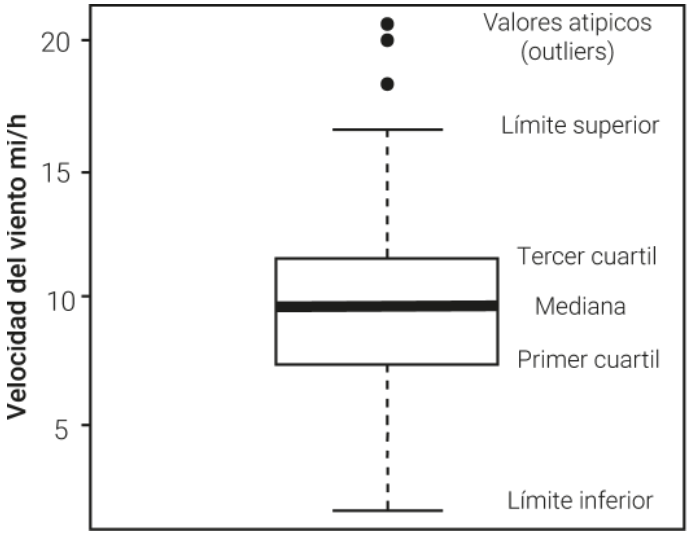
Cuando el valor medio es mayor que la mediana, también habría indicios de sesgo positivo. La presencia de valores atípicos no necesariamente indica que se son datos u observaciones erróneos, de hecho, son muy importantes porque poseen información valiosa del conjunto de datos; no deben ser removidos 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.

Figura 8

Distribución de la velocidad del viento mediante diagrama de cajas



Nota. Tomado de Graphical Methods for Data Analysis (p. 147), por Chambers et al., 1983, Belmont, CA: Wadsworth. CC BY 4.0.

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

Por ahora nos limitaremos a decir que el valor central de la velocidad está alrededor de 10mi/h, y que hay presencia de valores extremos de la velocidad por arriba del límite superior. Estos valores atípicos corresponden a valores muy altos de la velocidad (días con viento muy fuerte alrededor de 20mi/h). 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:

Límite inferior = Q1 – (1.5*IQR) Límite superior = Q3 + (1.5*IQR)

Donde Q1: cuartil de orden 1 (percentil 25)

Q3: cuartil de orden 3 (percentil 75)

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.

Actividades de aprendizaje recomendadas

Estimado estudiante, con el propósito de reforzar sus conocimientos sobre las gráficas numéricas, realice las siguientes actividades:

Actividad 1:

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

Ejercicio práctico: agrupar los datos de la variable ozono (Ozone) de la tabla “airquality” en siete clases y reportar los resultados conforme a la Tabla 1, con la gráfica respectiva.

Retroalimentación: La variable Ozone está dada en partes por billón (ppb). Similar al ejemplo de la temperatura, ahora puede crear 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.

Actividad 2:

Analizar la distribución de datos de la variable ozono (columna Ozone de la tabla "*airquality*") mediante: un histograma, un polígono de frecuencias y una curva de densidad.

Actividad 3:

Realice los siguientes ejercicios, (tomados de Mendenhall et al. (2015):

- Cincuenta personas se agrupan en 4 categorías A, B, C y D, el número de personas que caen en cada categoría son: 11, 14, 20, y 5 respectivamente. (a) ¿Cuál es la unidad experimental? (b) ¿Cuál y qué tipo de variable es la que se mide? (c) Elabore una gráfica circular y de barras para describir los datos (d) ¿Qué porcentaje de personas están en la categoría B?
- Un fabricante de Jeans tiene plantas en California, Arizona y Texas. Un grupo de 25 pares de jeans se seleccionan al azar de la base de datos, registrándose el estado en que se produjo cada uno: CA, CA, AZ, CA, CA, AZ, CA, AZ, AZ, AZ, AZ, TX, CA, TX, AZ, TX, TX, AZ, TX, CA, CA, TX, TX, TX, CA. (a) ¿Cuál es la unidad experimental? (b) ¿Qué variable se mide y de qué tipo es? (c) Elabore gráfica circular y de barras (d) ¿Qué proporción de jeans se hizo en Texas, TX? (e) ¿Cuál estado produjo más jeans?

Retroalimentación: las tres formas de representación gráfica 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 estas preguntas le permitirá asimilar la idea intuitiva de la propiedad de normalidad, muy importante en estadística.

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 4

Unidad 2. Exploración de datos

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, y en adelante las llamaremos también correlacionales.

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 dos variables categóricas (categórica vs. categórica)

Para ello es necesario partir de una tabla resumen (tabla 1) de dos variables cualitativas (X, Y), denominada tabla de doble entrada (o tabla de contingencia).

Tabla 1
Esquema de tabla de doble entrada (tabla de contingencia)

X \Y	B1	B2	...	Bk	Total fila
A1	n11	n12	...	n1k	TF1
A2	n21	n22	...	n2k	TF2
...	...	...		...	...
Al	nl1	nl2	...	nlk	TFl
Total columna	TC1	TC2	...	TCk	**N**

Nota. Ramón, P., 2020.

Donde $\boldsymbol{A}_1$, ..$\boldsymbol{A}_l$ *y* $\boldsymbol{B}_1$, ..., $\boldsymbol{B}_k$ son las categorías de X e Y respectivamente, N el número total de individuos observados, n_{lk} es la frecuencia absoluta del par ($\boldsymbol{A}_l$, $\boldsymbol{B}_k$) de entre los N individuos que poseen la categoría $\boldsymbol{A}_l$ de X y la categoría $\boldsymbol{B}_k$ 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 ($\boldsymbol{A}_l$, $\boldsymbol{B}_k$) 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

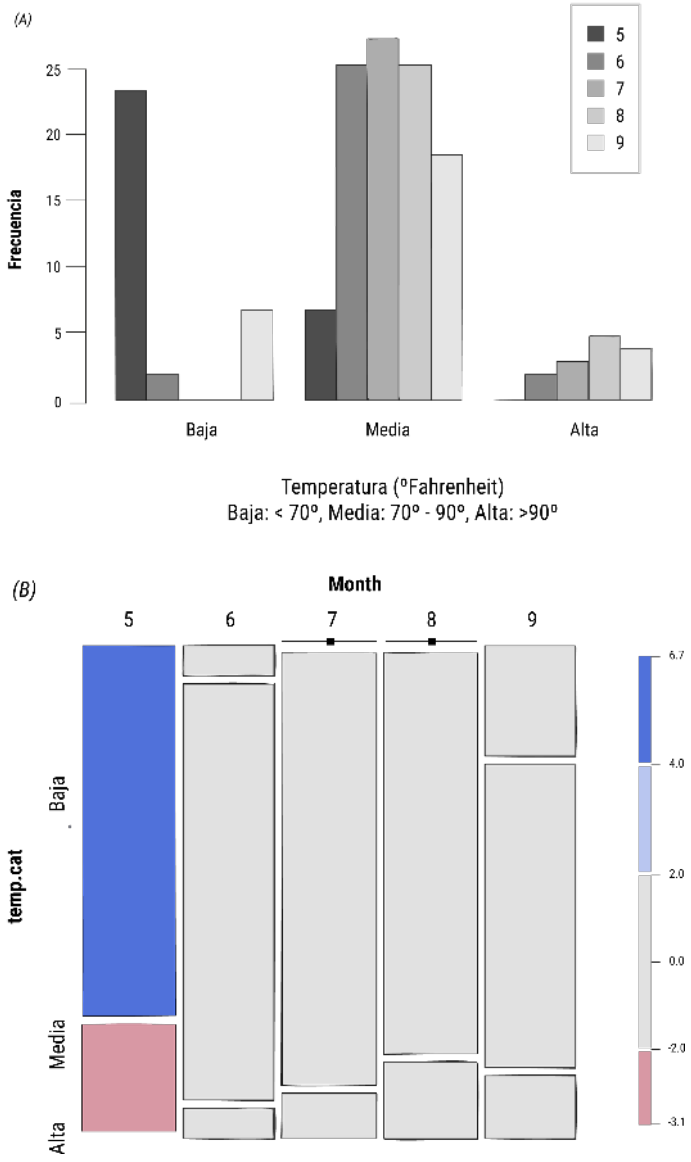
Este diagrama es una representación gráfica de una tabla de contingencia, donde las filas y las columnas representan las dos variables categóricas respectivamente. Esta gráfica permite examinar la relación entre dichas variables. Sobre el eje Y se presentan las categorías (modalidades) de una de las 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 divide en 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 con los datos de la tabla “airquality”.

Figura 9

Relación entre la temperatura (baja: <70°F, media: 70-90°F-15m/h, alta: >90°F) y el mes observado: (A) diagrama de barras, (B) diagrama de mosaico



Nota. Tomado de Graphical Methods for Data Analysis (p. 147), por Chambers et al., 1983, Belmont, CA: Wadsworth. CC BY 4.0.

En la Figura 9-A , se observa que la temperatura máxima se presentó en los meses de junio, julio y agosto (6, 7 y 8 respectivamente), en julio y agosto no se observó niveles de temperatura baja (Figura 9-A). Por otro lado, en el mes de mayo (mes 5) se observó más registros de temperatura baja que lo esperado por azar y así mismo, menos registros de temperatura media de lo esperado (Figura 9-B). En los meses restantes no se observó diferencias significativas entre los niveles de temperatura observados y esperados, eso lo ratifica el color gris de los rectángulos para junio, julio, agosto y septiembre (Figura 9-B).

Diagrama de cajas (numérica vs categórica)

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 dos variables (numérica dependiente vs cualitativa 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: (1) la variación de la edad de los empleados de una empresa dependiendo de la sección donde labora; (2) el área de hoja de cierta especie vegetal por efecto de la altitud; (3) la variación de la presión arterial por efecto del medicamento; (4) 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, se desea saber cómo cambia la temperatura en función del mes observado.

B. Con los datos de la tabla "*airquality*", se quiere conocer gráficamente la relación entre la radiación solar (Solar.R) en función de la temperatura categorizada.

Visualmente estas relaciones se pueden identificar mediante un diagrama de cajas bi-variado (Figura 11–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.

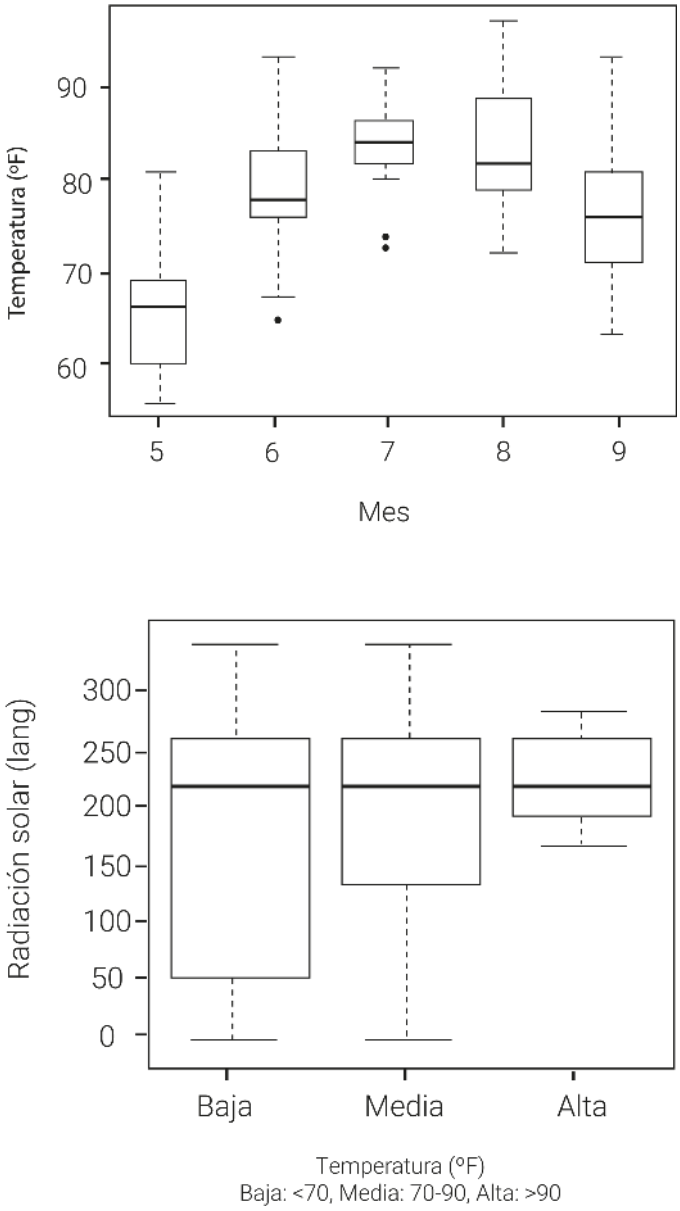
Los meses 6 y 7 presentan valores anómalos, registros muy bajos de la temperatura. Consecuentemente distribución sesgada de la temperatura.

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 10

(A) Relación de la temperatura en Nueva York y el mes observado. (B)Relación entre la temperatura y la radiación solar. Tabla de datos "airquality"



Nota. Tomado de Graphical Methods for Data Analysis (p. 147), por Chambers et al., 1983, Belmont, CA: Wadsworth. CC BY 4.0.

Se observó un máximo de temperatura en el mes de julio (Figura 10-A). Por otro lado se observó una ligera relación creciente entre la radiación solar y la temperatura (Figura 10-B). , Así mismo observando el ancho de caja, se observó mayor variación de la temperatura en la categoría por debajo de 70°F.

Relación de dos variables numéricas

Diagrama de dispersión (scatterplot)

Es un conjunto de puntos ubicados en un plano de coordenadas cartesianas, donde cada eje del plano representa una variable numérica, por ejemplo, la relación estatura vs peso de un grupo de personas de una localidad. 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 (eje X) puede ser considerada explicativa (por ejemplo, la estatura) y la otra puede ser considerada variable de respuesta (por ejemplo, el peso de las personas). Cuando la tendencia es creciente en ambos ejes, se habla de asociación positiva, si es creciente en un eje y decreciente en otro, entonces la asociación será negativa. En el caso de no existir una tendencia (puntos dispersos aleatoriamente en el plano), las variables no están relacionadas.

Algunas consideraciones adicionales sobre el 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 (correlación no implica causalidad).
2. Cuánto 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 el coeficiente de correlación (esta técnica se revisará 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. Es decir, puede ser necesario

incluir una tercera variable cualitativa en el diagrama con la finalidad de observar agregaciones por categorías.

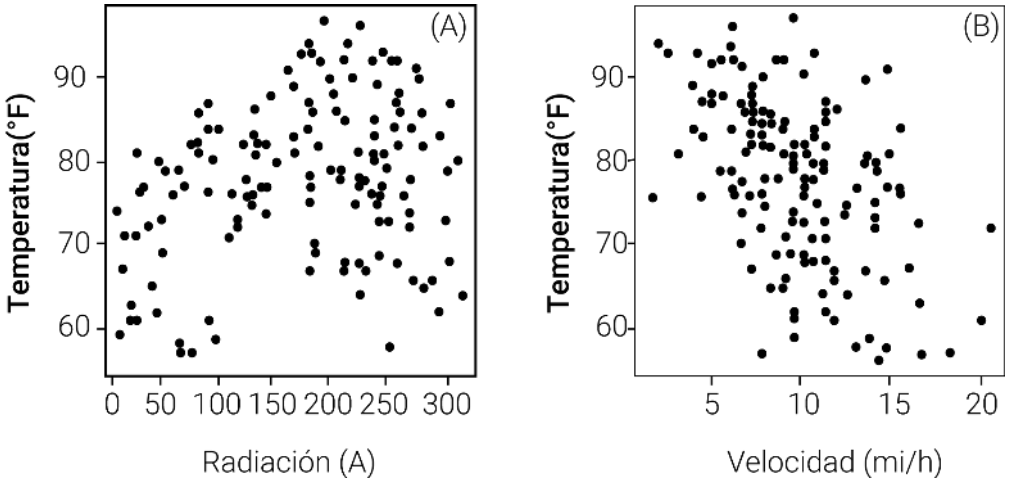
5. Dibujar un diagrama de dispersión es el primer paso en el análisis de correlación y regresión entre variables numéricas.

Ejemplo 2.4.3

Mediante diagramas de dispersión, analice 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*".

Figura 11

Relaciones bivariadas: (A) Temperatura vs radiación solar, (B) Temperatura vs velocidad del viento



Nota. Tomado de Graphical Methods for Data Analysis (p. 147), por Chambers et al., 1983, Belmont, CA: Wadsworth.

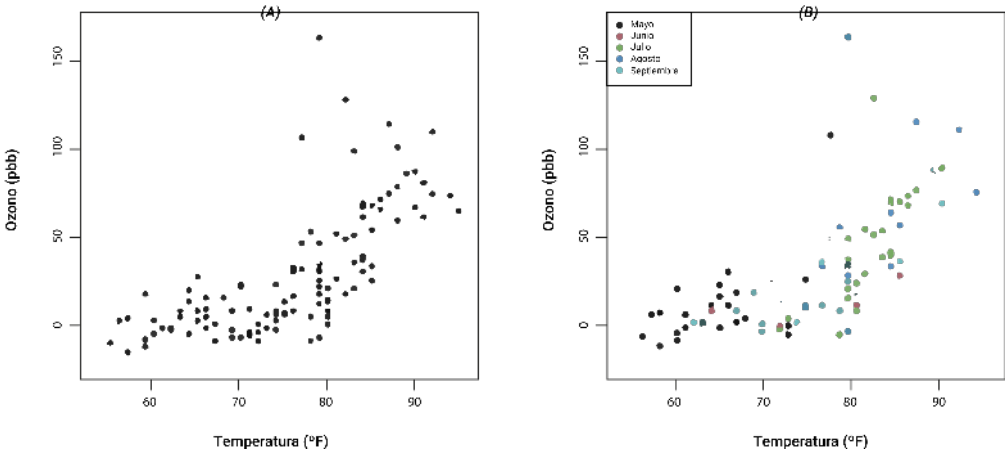
Se observa una ligera relación positiva entre la temperatura del ambiente y la radiación solar (figura 11-A), sin embargo, si se observa con detalle, la tendencia parece ser cuadrática (forma parabólica) puesto que forma un máximo para una radiación aproximada de 200; por otro lado, se evidencia una relación negativa (inversa) entre la temperatura y la velocidad del viento (figura 11-B). Hay una clara diferencia entre los dos diagramas, el segundo (figura 11-

B) muestra una relación más definida que el primero. Esta fuerza de relación puede ser cuantificada mediante el 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 gráficas si presentan indicios de correlación.

Figura 12

Relación entre la temperatura del ambiente y la concentración de ozono: (A) Diagrama de dispersión inicial, (B) Diagrama estratificado por el mes



Nota. Tomado de Graphical Methods for Data Analysis (p. 147), por Chambers et al., 1983, Belmont, CA: Wadsworth. CC BY 4.0.

En la Figura 12-A se observa relación creciente entre la temperatura y la concentración de Ozono, sin embargo, no todos los meses presentan la misma trayectoria (Figura 12-B). Los meses que mantienen la trayectoria del patrón global son julio, agosto y septiembre. También se observan algunos valores extremos de Ozono en julio y agosto. Mientras que el mes de mayo presentó los valores más bajos de Ozono, con relación levemente creciente.

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

Actividades de aprendizaje recomendadas

Estimado estudiante, luego de haber revisado las tres formas de gráficas correlacionales, le invito a realizar las siguientes actividades:

1. Utilizando la tabla "*airquality*" (Chambers et al. 1983), crear una variable llamada "temp.cat" a partir de categorizar la variable numérica "temperatura" en tres clases: (0-70], (70-90] y (> 90), asignando las etiquetas baja, media y alta respectivamente. Luego tabular conjuntamente las variables temp.cat y Month. Con el resultado de la tabulación, construir un diagrama de barras y uno de mosaico. Escriba una interpretación de cada 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.

2. 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 bivariadas, 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 el ancho de cada caja, y con base en las diferencias que observe, escriba una interpretación.

3. Utilizando los datos de la tabla "*trees*", 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: en la tabla observará que hay tres variables numéricas, las cuales le servirán para construir los diagramas de dispersión. Las gráficas presentan cierta dispersión de los puntos, es decir, no hay relación perfecta.

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

Autoevaluación 2

Con el propósito de cuantificar el aprendizaje de lo revisado en la unidad 2, acerca de las formas básicas de representación gráfica de datos estadísticos, propongo el siguiente cuestionario.

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. Con base en sus características podemos decir que son equivalentes:

   a. Diagrama de barras y diagrama de cajas.
   b. Diagrama de barras y diagrama circular.

c. Diagrama de cajas y diagrama de dispersión.

4. La gráfica que permite identificar la variación y el valor central es:

   a. Histograma.
   b. Nube de puntos.
   c. Diagrama de cajas.

5. Se quiere analizar la relación entre el tipo de empresa y la edad de los empleados, sería adecuado utilizar:

   a. Diagrama de cajas.
   b. Diagrama de dispersión.
   c. Diagrama de densidad.

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 representa la _____ 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

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 5

Unidad 3. Estadísticos descriptivos

En la semana cinco se abordan temas relacionados con el cálculo de estadísticos descriptivos, los cuales se clasifican en medidas **centrales**, medidas de **dispersión** y medidas de **posición**. Las medidas centrales sirven para encontrar un elemento que represente al conjunto de datos, mientras que las medidas de dispersión son cuantificadores de la variación de los datos, finalmente las medidas de posición representan valores de la variable detrás de los cuales se ubica cierto porcentaje de observaciones.

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 cualquiera de los campos profesionales, conlleva tratar con información muy variable, por ello es necesario definir los tipos de medidas (estadísticos) que sinteticen la información. En la presente unidad revisaremos las medidas más utilizadas en la práctica para resumir datos: tanto centrales, así como de dispersión o variación y también de posición.

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, así como en el proceso de cálculo.

3.1 Medidas de centralización

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 utilizada 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 para su cálculo según el tipo de datos se presenta a continuación:

Datos no agrupados

$$\bar{x} = \frac{\sum_{l=1}^{n} x_l}{n}$$

Datos agrupados

$$\bar{x} = \frac{\sum_{l=1}^{n} f_l x_l}{n}$$

Donde x_i representa el i-ésimo valor de la variable X y n el total de valores de la variable, también conocido como tamaño de muestra. Para el caso de datos agrupados, x_i representa el valor medio de cada 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 del grupo; caso contrario, se tornará inestable. Por ello, se recomienda tener en cuenta:

- La media aritmética es sensible ante la presencia de valores extremos o atípicos, sobre todo cuando la muestra es pequeña.
- No es recomendable usar la media aritmética como medida central en distribuciones muy asimétricas (o sesgadas).

- Cuando la variable es numérica discreta (por ejemplo, número de empleados en una sección), la media aritmética no refleja el valor exacto de la variable, porque se podría obtener valores fraccionarios (por ejemplo, media= 20.5); en tales casos se recomienda redondear al entero más próximo (Ejemplo: 20.5=21; 20.2=20).

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

Hay otras medias generalizadas como la media geométrica, la media 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 emplean 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 la 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 la 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*) o extremos.

La mediana es el dato central de todos los datos ordenados. Para su cálculo partimos de una serie ordenada y creciente de datos (x1, x2, …, xn), entonces la mediana muestral será:

Si n es impar: se toma el dato central.

Si n es par: se promedian los dos datos del centro.

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

Representa 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 (máximo), entonces se dice que la distribución es polimodal.

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*", calcular la media aritmética, la mediana y la moda de la variable concentración de ozono (Ozone), y representar gráficamente. Los resultados se reportan en la Tabla 2.

Tabla 2
Estadísticos centrales de la variable ozono (ppb)

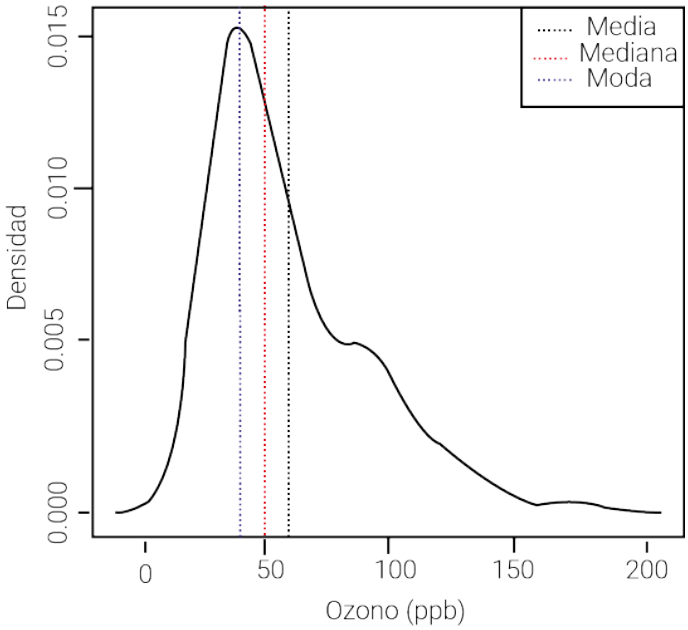
Variable	Media aritmética	Mediana	Moda estimada
Ozono	42.12	31.50	20.58

Nota. Ramón, P., 2020.

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

Figura 13

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



Nota. Tomado de Graphical Methods for Data Analysis (p. 147), por Chambers et al., 1983, Belmont, CA: Wadsworth. CC BY 4.0.

La distribución de la variable ozono es claramente sesgada a la derecha (figura 13), 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.

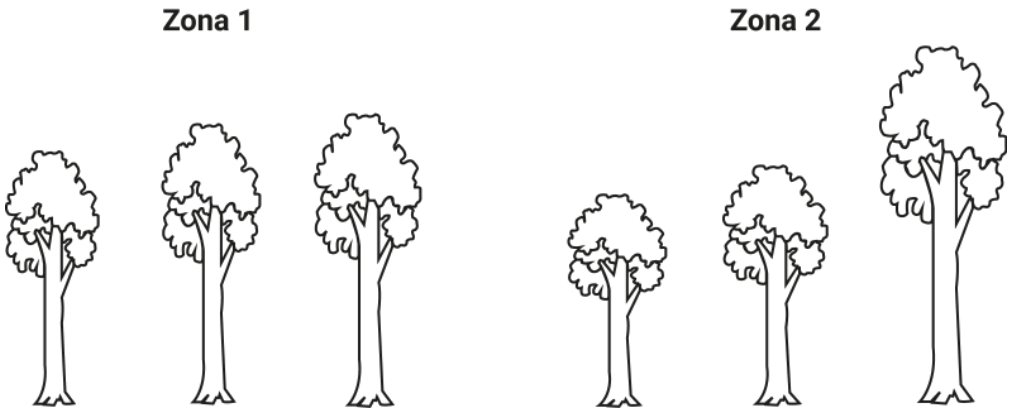
3.2 Medidas de variación

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 14, donde se compara la longitud de los árboles en dos zonas diferentes (tres árboles en cada zona), la longitud promedio 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 encontramos 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 concepto y aplicación diferente en situaciones específicas.

Figura 14

Comparación de la longitud de árboles de cierta especie en dos zonas diferentes



Nota. Ramón, P., 2020.

A continuación, una breve descripción de las medidas de dispersión más utilizadas en la práctica.

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 del rango es la siguiente:

$$R = X_{máx} - X_{min}$$

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 primero 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 en 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.

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), gráficamente estaría representado 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 se define a continuación:

Varianza poblacional

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

Varianza muestral

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

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 en su lugar 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 que 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, revise 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 siguiente:

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

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*", calcular: la amplitud, el rango inter-cuartil, la varianza y la desviación estándar.

La amplitud se define como la diferencia entre el máximo y el mínimo.

Máximo= 97; Mínimo= 56. Entonces Amplitud=41

El rango inter-cuartil (IQR) es la diferencia entre los cuartiles 3 y 1.

Cuartil 3 = 85; Cuartil 1=72. Entonces IQR=13

La Varianza = 89.59

La desviación estándar= 9.47

Ejemplo 3.2.2

Para la variable temperatura de la tabla “airquality” calcule el coeficiente de variación (dispersión relativa) de la temperatura por cada mes observado, e identifique 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. Los resultados se observan en la tabla 3, la cual muestra que la variación máxima (10.9%) de la temperatura se presentó en el mes de septiembre (mes 9).

Tabla 3
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

Nota. Ramón, P., 2020.

Actividades de aprendizaje recomendadas

Estimado estudiante, una vez que ha revisado lo concerniente a las medidas de centralización, se sugiere realizar las siguientes actividades:

Actividad 1:

Utilizando los datos de la tabla "*mdeaths*" (Diggle 1990) correspondiente al número de muertes por enfermedades pulmonares en UK, calcular las medidas centrales para cada año, además representar gráficamente la serie completa y realizar una pequeña interpretación de la gráfica.

Retroalimentación: Si observa, las filas de la tabla corresponden a los años empezando en 1974 hasta 1979. Para la representación gráfica, seleccione cualquiera de las gráficas descritas en la unidad 2, que estén relacionadas con variables numéricas. Si quiere identificar valores atípicos, lo más adecuado sería emplear un diagrama de cajas.

Actividad 2:

Con el propósito de aclarar las diferencias entre varianza poblacional y varianza muestral, investigue sobre el tema "Medidas de variabilidad".

Actividad 3:

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 intercuartil (IQR) es calcular por separado los cuartiles 1 (Q1) y 3 (Q3) y luego establecer la diferencia: IQR=Q3-Q1.

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 6

Unidad 3. Estadísticos descriptivos

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 las medidas de posición denominadas *percentiles*. Es así como la mediana divide a la distribución de 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 con los *cuantiles*. Un cuantil del 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. En este sentido, 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 un 0 % y 100 %, lo que se conoce como rango percentil. Casos particulares 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 percentil como Pk = k*(n+1)/100.
- Si el valor de la posición es un número decimal, por ejemplo, 6.67, ubicamos los valores que están en las posiciones 6 (A) y 7 (B). Luego, mediante el uso de la fórmula, calculamos el percentil deseado.

Para nuestro ejemplo, el percentil de orden k será: Pk = A + 0.67(B-A)= **65**

Para entender mejor, vamos a ejemplificar este procedimiento.

Ejemplo 3.3

Utilizando los datos de la altura de los árboles (Tabla 4), calcular el percentil 45.

Tabla 4

Diámetro, altura y volumen de los árboles cerezos negros

Diámetro	8.3, 8.6, 8.8, 10.5, 10.7, 10.8, 11, 11, 11.1, 11.2, 11.3, 11.4, 11.4, 11.7, 12, 12.9, 12.9, 13.3, 13.7, 13.8, 14, 14.2, 14.5, 16, 16.3, 17.3, 17.5, 17.9, 18, 18, 20.6
Altura	70, 65, 63, 72, 81, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75, 74, 85, 86, 71, 64, 78, 80, 74, 72, 77, 81, 82, 80, 80, 80, 87
Volumen	10.3, 10.3, 10.2, 16.4, 18.8, 19.7, 15.6, 18.2, 22.6, 19.9, 24.2, 21, 21.4, 21.3, 19.1, 22.2, 33.8, 27.4, 25.7, 24.9, 34.5, 31.7, 36.3, 38.3, 42.6, 55.4, 55.7, 58.3, 51.5, 51, 77

Nota. Datos tomados de The Minitab Student Handbook, por T. A. Ryan, B. L. Joiner, & B. F. Ryan (1976), Duxbury Press.

El percentil 45 estaría en la posición: 45(31+1)/100 = 14.4

Luego el percentil 45 será: 75+(0.4(76-75)) = 75.4

Redondeando el valor al entero más próximo podemos decir que P_{45} = 75

Este valor nos indica que el 45% de los árboles poseen altura inferior o igual a 75 pies, consecuentemente, el 55% posee altura mayor que 75 pies.

Con los datos de la altura podemos también calcular los cuartiles 2 y 3, mismos que equivalen a los percentiles 50 y 75 respectivamente.

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, podemos hacer un resumen de la variable “altura” conforme se muestra a continuación.

Mínimo	=	63
Q1	=	72
Mediana	=	76
Media aritmética	=	76
Q3	=	80
Máximo	=	87

La altura mínima es 63 y la máxima 87 pies. 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.

Actividades de aprendizaje recomendadas

Estimado estudiante, le invito a realizar las actividades que se proponen a continuación:

Actividad 1:

Ejercicio práctico: utilizando los datos de la tabla "trees" (Atkinson 1985), para la variable diámetro, determine las siguientes medidas de posición: percentil 10, percentil 90, cuartiles 1, 2 y 3. Verifique que el cuartil 2 coincide con la mediana.

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

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 los percentiles, usted debe obtener los siguientes valores:

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

Actividad 2:

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

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.

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

1. ( ) La amplitud es una 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.

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.

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 del 90 % de las observaciones.
   c. Deja 10 observaciones a su derecha.

Ir al solucionario

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

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 7

Actividades finales del bimestre

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, evalúen 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, le propongo realizar los siguientes ejercicios prácticos.

- **De la bibliografía complementaria de Mendenhall et al. (2015), capítulo 2, resolver los ejercicios suplementarios 2.54 y 2.65.**

  Para el ejercicio 2.54 intente representar la información mediante un diagrama de cajas. **Del capítulo 3, sección 3.2, resolver el ejercicio 3.3 (gasto de consumidores).**

Retroalimentación: el cálculo de los estadísticos descriptivos puede hacerlo mediante fórmulas, usando la hoja electrónica Excel (ver Anexo 1. Estadística usando Excel). Revise los ejemplos resueltos en la guía didáctica, sección de gráficas.

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 8

Actividad de aprendizaje recomendada

Se recomienda que revise los diferentes recursos educativos relacionados con las temáticas de las unidades 1, 2 y 3, Mendenhall et al. (2015), anexo y REA.

Segundo bimestre

Resultado de aprendizaje 1:

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

Calcula probabilidades mediante técnicas de conteo. Identifica e interpreta las distribuciones de probabilidad para variables aleatorias en casos reales. Emplea programas informáticos para el cálculo de intervalos de confianza.

Contenidos, recursos y actividades de aprendizaje recomendadas

Recuerde revisar de manera paralela los contenidos con las actividades de aprendizaje recomendadas y actividades de aprendizaje evaluadas.

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 visto pertinente 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 cambiaran 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 como base 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 un 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.

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 quisiéramos 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 solo 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 siguiente infografía.

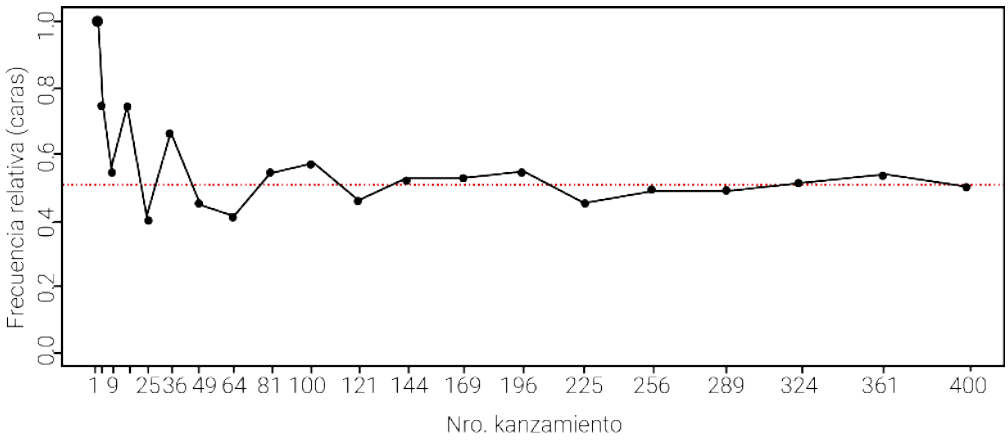
Tipos de probabilidad y sus características principales

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.

Figura 15

Simulación del experimento "lanzamiento de la moneda"



Nota. Ramón, P., 2020.

El experimento se realizó 20 veces, incrementando el número de lanzamientos en cada ensayo. 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 15).

Para que resulte más sencillo entender este tema relativo a la probabilidad, a continuación, les propongo algunos (de entre muchos) términos y conceptos básicos, necesarios para familiarizarse con el lenguaje probabilístico.

Conceptos previos

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 la bibliografía complementaria de Mendenhall et al. (2015).

4.2 Propiedades operacionales

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

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

1. Toda probabilidad es no negativa: $P(A) \geq 0$
2. Toda probabilidad se define en el intervalo [0,1]: $0 \leq P(A) \leq 1$
3. La probabilidad del espacio muestral es 1: $P(\mathbf{\Omega}) = 1$, siempre que **Ω** esté compuesto de sucesos excluyentes.
4. P es aditiva, es decir, si $A_1, A_2, \ldots, A_n$ son n sucesos disjuntos o excluyentes entonces:

$$P(U_{i=1}^{n} A_i) = \sum_{i=1}^{n} \square P(A_i)$$

5. $P(\varnothing) = 0$
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 5 para el cálculo de probabilidades de un suceso.

Tabla 5

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**

Nota. Adaptado de Propuesta preliminar de un sistema de clasificación de vegetación para Ecuador continental (p. 2), por R. Sierra (1999). Reimpresión, Universidad Técnica Particular de Loja (UTPL).

Con estos datos, si se escoge un individuo al azar, calcular las siguientes probabilidades: Suceso simple: ¿Cuál es la probabilidad de que sea una especie endémica?

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

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

$$P(A \cap N) = 45/531 = 0.085$$

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

$$P(RoE) = (163 + 121 + 12)/531 = 296/531 = 0.557$$
$$P(R \cup E) = P(R) + P(E) - P(R \cap E)$$
$$= (133/531) + (346/531) - (121/531)$$
$$= 358/531 = 0.674$$

Actividad de aprendizaje recomendada

Estimado estudiante, con el propósito de reforzar sus conocimientos sobre la temática relacionada con la probabilidad, le invito a realizar la siguiente actividad:

Realice la lectura "Axiomas, interpretaciones y propiedades de la probabilidad", en la bibliografía complementaria de Mendenhall et al. (2015). Donde podrá encontrar información relacionada con las reglas básicas de la probabilidad y sus propiedades operacionales.

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

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 10

Unidad 4. Probabilidad

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= (número de eventos simples favorables) / (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, …, n_k. Si un elemento es elegido al azar de cada conjunto, entonces el número total de diferentes resultados es: $(n_1)(n_2)…(n_k)$ (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 se define por la operación factorial dada simbólicamente por n!, se lee como el factorial de n y representa el producto de todos los números naturales de 1 hasta n.

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

P(3)= 3! = 3*2*1 = 6.

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:

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

Nótese que, en este caso, el orden de los subconjuntos es importante. Por ejemplo, de un conjunto de 10 empleados de una empresa, cuántas parejas de personas se pueden formar, considerando que el orden de los pares es importante.

Reemplazando en la fórmula sería:

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

Esto nos dice que podemos formar 90 parejas de personas 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:

$$(n,k) = \frac{n!}{k!(n-k)!} = \frac{n(n-1)...(n-k+1)}{k!}$$

Considerando el mismo conjunto anterior, de 10 empleados de una empresa, se quiere combinaciones de personas de dos en dos, donde el orden no importa, entonces se tendrían 45 combinaciones conforme al cálculo que se indica a continuación:

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

En Excel podemos realizar la operación así: =COMBINAT(10; 2).

Hay 45 formas diferentes de parejas de personas que se pueden combinar.

4.4 Teoremas básicos de la probabilidad

Probabilidad condicional

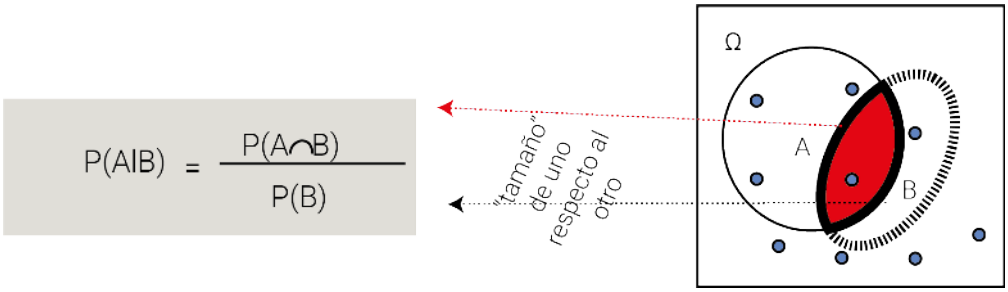
Intervienen dos sucesos (A,B), donde la probabilidad de ocurrencia del primer suceso (A) está condicionada a la ocurrencia del segundo suceso (B). La fórmula la puede observar en la sección 2.4 (Probabilidad condicional) de la bibliografía complementaria de Mendenhall et al. (2015**).**

Representación gráfica de probabilidad condicional

Para entender mejor el concepto de probabilidad condicional, ilustramos la definición con ayuda de un diagrama de conjuntos.

Figura 16

Relación entre la probabilidad condicional y la teoría de conjuntos



Nota. Adaptado de Apuntes y Videos de Bioestadística (2013), Universidad de Málaga. bioestadistica.

A partir de la figura 16 podemos destacar algunas observaciones importantes:

1. Para que la probabilidad condicional exista, la probabilidad del suceso B 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 en la medida en que A y B tengan más elementos en común.

Ejemplo 4.4.1

Utilizando los datos de la Tabla 5, 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 condicionada, 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:

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

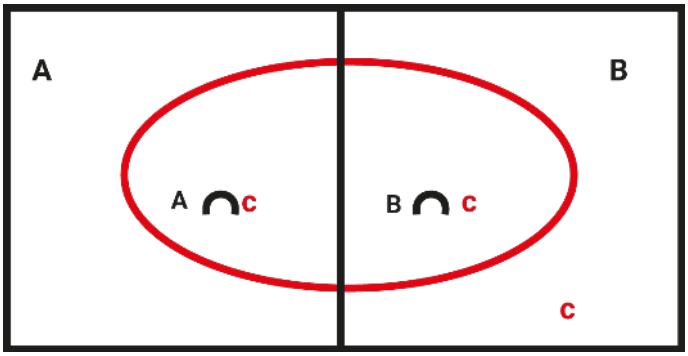
Teorema de la probabilidad total

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 servirá para determinar la probabilidad de encontrar especies caducifolias tanto en A como en B.

Figura 17

Representación gráfica el teorema de la probabilidad total



Nota. Ramón, P., 2020.

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

$$C = (A \cap C) \cup (B \cap C), \textit{ entonces :}$$
$$P(C) = P(C \cap A) + P(C \cap B)$$

Luego, aplicando la regla de la multiplicación tenemos:

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

Ejemplo 4.4.2

Utilizando el esquema de la figura 17, 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:

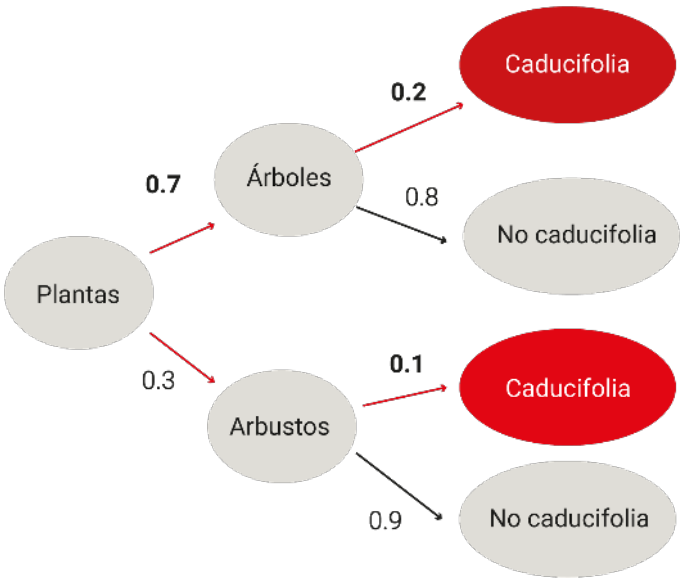
$$P(C) = P(C|A)P(A) + P(C|B)P(B)$$
$$P(C) = (0.70) * (0.20) + (0.30) * (0.10)$$
$$P(C) = 0.17$$

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 4.4.2 tenemos el siguiente diagrama:

Figura 18

Diagrama de árbol para los datos del ejemplo 4.4.2



Nota. Ramón, P., 2020.

Los caminos a través de los nodos representan intersecciones, y las bifurcaciones representan uniones disjuntas. En la figura 18 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.

Actividades de aprendizaje recomendadas

Estimado estudiante, con el propósito de profundizar los conceptos y propiedades de la probabilidad, se sugiere realizar las siguientes actividades:

Actividad 1:

Dar lectura a los siguientes documentos:

- Para mayores detalles, se le recomienda leer el tema relativo a **Técnicas de muestreo,** en la bibliografía complementaria de Mendenhall et al. (2015)**.**
- Para profundizar sobre este tema de la probabilidad condicional, le sugiero leer el tema **Probabilidad condicional,** en la bibliografía complementaria de et al. (2015)**.**
- Para conocer más acerca de sucesos independientes, le recomiendo leer el tema de **Independencia** en la bibliografía complementaria de Mendenhall et al. (2015)**.**

Actividad 2:

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 condicionado 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.

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

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 11

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

El tema que nos corresponde revisar esta semana 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 Excel.

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 de la bibliografía complementaria de Mendenhall et al. (2015).

Una vez que haya revisado el contenido indicado en la bibliografía mencionada, 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 será aleatorio cuando no existe una certeza a priori sobre las observaciones que ocurrirán (Barragués et al. 2014). En este curso nos interesa 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 presentes 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 asumir cualquier valor real. Para facilitar la comprensión de las variables aleatorias, les propongo algunos ejemplos en la Tabla 6.

Tabla 6

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	$0 \leq x \leq 100$

Variables aleatorias discretas		
Experimento	**Variable aleatoria (v.a.)**	**Valores posibles de la v.a.**
Ensayar un nuevo proceso químico	Temperatura cuando se lleva a cabo la reacción deseada (min **150º *F***; máx **212**o***F***)	150 ≤ x ≤ 212

Nota. Ramón, P., 2020.

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 6). 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 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 más 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 contenido 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.

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: **Ω** = {SS, CS, SC, CC}

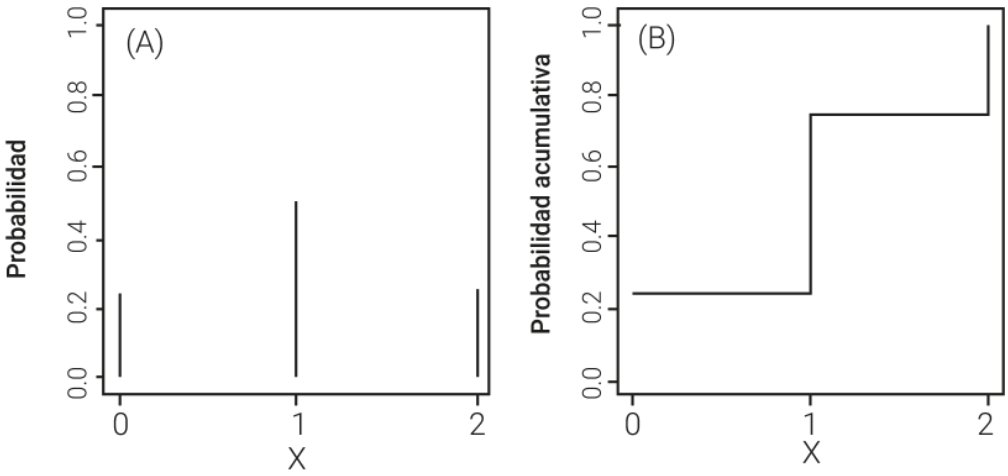
Luego, a partir del espacio muestral, cuantificamos los valores (frecuencias) que la variable aleatoria X aparece en cada uno de los cuatro resultados de **Ω**. Estos valores son {0,1,1,2}. Entonces el conjunto de valores posibles de X es {0, 1, 2}.

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

$$P(X=0)=1/4, P(X=1)=2/4, P(X=2)=1/4, \textit{donde 4 es el número de resultados en } \Omega.$$

Figura 19

(A) Función de probabilidad, (B) Función de distribución acumulada



Nota. Ramón, P., 2020.

La recta vertical con mayor altura (Figura 19-A) representa el valor esperado de la variable aleatoria X, es decir, aquel valor de la variable con mayor probabilidad de ocurrencia, en este caso X=1: P(X=1) =0.5. La figura 19-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 19-B) es aproximadamente 0.80 en el eje Y, este valor representa la probabilidad P(X≤ 1), es decir, la suma de las probabilidades: P(X=0) + P(X=1).

Complementariamente, a la distribución de probabilidades de una variable aleatoria, podemos calcular el valor esperado (aquel valor de la v.a. con mayor posibilidad de ocurrencia) y la desviación estándar como 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.

Los resultados que se obtienen respectivamente son: 1; 0.50; 0.71

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 variable 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~B(n,p) . Una forma de escribir la ecuación matemática de la distribución binomial es:

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Donde k es el número de aciertos, n el número de ensayos o resultados del experimento, p la probabilidad de éxito y q 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: 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 7
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

Nota. Ramón, P., 2020.

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:

   O también podemos usar Excel y calcular usando el siguiente comando: =DISTR.BINOM.N(5;20;0.08;FALSO)

Y obtenemos el resultado: 0.0145

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 una de las cinco igualdades en la ecuación de la ley binomial, y finalmente sumar todas las respuestas. Esta operación la realizamos en Excel con el comando "DISTR.BINOM.N", de la siguiente forma:

DISTR.BINOM.N(4;20;0.08;VERDADERO)

Obtenemos el resultado: 0.98166

De esta forma se observa que hay una probabilidad del 98% de que menos de 5 frascos estén contaminados.

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

Así la probabilidad de identificar más de 5 frascos contaminados es de

1 – 0.98166 = 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).

$P(X \leq 5)$ = DISTR.BINOM.N(5; 20; 0.08; VERDADERO) = 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 se incluyen en el intervalo y en la operación. En Excel, esta operación se realiza de la siguiente forma:

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

P(X=4)= DISTR.BINOM.N(4; 20; 0.08; FALSO) = 0.0523

P(X=5)= DISTR.BINOM.N(5; 20; 0.08; FALSO) = 0.0145

P(X=6)= DISTR.BINOM.N(6; 20; 0.08; FALSO) = 0.0032

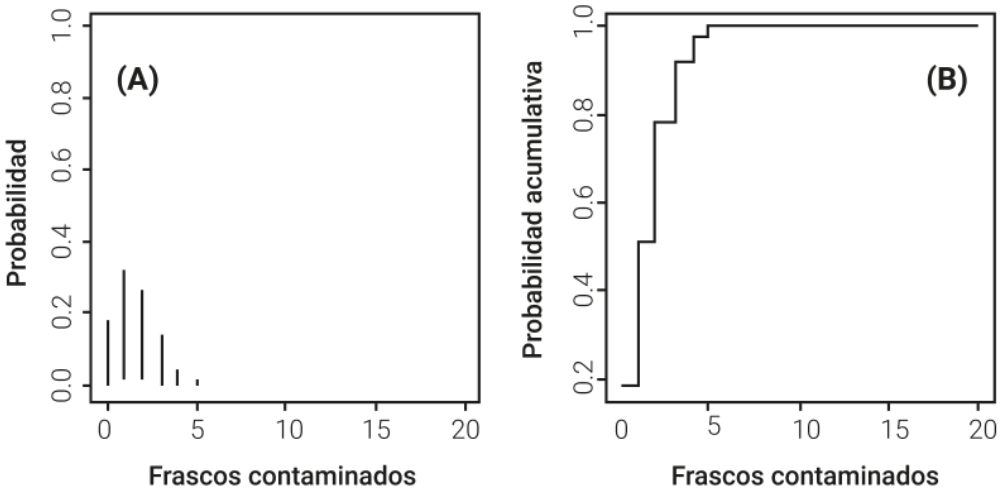
Entonces: $P(4 \leq X \leq 6) = 0.0523+0.0145+0.0032 = 0.07$

De esta forma se obtiene una probabilidad de 7% 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 variable X, asociada al experimento completo para un lote de 20 frascos observados. Esto lo realizamos de la siguiente forma:

Figura 20

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



Nota. Ramón, P., 2020.

En la Figura 20-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 20-B). Dicho en términos del problema, la probabilidad de identificar no más de 5 frascos contaminados es aproximadamente de 1 (o 100%).

Actividades de aprendizaje recomendadas

Estimado estudiante, le invito a realizar las siguientes actividades para reforzar los conocimientos de esta semana:

Actividad 1:

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

Ω={NNN, PNN, NPN, NNP, PPN, PNP, NPP, PPP}

Si la variable aleatoria es X= "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, misma que corresponderá al valor esperado de X.

Actividad 2:

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

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

Retroalimentación: Se sugiere primero simbolizar la variable aleatoria y luego la probabilidad que se busca en cada caso. Recuerde que el argumento lógico en Excel se modifica según el tipo de probabilidad, así para probabilidad acumulada será VERDADERO y para probabilidad simple o puntual será FALSO. Identifique los valores de los parámetros n y p antes de usar las fórmulas en Excel.

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 12

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

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 es 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: Mendenhall, W., Beaver, R. & Beaver, B. (2015). Introducción a la Probabilidad y Estadística. 14.ª edición. México: CENGAGE LEARNING.

De esta bibliografía, en el capítulo 5, sección 5.3, 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 el capítulo 5 de la bibliografía indicada puede encontrar la descripción y aplicación de algunas distribuciones discretas útiles.

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 el capítulo 6 de la bibliografía indicada anteriormente**.**

5.3 Distribución de Poisson

Otra de las distribuciones muy utilizadas para resolver problemas de 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 de 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 con 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 variables de tipo Poisson tiene las siguientes características fundamentales: (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*** es grande y la probabilidad de éxito (***p***) es pequeña (<10 %). 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. A continuación, la ecuación que rige al modelo de Poisson.

$$P(X = k) = \frac{(\lambda^k e^{-\lambda})}{k!}$$

En la ecuación de la distribución de Poisson se identifica el parámetro lambda (λ), el cual representa el valor esperado de la variable aleatoria 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: **λ** = np.

Una característica propia de la distribución de Poisson es que coincide con el valor esperado con la varianza de la variable. X, es decir: E[X] = V[X] = **λ.**

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

1. La probabilidad de 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 de la bibliografía de Mendenhall et al. (2015)).

En el agua de lastre que es descargada de un barco hay organismos con una concentración de 10 organismos/m^3, de acuerdo con un proceso de Poisson.

a. ¿Cuál es la probabilidad de que 1 m^3 de descarga tenga al menos 8 organismos?
b. ¿Cuál es la probabilidad de que el número de organismos en 1.5 m^3 de agua de descarga exceda su valor medio por más de una desviación estándar?
c. ¿Para qué cantidad de descarga la probabilidad de que haya menos de un organismo sería igual a 0.999?
d. Graficar la distribución de probabilidades de la variable aleatoria.

Solución:

Definimos la variable aleatoria X= "Número de organismos", y el valor medio λ = 10 por cada metro cúbico de volumen.

Para calcular la distribución de Poisson en Excel, podemos usar el comando POISSON.DIST().

A continuación, damos respuesta a cada ítem. 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:

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

Esta probabilidad se puede calcular en Excel de la siguiente forma:

$P(X < 8) = P(X \leq 7)$ = POISSON.DIST(7;10;VERDADERO) = 0.2202

b. $- P(X < 8) = 0.7798$.

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

El valor esperado es 15 y una desviación estándar es 3.87, podemos tomar el valor redondeado de 4. Así, simbólicamente lo que se quiere determinar es:

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

POISSON.DIST(19;15;VERDADERO) = 0.8752

$1 - P(X \leq 19) = 1 - 0.8752 = 0.1248$.

(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.

Implica hallar el valor de X tal que: $P(X < 1) = 0.999$

El valor será X=0.

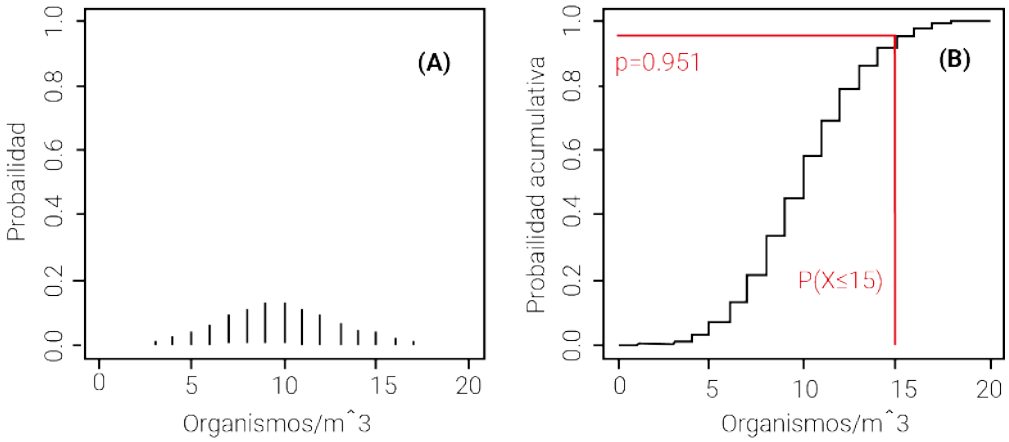
Cero organismos ocurren cuando tenemos 0.001 organismos en promedio, y esto se cumple para una descarga= $0.001/10 = 0.0001$ m^3.

(a) Finalmente, graficamos la distribución de probabilidades para un valor medio de 10 organismos por m^3.

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

Figura 21

(A) Función de probabilidad de Poisson de la v.a. X="Número de organismos por metro cúbico de descarga"; (B) Función de distribución acumulada de X



Nota. Ramón, P., 2020.

En la gráfica de la probabilidad absoluta (figura 21-A), se observa que la probabilidad de encontrar organismos en $1m^3$ 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 $1m^3$ de descarga es mayor al 95% (figura 21-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 la bibliografía**.**

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 variable 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.

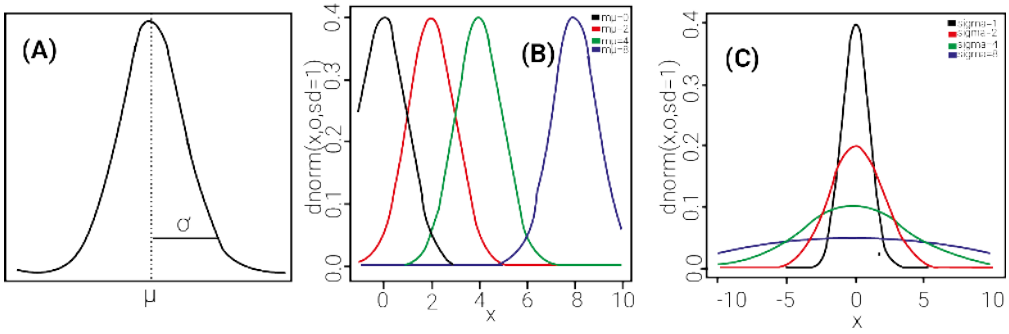
5.4 La distribución normal

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 árboles de cierta especie (altura, peso, diámetro, 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 de campana y por esto también se le denomina "campana de Gauss" (Figura 22-A).

Figura 22

(A) Curva normal N(μ,σ). (B) Efecto del parámetro de centralización (μ). (C) Efecto del parámetro de escala (σ)



Nota. Ramón, P., 2020.

La curva normal se caracteriza por sus parámetros poblacionales media **μ** y varianza σ^2 , 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 variable X sigue una distribución normal se emplea la expresión X~N(μ,σ).

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(60cm, 8cm), 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

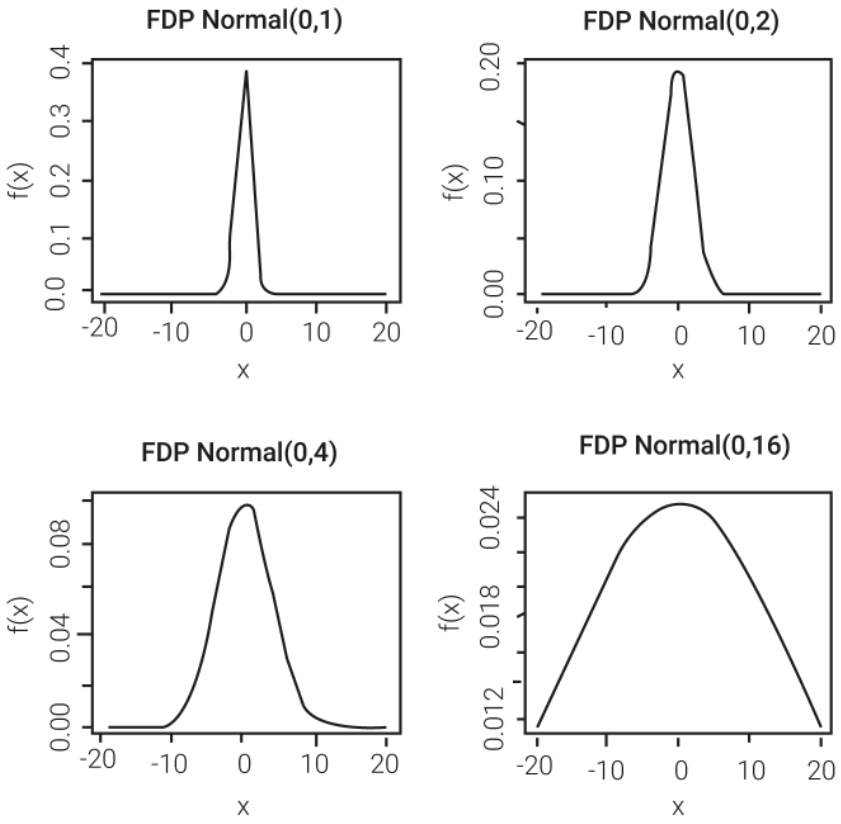
La distribución normal tiene propiedades muy particulares que la hacen importante al momento de realizar análisis estadísticos, entre ellas se destacan:

- Es simétrica respecto de la media aritmética, mediana y moda.
- Es asintótica respecto del eje X (o eje Z si se trata de la normal estándar)
- Puede tomar cualquier valor en los números reales (- ∞ , + ∞).
- Hay más probabilidad de ocurrencia para los valores próximos a la media aritmética.
- Conforme nos alejamos de la media aritmética, 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 22-B, 22-C).

Aunque la forma de la curva normal siempre es simétrica, sin embargo, no siempre tiene la misma dispersión (figura 23).

Figura 23

Curvas normales con igual media y dispersión variante



Nota. Ramón, P., 2020.

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 variable aleatoria 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: ¿cómo podemos obtener una distribución normal estándar?

Partimos de una distribución continua X (por ejemplo, el peso de los empleados de una empresa), 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:

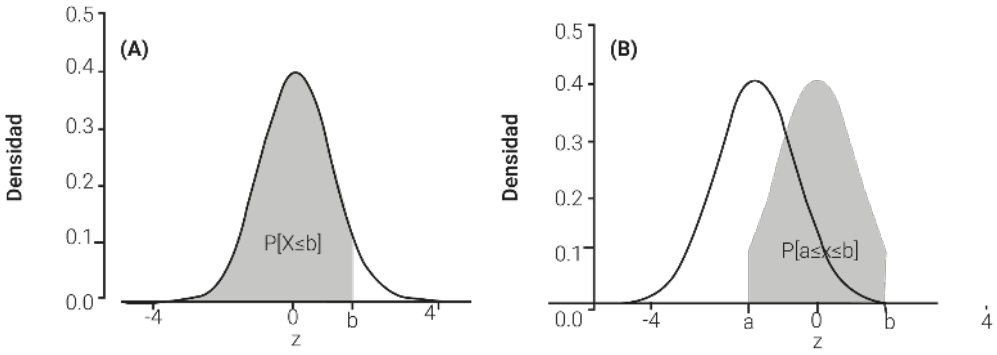
- Es la única distribución normal donde los parámetros son conocidos μ=0 y σ=1. Por ello generalmente se denota por N(0,1).
- Las unidades de medida de la variable normal estándar (Z) están dadas en términos de desviaciones estándar.
- El área total bajo la curva es igual a 1 (esto similar a otras distribuciones continuas).
- Esta distribución sirve para el cálculo de probabilidades.

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 24).

Figura 24

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]



Nota. Ramón, P., 2020.

En las figuras 24-A y 24-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.

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^{o}C y desviación estándar 5^{o}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^{o}
b. La probabilidad de que la temperatura en septiembre exceda los 21^{o}
c. La probabilidad de que la temperatura en septiembre esté comprendida entre 14 y 22^{o}
d. Halle el valor de la temperatura tal que solo el 20% de las observaciones excedan dicho valor.

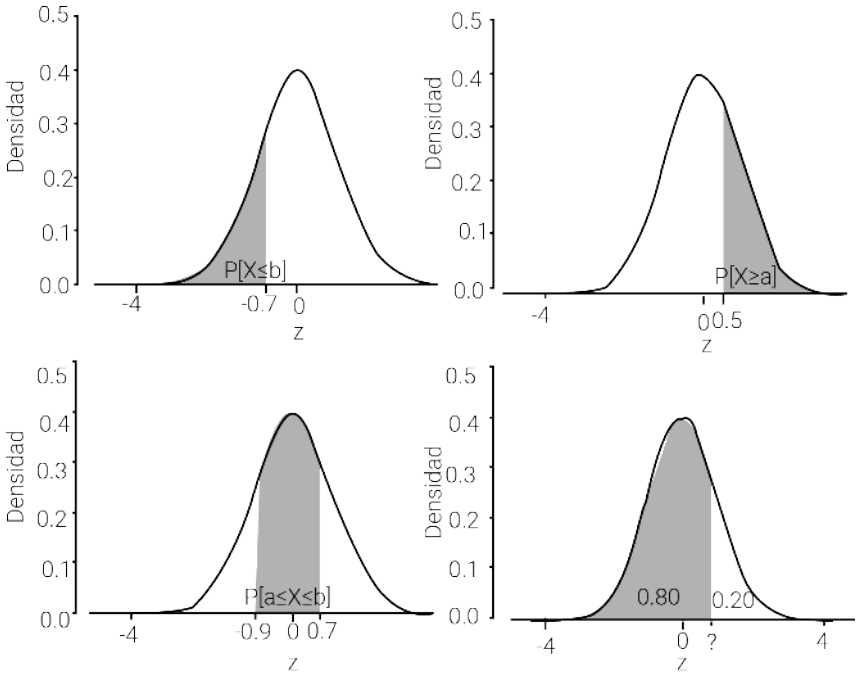
Solución:

Definimos la v.a. X= "Temperatura del mes de septiembre expresada en ^{o}C". O simplemente "Temperatura".

Hacemos un bosquejo (si es posible) del área correspondiente a la probabilidad que se desea calcular en cada literal (figura 25).

Figura 25

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



Nota. Ramón, P., 2020.

a. 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 ≤ 15)

   Usando Excel, realizamos el cálculo de la siguiente forma:

   DISTR.NORM.ESTAND.N(-0.7; VERDADERO) = 0.24

   Donde el valor -0.7 resulta de estandarizar x=15, así: (15-18.5)/5

   Por tanto, hay una probabilidad de 24% de que se presenten temperaturas menores o iguales a 15^{o}C durante septiembre.

   Importante: tener en cuenta que para el cálculo del área, se debe ingresar el valor de la variable X tipificado.

El valor -0.74 en la abscisa de la figura 25-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 o de la función de Excel.

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

Podría también simbolizarse P(X > 21) (figura 25-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≤ 21).

Estandarizamos el valor de x=21 y obtenemos Z=0.5

Luego en Excel: DISTR.NORM.ESTAND.N(0.5; VERDADERO) = 0.6915

Finalmente P(X > 21) = 1 – 0.6915 = 0.3085

Entonces, hay una probabilidad de 0.31 (redondeado a 2 decimales) de que se presenten temperaturas superiores a 21^{o}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 25-C):

P(X ≤ 22) = DISTR.NORM.ESTAND.N(0.7; VERDADERO) = 0.7580

P(X ≤ 14) = DISTR.NORM.ESTAND.N(-0.9; VERDADERO) = 0.1841

P(14 ≤ X ≤ 22) = 0.7580 – 0.1841 = 0.5739

Deducimos que es bastante probable (57%) que se registren temperaturas entre 14^{o}C y 22^{o}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 25-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 nos arroja el valor de 22.708

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

Actividades de aprendizaje recomendadas

Estimado estudiante, le invito a realizar las actividades que se describen a continuación:

Actividad 1:

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 μ=3.2 cm y desviación

estándar σ=1.8 cm. ¿Qué proporción de la población se esperaría que tuviera longitud de pétalo?

a. ¿Mayor a 4.5 cm?
b. ¿Superior a 1.78 cm?
c. ¿Entre 2.9 y 3.6 cm?
d. ¿Menor a 2 cm?
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. La P(X < k) en Excel usamos el comando DISTR.NORM.ESTAND.N(Z, VERDADERO), donde Z es el valor estandarizado de k. Para probabilidades del tipo P(X > k)=1-P(X<k), usamos la propiedad del complemento. Para probabilidades de intervalo P(a < X < b)=P(X < b) – P(X < a), por ser una variable continua.

- ***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 de Excel, debe simbolizar las probabilidades en cada literal. Tenga en cuenta cuando las preguntas incluyen la desigualdad absoluta, como, por ejemplo, en el literal (c).

Actividad 2:

Finalmente, para "cuantificar" el aprendizaje en el tema de las distribuciones probabilísticas discretas, le propongo responder la autoevaluación que se presenta a continuación .

Autoevaluación 5

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:

   a. $X \geq 0$
   b. $X = 0, 1, 2,\ldots$
   c. $X = 1, 2, 3,\ldots$

5. Una variable aleatoria establece una relación entre:

   a. Dos conjuntos cualesquiera de los números reales.
   b. Dos sucesos del espacio muestral.
   c. Elementos del espacio muestral con números reales.

En los enunciados del 6 al 10, escriba dentro de los 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: $X \geq 0$ mm.

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

Contenidos, recursos y actividades de aprendizaje recomendadas

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. Los animo a que trabajen resolviendo ejercicios y compartiendo entre compañeros sus ideas y dudas a través del entorno virtual canvas.

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

- **Mendenhall, W., Beaver, R., & Beaver, B. (2015). Introducción a la Probabilidad y Estadística. 14.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"** de la bibliografía indicada.

Durante el proceso de estimación surgen algunas interrogantes, por ejemplo: ¿cómo se calculan los límites de intervalo?, ¿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 la bibliografía complementaria deMendenhall et al. (2015)**.** 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 de la bibliografía complementaria de Mendenhall et al. (2015). En esa figura se identifican dos regiones: la región sombreada, denominada región crítica, y la región en blanco, denominada región de confianza.

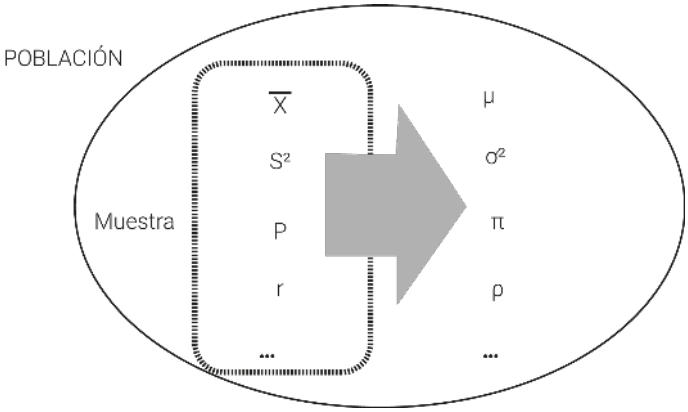
Para conocer más acerca de la distribución t-Student, le sugiero revisar el tema relacionado con intervalos **basados en una distribución de población normal.** Ahí usted puede establecer diferencias entre la distribución de Z y la distribución de t.

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 de 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 27), mientras que el contraste de hipótesis implica tomar una decisión con base en la prueba de un supuesto o afirmación acerca del parámetro.

Figura 26

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



Nota. Ramón, P., 2020.

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 profesional de salud 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 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

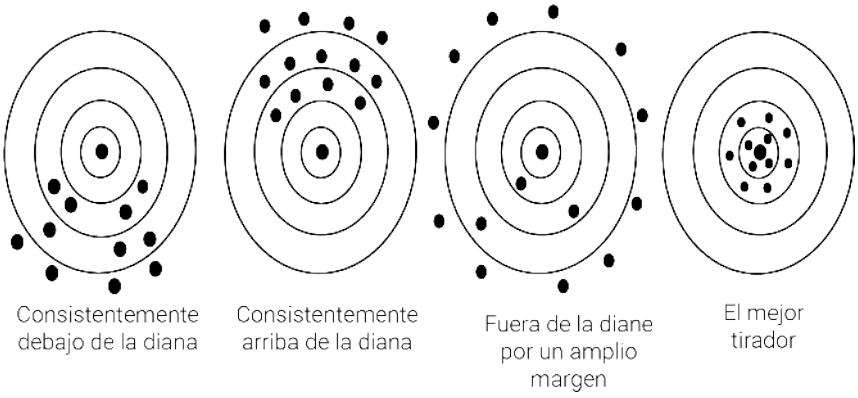
En la literatura estadística se definen básicamente dos formas diferentes de calcular estimadores, estas son: ***estimación puntual y estimación de intervalo***.

Les comento que todos los indicadores estadísticos que hemos generado en las unidades anteriores se tratan de estimaciones puntuales (la media aritmética, 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.

Figura 27

Semejanza entre la estimación puntual y el tiro al blanco



Nota. Tomado de Introducción a la probabilidad y estadística. 14ª edición (p. 284), por Mendenhall, W. y Beaver, B., 2015, México: CENGAGE Learning. CC BY 4.0.

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?

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, algunos autor esproponen una de las interpretaciones más aceptadas de intervalo de confianza, y la relacionan 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 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: $m \pm (Z*EE)$, donde m representa la media muestral y el error estándar de la media está dado por EE. 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 llamada t-Student con n-1 grados de libertad. Entonces el intervalo de confianza quedará definido como , $\bar{x} \pm (t*EE)$, 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 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.

Ejemplo 6.2.1

Utilizando la tabla de datos 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.

A continuación la versión de la solución mediante el uso de fórmulas estadísticas .

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:

m=5.84 cm, S = 0.828, n = 150, t = 1.976, EE = 0.0676

LI=5.84-(1.976*0.0676)=5.71cm LS=5.84+(1.976*0.0676)=5.98cm

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

De esto se puede concluir que con el 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.

Complete el ejercicio calculando el IC para el 99% de confianza. Para este caso deberá usar el valor t= 2.609

Note 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).

NOTA: para tamaños de muestra grandes (n>100), la puntuación t se aproxima a la puntuación Z.

En resumen: podemos emplear la puntuación Z cuando conocemos la varianza poblacional (es decir en pocas ocasiones), al contrario, empleamos la puntuación t cuando se desconoce la varianza poblacional y esta 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-α)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-α)100% de confianza que el error no excederá una cierta cantidad E, cuando el tamaño de muestra sea al menos:

$$n = (\frac{Z*\sigma}{E})^2$$

Donde 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?

$$n = (\frac{1.96*0.^28}{0.25}) = 39.3 \approx 39$$

Por lo tanto, se requiere al menos una muestra de tamaño 39 para una precisión máxima de 0.25cm.

Actividad de aprendizaje recomendada

Estimado estudiante, le invito a reforzar su conocimiento mediante el desarrollo de la siguiente actividad, la cual, 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 produjo 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 del 95 % para la vida útil promedio.
b. Obtenga un IC al 99 %.

Retroalimentación: si va a utilizar Excel, primero debe ingresar el vector de valores en una columna.

X= 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 columna de Excel, y está listo para utilizar.

Respuesta al 95 %: [1.92 – 6.50].

Complete el proceso para el 99 % de confianza.

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 14

Unidad 6. Estimación estadística - intervalos de confianza

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 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 la 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), queremos aproximar la proporción de ocurrencia de una determinada categoría.

Si queremos calcular un IC para la proporción poblacional empleando la distribución normal estándar Z, debemos utilizar la expresión siguiente:

$$\hat{p} \pm (Z * EE_p)$$
$$EE_p = \sqrt{\frac{(p*q)}{n}}$$

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 la versión de la solución, una mediante el uso de fórmulas.

Con el 99 % de confianza, calculamos el límite inferior y superior (LI, LS) respectivamente.

$$\hat{p} = \frac{24}{100} = 0.24,\ n = 100,\ Z = 2.58,\ EE_p = 0.0427$$
$$LI = 0.24 - (1.96 * 0.0427) = 0.13$$
$$LS = 5.84 + (1.96 * 0.0427) = 0.35$$

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 13% y 35%. Con el 1% de posibilidad de que dicho intervalo no contenga al valor de la proporción real.

Como podrán darse cuenta, la estimación para la proporción consiste en 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 se emplea como estimación de *p*, podemos tener (1-α)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 6.3.2 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 debería 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 realizar las siguientes actividades con el propósito de que le sirva como un indicador de aprendizaje y comprensión de las temáticas estudiadas.

Actividades de aprendizaje recomendadas

Actividad 1:

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 realizó un muestreo aleatorio de 60 vehículos y se observó que solamente 50 cumplieron con el requerimiento. Estime la proporción de unidades que cumplen con el requerimiento, al 95 % y el 99 % de confianza.

Retroalimentación: sea que resuelva mediante fórmulas o usando Excel, primero debe calcular los insumos 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 el límite inferior (LI) y superior (LS).

Actividad 2:

Con el propósito de fortalecer el fundamento teórico de los temas relacionados con estimación estadística mediante intervalos de confianza, se propone la siguiente autoevaluación.

Autoevaluación 6

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 % seguro que contenga a la proporción buscada.
   c. Se espera que el 90 % de las proporciones muestrales estimadas estén contenidas en el intervalo.

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.

10. ( ) Cuando no se conoce la varianza poblacional, para estimar la media utilizamos la puntuación normal estándar Z.

Ir al solucionario

Contenidos, recursos y actividades de aprendizaje recomendadas

Semana 15

Actividades finales del bimestre

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, evalúen su aprendizaje realizando la actividad suplementaria.

Actividades de aprendizaje recomendadas

Estimado estudiante, compilando todo lo revisado en las unidades 4, 5 y 6, se le propone realizar los siguientes ejercicios prácticos:

Actividad 1:

Ejercicio práctico 1 (probabilidad): una fábrica recibe lotes de material de tres proveedores en proporciones del 50 %, el 30 % y 20%. Se sabe que el 0.1 % de los lotes del primer proveedor es rechazado, 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 Venn o diagrama de árbol.

Actividad 2:

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 usando Excel; para ello, revise el ejercicio 5.3.1 resuelto en la guía didáctica.

Actividad 3:

Ejercicio práctico 3 (intervalos de confianza): en un reporte de prensa (El Comercio, 1 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. Puede apoyarse con el ejercicio resuelto en la unidad 6 de la guía didáctica.

Contenidos, recursos y actividades de aprendizaje recomendadas

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 completo en la guía didáctica y responda las autoevaluaciones.

4. Autoevaluaciones

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 con 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 direccionada, se denomina muestra de conveniencia.

Ir a la autoevaluación

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	Polígonos de frecuenias	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

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 el 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 solo 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

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 compartan elementos.
5	Producto	Sucesos independientes son equivalentes a sucesos excluyentes, y su intersección no involucra elementos. Por la regla del producto: P(A 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(Ac).
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

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 (≤ 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 = 38.5 mm.
10	F	Suceso raro es aquel que tiene pocas posibilidades de ocurrencia.

Ir a la autoevaluación

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

Atkinson, A.C. (1985) Plots, Transformations and Regression. Oxford University Press.

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). *Graph ical 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.

Diggle P. (1990) Time Series: A Biostatistical Introduction. Oxford Goos, P. y Meintrup, D. (2015). Statistics with JMP: graphs, Descriptive *Stati stics, and Probability*. Chichester: John Wiley & Sons Ltd.

Gutiérrez Banegas, A. (2012). Probabilidad y estadística. Enfoque por *c ompetencias*. 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. Ph armacol Pharmacoter, 2(2), 140-142.*

Mendenhall, W., Beaver, R. J. y Beaver, B. M. (2015). *Introducción a la probabilidad y estadística*. 14ª 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.

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.

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 Yersey: Pearson Prentice Hall.

6. Anexos

Anexo 1. Estadística usando Excel

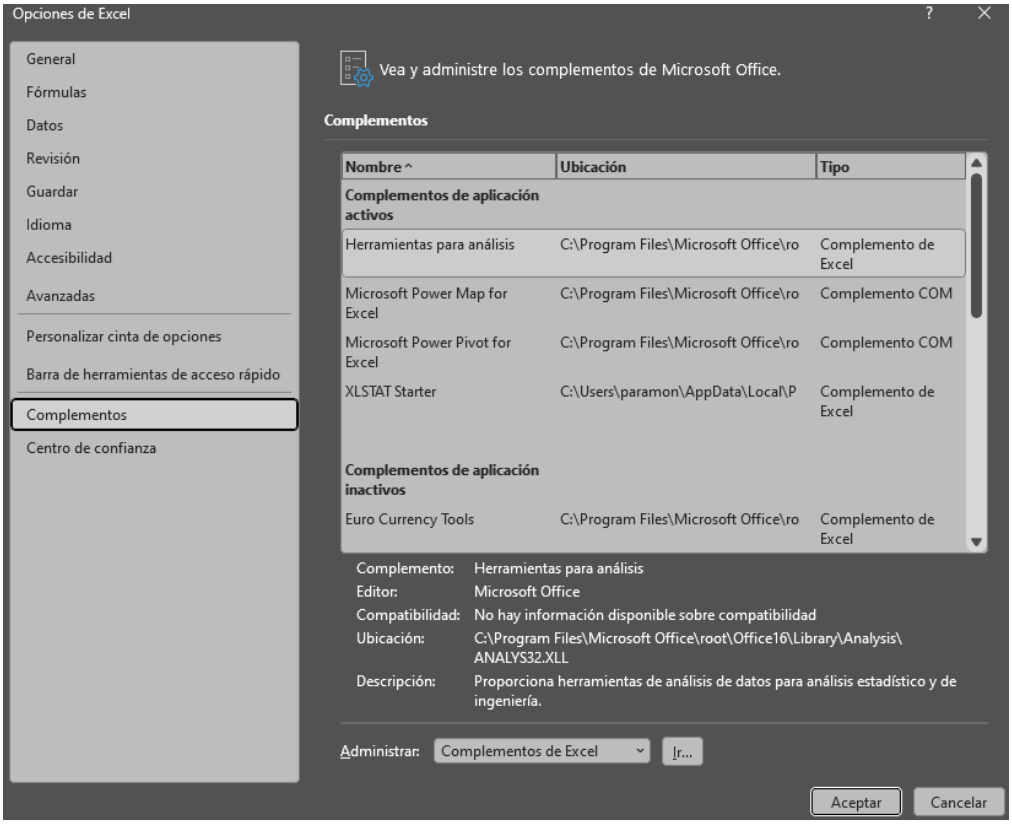
PRIMER BIMESTRE

Cálculo de estadísticos descriptivos usando Excel

(1) Debemos activar la barra análisis de datos en la sección “Datos” de la barra de tareas. En caso de que no esté activa, vamos a “Opciones” de Excel --- Complementos, y activamos “Herramientas para análisis” conforme se muestra en la figura a continuación:

Figura 1

Configuración de complementos en Excel

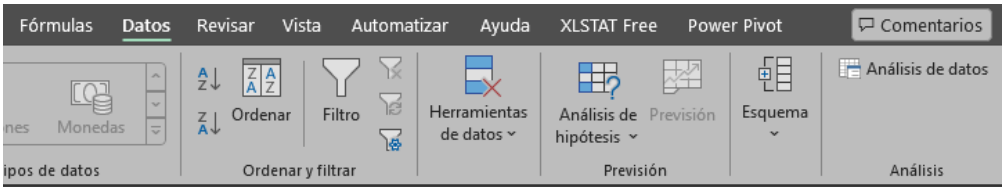


Nota. Elaboración propia.

(2) Luego en el menú datos deberá aparecer la barra de “Análisis de datos” conforme la figura:

Figura 2

Barra de herramientas de Datos en Excel



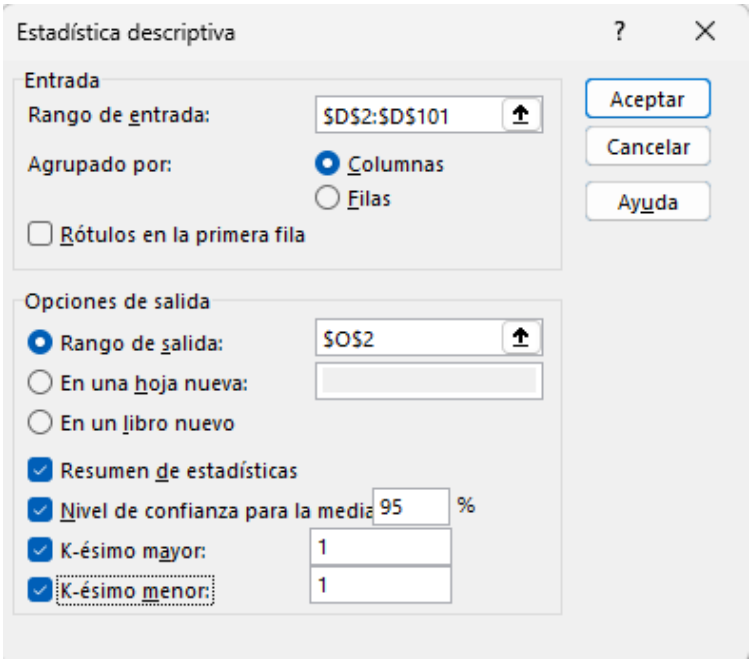
Nota. Elaboración propia.

(3) Clic en la barra y seleccionamos la opción "Estadística descriptiva" (ver figura a continuación), ahí seleccionamos los datos de entrada que deben estar en fila o columna, luego señalamos la celda donde van a salir los resultados (Rango de salida) y lo que deseamos imprimir (Resumen de estadísticas, etc.). Este resumen incluye:

Por ejemplo, las siguientes medidas: Media aritmética, mediana, moda, error típico, desviación estándar, varianza muestral, curtosis, coeficiente de asimetría, rango, mínimo, máximo, suma (total), cuenta (tamaño de muestra), etc.

Figura 3

Ventana de Estadística Descriptiva en Excel



Nota. Elaboración propia.

Cálculo de percentiles

En la opción “Insertar función”, usamos la función “PERCENTIL.EXC”. Esta función utiliza dos argumentos: los datos numéricos y el percentil expresado en fracción. Así, por ejemplo, para el percentil de orden 25 (o primer cuartil), cuyos datos están distribuidos en la columna D desde la celda D2 hasta la celda D101, sería: PERCENTIL.EXC(D2:D101; 0.25).

Tabla de frecuencias

1. Disponemos del vector de datos iniciales. Por ejemplo, la columna “Month” de la tabla “airquality”.

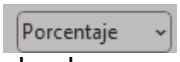
2. Determinamos las clases. En este ejemplo sería el vector de datos: 5,6,7,8,9.

3. Calculamos las frecuencias, ubicamos el cursor en la celda en la que deseamos los resultados.

Por ejemplo, en la columna C, en la celda C1, ubicamos la expresión: =FRECUENCIA(A2:A154; B2:B6).

4. Obtenemos los resultados, en este ejemplo, los valores: 31,30,31,31,30.

5. Calculamos las frecuencias relativas dividiendo cada frecuencia para el número total de datos, en este caso 153. Para este ejemplo usamos la fórmula “=C10/C15”. Usamos el doble signo de dólar para fijar el valor del total que está en la celda C15. Los resultados de este ejemplo serán: 0.20, 0.19, 0.20, 0.20, 0.19.

6. Convertimos a porcentaje los valores de la frecuencia relativa, seleccionamos la opción “Porcentaje” 

en la barra de menú superior, seleccionando la celda donde se van a imprimir los resultados.

UNIDAD 2: GRÁFICAS ESTADÍSTICAS

GRÁFICAS DE VARIABLES CATEGÓRICAS (CUALITATIVAS)

Gráfica de sectores (pastel)

1. Marcamos los datos de porcentajes de la tabla anterior, seleccionamos “Insertar” en la barra de menú, luego en la opción “*Gráficos*”—“*Gráfico 2D circular*”.

2. Podemos incluir los datos en la gráfica, clic derecho sobre la gráfica y “*Agregar etiqueta de datos*”, y tenemos una gráfica de sectores que incluye los porcentajes.

Prueba de hipótesis Chi-cuadrado:

1. En la barra de menú seleccionamos: Fórmulas - Más funciones – estadísticas – PRUEBA. CHICUAD. Seleccionamos las frecuencias observadas y luego las esperadas.

2. Con base en el valor p obtenido, concluimos si se rechaza o no la hipótesis nula.

Diagrama de cajas

1. Marcamos la columna de datos en la hoja Excel.

2. Insertar gráfico de estadística, en la sección donde está el histograma. En la parte inferior seleccionamos “Cajas y Bigotes”.

SEGUNDO BIMESTRE

En este bimestre abordamos las técnicas de inferencia como son estimación y prueba de hipótesis. Para ello es necesario revisar algunas operaciones básicas, por ejemplo, la función factorial. En Excel nos ubicamos en una celda vacía donde queremos el resultado y usamos la función “FACT”. Por ejemplo, el factorial de 5 sería: =FACT(5).

UNIDAD 4: PROBABILIDAD

Técnicas de conteo.

Número de combinaciones de 2 tomadas de un conjunto de 10 elementos:

=COMBINA(10; 2).

El número de permutaciones posibles de un grupo de 3 objetos en los que 2 se escogen:

=PERMUTACIONES(3; 2)

UNIDAD 5: DISTRIBUCIONES DE VARIABLES ALEATORIAS

DISTRIBUCIÓN BINOMIAL

PROBABILIDAD EXACTA O ABSOLUTA P(X=K), donde K es número entero, incluido el cero. Utilizamos la siguiente función:

=DISTR.BINOM.N(B3;B4;B5;FALSO).

Donde B3: celda que contiene el número de éxitos (K).

B4: celda que contiene el número total de ensayos n.

B5: celda que contiene el valor de la probabilidad de éxito, valor fraccionario en el intervalo [0,1].

FALSO: valor lógico que indica que se debe calcular la probabilidad absoluta. Por ejemplo, P(X=K).

PROBABILIDAD ACUMULADA P(X ≤ K) Utilizamos la siguiente función:

=DISTR.BINOM.N(B3;B4;B5;VERDADERO)

Aquí el valor lógico “VERDADERO” indica que el cálculo debe ser la suma de todas las probabilidades hasta el valor de K que está en la celda B3. Donde K=0,1,2,…,n.

PROBABILIDAD BINOMIAL EN UN INTERVALO [A; B]

A continuación les propongo dos formas equivalentes de ingresar la fórmula

Forma 1. Fijando las celdas de los valores de los parámetros.

=DISTR.BINOM.N(B8;B4;B5;FALSO)

La celda B8 indica el valor A de la serie A,...,B del cual se busca la probabilidad. Los parámetros de la binomial n y p están fijados en las celdas B4 y B5 respectivamente. El valor lógico FALSO se indica para que calcule la probabilidad absoluta en cada caso.

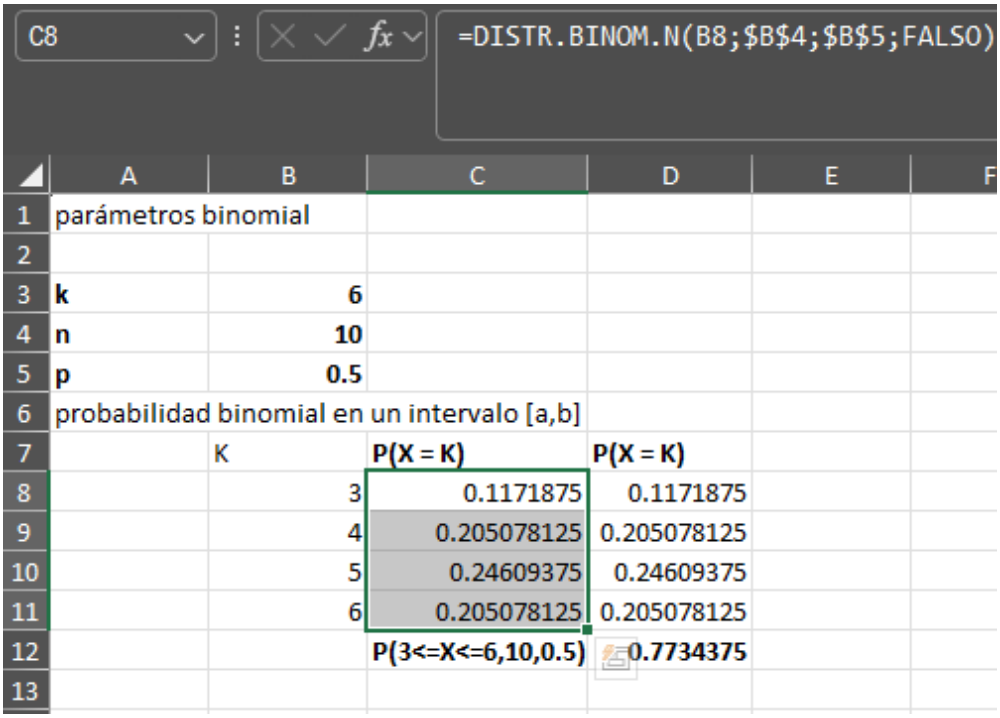
Forma 2. Ingresando el valor de los parámetros en la fórmula:

=DISTR.BINOM.N(B8;10;0.5; FALSO).

Los valores de 10 y 0.5 corresponden a n y p, respectivamente.

Figura 4

Cálculo de la probabilidad binomial en Excel, fijando las celdas donde están los valores de los parámetros n y p.



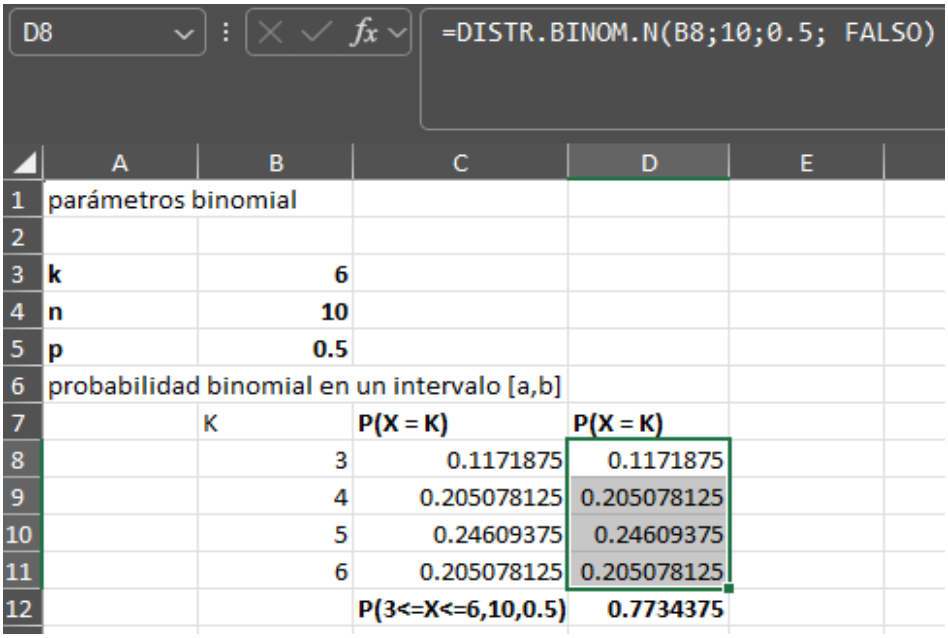
	A	B	C	D	E	F
1	parámetros binomial					
2						
3	**k**	**6**				
4	**n**	**10**				
5	**p**	**0.5**				
6	probabilidad binomial en un intervalo [a,b]					
7		K	**P(X = K)**	**P(X = K)**		
8		3	0.1171875	0.1171875		
9		4	0.205078125	0.205078125		
10		5	0.24609375	0.24609375		
11		6	0.205078125	0.205078125		
12			**P(3<=X<=6,10,0.5)**	**0.7734375**		
13						

Nota. Elaboración propia.

Forma 2: Describiendo los valores de los parámetros en la fórmula.

Figura 5

Cálculo de la probabilidad binomial en Excel, incluyendo los valores de los parámetros n y p en la fórmula.



	A	B	C	D	E
1	parámetros binomial				
2					
3	**k**	**6**			
4	**n**	**10**			
5	**p**	**0.5**			
6	probabilidad binomial en un intervalo [a,b]				
7		K	**P(X = K)**	**P(X = K)**	
8		3	0.1171875	0.1171875	
9		4	0.205078125	0.205078125	
10		5	0.24609375	0.24609375	
11		6	0.205078125	0.205078125	
12			**P(3<=X<=6,10,0.5)**	**0.7734375**	

Nota. Elaboración propia.

Al tratarse de eventos independientes, la probabilidad total será la suma de las probabilidades individuales. Así:

P(3 ≤ X ≤ 6) = P(X=3) + P(X=4) + P(X=5) + P(X=6).

DISTRIBUCIÓN DE POISSON

PROBABILIDAD EXACTA O ABSOLUTA P(X=k), donde k es número entero, incluido el cero.

Utilizamos la siguiente función:

=POISSON.DIST(B9; C3; FALSO).

Donde B9: valor de x para el cual se quiere calcular la probabilidad

C3: valor esperado (o media) de la variable aleatoria de Poisson (en este caso está fijado C3 porque se quiere calcular la probabilidad para algunos valores de x.

FALSO: para que imprima el valor de la probabilidad absoluta.

PROBABILIDAD ACUMULADA P(X ≤ k) Utilizamos la siguiente función:

=POISSON.DIST(B9; C3; VERDADERO).

Aquí el valor lógico “VERDADERO” indica que el cálculo debe ser la suma de todas las probabilidades hasta el valor k que está en la celda B9.

PROBABILIDAD DE POISSON EN UN INTERVALO [A; B].

=POISSON.DIST(B9;C3;FALSO).

Ejemplo: se desea calcular la probabilidad en el intervalo [1; 10], valores que se ubicaron en las celdas desde la B9 hasta la B18, y en la celda C3 el valor esperado (lambda). El valor lógico FALSO indica que se pide el valor de la probabilidad absoluta. Finalmente, la fórmula escrita en la celda C9 se arrastra hasta la celda C18 y se obtienen los 10 valores de probabilidad. Con estos dos vectores x,p se puede insertar una gráfica de barras verticales que representa la distribución de probabilidades de Poisson para este ejemplo en concreto.

DISTRIBUCIÓN NORMAL

PROBABILIDAD ACUMULADA P(X ≤ k) Utilizamos la siguiente función:

= DISTR.NORM.ESTAND.N(Z; VERDADERO).

Donde Z es el valor estandarizado de X=k, el valor lógico “VERDADERO” indica que el cálculo debe ser la suma de todas las probabilidades hasta el valor k.

Para probabilidades del tipo P(X > k) utilizamos la regla de complemento: 1-P(X ≤ k).

PROBABILIDAD NORMAL EN UN INTERVALO [A; B].

= DISTR.NORM.ESTAND.N(ZB; VERDADERO) – DISTR.NORM.ESTAND.N(ZA; VERDADERO).

Donde ZA y ZB son los valores estandarizados de A y B respectivamente.