Principios de Estadística

Manuel A. Zambrano-Monserrate
Alexia Berrús-Zhumi
Giuliana Goncalves Guillén

UEES

UEES
UNIVERSIDAD ESPÍRITU SANTO

PRINCIPIOS DE ESTADÍSTICA

Manuel Zambrano Monserrate
Alexia Berrús Zhumi
Giuliana Goncalves Guillén

2023

UNIVERSIDAD ESPÍRITU SANTO
Km. 2,5 Vía a Samborondón - Ecuador
Teléfono: (593-4) 5000950
ceninv@uees.edu.ec
www.uees.edu.ec

Autores:
Manuel Zambrano Monserrate
Alexia Berrús Zhumi
Giuliana Goncalves Guillén

Editor:
Fernando Espinoza Fuentes

Coordinadora editorial:
Natascha Ortiz Yánez

Cita:
(Zambrano Monserrate, Berrús Zhumi & Goncalves Guillén, 2023)

Referencia Bibliográfica:
Zambrano Monserrate, M., Berrús Zhumi, A. & Goncalves Guillén, G. (2023). Principios de Estadística. Universidad Espíritu Santo - Ecuador.

Portada:
Universidad Espíritu Santo

Diseño e Impresión:
TRIBU Soluciones Integrales
Urdesa Norte Av. 2da. #315
Teléfono: (593-4) 2383926
eperalta@tribuec.net

Edición:
Primera, Octubre 2023

ISBN-E:
978-9978-25-238-3

DEDICATORIA

A mis padres
Manuel

A mi familia
Alexia

A mi familia
Giuliana

TABLA DE CONTENIDO

PRÓLOGO

La estadística es la ciencia que se ocupa de recopilar, analizar e interpretar datos cuantitativos y cualitativos para obtener información significativa. Es una disciplina útil para comprender y describir el mundo que nos rodea, desde la investigación científica, hasta la toma de decisiones empresariales.

En estadística, se utilizan técnicas matemáticas y computacionales para recopilar y analizar datos que luego se interpretan para obtener conclusiones y tomar decisiones informadas. La estadística se utiliza en una amplia variedad de campos, incluyendo la medicina, la economía, la psicología, la ingeniería, la ciencia política, entre otros.

En este sentido, la presente obra pretende brindar las herramientas estadísticas básicas para la comprensión adecuada de distintos fenómenos. Analizamos diversas técnicas descriptivas como tabla de frecuencias y contingencia, así como medidas tendencia central y dispersión. Además, abordamos también la parte inferencial estudiando conceptos de probabilidad; acá, analizamos el teorema de Bayes y las distintas distribuciones de probabilidad.

En todos los capítulos presentamos las definiciones conceptuales de cada tema y ejercicios resueltos para su mejor comprensión. En el capítulo 2, inclusive, recomendamos algunas herramientas de Excel relacionadas a los conceptos analizados. Al final de los capítulos proponemos ejercicios complementarios para ser resueltos.

Este libro está dirigido a estudiantes, profesionales y demás personas interesadas en aprender las herramientas básicas de estadística para una mejor toma de decisiones.

Capítulo I

EL PAPEL DE LA ESTADÍSTICA

I. EL PAPEL DE LA ESTADÍSTICA

1.1 Definición de estadística

La estadística es una disciplina científica que tiene como objetivo principal mejorar la comprensión de los hechos a partir de información disponible. La estadística facilita la toma de decisiones y la solución de problemas en distintos ámbitos.

Por ejemplo, a partir de datos de contagio de algún tipo virus, organismos competentes podrían establecer medidas de contención contra la enfermedad. También, al gobierno le puede interesar conocer la evolución de la tasa de desempleo en los últimos meses para determinar si las políticas empleadas han sido efectivas o no. Los directivos de una empresa pueden estar interesados en conocer el promedio (la media) de ventas de los últimos meses, lo cual facilitaría el análisis sobre el cumplimiento de metas propuestas.

En resumen, la estadística es la «ciencia de los datos», pues a partir de ellos se pueden analizar y pronosticar fenómenos observados. El alcance de aquello depende del tipo de estadística que se utilice.

1.2 Tipos de estadística

Dentro de la estadística existen dos grandes divisiones: estadística descriptiva y estadística inferencial; esta última se divide a su vez en paramétrica y no paramétrica.

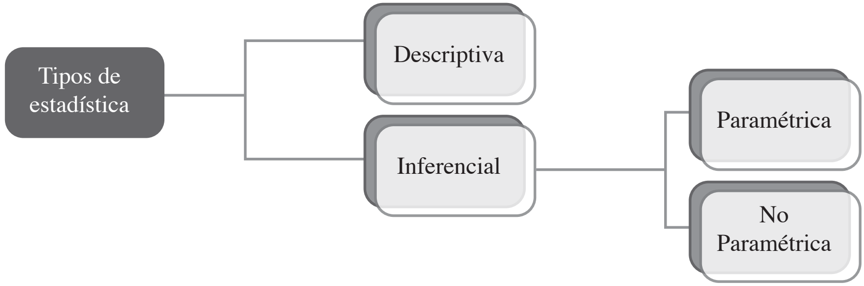


Gráfico 1.1 Tipos de estadística.

La estadística descriptiva es aquella que hace referencia a la recopilación, organización, síntesis y presentación de un conjunto de datos (Lind, Marchal, & Wathen, 2012). La estadística descriptiva se auxilia en gráficos, cuadros e indicadores para poder describir los aspectos más relevantes de los datos recolectados como su posición y dispersión. Algunas herramientas utilizadas por la estadística descriptiva incluyen la moda, la media, la mediana, tablas de frecuencias y contingencia, diagrama de barras *(gráfico 1.2)*, diagrama circular, entre otros.

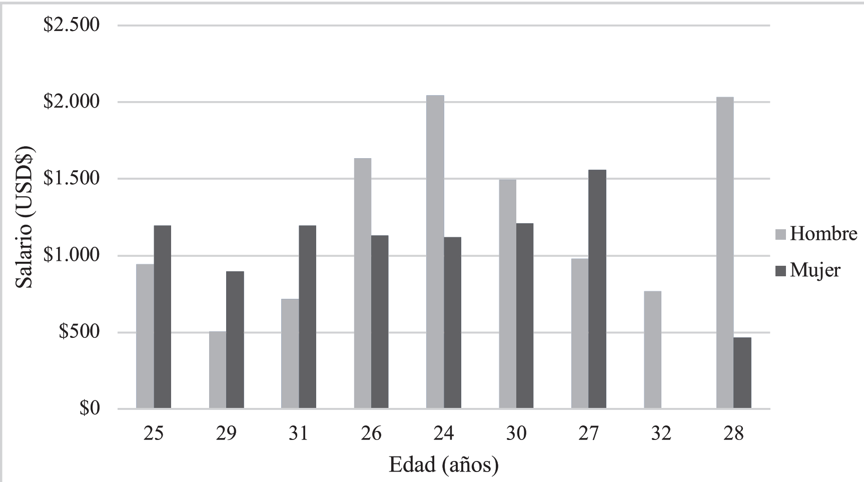


Gráfico 1.2 Salarios de hombres y mujeres por edades.

Por otro lado, la estadística inferencial va un paso más allá de la estadística descriptiva. Tiene como objetivo llegar a conclusiones sobre una población a partir de datos muestrales, usualmente. Permite obtener estimadores y probar hipótesis sobre ellos, es decir, mide la realidad de forma objetiva. La estadística inferencial implica todos aquellos métodos usados para estudiar las diferentes relaciones entre variables: diferencias entre grupos, asociación/regresión y causalidad.

Entre las subdivisiones de la estadística inferencial se encuentra la estadística inferencial paramétrica y la no paramétrica. La estadística paramétrica asume que los datos siguen una distribución específica y conocida. Por ejemplo, al realizar un análisis paramétrico se puede tomar el supuesto de que la población de análisis se distribuye de forma normal. En este caso habría que probar dicho supuesto y posteriormente obtener conclusiones bajo el cumplimiento de dicha premisa. Por otro lado, la estadística no paramétrica no asume a priori una distribución específica para los datos. En este sentido, es menos «exigente» que la estadística paramétrica, y normalmente se la toma como alternativa cuando los supuestos paramétricos no se cumplen. No obstante, y dependiendo del método no paramétrico, también se deben verificar ciertos supuestos para su aplicación.

1.3 Definiciones básicas de estadística

Al igual que en cualquier otra área de estudio o disciplina científica, la estadística cuenta con un propio vocabulario. Algunos de los términos más comúnmente utilizados en el análisis estadístico incluyen:

- Población y parámetros
- Muestras y estadísticos
- Variables

1.3.1 Población y parámetros

La población, en estadística, se refiere a la recolección completa de todas las observaciones de interés para el investigador. Por ejemplo, si se desea analizar datos sobre los fumadores en Ecuador, la población sería todos los fumadores del país. Por otro lado, los parámetros son medidas que ofrecen información sobre el centro de un conjunto de datos (medidas de tendencia central), sobre la dispersión o variabilidad (medidas de dispersión) o sobre la posición de un valor (medidas de posición como los percentiles). Lo esencial es comprender que un «parámetro es una característica numérica de una población» (Anderson, Sweeney, & Williams, 2008). Algunos ejemplos de parámetros son la mediana del salario de todos los trabajadores del sector privado en un determinado país o, la desviación típica entre los sueldos de los empleados del sector privado en Ecuador.

1.3.2 Muestras y estadísticos

En estadística la muestra es una porción o parte representativa de la población que es escogida para ser estudiada debido a que la población es demasiado grande, lo que dificulta estudiarla en su totalidad.

Gráfico 1.3 Muestra y población.

Un estadístico es aquel valor que describe alguna característica de la muestra de estudio y sirve como una estimación del parámetro de la población correspondiente. En términos simples, el estadístico es a la muestra lo que el parámetro es a la población.

1.3.3 Variables

Las variables son aquellas características o cualidades de una muestra o población a la cual se le puede aginar un valor. Existen algunas clasificaciones de variables.

a. Variables cuantitativas y variables cualitativas

Existen dos categorías principales de variables: cuantitativas y cualitativas. Las variables cuantitativas son aquellas que se pueden expresar en términos numéricos, y pueden ser de dos tipos: continuas o discretas. Una variable continua tiene un número infinito de valores o valores que son difíciles de contar, como la cantidad de granos de azúcar en una bolsa de 300 gramos. En cambio, una variable discreta tiene un número finito de valores, como la cantidad de estudiantes en una clase de estadística. Es importante destacar que, el concepto de variables continuas o discretas no debe asociarse a si la variable se representa por valores enteros o decimales.

Las variables cualitativas, por otro lado, no se pueden expresar en términos numéricos y se dividen en dos tipos: dicótomas y policótomas. Las variables dicótomas solo tienen dos valores, como el sexo de una persona (hombre o mujer); mientras que las variables politómicas pueden tener tres o más valores, como el nivel de educación de una persona (primaria, secundaria, universitaria, posgrado). En el grafico 1.4 se resume esta clasificación.

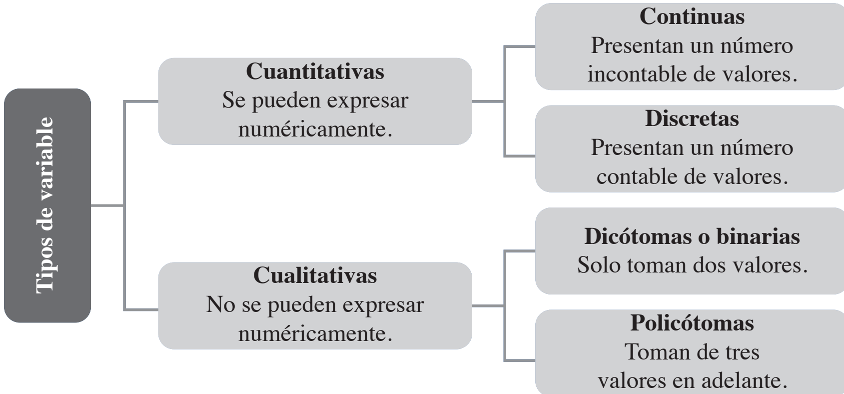


Gráfico 1.4 Variables cuantitativas vs variables cualitativas.

b. Variables nominales, ordinales, intervalo y de razón

A las variables también se las puede clasificar por su naturaleza. Primero, están las variables nominales. Estas son variables de carácter cualitativo y su categorización no obedece a un orden específico. Por ejemplo, el sexo o la nacionalidad de una persona. Por otro lado, las variables ordinales son cuasi cuantitativas y se caracterizan porque sus valores representan categorías que siguen algún orden, por ejemplo, el nivel de educación o nivel socioeconómico de una persona. Finalmente, las variables de intervalo y razón, ambas cuantitativas, se expresan numéricamente. Por un lado, en las variables de intervalo, el cero no indica la presencia o ausencia de un atributo (simplemente es un punto de referencia), por ejemplo, la medición de temperatura. En otros casos, el cero no se incluye, como en la medición del cociente intelectual. Contrariamente, en las variables de razón, el cero tiene relevancia y toma un valor absoluto, por ejemplo, la edad, estatura o peso de un individuo.

CARACTERÍTICAS	Cualitativas	Cuasi cuantitativas	Cuantitativas	
	Nominal	Ordinal	Intervalo	Razón
Para qué sirve	Categorizar	Ordenar	Medir por intervalos iguales	Todas las anteriores
Propiedades numéricas	De clasificación	De ordenamiento	De medición sin cero absoluto	Todas las anteriores, pero con cero absoluto.
Ejemplos	Sexo, Nacionalidad, preferencia política	Nivel socioeconómico, nivel de educación, etapa del desarrollo.	Temperatura, nivel de ansiedad, autoestima, cociente intelectual.	Edad, estatura, peso.

Tabla 1.1 Clasificación y ejemplos de variables.

1.4 Tipo de relación entre variables

Como se mencionó previamente, la estadística inferencial abarca, en forma general, el estudio de tres tipos de relación entre variables: diferencias entre grupos, asociación/regresión y causalidad.

La diferencia entre grupos se refiere a determinar si existe o no una diferencia estadísticamente significativa entre dos grupos o más, en relación a una variable de interés. Por ejemplo, se puede determinar si existen diferencias entre hombres y mujeres en relación a los promedios obtenidos en un curso de estadística. En este caso, los grupos son los «hombres» y las «mujeres» y, la variable de interés, los «promedios». Se pueden tener «n» grupos y la variable de interés puede ser cuantitativa o cualitativa. Las pruebas para determinar las diferencias entre los grupos pueden ser paramétricas y no paramétricas.

Por otro lado, los estudios de asociación/regresión representan un nivel más alto que el análisis de diferencias entre grupos. Por un lado, en los estudios de asociación o correlación se busca determinar si dos o más variables se relacionan significativamente. Por ejemplo, se puede estar interesado en determinar el grado de correlación entre las calificaciones de estadística 1 y estadística 2. Se esperaría que la correlación sea positiva (los que obtuvieron calificaciones más altas en estadística 1 seguramente también tendrán calificaciones más altas en estadística 2 y viceversa). Las variables a correlacionar pueden ser cuantitativas o cualitativas y se pueden utilizar técnicas paramétricas y no paramétricas para determinar la significancia de la correlación. Por otro lado, en el análisis de regresión se busca habitualmente estimar o predecir el valor promedio de una variable llamada «dependiente» en función de otras variables llamadas «independientes». Por ejemplo, se puede estar interesado en analizar el consumo de las familias en función de los ingresos familiares, los miembros del hogar, la educación promedio de la familia y la zona de residencia (urbana o rural). Para estimar esta relación se pueden utilizar métodos paramétricos (como el análisis de regresión por Mínimos Cuadrados Ordinarios) o no paramétricos.

Finalmente, se encuentran los estudios de causalidad. Dentro del análisis inferencial, este es el nivel más alto y, por tanto, el más complejo. Causalidad en términos simples significa que una variable influye directamente en el comportamiento de otra. Esto no es fácil de determinar, ya que, el comportamiento de una variable puede estar afectado por muchas variables a la vez. El mayor reto consiste en aislar el efecto de otras variables y cuantificar la influencia de una sola. Para esto se han desarrollado métodos cuasiexperimentales como «diferencias en diferencias» o «regresiones discontinuas». En este punto es importante señalar que correlación no implica necesariamente causalidad. Dos variables pueden estar correlacionadas fuertemente pero no necesariamente implicar causalidad. En el ejemplo de las calificaciones de estadística 1 y 2, puede existir una correlación fuerte entre ambas variables, pero las calificaciones del curso de estadística 2 pueden no necesariamente depender de las calificaciones del curso estadística 1. El rendimiento académico se explica por muchos otros factores como la motivación, la educación de los padres, el índice de inteligencia, entre otros. Comprender que correlación no implica necesariamente causalidad evita que lleguemos a conclusiones equivocadas en el análisis inferencial.

1.5 La importancia del muestreo

Como se mencionó previamente, obtener información de toda la población es tarea difícil. Por tanto, las muestras constituyen en los principales insumos para el análisis de datos. En este sentido, la información más importante de la muestra estará contenida en los estadísticos. Para obtener un buen estadístico, entre otros factores, la muestra debe ser representativa de la población. Consecuentemente, el proceso de muestreo es clave.

Existen dos tipos principales de muestreo: probabilístico y no probabilístico. En el muestreo probabilístico, todos los objetos o individuos tienen la misma probabilidad de ser seleccionados. Existen varios métodos de muestreo probabilístico, como: aleatorio simple, sistemático, estratificado y conglomerado. Por otro lado, en el muestreo no probabilístico, los objetos o individuos no son seleccionados de manera aleatoria. Algunos de los métodos de muestreo no probabilístico más conocidos son el propositivo, por cuotas, por conveniencia y bola de nieve. Sin embargo, una gran desventaja del muestreo no probabilístico es que los resultados no se pueden generalizar ni formular teorías a partir de ellos, esto debido a que la muestra generalmente no es representativa.

Por otro lado, para determinar el tamaño de la muestra se debe definir previamente el tamaño de la población universo, ya que para cada estudio existe un distinto «tamaño muestral idóneo» (Mira, Gomez, Aranaz, & Perez, 1997). Se pueden emplear fórmulas para población finita o infinita. La definición de estas fórmulas está fuera del alcance de este libro, por tanto, se recomienda literatura especializada de consulta como Malhotra (2008).

Ejercicios propuestos del capítulo

Ejercicio 1.1 Clasifique las siguientes variables como continuas, discretas, dicótomas o policótomas:

a. Los granos de azúcar en una bolsa de 1kg.

b. ¿El café con o sin azúcar?: SI, NO

c. Situación laboral actual: empleado, desempleado, jubilado, inactivo

d. La temperatura: 30.8 °C; 35.9 °C.

e. Número de dormitorios en su casa: 2, 3 ,4.

f. La distancia de su casa a la universidad: 20.3 km; 15.5 km.

g. Número de personas en un centro comercial: 100, 150, 200.

h. Estado civil: soltero, casado, viudo, separado.

i. Gusto por las películas de terror: SI, NO.

Ejercicio 1.2 En estadística, ¿Cuál es la diferencia entre población y muestra?

Ejercicio 1.3 Mencione dos de las herramientas aplicadas en la estadística descriptiva.

Ejercicio 1.4 Responda Verdadero (V) o Falso (F) según corresponda:

a. La estadística descriptiva se divide en paramétrica y no paramétrica.( )

b. El estadístico es a la muestra lo que el parámetro a la población. ( )

c. La temperatura es un ejemplo de una variable nominal. ( )

d. Las variables dicótomas son aquellas que toman más de dos valores. ( )

e. Una de las herramientas usadas en estadística inferencial son los diagramas de barras. ( )

Ejercicio 1.5 Describa la principal diferencia entre estadística descriptiva e inferencial.

Ejercicio 1.6 ¿Cuáles son los tipos de muestreo probabilístico más conocidos?

Ejercicio 1.7 ¿Cuáles son los tipos de muestreo no probabilístico más conocidos?

Ejercicio 1.8 Mencione 5 ejemplos de variables ordinales.

Ejercicio 1.9 Explique la diferencia entre variables de intervalo y razón.

Ejercicio 1.10 A través de un ejemplo realista, explique cómo la estadística puede ayudar a comprender de mejor forma un fenómeno.

Capítulo II

DESCRIPCIÓN DE UN CONJUNTO DE DATOS

II. DESCRIPCIÓN DE LOS CONJUNTOS DE DATOS

2.1 Introducción

Para realizar investigación científica, el primer requisito es la recolección de datos de interés, el procesamiento de los mismos y su análisis posterior. Los datos que se recolectan son mediciones que suelen derivarse usualmente de observaciones reales obtenidas al muestrear una población.

Los datos recolectados ya sean cualitativos o cuantitativos, como lo vimos en el primer capítulo, pueden provenir de distintas fuentes, como encuestas, revisiones bibliográficas, fuentes gubernamentales, entrevistas, cuestionarios, etc. La importancia radica en la organización, distribución y descripción que le demos; desde el punto de vista de la estadística, los datos se deben agrupar de tal forma que esa agrupación de sentido a la idea que queremos plantear.

En este capítulo, veremos distintas herramientas para organizar, modificar y simplificar los datos de una población/muestra tanto de forma tabular y/o gráfica. Estas diversas técnicas facilitan una mejor comprensión e interpretación de los datos para la toma de decisiones. En el gráfico 2.1 se resumen estas herramientas.

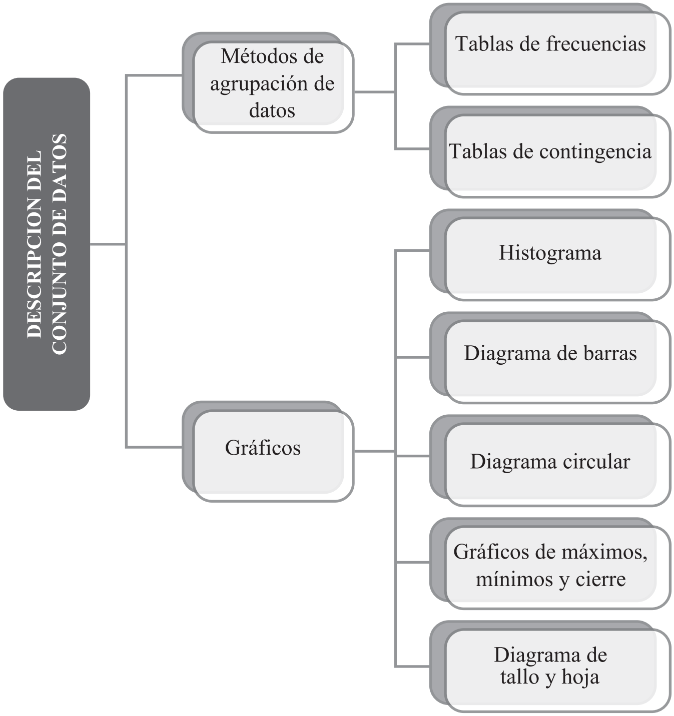


Gráfico 2.1 Métodos para la descripción de los conjuntos de datos.

2.2 Métodos de agrupación de datos

En los métodos para organizar datos estadísticos se pueden emplear una gran variedad de herramientas para describir y resumir un gran conjunto de datos; una de las más simples es la serie ordenada. Con esta herramienta los datos se deben colocar en orden ascendente o descendente de forma que se pueda determinar valores extremos, como el mínimo y el máximo. A pesar de esto, la información que proporciona una serie ordenada es limitada, ya que en algunas situaciones un ordenamiento de datos no resultará muy útil.

***Ejemplo 2.1*:**

Se tiene diez valores de puntuaciones de examen de estadística:

7 5 2 6 9 10 4 20 15 11

Al ordenar de forma ascendente se tiene:

2 4 5 6 7 9 10 11 15 20

Esto permite ver claramente cuál fue el valor máximo **(20)** y el valor mínimo **(2)**.

2.3 Tablas de frecuencias

Una tabla de frecuencia permite resumir cualquier tipo de variable cuantitativa. La estructura de una tabla de frecuencias consta de una serie de filas y columnas, que permiten analizar a profundidad los valores de la variable de interés (Salazar y Castillo, 2018). En pocas palabras, una tabla de frecuencias es una herramienta para organizar los datos de manera que las características de la distribución de un conjunto de datos o muestras se presentan numéricamente, este proceso se denomina tabulación.

2.3.1 Elementos de una tabla de frecuencias

Antes de construir una tabla de frecuencias, se debe definir los elementos que la conforman.

- ***Rango***

  Es un valor numérico que indica la diferencia entre el valor máximo y el mínimo de una población o muestra estadística. Este proporciona una idea de la dispersión de los datos, cuanto mayor es el rango, más dispersos estarán.

  Rango = Valor Máximo - Valor Mínimo

- ***Número de clases***

  Es la división de la muestra en diferentes rangos, es decir, es el número de intervalos o clases en las que están distribuidos los datos. El número de clases puede ser arbitrario, sin embargo, se pueden seguir ciertas reglas como de *Sturges*. La regla de *Sturges* es un método empírico muy utilizado para determinar el número de clases que deben existir en una tabla de frecuencias. Para determinar el número de clases a través de la regla de *Sturges* se sigue la siguiente fórmula:

  Número de clases = 1+3,3*[Log(Datos)]

Donde «Datos» es la cantidad de observaciones.

- ***Amplitud***

Se refiere al ancho del número de clases. Se la calcula de la siguiente forma:

$$Amplitud = \frac{Rango}{Número\ de\ clases}$$

- ***Marca de clase (mi)***

Es el promedio entre el intervalo inferior y superior de cada clase. Se la calcula de la siguiente forma:

$$Marca\ de\ clase\ (mi) = \frac{(Intervalo\ inferior + Intervalo\ superior)}{2}$$

- ***Frecuencia absoluta (ni)***

Indica cuántos valores del conjunto de datos hay en cada una de las clases en las que se han dividido las observaciones.

- ***Frecuencia absoluta acumulada (Ni)***

Indica cuántos valores del conjunto de datos se van acumulando en cada clase.

- ***Frecuencia relativa (fi)***

Indica el porcentaje del conjunto de datos que se distribuyen en cada clase. Se la calcula de la siguiente forma:

$$fi = \frac{Frecuencia\ absoluta}{\#Total\ de\ datos\ de\ la\ población\ o\ muestra}$$

- ***Frecuencia relativa acumulada (Fi)***

$$Fi\ (\%) = \left(\frac{Frecuencua\ acumulada}{\#\ total\ de\ datos\ de\ la\ población\ o\ muestra}\right) * 100$$

Indica el porcentaje del conjunto de datos que se va acumulando en cada clase.

Número de clases	Intervalo Inferior (I)	Intervalo Superior (S)	Marca de clase (mi)	Frecuencia absoluta acumulada (Ni)	Frecuencia absoluta (ni)	Frecuencia relativa (fi)	Frecuencia relativa acumulada (Fi)	Marca de clase (mi)
clase 1								
clase 2								
clase 3								
…								
clase k								

Tabla 2.1 Elementos de una tabla de distribución de frecuencias.

Ejercicio 2.1 Elaboración de una tabla de frecuencia para variables discretas.

Imagine que se tiene información de las calificaciones de estadística de 50 estudiantes de UEES.

1. *¿Cuáles son las calificaciones máximas y mínimas del curso?*
2. *Calcule e interprete el rango.*
3. *Interprete las frecuencias absolutas y relativas obtenidas. ¿Qué se puede concluir?*

68	71	77	83	79
72	74	57	67	69
50	60	70	66	76
70	84	59	75	94
65	44	85	79	45
83	84	74	35	97
77	73	78	93	95
78	81	79	90	83
80	84	91	100	40
93	92	99	80	69

Tabla 2.2 Datos brutos sobre las calificaciones de estudiantes de UEES.

1. Máximo y mínimo

Para resolver este ejercicio se utilizó la herramienta Excel. Los datos no se encuentran ordenados, pero no influye con el desarrollo, ya que con la función **CONTAR** nos dará la cantidad de datos en cada categoría; también se utilizará las funciones de **MÍNIMO** y **MÁXIMO** para la obtención del rango.

Valor Máximo = *100*

Valor Mínimo = *35*

Datos = *50*

2. Determinación del rango

Para el cálculo del rango y el número de clases se utiliza la fórmula previamente señalada.

Rango = 100 - 35

Rango = 65

3. Número de clases (k)

Aplicando la regla de *Sturges:*

*Número de clases = 1 +3.3*log (50)*

Número de clases = 6.60

Número de clases = 7

Para el número de clases es recomendable redondear el número con la función REDONDEAR.MAS de Excel.

4. Amplitud de clases

Para la amplitud, dividimos el rango para el número de clases y obtenemos el valor de 10. Luego, para la primera clase, se escoge como intervalo inferior el número 35 (que es el valor mínimo de la serie de dato) y a ese valor se le suma la amplitud o el ancho que es 10, dándonos un valor de 45, este valor será el intervalo superior de dicha clase. Para la siguiente clase, tomamos el intervalo superior de la anterior y sumamos nuevamente 10. Este proceso se repite hasta llegar a la séptima clase.

Número de clases	Intervalo Inferior (I)	Intervalo Superior (S)
1	35	45
2	45	55
3	55	65
4	65	75
5	75	85
6	85	95
7	95	105

Tabla 2.3 Amplitud de clases.

5. Distribución de frecuencias

Para calcular las frecuencias, primero es necesario que estén seleccionados todas las celdas de los datos brutos, posteriormente, con la función **FRECUENCIAS,** calculamos la frecuencia absoluta acumulada (Ni). La frecuencia absoluta la calculamos al ir restando las frecuencias absolutas acumuladas de las observaciones o clases de las muestras, y se la representa por las siglas ni. Para las frecuencias relativa y relativa acumulada se utilizan las fórmulas expuestas anteriormente.

Número de clases	Intervalo Inferior (I)	Intervalo Superior (S)	Frecuencia absoluta acumulada (Ni)	Frecuencia absoluta (ni)
1	35	45	4	4
2	45	55	5	1
3	55	65	9	4
4	65	75	22	13
5	75	85	40	18
6	85	95	47	7
7	95	105	50	3

Tabla 2.4 Cálculo de frecuencias.

Número de clases	Intervalo Inferior (I)	Intervalo Superior (S)	Frecuencia absoluta acumulada (Ni)	Frecuencia absoluta (ni)	Frecuencia relativa (fi)	Frecuencia relativa acumulada (Fi)	Marca de clase (mi)
1	35	45	4	4	8%	8%	40
2	45	55	5	1	2%	10%	50
3	55	65	9	4	8%	18%	60
4	65	75	22	13	26%	44%	70
5	75	85	40	18	36%	80%	80
6	85	95	47	7	14%	94%	90
7	95	105	50	3	6%	100%	100

Tabla 2.5 Distribución de frecuencia (absolutas, acumuladas y relativas) de la variable: Calificación de estudiantes de UEES.

Con esta información se pueden responder los enunciados propuestos. La frecuencia absoluta nos indica que hay 7 datos en la clase número 6, y la frecuencia absoluta acumulada indica que existen 22 datos acumulados hasta la clase 4. En la clase 5 se encuentra el mayor porcentaje de los datos con un 36%, y hasta la clase 2 están acumulados el 10% de los datos.

2.4 Tablas de contingencias

Como se mencionó previamente, las tablas de frecuencia sirven como un ordenamiento o recuento de datos de una sola variable. No obstante, si se desea examinar o comparar dos variables, sobre todo variables cualitativas (nominal u ordinal), una tabla de contingencia es lo más ideal. La tabla de contingencia o también llamada tabla cruzada, es una de las técnicas estadísticas más utilizadas para resumir datos cualitativos o categóricos. Por ejemplo, a un estudiante de UEES se le puede preguntar qué tipo de comida prefiere, comida rápida o saludable, y con la identificación de su género podemos realizar un desglose de las variables para hacer una comparación entre sus categorías.

El objetivo de este método estadístico es que, al comparar las variables mediante la distribución porcentual, podremos determinar si las variables están relacionadas, es decir, se trata de analizar si la distribución porcentual de una variable en las categorías de otra variable se repite de forma idéntica. Debido a que una variable se estudia en relación con otra, se debe diferenciar cuál es la variable dependiente e independiente. Esta distinción es importante porque las variables independientes se colocan en columnas, mientras que las variables dependientes se colocan en filas (Cárdenas, 2015). Para la resolución de un ejercicio de tablas de contingencia se utiliza la opción de tabla dinámica de Excel.

Ejercicio 2.2

Imagine que aparte de las calificaciones de estadística de 50 estudiantes de UEES se tiene información del género de la persona y su modalidad de estudio. Con la información de estas variables y a través de una tabla de contingencia, se pide responder las siguientes preguntas:

1. *¿Cuál es el perfil de estudiante más representativo de la muestra?*

Modalidad de estudio	**Género**		**Total, general**
	Femenino	***Masculino***	
Presencial	44%	24%	68%
Virtual	8%	24%	32%
Total, general	**52%**	**48%**	**100%**

Tabla 2.6 Tabla de contingencia del género y modalidad de estudio de los estudiantes de UEES.

Los resultados de la tabla 2.6 muestran que, de los 50 estudiantes, el 68% estudia presencialmente. El perfil más representativo de la muestra son las estudiantes del género femenino que estudian presencialmente (44%).

2. *¿Quiénes obtienen las calificaciones más altas, hombres o mujeres? ¿y las más bajas?*

Calificación	Género		Total, general
	Femenino	*Masculino*	
35-45	0%	8%	8%
45-55	0%	2%	2%
55-65	2%	6%	8%
65-75	10%	16%	26%
75-85	24%	12%	36%
85-95	10%	4%	14%
95-105	6%	0%	6%
Total, general	**52%**	**48%**	**100%**

Tabla 2.7 Tabla de contingencia del género y modalidad de estudio de los estudiantes de UEES.

A partir de la tabla 2.7 se puede identificar que, las calificaciones más altas (95-105), las obtienen los estudiantes del género femenino, mientras que las más bajas (35-45), las obtienen los estudiantes del género masculino.

3. *¿Qué porcentaje de los estudiantes de modalidad presencial obtienen calificaciones de 85-95? ¿Qué porcentaje de los estudiantes de modalidad virtual obtienen calificaciones de 35-45?*

Calificación	Modalidad de estudio		Total, general
	Presencial	*Virtual*	
35-45	2%	6%	8%
45-55	0%	2%	2%
55-65	2%	6%	8%
65-75	14%	12%	26%
75-85	30%	6%	36%
85-95	14%	0%	14%
95-105	6%	0%	6%
Total, general	**68%**	**32%**	**100%**

Tabla 2.8 Tabla de contingencia de la modalidad de estudio y calificaciones de los estudiantes de UEES.

El 14 % de estudiantes que asisten en modalidad presencial obtienen calificaciones entre 85-95, mientras que los estudiantes que cursan la modalidad virtual y obtienen calificaciones entre 35-45 representa el 6%.

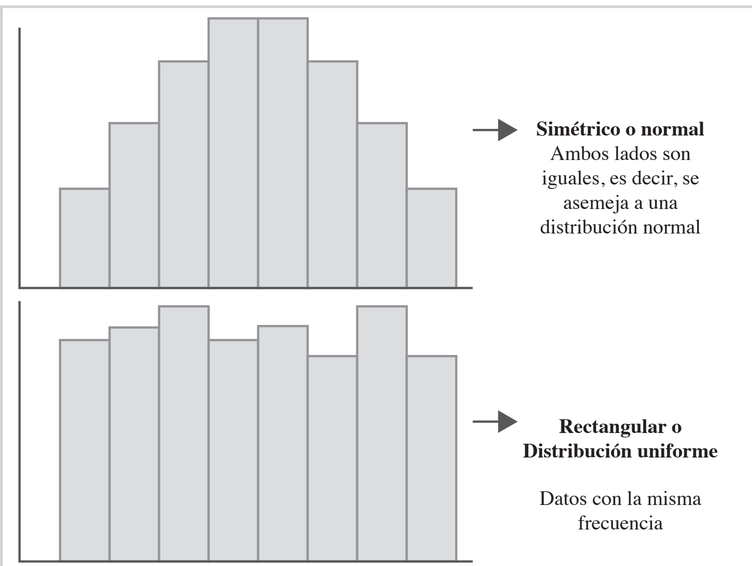
2.5 Método gráfico para variables continuas

2.5.1 Histograma

El histograma es la representación gráfica de un conjunto de datos. Muestra cómo se distribuyen los datos de una población o muestra mediante rangos numéricos o en valores absolutos (únicos). El histograma tiene cierto parecido al gráfico de barras, su principal diferencia es que el primero mide la frecuencia observada de datos continuos, su tendencia y su dispersión, mientras que en el gráfico de barras las variables son categóricas.

En el histograma la altura dependerá de la frecuencia absoluta o relativa de las distintas clases de la variable de estudio, está corresponderá al eje de las ordenadas (Y). Por otro lado, en el eje de las abscisas (X), normalmente, se colocan las marcas de clase de la variable de estudio. Se debe recordar que, en estadística, la frecuencia es la cantidad de veces que se repite un suceso. Por ejemplo, si queremos agrupar a un grupo en función de la edad, la frecuencia sería el número de individuos que tienen, por ejemplo, entre 18 y 25 años.

La simetría del histograma puede variar de acuerdo a las características que presentan los datos. En el gráfico 2.2 se presentan algunos ejemplos de las formas que suele tomar un histograma.



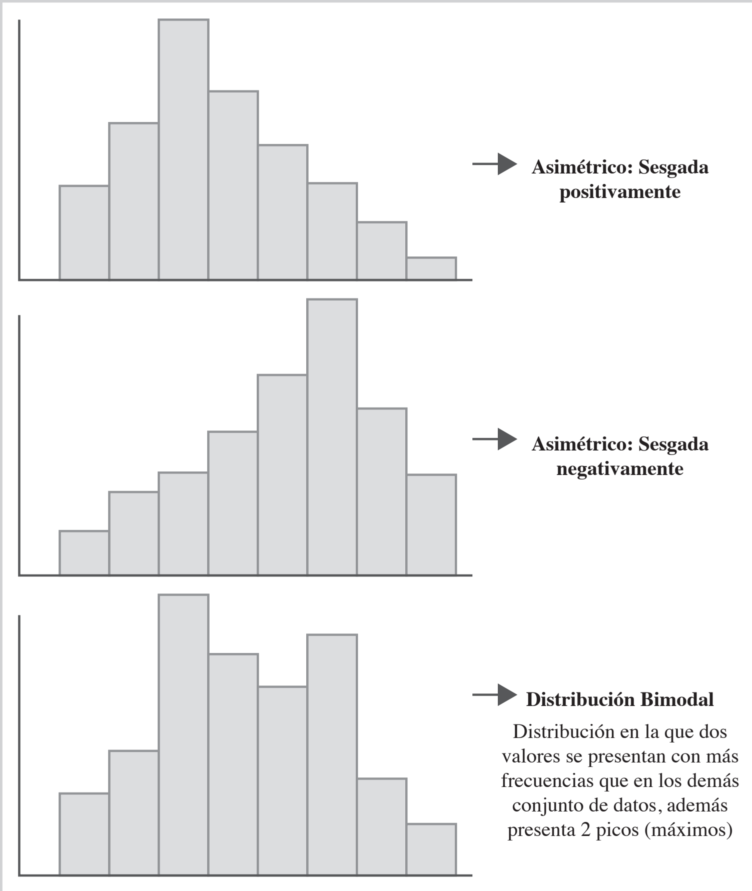


Gráfico 2.2 Tipos de histograma.

Ejercicio 2.3 Realizar el histograma de los datos de la tabla 2.2 que representan las calificaciones de los estudiantes de UEES.

Para graficar un histograma primero se necesita la elaboración de una tabla de frecuencias. En este caso utilizaremos la tabla anterior del ejercicio 2.1. Seleccionamos las celdas de la frecuencia relativa y marca de clase y en gráficos buscamos la opción de histograma.

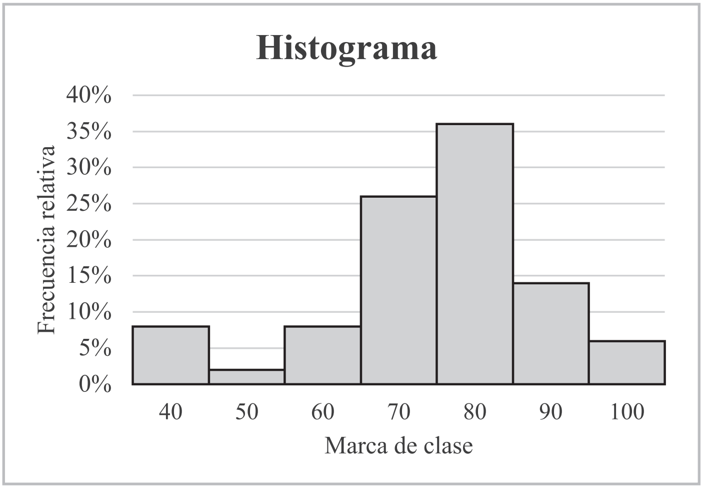


Gráfico 2.3 Histograma de las calificaciones de los estudiantes de UEES.

2.5.2 Diagrama de tallo y hoja

El diagrama de tallo y hojas es un gráfico que muestra la distribución de una variable numérica. Cada número del conjunto de datos se encuentra dividido por una hoja y un tallo, donde el último dígito (unidad) se lo denomina hoja y los números restantes (decena) se los denomina tallo (Serra, sf).

Tallo ← 3 2 → *Hoja*

Tallo ← 18 9 → *Hoja*

Ejercicio 2.4 Se han obtenido las calificaciones del examen de estadística de 24 estudiantes de UEES. Realizar un diagrama de tallo y hoja.

2.3	4.5	3.6	3.9
4.5	6.3	6.9	8.7
9.6	9.1	8.6	5.6
6.1	7.4	5.4	5.1
4.3	6.9	5.5	7.6
6.7	7.1	6.5	6.4

Tabla 2.9 Calificaciones de examen de estadística.

1. Se ordena los datos de menor a mayor.

2.3	3.6	3.9	4.3
4.5	4.5	5.1	5.4
5.5	5.6	6.1	6.3
6.4	6.5	6.7	6.9
6.9	7.1	7.4	7.6
8.6	8.7	9.1	9.6

2. Los tallos representan (en este ejemplo) los números enteros y las hojas los decimales. Si una hoja se repite, se la debe colocar nuevamente.

TALLO	*HOJAS*
2	3
3	6 9
4	3 5 5
5	1 4 5 6
6	1 3 4 5 7 9 9
7	1 4 6
8	6 7
9	1 6

Gráfico 2.4 Diagrama de tallo y hojas.

2.6 Gráficos para variables Categóricas

2.6.1 Diagrama de barras

Un diagrama de barras es una forma de representar gráficamente un conjunto de datos o valores utilizando barras que reflejan los conteos de frecuencia de valores de los distintos niveles de una variable categórica o nominal. Este tipo de gráfico, al igual que el histograma, se representa en ejes de coordenadas. En el eje horizontal se ubican las etiquetas o valores de la variable, mientras que en el eje vertical se indican las frecuencias correspondientes (Anderson et al, 2008). Es importante considerar que los valores de las variables estén en las mismas unidades.

Ejercicio 2.5 Se les ha preguntado a cinco personas sobre sus ingresos y gastos mensuales. A partir de estos datos realizar un diagrama de barras.

Personas	Ingresos mensuales	Gastos mensuales
Juan	1,000	800
Pedro	2,000	1,500
María	1,500	500
Raquel	2,500	1,000
Andrés	3,000	2,500

Tabla 2.10 Ingresos y gastos mensuales.

Para elaborar el diagrama de barras utilizamos la opción de Gráficos de columnas y seleccionamos los datos de estudio.

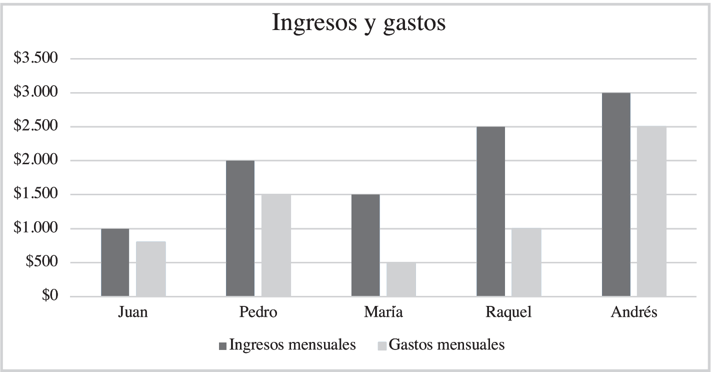


Gráfico 2.5 Diagrama de barras de los ingresos y gastos mensuales.

2.6.2 Diagrama circular o de pastel

El diagrama de pastel es una herramienta que permite presentar distribuciones de frecuencia absoluta y relativa de datos categóricos. Se utiliza la frecuencia relativa para subdividir el círculo en secciones correspondientes a la frecuencia relativa de cada clase (Anderson et al., 2008).

Los diagramas circulares son una de las herramientas básicas más importantes en la visualización de datos, y también son uno de los más fáciles de interpretar. Cada sección representa un porcentaje del total y la suma de todas las secciones siempre es igual a 100% (Gaskin, 2021).

Ejercicio 2.6 Se ha realizado una encuesta a un grupo de personas para conocer qué tan celosas son con su pareja. La muestra corresponde a 60 personas.

Nivel de celos	Frecuencia	Porcentaje
Poquito	10	17%
Algo	5	8%
Mucho	30	50%
Muchísimo	15	25%
TOTAL	**60**	**100%**

Tabla 2.11 ¿Qué tan celoso es con su pareja?.

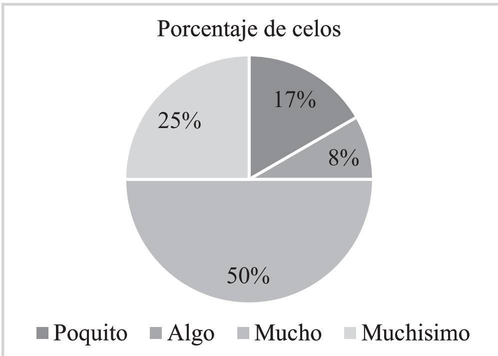


Gráfico 2.6. Porcentaje de celos.

Es importante destacar que un diagrama circular nunca debe usarse cuando se obtenga el 100% en una de las categorías de la variable de estudio. Por ejemplo, si se desea conocer el porcentaje de hombres y mujeres de una encuesta local, y se obtiene que el 100% son mujeres, representar esto en un diagrama circular como un todo, es estadísticamente incorrecto. Simplemente se debe mencionar que se obtuvo tal porcentaje sin el empleo de la gráfica.

2.7 Gráfico de máximos, mínimos y cierre

Este tipo de gráfico permite la visualización de tres valores: máximo, mínimo y de cierre de una variable, usualmente, financiera. El análisis se puede realizar durante un periodo de tiempo. Así, se puede analizar e interpretar las fluctuaciones o tendencias que presentan instrumentos financieros, como los precios de las acciones, bonos, monedas, etc.

Ejercicio 2.7 Crear un gráfico de máximos, mínimos y cierre con los siguientes datos.

Tiempo	Frecuencia	Porcentaje	Frecuencia
día 1	8	5	6.5
día 2	5	2	4
día 3	7	3	5
día 4	3	1	2
día 5	4	2	3.5

Tabla 2.12. Precio de acciones.

Para la elaboración de un gráfico de máximos, mínimos y cierres en Excel se seleccionan los datos de estudio y se inserta el gráfico de cotizaciones.

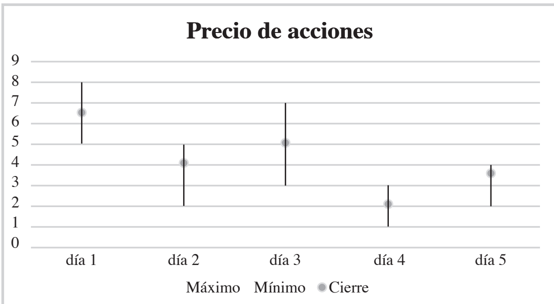


Gráfico 2.7. Máximo, mínimo, y cierre de precios de acciones.

Ejercicios propuestos del capítulo

Ejercicio 2.1 Se consultó a un grupo de 38 estudiantes sobre la asignatura que mayor dificultad les genera en la universidad. También se les preguntó su género y edad. Con base a los siguientes datos responda:

Genero	Edad	Asignatura
Femenino	17	Química
Femenino	22	Química
Femenino	25	Química
Masculino	23	Química
Masculino	25	Química
Masculino	20	Química
Masculino	18	Química
Masculino	18	Química
Masculino	21	Química
Masculino	21	Química
Femenino	21	Matemáticas
Masculino	23	Matemáticas
Masculino	24	Matemáticas
Femenino	24	Matemáticas
Femenino	20	Matemáticas
Femenino	21	Matemáticas
Femenino	22	Economía
Femenino	21	Economía
Masculino	24	Economía
Masculino	18	Economía
Masculino	18	Economía
Masculino	20	Economía
Masculino	21	Economía
Masculino	17	Economía
Femenino	25	Economía
Femenino	18	Economía

Femenino	25	Estadística
Femenino	24	Estadística
Masculino	25	Estadística
Masculino	18	Estadística
Masculino	18	Estadística
Masculino	21	Estadística
Femenino	18	Estadística
Femenino	19	Biología
Femenino	21	Biología
Masculino	24	Biología
Masculino	20	Biología
Masculino	18	Biología

a. ¿Cuál es la materia que más dificultad les genera a los hombres? ¿y a las mujeres?

b. De los estudiantes que se les dificulta economía, ¿Cuántos tienen 21 años?

c. ¿Cuál es el promedio de edad de los hombres?

Ejercicio 2.2 Con base a la siguiente información construya una tabla de frecuencias e interprete los principales resultados.

Datos (Altura en cm)			
155	168	165	170
192	178	157	193
192	161	185	168
188	186	152	152
189	185	156	186
173	170	170	171
182	164	200	175
164	164	195	160
161	185	154	170
198	189	180	193
173	169	187	174
156	153	165	177
183	169	170	159
168	158	190	164
184	183	175	172
154	200	164	200
178	152	165	190
182	188	173	197
150	176	182	189
161	191	193	168
153	190	180	199
184	193	187	174

Ejercicio 2.3 Se les preguntó a 25 empleados si estaban de acuerdo con una modificación en el plan de salud de la empresa. Realice un análisis descriptivo utilizando una tabla de contingencia.

Género	Opinión
Femenino	Si
Masculino	No
Femenino	Si
Femenino	Si
Femenino	Si
Femenino	Si
Masculino	No
Masculino	No
Masculino	No
Masculino	Si
Masculino	Si
Masculino	No
Femenino	Si
Femenino	No
Masculino	Si
Femenino	Si
Masculino	No
Femenino	Si
Masculino	No
Femenino	No
Masculino	Si
Masculino	No
Masculino	No
Femenino	Si
Femenino	No

Ejercicio 2.4 Se tiene información de cinco personas sobre sus promedios en estadística 1 y 2. Realice un análisis descriptivo utilizando un diagrama de barras.

Personas	Estadística 1	Estadística 2
Alberto	7	7.2
Alejandra	10	9
María	6.7	7.2
Raquel	9	9.5
Andrés	6	6.8

Ejercicio 2.5 A 95 personas se les pidió que señalen su centro comercial preferido en Guayaquil. Con base a esto realice un análisis descriptivo utilizando un diagrama circular.

Centro comercial	Frecuencia	Porcentaje
Mall del Sol	30	?
City Mall	15	?
San Marino	40	?
Policentro	10	?
TOTAL	**95**	**?**

Ejercicio 2.6 Se cuenta con el registro de los siguientes precios de acciones. Realice un análisis descriptivo utilizando un gráfico de máximos, mínimos y al cierre.

Días	Máximo	Mínimo	Cierre
Lunes	10	5	7
Martes	7	2	5
Miércoles	8	4	6
Jueves	3	1	2
Viernes	4	2	3.5
Sábado	9	5	8.5
Domingo	2	1	1.5

Ejercicio 2.7 Con base a los siguientes datos construya un diagrama de tallo y hojas. Analice los resultados.

15	73	1	65	16	3	42
36	42	3	61	19	36	47
20	45	29	73	69	34	23
22	21	33	27	55	58	17
4	17	48	25	36	11	4
54	70	51	3	34	26	10

Ejercicio 2.8 Se tiene información del salario de 40 trabajadores del sector privado del Ecuador. Realice un análisis descriptivo utilizando una tabla de frecuencias.

300	1,500	950	1,400
400	760	800	400
200	600	500	700
2,300	2,000	900	500
700	500	350	1,000
800	300	300	250
1,000	280	450	1,100
1,500	300	600	1,300
400	500	3,000	250
700	800	700	750

Ejercicio 2.9 Responda verdadero o falso según corresponda

a. En estadística, la frecuencia es la cantidad de veces que se repite un suceso. ( )

b. Las tablas de contingencia se utilizan principalmente para variables de razón. ( )

c. En el diagrama de barras es aconsejable que los valores de las variables a estudiar estén en las mismas unidades. ( )

d. Está mal visto usar un diagrama circular si el porcentaje relativo de la categoría de una variable es 100%. ( )

e. Los gráficos de máximos, mínimos y cierre se utilizan principalmente para analizar instrumentos financieros como los precios de las acciones. ( )

f. El diagrama de tallo y hoja puede ser empleado simultáneamente para dos variables o más. ( )

Capítulo III

MEDIDAS DE TENDENCIA CENTRAL Y DISPERSIÓN

III. MEDIDAS DE TENDENCIA CENTRAL Y DISPERSIÓN

3.1 Introducción

En el capítulo anterior, se utilizaron varias herramientas de estadística descriptiva para poder presentar y organizar datos cuantitativos o cualitativos. Herramientas como tablas de frecuencia, contingencia y algunos métodos gráficos fueron estudiados a profundidad.

En este capítulo, se estudiarán las medidas más importantes de tendencia central y dispersión para describir datos cuantitativos. Una medida de tendencia central identifica el punto alrededor del cual se centran los datos, incluso un gran conjunto de datos puede ser descrito sencillamente con un simple número. Por otro lado, las medidas de dispersión indican el punto hasta el cual las observaciones individuales se esparcen alrededor de su punto central, en pocas palabras, esta medida pretende evaluar la variabilidad y dispersión de los datos, y la tendencia de las observaciones individuales al desviarse de dicho punto central.

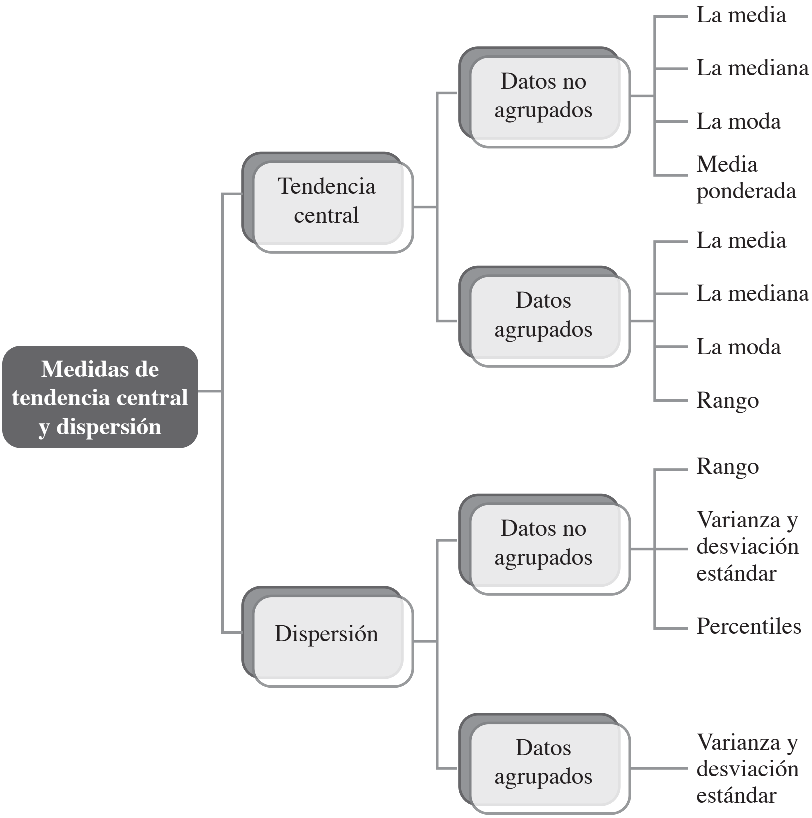


Gráfico 3.1 Medidas de tendencia central y dispersión.

En el grafico 3.1 se resumen las principales medidas de tendencia central y dispersión, tanto para datos no agrupados como agrupados. Los datos no agrupados son los datos «sueltos», es decir, aquellos datos brutos que se obtienen de fuentes primarias o secundarias. Por el contrario, los datos agrupados, son aquellos que se muestran resumidos en tablas, típicamente. Por su simpleza y facilidad de análisis, los datos no agrupados siempre serán preferibles a los datos agrupados.

3.2 Medidas de tendencia central a partir de datos no agrupados

3.2.1 La Media

La media es una medida de tendencia central que se considera como el promedio de un conjunto de datos. Se puede calcular la media poblacional, si contamos con datos de toda la población. En este caso, la media representa un parámetro y se representa con la letra μ. Por otro lado, si calculamos el promedio de un conjunto de datos provenientes de una muestra, obtenemos un estimador, que se representa con el símbolo .

$$Media\ poblacional: \mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$Media\ muestral: \bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Ejemplo 3.1 Suponga que las ventas (en miles de $) de cinco empresas del Ecuador en 2023 fueron: 50, 100, 70, 40, 10. ¿Cuál fue el promedio de ventas de estas empresas?

Solución

$$\bar{X} = \frac{50 + 100 + 70 + 40 + 10}{5} = 54$$

El promedio de ventas fue de $54.000.

3.2.2 La mediana

Es una medida que indica la observación de la mitad, después que se han ordenado los datos. Para conocer la posición de la mediana, se emplea la siguiente fórmula:

$$M_e = \frac{n + 1}{2}$$

Donde *n* es el número de datos. Del ejemplo anterior, primero debemos ordenar los datos de menor a mayor: *10, 40, 50, 70, 100*.

$$Posición\ de\ la\ mediana = \frac{5 + 1}{2} = 3$$

Dado que el tercer dato es 50, podemos concluir que este valor representa la mediana. Se puede interpretar como sigue: la mitad de las empresas alcanzaron ventas iguales o por debajo de 50 (mil) o ventas iguales o por encima de 50 (mil).

Cuando tenemos datos pares, obtener la posición de la mediana es un tanto diferente. Para ilustrar esto, al ejemplo anterior añadiremos otro valor: *10, 30, 40, 50, 70, 100*.

$$Posición\ de\ la\ mediana = \frac{6+1}{2} = 3.5$$

La mediana se encontrará entre los datos 4 y 5, por tanto, dichos valores se promediarán para obtener la mediana:

$$\frac{(40\ +\ 50)}{2} = 45$$

Significa que la mitad de las empresas obtuvieron ventas iguales o por debajo de 45 (mil) o ventas iguales o por encima de 45 (mil).

3.2.3 La moda

La observación modal es la observación que ocurre con mayor frecuencia.

Ejemplo 3.2 Se muestran las edades de 12 estudiantes universitarios:

$$\boldsymbol{18, 19, 19, 20, 21, 21, 21, 23, 23, 24, 26, 27}$$

La moda es el número 21, y se interpreta como sigue: la edad con más frecuencia es de 21 años. Hay que recalcar que utilizar la frase «la mayoría de…» es errónea.

De las tres medidas de tendencia central vistas, la media es la más común y utilizada. Desafortunadamente, la media se ve afectada por valores extremos, o valores atípicos (gráfica 3.2).

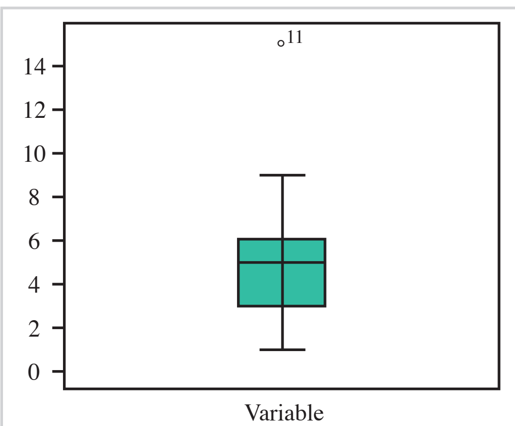


Gráfico 3.2 Dato atípico en un diagrama de caja.

Por lo contrario, la mediana tiene la ventaja de no ser afectada por datos atípicos. Consecuentemente, en presencia significativa de este tipo de datos, es preferible utilizar la mediana frente a datos atípicos como medida de tendencia central. Por otro lado, la moda tampoco se ve afectada por los valores atípicos. Sin embargo, si no hay moda, o si el conjunto de datos es bimodal, su uso puede ser confuso.

Por otro lado, las variables que presentan un comportamiento asimétrico[1] intrínseco, como los salarios ($,) no se ven bien representadas por la media, por lo contrario, la mediana, refleja de mejor forma la posición central de los datos en este tipo de variables.

La medida que se seleccione (media, mediana o moda) dependerá también de la naturaleza de los datos o del objetivo que se persigue. Por ejemplo, a un vendedor de zapatos le puede interesar poco o nada saber que la talla promedio que vendió de todos los zapatos el último mes fue 37.8521. De mayor utilidad para él sería conocer el tamaño modal (reconocer que vendió más zapatos de talla 39 que de cualquier otra talla).

Finalmente, la experiencia ha demostrado que la media sirve muy bien como medida de tendencia central cuando se trata de productos que están hechos para acomodarse a la estatura de las personas, por ejemplo, las puertas de un salón de clases de estadística.

3.2.4 La media ponderada

Para la media simple cada observación tiene igual importancia (peso). Sin embargo, en algunos casos, puede quererse dar mayor peso a algunas de las observaciones. La media ponderada permite asignar un peso específico a una o varias de las observaciones. Para calcularla, se debe multiplicar cada dato por su peso o ponderación y sumar los productos, después se lo divide por la suma del peso asignada a cada observación.

La fórmula para calcular la media ponderada es la siguiente:

$$\bar{X}_{w} = \frac{\sum XW}{\sum W}$$

[1]Analizaremos más adelante a profundidad este concepto.

En donde:

$\bar{X}_W$**:** la media ponderada.

X: Es la observación individual.

W**:** Es el peso o ponderación asignada a cada observación.

Ejemplo 3.3 El profesor de estadística menciona que el examen final valdrá el doble para determinar el promedio final. Un estudiante obtiene la calificación de 8.5 en el primer parcial y 8 en el segundo. Con base a los datos de la tabla 3.1, calcule la media ponderada.

Exámenes	Calificación (X)	Peso (W)	XW
Parcial	8.5	1	8.5
Final	8	2	16
Total		3	**24.5**

Tabla 3.1 Cálculo de la media ponderada.

Solución

$$\bar{X}_W = \frac{\sum XW}{\sum W} = \frac{24.5}{3} = 8.17$$

Ejemplo 3.4 Se tienen los siguientes datos de una empresa:

Departamento	Salarios ($)	Cantidad de empleados
RRHH	600	10
Administrativo	800	5
Marketing	700	12
Ventas	600	20
Gerencias	3,000	2

Tabla 3.2 Salarios de acuerdo a cada departamento de la empresa.

¿Cuál es el promedio de los salarios que paga la empresa?

Solución

$$Media\ simple\colon\ \bar{X} = \frac{600 + 800 + 700 + 600 + 3{,}000}{5} = 1{,}140$$

Inicialmente pensaríamos en obtener un promedio simple. El problema del promedio simple es que se da los mismos pesos a cada departamento. Al observar la tabla 3.4, vemos que en el departamento de gerencia solo hay dos personas y en el de ventas, 20 empleados. Por tanto, cecesitamos obtener una media ponderada para determinar una medida de tendencia central más acorde a los datos.

Departamento	Salarios(X)	Cantidad de empleados(W)	XW
RRHH	600	10	6,000
Administrativo	800	5	4,000
Marketing	700	12	8,400
Ventas	600	20	12,000
Gerencias	3	2	6,000
Total		**49**	**36,400**

Tabla 3.3 Salarios de acuerdo a cada departamento de la empresa.

$$Media\ ponderada: \bar{X}_W = \frac{\sum XW}{\sum W} = \frac{36{,}400}{49} = 742.85$$

Si se compara le media ponderada ($742.85) con la media simple ($1,140), nos damos cuenta que la primera se aproxima más a la realidad que la segunda

Ejercicio 3.1 Un estudiante de UEES obtiene las siguientes calificaciones en cada parámetro de evaluación de la materia de estadística. ¿Cuál es su calificación final?

Evaluación	Puntaje	Peso en la calificación
Trabajos	90	35%
Lecciones	80	25%
Examen	40	40%

Tabla 3.4 Calificaciones de estadística en cada parámetro evaluado.

$$Promedio\ ponderado: \bar{X}_W = (90 * 35\%) + (80 * 25\%) + (40 * 40\%) = 67.5$$

Note que, si se hubiese obtenido un promedio simple, la calificación final hubiese sido de 70, es decir, un promedio mayor. ¿Tiene sentido esto?

3.3 Medidas de dispersión

Las medidas de dispersión informan sobre las variaciones que presenta la variable, es decir, miden que tanto se dispersan las observaciones alrededor de su punto central.

Ejemplo 3.5 Observe el siguiente conjunto de datos y determine si son iguales.

Conjunto 1	Conjunto 2	Conjunto 3
0,5,10	4,5,6	5,5,5

Solución

Podemos observar que los tres conjuntos de datos tienen una media de 5, pero esto **no significa** que los datos sean iguales. Un diagnóstico más certero sería observar su grado de dispersión en relación a su punto central. Las observaciones del conjunto 1 están muy dispersas (por encima y por debajo del punto central).

Conjunto 1
0,5,10

Las observaciones del conjunto 2, en cambio, están muy cercanas.

Conjunto 2
4,5,6

Las observaciones del conjunto 3 no tienen dispersión por lo que todas las observaciones son iguales a su media.

Conjunto 3
5,5,5

Además, sería erróneo asumir cualquier similitud en los conjuntos de datos simplemente basándose en su media. En este caso, podemos concluir que las medidas de dispersión son más útiles e informativas.

3.3.1 Rango

Es la medida de dispersión más simple pero tal vez la menos útil. Es la diferencia entre los valores más altos y más bajos de las observaciones de una población o muestra.

Conjunto 1	Conjunto 2	Conjunto 3
0,5,10	4,5,6	5,5,5
Rango=10	Rango=2	Rango=0

La limitación de esta medida radica en que, solo considera dos de los cientos de observaciones que puede haber en un conjunto de datos. El resto de las observaciones se ignoran.

3.3.2 Varianza y desviación estándar de una población

La varianza: es una medida de dispersión que representa el promedio de las observaciones respecto a su media elevadas al cuadrado.

La fórmula para el cálculo de la varianza es la siguiente:

$$Varianza\ poblacional: \sigma^2 = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + (X_3 - \mu)^2 + \cdots + (X_N - \mu)^2}{N} = \frac{\sum(X_i - \mu)^2}{N}$$

En donde:

X_1 , X_1 , X_1 ... , X_N: Son las observaciones individuales.

μ: Es la media poblacional.

N: Es el número de observaciones.

Desviación estándar: Es la raíz cuadrada de la varianza. Su fórmula es la siguiente:

$$Desviación\ estándar: \sigma = \pm\sqrt{\sigma^2}$$

El concepto de desviación estándar es muy importante en los negocios y la economía. En finanzas la desviación estándar se utiliza como medida de riesgo relacionada con varias oportunidades de inversión.

Ejemplo 3.6 Se tienen los siguientes datos de las calificaciones de un curso.

Calificaciones	
8	2
7	9
4	10
8	7
5	9

Tabla 3.5 Calificaciones de los estudiantes.

Con base a esta información, calcule la varianza y desviación estándar poblacional.

Solución

$$Media\ poblacional: \mu = \frac{8+7+4+8+5+2+9+10+7+9}{10} = 6.6$$

$$Varianza\ poblacional: \sigma^2 = \frac{(8-6.6)^2 + (7-6.6)^2 + (4-6.6)^2 + \cdots}{10} = 5.24$$

$$Desviación\ estandar: \sigma = \pm\sqrt{5.24} = 2.29$$

Ejemplo 3.7 Juan quiere invertir en un plazo fijo. Para ello tiene dos opciones: cooperativa A y cooperativa B. Las tasas de rendimiento que ofrecen las cooperativas para los últimos 5 años se muestran a continuación:

Año	Cooperativa A	Cooperativa B
2017	9%	10%
2018	7%	9%
2019	10%	11%
2020	6%	7%
2021	8%	3%

Tabla 3.6 Tasas de rendimiento de acuerdo a cada cooperativa.

Si ambas cooperativas ofrecen un rendimiento promedio del 8%. ¿En qué cooperativa le recomendaría a Juan invertir?

Solución

$$\sigma^2(A) = \frac{(9-8)^2 + (7-8)^2 + (10-8)^2 + \cdots}{5} = 2$$

$$\sigma^2(B) = \frac{(10-8)^2 + (9-8)^2 + (11-8)^2 + \cdots}{5} = 8$$

$$\sigma(A) = \sqrt{2} = 1.41\%$$

$$\sigma(B) = \sqrt{8} = 2.83\%$$

Debido a que el fondo A presenta menos variabilidad en sus rendimientos y ofrece la misma tasa de rendimiento promedio que el fondo B, representa la opción más segura de las dos inversiones.

3.3.3 Varianza y desviación estándar para una muestra

El cálculo de la varianza para una muestra es ligeramente diferente a la de una población. La diferencia está en el denominador, en donde se resta el valor de 1. Esto se hace con el objetivo de «inflar» artificialmente la varianza.

Fórmulas	**Varianza**	**Desviación estándar**
Población	$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$	$\sigma = \pm\sqrt{\sigma^2}$
Muestra	$S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$	$s = \pm\sqrt{s^2}$

3.4 Medidas de tendencia central y dispersión para datos agrupados

Cuando se tienen datos agrupados, no se pueden aplicar los procedimientos mostrados anteriormente. Hay que buscar métodos alternativos los cuales nos darán aproximaciones, más no valores exactos. En las tablas 3.7 y 3.8 se muestra ejemplos de datos no agrupados y agrupados, respectivamente.

68	71	77	83	79
72	74	57	67	69
50	60	70	66	76
70	84	59	75	94
65	44	85	79	45
83	84	74	35	97
77	73	78	93	95
78	81	79	90	83
80	84	91	100	40
93	92	99	80	69

Tabla 3.7 Datos no agrupados.

Número de clases	Intervalo Inferior (I)	Intervalo Superior (S)	Frecuencia absoluta acumulada (Ni)	Frecuencia absoluta (ni)	Frecuencia relativa (fi)	Frecuencia relativa acumulada (Fi)	Marca de clase (mi)
1	35	45	4	4	8%	8%	40
2	45	55	5	1	2%	10%	50
3	55	65	9	4	8%	18%	60
4	65	75	22	13	26%	44%	70
5	75	85	40	18	36%	80%	80
6	85	95	47	7	14%	94%	90
7	95	105	50	3	6%	100%	100

Tabla 3.8 Datos agrupados.

3.4.1 La media

Como mencionamos anteriormente, la media es una medida de localización central y la fórmula para datos no agrupados es la siguiente:

$$\bar{X}_g = \frac{\sum fM}{n} = \frac{\sum fM}{\sum f}$$

En donde:

f: Es la frecuencia o el número de observaciones en cada clase.

M: Es el punto medio de cada clase.

n: Es el tamaño de cada clase y es igual a las frecuencias sumadas en todas las clases.

Intervalo Inferior (I)	Intervalo Superior (S)	Marca de clase (mi)	Frecuencia absoluta (Ni)	fM (ni*mi)
35	45	40	4	160
45	55	50	1	50
55	65	60	4	240
65	75	70	13	910
75	85	80	18	1,440
85	95	90	7	630
95	105	100	3	300
			50	3,730

Tabla 3.9 Distribución de frecuencias de calificaciones de estudiantes de la UEES.

Con base a los datos que se muestran en la tabla anterior, se puede calcular la media de la siguiente forma:

$$\bar{X}_g = \frac{\sum fM}{n} = \frac{3{,}730}{50} = 74.6$$

3.4.2 La mediana

La fórmula para calcular la mediana para datos agrupado es la siguiente:

$$Mediana\ para\ datos\ agrupados\text{:}\ L_{md} + \left[\frac{\frac{n}{2} - F}{fmd}\right](c)$$

En donde:

L_{md}: Es el intervalo inferior de la clase de la mediana[2].

F: Es la frecuencia acumulada de la clase que antecede a la clase de la mediana.

***fmd*:** Es la frecuencia absoluta de la clase de la mediana.

C: Es el intervalo de clase de la clase de la mediana.

[2]La clase de la mediana se la calcula como sigue: igual o mayor que n/2 en Ni.

Intervalo inferior	Intervalo superior	Frecuencia absoluta (ni)	Frecuencia absoluta acumulada (N_i)
35	45	4	4
45	55	1	5
55	65	4	9
65	75	13	22
75	85	18	40
85	95	7	47
95	105	3	50

Con los datos de la tabla anterior, la mediana sería:

$$Mediana = L_{md} + \left[\frac{\frac{n}{2} - F}{f_{md}}\right](C) = 7 + \left[\frac{\frac{50}{2} - 22}{18}\right](10) = 76.67$$

Significa que la mitad de las estudiantes obtuvieron calificaciones iguales o por debajo de 77 o calificaciones iguales o por encima de 77.

3.4.3 La moda

Como se describió anteriormente, la moda es el valor que más se repite, el cual se halla en la clase que tenga la frecuencia más alta, llamada clase modal. Cuando se trabaja con datos agrupados, se utiliza la siguiente fórmula para su cálculo:

$$Moda\ para\ datos\ agrupados: L_{mo} + \left[\frac{D_a}{D_b + D_a}\right](C)$$

En donde:

L_{mo}: Es el intervalo inferior de la clase modal.

D_a: Es la diferencia entre la frecuencia de la clase modal y la clase que la antecede.

D_b: Es la diferencia entre la frecuencia de la clase modal y la clase que la sigue.

***C*:** Es el intervalo de clase de la clase modal.

Intervalo inferior	Intervalo superior	Frecuencia absoluta (ni)	Frecuencia absoluta acumulada (N_i)
35	45	4	4
45	55	1	5
55	65	4	9
65	75	13	22
75	85	18	40
85	95	7	47
95	105	3	50

Utilizando los datos de la tabla anterior podemos obtener la moda como sigue:

$$Moda = L_{mo} + \left[\frac{D_a}{D_b + D_a}\right](C) = 75 + \left[\frac{18-13}{(18-7)+(18-13)}\right]10 = 78.13$$

3.4.4 Varianza y desviación estándar

Para datos agrupados, se debe utilizar la siguiente fórmula para el cálculo de la varianza y desviación estándar:

Varianza de la muestra de datos agrupados	$S^2 = \frac{\sum fM^2 - n\bar{X}^2}{n-1}$
Desviación estándar muestral para datos agrupados	$S = \sqrt{S^2}$

Tabla 3.10 Fórmulas para el cálculo de la varianza y desviación estándar para datos agrupados.

Con base a los datos mostrados en la siguiente tabla, se puede obtener la varianza muestral y posteriormente la desviación estándar.

Intervalo Inferior (I)	Intervalo Superior (S)	Marca de clase (mi)	Frecuencia absoluta (Ni)	fM (ni*mi)	M2 (ni*mi)	fM2
35	45	40	4	160	1,600	6,400
45	55	50	1	50	2,500	2,500
55	65	60	4	240	3,600	14,400
65	75	70	13	910	4,900	63,700
75	85	80	18	1,440	6,400	115,200
85	95	90	7	630	8,100	56,700
95	105	100	3	300	10,000	30,000
						288,900

$$S^2 = \frac{\sum fM^2 - n\bar{X}^2}{n-1} = \frac{288{,}900 - 50(74.6)^2}{50-1} = 217.184$$

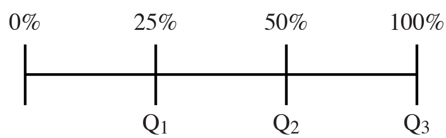
$$S = \sqrt{217.184} = 14.74$$

3.5 Otras medidas de dispersión

Aunque la varianza y la desviación estándar son las medidas de dispersión más utilizadas en el análisis estadístico, existen otras técnicas para determinar la dispersión de un conjunto de datos. Estos son los cuartiles, quintiles, deciles, percentiles, rango intercuartílico y diagrama de cajas.

3.5.1 Cuartiles

Los cuartiles separan a un conjunto de datos en cuatro subconjuntos iguales. Por lo general, se definen tres cuartiles denotados como Q_1, Q_2 y Q_3. El primer cuartil es ese valor debajo del cual se clasifica el 25% de las observaciones, el segundo cuartil es justo la mitad (la mediana) y el tercer cuartil es el valor debajo del cual está el 75% de las observaciones. Por ejemplo, muchas escuelas de posgrado admiten solo a aquellos estudiantes que estén en el 25% superior (tercer cuartil de los candidatos).



3.5.2 Quintiles

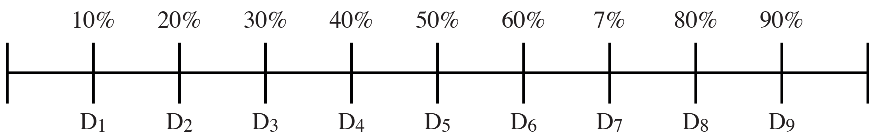
Un quintil es un quinto de un grupo, es decir, representa el 20% de una población determinada. Por ejemplo, se toma la población en un año y se ordena por niveles de renta, en forma ascendente. Luego, se divide en cinco partes iguales, para que así reduzcamos un número enorme de datos a cinco grupos que será más fácil de manejar.

Quintil	Calificaciones
1	Clase pobre
2	Clase Media Baja
3	Clase Media
4	Clase Media Alta
5	Clase Rica

Tabla 3.11 Clasificación socioeconómica en cinco grupos.

3.5.3 Deciles

Los deciles separan a un conjunto de datos en 10 subconjuntos iguales.



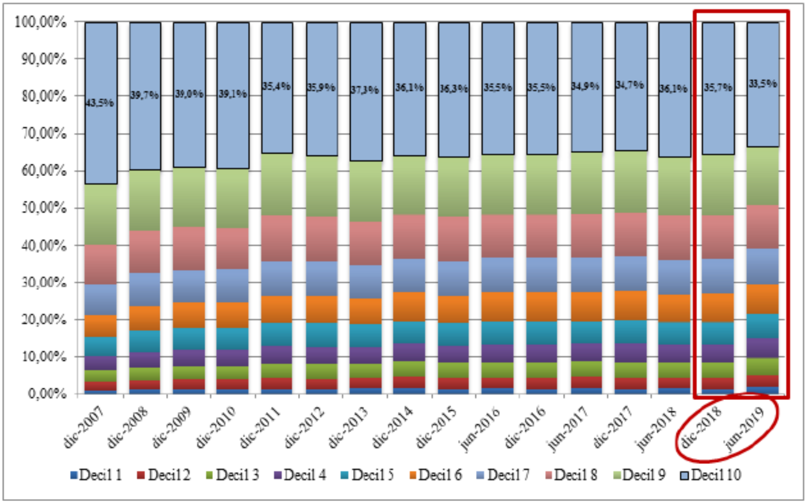


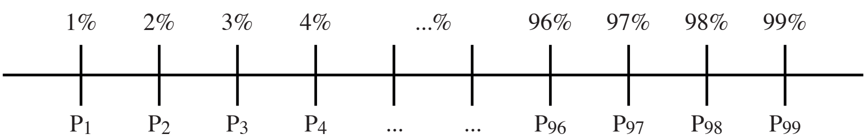
Gráfico 3.3 Participación en el ingreso per cápita por deciles 2007- 2019.

Elaboración: Báez (2019).

En el gráfico 3.3 se observa que el decil más rico (decil 10) aumenta su participación en el ingreso per cápita a junio de 2018 con un 36.1%, considerando que en junio de 2017 alcanzaba el 34.9%. Por un lado, eso implica una recuperación de cierto sector privilegiado de la población en sus ingresos.

3.5.4 Percentiles

Los percentiles separan a un conjunto de datos en 100 subconjuntos iguales, y dispone los datos de menor a mayor. El cálculo de los cuartiles, deciles y quintiles se puede determinar a partir de los percentiles correspondientes. Por ejemplo, el percentil 10 es el primer decil, el percentil 25 es el primer cuartil, el percentil 20 es el primer quintil y así.



Para su cálculo utilizamos la siguiente fórmula:

$$\text{Ubicación de un percentil: } L_p = (n+1)\frac{P}{100}$$

En donde:

L_p: Es el sitio del percentil deseado en una serie ordenada.

n: Es el número de observaciones

P: Es el percentil deseado.

Ejemplo 3.8 A partir de los datos que se muestran en la siguiente tabla 3.12, se desea calcular el percentil 25 y 35.

3	15	31	39	56	72
4	17	31	43	59	73
7	19	34	45	62	74
9	20	34	47	63	74
10	21	34	48	94	76
10	25	36	48	67	79
12	27	37	52	67	80
14	27	38	53	69	
	29	38	56		

Solución

$$L_{25} = (50+1)\frac{25}{100} = 12.75$$

$$P_{25} = 20 + 0.75(21-20) = 20.75$$

$$L_{35} = (50+1)\frac{35}{100} = 17.85$$

$$P_{35} = 29 + 0.85(31-29) = 30.7$$

El percentil 25 está ubicado al 75% del trayecto comprendido entre la duodécima observación (20) y la décima tercera observación (21). Además, el 25% de las observaciones está por debajo de 20.75 y el restante 75%, por encima de 20.75. Una interpretación similar se le puede dar al percentil 35.

3.5.5 Rango intercuartílico

Entendemos por rango intercuartílico a la diferencia entre el tercer cuartil y el primer cuartil . La mitad de las observaciones se clasifican dentro de este rango. Cuanto mayor es el rango intercuartílico, mayor la dispersión entre los datos.

El 50% intermedio de los datos esta entre 9.5 y 17.5.

3.5.6 Diagrama de cajas

El diagrama de cajas es una herramienta estadística que se utiliza para clasificar a un conjunto de datos en sus respectivos cuartiles. Además, el diagrama de cajas se utiliza para identificar datos atípicos.

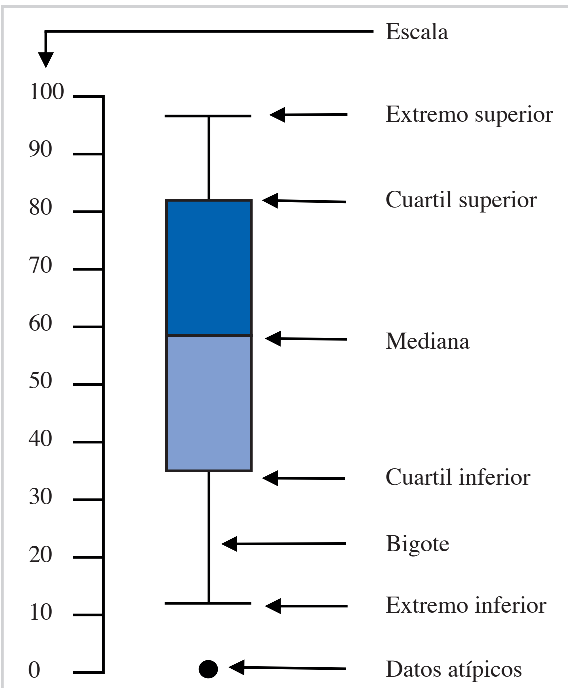


Gráfico 3.4 Elementos de un diagrama de cajas.

Ejercicio 3.2 En la siguiente tabla se muestra información sobre el número de cervezas que toma a la semana un grupo de encuestados. Con base a estos datos, realice un diagrama de cajas.

Encuestado	Cervezas
1	-3
2	1
3	1
4	1
5	2
6	2
7	3
8	3
9	3
10	4
11	4
12	4
13	4
14	5
15	5
16	5
17	6
18	10
19	10
20	15

Tabla 3.13 Número de cervezas que los participantes beben a la semana.

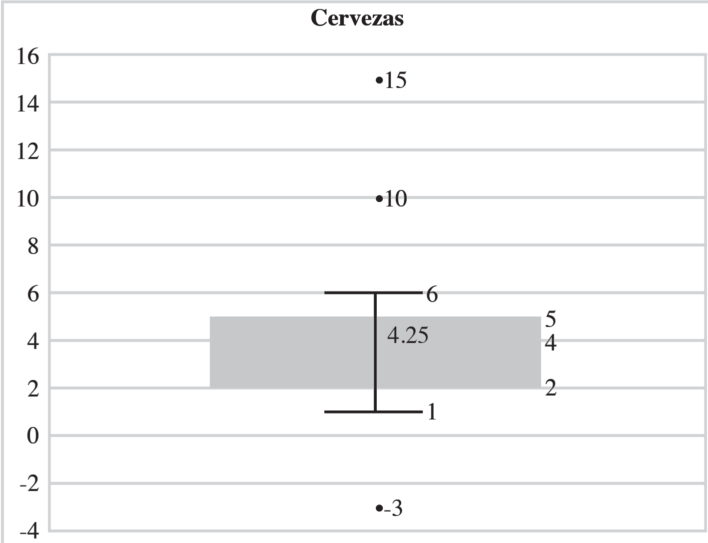


Gráfico 3.5 Número de cervezas tomadas a la semana.

La interpretación de este diagrama de cajas es la siguiente: el primer cuartil es 2, lo que nos indica que el 25% de los encuestados se toman 2 o menos cervezas a la semana. La mediana (segundo cuartil) es 4, es decir, el 50% de los participantes toman entre 2 a 5 cervezas a la semana, mientras que el tercer cuartil es 5. Por otro lado, el promedio de cervezas tomadas a la semana es de 4.25. Además, entre el 1 y 6 están los valores considerados «típicos».

Finalmente, los valores atípicos son aquellos que salen del rango típico, es decir, los números 15, 10 y -3. Algo importante de mencionar es que no todos los valores atípicos deben eliminarse, solo aquellos que se correspondan a información falsa o de mal registro o que estén contra la naturaleza de los datos. En este ejemplo, nadie puede tomarse -3 cervezas, por lo que esta observación debería eliminarse. Los restantes valores 15 y 10, debería contrastarse con información adicional[3].

[3]Por ejemplo, si los datos provienen de una encuesta, debería analizarse a profundidad dichas observaciones contrastando con otras preguntas del cuestionario para detectar anomalías en las respuestas de esos encuestados.

3.6 Usos frecuentes de la desviación estándar

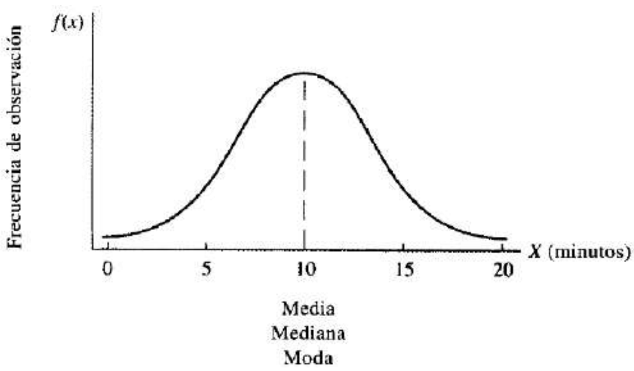
Veamos ahora algunos usos frecuentes de la desviación estándar.

3.6.1 La distribución normal y la regla empírica

La desviación estándar puede utilizarse para determinar si un conjunto de datos está distribuido normalmente. La distribución normal es una distribución de datos continuos (no discretos) que produce una curva simétrica en forma de campana. Ilustremos este concepto a través de un ejemplo.

Suponga que se tiene un gran número de observaciones para el tiempo (en minutos) que les toma a unos esquiadores del Chimborazo terminar un trayecto. Las observaciones extremas ocurrirán con poca frecuencia relativa en este tipo de distribuciones; mientras que las observaciones en la mitad, ocurrirán con mayor frecuencia. De hecho, la mitad de las observaciones está por encima de la media y el otro restante por debajo.

Por ejemplo, imagine que se tiene información de 1,000 esquiadores, el promedio que se toman en bajar una cuesta es 10 min (μ=10) y presenta una desviación estándar de 2 min (σ=2).



Gráfica 3.6 Distribución normal.

La regla empírica señala que si se incluyen todas las observaciones que están a una desviación estándar de la media, éstas serán 68.3% de todas las observaciones. Es decir, sin importar cuál sea la media y cuál sea la desviación estándar, si las observaciones se distribuyen normalmente, se puede estar seguro de que el 68.3 % de las observaciones están dentro de una desviación estándar de la media.

Debido a que en promedio a los esquiadores les toma 10 min en terminar el trayecto, mover una desviación estándar (es decir, 2 min) por encima y por debajo de esta media de 10, produce un rango de 8 a 12 min.

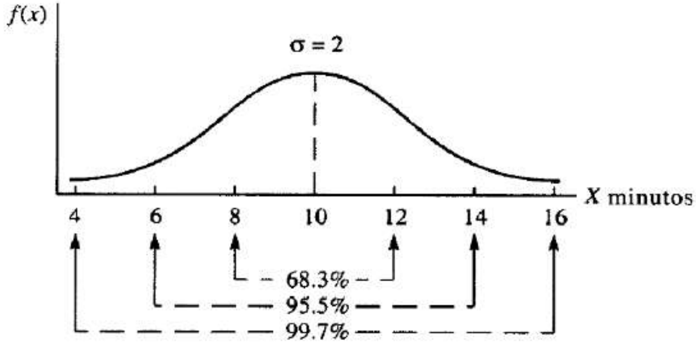


Gráfico 3.7 Interpretación de la distribución de datos en la distribución normal.

El grafico 3.7 indica que el 68.3% de las observaciones están dentro de más o menos una desviación estándar de la media. El 95.5% de las desviaciones están dentro de más o menos dos desviaciones estándar de la media, y el 99.7% de las desviaciones están dentro de más o menos tres desviaciones estándar de la media.

Por otro lado, si las observaciones están altamente dispersas (mayor desviación estándar), la curva en forma de campana se aplanará y se esparcirá. Por ejemplo, imagine que otro grupo también tardo 10 minutos en completar el trayecto, pero tuvo una desviación estándar de 4 minutos. En comparación con el primer grupo, el tiempo de esquí más rápido fue de menos de 10 min y el más lento estaban por encima de 10 min.

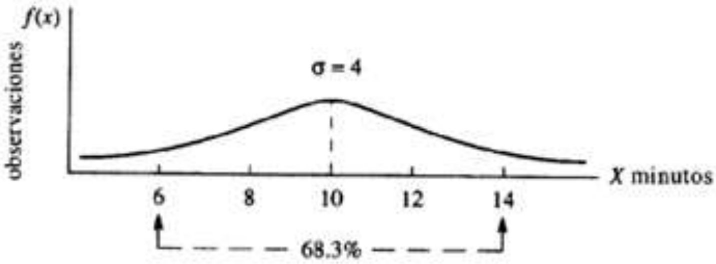


Gráfico 3.8 Distribución normal con mayor desviación estándar.

3.6.2 Curtosis

La curtosis es una medida estadística que mide la concentración de los valores de una variable alrededor de la región central de la distribución de frecuencia. Su interpretación se basa en el valor que toma el coeficiente de Fisher, así:

- Coeficiente de curtosis > 0 Distribución Leptocúrtica
- Coeficiente de curtosis = 0 Distribución Mesocúrtica
- Coeficiente de curtosis < 0 Distribución Platicúrtica

Gráficamente este tipo de distribuciones se representa como sigue:

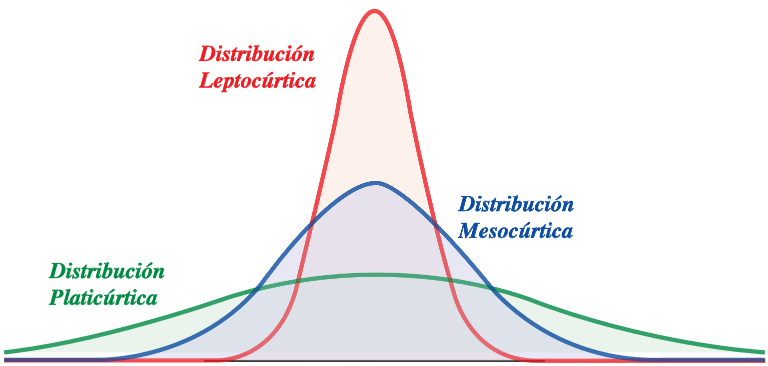


Gráfico 3.9 Diferentes tipos de curtosis.
Elaboración: Zapata (2022).

Para el cálculo formal del coeficiente de Fisher utilizamos la siguiente fórmula:

$$Coeficiente\ de\ Curtosis\text{:} \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{X_i-\bar{X}}{S}\right)^4\right]-\frac{3(n-1)^2}{(n-2)(n-3)}$$

Si el coeficiente da como resultado una distribución leptocúrtica, significa que la dispersión de los datos es baja, si el resultado es una distribución mesocúrtica, los datos poseen una dispersión media, y si es una distribución platicúrtica, los datos presentan una alta dispersión.

3.6.3 Sesgo

No todas las distribuciones son simétricas (normales), algunas están sesgadas a la izquierda y otras a la derecha. El sesgo nos permite determinar la simetría de una distribución de datos respecto a la media. Cuando una distribución presenta sesgo, significa que los datos no se distribuyen normalmente, es decir, son asimétricos. El sesgo en una variable puede obedecer a la presencia de valores extremos o atípicos o la naturaleza intrínseca de la variable (recuerde el ejemplo de los salarios).

Como se mencionó previamente, cuando los datos están sesgados, la medida de tendencia central más recomendada es la mediana, en lugar de la media.

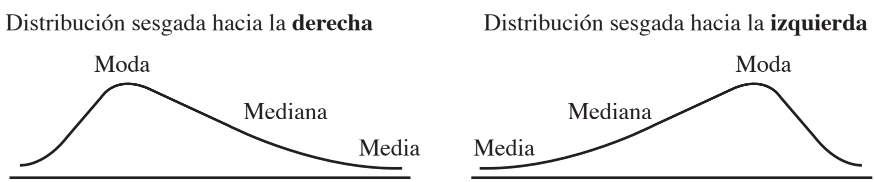


Gráfico 3.10 Características de una distribución normal.

Para calcular el coeficiente de sesgo utilizamos la siguiente fórmula:

$$\text{Coeficiente de sesgo (CS): } \frac{n}{(n-1)(n-2)}\sum\left(\frac{X_i-\bar{X}}{S}\right)^3$$

La interpretación es la siguiente:

- CS < 0, los datos están sesgados a la izquierda.

- CS > 0, los datos están sesgados a la derecha.
- CS = 0, los datos están distribuidos normalmente.

3.6.4 Coeficiente de variación

El coeficiente de variación mide qué tan disperso está un conjunto de datos en relación a su media. Cuando se comparan dos o más distribuciones con diferentes medias, no es oportuno utilizar la desviación estándar como medida comparativa.

Por ejemplo, tenemos dos conjuntos de datos con sus respectivas medias y desviaciones estándar, ¿Cuál de esos dos tiene mayor dispersión?

Datos A	$\longrightarrow \bar{X}_A = 78.7$; $S_A = 12.14$
Datos B	$\longrightarrow \bar{X}_B = 1{,}267.5$; $S_B = 152.7$

Inicialmente pensaríamos que los datos B tienen mayor dispersión debido a que su desviación estándar (152.7) es mayor que la de los datos A (12.14). Sin embargo, observamos que los datos poseen medias bastante diferentes, por lo que el coeficiente de variación sería una mejor alternativa para comparar su dispersión. Para calcular el coeficiente de variación utilizamos la siguiente fórmula:

$$Coeficiente\ de\ variación: CV = \frac{S}{\bar{X}}(100)$$

$$\boldsymbol{CV_A = \frac{12.14}{78.70}(100) = 15.43} \qquad \boldsymbol{CV_B = \frac{152.7}{1,267.5}(100) = 12.05}$$

Con estos cálculos podemos observar que los datos A son los que poseen mayor dispersión.

Ejercicios propuestos del capítulo

Ejercicio 3.1 Se tiene la siguiente información sobre los precios ($) de alquiler de viviendas en Guayaquil.

80	400	250	800
300	250	400	850
900	300	550	1,000
750	700	900	2,000
600	300	250	400

a. ¿Cuál es el precio promedio de alquiler?

b. ¿Cuál es la mediana? Interprete. ¿Tiene mayor sentido su uso en lugar de la media?

c. ¿Cuál es la moda?

d. Construya un diagrama de cajas e interprete

e. Calcule el segundo decil, cuarto quintil y el percentil 30. Interprete

f. Calcule el coeficiente de curtosis. Interprete

g. Calcule el coeficiente de asimetría. Interprete

Ejercicio 3.2 Su profesor de estadística en UEES menciona que la lección 1 tendrá un peso de tres en el promedio total de lecciones. Calcule la media ponderada del siguiente estudiante y compare con la media simple.

Lección 1	30
Lección 2	90
Lección 3	80

Ejercicio 3.3 Calcule el coeficiente de variación para el siguiente conjunto de datos. ¿Qué datos tienen mayor dispersión?

Gasto 1	Gasto 2
500	100
200	40
700	35
500	45
300	40
400	44
800	50
300	43
400	41
200	200
200	35

Ejercicio 3.4 Calcule el coeficiente de variación para el siguiente conjunto de datos. ¿Qué datos tienen mayor dispersión?

Gasto A	Gasto B
500	150
300	35
800	40
500	45
400	45
500	50
300	42
500	41
200	180
250	35

Ejercicio 3.5 Con base a los siguientes datos de consumo, calcule el coeficiente de curtosis y sesgo. Interprete.

Consumo		
100	100	100
150	550	500
150	150	150
150	100	200
300	250	800
250	150	900
250	700	1,000
200	200	600
100	350	100
450	150	100
100	200	300
350	200	200

Ejercicio 3.6 Se tiene información del gasto mensual de 10 personas de dos grupos. Calcule la media, mediana, moda, varianza, desviación estándar y coeficiente de variación e interprete los resultados.

Grupo 1	Grupo 2
200	600
300	850
350	900
200	1,000
400	800
250	600
200	400
450	700
150	1,200
250	600

Ejercicio 3.7 En el concurso para profesor de estadística, los puntos obtenidos en el examen teórico y práctico representan el 40% y 60%, respectivamente. La nota final corresponderá a la media ponderada.

Las calificaciones obtenidas por Juan y Mariana fueron:

	Juan	Mariana
Teórico Práctico	4.65	5.52
	6.15	4.95

¿Cuál de los dos candidatos obtendrá la única plaza de profesor por tener una calificación mayor?

Ejercicio 3.8 Dado los siguientes datos calcule la media y la mediana.

Tamaño tabla	Frecuencia acumulada
(3-6)	37
(6-11)	235
(11-16)	426
(16-21)	575
(21-26)	654
(26-31)	700
(31-41)	755
(41-51)	806
(51-76)	832
(76-101)	857
(101-201)	882
(201-501)	893
(501-1,000)	895

Capítulo IV

PRINCIPIOS DE PROBABILIDAD

IV. PRINCIPIOS DE PROBABILIDAD

4.1 Introducción

Como se mencionó en capítulos anteriores, la estadística se divide en dos grandes ramas: estadística descriptiva y estadística inferencial.

La estadística descriptiva es básicamente «resumir» el comportamiento de una variable (con información del pasado) a través de medidas de tendencia central y dispersión. Esto lo analizamos en los capítulos dos y tres. Por su parte, la estadística inferencial es sacar conclusiones de una población con base a una muestra; es calcular la probabilidad de que algo ocurra en el futuro con base a información del pasado. Se puede decir que, la estadística inferencial nace con la teoría de la probabilidad.

4.2 ¿Qué es la probabilidad?

La probabilidad es una medida cuantitativa de incertidumbre, y representa el grado de certeza o confianza que se tiene en la ocurrencia de un evento. La probabilidad de un evento es siempre un número entre 0 y 1, inclusive. Si la probabilidad de un evento está más cerca de 1, mayor es la probabilidad de que ocurra el evento; cuanto más se acerque a 0, menor será la probabilidad de que el evento suceda. Es común que una probabilidad se exprese en forma decimal, como 0.70; 0.27 o 0.50. No obstante, también se puede expresar en forma de fracción, como 7/10, 27/100 o 1/2.

Dentro del estudio de la probabilidad, se requieren comprender tres conceptos claves:

Experimento aleatorio: Un experimento aleatorio significa que el experimento tiene más de un resultado posible y no es posible predecir con certeza cuál será ese resultado. Por ejemplo, en un experimento de lanzar una moneda ordinaria, se puede predecir con certeza que la moneda caerá cara o cruz, pero no se sabe con certeza si saldrá cara o cruz.

Espacio muestral o resultado: conjunto de todos los resultados posibles de un experimento aleatorio.

Punto muestral o evento: Un evento hace referencia al conjunto de uno o más resultados dados dentro de un experimento.

Experimento aleatorio	Resultados	Eventos
Lanzar un dado	Existen 6 resultados posibles: $\{1, 2, 3, 4, 5, 6\}$	• Sacar un número par: {2,4,6} • Sacar un 3: {3} • Sacar un 1 o un 3: {1, 3} • Sacar un 1 y un 3: {} (Solo puede salir un número, por lo que esto es imposible. El evento no contiene resultados).

Los resultados pueden ser:

Mutuamente excluyentes: dos o más resultados son mutuamente excluyentes si la ocurrencia de uno impide la ocurrencia del otro. Es decir, los dos resultados no se pueden dar simultáneamente. Por ejemplo, la variable «género» da origen a resultados mutuamente excluyentes: hombre o mujer.

Colectivamente exhaustivos: Al menos uno de los sucesos debe ocurrir. Por ejemplo, en el experimento de lanzar un dado, cada resultado será par o impar. Por consiguiente, el conjunto es colectivamente exhaustivo.

4.3 Enfoques para asignar probabilidades

Probabilidad Clásica: El enfoque clásico, también denominado *enfoque a priori*, predice la probabilidad de un evento a partir del número de resultados favorables dividido para el número total de posibles resultados. Bajo este enfoque, cada elemento del experimento tiene la misma probabilidad de ser seleccionado. La probabilidad clásica de un evento se calcula de la siguiente forma:

$$Probabilidad\ de\ un\ evento = \frac{Número\ de\ resultados\ favorables}{Número\ \ total\ de\ posibles\ resultados}$$

Ejemplo 4.1 Considere el experimento de lanzar un dado. ¿Qué probabilidad hay que salga un número par?

Solución

Al lanzar el dado se pueden obtener seis resultados posibles, de los cuales tres, favorecen a la pregunta a números pares (el dos, el cuatro y el seis). Con base a este conjunto de resultados posibles, la probabilidad se calcula de la siguiente forma:

$$P = \frac{3}{6} = 0.5$$

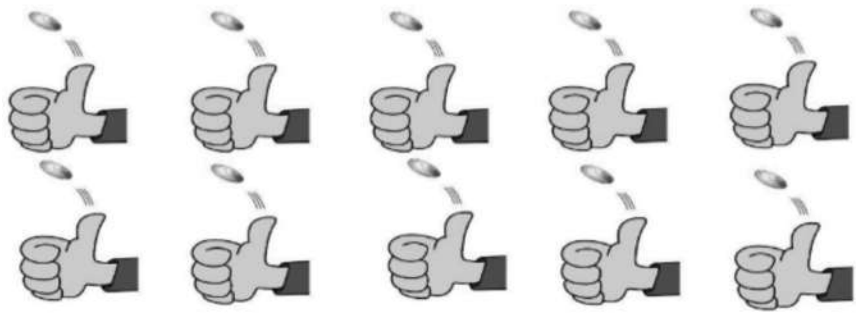
Si los eventos son mutuamente excluyentes y la secuencia de eventos es colectivamente exhaustiva, las probabilidades suman 1. Bajo este escenario, resulta innecesario llevar a cabo un experimento para determinar la probabilidad de un evento mediante el enfoque clásico, porque el número total de resultados se sabe antes de realizar el experimento.

Probabilidad empírica: El enfoque empírico, también denominado *enfoque de frecuencia relativa*, determina el valor de la probabilidad de un evento con base al número de intentos conocidos. Es decir, la probabilidad de un evento representa una fracción de los sucesos similares en el pasado. Consecuentemente, cada elemento del experimento no tiene la misma probabilidad de ser seleccionado. La fórmula para determinar la probabilidad de ocurrencia de un evento bajo el enfoque empírico es:

$$Probabilidad\ empírica = \frac{Número\ de\ veces\ que\ el\ evento\ ocurre}{Número\ \ total\ de\ observaciones}$$

El concepto de probabilidad empírica se basa en la ley de los grandes números, la cual indica que en un gran cantidad de intentos, la probabilidad empírica de un evento se aproxima a su probabilidad real.

Número de ensayos	Número de caras	Frecuencia relativa de las caras
1	0	0.00
10	3	0.30
50	26	0.52
100	52	0.52
500	236	0.47
1,000	494	0.49
10,000	5,027	0.50



La ley de los grandes números se puede explicar utilizando el ejemplo de lanzar una moneda común. En cada lanzamiento, el resultado puede ser cara o cruz. Si se lanza la moneda una sola vez, la probabilidad empírica de obtener una cara será cero o uno. Sin embargo, si se lanza la moneda repetidas veces, la probabilidad de que salga cara se aproximará a 0.5. Un experimento que involucre lanzar la moneda 1, 10, 50, 100, 500, 1,000 y 10,000 veces, muestra que la probabilidad empírica de obtener una cara se acerca a 0.5, lo cual coincide con el valor esperado según el enfoque clásico de la probabilidad.

Ejemplo 4.2 De los últimos 1,000 vuelos comerciales que partieron del aeropuerto de Guayaquil, hubo dos accidentes menores. Con base en esta información, ¿cuál es la probabilidad de que el siguiente vuelo no presente inconvenientes?

Solución

Con la fórmula descrita anteriormente obtenemos:

$$Probabilidad\ empírica = \frac{Número\ de\ veces\ que\ el\ evento\ ocurre}{Número\ \ total\ de\ observaciones}$$

$$P(A) = \frac{998}{1{,}000} = 0.998$$

La probabilidad de éxito del siguiente vuelo es aproximadamente 99%. También se puede decir que la probabilidad de que el siguiente vuelo presente inconvenientes es de 1%.

Probabilidad subjetiva: Este tipo de probabilidad indica la posibilidad de que un evento ocurra a partir de la experiencia o creencia (dogma) de una persona. Por ejemplo, un periodista podría determinar, a partir de su experiencia previa (o fanatismo), la probabilidad que tiene un equipo en ganar el campeonato ecuatoriano de fútbol.

En la gráfica 4.1 se resumen los enfoques de probabilidad.

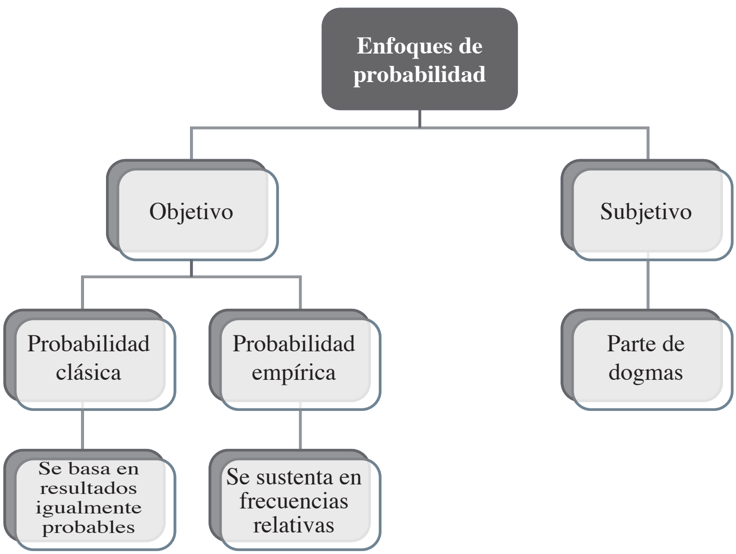


Gráfico 4.1 Enfoques para asignar probabilidades.

Ejemplo 4.3 Una encuesta a 40 estudiantes de UEES mostró que éstos siguen las siguientes carreras:

Contabilidad	12
Finanzas	6
Economía	5
Administración	7
Marketing	10

Tabla 4.1 Tabla de distribución de carreras de 40 estudiantes de UEES.

Suponga que se elige un estudiante al azar.

Solución

a. ¿Cuál es la probabilidad de que el estudiante siga la carrera de administración?

$$Probabilidad\ empírica = \frac{Número\ de\ veces\ que\ el\ evento\ ocurre}{Número\ \ total\ de\ observaciones}$$

$$P(E) = \frac{7}{40} = 0.18$$

b. ¿Qué concepto de probabilidad se emplea para hacer este cálculo?

Probabilidad empírica.

4.4 Regla de adición para calcular probabilidades

4.4.1 Regla especial de la adición

Si dos eventos A y B son mutuamente excluyentes, la regla especial de la adición establece que la probabilidad de que ocurra uno u otro es igual a la suma de sus probabilidades. Esta regla se expresa de la siguiente forma:

$$Regla\ especial\ de\ la\ adición\text{: } P(A\ o\ B) = P(A) + P(B)$$

En el caso de tres eventos mutuamente excluyentes (llámense A, B y C), la regla se expresa de la siguiente forma:

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

Ejemplo 4.4 Una máquina automática proveedora de paquetes con dulces contiene una combinación de figuritas que corresponden a diferentes frutas. La mayoría de estos paquetes contiene el peso correcto, aunque como consecuencia de la variación del tamaño de una figura de sandía o uva o banano, un paquete podría estar pesando menos o más. Una revisión de 3,800 paquetes que se llenaron el mes previo arrojó los siguientes datos:

Peso	Evento	Número de paquetes		Probabilidad de que ocurra el evento
Menos peso	A	200	$\leftarrow \frac{200}{3{,}800}$	0.053
Peso aceptable	B	3,400		0.895
Más peso	C	200		0.053
		3,800		1

Tabla 4.2 Tabla de datos de peso de paquetes de dulces en una máquina automática.

¿Cuál es la probabilidad de que un paquete en particular pese menos o más del peso aceptable?

Solución

El resultado «pesa menos» es el evento *A*; el resultado «pesa más» es el evento *C*. Al aplicar la regla especial de la adición se tiene:

$$P(A \text{ o } C)= P(A) + P(C)=0.053+0.053=0.106$$

Se puede observar que los eventos son mutuamente excluyentes, lo que significa que un paquete de dulces con figuras de frutas no puede pesar menos, tener el peso aceptable y pesar más al mismo tiempo. También, son colectivamente exhaustivos; es decir, un paquete seleccionado debe pesar menos, tener un peso aceptable o pesar más. Los eventos se pueden representar gráficamente a través del gráfico 4.2.

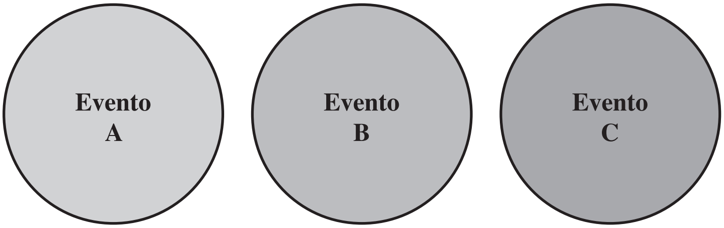


Gráfico 4.2 Interpretación de la Regla especial de adición para eventos mutuamente excluyentes.

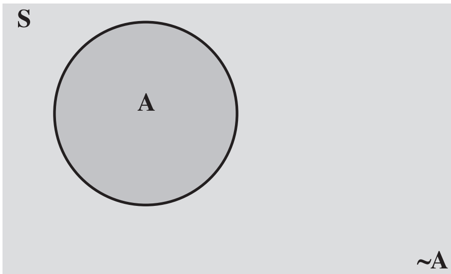


Gráfico 4.3 Diagrama de Venn.

Regla del complemento: esta regla señala que la probabilidad de que ocurra cualquier evento A será siempre igual a la unidad menos la probabilidad de que ocurra el evento contrario o complementario a A. Formalmente esto se representa de la siguiente manera:

$$Regla\ del\ complemento: P(A) = 1 - P(\sim A)$$

Con esta regla podemos calcular la probabilidad de ocurrencia de cualquier suceso si conocemos la probabilidad de su complemento, y viceversa. Esto es particularmente importante porque en muchas situaciones del mundo real donde se debe calcular la probabilidad de un evento, es mucho más fácil calcular directamente la probabilidad de su complemento (Parada, 2021). Gráficamente se puede representar como se muestra en el gráfico 4.3.

Utilizando los datos del ejemplo 4.4, la probabilidad de que un paquete de golosinas mixtas pese menos de lo ideal es de 0.053 y la probabilidad de que pese más del ideal es también de 0.053; con estos valores podemos obtener (por regla de complemento), la probabilidad de que un paquete tenga un peso aceptable.

$$P(B) = 1 - [P(A) + P(C)] = 1 - [0.053 + 0.053] = 0.894$$

El paquete tiene un peso aceptable si no tiene ni más ni menos peso. Gráficamente esto se ilustra en el gráfico 4.4.

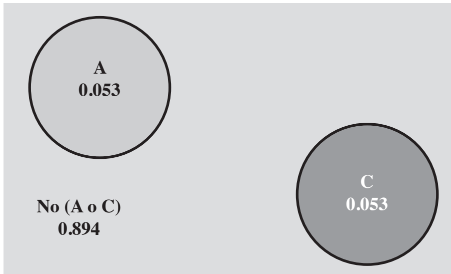


Gráfico 4.4 Interpretación de la Regla del complemento.

4.4.2 Regla general de la adición

Los resultados de un experimento pueden no ser mutuamente excluyentes. En este caso, no se debería aplicar la regla especial de la adición, y en su lugar se debería aplicar la siguiente regla:

$$Regla\ general\ de\ la\ adición: P(A\ o\ B) = P(A) + P(B) - P(A\ y\ B)$$

En donde:

- El conectivo «o» representa a la «unión» en la teoría de conjuntos, y se la puede representar con el símbolo «».
- El conectivo «y» representaría la «intersección» en la teoría de conjuntos, y se la puede representar con el símbolo «».

Ejemplo 4.5 Se levantó información de 200 turistas que visitaron Guayaquil el año pasado. La encuesta reveló que 120 personas fueron al *Malecón 2000* y 100 al barrio *Las Peñas*. Además, se sabe que 60 personas visitaron ambos sitios. ¿Cuál es la probabilidad de que una persona seleccionada al azar haya visitado el Malecón o Las Peñas?

Solución

Si se emplea la regla especial de la adición, la probabilidad de seleccionar a un turista que haya ido al malecón es 0.60 (120/200). La probabilidad de seleccionar a un turista que haya ido a Las Peñas es 0.50 (100/200). La suma de ambas probabilidades es 1.10, pero la probabilidad no puede ser mayor que 1. La razón de esto se puede explicar debido a que hubo turistas que visitaron ambos sitios. Por ende, se aplicará la regla general de la adición.

$$P(Malecón\ o\ Las\ Peñas) = P(Malecón) + P(Las\ Peñas) - P(Malecón\ y\ las\ Peñas)$$

$$P(Malecón\ o\ Las\ Peñas) = 0.60 + 0.50 - 0.30 = 0.80$$

Se observa que existe una probabilidad del 80 % de que una persona seleccionada al azar haya visitado el Malecón o Las Peñas. Gráficamente esto se puede representar como sigue:

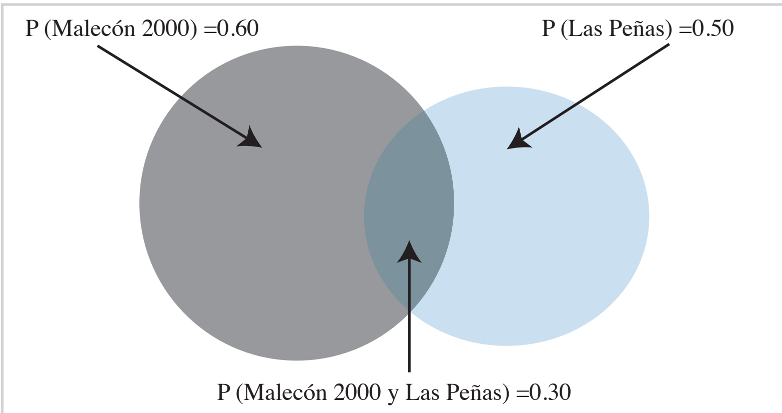


Gráfico 4.5 Diagrama de Venn que muestra la unión de dos eventos A y B.

Probabilidad conjunta: La probabilidad conjunta se refiere a la probabilidad de que dos o más eventos ocurran simultáneamente, por lo que no son eventos mutuamente excluyentes. En otras palabras, es la probabilidad de que dos o más variables aleatorias tomen valores específicos al mismo tiempo.

Ejercicio 4.1

Los eventos A y B son mutuamente excluyentes. Suponga que P(A) = 0.30 y P (B) = 0.20.

a. ¿Cuál es la probabilidad de que A o B ocurran?

$$P(A\ o\ B) = P(A) + P(B)$$

$$P(A\ o\ B) = 0.30 + 0.20 = 0.50$$

La probabilidad de que ocurra A o B es el 50%.

b. ¿Cuál es la probabilidad de que ni A ni B ocurran?

$$P(Ninguna) = 1 - P(A) + P(B)$$

$$P(Ninguna) = 1 - 0.50 = 0.50$$

La probabilidad de que no ocurra ningún evento es 50%.

Ejercicio 4.2 Un estudio de 300 empresas de publicidad reveló los siguientes ingresos después de impuestos:

Ingresos después de impuestos	P(X)	Número de empresas
Menos de $1 millón	A	160
De $1 millón a $20 millones	B	90
$20 millones o más	C	50
	Total	**300**

Tabla 4.3 Tabla de Ingresos después de Impuestos de 300 Empresas de Publicidad.

a. ¿Cuál es la probabilidad de que una empresa de publicidad seleccionada al azar tenga un ingreso después de impuestos menor a un millón de dólares?

$$P(A) = \frac{160}{300} = 0.53$$

La probabilidad de que una empresa de publicidad tenga menos de $1 millón es de 53%.

b. ¿Cuál es la probabilidad de que una empresa de publicidad seleccionada al azar tenga un ingreso después de impuestos entre 1 y 20 millones de dólares o un ingreso de 20 millones de dólares o más? ¿Qué regla de probabilidad aplicó?

Para resolver este inciso se utilizó la regla especial de la adición:

$$P(B\ o\ C) = P(B) + P(C)$$

$$P(B\ o\ C) = \frac{90}{300} + \frac{50}{300}$$

$$P(B\ o\ C) = 0.3 + 0.17 = 0.47$$

La probabilidad de que una empresa de publicidad tenga un ingreso entre 1 y 20 millones de dólares o un ingreso de 20 millones de dólares o más después de impuestos es de 46%. Se aplicó la regla especial de la adición.

Ejercicio 4.3 Las probabilidades de los eventos A y B son 0.20 y 0.30, respectivamente. La probabilidad de que A y B ocurran es de 0.15. ¿Cuál es la probabilidad de que A o B ocurran?

Datos:

P(A) = 0.20

P(B) = 0.30

$P(A\ y\ B)$=0.10

$$Regla\ general\ de\ la\ adición: P(A\ o\ B) = P(A) + P(B) - P(A\ y\ B)$$

$$P(A\ o\ B) = 0.20 + 0.30 - 0.10 = 0.40$$

La probabilidad de que suceda A o B es de 40%.

4.5 Regla de la multiplicación

4.5.1 Regla especial de la multiplicación

Esta regla se utiliza para calcular la probabilidad conjunta de dos o más eventos que sean independientes. Entendemos por independencia a que la ocurrencia de un evento no tiene efecto sobre la probabilidad de ocurrencia del otro. Es decir, la probabilidad de que ocurran ambos eventos al mismo tiempo es igual al producto de las probabilidades de cada evento por separado. Para ilustrar la independencia, suponga que se lanzan al aire dos monedas. El resultado del lanzamiento de una moneda (cara o cruz) no se altera por el resultado de cualquier moneda lanzada previamente.

$$Regla\ especial\ de\ la\ multiplicación: P(A\ y\ B) = P(A) * P(B)$$

Para tres eventos independientes, la regla de multiplicación especial para determinar la probabilidad de ocurrencia es:

$$P(A\ y\ B\ y\ C) = P(A) * P(B) * P(C)$$

Ejemplo 4.6 Una encuesta realizada el año pasado a estudiantes de UEES, reveló que el 40% utiliza el BUEES para trasladarse al campus universitario. Se seleccionaron al azar dos estudiantes (de los encuestados). ¿Cuál es la probabilidad de que ambos hayan utilizado el BUEES?

Solución

La probabilidad de que el primero haya utilizado el BUEES es P(A)=0.40. la probabilidad de que el segundo haya utilizado el BUEES también es P(B)=0.40. Como hay suficiente transporte, se entiende que el uso de uno no afecta al otro, por ende, A y B son independientes. Consecuentemente, la probabilidad de que ambos hayan utilizado el BUEES es:

$$P(A\ y\ B) = P(A) * P(B) = (0.40) * (0.40) = 0.16$$

4.5.2 Regla general de la multiplicación

Esta regla establece que la probabilidad conjunta de dos o más eventos es igual al producto de las probabilidades condicionales de cada evento dado que los eventos anteriores ya hayan acontecido.

Ejemplo 4.7 Imaginemos que en un refrigerador hay 20 latas de refresco, de las cuales 13 son normales y 7 son dietéticas. Si se saca una lata al azar, hay una probabilidad del 7/20 de que sea una lata de refresco dietético y una probabilidad del 13/20 de que sea una lata de refresco normal. Luego, se elige una segunda lata del refrigerador sin devolver la primera. La probabilidad de que la segunda lata sea de refresco dietético depende de que la primera lo haya sido o no. La probabilidad de que la segunda lata sea de refresco dietético es:

Solución

- 6/19 = si la primera bebida es dietética (solo quedan seis latas de refresco dietético en el refrigerador).
- 7/19 = si la primera lata elegida es normal (los siete refrescos aún están en el refrigerador).
- La fracción 6/19 (o 7/19) es una **probabilidad condicional** porque su valor se encuentra condicionado (o depende) del hecho de que un refresco regular o dietético haya sido el primero en ser seleccionado del

refrigerador.

En términos matemáticos, si A y B son dos eventos, entonces la probabilidad conjunta de que ambos eventos ocurran simultáneamente es igual a:

$$Regla\ general\ de\ la\ multiplicación: P(AyB) = P(A) * P(B|A)$$

Ejemplo 4.8 Suponga que P(A) = 0.40 y P(B|A) = 0.30. ¿Cuál es la probabilidad conjunta de A y B?

Solución

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

$$P(AyB) = 0.40 * 0.30 = 0.12$$

La probabilidad conjunta de A y B es de 0.12 (12%).

Ejemplo 4.9 El profesor de estadística mencionó que solo el 30 % de sus estudiantes aprobó el curso de manera directa el periodo anterior. ¿Cuál es la probabilidad de que tres estudiantes seleccionados al azar hayan aprobado el curso?

Solución

$$P(E1y\ E2\ y\ E3) = (0.30)(0.30)(0.30) = 0.027$$

La probabilidad de que tres estudiantes seleccionados al azar hayan aprobado el curso es de 2.7%.

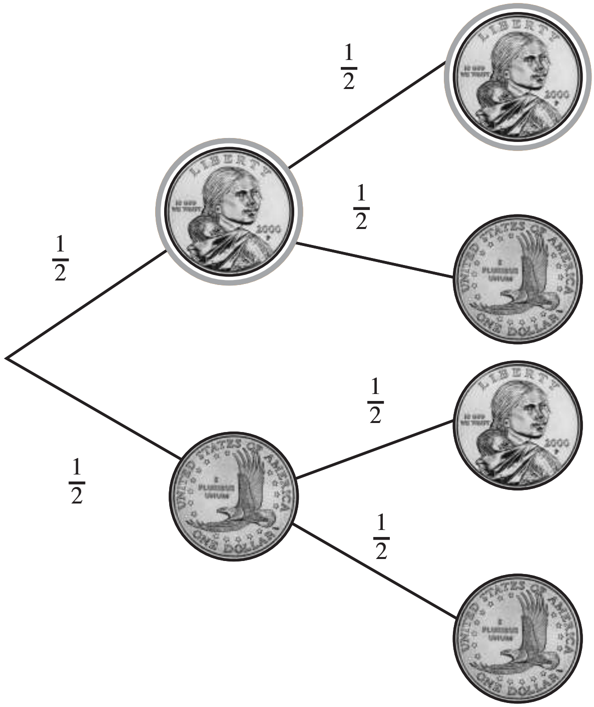
4.6 Diagramas de árbol

Un diagrama de árbol es un gráfico para representar la secuencia de eventos que pueden ocurrir en un experimento aleatorio. Consiste en una serie de ramas que representan cada resultado posible del experimento y nodos que representan cada evento o fase del experimento. La raíz del árbol representa el comienzo del experimento y cada rama que se extiende desde la raíz representa el posible resultado del primer evento. Cada rama se divide en ramas secundarias, que a su vez representan los posibles resultados del siguiente evento, y así sucesivamente hasta llegar al resultado final.

Ejemplo 4.10

a. ¿Cuáles son todos los resultados posibles que pueden ocurrir al lanzar una moneda 2 veces?

Solución



En un diagrama de árbol las «ramas» deben sumar 1 (en cada etapa). La respuesta a la pregunta es que existe cuatro resultados posibles al lanzar dos monedas.

b. ¿Cuál es la probabilidad de que al lanzar la moneda dos veces primero salga cara y luego águila?

$$\frac{1}{2} * \frac{1}{2} = \frac{1}{4} = 25\%$$

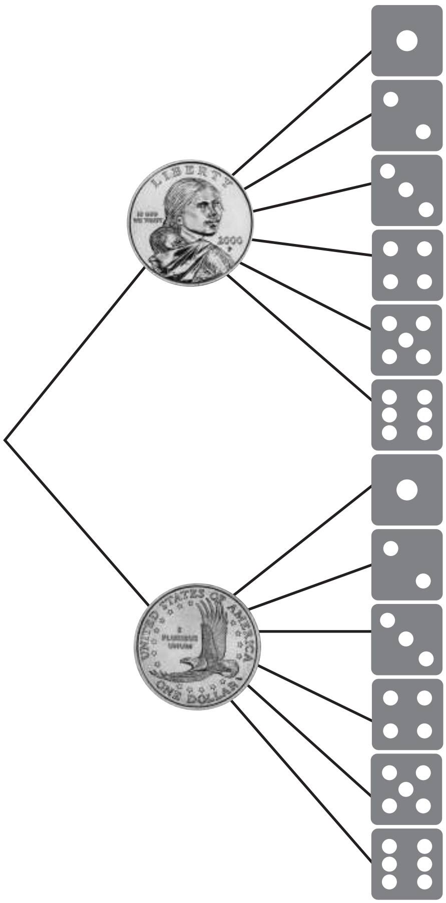
c. ¿Cuál es la probabilidad de que al lanzar la moneda dos veces primero salga cara y luego cara otra vez?

$$\frac{1}{2} * \frac{1}{2} = \frac{1}{4} = 25\%$$

Ejemplo 4.11: ¿Cuáles son todos los resultados posibles que pueden ocurrir al lanzar una moneda y luego un dado?

Solución

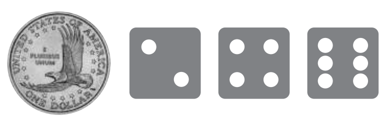
Hay 12 resultados posibles



a. ¿Cuál es la probabilidad de que al lanzar la moneda caiga cara y luego al lanzar el dado salga 2?

$$\frac{1}{2} * \frac{1}{6} = \frac{1}{12}$$

b. ¿Cuál es la probabilidad de que al lanzar la moneda caiga águila y luego al lanzar el dado salga un número par?



$$\frac{1}{2} * \frac{3}{6} = \frac{1}{4}$$

4.7 Teorema de Bayes

Durante el siglo XVIII, el reverendo Thomas Bayes, quien era un clérigo presbiteriano de origen inglés, se preguntó acerca de la existencia de Dios. Debido a su inclinación por las matemáticas, Bayes intentó desarrollar una fórmula que permitiera determinar la probabilidad de la existencia de Dios en función de la evidencia disponible en la Tierra. Posteriormente, Pierre-Simon Laplace mejoró el trabajo de Bayes y lo denominó teorema de Bayes.

El teorema de Bayes es utilizado para calcular la probabilidad de un suceso, teniendo información de antemano sobre ese suceso. La fórmula es la siguiente:

$$P(A_i|B) = \frac{P(A_i) * P(B|A_i)}{P(A_1) * P(B|A_1) + P(A_2) * P(B|A_2)}$$

En donde A_1 y A_2 son eventos mutuamente excluyentes y colectivamente exhaustivos. Por otro lado, A_i se refiere a cualquiera de ambos eventos

Suponga que, en el algún momento, el 5% de la población de Ecuador tuvo COVID-19, A_1 es el evento «padece la enfermedad» y A_2 es el evento «no padece la enfermedad». Por lo tanto, si se selecciona al azar a una persona de Ecuador, la probabilidad de que el individuo elegido padezca la enfermedad es el 5%, ya que la probabilidad a priori de ese evento es 0.05 ($P(A_1)$= 0.05). Se denomina probabilidad a priori a aquella probabilidad que está basada en el nivel de información actual. Por otro lado, la probabilidad de que una persona no padezca la enfermedad es 95% ($P(A_2)$= 0.95). Ahora, suponga que existe una técnica de diagnóstico de la enfermedad que tiene una eficacia del 90% para aquellos que la padecen ($P(B|A_1)$= 0.90). Además, la probabilidad de que la prueba indique la presencia de la enfermedad en una persona que en realidad no la padece es de 15% ($P(B|A_2)$= 0.15).

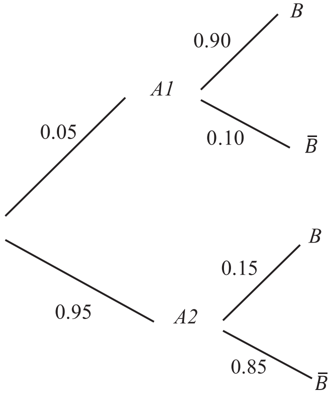
$$B = Prueba\ positiva \qquad \bar{B} = Prueba\ negativa$$

Se elige al azar a una persona y se aplica la prueba, los resultados indicaran que la enfermedad está presente. Ahora, se plantea la siguiente pregunta: ¿Cuál es la probabilidad de que la persona en realidad padezca de la enfermedad?

$$P(A_i|B) = \frac{P(A_i) * P(B|A_i)}{P(A_1) * P(B|A_1) + P(A_2) * P(B|A_2)}$$

$$P(A_i|B) = \frac{(0.05) * (0.90)}{(0.05) * (0.90) + (0.95) * (0.15)} = 0.24$$

Si una prueba da positivo, la probabilidad de que una persona tenga la enfermedad es del 24%. Sin embargo, si se toma al azar a alguien de la población, la probabilidad de que esa persona tenga la enfermedad es del 5%. Si esa persona se somete a la prueba y el resultado es positivo, entonces la probabilidad de que realmente tenga la enfermedad aumenta cinco veces (pasa del 5% al 24%).



$$P(A_i|B) = \frac{P(A_i) * P(B|A_i)}{P(A_1) * P(B|A_1) + P(A_2) * P(B|A_2)}$$

$$P(A_i|B) = \frac{(0.05) * (0.90)}{(0.05) * (0.90) + (0.95) * (0.15)} = 0.24$$

4.8 Principios de conteo

El principio de conteo establece que, si se tienen *m* formas de realizar una tarea y *n* formas de hacer otra tarea, entonces se tendrán *m x n* formas de realizar ambas tareas simultáneamente. Este principio es útil para determinar el número total de resultados posibles en un experimento cuando la cantidad de posibilidades es pequeña y se pueden contar fácilmente. De esta forma, se puede utilizar el principio de conteo para calcular la probabilidad de un evento específico en función de la cantidad total de resultados posibles.

Por ejemplo, si se lanza un dado se sabe que existen seis posibles resultados de lanzamiento. No obstante, cuando se trata de un experimento con una gran cantidad de posibles resultados, como por ejemplo el lanzamiento de una moneda 10 veces y contar el número de caras y águilas, el proceso de contar todas las posibilidades sería muy laborioso.

Para simplificar esta tarea, se pueden utilizar tres fórmulas de conteo: 1) la fórmula de la multiplicación (no debe confundirse con la regla de la multiplicación explicada en el capítulo 3), 2) la fórmula de las permutaciones, y 3) la fórmula de las combinaciones. De esta manera, se puede determinar el número de resultados posibles de manera más eficiente. La fórmula de multiplicación es la siguiente:

$$\textit{Número total de disposiciones} = (m)(n)$$

Esta fórmula se puede extender a más de dos eventos. Así, en el caso de tres eventos el número total de disposiciones será:

$$\textit{Número total de disposiciones} = (m)(n)(o)$$

Ejemplo 4.12 Se desea determinar la cantidad de opciones distintas que un distribuidor de automóviles puede ofrecer en su anuncio, donde por un precio de \$29,999 se puede adquirir un sedán de dos puertas o uno de cuatro puertas, y se puede elegir entre rines de acero o rines de aluminio.

Solución

Por supuesto, el distribuidor podría determinar el número total de disposiciones haciendo un diagrama y contando. Mediante la fórmula de la multiplicación se verifica el resultado, en la cual m es el tipo de modelo y n es el tipo de rin.

$$\textit{Número total de disposiciones} = (m)(n) = (3)(2) = 6$$

No resulta difícil contar todas las posibles combinaciones de modelos y rines en este ejemplo. Sin embargo, suponga que el distribuidor decide ofrecer 8 modelos y 6 tipos de rines. Resultaría tedioso representar y contar todas las posibles alternativas, por lo que es mejor aplicar la fórmula de la multiplicación.

$$\textit{Número total de disposiciones} = (m)(n) = (8)(6) = 48$$

Es importante destacar que la fórmula de la multiplicación se emplea para calcular el número de posibles disposiciones de dos o más grupos, mientras que la fórmula de permutaciones se utiliza para calcular el número posible de disposiciones cuando solo hay un grupo de objetos.

En estadística, una permutación se refiere a la disposición ordenada de un conjunto de objetos o eventos. La fórmula de permutaciones se utiliza para calcular el número total de permutaciones posibles de un conjunto de *n* elementos, tomados en grupos de *r* elementos.

$$Fórmula\ de\ las\ permutaciones\text{: } {}_{n}P_{r} = \frac{n!}{(n-r)!}$$

En donde:

n: Representa el total de objetos.

r: Representa el total de objetos seleccionados.

Por definición, 0! =1.

Ejemplo 4.13 Se tiene un recipiente que contiene tres pelotas de tres colores diferentes: morado, azul y turquesa. Se extraen las pelotas de manera secuencial, es decir, primero una y luego las otras dos. ¿Cuántas permutaciones diferentes se pueden generar al extraer las tres pelotas?

Solución

$$nPr = \frac{n!}{(n-r)!} = 3P3 = \frac{3!}{(3-3)!} = \frac{1*2*3}{0!} = 6$$

Por otro lado, si el orden de los objetos seleccionados no es importante, cualquier selección se denomina combinación. La fórmula para contar el número de *r* combinaciones de objetos de un conjunto de *n* objetos es:

$$nCr = \frac{n!}{r!\,(n-r)!}$$

Ejemplo 4.14 En una actividad de clasificación, se tienen tres letras: A, B y C. Se desea determinar, cuántas combinaciones de dos letras se pueden formar considerando el orden de las mismas y cuántas combinaciones de dos letras se pueden formar sin importar dicho orden.

a) ¿Cuántas combinaciones de dos letras se pueden formar considerando el orden?

Solución

Se debe emplear la fórmula de permutaciones en este caso, porque el orden es importante.

$$nPr = \frac{n!}{(n-r)!} = 3P3 = \frac{3!}{(3-3)!} = \frac{1*2*3}{0!} = 6$$

Por lo tanto, en este caso existen 6 permutaciones las cuales serían: AB; BA; AC; CA; BC; CB.

b) ¿Cuántas combinaciones de dos letras se pueden formar con las letras A, B y C sin considerar el orden de las mismas?

Se debe aplicar la fórmula de combinación porque el orden los objetos no es importante.

$$nCr = \frac{n!}{r!\,(n-r)!} = 3C2 = \frac{3!}{2!\,(3-2)!} = \frac{3*2*1}{(2*1)} = 3$$

Por lo tanto, en este caso hay tres combinaciones posibles: AB, AC y BC.

Ejemplo 4.15 La cafetería de UEES *Caramel Coffee* utiliza equipo de tres empleados para trabajar en la dulcería cada tarde-noche. Hay siete empleados disponibles. ¿Cuántos equipos diferentes pueden programarse para cubrir el turno?

Solución

Hay 35 posibles combinaciones, determinadas por:

$$7C3 = \frac{n!}{r!\,(n-r)!} = \frac{7!}{3!\,(7-3)!} = \frac{7!}{3!\,4!} = 35$$

Los siete empleados, en grupo de tres, crearían la posibilidad de 35 equipos diferentes.

Ejercicios propuestos del capítulo

Ejercicio 4.1 En cado uno de los siguientes enunciados, indique si se utilizó el concepto de probabilidad clásica, empírica o subjetiva.

Un jugador de Barcelona acertó 15 de los últimos 20 penales que cobró. La probabilidad de que anote un penal en el siguiente partido es de 75%.

a. Un estudiante después de rendir el examen final de estadística menciona que tiene mucha fe en aprobar a pesar de que no estudió nada.

b. Para analizar problemas ambientales se forma un comité de estudiantes con 7 miembros. ¿Cuál es la probabilidad de que cualquier de los 7 sea elegido vocero del equipo?

c. Usted compra 1 de los 100 mil boletos vendidos de la lotería nacional. ¿Cuáles son las posibilidades de que gane el premio mayor?

d. July descubre que José, el pretendiente de su mejor amiga, ha sido infiel a 4 de sus últimas 5 novias. Con base a esto, July le dice a su amiga que existe un 80% de probabilidad de que le sea infiel a ella si lo acepta. ¡Amiga, huye de ahí!

Ejercicio 4.2 En un ánfora hay 6 pelotas blancas y 4 amarillas. ¿Cuál es la probabilidad de sacar 3 pelotas amarillas sin reemplazo?

Ejercicio 4.3 Una empresa tiene tres líneas de producción de puertas para automóviles. La primera línea, produce el 38% de la producción, la segunda el 41% y la tercera el restante. De la primera línea, el 11 % de la producción son puertas defectuosas, de la línea dos el 8% y de la tercera el 14%. Si se elige al azar una puerta cualquiera de toda la producción, calcule la probabilidad de que (justifique su respuesta dibujando el árbol de probabilidades. Redondee a 4 decimales las probabilidades):

a. Haya sido producido por la línea 1 y que no sea defectuoso.

b. Sea una puerta no defectuosa.

c. Haya sido producido por la línea 3 dado que sea defectuoso.

d. No se haya producido por la línea 2 dado que no sea defectuoso.

Ejercicio 4.4 Se eligió una muestra de 40 ejecutivos de EP PETROECUADOR para someter a prueba un cuestionario. Una pregunta relacionada con cuestiones ambientales requería que se respondiera sí o no.

a. ¿En qué consiste el experimento?

b. Indique un posible evento.

c. Diez de los 40 ejecutivos respondieron que sí. Con base en estas respuestas de la muestra, ¿Cuál es la probabilidad de que un ejecutivo de EP PETROECUADOR responda de manera afirmativa?

d. ¿Qué concepto de probabilidad se ilustra?

e. ¿Los posibles resultados son igualmente probables y mutuamente excluyentes?

Ejercicio 4.5 En cada uno de los enunciados indique si es verdadero (V) o falso (F), según corresponda.

a. La propiedad de mutuamente excluyente significa que por lo menos uno de los eventos debe ocurrir cuando se lleva a cabo el experimento. ( )

b. El enfoque de probabilidad subjetivo se sustenta en frecuencias relativas. ( )

c. Si los eventos A y B son mutuamente excluyentes, y P(A) = 0.30 y P (B) = 0.20, entonces la probabilidad de que A o B ocurran es 0.10. ( )

d. Las probabilidades de los eventos A y B son 0.20 y 0.30, respectivamente. La probabilidad de que A y B ocurran es de 0.15. Entonces, la probabilidad de que A o B ocurran es 0.35. ( )

e. Si el orden de los objetivos no es importante, cualquier selección se denomina permutación. ( )

Ejercicio 4.6 El acuario *Akua* de Guayaquil contiene 140 peces espada. De estos, 80 son rojos (44 hembras y 36 machos), y 60 son amarillos (36 hembras y 24 machos). Si se captura uno al azar en este acuario.

a. ¿Cuál es la probabilidad de que sea rojo?

b. ¿Cuál es la probabilidad de que sea un macho?

c. ¿Cuál es la probabilidad de que sea un macho rojo?

d. ¿Cuál es la probabilidad de que sea macho o rojo?

Ejercicio 4.7 Se tiene una ruleta y un dado. Realice el diagrama de árbol de lo que puede ocurrir al girar la ruleta de 4 colores y tirar el dado.

a. ¿Cuáles son todos los resultados posibles?

b. ¿Cuál es la probabilidad de que caiga color amarillo y número 2?

c. ¿Cuál es la probabilidad de que caiga color azul y número par?

d. ¿Cuál es la probabilidad de que caiga color verde y número 5 o 6?

Ejercicio 4.8 Resuelva las siguientes operaciones:

a. $\frac{20!}{17!}$

b. ${}_7P_3$

c. ${}_7C_2$

Ejercicio 4.9 Una fábrica de celulares dispone de dos máquinas A_1 y A_2 que elaboran el 60% y el 40% de la producción. El porcentaje de celulares defectuosos que produce cada máquina es del 5% y del 10%, respectivamente.

a. ¿Cuál es la probabilidad que el celular haya sido fabricado por la máquina A_1, sabiendo que es defectuoso?

b. ¿Cuál es la probabilidad que el celular haya sido fabricado por la máquina A_2, sabiendo que no es defectuoso?

Ejercicio 4.10 Un encuestador local ha creado un cuestionario de 20 preguntas para evaluar el rendimiento del alcalde de la ciudad. El encuestador decide seleccionar 12 preguntas del cuestionario. ¿De cuántas maneras distintas se pueden elegir estas 12 preguntas, teniendo en cuenta su orden dentro del cuestionario?

Capítulo V

DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

V. DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

5.1 Introducción

En el capítulo anterior se discutieron conceptos introductorios a la teoría de probabilidad. Se analizaron, fundamentalmente, las reglas de adición y multiplicación. Asimismo, se estudió el teorema de Bayes y su relación con la teoría de probabilidad. En este capítulo se abordarán conceptos relacionados a variables aleatorias, distribución de probabilidad, tipos de distribución de probabilidad y medidas de tendencia central de una distribución de probabilidad discreta.

5.2 Espacio muestral y puntos muestrales

Para poder comprender el concepto de espacio muestral y puntos muestrales es importante recordar qué es una población. Se denomina población a aquel conjunto de todos los resultados posibles de un experimento aleatorio. A la población también se la puede definir como espacio muestral. Dentro de un espacio muestral a cada miembro se denomina punto muestral.

Por ejemplo, en un experimento en el que se lanza 2 monedas, el espacio muestral, o población, consiste de 4 resultados posibles: CC, CS, SC y SS, donde C representa cara y S representa sello. El resultado CC establece que se obtuvo cara el primer lanzamiento y cara nuevamente en el segundo, CS significa una cara en el primer lanzamiento y sello en el segundo, y así sucesivamente. Cada uno de los sucesos de este experimento representa un punto muestral.

5.3 Variables aleatorias

En estadística se denomina variable aleatoria a una variable cuyo valor se determina por el resultado de un experimento al azar. Suelen denotarse con letras mayúsculas como X, Y, Z, y cada uno de los sucesos posibles se representan con letras minúsculas, x, y, z.

Algunos ejemplos de variables aleatorias incluyen:

- Número de caras/sellos al lanzar cuatro monedas.
- Número de personas que estudian economía en una muestra de 100 individuos.
- Altura de una persona escogida al azar.
- Por otro lado, algunos ejemplos de variables que no son aleatorias son:
- Seleccionar una camiseta de un closet.
- Seleccionar un balón de un cesto que contiene balones azules, amarrillos y rojos.

5.3.1 Variables aleatorias discretas

Existen dos clasificaciones de variables aleatorias: discretas y continuas. Una variable aleatoria discreta adopta valores «claramente separados» (Lind, Marchal, y Wathen, 2012).

Un ejemplo de variable aleatoria discreta es el número de tarjetas de crédito y/o débito que posee cada uno de los clientes de un banco (Tabla 5.1)

Cantidad de tarjetas	Porcentaje de clientes (%)
0	5%
1	20%
2	10%
3	35%
4	12%
5 o más	18%
TOTAL	**100%**

Tabla 5.1 Número de tarjetas de clientes de un banco.

En este caso, los sucesos, son la cantidad de tarjetas que posee un cliente y solo pueden tomar valores enteros como 0, 1, 2, 3, etc. Es importante mencionar que en ciertas situaciones una variable aleatoria discreta puede tomar valores no enteros, como valores fraccionarios o decimales. Es importante recordar que el concepto de variable discreta o continua no está asociado a si los valores que toman son enteros o decimales.

5.3.2 Variables aleatorias continuas

Por otro lado, las variables aleatorias continuas son aquellas variables que pueden adoptar cualquier valor dentro de un intervalo de valores. Por ejemplo, la estatura de una persona puede ser considerada una variable continua, ya que esta puede adquirir cualquier valor en un intervalo determinado (Tabla 5.2).

Estatura	Persona
1.54	1
1.67	2
1.75	3
1.56	4
1.69	5
1.60	6
TOTAL	**100%**

Tabla 5.2 Estatura de seis personas escogidas al azar.

5.4 Distribución de probabilidad

La distribución de probabilidad de una variable aleatoria es una función que asigna a cada posible valor de la variable la probabilidad de que dicho valor ocurra o se cumpla. En otras palabras, es una medida de la probabilidad asociada a cada resultado posible de una variable aleatoria. La distribución de probabilidad es similar a una distribución de frecuencias relativas, sin embargo, en lugar de hacer referencia a eventos pasados, describe la probabilidad de que un evento suceda en el futuro.

En una distribución de probabilidad existen tres características esenciales que deben presentarse:

- La probabilidad de un resultado específico se encuentra entre 0 y 1, inclusive.
- Los resultados son eventos mutuamente excluyentes. Es decir, los eventos no pueden suceder al mismo tiempo.
- La lista es exhaustiva. Por lo que, la suma de las probabilidades de los distintos eventos es igual a 1.

Por ejemplo, suponga que se interesa conocer el número de caras que se obtienen al lanzar tres veces una moneda. En este caso, los posibles resultados del evento serían: obtener tres caras, dos caras, una cara o cero caras (tabla 5.3).

Posible resultado	Número de lanzamiento			Número de caras
	Primero	*Segundo*	*Tercero*	
1	Cara	Cara	Cara	3
2	Cara	Cara	Cruz	2
3	Cara	Cruz	Cara	2
4	Cara	Cruz	Cruz	1
5	Cruz	Cara	Cara	2
6	Cruz	Cara	Cruz	1
7	Cruz	Cruz	Cara	1
8	Cruz	Cruz	Cruz	0

Tabla 5.3 Posibles resultados en tres lanzamientos de una moneda.

Bajo este escenario, la distribución de probabilidad de los eventos sería la siguiente:

Número de caras, x	Probabilidad del resultado P (x)	
0	1/8	0,125
1	3/8	0,375
2	3/8	0,375
3	1/8	0,125
TOTAL	**1**	**1**

Tabla 5.4 Distribución de probabilidad de los eventos relativos a cero, una, dos y tres caras en lanzamiento de tres monedas.

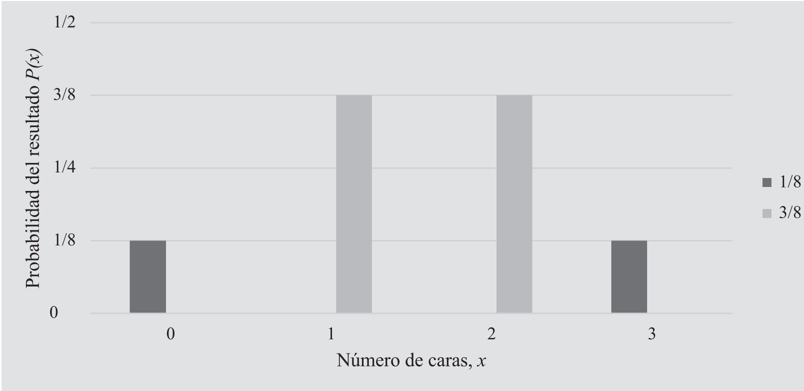


Gráfico 5.1 Distribución de probabilidades de los eventos relativos a cero, una, dos y tres caras en lanzamiento de tres monedas.

5.5 Tipos de distribución de probabilidad

Las distribuciones de probabilidad se clasifican en dos grandes grupos: continuas y discretas. Dentro la distribución de probabilidad de variables discretas se encuentran la distribución Binomial, Hipergeométrica y Poisson. Por otra parte, en la distribución de variables continuas tenemos a la distribución uniforme, normal y exponencial (gráfico 5.2).

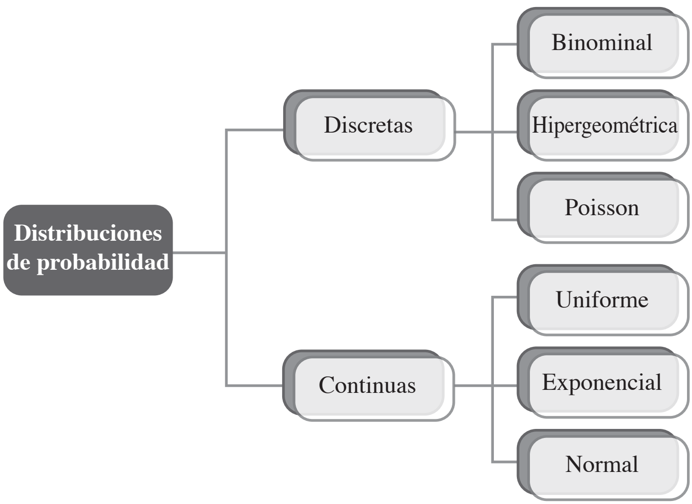


Gráfico 5.2 Tipos de distribución de probabilidad.

A lo largo de este capítulo se profundizará en la distribución de probabilidad de variables discretas y sus clasificaciones.

5.6 Media, varianza y desviación estándar de una distribución de probabilidad discreta

En el capítulo 3 se estudiaron las medidas de tendencia central y dispersión de una distribución de frecuencias. Se estableció que, en general, la media indica la posición central de los datos (salvo cuando los datos están sesgados) y la varianza indica la dispersión. En una distribución de probabilidad igualmente se pueden calcular estas medidas de tendencia central y dispersión. La interpretación es muy similar a las que se dió en el capítulo 3.

5.6.1 Media de una distribución de probabilidad discreta

En este caso, la media es «aquel valor típico para representar la posición central de una distribución de probabilidad» (Lind, Marchal, y Wathen, 2012); representa el valor promedio de la variable aleatoria. En una distribución de probabilidad discreta la media se calcula de la siguiente forma:

$$\mu = \Sigma[xP(x)]$$

Donde *P(x)* es la probabilidad de un valor particular *x*. Es decir, cada valor *x* se multiplica por las probabilidades del que mismo ocurra y se suman los productos de las multiplicaciones.

5.6.2 Varianza y desviación estándar de una distribución de probabilidad discreta

Como se mencionó previamente, la media no especifica el nivel de dispersión o variación de una distribución. Sin embargo, la varianza sí lo hace. La varianza indica que tan cercanos están las observaciones de la muestra en relación a la media. Si la varianza es alta, significa que existe gran dispersión en la muestra. Para el caso de una distribución de probabilidad discreta, la varianza se calcular de la siguiente forma:

$$\sigma^2 = \Sigma[(x-\mu)^2 P(x)]$$

Por otro lado, la desviación estándar es la raíz cuadrada de la varianza y se representa como sigue:

$$\sigma = \sqrt{\sigma^2}$$

Ejemplo 5.1 Juan trabaja como vendedor de automóviles nuevos en una concesionaria de Guayaquil. Debido a su experiencia, sabe que los sábados suele vender más autos que en otros días, por lo que ha elaborado una distribución de probabilidades en la que se muestra la cantidad de autos que espera vender en un determinado sábado.

Cantidad de automóviles vendidos, x	Probabilidad P(x)
0	0.1
1	0.2
2	0.3
3	0.3
4	0.1
TOTAL	**1**

Tabla 5.5 Ejemplo distribución de probabilidad en venta de automóviles.

Solución

a. ¿De qué tipo de distribución se trata?

Se trata de una distribución de probabilidad discreta.

b. ¿Cuántos automóviles espera vender Juan un sábado normal?

$$\mu = \Sigma[xP(x)]$$

$$= 0(0.1) + 1(0.2) + 2(0.3) + 3(0.3) + 4(0.1)$$

$$= 2.1$$

El promedio de venta de autos de Juan los sábados es de 2.1, lo que significa que, a lo largo de muchos sábados, se espera que venda un promedio de 2.1 autos por día. Aunque es imposible vender exactamente 2.1 autos en un sábado en particular, el valor esperado se utiliza para hacer una predicción sobre la media de la cantidad de autos vendidos a largo plazo. Por ejemplo, si Juan trabajara 25 sábados en un año puede esperar vender 52 automóviles solo durante los sábados. La media es también conocida como valor esperado.

c. ¿Cuál es la varianza de la distribución?

Cantidad de automóviles vendidos, x	Probabilidad P(x)	(x - μ)	(x - μ)2	(x - μ)2 P(x)
0	0.10	0 - 2.1	4.41	0.441
1	0.20	1 - 2.1	1.21	0.242
2	0.30	2 - 2.1	0.01	0.003
3	0.30	3 - 2.1	0.81	0.243
4	0.10	4 - 2.1	3.61	0.361
				σ2 = 1.290

Tabla 5.6. Ejemplo varianza de distribución de probabilidad.

La desviación estándar en este ejemplo sería la raíz cuadrada de 1.290, la cual es 1.136. ¿Cómo se interpreta una desviación estándar de 1.136 automóviles? Si existe otra vendedora, digamos María, que también vendió en promedio 2.1 automóviles los sábados y la desviación estándar de sus ventas fue de 1.82 automóviles, se concluye que existe mayor variabilidad en las ventas sabatinas de María que en las de Juan, ya que 1.82 es mayor que 1.136.

5.7 Distribución de probabilidad binomial

La distribución de probabilidad binomial es de probabilidad discreta y se suele presentar con mucha frecuencia. Para describir los resultados experimentales con una distribución binomial existen cuatro requisitos que deben cumplirse:

- El resultado de cada ensayo del experimento recae en una de dos categorías mutuamente excluyentes: éxito o fracaso.
- La variable aleatoria permite contar el número de éxitos en una cantidad de ensayos determinada.
- La probabilidad de éxito y fracaso es la misma en cada ensayo.
- Los ensayos son considerados como eventos independientes, lo que significa que el resultado de un ensayo no tiene ninguna influencia sobre el resultado de otro ensayo. El resultado de cada ensayo es completamente autónomo e independiente de cualquier resultado previo.

Para calcular la probabilidad de un evento bajo el enfoque binomial debemos utilizar la siguiente expresión:

$$P(x) =_n C_x . \pi^x (1-\pi)^{n-x}$$

En donde:

C: Es el símbolo de combinación; *n* es el número de ensayos; *x* es la variable aleatoria definida como el número de éxitos y; π es la probabilidad de éxito de cada ensayo.

Ejemplo 5.2 La aerolínea *EcuaAir* ofrece 5 vuelos diarios de Guayaquil a Quito. Suponga que la probabilidad de cualquier de estos vuelos llegue tarde es de 0.3.

Solución

a. ¿Cuál es la probabilidad de que ningún vuelo llegue tarde hoy?

La probabilidad de que uno de los vuelos llegue tarde es de 0.3, por lo que π=0.3. Luego, existen cinco vuelos al día por lo que *n=5*. El valor de *x* hace referencia a los casos de éxito y en el ejemplo propuesto un éxito sería que el avión no llegue tarde. Por lo tanto, *x=0*. Aplicando la fórmula de probabilidad binomial tenemos lo siguiente:

$$P(x) =_n C_x . \pi^x (1-\pi)^{n-x}$$

$$P(0) =_5 C_0 . (0.3)^0 (1-0.3)^{5-0}$$

$$= (1)(1)(0.16807) = 0.16807$$

En conclusión, la probabilidad de que ningún vuelo llegue tarde el día de hoy es 0.16807.

b. ¿Cuál es la probabilidad de que exactamente uno de los vuelos llegue tarde hoy?

Para calcular la probabilidad de que exactamente uno de los vuelos llegue tarde se define a *x=1*, entendiendo que éxito en este caso sería el hecho de llegar tarde. Por tanto, la probabilidad seria:

$$P(x) = {}_n C_x . \pi^x (1-\pi)^{n-x}$$

$$P(1) = {}_5 C_1 . (0.3)^1 (1-0.3)^{5-1}$$

$$= (5)(0.3)(0.2401) = 0.36015$$

5.7.1 Media y varianza de una distribución binomial

La media de una distribución binomial se puede calcular a través de la siguiente expresión:

$$\mu = n\pi$$

Mientras que la varianza es calculada mediante la siguiente expresión:

$$\sigma^2 = n\pi(1-\pi)$$

Si aplicamos ambas expresiones para calcular la media y varianza del ejemplo 5.2 de los vuelos *EcuaAir* y recordando que $\pi=0.3$ y $n=5$ obtendríamos lo siguiente:

Media

$$\mu = (5)(0.3)$$

$$\mu = 1.5$$

Varianza

$$\sigma^2 = 5(0.3)(1-0.3)$$

$$\sigma\text{^}2 = 1.05$$

5.8 Distribución de probabilidad binomial acumulada

Existen ocasiones en las que se desea conocer la probabilidad de, por ejemplo, acertar 4 de 10 preguntas en un examen final de estadística o matemática. En estos casos son necesarias las distribuciones de frecuencia acumulada.

La Función de Distribución Acumulativa (FDA) de una variable aleatoria discreta especifica la probabilidad de que una variable aleatoria sea menor o igual a un valor dado. Es decir:

$$F(x) = P(X \leq x)$$

Donde X es la variable aleatoria discreta (VAD); P(x) la función de probabilidad de la VAD; P(X) es la función de distribución acumulativa de la VAD.

Ejemplo 5.3 El departamento de transporte de Sambrondón realizó un estudio donde concluyó que el 69% de las personas que ocupaban los asientos delanteros de los vehículos usaban cinturón de seguridad, lo que implica que los dos pasajeros delanteros lo usaban. Ahora, suponga que se desea comparar esta información con el uso actual del cinturón de seguridad y para ello, selecciona una muestra aleatoria de 15 vehículos.

Solución

Los datos serían: n= 15; π = 0.69; x = 0, 1, 2, … ,15

a. ¿Cuál es la probabilidad que los ocupantes de la parte delantera en exactamente 7 de los 15 vehículos seleccionados utilicen cinturones de seguridad?

$$P(x = 7) =_{15} C_7 (0.69)^7 (1 - 0.69)^{15-7} = 0.0409$$

b. ¿Cuál es la probabilidad que los ocupantes de la parte delantera de por lo menos 7 de los 15 vehículos utilicen cinturón de seguridad?

$$P(x \geq 7) = P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10) + P(x = 11) + P(x = 12) + \cdots + P(x = 15) = 0.0409 + 0.0910 + 0.1575 + 0.2103 + 0.2128 + 0.1579 + \cdots + 0.0038 = 0.9810$$

5.9 Distribución de probabilidad hipergeométrica

Uno de los requisitos para utilizar la distribución de probabilidad binomial es que la probabilidad de éxito en cada ensayo debe permanecer constante y el muestreo debe incluir reemplazos. No obstante, la mayoría de veces, se realiza un muestreo sin reemplazo. Esto quiere decir que, por ejemplo, en el caso de que *n* sea igual a 20, la probabilidad de seleccionar un elemento es 1/20. La probabilidad de seleccionar un segundo elemento es de 1/19, y así sucesivamente.

Por ende, cuando se realiza un muestreo sin reemplazo en una población relativamente pequeña, la probabilidad de éxito no se mantiene igual en todos los eventos. Por tanto, no es apropiado utilizar la distribución binomial en estos casos. En su lugar, se utiliza la distribución hipergeométrica para calcular la probabilidad de éxito. La distribución hipergeométrica toma en cuenta que la probabilidad de que ocurra un éxito en un ensayo disminuirá a medida que se realicen más ensayos, ya que la población de la que se extrae la muestra será cada vez más pequeña.

La probabilidad de ocurrencia de un suceso bajo una distribución de probabilidad hipergeométrica se expresa de la siguiente manera:

$$P(x) = \frac{({}_{s}C_{x})({}_{N-S}C_{n-x})}{({}_{N}C_{n})}$$

Donde:

N: representa el tamaño de la población;

S: representa la cantidad de éxitos en la población;

X: indica el número de éxitos en la muestra; puede asumir los valores de 0,1,2,3…X

N: se refiere al tamaño de la muestra o al número de ensayos;

C: es la abreviación de combinación.

Una distribución de probabilidad hipergeométrica posee las siguientes características:

- Los resultados de cada ensayo de un experimento se clasifican en dos categorías: éxito o fracaso.
- La cantidad de éxitos en un número determinado de ensayos es la variable aleatoria.
- Los ensayos no son independientes.
- Realiza muestreos sin reemplazo en una población finita, lo que cambia la probabilidad de éxito en cada ensayo.

Ejemplo 5.4 PYCCA tiene 25 empleados que conforman el departamento de cobranzas. Solo 15 de ellos pertenecen al sindicato. Si se escogen aleatoriamente 5 empleados para formar una comitiva que hablará con la empresa sobre los horarios en días festivos.

Solución

¿Cuál es la probabilidad de que 4 de los 5 empleados elegidos para formar parte del comité pertenezcan al sindicato?

N=25 (número de empleados); S=15 (número de empleados sindicalizados); x=4 (número de empleados sindicalizados escogidos); n=5 (número de empleados escogidos)

$$P(4) = \frac{({}_{15}C_4)({}_{25-15}C_{5-4})}{({}_{25}C_5)} = \frac{(1{,}365)(10)}{53{,}130} = 0.2569$$

La probabilidad de que 4 de los 5 empleados elegidos para formar parte del comité pertenezcan al sindicato es de 25.69%.

5.10 Distribución de probabilidad de Poisson

La distribución de probabilidad de Poisson es una distribución de probabilidad discreta que describe el número de veces que ocurre un evento en un intervalo de tiempo o espacio específico, bajo la suposición de que los eventos ocurren de manera independiente y a una tasa constante promedio. La distribución de Poisson se utiliza comúnmente para modelar la ocurrencia de eventos raros o inusuales, como accidentes, errores o defectos en la producción, en un intervalo específico de tiempo o espacio. Por lo tanto, esta distribución es útil en entornos empresariales y de fabricación donde se desea calcular la probabilidad de que un cierto número de eventos raros ocurran en un período determinado. La distribución de probabilidad de Poisson cuenta con tres características principales:

- La variable se refiere a la frecuencia de ocurrencia de un evento dentro de un intervalo establecido.
- La probabilidad de que dicho evento ocurra es directamente proporcional al tamaño del intervalo.
- Los intervalos son mutuamente excluyentes e independientes.

La distribución de probabilidad de Poisson se expresa de la siguiente manera:

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

Donde:

µ (mu): representa la media de la frecuencia de ocurrencia de un evento dentro de un intervalo determinado

e: se refiere a la constante matemática 2.71828 (base de los logaritmos naturales);

x: representa la cantidad de veces que se presenta el evento;

P(x): corresponde a la probabilidad de que ocurra un valor específico de x.

Para calcular la media de la cantidad de éxitos, μ, se puede utilizar la fórmula nπ, en la que n representa el número total de ensayos y π indica la probabilidad de éxito.

Ejemplo 5.5 Suponga que en *Equair* es poco común que se pierda el equipaje. La mayoría de los vuelos no tienen maletas perdidas, algunos tienen una, pocos tienen dos, y es muy raro que haya tres o más maletas perdidas en un solo vuelo. Se toma una muestra aleatoria de 1,000 vuelos y se encuentra que en total se han perdido 200 maletas. Si la cantidad de maletas perdidas por vuelo sigue una distribución de Poisson. ¿Cuál es la probabilidad de que no se pierda ninguna maleta?

Solución

Primero, se calcula la media aritmética del número de maletas perdidas por vuelo al dividir 200 entre 1,000, lo que nos da como resultado 0.2. Luego se calcula la probabilidad.

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

$$P(0) = \frac{(0.2)^0(e^{-0.2})}{0!} = 0.819$$

En otras palabras, en el 81.9 % de los vuelos no existirán maletas perdidas. Para conocer cuál es la probabilidad de que en los vuelos se pierda exactamente una maleta se realiza la siguiente operación:

$$P(1) = \frac{(0.2)^1(e^{-0.2})}{1!} = 0.164$$

Se espera que se pierda exactamente una maleta en 16.4% de los vuelos

Ejercicios propuestos del capítulo

Ejercicio 5.1 Laura es la administradora de un restaurante de comida típica. Generalmente, el restaurante recibe más comensales los fines de semana, así que Laura desarrolló una distribución de probabilidad en la que se muestra la cantidad de comensales que espera recibir en el restaurante un fin de semana determinado. La distribución de probabilidad es la siguiente:

Cantidad de comensales, x	Probabilidad P(x)
8	0.10
9	0.20
10	0.30
11	0.30
12	0.10
TOTAL	**1.00**

a. ¿De qué tipo de distribución se trata?

b. ¿Cuántos comensales espera Laura recibir un fin de semana normal?

c. ¿Cuál es la varianza de la distribución?

Ejercicio 5.2 En Ecuador se recogió información con respecto a las llamadas diarias que recibió la línea telefónica para emergencias «ECU-911» en los últimos 60 días. Así, por ejemplo, existieron 20 días en los que se realizaron 2 llamadas, 13 días en los que se realizaron 3 llamadas, etc. La información recolectada fue la siguiente:

Cantidad de llamadas	Frecuencia
0	10
1	15
2	20
3	13
4	2
TOTAL	**60**

a. Convierta la información del número de llamadas en una distribución de probabilidad.

b. ¿Es un ejemplo de distribución de probabilidad discreta o continua?

c. ¿Cuál es la media de la cantidad de llamadas de emergencia al día?

d. ¿Cuál es la desviación estándar de la cantidad de llamadas diarias?

Ejercicio 5.3 Calcule los valores de la función de probabilidad para una variable aleatoria que tiene una distribución binomial. En donde:

X: veces que sale 2 al lanzar un dado 10 veces**.**

Calcule también su media y varianza.

Ejercicio 5.4 La aseguradora *Seguros Guayas* vende ocho seguros de vehículos al mes. Suponga que la probabilidad de que cualquiera de las personas que adquirieron dicho seguro sufra un siniestro de tránsito durante el siguiente mes es de 0.35.

a. ¿Cuál es la probabilidad de que ninguna persona sufra un siniestro de tránsito el mes siguiente al adquirir el seguro?

b. ¿Cuál es la probabilidad de que exactamente dos personas estén involucradas en un siniestro de tránsito al mes posterior de adquirir el seguro?

Ejercicio 5.5 El profesor de estadística de UEES determinó que, en promedio, el 82% de los estudiantes que lograban aprobar la materia asistían a todas las clases. Suponga que usted desea conocer si esta información es correcta y selecciona una muestra de 20 estudiantes.

a. ¿Cuál es la probabilidad que exactamente 12 de los 20 estudiantes seleccionados que aprobaron la materia hayan asistido a todas las clases?

b. ¿Cuál es la probabilidad que por lo menos 15 de los 20 estudiantes seleccionados que aprobaron la materia hayan asistido a todas las clases?

Ejercicio 5.6 Una fábrica de peluches tuvo una producción de 100 peluches durante el último mes. De estos 100 peluches, 24 se encuentran defectuosos. En una muestra de 15 peluches:

a. ¿Cuál es la probabilidad de que exactamente 2 sean defectuosos? Suponga que las muestras se toman sin reemplazo.

Ejercicio 5.7 En una distribución de Poisson el valor de μ=0.25.

a. ¿Cuál es la probabilidad de que $x = 0$?

b. ¿Cuál es la probabilidad de que $x = 3$?

Ejercicio 5.8 *Responda Verdadero (V) o Falso (F), según corresponda:*

a. Las distribuciones de probabilidad se clasifican en continuas y discontinuas. ( )

b. La distribución de probabilidad de Poisson es una distribución de probabilidad continua. ( )

c. En la distribución de probabilidad binomial la probabilidad de éxito y fracaso son iguales en cada ensayo. ( )

d. Los ensayos en la distribución de probabilidad hipergeométrica son independientes. ( )

Ejercicio 5.9 Define el concepto de variables aleatorias y de un ejemplo.

Ejercicio 5.10 *¿Cuáles son las principales características de la distribución de probabilidad de Poisson?*

Capítulo VI

DISTRIBUCIONES CONTINUAS DE PROBABILIDAD

VI. DISTRIBUCIONES CONTINUAS DE PROBABILIDAD

6.1 Introducción

En el capítulo 5 se estudió a profundidad las distribuciones de probabilidad discretas. A lo largo de este capítulo nos concentraremos en las distribuciones de probabilidad continuas, a saber: uniforme, normal y, exponencial.

6.2 Distribución de probabilidad uniforme

La distribución de probabilidad uniforme, también conocida como distribución rectangular. es utilizada para modelar un rango de valores con probabilidad constante. Esta tiene forma rectangular y está definida por valores máximos mínimos.

Por ejemplo, el tiempo que le toma a un estudiante de economía en resolver un examen de econometría varía de 60 a 80 minutos. La variable aleatoria dentro de este ejemplo sería el tiempo que le toma al estudiante resolver el examen dentro de ese intervalo.

La distribución de probabilidad uniforme toma la siguiente forma:

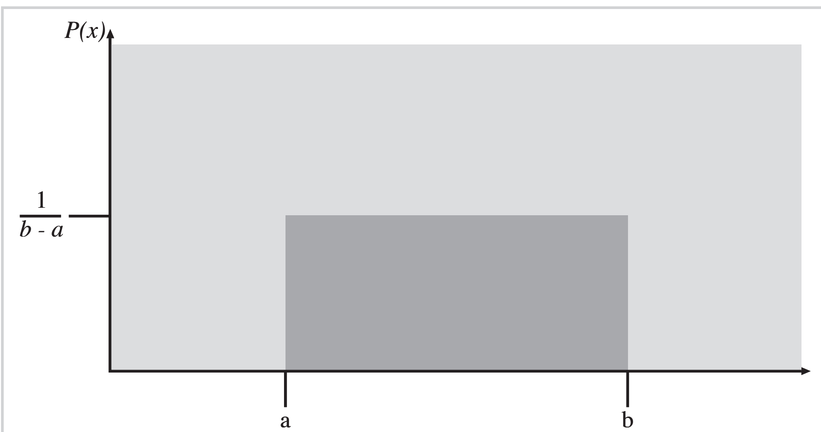


Gráfico 6.1 Distribución de probabilidad uniforme.

$$Prob\ (X_1 < X < X_2) = \frac{1}{b-a} * (X_2 - X_1)$$

Donde Prob (X_1 < X < X_2) es la probabilidad de que ocurra un evento dentro del intervalo (X_1,X_2). Por otro lado, *a* y *b* representan constantes y son los extremos de la distribución.

Por otro lado, para calcular la media de la distribución uniforme aplicamos la siguiente fórmula:

$$\mu = \frac{a+b}{2}$$

Y la desviación estándar se calcula como sigue:

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Ejemplo 6.1

La UEES ofrece servicio de transporte en autobús a sus empleados. Durante los días de la semana, el autobús llega a la parada cada 40 minutos desde las 16:00 hasta las 20:00. Los empleados llegan a la parada en tiempos aleatorios y el tiempo que tienen que esperar para tomar el autobús tiene una distribución uniforme que va desde 0 hasta 40 minutos.

Solución

a. ¿Cuál es la probabilidad de que un trabajador espere más de 30 minutos? En este caso, a = 0 y b = 40, ya que estos son el mínimo y máximo tiempo que espera un empleado. Por otro lado, X_1 = 30 y X_2 = 40. Por tanto:

$$Prob(30 < x < 40) = \frac{1}{40-0} * (40 - 30) = 0.25$$

b. ¿Cuál es la probabilidad de que un empleado espere entre 15 y 20 minutos?

$$X_1 = 15 \qquad\qquad X_2 = 20$$

$$Prob(15 < x < 20) = \frac{1}{40-0} * (20 - 15) = 0.125$$

c. ¿Cuál es el promedio de espera de un autobús? ¿Cuál es la desviación estándar de los tiempos de espera?

$$\mu = \frac{a+b}{2} = \frac{0+40}{2} = 20\ min$$

El promedio de espera de un autobús es de 20 minutos.

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(40-0)^2}{12}} = 11.55\ min$$

La desviación estándar de los tiempos de espera es de aproximadamente más/menos 12 minutos.

6.3 Distribución de probabilidad normal

La distribución de probabilidad normal presenta las siguientes características:

- Su forma es similar a una campana y tiene un punto máximo en la mitad de la distribución.
- Las medidas de tendencia central (media, mediana y moda) son iguales y se encuentran en el centro de la distribución.
- El área total debajo de la curva es igual a 1.00. Esta área se divide en dos mitades, una a la derecha de la curva y la otra mitad a la izquierda.
- Es simétrica respecto a la media. Las dos mitades bajo la curva son imágenes especulares.
- La distribución es asintótica, es decir, los extremos de la curva se extienden infinitamente en ambas direcciones.
- La ubicación de una distribución normal se determina a través de la media y la dispersión de dicha distribución se conoce a través de la desviación estándar.

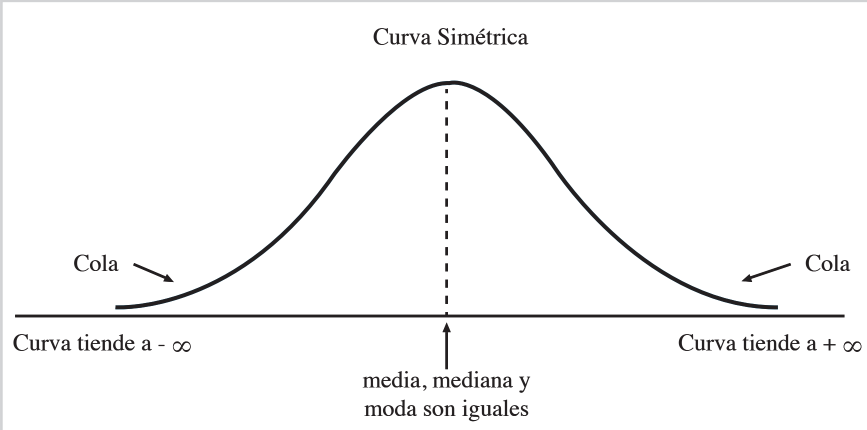


Gráfico 6.2 Distribución de probabilidad normal.

La expresión para determinar la probabilidad de un evento bajo una distribución de probabilidad normal es la siguiente:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-[\frac{(x-\mu)^2}{2\sigma^2}]}$$

6.3.1 Distribución de probabilidad normal estándar

Existen un número ilimitado de distribuciones normales, todas con distinta media, desviación estándar o ambas[4]. A pesar de que es posible realizar tablas de probabilidad de distribuciones discretas tales como la de Poisson o la Binomial, es poco práctico realizar tablas de probabilidad para la infinidad de distribuciones normales. Por esta razón, un solo miembro de la familia se aplica para calcular las probabilidades de todas las distribuciones de probabilidad normal; a esta distribución se le llama *distribución de probabilidad normal estándar.*

La *distribución de probabilidad normal estándar* se distingue de las demás distribuciones normales por tener una media de 0 y desviación estándar de 1. Cualquier distribución normal puede llegar a ser una distribución normal estándar, para esto se debe restar la media de cada observación y esta diferencia

[4]Una distribución normal siempre será simétrica (no tendrá sesgo), no obstante, podría tener diferentes curtosis.

dividirla para la desviación estándar. Los resultados obtenidos se denominan *valores z* o valores tipificados.

El valor de z indica la ubicación de una observación x en relación a la media y en términos de unidades de desviación estándar. Para su cálculo se emplea la siguiente fórmula:

$$z = \frac{x - \mu}{\sigma}$$

En donde x representa el valor de cualquier observación; μ es la media de la distribución y; σ es la desviación estándar de la distribución.

Ejemplo 6.2 Los salarios mensuales de los profesores de la Facultad de Emprendimiento, Negocios y Economía de UEES siguen una distribución de probabilidad normal con una media de \$2,000 y una desviación estándar de \$250.

a. ¿Cuál es el valor z del salario x de un profesor que tiene un salario de \$2,500 mensuales?

Solución

X= \$2.500

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{2{,}500 - 2{,}000}{250}$$

$$z = 2$$

Este valor z indica que un salario mensual de \$2,500 se encuentra a dos desviaciones estándar por encima de la media.

b. ¿Cuál es el valor z de un profesor que gana \$1,400 mensuales?

X= \$1,400

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{1{,}400 - 2{,}000}{250}$$

$$z = -2.4$$

Un valor z negativo de 2.4 indica que un salario mensual de \$1,400 se encuentra a más de dos desviaciones estándar por debajo de la media.

6.3.2 Determinación de áreas bajo la curva normal

El área total bajo la curva de la distribución de probabilidad es igual a 1, como se mencionó anteriormente. No obstante, es posible calcular áreas específicas bajo la curva para eventos o valores particulares.

Ejemplo 6.3 Como se mencionó en el ejemplo 6.2, el salario mensual de los profesores de la Facultad de Emprendimiento, Negocios y Economía sigue una distribución normal con una media de \$2,000 y una desviación estándar de \$250. ¿Cuál es la probabilidad de seleccionar un profesor de dicha facultad con un salario mensual que se encuentre entre los \$2,000 y \$2,500?

Solución

Por el ejemplo anterior sabemos que a \$2,000 le corresponde un valor z de 2.00.

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{2{,}500 - 2{,}000}{250}$$

$$z = 2.00$$

Luego, a partir de este valor, buscamos en la tabla de probabilidades asociadas al valor z de 2.00 (Apéndice A). El valor correspondiente es de 0.4772.

z	0.00	0.01	0.02
0	0	0.004	0.008
0.1	0.0398	0.0438	0.0478
0.2	0.0793	0.0832	0.0871
0.3	0.1179	0.1217	0.1255
0.4	0.1554	0.1591	0.1628
0.5	0.1915	0.195	0.1985
0.6	0.2257	0.2291	0.2324
0.7	0.258	0.2611	0.2642
0.8	0.2881	0.291	0.2939
0.9	0.3159	0.3186	0.3212
1.0	0.3413	0.3438	0.3461
1.1	0.3643	0.3665	0.3686
1.2	0.3849	0.3869	0.3888
1.3	0.4032	0.4049	0.4066
1.4	0.4192	0.4207	0.4222
1.5	0.4332	0.4345	0.4357
1.6	0.4452	0.4463	0.4474
1.7	0.4554	0.4564	0.4573
1.8	0.4641	0.4649	0.4656
1.9	0.4713	0.4719	0.4726
2	**0.4772**	0.4778	0.4783
2.1	0.4821	0.4826	0.483
2.2	0.4861	0.4864	0.4868
2.3	0.4893	0.4896	0.4898
2.4	0.4918	0.492	0.4922

Tabla 6.1 Ejemplo. Áreas bajo la curva normal estandarizada.

Este valor indica que el área bajo la curva normal entre $2,000 y $2,500 es de 0.4772. Este valor también se puede interpretar de la siguiente forma: el 47.72% de los profesores de Facultad de Emprendimiento, Negocios y Economía de UEES tienen un salario mensual entre $2,000 y $2,500 o la probabilidad de seleccionar a un profesor con un salario entre los $2,000 y $2,500 es de 0.4772.

Es importante recordar que, la mitad del área se encuentra sobre la media y la otra mitad por debajo de esta, por lo que la probabilidad de seleccionar a un profesor que gane menos de $2,000 mensuales es de 0.5000. Y si, sumamos esta probabilidad a la calculada en el ejemplo, obtenemos que 0.500+0.4772 = 0.9772. Esto quiere decir que, alrededor del 97.72% de profesores de Emprendimiento, Negocios y Economía de UEES, tienen un salario mensual menor a $2,500.

Ejemplo 6.4 La distribución de los salarios mensuales de un profesor de Derecho de la UEES sigue una distribución normal, con una media de $2,000 y una desviación estándar de $250. Determine la probabilidad de seleccionar a un profesor cuyo salario:

a. Oscile entre $2,100 y $2,500.

b. Sea menor a $2,100.

Solución

Primero, se debe calcular el valor *z* correspondiente a los valores de $2,100 y $2,500.

$$z_1 = \frac{x - \mu}{\sigma}$$

$$z_1 = \frac{2{,}100 - 2{,}000}{250}$$

$$z_1 = 0.4$$

$$z_2 = \frac{x - \mu}{\sigma}$$

$$z_2 = \frac{2{,}500 - 2{,}000}{250}$$

$$z_2 = 2$$

De acuerdo con el Apéndice A, las probabilidades asociadas a estos valores son 0.1554 y 0.4772, respectivamente. Para determinar la probabilidad que un profesor gane entre $2,100 y $2,500, simplemente restamos ambas probabilidades, esto es, P ($2,100 < salario mensual < $,2500) = 0.4772-0.1554= 0.3218.

Por otro lado, para determinar la probabilidad que un profesor gane menos de $2,100, simplemente sumamos 0.5000+0.1554= 0.6564. Esto es así porque sabemos que la otra mitad de la distribución tiene una probabilidad de 0.5000.

z	0.00	0.01	0.02	0.03
0	0	0.004	0.008	0.012
0.1	0.0398	0.0438	0.0478	0.0517
0.2	0.0793	0.0832	0.0871	0.091
0.3	0.1179	0.1217	0.1255	0.1293
0.4	0.1554	0.1591	0.1628	0.1664
0.5	0.1915	0.195	0.1985	0.2019
0.6	0.2257	0.2291	0.2324	0.2357
0.7	0.258	0.2611	0.2642	0.2673
0.8	0.2881	0.291	0.2939	0.2967
0.9	0.3159	0.3186	0.3212	0.3238
1.0	0.3413	0.3438	0.3461	0.3485
1.1	0.3643	0.3665	0.3686	0.3708
1.2	0.3849	0.3869	0.3888	0.3907
1.3	0.4032	0.4049	0.4066	0.4082
1.4	0.4192	0.4207	0.4222	0.4236
1.5	0.4332	0.4345	0.4357	0.437
1.6	0.4452	0.4463	0.4474	0.4484
1.7	0.4554	0.4564	0.4573	0.4582
1.8	0.4641	0.4649	0.4656	0.4664
1.9	0.4713	0.4719	0.4726	0.4732
2.0	0.4772	0.4778	0.4783	

Tabla 6.2. Ejemplo Áreas bajo la curva normal estandarizada.

Ejemplo 6.5 Suponga ahora que los profesores de Arquitectura de UEES ganan en promedio $2,500 mensuales con una desviación estándar de $250. **¿Cuál es** la probabilidad de que un profesor seleccionado al azar gane entre $2,100 y $3,000?

Solución

Como el caso anterior, primero debemos calcular los valores *z* asociados a 2,100 y 3,000.

$$z_1 = \frac{x - \mu}{\sigma}$$

$$z_1 = \frac{2{,}100 - 2{,}500}{250}$$

$$z_1 = -1.60$$

$$z_2 = \frac{x - \mu}{\sigma}$$

$$z_2 = \frac{3{,}000 - 2{,}500}{250}$$

$$z_2 = 2.00$$

El área bajo la curva de un valor *z* de -1.60 es 0.4452 (Apéndice A). El área bajo la curva de un valor *z* de 2.00 es 0.4772. Si se suma ambas áreas se obtiene: 0.4452+0.4772=0.9224. Por lo tanto, la probabilidad de que un profesor tenga un salario mensual entre $2,100 y $3,000 es de 0.9224. O, de igual manera, se puede establecer que el 92.24% de los profesores de Arquitectura de UEES poseen un ingreso mensual que oscila entre los $2,100 y $3,000.

6.4 Distribución de probabilidad exponencial

Este tipo de distribución de probabilidad continua suele describir los tiempos entre eventos que ocurren en secuencia.

Las acciones suceden independientemente a un ritmo constante por unidad o duración de tiempo. Como el tiempo nunca es un valor negativo, una variable aleatoria exponencial siempre será positiva. La distribución exponencial describe eventos tales como:

- Los tiempos de servicio al cliente en el módulo de información del municipio.
- Los tiempos de atención a un cliente en un cajero del supermercado.

La distribución de probabilidad exponencial presenta una inclinación hacia valores positivos, lo que la distingue de las distribuciones uniforme y normal que son simétricas. La distribución exponencial se caracteriza por un solo parámetro denotado con el símbolo λ(lambda).

Otra característica importante de la distribución exponencial es su relación con la distribución Poisson, la cual se describe en el capítulo 5. Ambas distribuciones tienen un solo parámetro (μ y λ) y se tratan de distribuciones con sesgo positivo.

Para comprender mejor la relación entre ambas distribuciones suponga que el ritmo al que llegan los clientes a un restaurante es de seis clientes por hora. La distribución Poisson se utiliza para determinar la probabilidad de que, en cualquier hora de la cena lleguen 2 clientes o 3 clientes o 4 clientes, y así sucesivamente.

Imaginemos que en lugar de analizar la cantidad de clientes que llegan en una hora, queremos estudiar el lapso de tiempo que transcurre entre cada llegada. En este caso, se utiliza la distribución de probabilidad Exponencial. La fórmula para calcular esta distribución es la siguiente:

$$P(x) = \lambda e^{-\lambda x}$$

Donde x es la unidad de tiempo, λ es el lapso de tiempo que transcurre entre cada evento y e se refiere a la constante matemática 2.71828 (base de los logaritmos naturales).

Ejemplo 6.6 Los pedidos en una cafetería llegan siguiendo una distribución de probabilidad exponencial, con una media de 1 cada 30 segundos. Encuentra la probabilidad de que el siguiente pedido llegue en menos de 10 segundos, o en más de 50 segundos.

Solución

Para esto, primero se determina el parámetro de ritmo λ, que en este caso es 1/30. Para encontrar la probabilidad, se inserta 1/30 en lugar de λ y 10 por x.

$$P(\textit{Tiempo del siguiente pedido} < 10) = 1 - e^{-\frac{1}{30}(10)} = 0.2835$$

Mientras que para calcular la probabilidad de que el siguiente pedido llegue en más de 50 segundos el proceso es el siguiente:

$$P(\textit{Tiempo del siguiente pedido} < 50) = 1 - e^{-\frac{1}{30}(50)} = 0.8111$$

$$P(\textit{Tiempo del siguiente pedido} > 50) = 1 - 0.8111 = 0.1889$$

Ejercicios propuestos del capítulo

Ejercicio 6.1 *¿Cuáles son las principales características de la distribución normal?*

Ejercicio 6.2 Responda Verdadero (V) o Falso (F), según corresponda:

a. El área total bajo la curva de la distribución normal es de 1. ( )

b. Toda distribución normal puede convertirse en una distribución normal estándar. ( )

c. La distribución normal estándar tiene una media de 2.5 y una desviación estándar de 0. ( )

d. La distribución de probabilidad exponencial tiene sesgo positivo. ( )

e. La distribución de probabilidad normal estándar es asimétrica. ( )

Ejercicio 6.3 Mencione algunos ejemplos en los que se aplicaría la distribución de probabilidad exponencial.

Ejercicio 6.4 *¿Qué es el valor tipificado z?*

Ejercicio 6.5 Se recolectó información en UEES sobre el gasto promedio diario de los estudiantes, el cual sigue una distribución normal. Los datos mostraron una media de $10 y una desviación estándar de $2. Con base a esta información, responda:

a. ¿Qué porcentaje de los estudiantes gastan menos de $7 al día?

b. ¿Cuál es la probabilidad de que un estudiante seleccionado al azar gaste más de $11 al día?

c. ¿Qué porcentaje de los estudiantes gastan entre $8 a $12 al día?

Ejercicio 6.6 Al aeropuerto de Quito llegan vuelos internacionales, de acuerdo a una distribución de probabilidad exponencial con una media de 1 cada 30 minutos.

a. Encuentra la probabilidad de que el siguiente vuelo internacional llegue en menos de 20 minutos o en más de 45.

Ejercicio 6.7 Los tiempos de espera para recibir la orden después de hacer el pedido en *Caramel Coffee* siguen una distribución exponencial con una media de 2 minutos. Calcule la probabilidad de que un cliente espere:

a. Menos de 1 minuto

b. Más de 4 minutos

c. Entre 1 minuto y 3 minutos

Referencias

Anderson, David R., Dennis J. Sweney y Thomas A. Williams. (2008). Estadística para administración y economía. México: Cengage Learning Editores.

Báez, J. (2019). Desigualdad en Ecuador: La participación en el ingreso del decil más rico aumentó así como el Índice de Palma. Instituto de investigaciones económicas universidad central del Ecuador. Recuperado de https://coyunturauceiie.wordpress.com/2019/09/12/desigualdad-en-ecuador-la-participacion-en-el-ingreso-del-decil-mas-rico-aumento-asi-como-el-indice-de-palma/

Cárdenas, J. (2015). *Tablas de contingencia: Cómo analizarlas fácilmente*. Networkianos: Blog de Sociología. Recuperado de https://networkianos.com/tablas-de-contingencia/

Douglas, A., Marchal, W., & A. Wathen, S. (2012). *ESTADISTICA APLICADA A LOS NEGOCIOS Y LA ECONOMÍA* (Decimoquinta edición ed.). New York, NY: McGraw-Hill.

Gaskin, J. (2021). *Todo lo que necesitas saber sobre las gráficas de pastel*. Venngage. Recuperado de https://es.venngage.com/blog/grafica-de-pastel/

Lifeder. (15 de diciembre de 2022). Curtosis: definición, tipos, fórmulas, para qué sirve, ejemplo. Recuperado de: https://www.lifeder.com/curtosis/.

Lind, D.A., Marchal, W.G. & Wathen, S.A. (2012) *Estadística Aplicada a Los Negocios y La Economía*. 15va edición. México: McGraw-Hill Interamericana.

Malhotra, N. (2008). Investigación de Mercados (Quinta edición ed.). Ecuador: Pearson Prentice Hall.

Mira, Gomez, J., Aranaz, J., & Perez, J. (1997). Auditoría de Historias Clínicas: ¿cuál es el tamaño adecuado de la muestra? España: Universidad de Alicante.

Parada, I. (2021). *La regla del complemento en estadística*. YuBrain. Recuperado de https://www.yubrain.com/matematicas/estadistica/regla-complemento-estadistica-ejemplo-probabilidad/

Salazar, C., & Castillo, S. D. (2018). Fundamentos básicos de estadística. Recuperado de http://up-rid2.up.ac.pa:8080/xmlui/handle/123456789/1570

Serra, B. (2014). *Diagrama de tallo y hojas*. Universo Formulas. Recuperado de https://www.universoformulas.com/estadistica/descriptiva/diagrama-tallo-hojas/

Apéndice

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499
3.1	0.499	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Tabla A. Áreas bajo la curva normal estandarizada.

Manuel A. Zambrano-Monserrate

manuelzambranom@uees.edu.ec

Economista de profesión. Cuenta con una maestría y doctorado en Economía. Ha sido profesor universitario de pregrado y posgrado por más de 14 años. Es autor de diversas publicaciones científicas en el área de Economía. Como consultor y capacitador ha colaborado para el Ministerio del Ambiente y el Banco del Desarrollo del Ecuador.

Alexia Berrús-Zhumi

aberrus@uees.edu.ec

Estudiante de Economía por la Universidad Espíritu Santo, comprometida con su formación e interesada por el análisis y comprensión de fenómenos económicos y sociales. A pesar de los desafíos que ha enfrentado, Alexia está enfocada en terminar su carrera con éxito y convertirse en una profesional destacada en el campo de la investigación económica.

Giuliana Goncalves Guillén

ggoncalvez@uees.edu.ec

Estudiante cursando su último año de la carrera de Economía de la Universidad Espíritu Santo. Giuliana se encuentra interesada por la microeconomía y la economía aplicada a los negocios. Actualmente, está centrada en su desarrollo académico y profesional, teniendo como objetivo principal culminar su carrera y seguir un posgrado en el área de negocios.

UEES Centro de Investigaciones

ISBN: 978-9978-25-238-3

uees_ec

universidadespiritusanto

www.uees.edu.ec

Km. 2,5 La Puntilla,
Samborondón

ceninv@uees.edu.ec

Teléfono: (593-4) 500 0950 Ext: 1319 - 1317