inBlue editorial
...nosotros lo editamos!

2023

SEGUNDO FABIAN SIZA M
MIGUEL ÁNGEL SÁEZ P
MICHAEL ESTEFANIA JÁTIVA B

PENSAMIENTO ESTADÍSTICO

EN LA TOMA DE DECISIONES

Sello Editorial

Publicado en:
https://inblueditorial.com
Teléfonos: 062015939 /
0986391700 / 0967646017
Mail: inblueedit@gmail.com
Esmeraldas - Ecuador

Título del libro:
Pensamiento estadístico en la toma de decisiones

Libro Digital
Primera Edición, Abril/ 2023

Editores
PhD. Ermel Viacheslav Tapia Sosa
PhD. Nayade Caridad Reyes Palau

Revisión de pares evaluadores:
Dr. C. Isabel Alonso Berenguer
Dr. C. Alexander Gorina Sánchez

Diseño y Maquetación
Lenin Wladimir Tapia Ortiz

Ilustraciones y fotografías
Archivo del autor y sitios web debidamente referidos

ISBN: 978-9942-44-273-4

DOI: 10.56168/ibl.ed.167894

Autores

© **Segundo Fabián Siza Moposita**
Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador
Facultad de Ciencias Pecuarias, Orellana, Ecuador
fabian.siza@espoch.edu.ec
https://orcid.org/0000-0001-8036-6974

© **Miguel Ángel Sáez Paguay**
Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador
Facultad de Recursos Naturales, Orellana, Ecuador
miguel.saez@espoch.edu.ec
https://orcid.org/0000-0003-3192-5084

© **Michael Estefania Játiva Brito**
Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador
Facultad de Informática y Electrónica, Orellana, Ecuador
estefania.jativa@espoch.edu.ec
https://orcid.org/0000-0002-6394-2586

Esta obra cumple con el requisito de evaluación por dos pares de expertos, bajo el sello editorial inBlue Editorial (ISBN y Doi).

PENSAMIENTO ESTADÍSTICO EN LA TOMA DECISIONES

2023

AUTORES

© Segundo Fabián Siza Moposita

Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador
Facultad de Ciencias Pecuarias, Orellana, Ecuador
fabian.siza@espoch.edu.ec
https://orcid.org/0000-0001-8036-6974

© Miguel Ángel Sáez Paguay

Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador
Facultad de Recursos Naturales, Orellana, Ecuador
miguel.saez@espoch.edu.ec
https://orcid.org/0000-0003-3192-5084

© Michael Estefania Játiva Brito

Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador
Facultad de Informática y Electrónica, Orellana, Ecuador
estefania.jativa@espoch.edu.ec
https://orcid.org/0000-0002-6394-2586

Como citar el libro: Siza Moposita, S., Sáez Paguay, M. y Játiva Brito, M. (2023). *Cultura, Razonamiento y Pensamiento Estadístico*. Primera edición. inBlue Editorial. ISBN: 978-9942-44-273-4 Doi: 10.56168/ibl.ed.167894

ÍNDICE

INTRODUCCIÓN 1
CAPÍTULO I. INTRODUCCIÓN A LA ESTADÍSTICA 4
Introducción 4
La estadística: conceptos y terminología 5
Conceptos y Terminología 7
Organización de la información 17
Tabla de frecuencia simple y agrupada 17
Regla de Sturges 22
Medidas de centralización 25
Medidas de posición 30
Medidas de dispersión 39
Medidas de forma 47
Puntuaciones diferenciales y puntuaciones típicas 50
Las gráficas estadísticas 52
CAPÍTULO II. PROBABILIDADES 62
Introducción 62
Conceptos básicos y operaciones entre sucesos 62
Conceptos básicos 63
Operaciones o relaciones entre eventos 65
Definición de probabilidad 68
Probabilidad clásica 70
Regla de la Suma 73
Regla de la probabilidad condicional 76
Regla de la multiplicación 78
Propiedades de la definición clásica de probabilidad 82
Muestreo con remplazo 83
Muestreo sin remplazo 85
Esquemas de distribución 87
Ejercicios propuestos sobre probabilidad clásica 91
Probabilidad frecuentista o estadística 96

Capítulo III. ELEMENTOS DE ESTADÍSTICA INFERENCIAL 100
Variables aleatorias 100
Función de probabilidad. Propiedades 102
Función de distribución acumulada. Propiedades 106
Características numéricas de las distribuciones 109
Distribuciones continuas clásicas 113
Estimación estadística 123
Estimación por intervalos 126
Técnicas de muestreo probabilístico 143
Conceptos básicos del muestreo estadístico 145
Muestreo aleatorio simple (M.A.S.) 148
Muestreo aleatorio estratificado 151
Muestreo sistemático 158
Muestreo por conglomerados 167
Prueba de hipótesis 174
Distribuciones bidimensionales 187
Tablas de doble entrada 187
Dependencia entre dos variables 188
Correlación lineal 189
Interpretación del coeficiente de correlación lineal r 196
Modelo de regresión simple 196
BIBLIOGRAFÍA 208

INTRODUCCIÓN

Aunque la estadística se enseña hoy día en todos los niveles educativos, al ser una herramienta fundamental en la vida personal y profesional, son muchos los estudiantes que finalizan los cursos de estadística sin comprender correctamente o ser capaces de aplicar los conceptos y procedimientos estadísticos, como se muestra en la amplia investigación sobre el tema (Batanero, Díaz, Contreras & Roa, 2013).

Una de las causas de esta situación se identifica con los problemas didácticos que todavía persisten en el proceso de enseñanza aprendizaje de la Estadística, que no favorecen que los estudiantes comprendan y apliquen los principales conceptos estadísticos (Batanero, 2001).

En tal dirección, los currículos actuales se sustentan en tres categorías analíticas básicas: la cultura, el razonamiento o el pensamiento estadístico, los cuales pueden comprenderse como metas deseables de la formación estadística en la actualidad.

- La *cultura estadística* implica comprender y utilizar el idioma y los instrumentos básicos de la Estadística, es decir, conocer lo que significan los términos estadísticos, utilizar apropiadamente los símbolos estadísticos, conocer e interpretar las representaciones de datos (Improving Statistical Thinking, 2007).
- El *razonamiento estadístico* es la forma en que las personas argumentan sobre las ideas estadísticas y el sentido que le dan a la información estadística. El razonamiento estadístico implica conectar un concepto a otro o combinar ideas acerca de los datos y la probabilidad. Significa entender y estar en capacidad de explicar los procesos estadísticos y de interpretar completamente los resultados estadísticos (Improving Statistical Thinking, 2007).

- El *pensamiento estadístico* implica la comprensión del por qué y de cómo se realizan las investigaciones estadísticas. Esto incluye reconocer y comprender el proceso investigativo completo (desde la pregunta de investigación a la recolección de datos, así como la selección de la técnica para analizarlos, probar las suposiciones, etc.), entendiendo cómo se utilizan los modelos para simular los fenómenos aleatorios, cómo los datos se producen para estimar las probabilidades, reconocimiento de cómo, cuándo, y por qué los instrumentos deductivos existentes se pueden utilizar, y son capaz de entender y utilizar el contexto de un problema para emitir conclusiones y planear investigaciones (Improving StatisticalThinking, 2007; Pfannkuch, & Wild, 2002).

Cabe señalar que independientemente del avance experimentado en la formación estadística y del desarrollo experimentado por la Didáctica de la Estadística, todavía existen ambientes de formación en los que existe una insuficiente comprensión de las herramientas estadísticas por parte de los estudiantes.

En tal dirección, cabe señalar que en universidades de la región amazónica se presentan dificultades con la formación estadística de los estudiantes. Las causas de estas dificultades no solo se limitan a la complejidad que encierra alcanzar las metas trazadas para este tipo de formación, pues además en este ambiente de formación sobresale la carencia de seguridad para estudiantes y profesores, una insuficiente disponibilidad de aulas y laboratorios especializados, y dificultades tecnológicas para establecer la comunicación entre estudiantes y profesores. Además, prevalece un ambiente lluvioso que no favorece la asistencia de los estudiantes y profesores.

Por ejemplo, la inseguridad los estudiantes imposibilita que ellos lleven consigo su computadora portátil (quienes la tienen) para la práctica con programas como SPSS o el Microsoft Excel. Se limitan las clases a temas teóricos y de

manera reducida al uso de Excel en los temas como gráficas de barras, circulares, nube de puntos, recta de regresión, entre otros.

Los aspectos adversos del ambiente de las universidades de la región amazónica limitan la formación de competencias estadísticas en los estudiantes. Por lo que se hace necesario buscar alternativas didácticas viables que puedan contribuir al logro de mejores resultados en la formación estadística de estos estudiantes universitarios.

En consecuencia, las carreras de Ingeniería Ambiental, Zootecnia, Tecnologías de la Información y Turismo de la Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador, que disponen de sílabos que coinciden en la mayoría de los contenidos estadísticos, podrían utilizar el presente libro para tratar los contenidos generales que se analizan en los cursos de estadística en cada una de estas carreras.

De esta forma, los estudiantes podrían disponer de un material docente que se ajuste más a sus necesidades de formación estadística, contribuyendo de esta forma a su formación profesional y ciudadana.

CAPÍTULO I. INTRODUCCIÓN A LA ESTADÍSTICA

Introducción

Un tema recurrente de la actual sociedad de la información es la creciente y determinante importancia que tiene la información para los individuos de la sociedad, en cualquier país, latitud, cultura o nivel de desarrollo. Aunque debe precisarse que aquellos individuos que hayan obtenido mayor nivel educacional y cultural, y que se encuentren en mejores perspectivas de desarrollo, estarán estimulados a consumir más y superior información para satisfacer sus crecientes necesidades.

Precisamente, la Estadística es una de las ciencias que cobra especial relevancia en la sociedad de la información, pues aporta una base conceptual y técnica para el estudio de fenómenos complejos, en los que hay que comenzar por definir el objeto de estudio y las variables relevantes, así como tomar datos de las mismas, interpretarlos y analizarlos con el fin de emitir conclusiones.

De forma general, se considera que la Estadística es la parte de las Matemáticas que estudia los fenómenos aleatorios, siendo un área de conocimiento clave para la toma de decisiones en entornos de riesgo e incertidumbre. Las contribuciones de esta ciencia son cruciales para el avance del conocimiento científico, el progreso tecnológico y la participación activa de la ciudadanía en el debate político, económico y social de la sociedad informacional actual.

Ahora bien, una definición de Estadística que ayuda a introducirse en sus rasgos característicos es la siguiente:

Estadística: ciencia que se ocupa del estudio de fenómenos de tipo genérico, normalmente complejos y enmarcados en un universo variable, mediante el empleo de modelos de reducción de la información y de análisis de validación de los resultados en términos de la representatividad. La información puede ser numérica, alfabética o simbólica. El proceso estadístico consiste en las fases de recogida de información, de análisis y de presentación e interpretación de los resultados y elaboración de métodos. (Sánchez y Manzano, 2002, p. 10)

En general, la Estadística tiene como objetivo el desarrollo de técnicas para el conocimiento numérico de un conjunto numeroso de datos empíricos (recogidos mediante experimentos o encuestas). Es decir, se ocupa de *recoger*, *organizar*, *resumir* y *analizar* una gran cantidad de datos obtenidos de la realidad para hacer visible lo invisible, e inferir conclusiones respecto de ellos.

La estadística: conceptos y terminología

Cabe señalar que la Estadística ofrece una base conceptual y técnica que posibilitan extraer información de los datos, muchas veces esta información no es posible conocerla con una simple observación de los datos porque generalmente estos son muy numerosos.

Las principales ***funciones de la Estadística*** son la descripción, relación y comparación de variables:

- **Descripción**: técnicas donde no se infiere, sólo se analiza la información que se posee, habitualmente a través de muestras. Se calculan descriptores que captan aspectos relevantes de la información de los datos y se dibujan gráficos que la resumen. En este caso la muestra es el fin, no un medio para otros análisis.

- **Relación**: técnicas que buscan relaciones entre variables, entre diferentes características medidas a una serie de individuos. Se busca la existencia de relación entre ellas y se pretende establecer sus relaciones matemáticas.
- **Comparación**: técnicas que comparan poblaciones de individuos. El objetivo es poder hablar de la igualdad o de la diferencia entre esos grupos, entre esas poblaciones.

Para llevar a cabo la *relación* y *comparación* se aplican técnicas inferenciales. En estos casos el trabajo con la muestra es un medio, no un fin en sí misma, pues a través de la misma se pretende extraer conclusiones poblacionales. Por lo tanto, desde la relación entre variables a nivel muestral o desde la comparación de dos o más muestras, se busca hacer afirmaciones poblacionales que van más allá de lo que se puede observar directamente en la muestra, para lo cual se realizan inferencias basadas en la probabilidad.

Es muy importante situar desde el principio cuál es el papel básico de la Estadística. Para lo cual es necesario utilizar conceptos probabilísticos: variable aleatoria, función de distribución, modelización matemática, entre otros. Estos conceptos son la base para la aplicación de técnicas estadísticas para describir la muestra y para inferir acerca de las características que no son directamente observables en la misma.

Según el colectivo a partir del cual se obtenga la información y el objetivo que persiga al analizar esos datos, la estadística se llama descriptiva o inferencial.

Estadística Descriptiva: se fundamenta en la descripción y análisis de las características de un conjunto de datos, de donde se extrae información y conclusiones sobre el comportamiento de los datos y relaciones existentes con entre ellos o de ellos con otras poblaciones con las cuales se comparan. Se trata de estimar, pronosticar y definir comportamientos que se puedan reproducir bajos similares condiciones de experimentación.

Estadística Inferencial: está fundamentada en los resultados obtenidos del análisis de una muestra de población, con el fin de inferir el comportamiento o característica de la población, de donde procede, por lo que recibe también el nombre de Inferencia estadística. El objetivo de la inferencia en investigación científica y tecnológica radica en conocer clases numerosas de objetos, personas o eventos a partir de otras relativamente pequeñas compuestas por los mismos elementos.

Los problemas por los que se ocupa la Estadística Inferencial se relacionan con la estimación de parámetros tanto muestrales como poblacionales y la definición de criterios para verificar si lo que se ha hecho u obtenido tiene la suficiencia en calidad estadística, y si se puede utilizar como elemento de pronóstico o de representación del fenómeno estudiado, con los cual se pueda tomar una decisión objetiva y lo más aproximada a la realidad.

Conceptos y Terminología

Población: es el conjunto de todos los elementos a los que se somete a un estudio estadístico.

Individuo o unidad estadística: es cada uno de los elementos que componen la población.

Observación: no debe confundirse la población en sentido demográfico y la población en sentido estadístico. La población en sentido demográfico es un conjunto de individuos (por ejemplo todos los habitantes de un país), mientras que una población en sentido estadístico es un conjunto de datos referidos a determinada característica o atributo de los individuos (por ejemplo las edades de todos los individuos de un país).

Muestra: es un conjunto representativo de la población de referencia, el número de individuos de una muestra es menor que el de la población.

Muestreo: el muestreo es una técnica que sirve para obtener una o más muestras de población. Éste se realiza una vez que se ha establecido un marco muestral representativo de la población, se procede a la selección de los elementos de la muestra aunque hay muchos diseños de la muestra.

Tipos de muestreo: existen dos métodos para seleccionar muestras de poblaciones; el muestreo no probabilístico y probabilístico aleatorio. En este último todos los elementos de la población tienen la oportunidad de ser escogidos en la muestra.

Muestreo no probabilístico: muchas veces se recurre a estos métodos, aun siendo conscientes de que no son los idóneos para realizar generalizaciones, dado que no se tiene certeza de que la muestra extraída sea representativa de la población. En general, en este tipo de muestreo, las unidades se seleccionan por conveniencia, de manera secuencial, siguiendo determinados criterios subjetivos o porque simplemente están disponibles.

Son tres los métodos principales de muestreo no probabilístico:

1. *Muestreo por conveniencia*: implica el empleo de una muestra integrada por las personas o los objetos cuya disponibilidad como sujetos de estudio sea más conveniente. Cuando los fenómenos que se investigan son suficientemente homogéneos en la población, se reduce el riesgo de sesgo.

2. *Muestreo por cuotas*: el investigador identifica estratos de la población y establece las proporciones de elementos necesarias a partir de los distintos segmentos estratificados. Con base en información previa acerca de la composición de la población, el investigador se asegura de que los diversos segmentos o sectores estén representados en la muestra en las mismas proporciones en que se presentan en la población. El muestreo por cuotas no requiere de la aplicación de técnicas complejas ni la inversión de una cantidad

extraordinaria de tiempo o esfuerzos, salvo por la identificación de estratos y la representación proporcional correspondiente, esta técnica es muy semejante a la de muestreo por conveniencia, por lo que comparte muchas de sus deficiencias.

3. *Muestreo intencional*: se basa en la idea de que el investigador puede usar sus conocimientos acerca de la población para elegir los casos que incluirá en la muestra. Quizá decida deliberadamente seleccionar la variedad más amplia posible de personas o los sujetos que a su juicio son característicos de la población que le interesa o que disponen de mayor información acerca del tema de estudio. Si bien esta forma subjetiva de muestreo no ofrece un método externo y objetivo para evaluar cuán típicos de la población son los sujetos seleccionados, puede representar ciertas ventajas con base en la técnica del informante clave.

Censo: se entiende por censo aquella numeración que se efectúa a todos y cada uno de los caracteres componentes de una población.

Encuesta: se entiende por encuesta las observaciones realizadas por muestreo, es decir son observaciones parciales.

Datos estadísticos: son los resultados del experimento o mediciones de las observaciones realizadas, son en general, el producto de las observaciones efectuadas en los cuales se produce el fenómeno que queremos estudiar. Un dato es cada uno de los valores que se ha obtenido al realizar un estudio estadístico.

Valor: un valor es cada uno de los distintos resultados que se pueden obtener en un estudio estadístico. Por ejemplo si lanzamos una moneda al aire 5 veces obtenemos dos valores: cara y cruz[1].

[1] Para mayor simplicidad en el análisis en el presente texto las dos caras de una moneda se clasificarán en cara y escudo, como habitualmente se utiliza en la literatura de Estadística con fines docentes.

Clasificación de los datos

Los datos estadísticos pueden ser clasificados en cualitativos, cuantitativos, cronológicos (series de tiempo), espaciales (series de espacios), entre otros.

- **Cuantitativos**: cuando son representados por un número.
- **Cualitativos**: cuando señalan cualidades y no están representados numéricamente.
- **Cronológicos**: cuando los valores de los datos varían en diferentes instantes o períodos de tiempo.
- **Espaciales**: cuando los datos están referidos a una localidad, espacio o área geográfica determinada.

Fuentes de datos estadísticos

Los datos estadísticos necesarios para la comprensión de los hechos pueden obtenerse a través de fuentes primarias o secundarias:

- **Fuentes primarias**: es el material de primera mano relativo a un fenómeno que se desea investigar, o sea, cuando se va al origen mismo de la información o experimento y se toman los datos directamente.
- **Fuentes secundarias**: es un texto basado en fuentes primarias, que implica un tratamiento de generalización, análisis, síntesis, interpretación o evaluación, por lo que la información se obtienen indirectamente del experimento u observación directa.

Métodos de recolección de datos

Método de recolección de datos: son los distintos tipos de procesos sistemáticos para recabar información de fuentes relevantes con el fin de encontrar respuestas a los problemas de investigación.

En función de la fuente de la que recogen la información, los métodos de recolección de datos pueden dividirse en dos categorías: métodos primarios o métodos secundarios.

Los métodos primarios de recolección de datos recopilan información directamente a través de fuentes de datos primarias, por lo que son datos de origen. Mientras que los métodos secundarios de recolección de datos recopilan información a partir de fuentes de datos secundarias.

La estadística emplea una variedad de métodos primarios de recolección de los datos que se desea investigar. Entre ellos sobresalen los cuestionarios, encuestas, entrevistas, observación y registros.

A continuación se presentan métodos primarios de recolección de datos que son de amplio uso.

Cuestionarios: consiste en un conjunto de preguntas respecto a una o más variables a medir. El contenido de las preguntas puede ser tan variado como los aspectos que mida y básicamente se puede hablar de dos tipos de preguntas: *cerradas* y *abiertas*.

Las preguntas cerradas contienen categorías o alternativas de respuestas que han sido delimitadas. Es decir, se presentan a los sujetos las posibilidades de respuestas y ellos deben circunscribirse a ellas. Pueden ser dicotómicas (dos alternativas de respuestas) o incluir varias alternativas de respuestas.

Por su parte, las preguntas abiertas son aquellas en las cuales se le pide al interrogado que responda él mismo a la pregunta formulada. Esto le otorga mayor libertad al entrevistado y posibilita adquirir respuestas más profundas, así como también preguntar sobre el por qué y cómo de las respuestas realizadas. A su vez, posibilita adquirir respuestas que no habían sido tenidas en cuenta a la hora de hacer los formularios y pueden crear así relaciones nuevas con otras variables y respuestas.

Encuestas: es un procedimiento dentro de los diseños de una investigación descriptiva en el que el investigador busca recopilar datos por medio de un cuestionario previamente diseñado en dar una entrevista a alguien, sin modificar el entorno ni el fenómeno donde se recoge la información, ya sea

para entregarlo en forma de tríptico, gráfica o tabla. Los datos se obtienen realizando un conjunto de preguntas normalizadas dirigidas a una muestra representativa o al conjunto total de la población estadística en estudio, integrada a menudo por personas, empresas o entes institucionales, con el fin de conocer estados de opinión, ideas, características o hechos específicos.

La encuesta es un método de trabajo relativamente económico y rápido. Si se cuenta con un equipo de entrevistadores y codificadores convenientemente entrenado, resulta fácil llegar rápidamente a una multitud de personas y obtener una gran cantidad de datos en poco tiempo. Su costo, para los casos simples, es sensiblemente bajo.

Entrevista: desde el punto de vista del método, es una forma específica de interacción social que tiene por objeto recolectar datos para una indagación. El investigador formula preguntas a las personas capaces de aportarle datos de interés, estableciendo un diálogo peculiar, asimétrico, donde una de las partes busca recoger informaciones y la otra es la fuente de esas informaciones.

La ventaja esencial de la entrevista reside en que son los mismos actores sociales quienes proporcionan los datos relativos a sus conductas, opiniones, deseos, actitudes y expectativas, que por su misma naturaleza es casi imposible de observar desde fuera. Nadie mejor que la misma persona involucrada para hablarnos acerca de todo aquello que piensa y siente, de lo que ha experimentado o proyecta hacer.

Observación: permite conocer la realidad mediante la sensopercepción directa de entes y procesos, para lo cual debe poseer algunas cualidades que le dan un carácter distintivo.

La observación es el método clásico más característico en las ciencias descriptivas y fue el primer método utilizado por los científicos, que en la actualidad continúa siendo su instrumento universal. Puede asumir muchas

formas; puede ser simple en la cual tanto el observador como los observados participan de la manera más natural posible, y en este caso el observador deberá tener un plan previo para la información a partir de las notas que vaya levantando a lo largo de la observación.

Pero en muchos casos es necesario una observación más sistemática con controles tanto para el observador como para el observado, para aumentar la precisión de su trabajo y protegerse de las críticas; no se pretende limitar en ningún grado las actividades de los individuos sino sistematizar el proceso de observación por medio de dispositivos sincronizadores mecánicos, observación en equipo, películas y grabaciones, planes e inventarios, casi a un paso de la situación que se vive en un laboratorio. Lo que depende del grado de conciencia que tengan los observados respecto a lo que se está realizando, y si se introduce el concepto de variables experimentales. Es muy importante señalar que la observación en sí puede conducir a una alteración de las condiciones de la realidad que se procura observar.

Variables estadísticas

Variable cualitativa, de atributos, o categórica: es una variable que clasifica o describe a un elemento de una población.

Variable cuantitativa o numérica: es aquella que cuantifica un elemento de una población.

Ejemplos:

1) Una muestra de cuatro clientes de un restaurante de Quito, Ecuador, fue encuestada en cuanto a: «nivel satisfacción por el servicio recibido» y «ciudad de residencia actual». Estas dos variables son cualitativas (de atributos), ya que describen alguna característica de la persona, y todas las personas con el mismo atributo pertenecen a la misma categoría. Los datos recolectados

fueron {muy satisfecho, satisfecho, algo satisfecho} y {Quito, Esmeraldas, Santa Elena, Guayaquil}.

2) El «costo total» de los libros de texto adquiridos por cada estudiante para las clases de este semestre es un ejemplo de variable cuantitativa (numérica). Se obtuvo una muestra con los datos siguientes: $238.87, $94.57, $139.24. [Para determinar el "costo promedio", simplemente se suman los tres números y el resultado se divide entre tres: (238.87 + 94.57+ 1 39.24)/3 = $157.56]

Observación: Algunas operaciones aritméticas, como sumar y promediar, tienen sentido para los datos que resultan de una variable cuantitativa.

Cada uno de estos tipos de variables puede subdividirse aún más, como se ilustra en la figura 1.

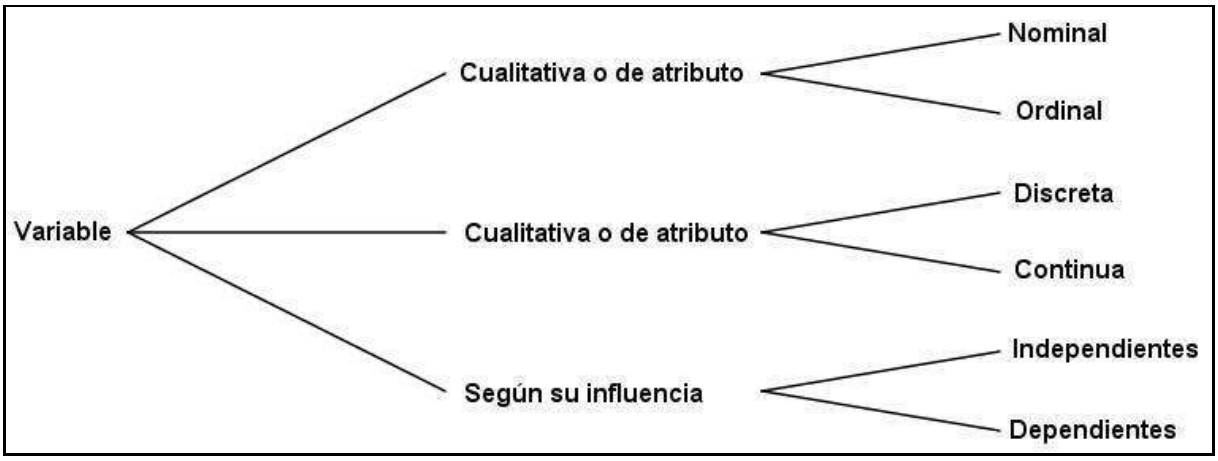


Figura 1. Clasificación de los tipos de variables y sus subdivisiones.

Las variables cualitativas pueden caracterizarse como nominales u ordinales.

Variable nominal: es una variable cualitativa que caracteriza (describe o identifica) a un elemento de una población. Para los datos resultantes de una variable nominal, las operaciones aritméticas no sólo carecen de sentido sino que tampoco se puede asignar un orden a las categorías.

En la encuesta anterior que se aplicó a los cuatro clientes de un restaurante, la variable «ciudad de residencia actual» es un ejemplo de variable nominal,

ya que identifica una característica de la persona y carece de sentido encontrar el promedio muestral al sumar y dividir entre cuatro. Por ejemplo, (Quito + Esmeraldas + Santa Elena + Guayaquil)/4 no está definido. Además, la ciudad de residencia actual no tiene un orden en sus categorías.

Variable ordinal: es una variable cualitativa que presenta una posición, o clasificación, ordenada.

Ejemplos:

1) En la encuesta anterior de cuatro clientes de un restaurante, la variable «nivel satisfacción por el servicio recibido» es un ejemplo de variable ordinal, ya que presenta una clasificación ordenada, por ejemplo «muy satisfecho» está antes que «satisfecho», que se encuentra antes que «algo satisfecho».

2) Otro ejemplo de una variable ordinal sería la clasificación de cinco fotografías de paisaje según la preferencia de alguien: primera elección, segunda elección, tercera elección, cuarta elección.

Las variables cuantitativas o numéricas también pueden subdividirse en dos clasificaciones: variables discretas y variables continuas.

Variable discreta: es una variable cuantitativa que puede asumir un número contable o finito de valores. Intuitivamente, la variable discreta puede asumir los valores correspondientes a puntos aislados a lo largo de un intervalo de recta, por lo que entre dos valores cualesquiera no se pueden fraccionar.

Variable continua: es una variable cuantitativa que puede asumir una cantidad incontable de valores. Intuitivamente, la variable continua puede asumir cualquier valor a lo largo de un intervalo de recta, incluyendo cualquier valor posible entre dos variables determinadas (se pueden fraccionar).

En muchos casos, es posible distinguir los dos tipos de variables decidiendo si las variables están relacionadas con un conteo o una medición.

Ejemplos:

1) La variable «edad de un estudiante» es un ejemplo de una variable discreta; sus valores se determinan al contar los años que tiene el mismo. Al contar, no es posible que ocurran valores fraccionarios; en consecuencia, entre los valores que puedan ocurrir hay huecos (números fraccionarios).

2) La variable «peso del estudiante» es un ejemplo de variable aleatoria continua; los valores de la variable se encuentran midiendo el peso. Al medir, puede ocurrir cualquier valor fraccionario; así, a lo largo de la recta es posible obtener cualquier valor.

3) Considere la variable «calificación asignada por un juez» en una competencia de patinaje artístico. Si se consideran algunas calificaciones que ya se han asignado: 9.9, 9.5, 8.8, 10.0, y se observa la presencia de cifras decimales, podría pensarse que todas las fracciones son posibles y concluir que la variable es continua. Sin embargo, esto no es cierto; de hecho, entre los valores posibles hay huecos y la variable es discreta.

Observación: cuando intente determinar si una variable es continua o discreta, recuerde analizar la naturaleza de la variable y piense en los valores que podría alcanzar. Fijar la atención solamente en los valores de datos que se han registrado puede ser engañoso.

Existe además una clasificación de variables según su influencia y estas son:

Variable independiente: es aquella característica o propiedad que se supone ser la causa del fenómeno estudiado; en investigación experimental se llama así a la variable que el investigador manipula.

Un tipo especial de variables son las de control, que modifican al resto de las variables independientes y que de no tenerse en cuenta adecuadamente pueden alterar los resultados por medio de un sesgo.

Variables dependientes: son las variables de respuesta que se observan en el estudio y que podrían estar influenciadas por los valores de las variables independientes. La variable dependiente es el factor que es observado y medido para determinar el efecto de la variable independiente.

Las variables, también suelen ser llamados caracteres cuantitativos, son aquellos que pueden ser expresados mediante números. Ejemplos de caracteres susceptibles de medición son la estatura, el peso, el salario y la edad.

Organización de la información

Tabla de frecuencia simple y agrupada

Una vez que se ha aplicado algún método de colección de datos, entonces estos datos se deben añadir a una matriz de datos de forma tal que se conforma la base de datos estadísticos.

Una forma de organizar la información de la base de datos estadísticos es a través de las tablas de frecuencias, que es una ordenación en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente.

Frecuencia absoluta: es el número de veces que aparece un determinado valor en un estudio estadístico y se representa por f_i.

Observación: la suma de las frecuencias absolutas es igual al número total de datos, que se representa por N.

$$f_1 + f_2 + f_3 + \cdots + f_n = N$$

Para indicar resumidamente estas sumas se utiliza la letra griega Σ (sigma mayúscula) que se lee suma o sumatoria.

$$\sum_{i=1}^{n} f_i = N$$

Frecuencia relativa: la frecuencia relativa es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos. Se puede expresar en tantos por ciento y se representa por n_i.

Donde: $n_i = \frac{f_i}{N}$

Observación: la suma de las frecuencias relativas es igual a 1.

Frecuencia acumulada: la frecuencia acumulada es la suma de las frecuencias absolutas de todos los valores inferiores o iguales al valor considerado y se representa por n_i.

Frecuencia relativa acumulada: la frecuencia relativa acumulada es el cociente entre la frecuencia acumulada de un determinado valor y el número total de datos y se representa por N_i. Se puede expresar en tantos por ciento.

Ejemplo

Durante el mes de julio, en una ciudad se han registrado las siguientes temperaturas máximas:

32, 31, 28, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33, 29, 29.

En la primera columna de la tabla 1 colocamos la variable ordenada de menor a mayor, en la segunda hacemos el recuento y en la tercera anotamos la frecuencia absoluta.

Sin embargo, lo más común es utilizar una distribución de frecuencias agrupadas o tabla con datos agrupados, que se emplea cuando las variables toman un número grande de valores o la variable estudiada es continua. Se agrupan los valores en intervalos que tengan la misma

amplitud denominados clases. A cada clase se le asigna su frecuencia correspondiente.

Tabla 1. Se representa una tabla de frecuencia simple.

x_i	f_i	F_i	n_i	N_i
27	1	1	0.032	0.032
28	2	3	0.065	0.097
29	6	9	0.194	0.290
30	7	16	0.226	0.0516
31	8	24	0.258	0.774
32	3	27	0.097	0.871
33	3	30	0.097	0.968
34	1	31	0.032	1
	31		1	

Para lograr obtener una distribución de frecuencias agrupadas es necesario determinar:

- **Límites de la clase:** cada clase está delimitada por el límite inferior de la clase y el límite superior de la clase (o sea, el menor y mayor valor que se alcanza en los datos en la clase).
- **Amplitud de la clase:** la amplitud de la clase es la diferencia entre el límite superior e inferior de la clase y se denota por a.
- **Marca de clase:** la marca de clase es el punto medio de cada intervalo y es el valor que representa a todo el intervalo para el cálculo de algunos parámetros y se denota por C_i.
- **Rango o recorrido:** es la diferencia que existe entre el mayor valor y el menor valor del número total de datos u observaciones; se denota por R y se calcula como $R = X_{max} - X_{min.}$

Ejemplo

A continuación se muestran los siguientes datos:

3, 15, 24, 28, 33, 35, 38, 42, 43, 38, 36, 34, 29, 25, 17, 7, 34, 36, 39, 44, 31, 26, 20, 11, 13, 22, 27, 47, 39, 37, 34, 32, 35, 28, 38, 41, 48, 15, 32, 13.

1er paso: se localizan los valores mínimo y máximo de la distribución. En este caso son 3 y 48.

2do paso: se restan y se busca un número entero un poco mayor que la diferencia y que sea divisible por el número de intervalos de queramos poner.

Es conveniente que el número de intervalos oscile entre 6 y 15.

En este caso, $R = X_{max} - X_{min} = 48 - 3 = 45$, si deseamos trabajar con intervalos de amplitud $a = 5$, entonces por conveniencia podemos considerar el número R igual a 50. De esta forma el número de intervalos es $C = R/a = 50/5 = 10$.

Se forman los 10 intervalos teniendo presente que el límite inferior de una clase pertenece al intervalo, pero el límite superior no pertenece al intervalo, pues se considera del siguiente intervalo.

Luego, se resume la información en una tabla de frecuencia agrupada, como se muestra en la tabla 2.

Tabla 2. Se representa una tabla de frecuencia agrupada.

Clases	C_i	f_i	F_i	n_i	N_i
[0-5)	2.5	1	1	0.025	0.025
[5-10)	7.5	1	2	0.025	0.050
[10-15)	12.5	3	5	0.075	0.125
[15-20)	17.5	3	8	0.075	0.200
[20-25)	22.5	3	11	0.075	0.2775
[25-30)	27.5	6	17	0.150	0.425
[30-35)	32.5	7	24	0.175	0.600
[35-40)	37.5	10	34	0.250	0.850
[40-45)	42.5	4	38	0.100	0.950

[45-50)	47.5	2	40	0.050	1
		40		1	

Para representar la información estadística resumida en las tablas de frecuencia es habitual que se utilicen diferentes tipos de gráficos. Uno de los más usados son los histogramas de frecuencia, los que pueden construirse tomando como base la frecuencia absoluta o relativa, así como las observaciones o clases correspondientes en dependencia que se tome como base una tabla de frecuencia simple o agrupada.

En la figura 2 se muestra un histograma de frecuencias absolutas que tomó como base la información de la tabla 2.

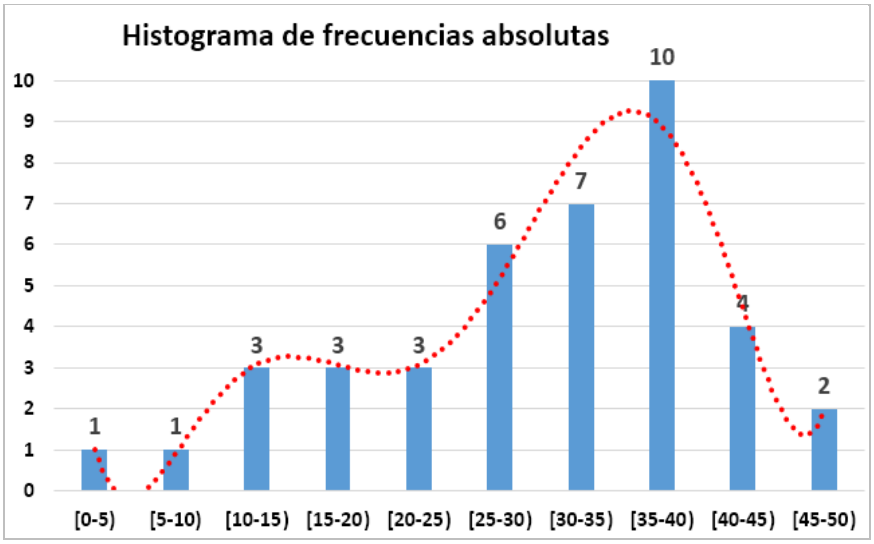


Figura 2. Histograma de frecuencias absolutas.

Por su parte, en la figura 3 se muestra un histograma de frecuencias relativas que también tomó como base la información de la tabla 2.

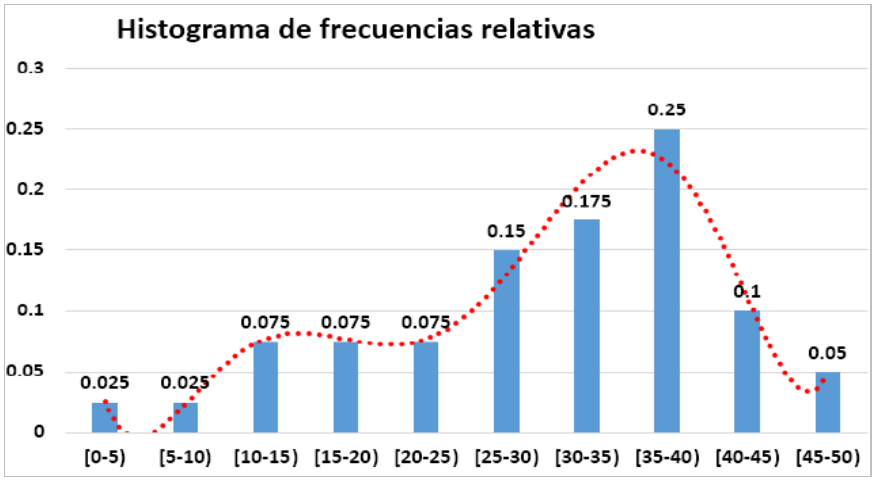


Figura 3. Histograma de frecuencias relativas.

Regla de Sturges

La regla de Sturges es una regla que sirve para calcular el número razonable de clases o intervalos en los que se debe dividir un conjunto de datos. La fórmula de la regla de Sturges establece que el número de clases es igual a uno más el logaritmo en base dos del número total de datos:

$$C = 1 + \log_2 N$$

Donde C es el número de clases o intervalos y N es el número total de observaciones de la muestra.

La mayoría de calculadoras solo permiten hacer cálculos con logaritmos de base 10. En tal caso, puedes utilizar esta fórmula equivalente:

$$C = 1 + \frac{\log(N)}{\log(2)}$$

Ahora que ya sabemos en qué consiste la regla de Sturges, vamos a mostrar un ejemplo que ilustre su aplicación.

Ejemplo

Se ha medido la altura a una muestra de 50 personas diferentes y se han registrado todos los valores en la tabla 3 de datos. Aplica la regla de Sturges para dividir el conjunto de datos en intervalos y luego represente los datos en un histograma de frecuencias absolutas.

Tabla 3. Altura de una muestra de 50 personas diferentes.

Tabla de datos estadísticos				
158	165	174	153	192
183	145	189	159	157
172	189	175	167	205
149	173	180	182	188
177	175	159	175	170
162	181	168	193	163
151	163	156	156	159
159	162	171	155	151
162	150	184	178	168
188	201	196	174	153

En primer lugar, tenemos que separar los datos en intervalos. En total hay 50 datos, por lo tanto, usamos la regla de Sturges con este valor:

$$C = 1 + \log_2 N$$

$$C = 1 + \log_2 50$$

$$C = 1 + 5{,}64$$

$$C \approx 7$$

De modo que debemos separar los datos y agruparlos en siete intervalos. Ahora necesitamos saber la amplitud de cada intervalo, para ello, simplemente tenemos que dividir el valor máximo menos el valor mínimo entre el número total de intervalos:

$$a = \frac{Xmax - Xmin}{C} = \frac{205 - 145}{7} = \frac{60}{7} = 8{,}571 \approx 9$$

En definitiva, tienen que haber 7 intervalos con una amplitud de 9. Una vez hemos calculado los intervalos, tenemos que contar el número de veces que aparece un dato en cada intervalo y construir la tabla de frecuencias:

Tabla 4. Clases y frecuencias absolutas de la variable altura.

Clases	f_i
[145-154)	7
[154-163)	12
[163-172)	8

[172-181)	10
[181-190)	8
190-199)	3
[199-208)	2

Luego, en la figura 4 se muestra el histograma de frecuencias relativas de la variable altura en una muestra de 50 personas diferentes.

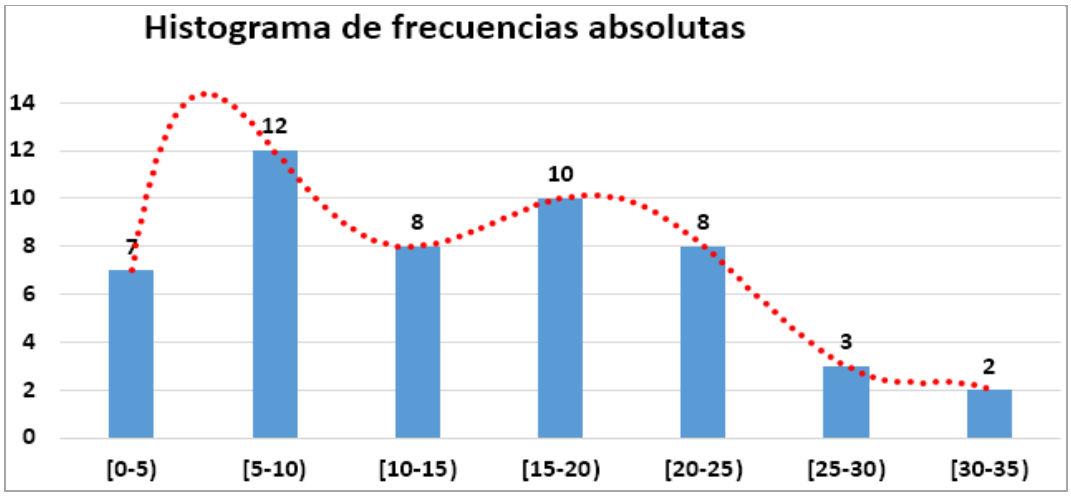


Figura 4. Histograma de frecuencias absolutas para la variable peso.

Tarea en línea:

En el sitio web ProbabilidadyEstadística.net hay disponible una «Calculadora de la regla de Sturges»: https://www.probabilidadyestadistica.net/regla-de-sturges/

Resumen y caracterización de la información

Estamos ahora en condiciones de caracterizar la información, resumiendo la misma mediante un conjunto reducido de valores que describan las características generales de la distribución de frecuencias. Para realizar esta descripción se utilizarán los parámetros estadísticos.

Parámetro estadístico: es un número que se obtiene a partir de los datos de una distribución estadística que sirve para sintetizar la información.

Hay tres tipos de parámetros estadísticos: de centralización, de posición, de dispersión.

Medidas de centralización

Al describir grupos de diferentes observaciones, con frecuencia es conveniente resumir la información con un solo número. Este número que, para tal fin, suele situarse hacia el centro de la distribución de datos se denomina medida o parámetro de tendencia central o de centralización. Dicho número indican en torno a qué valor (centro) se distribuyen los datos.

Las medidas de centralización que estudiaremos son la media aritmética y la moda.

Media aritmética: es el valor promedio de la distribución que se obtiene al sumar todos los datos y dividir el resultado entre el número total de datos; se representa por $\overline{X}$ es el símbolo de la media aritmética y se calcula a partir de la siguiente expresión

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N}$$

Ejemplo

Los pesos de seis amigos son: 84, 91, 72, 68, 87 y 78 kg. Hallar el peso medio.

Solución:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{84 + 91 + 72 + 68 + 87 + 78}{6} = 80 \text{ kg}$$

Media aritmética para datos agrupados: cuando los datos están agrupados en una tabla de frecuencias, la expresión de la media es:

$$\overline{X} = \frac{\sum_{i=1}^{N} C_i f_i}{N} = \frac{C_1 f_1 + C_2 f_2 + C_3 f_3 + \cdots + C_N f_N}{N}$$

Donde

C_i es la marca de clase del intervalo i-ésimo.

f_i es la frecuencia absoluta de cada intervalo i-ésimo.

Ejemplo

En un test realizado a un grupo de 42 personas se han obtenido las puntuaciones que muestra la tabla 5. Calcula la puntuación media.

Tabla 5. Resultados agrupados de un test aplicado a 42 personas.

Clases	Ci	f_i	$C_i \cdot f_i$
[10, 20)	15	1	15
[20, 30)	25	8	200
[30,40)	35	10	350
[40, 50)	45	9	405
[50, 60)	55	8	440
[60,70)	65	4	260
[70, 80)	75	2	150
		42	1 820

Solución

$$\overline{X} = \frac{\sum_{i=1}^{C} C_i f_i}{N} = \frac{C_1 f_1 + C_2 f_2 + C_3 f_3 + \cdots + C_C f_C}{N}$$

$$\overline{X} = \frac{15 + 200 + 350 + 405 + 440 + 260 + 150}{42}$$

$$\overline{X} = \frac{1820}{42} = 43.333$$

Propiedades de la media aritmética

1. La suma de las desviaciones de todas las puntuaciones de una distribución respecto a la media de la misma igual a cero.

$$\sum_{i=1}^{N}(X_i - \overline{X}) = 0$$

Ejemplo:

Las suma de las desviaciones de los números 8, 3, 5, 12, 10 de su media aritmética 7.6 es igual a 0:

$\sum_{i=1}^{5}(X_i - \overline{X}) = (8 - 7.6) + (3 - 7.6) + (5 - 7.6) + (12 - 7.6) + (10 - 7.6)$

$\sum_{i=1}^{5}(X_i - \overline{X}) = 0.4 - 4.6 - 2.6 + 4.4 + 2.4$

$\sum_{i=1}^{5}(X_i - \overline{X}) = 0$

2. La media aritmética de los cuadrados de las desviaciones de los valores de la variable con respecto a un número cualquiera se hace mínima cuando dicho número coincide con la media aritmética.
3. Si a todos los valores de la variable se les suma un mismo número, la media aritmética queda aumentada en dicho número.
4. Si todos los valores de la variable se multiplican por un mismo número la media aritmética queda multiplicada por dicho número.

Observaciones:

1. La media se puede hallar sólo para variables cuantitativas.
2. La media es independiente de las amplitudes de los intervalos.
3. La media es muy sensible a las puntuaciones extremas.

Ejemplo:

Si tenemos una distribución con los siguientes pesos: 65 kg, 69kg, 65 kg, 72 kg, 66 kg, 75 kg, 70 kg, 110 kg.

La media es igual a 74 kg, que es una medida de centralización poco representativa de la distribución.

4. La media no se puede calcular si hay un intervalo con una amplitud indeterminada, porque no podemos calcular la marca de clase correspondiente.

Moda: es el valor que más se repite en una distribución y se representa por Mo. Se puede hallar la moda para variables cualitativas y cuantitativas.

Ejemplos:

1) Hallar la moda de la distribución: 2, 3, 3, 4, 4, 4, 5, 5 Mo = 4

2) Si en un grupo hay dos o varias puntuaciones con la misma frecuencia y esa frecuencia es la máxima, la distribución es bimodal o multimodal, es decir, tiene varias modas.

1, 1, 1, 4, 4, 5, 5, 5, 7, 8, 9, 9, 9 Mo = 1, 5, 9

3) Cuando todas las puntuaciones de un grupo tienen la misma frecuencia, no hay moda.

2, 2, 3, 3, 6, 6, 9, 9

4) Si dos puntuaciones adyacentes tienen la frecuencia máxima, la moda es el promedio de las dos puntuaciones adyacentes.

0, 1, 3, 3, 5, 5, 7, 8 Mo = 4

Para el caso de datos agrupados el cálculo de la moda se hace para dos casos diferentes, cuando los intervalos de clase tienen la misma amplitud y cuando tienen amplitudes diferentes.

I) Cuado los intervalos de clase tienen la misma amplitud:

$$M_o = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i$$

Donde:

L_i es el límite inferior de la clase modal.

f_i es la frecuencia absoluta de la clase modal.

f_{i-1} es la frecuencia absoluta inmediatamente inferior a la clase modal.

f_{i+1} es la frecuencia absoluta inmediatamente posterior a la clase modal.

a_i es la amplitud de clase.

Ejemplo

Calcule la moda de una distribución estadística dada por la tabla 6.

Tabla 6. Distribución estadística de una variable.

Clases	fi
[60, 63)	5
[63, 66)	18
[66, 69)	42
[69, 72)	27
[72, 75)	8
	100

Solución

$$M_o = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \cdot a_i = 66 + \frac{42-18}{(42-18)+42-27)} \cdot 3 = 67.846$$

II) Cuando los intervalos tienen amplitudes distintas:

En primer lugar tenemos que hallar las alturas.

$$h_i = \frac{f_i}{a_i}$$

La clase modal es la que tiene mayor altura.

$$M_o = L_i + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot a_i$$

Ejemplo

En la tabla 7 se muestran las calificaciones (suspenso, aprobado, notable y sobresaliente) obtenidas por un grupo de 50 alumnos. Calcular la moda.

Tabla 7. Distribución estadística de las calificaciones de 50 alumnos.

Clases	fi	hi
[0, 5)	15	3
[5, 7)	20	10
[7, 9)	12	6
[9, 10)	3	3
	50	

$$M_o = L_i + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (h_i - h_{i+1})} \cdot a_i$$

$$M_o = 5 + \frac{10 - 3}{(10 - 3) + (10 - 6)} \cdot 2 = 6.273$$

Medidas de posición

Medidas de posición: dividen un conjunto de datos en grupos con el mismo número de individuos. Para calcular las medidas de posición es necesario que los datos estén ordenados de menor a mayor.

Las medidas de posición que estudiaremos son:

- **La mediana**: divide la serie de datos en dos partes iguales.
- **Los cuartiles**: dividen la serie de datos en cuatro partes iguales.
- **Deciles**: dividen la serie de datos en diez partes iguales.
- **Percentiles**: dividen la serie de datos en cien partes iguales.

La mediana: es la puntación de la escala que separa la mitad superior de la distribución y la inferior, es decir divide la serie de datos en dos partes iguales.

Es el valor que ocupa el lugar central de todos los datos cuando éstos están ordenados de menor a mayor. Se representa por Me y se determina sólo para variables cuantitativas.

A continuación se muestran los pasos a seguir para el cálculo de la mediana:

1. Ordenamos los datos de menor a mayor.
2. Si la serie tiene un número impar de medidas la mediana es la puntuación central de la misma.

   Ejemplo: 2, 3, 4, 4, 5, 5, 5, 6, 6 Me= 5
3. Si la serie tiene un número par de puntuaciones la mediana es la media entre las dos puntuaciones centrales.

   Ejemplo: 6, 7, 8, 9, 10, 11, 12, 13 Me= 9.5

Cuando los datos están agrupados la mediana se encuentra en el primer intervalo donde la frecuencia acumulada llega al menos a representar la mitad de la suma de las frecuencias absolutas.

Es decir tenemos que buscar el intervalo en el que se encuentre $N/2$.

$$M_e = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot a_i$$

Donde:

L_i es el límite inferior de la clase donde se encuentra la mediana.

$\frac{N}{2}$ es la semisuma de las frecuencias absolutas.

F_{i-1} es la frecuencia acumulada anterior a la clase mediana.

a_i es la amplitud de la clase.

f_i es la frecuencia absoluta asociada a la clase mediana.

Debe tenerse en cuenta que la mediana es independiente de las amplitudes de los intervalos.

Ejemplo

Calcule la mediana de una distribución estadística que está dada en la tabla 8:

Tabla 8. Distribución estadística de una variable.

Clases	fi	Fi
[60, 63)	5	5
[63, 66)	18	23
[66, 69)	42	65
[69, 72)	27	92
[72, 75)	8	100
	100	

Solución:

Para este caso

$N/2 = 100/2 = 50$

Clase modal: [66, 69)

$$M_e = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot a_i = 66 + \frac{50-23}{42} \cdot 3 = 67.929$$

Los cuartiles: son los tres valores de la variable que dividen a un conjunto de datos ordenados en cuatro partes iguales. Q_1, Q_2 y Q_3 determinan los valores correspondientes al 25%, al 50% y al 75% de los datos; Q_2 coincide con la mediana.

Cálculo de los cuartiles

1. Ordenamos los datos de menor a mayor.
2. Buscamos el lugar que ocupa cada cuartil mediante la expresión

$\frac{k \cdot N}{4}$ donde $k = 1, 2, 3$

Caso de un número impar de datos

Ejemplo: 2, 5, 3, 6, 7, 4, 9

Al ordenar de mayor a menor se obtiene:

2, 3, 4, 5, 6, 7, 9,

↓ ↓ ↓

Q_1 Q_2 Q_3

- Caso de un número par de datos

Ejemplo: 2, 5, 3, 4, 6, 7, 1, 9

Al ordenar de mayor a menor se obtiene:

1, 2, 3, 4, 5 6, 7 9

2.5 4.5 6.5

↓ ↓ ↓

Q_1 Q_2 Q_3

- Cálculo de los cuartiles para datos agrupados

En primer lugar buscamos en la tabla de frecuencias acumuladas la clase donde se encuentra $\frac{k \cdot N}{4}$ donde $k = 1, 2, 3$

Luego, calculamos los cuartiles a través de la siguiente expresión:

$$Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i, \quad \text{para} \quad k = 1, 2, 3.$$

Donde:

Li es el límite inferior de la clase donde se encuentra la mediana.

N es la suma de las frecuencias absolutas.

Fi-1 es la frecuencia acumulada anterior a la clase mediana.

ai es la amplitud de la clase.

Ejemplo

Calcular los cuartiles de la distribución de la tabla 9.

Tabla 9. Distribución estadística de una variable.

clases	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18

[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	65	

Solución:

- Cálculo del primer cuartel $\frac{65 \cdot 1}{4} = 16.25$

$$Q_1 = 60 + \frac{\frac{16 \cdot 25}{4} - 8}{10} \cdot 10 = 68.25$$

- Cálculo del segundo cuartil

$(65 \cdot 2)/4 = 32.5$

$$Q_2 = 70 + \frac{32.5 - 18}{16} \cdot 10 = 79.0625$$

- Cálculo del tercer cuartil

$(65 \cdot 3)/4 = 48.75$

$$Q_3 = 90 + \frac{48.75 - 48}{10} \cdot 10 = 90.75$$

Los deciles: son los nueve valores que dividen la serie de datos en diez partes iguales. Los deciles dan los valores correspondientes al 10%, al 20%... y al 90% de los datos; D_5 coincide con la mediana.

Cálculo de los deciles

1. En primer lugar buscamos en la tabla de las frecuencias acumuladas la clase donde se encuentra $\frac{k \cdot N}{10}$, k = 1, 2, 3,…, 9

2. En segundo lugar aplicamos la fórmula

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i,\ k = 1, 2, 3,\ldots, 9$$

Donde:

Li es el límite inferior de la clase donde se encuentra la mediana.

N es la suma de las frecuencias absolutas.

Fi-1 es la frecuencia acumulada anterior a la clase mediana.

ai es la amplitud de la clase.

Ejemplo

Calcular los deciles de la distribución de la tabla 10.

Tabla 10. Distribución estadística de una variable.

Clases	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	65	

Solución:

- Cálculo del primer decil:

$$\frac{k \cdot N}{10} = \frac{1 \cdot 65}{10} = 6.5$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_1 = 50 + \frac{\frac{1 \cdot 65}{10} - 0}{8} \cdot 10 = 58.125$$

- Cálculo del segundo decil:

$$\frac{k \cdot N}{10} = \frac{2 \cdot 65}{10} = 13$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_2 = 60 + \frac{\frac{2 \cdot 65}{10} - 8}{10} \cdot 10 = 65$$

- Cálculo del tercer decil:

$$\frac{k \cdot N}{10} = \frac{3 \cdot 65}{10} = 19.5$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_3 = 70 + \frac{\frac{3 \cdot 65}{10} - 18}{16} \cdot 10 = 70.938$$

- Cálculo del cuarto decil

$$\frac{k \cdot N}{10} = \frac{4 \cdot 65}{10} = 26$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_4 = 70 + \frac{\frac{4 \cdot 65}{10} - 18}{16} \cdot 10 = 75$$

- Cálculo del quinto decil

$$\frac{k \cdot N}{10} = \frac{5 \cdot 65}{10} = 32.5$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_5 = 70 + \frac{\frac{5 \cdot 65}{10} - 18}{16} \cdot 10 = 79.063$$

- Cálculo del sexto decil

$$\frac{k \cdot N}{10} = \frac{6 \cdot 65}{10} = 39$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_6 = 80 + \frac{\frac{6 \cdot 65}{10} - 34}{14} \cdot 10 = 83.571$$

- Cálculo del séptimo decil

$$\frac{k \cdot N}{10} = \frac{7 \cdot 65}{10} = 45.5$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_7 = 80 + \frac{\frac{7 \cdot 65}{10} - 34}{14} \cdot 10 = 88.214$$

- Cálculo del octavo decil

$$\frac{k \cdot N}{10} = \frac{8 \cdot 65}{10} = 52$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_8 = 90 + \frac{\frac{8 \cdot 65}{10} - 48}{10} \cdot 10 = 94$$

- Cálculo del noveno decil

$$\frac{k \cdot N}{10} = \frac{9 \cdot 65}{10} = 58.5$$

$$D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i \qquad D_9 = 100 + \frac{\frac{9 \cdot 65}{10} - 58}{5} \cdot 10 = 101$$

Los percentiles: son los 99 valores que dividen la serie de datos en 100 partes iguales. Los percentiles dan los valores correspondientes al 1%, al 2%... y al 99% de los datos. P_{50} coincide con la mediana.

Cálculo de los percentiles:

1. En primer lugar buscamos en la tabla de las frecuencias acumuladas la clase donde se encuentra $\frac{k \cdot N}{100}$, k = 1, 2, 3,…, 99.

2. En segundo lugar aplicamos la fórmula

$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i$, k = 1, 2, 3,…, 99

Donde:

Li es el límite inferior de la clase donde se encuentra la mediana.

N es la suma de las frecuencias absolutas.

Fi-1 es la frecuencia acumulada anterior a la clase mediana.

ai es la amplitud de la clase.

Ejemplo

Calcular el percentil 35 y 60 de la distribución de la tabla 11.

Tabla 11. Distribución estadística de una variable.

Clases	f_i	F_i
[50, 60)	8	8
[60, 70)	10	18
[70, 80)	16	34
[80, 90)	14	48
[90, 100)	10	58
[100, 110)	5	63
[110, 120)	2	65
	65	

Solución:

- Percentil 35

$$\frac{k \cdot N}{100} = \frac{35 \cdot 65}{100} = 22.75$$

$$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i \qquad P_{35} = 70 + \frac{\frac{35 \cdot 65}{100} - 18}{16} \cdot 10 = 72.969$$

- Percentil 60

$$\frac{k \cdot N}{100} = \frac{60 \cdot 65}{100} = 39$$

$$P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i \qquad P_{60} = 80 + \frac{\frac{60 \cdot 65}{100} - 34}{14} \cdot 10 = 83.571$$

Medidas de dispersión

A continuación estudiaremos las medidas de dispersión

Las medidas de dispersión: nos informan sobre cuánto se alejan del centro los valores de la distribución.

Las medidas de dispersión que estudiaremos son:

- **El Rango o recorrido**: es la diferencia entre el mayor y el menor de los datos de una distribución estadística.
- **Desviación media**: es la media aritmética de los valores absolutos de las desviaciones respecto a la media.
- **Varianza**: es la media aritmética del cuadrado de las desviaciones respecto a la media.
- **Desviación típica**: es la raíz cuadrada positiva de la varianza.

Rango o recorrido: es la diferencia entre el mayor (X_{max}) y el menor (X_{min}) de los datos de una distribución estadística, se denota por R y se calcula a partir de la siguiente expresión: $R = X_{max} - X_{min}$

Ejemplo

Calcular el rango o recorrido de la distribución siguiente:

9, 3, 8, 8, 2, 10, 9, 8, 9, 18

Solución:

En este caso $X_{max} = 18$ y $X_{min} = 2$

Entonces $R = X_{max} - X_{min} = 18 - 2 = 16$

Desviación media: es la media aritmética de los valores absolutos de las desviaciones respecto a la media, se representa por D_m y se calcula a partir de las siguientes expresiones:

1) Para datos no agrupados

$$D_m = \frac{\sum_{i=1}^{N}|x_i - \bar{x}|}{N} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + |x_3 - \bar{x}| + \cdots + |x_N - \bar{x}|}{N}$$

2) Para datos agrupados

$$D_m = \frac{\sum_{i=1}^{C} f_i \cdot |C_i - \bar{x}|}{N} = \frac{f_1 \cdot |C_1 - \bar{x}| + f_2 \cdot |C_2 - \bar{x}| + f_3 \cdot |C_3 - \bar{x}| + \cdots + f_C \cdot |C_C - \bar{x}|}{N}$$

Donde:

N es la cantidad de observaciones de la muestra.

C_i es la marca de clase del intervalo i-ésimo.

f_i es la frecuencia absoluta de cada intervalo i-ésimo.

Ejemplos:

1) Calcular la desviación media de la distribución:

9, 3, 8, 8, 9, 8, 9, 18

Solución:

Primero se determinará la media

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$D_m = \frac{\sum_{i=1}^{N}|x_i - \bar{x}|}{N} = \frac{|9-9|+|3-9|+|8-9|+|8-9|+|9-9|+|8-9|+|9-9|+|18-9|}{8} = 2.25$$

2) Calcular la desviación media para datos agrupados si los datos vienen agrupados en la tabla 12.

Tabla 12. Distribución estadística de una variable para datos agrupados.

Clases	C_i	f_i

[10, 15)	12.5	3
[15, 20)	17.5	5
[20, 25)	22.5	7
[25, 30)	27.5	4
[30, 35)	32.5	2
		21

Solución:

$$\bar{x} = \frac{C_1 f_1 + C_2 f_2 + C_3 f_3 + \cdots + C_C f_C}{N} = \frac{12.5 \cdot 3 + 17.5 \cdot 5 + 22.5 \cdot 7 + 27.5 \cdot 4 + 32.5 \cdot 2}{21}$$

$$\bar{x} = 21.786$$

$$D_m = \frac{\sum_{i=1}^{C} f_i \cdot |C_i - \bar{x}|}{N} = \frac{f_1 \cdot |C_1 - \bar{x}| + f_2 \cdot |C_2 - \bar{x}| + f_3 \cdot |C_3 - \bar{x}| + f_4 \cdot |C_4 - \bar{x}| + f_5 \cdot |C_5 - \bar{x}|}{N}$$

$$= \frac{3 \cdot |12.5 - 21.786| + 5 \cdot |17.5 - 21.786| + 7 \cdot |22.5 - 21.786| + 4 \cdot |27.5 - 21.786| + 2 \cdot |32.5 - 21.786|}{21}$$

$$D_m = 4.694$$

La varianza: es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística; se representa por $\mathrm{Var}\, x$ y se calcula a partir de las siguientes expresiones:

1) Para datos no agrupados

$$\mathrm{Var}\, X = \frac{\sum_{i=1}^{N} (C_i - \bar{X})^2}{N} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2}{N}$$

2) Para datos agrupados

$$\mathrm{Var}\, X = \frac{\sum_{i=1}^{C} f_i \cdot (C_i - \bar{X})^2}{N} = \frac{f_1 \cdot (C_1 - \bar{X})^2 + f_2 \cdot (C_2 - \bar{X})^2 + f_3 \cdot (C_3 - \bar{X})^2 + \cdots + f_C \cdot (C_C - \bar{X})^2}{N}$$

Para simplificar el cálculo de la varianza podemos utilizar las siguientes expresiones que son equivalentes a las anteriores:

1) Para datos no agrupados

$$\text{Var } x = \sum_{i=1}^{N} \frac{X_i{}^2}{N} - \overline{X}^2 = \frac{X_1{}^2 + X_2{}^2 + X_3{}^2 + \cdots + X_N{}^2}{N} - \overline{X}^2$$

2) Para datos agrupados

$$\text{Var } X = \sum_{i=1}^{C} \frac{f_i \cdot X_i{}^2}{N} - \overline{X}^2 = \frac{f_1 \cdot C_1{}^2 + f_2 \cdot C_2{}^2 + f_3 \cdot C_3{}^2 + \cdots + f_C \cdot C_C{}^2}{N} - \overline{X}^2$$

Ejemplos

1) Calcular la varianza de la distribución: 9, 3, 8, 8, 9, 8, 9, 18

Solución:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$\text{Var } X = \frac{\sum_{i=1}^{N} (C_i - \overline{X})^2}{N} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + \cdots + (X_N - \overline{X})^2}{N}$$

$$\text{Var } X = \frac{(9-9)^2 + (3-9)^2 + (8-9)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (9-9)^2 + (18-9)^2}{8}$$

$$\text{Var } X = 15$$

2) Calcular la varianza de la distribución de la tabla 13.

Tabla 13. Distribución estadística de una variable para datos agrupados.

Clases	C_i	f_i
[10, 20)	15	1
[20, 30)	25	8
[30,40)	35	10
[40, 50)	45	9
[50, 60	55	8
[60,70)	65	4
[70, 80)	75	2
		42

Solución:

$$\bar{x} = \frac{C_1 f_1 + C_2 f_2 + C_3 f_3 + \cdots + C_C f_C}{N}$$

$$\bar{x} = \frac{15 \cdot 1 + 25 \cdot 8 + 35 \cdot 10 + 45 \cdot 9 + 55 \cdot 8 + 65 \cdot 4 + 75 \cdot 2}{42} = 43.333$$

$$\text{Var } X = \frac{\sum_{i=1}^{C} f_i \cdot (C_i - \bar{X})^2}{N} = 218.68$$

Propiedades de la varianza

1. La varianza será siempre un valor positivo o cero, en el caso de que las puntuaciones sean iguales.
2. Si a todos los valores de la variable se les suma un número la varianza no varía.
3. Si todos los valores de la variable se multiplican por un número la varianza queda multiplicada por el cuadrado de dicho número.
4. Si tenemos varias distribuciones con la misma media y conocemos sus respectivas varianzas se puede calcular la varianza total.

   a) Si todas las muestras tienen el mismo tamaño:

   $$\text{Var } X = \frac{\text{Var } X_1 + \text{Var } X_2 + \text{Var } X_3 + \cdots + \text{Var } X_N}{N}$$

   b) Si las muestras tienen distinto tamaño $k_1, k_2, k_3, \ldots, k_N$:

   $$\text{Var } X = \frac{k_1 \cdot \text{Var } X_1 + k_2 \cdot \text{Var } X_2 + k_3 \cdot \text{Var } X_3 + \cdots + k_p \cdot \text{Var } X_N}{k_1 + k_2 + k_3 + \cdots + k_N}$$

Observaciones sobre la varianza

1. La varianza, al igual que la media, es un índice muy sensible a las puntuaciones extremas.
2. En los casos que no se pueda hallar la media tampoco será posible hallar la varianza.

3. La varianza no viene expresada en las mismas unidades que los datos, ya que las desviaciones están elevadas al cuadrado.

La desviación típica: es la raíz cuadrada de la varianza, o sea, es la raíz cuadrada de la media de los cuadrados de las puntuaciones de desviación; se representa por S y se calcula a partir de las siguientes expresiones:

1) Para datos no agrupados

$$S = \sqrt{VarX} = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}} = \sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + \cdots + (X_N - \overline{X})^2}{N}}$$

2) Para datos agrupados

$$S = \sqrt{VarX} = \sqrt{\frac{\sum_{i=1}^{N} f_i \cdot (C_i - \overline{X})^2}{N}}$$

$$S = \sqrt{VarX} = \sqrt{\frac{f_1 \cdot (C_1 - \overline{X})^2 + f_2 \cdot (C_2 - \overline{X})^2 + f_3 \cdot (C_3 - \overline{X})^2 + \cdots + f_N \cdot (C_N - \overline{X})^2}{N}}$$

Para simplificar el cálculo vamos o utilizar las siguientes expresiones que son equivalentes a las anteriores.

1) Para datos no agrupados

$$S = \sqrt{\sum_{i=1}^{N} \frac{X_i{}^2}{N} - \overline{X}^2} = \sqrt{\frac{X_1{}^2 + X_2{}^2 + X_3{}^2 + \cdots + X_N{}^2}{N} - \overline{X}^2}$$

2) Para datos agrupados

$$S = \sqrt{\sum_{i=1}^{C} \frac{f_i \cdot X_i{}^2}{N} - \overline{X}^2} = \sqrt{\frac{f_1 \cdot C_1{}^2 + f_2 \cdot C_2{}^2 + f_3 \cdot C_3{}^2 + \cdots + f_C \cdot C_C{}^2}{N} - \overline{X}^2}$$

Ejemplos

1) Calcular la desviación típica de la distribución: 9, 3, 8, 8, 9, 8, 9, 18.

Solución

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$S = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}} = \sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + \cdots + (X_N - \overline{X})^2}{N}}$$

$$S = \frac{(9-9)^2 + (3-9)^2 + (8-9)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (9-9)^2 + (18-9)^2}{8}$$

S =3.873

2) Calcular la desviación típica de la distribución de la tabla 13.

Solución

$$\overline{x} = \frac{C_1 f_1 + C_2 f_2 + C_3 f_3 + \cdots + C_C f_C}{N}$$

$$\overline{x} = \frac{15 \cdot 1 + 25 \cdot 8 + 35 \cdot 10 + 45 \cdot 9 + 55 \cdot 8 + 65 \cdot 4 + 75 \cdot 2}{42} = 43.333$$

$$S = \sqrt{\frac{\sum_{i=1}^{C} f_i \cdot (C_i - \overline{X})^2}{N}} = \sqrt{218.68} = 14.797$$

Propiedades de la desviación típica

1. La desviación típica será siempre un valor positivo y es igual a cero cuando todas las puntuaciones son iguales.
2. Si a todos los valores de la variable se les suma un número la desviación típica no varía.
3. Si todos los valores de la variable se multiplican por un número la desviación típica queda multiplicada por dicho número.
4. Si tenemos varias distribuciones con la misma media y conocemos sus respectivas desviaciones típicas se puede calcular la desviación típica total.

   a) Si todas las muestras tienen el mismo tamaño:

$$S = \sqrt{\frac{\text{Var } X_1 + \text{Var } X_2 + \text{Var } X_3 + \cdots + \text{Var } X_N}{N}}$$

b) Si las muestras tienen distinto tamaño $k_1, k_2, k_3, \ldots, k_N$:

$$S = \sqrt{\frac{k_1 \cdot \text{Var}\, X_1 + k_2 \cdot \text{Var}\, X_2 + k_3 \cdot \text{Var}\, X_3 + \cdots + k_p \cdot \text{Var}\, X_N}{k_1 + k_2 + k_3 + \cdots + k_N}}$$

Observaciones sobre la desviación típica

1. La desviación típica, al igual que la media y la varianza, es un índice muy sensible a las puntuaciones extremas.
2. En los casos que no se pueda hallar la media tampoco será posible hallar la desviación típica.
3. Cuanta más pequeña sea la desviación típica mayor será la concentración de datos alrededor de la media.

El coeficiente de variación: es la relación entre la desviación típica de una muestra y su media; lo representaremos por CV y se calculará a partir de la siguiente expresión:

$$CV = \frac{\text{Desviación Típica}}{\text{Media Aritmética}} = \frac{S}{\overline{X}}$$

El coeficiente de variación se suele expresar en porcentajes:

$$CV = \frac{S}{\overline{X}} \cdot 100$$

El coeficiente de variación permite comparar las dispersiones de dos distribuciones distintas, siempre que sus medias sean positivas. Se calcula para cada una de las distribuciones y los valores que se obtienen se comparan entre sí. La mayor dispersión corresponderá al valor del coeficiente de variación mayor.

Ejemplo

Una distribución tiene x = 140 y S = 28 y otra x = 150 y S = 24 ¿Cuál de las dos presenta mayor dispersión?

Solución

$$CV_1 = \frac{S}{\bar{X}} \cdot 100 = \frac{28}{140} \cdot 100 = 20\%$$

$$CV_2 = \frac{S}{\bar{X}} \cdot 100 = \frac{24}{150} \cdot 100 = 16\%$$

La primera distribución presenta mayor dispersión.

Medidas de forma

Este tipo de medidas permite conocer la forma de la distribución sin necesidad de recurrir a su representación gráfica. Existen dos tipos de medidas de forma: asimetría y curtosis. Para clasificar la distribución según estas medidas, se establece en ambos casos una tipología de distribuciones.

Asimetría: es una medida que indica el grado de simetría (o asimetría) de una distribución respecto a su media. Es decir, la asimetría es un parámetro estadístico que sirve para determinar cuánto de simétrica (o asimétrica) es una distribución sin necesidad de representarla gráficamente.

Como eje de simetría consideramos una recta paralela al eje de ordenadas que pasa por la media de la distribución.

Así pues, una distribución asimétrica es aquella que tiene un número de valores a la izquierda de la media diferente de los que tiene a su derecha. En cambio, en una distribución simétrica hay el mismo número de valores a la izquierda y a la derecha de la media.

Se distinguen tres tipos de asimetría:

- **Asimetría positiva**: la distribución tiene más valores diferentes a la derecha de la media que a su izquierda.
- **Simetría**: la distribución tiene el mismo número de valores a la izquierda que a la derecha de la media.
- **Asimetría negativa**: la distribución tiene más valores diferentes a la izquierda de la media que a su derecha.

En la figura 5 se representa genéricamente los tipos de simetría (o asimetría) de una distribución respecto a su media.

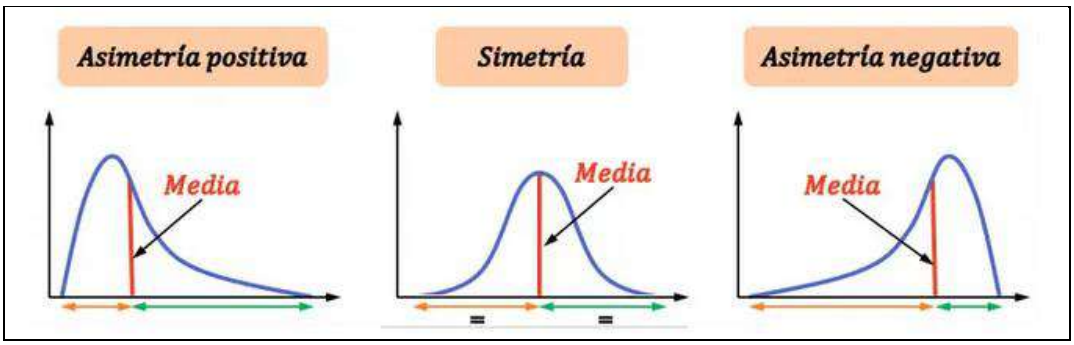


Figura 5. Tipos de simetría (o asimetría) de una distribución estadística.

Fuente: Asimetría y curtosis. ProbabilidadyEstadística.net.

https://www.probabilidadyestadistica.net/asimetria-y-curtosis/

La medida más popular de asimetría es el coeficiente de asimetría de Fisher, que para los datos agrupados viene dado por la siguiente expresión:

$$g_1 = \frac{m_3}{S^3} = \frac{\frac{1}{n}\sum_{i=1}^{C}(C_i - \bar{X})^3 f_i}{\left[\frac{1}{n}\sum_{i=1}^{C}(C_i - \bar{X})^2 f_i\right]^{\frac{3}{2}}}$$

Cuya interpretación es:

- Si $g_1 > 0$, la distribución es asimétrica positiva
- Si $g_1 = 0$, la distribución es simétrica.
- Si $g_1 < 0$, la distribución es asimétrica negativa.

Observación: el coeficiente de asimetría de Fisher es invariante frente a cambios de origen y de escala.

La curtosis: (también llamada *apuntamiento*), indica el grado de concentración de una distribución alrededor de su media.

Es decir, la curtosis muestra si una distribución es escarpada o achatada. En concreto, cuanto mayor sea la curtosis de una distribución más escarpada (o apuntada) es.

Hay tres tipos de curtosis:

- **Leptocúrtica**: la distribución es muy apuntada, es decir, los datos están muy concentrados alrededor de la media. En concreto, las distribuciones leptocúrticas se definen como aquellas distribuciones más apuntadas que la distribución normal.
- **Mesocúrtica**: la curtosis de la distribución es equivalente a la curtosis de la distribución normal. Por tanto, no se considera ni apuntada ni achatada.
- **Platicúrtica**: la distribución es muy achatada, es decir, la concentración en torno a la media es baja. Formalmente, las distribuciones platicúrticas se definen como aquellas distribuciones más achatadas que la distribución normal.

En la figura 6 se representa genéricamente los tipos de curtosis de una distribución respecto a su media.

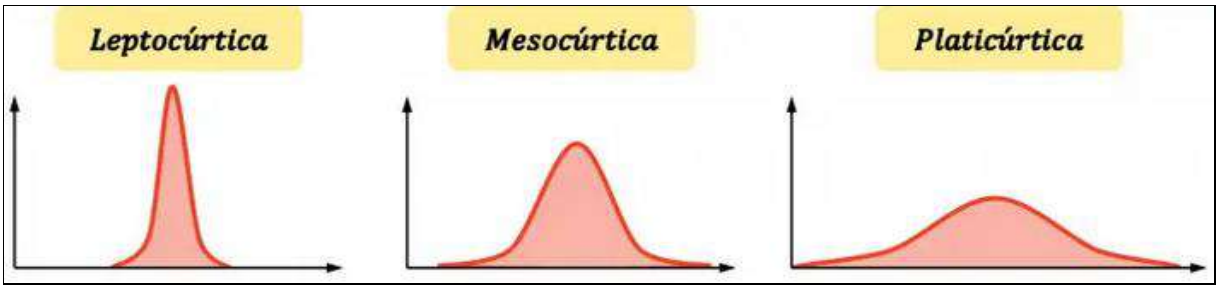


Figura 6. Tipos de curtosis de una distribución estadística.

Fuente: Asimetría y curtosis. ProbabilidadyEstadística.net.

https://www.probabilidadyestadistica.net/asimetria-y-curtosis/

La comparación se realiza respecto a una distribución «moderada» como es la distribución normal (mesocúrtica).

El coeficiente de curtosis viene dado por:

$$g_2 = \frac{m_4}{S^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{C}(C_i - \bar{X})^4 f_i}{\left[\frac{1}{n}\sum_{i=1}^{C}(C_i - \bar{X})^2 f_i\right]^2}$$

Que se interpreta del siguiente modo:

- Si $g_2 = 0$, la distribución es mesocúrtica o normal.
- Si $g_2 > 0$, la distribución es leptocúrtica o por encima de lo normal.
- Si $g_2 < 0$, la distribución es platicúrtica o por debajo de la normal.

Al igual que el coeficiente de asimetría de Fisher, el coeficiente de curtosis es invariante frente a cambios de origen y de escala.

Puntuaciones diferenciales y puntuaciones típicas

En ocasiones para llevar a cabo el análisis de los datos realizamos algunas transformaciones a los mismos que posibilitan extraer nuevas representaciones o conclusiones.

A partir de las puntuaciones diferenciales y las puntuaciones típicas pueden lograrse algunas transformaciones a los datos, lo cual facilita su análisis.

Puntuaciones diferenciales: resultan de restarles a las puntuaciones directas la media aritmética y se calculan a través de la siguiente expresión:

$$x_i = X_i - \bar{X}$$

Ejemplo

1) Encuentre las puntuaciones diferenciales a partir de la distribución: 9, 3, 8, 8, 9, 8, 9, 7, 10.

Solución:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{9+4+8+8+9+8+9+7+10}{9} = 8$$

Puntuaciones directas: X_i	Transformación $x_i = X_i - \overline{X}$	Puntuaciones diferenciales: x_i
9	9 – 8	1
5	5 – 8	-3
8	8 – 8	0
8	8 – 8	0
9	9 – 8	1
8	8 – 8	0
9	9 – 8	1
7	7 – 8	-1
9	9 – 8	1

Las puntuaciones típicas: son el resultado de dividir las puntuaciones diferenciales entre la desviación típica, a este proceso se llama tipificación. Las puntuaciones típicas se representan por z y se calculan a partir de la siguiente expresión

$$Z = \frac{X_i - \overline{X}}{S}$$

Observaciones sobre puntuaciones típicas

1. La media aritmética de las puntuaciones típicas es 0.
2. La desviación típica de las puntuaciones típicas es 1.
3. Las puntuaciones típicas son adimensionales, es decir, son independientes de las unidades utilizadas.
4. Las puntuaciones típicas se utilizan para comparar las puntuaciones obtenidas en distintas distribuciones.

Ejemplo

En una clase hay 15 alumnos y 20 alumnas. El peso medio de los alumnos es 58.2 kg y el de las alumnas y 54.4 kg. Las desviaciones típicas de los dos grupos son, respectivamente, 3.1 kg y 5.1 kg. El peso de José es de 70 kg y

el de Ana es 65 kg ¿Cuál de ellos puede, dentro del grupo de alumnos de su sexo, considerarse más grueso?

Solución:

$$Z_1 = \frac{X_i - \overline{X}}{S} = \frac{70 - 58.2}{3.1} = 3.806$$

$$Z_2 = \frac{X_i - \overline{X}}{S} = \frac{65 - 52.4}{5.1} = 2.471$$

José es más grueso respecto de su grupo que Pilar respecto al suyo.

Las gráficas estadísticas

La representación gráfica o visualización de los datos es importante para cualquier análisis de datos. Es una herramienta muy eficaz, ya que un buen gráfico capta la atención del lector, presenta la información de forma sencilla, clara y precisa, no induce a error y facilita la comparación de datos. Además, destaca las tendencias y las diferencias, y ayuda a ilustrar el mensaje, tema o trama del texto al que acompaña.

Debe señalarse que unos de los gráficos estadísticos de uso más frecuentes son los de barras, circulares, de caja, de bigote y de tallo y hojas. A continuación se brinda una explicación de sus características.

Gráfico de barras

Un gráfico de barras (*bar graph*) es una representación gráfica de los resultados de un análisis estadístico. El gráfico consta de barras para cada dato representado. Las anchuras de estas barras son iguales, pero las longitudes varían según la importancia del valor.

Estas barras se colocan generalmente en 2 ejes que pueden invertirse dependiendo de si se quiere hacer un gráfico de barras horizontal o vertical.

Características de un diagrama de barras:

- Se compone de columnas o barras de diferentes alturas, estas pueden ser horizontales o verticales.
- Tiene un eje horizontal o eje x, donde se ubica una variable.

- Tiene un eje vertical o eje y, donde se ponen los valores que determinan la altura de las barras. A estos números se les conoce como frecuencia.
- El ancho de las barras y el espacio entre cada una debe ser el mismo.
- Las barras también sirven para comparar valores.

Ejemplo

La siguiente distribución de datos representa la ganancia mensual expresada en miles de dólares de cinco empresas de Ecuador: 160, 20, 40, 120, 60. Represente mediante un gráfico de barra esta distribución de datos.

Solución

En este caso nos apoyaremos del software Microsoft Excel. Es por ello que la tabla está diseñada con la estructura que necesita este software para obtener el gráfico circular.

Empresas	Ganancia mensual (miles de dólares)
Empresa 1	160
Empresa 2	20
Empresa 3	40
Empresa 4	120
Empresa 5	60

Una vez seleccionado el tipo de gráfico de barra en este software, al tomar como base los datos de la tabla, se obtuvo la figura 7.

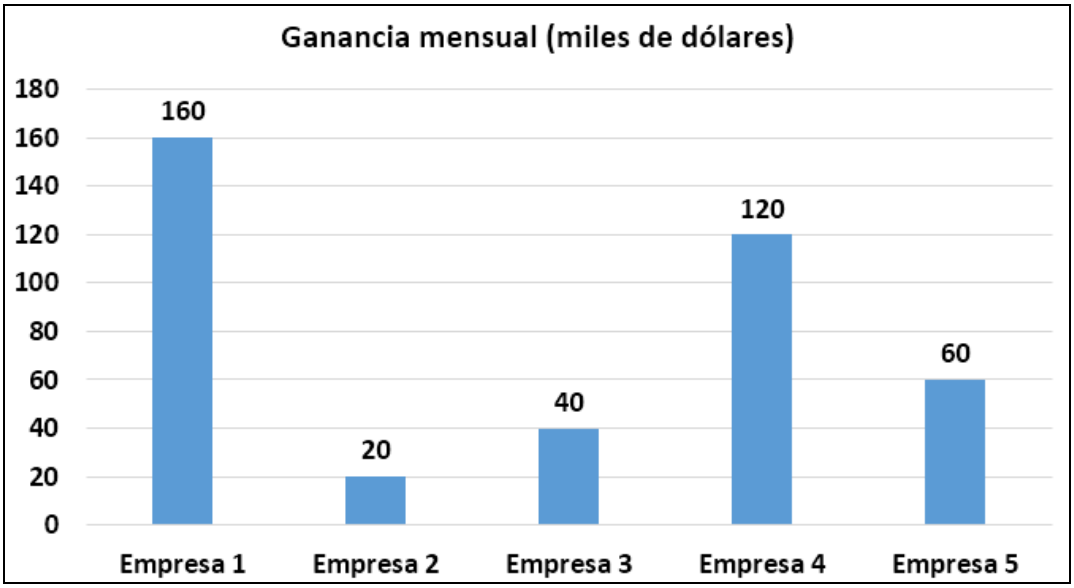


Figura 7. Ejemplo de gráfico de barras para una distribución de datos.

Otros ejemplos de gráficos de barra son los representados en las figuras 2, 3 y 4. En este caso tienen la peculiaridad de ser histogramas de frecuencia.

Gráfico circular

Una gráfica circular (*pie chart*), también llamada gráfica de pastel, es una forma de resumir un conjunto de datos nominales o de presentar los diferentes valores de una variable determinada, por ejemplo, una distribución porcentual. Los porcentajes se utilizan porque es más fácil representar diversos subconjuntos sobre determinada categorías respecto al conjunto total que representa el 100%.

Este tipo de gráfico está formado por un círculo dividido en sectores. Cada sector representa una categoría particular. El área de cada sector es la misma proporción del círculo que la categoría es del total de los datos.

Las gráficas circulares suelen mostrar una parte de un conjunto. A veces se separa una zona del resto del círculo para destacar la importancia de la información. Esto se llama un gráfico circular desglosado. Este es uno de los tipos de gráficos más populares para la visualización de datos.

Los gráficos circulares deben emplearse con cuidado por dos razones:

- Son útiles para presentar la información cuando sólo hay un máximo de cinco o seis elementos. Si hay más elementos, la figura creada será demasiado difícil de entender.
- Los gráficos circulares no son útiles cuando los valores de los componentes son demasiado similares porque puede ser difícil ver las diferencias de tamaño.

Ejemplo

Para la propia distribución de datos del ejemplo anterior, represente mediante un gráfico circular la misma.

Solución

En este caso nos apoyaremos también en las facilidades que ofrece el software Microsoft Excel. La tabla está diseñada con la estructura que necesita este software para obtener el gráfico circular.

Empresas	Ganancia mensual (miles de dólares)
Empresa 1	160
Empresa 2	20
Empresa 3	40
Empresa 4	120
Empresa 5	60

Una vez seleccionado el tipo de gráfico circular en este software, al tomar como base los datos de la tabla, se obtuvo la figura 8.

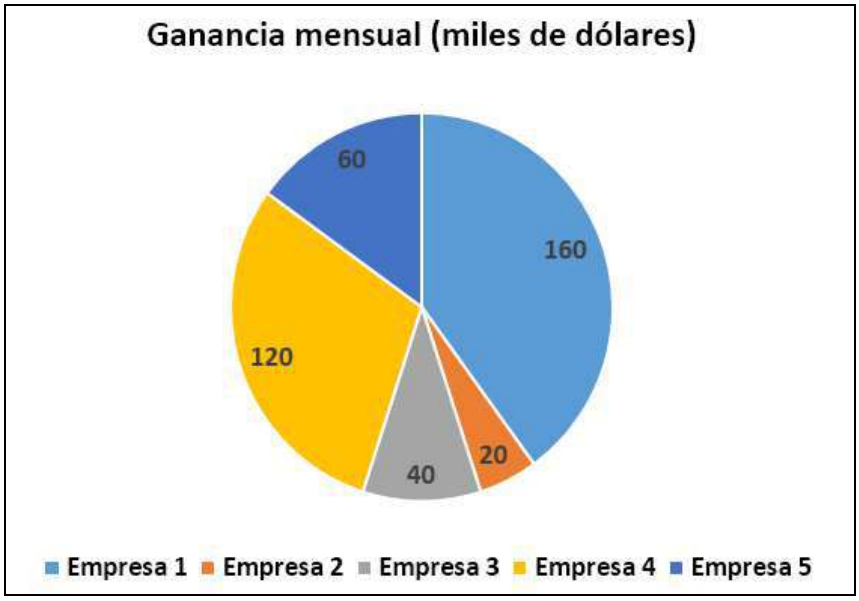


Figura 8. Ejemplo de gráfico circular para una distribución de datos.

Si quisiéramos comparar la ganancia de las empresas tomando como base la distribución porcentual, en el software Microsoft Excel solo habría que seleccionar esta opción en las opciones de etiqueta. Para este caso se diseñó el figura 9.

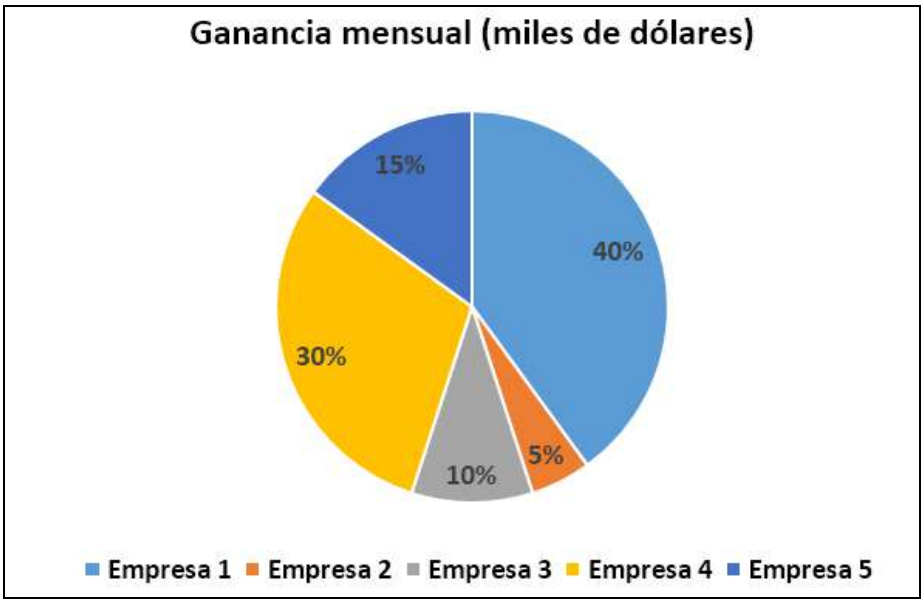


Figura 9. Ejemplo de gráfico circular para una distribución porcentual.

Gráfico de caja y bigotes

El gráfico de caja y bigotes (*box and whisker plot*) es una forma de presentación estadística destinada, fundamentalmente, a resaltar aspectos de la distribución de las observaciones en una o más series de datos cuantitativos. Reemplaza, en consecuencia, al histograma y a la curva de distribución de frecuencias sobre los que tiene ventajas en cuanto a la información que brinda y a la apreciación global que surge de la lectura. Fue ideado por John Tukey, de la Universidad de Princeton (U.S.A.) en 1977.

Este gráfico utiliza una sola escala: *la correspondiente a la variable de los datos que se presentan*. Es decir, no utiliza escala de frecuencias. Por lo tanto, no corresponde asociarlo a los que utilizan el sistema de coordenadas cartesianas. Los elementos que los constituyen son:

- **La caja**: es un rectángulo que abarca el recorrido (o rango, o intervalo) intercuartílico (RIC) de la distribución; o sea, el tramo de la escala que va desde el primer cuartil (Q_1) al tercer cuartil (Q_3). Esto incluye el 50 % de las observaciones centrales.
- **Mediana**: se dibuja mediante una línea (algunos lo marcan con un asterisco, otros con una cruz) dentro de la caja y a la altura de la escala que corresponde al valor de esa medida.
- **Bigotes**: son líneas que salen a los costados de la caja y que sirven como referencia para ubicar las observaciones que están por fuera del 50 % central de la distribución. (Para determinar su longitud: ver explicación más adelante).
- **Cercados interiores**: indica la finalización de los bigotes. A veces no se dibujan.
- **Cercados exteriores**: ubicados más periféricamente en la distribución. Casi nunca se dibujan.
- **Periféricos (o periféricos próximos)**: señalamiento de las observaciones que se encuentran entre el cercado interior y el cercado exterior.

- **Periféricos lejanos (o periféricos extremos)**: señalamiento de las observaciones que se encuentran fuera del cercado exterior.

Ejemplo

Dada la distribución de frecuencias de la edad de un colectivo de 20 personas.

36 25 37 24 39 20 36 45 31 31

39 24 29 23 41 40 33 24 34 40

Encuentre el gráfico de caja y bigotes.

Solución

Para calcular los parámetros estadístico, lo primero es ordenar la distribución:

20 23 24 24 24 25 29 31 31 33 34 36 36 37 39 39 40 40 41 45

Luego, se procede al cálculo de cuartiles:

Q_1, el cuartil primero es el valor mayor que el 25% de los valores de la distribución. Como N = 20 resulta que N/4 = 5; el primer cuartil es la media aritmética de dicho valor y el siguiente:

$Q_1=(24 + 25) / 2 = 24,5$

Q_2, el segundo cuartil es, evidentemente, la mediana de la distribución, es el valor de la variable que ocupa el lugar central en un conjunto de datos ordenados. Como N/2 =10; la mediana es la media aritmética de dicho valor y el siguiente:

$me = Q_2 = (33 + 34)/ 2 =33,5$

Q_3 , el tercer cuartil, es el valor que sobrepasa al 75% de los valores de la distribución. En nuestro caso, como 3N / 4 = 15, resulta

$Q_2=(39 + 39) / 2 = 39$

Luego se procede a dibujar el gráfico de caja y bigotes (ver figura 10).

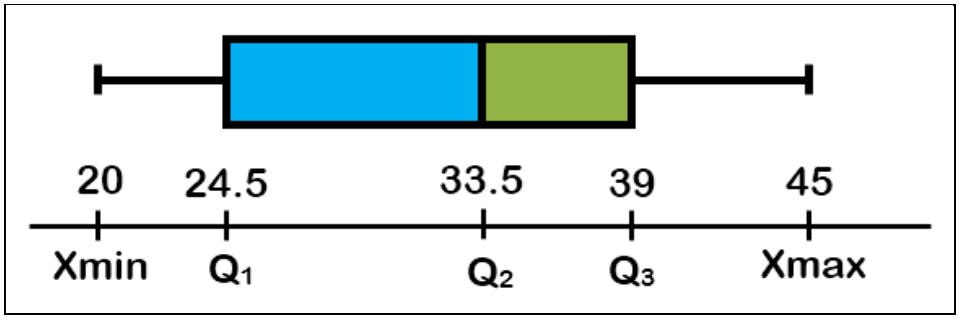


Figura 10. Gráfico de gráfico de caja y bigotes.

- El bigote de la izquierda representa al colectivo de edades (Xmin, Q_1).
- La primera parte de la caja a (Q_1, Q_2).
- La segunda parte de la caja a (Q_2, Q_3)
- El bigote de la derecha viene dado por (Q_3, Xmax).
- Puede observarse que los datos están algo más agrupados a la derecha de la mediana me = Q_2 = 33,5 que a su izquierda.

Existen varios softwares estadísticos que se pueden utilizar para diseñar un gráfico de caja y bigotes, entre ellos están el IBM SPSS, RStudio y el STATISTICA.

Gráfico de tallo y hojas

El gráfico de tallo y hojas (*stem-and-leaf graphic*) permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica. Para construirlo basta separar en cada dato el último dígito de la derecha (que constituye la hoja) del bloque de cifras restantes (que formará el tallo).

Esta representación de los datos es semejante a la de un histograma pero además de ser fáciles de elaborar, presentan más información que estos.

Ejemplo

Horarios de trenes

Basándome en un artículo de Juan C. Dürsteler en InfoVis.net, tomamos como ejemplo un horario de trenes confeccionado a partir de un díptico de la línea Castelldefels-Barcelona/Sants recogido en la estación de Renfe (Tabla 14). Originalmente el horario ocupa una tabla de 10 filas y 9 columnas más una columna «viuda» con el tren de las 22:38. Un total de 91 campos con formato hh.mm (hora:minuto) cada uno, 455 caracteres.

Tabla 14. Díptico original Trayecto Castelldefels -> Barcelona-Sants

5.03	7.32	9.02	11.07	13.32	15.07	16.50	18.32	20.07	22.38
6.02	7.37	9.07	11.32	13.37	15.20	17.02	18.37	20.20	
6.18	7.50	9.24	11.37	13.50	15.32	17.07	18.50	20.32	
6.37	8.02	9.32	12.02	14.02	15.37	17.20	19.02	20.37	
6.48	8.05	9.37	12.07	14.07	15.50	17.32	19.07	20.50	
6.55	8.20	10.02	12.32	14.20	16.02	17.37	19.20	21.02	
7.02	8.24	10.07	12.37	14.32	16.07	17.50	19.32	21.07	
7.07	8.32	10.32	13.02	14.37	16.20	18.02	19.37	21.20	
7.20	8.37	10.37	13.07	14.50	16.32	18.07	19.50	21.32	
7.25	8.51	11.02	13.20	15.02	16.37	18.20	20.02	21.37	

Fuente: Juan C. Dürsteler en InfoVis.net.

Solución

En el gráfico de tallo y hojas se representa la hora a la izquierda de la barra de separación | y los minutos de la salida de cada tren a la derecha (figura 11). La frecuencia de los trenes se deduce fácilmente de la longitud de las filas y es, además, muy fácil ver en qué minutos de cada hora pasan típicamente los mismos.

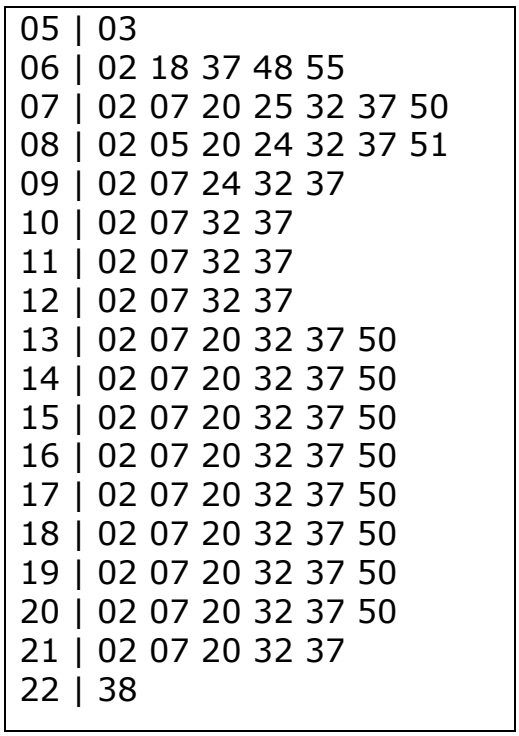


Figura 11. Gráfico de tallo y hojas basado en el horario de trenes. Fuente: Juan C. Dürsteler en InfoVis.net.

Por otra parte, dado que a algunas horas se repite exactamente el horario de los trenes se puede reducir aún más el tamaño del gráfico, sin perder información y ganando en claridad (ver figura 12).

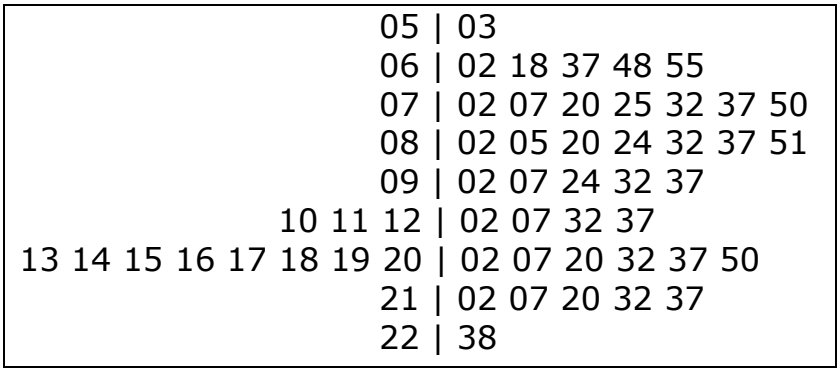


Figura 12. Gráfico de tallo y hojas reducido basado en el horario de trenes. Fuente: Juan C. Dürsteler en InfoVis.net.

CAPÍTULO II. PROBABILIDADES

Introducción

La teoría de la probabilidad se usa extensamente en áreas como la Estadística, la Física, la Matemática, las ciencias y la Filosofía para extraer conclusiones sobre la probabilidad discreta de sucesos potenciales y la mecánica subyacente discreta de sistemas complejos, por lo tanto, es la rama de las matemáticas que estudia, mide o determina a los experimentos o fenómenos aleatorios.

Se puede decir razonablemente que el descubrimiento de métodos rigurosos para calcular y combinar los cálculos de probabilidad ha tenido un profundo efecto en la sociedad moderna. Por consiguiente, puede ser de alguna importancia para la mayoría de los ciudadanos entender cómo se calculan los pronósticos y las probabilidades, y cómo contribuyen a la reputación y a las decisiones, especialmente en una democracia.

Otra aplicación significativa de la teoría de la probabilidad en el día a día es en la fiabilidad. Muchos bienes de consumo, como los automóviles y la electrónica de consumo, utilizan la teoría de la fiabilidad en el diseño del producto para reducir la probabilidad de avería. De forma general, elementos de esta teoría pueden utilizarse prácticamente para el estudio de cualquier fenómeno aleatorio.

Conceptos básicos y operaciones entre sucesos

La teoría de las probabilidades se originó fundamentalmente como una teoría de los juegos de azar, dado que aunque estos juegos eran conocidos desde la antigüedad, las probabilidades se originan cuando estos juegos se popularizaron.

Diversos hombres de ciencia lograron hacer importantes aportes a la teoría de las probabilidades en sus orígenes, entre ellos:

- Pascal (1623-1662)
- Fermat (1601-1665)
- Galileo (1564-1642)
- Huygers (1629-1695)
- Bernoulli (1654-1705)

Conceptos básicos

En los más variados campos del saber humano se presentan experimentos en los que es imposible conocer de antemano el resultado que se va a obtener. Por ejemplo:

a) El lanzamiento de un dado ordinario (no se conoce el número que se va a obtener en cada lanzamiento).
b) El lanzamiento de una moneda (no sabemos de antemano si va a salir cara o cruz).
c) El número de niños y niñas que nacerán a una hora determinada en cierta región (este número también es variable y desconocido).

Definición: en teoría de la probabilidad un ***experimento aleatorio*** (lo denotaremos por ε) es aquel que bajo el mismo conjunto aparente de condiciones iniciales, puede presentar resultados diferentes, es decir, no se puede predecir o reproducir el resultado exacto de cada experiencia particular.

Definición: si dado un experimento aleatorio ε, un evento A asociado a ese experimento puede ocurrir o no, se dirá que A es un ***evento aleatorio***. Se denotará con letras mayúsculas.

A continuación se presentan algunos ejemplos de experimentos y eventos aleatorios.

Ejemplos:

a) En el lanzamiento de una moneda consideramos el evento: obtener cara (puede ocurrir o no).
b) En el lanzamiento de un dado en el cual el evento aleatorio consiste en obtener el número 2 (puede ocurrir o no).

Otras definiciones y ejemplos relacionados con los experimentos y eventos aleatorios son las siguientes:

Definición: si dado un experimento aleatorio ε, un evento A asociado a este experimento siempre ocurre, se dirá que A es el ***evento cierto*** y será denotado por Ω.

Ejemplos:

a) En el lanzamiento de una moneda al considerar el evento: *obtener cara o cruz*.

b) En el lanzamiento de un dado al considerar el evento: obtener un número entre 1 y 6.

Definición: si dado un experimento aleatorio ε, un evento A asociado a este experimento nunca ocurre, se dirá que A es un ***evento imposible*** y se denotará por la letra V.

Ejemplos:

a) Si se examinan 20 alumnos y se considera el evento: que aprueben 22 alumnos.

b) Si en el lanzamiento de un dado se considera el evento: obtener un 8.

Definición: a cada posible resultado del experimento aleatorio se le llamará ***evento elemental***.

Definición: al conjunto de todos los eventos elementales asociados a un experimento aleatorio se le llamará ***espacio muestral*** (Ω).

Definición: a cada uno de los eventos del espacio muestral se le llamará ***evento simple***.

Definición: un ***evento compuesto*** se define como una colección de eventos simples.

Por lo tanto, los eventos compuestos se forman combinando dos o más eventos simples. A continuación se presenta un ejemplo ilustrativo.

Ejemplo: si se considera el lanzamiento de 3 monedas del mismo tipo al aire y el interés es obtener 2 caras (c_1).

Para este caso el espacio muestral, conformado por todos los resultados posibles del experimento aleatorio, es el siguiente:

$\Omega = \{(c_1, c_1, c_1), (c_1, c_1, c_2), (c_1, c_2, c_1), (c_2, c_1, c_1), (c_1, c_2, c_2), (c_2, c_1, c_2), (c_2, c_2, c_1), (c_2, c_2, c_2)\}$

El evento de interés en este ejemplo es un evento compuesto que está conformado por los 3 eventos simples (c_1, c_1, c_2), (c_1, c_2, c_1), (c_2, c_1, c_1).

Existen tres operaciones básicas con conjuntos mediante las cuales se pueden generar eventos compuestos, lo que presupone además del conocimiento de los tipos de relaciones que se pueden establecer entre conjuntos. A continuación se presentarán las operaciones y relaciones básicas.

Operaciones o relaciones entre eventos

Debe recordarse que un conjunto es una colección de objetos considerada como un objeto en sí. Un conjunto está definido únicamente por los elementos que lo componen y no por la manera en la que se lo representa.

Existe una serie de relaciones básicas entre conjuntos y sus elementos:

- **Pertenencia**: la relación relativa a conjuntos más básica es la relación de pertenencia. Dado un elemento x, éste puede o no pertenecer a un conjunto dado A. Esto se indica como $x \in A$.
- **Igualdad**: dos conjuntos son iguales si y sólo si tienen los mismos elementos.
- **Inclusión**. Dado un conjunto A, cualquier subcolección B de sus elementos es un subconjunto de A, y se indica como $B \subset A$.

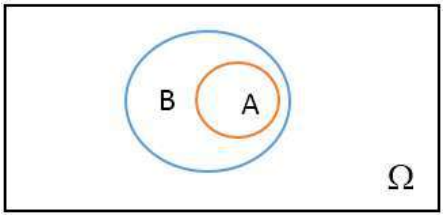
El conjunto vacío es el conjunto sin ningún elemento, y se denota por $\emptyset$. El conjunto universal es el conjunto que contiene todos los elementos posibles,

dentro del contexto considerado. Por ejemplo, si se estudian los números naturales, el conjunto universal es el conjunto de todos ellos, N. En este libro el conjunto universal se denota por el espacio muestral Ω.

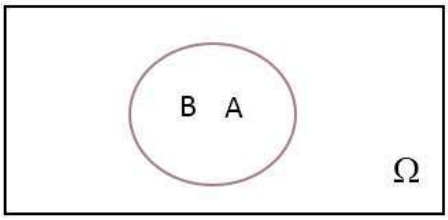
Una de las gráficas más utilizadas para representar eventos compuestos y que facilita la interpretación y el cálculo de probabilidades es el diagrama de Venn. En general, los diagramas de Venn son gráficas pictóricas en las que comúnmente se utiliza un rectángulo para representar al espacio muestral (Ω) y círculos dentro del rectángulo para representar los eventos. En ocasiones, los elementos de cada evento se insertan dentro del evento a que pertenecen y en otras ocasiones lo que se inserta dentro del evento es su probabilidad.

A continuación se presentan eventos compuestos como resultado de las operaciones con eventos simples o sus relaciones.

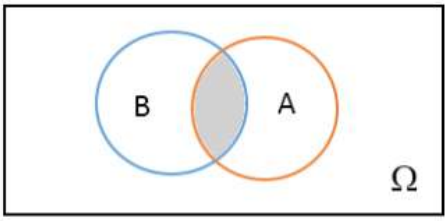
1) $A \Rightarrow B$ ó $A \subset B$: siempre que ocurre A ocurre B (implicación).



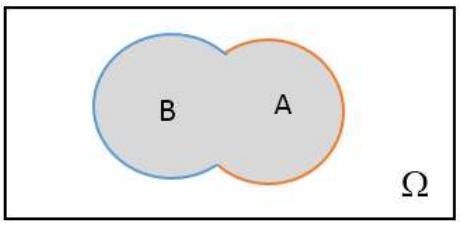
2) $A = B$ cuando $A \Rightarrow B$ y $B \Rightarrow A$ ó $A \subset B$ y $B \subset A$ (equivalencia).



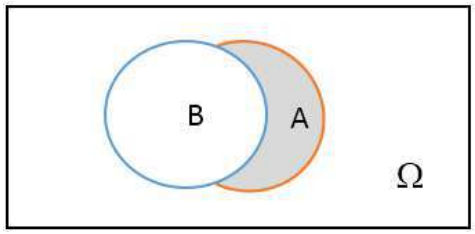
3) AB ó $A \cap B$: cuando A y B ocurren simultáneamente (producto).



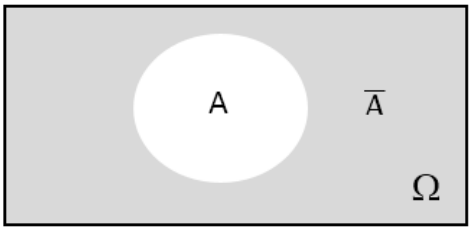
4) $A + B$ ó $A \cup B$: cuando ocurre al menos A o al menos B o ambos (suma)



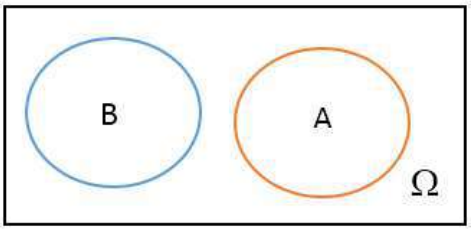
5) A - B: cuando ocurre A y no ocurre B (diferencia).



6) $\overline{A} = \Omega$ - A: evento complementario.



7) Eventos excluyentes: Si $A \cap B = \emptyset$



Definición: Una familia S de eventos forma un ***campo de eventos*** si se cumplen las siguientes condiciones:

1) Se contiene al evento cierto y al imposible.
2) Si $A \in S \Rightarrow \overline{A} \in S$
3) Si $A, B \in S \Rightarrow A+B \in S$

Ejemplo:

Si se considera el lanzamiento de una moneda 3 veces:

C_1: que salga una cara.

C_2: que salga cruz.

Solución:

El espacio muestral en este caso es:

$\Omega = \{(c_1, c_1, c_1), (c_1, c_1, c_2), (c_1, c_2, c_1), (c_2, c_1, c_1), (c_1, c_2, c_2), (c_2, c_1, c_2), (c_2, c_2, c_1), (c_2, c_2, c_2)\}$

Pueden definirse los siguientes eventos:

A: obtener un número impar de caras.

B: obtener una cara.

C: obtener al menos una cara.

D: obtener tres cruces.

Es fácil verificar que se cumplan las siguientes relaciones:

$B \subset A$ $A \subset C$ $A \cap C = A$ $A + B = A$ $A + C = C$

$\bar{C} = D$ $C \cap B = B$ $C + D = \Omega$ $B + C = C$

A continuación se proponen dos ejercicios para que los resuelva de forma independiente.

Ejercicios propuestos

1) Supongamos que se lanza una moneda 4 veces. Construya el espacio muestral y un campo de sucesos.

2) Sean los eventos:

A: obtener un número par de caras en los 4 lanzamientos.

B: obtener un número impar de caras en los 4 lanzamientos.

C: obtener una cara.

D: obtener 3 caras.

Determine los eventos $A + B$, $A \cap B$, $B + C + D$, $B \cap D$, $B \cap C$

Definición de probabilidad

La definición de probabilidad surge debido al deseo del ser humano por conocer con certeza los eventos que sucederán en el futuro. Es por eso que a través

de la historia se han desarrollado diferentes enfoques para tener un concepto de la probabilidad y determinar sus valores.

Por lo que la definición de probabilidad ha llevado un largo proceso histórico, pasando por diferentes etapas de su desarrollo y formación matemática. Actualmente, se ha logrado resumir en tres grupos fundamentales los diferentes enfoques de la definición de probabilidad:

- **Grupo 1**: los que consideran la definición de probabilidad como la medida cuantitativa del grado de certeza del observador.

  En este caso se asume un enfoque subjetivo, que define la probabilidad de un evento a base del grado de confianza que una persona tiene sobre la ocurrencia de un evento, al tener evidencia disponible que se fundamenta en la intuición, opiniones, creencias personales y otra información indirecta relevante.
- **Grupo 2**: los que reducen la definición de probabilidad a los eventos con igual probabilidad de ocurrencia, llamada Probabilidad Clásica.

  En este caso se asume un enfoque que se sustenta en la formulación clásica de la teoría de las probabilidades basado en el concepto de resultados igualmente verosímiles, que se ve limitado a situaciones en las que hay un número finito de resultados igualmente probables.
- **Grupo 3**: los que definen la probabilidad basándose en la frecuencia de ocurrencia del evento en la repetición de un gran número de experimentos, llamada Probabilidad Estadística.

  En este caso se asume un enfoque empírico, en el cual para determinar los valores de probabilidad se requiere de la observación y de la recopilación de datos. La definición empírica se basa en la frecuencia relativa de ocurrencia de un evento con respecto a un gran número de repeticiones del experimento aleatorio.

En este capítulo se estudiarán los 2 últimos grupos de enfoques sobre la definición de probabilidad, por ser más objetivos para fines prácticos y

coherentes con el proceso de investigación científica, a diferencia del primero que presenta un grado de subjetividad muy elevado.

Probabilidad clásica

A partir de la segunda mitad del siglo XVII, el desarrollo alcanzado en las ideas básicas de las probabilidades condujo a la formulación clásica de la Teoría de las Probabilidades.

Esta formulación dio lugar al concepto de probabilidad clásica basado en la concepción de resultados igualmente verosímiles, que se basa en el denominado *Principio de la Razón Insuficiente*, el cual postula que: si no existe un fundamento para preferir una entre varias posibilidades, todas deben ser consideradas equiprobables.

En consecuencia, en el lanzamiento de una moneda perfecta, la probabilidad de que se obtenga como resultado cara es igual a que se obtenga cruz, por tanto, ambas probabilidades son iguales a ½, o sea, 0.5.

De la misma manera, la probabilidad de cada uno de los seis sucesos elementales asociados al lanzamiento de un dado perfectamente balanceado debe ser 1/6.

Laplace recogió esta idea y formuló la regla clásica del cociente entre casos favorables y casos posibles, supuestos éstos igualmente verosímiles.

El enfoque clásico o «a priori» para definir la probabilidad es de uso limitado puesto que descansa sobre la base de las siguientes dos condiciones:

1. El espacio muestral (S) del experimento es finito (su número total de elementos es un número natural n = 1, 2, 3,...).
2. Los resultados del espacio muestral deben ser igualmente probables (tienen la misma posibilidad de ocurrir).

Bajo estas condiciones, suponga que se realiza un experimento aleatorio. El número total de elementos del espacio muestral del experimento es denotado como n(S). Dicho de otro modo, n(S) representa el número total de eventos simples distintos posibles al realizar un experimento. Además, si A es un

evento de este experimento, el número total de elementos del espacio muestral contenidos en A es denotado como n(A). Es decir, n(A) representa el número total de formas distintas en que A puede ocurrir.

En consecuencia, surge la definición clásica de probabilidad.

Definición: en el enfoque clásico la probabilidad de que el evento A ocurra se define como:

$$P(A) = \frac{n(A)}{n(S)} = \frac{\text{número de formas distintas en que A puede ocurrir}}{\text{número total de eventos simples distintos posibles}}$$

Ejemplos:

1) Al lanzar un dado al azar, ¿cuál es la probabilidad de obtener un número par?

Solución: Suponga que A es el evento de obtener un número par al lanzar un dado al azar. Notemos que S = {1, 2, 3, 4, 5, 6} y todos los resultados igualmente probables. Además, A puede ocurrir de tres formas distintas (2, 4 ó 6).

Por lo tanto, $n(A) = 6$ y $n(S) = 3$, entonces:

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2} = 0.5$$

2) ¿Cuál es la probabilidad de que en una familia que tiene tres hijos, haya dos niñas y un niño, si se considera igualmente probable el nacimiento de un niño o niña?

Solución: Usando "a" para niña y "o" para niño, el espacio muestral es:

S = {aaa, aao, aoa, aoo, oaa, oao, ooa, ooo} por lo que n(S) = 8.

Se define el evento A como que haya dos niñas y un niño, entonces:

A = {aao, aoa, oaa} y n(A) = 3

Por lo tanto: $P(A) = \frac{n(A)}{n(S)} = \frac{3}{8} = 0.375$

3) Supongamos que se lanzan 2 dados y se quiere hallar la probabilidad de que:

a) Cada uno de los eventos simples.

b) La suma de los resultados sea igual a 5.

Solución:

a) El espacio muestral está constituido por los siguientes elementos:

(1,1) (2,1) (3,1) (4,1) (5,1) (6,1)

(1,2) (2,2) (3,2) (4,2) (5,2) (6,2)

(1,3) (2,3) (3,3) (4,3) (5,3) (6,3)

(1,4) (2,4) (3,4) (4,4) (5,4) (6,4)

(1,5) (2,5) (3,5) (4,5) (5,5) (6,5)

(1,6) (2,6) (3,6) (4,6) (5,6) (6,6)

El primer número de cada par denota el número que salió en el primer dado y el segundo número representa el que salió en el segundo dado.

El total de casos es igual a 36 (n(S)). Entonces un evento simple A cualquiera según la definición anterior tiene una probabilidad de salir cada vez que se tiren los dados igual a $\mathrm{P(A)} = \frac{\mathrm{n(A)}}{\mathrm{n(S)}} = \frac{1}{36} = 0.278$ (evento equiprobable).

b) Sea A el evento en que consiste en que la suma de los resultados sea 5, entonces los resultados favorables al evento son los pares (1,4), (2,3), (3,2), (4,1), es decir en total 4. Luego $P(A) = \frac{4}{36} = \frac{1}{9}$

Observación: para todo lanzamiento de dados el total de casos posibles es

a) 1 dado------6^1

b) 2 dados-----6^2

c) 3 dados-----6^3

d) n dados-----6^n

Definición: la probabilidad de que un evento compuesto A ocurra, es la suma de las probabilidades de los eventos simples de los cuales está compuesto A.

Ejemplos:

a) Retomando el ejemplo del experimento aleatorio de obtener 2 caras al lanzar 3 monedas al aire. Si este evento compuesto A estará conformado por los eventos simples:

(c_1, c_1, c_2), (c_1, c_2, c_1), (c_2, c_1, c_1).

Al aplicar la definición anterior la probabilidad de obtener 2 caras es:

$$P(A) = P(c_1, c_1, c_2) + P(c_1, c_2, c_1) + P(c_2, c_1, c_1) = \frac{3}{8} = 0.375$$

b) En el espacio muestral de dígitos conformados por los números del 0 al 9, si B es el evento de obtener un dígito menor que 4 al seleccionar un dígito al azar, entonces se tiene que:

$$P(B) = P(0) + P(1) + P(2) + P(3) = \frac{4}{10} = \frac{2}{5} = 0.4$$

c) Sea el experimento aleatorio que consiste en realizar dos extracciones al azar de una caja que contiene 3 bolas rojas, 2 negras y 1 verde. Si se desea calcular la probabilidad del evento compuesto C que consiste en obtener una bola roja y una bola verde, entonces:

$$P(C) = P(r) + P(v) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} = 0.666$$

Regla de la Suma

La regla de la suma es un método para calcular probabilidades que pueden expresarse de la forma P(A + B) o $P(A \cup B)$, es decir, la probabilidad de que A ocurra o de que ocurra B (o de que ambos ocurran). Para calcular la probabilidad de que ocurra el evento A o el evento B, primero se debe determinar de cuántas formas distintas A ocurre y de cuántas formas distintas B ocurre, pero al computar el número total de formas no se puede contar un mismo resultado más de una vez.

Regla de la Suma

Sean A y B dos eventos de un mismo espacio muestral Ω, entonces

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

Donde $P(A \cap B)$ denota la probabilidad de que A y B ocurran al mismo tiempo.

Ejemplo:

A un grupo docente de 40 estudiantes se le administra un examen para cualificarlos en la asignatura de Estadística. La siguiente tabla resume los resultados divididos por género:

	Masculino (M)	Femenino (F)
Aprobado (A)	7	2
Desaprobado (D)	18	13

Si un estudiante del grupo se selecciona al azar, halle la probabilidad de que:

a) Sea masculino o aprobó el examen.

b) Sea femenino o fracasó.

Solución:

Al utilizar la regla de la suma se tiene que:

a) $P(M+A) = P(M) + P(A) - P(M \cap A)$

$$P(M+A) = \frac{25}{40} + \frac{9}{40} - \frac{7}{40} = \frac{27}{40} = 0.675$$

b) $P(F+D) = P(F) + P(D) - P(M \cap D)$

$$P(F+D) = \frac{15}{40} + \frac{31}{40} - \frac{13}{40} = \frac{33}{40} = 0.825$$

Cuando fueron presentadas las operaciones o relaciones entre eventos, se analizó la propiedad que tienen los eventos aleatorios mutuamente excluyentes, en los cuales la ocurrencia de uno evita la ocurrencia del otro, o sea, su intercepción es vacía.

Al tener en cuenta esta propiedad de los eventos mutuamente excluyentes entonces es cobra sentido la siguiente definición.

Definición: Sean A y B dos eventos de un mismo espacio muestral Ω. Decimos que A y B son eventos mutuamente excluyentes si no pueden ocurrir simultáneamente, es decir, $A \cap B = \emptyset$. Por lo tanto, si A y B son eventos mutuamente excluyentes, entonces $P(A \cap B) = 0$

Ejemplo:

En el experimento aleatorio de lanzar un dado al azar, el espacio muestral es $\Omega = \{1, 2, 3, 4, 5, 6\}$. Si A es el evento de obtener un número par y B es el evento de obtener un número impar, entonces A y B son eventos mutuamente excluyentes pues no pueden ocurrir simultáneamente (si uno ocurre el otro no puede ocurrir).

La regla de la suma ya presentada establece que:

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

Sin embargo, si A y B son eventos mutuamente excluyentes, entonces la regla de la suma se reduce a

$P(A + B) = P(A) + P(B)$, ya que $P(A \cap B) = 0$.

Claramente un evento cualquiera A y su complemento $\overline{A}$ son mutuamente excluyentes. Además, en todo experimento uno de ellos ocurre. Esto es debido que el evento A ocurre o no ocurre (lo que implica que complementariamente $\overline{A}$ no ocurra o ocurra). Entonces se tiene que:

$P(A + \overline{A}) = P(A) + P(\overline{A}) = 1$

Este resultado de la regla de la suma da lugar a las siguientes tres formas equivalentes:

1. $P(A) + P(\overline{A}) = 1$
2. $P(A) = 1 - P(\overline{A})$
3. $P(\overline{A}) = 1 - P(A)$

Ejemplo:

¿Cuál es la probabilidad de obtener un total de por lo menos 10 puntos al tirar dos dados?

Solución:

Sean A, B y C los eventos de obtener exactamente un total de 10 puntos, 11 puntos y 12 puntos, respectivamente.

(1,1) (2,1) (3,1) (4,1) (5,1) (6,1)
(1,2) (2,2) (3,2) (4,2) (5,2) (6,2)
(1,3) (2,3) (3,3) (4,3) (5,3) (6,3)
(1,4) (2,4) (3,4) (4,4) (5,4) (6,4)
(1,5) (2,5) (3,5) (4,5) (5,5) (6,5)
(1,6) (2,6) (3,6) (4,6) (5,6) (6,6)

El 1er número de cada par denota el número que salió en el primer dado y el segundo número representa la cantidad de puntos que salieron en el segundo. En la tabla puede verse con claridad que dichos eventos no tienen puntos en común y que sus probabilidades están dadas por P(A)=3/36, P(B)=2/36 y P(C)=1/36.

Al determinar la probabilidad de que por lo menos ocurra uno de dichos eventos mutuamente excluyentes se obtiene:

$$P(A + B + C) = P(A) + P(B) + P(C) = \frac{3}{36} = \frac{2}{36} = \frac{1}{36} = \frac{6}{36} = \frac{1}{6} = 0.166$$

Regla de la probabilidad condicional

Ahora consideraremos el principio de que la probabilidad de un evento suele afectarse por el conocimiento previo de las circunstancias. Por ejemplo, si se selecciona a un estudiante al azar de la población de una universidad, la probabilidad de obtener un varón podría suponerse cercana a 0.5 pero si usted ya sabe que además el estudiante a seleccionar practica fútbol, entonces la

probabilidad de que sea varón aumenta drásticamente, puesto que la gran mayoría de los estudiantes que practican este deporte son varones.

Definición:

La probabilidad condicional de un evento B es la probabilidad de que B ocurra cuando ya se sabe que otro evento A ocurrió. Esta probabilidad se denota por P(B | A) y se lee «la probabilidad de B dado A».

Para ilustrar la definición de probabilidad condicional de un evento B a continuación se mostrará un ejemplo.

Ejemplo: Al lanzar un dado, halle la probabilidad de obtener:

a) El número 4 dado que se obtuvo un número par.

b) Un número impar dado que se obtuvo un número menor que 6.

Solución:

Al utilizar la definición de probabilidad condicional, se tiene que:

a) Se sabe que se obtuvo un número par, por lo que hay 3 posibles resultados: 2, 4 ó 6, de los cuales en sólo uno de ellos ocurre el 4.

Por lo tanto, P(4 | par) = 1/3 = 0.333

b) Se sabe que se obtuvo un número menor que 6, por lo que hay 5 posibles resultados: 1, 2, 3, 4 ó 5, de los cuales en tres de ellos obtenemos un número impar.

Por lo tanto, P(impar | menor que 6) = 3/5 = 0.6

Regla de probabilidad condicional:

Sean A y B dos eventos de un mismo espacio muestral Ω, entonces:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{donde} \quad P(B) > 0$$

y $$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad \text{donde} \quad P(A) > 0$$

Ejemplo:

¿Cuál es la probabilidad de que una carta de póker escogida al azar de un paquete completo de cartas sea un as, sabiendo que, la carta es roja?

Solución:

Las cartas de póker son 52. Estas se componen de 26 cartas rojas y 26 cartas negras. Además, las 26 cartas rojas se dividen en 13 cartas de corazones y 13 de diamantes. Las 26 cartas negras se dividen en 13 cartas de espada y 13 de trébol. Cada uno de los grupos de 13 cartas tiene una carta de cada uno de los siguientes caracteres A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K.

Entonces

$$P(\text{as} \mid \text{roja}) = \frac{P(\text{as} \cap \text{roja})}{P(\text{roja})} = \frac{\frac{2}{52}}{\frac{26}{52}} = \frac{1}{13} = 0.077$$

Regla de la multiplicación

La regla de la suma presentada se utiliza para calcular P(A + B) o P(A ∪ B). Ahora se presentará la regla de la multiplicación, la cual provee de un método para calcular probabilidades que pueden expresarse de la forma P(A ∩ B). Es decir, la probabilidad de que A ocurra y de que ocurra B al mismo tiempo.

Para calcular la probabilidad de que ocurra el evento A y el evento B al mismo tiempo, primero se debe determinar de cuántas formas distintas A ocurre y luego determinar de cuántas formas distintas B ocurre, dado que ya ocurrió A. Otra forma de hacerlo es, primero determinar de cuántas formas distintas B ocurre y luego determinar de cuántas formas distintas A ocurre, dado que ya ocurrió B.

Regla de la Multiplicación

Sean A y B dos eventos de un mismo espacio muestral Ω, entonces:

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

o $$P(A \cap B) = P(B) \cdot P(A|B)$$

Esta regla de multiplicación ayuda a calcular probabilidades cuando existe relación entre dos eventos aleatorios. A continuación se presenta un ejemplo que ilustra su aplicación.

Ejemplo: En un grupo de 25 personas hay 16 de ellas casadas y 9 solteras. ¿Cuál es la probabilidad de que si dos de estas personas son seleccionadas aleatoriamente sean ambas casadas?

Solución:

Nótese que al seleccionar la primera persona, la probabilidad de que sea casada es 16/25. Pero al seleccionar la segunda persona quedan 24 personas posibles ya que una ha sido seleccionada y el muestreo es sin remplazo.

En consecuencia, dado que la primera persona seleccionada es casada (evento A), quedan 15 casadas posibles para la segunda selección. Por lo tanto, la probabilidad de que la segunda persona seleccionada sea casada (evento B) dado que la primera fue casada, es igual a 15/24.

Luego, al utilizar la regla de la multiplicación, se tiene que:

$$P(A \cap B) = P(A) \cdot P(B|A) = \frac{16}{25} \cdot \frac{15}{24} = 0.4$$

Otro teorema que puede ayudar a determinar la probabilidad de un evento aleatorio que está relacionado con una serie de eventos mutuamente excluyentes entre sí, es el de la probabilidad total, el cual se presenta a continuación.

Teorema de la probabilidad total

Dado un experimento aleatorio ε y eventos asociados B, A_1, A_2,..., A_n. Si el evento B ocurre con uno y solamente uno de los eventos A_1, A_2,..., A_n mutuamente excluyentes, entonces la probabilidad de ocurrencia del evento B puede calcularse a partir de la siguiente expresión

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

A continuación se ilustrara la utilización de este teorema a través de un ejemplo que permite valorar su utilidad práctica.

Ejemplo:

Se tiene tres urnas con bolas. La primera contiene cuatro bolas blancas y tres azules; la segunda contiene cinco bolas blancas y tres azules; la tercera contiene tres bolas blancas, cuatro rojas y una azul. Se elige una urna al azar y se extrae una bola. Hallar la probabilidad de que la bola extraída sea blanca.

Solución:

Se definen los eventos aleatorios:

B: que la bola extraída sea blanca.

A_1: que la urna seleccionada sea la primera.

A_2: que la urna seleccionada sea la segunda.

A_3: que la urna seleccionada sea la tercera.

Como el evento B ocurre con uno y solo uno de los eventos mutuamente excluyentes A_1, A_2 y A_3, entonces se puede aplicar el teorema de probabilidad total, obteniéndose:

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_2)P(B|A_2)$$

$$P(B) = \frac{1}{3}\cdot\frac{4}{7} + \frac{1}{3}\cdot\frac{5}{8} + \frac{1}{3}\cdot\frac{3}{8} = \frac{11}{21} = 0.524$$

Supóngase en este ejemplo que se extrae la bola de la urna seleccionada al azar y resulta blanca, y se quiere conocer la probabilidad de que la bola haya sido extraída de la primera urna, o sea, la probabilidad de interés es $P(A_1|\,B)$.

Esta probabilidad se puede calcular fácilmente si se tiene en cuenta el teorema de la multiplicación:

$P(A_1 \cap B) = P(B) \cdot P(A_1|B)$

Al despejar $P(A_1|B)$ en la expresión anterior se obtiene

$$P(A_1|B) = \frac{P(A_1)\,P(B \mid A_1)}{P(B)}$$

Como se está trabajando bajo la suposición de que se cumplen las condiciones para aplicar el teorema de la probabilidad total, entonces

$$P(A_1|B) = \frac{P(A_1)\,P(B \mid A_1)}{\sum_{i=1}^{3} P(A_i)P(B|A_i)} = \frac{P(A_1)\,P(B \mid A_1)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+P(A_3)P(B|A_3)} = \frac{4}{11}$$

Teorema de Bayes

Si se tiene un evento B que ocurre con uno y solo uno de los n eventos, A_1, A_2,..., A_n, entonces se pueden calcular las probabilidades de estos n eventos después que se conoce la ocurrencia del evento B, a través de la siguiente expresión

$$P(A_i|B) = \frac{P(A_i)\,P(B \mid A_i)}{\sum_{j=1}^{n} P(A_j)P(B|A_j)}$$

Ejemplo

Supóngase que una primera caja contiene dos bolas rojas y una blanca y la segunda caja contiene dos bolas rojas y dos blancas. Se selecciona al azar una caja y se extrae una bola de ella. Si la bola extraída es roja, ¿cuál es la probabilidad de que provenga de la primera caja?

Solución

Si se supone que A_1 es el evento de seleccionar la primera caja y A_2 el evento de coger la 2da caja, donde $P(A_1) = P(A_2) = \frac{1}{2}$. Además, el evento B se define como haber obtenido una bola roja, entonces:

$$P(B|A_1) = \frac{2}{3} \quad \text{y} \quad P(B|A_2) = \frac{2}{4} = \frac{1}{2}$$

A sustituir en la fórmula de Bayes, se tiene que

$$P(A_1|\,B) = \frac{P(A_1)\,P(B\,|A_1)}{P(A_1)\,P(B\,|A_1) + P(A_2)\,P(B\,|A_2)} = \frac{\frac{1}{3}}{\frac{1}{2}\cdot\frac{2}{3}+\frac{1}{2}\cdot\frac{2}{4}} = \frac{\frac{1}{3}}{\frac{7}{12}} = \frac{4}{7} = 0.571$$

$P(A_1|\,B) = 0.571$ Probabilidad de que provenga de la primera caja.

Propiedades de la definición clásica de probabilidad

Si se supone que el espacio muestral Ω está constituida por n eventos mutuamente excluyentes e igualmente probables que son todos los resultados posibles del experimento A.

1) Para cada $A \subset \Omega$ se tiene que $P(A) \geq 0$.

   Demostración: $P(A) = m/n \geq 0$.

2) Para el evento cierto Ω se tiene que $P(\Omega) = 1$

   Demostración: $P(\Omega) = n/n = 1$

3) Si el evento A se puede expresar como la suma de 2 eventos B y C mutuamente excluyentes, entonces P(A)=P(B)+P(C).

   Demostración:

   Sean m: número de eventos para A.

   m_1: número de eventos para B.

   m_2: número de eventos para C.

   Como B y C son excluyentes, $m = m_1 + m_2$

   Luego $P(A) = m/n = m_1/n + m_2/n = P(B) + P(C)$.

4) $P(A)=1 - P(A^c)$

Demostración: $\Omega = A + A^c$ aplicando las propiedades (2) y (3) se tiene que:

$P(\Omega) = 1 = P(A) + P(A^c) \Rightarrow P(A) = 1 - P(A^c)$

5) La probabilidad de ϕ es cero.

Demostración: $\phi + \Omega = \Omega$

Como $\phi \cap \Omega = \phi \Rightarrow P(\phi) + P(\Omega) = P(\Omega) \Rightarrow P(\phi) = 0$

6) Si $A \subset B \Rightarrow P(A) \leq P(B)$

Demostración: B se puede escribir como la suma de A y $A^c \cap B$.

$B = A + A^c \cap B$

Por las propiedades (1) y (3) tenemos que

$P(B) = P(A) + P(A^c \cap B) \geq P(A)$.

7) Para cualquier evento A: $0 \leq P(A) \leq 1$

Demostración: $\phi \subset A \subset \Omega$.

Aplicando (6) se tiene que $P(\phi) = 0 \leq P(A) \leq 1 = P(\Omega)$

Muestreo con remplazo

Para el cálculo de las probabilidades de ciertos eventos se hace imprescindible conocer algunas formas especiales de selección de diferentes objetos en diferentes experimentos aleatorios.

Una de estas formas especiales de seleccionar los objetos es a través del esquema que brinda el muestreo con remplazo.

Definición: el ***muestreo con remplazo*** en una población constituida por N objetos es aquel procedimiento que tiene como objetivo formar muestras ordenadas de n objetos mediante el siguiente procedimiento:
Se extrae un objeto al azar de los N objetos que conforman la población, se anota el resultado y se devuelve el objeto a la población, se realiza otra extracción de la población y se anota el segundo resultado a continuación del

primero, y se devuelve el objeto a la población. Este procedimiento se continúa hasta que se tienen n objetos ordenados que han sido extraídos de la población de N objetos.

Observación: el número total de muestras que se pueden formar es N^n y la probabilidad de seleccionar una muestra ordenada de n elementos con remplazo cuando se tienen N elementos en la población es igual a $1/N^n$.

Ejemplo:

Si se lanzan tres 3 monedas y se desea determinar la probabilidad del evento A, que consiste en obtener 2 caras, entonces: N=2 y n=3 (población de tamaño 2 con 3 extracciones).

- El número de casos o muestras posibles a formar de tamaño 3 es 2^3.
- El número de casos o muestras favorables donde aparecen 2 veces las caras es: $\binom{3}{2} = \frac{3!}{2!1!} = \frac{3.2.1}{2.1.1} = 3$

Luego la probabilidad de obtener 2 caras del mismo tipo es

$$P(A) = \frac{3}{2^3} = \frac{3}{8} = 0.375$$

Ahora bien, si se tiene una población de N objetos que se puede dividir en dos clases mutuamente excluyentes de N_1 objetos del tipo I y N_2 objetos del tipo II. Si se toma una muestra aleatoria de tamaño n con remplazo es posible calcular la probabilidad de que hallan exactamente k objetos de tipo I.

Para calcular esta probabilidad se definirá primeramente:

- N^n: número total de casos posibles o muestras de tamaño n.
- ${N_1}^k$: número de casos favorables de objetos del tipo I.
- ${N_2}^{n-k}$: número de casos favorables de objetos del tipo II.
- ${N_1}^k \cdot {N_2}^{n-k}$: total de muestras.

Como existen varias posibilidades de que los k objetos del tipo I puedan variar en los n lugares de la muestra, entonces esto se pueden determinar el número

de estas posibilidades a partir de las combinaciones de n elementos tomados k, de modo que la probabilidad que se desea calcular resulta:

$$P = \binom{n}{k}\left(\frac{N_1}{N}\right)^k\left(\frac{N_2}{N}\right)^{n-k} = \binom{n}{k}\left(\frac{N_1}{N}\right)^k\left(\frac{N-N_1}{N}\right)^{n-k}$$

Donde $\binom{n}{k} = \frac{n!}{(n-k)!k!}$

Ejemplo:

Considere una urna con 20 bolas, 12 de las cuales son azules y 8 son rojas. Si se seleccionan 5 bolas al azar mediante un muestreo con remplazo, o sea, en cada extracción se anota el color de la bola y se devuelve a la urna ¿Cuál es la probabilidad de seleccionar 2 bolas azules y 3 rojas?

Solución:

N=20
n = 5
N_1 = 12
N_2 = 8
k = 2 (seleccionar dos bolas azules)
n – k = 3 (seleccionar tres bolas rojas)

Sea A el evento de que al seleccionar cinco bolas dos sean azules y tres rojas, entonces:

$$P(A) = \binom{n}{k}\left(\frac{N_1}{N}\right)^k\left(\frac{N_2}{N}\right)^{n-k} = \binom{n}{k}\left(\frac{N_1}{N}\right)^k\left(\frac{N-N_1}{N}\right)^{n-k}$$

$$P(A) = \binom{5}{2}\left(\frac{12}{20}\right)^2\left(\frac{8}{20}\right)^3 = 10 \cdot 0.36 \cdot 0.064 = 0.2304 \quad \text{Donde } \binom{5}{2} = \frac{5!}{(5-2)!2!} = \frac{5.4.3!}{3!.2} = 10$$

$$P(A) = 0.2304$$

Muestreo sin remplazo

Supongamos que tenemos una población de N objetos y queremos formar muestras de tamaño n (n ≤ N). En este caso los elementos que se extraen de

la población de N elementos no se devuelven a ella, es decir, no se remplazan. En el muestreo sin remplazo el tamaño de la muestra tiene que ser obligatoriamente menor o igual que el tamaño de la población y las muestras pueden extraerse en forma ordenada según el orden de selección o sin considerar el orden de selección.

Además, en el muestreo sin remplazo se puede tener en cuenta que la muestra sea sin ordenar u ordenada, o sea, que no se tiene en cuenta el orden de selección o puede tenerse en cuenta.

- *Muestras sin ordenar*

Todas las muestras de tamaño n que se pueden formar de una población de N elementos con N ≥ n, sin remplazar los elementos y sin tener en cuenta el orden de selección, son los subconjuntos de n elementos tomados n a n de un conjunto de N elementos, que son las *combinaciones* $\binom{N}{n}$, luego la probabilidad de seleccionar una muestra particular de tamaño n sin remplazo de la población de tamaño N y sin tener en cuenta el orden de selección es $P = \frac{(N-n)! \cdot n!}{N!}$

- *Muestras ordenadas*

Supongamos ahora que el muestreo se realiza sin remplazo y ordenando la muestra de tamaño n. Todas la muestras de tamaño n que se pueden formar de una población de tamaño N, sin remplazar los elementos de la población y teniendo en cuenta al orden de selección de los elementos, es lo mismo que si consideramos todos los grupos de tamaño n, tomados n a n de una población de N elementos, donde los grupos de n elementos se diferencian en el orden en que aparecen y en los elementos que son diferentes y que es igual a las *variaciones* de N elementos tomados n a n que se pueden escribir como $\binom{N}{n} \cdot n!$

Recordar que:

a) n! define el número de permutaciones sin repetición de n elementos.

b) $\binom{m}{n}$ combinaciones sin repetición de m objetos tomados de n en n.

Ejemplo

Si se tiene una población compuesta por los números 1, 2, 3 y 4 y se quiere formar muestras de tamaño 2.

a) Si las muestras son con remplazo y teniendo en cuenta el orden se obtiene que N = 4 y n = 2, por lo tanto: $N^n = 4^2 = 16$.

(1,1)	(2,1)	(3,1)	(4,1)
(1,2)	(2,2)	(3,2)	(4,2)
(1,3)	(2,3)	(3,3)	(4,3)
(1,4)	(2,4)	(3,4)	(4,4)

Cada par de la tabla denota el primer y segundo números seleccionado al azar, en ambos casos pueden seleccionarse uno de los cuatro números porque el muestreo es con remplazo.

b) Si las muestras son sin remplazo:

- Al no tener en cuenta el orden de selección, el número de muestra será:

$\binom{4}{2} = 6$ Posibilidades, estas son:

(1,2), (1,3), (1,4), (2,3), (2,4), (3,4)

- Al tener en cuenta el orden de selección, el número de muestras total será:

$\binom{N}{n} \cdot n! = \binom{4}{2} \cdot 2! = 6 \cdot 2 = 12$ posibilidades, estas son:

(1,2), (2,1), (1,3), (3,1), (1,4), (4,1), (2,3), (3,2), (2,4), (4,2), (3,4), (4,3)

Esquemas de distribución

- **Bernoulli**

Se considera un experimento aleatorio que solo tiene 2 posibles resultados: éxito o fallo. Si se supone que el experimento se repita de tal forma que cada

repetición no depende de otra (independiente) ocurren N casos favorables al fallo. Este tipo de problemas se conocen como ensayos de Bernoulli.

Denotaremos:

A_n: números de casos favorables al evento A, dentro de todos los posibles resultados (igualmente probables) en n ensayos de Bernoulli.

B_n: número de casos desfavorables al evento A dentro de todos los posibles resultados (igualmente probables) en n ensayos de Bernoulli.

Teorema de Bernoulli: $\frac{A_n}{B_n}$ se hace tan grande como se quiera a medida que n crece si y solo si el número de casos favorables al evento A, se hace más grande que el número de casos desfavorables al evento A, mientras más se repita el experimento.

Aplicación de Bernoulli para pruebas independientes

Teorema: Si n pruebas son independientes, entonces un número m cualquiera de ellas son también independientes.

$P_n(m) = \binom{n}{m} p^m q^{n-m}$ donde p: probabilidad de éxito

q: probabilidad de fallo

n: pruebas independientes (total)

m: pruebas independientes (favorables)

Ejemplo

En una industria la probabilidad de que el consumo de petróleo sea considerado normal (es decir, que no rebase cierta cantidad en 24 horas) es igual a ¾. Se desea determinar cuáles serán las probabilidades de que, en el curso de los días laborables de la semana siguiente, el consumo del petróleo sea normal durante 0, 1, 2, 3, 4, 5 y 6 días.

Solución:

Denotaremos por A el evento que consiste en que el consumo de petróleo en el central sea normal cuya probabilidad es ¾ y por $P_6(m)$ la probabilidad de que durante *m* días de 6 el consumo sea normal.

Entonces

$$P_6(6) = \left(\frac{3}{4}\right)^6 \approx 0.18$$

$$P_6(5) = \binom{6}{5}\left(\frac{3}{4}\right)^6\left(\frac{1}{4}\right) \approx 0.36$$

$$P_6(4) = \binom{6}{4}\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^2 \approx 0.30$$

$$P_6(3) = \binom{6}{3}\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right)^3 \approx 0.13$$

$$P_6(2) = \binom{6}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^4 \approx 0.03$$

$$P_6(1) = \binom{6}{1}\left(\frac{3}{4}\right)\left(\frac{1}{4}\right)^5 \approx 0$$

$$P_6(0) = \left(\frac{1}{4}\right)^6 \approx 0$$

Obsérvese que se cumple que $\sum_{m=0}^{6} P_6(m) = 1$

- **Binomial**

Supongamos que se realizan n experimentos independientes, en cada uno de los cuales el suceso A puede ocurrir o no ocurrir. La probabilidad de que ocurra el suceso A en todas las pruebas es constante e igual a p, por lo tanto la probabilidad de que no ocurra es q=1-p. Evidentemente el suceso A en esta prueba puede no aparecer o aparece una vez o 2 veces,..., o n veces. Luego para cada k =0,..., n $P_n(k) = C_n^k p^k q^{n-k}$

Se llama Binomial porque el segundo miembro de la igualdad puede considerarse como término general de la descomposición del binomio de Newton.

$$(p+q)^n = C_n^n p^n + C_n^{n-1} p^{n-1} q + \cdots + C_n^k p^k q^{n-k} + \cdots + C_n^0 q^n$$

Donde

$C_n^n p^n$ es la probabilidad de que el suceso ocurra n veces de n veces.

$C_n^{n-1} p^{n-1} q$ es la probabilidad de que el suceso ocurra n-1 veces de n veces.

$C_n^k p^k q^{n-k}$ es la probabilidad de que el suceso ocurra k veces de n veces.

$C_n^0 q^n$ es la probabilidad de que el suceso no ocurra nunca en n veces.

Ejemplo

a) Una moneda se arroja 2 veces.

A_1: aparezca cara (1)

A_2: aparezcan caras (2)

A_0: no aparezca cara (0)

Solución:

P= ½ en que en una salida salga cara o escudo.

$$P_2(2) = C_2^2 p^2 q^0 = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

$$P_2(1) = C_2^1 pq = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$$

$$P_2(0) = C_2^0 p^0 q^2 = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

Se tienen N objetos, de ellos N_1 de tipo I y N_2 de tipo II, se toma una muestra de tamaño n y se desea saber la probabilidad de que esa muestra de tamaño n, haya k objetos de tipo I. (con remplazo)

$$P = \binom{n}{k} \left(\frac{N_1}{N}\right)^k \left(\frac{N_2}{N}\right)^{n-k} \quad \text{donde} \quad p = \frac{N_1}{N} \quad \text{y} \quad q = \frac{N_2}{N}$$

- **Hipergeométrica**

Se tiene N objetos, de ellos N_1 de tipo I y N2 de tipo II, se toma una muestra de tamaño n y se desea saber la probabilidad de que esa muestra de tamaño n, haya k objetos de tipo I (sin remplazo).

$$P = \frac{\binom{N_1}{k}\binom{N - N_1}{n-k}}{\binom{N}{n}}$$

Donde:

$\binom{N_1}{k}\binom{N - N_1}{n-k} \Rightarrow$ número de casos favorables (que haya k objetos de tipo I)

$\binom{N}{n} \Rightarrow$ número de casos posibles.

- **Binomial negativa**

Obtengamos la probabilidad de obtener exactamente x fallos antes del r-ésimo éxito. Para que esto suceda debemos obtener un éxito sobre el ensayo (*x+r*), precedido exactamente de r-1 éxito y x fallos en los 1ros *r+x-1* ensayos. La probabilidad de obtener *r-1* éxitos en los 1ros *r+x-1* ensayos se obtiene de $P_{r-1}(r+x-1) = \binom{r+x-1}{r-1} p^r (1-p)^x$, *x = 0, 1, 2, ...*

- **Poisson**

Supongamos que se realizan *n* experimentos independientes, en cada uno de los cuales la probabilidad de que aparezca el suceso A es igual a *p*. En el caso de *n* grande y *p* pequeño ($p \leq 0,1$) se emplea la fórmula de Poisson.

El caso es hallar la probabilidad de que para un número muy grande de pruebas, en cada una de las cuales la probabilidad del suceso es muy pequeña, el suceso ocurrirá exactamente *k* veces.

$$P_n(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Ejercicios propuestos sobre probabilidad clásica

1. Existen dos talleres que ensamblan televisores marca SONY. El equipo de auditoría ha determinado que en el taller 1 solo el 2% de sus producciones son defectuosas y que en el taller 2 solo el 1%. Estos talleres son los proveedores de la tienda «La Fantasía». Se conoce que el taller 1 envió 10

televisores a la tienda y el taller 2 envió 15 televisores. Si un cliente hizo una reclamación por haber comprado un televisor defectuoso ¿Cuál es la probabilidad que haya sido ensamblado en el taller 1?

Solución 1

Se definen los eventos:

A_1: Que un televisor haya sido ensamblado en el primer taller.

A_2: Que un televisor haya sido ensamblado en el segundo taller.

B: Que el televisor sea defectuoso.

Luego se aplica la fórmula de Bayes:

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{\frac{10}{25} \cdot (0.01)}{\frac{10}{25} \cdot (0.01) + \frac{15}{25} \cdot (0.02)} = 0.25$$

La probabilidad que haya sido ensamblado en el taller 1 es igual a 0.25.

2. Se ha observado que un aprendiz logra tener éxito en el 90% de los tabacos que ha torcido durante su periodo de adiestramiento laboral. Si para una evaluación de desempeño tiene que torcer 5 tabacos, halle la probabilidad de que al menos uno sea defectuoso.

Solución 2

En este caso estamos en presencia de un esquema binomial:

Definimos la probabilidad de éxito (p = 0.9) y la de fallo (q = 0.1)

$$P_n(m) = \binom{n}{m} p^m q^{n-m}$$

El evento compuesto que al menos uno de los tabacos sea defectuoso es complementario a que ningún tabaco sea defectuoso, por eso es más fácil calcular la probabilidad utilizando la siguiente expresión:

$$1 - P_5(5) = 1 - \binom{5}{5}(0.9)^5(0.1)^0 = 1 - (0.59049) \approx 0.41$$

La probabilidad de que al menos un tabaco sea defectuoso es igual a 0.41.

3. Hay una caja que contiene 20 clavos cortos y 10 medianos y otra caja que contiene 20 medianos y 20 cortos. Se selecciona una caja al azar y se extrae un clavo de ella. Si el clavo extraído es corto ¿Cuál es la probabilidad de que sea de la primera caja?

Solución 3

Se definen los eventos:

A_1: seleccionar la primera caja.

A_2: seleccionar la segunda caja.

B: que el clavo seleccionado sea corto.

Luego se aplica la fórmula de Bayes:

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{\frac{1}{2} \cdot (\frac{20}{30})}{\frac{1}{2} \cdot (\frac{20}{30}) + \frac{1}{2} \cdot (\frac{20}{40})} = \frac{4}{7} \approx 0.57$$

La probabilidad de que el clavo corto seleccionado sea de la primera caja es igual a 0.57.

4. Un pescador captura 10 peces, 3 de los cuales están por debajo del peso que permite la ley. Un inspector realiza un control y para ello selecciona aleatoriamente 2 pescados. Calcule la probabilidad de que ambos pescados estén por encima del peso que permite la ley.

Solución 4

En este caso estamos en presencia de un esquema binomial:

Definimos la probabilidad de éxito (p=0.7), estar por encima del peso que permite la ley, y la de fallo (q=0.3), que el peso esté por debajo.

El esquema es binomial:

$$P_n(m) = \binom{n}{m} p^m q^{n-m}$$

$$P_2(2) = \binom{2}{2}(0.7)^2(0.3)^0 = 0.49$$

La probabilidad de que ambos pescados estén por encima del peso que permite la ley es igual a 0.49.

5. Hay una oficina que tiene dos sillas nuevas y una silla vieja y otra oficina que tiene dos sillas nuevas y dos sillas viejas. Se selecciona una oficina al azar y se extrae una silla. Si la silla extraída es vieja ¿Cuál es la probabilidad de que sea de la primera oficina?

Solución 5

Se definen los eventos:

A_1: Seleccionar la primera oficina.

A_2: Seleccionar la segunda oficina.

B: Que la silla seleccionada sea vieja.

Luego se aplica la fórmula de Bayes:

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{\frac{1}{2}\cdot(\frac{2}{3})}{\frac{1}{2}\cdot(\frac{2}{3}) + \frac{1}{2}\cdot(\frac{2}{4})} = \frac{4}{7} \approx 0.57$$

La probabilidad de que la silla vieja seleccionada sea de la primera oficina es igual a 0.57.

6. Una caja contiene 50 piezas de las cuales 5 son usadas y el resto son nuevas. Se escoge una muestra de 5 piezas sin reemplazamiento. Calcule la probabilidad de que las cinco piezas sean nuevas.

Solución 6

En este caso estamos en presencia de un esquema binomial:

Definimos la probabilidad de éxito (p=0.9), que la pieza seleccionada sea nueva, y la de fallo (q=0.1), que la pieza seleccionada sea vieja.

El esquema es binomial:

$$P_n(m) = \binom{n}{m} p^m q^{n-m}$$

$$P_5(5) = \binom{5}{5}(0.9)^5(0.1)^0 = 0.81$$

La probabilidad de que las cinco piezas seleccionadas sean nuevas es igual a 0.81.

7. Una caja contiene dos bolas rojas y una bola blanca y otra caja contiene una bola roja y dos bolas blancas. Se selecciona una caja al azar y se extrae una bola. Si la bola extraída es roja ¿Cuál es la probabilidad de que sea de la segunda caja?

Solución 7

Se definen los eventos:

A_1: seleccionar la primera caja.

A_2: seleccionar la segunda caja.

B: que la bola seleccionada sea roja.

Luego se aplica la fórmula de Bayes:

$$P(A_2|B) = \frac{P(A_2)P(B|A_2)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{\frac{1}{2}\cdot(\frac{1}{3})}{\frac{1}{2}\cdot(\frac{2}{3}) + \frac{1}{2}\cdot(\frac{1}{3})} = \frac{1}{3} \approx 0.33$$

La probabilidad de que la bola roja sea de la segunda caja es igual a 0.33.

8. En una industria la probabilidad de que el consumo de petróleo sea considerado normal (es decir, que no rebase cierta cantidad en 24 horas) es igual a ¾. Se desea determinar cuál es la probabilidad de que el consumo del petróleo sea normal durante 3 días consecutivos.

Solución 8

En este caso estamos en presencia de un esquema binomial:

Definimos la probabilidad de éxito (p=0.75), que el consumo de petróleo sea considerado normal, o sea, que no rebase la cantidad fijada, y la de fallo (q=0.25), que rebase dicha cantidad.

El esquema es binomial:

$$P_n(m) = \binom{n}{m} p^m q^{n-m}$$

$$P_3(3) = \binom{3}{3} (0.75)^3 (0.1)^0 = 0.422$$

La probabilidad de que el consumo se normal durante tres días consecutivos es igual a 0.422.

Probabilidad frecuentista o estadística

Cabe señalar que el enfoque de la probabilidad clásica tuvo éxito al ser aplicado a problemas relacionados con los juegos de azar, pero presentaba series limitaciones cuando se intentaba utilizar en otros tipos de problemas que sus condiciones no se ajustaban a sus supuestos básicos.

En consecuencia, surgieron métodos más relacionados con la experimentación y se sustentaron en una base empírica de la interpretación frecuentista o estadística de la probabilidad, a partir del análisis de las frecuencias de los fenómenos aleatorios.

El problema aquí surge porque en definitiva igualmente verosímil es lo mismo que igualmente probable, es decir, se justifica la premisa con el resultado. Además ¿qué ocurre cuando estamos considerando un experimento donde no se da esa simetría?, ¿qué hacer cuando el número de resultados posibles es infinito?

Es un hecho empíricamente comprobado que la frecuencia relativa de un suceso tiende a estabilizarse cuando la frecuencia total aumenta. Pues, cuando se repite un experimento bajo las mismas condiciones, un gran número de veces, la presencia de eventos mantiene cierta regularidad y así, cuando se

calcula el cociente del número de veces que ocurra el evento por el número de veces que se realiza el experimento, este número o frecuencia tiende a estabilizarse alrededor de determinado valor constante p = P(A).

En consecuencia, surge así la definición de probabilidad estadística o frecuentista.

Definición: la ***probabilidad estadística o frecuentista*** de un suceso es un número ideal al que converge su frecuencia relativa cuando la frecuencia total tiende a infinito.

Por ejemplo, la probabilidad de que salga un seis al tirar un dado balanceado es 1/6 porque al hacer un gran número de tiradas su frecuencia relativa es aproximadamente esa.

El problema radica en que al no poder repetir la experiencia infinitas veces, la probabilidad de un suceso ha de ser aproximada por su frecuencia relativa para un número suficientemente grande.

Se dirá que un evento A tiene una probabilidad P si se cumplen las siguientes propiedades:

1. Bajo las mismas condiciones es posible repetir un experimento aleatorio ε un número ilimitado de veces.
2. Como resultado de un número suficientemente grande de repeticiones del experimento aleatorio ε, la frecuencia de ocurrencia de un evento A difiere ligeramente de una cierta constante que generalmente es desconocida.

Al repetir un experimento un gran número de veces, como valor numérico de la constante se puede tomar la frecuencia o un número cercano. La probabilidad de un evento aleatorio así definido se llama probabilidad estadística.

Otra forma de enunciarlo es:

Si se hacen n repeticiones de un experimento aleatorio ε y el suceso A es observado n_A veces, $f_A = \frac{n_A}{n}$ es la probabilidad de A.

Ejemplos

a) Si no sabemos qué dolencia aqueja a los enfermos que acuden a la consulta de un hospital, debemos aplicar la definición anterior.

   Se debe tomar una muestra:

   - Seleccionamos un nombre y se anota la enfermedad del paciente.
   - Este proceso se repite n veces después de reponer la tarjeta del paciente seleccionado.

   Si n = 100 y obtenemos que de ellos 30 padecen de gripe, 10 de la circulación, 40 de reuma y 20 de la presión, entonces podemos decir que los pacientes que acuden a un hospital tiene una probabilidad de 0,3 de padecer de gripe, de 0,1 de padecer de la circulación, de 0,4 de padecer de reuma y de 0,2 de padecer de la presión.

b) Repetidamente se realizaron pruebas de arrojamiento de la moneda en las que se contaron el número de aparición de cara. Los resultados se muestran en la siguiente tabla:

	número de arrojamientos	número de caras	Frecuencia relativa
Buffon	4040	2048	0,5069
K. Pearson	12000	6019	0,5016
K. Pearson	24000	12012	0,5005

En este ejemplo las frecuencias se desvían un poco del número 0,5. La desviación es menor mientras mayor es el número de pruebas.

Como la probabilidad de aparición de cara al arrojar la moneda es 0,5 nuevamente comprobamos que la frecuencia relativa oscila alrededor de la probabilidad.

Esta definición salva las deficiencias de la definición clásica, pues para aplicarla no es necesario que los eventos sean equiprobables, ni que haya un número finito (fijo) de estos. Sin embargo, esta definición también tiene deficiencias:

- Es más bien descriptiva que formalmente matemática;
- No pone al descubierto la esencia de los eventos a los cuales se le aplica;
- No es aplicable a todas las situaciones de la vida real.

Capítulo III. ELEMENTOS DE ESTADÍSTICA INFERENCIAL

Variables aleatorias

Para introducir el concepto de variable aleatoria relacionado con los conceptos del tema 2 se utilizará el experimento aleatorio consistente en el lanzamiento de dos monedas representando sus puntos muestrales, cara (△) y cruz (✯), por los símbolos señalados.

△△

✯△

△✯

✯✯

Se puede observar que es posible referirse a esos mismos puntos muestrales, definiendo una variable X que «mida» la cantidad de cruces que se obtienen al lanzar dos monedas.

A cada punto muestral es posible asociarle un valor de la variable X:

△△ X tomaría el valor 0

✯△ X tomaría el valor 1

△✯ X tomaría el valor 1

✯✯ X tomaría el valor 2

Sobre la base de la variable X definida, que toma esos valores de manera aleatoria, puede expresarse el espacio muestral S como:

$S= \{x \mid x = 0, 1, 2\}$

Precisamente, a esa X se le denomina variable aleatoria.

Variable aleatoria: es una función que asocia un valor, usualmente numérico, al resultado de un fenómeno aleatorio.

Intuitivamente, una variable aleatoria puede tomarse como una cantidad cuyo valor no es fijo pero puede tomar diferentes valores; una distribución de probabilidad se usa para describir la probabilidad de que se den los diferentes valores. En términos formales una variable aleatoria es una función definida sobre un espacio de probabilidad.

Mediante la utilización del concepto de variable aleatoria se puede expresar de una manera más simple a los eventos:

- A: que al menos se obtenga una cruz $A = \{X \geq 1\}$
- B: que exactamente se obtengan dos estrellas $B = \{X = 2\}$
- C: que a lo sumo se obtenga una estrella $C = \{X \leq 1\}$

La representación de cualquier evento siempre vendrá dada por un conjunto de números reales y ello hará mucho más fácil las operaciones entre los mismos.

Ejemplo

$A = (X \geq 1)$ $B = (X = 2)$ y entonces $(A.B) = (X=2)$; ya que 2 es el único número que tanto pertenece a A como a B.

Aunque el concepto de variable aleatoria se ha explicado sobre la base de un fenómeno aleatorio cuyo espacio muestral es finito; este concepto también es válido para S infinitos.

Ejemplos

1) X: Cantidad de personas que llegan a una cola en un tiempo t.

   Sus posibles valores serían X = 0,1, 2... (infinito numerable)

2) Y: Tiempo de trabajo sin fallo de cierto equipo.

   Sus posibles valores $Y \geq 0$ (infinito no numerable)

Para comprender de una manera más amplia y rigurosa los tipos de variables, es necesario conocer la definición de conjunto discreto.

Conjunto discreto: si está formado por un número finito de elementos, o si sus elementos se pueden enumerar en secuencia de modo que haya un primer elemento, un segundo elemento, un tercer elemento, y así sucesivamente (es decir, un conjunto infinito numerable sin puntos de acumulación).

En consecuencia, las variables aleatorias pueden clasificarse en discretas o continuas.

Variable aleatoria discreta: es aquella variable aleatoria que toma determinados valores en un conjunto discreto.

Por ejemplo, las siguientes son variables aleatorias discretas: «cantidad de cruces obtenidas al lanzar dos monedas» y «cantidad de personas que llegan a una cola en un tiempo t».

Variable aleatoria continua: es aquella que toma todos los valores en un conjunto no numerable.

Por ejemplo, la siguiente es una variable aleatoria continua: «tiempo de trabajo sin fallo de cierto equipo».

Observación: las variables aleatorias se acostumbran a denotar mediante las letras del alfabeto en mayúscula (X, Y, Z) y para referirse a sus posibles valores se utilizan las minúsculas correspondientes: $X = x_1; x_2; x_3$; se entiende que la variable aleatoria X toma los valores x_1, x_2 y x_3.

Función de probabilidad. Propiedades

El objetivo central al estudiar los fenómenos aleatorios es conocer las probabilidades de ocurrencia de diferentes eventos en él definidos; resulta de interés entonces determinar o conocer el comportamiento probabilístico de las variables aleatorias que se definan para estudiar esos fenómenos; en otras palabras no tiene sentido conocer sólo los valores que puede tomar una

variable aleatoria, sino que, unido a ello es imprescindible conocer las probabilidades con que la variable aleatoria discreta toma cada uno de sus posibles valores, o conjunto de sus valores.

En el caso de las variables discretas como en el ejemplo del lanzamiento de una moneda dos veces para el cual se definió la variable X: «cantidad de cruces que se obtienen al hacer el lanzamiento de dos monedas» se puede determinar la probabilidad de ocurrencia de cada uno de sus posibles valores:

Probabilidad de ocurrencia

△△	1/2 * 1/2 =1/4	⟶	P(X = 0) = ¼
✯△	1/2 * 1/2 =1/4	}	P(X = 1) = ¼ + ¼ = ½
△✯	1/2 * 1/2 =1/4		
✯✯	1/2 * 1/2 =1/4	⟶	P(X = 2) = ¼

Observe que los eventos (X = 0); (X = 1) y (X = 2) forman para este experimento un «grupo completo de eventos» y por consiguiente, al obtener las probabilidades de esos eventos, se ha obtenido una distribución de probabilidad del referido fenómeno. Específicamente, este tipo de distribución de probabilidad asociada a variable aleatoria discreta recibe el nombre de *función de probabilidad*.

Función de probabilidad

Sea X una variable aleatoria discreta que conforma el espacio muestral E de la variable aleatoria X de puntos x_1, x_2,..., x_k y sea f una función de R en el conjunto cerrado [0,1], subconjunto de R; entonces la función de probabilidad se define como:

f: R $\rightarrow$ [0,1]

X $\rightarrow$ f (x) y la misma cumple con las siguientes propiedades:

1) $f(x_i) \geq 0 \quad \forall\, x_i \in E$

2) $\sum_{i=1}^{k} f(x_i) = 1, \quad \forall\, x_i \in E$

En otras palabras, la función de probabilidad es una función f(x) que asocia una probabilidad a cada valor de la variable aleatoria X.

f (x) = P (X = x)

Ejemplo

La función de probabilidad de la variable aleatoria X: «número de cruces al lanzar dos monedas», puede expresarse como:

X	0	1	2
f(x)	1/4	1/2	1/4

Comprobando si cumple las propiedades:

1. $f(x) \geq 0$

   f (0) = P (X = 0) = ¼ >0

   f (1) = P (X = 1) = ½ >0

   f (2) = P (X = 2) = ¼ >0, entonces se cumple la primera propiedad.

2. $\sum_{i=1}^{k} f\ (x_i) = 1$

$$\sum_{i=1}^{k} f\ (x_i) = f(0) + f(1) + f(2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

La representación de esta función de probabilidad se muestra en la figura 13.

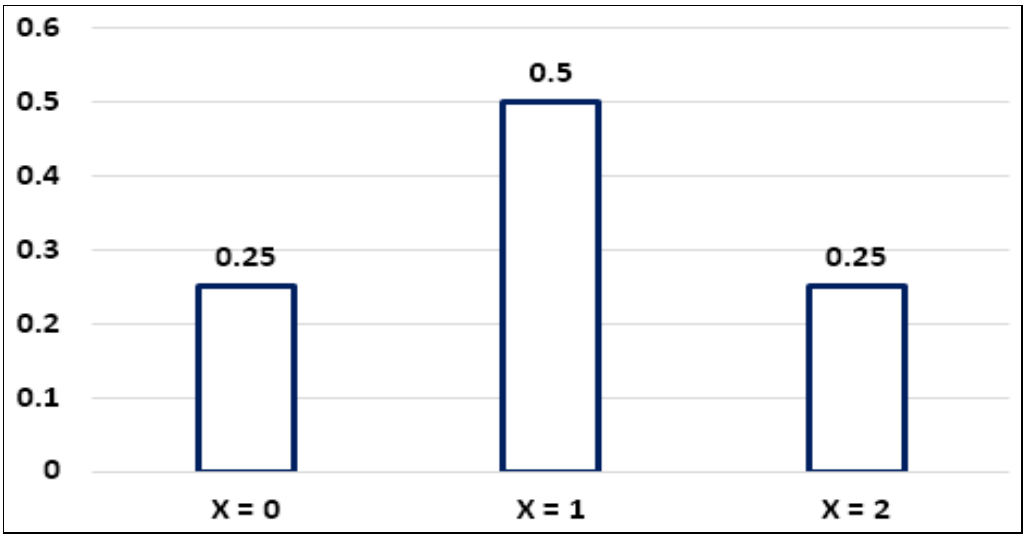


Figura 13. Representación gráfica de una función de probabilidad.

La representación gráfica de una función de probabilidad son puntos en el plano, pero para lograr una más rápida y mejor visión del comportamiento

de la variable aleatoria se acostumbra a utilizar barras verticales similares al de la figura 13.

Por otro lado, es preciso señalar que al conocer la función de probabilidad de una variable aleatoria se puede calcular la probabilidad de cualquier evento que se defina asociado al fenómeno que se estudia.

Ejemplo

Para la función de probabilidad

X	0	1	2
f(x)	1/4	1/2	1/4

Pueden definirse los eventos:

1) A: al menos se obtenga una cruz [A = (X ≥ 1)]

P (A) = P (X ≥ 1) = P(X = 1) + P (X = 2) = f (1) + f (2) = ½ + ¼ = ¾

2) B: no se obtengan cruces

P (B) = P (X = 0) = f (0) = ¼

Ejemplo

Sea la variable aleatoria X: «número de clientes que arriban a un restaurante en un periodo de tiempo de 5 minutos». Se determinó que la función de probabilidad de X es la siguiente:

$$f(x) = \begin{cases} \frac{1}{8} & \text{para } X = 0 \\ \frac{3}{8} & \text{para } X = 1 \\ \frac{3}{8} & \text{para } X = 2 \\ a & \text{para } X = 3 \\ 0 & \text{otros valores} \end{cases}$$

a) Determine el valor de a.

b) Interprete el valor anterior en relación a la frecuencia de veces que arriban 3 clientes en períodos de 5 minutos.

c) Calcule la probabilidad de que arriben al restaurante al menos 2 clientes en un periodo de 5 minutos.

d) ¿Cuál es la probabilidad de que en un periodo de 5 minutos considerado arriben entre 1 y 3 clientes?

Solución

X: número de clientes que arriban a un restaurante en un periodo de tiempo de 5 minutos.

a) $\sum_{i=1}^{4} f\ (x_i) = f(0) + f(1) + f(2) + f(3) + f(\text{otros valores}) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + a + 0 = 1$

Luego se obtiene que $a = 0.125$

b) En el 12.5 % de los intervalos de 5 minutos arriban 3 clientes al restaurante.

c) P(X ≥ 2) = f (2) + f (3) + f (otros valores) = 3/8 + 1/8 + 0 = 4/8

P(X ≥ 2) = 0.5

d) P(1 ≤ X ≤ 3) = f(1) + f(2) + f(3) = 3/8 + 3/8 + 1/8 = 7/8 = 0.875

Función de distribución acumulada. Propiedades

Con frecuencia para expresar el comportamiento de las variables aleatorias también se utiliza otra distribución de probabilidad denomina *función de distribución acumulada*.

Función de distribución acumulada:

Esta función se denota como $F_x(t)$ y es una función tal que:

$F_x(t) = P(X \leq t) \quad \forall\, t \in R$

La $F_x(t)$ en variables aleatorias discretas se obtiene a partir de la función de probabilidad mediante $F_x(t) = \sum_{i=1}^{k} f\ (x_i)$ para toda $x_i \leq t$

Ejemplo

La función de distribución acumulada $F_x(t)$ del ejemplo de la función de probabilidad de la variable aleatoria X: «número de cruces al lanzar dos monedas», sería:

$F_x(0) = P(X \leq 0) = P(X = 0) = ¼$

$F_x(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = ¾$

$F_x(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 1$

x	0	1	2
F(x)	¼	¾	1

Con esta función se puede calcular P(X = 1) de la siguiente forma:

$P(X = 1) = P(X \leq 1) - P(X \leq 0) = F_x(1) - F_x(0) = ¾ - ¼ = ½.$

La representación gráfica de esta función de distribución acumulada es la que se puede observar en la figura 14.

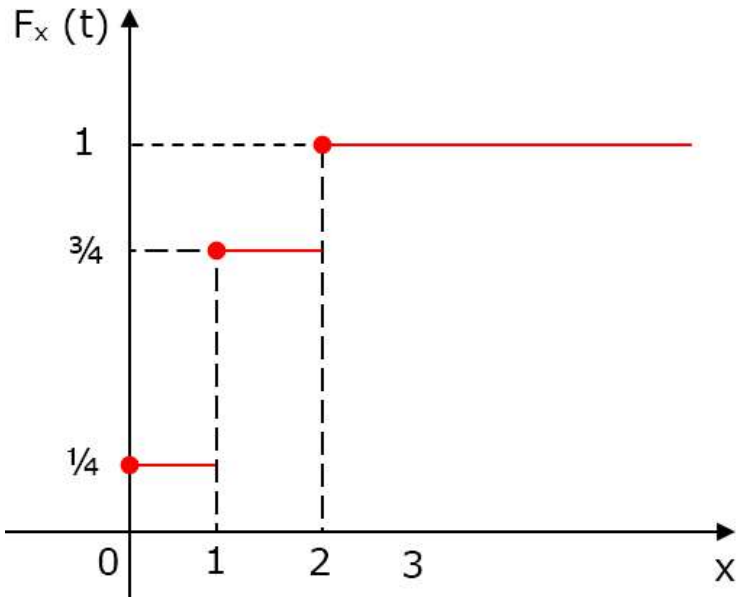


Figura 14. Representación gráfica de una función de distribución acumulada.

Como se puede ver en la figura 14, la gráfica la función es no decreciente y en este caso de variable discreta se presenta de forma escalonada debido a que hay intervalos en que la variable no toma valor y, por tanto, mantiene su probabilidad en los mismos.

Función de densidad probabilística. Propiedades

La función de densidad probabilística permite calcular:

$P(a < X \leq b) = F_x(b) - F_x(a)$ tanto para variables aleatorias discretas como para variables aleatorias continuas. Para este último caso esta función siempre será continua y si además es derivable, podemos apoyarnos en el segundo teorema fundamental del cálculo que plantea:

$$F(b) - F(a) = \int_a^b f(x)dx$$

Es decir, la integral definida de f(x) en (a, b) puede hallarse mediante la diferencia de su primitiva evaluada para los extremos del intervalo, podemos plantear:

$$P(a < x \leq b) = F(b) - F(a) = \int_a^b f(x)dx$$

Siendo la primitiva la función de distribución acumulada. A la función f(x) se le denomina *función de densidad probabilística*.

Definición:

Dada una función f(x); ésta será una *función de densidad probabilística* si

$$P(a < X \leq b) = \int_a^b f(x)dx$$

y tiene que cumplir con dos propiedades:

1. $f(x) \geq 0$ para $\forall x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Si se quiere hallar $F_x(t)$ a partir de la función de densidad se tendría que proceder como sigue

$$F_x(t) = \int_{-\infty}^{t} f(x)dx$$

a partir de esa expresión se concluye que

$$\frac{dF_x(t)}{dt} = f(x)$$

Ejemplo

Se ha decidido construir un servicentro en cierta avenida y por los estudios realizados se concluyó que la demanda de gasolina semanal, en miles de litros, se ajusta a la función de densidad probabilística siguiente:

$$f(X) = \begin{cases} X-1 & \text{si } 1 \le X \le 2 \\ 3-X & \text{si } 2 \le X \le 3 \\ 0 & \text{otros valores} \end{cases}$$

Determine:

a) La probabilidad de que la demanda de gasolina en una semana supere los 2.2 miles de litro. Interprete el resultado.

b) Si en los primeros 3 días de una semana se demandan 1500 litros, ¿qué probabilidad hay de que en esa semana la demanda sea inferior a 2500 litros.

Solución:

X: demanda de gasolina semanal en miles de litros

a) $P(X > 2.2) = \int_{2.2}^{3} (3 - X)dx = 0.32$

Interpretación: el 32 % de las semanas la demanda supera los 2.2 litros

b) $$\frac{P(1.5 \le X < 2.5)}{P(X \ge 1.5)} = \frac{\int_{1.5}^{2} (X-1)dx + \int_{2}^{2.5} (3-X)dx}{\int_{1.5}^{2} (X-1)dx + \int_{2}^{3} (3-x)dx} = \frac{0.75}{0.875} = 0.8571$$

Características numéricas de las distribuciones

Las distribuciones de probabilidad proporcionan una información completa sobre el comportamiento aleatorio de una variable aleatoria, pero sobre la

base de ese comportamiento es posible definir y calcular, algunos valores numéricos capaces de resumir las características generales del fenómeno aleatorio que se estudia a partir de esa variable aleatoria. Se tratarán así, en esta actividad, las características numéricas fundamentales: valor esperado y varianza.

- **Valor esperado de una variable aleatoria**

El valor esperado de una variable aleatoria (o de la distribución de probabilidad de X), es una medida de tendencia central de una variable aleatoria e informa alrededor de que valor se mueven los valores de la variable aleatoria.

Por ejemplo en el lanzamiento de una moneda dos veces para el cual se define la variable discreta X: «cantidad de caras que se obtienen al hacer el lanzamiento de dos monedas», supongamos que se repite el experimento 16 veces y se registran los resultados de cada repetición si los resultados observados muestran que se obtiene:

Ninguna cara (X=0) en 4 ocasiones

Una sola cara (X=1) en 7 ocasiones

Dos caras (X=2) en 5 ocasiones

Entonces si se calcula

$$\bar{X}=\frac{4\cdot 0+7\cdot 1+5\cdot 2}{16}=1.06$$

Si se reestructura este cálculo convenientemente se tendrá

$$\bar{X}=\frac{4}{16}(0)+\frac{7}{16}(1)+\frac{5}{16}(2)=1.06$$

Donde

$\frac{4}{16}, \frac{7}{16}, \frac{5}{16}$ son valores de frecuencias relativas de cada valor de los variable para las 16 repeticiones del experimento.

Se conoce que las frecuencias relativas son valores de probabilidad, de manera que, para variables aleatorias discretas se podrá obtener su media sin necesidad de realizar el experimento a partir de la siguiente expresión:

$$E(X) = \mu = \sum_{\forall x} xf(x)$$

En el ejemplo para el cual se había obtenido:

X	0	1	2
f(x)	1/4	1/2	1/4

Se determina

$$E(X) = \mu = \sum_{x} xf(x) = 0 \cdot f(0) + 1 \cdot f(1) + 2 \cdot f(2) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

Esto implica que si se repite un gran número de veces el experimento de lanzar una moneda dos veces se obtendrá como promedio una cara.

Nótese que el valor esperado de una variable discreta no necesariamente tiene que coincidir con un valor de la variable.

En el caso de las variables aleatorias continuas, la concepción es la misma, pero se obtiene según:

$$E(X) = \mu = \int_{\forall x} x\, f(x)\, dx$$

En el ejemplo de la construcción del servicentro, si queremos hallar cuál es la demanda promedio para una semana cualquiera

X: demanda de gasolina semanal en miles de litros

$$f(X) = \begin{cases} X-1 & \text{si } 1 \leq X \leq 2 \\ 3-X & \text{si } 2 \leq X \leq 3 \\ 0 & \text{en otros valores} \end{cases}$$

$$E(X) = \mu = \int_1^2 x(x-1)dx + \int_2^3 x(3-x)dx = \frac{5}{6} + \frac{7}{6} = 2$$ (miles de litro semanales)

Las propiedades más importantes del valor esperado son:

- E (C) = C
- E (CX) = C · E (X)
- E (X + Y) = E (X) + E (Y)

Si se requiere calcular el valor esperado de una función de una variable aleatoria g(X) con función de probabilidad conocida f(X) entonces:

Para variables discretas

$$E[g(x)] = \sum_{\forall x} g(x) f(x)$$

Para variables continuas

$$E[g(x)] = \int_{\forall x} g(x) f(x) dx$$

- **Varianza de una variable aleatoria**

El valor esperado por sí solo no brinda toda la información sobre el comportamiento de la variable aleatoria; conviene saber además si ese valor promedio representa adecuadamente al conjunto de valores de la variable, es decir, si es un promedio de valores cercanos uno de otro, o si por el contrario es un promedio de valores alejados unos de otros.

Esa información nos la brinda la varianza V(X), que es una medida de la variabilidad o dispersión de los valores de la variable aleatoria con respecto a su valor esperado, complementa así la caracterización de la distribución, dando una idea de la forma de la distribución.

Conceptualmente se expresa como:

$$V(X) = E\left[\left(x - E(x)\right)^2\right]$$

Y se calculará mediante la expresión:

$$V(X) = E(x^2) - E^2(x)$$

La varianza nos informa cuan concentrados o dispersos están los valores de la variable aleatoria alrededor de la media o valor esperado.

Si V (X_1) < V (X_2) hay mayor concentración en los valores de la variable X_1 que en los valores de la variable X_2.

Siguiendo el ejemplo del lanzamiento de una moneda dos veces para hallar la varianza de X se hallaría:

$$E(x^2) = \sum_{\forall x} x^2 f(x) = 0^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{4} = 1.5$$

$$V(X) = E(x^2) - E^2(x) = 1.5 - 1^2 = 0.5 \text{ (caras)}$$

Las propiedades más importantes de la varianza son:

- V (C) = 0
- V (CX) = C^2 V (X)

Distribuciones continuas clásicas

- **Distribución Uniforme**

Definición: Se dice que la variable aleatoria continua *X,* tiene una distribución uniforme si su función de probabilidad está dada por la expresión:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b \\ 0 & \text{otro caso} \end{cases}$$

La notación para esta distribución es *X~U(a, b)* y se lee «la variable aleatoria *X* se distribuye uniformemente en el intervalo (*a, b*)».

La función de probabilidad acumulada para esta distribución uniforme está dada por:

$$F_X(x) = \frac{x-a}{b-a} \qquad \text{si } a < x < b.$$

Valor esperado y varianza:

$$E(X) = \frac{b+a}{2} \qquad V(X) = \frac{(a-b)^2}{12}$$

- **Distribución Exponencial**

Definición: La variable aleatoria *X* tiene una distribución exponencial con parámetro $\beta > 0$, y se escribe X~exp(x; β) si su f función de probabilidad es de la forma:

$$f_X(x) = \begin{cases} \beta e^{-\beta x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

La función de distribución de la variable aleatoria es:

$$F_X(x) = 1 - e^{-\beta x}$$

Valor esperado y varianza:

$$E(X)=\frac{1}{\beta} \qquad V(X)=\frac{1}{\beta^2}$$

Esta distribución tiene la propiedad de falta de memoria pues para t >0 y h>0

$$P(X\geq t+h \mid X\geq t)=P(X\geq h)$$

Teorema: Supóngase que las variables aleatorias $X_1, X_2, \ldots, X_n$ constituyen una muestra aleatoria (esto quiere decir que son independientes y con la misma función de densidad de probabilidad) de una distribución exponencial con parámetro β. Entonces la distribución de $Y=min(X_1, X_2, \ldots, X_n)$ es una distribución exponencial con parámetro $n\beta$.

- **Distribución Gamma.**

Definición: La función gamma para α>0 se define como

$\Gamma(\alpha)=\int_0^{\infty} x^{\alpha-1}e^{-x}dx$, para que la integral converja se requiere x > 0.

La función gamma tiene las siguientes propiedades:

1) Si $\alpha>1$ entonces $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$

2) Si $n\in Z^+$ entonces $\Gamma(n)=(n-1)!$

Definición: Se dice que la variable aleatoria X tiene una distribución de probabilidad Gamma con parámetros α >0 y β >0 y se escribe X~ Gamma $(x;\alpha,\beta)$ si su función de probabilidad está definida como:

$$f_X(x)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} & x>0\\ 0 & \text{otro caso}\end{cases}$$

La media y la varianza para esta distribución Gamma están dadas por:

$$E(X)=\frac{\alpha}{\beta} \qquad V(X)=\frac{\alpha}{\beta^2}$$

Propiedades:

1) Si las variables aleatorias $X_1, X_2, \ldots, X_\alpha$ tienen una distribución exponencial con parámetro β y son independientes unas de otras. Entonces la variable aleatoria $X = \sum_{i=1}^{\alpha} X_i$ se distribuye como $\Gamma(x;\alpha,\beta)$. Es decir,

La suma de α variables aleatorias, independientes y distribuidas exponencialmente con parámetro β tiene una distribución Gamma $\Gamma(x;\alpha,\beta)$.

2. Si las variables aleatorias $X_1, X_2, \ldots, X_k$ son independientes y $X_i \sim \Gamma(x_i;\alpha_i,\beta)$ para *i*=1,2,...,*k*. Entonces la variable aleatoria $X = \sum_{i=1}^{k} X_i$ se distribuye como $X \sim \Gamma(x;\alpha = \sum_{i=1}^{k} \alpha_i, \beta)$.

- **Distribución Normal**

Definición: La variable aleatoria cuya ley es normal con parámetros μ y σ^2 tiene la siguiente función de densidad

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \Re, \ \mu \in \Re, \ y \ \sigma > 0$$

La ley normal desempeña un papel importante en la teoría de las probabilidades y es usual que sirva para modelar una variedad de fenómenos aleatorios reales.

No es difícil comprobar que la función de densidad de una variable aleatoria normal tiene un máximo en $x=\mu$, tiene puntos de inflexión en $x=\mu\pm\sigma$ y el eje de las abscisas sirve de asíntota cuando $x\to\pm\infty$. Además:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = 1$$

La gráfica de la función de densidad normal se llama curva normal (curva de Gauss).

Examinemos la función $y = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

1) La función está definida en todo el eje x.
2) $\forall$x la función toma valores positivos, es decir, la curva normal está situada encima del eje x.
3) El límite de la función, al crecer ilimitadamente x (en valor absoluto), es igual a cero: $\lim_{|x|\to\infty} y = 0$, el eje x sirve de asíntota horizontal de la gráfica.

   Examinemos la función respecto al extremo. Hallemos la primera derivada:

   $y' = -\frac{x-\mu}{\sigma^3\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

   y´= 0 cuando x = μ

   y´> 0 cuando x < μ

   y´< 0 cuando x > μ

   Por lo tanto, cuando x=μ la función tiene un máximo igual a: $1/\sigma(2\pi)^{1/2}$
4) La diferencia x-μ está contenida en la expresión analítica de la función al cuadrado, es decir, la gráfica de la función es simétrica respecto a la recta x=μ.
5) Examinemos la función en el punto de inflexión. Hallemos la 2da derivada:

   $y'' = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]$ para x=μ+σ y x=μ-σ la derivada segunda es igual a cero y al pasar por estos puntos, cambia de signo (en ambos puntos el valor de la función es igual a $1/\sigma(2\pi e)^{1/2}$).

   Por consiguiente los puntos $\left(\mu-\sigma; \frac{1}{\sigma\sqrt{2\pi e}}\right)$ y $\left(\mu+\sigma; \frac{1}{\sigma\sqrt{2\pi e}}\right)$ de la gráfica son puntos de inflexión.

Influencia de los parámetros de la distribución normal sobre la fórmula de la curva normal (ver figura 17):

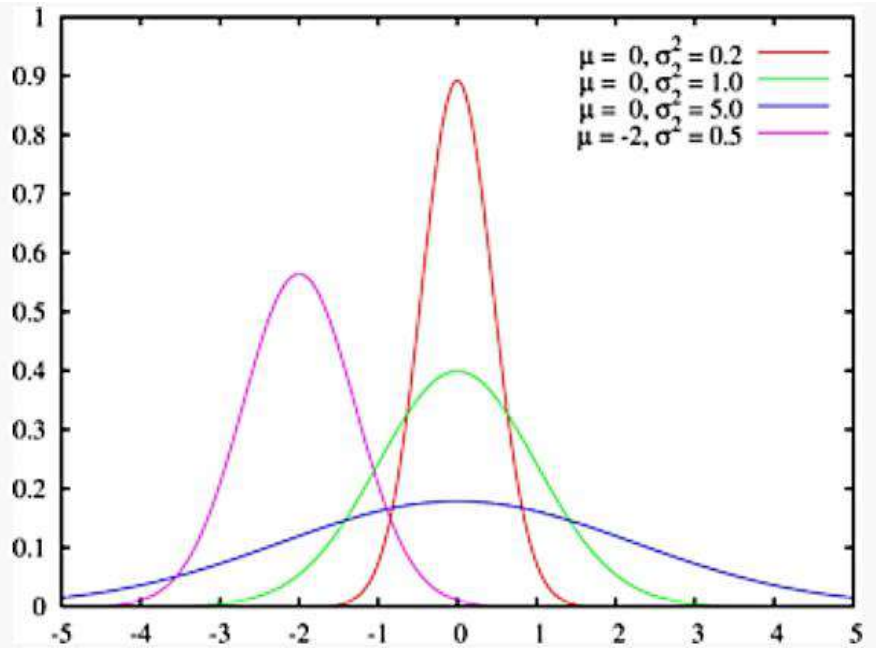


Figura 17. Distribución normal. Forma acampanada y simétrica

- (Parámetro μ). Las gráficas de las funciones de *f(x)* y *f(x-μ)* tienen igual forma, desplazando la gráfica de *f(x)* en el sentido positivo del eje *x* en μ unidades de la escala para $\mu > 0$, obtenemos la gráfica de *f(x-μ)*.

  Por lo tanto, la variación de la magnitud del parámetro μ no altera la forma de la curva normal, sino da lugar solamente a su desplazamiento a lo largo del eje x, hacia la derecha si μ crece y hacia la izquierda si μ decrece.

- (Parámetro σ). Ya sabemos que el máximo de la función es igual a $\frac{1}{\sigma\sqrt{2\pi}}$.

  Por lo tanto, al crecer σ la ordenada máxima de la curva normal decrece, mientras que la propia curva decrece más suavemente, es decir, se aproxima al eje x; cuando decrece σ, la curva normal decrece más agudamente y se alcanza en sentido positivo del eje y.

Para todos los valores de los parámetros σ y μ el área limitada por la curva normal y el eje x se mantiene igual a la unidad (propiedad de la función de densidad).

Probabilidad de que una magnitud aleatoria X está en un intervalo dado:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f_x(x)\,dx$$

Si suponemos que $X \propto N(\mu, \sigma^2)$, entonces:

$$P(x_1 < X < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$ haciendo un cambio de variable

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

$$dx = \sigma\,dz$$

Obtenemos los nuevos límites de integración:

si x = x1 $\Rightarrow$ $z = \frac{x_1 - \mu}{\sigma}$

si x = x2 $\Rightarrow$ $z = \frac{x_2 - \mu}{\sigma}$

Entonces

$$P(x_1 < X < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\frac{x_1-\mu}{\sigma}}^{\frac{x_2-\mu}{\sigma}} e^{-\frac{z^2}{2}} \sigma\,dz = \frac{1}{\sqrt{2\pi}} \int_{\frac{x_1-\mu}{\sigma}}^{0} e^{-\frac{z^2}{2}}\,dz + \frac{1}{\sqrt{2\pi}} \int_{0}^{\frac{x_2-\mu}{\sigma}} e^{-\frac{z^2}{2}}\,dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{\frac{x_2-\mu}{\sigma}} e^{-\frac{z^2}{2}}\,dz - \frac{1}{\sqrt{2\sigma}} \int_{0}^{\frac{x_1-\mu}{\sigma}} e^{-\frac{z^2}{2}}\,dz$$

utilizando la función de Laplace $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}}\,dz$

$$\Rightarrow\ p(x_1 < X < x_2) = \Phi\left(\frac{x_2-\mu}{\sigma}\right) - \Phi\left(\frac{x_1-\mu}{\sigma}\right)$$

Ejemplo

Sea $X \sim N(\mu, \sigma^2)$ donde μ = 30 y σ = 20. Halle la probabilidad de que X tome un valor correspondiente al intervalo (10,50).

$$P(10 < X < 50) = \Phi\left(\frac{50-30}{10}\right) - \Phi\left(\frac{10-30}{10}\right) = \Phi(2) - \Phi(-2)$$ = 0.977 – 0.023 = 0.954

Habitualmente los valores de la función de distribución acumulada para la distribución normal aparecen en las tablas estadísticas o se pueden determinar en varios softwares estadísticos.

Teorema: Si *X* tiene una distribución normal con media μ y varianza σ^2 y si *Y = aX + b,* donde *a* y *b* son constantes con *a* diferente de cero, entonces la variable aleatoria *Y* tiene una distribución normal con media $a \cdot \mu + b$ y varianza $a^2 \cdot \sigma^2$.

Teorema: Si las variables aleatorias $X_1,...,X_k$ son independientes y las $X_i \sim N(x_i;\ \mu_i,\ \sigma_i^2)$ entonces $X_1 + \cdots + X_k$ tiene una distribución normal con media $\mu = \mu_1 + \cdots + \mu_k$ y varianza $\sigma_1^2 + \cdots + \sigma_k^2$.

Corolario (1): Si las variables aleatorias $X_1,...,X_k$ son independientes con $X_i \sim N(x_i;\ \mu_i,\ \sigma_i^2)$ y si $a_1,...,a_k$ y *b* son constantes para las que al menos uno de los valores $a_1,...,a_k$ es distinto de cero, entonces la variable $X = a_1X_1 + \cdots + a_kX_k + b$ tiene una distribución normal con media $\mu = a_1\mu_1 + \cdots + a_k\mu_k + b$ y varianza $\sigma^2 = a_1^2\sigma_1^2 + \cdots + a_k^2\sigma_k^2$.

Corolario (2): Si las variables aleatorias $X_1,...,X_k$ constituyen una muestra aleatoria de una distribución normal con media μ y varianza σ^2. Sea la variable aleatoria $\bar{X}$ conocida como la media muestral.

Entonces $\bar{X} \sim N(x;\mu,\sigma^2/n)$.

Teorema Central del límite

Sean X_1, X_2,..., X_n una muestra aleatoria de una distribución con media μ y varianza σ^2. Entonces, si n es lo suficientemente grande, $\overline{X}$ tiene una distribución normal aproximada con $\mu_{\bar{X}} = \mu$ y $\sigma^2{}_{\bar{X}} = \frac{\sigma^2}{n}$

Es decir: $Z = \frac{X - \mu}{\sigma}$ se convierte en $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$, y entonces tenemos que

$\overline{X} \to N(\bar{x}; \mu, \frac{\sigma^2}{n})$

De forma equivalente, $X = \sum_{i=1}^{n} X_i$ se distribuye aproximadamente normal con

media $\mu_X = n\mu$ y varianza $\sigma_X^2 = n\sigma^2$.

Es decir: $Z = \frac{X - \mu}{\sigma}$ se convierte en $Z = \frac{\overline{X} - n\mu}{\sqrt{n\sigma^2}}$, y entonces tenemos que

$X \to N(x; n\mu, n\sigma^2)$

Aproximación de la distribución Normal a la Binomial: Si repetimos *n* ensayos Bernoulli de manera independiente, se tiene que $E(X_i)=p$ y $V(X_i)=p\cdot q$

con $X_i = \begin{cases} 1 & \text{Si hay fracaso en el i} - \text{ésimo ensayo} \\ 0 & \text{Si hay éxito en el i} - \text{ésimo ensayo} \end{cases}$

el Teorema Central del Límite garantiza que para *n* grande que:

$$\overline{X} \sim N(x; \mu_{\bar{X}} = \mu = p, \sigma^2_{\bar{X}} = \sigma^2 / n = pq / n)$$

y que: $X = \sum_{i=1}^{n} X_i \sim N(x; \mu_X = n\mu = np, \sigma_X = n\sigma^2 = npq)$

Una aplicación práctica cuando *n* es grande, y no podemos usar tablas binomiales para encontrar probabilidades de este tipo, es aproximar la distribución binomial a una normal.

Si se toma una muestra aleatoria de tamaño *n* de una población de tamaño *N* con función de probabilidad $f_X(x)$, cada una de las variables aleatorias $X_1, X_2, \ldots, X_n$ que conforman la muestra son independientes y con la misma distribución.

Cualquier función de estas variables sigue siendo una variable aleatoria que debe tener una función de probabilidad propia, que se puede deducir de la distribución conjunta de las variables aleatorias $X_1, X_2, \ldots, X_n$, que conforman la muestra aleatoria, por esta razón se denomina a esta función de probabilidad como distribución muestral.

Definición: Un estadístico es cualquier función de una muestra aleatoria $X_1, X_2, \ldots, X_n$.

Definición: Los grados de libertad (g.l.) para cualquier estadístico es el número de datos que pueden variar libremente al calcular ese estadístico.

- **Distribución Chi-cuadrada**

Definición: Si *X* es una variable aleatoria con distribución gamma con parámetros $\beta = \frac{1}{2}$ y $\alpha = \frac{n}{2}$ con *n* entero positivo, entonces se dice que *X* tiene una distribución llamada chi-cuadrada con *n* grados de libertad y se denota como $X \sim \chi^2_n$.

Su función de probabilidad es de la forma $f_X(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}, \quad x > 0$.

La media y la varianza de la distribución chi-cuadrada son:

$E(X) = n \qquad V(X) = 2n$

En esta distribución *n* es el parámetro de forma. Los valores de la distribución de probabilidad acumulada se encuentran tabulados en tablas.

Esta distribución es de gran importancia en estadística inferencial, por ejemplo en la construcción de intervalos de confianza, al probar una hipótesis, para encontrar relación entre variables, etc.

- **Distribución t de Student**

Definición: Si *Y* y *Z* son dos variables aleatorias independientes tales que *Y*~N(y;0,1) y $Z \sim \chi^2_n$, entonces la variable aleatoria $t = \frac{Y}{\sqrt{Z/n}}$ tiene una distribución que se llama *t* de Student con *n* grados de libertad, con función de probabilidad dada por la expresión

$$t \sim t_n(t;n) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{\pi n}}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \qquad -\infty < t < \infty$$

Esta distribución también es simétrica y cuando *n* tiende a infinito la distribución t tiende a la normal estándar, la diferencia entre estas distribuciones es que la t de Student tiene colas más pesadas, es decir se acumula más probabilidad en las colas.

La distribución t de Student está estrechamente relacionada con muestras aleatorias de una distribución normal.

- **Distribución F**

Definición: Si *U* y *V* son dos variables aleatorias independientes tales que $U \sim \chi^2_m$ y $V \sim \chi^2_n$ entonces la variable aleatoria $X = \frac{U/m}{V/n}$ tiene una distribución llamada F con m grados de libertad en el numerador y n grados de libertad en el denominador y se escribe $X \sim F(x; m, n) = F_{m,n}$.

La distribución F tiene gran aplicación al probar hipótesis sobre dos o más distribuciones normales. Esta distribución se usa para hacer inferencias sobre las varianzas poblacionales cuando se tienen dos muestras aleatorias.

Los valores de $P(X \geq x) = \alpha = P(X \geq F_{\alpha,m,n})$ se encuentran tabulados. Esto significa que con los grados de libertad para el numerador y para el

denominador, la tabla brinda información sobre la probabilidad a la derecha del valor F_{α,v_1,v_2}.

La función de distribución es de la forma:

$$f_X(x;m,n)=\frac{\Gamma\left[\frac{1}{2}(m+n)\right]m^{m/2}n^{n/2}}{\Gamma\left(\frac{1}{2}m\right)\Gamma\left(\frac{1}{2}n\right)}\frac{x^{\frac{m}{2}-1}}{(mx+n)^{(m+n)/2}}$$

Observación: $F_{1-\alpha,m,n}=\frac{1}{F_{\alpha,n,m}}$

Estimación estadística

En el capítulo anterior se dieron elementos de Teoría de las Probabilidades que constituyen la base para la *Inferencia Estadística*, que es una rama de la Estadística que tiene como objeto de estudio aquellos procedimientos estadísticos para realizar inferencias sobre las características de diversos fenómenos aleatorios.

En general existen dos tipos de inferencias:

La inferencia deductiva	**Es un juicio o generalización basado en axiomas, teoremas y supuestos teóricos. De esta forma se obtienen conclusiones a partir de un razonamiento o proceso lógico a *priori*.**
La inferencia inductiva	**Es un juicio o generalización derivado de observaciones experimentales o empíricas. Se obtienen conclusiones a partir de un razonamiento a *posteriori*.**

Ejemplo

1) La *definición clásica de probabilidad* es un ejemplo de *inferencia deductiva*, ésta es una definición a priori, se llega a la probabilidad de un suceso a partir de axiomas, sin efectuar ninguna experiencia, sólo a través de razonamientos abstractos.

2) La *definición frecuencial de probabilidad* es un ejemplo de *inferencia inductiva*, sólo se llega a la probabilidad de un suceso después que se ha repetido el experimento un número grande de veces.

Sin embargo, la *Inferencia Estadística* es esencialmente inductiva. En ella se alcanzan conclusiones, generalizaciones respecto a las características de una población, a partir de observaciones empíricas realizadas a una muestra.

En la práctica se dispone de una o varias muestras y a partir de ellas se pretende obtener conclusiones referentes a la población o poblaciones de donde proceden dichas muestras.

Inferencia Estadística: se trata de un proceso inductivo que parte de lo particular, la muestra, para llegar a lo general, la población. Tal proceso, envuelve cierto grado de incertidumbre y supone riesgos o pérdidas debido a decisiones incorrectas.

De acuerdo a los objetivos que se persigan, la Inferencia Estadística comprende dos tipos de técnicas básicas, la *estimación* y *dócima de hipótesis*.

La teoría de la estimación fue fundada por R. A. Fisher y su objetivo esencial es obtener un conocimiento de cierto aspecto de la población sobre el cual se desconocen las características fundamentales. Por ejemplo, determinar la temperatura media de una región, la producción media de cierto artículo, etc.

Existen dos formas de realizar la estimación, la estimación puntual y la estimación por intervalos:

- **La estimación puntual**: como su nombre lo indica, es la designación de un valor o punto, como estimado del aspecto de interés en la población. Por ejemplo, de la media poblacional, de la varianza.
- **La estimación por intervalos**: ofrece un conjunto de valores posibles como estimado. Por ejemplo, puede estimarse que la temperatura media anual de una región se encuentra entre los valores 25 C^0 y los 38 C^0.

Para estudiar con mayor detalle el concepto de estimación, debemos conocer los conceptos de espacio muestral y espacio paramétrico.

El *espacio muestral* (Ω), recordemos que se definía como el conjunto formado por todos los posibles resultados del experimento, por ejemplo, si tenemos una población normal y tomamos de ella una muestra de tamaño n, puede afirmarse que el espacio muestral será $\Re^n$.

El *espacio paramétrico* está relacionado con la distribución de probabilidad de la variable aleatoria que se esté estudiando. Vimos que las leyes o funciones de distribución de probabilidad, dependen de constantes, usualmente llamadas parámetros. Así, por ejemplo, en la distribución Binomial, los parámetros serán n y p, en la distribución Normal, los parámetros serán μ y σ^2.

Entonces, podemos decir que:

Espacio paramétrico: es el conjunto formado por todos los posibles valores de la distribución de probabilidad de una variable aleatoria, que puede tomar sus parámetros y se denotará por la letra griega Θ.

Ejemplo

Los siguientes son espacios paramétricos:

- Distribución Normal: $\Theta = \Re x \Re_+^*$
- Distribución Poisson: $\Theta = \Re_+^*$
- Distribución Bernoulli: $\Theta =]0,1[$

Debemos aclarar que al hacer variar el parámetro de una distribución por todo el espacio paramétrico, se obtienen todas las posibles distribuciones de la familia, es decir, si conocemos que una variable aleatoria X sigue una distribución Normal con parámetros desconocidos, la distribución real de esta variable pertenecerá al conjunto formado por todas las distribuciones normales (familia de distribuciones normales), la cual se denotará como: $\{P_\theta : \theta \in \Theta\}$,

siendo P_θ la distribución de probabilidad que depende del parámetro θ y Θ el conjunto formado por todos los valores que puede tomar θ, es decir, el **espacio paramétrico**.

Como cada variable aleatoria tiene una sola distribución de probabilidad, es obvio que solo una distribución de la familia será la verdadera, y el conocimiento de esta distribución permitirá conocer el valor real del parámetro θ y viceversa.

Estadígrafo: a un valor calculado a partir de los valores observados de una muestra, usualmente, pero no necesariamente como algún estimador de un parámetro poblacional, una función de los valores muestrales.

Estimación por intervalos

Estimación por intervalo: establece un intervalo dentro del cual es muy probable que se encuentre el parámetro poblacional. El coeficiente de confianza se usa para indicar la probabilidad de que una estimación por intervalo contenga al parámetro poblacional. El nivel de confianza es el coeficiente de confianza expresado como un porcentaje.

- **Límites de confianza para la media de una población normal**

Partiendo de que Z~ N(0,1), lo que se busca es encontrar un valor de z_1 tal que el intervalo ($-z_1 < Z < z_1$) tenga alta probabilidad de ocurrencia ($1-\alpha$), es decir:

$$P(-z_1 < Z < z_1) = 1-\alpha \quad \text{(a)}$$

ahora si X~N(μ, σ^2) entonces:

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad \text{(b)}$$

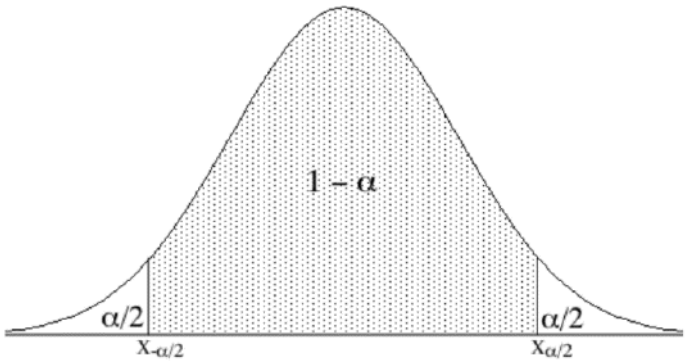
Reemplazando (b) en (a) tenemos que:

$$P(-z_1 < \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} < z_1)$$

$$P(\underbrace{\overline{x} - z_1\sigma_{\overline{x}}}_{l_1} < \mu < \underbrace{\overline{x} + z_1\sigma_{\overline{x}}}_{l_2})$$

Como X es una variable aleatoria, entonces los límites del intervalo, l_1 y l_2, serán también variables aleatorias mientras no se reemplacen los valores obtenidos en una muestra.

Los límites l_1 y l_2 pueden variar de una muestra a otra pudiéndose obtener una situación como la que ilustra el siguiente gráfico.



A. *Intervalo de confianza para* μ *con varianza* σ^2 *conocida*.

Si $\overline{x}$ es la media de una muestra aleatoria de tamaño n de una población normal con varianza σ^2 conocida, un intervalo de confianza de $(1-\alpha)\cdot 100\%$ para μ está dado por:

$$\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

donde $z_{1-\alpha/2}$ es el valor que deja un área de $1-\alpha/2$ a la izquierda.

Al término $\frac{\sigma}{\sqrt{n}}$ se le conoce como el error estándar o desviación estándar del promedio muestral cuando la selección de la muestra es con reemplazo.

Si el muestreo es sin reemplazo y la fracción de muestreo $\frac{n}{N} \geq 0.05$, el error estándar será:

$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ y los límites de confianza se calculan con la siguiente expresión:

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} < \mu < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Ejemplo

Una empresa fabrica focos que tienen una duración aproximadamente normal con desviación estándar de 40 horas. Si una muestra de 30 focos tiene una duración promedio de 780 horas, encuentre un intervalo de confianza de 96% para la media de la población de todos los focos que produce esta empresa.

Solución

La estimación puntual de μ es $\bar{x} = 780$. El valor z, que deja un área 0.980 a la izquierda, es $z_{0.98} = 2.05$.

De aquí que el intervalo de confianza del 96% sea:

$$780 - (2{,}05)\left(\frac{40}{\sqrt{30}}\right) < \mu < 780 + (2{,}05)\left(\frac{40}{\sqrt{30}}\right)$$

Efectuando las operaciones indicadas se tiene: $765 < \mu < 795$

O sea, con un 96% de confianza entre 765 y 795 horas se encontrará la media de la duración de la población de todos los focos que produce la empresa.

B. *Intervalo de confianza para* μ *con varianza* σ^2 *desconocida* (muestra pequeña $n < 30$)

Si $\bar{x}$ y S son la media y la desviación estándar de una muestra aleatoria de tamaño n de una población normal con varianza σ^2 desconocida, un intervalo de confianza de $(1-\alpha)\cdot 100\%$ para μ está dado por:

$$\bar{x} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{S}{\sqrt{n}}$$

donde $t_{\alpha/2}$ es el valor t con (n – 1) grados de libertad, que deja un área de $\alpha/2$ a la derecha.

Al término $\frac{S}{\sqrt{n}}$ se le conoce como la estimación del error estándar o desviación estándar del promedio muestral cuando la selección de la muestra es con reemplazo.

Si el muestreo es sin reemplazo y la fracción de muestreo $\frac{n}{N} \geq 0.05$, el error estándar será:

$\frac{S}{n}\sqrt{\frac{N-n}{N-1}}$ y los límites de confianza se calculan con la siguiente expresión.

$$\bar{x} - t_{\alpha/2}\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} < \mu < \bar{x} + t_{\alpha/2}\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Ejemplo

Una máquina produce piezas metálicas de forma cilíndrica. Se toma una muestra de las piezas y los diámetros son 1,01; 0,97; 1,03; 1,04; 0,99; 0,98; 0,99; 1,01 y 1,03 centímetros. Encuentre un intervalo de confianza de 99% para el diámetro medio de las piezas de esta máquina, suponga una distribución aproximadamente normal.

Solución:

Primero se calculará las estimaciones puntuales de μ y σ^2, es decir el promedio y desviación estándar muestral.

$$x = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1.01 + 0.97 + ... + 1.03}{9} = 1.006$$

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - x)^2}{n-1}} = \sqrt{\frac{(1.01-1.006)^2 + (0.97-1.006)^2 + ... + (1.03-1.006)^2}{8}} = 0.025$$

El valor t con 8 grados de libertad, que deja un área de 0.005 a la derecha es $t_{0.005} = 3.355$. De aquí que el intervalo de confianza del 99% sea:

$$1.006 - (3.355)\left(\frac{0.025}{\sqrt{9}}\right) < \mu < 1.006 + (3.355)\left(\frac{0.025}{\sqrt{9}}\right)$$

Efectuando las operaciones indicadas se tiene: $0{,}98 < \mu < 1{,}03$

O sea, con 99% de confianza entre 0.98 y 1.03 horas se encontrará el diámetro medio de las piezas de la máquina.

C. *Intervalo de confianza para* μ *con varianza* σ^2 *desconocida* (muestra grande $n \geq 30$)

En una variable que se distribuye normal, cuando σ^2 es desconocida y $n \geq 30$, S^2 puede reemplazar a σ^2 y utilizar el intervalo de confianza siguiente:

$$\bar{x} - z_{1-\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2} \frac{S}{\sqrt{n}}$$

donde $z_{1-\alpha/2}$ es el valor que deja un área de $1-\alpha/2$ a la izquierda.

De igual forma $\frac{S}{\sqrt{n}}$ es la estimación del error estándar o desviación estándar del promedio muestral cuando la selección de la muestra es con reemplazo.

Si el muestreo es sin reemplazo y la fracción de muestreo $\frac{n}{N} \geq 0.05$, el error estándar será:

$\frac{S}{n}\sqrt{\frac{N-n}{N-1}}$ y los límites de confianza se calculan con la siguiente expresión:

$$\bar{x} - z_{1-\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} < \mu < \bar{x} + z_{1-\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Ejemplo

Una muestra aleatoria de 100 propietarios de automóviles muestra que, en el estado de Virginia, un automóvil se maneja, en promedio, 23 500 kilómetros por año con una desviación estándar de 3 900 kilómetros. Construya un intervalo de confianza de 99% para el número promedio de kilómetros que se maneja un automóvil anualmente en Virginia.

Solución:

Las estimaciones puntuales de μ y σ^2 son respectivamente $\bar{x} = 23500$ y $S = 3900$. El valor z, que deja un área de 0.005 a la derecha y por lo tanto tiene un

área de 0.995 a la izquierda, es $z_{0.995} = 2.58$. De aquí que el intervalo de confianza del 99% sea:

$$23500 - (2{,}58)\left(\frac{3900}{\sqrt{100}}\right) < \mu < 23500 + (2.58)\left(\frac{3900}{\sqrt{100}}\right)$$

Efectuando las operaciones indicadas se tiene: $22\,493.8 < \mu < 24\,506.2$

O sea, con 99% de confianza entre 22 493.8 y 24 506.2 se encontrará el número promedio de kilómetros manejados por los propietarios de automóviles en Virginia.

Teorema: *Tamaño de muestra cuando la varianza poblacional es conocida*

Si $\bar{x}$ se usa como estimación de μ, podemos tener $(1-\alpha)\cdot 100\%$ de confianza de que el error no exceda una cantidad específica e cuando el tamaño de la muestra es:

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2$$

Si el cálculo del tamaño de muestra resulta un valor con decimales, se debe redondear al siguiente número entero.

Observación: si el muestreo es sin reemplazo, el tamaño de muestra se calcula con la siguiente expresión:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{donde } n_0 = \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2 \text{ y N es el tamaño de la población.}$$

Si la fracción de muestreo $\frac{n_0}{N} < 0.05$ entonces n ≈ n_0

Ejemplo

¿De qué tamaño se necesita una muestra si deseamos tener 98% de confianza que la media muestral esté dentro de 10 horas de la media real?

Solución

Como ya se calculó el valor de $Z_{0.98}$ = 2,05 y se tiene el dato que la desviación estándar poblacional es 40, entonces el tamaño de muestra para un error de 10 horas es

$$n = \left(\frac{(2.05)(40)}{10} \right)^2 = 67.24$$

entonces, el tamaño de muestra para las condiciones solicitadas será 68.

Teorema: *Tamaño de muestra cuando la varianza poblacional es desconocida*
Si $\overline{x}$ y S son las estimaciones de μ y σ^2, repectivamente, podemos tener $(1-\alpha)\cdot 100\%$ de confianza de que el error no exceda una cantidad específiva *e* cuando el tamaño de la muestra es:

$$n = \left(\frac{z_{1-\alpha/2} S}{e} \right)^2$$

El valor de S puede ser obtenido a partir de una muestra preliminar de por lo menos 30 elementos.

Observación: si el valor del tamaño de muestra es decimal se debe redondear al siguiente número entero.

- **Intervalo de confianza para la varianza de una población normal**

Si S^2 es la varianza de una muestra aleatoria de tamaño n de una población normal, un intervalo de confianza de $(1-\alpha)\cdot 100\%$ para σ^2 es:

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

donde $x^2_{\alpha/2}$ y $x^2_{1-\alpha/2}$ son valores X^2 con v = n – 1 grados de libertad.

Ejemplo

Un fabricante de baterías para automóviles afirma que sus baterías durarán en promedio tres años, con una varianza de un año. Si cinco de estas baterías tienen duraciones de 1,9; 2,4; 3,0; 3,5 y 4,2 años, construya un intervalo de confianza del 95% para σ^2 y decida si la afirmación del fabricante de que $\sigma^2 = 1$ es válida. Suponga que la población de duraciones de las baterías es de forma aproximadamente normal.

Solución

La estimación puntual de σ^2 es S^2 = 0,815.

El valor $X^2_{1-\alpha/2}$ es $X^2_{0.025} = 11.143$; y el valor $X^2_{\alpha/2}$ es igual a $X^2_{0.975} = 0.484$

De aquí que el intervalo de confianza del 95% sea:

$$\frac{(5-1)\,(0.815)}{11.143} < \sigma^2 < \frac{(5-1)\,(0.815)}{0.484}$$

Efectuando las operaciones indicadas se tiene: $0.3 < \sigma^2 < 6.7$

O sea, con 95% de confianza entre 0.3 y 6.7 se encontrará la varianza de la duración de la baterías. Por lo tanto es válida la afirmación del fabricante porque el intervalo hallado contiene a la unidad.

- **Intervalo de confianza para la proporción poblacional.**

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n y $\hat{q} = 1 - \hat{p}$. Un intervalo de confianza aproximado de $(1-\alpha)\cdot 100\%$ para el parámetro binomial p está dado por:

$$\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

donde $z_{1-\alpha/2}$ es el valor z que deja un área de $1-\alpha/2$ a la izquierda.

Si el muestreo es sin reemplazo y la fracción de muestreo $\frac{n}{N} \geq 0.05$, los límites de confianza se calculan con la siguiente expresión:

$$\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}$$

Ejemplo

Un genetista se interesa en la proporción de hombres africanos que tienen cierto trastorno sanguíneo menor. En una muestra aleatoria de 100 hombres africanos, se encuentran que 24 lo padecen.

a) Calcule un intervalo de confianza de 99% de confianza para la proporción de hombres africanos que tienen este desorden sanguíneo.

b) ¿Qué se puede asegurar con 99% de confianza acerca de la posible magnitud de nuestro error si estimamos que la proporción de hombres africanos con este trastorno sanguíneo es 0.24?

Solución

a) La estimación puntual de p es $\hat{p} = \frac{24}{100} = 0.24$. El valor z, que deja un área de 0.005 a la derecha y por lo tanto un área de 0.995 a la izquierda, es $z_{0.995} = 2.58$. De aquí que el intervalo de confianza del 99% sea:

$$0.24 - (2.58)\sqrt{\frac{(0.24)(0.76)}{100}} < p < 0.24 + (2.58)\sqrt{\frac{(0.24)(0.76)}{100}}$$

Efectuando las operaciones indicadas se tiene: $0.13 < \mathrm{p} < 0.35$

O sea, con 99% de confianza entre 0.13 y 0.35 se encontrará la proporción de hombres africanos que tienen este desorden sanguíneo.

b) Si la proporción de hombres africanos con trastorno sanguíneo menor es 0.24, la magnitud del error es:

$$e = (2.58)\sqrt{\frac{(0.24)(0.76)}{100}} = 0.11$$

A. *Calculo del tamaño de muestra para estimar un proporción a partir de la información muestral*

Si $\hat{p}$ se utiliza como una estimación de p, podemos tener una confianza del $(1-\alpha)\cdot 100\%$ de que el error será menor de una cantidad específica e cuando el tamaño de la muestra es aproximadamente:

$$n = \frac{z^2_{\alpha/2}\hat{p}\hat{q}}{e^2}$$

Observación: si el muestreo es sin reemplazo, el tamaño de muestra se calcula con la siguiente fórmula:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

donde $n_0 = \frac{z^2_{1-\alpha/2}\hat{p}\hat{q}}{e^2}$ y N es el tamaño de la población.

Si la fracción de muestreo $\frac{n_0}{N} < 0.05$ entonces n ≈ n_0

Ejemplo

Fueron encuestados 1600 adultos sobre un artículo del periódico, 32% dijeron que el programa espacial debe enfatizar la exploración científica ¿Qué tan grande se necesita que sea la muestra de adultos en la encuesta si se desea tener una confianza de 95% de que el porcentaje estimado esté dentro de 2% del porcentaje real?

Solución

El valor de $Z_{1-\alpha/2}=1.96$ y la estimación del porcentaje de adultos que manifiestan se debe enfatizar en la exploración científica es 32%, entonces el tamaño de muestra para un error de 2% es

$$n=\frac{1.96^2(0.32)(0.68)}{(0.02)^2}=2089.8$$

Entonces el tamaño de muestra con las condiciones solicitadas será 2090.

B. *Cálculo del tamaño de muestra para estimar un proporción sin utilizar la información muestral*

El valor de $\hat{p}\hat{q}$ se hace máximo cuando $\hat{p}=\frac{1}{2}$, por lo tanto la fórmula para calcular el tamaño de muestra queda de la siguiente manera:

$$n=\frac{z^2_{1-\alpha/2}}{4e^2}$$

Ejemplo

Se lleva a cabo un estudio para estimar el porcentaje de ciudadanos de una ciudad que están a favor de que el agua se trate con flúor. ¿Qué tan grande se necesita que sea la muestra si se desea tener una confianza de 95% de que la estimación esté dentro del 1% del porcentaje real?

Solución

El valor de $Z_{1-\alpha/2}=1.96$, por lo que el tamaño de muestra para un error de 1% es

$$n = \frac{1.96^2(0.5)(0.5)}{(0.01)^2} = 9604$$

Entonces, el tamaño de muestra para las condiciones solicitadas será 9604.

- **Intervalos de confianza para la diferencia entre dos medias de una variable normal**

A. *Cuando las varianzas poblacionales son conocidas*

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes de tamaño n_1 y n_2 de poblaciones normales con varianzas conocidas σ_1^2 y σ_2^2, respectivamente, un intervalo de confianza de $(1-\alpha)\cdot 100\%$ para $\mu_1 - \mu_2$ está dado por:

$$\left(\bar{X}_1 - \bar{X}_2\right) - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \left(\bar{X}_1 - \bar{X}_2\right) + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

donde $z_{1-\alpha/2}$ es el valor que deja un área de $1-\alpha/2$ a la izquierda.

Al término $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ se le conoce como el error estándar o desviación estándar de la diferencia entre dos promedios muestrales cuando la selección de la muestra es con reemplazo.

Si *el muestreo es sin reemplazo* y las fracciones de muestreo $\frac{n_1}{N_1} \geq 0.05$ y $\frac{n_2}{N_2} \geq 0.05$, el error estándar será:

$$\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}$$

y los límites de confianza se calculan con la siguiente expresión:

$$LC(\mu_1 - \mu_2) = \left(\bar{X}_1 - \bar{X}_2\right) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2 - 1}\right)}$$

Ejemplo

Para comparar dos métodos de la enseñanza de las matemáticas, se aplicaron a 200 alumnos elegidos al azar el método tradicional y a otra muestra de 250 alumnos el método nuevo resultando las calificaciones promedio respectivos de 13 y 15. Suponga que las varianzas poblacionales respectivas son 9 y 16. Utilizando un intervalo de confianza del 95% para la diferencia de las medias, ¿podemos afirmar que no hay diferencias significativas entre los dos métodos?, si hay diferencias, ¿podemos afirmar que el método nuevo es mejor que el método tradicional? Suponga que la distribución de las variables poblacionales es normal.

Solución:

La estimación puntual de $\mu_1 - \mu_2$ es $\overline{X}_1 - \overline{X}_2 = 13 - 15 = -2$. Con 0,05 se encuentra el valor z, que deja un área de 0,025 a la derecha y por lo tanto un área de 0.975 a la izquierda, es $z_{0.975} = 1.96$. De aquí que el intervalo de confianza del 96% sea:

$$(-2) - 1.96\sqrt{\frac{9}{200} + \frac{16}{250}} < \mu_1 - \mu_2 < (-2) + 1.96\sqrt{\frac{9}{200} + \frac{16}{250}}$$

Efectuando las operaciones indicadas se tiene: $-2.6 < \mu_1 - \mu_2 < -1.3$

O sea, con 95% de confianza entre -2,6 y -1,3 se encontrará la diferencia de calificaciones medias obtenidas con los métodos evaluados. Como el intervalo calculado contiene valores negativos, se puede concluir que el método nuevo es mejor que el tradicional.

B. *Cuando las varianzas poblacionales son iguales pero desconocidas*

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes de tamaño n_1 y n_2 respectivamente, de poblaciones aproximadamente normales con varianzas iguales pero desconocidas, un intervalo de confianza de $(1-\alpha)\cdot 100\%$ para $\mu_1 - \mu_2$ está dado por:

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

donde S_p es la estimación de unión de la desviación estándar poblacional y se calcula mediante la siguiente expresión

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

y $t_{\alpha/2}$ es el valor t con v = n_1 + n_2 – 2 grados de libertad, que deja un área de $\alpha/2$ a la derecha.

Ejemplo

Los siguientes datos, registrados en días, representan el tiempo de recuperación para pacientes que se tratan al azar con uno de dos medicamentos para curar infecciones graves de la vejiga:

Medicamento 1	Medicamento 2
$n_1 = 14$	$n_2 = 16$
$x_1 = 17$	$x_2 = 19$
$s_1^2 = 1.5$	$s_2^2 = 1.8$

Encuentre un intervalo de confianza de 99% para la diferencia $\mu_1 - \mu_2$ del tiempo promedio de recuperación para los dos medicamentos, suponga poblaciones normales con varianzas iguales.

Solución:

La estimación puntual de $\mu_1 - \mu_2$ es $\bar{x}_1 - \bar{x}_2 = 19 - 17 = 2$.

La estimación de la varianza común, S^2_p, es

$$S_p^2 = \frac{(14-1)(1.5) + (16-1)(1.8)}{14 + 16 - 2} = 1.661$$

Al tomar la raíz cuadrada obtenemos S_p = 1.289. Con el uso de $\alpha = 0.01$, encontramos que $t_{0.005}$ =2.763 para v = 14 + 16 – 2 = 28 grados de libertad, y por lo tanto el intervalo de confianza del 99% es:

$$2 - 2.763\,(1{,}2887)\sqrt{\frac{1}{14} + \frac{1}{16}} < \mu_2 - \mu_1 < 2 + 2.763\,(1{,}2887)\sqrt{\frac{1}{14} + \frac{1}{16}}$$

Efectuando las operaciones indicadas se tiene: $0.7 < \mu_2 - \mu_1 < 3.3$

O sea, con 99% de confianza entre 0,7 y 3,3 días se encontrará la diferencia de tiempos promedios de recuperación para los dos tipos de medicamentos.

C. *Cuando las varianzas poblacionales son desconocidas y diferentes*

Si $\bar{X}_1 \; y \; S_1^2 \;\; y \;\; \bar{X}_2 \; y \; S_2^2$ son las medias y varianzas de muestras pequeñas e independientes de distribuciones aproximadamente normales con varianzas desconocidas y diferentes, un intervalo de confianza de $(1-\alpha)\cdot 100\%$ para $\mu_1 - \mu_2$ está dado por:

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

donde $t_{\alpha/2}$ es el valor t con $v = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\frac{S_1^2}{n_1}\right)^2}{(n_1 - 1)} + \dfrac{\left(\frac{S_2^2}{n_2}\right)^2}{(n_2 - 1)}}$ grados de libertad, que deja un área de $\alpha/2$ a la izquierda.

Ejemplo

Una compañía de taxis trata de decidir si comprar neumáticos de la marca A o de la B para su flotilla de taxis. Se lleva a cabo un experimento utilizando 12 de cada marca. Los neumáticos se utilizaron hasta que se gastan. Los resultados son:

Marca A	Marca B
$\bar{x}_1 = 36\,300$ kilómetros	$\bar{x}_2 = 38\,100$ kilómetros
$s_1^2 = 5\,000$ kilométros	$s_2^2 = 6\,100$ kilométros

Calcule un intervalo de confianza de confianza de 90% para la diferencia de rendimiento promedio de ambas marcas de neumáticos. Suponga que la diferencia de kilómetros de rendimiento se distribuyen de forma aproximadamente normal con varianzas distintas.

Solución:

Representamos con $\mu_1 \text{ y } \mu_2$ las medias poblacionales, respectivamente, para los tiempos promedios de duración de las películas que producen las compañía A y B.

La estimación puntual de $\mu_1 - \mu_2$ es $\bar{x}_1 - \bar{x}_2 = 36\,300 - 38\,100 = -1\,800$

Como las varianzas son desconocidas y diferentes, debemos encontrar un intervalo de confianza de 90% aproximado basado en la distribución t con v grados de libertad, donde

$$v = \frac{\left(\frac{5000}{12} + \frac{6100}{12}\right)^2}{\frac{\left(\frac{5000}{12}\right)^2}{(12-1)} + \frac{\left(\frac{6100}{12}\right)^2}{(12-1)}} = 21.79 \approx 22$$

Con el uso de $\alpha = 0.10$, encontramos que $t_{0.05} = 1.717$ para v = 22 grados de libertad, y por lo tanto el intervalo de confianza del 90% es:

$$-1800 - 1.717\sqrt{\frac{5000}{12} + \frac{6100}{12}} < \mu_1 - \mu_2 < -1800 + 1.717\sqrt{\frac{5000}{12} + \frac{6100}{12}}$$

Efectuando las operaciones indicadas se tiene: $-1852.2 < \mu_1 - \mu_2 < -1747.8$

O sea, con 90% de confianza entre -1852 y -1748 días se encontrará la diferencia de rendimiento promedio de ambas marcas de neumáticos.

- **Muestras relacionadas**

La prueba de dos medias puede llevarse a cabo cuando los datos están en forma de observaciones pareadas.

Un intervalo de $(1-\alpha)\cdot 100\%$ de confianza para la diferencia de medias cuando las muestras están relacionadas es:

$$\bar{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_1 - \mu_2 < \bar{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

donde $t_{\alpha/2}$ es el valor t con n-1 grados de libertad, que deja un área de $\alpha/2$ a la derecha.

- **Intervalos de confianza para el cociente de varianzas**

Si S^2_1 y S^2_2 son las varianzas de muestras independientes de tamaño n_1 y n_2 respectivamente, de poblaciones normales, entonces un intervalo de confianza de $(1-\alpha)\cdot 100\%$ para σ_1^2/σ_2^2 es:

$$\frac{S_1^2}{S_2^2}\cdot\frac{1}{f_{(v_1,v_2,,\alpha/2)}}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{S_1^2}{S_2^2}\cdot f_{(v_2,v_1,\alpha/2)}$$

donde $f_{(v_2,v_1,\alpha/2)}$ es un valor f con $v_2 = n_2 - 1$ y $v_1 = n_1 - 1$ grados de libertad que deja un área de $\alpha/2$ a la derecha, $f_{(v_1,v_2,\alpha/2)}$ es un valor f similar con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

Ejemplo

Construya un intervalo de confianza de 98% para σ_1^2/σ_2^2 en el ejercicio 7 de la página 263, donde σ_1^2 y σ_2^2 son, respectivamente, las varianzas para los tiempos de recuperación de pacientes que recibieron los medicamentos 1 y 2.

Solución

Tenemos que $n_1 = 14$, $n_2 = 16$, $S_1^2 = 1.5$ y $S_2^2 = 1.8$. Para un intervalo de confianza de 98%. Al observar la tabla de valores f no se encuentra el valor correspondiente a $f_{(13,15,0.01)}$, por lo tanto se deberá interpolar de la siguiente manera:

$$\begin{array}{l} f_{0.99(12,15)} = 3.67 \\ f_{0.99(13,15)} = ? \\ f_{0.99(15,15)} = 3.52 \end{array} \Rightarrow \begin{array}{l} \frac{1}{12}\ldots\ldots\ldots\ldots 3.67 \\ \frac{1}{13}\ldots\ldots\ldots\ldots ? \\ \frac{1}{15}\ldots\ldots\ldots\ldots 3.52 \end{array} \Rightarrow \frac{\left(\frac{1}{12}-\frac{1}{15}\right)}{\left(\frac{1}{12}-\frac{1}{13}\right)} = \frac{(3.67-3.52)}{(3.67-?)} \Rightarrow ? = 3.61$$

Entonces $f_{0.99(13,15)} = 3.61$ y $f_{(15,13,0.99)} = 3.82$

El intervalo será: $\frac{1.5}{1.8}\cdot\frac{1}{3.61}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{1.5}{1.8}\cdot 3.82$

Efectuando las operaciones indicadas se tiene: $0.23<\frac{\sigma_1^2}{\sigma_2^2}<3.18$

O sea, con 98% de confianza entre 0.23 y 3.18 se encontrará el cociente de varianzas de los tiempos de recuperación para los dos tipos de medicamentos.

Observación: en este intervalo de confianza si aproximadamente $\frac{\sigma_1^2}{\sigma_2^2}=1$ entonces se podría pensar que existe homogeneidad de varianzas.

- **Intervalos de confianza para la diferencia de proporciones**

Si $\hat{p}_1$ y $\hat{p}_2$ son las proporciones de éxitos en muestras aleatoria de tamaño n_1 y n_2, respectivamente, además $\hat{q}_1=1-\hat{p}_1$ y $\hat{q}_2=1-\hat{p}_2$, un intervalo de confianza aproximado de $(1-\alpha)\cdot 100\%$ para la diferencia de dos parámetros binomiales $p_1 - p_2$, está dado por:

$$(\hat{p}_1-\hat{p}_2)-z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}<p_1-p_2<(\hat{p}_1-\hat{p}_2)+z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$$

donde $z_{\alpha/2}$ es el valor z que deja un área de $\alpha/2$ a la derecha.

Ejemplo

Una encuesta de 1000 estudiantes concluye que 274 eligen al equipo profesional de beisbol A como su equipo favorito. En 1991, se realizó la misma encuesta con 760 estudiantes. Concluyó que 240 de ellos también eligieron al equipo A como su favorito. Calcule un intervalo de confianza del 95% para la diferencia entre la proporción de estudiantes que favoren al equipo A entre las dos encuestas ¿Hay una diferencia significativa?

Solución:

Sean p_1 y p_2 las proporciones reales de de estudiantes que eligieron al equipo A como su favorito para la encuesta actual y la de 1991, respectivamente. De

aquí $\hat{p}_1 = \frac{274}{1000} = 0.274$ y $\hat{p}_2 = \frac{240}{760} = 0.316$. El valor z, que deja un área de 0.975 a la izquierda, es $z_{0.975} = 1.96$.

De aquí que el intervalo de confianza de 95% sea:

$$IC(p_1 - p_2) = (0.274 - 0.316) \pm (1.96)\sqrt{\frac{(0.274)(0.726)}{1000} + \frac{(0.316)(0.684)}{760}}$$

Efectuando las operaciones indicadas se tiene: $-0.085 < p_1 - p_2 < -0.042$

O sea, con 95% de confianza entre 8.5% y 4% habrá aumentado la preferencia por elegir al equipo A.

Técnicas de muestreo probabilístico

Se puede considerar la teoría de muestreo como coexistente con los métodos estadísticos modernos. Casi todos los adelantos estadísticos modernos se refieren a inferencias que se pueden efectuar respecto a la población, cuando se dispone de información sólo de una muestra de dicha población. A continuación se mencionan algunas de las formas en que esto se refleja:

a) *Trabajo de investigación*. En la mayoría de los trabajos de investigación la población se compone de todas las personas (o establecimientos industriales, viviendas, hogares, etc.) en una ciudad u otras áreas. Se obtiene o se desea información de una muestra de la población, pero se requieren inferencias sobre las características de toda la población.

b) *Diseño y análisis de experimentos*. En el diseño y análisis de experimentos, la población representa todas las posibles aplicaciones de varias técnicas alternativas que pueden usarse. Por ejemplo, el experimento puede ser agrícola en el cuál se investiga el efecto de varios fertilizantes. La población es infinita, debido a que representa el uso de fertilizantes en todas las unidades agrícolas de cualquier época. El problema consiste en diseñar experimentos de

modo que se disponga del máximo de información para realizar inferencias respecto a la población total a base de una muestra de tamaño limitado.

c) *Control de Calidad*. Para la aplicación de los métodos de control de calidad en un establecimiento industrial, por ejemplo, la población de todo producto que sale de una máquina. Se necesita inferencias sobre la forma como los productos cumplen con las especificaciones. El término de «control de calidad» se aplica también a una verificación por muestreo, sobre la calidad de trabajo de campo efectuado en una encuesta por muestreo; la verificación por muestreo se ejecuta después que se ha completado la muestra.

Las siguientes son las principales ventajas del uso de muestras:

- Puede ahorrar dinero (comparado con el costo de un censo de enumeración completa) cuando no se requiere gran precisión.
- Ahorra tiempo, cuando los datos se necesitan oportunamente.
- Permite concentrar la atención en casos particulares o individuales.
- En ensayos destructivos solamente se utiliza una muestra de ítems.
- Algunas poblaciones consideradas infinitas solo pueden ser muestreadas.

Sin embargo, el muestreo presenta las siguientes limitaciones:

- En áreas muy pequeñas se requieren muestras desproporcionadamente grandes, pues la precisión de una muestra depende fuertemente del tamaño de la muestra y no de la tasa de muestreo. En este caso, el muestreo puede ser tan costoso como un censo completo.
- Si los datos se necesitan a intervalos regulares de tiempo, y es importante medir cambios muy pequeños de un período a otro, se necesitarán muestras grandes.

Ahora bien, existen criterios para la aceptabilidad del muestreo, pues los métodos modernos de muestreo pueden proporcionar datos de confiabilidad conocida con eficiencia y economía. Para aceptar una muestra es necesario

que ésta **represente a la población**, que tenga una confiabilidad que se pueda medir y que responda a un plan práctico y eficaz.

En tal dirección, los principales criterios a tener en cuenta son los siguientes:

a) *Probabilidad de selección*. La muestra debe seleccionarse de todo que represente adecuadamente a la población. Cada Unidad debe tener una probabilidad conocida de ser elegida y esta probabilidad debe ser siempre distinta de cero, o sea, $0 < p \leq 1$.

b) *Confiabilidad medible.* Los valores de la muestra deben proporcionar medidas de la confiabilidad de las estimaciones que se calculan y de la precisión que se espera o se desea que tengan.

c) *Viabilidad o factibilidad*. El plan de muestreo adoptado debe ser práctico y permitir que sea realizado en la forma proyectada.

d) *Economía y eficiencia.* El diseño muestral debe ser eficiente, o sea que sea capaz de proporcionar la mayor cantidad de información al menor costo, para lo cual debe hacerse el uso más efectivo de los recursos disponibles.

Conceptos básicos del muestreo estadístico

Existen varios conceptos que posibilitan comprender la teoría de muestreo y aplicar sus técnicas. A continuación se presentan los más importantes:

a) *Unidad de análisis*: *son las unidades para las cuales se desea obtener información estadística.*

   Estas pueden ser personas, hogares, empresas, establecimientos, unidades agropecuarias, etc. también pueden ser fichas personales o los productos que salen de un proceso mecánico para otros tipos de análisis. La unidad de análisis se llama frecuentemente elemento de la población.

b) *Universo: conjunto de individuos, objetos o unidades.*

   Ejemplo*: los trabajadores de una empresa, los individuos de un país.*

c) *Población: conjunto de todos los individuos o elementos que cumplen o satisfacen la o las características en estudio. Un mismo universo, puede contener varias poblaciones. La población puede ser finita o infinita.*

d) *Muestra*: está constituida por una parte de los individuos o elementos que componen la población. Una muestra será representativa, si nos proporciona una información global acerca de algunas características observables en la población. Sólo cuando la muestra es representativa, podrá inferirse importantes conclusiones en la población a partir de su análisis.

e) *Unidad de muestreo*: es una unidad seleccionada del Marco Muestral. Puede ser igual a la unidad de análisis, aunque no necesariamente.

f) *Marco muestral: es la totalidad de las unidades de muestreo entre las cuales se seleccionará la muestra.*

El marco puede ser una lista de personas o de unidades de vivienda, un archivo de registros o un conjunto de fichas personales, puede ser un mapa subdividido, o puede ser un directorio de nombres y direcciones en cinta magnética para computadora.

g) *Diseño muestral:* es el conjunto de métodos muestrales más apropiados elegidos para la determinación del tamaño de muestra óptima. La mecánica de selección de la muestra, la obtención de los coeficientes de expansión de los valores muestrales, estimaciones así como la determinación de los errores muestrales y la evaluación de los errores no muestrales.

h) *Coeficiente de expansión*: es el valor inverso de la fracción de muestreo, se utiliza como coeficiente de los datos muestrales para estimar totales poblacionales.

i) *Estimador de un parámetro*: es una función matemática de las observaciones muestrales cuya distribución de probabilidad se concentra al rededor del parámetro.

j) *Estimación de un parámetro*: es el valor experimental que toma el estimador de una muestra determinada y sirve como aproximación al valor del Parámetro.

k) *Parámetro*: es una función matemática de las observaciones de las unidades muestrales de toda la población.

l) Desviación estándar o típica (σ): Medida de la variabilidad en la población. Su cuadrado es la varianza (σ^2).

m) *Error muestral (error estándar de estimación)*: representa una medida de la variación media de las estimaciones de muestreo alrededor de su valor teórico. En tal sentido dicho error muestral está ligado a la confiabilidad de las estimaciones obtenidas.

El valor del error muestral indica la precisión de la estimación de los parámetros. Mientras más pequeño es el error muestral, mayor es la precisión de la estimación. El error muestral es controlado con el diseño muestral utilizado.

n) Errores no muestrales: son los errores que se generan durante el recojo, crítica, codificación, procesamiento de la información. La magnitud de los errores no muestrales debe ser evaluado a fin de medir el alcance y cobertura de una encuesta.

ñ) Intervalo de confianza: es el intervalo alrededor del cual se espera que esté el verdadero valor poblacional con más probabilidad fija de acertar.

En los problemas prácticos de muestreo, cuando se usa una muestra razonablemente grande (por lo general 100 ó más), la distribución de los resultados muestrales a través de todas las muestras posibles se aproxima bastante fielmente a la distribución normal (curva con forma de campana). En esta distribución son conocidas las probabilidades de quedar a una distancia fija del valor medio. Esas probabilidades dependen del valor del error estándar. Así, la probabilidad de estar dentro de una unidad el error estándar

es de 68 %; de 2 unidades el error estándar, es de 95 %; de tres unidades el error estándar es de 99.7 %.

$$\hat{\theta} \pm Z\sqrt{Var(\hat{\theta})}\ 1$$

Z = 1.64 para un 90% de confianza

Z = 1.96 para un 95% de confianza

Z = 2.58 para un 99% de confianza

o) Evaluación de la muestra: consiste en el análisis estadístico del comportamiento de la muestra en el terreno en función de las unidades muestrales.

p) Estadígrafos: un estadígrafo es una cantidad obtenida a partir de una muestra de observaciones de una característica, generalmente con el propósito de realizar una inferencia sobre la población.

La característica puede ser cualquier variable asociada con un miembro de la población tal como: edad, ingreso, nivel de educación, situación en el empleo, etc.; la cantidad puede ser un total, un promedio, una mediana u otro percentil. También puede ser una tasa de cambio, un porcentaje, una desviación estándar, o puede ser cualquier otra cantidad cuyo valor se desea estimar para la población.

Muestreo aleatorio simple (M.A.S.)

Muestreo Aleatorio Simple: es un procedimiento de selección aleatoria simple de una muestra por el cual todos y cada uno de los ítems de la población tienen una oportunidad igual e independiente de ser incluidos en la misma.

Teóricamente, todos y cada uno de los ítems individuales extraídos deben ser medidos, registrados y devueltos a la población durante la selección de una muestra aleatoria simple antes que se realice otra selección.

El muestreo aleatorio simple se usa en poblaciones suficientemente homogéneas, es decir, cuya varianza poblacional tienda a cero, exige de disponer de una lista enumerada de 1 a N y de allí, mediante un experimento aleatorio, seleccionar a cada uno de los n elementos de la muestra.

Dos factores afectan la cantidad de información contenida en la muestra y por tanto, la precisión:

- ✓ El tamaño de la muestra.
- ✓ La cantidad de variación que se controla por el tipo de muestreo.

Notación básica:

y_i medida de interés en el i-ésimo elemento de la muestra.

N tamaño de la población total.

u_i elemento genérico de la población.

$Y = \sum_{i=1}^{N} y_i$ total poblacional

$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ media poblacional

$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ Media muestral

Suponga que $y_1, y_2, \ldots, y_n$ es una muestra aleatoria simple de una población de valores $u_1, u_2, \ldots, u_N$, considere que y_i es una muestra aleatoria de tamaño uno. Entonces:

- Media poblacional μ:

$$\mu = E(y_i) = \sum_{i=1}^{N} \frac{u_i}{N} = \overline{Y}$$

- Varianza poblacional σ^2:

$$\sigma^2 = E[Y_i - \mu]^2 = \sum_{i=1}^{N} (y_i - \overline{Y})^2 \frac{1}{N} = \frac{1}{N}\left(\sum_{i=1}^{N} y_i^2 - N\overline{Y}^2\right)$$

Se emplea como varianza poblacional la siguiente función de σ^2:

$$S^2 = \frac{N}{N-1}\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2$$

- La varianza muestral es: $S^2 = \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{Y})^2$

Estimación de la media poblacional:

- Estimador de la media poblacional μ, $\hat{\mu} = \overline{y}$ donde $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

  1) La media muestral es un estimador insesgado, es decir $E(\overline{y}) = \mu$.

  2) La varianza de la media es $V(\overline{y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right)$ y su estimador es $\hat{V}(\overline{y}) = \frac{\hat{S}^2}{n}\left(\frac{N-n}{N}\right)$ que también es insesgado.

- El límite de error de estimación $l = e = t_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(\overline{y})}$

- Tamaño de la muestra para estimar la media:

$$n = \frac{S^2}{\frac{e^2}{t_{1-\alpha}^2} + \frac{S^2}{N}}$$

Observación: tanto en el caso de muestras para estimar el total o la media se supone que el investigador debe conocer S^2.

Ejemplo

Se desea estimar el tiempo medio en recorrer una vuelta a una pista de 400 metros por fumadores de más de 20 cigarrillos diarios, con edades comprendidas entre 35 y 40 años, con una precisión de 5 segundos. Ante la ausencia de cualquier información acerca de la variabilidad del tiempo medio de la extracción de sangre es este tipo de individuos, se tomó una muestra preliminar de 5 individuos, de una población de tamaño 100, en los que se obtuvieron los siguientes tiempos (en segundos): 97, 80, 67, 91, 73.

Determinar el tamaño mínimo de muestra, al 95%, para cumplir el objetivo anterior.

Solución:

$$\hat{S}^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}\left(y_i - \bar{y}\right)^2 = 153{,}8$$

$$n = \frac{S^2}{\frac{e^2}{t_{1-\alpha}^2} + \frac{S^2}{N}} = \frac{153{,}8}{\left(\frac{5}{1{,}64}\right)^2 + \frac{153{,}8}{100}} = \frac{12{,}40}{9{,}30 + 1{,}538} = 14{,}19$$

Por lo tanto: n = 15

- Estimación del total poblacional

  Estimador de Y: $\hat{Y} = N\bar{y}$

  1) El estimador del total es un estimador insesgado, es decir $E\left(N\bar{y}\right) = Y$.

  2) La varianza de $\hat{T}$ es $V\left(\hat{T}\right) = V\left(N\bar{y}\right) = N^2 V\left(\bar{y}\right) = N^2\frac{S^2}{n}\left(\frac{N-n}{N}\right)$ y su estimador es

  $\hat{V}\left(\hat{T}\right) = N^2\frac{\hat{S}^2}{n}\left(\frac{N-n}{N}\right)$ que también es insesgado.

- El límite de error de estimación $e = t_{1-\alpha} * \sqrt{\hat{V}\left(\hat{T}\right)}$
- Tamaño de la muestra para estimar el total:

$$n = \frac{N^2 S^2}{\frac{e^2}{t_{1-\alpha}^2} + NS^2}$$

Muestreo aleatorio estratificado

En el muestreo aleatorio simple la varianza del estimador depende del tamaño de la muestra y de la dispersión de la variable en estudio. Sin embargo, si la población es muy heterogénea y las consideraciones de costos limitan el tamaño de la muestra, podría ser imposible obtener una estimación lo suficientemente precisa tomando una muestra aleatoria simple. Es decir, el tamaño de la muestra aumenta para una precisión dada.

Pero si podemos clasificar los elementos de la población en grupos (estratos) de manera que se reduzca la variación de la variable Y dentro de cada estrato, entonces puede hacerse una mejor estimación.

Una muestra aleatoria estratificada: es la obtenida mediante la división de la población en subpoblaciones denominadas estratos, en la cual, dentro de cada estrato se selecciona en forma independiente una muestra irrestricta aleatoria. Calculándose para cada estrato sus estimadores y el estimador de la población se calcula como una ponderación adecuada de las estimaciones por estrato.

Las principales razones para estratificar son las siguientes:

Aumentar la precisión de las estimaciones al disminuir la variación dentro de los estratos. La estratificación puede producir un límite más pequeño para el error de estimación que el que se produciría con un muestreo aleatorio simple.

Disminuir los costos al estratificar y variar las fracciones de muestreo dentro de los estratos.

Permitir definir los estratos como dominios de estudio y obtener estimaciones con precisión conocida para los estratos.

¿Cómo seleccionar una muestra aleatoria estratificada?

Dividir la población en estratos de acuerdo a las razones para estratificar, ubicar cada unidad muestral en su respectivo estrato, asignar el tamaño muestral de cada estrato n_i de modo que si L es la cantidad de estratos y n es el tamaño de la muestra, $n = \sum_{i=1}^{L} n_i$ y seleccionar muestras aleatorias simples en cada estrato de forma independiente.

La estratificación se realiza de acuerdo a:

- La distribución de la variable en estudio.
- Una variable X altamente correlacionada con la variable en estudio.
- Un criterio de disminución de los costos.

En general, la precisión aumenta con el número de estratos si estos están bien elegidos, pero no es conveniente aumentar mucho el número de estratos si tal aumento no compensa las complicaciones de cálculo y la disminución del tamaño de la muestra dentro de los estratos.

Notación básica:

N tamaño de la población.

L número de estratos.

N_i tamaño del i-ésimo estrato i = 1,2,...,L

n tamaño de la muestra.

$N = \sum_{i=1}^{L} N_i$, $n = \sum_{i=1}^{L} n_i$

$W_i = \frac{N_i}{N}$ peso del estrato i, $\sum_{i=1}^{L} W_i = 1$

$w_i = \frac{n_i}{n}$ proporción de la muestra en el estrato i, $\sum_{i=1}^{L} w_i = 1$

$f_i = \frac{n_i}{N_i}$ fracción de muestreo en el estrato i.

- Estimación de la media

Para estimar la media poblacional μ el estimador es $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = \sum_{i=1}^{L} W_i \bar{y}_i$,

donde la media muestral del i-ésimo estrato es: $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$.

- La Varianza poblacional de $\bar{y}_{st}$ es:

$$V(\bar{y}_{st}) = V\left[\frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i\right] = \frac{1}{N^2}\left[\sum_{i=1}^{L} N_i^2 V(\bar{y}_i)\right]$$

$$V(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{S_i^2}{n_i}\right)\right]$$

- La varianza estimada de $\bar{y}_{st}$ es:

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{S}_i^2}{n_i}\right)\right],$$

donde $\hat{S}_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$

Ejemplo

Gasto promedio de estadía de pacientes operados mediante una cirugía menor y cirugía mayor.

Estrato I (Cirugía menor)	Estrato II (Cirugía Mayor)
$N_1 = 110$	$N_2 = 168$
$n_1 = 20$	$n_2 = 30$
$\sum_{j=1}^{20} y_{1j} = 240\,000$	$\sum_{j=1}^{30} y_{2j} = 420\ 000$
$\sum_{j=1}^{20} y_{1j}^2 = 2\,980\,000\,000$	$\sum_{j=1}^{30} y_{2j}^2 = 6\,010\,000\,000$

Determine la media, el error de estimación y los límites de confianza.

Solución

Sabemos que: $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$, $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ y $\hat{S}_i^2 = \frac{1}{n_i - 1}\left(\sum_{j=1}^{n_i} y_{ij}^2 - n_i \bar{y}_i^2\right)$

$\bar{y}_1 = \frac{240\,000}{20} = 12\,000$ $\quad$ $\bar{y}_2 = \frac{420\,000}{30} = 14\,000$

$\hat{S}_1^2 = 5\,263157.895$ $\quad$ $\hat{S}_2^2 = 4\,482\,758.620$

$\bar{y}_{st} = \frac{1}{278}[110*12\,000 + 168*14\,000] = 13\,208.63$

es el valor promedio de los gastos en estas cirugías.

Al sustituir los valores respectivos en la varianza estimada tenemos:

$\hat{V}(\bar{y}_{st}) = 7\,853.52$

El error de estimación es: $e = t_{1-\frac{\alpha}{2}} * \sqrt{\hat{V}[\bar{y}_{st}]} = 2 * \sqrt{7\,853.52} = 560.48$

Los límites de confianza son: $\bar{y}_{st} \pm e$ luego, en este ejemplo:

(13 208.63 – 560.48; 13 208.63+560.48) = (12 648.15; 13 769.11)

Como en este tipo de muestreo, las muestras en cada estrato son independientes, entonces se puede realizar estimaciones separadas, así:

Estrato 1	**Estrato 2**
$\bar{y}_1 \pm t_{1-\frac{\alpha}{2}} * \sqrt{\hat{V}[\bar{y}_1]}$	$\bar{y}_2 \pm t_{1-\frac{\alpha}{2}} * \sqrt{\hat{V}[\bar{y}_2]}$
$12\,000 \pm 928.03$	$14\,000 \pm 700.69$
$(11\,071.97;\ 12\,928.03)$	$(13\,299.31;\ 14\,700.69)$

- Estimación del total

  El estimador del total es: $\hat{T}_{st} = N\bar{y}_{st} = N\sum_{i=1}^{L} W_i \bar{y}_i = N\frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i = \sum_{i=1}^{L} N_i \bar{y}_i$

  La Varianza poblacional de T: $V(\hat{T}_{st}) = V[N\bar{y}_{st}] = N^2 V[\bar{y}_{st}] = \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{S_i^2}{n_i}\right)$

  La Varianza estimada de T: $\hat{V}(\hat{T}_{st}) = \hat{V}[N\bar{y}_{st}] = N^2 \hat{V}[\bar{y}_{st}] = \sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{S}_i^2}{n_i}\right)$

- Asignación o afijación de la muestra

Asignación o afijación de la muestra: es el reparto o distribución del tamaño de la muestra n entre los diferentes estratos, es decir, la determinación de los L valores n_i de modo que $n = n_1 + n_2 + \ldots + n_L$.

Cada asignación puede originar una varianza diferente al estimador, nuestro objetivo es determinar un esquema de asignación que aumente la precisión y minimice los costos.

Los factores que influyen en la asignación son:

1. El número total de elementos en cada estrato.
2. La dispersión en cada estrato.
3. El costo de observación en cada estrato.

Tipos de asignación:

1. *Igual*: $n_i = \frac{n}{L}$
2. *Proporcional*: en la asignación proporcional se tiene que $\frac{n_i}{N_i} = \frac{n}{N}$ o equivalentemente a la fracción muestral es la misma en cada uno de los estratos, esto es: $f_i = f, \ \forall i$
3. *Óptima*: los valores de los tamaños de la muestra por estrato pueden ser asignados con la finalidad de minimizar la variabilidad del estimador para un costo fijo o para minimizar el costo para un valor específico de la varianza de la media $V(\bar{y}_{st})$.

En este último tipo de asignación, la función de costo fijo más sencilla es $C = c_0 + \sum_{i=1}^{L} c_i * n_i$. Dentro de cualquier estrato el costo es proporcional al tamaño de la muestra, pero el costo por cada unidad c_i puede variar entre los estratos. Por tanto, c_0 representa un costo general y c_i el costo por unidad encuestada en el estrato i. Entonces, se pueden analizar dos casos:

Caso a: minimizar la varianza $V(\bar{y}_{st})$ sujeto a la restricción $C - c_0 = \sum_{i=1}^{L} c_i * n_i$.

Usando el método de los multiplicadores de Lagrange debemos minimizar, la función:

$$\Phi(n_i) = \hat{V}(\bar{y}_{st}) + \lambda\left(\sum_{i=1}^{L} c_i n_i - C + c_0\right)$$

El resultando que se obtiene es: $n_i = n \frac{N_i \hat{S}_i / \sqrt{c_i}}{\sum_{j=1}^{L} N_j \hat{S}_j / \sqrt{c_j}}$

Este resultado nos indica que en un estrato dado se debe tomar una muestra grande si:

- El estrato es grande.
- El estrato es más variable internamente.
- El muestreo es más barato en el estrato.

Caso b: minimizar el costo $C = c_0 + \sum_{i=1}^{L} c_i * n_i$ para un valor específico de la varianza del estimador de la media poblacional $V = V(\bar{y}_{st})$

Usando el método de los multiplicadores de Lagrange debemos minimizar, la función:

$$\Psi(n_i) = c_0 + \sum_{i=1}^{L} c_i n_i - \lambda\left(V - \hat{V}(\bar{y}_{st})\right)$$

El resultando que se obtiene es:

$n_i = n \dfrac{N_i \hat{S}_i / \sqrt{c_i}}{\sum_{j=1}^{L} N_j \hat{S}_j / \sqrt{c_j}}$ el mismo valor obtenido en el caso a.

Además, podemos determinar el tamaño de la muestra para la asignación óptima:

Caso a: si el costo es fijo, entonces, se minimiza $V(\bar{y}_{st})$. Quiere decir que en la función de costos sustituimos el valor de n_i obtenido y se despeja el valor de n,

$$C - c_0 = \sum_{i=1}^{L} c_i * n \frac{N_i \hat{S}_i / \sqrt{c_i}}{\sum_{j=1}^{L} N_j \hat{S}_j / \sqrt{c_j}}$$

Entonces el tamaño de la muestra para la asignación óptima es:

$$n = \frac{(C - c_0)\left(\sum_{i=1}^{L} N_i \hat{S}_i / \sqrt{c_i}\right)}{\sum_{i=1}^{L} \sqrt{c_i} N_i \hat{S}_i}$$

Caso b: sustituyendo el valor de n_i en la expresión de la varianza (V), obtenemos:

$$V = \sum_{i=1}^{L} \frac{W_i^2 \hat{S}_i^2}{n_i} - \frac{1}{N} \sum_{i=1}^{L} W_i \hat{S}_i^2$$

Entonces el tamaño de la muestra para la asignación óptima es:

$$n = \frac{\left(\sum_{i=1}^{L} N_i \hat{S}_i / \sqrt{c_i}\right)\left(\sum_{i=1}^{L} \sqrt{c_i} N_i \hat{S}_i\right)}{N^2 V + \sum_{i=1}^{L} N_i \hat{S}_i^2}$$

Muestreo sistemático

Muestreo sistemático: es una técnica de muestreo en la que se selecciona un elemento de la población a intervalos regulares. Esto significa que se elige un elemento inicial al azar y luego se selecciona un elemento adicional cada cierto número de elementos en la lista de la población.

Se utiliza cuando el universo o población es de gran tamaño, o ha de extenderse en el tiempo.

Procedimiento del muestreo sistemático:

a) Primero hay que identificar las unidades y relacionarlas con el calendario (cuando proceda). Luego hay que calcular una constante, denominada *coeficiente de elevación* (K):

$$K = N / n$$

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

b) Luego se genera un número aleatorio A que debe estar comprendido entre 1 y el tamaño de la muestra n.

c) Los demás números aleatorios se calculan mediante la fórmula A + K, multiplicando a K desde 1 hasta el número que complete el tamaño de la muestra. Es decir, A + K, A + 2K, A + 3K,…, A + (n – 1) K.

¿En qué situaciones se utiliza el muestreo sistemático?

En general, el muestreo sistemático es un procedimiento que puede ser implementado en cualquier tipo de investigación. Sin embargo, en algunos casos su uso es más recomendable y provechoso que en otros. Principalmente,

se refiere a las investigaciones donde los elementos ya cuentan con un ordenamiento predeterminado, de manera que el proceso de selección y conteo se pueda realizar fácilmente.

Un ejemplo de ello es una lista de estudiantes, productos o cualquier tipo de inventario cuyos elementos se encuentren enumerados.

Principales características del muestreo sistemático:

- Hace parte del grupo de muestreos probabilísticos, los cuales se basan en la selección aleatoria de los elementos de la muestra.
- Utiliza ciertas fórmulas y operaciones específicas, las cuales constituyen el sistema que lo caracteriza.
- Abarca todo el rango de la población estudiada.

Principales ventajas del muestreo sistemático

- Es un método sencillo de aplicar.
- Garantiza un alto grado de representatividad por parte de la muestra.
- Toma en cuenta todo el rango de la población, obteniendo una mayor representatividad de la misma.
- No requiere realizar cálculos complejos.
- Evita posibles influencias y errores sobre los cálculos y estimaciones realizadas.

Principales desventajas del muestreo sistemático:

- En algunos casos, donde los elementos se encuentran ordenados con cierta periodicidad, los resultados de las operaciones pueden coincidir con la misma, lo que no permitiría considerar otros valores con otras características, afectando negativamente la representatividad de la muestra.

Dos procedimientos para seleccionar muestras

Cuando seleccionamos una muestra mediante un muestreo sistemático, la muestra es expresada en términos de «una muestra de 1 en 10» o «de 1 en 20». En general diríamos «una muestra de 1 en k», lo cual significa que la fracción muestral es 1/k.

Método A

Supongamos que tenemos una población:

$$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}$$

y deseamos seleccionar una 1 en k = 3 muestra sistemática. Un método de selección consiste en seleccionar una unidad muestral de entre las 3 primeras unidades muestrales (supongamos que salió la unidad j=2) y luego seleccionar cada tercera unidad a partir de la 2. Aplicando este procedimiento, obtenemos:

$$X_1, X_2, X_3 \big| X_4, X_5, X_6 \big| X_7, X_8, X_9 \big| X_{10}, X_{11}, X_{12}$$

$$X_2, X_5, X_8, X_{11}$$

Como acabamos de ver, la población ha sido dividida en:

$$n = \frac{N}{k} = \frac{12}{3} = 4$$

grupos y de cada grupo hemos seleccionada una unidad muestral. A estos grupos lo denominaremos zonas o estratos. La característica de este método es que $N = nk$, el tamaño de la población es un múltiplo de k.

En este caso k=3 es el tamaño del estrato y n=4 el tamaño de la muestra. Debido a que se selecciona una unidad de cada estrato, k=3 es el número de posibles muestras sistemáticas que pueden ser seleccionadas. Para este ejemplo, las muestras sistemáticas que pueden obtenerse serían:

Muestra 1	Muestra 2	Muestra 3
X_1	X_2	X_3
X_4	X_5	X_6
X_7	X_8	X_9
X_{10}	X_{11}	X_{12}

Los elementos de la muestra pueden ser denotamos por:

j, j + k, j + 2k, j + 3k, con j = 1,2,3 y k=3

Debido a que la unidad muestral inicial se selecciona aleatoriamente de las 3 primeras, la probabilidad de ser seleccionada será 1/3, y la probabilidad de seleccionar una muestra sistemática será también 1/3.

En general los elementos de la muestra se pueden representar por:

$$j, j + k, j + 2k, j + 3k,\ldots, j+(n-1)k$$

y la probabilidad de seleccionar una cualquiera de estas muestras será 1/k.

Observación: en el caso en que $N \neq nk$ todas las muestras no serán de igual tamaño.

Ejemplo:

Supongamos que N = 37 y una muestra 1 en 8 aleatoria sistemática debe ser seleccionada.

Solución:

Para buscar el tamaño de la muestra hacemos lo siguiente:

$$\frac{N}{k} = \frac{37}{8} = 4\frac{5}{8}$$

por lo que el tamaño de la muestra será:

$$4 < 4\frac{5}{8} < 5$$

esto es, el tamaño de muestra es 4 ó 5. Las k = 8 posibles muestras se presentan a continuación.

1	2	3	4	5	6	7	8
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}
X_{25}	X_{26}	X_{27}	X_{28}	X_{29}	X_{30}	X_{31}	X_{32}
X_{33}	X_{34}	X_{35}	X_{36}	X_{37}			

La probabilidad de seleccionar una cualquiera de estas muestras es 1/8.

Método B

Supongamos que N = nk = 12 y deseamos seleccionar una muestra 1 en k=3. Una unidad de muestreo es seleccionada aleatoriamente de la población, digamos que salió la j-ésima unidad de muestreo, con j=8.

Entonces:

$$\frac{j}{k}=\frac{8}{3}=2 \text{ con un resto r=2}$$

Notemos que $r<k$, y que r tomará los valores 0, 1 y 2. Cuando r=1 seleccionamos X_1, cuando r=2 seleccionamos X_2 y cuando r=0, X_3 como punto de partida. Luego se selecciona la k-ésima unidad de muestreo, a partir de la unidad de partida. Para este ejemplo, las muestras posibles son:

Muestra 1	Muestra 2	Muestra 3
X_1	X_2	X_3
X_4	X_5	X_6
X_7	X_8	X_9
X_{10}	X_{11}	X_{12}

Como podemos observar las muestras obtenidas por ambos métodos coinciden.

Apliquemos ahora el método B cuando $N \neq nk$. Para ello asumamos que N=11.

Buscamos $\frac{j}{k}=\frac{8}{3}=2$ con un resto r=2

y desarrollamos el mismo procedimiento de selección que en el caso anterior, con el cual obtenemos:

Muestra 1	Muestra 2	Muestra 3
X_1	X_2	X_3
X_4	X_5	X_6
X_7	X_8	X_9
X_{10}	X_{11}	

La característica de este procedimiento de selección es que la probabilidad de seleccionar la muestra sistemática será n/N y no $1/k$.

La probabilidad de seleccionar X_2, X_5, X_8 o X_{11} es $1/11$. Cuando cualquiera de estas unidades es seleccionada, obtenemos r=2, por lo tanto la probabilidad de seleccionar la muestra sistemática correspondiente a r=2 es:

$$P(S_2)+P(S_5)+P(S_8)+P(S_{11})=\frac{1}{11}+\frac{1}{11}+\frac{1}{11}+\frac{1}{11}=\frac{4}{11}=\frac{n}{N}$$

donde: S_2 es el evento que consiste en seleccionar la unidad 2,

S_5 es el evento que consiste en seleccionar la unidad 5,

S_8 es el evento que consiste en seleccionar la unidad 8,

S_{11} es el evento que consiste en seleccionar la unidad 11.

Similarmente, la probabilidad de seleccionar X_3, X_6 y X_9 será $3/11$.

Diferencias entre los métodos A y B

1) Aunque las muestras sistemáticas obtenidas por ambos métodos son las mismas, existe una diferencia en la probabilidad de seleccionar éstas.
2) La segunda diferencia se ilustrará a través de un ejemplo.

Ejemplo

Supongamos que deseamos seleccionar una 1 en 20 muestra sistemática de usuarios que acuden a un cierto restaurante entre las 7 y 10 a.m., para estimar su gasto promedio.

Solución:

En este caso como N es desconocido, no podemos emplear el método B. Usando el método A. seleccionamos un número aleatorio entre 1 y 20, supongamos que obtuvimos el 7, entonces seleccionamos a todo usuario 20-ésimo a partir del 7mo.

La muestra hasta las 10 a.m. será:

$$7,\ 7+20,\ 7+2\cdot 20,\ 7+3\cdot 20,\ldots$$

- Estimador de la media poblacional

En dependencia del método empleado para seleccionar la muestra sistemática, el estimador de la media poblacional será sesgado o insesgado. Cuando $N \neq nk$ y el método empleado es el A, el estimador será sesgado. Sin embargo, si el método empleado es el B la media muestral será siempre insesgada sin importar si $N \neq nk$ o no.

Caso I: método A y $N = nk$

Sea $\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$ la media muestral para la i-ésima muestra sistemática, entonces:

$$E[\bar{x}_s] = \frac{1}{k}\sum_{j=1}^{k}\bar{x}_i = \frac{1}{k}\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij} = \bar{X}$$

Caso II: método A y $N \neq nk$

Veámoslo a través del ejemplo con N=11. Ya vimos que en este ejemplo las muestras a formar serían:

Muestra 1	Muestra 2	Muestra 3
X_1	X_2	X_3
X_4	X_5	X_6
X_7	X_8	X_9
X_{10}	X_{11}	

Luego, hay 2 muestras de tamaño 4 y una de tamaño 3.

Entonces:

$$E[\bar{x}_s] = \frac{1}{3}(\bar{x}_1 + \bar{x}_2 + \bar{x}_3) = \frac{1}{3}\left(\frac{1}{4}\sum_{j=1}^{4} x_{1j} + \frac{1}{4}\sum_{j=1}^{4} x_{2j} + \frac{1}{3}\sum_{j=1}^{3} x_{3j}\right) \neq \bar{X}$$

Caso III: Método B y $N = nk$

Sea $\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$ la media muestral para la i-ésima muestra sistemática, entonces:

$$E[\bar{x}_s] = \frac{n}{N}\sum_{j=1}^{k}\bar{x}_i = \frac{n}{N}\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij} = \bar{X}$$

Caso IV: Método B y $N \neq nk$

En este caso como no todas las muestras tienen igual tamaño, denotemos por n_i el tamaño de la i-ésima muestra sistemática, entonces:

$$E[\bar{x}_s] = \sum_{j=1}^{k}\frac{n_i}{N}\bar{x}_i = \sum_{i=1}^{k}\frac{n_i}{N}\left(\frac{1}{n_i}\sum_{j=1}^{n}x_{ij}\right) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij} = \bar{X}$$

- Varianza de la media muestral

De la definición de varianza tenemos que:

$$V(\bar{x}_s) = E(\bar{x}_s - \bar{X})^2 = \frac{1}{k}\sum_{i=1}^{k}(\bar{x}_i - \bar{X})^2$$

para el caso I se obtiene que

$$V(\bar{x}_s) = \frac{N-1}{N}S^2 - \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$$

donde $S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{X})^2$

Esto es, la varianza de la media muestral se ha dividido en dos partes, el primer término muestra la varianza para la población como un todo, y el segundo la variación dentro de las muestras sistemáticas.

En consecuencia:

- Mientras mayor sea la variación entre las muestras, menor será la varianza de la media muestral.
- Una variación grande entre las muestras indica que la muestra es heterogénea, por lo tanto, cuando las unidades muestrales dentro de la muestra sistemática son heterogéneas, la precisión del muestreo sistemático se incrementará.

Ejemplo:

Sea la población de tamaño N = 9, 1, 2, 3, 4, 5, 6, 7, 8, 9 y supongamos que debemos seleccionar una muestra 1 en 3 sistemática.

Solución:

Muestra 1		Muestra 2		Muestra 3	
x_{1j}	x_{1j}^2	x_{2j}	x_{2j}^2	x_{3j}	x_{3j}^2
1	1	8	4	3	9
4	16	5	25	6	36
7	49	8	64	9	81
12	66	15	93	18	126

$S^2 = 7.5$, $\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2 = 6$, $V(\bar{x}_s) = 0.667$

Veamos ahora como la varianza dentro de las muestras $\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2$ afecta la varianza de la media muestral. Para ello consideremos el siguiente caso extremo, en el cual la población es:

1 2 3 4 5 1 2 3 4 5 1 2 3 4 5

en la cual podemos percatarnos que existe periodicidad en las unidades. Si una muestra sistemática 1 en 5 es tomada y la unidad de partida es 2, la muestra estaría conformada por:

2 2 2

la cual es homogénea y no representativa de la población. La varianza dentro de las muestras es cero y la varianza de la media muestral será grande.

Una pregunta usual que debemos hacernos es ¿cómo medir esta homogeneidad o heterogeneidad?

Una medida que expresa el grado de homogeneidad en una muestra sistemática es el coeficiente de correlación intraclase ρ entre pares de unidades en la misma muestra sistemática.

$$\rho = \frac{E\left(x_{ij} - \bar{X}\right)\left(x_{il} - \bar{X}\right)}{E\left(x_{ij} - \bar{X}\right)^2}$$

Luego $V(\bar{x}_s) = \frac{S^2}{n}\frac{N-1}{N}[1+(n-1)\rho]$

La expresión $V(\bar{x}_s)$ muestra que:

- Cuando ρ es grande y positivo, la varianza de la media muestral será grande.
- Cuando ρ es pequeño y positivo o negativo, la varianza de la media muestral será pequeña.
- Cuando $\rho = 0$, la varianza de la media muestral será igual a la varianza de la media muestral bajo el muestreo aleatorio simple.

Un valor de ρ grande y positivo se obtendrá cuando las unidades sean homogéneas, y pequeño y positivo o negativo, cuando las unidades sean heterogéneas.

Muestreo por conglomerados

Se caracteriza porque las unidades de muestreo contienen a dos o más unidades primarias (últimas). La población se subdivide en subpoblaciones y algunas de ellas, que denominaremos conglomerados, pero no todas serán incluidas en la muestra.

El muestreo por conglomerados es similar al muestreo aleatorio simple, pero se diferencian en que la unidad de muestreo es un conjunto de unidades primarias o elementales.

A diferencia con el muestreo estratificado, donde la población también se subdivide en subpoblaciones, pero siempre todos los estratos están representados en la muestra. Mientras que el muestreo estratificado es diseñado y utilizado fundamentalmente con el objeto de reducir la varianza de los estimadores, el muestreo por conglomerados es utilizado debido a que muestrear directamente sobre las unidades primarias, el costo es exageradamente alto.

Este muestreo es, en muchos casos, un muestreo efectivo para obtener la información deseada a un menor costo, aunque el uso de los conglomerados conlleve en algunos casos a una varianza mayor de los estimadores.

Los casos en los cuales se justifica la aplicación de este diseño muestral son:

1) Donde existe un alto costo por la movilización o traslado entre las unidades primarias; el muestreo por conglomerado permite disminuir las distancias; pues por lo general, los conglomerados son áreas físicas o geográficas, donde las unidades primarias están contiguas.

2) Cuando no existe lista de las unidades primarias (o últimas) sobre los cuales hay que tomar las observaciones, y el costo de levantar un marco muestral de estas unidades es alto, en comparación con el costo de muestrear sobre conglomerados, los cuales si pueden disponer de un marco o directorio.

3) Para pequeñas unidades donde puede ser difícil fijar con precisión sus límites, sin embargo, puede ser posible y fácil, dividir con población en unidades mayores y luego muestrear y medir aquellas unidades mayores seleccionadas. Ejemplo: animales.

4) También, pueden existir consideraciones administrativos que jueguen papel importante en la colección del diseño a utilizar.

La diferencia de objetivos entre estratificación y conglomerados conduce a diferentes criterios para establecer los conglomerados o los estratos. En contraste, con el estratificado, la varianza del estimador se hace pequeña al hacer el conglomerado, tanto como sea posible, representativo de la diversidad de toda población, y todas los conglomerados deben ser en lo posible construidos de modo que sean lo más semejante entre sí. A diferencia del muestreo estratificado, donde los estratos deben ser homogéneos dentro de sí y heterogéneos entre sí.

¿Cómo seleccionar una muestra por conglomerados?

a) Definir el conglomerado tipo (tamaño del conglomerado). El número de elementos que integran un conglomerado se denomina tamaño. En la mayoría de los métodos por conglomerados, los conglomerados son de tamaños diferentes unas de otras, los conglomerados de igual tamaño, rara vez se logran en la práctica, pero se constituyen una introducción sencilla al estudio del método por muestreo, y pueden resultar en situaciones practicas donde las condiciones fueran las indicadas, tales como: procesos de producción (control de calidad).

El problema de elegir un tamaño de conglomerado (m_i) apropiado puede ser un proceso un tanto complicado. El tamaño óptimo de los conglomerados no es una característica que depende exclusivamente de la población, sino también de la estructura de costos de la investigación. El tamaño del conglomerado óptimo es aquel para el cual la varianza del estimador es mínimo donde el costo de la investigación o el costo de la encuesta es mínimo dada la varianza.

Así, por ejemplo, el tamaño del conglomerado se hace más pequeño cuando aumenta la duración de la entrevista, cuando el traslado entre las unidades primarias es barato, cuando la densidad del conglomerado es mayor y cuando el presupuesto del gasto aumenta.

b) Formar el marco muestral, listando los conglomerados en los cuales se ha particionando la población. Resolviendo las imperfecciones que el marco pueda tener y garantizando que todos las unidades primarias que están en los conglomerados esta en uno y solo uno de los conglomerados.

c) Seleccionar los conglomerados que van en la muestra utilizando un muestreo irrestricto aleatorio.

Si hacemos un muestreo o encuestamos todas las unidades de los conglomerados seleccionados, estamos en presencia de un *muestreo por conglomerados monoetápico*.

Si en vez de entrevistar u observar a todos los individuos o unidades primarias del conglomerado observado en la muestra a su vez tomamos muestras de estas unidades primarias de los conglomerados seleccionados, estamos en presencia de un *muestreo por conglomerados bietápico*, pues la muestra se selecciona en dos etapas.

Este proceso se puede generalizar a más de dos etapas y en estos casos estaremos estamos en presencia de un *muestreo por conglomerados polietápico.*

Notación básica:

N números de conglomerados en la población.

n números de conglomerados en la muestra.

m_i números de unidades elementales (primarias) en el i-ésimo conglomerado.

M total de elementos en la población, $M = \sum_{i=1}^{N} m_i$

$\bar{M}$ tamaño promedio del conglomerado en la población, $\bar{M} = \frac{M}{N}$

y_i total del conglomerado i-ésimo.

$\bar{m}$ tamaño promedio del conglomerado en la muestra, $\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}$

- Estimación de la media poblacional

Por definición la media poblacional es:

$$\mu = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} m_i}$$

Luego, la estimación de la media poblacional es:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

Este estimador de la media tiene la forma de un estimador de razón, por lo tanto, la varianza de la media tiene la forma de la varianza del estimador de razón, así:

$$\hat{\mathrm{V}}(\bar{\mathrm{y}}) = \left(\frac{\mathrm{N}-\mathrm{n}}{\mathrm{Nn}\bar{\mathrm{M}}^2}\right)\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\bar{\mathrm{y}}\mathrm{m_i})^2}{\mathrm{n}-1}$$

Si se desconoce el total de elementos en la población M, entonces $\bar{\mathrm{M}}$ puede ser estimado con:

$$\bar{\mathrm{m}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{m_i}}{\mathrm{n}}$$

El límite para el error de estimación es:

$$\mathrm{e} = \mathrm{B} = \mathrm{t_k}\sqrt{\hat{\mathrm{V}}(\bar{\mathrm{y}})} = \mathrm{t_k}\sqrt{\left(\frac{\mathrm{N}-\mathrm{n}}{\mathrm{Nn}\bar{\mathrm{M}}^2}\right)\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\bar{\mathrm{y}}\mathrm{m_i})^2}{\mathrm{n}-1}}$$

Los límites de confianza son: $\bar{\mathrm{y}} \pm e$

En el muestreo por conglomerados monoetápico distinguiremos dos casos:

1. Todos los conglomerados son de igual tamaño.

2. Todos los conglomerados son de tamaño diferentes.

- Estimación del total poblacional

El total poblacional τ puede ser determinado por porque denota el total de elementos en la población. Por lo tanto, así como en el muestreo aleatorio simple, el total puede ser estimado por:

$$\hat{\tau} = \mathrm{M}\bar{\mathrm{y}} = \mathrm{M}\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{y_i}}{\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{m_i}}$$

La varianza estimada de $\hat{\tau} = \mathrm{M}\bar{\mathrm{y}}$ es:

$$\hat{V}(\mathrm{M}\bar{\mathrm{y}}) = \hat{\tau} = \mathrm{M}^2\hat{V}(\bar{\mathrm{y}}) = N^2\left(\frac{N-n}{Nn}\right)\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\bar{\mathrm{y}}\mathrm{m_i})^2}{\mathrm{n}-1}$$

El límite para el error de estimación es:

$$e = B = t_k\sqrt{\widehat{V}(\bar{y})} = t_k\sqrt{N^2\left(\frac{N-n}{Nn}\right)\frac{\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2}{n-1}}$$

Observe que el estimador $\hat{\tau} = M\bar{y}$ solo es útil solo cuando se conoce el total de elementos de la población.

Sin embargo, a menudo ese número de elementos de la población no se conoce, por tanto se debe utilizar otro tipo de estimador, el cual no depende de M:

$$\hat{\tau} = N\bar{y}_t = \frac{N}{n}\sum_{i=1}^{n} y_i$$

Donde:

$\frac{N}{n}$ es el factor de expansión

$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_i$ es el promedio de totales de conglomerado para la muestra seleccionada.

La varianza estimada de $\hat{\tau} = N\bar{y}_t$ es:

$$\widehat{V}(N\bar{y}_t) = N^2\widehat{V}(\bar{y}_t) = N^2\left(\frac{N-n}{Nn}\right)\frac{\sum_{i=1}^{n}(y_i - \bar{y}_t)^2}{n-1}$$

El límite para el error de estimación es:

$$e = B = t_k\sqrt{\widehat{V}(\bar{y})} = t_k\sqrt{N^2\left(\frac{N-n}{Nn}\right)\frac{\sum_{i=1}^{n}(y_i - \bar{y}_t)^2}{n-1}}$$

Este estimador τ tiene a menudo el inconveniente de ser poco preciso, pues por lo general, las medias de los conglomerados varían poco y los m_i varían mucho. En este caso el total del conglomerado ($y_i = m_i\bar{y}_i$) también varía mucho de unidad a unidad y entonces $\hat{V}(\hat{\tau})$ es muy grande. Sin embargo, este estimador es a veces utilizado, pues tiene la ventaja de que no es necesario conocer el tamaño de la población $M = \sum_{i=1}^{N} m_i$

Los estimadores μ y τ poseen propiedades especiales cuando los tamaños de los conglomerados son de igual tamaño $m_1 = m_2 = \cdots = m_N = m$, es decir:

a) El estimador $\bar{y}$ es un estimador insesgado de μ.

b) La varianza estimada $\hat{V}(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(y_i - \bar{y}m_i)^2}{n-1}$ es un estimador insesgado de la varianza poblacional $V(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{N}(y_i - \bar{y}m_i)^2}{n-1}$

c) Los estimadores del total poblacional $\hat{\tau} = M\bar{y}$ y $\hat{\tau} = N\bar{y}_t$ son equivalentes.

- Selección del tamaño de muestra

1. *Para estimar la media poblacional*

Por definición el error de estimación es:

$$e = B = t_k\sqrt{V(\bar{y})} = t_k\sqrt{\left(\frac{N-n}{Nn}\right)\sigma_c{}^2}$$

donde:

$V(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)\sigma_c{}^2$ es la varianza poblacional

$\hat{V}(\bar{y}) = \left(\frac{N-n}{Nn\bar{M}^2}\right)S_c{}^2$ es la varianza estimada.

Al despejar de la fórmula del error de estimación el valor de n, se tiene que el tamaño de muestra es:

$$n = \frac{N\sigma^2 c}{ND + \sigma^2 c}$$

donde:

$D = \left(\frac{e^2}{t^2\alpha}\right)\bar{M}^2$ es la varianza anticipada.

2. Para estimar el total poblacional.

En este caso tenemos dos tipos de estimadores:

a) $\hat{\tau} = M\bar{y}$

$$n = \frac{N\sigma^2 c}{ND+\sigma^2 c} \quad \text{donde} \quad D = \frac{e^2}{t^2 \alpha N^2}$$

b) $\hat{\tau} = N\bar{y}_t$

$$n = \frac{N{\sigma_t}^2}{ND+{\sigma_t}^2}$$

donde

$$D = \frac{e^2}{t^2 \alpha N^2}$$

${\sigma_t}^2$ esta varianza es estimada por ${S_t}^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y}_t)^2}{n-1}$ que es la cuasivarianza de tales de conglomerados en la muestra.

Prueba de hipótesis

La prueba de hipótesis involucra una suposición elaborada sobre algún parámetro de la población. A partir de la información proporcionada por la muestra se verificará la suposición sobre el parámetro estudiado. La hipótesis que se contrasta se llama hipótesis nula (H_o).

Partiendo de los resultados obtenidos de la muestra, o bien rechazamos la hipótesis nula a favor de la alternativa, o bien no rechazamos la hipótesis nula y suponemos que nuestra estimación inicial del parámetro poblacional podría ser correcto.

El hecho de no rechazar la hipótesis nula no implica que ésta sea cierta. Significa simplemente que los datos de la muestra son insuficientes para inducir un rechazo de la hipótesis nula.

Contraste de hipótesis: es el procedimiento mediante el cual la hipótesis que se contrasta es rechazada o no en función de la información muestral. La hipótesis alternativa se especifica como opción posible si se rechaza la nula.

En el contraste de hipótesis se pueden cometer errores de tipo I y de tipo II.

		Información muestral	
		Aceptar H_0	Rechazar H_0
La realidad	H_0 es cierta	**No hay error**	**Error tipo I**
	H_0 es falsa	**Error tipo II**	**No hay error**

Error tipo I: ocurre cuando se rechaza una hipótesis H_0 que es verdadera. La probabilidad de error tipo I viene a ser la probabilidad de rechazar H_0 cuando ésta es cierta.

P(error de tipo I) = α

El valor α es fijado por la persona que realiza la investigación (por lo general varía entre 1 – 10%)

Error tipo II: ocurre cuando se acepta una hipótesis H_0 que es falsa, la probabilidad de error tipo II es la probabilidad de aceptar H_0 cuando ésta es falsa.

P(error de tipo II) = β

Debido a que el valor real del parámetro es desconocido este error no puede ser fijado.

Potencia de prueba o poder de prueba: es la probabilidad de rechazar una hipótesis planteada cuando esta es falsa.

Potencia de prueba = 1- β

Como el valor de β depende del valor del parámetro la potencia de prueba tampoco pude ser fijado, sin embargo se puede asumir un conjunto de valores del parámetro y para cada uno de ellos hallar el valor de la potencia de prueba. La curva que se genera se conoce como curva de potencia.

Pasos a seguir en una prueba de hipótesis:

Paso 1: Planteo de hipótesis.

Paso 2: Nivel de significación.

Paso 3: Prueba estadística.

Paso 4: Suposiciones.

Paso 5: Regiones críticas. Criterios de decisión.

Paso 6: Realización de la prueba.

Paso 7: Resultados y conclusiones.

Procedimiento general

Sea ρ el parámetro que representa: $(\mu, \sigma^2, p, \mu_1 - \mu_2, p_1 - p_2, \sigma_1^2/\sigma_2^2)$

1. Planteo de las hipótesis.

$$\begin{cases} H_0 : \rho \geq \rho_0 \\ H_1 : \rho < \rho_0 \end{cases} \quad \begin{cases} H_0 : \rho \leq \rho_0 \\ H_1 : \rho > \rho_0 \end{cases} \quad \begin{cases} H_0 : \rho = \rho_0 \\ H_1 : \rho \neq \rho_0 \end{cases} \quad \begin{cases} H_0 : \rho = \rho_0 \\ H_1 : \rho < \rho_0 \end{cases} \quad \begin{cases} H_0 : \rho = \rho_0 \\ H_1 : \rho > \rho_0 \end{cases}$$

2. Fijar el nivel de significación α

3. Pruebas estadísticas

- Distribución simétrica: Z, t
- Distribución asimétrica positiva: χ^2, F

4. Supuestos

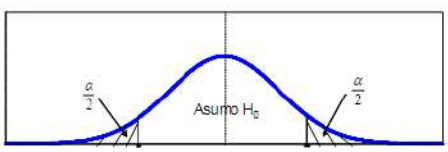
a) Supuestos para: $(\mu, \sigma^2, \mu_1 - \mu_2, \sigma_1^2/\sigma_2^2)$

- Población o poblaciones normalmente distribuidas.
- Muestra o muestras tomada al azar.

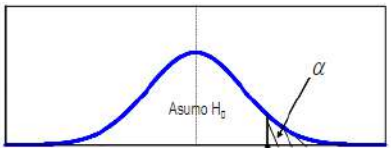
b) Supuestos para: $p, p_1 - p_2$

- Muestra o muestras tomada al azar.
- Muestra o muestras grandes

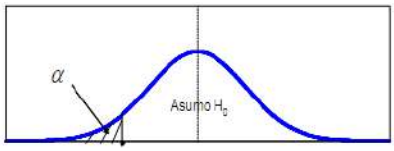
5. Definir región crítica y criterios de decisión.



Bilateral



Unilateral



Unilateral

6. Cálculos.

7. Resultados y conclusiones.

- **Prueba de hipótesis para una media poblacional**

Ejemplo

Una empresa eléctrica fabrica focos cuya duración se distribuye de forma aproximadamente normal con media de 800 horas y desviación estándar de 40 horas. Pruebe la hipótesis de que $\mu = 800$ horas contra la alternativa $\mu \neq 800$ horas si una muestra aleatoria de 28 focos tiene una duración promedio de 784 horas. Utilice un nivel de significancia de 0.05.

Solución

Sea X: Duración de los focos (horas)

$X \sim N(800, 40^2)$

1. Planteo de hipótesis.

$$\begin{cases} H_0 : \mu = 800 \\ H_1 : \mu \neq 800 \end{cases}$$

2. Nivel de significación.

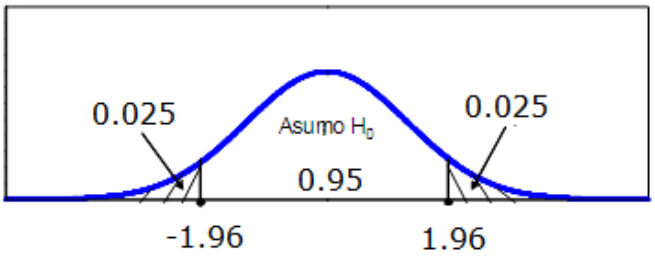
$\alpha = 0.05$

3. Prueba estadística.

$$Z_c = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0.1)$$

4. Supuestos.
   - Población normal.
   - Muestra tomada al azar.
5. Región crítica y criterios de decisión.

Región crítica:



Criterios de decisión:

- Si $-1.96 \leq Z_c \leq 1.96$ No se rechaza H_o

- Si $Z_c < -1.96$ o $Z_c > 1.96$ Se rechaza H_o

6. Cálculos

$$Z = \frac{784 - 800}{40/\sqrt{28}} = -2.12$$ (Z < -1.96 ⇒ Se rechaza H_o)

7. Conclusiones.

   Con 5% de nivel de significación y a partir de la información muestral, el tiempo promedio de duración de los focos es diferente de 800 horas.

- **Pruebas de hipótesis para una varianza poblacional**

Ejemplo

Se reporta que la desviación estándar de la resistencia al rompimiento de ciertos cables producidos por una compañía es 240 lb. Después de que se introdujo un cambio en el proceso de producción de estos cables, la resistencia al rompimiento de una muestra de 8 cables mostró una desviación estándar de 300 lb. Investigue la significancia del aumento aparente en la variación usando un nivel de significancia de 0.05. Asuma normalidad.

Solución

Sea X: Resistencia al rompimiento de cierto tipo de cable

$X \sim N(\mu, 240^2)$

1. Planteo de hipótesis.

$$\begin{cases} H_0 : \sigma^2 = 240^2 \\ H_1 : \sigma^2 > 240^2 \end{cases}$$

2. Nivel de significación.

$\alpha = 0.05$

3. Prueba estadística.

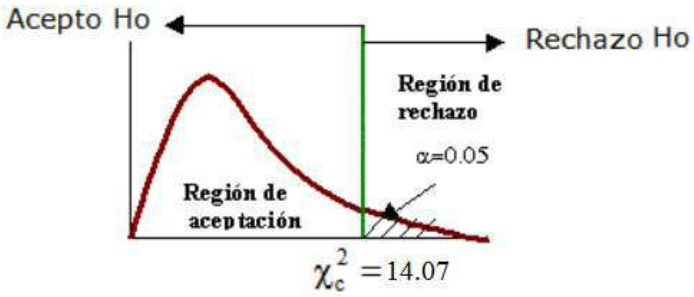
$$\chi_c^2 = \frac{(n-1)\, s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

4. Supuestos.

   - Población normal.
   - Muestra tomada al azar.

5. Región crítica y criterios de decisión.

Región crítica:



Criterios de decisión:

- Si $\chi_c^2 \leq 14.07$ No se rechaza H_o
- Si $\chi_c^2 > 14.07$ Se rechaza H_o

6. Cálculos

$$\chi^2 = \frac{(8-1)\ 300^2}{240^2} = 10.938$$ (χ^2 < 14.07 ⇒ No se rechaza H_o)

7. Conclusiones.

   Con 5% de nivel de significación y la información muestral es insuficiente para afirmar que la variación de la resistencia al rompimiento ha aumentado.

- **Pruebas de hipótesis para una proporción poblacional**

Ejemplo

Una empresa minorista de electrodomésticos anunció que vende el 21% de todos los computadores caseros ¿Esta afirmación se confirma si 120 de los 700 propietarios de computadores caseros se los compraron a esta empresa? Tome $\alpha = 0.05$.

Solución

Sea p la proporción de propietarios de computadores caseros que compraron en la empresa minorista de electrodomésticos.

1. Planteo de hipótesis.

$$\begin{cases} H_0 : p = 0.21 \\ H_1 : p \neq 0.21 \end{cases}$$

2. Nivel de significación.

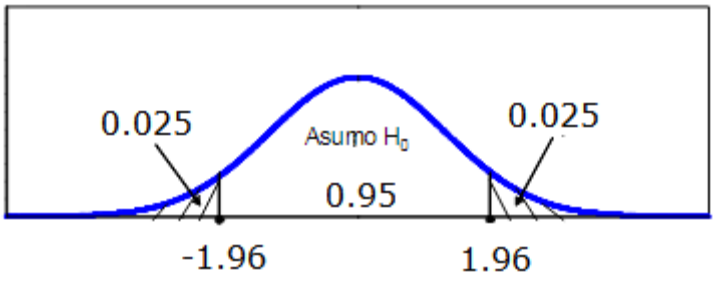
$\alpha = 0.05$

3. Prueba estadística

$$Z_c = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0.1)$$

4. Supuestos.
   - Muestra tomada al azar.
   - Muestra grande.
5. Región crítica y criterios de decisión.

   Región crítica:



   Criterios de decisión:
   - Si $-1.96 \leq Z_c \leq 1.96$ No se rechaza H_o
   - Si $Z_c < -1.96$ o $Z_c > 1.96$ Se rechaza H_o
6. Cálculos

$$Z = \frac{\frac{120}{700} - 0.21}{\sqrt{\frac{0.21(1-0.21)}{700}}} == -2.505$$ (Z=-2.505 < -1.96 ⇒ Se rechaza H_o)

7. Conclusiones.

   Con 5% de nivel de significación y a partir de la información muestral, la empresa minorista de electrodomésticos vende el 21% de todos los computadores caseros.

- **Pruebas de hipótesis para dos varianzas poblacionales**

Ejemplo

Diecisiete latas de gaseosa presentan una media de 17.2 onzas, con una desviación estándar de 3.2 onzas, y 13 latas de malta producen una media de 18.1 onzas y S = 2.7 onzas. Asumiendo varianzas iguales y distribuciones normales en los pesos de la población, ¿se puede afirmar con 5% de significación que las varianzas de los pesos son iguales?

Solución:

Sean X_1: contenido de una lata de gaseosa (onzas) $X_1 \sim N(\mu_1, \sigma_1^2)$

X_2: Contenido de una lata de malta (onzas) $X_2 \sim N(\mu_2, \sigma_2^2)$

1. Planteo de hipótesis.

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

2. Nivel de significación.

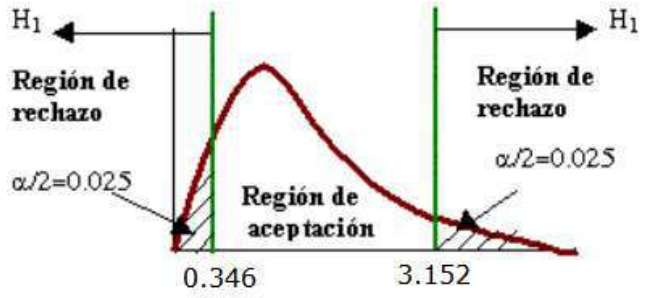
$$\alpha = 0.05$$

3. Prueba estadística

$$F_c = \frac{S_1^2}{S_2^2} \cdot \frac{1}{\frac{\sigma_1^2}{\sigma_2^2}} \sim \quad F_{(n_1-1, n_2-1)}$$

4. Supuestos.
   - Poblaciones normales.
   - Muestras tomadas al azar.

5. Regiones críticas y criterios de decisión.



Criterios de decisión:

- Si $0.346 \leq F_c \leq 3.152$ No se rechaza H_o
- Si $F_c < 0.346$ o $F_c > 3.152$ Se rechaza H_o

6. Cálculos

$$F = \frac{(3.2)^2}{(2.7)^2} = 1.405$$ (0.346 ≤ (F = 1.405) ≤ 3.152 ⇒ No se rechaza H_o)

7. Conclusiones.

Con 5% de nivel de significación la información muestral es insuficiente para rechazar que las varianzas de los pesos son iguales.

- **Pruebas de hipótesis para dos medias poblacionales.**

A. *Muestras independientes*

Ejemplo

Diecisiete latas de gaseosa presentan una media de 17.2 onzas, con una desviación estándar de 3.2 onzas, y 13 latas de malta producen una media de 18.1 onzas y S = 2.7 onzas. Asumiendo varianzas iguales y distribuciones normales en los pesos de la población, ¿se puede afirmar con 5% de significación que los pesos promedio son iguales?

Solución:

Sean X_1: Contenido de una lata de gaseosa (onzas) $X_1 \sim N(\mu_1, \sigma^2)$

X_2: Contenido de una lata de malta (onzas) $X_2 \sim N(\mu_2, \sigma^2)$

1. Planteo de hipótesis.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

2. Nivel de significación.

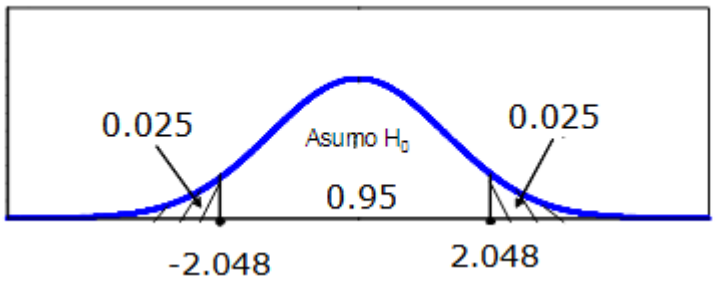
$$\alpha = 0.05$$

3. Prueba estadística

$$t_c = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)} \quad \text{donde:} \quad S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

4. Supuestos.
   - Poblaciones normales.
   - Muestras tomadas al azar.
5. Región crítica y criterios de decisión.

Regiones críticas:



Criterios de decisión:

- Si $-2.048 \leq t_c \leq 2.048$ No se rechaza H_o
- Si $t_c < -2.048$ o $t_c > 2.048$ Se rechaza H_o

6. Cálculos

$$t = \frac{(17.2 - 18.1) - (0)}{\sqrt{8.976\left(\frac{1}{17} + \frac{1}{13}\right)}} = -0.815$$

$(-2.048 \leq (t = -0.815) \leq 2.048 \Rightarrow$ No se rechaza $H_0)$

7. Conclusiones.

Con 5% de nivel de significación la información muestral es insuficiente para rechazar que los pesos promedios de los dos tipos de gaseosas son iguales.

- **Muestras relacionadas**

Ejemplo

Un gimnasio afirma que un nuevo programa de ejercicio reducirá la medida de la cintura de una persona en promedio dos centímetros en un período de cinco días. Las medidas de cinturas de seis hombres que participaron en este programa de ejercicios se registraron antes y después del período de cinco días en la siguiente tabla:

	Hombres					
	1	2	3	4	5	6
Medida de cintura antes	90.4	95.5	98.7	115.9	104.0	85.6
Medida de cintura después	91.7	93.9	97.4	112.8	101.3	84.0
Diferencias	-1.3	1.6	1.3	3.1	2.7	1.6

¿La afirmación del gimnasio es válida al nivel de significación de 5%? Suponga que la distribución de las diferencias de medidas de cintura antes y después del programa es aproximadamente normal.

Solución

Sean X_1: Medida de cintura antes (cm.)

X_2: Medida de cintura después (cm.)

1. Planteo de hipótesis.

$$\begin{cases} H_0 : D = 2 \\ H_1 : D \neq 2 \end{cases}$$

2. Nivel de significación.

$\alpha = 0.05$

3. Prueba estadística

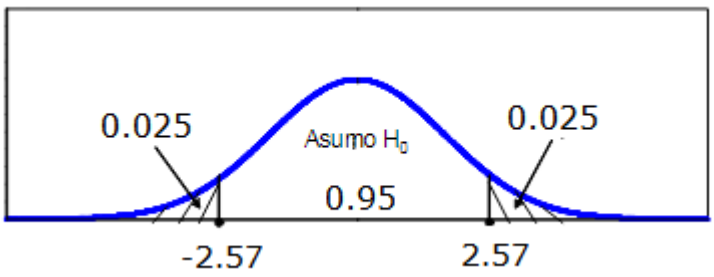
$$t_c = \frac{\bar{d} - D}{S_d / \sqrt{n}} \quad \sim \quad t_{(n-1)}$$

4. Supuestos.

- Las diferencias tienen distribución normal.

5. Región crítica y criterios de decisión.

Regiones críticas:



Criterios de decisión:

- Si $-2.57 \leq t_c \leq 2.57$ No se rechaza H_o
- Si $t_c < -2.57$ o $t_c > 2.57$ Se rechaza H_o

6. Cálculos

$$t = \frac{1.5 - 2}{1.543/\sqrt{6}} = -0.794$$ ($-2.57 \leq (t = -0.749) \leq 2.57 \Rightarrow$ No se rechaza H_0)

7. Conclusiones.

Con 5% de nivel de significación la información recogida resulta insuficiente para contradecir lo que afirma el gimnasio.

- **Prueba de hipótesis para dos proporciones poblacionales.**

Ejemplo

En una prueba de calidad de dos comerciales de televisión se pasó cada uno en un área de prueba seis veces, durante un período de una semana. La semana siguiente se llevó a cabo una encuesta telefónica para identificar a quienes habían visto esos comerciales. A las personas que los vieron se les pidió definieran el principal mensaje en ellos. Se obtuvieron los siguientes resultados:

Comercial	Personas que lo vieron	Personas que recordaron el mensaje principal
A	150	63
B	200	60

Use $\alpha = 0.05$ para probar la hipótesis que no hay diferencia en las proporciones que recuerdan los dos comerciales.

Solución:

Sea p_1 la proporción de personas que recordaron el mensaje principal del comercial A.

Sea p_2 la proporción de personas que recordaron el mensaje principal del comercial B.

1. Planteo de hipótesis.

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$$

2. Nivel de significación.

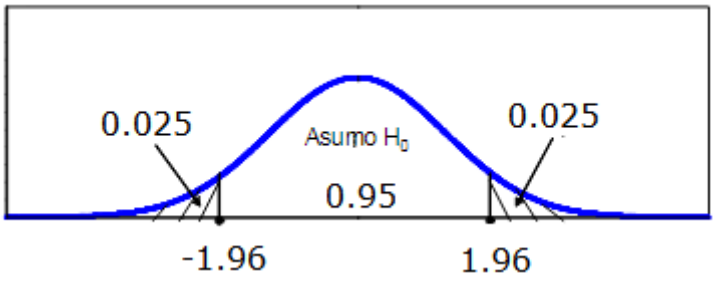
$$\alpha = 0.05$$

3. Prueba estadística

$$Z_c = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

4. Supuestos.
   - Muestra tomada al azar.
   - Muestra grande.
5. Región crítica y criterios de decisión.

   Región crítica:



   Criterios de decisión:
   - Si -1.96 $\leq Z_c \leq$ 1.96 No se rechaza H_o
   - Si Z_c < -1.96 o Z_c > 1.96 Se rechaza H_o
6. Cálculos

$$Z = \frac{\frac{63}{150} - \frac{60}{200}}{\sqrt{(0.351)(0.649)\left(\frac{1}{150} + \frac{1}{200}\right)}} = 2.328$$

   (1.96 $\leq$ (Z = 2.328) $\Rightarrow$ Se rechaza H_0)
7. Conclusiones.

   Con 5% de nivel de significación y a partir de la información muestral, hay diferencias significativas en las proporciones que recuerdan los dos comerciales.

Distribuciones bidimensionales

Tablas de doble entrada

Dato apareado: es un par de valores con dos componentes, uno perteneciente a la variable $X: X_1, X_2, \ldots, X_i, \ldots, X_k$ y el otro a la variable $Y: Y_1, Y_2, \ldots, Y_j, \ldots, Y_p$, que se simboliza por (X_i, Y_j) y que surge de contar, medir u observar dos características simultáneamente en cada uno de los elementos unitarios de la población.

Con la intención de reunir en una sola estructura toda la información disponible, creamos una tabla de doble entrada conformada por $k \cdot p$ casillas, organizadas de forma que se tengan k filas y p columnas. La celda denotada de forma genérica mediante la posición *ij* hará referencia a los elementos de la muestra que presentan simultáneamente las modalidades x_i e y_j respectivamente. De este modo, para *i = 1,..., k,* y para *j = 1,..., p*, se tiene que n_{ij} es el número de individuos o frecuencias absolutas, que presentan a la vez las modalidades X_i *e* Y_j.

Y / **X**	**y_1**	**y_2**	**...**	**y_j**		**y_p**	
x_1	n_{11}	n_{12}	...	n_{1j}		n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2j}		n_{2p}	$n_{2\bullet}$
	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}		n_{ip}	$n_{i\bullet}$
	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}		n_{kp}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet j}$	...	$n_{\bullet p}$	$n_{\bullet\bullet}$

El número de individuos que presentan la modalidad X_i, es lo que llamamos

frecuencia absoluta marginal de X_i y se representa como $n_{i\bullet}$.

Es evidente la igualdad

$$n_{i\bullet} = n_{i1} + n_{i2} + \ldots + n_{ip} = \sum_{j=1}^{p} n_{ij}$$

Obsérvese que hemos escrito un símbolo "•" en la "parte de las jotas" que simboliza que estamos considerando los elemento que presentan la modalidad X_i, independientemente de las modalidades que presente la variable *Y*.

De forma análoga se define la frecuencia absoluta marginal de la modalidad Y_j como

$$n_{\bullet j} = n_{1j} + n_{2j} + \ldots + n_{kj} = \sum_{i=1}^{k} n_{kj}$$

Estas dos distribuciones de frecuencias $n_{i\bullet}$ para *i = 1,..., k*, y $n_{\bullet j}$ para *j=1,..., p* reciben el nombre de distribuciones marginales de *X* e *Y* respectivamente.

El número total de elementos de la población lo obtendremos:

$$n = n_{\bullet\bullet} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{p} n_{\bullet j} = \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}$$

Dependencia entre dos variables

La dependencia funcional que nos refleja cualquier fórmula matemática o física es la que estamos normalmente más habituados. Si existe una dependencia funcional entre las variables *X* e *Y, sea esta por ejemplo Y=3X,* entonces si se ha observado la medida *X=5*, no es necesario practicar la de *Y,* pues la relación entre ambas es exacta, para este caso *Y=15.*

Existe un concepto que es radicalmente opuesto a la *dependencia funcional*, que es el de *independencia*. Se dice que dos variables X e Y son independientes si la distribución marginal de una de ellas es la misma que la condicionada por cualquier valor de la otra. Esta es una de entre muchas maneras de expresar el concepto de independencia, y va a implicar una estructura muy particular

de la tabla bidimensional, en el que todas las filas y todas las columnas van a ser proporcionales entre sí.

Independencia entre variables: dos variables X e Y, son independientes entre sí, cuando

$$n_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n_{\bullet\bullet}} \qquad \forall i, j$$

Existen características que ni son independientes, ni se da entre ellas una relación de dependencia funcional, pero sí se percibe una cierta relación de dependencia entre ambos; se trata de una *dependencia estadística*.

Si cambios en una variable X , son acompañados sistemáticamente por cambios en otra variable Y , decimos que el conjunto de pares de datos apareados (X_i, Y_j) están *asociados* o *correlacionados*.

Cuando los caracteres son de tipo cuantitativo, el estudio de la dependencia estadística se conoce como el problema de regresión, y el análisis del grado de dependencia que existe entre las variables se conoce como el problema de correlación.

Correlación lineal

Un primer aspecto del análisis de asociación se conoce por *análisis de correlación*, el cual se ocupa de determinar el grado de relación entre las variables. En el estudio de la correlación la designación de las variables dependientes e independientes es una elección estrictamente personal y no tiene significación práctica.

La correlación se presenta cuando nos preguntamos, por ejemplo, ¿existe alguna relación entre el hábito de fumar y los problemas circulatorios?

Supongamos que la intensidad del hábito de fumar se pueda medir en una escala de intervalo que varía de 0 a 15, y de igual forma, que la intensidad de las afectaciones circulatorias pueden medirse en esta misma escala. Entonces en un grupo de ocho pacientes que presentan problemas circulatorios obtenemos los

siguientes datos:

Tabla 15. Intensidad del hábito de fumar (X), intensidad de las afectaciones circulatorias (Y).

X	Y
1	1
3	2
4	4
6	4
8	5
9	7
11	8
14	9

Fuente: Estos son datos ficticios.

La investigación de la relación entre las dos variables, se comienza generalmente con un intento de descubrir la forma aproximada de la relación, para esto se presentan los datos en un sistema de coordenadas.

El gráfico representado en la figura 15 recibe el nombre de *diagrama de dispersión*, el cual muestra la ubicación de los puntos (X_i, Y_j) en un sistema de coordenadas rectangulares. En la gráfica se puede observar si existe o no una relación acentuada y si se puede tratar de forma lineal o en otra forma.

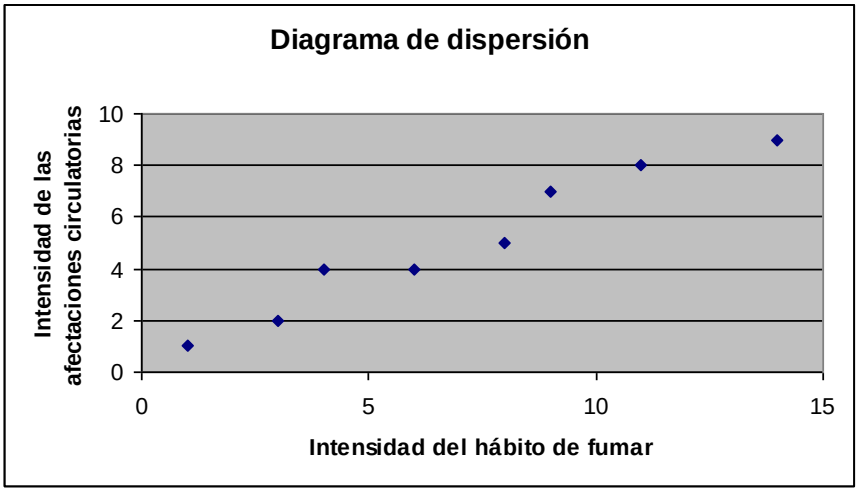


Figura 15. Diagrama de dispersión utilizando Microsoft Excel

Fuente: datos son ficticios

Si observamos este diagrama de dispersión, se evidencia que existe una tendencia a que los valores menores de X estén asociados a los valores menores de Y, así como que los valores mayores de X estén asociados a los valores mayores de Y. Además, en términos generales el aspecto del diagrama de dispersión es el de una línea recta.

Es usual en la búsqueda de la tendencia que manifiesta la información, analizar cualquier comprensión de los puntos en agruparse sobre ambos lados de alguna curva no compleja. Por ejemplo, de una línea recta.

Es conveniente utilizar una medida descriptiva que sirva para medir o cuantificar el grado de relación lineal entre ambas variables, en esta dirección podría utilizarse en primera instancia la covarianza entre X e Y: COV (X, Y).

Esta medida estadística expresa la variabilidad conjunta de dos variables aleatorias (cuantitativas) que varían de forma conjunta respecto a sus medias. Es decir, sabremos cómo se comporta una variable dependiendo de cómo lo haga la otra. O sea, la covarianza entre las variables aleatorias X e Y se determina por: $\mathrm{COV}(\mathrm{X},\mathrm{Y}) = \mathrm{E}[(\mathrm{X}-\mathrm{E}[\mathrm{X}])(\mathrm{Y}-\mathrm{E}[\mathrm{Y}])]$

El estimador insesgado de $\mathrm{COV}(\mathrm{X},\mathrm{Y})$ denotado por Sxy se determina por la expresión: $Sxy = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)(y_i - y)$, donde *n* es el tamaño de la muestra.

Para calcular la covarianza en una población es análoga pero en este caso se sustituye n por N en la expresión anterior, donde N representa la cantidad de elementos de la población, análogamente se sustituye $\overline{X}$ por $\overline{x}$, así como $\overline{Y}$ por $\overline{y}$.

Para realizar una interpretación geométrica de la covarianza, es conveniente considerar la nube de puntos formadas por las *n* parejas de datos (x_i, y_i). En la figura 16 se representan casos para los que la covarianza es positiva, negativa y cero:

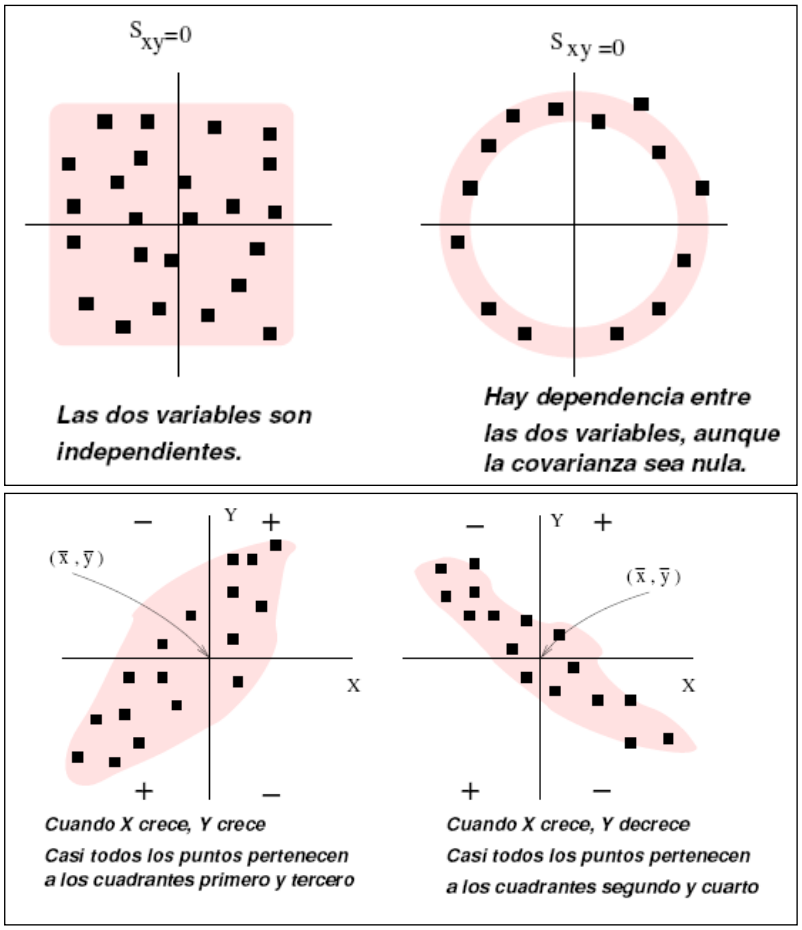


Figura 16. Casos para los que la covarianza es positiva, negativa y cero.

- Si hay mayoría de puntos en el tercer y primer cuadrante, ocurrirá que $S_{xy} \geq 0$, lo que se puede interpretar como que la variable Y tiende a aumentar cuando lo hace X.
- Si la mayoría de puntos están repartidos entre el segundo y cuarto cuadrante entonces $S_{xy} \leq 0$, es decir, las observaciones Y tienen tendencia a disminuir cuando las de X aumentan.
- Si los puntos se reparten con igual intensidad alrededor de (x, y), entonces se tendrá que $S_{xy}=0$.

Observación: cuando los puntos se reparte de modo más o menos homogéneo entre los cuadrantes primero y tercero, y segundo y cuarto, se tiene que $S_{xy} \approx 0$. Eso no quiere decir de ningún modo que no pueda existir ninguna relación

entre las dos variables, ya que esta puede existir como se aprecia en la figura de la derecha.

Propiedades de la Covarianza:

- Si $S_{xy} > 0$ las dos variables crecen o decrecen a la vez (nube de puntos creciente).
- Si $S_{xy} < 0$ cuando una variable crece, la otra tiene tendencia a decrecer (nube de puntos decreciente).
- Si los puntos se reparten con igual intensidad alrededor de (x, y), $S_{xy} = 0$ (no hay relación lineal).

Para los datos de la tabla 15 se puede verificar que $S_{xy}=10,5$. Pero si se hubiese utilizado una escala de intervalo de 0 a 150 para medir los valores de la variable X e Y respectivamente, el valor fuese igual a $S_{xy}=1050$.

En consecuencia, para medir el grado en que ambas variables X e Y se encuentran relacionadas linealmente, es conveniente tener en cuenta algunas propiedades deseables, como son:

1. Que la medida de la relación sea independiente de la elección del origen para las variables
2. Que sean independientes de la escala de medida para X e Y.
3. Que esté acotada.

La covarianza es una medida de la variabilidad común de dos variables (crecimiento de ambas al tiempo o crecimiento de una y decrecimiento de la otra), pero está afectada por las unidades en las que cada variable se mide. Así pues, es necesario definir una medida de la relación entre dos variables, y que no esté afectada por los cambios de unidad de medida.

Una forma de conseguir este objetivo es dividir la covarianza por el producto de las desviaciones típicas de cada variable, así se obtiene el coeficiente adimensional *r*, denominado como coeficiente de correlación lineal de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)Sx\,Sy}$$

Note que en el caso que se hayan observado todos los elementos de la población en la expresión anterior se divide por N (tamaño poblacional). En el caso anterior se trabaja con una muestra de dicha población por tanto debe dividirse por n-1, pues desde el punto de vista teórico se garantiza que la esperanza matemática del estimador coincida con el parámetro estimado (o sea, es un estimador insesgado).

Para el caso de los datos ficticios que se presentan en la tabla 15 el coeficiente de correlación lineal r es 0,977, puede verificarse que para el cambio de escala antes señalado este coeficiente de correlación mantiene el mismo valor (adimensional).

Aquí S_x y S_y son las desviaciones típicas muestrales correspondientes a las variables X e Y respectivamente.

$$Sx = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}, \qquad Sy = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n-1}}$$

Pueden verificarse tres expresiones equivalentes para determinar el coeficiente de correlación lineal de Pearson para el caso de una muestra:

a) $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)Sx\,Sy}$$

b) $$r = \frac{n\sum_{i=1}^{n} x_i\,y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]}}$$

c) $r = \frac{\sum_{i=1}^{n} x_i' y_i'}{\sqrt{\sum_{i=1}^{n} x_i'^2 \sum_{i=1}^{n} y_i'^2}}$, con el cambio de variable $x_i' = X_i - \bar{X}$ e $y_i' = y_i - \bar{Y}$

Ejemplo

Utilizando la expresión c), determine el coeficiente de correlación lineal de Pearson para los datos de la tabla 15.

Solución:

X	Y	$x_i' = X_i - \overline{X}$	$y_i' = y_i - \overline{Y}$	$x_i'^2$	$x_i' y_i'$	$y_i'^2$
1	1	-6	-4	36	24	16
3	2	-4	-3	16	12	9
4	4	-3	-1	9	3	1
6	4	-1	-1	1	1	1
8	5	1	0	1	0	0
9	7	2	2	4	4	4
11	8	4	3	16	12	9
14	9	7	4	49	28	16
$\sum_{i=1}^{8} x_i = 56$	$\sum_{i=1}^{8} y_i = 40$	---	---	$\sum_{i=1}^{8} x_i'^2 = 132$	$\sum_{i=1}^{8} x_i' y_i' = 84$	$\sum_{i=1}^{8} y_i'^2 = 56$
$\overline{X} = 7$	$\overline{Y} = 5$	---	---			

$$r = \frac{\sum_{i=1}^{n} x_i' y_i'}{\sqrt{\sum_{i=1}^{n} x_i'^2 \sum_{i=1}^{n} y_i'^2}} = \frac{84}{\sqrt{(132)(56)}} = 0.977$$

Propiedades del coeficiente de correlación lineal r:

- Carece de unidades de medida (adimensional).
- Es invariante para transformaciones lineales (cambio de origen y escala) de las variables.
- Sólo toma valores comprendidos entre −1 y 1. Cuando |r| esté próximo a uno, se tiene que existe una relación lineal muy fuerte entre las variables.
- Cuando r≈0, puede afirmarse que no existe relación lineal entre ambas variables. Se dice en este caso que las variables son *incorreladas*.

Interpretación del coeficiente de correlación lineal r

Si algún concepto estadístico se usa y se abusa de él es el coeficiente de correlación lineal r; por ello, es importante precisar su interpretación.

La interpretación de un coeficiente de correlación como medida del grado de relación lineal entre dos variables es una interpretación matemática pura y está completamente desprovista de implicaciones de causa y efecto. Es decir, que dos variables tienden a aumentar o a disminuir al mismo tiempo no implica que una tenga algún efecto directo o indirecto sobre la otra ya que puede suceder que ambas estén sujetas a la influencia de otras variables.

Es fácil imaginarse parejas de variables que pudiesen dar un alto valor de un coeficiente de correlación y que no se deba realmente a una estrecha relación entre ellas, sino al efecto común sobre estas de una tercera variable, y entonces este valor del coeficiente de correlación refleja solo este efecto común.

Reiteramos que los coeficientes de correlación se deben de manejar con sumo cuidado ya que de no ser así, puede llevarnos a conclusiones totalmente erróneas. Luego, para usarlos correctamente, se debe tener conocimiento del campo donde se esté utilizando.

Son muchos los campos donde el coeficiente de correlación r ha encontrado una aplicación útil.

Modelo de regresión simple

En este tipo de regresión se desea caracterizar el efecto lineal de una única variable explicativa sobre la variable respuesta. Los pasos para efectuar un análisis son los siguientes:

1. Representación gráfica de datos
2. Planteamiento del modelo
3. Estimación de la ecuación de predicción
4. Examen de la adecuación del modelo lineal
5. Intervalos de confianza para la estimación

En este apartado se explica el modelo de regresión lineal simple, un modelo con un solo regresor x que tiene una relación con una respuesta y, donde la relación es una línea recta. Este modelo de regresión lineal simple es:

$y = \beta_o + \beta_1 x + \varepsilon$ (Modelo poblacional de regresión)

Donde la ordenada al origen β_o y la pendiente β_1 son constantes desconocidas, y ε es una componente aleatorio del error. Se supone que los errores tienen promedio cero y varianza σ^2 desconocida. Además se suele suponer que los errores no están correlacionados. Esto quiere decir que el valor de un error no depende del valor de cualquier otro error.

- **Estimación de los parámetros por mínimos cuadrados**

Los parámetros β_o y β_1 son desconocidos, y se debe estimar con los datos de la muestra. Supongamos que hay n pares de datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Estos datos pueden obtenerse en un experimento controlado, diseñado en forma específica para recolectarlos, o en un estudio observacional, o a partir de registros históricos existentes (lo que se llama un estudio retrospectivo).

Para estimar β_o y β_1 se utiliza el método de mínimos cuadrados. Esto es, se estima β_o y β_1 tales que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea mínima. La ecuación se puede escribir $y_i = \beta_o + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \ldots, n$ (Modelo muestral de regresión), escritos en términos de los n pares de datos (y_i, x_i), $i = 1, 2, \ldots, n$. Así el criterio de mínimos cuadrados es:

$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$. Los estimadores por mínimos cuadrados de β_o y β_1, que se designarán por $\hat{\beta}_0$ y $\hat{\beta}_1$, deben satisfacer

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \qquad \text{y} \qquad \left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0$$

Simplificando estas dos ecuaciones se obtiene:

$$n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i \text{ ;} \qquad \hat{\beta}_0\sum_{i=1}^{n}x_i + \hat{\beta}_1\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_i y_i$$

Que son las llamadas ecuaciones normales de mínimos cuadrados. Su solución es la siguiente:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \text{ ;}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}y_i x_i - \frac{\left(\sum_{i=1}^{n}y_i\right)\left(\sum_{i=1}^{n}x_i\right)}{n}}{\sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}},$$

donde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$ y $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ son los promedios de y_i y x_i respectivamente.

Por consiguiente, $\hat{\beta}_o$ y $\hat{\beta}_1$, son los estimadores por mínimos cuadrados.

El modelo ajustado de regresión lineal simple es entonces: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ .Esta ecuación produce un estimado puntual de la media de y para x.

Otra forma más compacta de escribir $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, donde:

$$S_{xx} = \sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n}y_i x_i - \frac{\left(\sum_{i=1}^{n}y_i\right)\left(\sum_{i=1}^{n}x_i\right)}{n} = \sum_{i=1}^{n}y_i(x_i - \bar{x})$$

La diferencia entre el valor observado y_i y el valor ajustado correspondiente $\hat{y}_i$ se llama residual, matemáticamente el

i-ésimo residual es: $e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \quad i = 1,2,....n$

Aquí e_i tiene un papel importante para investigar la adecuación del modelo de regresión ajustado.

- **Propiedades de los estimadores por mínimos cuadrados y el modelo ajustado de regresión**

Tenemos que:

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{S_{xx}}$ Son combinaciones lineales de y_i, entonces se puede escribir:

$E\left(\hat{\beta}_1\right) = \beta_1$ $\qquad E(y_i) = \beta_0 + \beta_1 x_i$

$E\left(\hat{\beta}_0\right) = \beta_0$; $\qquad Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{s_{xx}}$ $\qquad Var\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)$

Propiedades útiles:

1. $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n} e_i = 0$
2. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$
3. La línea de regresión de mínimos cuadrados siempre pasa por el centroide de los datos que es el punto $(\bar{y}, \bar{x})$
4. $\sum_{i=1}^{n} x_i e_i = 0$
5. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

Además de estimar β_0 y β_1, se requiere estimar σ^2. Se obtiene de la suma de cuadrados residuales, o suma de cuadrados del error.

$$\mathrm{SC}_{\mathrm{Res}} = \sum_{i=1}^{n} e_i^{\,2}$$

$$= \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i \right)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} \left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) + \hat{\beta}_1^{\,2} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

$$= \sum_{i=1}^{n} \left(y_i^{\,2} - 2y_i\bar{y} + \bar{y}^2\right) - 2\hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + 2\hat{\beta}_1 \sum_{i=1}^{n} y_i \bar{x} + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i \bar{y} - 2\hat{\beta}_1 \sum_{i=1}^{n} \bar{x}\bar{y} + \hat{\beta}_1^{\,2} s_{xx}$$

$$= \sum_{i=1}^{n} y_i^{\,2} - 2n \frac{\sum_{i=1}^{n} y_i}{n} \frac{\sum_{i=1}^{n} y_i}{n} + n\bar{y}^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + 2\hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$+ 2\hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - 2\hat{\beta}_1 n \frac{1}{n} \sum_{i=1}^{n} x_i \frac{1}{n} \sum_{i=1}^{n} y_i + \hat{\beta}_1 . \frac{s_{xx}}{s_{xx}} . s_{xx}$$

$$= \sum_{i=1}^{n} y_i^{\,2} - 2n\bar{y}^2 + n\bar{y}^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} x_i y_i + 2\hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + \hat{\beta}_1 S_{xy}$$

$$= \sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2 - 2\hat{\beta}_1 \left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right] + \hat{\beta}_1 S_{xy}$$

$$= \sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2 - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1 S_{xy}$$

$$\mathrm{SC}_{\mathrm{Res}} = \sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

Pero $\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^{\,2} = SCT$

$$\mathrm{SC}_{\mathrm{Res}} = SCT - \hat{\beta}_1 S_{xy}$$

$$E\left(\mathrm{SC}_{\mathrm{Res}}\right) = \left(n-2\right)\sigma^2$$

La suma de cuadrados residuales tiene n-2 grados de libertad, 2 grados de libertas se asocian con los estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ que se usan para obtener $\hat{y}_i$.

El estimador insesgado de σ^2 es:

$\hat{\sigma}^2 = \frac{\mathrm{SC}_{\mathrm{Res}}}{\mathrm{n}-2} = \mathrm{CM}_{\mathrm{Res}}$ cuadrado medio residual

$\sqrt{\hat{\sigma}^2}$ error estándar de regresión.

$\hat{\sigma}^2$ es un estimado de σ^2 dependiente del modelo.

- **Intervalos de confianza**

Intervalos de confianza de β_0 y β_1 .

Como los errores se distribuyen en forma normal e independientes, entonces la distribución de muestreo de $\frac{\hat{\beta}_1-\beta_1}{S_e\left(\hat{\beta}_1\right)}$ y de $\frac{\hat{\beta}_0-\beta_0}{S_e\left(\hat{\beta}_0\right)}$ se distribuye t con n-2 grados de libertad.

Intervalos de confianza para β_0 y β_1

Un intervalo de confianza de $100\cdot(1-\alpha)\%$ para β_1 es:

$$\hat{\beta}_1 - t_{\alpha/2,n-2} S_e\left(\hat{\beta}_1\right) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2} S_e\left(\hat{\beta}_1\right)$$

Un intervalo de confianza de $100\cdot(1-\alpha)\%$ para β_0 es:

$$\hat{\beta}_0 - t_{\alpha/2,n-2} S_e\left(\hat{\beta}_0\right) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2} S_e\left(\hat{\beta}_0\right)$$

- **Estimación de intervalos de la respuesta media**

Una aplicación importante de un modelo de regresión es estimar la respuesta media $E(y)$, para determinado valor de la variable de regresión x.

Sea x_0 el valor o nivel de la variable de regresión para el que se desea estimar la respuesta media, es decir, $\mathrm{E}\left(\frac{\mathrm{y}}{\mathrm{x}_0}\right)$. Se supone que x_0 es cualquier valor de la variable de regresión dentro del intervalo de los datos originales de x que se usaron para ajustar el modelo.

Un estimador insesgado de $E\left(\frac{\mathrm{y}}{\mathrm{x}_0}\right)$ se determina a partir del modelo ajustado como sigue:

$$E\left(\frac{\hat{\mathrm{y}}}{\mathrm{x}_0}\right) = \hat{\mu}_{y/x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Para obtener un intervalo de confianza de $100\cdot(1-\alpha)\%$ para $E\left(\frac{\mathrm{y}}{\mathrm{x}_0}\right)$, se debe notar primero que $\hat{\mu}_{y/x_0}$ es una variable aleatoria normalmente distribuida, porque es una combinación lineal de las observaciones y_i.

La varianza de $\hat{\mu}_{y/x_0}$ es:

$$Var\left(\hat{\mu}_{y/x_0}\right) = \sigma^2\left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]$$

La distribución de muestreo de es una distribución t, con n-2 grados de libertad.

$$\frac{\hat{\mu}_{y/x_0} - E\left(y / x_0\right)}{\sqrt{CM\,\mathrm{Res}\left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right)}}$$

Un intervalo de confianza de $100\cdot(1-\alpha)\%$ para la respuesta media en el punto $x = x_0$ es:

$$\hat{\mu}_{y/x_0} - t_{\alpha/2,n-2}\cdot\sqrt{CMRes\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le E\left(\frac{y}{x_0}\right) \le \hat{\mu}_{y/x_0} + t_{\alpha/2,n-2}\cdot\sqrt{CMRes\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

El ancho del intervalo de confianza para $E\left(\frac{y}{x_0}\right)$ es una función de x_0. El ancho del intervalo es mínimo para $x_0 = \bar{x}$, y crece a medida que aumenta $|x_0 - \bar{x}|$.

- **Prueba de hipótesis de la pendiente y de la ordenada al origen**

Supongamos que deseamos probar que la pendiente es igual a una constante.

$H_0 : \beta_1 = \beta_{10}$
$H_1 : \beta_1 \neq \beta_{10}$

e_i son independientes y distribuidos $N(0,\sigma^2)$

$\hat{\beta}_1$ es una combinación lineal de las observaciones, y está distribuida normalmente.

$E\left(\hat{\beta}_1\right) = \beta_1$ (Promedio de β_1); $Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}}$ (Varianza de β_1)

Denótese a $S_e\left(\hat{\beta}_1\right) = \sqrt{\frac{CM_{Res}}{S_{xx}}}$ como el error estándar estimado o error estándar de la pendiente, entonces

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{S_e\left(\hat{\beta}_1\right)}.$$

Se rechaza la hipótesis nula si: $|t_0| > t_{\alpha/2,n-2}$.

- **Prueba de significancia de la regresión (Caso particular del test anterior: la conste es igual a cero).**

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

El no rechazar $H_0 : \beta_1 = 0$ implica que no hay relación lineal entre x e y. O sea:

- «x tiene muy poco valor para explicar la variación de y, por lo tanto el mejor estimador para cualquier x es $\hat{Y} = \bar{Y}$».
- «La verdadera relación entre x e y no es lineal».

Si se rechaza $H_0 : \beta_1 = 0$, explica que x tiene valor para explicar la variabilidad de y.

Rechazar $H_0 : \beta_1 = 0$ podría equivaler a que:

- «El modelo de línea recta es adecuado»
- «Aunque hay un efecto lineal en x se podrían obtener mejores resultados agregando términos polinomiales en x».

- **Predicción de nuevas observaciones**

Una aplicación importante del modelo de regresión es predecir nuevas observaciones y que correspondan a un nivel especificado de la variable de regresión x. Si x_0 es el valor de interés de la variable de regresión, entonces:

$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ es un estimador puntual del nuevo valor de la respuesta y_0.

Una aplicación importante del modelo de regresión es predecir nuevas observaciones y que correspondan a un nivel especificado de la variable de regresión x.

Si x_0 es el valor de interés de la variable de regresión, entonces:

$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ es un estimador puntual del nuevo valor de la respuesta y_0.

A continuación se obtendrá un estimado del intervalo para esta observación futura y_0.

Sea $\psi = y_0 - \hat{y}_0$, con distribución normal y media cero. Entonces su varianza es:

$$Var(\psi) = Var\left(y_0 - \hat{y}_0 \right)$$

$$= Var(y_0) + Var\left(\hat{y}_0 \right) - 2Cov\left(y_0, \hat{y}_0 \right)$$

$$= Var(y_0) + Var\left(\hat{\beta}_0 + \hat{\beta}_1 x_0 \right)$$

$$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$Var(\psi) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

El resultado de predicción de $100 \cdot (1-\alpha)\%$ de confianza para una observación futura en x_0 es:

$$\hat{y}_0 \pm t_{\alpha/2, n-2} \cdot \sqrt{CM\,\mathrm{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

- **Coeficiente de determinación**

La cantidad

$R^2 = \frac{SCR}{SCT} = 1 - \frac{SC\mathrm{Res}}{SCT}$. Se llama coeficiente de determinación.

Y su valor esperado: $E(R^2) = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\beta}_1^2 S_{xx} + \sigma^2}$

$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$: es una medida de variabilidad de y sin considerar el efecto de la variable de regresión x.

$SCRes = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$: es una medida de variabilidad de y que queda después de haber tenido en consideración a x.

R^2: Proporción de la variación explicada por la variable de regresión x.

Ya que $0 \le SCR \le SCT$: $0 \le \frac{SCR}{SCT} \le 1$; $0 \le R^2 \le 1$.

El coeficiente de determinación R^2 se expresa también en porciento para expresar el porciento de explicación de la variabilidad de los datos a través del modelo lineal.

Ejemplo

Se utilizan las pruebas de aptitud a 5 niños, antes y después de ser ayudados por sicólogos. Los resultados se presentan en la siguiente tabla

Niño	1	2	3	4	5	Total	Media
Antes	109	104	101	104	92	510	102
Después	110	103	100	105	92	510	102

Halle la ecuación de regresión lineal y diga si esta puede considerarse adecuada.

Solución:

	X	Y	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2(y_i - \bar{y})^2$
1	109	110	49	64	56
2	104	103	4	1	2
3	101	100	1	4	2
4	104	105	4	9	6
5	92	92	100	100	100
Total	510	510	158	178	166

A partir de los resultados de la tabla determinemos el coeficiente de correlación de Pearson:

$$r = \frac{\sum_{i=1}^{n} x_i' y_i'}{\sqrt{\sum_{i=1}^{n} x_i'^2 \sum_{i=1}^{n} y_i'^2}} = \frac{166}{158 \cdot 178} = 0.993$$ por lo que el modelo lineal es adecuado.

Estimemos ahora $\hat{\beta}_0$ y $\hat{\beta}_1$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$ Donde $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ y $S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$

$$\hat{\beta}_1 = \frac{166}{158} = 1.051$$

$$\hat{\beta}_0 = 102 - 1.051 \cdot (102) = -5.164$$

Luego, la ecuación de regresión es:

$$\hat{y} = -5.164 + 1.051\, x$$

Los intervalos de confianza del 95% para $\hat{\beta}_0$ y $\hat{\beta}_1$ son:

$$0.783 \leq \beta_1 \leq 1.324$$

$$-32.476 \leq \beta_0 \leq 22.146$$

El coeficiente de determinación para este caso es 0.976, lo que soporta que la ecuación obtenida es adecuada.

BIBLIOGRAFÍA

Abad Montes, F., Huete Morales, M.D., & Vargas, M. (2001). *Estadística para las ciencias sociales y laborales*. Ed. José Carlos Urbano Delgado, Granada.

Amor, R., Aguilar, C., & Morales, A. (2005). *Estadística Aplicada*. Grupo Editorial Universitario.

Batanero, C. (2001). *Didáctica de la Estadística*. Granada: Universidad de Granada. https://www.ugr.es/~batanero/pages/ARTICULOS/didacticaestadistica.pdf

Batanero, C., Díaz, C., Contreras, J. M., & Roa, R. (2013). El sentido estadístico y su desarrollo. *Números. Revista de didáctica de las Matemáticas*, 83, 7-18. http://funes.uniandes.edu.co/3651/1/Batanero2013ElNumeros83.pdf

Berenson, L. (1998). *Estadística para Administración y Economía*. Editorial Mc Graw – Hill.

Canavos (1987). *Probabilidad y Estadística: Aplicaciones y métodos*. Ed. McGraw Hill.

Casas, J. M. (1996). *Inferencia Estadística para la Economía y la Administración de Empresas*. Ed. Centro de Estudios Ramón Areces, S.A.

Casas, J.M., García, C., Rivera, L.F., & Zamora, A.I. (1998). *Problemas de Estadística. Descriptiva, probabilidad e inferencia*. Ed. Pirámide.

Choque T. J. (2008). *Fundamentos de Inferencia Estadística*. Oruro- Bolivia: Editorial Latina Editores

Choque, T. J. (2003). *Fundamentos de Cálculo de probabilidades*. Oruro-Bolivia: Editorial Latina Editores (Segunda edición).

Gutierrez Jáimez, R. Martínez Almécija, A., & Rodríguez Torreblanca, C. (1993) *Curso básico de probabilidad*. Ed. Pirámide.

Herrerías, R. (2004). *Ejercicios resueltos de inferencia estadística y del modelo lineal simple*. Ed. Delta. Publicaciones Universitarias. Madrid.

Herrerías, R., & Palacios, F. (2007) *Curso de inferencia estadística y del modelo lineal simple*. Ed. Delta. Publicaciones Universitarias. Madrid.

Ibagué, J. E. A. (2021). *Estadística descriptiva, regresión y probabilidad con aplicaciones*. Ediciones de la U.

Kazmier (1999). *Estadística aplicada a la Administración y Economía*. Editorial Mc Graw – Hill.

Martín, D. R. (2022). *Estadística inferencial aplicada*: Segunda edición revisada y aumentada. Universidad del Norte.

Martinez B. (2003). *Estadística y muestreo*. Editorial Ecoe ediciosnes.

Mood, A. M., Graybil, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statisitics* (3th edition). McGraw-Hill. Japan.

Morán, L. L., & Alonso, J. H. (2019). *Estadística descriptiva*. Sanz y Torres.

Newbold, P. (1997). *Estadística para los negocios y la economía*. Prentice Hall.

Palacios, F., & Callejón, J. (2002). *Mapas conceptuales, formulario y tablas de Técnicas cuantitativas II*. Ed. Plácido Cuadros, S.L. Granada.

Peña, C. G., & Fernández, C. A. M. (2019). *Estadística descriptiva y probabilidad*. Editorial Bonaventuriano.

Pereyra, L. E. (Ed.). (2021). *Probabilidad y estadística*. Klik.

Rohatgi (1976). *An introduction to Probability Theory and Mathematical Statistics*. Ed. John Wiley and Sons.

Salcedo, A., González, J., & González, J. (2021). Lectura e interpretación de gráficos estadísticos, ¿cómo lo hace el ciudadano?. *Revista Paradigma*, 42(Extra 1), 61-88.

Sánchez, G., & Manzano, V. (2002). Hipótesis Alternativa. Boletín de IASE para España, México y Venezuela, 3(2), 6-11. http://portal.ucv.ve/fileadmin/user_upload/cies/Documentos/Hipotesis_alternativa_N7.pdf

Sánchez, G., & Manzano, V. (2002). Hipótesis Alternativa. Boletín de IASE para España, México y Venezuela, 3(2), 6-11. http://portal.ucv.ve/fileadmin/user_upload/cies/Documentos/Hipotesis_alternativa_N7.pdf

Támara, L. G. (2013). *Estadística descriptiva y probabilidad*. Universidad Jorge Tadeo Lozano.

Utts, J. M. (2005). *Seeing through statistics*. Ed. Belmont, CA: Thomson.

Wild, C.J. (2000). *Chance encounters: a first course in data analysis and inference*. Ed. John Wiley and Sons. New York.

ISBN: 978-9942-44-273-4