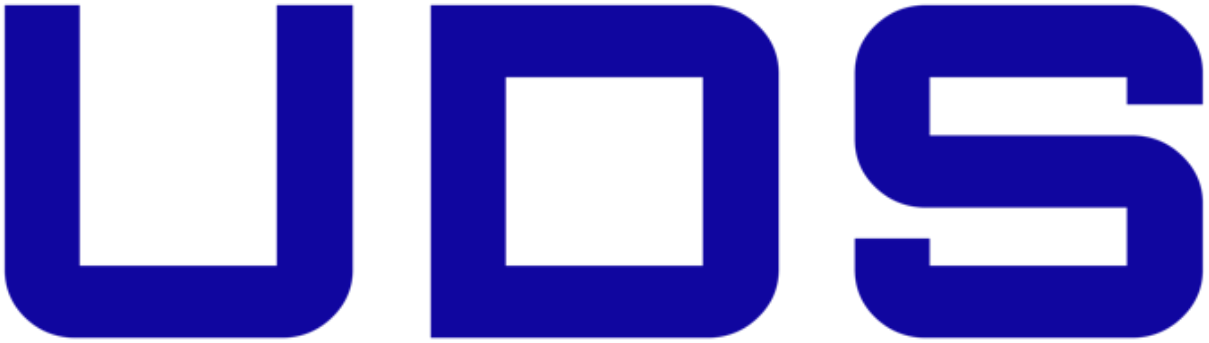


ANTOLOGIA

ESTADÍSTICA INFERENCIAL EN NUTRICIÓN

LICENCIATURA EN NUTRICIÓN

4º CUATRIMESTRE

Marco Estratégico de Referencia

ANTECEDENTES HISTORICOS

Nuestra Universidad tiene sus antecedentes de formación en el año de 1979 con el inicio de actividades de la normal de educadoras "Edgar Robledo Santiago", que en su momento marcó un nuevo rumbo para la educación de Comitán y del estado de Chiapas. Nuestra escuela fue fundada por el Profesor de Primaria Manuel Albores Salazar con la idea de traer Educación a Comitán, ya que esto representaba una forma de apoyar a muchas familias de la región para que siguieran estudiando.

En el año 1984 inicia actividades el CBTiS Moctezuma Ilhuicamina, que fue el primer bachillerato tecnológico particular del estado de Chiapas, manteniendo con esto la visión en grande de traer Educación a nuestro municipio, esta institución fue creada para que la gente que trabajaba por la mañana tuviera la opción de estudiar por las tarde.

La Maestra Martha Ruth Alcázar Mellanes es la madre de los tres integrantes de la familia Albores Alcázar que se fueron integrando poco a poco a la escuela formada por su padre, el Profesor Manuel Albores Salazar; Víctor Manuel Albores Alcázar en septiembre de 1996 como chofer de transporte escolar, Karla Fabiola Albores Alcázar se integró como Profesora en 1998, Martha Patricia Albores Alcázar en el departamento de finanzas en 1999.

En el año 2002, Víctor Manuel Albores Alcázar formó el Grupo Educativo Albores Alcázar S.C. para darle un nuevo rumbo y sentido empresarial al negocio familiar y en el año 2004 funda la Universidad Del Sureste.

La formación de nuestra Universidad se da principalmente porque en Comitán y en toda la región no existía una verdadera oferta Educativa, por lo que se veía urgente la creación de una institución de Educación superior, pero que estuviera a la altura de las exigencias de los jóvenes que tenían intención de seguir estudiando o de los profesionistas para seguir preparándose a través de estudios de posgrado.

Nuestra Universidad inició sus actividades el 18 de agosto del 2004 en las instalaciones de la 4ª avenida oriente sur no. 24, con la licenciatura en Puericultura, contando con dos grupos de cuarenta alumnos cada uno. En el año 2005 nos trasladamos a nuestras propias instalaciones en la carretera Comitán – Tzimol km. 57 donde actualmente se encuentra el campus Comitán y el Corporativo UDS, este último, es el encargado de estandarizar y controlar todos los procesos operativos y Educativos de los diferentes Campus, Sedes y Centros de Enlace Educativo, así como de crear los diferentes planes estratégicos de expansión de la marca a nivel nacional e internacional.

Nuestra Universidad inició sus actividades el 18 de agosto del 2004 en las instalaciones de la 4ª avenida oriente sur no. 24, con la licenciatura en Puericultura, contando con dos grupos de cuarenta alumnos cada uno. En el año 2005 nos trasladamos a nuestras propias instalaciones en la carretera Comitán – Tzimol km. 57 donde actualmente se encuentra el campus Comitán y el corporativo UDS, este último, es el encargado de estandarizar y controlar todos los procesos operativos y educativos de los diferentes campus, así como de crear los diferentes planes estratégicos de expansión de la marca.

MISIÓN

Satisfacer la necesidad de Educación que promueva el espíritu emprendedor, aplicando altos estándares de calidad Académica, que propicien el desarrollo de nuestros alumnos, Profesores, colaboradores y la sociedad, a través de la incorporación de tecnologías en el proceso de enseñanza-aprendizaje.

VISIÓN

Ser la mejor oferta académica en cada región de influencia, y a través de nuestra Plataforma Virtual tener una cobertura Global, con un crecimiento sostenible y las ofertas académicas innovadoras con pertinencia para la sociedad.

VALORES

- Disciplina
- Honestidad
- Equidad
- Libertad

ESCUDO

El escudo de la UDS, está constituido por tres líneas curvas que nacen de izquierda a derecha formando los escalones al éxito. En la parte superior está situado un cuadro motivo de la abstracción de la forma de un libro abierto.

ESLOGAN

"Mi Universidad"

ALBORES

Es nuestra mascota, un Jaguar. Su piel es negra y se distingue por ser líder, trabaja en equipo y obtiene lo que desea. El ímpetu, extremo valor y fortaleza son los rasgos que distinguen.

Nombre de la materia

Objetivo de la materia:

Que el alumno consolide la competencia habilitante de la lectura y escritura al reconocer y ejercer las cuatro habilidades de la lengua: escuchar, leer, hablar y escribir, con el fin de aplicarlas a diversas situaciones de su vida, académicas y cotidianas.

Criterios y procedimientos de evaluación y acreditación:

Actividad en plataforma	30%
Tareas	10%
Examen	60%
Total	100%
Escala de calificaciones	7-10
Mínima aprobatoria	7

INDICE

UNIDAD I

INTRODUCCIÓN A LA ESTADÍSTICA INFERENCIAL EN NUTRICIÓN

1.1 Breve historia de la estadística

La palabra Estadística procede del vocablo "Estado", pues era función principal de los Gobiernos de los Estados establecer registros de población, nacimientos, defunciones, impuestos, cosechas... La necesidad de poseer datos cifrados sobre la población y sus condiciones materiales de existencia han debido hacerse sentir desde que se establecieron sociedades humanas organizadas.

Es difícil conocer los orígenes de la Estadística. Desde los comienzos de la civilización han existido formas sencillas de estadística, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o ciertas cosas.

Su origen empieza posiblemente en la isla de Cerdeña, donde existen monumentos prehistóricos pertenecientes a los Nuragas, los primeros habitantes de la isla; estos monumentos constan de bloques de basalto superpuestos sin mortero y en cuyas paredes de encontraban grabados toscos signos que han sido interpretados con mucha verosimilidad como muescas que servían para llevar la cuenta del ganado y la caza.

Hacia el año 3.000 a.C. los babilonios usaban ya pequeñas tablillas de arcilla para recopilar datos en tablas sobre la producción agrícola y los géneros vendidos o cambiados mediante trueque.

Los egipcios ya analizaban los datos de la población y la renta del país mucho antes de construir las pirámides. En los antiguos monumentos egipcios se encontraron interesantes documentos en que demuestran la sabia organización y administración de este pueblo; ellos llevaban cuenta de los movimientos poblacionales y continuamente hacían censos.

Tal era su dedicación por llevar siempre una relación de todo que hasta tenían a la diosa Safnkit, diosa de los libros y las cuentas. Todo esto era hecho bajo la dirección del Faraón y fue a partir del año 3050 a.C.

En la Biblia observamos en uno de los libros del Pentateuco, bajo el nombre de Números, el censo que realizó Moisés después de la salida de Egipto. Textualmente dice: "Censo de las tribus: El día primero del segundo año después de la salida de Egipto, habló Yavpe a Moisés en el desierto de Sinaí en el tabernáculo de la reunión, diciendo: "Haz un censo general de toda la asamblea de los hijos de Israel, por familias y por linajes, describiendo por cabezas los nombres de todos los varones aptos para el servicio de armas en Israel. En el libro bíblico Crónicas describe el bienestar material de las diversas tribus judías.

En China existían los censos chinos ordenados por el emperador Tao hacia el año 2.200 a.C.

Posteriormente, hacia el año 500 a.C., se realizaron censos en Roma para conocer la población existente en aquel momento. Se erigió la figura del censor, cuya misión consistía en controlar el número de habitantes y su distribución por los distintos territorios.

En la Edad Media, en el año 762, Carlomagno ordenó la creación de un registro de todas sus propiedades, así como de los bienes de la iglesia.

Después de la conquista normanda de Inglaterra en 1.066, el rey Guillermo I, el Conquistador, elaboró un catastro que puede considerarse el primero de Europa.

Los Reyes Católicos ordenaron a Alonso de Quintanilla en 1.482 el recuento de fuegos (hogares) de las provincias de Castilla.

En 1.662 un mercader de lencería londinense, John Graunt, publicó un tratado con las observaciones políticas y naturales, donde Graunt pone de manifiesto las cifras brutas de nacimientos y defunciones ocurridas en Londres durante el periodo 1.604-1.661, así como las influencias que ejercían las causas naturales, sociales y políticas de dichos acontecimientos. Puede considerarse el primer trabajo estadístico serio sobre la población.

Curiosamente, Graunt no conocía los trabajos de B. Pascal » (1.623-1.662) ni de C. Huygens (1.629-1.695) sobre estos mismos temas. Un poco más tarde, el astrónomo Edmund Halley (1.656- 1.742) presenta la primera tabla de mortalidad que se puede considerar como base de los estudios contemporáneos. En dicho trabajo se intenta establecer el precio de las anualidades a satisfacer a las compañías de seguros. Es decir, en Londres y en París se estaban construyendo, casi de manera simultánea, las dos disciplinas que actualmente llamamos estadística y probabilidad.

En el siglo XIX, la estadística entra en una nueva fase de su desarrollo con la generalización del método para estudiar fenómenos de las ciencias naturales y sociales. Galton » (1.822-1.911) y Pearson (1.857-1936) se pueden considerar como los padres de la estadística moderna, pues a ellos se debe el paso de la estadística deductiva a la estadística inductiva.

Los fundamentos de la estadística actual y muchos de los métodos de inferencia son debidos a R. A. Fisher. Se interesó primeramente por la eugenesia, lo que le conduce, siguiendo los pasos de Galton a la investigación estadística, sus trabajos culminan con la publicación de la obra Métodos estadísticos para investigaciones. En él aparece la metodología estadística tal y como hoy la conocemos.

A partir de mediados del siglo XX comienza lo que podemos denominar la estadística moderna, uno de los factores determinantes es la aparición y popularización de los computadores. El centro de gravedad de la metodología estadística se empieza a desplazar

técnicas de computación intensiva aplicadas a grandes masas de datos, y se empieza a considerar el método estadístico como un proceso iterativo de búsqueda del modelo ideal

Las aplicaciones en este periodo de la Estadística a la Economía conducen a una disciplina con contenido propio: la Econometría. La investigación estadística en problemas militares durante la segunda guerra mundial y los nuevos métodos de programación matemática, dan lugar a la Investigación Operativa

1.2 Concepto de estadística

La estadística se ocupa de la sistematización, recogida, ordenación y representación de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de hacer previsiones sobre los mismos, tomar decisiones u obtener conclusiones. Teniendo en cuenta las funciones podemos considerar dos grandes áreas:

Estadística descriptiva: se organizan y resúmenes conjuntos de observaciones procedentes de una muestra o de la población total, en forma cuantitativa. Los procedimientos para una variable: índices de tendencia general, estadísticos de variabilidad y estadísticos de asimetría; y para dos variables: coeficientes de correlación y ecuaciones de regresión.

Estadística inferencial: se realizan inferencias acerca de una población basándose en los datos obtenidos a partir de una muestra. Los procedimientos: el cálculo de probabilidades.

Conceptos importantes: población es el conjunto de todos los elementos que cumplen una determinada característica objeto de estudio. Muestra es un subconjunto de una población.

Parámetro es una propiedad descriptiva (medida) de una población. Estadístico es una propiedad descriptiva (medida) de una muestra.

Las conclusiones obtenidas de una muestra sólo servirán para el total de una población si la muestra es representativa. Para asegurarnos que la muestra es representativa se utilizan métodos de muestreo probabilístico.

También existes las muestras no probabilísticas como por ejemplo la muestra de conveniencia o incidental.

1.3 Estadística descriptiva

La estadística descriptiva es la rama de las Matemáticas que recolecta, representa y caracteriza un conjunto de datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, en los meses de verano, etc.) con el fin de describir apropiadamente las diversas características de ese conjunto.

La estadística descriptiva: se dedica a la descripción, visualización y resumen de datos originados a partir de los fenómenos de estudio. Los datos pueden ser resumidos numérica o gráficamente. Ejemplos básicos de parámetros estadísticos son: la media y la desviación estándar. Algunos ejemplos gráficos son: histograma, pirámide poblacional, gráfico circular, entre otros.

1.4 Estadística Inferencial

Se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones. Se usa para modelar patrones en los datos y extraer inferencias acerca de la población bajo estudio. Estas inferencias pueden tomar la forma de respuestas a preguntas si/no (prueba de hipótesis), estimaciones de unas características numéricas (estimación), pronósticos de futuras observaciones, descripciones de asociación (correlación) o modelamiento de relaciones entre variables (análisis de regresión). Otras técnicas de modelamiento incluyen a nova, series de tiempo y minería de datos.

IMPORTANCIA DE LA ESTADISTICA INFERENCIAL

La Estadística Inferencial puede dar respuesta a muchas de las necesidades que la sociedad actual puede requerir. Su tarea fundamental es el análisis de los datos que se obtienen a partir de experimentos, con el objetivo de representar la realidad y conocerla. Permite la recolección de datos importantes para el estudio de situaciones que se presentan a diario y permite dar respuesta a los problemas de una forma útil y significativa.

La Estadística Inferencial se centra en tomar una pequeña muestra representativa de la población y a partir de ésta, infiere que el resto de la población tiene el mismo comportamiento.

En caso de que no sea factible realizar un estudio completo por cuestiones de tiempo, recursos o costo, se puede calcular un tamaño de muestra para medir solo algunos elementos de la población, posteriormente se infiere que el resto de la población se comporta igual que la muestra tomada

1.5 Breve introducción a la inferencia estadística

El principal objetivo de la Estadística es inferir o estimar características de una población que no es completamente observable (o no interesa observarla en su totalidad) a través del análisis de una parte de ella a la que llamamos muestra. Las razones por las que generalmente se trabaja con muestras son principalmente:

- Económicas.
- Tiempo: si la población es muy grande llevaría tanto tiempo analizarla que incluso la característica de interés podría variar en ese período. Por ejemplo, la tasa de paro.
- Destrucción: la medición de cierta característica podría llevar a la destrucción del individuo. Por ejemplo, al estudiar la supervivencia de ciertos animales a un tratamiento.

Lo que se hace entonces es analizar la muestra y las conclusiones desde la muestra a la población. Ahora bien, para considerar válidas en la población las conclusiones obtenidas en la muestra, ésta ha de representar bien a la población (representativa). Por lo tanto, la selección de la muestra es de suma importancia, y para ello hay diversos métodos (métodos de muestreo). Cuando se intuye que la característica en estudio puede presentar valores homogéneos en la población, una forma de obtener una muestra representativa es eligiéndola al azar. A este método de selección de la muestra se le llama muestreo aleatorio simple y es el más sencillo.

La Inferencia Estadística se puede clasificar en inferencia paramétrica e inferencia no paramétrica.

La inferencia paramétrica tiene lugar cuando se conoce la distribución de la variable de estudio en la población, y el interés recae sobre los parámetros desconocidos de la misma.

La inferencia no paramétrica tiene lugar si no se conoce la distribución y sólo se suponen propiedades generales de la misma

1.6 Teoría de decisión en estadística

Estudio formal sobre la toma de decisiones. Los estudios de casos reales, que se sirven de la inspección y los experimentos, se denominan teoría descriptiva de decisión; los estudios de la toma de decisiones racionales, que utilizan la lógica y la estadística, se llaman teoría preceptiva de decisión. Estos estudios se hacen más complicados cuando hay más de un individuo, cuando los resultados de diversas opciones no se conocen con exactitud y cuando las probabilidades de los distintos resultados son desconocidas. La teoría de decisión comparte características con la teoría de juegos, aunque en la teoría de decisión el „adversario" es la realidad en vez de otro jugador o jugadores.

Al hacer un análisis sobre esta teoría, y mirándola desde el punto de vista de un sistema, se puede decir que al tomar una decisión sobre un problema en particular, se debe tener en cuenta los puntos de dificultad que lo componen, para así empezar a estudiarlos uno a uno hasta obtener una solución que sea acorde a lo que se está esperando obtener de este, y si no, buscar otras soluciones que se acomoden a lo deseado.

La teoría de decisión, no solamente se puede ver desde el punto de vista de un sistema, sino en general, porque esta se utiliza a menudo para tomar decisiones de la vida cotidiana, ya que muchas personas piensan que la vida es como una de las teorías; La teoría del juego, que para poder empezarlo y entenderlo hay que saber jugarlo y para eso se deben conocer las reglas de este, para que no surjan equivocaciones al empezar la partida

1.7 Componentes de una investigación estadística

El estudio estadístico de una situación con propósitos inferenciales se centra en dos conceptos fundamentales: población y muestra, los cuales serán definidos a continuación:

Población. Es el conjunto formado por todos los valores posibles que puede asumir, la variable objeto de estudio. Así por ejemplo, en un estudio sobre la preferencia de los votantes en una elección presidencial, la población consiste en todas las respuestas de los votantes registrados. Pero el término no sólo está asociado a la colección de seres humanos u organismos vivos; y tenemos así que, si se va a hacer una investigación de las ventas anuales de los supermercados, entonces las ventas anuales de todos los supermercados constituyen así mismo la población.

Es bueno tener en cuenta que el término población se interpreta de dos maneras cuando se hace un estudio estadístico, a saber:

1. La interpretación propia en el Análisis Estadístico, que corresponde a la que hemos presentado anteriormente.
2. Como el conjunto de objetos sobre los cuales actúa la variable considerada. Por tanto, no es extraño escuchar expresiones tales como, "se hizo un estudio de los niveles de ingreso de la población trabajadora colombiana", entendiéndose con ello que el elemento estadístico objeto de análisis fue el registro numérico de los ingresos.

Muestra. Es cualquier subconjunto de la población, escogido al seguir ciertos criterios de selección. La muestra es el elemento básico sobre el cual se fundamenta la posterior inferencia acerca de la población de donde se ha tomado. Por ello, su escogencia y selección debe hacerse siguiendo ciertos procedimientos que son ampliamente tratados en la parte de la estadística llamada Teoría de muestreo. El concepto de muestra tiene también las dos connotaciones que hemos señalado para la población. Las características de una población se resumen para su estudio generalmente irá mediante lo que se denominan parámetros; éstos a su vez se toman o consideran como valores verdaderos de la característica estudiada. Por ejemplo, la proporción de todos los clientes que declaran cierta preferencia por una marca particular de un producto dado, es un parámetro de la población de todos los clientes; es la verdadera proporción de la población. Igualmente, la media aritmética de las cuentas corrientes de los clientes de un banco determinado constituye un parámetro de la población de las cuentas de los clientes de ese banco.

Cuando la característica de la población estudiada se reduce a una muestra el resumen de esa característica se hace mediante una está (medida) o estadígrafo. Así por ejemplo. si se toman 100 de todos los posibles clientes y se les entrevista hará ver si están a favor de una marca particular de un producto, estos 100 clientes la constituyen una muestra.. Si hay 70 clientes que prefieren dicha marca entonces la proporción maestral será 0.70 y constituirá un estadígrafo; de igual manera si se escogen 1,000 cuentas del total de las cuentas comentes; las 1,000 observaciones conforman una muestra y el promedio aritmético de estas cuentas un estimador. La inferencia estadística se orienta a sacar conclusiones acerca del parámetro o parámetros poblacionales con base en el valor de un estimador obtenido a partir de los datos muéstrales extraídos de esa población. Para llegar a ese objetivo a

través de un proceso racional y eficaz, se aconseja que se tengan en cuenta los siguientes pasos:

1. Formulación del problema. En este punto se debe especificar de manera clara la pregunta que se debe responder y la población de datos asociada a la pregunta. Los conceptos deben ser precisos y deben ponerse limitaciones adecuadas al problema motivadas por el tiempo, dinero disponible y la habilidad de los Investigadores. Algunos conceptos como, artículo defectuoso, económico, salario, pueden variar en cada caso y para cada problema debemos coincidir con las ideas señaladas en el estudio.
2. Diseño del experimento. Este aspecto es de gran importancia, puesto que la recolección de datos requiere dinero y tiempo. Es siempre nuestro deseo obtener máxima Información con el mínimo costo (dinero y tiempo) posible. Incluir excesiva Información en la muestra es a menudo costoso y antieconómico. Incluir poca también es poco satisfactorio. Esto implica, entre otras cosas, que debemos determinar el tamaño de la muestra o la cantidad o tipo de datos que nos permita resolver el problema de la manera más eficiente.
3. Recolección de datos. Esta parte, por lo general, es la que exige más tiempo en la Investigación. Esta recolección debe ajustarse a reglas estrictas ya que de los datos esperamos extraer la Información deseada.
4. Tabulación y descripción de los resultados. En esta etapa, los datos muéstrales se exponen de manera clara y se ilustran con representaciones tabulares y gráficas (diagramas. histogramas, etc.); además se calculan las medidas estadísticas apropiadas al proceso inferencial que haya sido escogido.
5. Inferencia estadística y conclusiones. Este último paso constituye tal vez la contribución más importante de la estadística al proceso inferencial. Aquí se fija el nivel de confiabilidad para la inferencia; esto es debido a que las conclusiones derivadas de inferencias estadísticas jamás se pueden tomar con un 100% de certeza, pero sí se les puede asociar un nivel de confiabilidad; en términos de probabilidad denominados nivel de confianza y nivel de significancia. El proceso Inferencial nos llevará a una conclusión estadística que servirá de orientación a quien o quienes deban tomar la decisión (administrativa o clínica) sobre el tema objeto de estudio.

1.8 Recolección de datos

La recolección de datos se refiere al uso de una gran diversidad de técnicas y herramientas que pueden ser utilizadas por el analista para desarrollar los sistemas de información, los cuales pueden ser la entrevistas, la encuesta, el cuestionario, la observación, el diagrama de flujo y el diccionario de datos.

Para el caso de la materia de control estadístico de la calidad la recolección de datos se realiza mediante la utilización de hojas de verificación o comprobación, estos son formatos especialmente constituidos para colectar datos fácilmente, en la que todos los artículos o factores necesarios son previamente establecidos y en la que los registros de pruebas, resultados de inspección o resultados de operaciones son fácilmente descritos con marcas utilizadas para verificar.

EJEMPLO:

Hoja de verificación		
Fecha: 12-02-2012	Fabrica: Estación de Servicio "Virgen del Valle"	Inspector: Grupo de Trabajo
Tipo de defectos: varios		

Tipo de defectos	Verificación	subtotal
El acondicionamiento de los surtidores.	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	40
Las altas temperaturas producidas por la máquina.	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 II	50
Fallas en los componentes de los surtidores.	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	60
Falta de materia prima.	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	120
Los operarios no respetan su hora de descanso.	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	78
La Estructura.	𝍸 𝍸 𝍸 𝍸 I	25
Tiempo de ocio por parte de los operarios al manejar los surtidores.	𝍸 𝍸	10
Otros.	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	70
		Total = 453

1.9 Estadística paramétrica

La estadística paramétrica es una rama de la estadística inferencial que comprende los procedimientos estadísticos y de decisión que están basados en distribuciones conocidas. Estas son determinadas usando un número finito de parámetros. Esto es, por ejemplo, si conocemos que la altura de las personas sigue una distribución normal, pero desconocemos cuál es la media y la desviación de dicha normal. La media y la desviación típica de la distribución normal son los dos parámetros que queremos estimar. Cuando desconocemos totalmente qué distribución siguen nuestros datos entonces deberemos aplicar primero un test no paramétrico, que nos ayude a conocer primero la distribución.

La mayoría de procedimientos paramétricos requiere conocer la forma de distribución para las mediciones resultantes de la población estudiada. Para la inferencia paramétrica es requerida como mínimo una escala de intervalo, esto quiere decir que nuestros datos deben tener un orden y una numeración del intervalo. Es decir nuestros datos pueden estar categorizados en: menores de 20 años, de 20 a 40 años, de 40 a 60, de 60 a 80, etc, ya que hay números con los cuales realizar cálculos estadísticos. Sin embargo, datos categorizados en: niños, jóvenes, adultos y ancianos no pueden ser interpretados mediante la estadística paramétrica ya que no se puede hallar un parámetro numérico (como por ejemplo la media de edad) cuando los datos no son numéricos

1.10 Población

Quizá, la definición teórica de población estadística sea un poco abstracta. Por eso, sin renunciar a la rigurosidad y precisión que requieren las variables cuantitativas, vamos a intentar abordar el concepto de población estadística de la forma más sencilla posible.

Empezaremos por la palabra población. ¿En qué piensas cuando lees o escuchas la palabra población? Muy probablemente en un número de personas. Por ejemplo, la población de Argentina, la población de Chile, la población de Nueva York o la población mundial. Y

dirás, ¿qué tiene que ver la población con la estadística? Pues tiene que ver mucho. Todo se remonta a los orígenes de la palabra estadística.

Con esto en mente, seguiremos la siguiente secuencia para entender el concepto: origen de la palabra, principales tipos de población y un ejemplo de población estadística.

Tipos de población estadística

Dentro de las poblaciones estadísticas, fundamentalemente dos tipos de poblaciones:

Población estadística finita: Es aquella en la que el número de valores que la componen tiene un fin. Por ejemplo, la población estadística que nos indica la cantidad de árboles de una ciudad es finita. Es cierto que puede variar con el tiempo, pero en un instante determinado es finita, tiene fin.

Población estadística infinita: Se trata de aquella población que no tiene fin. Por ejemplo, el número de planetas que existen en el universo. Aunque puede que sea finito, el número es tan grande y desconocido que estadísticamente se asume como infinito.

Adicionalmente, dentro de esta gran clasificación, existen otros tipos de poblaciones. Poblaciones según la distribución de los datos, según el tipo de dato (cualitativo o cuantitativo), etc.

1.11 Muestra aleatoria

Con el muestreo aleatorio, por tanto, lo que hacemos es plantear un método de elección. Un método que tiene en cuenta diferentes probabilidades. Esto lo diferencia de los métodos no aleatorios en que es la subjetividad del investigador la que decide la selección de la muestra.

A su vez, en este caso, el azar juega un papel significativo; ya que eliminamos la discreción.

¿Por qué utilizar el muestreo aleatorio?

Este tipo de muestreo es uno de los más utilizados en el método científico. Las razones son varias, pero las más relevantes serían las siguientes:

En primer lugar, es el único que permite hacer análisis confirmatorios e inferencia estadística. De hecho, la segunda se realiza también en muestreos no aleatorios, pero no podremos confirmar los resultados. En este caso, la investigación es de tipo exploratorio.

Por otro lado, relacionado con el apartado anterior, este método reduce el sesgo. Es decir, al tener cierta probabilidad (conocida) de escoger un determinado individuo de la población, evitamos la subjetividad inherente en la selección no aleatoria.

Por último, permite utilizar muestras de pequeño tamaño en poblaciones grandes. Eso sí, hay fórmulas para calcular esas muestras mínimas con poblaciones conocidas o desconocidas.

¿Cómo llevarlo a cabo?

- Como toda técnica utilizada en la ciencia, esta también se lleva a cabo siguiendo un proceso. Este permite replicar el experimento y reduce el sesgo y la subjetividad.
- El primer paso, y muy determinante, es la selección de la población. De hecho, tenemos que obtener toda la información que podamos. Sobre todo, nos interesa su composición por ciertas variables sociodemográficas como el sexo, la edad o la ocupación.
- Después, hay que elegir un muestreo aleatorio concreto. En el siguiente apartado veremos los más relevantes. La decisión dependerá de las características de la población.
- Una vez escogido el método, hay que calcular la muestra mínima. Para hacerlo, debemos tener en cuenta si conocemos, o no, el tamaño poblacional. Como hemos comentado, hay fórmulas para calcular este tamaño muestral.
- Por último, procedemos a obtener la muestra y a realizar en ella los análisis estadísticos pertinentes. Una vez hechos, podremos llevar a cabo un contraste de hipótesis u otros métodos de inferencia. El objetivo es extrapolar los resultados a la población.

Tipos de muestreo aleatorio

Existen varios tipos de muestreo aleatorio en función de las características poblacionales.

Veamos los más relevantes:

- Muestreo aleatorio simple: Es uno de los más utilizados. Consiste en asignar un número aleatorio a la población y, luego, a partir de este, escoger la muestra. Es muy útil en poblaciones con cierta homogeneidad. Por ejemplo, es muy utilizado en geología.
- Muestreo estratificado: En este caso, estamos ante una población que, si bien es heterogénea, se puede separar en grupos homogéneos (sexo, edad, etc.). En cada grupo se realiza una muestra aleatoria simple. Es muy utilizado en ciencias sociales, como la psicología.
- Muestreo por conglomerados: En este caso, el objetivo es crear una serie de bloques o conglomerados. Estos son elegidos al azar de entre toda la población. En este caso, existe una heterogeneidad dentro de ellos, así como una homogeneidad fuera. Las investigaciones de mercado suelen usar este muestreo aleatorio.
- Muestreo sistemático: En este caso, se divide el número de individuos de la población entre los de la muestra que queremos obtener. Después, escogemos uno al azar y vamos contando, utilizando ese valor. Los sujetos elegidos serán los que correspondan a ese recuento. Este tipo reduce el problema de autocorrelación.

0UNIDAD II

INFERENCIA ESTADÍSTICA: ESTIMACIÓN MUESTREO

2.1 Teoría de conjuntos

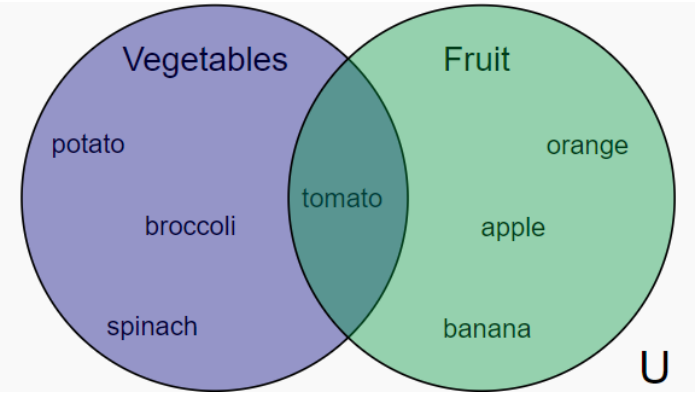
La teoría de conjuntos es una rama de las matemáticas que estudia conjuntos. Los conjuntos son una colección de objetos (normalmente) bien definidos. A continuación, se muestran algunos ejemplos:

{a, b, c, d, e}

{n | n ∈ ℕ, 1 ≤ n ≤ 10}

{verde, rojo, azul, amarillo, blanco, negro, violeta}

El diagrama de Venn muestra un conjunto compuesto por frutas y verduras.



Tenga en cuenta que un conjunto puede estar formado por prácticamente cualquier tipo de objeto. Si bien esto es cierto, la teoría de conjuntos se ocupa principalmente de objetos que son relevantes para las matemáticas. Los objetos de un conjunto se denominan elementos. Por ejemplo, en el diagrama de Venn, cada fruta o verdura es un elemento de su conjunto respectivo, y tanto las verduras como las frutas son parte del conjunto universal . La porción central donde se cruzan el conjunto de frutas y verduras contiene solo tomates, que se consideran una fruta botánicamente, pero comúnmente se consideran

una verdura en el contexto de la cocción. Este tipo de relaciones son la base de la teoría de conjuntos básica.

Notación y conceptos básicos de la teoría de conjuntos

En su nivel más básico, la teoría de conjuntos describe la relación entre objetos y si son elementos (o miembros) de un conjunto dado. Los conjuntos también son objetos y, por lo tanto, también se pueden relacionar entre sí normalmente mediante el uso de varios símbolos y notaciones.

Aunque la teoría de conjuntos puede parecer arbitraria y no necesariamente útil por sí sola, se usa en todas las matemáticas y se puede considerar como un bloque de construcción fundamental. Muchos conceptos matemáticos serían difíciles de definir con precisión (y concisión) sin el uso de la teoría de conjuntos. Como tal, es importante estar familiarizado con los diversos símbolos y notaciones que se utilizan en la teoría de conjuntos para comprender y comunicar conceptos matemáticos de manera eficaz. La siguiente tabla incluye algunos de los símbolos más comunes.

Símbolo	Definición/significado	Ejemplo
{}	Indica una colección de elementos	{1, 3, 7, 9}
∅	Conjunto vacío : el conjunto no contiene elementos	{}
…	Indica que el conjunto continúa el patrón en la dirección correspondiente hacia el infinito negativo (izquierda) o positivo (derecha)	{…, -9, -7, -3, -1, 1, 3, 7, 9, …}
|	«Tal que»	a∈ℤ | a & gt; 3 – «a es un número entero tal que a es mayor que 3»

∩	«y» o «intersección»	A={1, 2, 3, 4} B={4, 5, 6} A ∩ B={4}
∪	«o» o «unión»	A={1, 2, 3, 4} B={4, 5, 6} A ∪ B={1, 2, 3, 4, 5, 6 }
⊆	Subconjunto : A es un subconjunto de B si todos sus elementos están incluidos en B	{1, 2} ⊆ {1, 2, 3, 4, 5} {1, 2, 3, 4, 5} ⊆ {1, 2, 3, 4, 5 }
⊂	Subconjunto adecuado/estricto : A es un subconjunto adecuado de B si A es un subconjunto de B, pero no es igual a B	{1} ⊂ {1, 2, 3, 4, 5} {1, 2} ⊂ {1, 2, 3, 4, 5} {1, 2 , 3} ⊂ {1, 2, 3, 4, 5} {1, 2, 3, 4} ⊂ {1, 2, 3, 4, 5}
∈	Elemento de : indica que el objeto a la izquierda del símbolo es un elemento del objeto a la derecha	x∈ℚ – «x es un elemento de los números racionales»
	Conjunto universal : el conjunto de todos los valores posibles	A={1, 2} B={3, 4, 5} ={1, 2, 3, 4, 5}
A c o A ‘	Complemento : todos los elementos que no están en el conjunto A	={1, 2, 3, 4, 5} A={1, 2, 3} A c ={4,5 }
[a, b]	Intervalo cerrado : valores entre ayb, incluidos ayb	[1,4]={1, 2, 3, 4} si solo incluye números enteros
(a, b)	Intervalo abierto : valores entre ayb sin incluir ayb	(1,4)={2, 3} si solo incluye números enteros
| A |	Orden/cardinalidad – número de elementos en el conjunto	A={3, 6, 7, 9} | A |=4

$\mathbb{N}$	Números naturales : solo números positivos sin decimales ni fracciones	{1, 2, 3, …}
$\mathbb{Z}$	Enteros : todos los números positivos y negativos sin decimales ni fracciones, incluido 0	{…, -3, -2, -1, 0, 1, 2, 3, …}
$\mathbb{Q}$	Números racionales : un número que se puede representar como una fracción compuesta por números enteros	$\frac{2}{3}$
$\mathbb{R}$	Números reales : números racionales y números irracionales	π, e , 3, ½, 0.25
$\mathbb{C}$	Números complejos : números formados por un componente real e imaginario	$4 + 2i$

Orden e igualdad de conjuntos

El orden de un conjunto se refiere al tamaño de un conjunto. También se conoce como cardinalidad del conjunto. Los conjuntos pueden tener un orden finito o infinito. Si un conjunto tiene un orden finito, el orden de un conjunto está determinado por el número de elementos en el conjunto. Por ejemplo, el conjunto A={1, 2, 5, 7, 9} tiene un orden de 5, ya que contiene 5 elementos. Usando la notación de conjuntos, podríamos expresar el orden de A como:

$|A|=4$

Tenga en cuenta que el orden de los elementos en un conjunto no importa. Por ejemplo, dados los conjuntos

A={1, 2, 5, 7, 9}

B={1, 5, 2, 9, 7}

Diríamos que A y B son conjuntos iguales, o A=B. Esto se debe a que la igualdad de conjuntos está determinada por los elementos dentro del conjunto, no por el orden en que se enumeran los elementos.

Ejemplo

Dados los conjuntos

A={5, 3, 1}

B={3, 1, 5, 13, 10,}

C={2, 10, 6, 4}

consulte la tabla según sea necesario y determine los resultados de las siguientes operaciones:

A ∩ B

A ∩ C

B ∪ C

i. A ∩ B={1, 3, 5}

Debido a que todos los elementos de A también están en B, también podemos decir que A es un subconjunto propio de B, o A ⊂ B. Además, | A |=3 y | B |=5.

ii. A ∩ C=∅

A y C no tienen elementos comunes, por lo que su intersección es un conjunto vacío. También podemos decir que A y B son conjuntos mutuamente excluyentes.

iii. B ∪ C={1, 2, 3, 4, 5, 6, 10, 13}

La unión de B y C es el conjunto que contiene todos los elementos de B y C.

Conjuntos infinitos

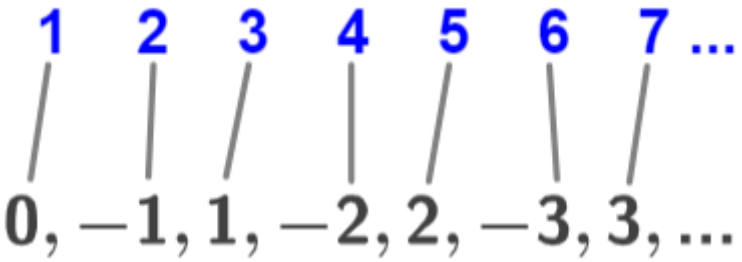
Los conjuntos pueden ser finitos o infinitos. Además, los conjuntos infinitos pueden ser contables o incontables.

Contable

Cualquier conjunto infinito que pueda emparejarse con los números naturales en una correspondencia uno a uno, de modo que cada uno de los elementos del conjunto pueda identificarse uno a la vez es un conjunto infinito numerable. Por ejemplo, dado el conjunto

{0, -1, 1, -2, 2, -3, 3, …}

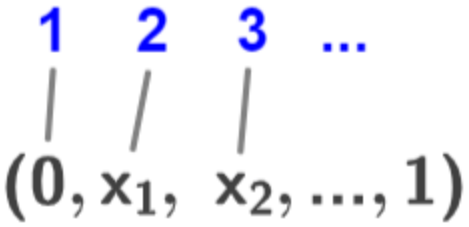
sus elementos se pueden emparejar con un número natural de la siguiente manera:



Por lo tanto, es posible identificar el elemento n como el número natural n .

Incontable

Los conjuntos innumerables no se pueden organizar de la misma manera que los conjuntos infinitos contables. El conjunto de números reales de cero a uno, o (0, 1), no es contable porque no es posible emparejar cada uno de los elementos del conjunto con un elemento único en el conjunto de números naturales. Por ejemplo, sea {0, x_1, x_2, …, 1} el conjunto de números reales de cero a uno donde $x_1 \neq x_2$. La figura siguiente muestra que los elementos no se pueden mapear de la misma manera que la figura anterior (con conjuntos infinitos contables):



Los números naturales se pueden asignar a los elementos, como se muestra en la figura, pero debido a que hay números reales entre dos números reales distintos, no hay números naturales para asignar a los números reales entre 0 y x_1, x_1 y x_2, y así sucesivamente. Por lo tanto, el conjunto es incontablemente infinito.

Establecer operaciones

Algunas de las operaciones básicas de conjuntos (unión e intersección) se discutieron anteriormente. A continuación se muestran algunas otras operaciones.

Producto cartesiano

El producto cartesiano de A y B, denotado A × B, es el conjunto compuesto por todos los pares ordenados (a, b) de manera que a es un elemento de A y b es un elemento de B. Usando set-builder notación, esto se puede denotar como:

Tenga en cuenta que el orden en el que se escriben los seis elementos del conjunto no importa, pero sí el orden de los pares ordenados. Por ejemplo, (a, i) no es el mismo par ordenado que (i, a). Generalmente, si hay m elementos en A y n elementos en B, hay m · n elementos en A × B.

Conjunto de energía

Un conjunto de potencias es un conjunto que se compone de todos los posibles subconjuntos de un conjunto. Sea A={1, 2, 3}. El conjunto de potencias de A, denotado ℘ (A), es:

℘ (A)={{}, {1}, {2}, {3}, {1, 2}, {1, 3}, {2, 3}, {1, 2, 3}}

Observe que tanto el conjunto vacío como el conjunto A se consideran subconjuntos de A. En general, si hay n elementos en A, hay 2 n subconjuntos en ℘ (A).

Leyes de De Morgan

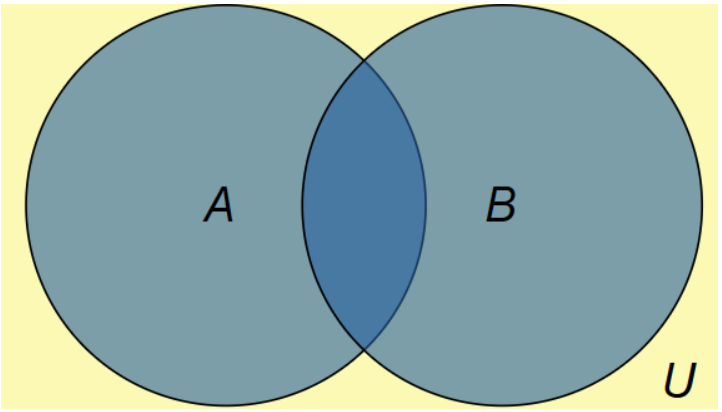
En la teoría de conjuntos, las leyes de De Morgan son un conjunto de reglas que relacionan la unión y la intersección de conjuntos a través de sus complementos.

Unión de conjuntos:

El complemento de la unión de dos conjuntos es igual a la intersección de sus complementos:

(A ∪ B) C =A C ∩ B C

Dado que A y B son subconjuntos del conjunto universal , esta relación se puede ver en la siguiente figura:



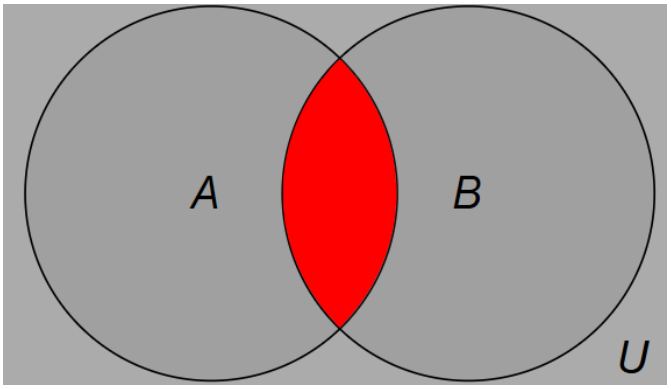
La unión de A y B, A ∪ B, está sombreada en azul. Su complemento, (A ∪ B) C está sombreado en amarillo. La intersección de los complementos de A y B, A C ∩ B C también está sombreada en amarillo.

Intersección de conjuntos:

El complemento de la intersección de dos conjuntos es igual a la unión de sus complementos:

A ∩ B=A C ∪ B C

Dado que A y B son subconjuntos del conjunto universal , esta relación se puede ver en la siguiente figura:



La intersección de A y B, A ∩ B, está sombreada en rojo. Su complemento, (A ∩ B) C está sombreado en gris. La unión de los complementos de A y B, A C ∪ B C , también está sombreada en gris.

2.2 Distribución de muestreo

El muestreo probabilístico es un método de muestreo (muestreo se refiere al estudio o el análisis de grupos pequeños de una población) que utiliza formas de métodos de selección aleatoria.

El requisito más importante del muestreo probabilístico es que todos en una población tengan la misma oportunidad de ser seleccionados.

Por ejemplo, si tienes una población de 100 personas, cada persona tendría una probabilidad de 1 de 100 de ser seleccionado. El método de muestreo probabilístico te ofrece la mejor oportunidad de crear una muestra representativa de la población.

Este método utiliza la teoría estadística para seleccionar al azar un pequeño grupo de personas (muestra) de una gran población existente y luego predecir que todas las respuestas juntas coincidirán con la población en general.

Por ejemplo, es prácticamente imposible enviar una encuesta a cada una de las personas de todo un país para recabar información, pero lo que puedes hacer utilizar el método de muestreo de probabilidad para obtener datos que pueden ser muy buenos (incluso aunque se obtengan de una población más pequeña).

Tipos de muestreo probabilístico

El muestreo aleatorio simple, tal y como su nombre lo indica, es un método completamente aleatorio que se utiliza para seleccionar una muestra. Este método de muestreo es tan fácil como asignar números a los individuos (muestra) y luego elegir de

manera aleatoria números entre los números a través de un proceso automatizado. Finalmente, los números que se eligen son los miembros que se incluyen la muestra.

Existen dos formas en que las muestras se eligen: A través de un sistema de lotería y uso de software de generación de números aleatorios. Esta técnica de muestreo funciona generalmente en grandes poblaciones y tiene tanto ventajas como desventajas.

Muestreo estratificado: este es un método en el cual una población grande se divide en dos grupos más pequeños, que generalmente no se superponen, sino que representan a toda la población en conjunto.

Durante el muestreo, estos grupos pueden organizarse y luego de estos se puede obtener una muestra de cada grupo por separado.

Algo común en este tipo de método es organizar o clasificar las muestras por sexo, edad, etnia, etc. Este método divide sujetos en grupos mutuamente exclusivos y luego utiliza un muestreo aleatorio simple para elegir miembros de los grupos.

Los miembros de cada uno de estos grupos deben ser distintos para que todos los miembros de todos los grupos tengan la misma oportunidad de ser seleccionados utilizando la probabilidad simple.

Muestreo por conglomerados: este es un método que selecciona de manera aleatoria a los participantes cuando están dispersos geográficamente.

Por ejemplo, tenemos a 1000 participantes de toda la población de México, supongamos que es probable que no sea posible obtener una lista completa de todos estos. Pero en

cambio, lo que hace el investigador es seleccionar áreas de manera aleatoria (es decir, ciudades, comunidades, etc), y selecciona al azar dentro de esos límites.

El muestreo por conglomerados por lo general analiza a una población particular en la que la muestra consiste en varios elementos, por ejemplo, ciudad, familia, universidad, etc. Los conglomerados se seleccionan básicamente dividiendo la población mayor en varias secciones más pequeñas.

Muestreo sistemático: este se enfoca en elegir a cada "enésima" persona para que sea parte de la muestra. Por ejemplo, puedes elegir que cada quinta persona sea parte de la muestra, o que cada décima persona sea parte de ella.

El muestreo sistemático es una implementación extendida de la mismísima técnica de probabilidad en la que cual, cada miembro de un grupo es seleccionado en periodos regulares para formar una muestra. Cuando se utiliza este método de muestreo, existe una oportunidad igual para que cada miembro de una población sea seleccionado.

¿Cuáles son los pasos para llevar a cabo un muestreo probabilístico?

1.- Elige cuidadosamente tu población de interés: piensa detenidamente y elige entre la población de manera correcta. Las personas que crees que tienen opiniones que deban recopilarse son las que tienes que incluir en tu muestra.

2.- Determina un marco de muestra adecuado: tu marco debe incluir una muestra de tu población de interés y nadie del exterior. Esto es importante si quieres recopilar datos precisos y que te sirvan.

3.-Selecciona tu muestra y comienza tu encuesta: a veces puede ser difícil encontrar la muestra correcta y determinar el marco de muestra adecuado. Incluso cuando todos los factores están a nuestro favor, muchas veces pueden haber problemas imprevistos como el factor de costo, la calidad de los encuestados y la rapidez de estos en responder.

Obtener una muestra para responder a una verdadera encuesta de probabilidad puede ser difícil, pero no imposible.

En la mayoría de los casos, utilizar la técnica de muestreo probabilístico te ahorrará tiempo, dinero y mucha frustración. Probablemente no puedas enviar encuestas a todas las personas, pero siempre puedes darles a todos la oportunidad de participar, de esto es de lo que se trata la técnica de muestreo de probabilidad.

Toma en cuenta estas consideraciones para tener el mejor muestreo.

¿Cuándo utilizar el muestreo probabilístico?

1.- Cuando se tiene que reducir el sesgo en el muestreo: este método de muestreo se utiliza comúnmente cuando el sesgo debe ser mínimo.

La selección de la muestra determina en gran medida la calidad de la investigación. Y la forma en la que los investigadores seleccionan su muestra determina la calidad de sus hallazgos.

El muestreo probabilístico proporciona en gran medida calidad en los hallazgos del investigador, esto sucede porque se trata de investigar a una representación imparcial de la población. Esto es de especial importancia para eliminar el sesgo en tus encuestas.

2.- Cuando la población es diversa: cuando el tamaño de la población es grande y diversa, este método de muestreo es útil ya que ayuda a los investigadores a crear muestras que representan completamente a la población.

Supongamos que queremos saber cuántas personas prefieren el turismo médico antes de recibir un tratamiento en su propio país, este método de muestreo puede ayudarle al

investigador a recoger muestras de diversos estratos socioeconómicos, antecedentes, etc., para representar a la población general.

Conoce más de la importancia de una muestra representativa para una investigación eficaz.

3.- Para crear una muestra precisa: el muestreo probabilístico ayuda a los investigadores a crear una muestra precisa de su población. Los investigadores pueden utilizar este método para crear un tamaño de muestra preciso que les pueda ayudar a obtener datos bien definidos.

Ventajas del muestreo probabilístico

1.- Es rentable: este proceso es rentable y efectivo en relación al tiempo y costo.

2.- Es simple y fácil: el muestreo de probabilidad es un método fácil ya que no implica un proceso complicado. Es rápido y ahorra tiempo.

3.- No es técnico: este método de muestreo no requiere ningún conocimiento técnico debido a la simplicidad con la que puede realizarse. Este método no requiere ningún tipo de conocimiento complejo y por suerte, no es nada largo.

Formula del Muestreo Probabilístico

Existe una gran cantidad de fórmulas para realizar un muestreo probabilístico, una de las más comunes por su sencillez es la del muestreo estatificado, sin embargo te recomendamos leer e investigar los diversos métodos de muestre que hemos mencionado anteriormente.

Ejemplo muestreo estratificado

Expresamos en una tabla todos los datos

	Hombres	Mujeres	Niños	TOTAL
Población	700	800	500	2000
Muestra	x	y	z	80

Expresamos la proporcionalidad:

$$\frac{700}{x} = \frac{800}{y} = \frac{500}{z} = \frac{2000}{80}$$

Para calcular cualquiera de las incógnitas, buscamos una proporción donde conozcamos 3 de los 4 datos:

$$\frac{700}{x} = \frac{2000}{80} \Rightarrow x = \frac{700 \cdot 80}{2000} = 28$$

$$\frac{800}{y} = \frac{2000}{80} \Rightarrow y = \frac{800 \cdot 80}{2000} = 32$$

$$\frac{500}{z} = \frac{2000}{80} \Rightarrow x = \frac{500 \cdot 80}{2000} = 20$$

2.3 Muestreo aleatorio simple

Las encuestas por muestreo consisten en extraer de una población finita de N unidades, subpoblaciones de un tamaño fijado de antemano. Si todas las unidades son indistinguibles, el número de muestras de tamaño n viene dado por:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = {}_{n}C_{n}$$

Por ejemplo, si la población contiene 5 unidades A, B, C, D, E; existen 10 muestras diferentes de tamaño 3, que son:

ABC, ABD, ABE, ACD, ACE

ADE, BCD, BCE, BDE. CDE

Debe notarse que la misma letra no ocurre dos veces en la misma muestra; y, también, que el orden de los elementos no tiene importancia, las seis muestras ABC, ACB, BAC, BCA, CAB, CBA son consideradas como iguales.

El muestreo aleatorio simple es un método de selección de n unidades sacadas de N, de tal manera que cada una de las muestras tiene la misma probabilidad de ser elegida.

En la práctica una muestra aleatoria simple es extraída de la siguiente forma:

Se numeran las unidades de la población del 1 al N, y por medio de una tabla de números aleatorios o colocando los números 1 a N en una urna, se extraen sucesivamente n números. Las unidades que llevan estos números constituyen la muestra.

El método elegido debe de verificar que en cualquier fase de la obtención de la muestra cada individuo que no ha sido sacado previamente, tiene la misma probabilidad de ser elegido[1].

Es fácil ver que cada una de las ${}_{N}C_{n}$ muestras tiene igual posibilidad de obtenerse.

Cuando un número ha sido sacado de la urna, éste no es reemplazado, ya que esto daría lugar a que la misma unidad entrara en la muestra más de una vez. Por esta razón el muestreo es descrito como sin reemplazo. El muestreo con reemplazo, es totalmente factible, aunque rara vez es usado, ya que no se ve la conveniencia de tener el mismo individuo dos veces en la misma muestra.

2.4 Muestreo aleatorio estratificado simple

El **muestreo aleatorio estratificado** es una técnica de muestreo que se utiliza cuando en la población se pueden distinguir subgrupos o subpoblaciones claramente identificables. Mediante este método de muestreo, la selección de los elementos que van a formar parte de la muestra se realiza por separado dentro de cada estrato, sin dejar ningún estrato sin muestrear. En la práctica esta técnica presenta dos ventajas importantes:

Puede facilitar la implementación física del muestreo (organización de la campaña de toma de datos, lugares a visitar, etc.)

Permite aplicar el esfuerzo de muestreo de forma "inteligente", tomando muestras de mayor tamaño en aquellos estratos que así lo requieran, y menos en donde no haga falta. Por poner un ejemplo extremo, si todos los sujetos de un estrato son clónicos, posiblemente bastaría con medir a uno de ellos para tener toda la información necesaria. Si los sujetos de un estrato son extremadamente heterogéneos, habrá que tomar una muestra grande para poder captar bien el efecto de esa variabilidad.

2.5 Muestreo por conglomerado

La población está dividida en áreas lo más heterogenias posibles internamente y lo más homogéneas posibles entre sí. Selecciona al azar un conglomerado que será el que formará la muestra.

Hay dos razones principales para la extensa aplicación del muestreo por conglomerado. En muchos países no hay listas completas ni al día de las personas, fincas, casas, etc en una región geográfica grande. Sin embargo, a partir de mapas de la región, la misma puede ser subdividida en segmentos de tierra con límites fácilmente identificables en las zonas rurales, o en unidades de superficie como manzanas en zonas urbanas. En EE.UU y Europa se toman a menudo estos conglomerados, porque resuelven el problema de construir una lista de unidades de muestreo.

Aun cuando se dispongan de listas consideraciones económicas pueden apuntar hacia la elección de una unidad conglomerada mayor. Para un tamaño de muestra dado una unidad pequeña usualmente da resultados más precisos que una unidad grande. Por ejemplo, una simple muestra al azar de 600 casas cubre una ciudad más uniformemente que 20 manzanas de 30 casas cada una. Pero obviamente se incurren en más gasto seleccionando 600 casas al azar y viajando por ellas que localizando 20 manzanas y la visita de todas las casas de las mismas. Cuando el costo es contrapesado con la precisión, la unidad mayor puede ser superior. En muchas decisiones prácticas el tipo de unidad puede tener alguna conveniencia o desventaja especial. Por ejemplo, elegir unidades pequeñas al muestrear una cosecha puede introducir un sesgo debido a la incertidumbre de los límites exactos de la unidad.

2.6 Intervalo de confianza para diferencia entre medias

El intervalo de confianza para la diferencia de medias es un intervalo que proporciona un valor máximo y un valor mínimo entre los cuales se encuentra el valor de la diferencia de las medias de dos poblaciones con un determinado nivel de confianza.

Por ejemplo, si el intervalo de confianza para la diferencia de las medias de dos poblaciones con un nivel de confianza del 95% es (3,5), significa que la diferencia entre las dos medias poblacionales estará entre 3 y 5 con una probabilidad del 95%.

Por lo tanto, en estadística el intervalo de confianza para la diferencia de medias se usa para estimar dos valores entre los cuales se encuentra la diferencia entre dos medias poblacionales. De manera que a partir de los datos de dos muestras, se puede aproximar cuál es la diferencia entre las medias de las poblaciones.

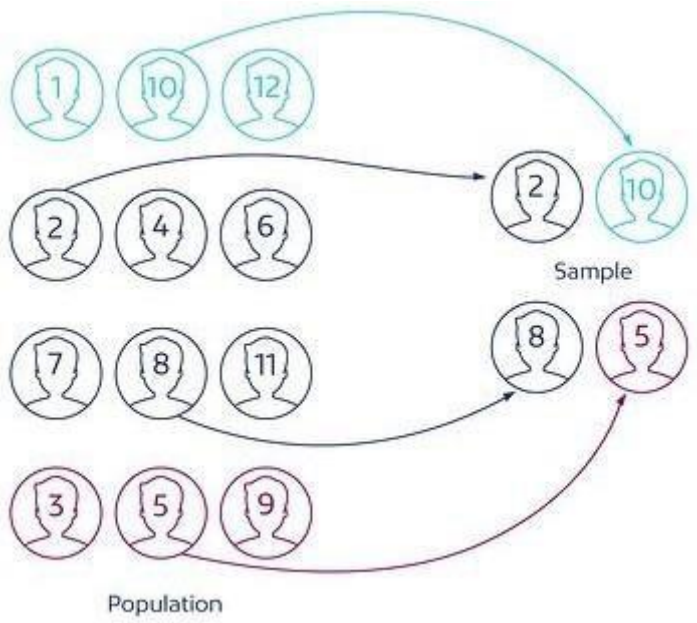
Fórmula del intervalo de confianza para la diferencia de medias

La fórmula del intervalo de confianza para la diferencia de medias depende de si se conocen las varianzas de las poblaciones y, en caso contrario, de si se puede suponer que las varianzas poblacionales son iguales o no. Así pues, a continuación, veremos cómo se calcula el intervalo de confianza para la diferencia de medias en cada caso.

2.7 Muestreo estratificado

Vimos en un post anterior la definición, ventajas e inconvenientes del muestreo aleatorio simple. Veamos ahora el muestreo estratificado.

Esta técnica, perteneciente a la familia de muestreos probabilísticos, consiste en dividir toda la población objeto de estudio en diferentes subgrupos o estratos disjuntos, de manera que un individuo sólo puede pertenecer a un estrato. Una vez definidos los estratos, para crear la muestra se seleccionan individuos empleando una técnica de muestreo cualquiera a cada uno de los estratos por separado. Si por ejemplo empleamos muestreo aleatorio simple en cada estrato, hablaremos de **muestreo aleatorio estratificado** (M.A.E. en adelante). Del mismo modo, podríamos usar otras técnicas de muestreo en cada estrato (muestreo sistemático, aleatorio con reposición, etc.).



Los estratos suelen ser grupos homogéneos de individuos, que a su vez son heterogéneos entre diferentes grupos. Por ejemplo, si en un estudio esperamos encontrar un comportamiento muy diferente entre hombres y mujeres, puede ser conveniente definir dos estratos, uno por cada sexo. Si la selección de estos estratos es correcta:

Los hombres deberían comportarse de forma parecida entre ellos.

Las mujeres deberían comportarse de forma muy similar entre ellas.

Hombres y mujeres deberían mostrar comportamientos dispares entre sí.

Si la anterior condición se cumple (estratos homogéneos internamente, heterogéneos entre sí) el uso del muestreo aleatorio estratificado reduce el error muestral, mejorando la precisión de nuestros resultados al realizar un estudio sobre la muestra.

Es relativamente habitual definir estratos de acuerdo a algunas variables características de la población como son la edad, sexo, clase social o región geográfica. Estas variables permiten

dividir fácilmente la muestra en grupos mutuamente excluyentes y con bastante frecuencia, permiten discriminar comportamientos diferentes dentro de la población.

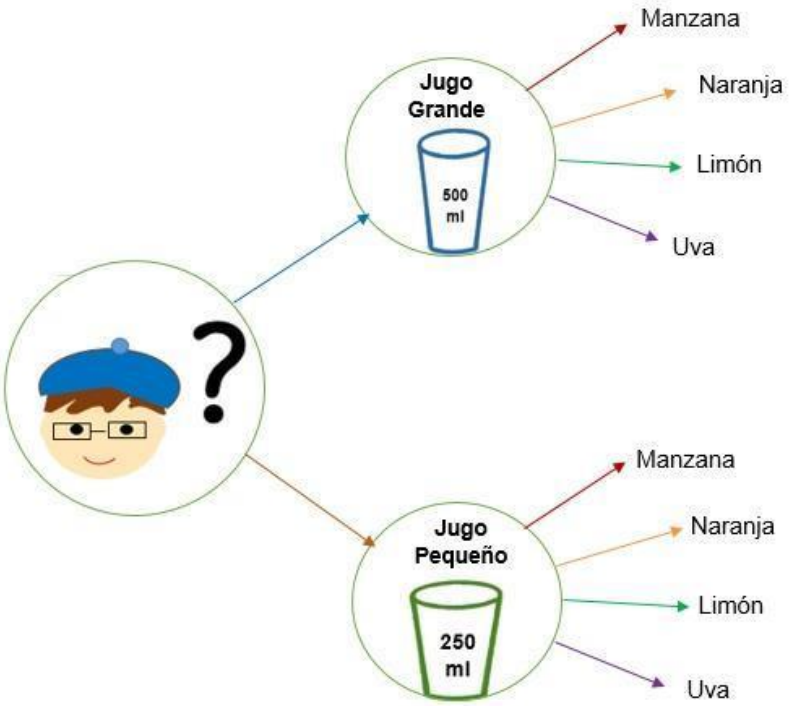
2.8 Principio aditivo, multiplicativo y arreglo rectangular

El **principio aditivo** es una técnica de conteo en probabilidad que permite medir de cuántas maneras se puede realizar una actividad que, a su vez, tiene varias alternativas para ser realizada, de las cuales se puede elegir solo una a la vez. Un ejemplo clásico de esto es cuando se quiere escoger una línea de transporte para ir de un lugar a otro.

En este ejemplo, las alternativas corresponderán a todas las líneas de transporte posibles que cubran el recorrido deseado, bien sea aéreas, marítimas o terrestres. No podemos ir a un lugar usando dos medios de transporte simultáneamente; es necesario que elijamos solo uno.

El principio aditivo nos dice que la cantidad de maneras que tenemos para realizar este viaje corresponderá a la suma de cada alternativa (medio de transporte) posible que exista para ir al lugar deseado, esto incluirá aun los medios de transporte que hagan escala en algún lugar (o lugares) intermedio.

Obviamente, en el ejemplo anterior siempre escogeremos la alternativa más cómoda y que más se ajuste a nuestras posibilidades, pero probabilísticamente es de suma importancia conocer de cuántas maneras se puede realizar un evento.



2.9 Diagrama de árbol, principio multiplicativo

Sabemos que para poder determinar la probabilidad de ocurrencia de un evento es necesario conocer el espacio muestral y específicamente su cardinalidad.

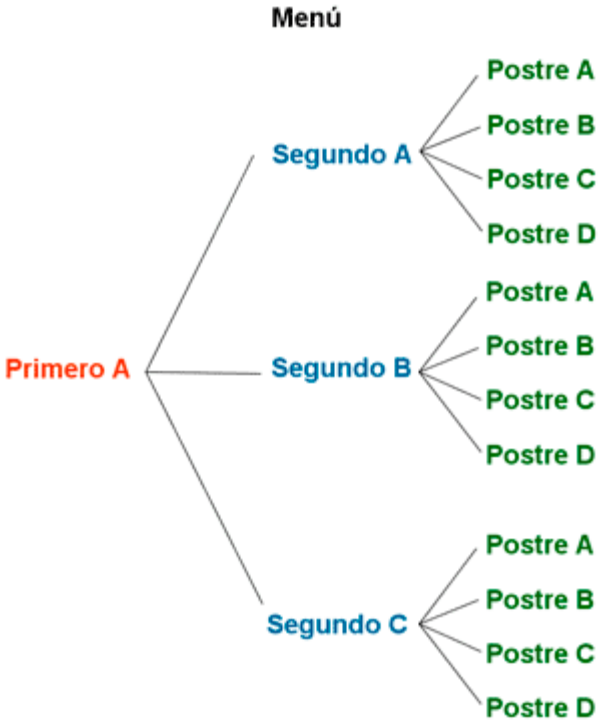
Una de las técnicas de conteo es el principio multiplicativo, el cual se usa para determinar la cardinalidad de un espacio muestral, es una forma de contar eficientemente.

Pensemos en el lanzamiento de dos dados:

En el primer lanzamiento tenemos 6 posibilidades, y en el segundo tenemos las mismas 6 posibilidades. Entonces tendremos en total $6 \cdot 6 = 36$ posibilidades.

Ahora bien, si tenemos 3 pantalones y 4 camisetas ¿Cuántas tenidas distintas podemos formar? $3 \cdot 4 = 12$ Posibles tenidas.

Es decir si tenemos un experimento el cual puede ocurrir de «n» maneras diferentes, un segundo experimento que tiene «m» maneras diferentes y el experimento es uno seguido del otro entonces tenemos $n \cdot m$ posibilidades de que este pueda ocurrir. Este principio puede generalizarse a cualquier número de experimentos.



Veremos el caso del lanzamiento de los dados: Supongamos que en el primer lanzamiento se obtiene 1, en el segundo lanzamiento podemos obtener 1, 2, 3, 4, 5, 6 lo representamos y obtenemos:

Segundo lanzamiento:

Tenemos por lo tanto 36 posibilidades.

2.10 Permutaciones

En matemáticas, una permutación de un conjunto es, en términos generales, una disposición de sus miembros en una secuencia u orden lineal, o si el conjunto ya está ordenado, una variación del orden o posición de los elementos de un conjunto ordenado o

una tupla. La palabra "permutación" también se refiere al acto o proceso de cambiar el orden lineal de un conjunto ordenado.1

Las permutaciones difieren de las combinaciones, que son selecciones de algunos miembros de un conjunto sin importar el orden. Por ejemplo, escritas como tuplas, hay seis permutaciones del conjunto {1, 2, 3}, a saber (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2) y (3, 2, 1). Estas son todas las ordenaciones posibles de este conjunto de tres elementos. Los anagramas de palabras cuyas letras son diferentes también son permutaciones: las letras ya están ordenadas en la palabra original, y el anagrama es una reordenación de las letras. El estudio de las permutaciones de conjuntos finitos es un tema importante en los campos de la combinatoria y la teoría de grupos.

Las permutaciones se utilizan en casi todas las ramas de las matemáticas y en muchos otros campos de la ciencia. En informática, se utilizan para analizar algoritmos de ordenación; en física cuántica, para describir estados de partículas; y en biología, para describir secuencias de ARN.

El número de permutaciones de n objetos distintos es n factorial, normalmente escrito como n!, que significa el producto de todos los enteros positivos menores o iguales a n.

Técnicamente, una permutación de un set S se define como una biyección de S a sí mismo. 23 Es decir, es una función de S a S para la cual cada elemento ocurre exactamente una vez como un valor de imagen. Esto está relacionado con el reordenamiento de los elementos de S en el que cada elemento s es reemplazado por el correspondiente f(s). Por ejemplo, la permutación (3, 1, 2) mencionada anteriormente es descrita por la función.

El conjunto de todas las permutaciones de un conjunto forman un grupo llamado grupo simétrico del conjunto. La operación de grupo es la composición (realizar dos reordenamientos dados sucesivamente), que da como resultado otro reordenamiento.

2.11 Combinaciones

Se llama **combinaciones** de m elementos tomados de n en $\mathrm{n}\,(\mathrm{m}\geq n)$ a todas las agrupaciones posibles que pueden hacerse con los m elementos de forma que:

No entran todos los elementos

No importa el orden

No se repiten los elementos

$$\mathrm{C}_m^n = \frac{V_m^n}{P_n}$$

También podemos calcular las combinaciones mediante **factoriales**:

$$\mathrm{C}_m^n = \frac{m!}{n!\,(m-n)!}$$

Las combinaciones se denotan por C_m^n o $\mathrm{C}_{m,n}$

Ejemplos de ejercicios de combinaciones

1 Calcular el número de combinaciones de 10 elementos tomados de 4 en 4

$$\mathrm{C}_{10}^4 = \frac{10\cdot 9\cdot 8\cdot 7}{4\cdot 3\cdot 2\cdot 1} = 210$$

O utilizando factoriales:

$$\mathrm{C}_{10}^4 = \frac{10!}{4!\,(10-4)!} = \frac{10\cdot 9\cdot \not{8}\cdot 7\cdot \not{6}!}{\not{4}\cdot 3\cdot \not{2}\cdot 1\cdot \not{6}!} = 10\cdot 3\cdot 7 = 210$$

2 En una clase de 35 alumnos se quiere elegir un comité formado por tres alumnos. ¿Cuántos comités diferentes se pueden formar?

No entran todos los elementos

No importa el orden: Juan, Ana, etc.

No se repiten los elementos

$$C_{35}^{3} = \frac{35 \cdot 34 \cdot 33}{3 \cdot 2 \cdot 1} = 6545$$

UNIDAD III

ASOCIACIÓN ESTADÍSTICA ENTRE VARIABLES

3.1 Concepto de asociación entre variables

El análisis estadístico de la asociación (relación, covarianza, correlación) entre variables representa una parte básica del análisis de datos en cuanto que muchas de las preguntas e

hipótesis que se plantean en los estudios que se llevan a cabo en la práctica implican analizar la existencia de relación entre variables.

La existencia de algún tipo de asociación entre dos o más variables representa la presencia de algún tipo de tendencia o patrón de emparejamiento entre los distintos valores de esas variables. A modo de ejemplo esquemático, si tenemos una variable X [a, b, c] y otra variable Y [m, n, p], de modo que los datos empíricos evidencian que las entidades que en X son a en Y tienden a ser n (o viceversa), que las que son b tienden a ser p, y que las que son c tienden a ser m, se pone de manifiesto que existe cierta asociación entre ambas variables.

Más formal que ésta, Solanas et al. (2005) ofrecen otra propuesta de definición general de lo que significa la asociación entre 2 variables: la existencia de asociación entre dos variables indicaría que la distribución de los valores de una de las dos variables difiere en función de los valores de la otra.

3.2 Midiendo la asociación entre dos variables

El caso de dos variables categóricas

¿Qué se puede decir acerca de la asociación entre las dos variables de la tabla de contingencia ("Estado de ánimo" y "Vivir en residencia")?

Para evaluar si ambas variables están relacionadas hay que observar si la distribución de los valores de una de las variables difiere en función de los valores de la otra, esto es, hay que comparar las distribuciones condicionadas de una de las dos variables agrupada en función de los valores de la otra. Si no hay relación entre las variables estas distribuciones deberían ser iguales. Por ejemplo, podemos comparar las distribuciones de frecuencias absolutas de "Estado de ánimo" condicionadas a vivir en una residencia (48, 42, 60) y a no vivir en una residencia (70, 105, 175).

Si nos fijamos en las distribuciones de frecuencias absolutas de "Estado de ánimo" condicionadas a los valores de "Vivir en residencia", se observa que estas distribuciones no son iguales, sin embargo, esto puede ser debido a que hay más sujetos que no viven en una residencia (350) que sujetos que sí viven en ella (150). En conclusión, no se deben comparar las distribuciones condicionadas en frecuencias absolutas si el número de casos difiere en las

categorías de la variable condicionante o agrupadora.

La asociación entre dos variables categóricas aparece más explícita en una tabla de frecuencias relativas condicionadas, pues de ese modo se relativiza el posible diferente tamaño de los subgrupos definidos por cualquiera de las dos variables. Este tipo de tabla se puede obtener de 2 formas alternativas, bien dividiendo las celdas de cada fila entre el respectivo marginal (total) de fila, bien cada columna entre el total de columna. Ambas tablas permitirán llegar al mismo tipo de conclusiones respecto a la asociación entre las 2 variables.

Si la relación entre las variables es asimétrica, la variable agrupadora o condicionante sería la que sea considerada la variable explicativa (predictora, independiente). Por ejemplo, en un estudio en que se evalúa la influencia del "Nivel de estudios" [primarios, secundarios, superiores] sobre la "Percepción de la influencia de la ciencia en la sociedad" [negativa, indiferente, positiva], dado que el nivel de estudios sería la variable explicativa, deberíamos comparar las distribuciones de la percepción de la influencia de la ciencia condicionadas al nivel de estudios, es decir, en cada categoría de nivel de estudios. En nuestro ejemplo sobre "Estado de ánimo" y "Vivir en residencia", dado que la relación es asimétrica y la variable explicativa es "Vivir en residencia" debemos comparar las distribuciones de "Estado de ánimo" condicionadas a "Vivir en residencia"

3.3 El caso de dos variables categóricas

¿Qué se puede decir acerca de la asociación entre las dos variables de la tabla de contingencia ("Estado de ánimo" y "Vivir en residencia")?

	−	±	+	Total
Sí	48	42	60	150
No	70	105	175	350
Total	118	147	235	500

Para evaluar si ambas variables están relacionadas hay que observar si la distribución de los

valores de una de las variables difiere en función de los valores de la otra, esto es, hay que comparar las distribuciones condicionadas de una de las dos variables agrupada en función de los valores de la otra. Si no hay relación entre las variables estas distribuciones deberían ser iguales. Por ejemplo, podemos comparar las distribuciones de frecuencias absolutas de "Estado de ánimo" condicionadas a vivir en una residencia (48, 42, 60) y a no vivir en una residencia (70, 105, 175).

Si nos fijamos en las distribuciones de frecuencias absolutas de "Estado de ánimo" condicionadas a los valores de "Vivir en residencia", se observa que estas distribuciones no son iguales, sin embargo, esto puede ser debido a que hay más sujetos que no viven en una residencia (350) que sujetos que sí viven en ella (150). En conclusión, no se deben comparar las distribuciones condicionadas en frecuencias absolutas si el número de casos difiere en las categorías de la variable condicionante o agrupadora.

La asociación entre dos variables categóricas aparece más explícita en una tabla de frecuencias relativas condicionadas, pues de ese modo se relativiza el posible diferente tamaño de los subgrupos definidos por cualquiera de las dos variables. Este tipo de tabla se puede obtener de 2 formas alternativas, bien dividiendo las celdas de cada fila entre el respectivo marginal (total) de fila, bien cada columna entre el total de columna. Ambas tablas permitirán llegar al mismo tipo de conclusiones respecto a la asociación entre las 2 variables.

Si la relación entre las variables es asimétrica, la variable agrupadora o condicionante sería la que sea considerada la variable explicativa (predictora, independiente). Por ejemplo, en un estudio en que se evalúa la influencia del "Nivel de estudios" [primarios, secundarios, superiores] sobre la "Percepción de la influencia de la ciencia en la sociedad" [negativa, indiferente, positiva], dado que el nivel de estudios sería la variable explicativa, deberíamos comparar las distribuciones de la percepción de la influencia de la ciencia condicionadas al nivel de estudios, es decir, en cada categoría de nivel de estudios. En nuestro ejemplo sobre "Estado de ánimo" y "Vivir en residencia", dado que la relación es asimétrica y la variable explicativa es "Vivir en residencia" debemos comparar las distribuciones de "Estado de ánimo" condicionadas a "Vivir en residencia":

	−	±	+	Total
Sí	0,32 (48/150)	0,28 (42/150)	0,40 (60/150)	1
No	0,20 (70/350)	0,30 (105/350)	0,50 (175/350)	1
Total	0,236 (118/500)	0,294 (147/500)	0,470 (235/500)	1

En la tabla anterior, la comparación de las distribuciones de frecuencias relativas condicionales con la distribución marginal de la variable de respuesta, nos permitirá comprobar la existencia de asociación entre las dos variables y, en el caso de que exista, la naturaleza de la misma.

A modo de ejemplo, si no hubiera relación entre ambas variables, las distribuciones de frecuencias relativas de "Estado de ánimo" condicionadas a "Vivir en residencia" serían iguales a la distribución marginal de la variable "Estado de ánimo", esto es:

	−	±	+	Total
Sí	0,236	0,294	0,470	1
No	0,236	0,294	0,470	1
Total	0,236	0,294	0,470	1

Como se puede comprobar, las distribuciones de frecuencias relativas de "Estado de ánimo" condicionadas a "Vivir en residencia" difieren bastante de las de la tabla anterior. Así, por ejemplo, se observa que la proporción de sujetos que tienen un estado de ánimo negativo entre los que viven en una residencia (0,32) es superior a la que cabría esperar si no hubiera relación entre ambas variables (0,236). Esto parece indicar que sí existe una relación entre ambas variables. • Si la relación entre las variables es simétrica es indiferente qué variable se elige como agrupadora o condicionante. Así, por ejemplo, si deseamos valorar si hay relación entre el lugar de residencia (rural o urbano) y la rama de bachiller cursada (ciencias, sociales, salud o humanidades) y no consideramos a priori que una de las variables sea la variable explicativa, podríamos comparar, o bien, las distribuciones de frecuencias relativas de "Lugar de residencia" condicionadas a "Bachiller", o bien, las distribuciones de frecuencias relativas de "Bachiller" condicionadas a "Lugar de residencia".

3.4 El caso de una variable categórica y una cuantitativa

De nuevo, el análisis de este tipo de asociación supone comparar las distribuciones

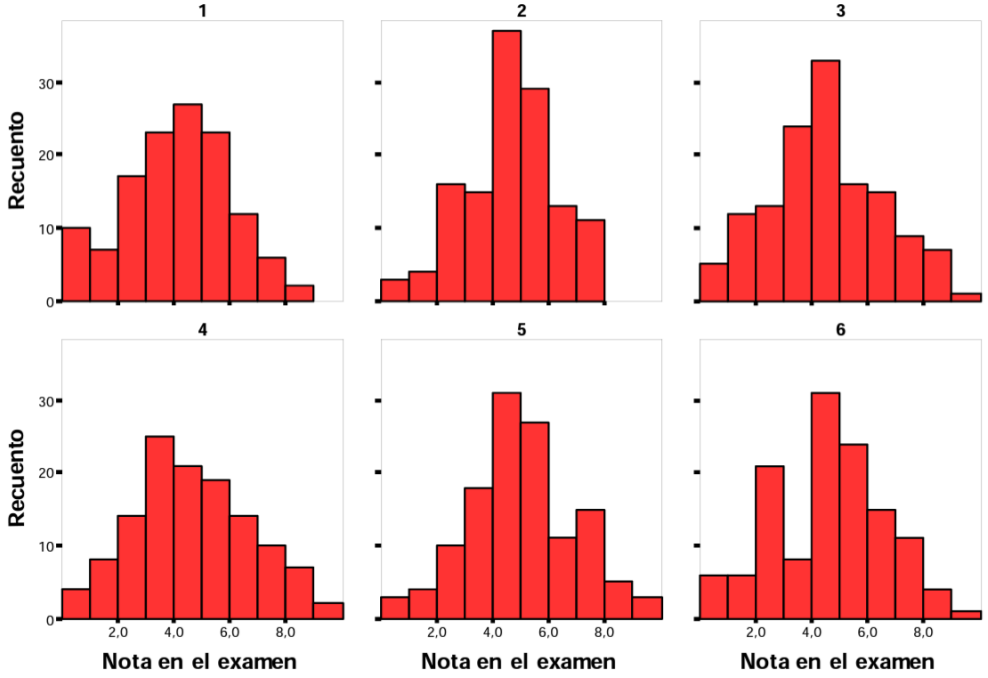
condicionales de una variable para los distintos valores que toma la otra. Normalmente, se suele tomar como condicionada a la cuantitativa y como condicionante a la categórica, si bien, las conclusiones a las que llegaríamos serían las mismas si se hiciese al revés. Si no hay diferencias entre las distribuciones condicionales, ello indicará que no hay asociación entre ambas variables.

Ejemplo del caso en que se quiera analizar la asociación entre las variables "Nota en un examen de una asignatura [0 a 10]" y "Grupo en el que se está matriculado [1 a 6]", disponiéndose de los datos de un total de 768 estudiantes de 6 grupos:

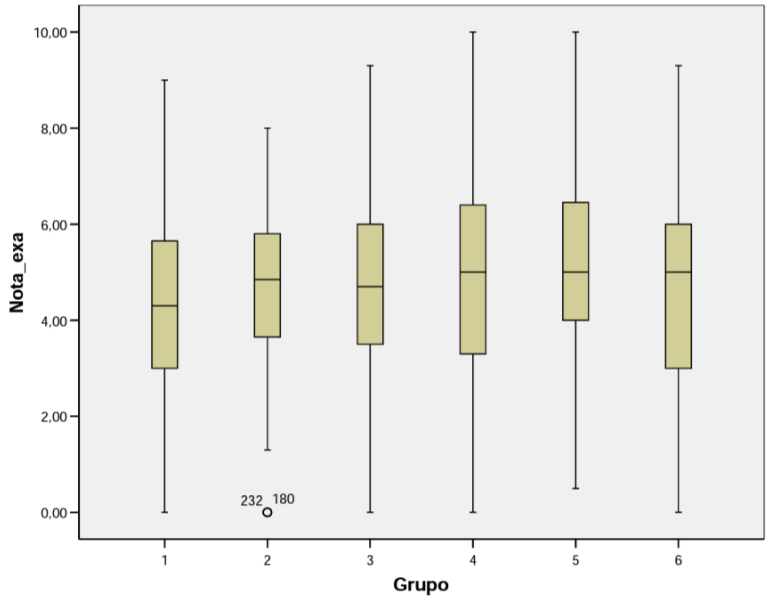
Grupo 1		-----	Grupo 2		-----	Grupo 3		-----	Grupo 4		-----	Grupo 5		-----	Grupo 6	
x_i	n_i		x_i	n_i		x_i	n_i		x_i	n_i		x_i	n_i		x_i	n_i
0	2		0	3		0	2		...	...		...	...		...	...
,3	3		1,3	1		,5	1									
,5	1		1,8	1		,8	2									
,8	3		2,0	2		1,3	2									
1,0	1		2,3	4		1,5	2									
1,3	1		2,5	2		1,8	1									
1,5	2		2,8	5		2,0	7									
1,8	2		2,9	1		2,3	3									
2,0	2		3,0	4		2,5	2									
2,3	3		3,3	3		2,6	3									
2,5	2		3,5	6		2,8	2									
2,6	1		3,8	3		2,9	1									
2,8	6		3,9	1		3,0	2									
3,0	5		4,0	2		3,3	3									
3,3	3		4,3	5		3,5	7									
3,5	5		4,5	6		3,8	9									
3,8	7		4,6	1		3,9	1									
4,0	8		4,7	8		4,0	4									
4,3	7		4,8	6		4,1	1									
4,5	5		4,9	6		4,3	3									
4,7	4		5,0	5		4,5	9									
4,8	3		5,1	1		4,7	6									
4,9	5		5,3	5		4,8	7									
5,0	3		5,4	1		4,9	4									
5,1	1		5,5	6		5,0	3									
5,3	6		5,6	1		5,3	4									
5,5	4		5,8	9		5,5	4									
5,8	7		5,9	1		5,6	1									
6,0	5		6,0	5		5,8	1									
6,1	1		6,3	2		5,9	2									
6,3	5		6,5	4		6,0	4									
6,5	3		6,8	4		6,3	3									
6,8	2		6,9	1		6,5	6									
7,0	1		7,0	2		6,8	2									
7,5	3		7,3	4		7,0	4									
7,6	1		7,5	3		7,1	1									
8,0	2		7,8	2		7,3	2									
9,0	2		8,0	2		7,5	5									
						8,0	1									
						8,3	2									
						8,4	1									
						9,0	4									
						9,3	1									

Dada la dificultad que puede representar comparar las distribuciones condicionales de una variable cuantitativa, se puede recurrir a representaciones gráficas que faciliten la realización

de este tipo de comparación. A modo de ejemplo, las dos siguientes obtenidas para los datos anteriores con el paquete estadístico SPSS o, también, el diagrama de dispersión presentado en el capítulo anterior para estos mismos datos:

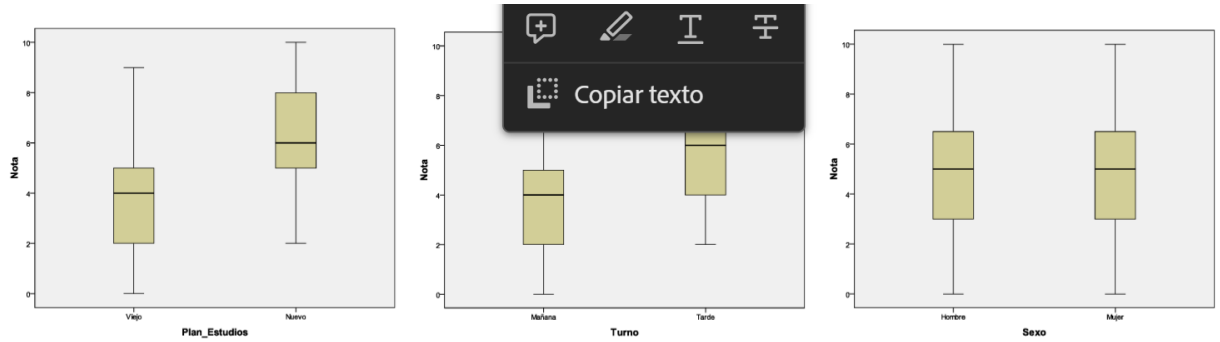


Ejemplo de diagrama de caja y bigotes con la distribución de la variable "Nota en un examen de una asignatura" condicionada a la variable "Grupo en el que se está matriculado":



Estos gráficos nos permiten comparar el grado de solapamiento (coincidencia) de las distribuciones condicionales. En general, cuanto mayor sea el solapamiento, menor será la intensidad de la asociación entre las dos variables y, viceversa, cuanto menor sea el solapamiento, mayor será el tamaño del efecto de la relación. En el ejemplo anterior existe bastante solapamiento entre las 6 distribuciones condicionales, poniendo de manifiesto una escasa relación entre ambas variables.

Ejemplo de diferente intensidad de asociación en 3 pares de variables (cada par constituido por una variable categórica dicotómica y una misma variable cuantitativa) basado en la visualización de las distribuciones condicionales mediante diagramas de caja y bigotes. Obsérvese como la relación entre las variables aumenta desde el gráfico de la izquierda (mayor solapamiento) hasta el de la derecha (menor solapamiento).



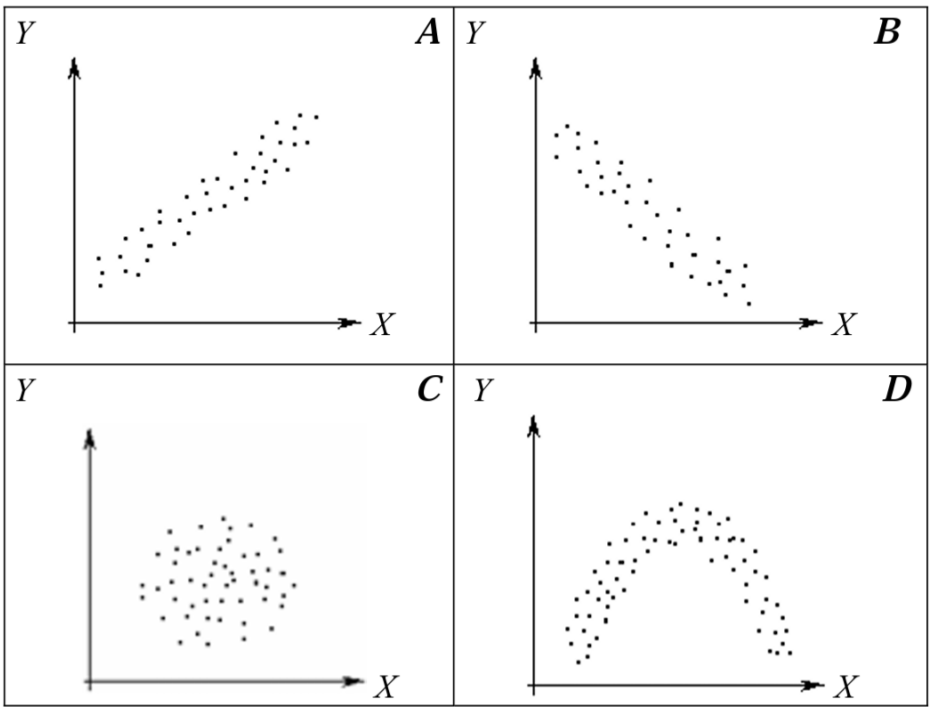
Otro gráfico que se suele utilizar para comparar el grado de solapamiento de las distribuciones condicionadas, y que ya se ha visto en temas anteriores, es el polígono de frecuencias superpuesto.

3.5 El caso de dos variables cuantitativas

Al igual que en los casos anteriores, la existencia de correlación o asociación entre 2 variables cuantitativas viene determinada por la presencia de diferencias en las distribuciones condicionales de una variable para los distintos valores de la otra.

Sin embargo, dado el número tan amplio de distribuciones condicionales que se pueden llegar a obtener en este caso, es más habitual analizar la asociación directamente sobre un diagrama

de dispersión, observando la disposición de la nube de puntos que representa la distribución conjunta de ambas variables. Así, ¿qué podríamos decir acerca de la asociación entre los 4 pares de variables cuyos diagramas de dispersión se muestran a continuación?



Un aspecto relevante del análisis de la correlación entre dos variables cuantitativas es que la presencia de ésta se puede plantear de acuerdo a diferentes modelos o patrones de asociación, por ejemplo, en forma de línea recta, tal como en los ejemplos A (relación lineal directa o positiva) y B (relación lineal inversa o negativa) de arriba, o en forma curvilínea tal como en D (relación parabólica o cuadrática). Así, la forma de evaluar la intensidad de la correlación suele consistir en analizar el ajuste de la nube de puntos al modelo de asociación que se considere que representa más adecuadamente a la distribución conjunta de ambas variables.

En la cuantificación de la asociación entre 2 variables cuantitativas nos vamos a ceñir al supuesto de que un modelo de relación lineal subyace a la asociación entre ambas. Subrayar que con frecuencia se obvia en los textos estadísticos que la relación que se analiza es en realidad una relación de tipo lineal. Los índices más utilizados en la práctica estadística a la hora de analizar la intensidad o tamaño del efecto de la relación lineal entre dos variables son

los tres siguientes:

$$S_{XY} = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n}$$

Al numerador de esta expresión se le conoce en la literatura estadística como suma de productos cruzados (SPXY), por lo que la anterior expresión queda como: sXY = SPXY / n –

Desarrollando algebraicamente la fórmula de la covarianza se puede llegar a una fórmula que se considera más conveniente cuando el cálculo de la misma se ha de realizar de forma manual:

$$S_{XY} = \frac{\sum X_i Y_i}{n} - \bar{X}\bar{Y}$$

Ejemplo para las variables Calificaciones en música (X) y Calificaciones en matemáticas (Y) obtenidas por un grupo de 10 niños.

X	Y	X^*Y
5	6	30
7	8	56
8	7	56
5	6	30
9	10	90
4	5	20
5	5	25
5	7	35
7	6	42
8	9	72
$\bar{X} = 6{,}3$	$\bar{Y} = 6{,}9$	$\Sigma\,(X^*Y) = 456$

$$S_{XY} = \frac{\sum X_i Y_i}{n} - \bar{X}\bar{Y} = \frac{456}{10} - (6{,}3 \cdot 6{,}9) = 2{,}13$$

La covarianza puede tomar valores tanto positivos como negativos. A nivel interpretativo, un mayor valor de la covarianza en valor absoluto indicará una relación lineal más intensa entre las dos variables. Un valor positivo pone de manifiesto una relación lineal directa; uno negativo, una relación lineal inversa; y si igual o muy próximo a 0, la inexistencia de relación

lineal entre las dos variables.

3.6 El modelo de regresión lineal

La regresión lineal es una técnica de análisis de datos que predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido. Modela matemáticamente la variable desconocida o dependiente y la variable conocida o independiente como una ecuación lineal. Por ejemplo, supongamos que tiene datos sobre sus gastos e ingresos del año pasado. Las técnicas de regresión lineal analizan estos datos y determinan que tus gastos son la mitad de tus ingresos. Luego calculan un gasto futuro desconocido al reducir a la mitad un ingreso conocido futuro.

Los modelos de regresión lineal son relativamente simples y proporcionan una fórmula matemática fácil de interpretar para generar predicciones. La regresión lineal es una técnica estadística establecida y se aplica fácilmente al software y a la computación. Las empresas lo utilizan para convertir datos sin procesar de manera confiable y predecible en inteligencia empresarial y conocimiento práctico. Los científicos de muchos campos, incluidas la biología y las ciencias del comportamiento, ambientales y sociales, utilizan la regresión lineal para realizar análisis de datos preliminares y predecir tendencias futuras. Muchos métodos de ciencia de datos, como el machine learning y la inteligencia artificial, utilizan la regresión lineal para resolver problemas complejos.

En esencia, una técnica de regresión lineal simple intenta trazar un gráfico lineal entre dos variables de datos, x e y. Como variable independiente, x se traza a lo largo del eje horizontal. Las variables independientes también se denominan variables explicativas o variables predictivas. La variable dependiente, y, se traza en el eje vertical. También puede hacer referencia a los valores y como variables de respuesta o variables pronosticadas.

Pasos en la regresión lineal

Para esta visión general, tenga en cuenta la forma más simple de la ecuación de gráfico de líneas entre y y x; y=c*x+m, donde c y m son constantes para todos los valores posibles de x e y. Así, por ejemplo, supongamos que los datos de entrada para (x, y) era (1,5), (2,8) y (3,11).

Para identificar el método de regresión lineal, debe seguir los siguientes pasos:

1. Trace una línea recta y mida la correlación entre 1 y 5.
2. Siga cambiando la dirección de la línea recta para los nuevos valores (2,8) y (3,11) hasta que se ajusten todos los valores.
3. Identifique la ecuación de regresión lineal como y = 3*x + 2.
4. Extrapola o predice que y es 14 cuando x es

3.7 Conceptos básicos sobre el análisis de regresión lineal

En el machine learning, los programas de computación denominados algoritmos analizan grandes conjuntos de datos y trabajan hacia atrás a partir de esos datos para calcular la ecuación de regresión lineal. Los científicos de datos primero entrenan el algoritmo en conjuntos de datos conocidos o etiquetados y, a continuación, utilizan el algoritmo para predecir valores desconocidos. Los datos de la vida real son más complicados que el ejemplo anterior. Es por eso que el análisis de regresión lineal debe modificar o transformar matemáticamente los valores de los datos para cumplir con los siguientes cuatro supuestos.

Relación lineal

Debe existir una relación lineal entre las variables independientes y las dependientes. Para determinar esta relación, los científicos de datos crean una gráfica de dispersión (una colección aleatoria de valores x e y) para ver si caen a lo largo de una línea recta. De lo contrario, puede aplicar funciones no lineales, como la raíz cuadrada o el registro, para crear matemáticamente la relación lineal entre las dos variables.

Independencia residual

Los científicos de datos utilizan residuos para medir la precisión de la predicción. Un residuo es la diferencia entre los datos observados y el valor previsto. Los residuos no deben tener un patrón identificable entre ellos. Por ejemplo, no querrá que los residuos crezcan con el tiempo. Puede utilizar diferentes pruebas matemáticas, como la prueba de Durbin-Watson, para determinar la independencia residual. Puede usar datos ficticios para reemplazar

cualquier variación de datos, como los datos estacionales.

Normalidad

Las técnicas de representación gráfica, como las gráficas Q-Q, determinan si los residuos se distribuyen normalmente. Los residuos deben caer a lo largo de una línea diagonal en el centro de la gráfica. Si los residuos no están normalizados, puede probar los datos para detectar valores atípicos aleatorios o valores que no sean típicos. Eliminar los valores atípicos o realizar transformaciones no lineales puede solucionar el problema.

Homocedasticidad

La homocedasticidad supone que los residuos tienen una variación constante o desviación estándar de la media para cada valor de x. De lo contrario, es posible que los resultados del análisis no sean precisos. Si no se cumple esta suposición, es posible que tenga que cambiar la variable dependiente. Dado que la variación se produce de forma natural en grandes conjuntos de datos, tiene sentido cambiar la escala de la variable dependiente. Por ejemplo, en lugar de usar el tamaño de la población para predecir la cantidad de estaciones de bomberos en una ciudad, podría usar el tamaño de la población para predecir la cantidad de estaciones de bomberos por persona.

Algunos tipos de análisis de regresión son más adecuados que otros para gestionar conjuntos de datos complejos. A continuación se muestran algunos ejemplos.

Regresión lineal simple

La regresión lineal simple se define mediante la función lineal:

$Y= \beta 0*X + \beta 1 + \varepsilon$

β0 y β1 son dos constantes desconocidas que representan la pendiente de regresión, mientras que ε (épsilon) es el término de error.

Puede utilizar la regresión lineal simple para modelar la relación entre dos variables, como las siguientes:

Lluvia y rendimiento de los cultivos

Edad y estatura en niños

Regresión lineal múltiple

En el análisis de regresión lineal múltiple, el conjunto de datos contiene una variable dependiente y múltiples variables independientes. La función de línea de regresión lineal cambia para incluir más factores, de la siguiente manera:

$Y = \beta 0*x0 + \beta 1x1 + \beta 2x2+\ldots\ldots \beta NxN+ \varepsilon$

A medida que aumenta el número de variables predictivas, las constantes β también aumentan en consecuencia.

La regresión lineal múltiple modela múltiples variables y su impacto en un resultado:

Lluvia, temperatura y uso de fertilizantes en el rendimiento de los cultivos

Dieta y ejercicio sobre enfermedades cardíacas

Crecimiento salarial e inflación en las tasas de préstamos hipotecarios

Regresión logística

Los científicos de datos utilizan la regresión logística para medir la probabilidad de que se produzca un evento. La predicción es un valor entre 0 y 1, donde 0 indica un evento que es poco probable que ocurra y 1 indica una probabilidad máxima de que suceda. Las ecuaciones logísticas usan funciones logarítmicas para calcular la línea de regresión.

A continuación, se indican varios ejemplos:

La probabilidad de ganar o perder en un partido deportivo

La probabilidad de aprobar o reprobar una prueba

La probabilidad de que una imagen sea una fruta o un animal

3.8 Ajuste de la recta de regresión

En las distribuciones bidimensionales que siguen una dependencia estadística se utilizan gráficas de puntos para representar sus tendencias. No obstante, dichas tendencias pueden apuntar a una ley de tipo funcional, que pueda explicar el comportamiento global de la distribución. Para hallar esta ley se utilizan métodos de regresión y correlación entre las variables.

Con frecuencia, las variables que constituyen una **distribución bidimensional** (ver t61) muestran un cierto grado de dependencia entre ellas. Un ejemplo típico de esta relación aparece en las tablas de peso y altura de los grupos de población: aunque no existe una ley causal que relacione ambas variables, en términos estadísticos se aprecia una dependencia entre ellas (cuando aumenta la altura, suele hacerlo también el peso). Esta dependencia se refleja en la nube de puntos que representa a la distribución, de modo que los puntos de esta gráfica aparecen condensados en algunas zonas.

La concentración de puntos en algunas regiones de la nube refleja la existencia de una dependencia estadística, y la posibilidad de definir una ecuación de regresión.
En tales casos, se pretende definir una ecuación de regresión que sirva para relacionar las dos variables de la distribución. La representación gráfica de esta ecuación recibe el nombre de línea de regresión, y puede adoptar diversas formas: lineal, parabólica, cúbica, hiperbólica, exponencial, etcétera.

Cuando la línea de regresión se asemeja a una recta (regresión lineal), puede ajustarse a esta forma geométrica por medio de un método general conocido como método de los mínimos cuadrados. La recta de ajuste tendrá por ecuación y = ax + b, donde los coeficientes a y b se calculan teniendo en cuenta que:
La recta debe pasar por el punto $(\bar{x}, \bar{y})$.

La separación de los puntos de la gráfica de dispersión con respecto a la recta de regresión debe ser mínima.

Estas dos condiciones conducen a una recta de ajuste expresada por la ecuación:

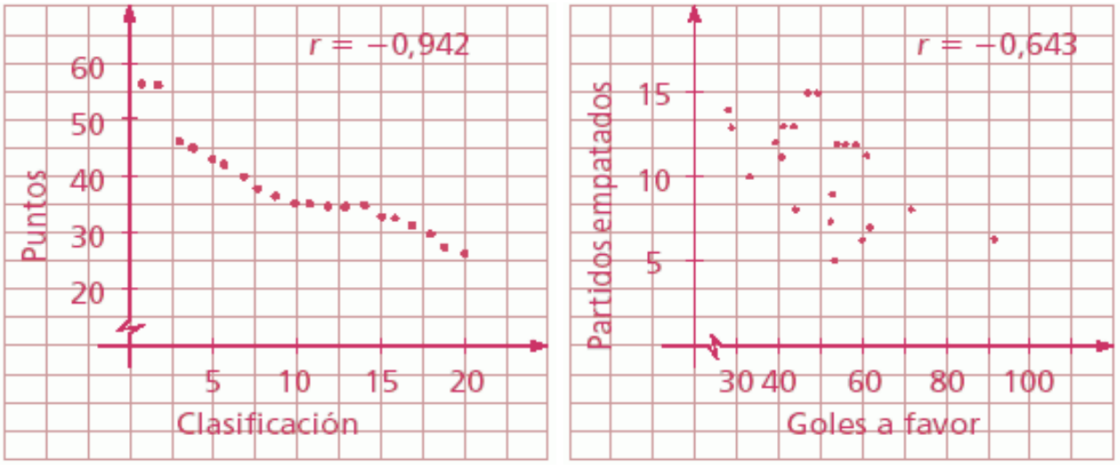
$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

donde $\bar{x}$ es la **media aritmética** de la primera variable, $\bar{y}$ la media aritmética de la segunda variable, **sx la desviación típica** de la primera variable y **sxy** un valor denominado **covarianza**, que se define por la expresión:

$$\sigma_{xy} = \frac{\sum x_i y_i}{n} - \bar{x} \cdot \bar{y}$$

En una distribución bidimensional, se define correlación, denotada por r, como el grado de dependencia que existe entre las dos variables del modelo, de modo que:
Cuando al aumentar el valor de una variable crece también el de la otra, la correlación es directa, e inversa en caso contrario.

Si no existe dependencia entre las variables, la correlación es nula.
Para conocer si una correlación es directa o inversa, basta con determinar su covarianza:
Si la covarianza es positiva, la correlación es directa.
Cuando la covarianza es negativa, existe una correlación inversa entre las variables.



3.9 Bondad de ajuste del modelo de regresión

En estadística, la bondad de ajuste es el grado de ajuste de un modelo de regresión a la muestra de datos. Es decir, la bondad de ajuste de un modelo de regresión se refiere al nivel de acoplamiento entre el conjunto de observaciones y los valores obtenidos mediante la regresión.

Por lo tanto, cuanto mayor sea la bondad de ajuste de un modelo de regresión, significa que mejor explica los datos estudiados. Así pues, nos interesa que el modelo estadístico esté cuanto más ajustado mejor.

Como puedes ver en la imagen anterior, generalmente el valor de una observación no se puede explicar totalmente mediante el modelo de regresión. Pero, lógicamente, cuanto más pueda explicar el modelo de regresión del conjunto de datos, mejor ajustado estará el modelo. En definitiva, nos interesa un modelo de regresión lo más ajustado posible.

Para determinar la bondad de ajuste de un modelo de regresión se suele utilizar el coeficiente de determinación, que es un coeficiente estadístico que indica el porcentaje explicado por el modelo de regresión. Así pues, cuanto mayor sea el coeficiente de determinación de un modelo, más ajustado estará el modelo a la muestra de datos.

$$R^2 = \text{Coeficiente de determinación}$$

No obstante, cabe destacar que cuantas más variables tenga un modelo de regresión, mayor será su coeficiente de determinación. Por eso también se suele utilizar el coeficiente de determinación ajustado para medir la calidad del ajuste de un modelo. El coeficiente de determinación ajustado es una variación del coeficiente anterior que indica el porcentaje explicado por el modelo de regresión penalizando por cada variable explicativa incluida en el modelo.

Por lo tanto, es mejor usar el coeficiente de determinación ajustado para comparar dos modelos que tienen un número de variables diferentes, ya que tiene en cuenta el número de variables incluidas en el modelo

Por último, cabe destacar que también se puede emplear la prueba chi cuadrado para medir la calidad del ajuste de un modelo de regresión, aunque normalmente se suelen utilizar los valores de los dos coeficientes anteriores.

Ejemplo resuelto de la bondad de ajuste

Para terminar, vamos a ver un ejercicio resuelto de la bondad de ajuste para acabar de asimilar este concepto estadístico.

Con la misma serie de datos, se realizan dos modelos de regresión lineales diferentes cuyos resultados puedes ver en la siguiente table ¿Cuál es el modelo que más nos conviene utilizar?

	Modelo de regresión 1	Modelo de regresión 2
Coeficiente de determinación	57%	64%
Coeficiente de determinación ajustado	49%	43%
Número de variables explicativas	3	7

En este caso, suponemos que los dos modelos cumplen con los supuestos previos de los modelos de regresión lineales y, por tanto, solo debemos analizar la bondad de ajuste de los modelos.

El modelo de regresión 2 tiene un coeficiente de determinación mayor que el modelo de regresión 1, por lo que a priori parece un modelo de regresión mejor pues es capaz de explicar mejor la muestra de datos.

Sin embargo, el modelo de regresión 2 tiene 7 variables independientes en el modelo, mientras que el modelo de regresión 1 solamente tiene 3. De modo que el modelo 2 será mucho más complicado y más difícil de interpretar que el primer modelo.

Además, si nos fijamos en el coeficiente de determinación ajustado, el cual tiene en cuenta el número de variables del modelo, el modelo de regresión 1 tiene un coeficiente de determinación ajustado mayor que el modelo de regresión 2.

En conclusión, es mejor utilizar el modelo de regresión 1, ya que su coeficiente de determinación ajustado es mayor que el del modelo de regresión 2. Aunque el modelo de regresión 2 tiene un coeficiente de determinación sin ajustar mayor, esto se debe a que se han

incluido muchas más variables en el modelo, lo que hace aumentar el valor de dicho coeficiente pero dificulta la interpretación del modelo y, seguramente, provoque que la predicción de un valor nuevo sea peor.

Para comparar modelos con diferentes números de variables, es mejor utilizar el coeficiente de determinación ajustado porque penaliza por cada variable añadida al modelo. Como has podido ver en este ejemplo, según el coeficiente de determinación sin ajustar es mejor el modelo de regresión 2, no obstante, gracias al coeficiente de determinación ajustado podemos saber que el modelo de regresión 1 es en realidad mejor.

3.10 Teoría de la probabilidad

En cualquier caso, sin desviarnos del concepto de teoría de la probabilidad, diremos que está formada por un conjunto de técnicas que nos permiten asignar un número a la posibilidad de que un evento ocurra.

Así, en el caso de una moneda, sabemos que al tirarla sobre un tablero el resultado puede ser cara o cruz. Suponiendo que la moneda y el tablero son perfectos y que las condiciones de lanzamiento no cambian, la probabilidad debe ser de 50% cara y 50% cruz.

En este punto nace el concepto de probabilidad. La probabilidad es un número entre 0 y 1, habitualmente expresado en % entre 0 y 100 que nos dice en cuántas ocasiones, de media, ocurrirá un suceso cada 100 veces.

Con esto en mente, llegamos a la conclusión, de que la teoría de la probabilidad se encarga de estudiar qué número entre 0 y 1 debemos asignar a un determinado suceso. Es decir, se encarga, de estudiar las probabilidades de suceder de un evento.

Historia de la teoría de la probabilidad

La probabilidad como concepto existe desde hace miles de años. Hay evidencias históricas, que nos indican que la primera civilización (Sumeria) era capaz de construir dados de 4 caras trabajando con huesos. Más tarde, primero en Egipto y luego en Grecia y Roma, los juegos de azar se hicieron populares.

Sin embargo, y a pesar de todo, las primeras publicaciones que acuñaban o intuían el concepto de probabilidad se escribieron a mediados del siglo XVI. Concretamente, fue Gerolamo Cardano quién escribió en 1553 un tratado sobre el juego de dados. Aunque su obra no sería publicada hasta 110 años más tarde, en 1663.

Posteriormente, han surgido avances gracias a diversos intelectuales que han permitido con sus publicaciones aumentar y mejorar el conocimiento de la probabilidad. Ejemplo de ello son intelectuales como Laplace, Gauss o Kolmogorov.

Diferencia entre estadística y probabilidad

Como decíamos al inicio estadística y probabilidad son habitualmente confundidos. Son conceptos relacionados pero bajo ningún concepto son sinónimos. La diferencia puede parecer, en un principio, algo sin importancia. Nada más lejos de la realidad. Conocer la diferencia entre uno y otro concepto nos ayudará a entenderlos mejor y obtener un conocimiento más preciso de la materia.

Así pues, la teoría de la probabilidad constituye un aparato matemático. Una herramienta que proviene de la ciencia matemática. La estadística, por decirlo de alguna manera, se sirve de dicha herramienta para poder llegar a conclusiones más precisas. Por tanto, probabilidad no es lo mismo que estadística. Y de hecho, es más, ni siquiera es una rama de la estadística.

3.11 Modelos teóricos de distribución de probabilidad

En ocasiones, algunas variables aleatorias siguen distribuciones de probabilidad muy concretas, como por ejemplo el estudio a un colectivo numeroso de individuos que se modelizan por la distribución “Normal”. Estudiaremos algunas de las distribuciones o modelos de probabilidad más importantes y que después nos resultarán muy útiles para el tema de la Estimación. Como hemos visto, las variables pueden ser discretas o continuas; por ello, también las distribuciones podrán ir asociadas a variables aleatorias discretas o continuas

Distribución uniforme discreta

Sea X una variable aleatoria discreta que toma valores x1.....xn tales la probabilidad de tomar

cada uno de los valores es () n xX i 1 == P . Cuando esto ocurre se dice que X se distribuye como una variable aleatoria Uniforme discreta. Esta es la distribución discreta más sencilla, la cual asigna la misma probabilidad a cada una de las soluciones.

Distribución de Bernouilli Considerado un experimento aleatorio en el cual solo hay dos posibles resultados incompatibles a los que se les puede denominar éxito o fracaso, entonces se dice que X es una variable aleatoria discreta que se distribuye como parámetro "p" donde "p" es la probabilidad de obtener éxito., y se expresa)

3.12 La distribución binominal

Existen una gran diversidad de experimentos o sucesos que pueden ser caracterizados bajo esta distribución de probabilidad. Imaginemos el lanzamiento de una moneda en el que definimos el suceso "sacar cara" como el éxito. Si lanzamos 5 veces la moneda y contamos los éxitos (sacar cara) que obtenemos, nuestra distribución de probabilidades se ajustaría a una distribución binomial.

Por lo tanto, la distribución binomial se entiende como una serie de pruebas o ensayos en la que solo podemos tener 2 resultados (éxito o fracaso), siendo el éxito nuestra variable aleatoria.

Propiedades de la distribución binomial

Para que una variable aleatoria se considere que sigue una distribución binomial, tiene que cumplir las siguientes propiedades:

En cada ensayo, experimento o prueba solo son posibles dos resultados (éxito o fracaso).

La probabilidad del éxito ha de ser constante. Esta se representa mediante la letra p. La probabilidad de que salga cara al lanzar una moneda es 0,5 y esta es constante dado que la moneda no cambia en cada experimento y las probabilidades de sacar cara son constantes.

La probabilidad de fracaso ha de ser también constate. Esta se representa mediante la letra q = 1-p. Es importante fijarse que mediante esa ecuación, sabiendo p o sabiendo q, podemos obtener la que nos falte.

El resultado obtenido en cada experimento es independiente del anterior. Por lo tanto, lo que ocurra en cada experimento no afecta a los siguientes.

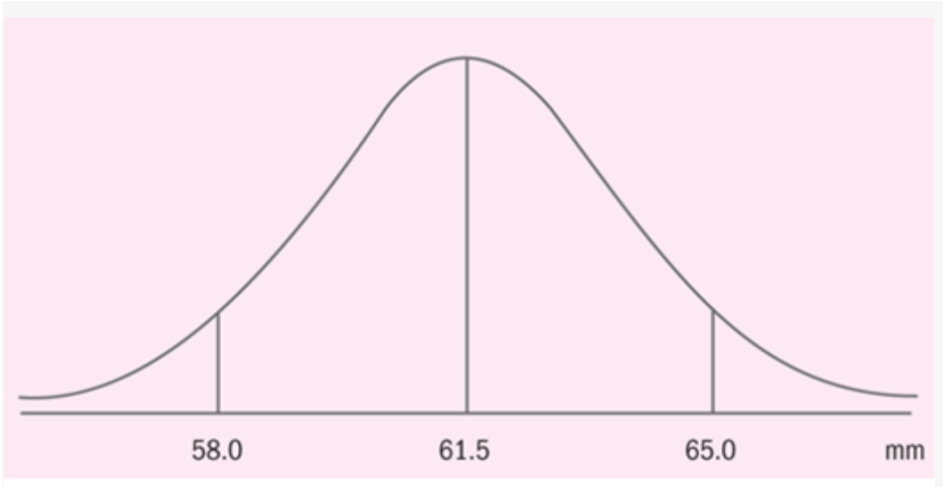
Los sucesos son mutuamente excluyentes, es decir, no pueden ocurrir los 2 al mismo tiempo. No se puede ser hombre y mujer al mismo tiempo o que al lanzar una moneda salga cara y cruz al mismo tiempo.

Los sucesos son colectivamente exhaustivos, es decir, al menos uno de los 2 ha de ocurrir. Si no se es hombre, se es mujer y, si se lanza una moneda, si no sale cara ha de salir cruz.
La variable aleatoria que sigue una distribución binomial se suele representar como X~(n,p), donde n representa el número de ensayos o experimentos y p la probabilidad de éxito.

3.13 La distribución o curva normal

La distribución normal o distribución de Gauss representa la forma en la que se distribuyen en la naturaleza los diversos valores numéricos de las variables continuas, como pueden ser estatura, peso, etc.

Para el caso de una variable de origen biológico, como es la distancia interpupilar (DIP) en los adultos sanos, se sabe que existen muchos individuos con una DIP cercana a 61.5 mm. También hay muchos con una DIP de 61 o 62 mm pero ya no son tan numerosos como los de 61.5 mm. Asimismo es posible encontrar personas con DIP de 58 o 65 mm, pero la frecuencia de este tipo de valores es muy escasa. La forma en que se distribuyen naturalmente los valores numéricos de la DIP se ilustra en la figura



Cuando se calcula la desviación estándar para una serie de datos no siempre es evidente el significado del resultado obtenido, y menos aún si no se compara con la desviación estándar de otra serie diferente de datos.

Para muchas personas podría tener significado el hecho de que el promedio de peso de un grupo de 300 personas fue de 80 kg, pues de acuerdo con la definición del promedio, imaginarían que si todos los individuos tuvieran el mismo peso éste sería de 80 kg; sin embargo, para quienes no tienen conocimiento de las características básicas del modelo de la curva normal podría carecer de significado que les mencionaran que la desviación estándar del peso de las mismas personas fue de 5 kg.

Interpretar la desviación estándar y comprender lo que significa en relación con los datos cuantitativos que se estén manejando sólo es posible a la luz del conocimiento del modelo de la curva normal.

Propiedades de la curva normal

La curva normal es un polígono de frecuencias en forma de campana para el que están calculadas sus áreas en función de los diversos valores del eje horizontal o abscisa

En la abscisa se encuentran valores de tipo cuantitativo continuo, denominados genéricamente como valores z, cuyas magnitudes ...

3.14 La selección de la muestra

Los métodos de selección de muestras son los métodos específicos que se utilizan para seleccionar los registros contenidos en una muestra.

Para el muestreo por registros y el muestreo por unidad monetaria, Analytics admite tres métodos de selección:

- Intervalo fijo
- celda
- aleatorio

Para el muestreo de variables clásicas, la única posibilidad es usar el método de selección aleatorio.

Tipo de muestreo en relación con el método de selección de muestras

Es importante comprender la diferencia entre el tipo de muestreo y el método de selección de muestras.

El **Tipo de muestreo** hace referencia al método estadístico general que se utiliza para llegar a un cálculo aproximado acerca de una población.

El **Método de selección de muestras** hace referencia a la manera en la que se extraen los registros de una población para incluirlos en una muestra.

UNIDAD IV

PRUEBA DE HIPÓTESIS CON UNA, DOS Y VARIAS MUESTRAS DE DATOS NUMÉRICOS

4.1 Metodología para la prueba de hipótesis

Una hipótesis se define como una afirmación transitoria que debe ser sometida a prueba. La inferencia estadística propone un procedimiento para llevar a cabo la prueba de las hipótesis. Propone, primero, enunciarlas formalmente y luego contrastarlas con la evidencia de los datos. Son los datos, entonces, con su coro de características, los que dirán si una hipótesis es falsa o verdadera.

Este procedimiento se realiza considerando a los parámetros, que ya sabemos corresponden al universo, como los objetos para los cuales se enuncian las hipótesis. Dicho de otro modo, una hipótesis se enuncia para una característica del universo o población y se origina en la observación del comportamiento de la misma característica en un grupo restringido o muestra.

Una hipótesis por ejemplo, al decir: "estos enfermos demoran en promedio 25 días en recuperarse" está afirmando que, en el universo, el promedio de los pacientes tardan 25 días en mejorar. Será tarea del investigador probar la veracidad o falsedad de dicha afirmación contrastando el valor propuesto para el parámetro del universo (25 días), con los datos reales

provenientes de una muestra cualquiera. Si luego de esta comparación resulta que el promedio obtenido en la muestra es de 22 días, se le encarga a la estadística que resuelva el dilema de si la diferencia entre el promedio muestral (22 días) y el poblacional (25 días) permite aceptar como verdadera la hipótesis planteada. Será el método estadístico el que permita en definitiva resolver este dilema, evaluando la significación de la diferencia entre 22 y 25.

¿Azar o no?

El método de las pruebas de hipótesis consiste fundamentalmente en establecer la probabilidad de que sea consecuencia del azar la diferencia existente entre dos cantidades. Se pueden distinguir dos situaciones:

a) Diferencia entre un valor muestral y un valor poblacional, o valor teórico.

b) Diferencia entre dos o más valores muestrales.

En el caso **a** se tratará de evaluar la diferencia entre un valor obtenido en la muestra (estadístico) y un valor correspondiente en el universo (parámetro), y en el caso **b** se evaluará la diferencia entre dos valores provenientes de dos muestras (estadísticos). Los valores que se comparen, ya sean de la muestra o del universo, pueden ser promedios, porcentajes u otros. Nosotros nos ocuparemos sólo de promedios y porcentajes.

En general, lo que hace una prueba estadística es evaluar la diferencia entre dos o más valores (dos promedios, dos porcentajes). Respecto de esta diferencia se elabora una hipótesis previa y se plantea formalmente en términos estadísticos.

Luego, usando la distribución de probabilidad adecuada, se calcula la probabilidad de la diferencia entre los valores comparados. Si la probabilidad de obtener tal diferencia es pequeña, diremos que dicha diferencia es significativa.

4.2 Hipótesis nula y alternativa

Las hipótesis nula y alternativa son dos enunciados mutuamente excluyentes acerca de una población. Una prueba de hipótesis utiliza los datos de la muestra para determinar si se puede

rechazar la hipótesis nula.

Hipótesis nula (H0)

La hipótesis nula indica que un parámetro de población (tal como la media, la desviación estándar, etc.) es igual a un valor hipotético. La hipótesis nula suele ser una afirmación inicial que se basa en análisis previos o en conocimiento especializado.

Hipótesis alternativa (H1)

La hipótesis alternativa indica que un parámetro de población es más pequeño, más grande o diferente del valor hipotético de la hipótesis nula. La hipótesis alternativa es lo que usted podría pensar que es cierto o espera probar que es cierto.

Hipótesis unilaterales y bilaterales

La hipótesis alternativa puede ser unilateral o bilateral.

Bilateral

Utilice una hipótesis alternativa bilateral (también conocida como hipótesis no direccional) para determinar si el parámetro de población es mayor que o menor que el valor hipotético. Una prueba bilateral puede detectar cuándo el parámetro de población difiere en cualquier dirección, pero tiene menos potencia que una prueba unilateral.

Unilateral

Utilice una hipótesis alternativa unilateral (también conocida como hipótesis direccional) para determinar si el parámetro de población difiere del valor hipotético en una dirección específica. Usted puede especificar la dirección para que sea mayor que o menor que el valor hipotético. Una prueba unilateral tiene mayor potencia que una prueba bilateral, pero no puede detectar si el parámetro de población difiere en la dirección opuesta. Ejemplos de hipótesis bilaterales y unilaterales

Bilateral

Un investigador tiene los resultados de una muestra de estudiantes que presentaron un examen nacional en una escuela secundaria. El investigador desea saber si las calificaciones de esa escuela difieren del promedio nacional de 850. Una hipótesis alternativa bilateral (también conocida como hipótesis no direccional) es adecuada porque el investigador está interesado en determinar si las calificaciones son menores que o mayores que el promedio nacional. (H0: $\mu = 850$ vs. H1: $\mu \neq 850$)

Unilateral

Un investigador tiene los resultados de una muestra de estudiantes que tomaron un curso de preparación para un examen nacional. El investigador desea saber si los estudiantes preparados tuvieron puntuaciones por encima del promedio nacional de 850. Una hipótesis alternativa unilateral (también conocida como hipótesis direccional) se puede utilizar porque el investigador plantea la hipótesis de que las puntuaciones de los estudiantes preparados son mayores que el promedio nacional. (H0: $\mu = 850$ vs. H1: $\mu > 850$)

4.3 Error tipo I y tipo II

Ninguna prueba de hipótesis es 100% cierta. Puesto que la prueba se basa en probabilidades, siempre existe la posibilidad de llegar a una conclusión incorrecta. Cuando usted realiza una prueba de hipótesis, puede cometer dos tipos de error: tipo I y tipo II. Los riesgos de estos dos errores están inversamente relacionados y se determinan según el nivel de significancia y la potencia de la prueba. Por lo tanto, usted debe determinar qué error tiene consecuencias más graves para su situación antes de definir los riesgos.

Error de tipo I

Si usted rechaza la hipótesis nula cuando es verdadera, comete un error de tipo I. La

probabilidad de cometer un error de tipo I es α, que es el nivel de significancia que usted establece para su prueba de hipótesis. Un α de 0.05 indica que usted está dispuesto a aceptar una probabilidad de 5% de estar equivocado al rechazar la hipótesis nula. Para reducir este riesgo, debe utilizar un valor menor para α. Sin embargo, usar un valor menor para alfa significa que usted tendrá menos probabilidad de detectar una diferencia si esta realmente existe.

Error de tipo II

Cuando la hipótesis nula es falsa y usted no la rechaza, comete un error de tipo II. La probabilidad de cometer un error de tipo II es β, que depende de la potencia de la prueba. Puede reducir el riesgo de cometer un error de tipo II al asegurarse de que la prueba tenga suficiente potencia. Para ello, asegúrese de que el tamaño de la muestra sea lo suficientemente grande como para detectar una diferencia práctica cuando esta realmente exista.

La probabilidad de rechazar la hipótesis nula cuando es falsa es igual a 1–β. Este valor es la potencia de la prueba.

	Verdad acerca de la población	
Decisión basada en la muestra	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta (probabilidad = 1 - α)	**Error tipo II** - no rechazar H_0cuando es falsa (probabilidad = β)
Rechazar H_0	**Error tipo I** - rechazar H_0 cuando es verdadera (probabilidad = α)	Decisión correcta (probabilidad = 1 - β)

Ejemplo de error de tipo I y tipo II

Para entender la interrelación entre los errores de tipo I y tipo II, y para determinar cuál error tiene consecuencias más graves para su situación, considere el siguiente ejemplo.

Un investigador médico desea comparar la efectividad de dos medicamentos. Las hipótesis nula y alternativa son:

Hipótesis nula (H_0): $\mu_1 = \mu_2$

- Los dos medicamentos tienen la misma eficacia.
- Hipótesis alternativa (H_1): $\mu_1 \neq \mu_2$
- Los dos medicamentos no tienen la misma eficacia.

Un error de tipo I se produce si el investigador rechaza la hipótesis nula y concluye que los dos medicamentos son diferentes cuando, en realidad, no lo son. Si los medicamentos tienen la misma eficacia, el investigador podría considerar que este error no es muy grave, porque de todos modos los pacientes se beneficiarían con el mismo nivel de eficacia independientemente del medicamento que tomen. Sin embargo, si se produce un error de tipo II, el investigador no rechaza la hipótesis nula cuando debe rechazarla. Es decir, el investigador concluye que los medicamentos son iguales cuando en realidad son diferentes. Este error puede poner en riesgo la vida de los pacientes si se pone en venta el medicamento menos efectivo en lugar del medicamento más efectivo.

Cuando realice las pruebas de hipótesis, considere los riesgos de cometer errores de tipo I y tipo II. Si las consecuencias de cometer un tipo de error son más graves o costosas que cometer el otro tipo de error, entonces elija un nivel de significancia y una potencia para la prueba que reflejen la gravedad relativa de esas consecuencias.

PRUEBAS DE HIPÓTESIS

Si queremos decidir entre dos hipótesis que afectan a un cierto parámetro de la población, a partir de la información de la muestra usaremos el contraste de hipótesis, cuando optemos

por una de estas dos hipótesis, hemos de conocer una medida del error cometido, es decir, cuantas veces de cada cien nos equivocamos.

En primer lugar, veremos cómo se escribirían las hipótesis que queremos contrastar:

H0 se llama hipótesis nula y es lo contrario de lo que sospechamos que va a ocurrir (suele llevar los signos igual, mayor o igual y menor o igual)

H1 se llama hipótesis alternativa y es lo que sospechamos que va a ser cierto (suele llevar los signos distinto, mayor y menor)

Los contrastes de hipótesis pueden ser de dos tipos:

Bilateral: En la hipótesis alternativa aparece el signo distinto.

Unilateral: En la hipótesis alternativa aparece o el signo > o el signo <.

Podemos aceptar una hipótesis cuando en realidad no es cierta, entonces cometeremos unos errores, que podrán ser de dos tipos:

Error de tipo I: Consiste en aceptar la hipótesis alternativa cuando la cierta es la nula.

Error de tipo II: Consiste en aceptar la hipótesis nula cuando la cierta es la alternativa.

Estos errores los aceptaremos si no son muy grandes o si no nos importa que sean muy grandes.

Alfa: Es la probabilidad de cometer un error de tipo I.

Beta: Es la probabilidad de cometer un error de tipo II.

De los dos, el más importante es alfa que llamaremos nivel de significación y nos informa de la probabilidad que tenemos de estar equivocados si aceptamos la hipótesis alternativa. Debido a que los dos errores anteriores a la vez son imposibles de controlar, vamos a fijarnos

solamente en el nivel de significación, este es el que nos interesa ya que la hipótesis alternativa que estamos interesados en probar y no queremos aceptarla si en realidad no es cierta, es decir, si aceptamos la hipótesis alternativa queremos equivocarnos con un margen de error muy pequeño.

El nivel de significación lo marcamos nosotros. Si es grande es más fácil aceptar la hipótesis alternativa cuando en realidad es falsa. El valor del nivel de significación suele ser un 5%, lo que significa que 5 de cada 100 veces aceptamos la hipótesis alternativa cuando la cierta es la nula.

Solamente vamos a estudiar el contraste bilateral para la media.

CONTRASTE DE HIPÓTESIS BILATERAL PARA LA MEDIA

Si se cumple una de las siguientes hipótesis:

El tamaño de la muestra es mayor de 30 y la variable sigue un modelo normal.

El tamaño de la muestra es mayor de 100.

Estudiaremos el siguiente contrate de hipótesis bilateral:

Calculamos los siguientes valores:
Valor experimental que se calcula a partir de la muestra.

Valor teórico y es el valor que en la distribución N (0,1) deja a su derecha un área de alfa/2 para un nivel de significación alfa. Es el valor z que definíamos ala principio del tema.

La regla de decisión fijado el nivel de significación, alfa, es la siguiente:

Si se acepta la hipótesis alternativa, llegamos a la conclusión de que la hipótesis es cierta. Si se acepta la hipótesis nula, en realidad no podemos afirmar que sea cierta, sino que la hipótesis alternativa no es cierta, ya que el margen de error con el que se acepta la hipótesis nula es

muy grande.

Actividad 21. Un equipo de psicólogos han comprobado que en cierta población infantil, el tiempo (en minutos) empleado en realizar determinada actividad manual, sigue un modelo Normal de probabilidad. Un grupo de 36 niños, seleccionados aleatoriamente en dicha población, realizaron esa actividad manual en un tiempo medio de 6,5 minutos con una desviación típica muestral de 1,5 minutos. A partir de esta información, para un nivel de significación del 1% (alfa=0,01) ¿podíamos rechazar la hipótesis de que el tiempo medio en la población es de 7 minutos? Utiliza la escena siguiente.

No podemos aceptar la hipótesis alternativa, y por lo tanto no podemos rechazar la hipótesis de que el tiempo medio en la población es de 7 minutos.

Actividad 22. El gerente de una empresa selecciona aleatoriamente entre sus trabajadores una muestra de 169 y anota el número de horas de trabajo que cada uno de ellos ha perdido por causa de accidentes laborales en el año 2001. A partir de la información obtenida determina, en esos 169 trabajadores, un número medio de horas perdidas por accidentes laborales en el 2001 de 36,5 horas. Sabiendo que:

Donde representa el número de horas perdidas por el i-ésimo trabajador.

¿Podríamos rechazar, con un nivel de significación del 1% la hipótesis de que el número medio de horas perdidas a causa de accidentes laborales en esa empresa durante el año 2001 fue de 35 horas?

¿Y para un nivel de significación del 5%?

4.4 Prueba de hipótesis Z para la media

Dentro del estudio de la inferencia estadística, se describe como se puede tomar una muestra aleatoria y a partir de esta muestra estimar el valor de un parámetro poblacional en la cual se puede emplear el método de muestreo y el teorema del valor central lo que permite explicar cómo a partir de una muestra se puede inferir algo acerca de una población, lo cual

nos lleva a definir y elaborar una distribución de muestreo de medias muestrales que nos permite explicar el teorema del límite central y utilizar este teorema para encontrar las probabilidades de obtener las distintas medias maestrales de una población.

Pero es necesario tener conocimiento de ciertos datos de la población como la media, la desviación estándar o la forma de la población.

En este caso es necesario hacer una estimación puntual que es un valor que se usa para estimar un valor poblacional. Pero una estimación puntual es un solo valor y se requiere un intervalo de valores a esto se denomina intervalo de confianza y se espera que dentro de este intervalo se encuentre el parámetro poblacional buscado. También se utiliza una estimación mediante un intervalo, el cual es un rango de valores en el que se espera se encuentre el parámetro poblacional

En nuestro caso se desarrolla un procedimiento para probar la validez de una aseveración acerca de un parámetro poblacional este método es denominado Prueba de hipótesis para una muestra.

4.5 Varianza

La varianza es una medida de dispersión que representa la variabilidad de una serie de datos con respecto a su media. Formalmente, se calcula como la suma de los cuadrados de los residuos dividida por las observaciones totales.

También puede calcularse como la desviación estándar al cuadrado. Por cierto, entendemos el residuo como la diferencia entre el valor de una variable a la vez y el valor medio de toda la variable.

El cálculo de la varianza es necesario para calcular la desviación estándar.

La varianza se utiliza para ver cómo se relacionan los números individuales dentro de un

conjunto de datos, en lugar de utilizar técnicas matemáticas más amplias.

También se distingue por tratar a todas las desviaciones de la media como si fueran iguales, independientemente de su dirección. Las desviaciones al cuadrado no pueden sumar cero y dar la apariencia de que no hay variabilidad en los datos.

Sin embargo, un inconveniente es que da más peso a los valores atípicos. Estos son números alejados de la media. Elevar al cuadrado estos números puede sesgar los datos.

Otro inconveniente del uso de la varianza es que no es fácil de interpretar. Se suele emplear principalmente para sacar la raíz cuadrada de su valor, que indica la desviación estándar de los datos.

Ejemplo de varianza

He aquí un ejemplo hipotético para demostrar cómo funciona la varianza, es este caso en el rubro de finanzas. Supongamos que los rendimientos de las acciones de la empresa ABC son del 10% en el primer año, del 20% en el segundo y del -15% en el tercero. La media de estas tres rentabilidades es del 5%. Las diferencias entre cada rendimiento y la media son del 5%, 15% y -20% para cada año consecutivo.

Al elevar al cuadrado estas desviaciones se obtiene un 0,25%, un 2,25% y un 4,00%, respectivamente. Si sumamos estas desviaciones al cuadrado, obtenemos un total del 6,5%. Si dividimos la suma del 6,5% entre uno menos el número de rendimientos del conjunto de datos, ya que se trata de una muestra (2 = 3-1), nos da una varianza del 3,25% (0,0325). Si se saca la raíz cuadrada de la varianza, se obtiene una desviación estándar del 18% ($\sqrt{0,0325}$ = 0,180) para los rendimientos.

Cómo se calcula la varianza

Siga estos pasos para calcular la varianza:

Calcula la media de los datos.

Encuentra la diferencia de cada punto de datos con respecto al valor medio.

Eleva al cuadrado cada uno de estos valores.

Suma todos los valores elevados al cuadrado.

Divide esta suma de cuadrados entre n – 1 (para una muestra) o N (para la población).

Fórmula para calcular la varianza

Antes de ver la fórmula, hay que decir que la varianza en estadística es muy importante. Porque, aunque es una medida sencilla, puede aportar mucha información sobre una variable concreta.

La unidad de medida será siempre la unidad de medida correspondiente a los datos pero al cuadrado. La varianza es siempre mayor o igual a cero. Como los residuos se elevan al cuadrado, es matemáticamente imposible que la varianza sea negativa. Y, por tanto, no puede ser inferior a cero.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

4.6 Desviación estándar

La desviación estándar es una medida de extensión o variabilidad en la estadística descriptiva. Se utiliza para calcular la variación o dispersión en la que los puntos de

datos individuales difieren de la media.

Una desviación baja indica que los puntos de datos están muy cerca de la media, mientras que una desviación alta muestra que los datos están dispersos en un rango mayor de valores.

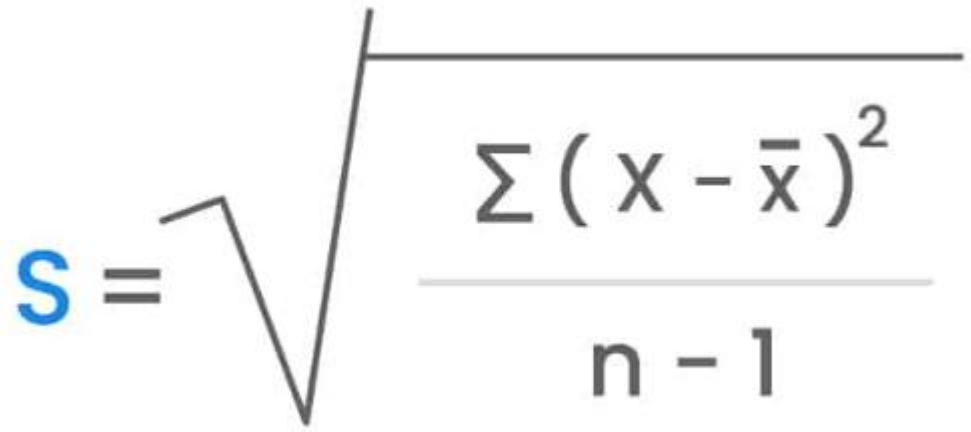
En el ámbito del marketing, la desviación puede ayudar a tener en cuenta la gran variación de los costes o las ventas.

La desviación estándar también ayuda a determinar la dispersión de los precios de los activos con respecto a su precio medio y la volatilidad en el mercado.

Conoce también qué es la desviación media.

Fórmula de la desviación estándar de una muestra

La desviación estándar es un componente fundamental para calcular el tamaño de la muestra de investigación. La fórmula para calcularla es la siguiente:



4.7 Pruebas para producciones

Frente a dos posibilidades reales, no hay diferencias (H0) o bien sí las hay (H1), las pruebas de hipótesis pueden dar dos resultados: rechazar o aceptar H0. En estas circunstancias, en forma análoga a lo que sucede con los exámenes de laboratorio diagnósticos, las alternativas son cuatro. Dos no constituyen más que la coincidencia entre la realidad y el resultado de las pruebas:

Se rechaza H0 cuando ésta es falsa, una diferencia verdadera es declarada estadísticamente significativa. Es un verdadero positivo.

Se acepta H0 cuando ésta es verdadera, no hay una diferencia estadísticamente significativa y en realidad no la hay. Un verdadero negativo.

Las otras alternativas implican una incongruencia entre la realidad y los resultados y, por lo tanto, constituyen errores.

Se rechaza H0 cuando ésta es verdadera, concluyendo que hay una diferencia que en realidad no existe, un falso positivo. Se ha cometido un error que se denomina de tipo I (α). La probabilidad de que ocurra este tipo de error es la que se controla al establecer α y normalmente no va más allá del 5%. Sin embargo, inadvertidamente puede ser mayor cuando no se cumplen los requisitos necesarios para aplicar la prueba de hipótesis elegida: usar un test paramétrico cuando en realidad se debió usar uno no paramétrico, una prueba de una cola en vez de una de dos colas o comparaciones múltiples con tests diseñados para comparar sólo dos medias o medianas.

Se acepta H0 cuando en realidad es falsa, un falso negativo, concluyendo que no hay diferencia cuando en realidad existe. Este es el error tipo II (β), que la mayoría de las veces se debe a un tamaño insuficiente de la muestra. La probabilidad de cometer un error tipo II es β cuyo valor depende de la magnitud del efecto de interés y del tamaño de la muestra. Sin embargo, es más

frecuente hablar de la potencia de la prueba para detectar un efecto de un tamaño determinado.

Estos dos errores deben ser considerados al evaluar el resultado de un trabajo de investigación que haya empleado pruebas de hipótesis, considerando la posibilidad de un error I cuando los resultados son significativos y de un error tipo II cuando son no significativos.

4.8 Distribución y T de Student

Descrita por William S. Gosset en 1908. Publicaba bajo el pseudónimo de "Student" mientras trabajaba para la cervecería Guinnes en Irlanda. Está diseñada para probar hipótesis en estudios con muestras pequeñas (menores de 30)

La fórmula general para la T de Student es la siguiente:

$$t = \frac{X - \mu}{s/\sqrt{n}}$$

En donde el numerador representa la diferencia a probar y el denominador la desviación estándar de la diferencia llamado también Error Estándar. En esta fórmula t representa al valor estadístico que estamos buscando X barra es el promedio de la variable analizada de la muestra, y miu es el promedio poblacional de la variable a estudiar. En el denominador tenemos a s como representativo de la desviación estándar de la muestra y n el tamaño de ésta.

Grados de libertad: El número de grados de libertad es igual al tamaño de la muestra (número de observaciones independientes) menos 1.

gl = df = (n – 1)

Si pudiera expresar en un cierto número de pasos para resolver un problema de t de student tendría que declarar los siguientes:

Paso 1. Plantear la hipótesis nula (H_0) y la hipótesis alternativa (H_1). La hipótesis alternativa plantea matemáticamente lo que queremos demostrar, en tanto que la hipótesis nula plantea

exactamente lo contrario.

Paso 2. Determinar el nivel de significancia (rango de aceptación de la hipótesis alternativa), a.

Se considera un nivel alfa de: 0.05 para proyectos de investigación; 0.01 para aseguramiento de la calidad; y 0.10 para estudios o encuestas de mercadotecnia.

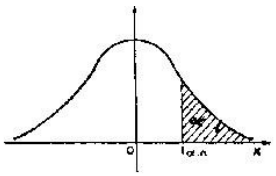
Paso 3. Evidencia muestral, se calcula la media y la desviación estándar a partir de la muestra.

Paso 4. Se aplica la distribución T de Student para calcular la probabilidad de error por medio de la fórmula general presentada al principio y se contrasta con el valor T obtenido de la tabla correspondiente.

Paso 5. En base a la evidencia disponible se acepta o se rechaza la hipótesis alternativa. Si la probabilidad de error (p) es mayor que el nivel de significancia se rechaza la hipótesis alternativa. Si la probabilidad de error (p) es menor que el nivel de significancia se acepta la hipótesis alternativa.

Por supuesto que al final lo que tenemos que contrastar es el valor de T que hayamos obtenido en el problema contra el valor T crítico que obtenemos de la tabla de T de Student.

Aquí puedes bajar la tabla de T de Student: Tabla T de Student un extremo y Tabla T de Student de dos extremos: Tabla T de Student dos extremos.



gl \ α/2	0,40	0,30	0,20	0,10	0,050	0,025	0,010	0,005	0,001	0,0005
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,192	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,259	0,851	1,303	1,648	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,262	3,495
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,256	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Si el resultado del problema cae en la región de H0 se acepta ésta, de lo contrario se rechaza. Por supuesto, si rechazas H0 aceptarás H1.

Ejercicio 1: Se aplica una prueba de autoestima a 25 personas quienes obtienen una calificación promedio de 62.1 con una desviación estándar de 5.83. Se sabe que el valor correcto de la prueba debe ser mayor a 60. ¿Existe suficiente evidencia para comprobar que no hay problemas de autoestima en el grupo seleccionado?

Paso 1. Hipótesis alternativa: la que se va a comprobar. El grupo no tiene problemas de autoestima. Valor de prueba para determinar autoestima mayor a 60. Hipótesis nula, lo contrario a la hipótesis alternativa.

Paso 2. Determinar el nivel de significancia alfa: alfa = 0.05. Paso 3. Resultados de la evidencia muestral: X = 62.1; s = 5.83 Paso 4. Aplicar la distribución de probabilidad calculando T:

$$T = \frac{62.1 - 60}{5.83/\sqrt{25}}$$

El resultado de la ecuación es 1.8. Dado que 1.8 es mayor que 1.7109 cae en la región de H1 y se acepta la hipótesis alternativa. Si buscamos el valor de 1.8 bajo la curva normal encontraremos que es de 0.0359 el cual es menor que 0.05. La conclusión es que no hay problemas de autoestima en el grupo estudiado. Esto con el diseño de la investigación presentado.

Veamos en video como resolver problemas de T de student con Excel y con SPSS:

Ejemplo 2: Suponga que Ud. tiene una técnica que puede modificar la edad a la cual los niños comienzan a hablar. En su localidad, el promedio de edad en la cual un niño emite su primera palabra es de 13.0 meses. No se conoce la desviación estándar poblacional. Usted aplica dicha técnica a una muestra aleatoria de 15 niños. Los resultados arrojan que la edad media muestral en la que se pronuncia la primera palabra es de 11.0 meses, con una desviación estándar de 3.34. Pruebe la hipótesis de que la técnica afecta la edad en que los niños empiezan a hablar con un nivel de significancia alfa del 0.05.

Aquí las preguntas de la investigación serían ¿Cuáles son la hipótesis nula y la alternativa? y si con el procesamiento estadístico se puede afirmar que la técnica es efectiva para modificar la edad en que los niños empiezan a hablar.

Hipótesis nula: La técnica no afecta la edad en que los niños comienzan a hablar, matemáticamente sería, H0 = 13.0

Hipótesis alternativa: La técnica afecta la edad en que los niños comienzan a hablar, matemáticamente sería, H1 ≠ 13.0

$$Tp = \frac{11 - 13}{3.34/\sqrt{15}}$$

El resultado de Tp es -2.32. Si lo comparamos con el resultado de T crítico o Tc obtenido de tablas con un nivel de significancia alfa de 0.05 y 14 grados de libertad para dos extremos, el resultado de Tc es 2.145

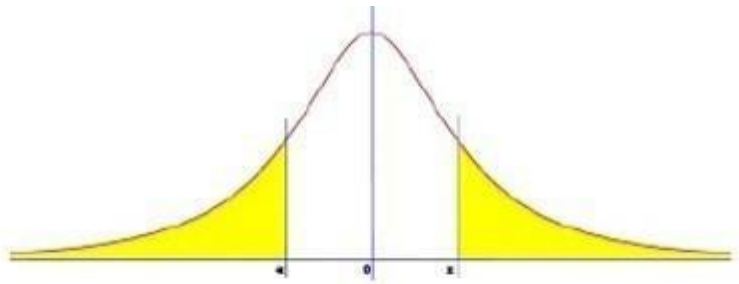


Tabla distribución t. Dos colas, probabilidad dentro(%)/fuera(0.00) del intervalo $\mu \pm t_{n-1}\sigma/\sqrt{n}$

Valor de t para un intervalo de confianza de *Valor crítico de | t | para valores de P de número de grados de libertad*	*90%* *0.10*	*95%* *0.05*	*98%* *0.02*	*99%* *0.01*
1	6.31	12.71	31.82	63.66
2	2.92	4.30	6.96	9.92
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71
7	1.89	2.36	3.00	3.50
8	1.86	2.31	2.90	3.36
9	1.83	2.26	2.82	3.25
10	1.81	2.23	2.76	3.17
12	1.78	2.18	2.68	3.05
14	1.76	2.14	2.62	2.98
16	1.75	2.12	2.58	2.92
18	1.73	2.10	2.55	2.88
20	1.72	2.09	2.53	2.85
30	1.70	2.04	2.46	2.75
50	1.68	2.01	2.40	2.68
∞	1.64	1.96	2.33	2.58

Con los resultados anteriores se rechaza la hipótesis nula y se decide que, la técnica afecta la edad en que los niños comienzan a hablar con un nivel de significancia de 0.05. El valor P correspondiente si lo buscamos en la curva normal de probabilidades sería de 0.010, por debajo del nivel de significancia.

Tabla de valores bajo la curva normal.

Ejemplo 3. Una profesora del programa de estudios para la mujer cree que la cantidad de cigarrillos fumados por las mujeres se ha incrementado en años recientes. Un censo realizado hace dos años con mujeres de una ciudad vecina mostró que el número promedio de cigarrillos fumados diariamente por una mujer era de 5.4 con una desviación estándar de 2.5. Para evaluar esta hipótesis, la profesora determinó el número de cigarrillos fumados diariamente por una muestra aleatoria de 120 mujeres que viven actualmente en la ciudad donde habita. Los datos muestran que el número de cigarrillos fumados diariamente por las 120 mujeres tiene una media de 6.1 y una desviación estándar de 2.7. Con esa información y un nivel de significancia de 0.05, ¿tiene razón la profesora al afirmar que la cantidad de cigarrillos fumados por las mujeres se ha incrementado?

$$Tp = \frac{6.1 - 5.4}{2.7/\sqrt{120}}$$

Los resultados de la ecuación muestran una Tp de 2.9 que, contrastada con la Tc obtenida de tablas para un extremo que resulta en 1.6449 cae en la región de rechazo de H0. Si calculamos P en tablas para 2.90 es 0.002, muy por debajo del 0.05 del nivel de significancia.

Ejercicio 4. La siguiente es una tabla de resultados del coeficiente intelectual entre niños que tienen buenas calificaciones en lectura y de aquellos que tienen bajas calificaciones en lectura. A un nivel de significancia del 0.05, ¿hay diferencia significativa entre el coeficiente intelectual entre los grupos? Utilice la prueba de T de Student contrastando las hipótesis contra el valor crítico.

4.9 Prueba de significancia

II.A) Pruebas paramétricas

1. Prueba t de Student

Con esta prueba se pretende averiguar si dos muestras que tienen medias iguales, provienen de la misma población.

Hipótesis nula "H_0" → $\mu_1 = \mu_2$;

Hipótesis alternativa "H_1" → $\mu_1 \neq \mu_2$

La prueba permite comparar la media con su valor verdadero o bien las medias de dos poblaciones. Se basa en los límites de confianza "LC" para el promedio x de n mediciones repetidas (Ec. 2.1). A partir de dicha ecuación tenemos:

$\mu = x \pm t(s/\sqrt{n})$ (Ec. 2.1) → $x - \mu = \pm t\, s/\sqrt{n}$ (Ec. 2.2)

$s/\sqrt{n}$: error estándar "EE" o desviación estándar "DE" de la distribución muestral de medias. Como las medias son $\sqrt{n}$ veces más probables que los resultados aislados, la DE de las medias es $\sqrt{n}$ veces menor que la DE de resultados aislados, siendo **n** el número de determinaciones con las que se calcula la media.

t: "t de student" (tabla 2). Es un parámetro tabulado que depende de los grados de libertad de la muestra (n-1) "gl" y del intervalo de confianza que se quiera (generalmente 95%).

Tabla 2. Valores críticos de t (p= 0.05, entre paréntesis p=0.1).

gl	1	2	3	4	5	6	7	8	9
t	12.71	4.3	3.18	2.78	2.57	2.45	2.36	2.31	2.26
	(6.31)	(2.92)	(2.35)	(2.13)	(2.02)	(1.94)	(1.89)	(1.86)	(1.83)
gl	10	12	14	16	18	20	30	50	α
T	2.23	2.18	2.14	2.12	2.1	2.09	2.04	2.01	1.96
	(1.81)	(1.78)	(1.76)	(1.75)	(1.73)	(1.72)	(1.70)	(1.68)	(1.64)

Nota: *los valores críticos de t son estimados para una prueba de 2 colas. Para una prueba de una cola se toma el valor que corresponde a p=0.1, es decir, el doble del valor de p deseado (0.05).*

Si x - μ obtenida en la muestra a comparar es menor que la calculada para un cierto nivel de probabilidad, no se rechaza la hipótesis nula de que x y μ sean iguales; es decir, sus diferencias son debidas a errores aleatorios y no existe un error sistemático significativo.

Para comparar 2 medias experimentales el proceso es semejante. Se ha de tener en cuenta si los datos de las 2 muestras están apareados o no (figura 1):

Figura 1. Esquema del cálculo de t para datos apareados y no apareados

A) t student con datos apareados

X_{1i}	X_{2i}	D_i
X_{11}	X_{21}	D_1
X_{12}	X_{22}	D_2
...	...	...
$\bar{X}_{1i}$	$\bar{X}_{2i}$	$\bar{D}_i$

$\bar{D}_i = \sum D_i / i$

$S = \sqrt{\sum (D_i - \bar{D}_i)^2/(n-1)}$

$t = \bar{D}_i / EE \Rightarrow EE = s / \sqrt{n}$

Si p≤0.05 rechazar H_0

B) t student con datos no apareados

X_{1i}	X_{2i}
X_{11}	X_{21}
X_{12}	X_{22}
...	...
$\bar{X}_{1i}$	$\bar{X}_{2i}$
$\bar{X}1$	$\bar{X}2$
S_1	S_2

$EE_{(x_1-x_2)} = \sqrt{(\bar{s})^2/(1/n_1 + 1/n_2)}$

siendo $(\bar{s})^2 = (n_1-1)\, s_1^2 + n_2 - 1) s_2^2)/(n_1+n_2-2)$

$t = (x_1-x_2)/EE$ [t tiene (n_1+n_2-2)° de libertad]

Si p≤0.05 rechazar H_0

Datos apareados: tienen la ventaja de permitir trabajar simplificando a una sola muestra (cuyos valores corresponden a la diferencia "D_i" entre cada par de datos apareados). Sustituimos x - μ (Ec. 2.2) por Di - 0 porque el valor real de las diferencias, suponiendo que las dos muestras tienen la misma media, es 0. La DE se calcula con la muestra de diferencias.

Datos no apareados: como no se puede simplificar a una sola muestra, se ha de introducir el concepto de desviación estándar ponderada "sp" (Ec. 2.3). En la **ecuación 2.2** se sustituye s por s_p y x - μ por $x_1 - x_2$ y el tamaño de muestra "n" se sustituye por N ponderado "$(N_1 + N_2)/ N_1N_2$".

$S_p=\sqrt{[S(x_1 - x_1)2+S(x_2 - x_2)2+ ...]/(n_1+ n_2+ ... - N_s)}$ (Ec. 2.3)

n_1, n_2, ...: el tamaño de las muestras.

N_s: número de muestras. $(n_1+n_2+...-N_s)$: número de grados de libertad.

Ejemplo 1: se analizaron dos sueros control (A y B) para la determinación de la glucemia. Se realizó sobre cada uno de ellos 5 determinaciones (tabla 3a) y se quiere determinar si estos dos sueros control son diferentes en relación al nivel de glucosa.

Tabla 3a

	glucosa mg/dl				
Suero A	75	80	82	79	80
Suero B	81	90	85	83	87

Aunque el número de determinaciones es reducido podemos suponer que si realizáramos más determinaciones la distribución sería normal (teorema central del límite). Realizamos la prueba t de student de datos no apareados, ya que aunque las dos muestras tienen el mismo tamaño provienen del análisis de dos sueros supuestamente diferentes (tabla 3b):

Tabla 3b

	$\bar{X}$	$\bar{X}_B - \bar{X}_A$	t (8 g l*)	s ponderada	N ponderado	t $s_P N_P$
Muestra A	79.2					
Muestra B	85.2	6	2.31**	9.45	2.5	13.8

* grados de libertad (ver **Ec. 2.3**). ** valor de t correspondiente (**tabla 2**)

Como la diferencia de las medias es menor que 13.8, puede decirse que las dos muestras son significativamente iguales ($p< 0.05$).

4.10 Comparación de dos muestras independientes

El problema más sencillo es el de comparar la media de un parámetro obtenido en una muestra con un valor de referencia de una población conocida. Esto nos permitirá, por un lado, decidir si es razonable concluir que la muestra puede pertenecer a la población o bien, por otro lado, contrastar hipótesis sobre la media poblacional a partir de la obtenida en la muestra.

Aunque es posible comparar dos medias utilizando la distribución normal, para ello necesitamos conocer la desviación estándar poblacional del parámetro, que suele ser desconocida. En estos casos utilizamos, como aproximación a la desviación estándar poblacional (σ), la de la muestra (s), remplazando las probabilidades de la distribución normal por las de la t de Student, que varían en función del tamaño muestral (los grados de libertad). En cualquier caso, cuando el tamaño muestral es grande, el valor de probabilidad obtenido mediante la t de Student se aproxima al obtenido con la distribución normal.

En la práctica, calculamos el intervalo de confianza de la media poblacional, ya sea utilizando la distribución normal o la t de Student con n-1 grados de libertad (siendo n el tamaño de la muestra), y comprobamos si el intervalo incluye el valor de referencia. La fórmula para el cálculo del intervalo de confianza, utilizando la normal o la t de Student, respectivamente, sería:

$$\mu = Z_{\alpha/_2} \pm \frac{\overline{X}}{\sqrt{\frac{\sigma}{n}}}$$

$$\mu = t_{n\_1;\alpha/_2} \pm \frac{\overline{X}}{\sqrt{\frac{\sigma}{n}}}$$

Aunque el cálculo es relativamente sencillo, aconsejamos utilizar un programa estadístico.

En esta base de datos tenemos los registros de una serie de pacientes asmáticos. Queremos saber si la talla de estos pacientes es similar a la talla media de la población de la misma edad, que sabemos que es de 150 cm. Para ello, una vez abierto RCommander y cargado el conjunto de datos activos, seleccionamos en el menú las opciones Estadísticos/Medias/Test t para una muestra… (figura 1). En la nueva ventana debemos seleccionar la variable de la que queremos obtener el intervalo, la opción que elegimos para la hipótesis alternativa (en este caso, de igualdad de medias, contraste bilateral) y el nivel de significación estadística (lo dejamos en 0,05).

Figura 1. Comparación de una media con un valor de referencia mediante la prueba de la t de Student para una muestra. Mostrar/ocultar

R nos presenta la media de talla de nuestra muestra, 122,21 cm, y su intervalo de confianza del 95%, de 114,85 cm a 129,58 cm. Podemos concluir que la talla de los niños asmáticos es inferior a la talla media conocida de la población general, ya que el intervalo no incluye el valor 150 cm.

El programa nos muestra también el valor del estadístico t, sus grados de libertad (n-1 en este caso) y su significación (p <2,2 × 10-16). Entenderemos mejor estos parámetros al tratar el siguiente punto.

Comparación de dos medias independientes

El supuesto más habitual es el de contrastar si hay una diferencia significativa en la media de una variable de resultado entre dos poblaciones diferentes e independientes. En estos casos, lo habitual es utilizar la prueba de la t de Student para dos muestras independientes.

Esta prueba compara las dos medias de una variable de resultado cuantitativo continuo obtenidas en dos categorías definidas por una variable cualitativa. Se basa en el cálculo del estadístico t, que tiene en cuenta la diferencia de medias a comparar y su error estándar, según la siguiente fórmula:

$$t = \frac{[\overline{X}_1 - \overline{X}_2]}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Siendo

$$\overline{X}_1,\ S_1^2\ y\ \overline{X}_2,\ S_1^2$$

las medias y las varianzas de las dos muestras respectivamente.

Bajo el supuesto de la hipótesis nula, la diferencia de medias es igual a cero, con lo que el valor de t será también igual a cero. Cuanto más se aleje t de ese valor, menos probable será que la diferencia observada se deba al azar.

Para poder aplicar esta prueba, debemos verificar previamente que se cumplen tres condiciones:

Los dos grupos deben ser independientes. Esto quiere decir que cada participante debe pertenecer a solo uno de los dos grupos y no tiene relación con los participantes del otro grupo.

La variable de resultado debe ser continua y seguir una distribución normal en los dos grupos.

Debe cumplirse el supuesto de homocedasticidad, esto es, igualdad de varianzas en los dos grupos.

El supuesto de normalidad de la variable en los dos grupos puede verificarse mediante la prueba de Shapiro-Wilk, más adecuada para muestras pequeñas, menores de 50, o la de Kolmogorov-Smirnov (con la modificación de Lilliefors en el supuesto habitual de desconocer la media y desviación estándar poblacional). Estas dos pruebas tienen el inconveniente de que asumen una hipótesis nula de normalidad. Si el resultado es significativo, podremos descartar la hipótesis nula y concluir que la variable no sigue una distribución normal. Sin embargo, un resultado no significativo no nos permite asegurar que la hipótesis alternativa de normalidad sea cierta (solo que no podemos rechazar la hipótesis nula). Además, ambas son poco potentes cuando el tamaño de la muestra no es grande, precisamente el supuesto en el que la t de Student es más sensible a la ausencia de normalidad. Por estos motivos, se recomienda completar la prueba de contraste con un método gráfico (como el histograma o el gráfico de cuantiles teóricos) y tener en cuenta el tamaño de la muestra.

En el caso de no seguir una distribución normal, se nos plantearán tres posibilidades. La primera, hacer el contraste de hipótesis con una prueba no paramétrica que, en este caso, sería la de la U de Mann-Whitney. La segunda, podemos intentar alguna transformación y comprobar si la variable transformada se distribuye de forma normal. La tercera, realizar una t de Student a pesar de no cumplirse la condición de normalidad, aplicando una corrección. Esto solo será aconsejable si hay ligeras desviaciones de la normalidad y el tamaño muestral es grande (mínimo de 30 a 50 participantes por grupo).

Una vez comprobada la normalidad, determinaremos que las dos varianzas son iguales y se cumple el supuesto de homocedasticidad. Para ello, puede calcularse una F de Snedecor con el cociente de las dos varianzas, colocándose en el numerador la varianza

mayor de los dos grupos y en el denominador, la menor. Los grados de libertad son n-1 de cada grupo, siendo n el tamaño de cada grupo.

Bajo el supuesto de igualdad de varianzas, F valdrá 1. Cuanto mayor sea el valor de F, menos probable será que la diferencia observada entre las varianzas se deba al azar.

La prueba de la F de Snedecor es muy sensible a la falta de normalidad, por lo que en estos casos será recomendable recurrir a la prueba de Levene.

Una vez comprobado estos dos supuestos, procederemos a realizar el contraste mediante la prueba de la t de Student, aplicando la corrección de Welch si no existe homocedasticidad.

Vamos a ver un ejemplo práctico con la misma base de datos, comparando el peso al nacimiento entre niños y niñas.

Primero, calculamos la media para los dos valores. Con RCommander, seleccionamos el menú Estadísticos->Resúmenes->tabla de estadísticas…. En la ventana emergente señalamos "Peso.al.nacimiento" como variable y "Sexo" como factor que diferencia los grupos a comparar. Como estadístico, señalaremos la media.

Obtenemos un peso al nacimiento medio de 2458 g en niñas y 2737 g en niños.

Comprobemos primero el supuesto de normalidad. Para ello, decidimos hacer una prueba de Kolmogorov-Smirnof. En el menú de RCommander seleccionamos las opciones Estadísticos/Resúmenes/Test de normalidad… y, en la ventana emergente, marcamos la variable "Peso.al.nacimiento", elegimos la prueba y pulsamos el botón "Test por grupos…" para seleccionar la variable "Sexo" (figura 2). En la ventana de salida podemos

ver el resultado, siendo la p >0,05 para las dos categorías de la variable "Sexo", por lo que no rechazamos la hipótesis nula de normalidad.

Figura 2. Comprobación del supuesto de normalidad mediante la prueba de Kolmogorov-Smirnov. Mostrar/ocultar

Comprobemos ahora el supuesto de homocedasticidad. Seleccionamos las opciones Estadísticos/Varianzas/Test F para dos varianzas… En la ventana emergente seleccionamos "Peso.al.nacimiento" como variable explicada y la agrupamos por "Sexo" (figura 3). Si marcamos la pestaña de "Opciones", podemos seleccionar el tipo de contraste y el nivel de significación. Dejamos la opción por defecto, que es un contraste bilateral con un nivel de significación de 0,05.

Figura 3. Comprobación del supuesto de homocedasticidad mediante la prueba de la F de Snedecor. Mostrar/ocultar

Vemos que el programa nos da el valor de F (1,2052), los grados de libertad para el numerador y el denominador y la probabilidad de encontrar ese valor de F por azar (valor de p = 0,7695). Por tanto, no rechazamos la hipótesis nula de igualdad de varianzas. Vemos que R nos proporciona también el intervalo de confianza del 95% del valor de F (de 0,36 a 3,45), que incluye el valor 1 que corresponde a varianzas iguales.

Una vez comprobadas normalidad y homocedasticidad, ya podemos realizar la prueba de la t de Student. En este caso, al haber homocedasticidad, no será necesario aplicar la corrección de Welch. Seleccionamos Estadísticos/Medias/Test t para muestras independientes… En la ventana emergente marcamos "Peso.al.nacimiento" como variable explicada y la agrupamos por "Sexo" (figura 4). A continuación, pulsamos la pestaña opciones, donde podemos cambiar el tipo de contraste y la significación (bilateral y 0,05, por defecto) y donde tenemos que señalar si las varianzas son o no iguales. En el caso de que marquemos "No", R hará la prueba aplicando la corrección de Welch. En este caso, marcamos "Sí".

Figura 4. Comparación de dos medias independientes mediante la prueba de la t de Student para muestras independientes. Mostrar/ocultar

El programa nos ofrece el valor del estadístico t (-2,36), sus grados de libertad (n-2, 28) y su valor de *p* (0,02). Por si tenemos alguna duda en la dirección del contraste, también nos dice cuál es la hipótesis alternativa: que la verdadera diferencia entre las medias de los dos grupos es distinta de cero. Al ser el valor de p <0,05, rechazamos la hipótesis nula de igualdad de medias y concluimos que sí existe una diferencia entre los dos grupos en el peso al nacimiento.

Merece la pena comentar que R nos ofrece también las medias de los dos grupos y el intervalo de confianza de la diferencia de medias. En este caso es de -520,88 a -37,17. Viendo que el intervalo no incluye el 0 (valor nulo para la diferencia de medias), podemos también rechazar la hipótesis nula de igualdad sin necesidad de recurrir al valor de *p*.

Comparación de dos medias dependientes o relacionadas

En ocasiones se plantea el problema de comparar las medias de dos grupos que están relacionados, como puede ser el caso de medidas obtenidas del mismo participante en diferentes momentos, de diferentes localizaciones de la misma persona (por ejemplo, presión intraocular de ojo derecho e izquierdo) o cuando se comparan los datos de cada caso con su correspondiente control emparejado. Esto es muy típico de los estudios longitudinales, los estudios de antes-después de una intervención y los estudios de casos y controles.

En estos casos no existe una variable que defina los grupos, sino que la variable de resultado que se valorará será las diferencias entre los dos resultados de cada pareja, suponiendo la hipótesis nula que la media de estas diferencias es igual a cero. Así, en este tipo de análisis el interés no se centra en las diferencias entre individuos, sino en las que

puede haber en el mismo individuo en dos momentos diferentes o entre las observaciones de los individuos relacionados.

La prueba que empleamos en estos casos es la de la t de Student para medidas repetidas (datos apareados o relacionados). Para poder aplicarla, debe cumplirse que la variable de interés sea cuantitativa continua, que la muestra de pares de datos haya sido obtenida al azar de la población y que la diferencia entre las parejas se distribuya de forma normal. Lógicamente, en este caso no tiene sentido plantear si hay igualdad de varianzas, ya que se trata de los mismos participantes en los dos grupos.

El planteamiento de la prueba es similar al de la t de Student para medias independientes, solo que en este caso se genera una nueva variable a partir de las dos medidas a comparar:

$$d_i = x_{i_1} - x_{i_2}$$

donde di es la diferencia de resultado de cada pareja en dos instantes diferentes, x1 y x2.

En este análisis, el estadístico t se obtiene con la media y la desviación estándar de esta variable, según la siguiente ecuación:

$$t = \frac{\bar{d}}{S_{d/\sqrt{n}}}$$

El contraste bilateral plantea la hipótesis nula de que la diferencia es igual a 0. En este caso, los grados de libertad son n-1, siendo n el tamaño de la muestra. Como en el caso de muestras independientes, si la hipótesis nula es cierta el valor de la diferencia será cero, por lo que t valdrá cero. Cuanto mayor sea el valor de t, menos probable será que la diferencia observada se deba al azar.

Desde un punto de vista práctico, el procedimiento es similar al que hemos descrito previamente, solo que esta vez seleccionaremos la opción Estadístico/Medias/Test t para

datos relacionados… de RCommander. La lectura de los resultados será similar, proporcionándonos el programa el valor del estadístico t, su significación y el intervalo de confianza de la diferencia de medias. Animamos al lector a probar estas opciones.

4.11 Prueba de ficheros para varianza y de igualdad de dos poblaciones normales

Este capítulo introduce una nueva función de densidad de probabilidad: la distribución F. Se utiliza para muchas aplicaciones, incluso el ANOVA y para probar la igualdad entre varias medias. Comenzamos con la distribución F y la prueba de la hipótesis de las diferencias en las varianzas. A menudo es conveniente comparar dos varianzas en vez de dos promedios. Por ejemplo, a los administradores del instituto universitario les gustaría que dos profesores que califiquen exámenes tengan la misma variación en su calificación.

Para que una tapa se adapte a un recipiente, la variación en la tapa y del recipiente debería ser aproximadamente la misma. Un supermercado podría estar interesado en la variabilidad de los tiempos para procesar una compra en dos de sus cajas. En finanzas, la varianza es una medida de riesgo; por ende, sería interesante comprobar la hipótesis de que dos carteras de inversión diferentes tienen la misma varianza: la volatilidad.

Para realizar una prueba *F* de dos varianzas, es importante que ocurra lo siguiente:

Las poblaciones de las que se extraen las dos muestras tienen una distribución aproximadamente normal.

Las dos poblaciones son independientes entre sí.

A diferencia de la mayoría de las pruebas de hipótesis en este libro, la prueba *F* para la igualdad de dos varianzas es muy sensible a las desviaciones de la normalidad. Si las dos distribuciones no son normales, o se aproximan, la prueba puede dar un resultado sesgado para el estadístico de prueba.

Supongamos que tomamos una muestra aleatoria de dos poblaciones normales independientes. Supongamos que σ21�12 y σ22�22 son las varianzas poblacionales desconocidas y S21�12 y S22�22 sean las varianzas de la muestra. Supongamos que los tamaños de las muestras son n_1 y n_2. Como nos interesa comparar las dos varianzas de la muestra, utilizamos el cociente *F*:

$$F = \frac{\left[\frac{s_1^2}{\sigma_1^2}\right]}{\left[\frac{s_2^2}{\sigma_2^2}\right]}$$

F tiene la distribución $F \sim F(n_1 - 1, n_2 - 1)$

donde $n_1 - 1$ son los grados de libertad del numerador y $n_2 - 1$ son los grados de libertad del denominador.

Si la hipótesis nula es σ21=σ22�12=�22, entonces el cociente *F*, el estadístico de prueba, se convierte en

$$F_c = \frac{\left[\frac{s_1^2}{\sigma_1^2}\right]}{\left[\frac{s_2^2}{\sigma_2^2}\right]} = \frac{s_1^2}{s_2^2}$$

Las distintas formas de las hipótesis probadas son:

Prueba de dos colas	**Prueba de una cola**	**Prueba de una cola**
$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1^2 \leq \sigma_2^2$	$H_0: \sigma_1^2 \geq \sigma_2^2$

Prueba de dos colas	Prueba de una cola	Prueba de una cola
$H_1: \sigma_1^2 \neq \sigma_2^2$	$H_1: \sigma_1^2 > \sigma_2^2$	$H_1: \sigma_1^2 < \sigma_2^2$

Tabla 12.1

Una forma más general de las hipótesis nula y alternativa para una prueba de dos colas sería:

$$H_0 : \frac{{\sigma_1}^2}{{\sigma_2}^2} = \delta_0$$

$$H_a : \frac{{\sigma_1}^2}{{\sigma_2}^2} \neq \delta_0$$

Donde si $\delta_0 = 1$ es una simple prueba de la hipótesis de que las dos varianzas son iguales. Esta forma de la hipótesis tiene la ventaja de permitir pruebas que van más allá de las simples diferencias y puede dar cabida a pruebas de diferencias específicas, como hicimos con las diferencias de medias y las diferencias de proporciones. Esta forma de la hipótesis también muestra la relación entre la distribución F y la χ^2: la F es un cociente de dos distribuciones de chi-cuadrado, que vimos en el capítulo anterior. Esto sirve para determinar los grados de libertad de la distribución F resultante.

Si las dos poblaciones tienen varianzas iguales, entonces $S21\text{\textdiamond}12$ y $S22\text{\textdiamond}22$ están cerca en valor y el estadístico de prueba, $Fc=s12s22\text{\textdiamond\textdiamond}=\text{\textdiamond}12\text{\textdiamond}22$ está cerca de uno. Pero si las dos variantes de la población son muy diferentes, $S21\text{\textdiamond}12$ y $S22\text{\textdiamond}22$ también suelen ser muy diferentes. Al elegir $S21\text{\textdiamond}12$ ya que la mayor varianza de la muestra hace que el cociente $s12s22\text{\textdiamond}12\text{\textdiamond}22$ sea mayor que uno. Si $S21\text{\textdiamond}12$ y $S22\text{\textdiamond}22$ están muy separados, entonces $Fc=s12s22\text{\textdiamond\textdiamond}=\text{\textdiamond}12\text{\textdiamond}22$ es un número grande.

Por lo tanto, si *F* es cercano a uno, la evidencia favorece la hipótesis nula (las dos varianzas de la población son iguales). Pero si *F* es mucho mayor que uno, entonces la evidencia es contraria a la hipótesis nula. En esencia, nos preguntamos si el valor calculado del estadístico de prueba F es significativamente diferente de uno.

Para determinar los puntos críticos tenemos que calcular $F_{\alpha, df1, df2}$. Consulte la tabla F en el Apéndice A. Esta tabla F tiene valores para varios niveles de significación de 0,1 a 0,001, designados como "p" en la primera columna. Elija el nivel de significación deseado y siga hacia abajo y a través para encontrar el valor crítico en la intersección de los dos grados de libertad diferentes. La distribución F tiene dos grados de libertad diferentes, uno asociado al numerador, $df1$, y otro asociado al denominador, $df2$. Para complicar las cosas, la distribución F no es simétrica y cambia el grado de asimetría a medida que cambian los grados de libertad. Los grados de libertad en el numerador son n_1-1, donde n_1 es el tamaño de la muestra del grupo 1, y los grados de libertad en el denominador son n_2-1, donde n_2 es el tamaño de la muestra del grupo 2. $F_{\alpha, df1, df2}$ dará el valor crítico en el extremo **superior** de la distribución F.

Para calcular el valor crítico para el extremo **inferior** de la distribución, invierta los grados de libertad y divida el valor F de la tabla entre el número uno.

Valor crítico superior de la cola: $F_{\alpha,df1,df2}$

Valor crítico inferior de la cola: $1/F_{\alpha,df2,df1}$

Cuando el valor calculado de F está entre los valores críticos, no en la cola, no podemos rechazar la hipótesis nula de que las dos varianzas proceden de una población con la misma varianza. Si el valor F calculado está en cualquiera de las dos colas, no podemos aceptar la hipótesis nula, tal y como hemos hecho en todas las pruebas de hipótesis anteriores.

Una forma alternativa de calcular los valores críticos de la distribución F facilita el uso de la tabla F. Observamos en la tabla F que todos los valores de F son mayores que uno, por lo que el valor crítico de F para la cola de la izquierda siempre será menor que uno, porque para calcular el valor crítico en la cola de la izquierda dividimos un valor de F entre el número uno, como se muestra arriba.

También observamos que si la varianza de la muestra en el numerador del estadístico de prueba es mayor que la varianza de la muestra en el denominador, el valor F resultante será mayor que uno. El método abreviado para esta prueba consiste en asegurarse de que la mayor de las dos varianzas de la muestra se coloque en el numerador para calcular el estadístico de prueba. Esto significará que solo habrá que calcular el valor crítico de la cola derecha en la tabla F.

BIBLIOGRAFIA

Álvarez, D. (noviembre, 2014). Las Redes sociales y las "tecnologías del yo" de Foucault. Recuperado de http://sociologiayredessociales.com/2014/11/las-redessociales-y-las-tecnologias-del-yo-de-foucault/

Devore, Jay L. Probabilidad y estadística para ingeniería y ciencias. Internacional Thompson

Hildebrand, David K. & Ott, Lyman R. Estadística aplicada a la administración y la economía. Addison-Wesley Iberoamericana

Probabilidad y estadística de George Canavos Estadística de Murray R. Spiegel