Francisco Carruitero Lecca
Tula Benites Vásquez
Angel Hospinal Alvarez

Estadística para la Ciencia Jurídica

FONDO EDITORIAL DE LA UNIVERSIDAD PRIVADA ANTENOR ORREGO

FRANCISCO CARRUITERO LECCA
TULA BENITES VÁSQUEZ
ANGEL HOSPITAL ALVAREZ

Estadística para la Ciencia Jurídica

FONDO EDITORIAL DE LA UNIVERSIDAD PRIVADA ANTENOR ORREGO

Estadística para la Ciencia Jurídica

Editado por:
© UNIVERSIDAD PRIVADA ANTENOR ORREGO
Av. América Sur N° 3145,
Urb Monserrate, Trujillo, Perú
Teléfono (51) 44 604444, anexo 2087
wwww.upao.edu.pe

Primera edición digital del Fondo Editorial UPAO, Setiembre 2021
ISBN N° 978-612-4479-25-0

Libro revisado por pares externos.

ÍNDICE

PRESENTACIÓN

Este libro denominado *Estadística para la Ciencia Jurídica* tiene por finalidad mostrar el campo de la estadística como una asociación de procedimientos para recolectar, establecer mediciones, clasificar, aplicar modelos informáticos, establecer promedios, conocer desviaciones estándar, probabilidades, prueba de hipótesis, correlaciones y regresiones de los datos jurídicos obtenidos sistemáticamente.

Los abogados en su actividad profesional encuentran constantemente un cálculo de probabilidades sea para enfrentar un proceso judicial o cuando postulan para el acceso a un determinado centro laboral; es decir, las probabilidades permiten tomar decisiones importantes.

El conocimiento de la estadística es necesario para la investigación cuantitativa y cualitativa, el cual en el ámbito del derecho se denomina empírica o investigación jurídica social. Hoy en día es imposible llevar a cabo este tipo de investigación si no se conoce las herramientas estadísticas e informáticas.

En muchas universidades en América Latina se exigen tesis cuantitativas y cualitativas en derecho, pero lo contradictorio es que en las mallas curriculares no registran materias imprescindibles, como la matemática, estadística e informática. Por ejemplo, la variable es un concepto matemático que todo investigador jurídico debe conocer pues esta nos permite medir. Veamos rápidamente algunos tipos de variables: La variable discreta que es aquella que solo puede tomar valores dentro de un conjunto finito, como los números naturales. La variable continua que toma valores en uno o varios intervalos de la recta real. La variable proposicional que puede ser verdadera o falsa. La variable estadística mide diferentes sujetos y puede adoptar diferentes valores.

Por otro lado, en las investigaciones empíricas llamadas en del derecho socio jurídicas el cual no dejan de llevar un modelo estadístico, los valores de las variables dependientes dependen necesariamente de los valores de la variable independiente. Ciertamente, los variables dependientes vienen a representar el resultado de cuya variación se está investigando y la variable independiente son las causas de la variación. La variación es la razón de ser de la estadística. En un experimento cuando se manipula una variable independiente se prueban los efectos que estas tienen sobre las variables dependientes.

Ahora bien, en el ámbito de la investigación jurídica cuantitativa y cualitativa, la estadística permite contar con técnicas e instrumentos útiles para cuantificar con precisión los hechos

con relevancia jurídica tanto materiales como humanos de la estructura social.

La estadística, es la ciencia que aplica las leyes de la cantidad a los hechos sociales, mide su intensidad, deduce las leyes que los rigen y fundamentalmente predice: En el ámbito jurídico tiene una amplia importancia, pues ordena, resume y procesa los datos que son parte de una población, con la finalidad de comprender de manera didáctica, su contenido y características.

En este sentido, la estadística, aplicada al ámbito jurídico, presenta la información compleja, amplia, diversa y ordena a través de la presentación de gráficos obtenidos de los procesadores estadísticos, como el IBM SPSS o el Minitab que es el software que hemos utilizado en el presente libro.

Nuestra pretensión, es que este libro sea de utilidad a los investigadores, toda vez, que en los estudios de posgrado, maestría y doctorado, la investigación es una tarea principal en el cual el rol de la estadística es fundamental y contribuye a conseguir los objetivos de la investigación jurídica. En esta misma línea de pensamiento todo investigador tiene la obligación de comprender los conceptos y técnicas estadísticas para recolectar y analizar los datos y de esta manera construir instrumentos especializados para la investigación de los hechos y fenómenos jurídicos.

Este libro contiene doce capítulos:

Primer capítulo: Fundamentos de la ciencia e investigación jurídica.

Segundo capítulo: La estadística para la ciencia jurídica.

Tercer capítulo: Tablas y gráficas.

Cuarto capítulo: Estadística descriptiva

Quinto capítulo: Medidas de dispersión.

Sexto capítulo: Teoría de las probabilidades

Sétimo capítulo: Distribuciones

Octavo capítulo: La muestra en la investigación jurídica.

Noveno capítulo: Prueba de hipótesis.

Décimo capítulo: Intervalos de confianza.

Décimo primer capítulo: Correlación y regresión lineal simple.

Décimo segundo capítulo: Regresión lineal múltiple.

CAPÍTULO I

FUNDAMENTOS DE LA CIENCIA E INVESTIGACIÓN JURÍDICA

1.1 INTRODUCCIÓN

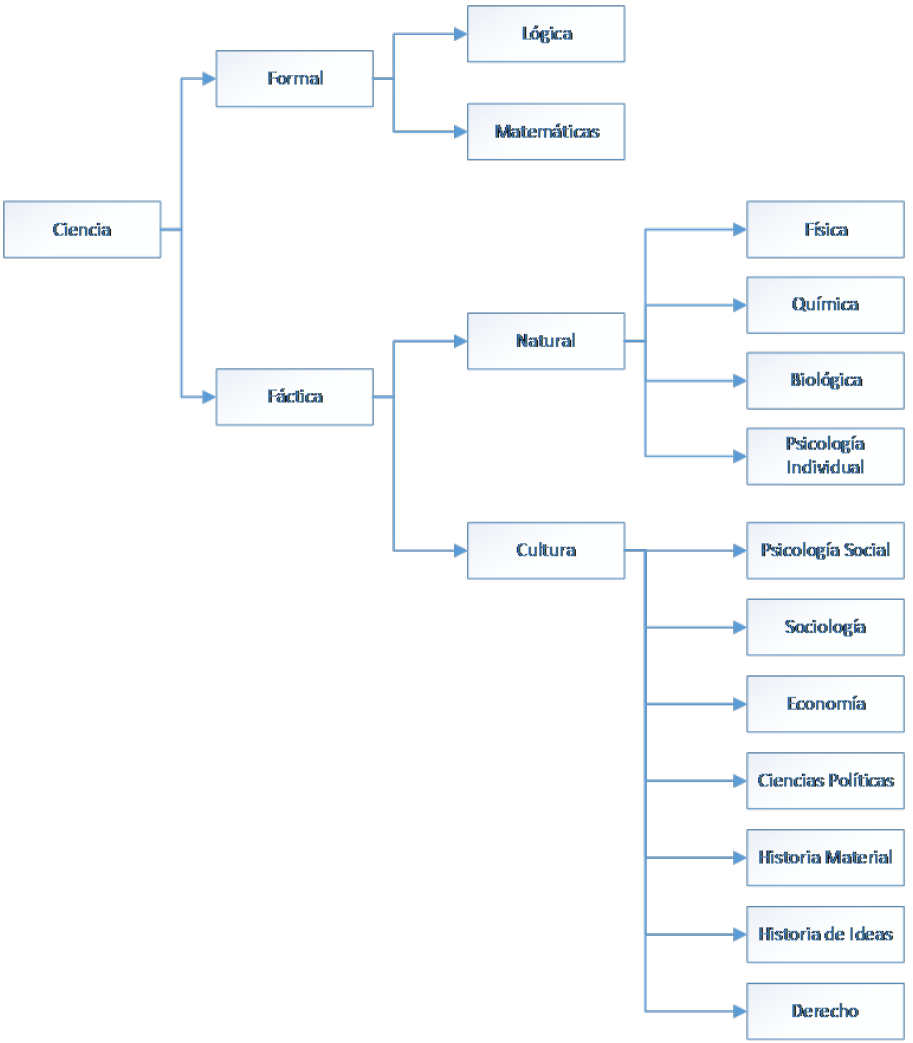
El capítulo I, denominado F*undamentos de la ciencia e investigación jurídica* trata de los siguientes temas: la ciencia, características de la ciencia, la ciencia jurídica, la investigación científica, la investigación jurídica, el problema vulgar en la investigación jurídica, el problema científico en la investigación jurídica, tipos de investigación científica: a) Investigaciones pura o básica, aplicada, profesional; b) Investigaciones exploratoria, descriptiva, correlacional y explicativa; c) Investigaciones teórica y empírica; d) Investigaciones cualitativa y cuantitativa; y e) Investigaciones primaria y bibliográfica.

1.2 LA CIENCIA

Una de las definiciones más importantes de la ciencia es la de acumulación sistemática de conocimiento obtenidos mediante el método científico. El marco de referencia colectivo que el científico jurídico social debe tener presente en la realización de su investigación se articula en dos momentos: la estructura lógica del itinerario de la investigación y la instrumentación técnica a utilizar.

Así, Blanco y Villapando (2012, p. 9) sostienen que "según su naturaleza sustantiva, la investigación científica tiene como finalidad realizar aportaciones a la ciencia en sus diferentes áreas del conocimiento. La ciencia se puede definir una en sentido estricto, como un conjunto sistemático de conocimientos sobre la realidad observable, obtenidos mediante el método de investigación científica. Según esta definición, la ciencia se analiza en cuanto a su contenido y a los elementos que configuran su naturaleza, un campo de actuación y un procedimiento con forma de actuar. Está constituida exclusivamente por un conjunto de conocimientos sobre la realidad, que forma el concepto y enunciados. Las ideas desde conjunto se hallan interrelacionadas entre sí o sistematizada y forman lo que se llama la teoría. El campo de actuación propio y único de la ciencia es la realidad observable, las realidades del mundo en que vivimos. Lo no empírico, digamos lo trascendente, que fuera del campo de la ciencia en sentido estricto. La ciencia utiliza el método de investigación científico, que es lo que la tipifica como procedimiento con forma de actuación en la formación de conocimientos que la integran."

Por su parte Bunge (1972, pp. 11 y ss) se refiere al objeto de estudio de cada ciencia, para él existe la ciencia formal y ciencia fáctica, pues afirma que no toda investigación científica procura el conocimiento objetivo. Así la lógica y la matemática, son racionales, sistemáticas y verificables, pero no objetivas, no nos dan informaciones acerca de la realidad. En las ciencias fácticas, la situación es diferente porque ellas no emplean símbolos vacíos sino tan sólo símbolos interpretados, por eso que al conocimiento factico se le llama ciencia empírica. Veamos el siguiente gráfico:



1.3 CARACTERÍSTICAS DE LA CIENCIA

Las características de la ciencia son:

- Es analítica pues descompone el todo en sus elementos.
- Es explicativa porque explica los hechos en términos de leyes y las leyes en términos de principios. Sus explicaciones son: causal, morfológica, cinemática, dinámica, de composición, de asociación, de tendencias globales, dialéctica, teológica, etc.
- Es abierta porque no reconoce barrera a priori que limite el conocimiento.
- Es útil pues busca la verdad y es eficaz en la provisión de herramientas para la acción.

1.4 LA CIENCIA JURÍDICA

Para Bolívar (2001, pp. 47-48) el derecho es una ciencia por el hecho de ser una disciplina autónoma e independiente, con una estructura sistemática y teórica propia. La autonomía de la ciencia jurídica se refiere a su independencia. Tiene el status científico porque ha alcanzado un nivel especializado, con un objeto de conocimiento autónomo respecto de otras disciplinas sociales. La ciencia jurídica no es una ciencia especulativa, sino más bien es una ciencia fáctica que se valida a partir de la contrastación de sus enunciados con la realidad.

Asimismo, Gómez (2013, p. 57) manifiesta que el derecho es el estudio de la acción humana denominada jurídica, con una larga tradición académica de cerca de dos milenios y medio, que se inicia en el mundo griego, y se viene transmitiendo hasta la actualidad y que requiere de propuestas y explicaciones en medio de un desarrollo cuantitativo y cualitativo, en sus actividades, su institucionalización y su profesionalización.

Es importante señalar también que en la ciencia jurídica los métodos de las ciencias naturales no son compatibles con ella, motivo por el cual es necesario utilizar métodos adecuados.

1.5 LA INVESTIGACIÓN CIENTÍFICA

La UNESCO (2011, p. 5) define a la investigación como el estudio ordenado de una materia con el fin de sumar conocimientos a aquellos ya adquiridos. La investigación puede implicar que la materia ya se conoce, pero que debido a una razón u otra, debe ser estudiada una vez más. También puede referirse a la investigación de un nuevo problema o fenómeno.

Como bien afirma Ander-Egg (1987, p. 43) la investigación científica busca un conocimiento general, amplio y profundo de la realidad aplicando el llamado método científico. Este último se identifica porque es fáctico, trasciende los hechos, se atiene a reglas, utiliza la verificación empírica, es autocorrectivo y progresivo, presenta formulaciones generales, y es objetivo.

1.6 LA INVESTIGACIÓN JURÍDICA

Para Chuliá y Agulló (2012, pp. 13-14) una investigación jurídica es un trabajo de investigación académica. Pues, comparte características principales con trabajos de investigación en otras disciplinas. Los principales temas son los siguientes:

- La investigación jurídica tiene características específicas en relación a investigaciones en otras disciplinas.
- La investigación jurídica puede clasificarse en dos categorías: elementos empíricos y teoría jurídica.
- La investigación jurídica-empírica tiene como finalidad describir y explicar los fenómenos.
- La investigación en teoría jurídica tiene como propósito la exposición y el análisis crítico de los principios normativos y los valores éticos que sustentan la teoría jurídica.

1.7 EL PROBLEMA VULGAR EN LA INVESTIGACIÓN JURÍDICA

Existen muchas formas de estudiar y hablar acerca de la vida jurídica, a veces es difícil distinguir cuál de ellas es investigación jurídica y cual no es. Para Mac Donald (1972, p.3), el punto de partida de la investigación es siempre un problema vulgar para distinguirlo del problema planteado en términos de una teoría.

> Enseña Mac Donald (1972, p.3) que "el problema vulgar puede ser de índole diferente: puede ser un problema muy práctico o puede ser un problema de conocimiento teórico acerca de un fenómeno determinado que no implica necesariamente consecuencias prácticas inmediatas; puede ser muy simple o muy complejo. Puede ser que solamente sea necesario describir el fenómeno o que haya que explicarlo. El tipo de problema vulgar tiene influencia sobre el planteamiento de las posibles maneras de resolverlo. En ello la investigación tiene un rol muy importante ya que sirve para evaluar empíricamente tales formas de solución además de verificar y dar conocimientos nuevos."

Mac Donald (1972, p.3), enseña que el problema vulgar debe ser reformulado en términos de una teoría científica. Esto significa, que hay que buscar los términos más adecuados para para poder captar el problema: "hay que utilizar aquí el conocimiento teórico que se tenga y hay que desarrollar y/o elaborar teoría. Esto depende mucho de la habilidad del investigador y de la madurez de la ciencia en cuanto a las teorías confirmadas de que disponga. Hay que cuidarse de no caer en el cientifismo, es decir, de complicar el aparato teórico sin necesidad. Cuanto más simple es el aparato teórico, mejor. En la mayoría de los casos será posible para el investigador escoger entre varias posibilidades teóricas en términos de las cuales pueda formular el problema. De todas maneras en cualquier caso, el investigador, tiene que justificar y fundamentar su preferencia. La formulación del problema en términos de una teoría se conoce como el planteo del problema." (Mac Donald, 1972, p.3).

Así, pues la investigación jurídica en concreto se inicia con la determinación del problema a investigar. El planteamiento del problema es la delimitación clara y precisa del objeto de investigación, realizada por medio de preguntas, lecturas, encuestas, entrevistas, etc., un problema de investigación bien planteado permite una buena investigación, por el contrario un problema de investigación mal planteado va a dificultar la elaboración de inferencias descriptivas o casuales.

No existe método de investigación que informe cómo tener nuevas ideas de investigación, ni cómo plantear bien un problema de investigación que dé origen, por ejemplo, a una tesis en ciencia jurídica que sea brillante. En la primera etapa de investigación jurídica hay un componente básico de creación personal y de imaginación, que no se rige por ningún método riguroso.

Por su parte, Anduiza, et al (2000, p.14), sostiene que la relevancia de un problema de investigación puede ser evaluada de acuerdo con algunos criterios: conviene que el tema elegido sea de interés personal del investigador. Las experiencias son una fuente de inspiración notable a la hora de elaborar proyectos de investigación.

En el mismo sentido Buendía y otros (1988, p. 10), afirman que un problema de investigación expresa lo que el investigador quiere hacer. Los problemas de investigación vienen expresados en forma de pregunta.

Ahora presentamos los siguientes ejemplos:

¿Qué piensan los magistrados de la Corte Superior de Justicia de La Libertad acerca de la reforma del sistema judicial? (Investigación por encuesta)

¿El aprendizaje a través del método reflexivo en grupo produce mejores resultados en la calidad del aprendizaje de los alumnos de la Facultad de Derecho de la Universidad Privada Antenor Orrego de Trujillo? (Investigación experimental)

¿Qué sucede en un aula del segundo año de la Facultad de Derecho de la Universidad Privada Antenor Orrego durante un semana del curso de Introducción a la ciencia jurídica? (Investigación observacional)

¿Cómo podemos predecir qué estudiantes de ciencia jurídica utilizarán enfoques de aprendizaje distintos según el tipo de materias? (Investigación correlacional).

¿Existe algún tipo de interacción profesor-alumno diferencial según el género de alumnos y los profesores de la Facultad de Derecho de la Universidad Privada Antenor Orrego (Investigación causal-comparativa).

1.8 EL PROBLEMA CIENTÍFICO EN LA INVESTIGACIÓN JURÍDICA

El proceso de investigación científica permite organizar las ideas en teoría, para realizar predicciones empíricas que colaboren con la teoría para reunir datos que puedan probar las predicciones. Todo proceso de investigación científica. Enseña Ritchey (2008, pp. 13-14) que este proceso comprende siete pasos:

1. Especifique la pregunta de investigación: planteamos una pregunta e identificamos la variable dependiente. Por ejemplo, podemos preguntar ¿por qué están ocurriendo disturbios en la ciudad de Trujillo?
2. Revise la literatura científica: hacemos esto para no desperdiciar tiempo y dinero recolectando datos que ya existen. Buscamos las fronteras del conocimiento, los límites exteriores de lo que ya ha sido aprendido, por ejemplo sobre los disturbios en la ciudad de Lima. La investigación bien informada y publicable extiende el conocimiento más allá de las fronteras.
3. Proponga una teoría y formule una hipótesis: la teoría involucra la organización de ideas en una forma lógica que pueda explicar la variación en la variable dependiente. Al desarrollar una teoría, identificamos las variables independientes y hacemos declaraciones predecibles respecto de cómo pensamos que afecta la variable dependiente, asumiendo que la teoría es comprensible. Las hipótesis generan y motivan por la teoría, ideas probadas que han sido encontradas en la literatura científica con motivaciones innovadoras. La teoría nos lleva a esperar ciertos resultados observados de los datos. Si estos datos se presentan, la teoría se corrobora.

   Por ejemplo, la teoría de los disturbios en la ciudad de Trujillo motiva la siguiente hipótesis:

   > H1: Las ciudades con alta incidencia de brutalidad policiaca (variable independiente) están sujetas a tener una elevada incidencia de desórdenes civiles (variable dependiente).

   En contraste, la teoría de la conspiración comunista en la conducta de los disturbios da origen a la siguiente hipótesis:

   > H2 Las ciudades con un gran número de comunistas (variable independiente) están sujetas a tener una elevada incidencia de desórdenes civiles (variable dependiente).

4. Seleccione un diseño de investigación: en el diseño de investigación como se medirán, muestrearan y reunirán los datos. Los métodos comunes para la investigación jurídica incluyen la observación directa del comportamiento, el experimento de laboratorio, la encuesta, el análisis de contenido de las comunicaciones y el análisis de datos existentes o secundarios como en los expedientes judiciales, censos, informes jurídicos, atestados policiales).
5. Recolecte datos: esta es normalmente la parte más costosa de la investigación. Se trata de entrar en el campo para informar a las personas sobre el estudio y recolectar datos utilizado en el diseño de investigación. También es una de las partes más agradables de la investigación, pues permite al investigador salir de la oficina y conocer nuevas personas.

6. Analice los datos y saque conclusiones: es aquí donde entra el análisis estadístico. Las hipótesis se prueban mediante las observaciones con predicciones teóricas. En el ejemplo, sobre los disturbios en la ciudad de Trujillo, los datos recolectados por la Policía Nacional del Perú del área de seguridad ciudadana, apoyaron la hipótesis I y no la hipótesis 2, otorgándole mayor credibilidad a la teoría de la protesta.
7. Difunda los resultados: difundir significa diseminar ampliamente y compartir. Los hallazgos científicos se comparten con el público en general y la comunidad científica.

1.9 TIPOS DE INVESTIGACIÓN CIENTÍFICA

Existen muchas clasificaciones que tiene por objeto construir una taxonomía de la investigación científica. Veamos algunas de ellas basadas en el objetivo de la investigación:

a. Investigaciones pura o básica, aplicada, profesional;
b. Investigaciones exploratoria, descriptiva, correlacional y explicativa;
c. Investigaciones teórica y empírica;
d. Investigaciones cualitativa y cuantitativa.

1.9.1 Investigaciones pura o básica, aplicada y profesional

De acuerdo a los propósitos inmediatos que se persiguen con la investigación, ésta se ha dividido en tres formas: la investigación pura o básica, la aplicada y la profesional. La investigación básica o fundamental es aquella que plantea teoría y por su parte la investigación aplicada confronta la teoría con la realidad. Y dentro de la investigación aplicada confronta la teoría con la realidad.

Enseña Cazau (2006, p. 17) que la investigación científica pura tiene como finalidad ampliar y profundizar el conocimiento de la realidad; la investigación científica aplicada se propone transformar ese conocimiento 'puro' en conocimiento utilizable; la investigación profesional suele emplear ambos tipos de conocimiento para intervenir en la realidad y resolver un problema puntual. Lo que habitualmente se llama investigación científica engloba solamente las dos primeras, en la medida en que ellas buscan obtener un conocimiento general, y no meramente casuístico, ya que la investigación pura (o básica) busca ampliar y profundizar el conocimiento de la realidad.

Para entender con claridad estos tipos de investigación Cazau (2006, p. 17) nos presenta los siguientes ejemplos:

Ejemplo 1) en psicología, la investigación pura investiga el mecanismo de la proyección. La investigación aplicada busca, utilizando como marco teórico el conocimiento puro, un saber general que pueda utilizarse prácticamente. Por ejemplo, investigar alguna técnica de diagnóstico sobre la base del concepto freudiano de proyección, como podría ser un test proyectivo. Finalmente, la investigación profesional consiste en intervenir en la realidad. Por ejemplo,

diagnosticar una situación usando la técnica proyectiva descubierta y validada en la investigación aplicada.

Ejemplo 2) un bioquímico estudia en su laboratorio la estructura molecular de ciertas sustancias (investigación pura); luego, otro investigador utiliza este conocimiento para probar la eficacia de ciertas sustancias como medicamentos (investigación aplicada); finalmente, el profesional hará un estudio para determinar si a su paciente puede o no administrarle el medicamento descubierto (investigación profesional).

Ejemplo 3) los matemáticos desarrollan una teoría de la probabilidad y el azar (investigación pura); luego, sobre esta base, los especialistas en diseño experimental y en estadística investigan diversos tipos de diseños experimentales y pruebas estadísticas (investigación aplicada); finalmente, un investigador indagará la forma de utilizar o adaptar estos diseños y pruebas a la investigación concreta que en ese momento esté realizando (investigación profesional).

Ejemplo 4) los antropólogos estudian el problema de la transculturación y su influencia en el aprendizaje escolar de niños que llegan a una nueva cultura (tales problemas suelen provenir, por ejemplo, cuando un maestro argentino enseña a un niño que viene de una cultura diferente, donde no coincide la lógica de la enseñanza con la lógica de apropiación del conocimiento idiosincrásica del niño, o sea, no coinciden la forma de enseñar con la de aprender). Un investigador aplicado diseña técnicas especiales para enseñar a estos niños facilitándoles el aprendizaje, y un profesional investiga la forma de aplicarlas a ese caso particular.

1.9.2 Investigaciones: exploratoria, descriptiva, correlacional y explicativa

Para Cazau (2006, p. 25), según su alcance, las investigaciones pueden ser exploratorias, descriptivas, correlacionales o explicativas. Estos tipos de investigación suelen ser las etapas cronológicas de todo estudio científico, y cada una tiene una finalidad diferente: primero se explora un tema para conocerlo mejor, luego se describen las variables involucradas, después se correlacionan las variables entre sí para obtener predicciones rudimentarias, y finalmente se intenta explicar la influencia de unas variables sobre otras en términos de causalidad.

1.9.2.1 Investigación exploratoria

El primer nivel de conocimiento científico sobre un problema de investigación se logra a través de estudios de tipo exploratorio; tienen por objetivo, la formulación de un problema para posibilitar una investigación más precisa o el desarrollo de una hipótesis. Permite al investigador formular hipótesis de primero y segundo grados.

Para definir este nivel, debe responder a algunas preguntas: ¿El estudio que propone tiene pocos antecedentes en cuanto a su modelo teórico o a su aplicación práctica? ¿Nunca se han realizado otros estudios sobre el tema? Busca hacer una recopilación de tipo teórico por la ausencia de un modelo específico referido a su problema de investigación? ¿Considera que su trabajo podría servir de base para

la realización de nuevas investigaciones? (Gestiopolis, 2019).

Ejemplos de investigación exploratoria:

¿Qué bibliografía existe y se encuentra sobre el derecho penal?

¿Cómo visualiza la Facultad de Derecho de la Universidad Privada Antenor Orrego el nuevo sistema de evaluación por competencias, qué innovaciones traerá, qué mejorará en comparación con la evaluación por objetivos?

¿Cuáles son los últimos avances en materia de litigación oral?

¿Cuáles son las últimas investigaciones que contribuyen al aprendizaje del derecho constitucional?

1.9.2.2 Investigación descriptiva

Para la Universidad Nacional Autónoma de México (2018) en la investigación descriptiva, se describe las características más importantes de un determinado objeto de estudio con respecto a su aparición y comportamiento, o simplemente el investigador buscará describir las maneras o formas en que éste se parece o diferencia de él mismo en otra situación o contexto dado. Los estudios descriptivos también proporcionan información para el planteamiento de nuevas investigaciones y para desarrollar formas más adecuadas de enfrentarse a ellas. De esta aproximación, al igual que de la del estudio exploratorio, tampoco se pueden obtener conclusiones generales, ni explicaciones, sino más bien descripciones del comportamiento de un fenómeno dado.

Ejemplo de investigación descriptiva

La investigación descriptiva proporciona información acerca de las condiciones, situaciones y sucesos que ocurren en el presente. Por ejemplo, un estudio sobre las condiciones de aprendizaje del curso de filosofía del derecho de los alumnos de la Facultad de Derecho de la Universidad Privada Antenor Orrego.

Al respecto Hernández, et. al (2014, p. 92) cuando se refieren a la investigación descriptiva toma como ejemplo a un censo nacional de población. Así el censo nacional es un estudio descriptivo cuyo propósito es medir una serie de conceptos en un país y momento específicos: aspectos de la vivienda (tamaño en metros cuadrados, número de pisos y habitaciones, si cuenta o no con energía eléctrica y agua entubada, tipo de techo y piso, combustible utilizado, tenencia o propiedad de la vivienda, ubicación de la misma, etc.), información sobre los ocupantes (número, medios de comunicación de que disponen y edad, género, bienes, ingreso, alimentación, lugar de nacimiento, idioma o lengua, religión, nivel de estudios, ocupación de cada persona) y otras dimensiones que se juzguen relevantes para el censo. En este caso, el investigador elige una serie de conceptos a considerar, que también se denominarán variables, los mide y los resultados le sirven para describir el fenómeno de interés (la población).

Hernández, et. al (2014, p. 92), desarrolla los siguientes ejemplos: 1. Una

investigación que determine cuál de los partidos políticos tiene más seguidores en una nación, cuántos votos ha conseguido cada uno de estos partidos en las últimas elecciones nacionales y locales, así como qué tan favorable o positiva es su imagen ante la ciudadanía.1 Observe que no nos dice los porqués (razones). 2. Una investigación que nos indicara cuántas personas asisten a psicoterapia en una comunidad específica y a qué clase de psicoterapia acuden.

Para Rojas (2015, p. 7), la investigación jurídica descriptiva, exhibe el conocimiento de la realidad jurídica tal como se presenta en una situación de espacio y de tiempo dado. Se observa y se registra, o se pregunta y se registra. Describe el fenómeno sin introducir modificaciones: tal cual. Las preguntas de rigor son: ¿Qué es?, ¿Cómo es?, ¿Dónde está?, ¿Cuándo ocurre?, ¿Cuantos individuos o casos se observan?, ¿Cuáles se observan? La expresión relacional es: "X". . . tal cual, como una foto "Y".

1.9.2.3 Investigación correlacional

La investigación correlacional entraña la búsqueda de la relación entre ciertas variables a través del uso de diversas mediciones de asociación estadística. Por ejemplo, la investigación de la relación existente entre los estilos de aprendizaje y el rendimiento académico por parte de los alumnos del tercer año de la Facultad de Derecho de la Universidad Privada Antenor Orrego de Trujillo.

Una correlación es una medida del grado en que dos variables se encuentran relacionadas. Un estudio correlacional puede intentar determinar si individuos con una puntuación alta en una variable también tiene puntuación alta en una segunda variable y si individuos con una baja puntuación en una variable también tienen baja puntuación en la segunda. Estos resultados indican una relación positiva. (Universidad de Jaén 2018).

Para Sánchez (2018), los estudios correlaciónales responden a preguntas de investigación tales como:

- ¿Conforme transcurre una psicoterapia orientada hacia el paciente, aumenta la autoestima de éste?
- ¿A mayor variedad y autonomía en el trabajo corresponde mayor motivación intrínseca respecto a las tareas laborales?
- ¿Los niños que dedican cotidianamente más tiempo a ver la televisión tienen un vocabulario más amplio que los niños que ven diariamente menos televisión?
- ¿Los campesinos que adoptan más rápidamente una innovación poseen mayor inteligencia que los campesinos que la adoptan después?

1.9.2.4 Investigación explicativa

Según el Centro Universitario Interamericano (2019), los estudios explicativos van más allá de la descripción de conceptos o fenómenos o del establecimiento de relaciones entre conceptos; están dirigidos a responder a las causas de los eventos

físicos o sociales. Como su nombre lo indica, su interés se centra en explicar por qué ocurre un fenómeno y en qué condiciones se da éste, o por qué dos o más variables están relacionadas. Ejemplo de las diferencias entre un estudio explicativo, uno descriptivo y uno correlacional. Los estudios explicativos responderían a preguntas tales como: ¿qué efectos tiene que los adolescentes peruanos- que viven en zonas urbanas y cuyo nivel socioeconómico es elevado - se expongan a videos televisivos musicales con alto contenido de sexo?, ¿a qué se deben estos efectos?, ¿qué variables mediatizan los efectos y de qué modo?, ¿por qué prefieren dichos adolescentes ver videos musicales con altos contenidos sexuales? Un estudio descriptivo solamente respondería a preguntas como ¿cuánto tiempo dedican dichos adolescentes a ver videos televisivos musicales y especialmente videos con alto contenido de sexo?, ¿en qué medida les interesa ver este tipo de videos?, en su jerarquía de preferencias por ciertos contenidos televisivos ¿qué lugar ocupan los videos musicales?

1.10 INVESTIGACIONES TEÓRICA Y EMPÍRICA

Enseña Cazau (2006, p.32) que "es posible distinguir dos actividades diferentes y complementarias en el ámbito de la investigación científica: la investigación teórica, que compara ideas entre sí, y la investigación empírica, que compara las ideas con la realidad".

1.11 INVESTIGACIONES CUALITATIVA Y CUANTITATIVA

Sierra (2009, p. 24), explica que aunque el método puede presentar diversas modalidades, especialmente en las ciencias sociales (incluye el derecho), se distingue ante todo según se centre por ejemplo en la observación de muchos casos particulares o en el estudio a fondo y globalmente cualquiera que sea su amplitud, de uno solo o unos pocos casos individuales. En el primer caso, se tiene el método cuantitativo predominantemente inductivo, que busca determinar las características externas generales de una población basándose en la observación de muchos casos individuales de la misma. En el segundo caso se trata de métodos científicos cualitativos o si se quiere fenomenológico que pretenden comprender, lo más profundamente posible, una entidad, fenómeno vital o situación determinada.

En la investigación cualitativa, se produce hallazgos, sin contar con procedimientos estadísticos ni instrumentos de cuantificación. Sus investigaciones pueden tratarse sobre los fenómenos culturales del derecho, la vida de los juristas, los movimientos sociales, algunos de sus datos pueden cuantificarse, pero la mayor fuerza de su análisis es interpretativo.

Para Strauss y Corbin (2002, p.11), la expresión investigación cualitativa produce confusión porque puede tener diferentes significados para personas diferentes. Algunos investigadores reúnen datos por medio de entrevistas y observaciones, técnicas normalmente asociadas con los métodos cualitativos. Sin embargo, los codifican de tal

manera que permiten hacerles un análisis estadístico. Lo que hacen es cuantificar los datos cualitativos. Al hablar sobre análisis cualitativo, nos referimos, no a la cuantificación de los datos cualitativos, sino al proceso no matemático de interpretación, realizado con el propósito de descubrir conceptos y relaciones en los datos brutos y luego organizarlos en un esquema explicativo teórico. Los datos pueden consistir en entrevistas y observaciones pero también pueden incluir documentos, películas, o cintas de videos.

Al respecto veamos la comparación entre investigación cuantitativa y cualitativa:

	Investigación cuantitativa	Investigación cualitativa
Planteamiento de la investigación		
Relación teórica-investigación	Estructura, fases lógicamente secuenciales. Deducción (la teoría procede a la observación)	Abierta, interactiva Inducción (la teoría surge de la observación)
Función de la teoría	Fundamental para la definición de las teorías y de las hipótesis.	Auxiliar
Conceptos	Operativos	Orientativos, abiertos, en construcción.
Relación con el ambiente	Enfoque manipulador	Enfoque naturalista
Interacción psicológica estudioso-estudiado	Observación científica, distanciada, neutral	Identificación empática con el objeto estudiado.
Interacción física estudioso-estudiado	Distancia, separación	Proximidad, contacto
Papel del sujeto estudiado	Pasivo	Activo
Recogida de datos		
Diseño de la investigación	Estructurado, cerrado, precede a la investigación	Desestructurado, abierto, construido en el curso de la investigación.
Representatividad/ inferencia	Muestra estadísticamente representativa	Casos individuales no representativos estadísticamente.
Instrumento de investigación	Uniforme para todos los sujetos. Objetivo: matriz de los datos	Varía según el interés de los sujetos. No se tiende a la estandarización.
Naturaleza de los datos	Hard, objetivos y estandarizados	Soft, ricos y profundos (profundidad frente a superficialidad)

Análisis de los datos		
Objeto del análisis	La variable (análisis por variables, impersonal)	El individuo (Análisis por sujetos)
Técnicas de las matemáticas y estadísticas	Uso intenso	Ningún uso
Resultados		
Presentación de los datos	Tablas (perspectiva relacional)	Fragmentos de entrevistas, de textos (perspectiva narrativa)
Generalizaciones	Correlaciones. Modelos causales. Leyes. Lógica de la causalidad.	Clasificaciones y tipologías. Tipos ideales. Lógica de la clasificación.
Alcance de los resultados	Se persigue generalizar (inferencia) (en último término, nomotética)	Especificidad (en último término, idiográfica)

1.12 EL MÉTODO ESTADÍSTICO

Para Reynaga (2019), el método estadístico consiste en una secuencia de procedimientos para el manejo de los datos cualitativos y cuantitativos de la investigación. Dicho manejo de datos tiene por propósito la comprobación, en una parte de la realidad, de una o varias consecuencias verificables deducidas de la hipótesis general de la investigación. Las características que adoptan los procedimientos propios del método estadístico dependen del diseño de investigación seleccionado para la comprobación de la consecuencia verificable en cuestión. El método estadístico tiene las siguientes etapas:

1. Recolección (medición)
2. Recuento (cómputo)
3. Presentación
4. Síntesis
5. Análisis

Los diseños de investigación

Según Vallejo (2002 p.1), los diseños de investigación se clasifican en dos grandes grupos de acuerdo al grado de control que tendrá el investigador sobre las variables y factores, tanto internos como externos en estudio, así, un diseño *puramente experimental* es aquel en el que el investigador tiene control total sobre todas las variables y factores en estudio; cuando esto no es posible, entonces se debe emplear un diseño *observacional*. Otra forma de clasificarlos se relaciona con el momento en que se llevará a cabo la obtención y el análisis de la información, cuando la información es captada en el pasado y analizada en el presente, se dice que el estudio es retrospectivo, pero si las variables se miden en el desarrollo de la investigación

y se analizan al concluirlo, entonces el diseño es *prospectivo*. El número de veces que se miden las variables en un estudio es otra forma de catalogar el diseño de investigación, cuando solamente se hace una medición de las variables el diseño es *transversal*. Es *longitudinal* cuando el investigador lleva a cabo un seguimiento de una cohorte de individuos en los que realiza mediciones a intervalos de tiempo definidos. Finalmente, cuando el estudio de investigación tiene por objetivo documentar las condiciones, actitudes o características de la población o poblaciones en estudio, el diseño de investigación es *descriptivo*. Por otro lado, un diseño *analítico* busca explicar la relación, generalmente causal, entre los factores en estudio.

1.13 PREGUNTAS DE REPASO

1. ¿Qué es el método estadístico?
2. ¿En qué campo se utiliza el método sintético?
3. ¿Qué diferencia existe entre el método inductivo y el deductivo?
4. ¿Cuál es el elemento principal del enfoque cuantitativo?
   a) Justificación
   b) Síntesis
   c) Hipótesis
   d) Desarrollo
5. Señalar lo correcto sobre la prueba de hipótesis:
   a) La hipótesis nula plantea la existencia de diferencias.
   b) Es un tipo de estadística descriptiva.
   c) La hipótesis alternativa plantea la no diferencia.
   d) La hipótesis nula y la alternativa pueden no ser excluyentes.
   e) Permite saber la probabilidad de equivocarse en una afirmación.

1.14 REPUESTAS

1. Es el procedimiento para manejar datos cuantitativos y cualitativos mediante técnicas de recolección, recuento, presentación, descripción y análisis. Además, permiten comprobar hipótesis y establecer relaciones de causalidad de un fenómeno.
2. Este método es utilizado en todas las ciencias experimentales, ya que mediante éste se extraen las leyes generalizadoras.
3. El método inductivo parte de los hechos para hacer inferencias de carácter general, mientras que el deductivo parte siempre de verdades generales y progresa con el razonamiento.
4. Hipótesis
5. Permite saber la probabilidad de equivocarse en una afirmación.

CAPÍTULO II

LA ESTADÍSTICA PARA LA CIENCIA JURÍDICA

2.1 INTRODUCCIÓN

El objetivo del capítulo II denominado La estadística para la ciencia jurídica, es presentar una visión de la realidad jurídica basada en el análisis estadístico. Esta visión estadística conocida como imaginación estadística en el cual se aprecia que tan usual o inusual es un evento, circunstancia o comportamiento, en relación con un conjunto mayor de eventos similares y una apreciación de las causas y consecuencias para el mismo. Así presentamos los siguientes temas: la estadística como método para examinar fenómenos jurídicos, clasificación de la estadística: la estadística descriptiva, la estadística inferencial. Problemas entre la estadística y la ciencia jurídica, la medición de los conceptos en la ciencia jurídica, la medición de las variables: variable nominal, variable ordinal, variable de intervalo, variable de proporción, variable independiente, variable dependiente, variable interviniente, variables cualitativas, cuantitativas: discretas y continúas, variabilidad y curtosis.

2.1.1 La estadística como método para examinar fenómenos jurídicos

Fue el sociólogo estadounidense Mills (1959 p. 2) quien entendió a la imaginación sociológica como un conocimiento de la relación del individuo con la sociedad y con la historia. La imaginación sociológica es el reconocimiento de que el comportamiento individual se rige en función de estructuras sociales más grandes; que la mayoría de las acciones individuales involucra apegarse a las reglas de la sociedad y no a la iniciativa personal y que, bien o mal, tales reglas se definen dentro de un contexto cultural.

Así para Mills (1959, pp. 7-12) la labor central de la imaginación social: "(…) era encontrar y articular las conexiones entre los entornos sociales de los individuos (también conocido como "medio"), con el contexto social más amplio y las fuerzas históricas en el que están inmersos. Este enfoque cuestiona un abordaje estructural funcionalista de la sociología, ya que abre nuevas posiciones para el individuo con respecto a una estructura social mayor. La función individual que reproduce las estructuras sociales es sólo una de muchas funciones posibles, y no necesariamente la más importante. Mills también escribió sobre el peligro de malestar que veía como inextricablemente incrustado en la creación y el mantenimiento de las sociedades modernas. Esto le llevó a la pregunta de si los individuos existen en las sociedades modernas, en el sentido en que "lo individual" es comúnmente entendido".

Al respecto, Ritchey (2008 p. 3) afirma que "para poseer la imaginación es entender que la mayoría de los eventos son predecibles es decir, éstos tienen una probabilidad de ocurrencia basada en tendencias y circunstancias a largo plazo. La imaginación

estadística es la habilidad para pensar a través de un problema y mantener un sentido de proporción o equilibrio cuando se pondera la evidencia contra nociones preconcebidas; es reconocer eventos muy raros por lo que son y no por la reacción ante ellos. Ser estadísticamente falto de imaginación es poner las cosas fuera de proporción, para pensar de manera reaccionaria en lugar de proporcional. La imaginación sociológica implica ver un detalle aislado –una parte– con respecto a una representación más amplia –el todo–; ver el bosque así como los árboles."

Claro está, que el campo de la estadística se funda en el conjunto de procedimientos para reunir, medir, clasificar, codificar, computar y resumir información numérica adquirida sistemáticamente. Por otro lado el error estadístico es el grado de imprecisión en los procedimientos utilizados para reunir y procesar información. Así la imaginación estadística no solo requiere un sentido de proporción acerca de la realidad jurídica, sino las también las diligencias para mantenerse al tanto los detalles para minimizar el error. (Ritchey 2008, 3).

La estadística aplicada a la ciencia jurídica busca un método para examinar los fenómenos socio-jurídicos para luego suministrar las bases para la toma de decisiones jurídica.

Enseña Ritchey (2008, p. 7), la estadística debe "(…) ser entendida como un conjunto de procedimientos para reunir, medir, clasificar, codificar, computar, analizar y resumir información numérica adquirida sistemáticamente. Ahora bien, las estadísticas y la recolección de datos no son actividades informales pero son empresas que requieren esfuerzos máximos. El análisis estadístico implica precisión, es decir, se refiere a seguir procedimientos, y realizar mediciones precisas y predicciones exactas sobré como ocurrirá los eventos en el mundo. Cuando el análisis estadístico se hace de manera apropiada el analista conoce las limitaciones del razonamiento y de los procedimientos matemáticos, y sabe cuándo las predicciones sobre eventos o conductas son menos que precisas; además, puede expresar el grado de confianza que tiene al hacer una conclusión. En cuanto a esto, el objetivo de la estadística consiste en controlar el error. Los errores estadísticos no son equivocaciones. El error estadístico se refiere al grado conocido de imprecisión en los procedimientos utilizados para reunir y procesar información. Controlar el error significa ser tan preciso como sea necesario para reforzar la confianza en las conclusiones derivadas".

La estadística es ciencia y arte que busca dar sentido a los datos proporcionando la teoría y los métodos para extraer información de estos y resolver problemas del mundo real. En el campo del Derecho, la estadística es usada para apoyar la toma de decisiones de los operadores jurídicos dentro del sistema jurídico.

Ciertamente, la estadística es una ciencia, pues trata de hallar regularidades en los fenómenos sean jurídicos, sociales, económicos, etc., de manera que sirvan para describir y predecir. Por ello, es la colección de métodos científicos que permiten el análisis e interpretación de la Investigación jurídica.

Claro está, que el campo de la estadística es un conjunto de procedimientos para reunir, medir, clasificar, codificar, computar, analizar y resumir información numérica adquirida sistemáticamente.

Por otro lado, un curso de estadística para los abogados suele ser percibido como aquel que involucra muchas fórmulas y cálculos. De hecho intervienen algunas operaciones matemáticas, pero no constituyen el catalizador de la estadística y por lo general las computadoras se encargan de esta parte. En realidad, la estadística implica aprender una nueva manera de ver las cosas, adquirir una visión de la realidad basada en el análisis cuidadoso de hechos, en lugar de reacciones emocionales ante experiencias aisladas. (Ritchey 2008, pp. 1-2).

En la actualidad hay una actitud poco favorable hacia la estadística porque esta se produce paradójicamente por el derecho a la transparencia, en una sociedad en el cual la información cuantitativa invade los aspectos más íntimos de nuestra vida: divorcio, querellas, derecho a la privacidad, índices de criminalidad, ingresos, gasto público, encuestas electorales, números de sentencias fundadas, infundadas, improcedentes, número de quejas a los magistrados, etc.

Además, a través de la historia, desde la óptica pagana o cristiana no había azar; todos los fenómenos obedecían a leyes divinas y no a la probabilidad. Hasta que Europa y Estados Unidos no superaran la teología y filosofía medieval no fue posible desarrollar el cálculo de las probabilidades.

Ahora bien, desde finales del siglo XIX se comienza a descubrir regularidades en disciplinas, como la genética, biología, meteorología, economía, psicología, antropología, sociología y el derecho e incluso en las artes.

La demografía ayudó a desarrollar la estadística en el Perú; el Primer Censo de Población de la Época Republicana, se levantó en 1836 durante el Gobierno del General Don Andrés de Santa Cruz, cuyo resultado indicó una población de 1'873,736 habitantes. Parte del desprestigio de la estadística es porque a veces se utilizan datos numéricos para apoyar razonamientos falsos.

En resumen, las aplicaciones de la estadística en la ciencia jurídica son muy diversas entre otras veamos los siguientes:

- Análisis de los datos y extracción de información relevante de los mismos, de las mediciones observadas en el Poder Judicial, Ministerio Público, Policía Nacional del Perú, Tribunal Constitucional, Jurisdicción Arbitral, Jurisdicción Militar, Jurisdicción Campesina, Jurado Nacional de Elecciones, Consejo Nacional de la Magistratura, operadores del derecho, etc.
- Búsqueda y evaluación de modelos y pautas que ofrecen los datos, pero que se encuentran ocultos por la inherente variabilidad aleatoria de los mismos en el sistema jurídico.

- Contribuir al diseño eficiente de experimentos y encuestas aplicados al ámbito de las instituciones jurídicas. Facilitar la comunicación entre los abogados, ya que será más fácil comprender la referencia a un procedimiento estándar sin necesidad de mayor detalle. Los datos bien organizados reducen al mínimo el error estadístico. Así, los juristas realizan predicciones empíricas para probar la exactitud de sus ideas.
- Para la elaboración de las tesis jurídicas, es de suma utilidad el conocimiento de la estadística, sobre todo el campo de las probabilidades.

Por ello, nos preguntamos ¿cuál es la probabilidad de que seas víctima de un delito de robo agravado en el Callao?, nos encontramos frente a tres tipos de predicción basados en la idea de que el riesgo de ser víctima del delito robo agravado puede reducirse por medio del estudio cuidadoso de actividades rutinarias.

Un primer factor de riesgo es la, exposición o vulnerabilidad circunstancial, es trabajar en un turno nocturno. Un segundo factor es la proximidad de trabajar en un lugar con un alto índice delictivo. Y un tercero factor desear la propiedad de una víctima. Si la persona se pone en riesgo innecesario, un robo o un asesinato no serían un suceso aleatorio o una equivocación, sería un error estadístico, el cual es un grado conocido de imprecisión en los procedimientos utilizados.

A manera de resumen, la estadística es la ciencia de la sistematización, el ordenamiento y la presentación de los datos en relación a un fenómeno que presenta variabilidad o incertidumbre para su estudio racionalizado con el objeto de deducir las leyes que rigen esos fenómenos y poder llevar a cabo previsiones y tomar decisiones y llegar a conclusiones.

2.2 CLASIFICACIÓN DE LA ESTADÍSTICA

La estadística cumple dos funciones fundamentales, que se van a definir a su vez en dos tipos de estadística:

2.2.1 La estadística descriptiva

Es el conjunto de técnicas para la reducción de datos cuantitativos y cualitativos de una población o una muestra a un número más pequeño y de lectura más simple, de modo que podamos caracterizar de forma resumida los valores adoptados por las variables de nuestro estudio. La principal característica es que las conclusiones no superan el límite del colectivo estudiado.

Así, la estadística descriptiva, es el conjunto de métodos estadísticos que se relacionan con el resumen y descripción de los datos como tablas gráficas y el análisis mediante algunos cálculos. (Córdova, 2009, p.1).

En el mismo sentido, para Mitacc (2014, p.1), la estadística descriptiva es la parte de la estadística que se encarga de la recolección, calificación, presentación, descripción

y simplificación de los datos. Resume el autor la estadística descriptiva en el siguiente diagrama:

2.1.2 La estadística inferencial

Es el conjunto de técnicas para tomar decisiones acertadas que ayuden a los investigadores jurídicos a hacer inferencias (deducciones) de las muestras a las poblaciones y, en consecuencia, a comprobar hipótesis relativas a la naturaleza de la realidad social mediante un proceso de deducción - inducción.

Se ocupa de la forma en la que se pueden obtener muestras fiables y los resultados obtenidos en ellas se pueden hacer extensibles a la población en general. La principal característica es que las conclusiones superan el límite del colectivo estudiado.

2.3 PROBLEMAS ENTRE LA ESTADÍSTICA Y LA CIENCIA JURÍDICA

Ciertamente, las dificultades que plantea el análisis estadístico del mundo jurídico se da por varios factores, por ejemplo, la medición de fenómenos jurídicos son eminentemente subjetivos.

En el análisis estadístico de los datos jurídicos se da el error numérico y la fiabilidad de la investigación. La ciencia estadística al dar el resultado de fenómenos jurídicos es acertada, si no falla el instrumento estadístico cuando es bien utilizado.

Si el problema jurídico que nos ocupa no está teóricamente bien definido, de poco servirá la utilización de un gran aparato estadístico. La estadística es siempre una buena ayuda, pero nunca un sustituto para un buen razonamiento teórico y un buen quehacer metodológico.

Los teóricos de la ciencia jurídica reconocieron la importancia de la obtención de información cuantitativa relevante sobre los fenómenos socio-jurídicos y de su tratamiento estadístico para construir una ciencia jurídica sobre la sociedad.

A finales del siglo XIX los juristas disponían de pocos datos, pero de mucho genio creador para las bases teóricas; hoy en día se dispone de un mar de datos socio-jurídicos, en el cual se hace necesario el análisis multivariable.

2.4 LA MEDICIÓN DE LOS CONCEPTOS EN LA CIENCIA JURÍDICA

La medición es la asignación de símbolos, tanto nombres como números, a las diferencias que observamos en las cualidades o cantidades de una variable. La medición de un sujeto particular de la muestra en una variable es la puntuación del sujeto para esa variable, o para usar terminología computacional.

Para medir los conceptos de la ciencia jurídica; hay que cuantificar las variables que definen el fenómeno jurídico. La medición es la fase intermedia:

1. razonamiento teórico
2. medición
3. introducción de los métodos estadísticos de investigación.

La combinación ponderada de valores que toman cada uno de los indicadores (por ejemplo, nivel de ingresos, años de escuela y ocupación) forman un índice (status socio económico), que tomará valores numéricos concretos.

El empleo de las herramientas estadísticas requiere que las variables jurídicas sean cuantificadas siguiendo el nivel de medición que las propiedades exigen.

El procedimiento de medición de las variables jurídicas se busca fijándose en dos aspectos:

- La fiabilidad. - es la propiedad del instrumento que le permite que al ser utilizado repetidas veces bajo idénticas circunstancias produzca iguales efectos.
- La validez. - es que el instrumento mida lo que realmente queremos medir.

Cualquier proceso de medición debe ser exhaustivo, con categorías suficientes en las que puedan clasificarse cada uno de los casos considerados. Las categorías deben ser mutuamente excluyentes, que debe ser posible clasificar cada caso individual tan sólo en una categoría. También debe ser lo más preciso posible que haya el mayor número de distinciones.

2.5 LA MEDICIÓN DE LAS VARIABLES

Los distintos niveles forman una escala acumulativa de tipo ascendente; el nivel ordinal posee las propiedades del nominal. Una de las metas más perseguidas por los juristas es la de obtener medidas, cuyas naturalezas admitan el nivel de medición intervalar.

Identifica las propiedades de medición de la variable y determina el tipo de operaciones matemáticas (suma, multiplicación etc.) que puede usarse apropiadamente con dicho nivel, así como las fórmulas estadísticas que utiliza para probar las hipótesis teóricas.

2.5.1 Variable nominal

Las variables nominales son aquellas en las que los códigos solo indican una diferencia en categoría, clase, calidad o tipo. La palabra nominal viene del vocablo latín para

nombre y estas variables tienen categoría de nombre, lugar de nacimiento, sabor favorito de una gaseosa, marca de una camioneta, profesión (abogado).

Así, es el nivel más bajo de medición y permite la clasificación, por ejemplo: religión, sexo, etc., sin que uno sea superior a otro. No puede teóricamente realizarse directamente operaciones matemáticas con ellas. Se sustituyen los objetos reales por números o símbolos indicando sólo la diferencia respecto a una cualidad dada, para poder realizar operaciones matemáticas.

2.5.2 Variable ordinal

Al igual que las variables nominales, las variables ordinales designan categorías, pero tienen la propiedad adicional de permitir clasificar las categorías desde la mayor a la menor, de la mejor a la peor o de la primera a la última.

Nos encontramos con un nivel que permite clasificación y orden de mayor a menor o viceversa, por ejemplo: ingresos medios según clase baja, media y alta. No ofrece ningún tipo de información sobre la magnitud de las diferencias entre las categorías, sólo que $3 > 2 > 1$.

2.5.3 Variable de Intervalo

Las variables de intervalo tienen las características de las variables nominales y además una unidad numérica de medición definida. Las variables de intervalo identifican las diferencias de monto, cantidad, grado o distancia y se les asigna puntuaciones numéricas útiles. Los ejemplos incluyen la temperatura, cuando se expresa la temperatura en grados.

Este nivel clasifica, ordena e indica la distancia entre distintas categorías. Lo característico es la existencia de una unidad de medida común y constante que permite asignar un número real a todos los pares de objetos del conjunto ordenado, por ejemplo: el coeficiente de inteligencia, grados de temperatura. En este nivel de medición el punto cero está arbitrariamente determinado y no representa ausencia completa de la característica que se mida.

2.5.4 Variable de proporción

Es similar al anterior, ya que permite clasificación, orden y distancia, pero el cero en este nivel es absoluto y representa la ausencia completa de la característica que mide, por ejemplo: el peso, la masa, el tiempo. La distinción con la anterior es puramente académica, ya que una vez establecida la magnitud de la unidad es casi siempre posible concebir 0 unidades. Si se utilizara un procedimiento estadístico poco apropiado para niveles bajos de medición con puntuaciones definidas a un nivel de medición más alto, no se cometería un error técnico, sino que se produciría una pérdida de información, dado que las propiedades de los niveles de medición son acumulativas.

2.5.5 Variable independiente

Son las que influyen en las dependientes; permiten conocer porqué varía la variable dependiente de la forma en la que lo hace.

2.5.6 Variable dependiente

Es la que atrae primordialmente la atención del investigador y cuya variación trata de explicar.

2.5.7 Variable interviniente

Se supone que tiene un efecto determinado sobre la variable dependiente que puede ser controlado por la variable independiente. Por ejemplo: estudio sobre las causas del divorcio. La situación matrimonial es la variable dependiente, que habría que explicar a partir de otras variables independientes.

2.5.8 Variables cualitativas

Estas no pueden adoptar valores numéricos.

2. 5.9. Variables cuantitativas

Son las que pueden adoptar valores numéricos y se clasifican en:

A. Variables discretas

Son valores con números enteros, número de hijos; continuas: con infinitos valores fraccionados, por ejemplo: temperaturas, la mayoría de las variables nominales son discretas. Es una variable que sólo puede tomar valores dentro de un conjunto numerable, es decir, no acepta cualquier valor sino sólo aquellos que pertenecen al conjunto. En estas variables se dan de modo inherente separaciones entre valores observables sucesivos. Dicho con más rigor, se define una variable discreta como la variable que hay entre dos valores observables (potencialmente), hay por lo menos un valor no observable (potencialmente). Como ejemplo, el número de animales en una granja (0, 1, 2, 3...).

B. Variables continuas

La variable continua puede tomar un valor cualquiera dentro de un intervalo predeterminado. Por ello, entre dos valores observables va a existir un tercer valor intermedio que también podría tomar la variable continua. Una variable continua toma valores a lo largo de una continuidad, en todo un intervalo de valores. Un atributo de una variable continua, nunca puede ser medida con exactitud; el valor observado depende en gran medida de la precisión de los instrumentos de medición. Ejemplo, la estatura de una persona (1.71m, 1.715m, 1.7154m....).

2.10. VARIABILIDAD

Para describir una variable se utiliza alguna medida de posición central y una medida de dispersión. El par de medidas usado es la media aritmética y la desviación estándar. Pero cuando la distribución de las observaciones es sesgada, la media no es una buena medida de posición central y preferimos la mediana. La mediana en general va acompañada del rango como medida de dispersión. Pero cuando observamos valores extraños (extremos) el rango se ve muy afectado, por lo que se prefiere usar el rango entre cuartiles.

2. 11. LA CURTOSIS

De una distribución de frecuencias no tiene un referente natural como en el caso de la simetría. La curtosis se sustenta en la comparación respecto a una distribución de referencia, es llamada también distribución normal o campana de Gauss.

Su obtención se basa en una distribución de frecuencias que sea similar a la de la curva normal. La curtosis señala el grado en que una distribución acumula casos en sus colas en comparación con los casos acumulados en las colas de una distribución normal cuya dispersión sea equivalente.

2. 12. PREGUNTAS DE REPASO

Definir si las siguientes variables son cualitativas o cuantitativas (discretas o continuas):

1. Cantidad de libros en una biblioteca.
2. La religión de una persona.
3. Volumen de agua en una piscina.
4. Suma de resultados obtenidos por el lanzamiento de un dado.
5. La marca de una gaseosa.

2.13 RESPUESTAS

1) Cuantitativa discreta.
2) Cualitativa.
3) Cuantitativa continúa.
4) Cuantitativa discreta.
5) Cualitativa.

CAPÍTULO III

TABLAS Y GRÁFICAS

3.1 INTRODUCCIÓN

En el capítulo III denominado *Tablas y gráficas*, presenta los siguientes temas: Resumen de grandes cantidades de información jurídica, guía para el estudiante, gráfico de pastel, gráfico de barras, gráfico de barras compuesto, gráfico radial, histograma, polígono de frecuencias, gráfico de frecuencia absoluta, gráfico de frecuencia absoluta acumulada, diagrama de Pareto, gráfico de cajas.

3.2 RESUMEN DE GRANDES CANTIDADES DE INFORMACIÓN JURÍDICA

En la actualidad en Perú, el análisis estadístico se refiere también entre otros aspectos a resumir grandes cantidades de información jurídica. Por ejemplo la media aritmética de las edades de los alumnos del tercer ciclo de la Facultad de Derecho de la Universidad Privada Antenor Orrego.

Es probable que hayan surgido muchas preguntas en tu mente cuando, el profesor de investigación jurídica te solicita que le expliques como vas a analizar y presentar los datos de tu investigación y como los vas a procesar, seguramente te vas a preguntar: ¿Cuántos son los datos de estudio? ¿Qué gráficos y cuadros estadísticos existen?

Esto nos lleva a pensar, que en nuestra vida cotidiana necesitamos simplificar y organizar nuestras percepciones. Para orientarnos a una nueva situación rápidamente buscamos nuevas generalizaciones que describen esta gran imagen. Necesitamos simplificar las generalidades de una manera apropiada y eficiente y no ser engañados.

No cabe duda que viene a nuestra mente, el concepto de proporción, describimos numéricamente la distribución de las puntuaciones de una variable con frecuencias porcentuales. Por ello los gráficos nos permiten afirmar que tiene un valor mayor a un conjunto de palabras.

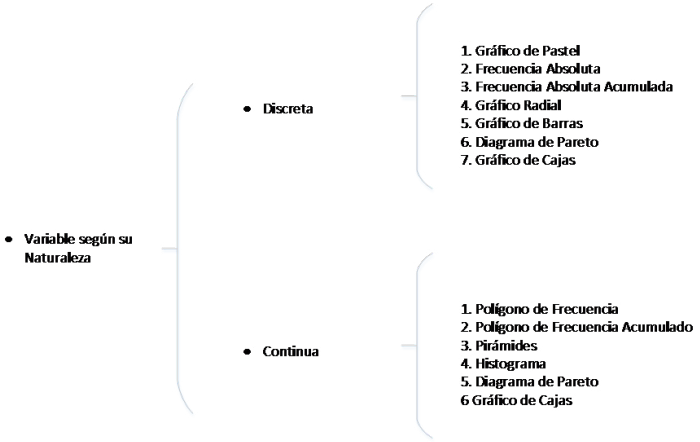
Los diseños gráficos y pictóricos se eligen con base en (I) el nivel de medición de una variable (2) los objetivos y los aspectos relevantes del estudio y (3) el público a quien se dirigen para los programas públicos los gráficos sencillos y a todo color funcionan mejor y brinda una perspectiva global de los estadísticos descriptivos, tales como porcentajes y promedios. En contraste los públicos compuestos por especialistas están acostumbrados a los estadísticos inferenciales aquellos disertados para explicar y probar hipótesis. Junto con las tablas estadísticas, los gráficos nos ayudan a discernir las formas de las distribuciones de frecuencia. Incluso los gráficos descriptivos alertan a un analista sobre fuentes de error potenciales que puedan incluir en el análisis realizado. (Ritchey 2008)

Las presentaciones gráficas deben cumplir con algunas reglas y lineamientos simples los cuales también se aplican a las tablas y a la elaboración de reportes. (Ritchey 2008).op

- Elige el diseño con base en a) el nivel de medición de una variable, b) los objetivos del estudio y c) el público a quien se dirige.
- Ante todo, una buena presentación grafica tiene que ser clara y entendible. Debe simplificar no complicar.
- Un gráfico o diagrama requiere explicarse por sí mismo y transmitir información sin hacer referencia a un texto o a alguien que lo explique. La selección cuidadosa de títulos, descripciones de la escala, subtítulos y otras leyendas contribuyen a lograr este objetivo. Somete cada grafico o tabla a la prueba de "perdido en el estacionamiento". Pregúntate si este grafico fuera abandonado en un estacionamiento ¿podría tomarlo un perfecto extraño e interpretarlo?
- Antes de decidirte sobre el tipo de presentación pictórica (por ejemplo, gráfico de pastel contra grafico de barras) elabora bosquejos con varias opciones. Los programas de cómputo hacen esto en forma relativamente fácil. Para ampliar las alternativas solicitan opiniones y consultas otros materiales, tales como informes organizacionales.
- Adhieres a los principios de inclusividad. Ahora al pie de página cualquier excepción.
- Si los datos no son tuyos indica la fuente de los mismos al final de la tabla.

3.3 GUÍA PARA EL ESTUDIANTE

Para elaborar los gráficos en el presente capítulo se empleara la clasificación de las variables según su naturaleza, las cuales pueden ser discretas o continuas. Los tipos de gráficos más comunes que se pueden emplear se listan a continuación, en algunos casos, los gráficos pueden ser empleados para ambos tipos de variables.



3.4 GRÁFICO DE PASTEL

También denominado gráfico de 360 grados, de sectores o circular, es empleado para presentar datos discretos (nominales u ordinales) y para representar porcentajes y proporciones que permiten visualizar qué parte del total representa cada categoría o sección. Se recomienda emplear este gráfico para mostrar hasta 5 categorías, debido a que si se emplea un mayor número de secciones se dificultará el entendimiento. Por otro lado, al ser una gráfica circular, se requiere conocer de manera aproximada el ángulo de la sección para interpretar el tamaño de la proporción de cada categoría, por lo que es recomendable acompañar el gráfico con una leyenda. La proporción se calcula multiplicando la frecuencia relativa por 360°, que corresponde con el total que no debe superar el 100%.

Forma Tradicional:

La Tabla 1 muestra las edades de los 48 estudiantes del 4to ciclo de la Facultad de Derecho y Ciencia Política de la UNMSM esta población será tomada como base la elaboración del gráfico pastel.

Tabla 1 : Edad de Estudiantes de la Facultad de Derecho y Ciencia Política de la UNMSM

Edad en Años de los Estudiantes de Derecho (48 alumnos)											
21	20	21	19	19	19	21	21	21	21	20	19
20	20	20	21	21	20	19	20	20	20	19	21
20	21	19	21	21	19	19	20	20	20	21	21
21	20	20	20	21	20	20	20	20	19	20	20

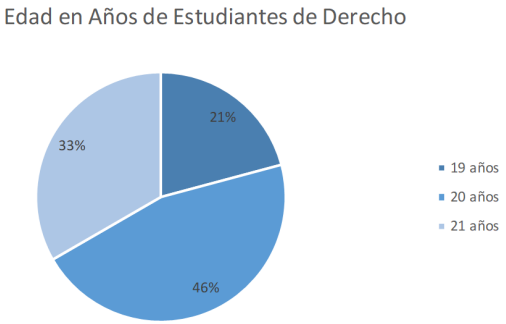
Se procede a realizar un cuadro de resumen con la información como se muestra en la Tabla 2. La frecuencia relativa se calcula de la siguiente manera:

$$Frecuencia\ Relativa = \frac{Cantidad\ de\ Estudiantes\ de\ 19\ Años}{Cantidad\ Total\ de\ Estudiantes} = \frac{10}{48}$$
$$= 0{,}21 = 21\%$$

$$Frecuencia\ Relativa = \frac{Cantidad\ de\ Estudiantes\ de\ 19\ Años}{Cantidad\ Total\ de\ Estudiantes} = \frac{22}{48}$$
$$= 0{,}46 = 46\%$$

$$Frecuencia\ Relativa = \frac{Cantidad\ de\ Estudiantes\ de\ 19\ Años}{Cantidad\ Total\ de\ Estudiantes} = \frac{16}{48}$$
$$= 0{,}33 = 33\%$$

Resumen de Edad en Años de Estudiantes de Derecho		
Edad	**Cantidad**	**Frecuencia Relativa**
19 años	10	21%
20 años	22	46%
21 años	16	33%
Total	48	100%

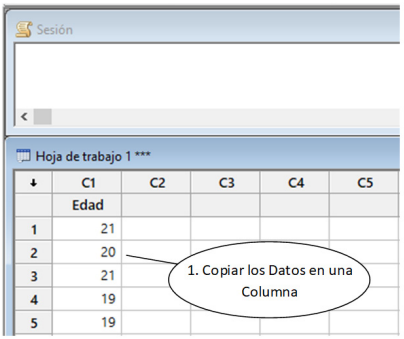


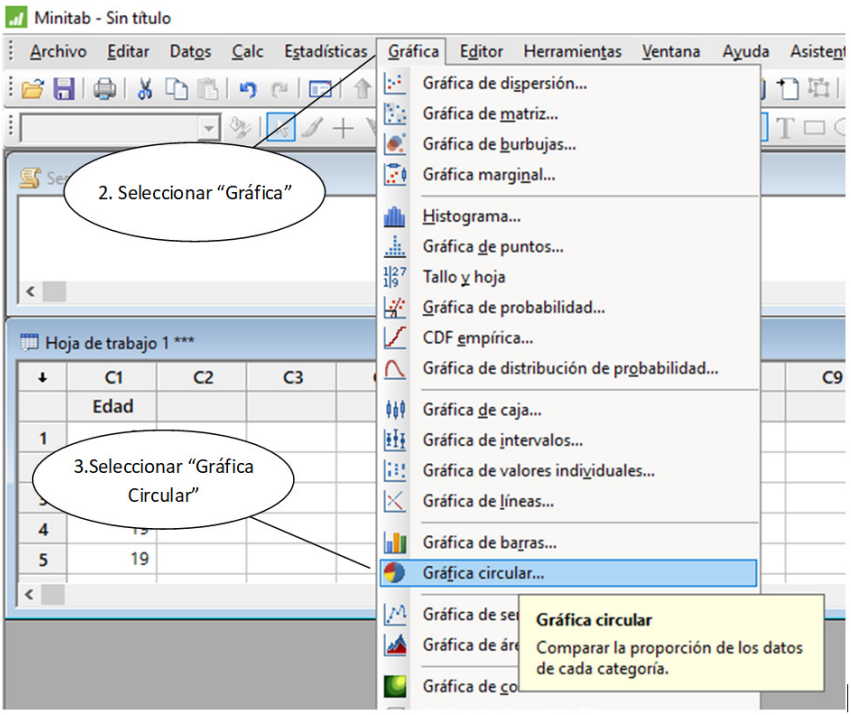
Aplicando Minitab:

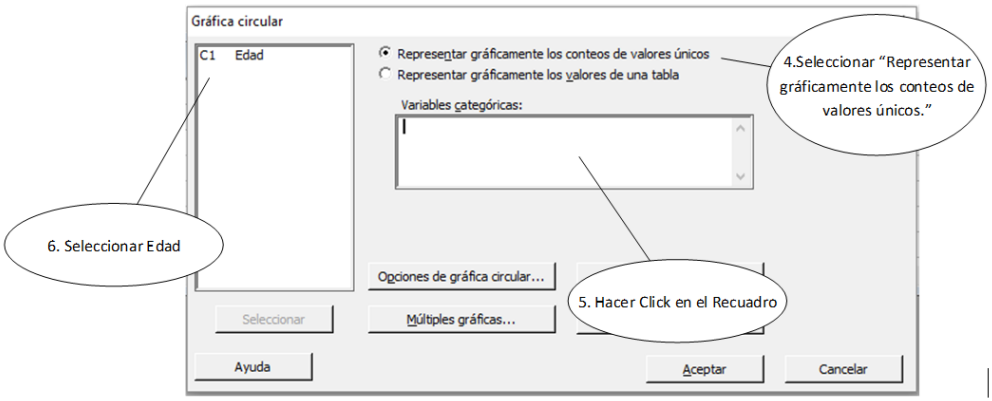
Se procede a copiar los datos en la hoja de trabajo de Minitab.

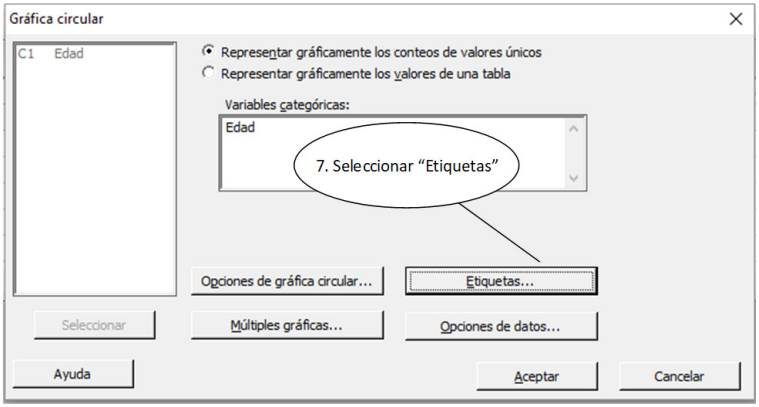
Tabla 1: Edad de Estudiantes de Derecho y Ciencia Política de la UNMSM

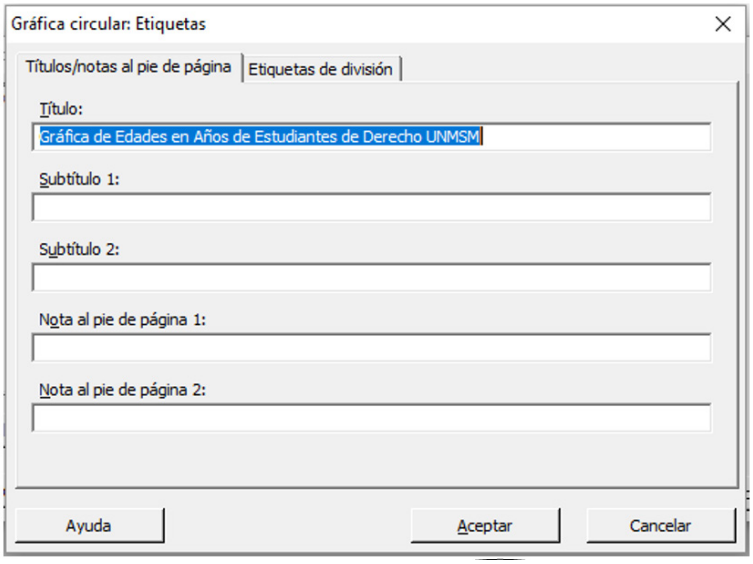
Edad en Años de los Estudiantes de Derecho (48 alumnos)											
21	20	21	19	19	19	21	21	21	21	20	19
20	20	20	21	21	20	19	20	20	20	19	21
20	21	19	21	21	19	19	20	20	20	21	21
21	20	20	20	21	20	20	20	20	19	20	20

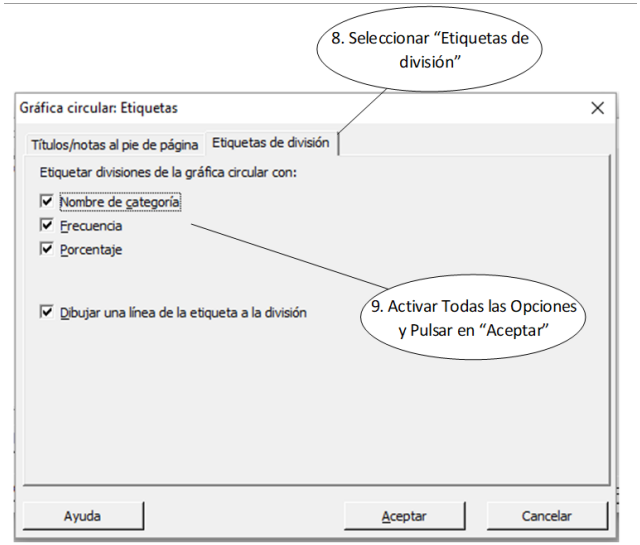


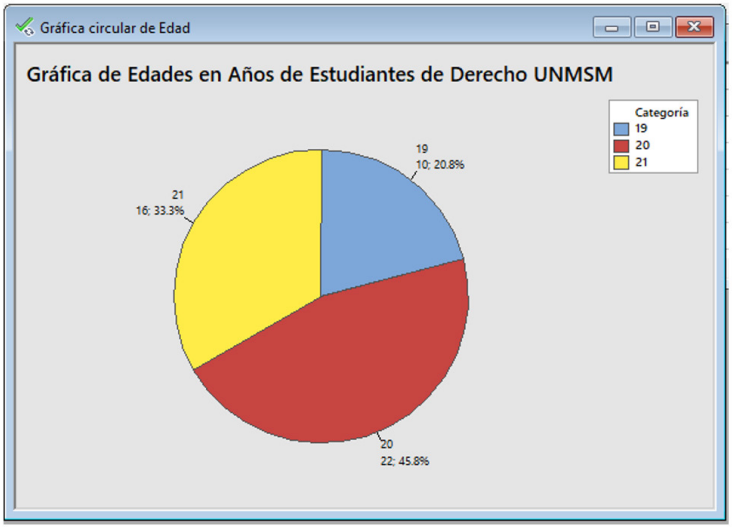

Minitab - Sin título
Archivo Editar Datos Calc Estadísticas Gráfica Editor Herramientas Ventana Ayuda Asistent
2. Seleccionar "Gráfica"
Gráfica de dispersión...
Gráfica de matriz...
Gráfica de burbujas...
Gráfica marginal...
Histograma...
Gráfica de puntos...
Tallo y hoja
Gráfica de probabilidad...
CDF empírica...
Gráfica de distribución de probabilidad...
Gráfica de caja...
Gráfica de intervalos...
Gráfica de valores individuales...
Gráfica de líneas...
Gráfica de barras...
Gráfica circular...
Hoja de trabajo 1 ***
C1 C2 C3
Edad
3.Seleccionar "Gráfica Circular"
Gráfica circular
Comparar la proporción de los datos de cada categoría.


Gráfica circular
C1 Edad
Representar gráficamente los conteos de valores únicos
Representar gráficamente los valores de una tabla
Variables categóricas:
4.Seleccionar "Representar gráficamente los conteos de valores únicos."
6. Seleccionar Edad
Opciones de gráfica circular...
Seleccionar
Múltiples gráficas...
5. Hacer Click en el Recuadro
Ayuda
Aceptar
Cancelar


Gráfica circular
C1 Edad
Representar gráficamente los conteos de valores únicos
Representar gráficamente los valores de una tabla
Variables categóricas:
Edad
7. Seleccionar “Etiquetas”
Opciones de gráfica circular...
Etiquetas...
Seleccionar
Múltiples gráficas...
Opciones de datos...
Ayuda
Aceptar
Cancelar


Gráfica circular: Etiquetas
Títulos/notas al pie de página
Etiquetas de división
Título:
Gráfica de Edades en Años de Estudiantes de Derecho UNMSM
Subtítulo 1:
Subtítulo 2:
Nota al pie de página 1:
Nota al pie de página 2:
Ayuda
Aceptar
Cancelar





Interpretación: Se puede observar que los estudiantes de Derecho de la UPAO de 20 años representan el 45% del total, los de 21 años el 33% y los de 19 años el 20%.

3.5 GRÁFICO DE BARRAS

Para presentar variables discretas (nominales u ordinales) se puede emplear el gráfico de barras. Este gráfico representa, mediante barras o rectángulos del mismo ancho, la frecuencia con la que se repite cada categoría. Es por ello, que la altura o longitud de la barra es proporcional al valor o frecuencia de la categoría de cada variable que se representa. Generalmente se emplean para comparar dos o más valores y pueden ser horizontales o verticales.

Forma Tradicional:

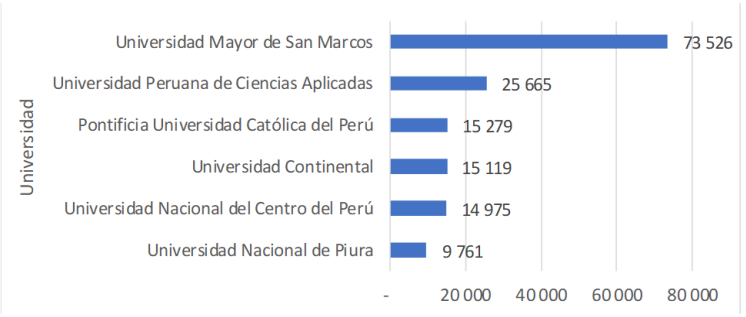
La Tabla 2 muestra la cantidad de postulantes por género en las diversas universidades del Perú tanto para pregrado, maestría y doctorado.

TABLA 2 – Cantidad de Postulantes 2016 I – Pregrado, Maestría y Doctorado.

Cantidad de Postulantes 2016 - I			
Universidad	**Femenino**	**Masculino**	**Total**
Universidad Mayor de San Marcos	36 157	37 369	73 526
Universidad Peruana de Ciencias Aplicadas	12 431	13 234	25 665
Pontificia Universidad Católica del Perú	6 790	8 489	15 279
Universidad Continental	5 829	9 290	15 119
Universidad Nacional del Centro del Perú	7 335	7 640	14 975
Universidad Nacional de Piura	5 731	4 030	9 761
Total	74 273	80 052	154 325

Fuente: https://www.sunedu.gob.pe/sibe/

Haciendo uso de los datos de la columna "Total" se iniciará el diseño del gráfico de barras. Para ello debemos definir si emplearemos el eje X o el eje Y para representar el nombre de las categorías y la frecuencia de cada una. Para este caso emplearemos el eje Y para asignar el nombre a la categoría y el eje X para expresar la frecuencia.

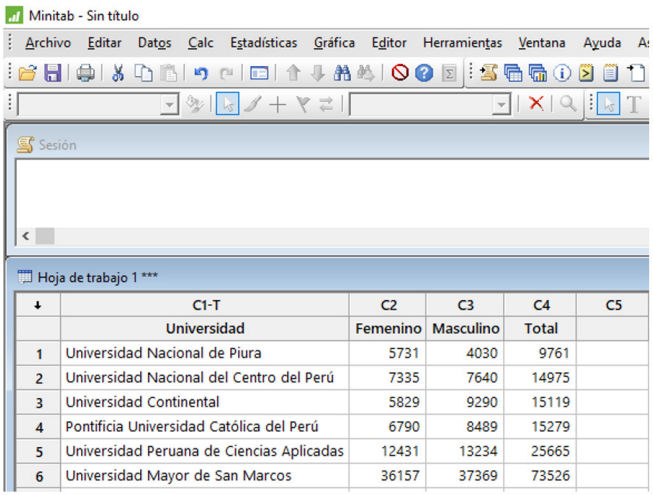


Interpretación:

Del gráfico mostrado, la Universidad Nacional Mayor de San Marcos es la universidad con mayor cantidad de postulantes, seguida de la Universidad de Ciencias Aplicadas con 25 665 y la Pontificia Universidad Católica del Perú con 15 279 postulantes.

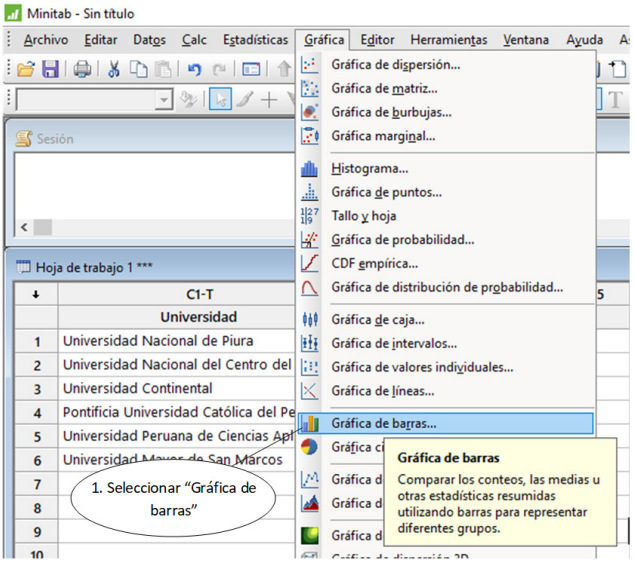
Aplicando Minitab:

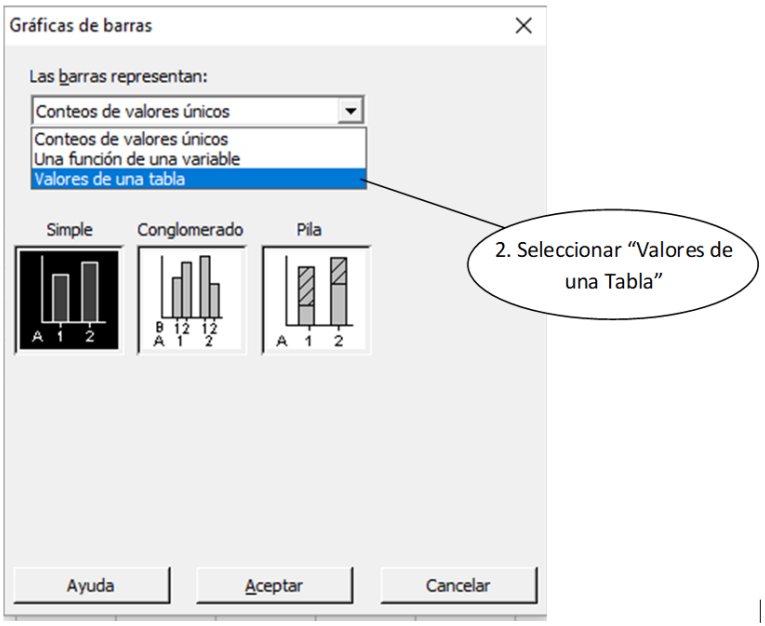
Se procede a copiar los datos en la hoja de trabajo de Minitab.

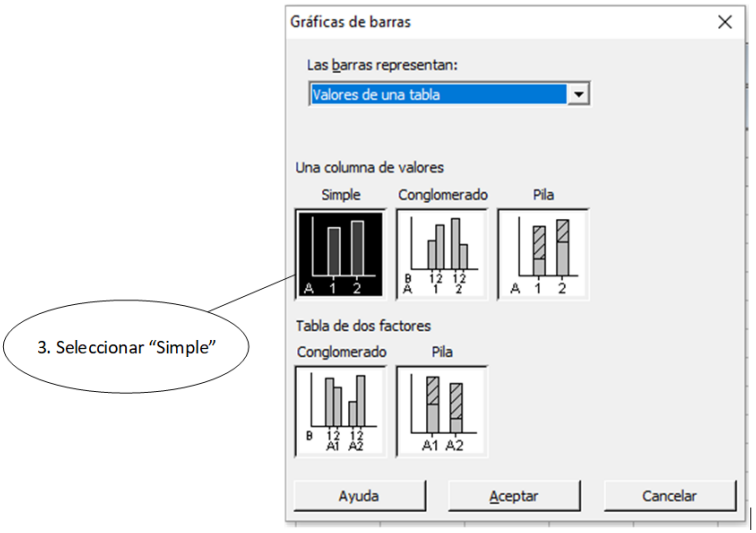


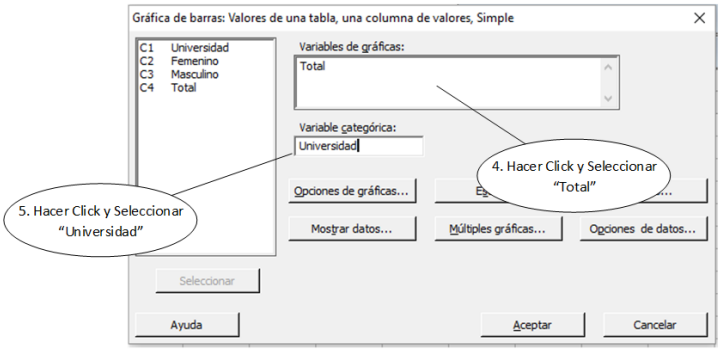
	C1-T	C2	C3	C4	C5
	Universidad	Femenino	Masculino	Total	
1	Universidad Nacional de Piura	5731	4030	9761	
2	Universidad Nacional del Centro del Perú	7335	7640	14975	
3	Universidad Continental	5829	9290	15119	
4	Pontificia Universidad Católica del Perú	6790	8489	15279	
5	Universidad Peruana de Ciencias Aplicadas	12431	13234	25665	
6	Universidad Mayor de San Marcos	36157	37369	73526	

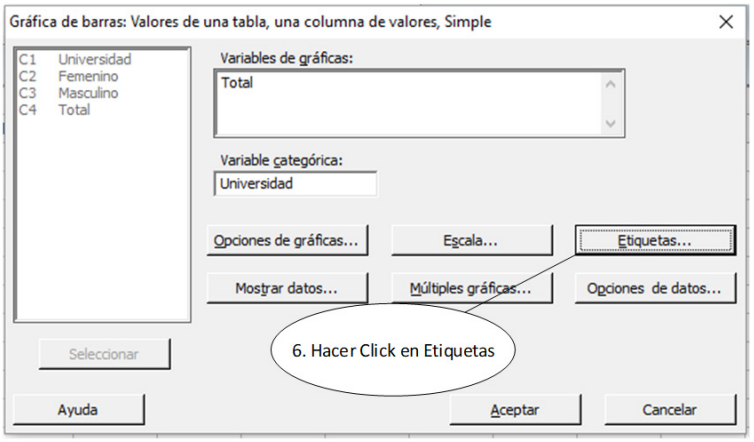
En la parte superior seleccionar “Gráfica” y posteriormente “Gráfica de barras”

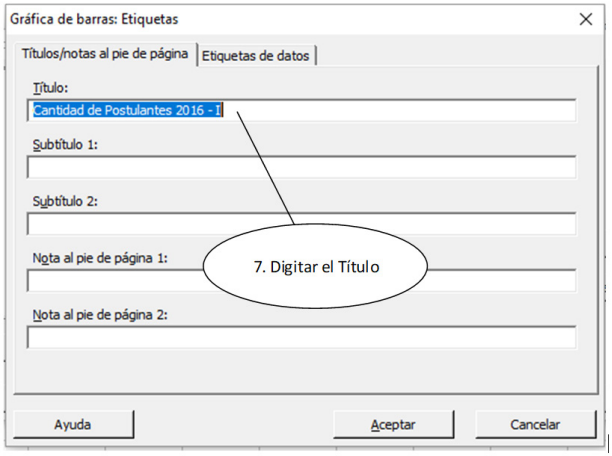


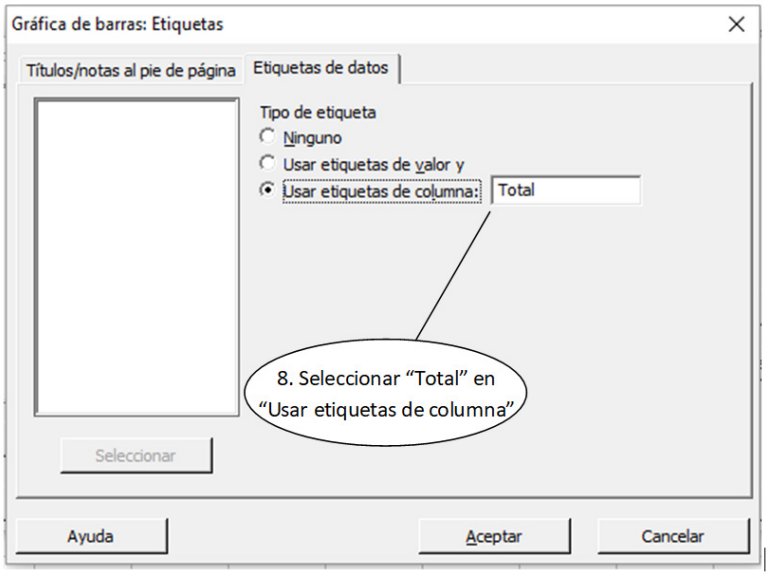

Gráficas de barras
Las barras representan:
Conteos de valores únicos
Conteos de valores únicos
Una función de una variable
Valores de una tabla
Simple
Conglomerado
Pila
2. Seleccionar "Valores de una Tabla"
Ayuda
Aceptar
Cancelar

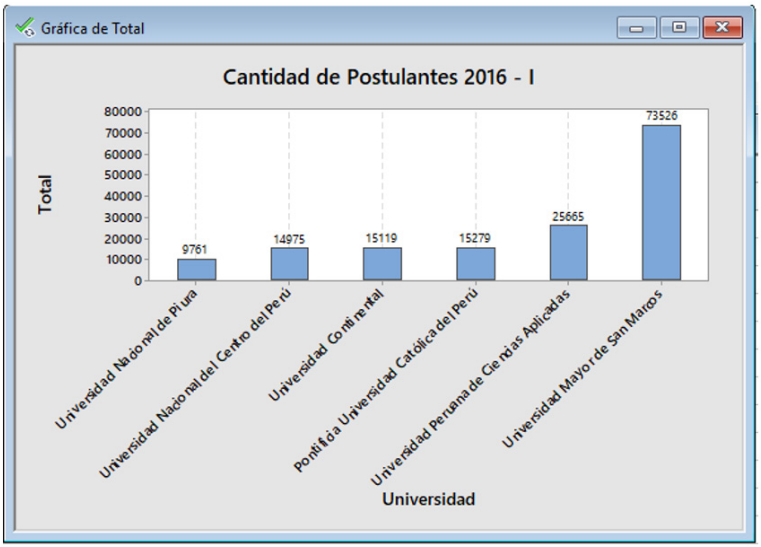

Gráficas de barras
Las barras representan:
Valores de una tabla
Una columna de valores
Simple
Conglomerado
Pila
Tabla de dos factores
Conglomerado
Pila
3. Seleccionar "Simple"
Ayuda
Aceptar
Cancelar


Gráfica de barras: Valores de una tabla, una columna de valores, Simple
C1 Universidad
C2 Femenino
C3 Masculino
C4 Total
Variables de gráficas:
Total
Variable categórica:
Universidad
4. Hacer Click y Seleccionar "Total"
5. Hacer Click y Seleccionar "Universidad"
Opciones de gráficas...
Mostrar datos...
Múltiples gráficas...
Opciones de datos...
Seleccionar
Ayuda
Aceptar
Cancelar


Gráfica de barras: Valores de una tabla, una columna de valores, Simple
C1 Universidad
C2 Femenino
C3 Masculino
C4 Total
Variables de gráficas:
Total
Variable categórica:
Universidad
Opciones de gráficas...
Escala...
Etiquetas...
Mostrar datos...
Múltiples gráficas...
Opciones de datos...
6. Hacer Click en Etiquetas
Seleccionar
Ayuda
Aceptar
Cancelar

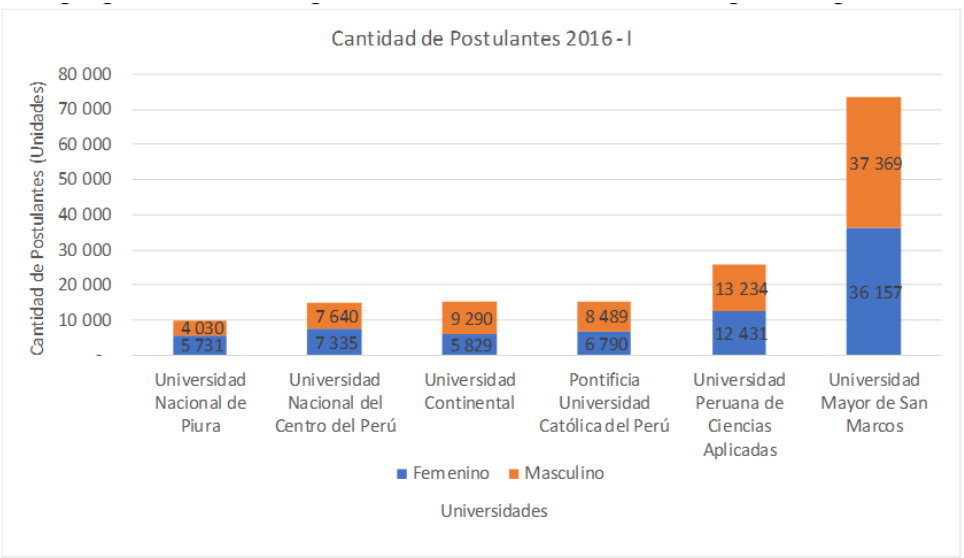

Gráfica de barras: Etiquetas
Títulos/notas al pie de página
Etiquetas de datos
Título:
Cantidad de Postulantes 2016 - T
Subtítulo 1:
Subtítulo 2:
Nota al pie de página 1:
7. Digitar el Título
Nota al pie de página 2:
Ayuda
Aceptar
Cancelar


Gráfica de barras: Etiquetas
Títulos/notas al pie de página
Etiquetas de datos
Tipo de etiqueta
Ninguno
Usar etiquetas de valor y
Usar etiquetas de columna:
Total
8. Seleccionar "Total" en "Usar etiquetas de columna"
Seleccionar
Ayuda
Aceptar
Cancelar


Gráfica de Total
Cantidad de Postulantes 2016 - I
Total
80000
70000
60000
50000
40000
30000
20000
10000
0
9761
14975
15119
15279
25665
73526
Universidad Nacional de Piura
Universidad Nacional del Centro del Perú
Universidad Continental
Pontificia Universidad Católica del Perú
Universidad Peruana de Ciencias Aplicadas
Universidad Mayor de San Marcos
Universidad

3.6 GRÁFICO DE BARRAS COMPUESTO

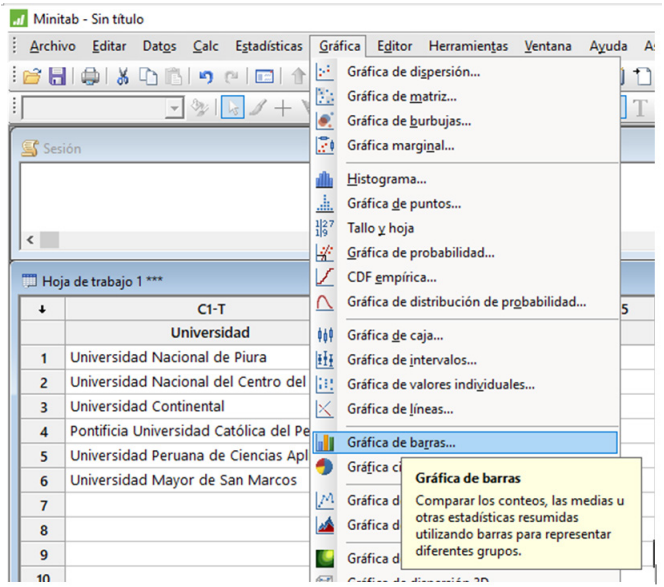
Cuando se requiere presentar una cantidad mayor de información se puede emplear el Grafico de Barras Compuesto, en el cual la categoría inicial (Universidad) se divide en subcategorías, en este caso, la subcategoría de género masculino y femenino. Con los datos de Tabla 2 se agrega esta subcategoría resultando finalmente el siguiente gráfico.

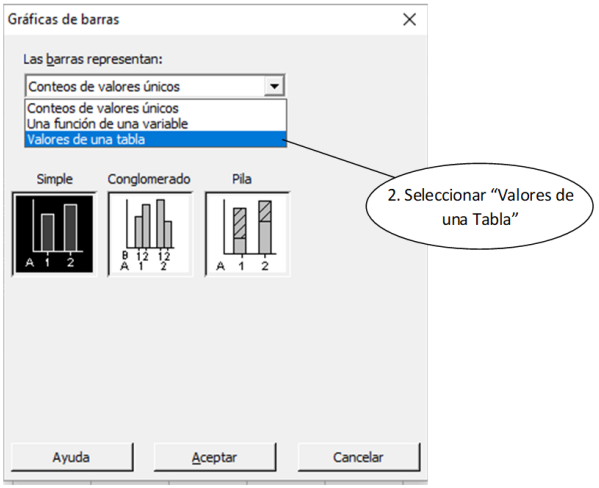


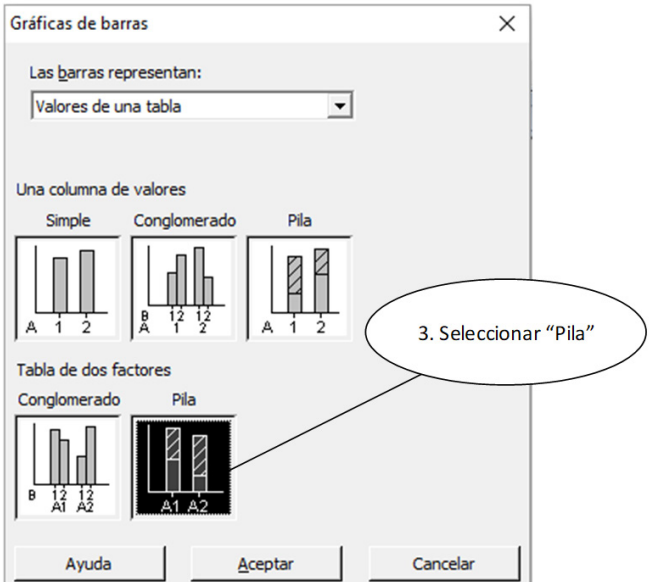
Interpretación:

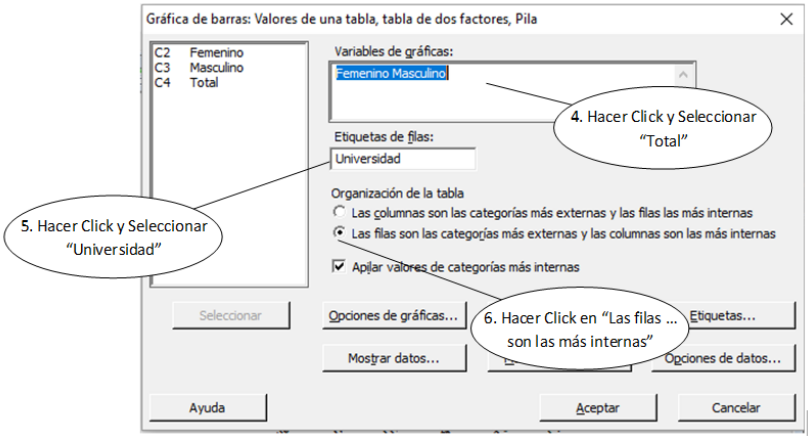
Se observa del gráfico que la Universidad Nacional Mayor de San Marcos ha recibido una mayor cantidad de postulantes en el año 2016 – I, en comparación con las demás universidades estudiadas. Así mismo, se puede observar que existe una cantidad similar de postulantes femeninas y masculinas para todas las universidades en el mismo periodo.

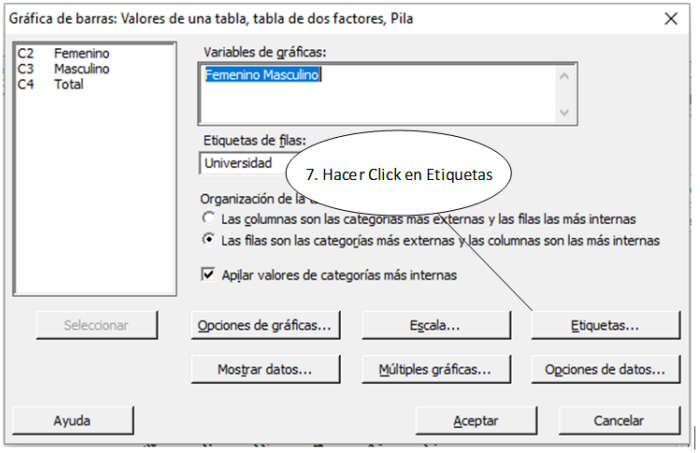
Aplicando MiniTab:

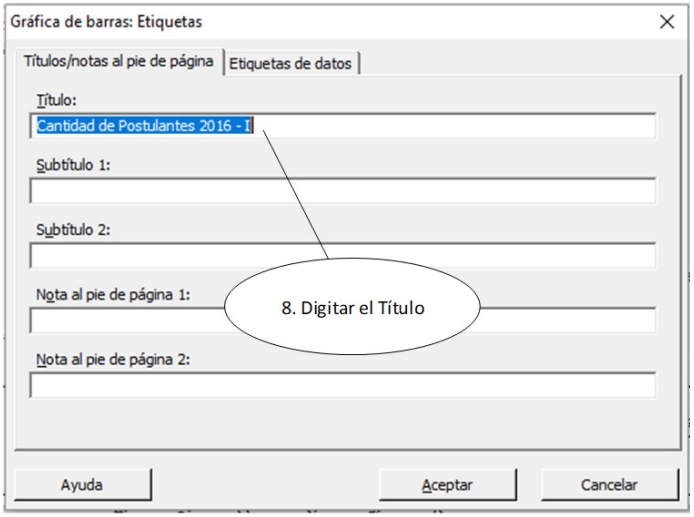


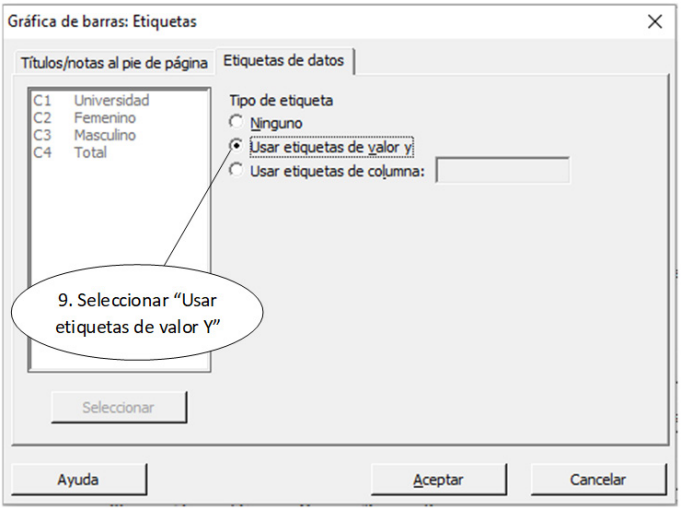


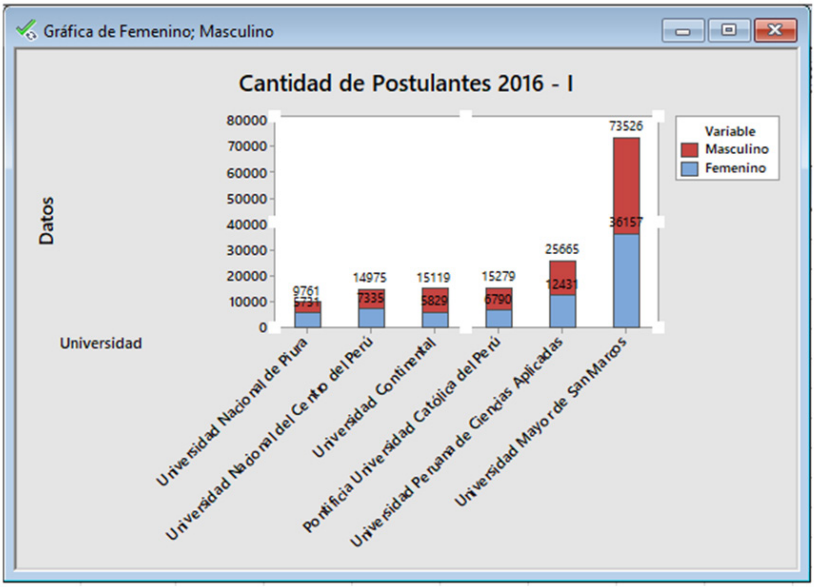

Gráficas de barras
Las barras representan:
Valores de una tabla
Una columna de valores
Simple
Conglomerado
Pila
Tabla de dos factores
Conglomerado
Pila
3. Seleccionar "Pila"
Ayuda
Aceptar
Cancelar


Gráfica de barras: Valores de una tabla, tabla de dos factores, Pila
C2 Femenino
C3 Masculino
C4 Total
Variables de gráficas:
Femenino Masculino
4. Hacer Click y Seleccionar "Total"
Etiquetas de filas:
Universidad
5. Hacer Click y Seleccionar "Universidad"
Organización de la tabla
Las columnas son las categorías más externas y las filas las más internas
Las filas son las categorías más externas y las columnas son las más internas
Apilar valores de categorías más internas
6. Hacer Click en "Las filas ... son las más internas"
Seleccionar
Opciones de gráficas...
Etiquetas...
Mostrar datos...
Opciones de datos...
Ayuda
Aceptar
Cancelar


Gráfica de barras: Valores de una tabla, tabla de dos factores, Pila
C2 Femenino
C3 Masculino
C4 Total
Variables de gráficas:
Femenino Masculino
Etiquetas de filas:
Universidad
7. Hacer Click en Etiquetas
Las columnas son las categorías más externas y las filas las más internas
Las filas son las categorías más externas y las columnas son las más internas
Apilar valores de categorías más internas
Seleccionar
Opciones de gráficas...
Escala...
Etiquetas...
Mostrar datos...
Múltiples gráficas...
Opciones de datos...
Ayuda
Aceptar
Cancelar


Gráfica de barras: Etiquetas
Títulos/notas al pie de página
Etiquetas de datos
Título:
Cantidad de Postulantes 2016 - I
Subtítulo 1:
Subtítulo 2:
Nota al pie de página 1:
8. Digitar el Título
Nota al pie de página 2:
Ayuda
Aceptar
Cancelar


Gráfica de barras: Etiquetas
Títulos/notas al pie de página
Etiquetas de datos
C1 Universidad
C2 Femenino
C3 Masculino
C4 Total
Tipo de etiqueta
Ninguno
Usar etiquetas de valor y
Usar etiquetas de columna:
9. Seleccionar "Usar etiquetas de valor Y"
Seleccionar
Ayuda
Aceptar
Cancelar


Gráfica de Femenino; Masculino
Cantidad de Postulantes 2016 - I
Datos
Universidad
Variable
Masculino
Femenino
80000
70000
60000
50000
40000
30000
20000
10000
0
9761
14975
15119
15279
25665
73526
36157
Universidad Nacional de Piura
Universidad Nacional del Centro del Perú
Universidad Continental
Pontificia Universidad Católica del Perú
Universidad Peruana de Ciencias Aplicadas
Universidad Mayor de San Marcos

3.7 GRÁFICO RADIAL

También llamado araña o estrella, es un gráfico que sirve para comparar datos discretos (nominales y ordinales). Usa la circunferencia o un polígono del gráfico como ordenadas, a cada variable se le otorga un eje que inicia en el centro del círculo, estos se disponen radialmente con iguales distancias y escalas entre sí. Cada valor se traza en su eje y se conectan los ejes entre diferentes variables para formar un polígono. Su utilidad radica en que permite visualizar qué variables tienen valores similares o atípicos.

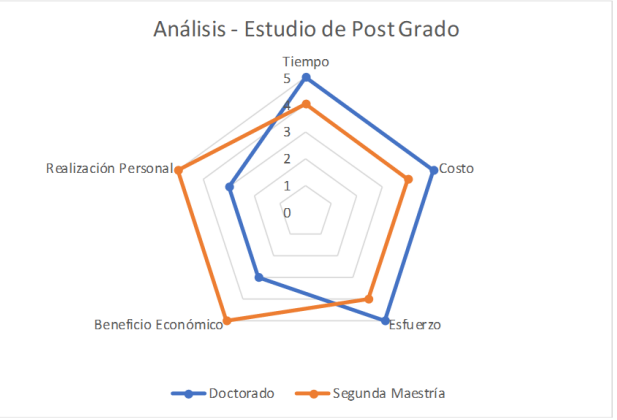
Aplicación Tradicional:

El Gerente de una prestigiosa institución del Estado Peruano está evaluando estudiar una segunda maestría o un doctorado, para ello ha recopilado información consultando a sus colegas y mentores en función a las cinco variables que más le interesan obteniendo el resultado de la Tabla 3. El Gerente ha establecido una un rango de puntuación siendo el valor más bajo igual a 0 puntos y el más alto igual a 5 puntos.

Tabla 3 – Puntuación de Variables – Doctorado vs Segunda Maestría

Variables	Doctorado	Segunda Maestría
Tiempo	5	4
Costo	5	4
Dificultad	5	4
Menor Beneficio Económico	3	5
Menor Realización Profesional	3	5

Con los datos de la Tabla 3, se diagrama el siguiente gráfico



Del gráfico se entiende que, la segunda maestría requerirá de un menor esfuerzo, tiempo y costo que el doctorado, sin embargo, el doctorado le permitirá lograr una mayor realización personal y obtener un mayor beneficio económico.

3.8 HISTOGRAMA

El Histograma es un gráfico utilizado para representar datos continuos de intervalo o razón. Este gráfico está compuesto por barras verticales donde el ancho de la base de cada barra es igual a la otra. En el eje horizontal se acumula la frecuencia de cada puntuación. Se debe tener en consideración que las barras se grafican juntas a diferencia del Gráfico de Barras donde se grafican espaciadas.

Modo tradicional:

Tabla 5

Peso de camiones (Toneladas)									
17,70	17,20	16,90	20,50	14,50	17,60	14,00	20,50	21,70	17,30
21,40	15,80	21,60	14,60	17,70	16,10	21,40	18,20	17,50	14,60

Para realizar el gráfico del Histograma se debe calcular los siguientes valores

Tabla 6 – Elementos del Histograma

Elemento	Valor	Fórmula
N	20,0 camiones	Número de valores
Valor Mínimo	14,0 toneladas	Valor Mínimo
Valor Máximo	21,7 toneladas	Valor Máximo
Rango	7,7 toneladas	Valor Máximo – Valor Mínimo
Número de Clases	6,0 clases	1 + 3.33 log (N=20)*
Amplitud	1,3 toneladas	Rango / Cantidad de Clases

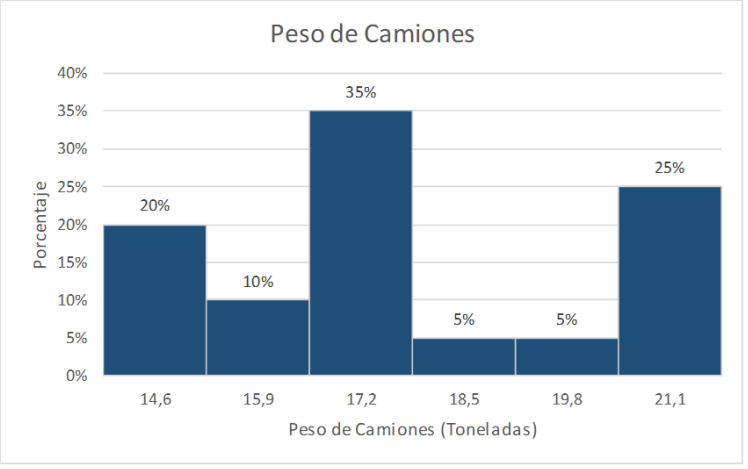
* El número de clases puede ser elegido empleado la Ley de Sturges (1 + 3,33*log(N)) o a criterio del lector.

Con la información de la Tabla 6 se puede interpretar que el gráfico contará con 6 barras verticales de un ancho o amplitud de 1,3 Toneladas.

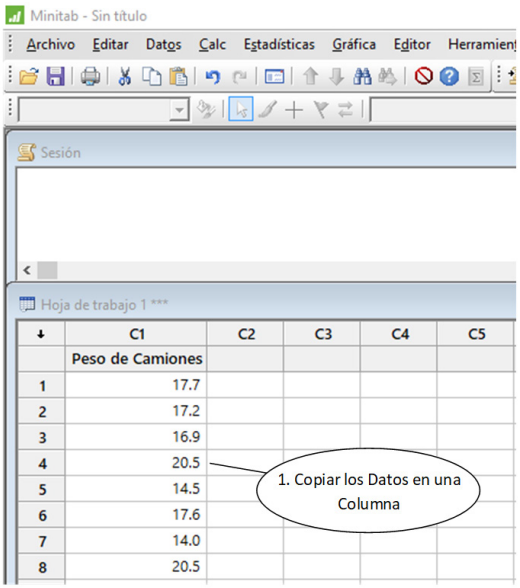
Para iniciar la construcción de la Tabla 7, se debe considerar el siguiente razonamiento:

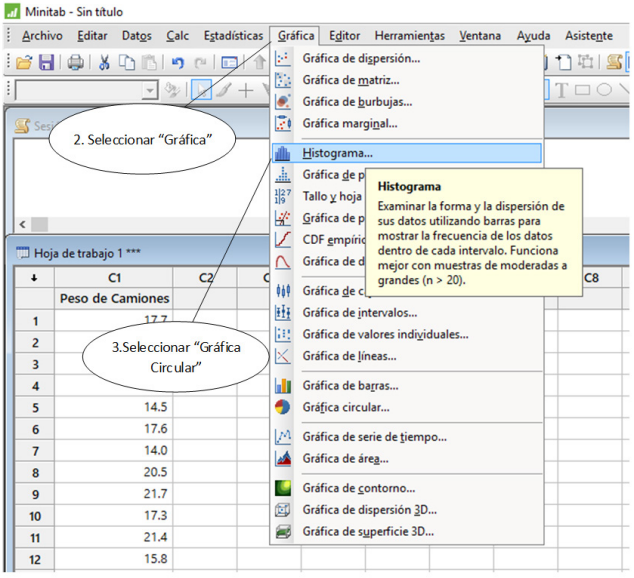
$Límite\ Inferior = Valor\ Mínimo$
$Límite\ Superior = Límite\ Inferior + Amplitud$
$Marca\ de\ Clase = \frac{(Límite\ Inferior + Limite\ Superior)}{2}$
$Cantidad\ de\ Camiones = Camiones\ con\ tonelaje\ entre\ el\ Límite\ Inferior\ y\ el\ Límite\ Superior$
$Frecuencia\ Relativa = \frac{Cantidad\ de\ Camiones\ por\ Clase}{Total\ de\ Camiones}$ x 100%

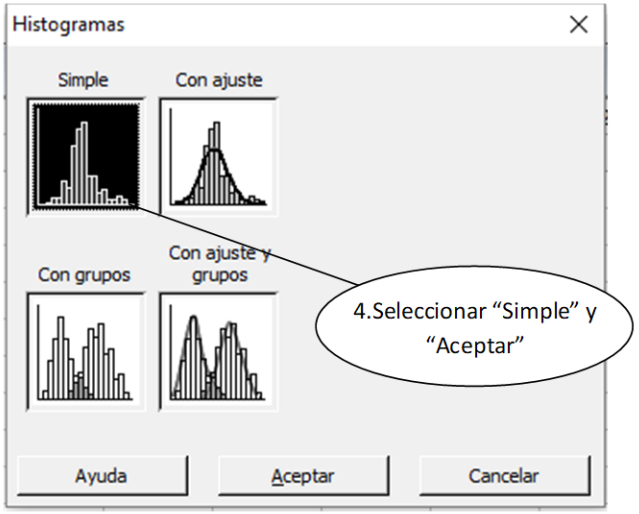
Peso de Camiones (Toneladas)					
Número de Clase	**Límite Inferior**	**Límite Superior**	**Marca de Clase**	**Cantidad de Camiones**	**Frecuencia Relativa**
1	14,0	15,3	14,6	4	20%
2	15,3	16,6	15,9	2	10%
3	16,6	17,9	17,2	7	35%
4	17,9	19,2	18,5	1	5%
5	19,2	20,5	19,8	1	5%
6	20,5	21,7	21,1	5	25%
Total				**20**	**100%**

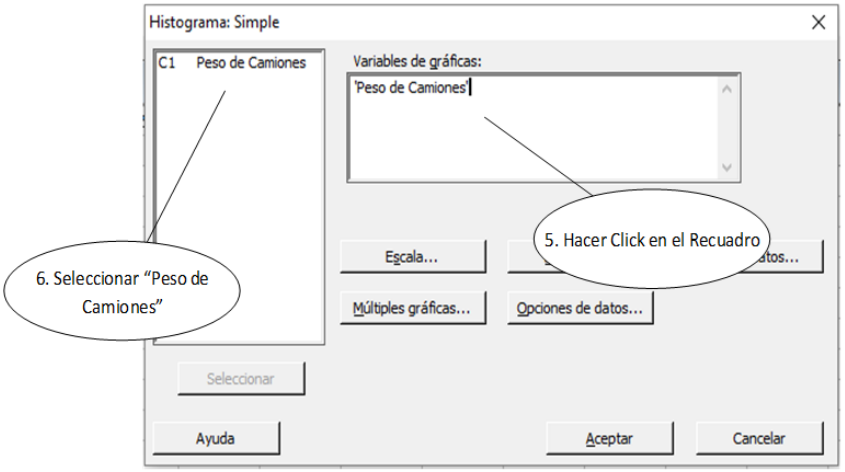


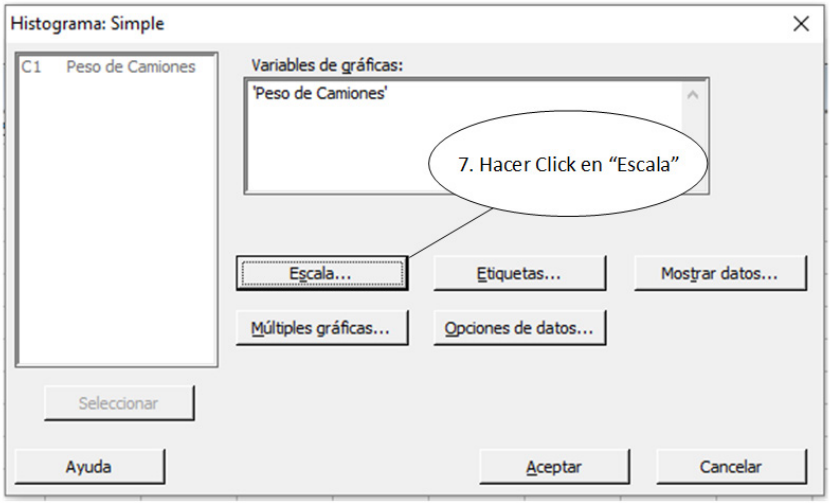
Aplicando Minitab

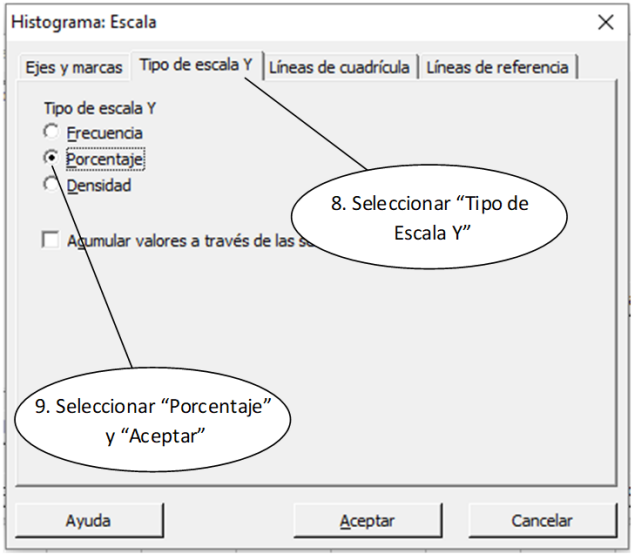


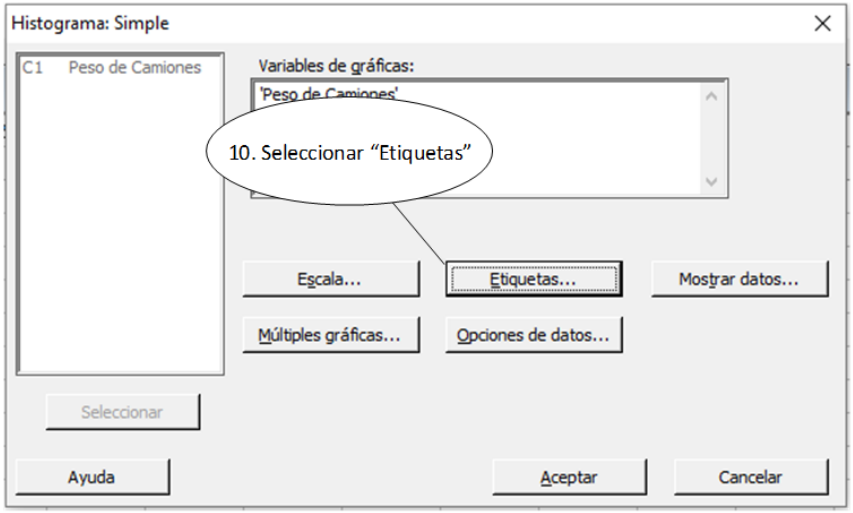


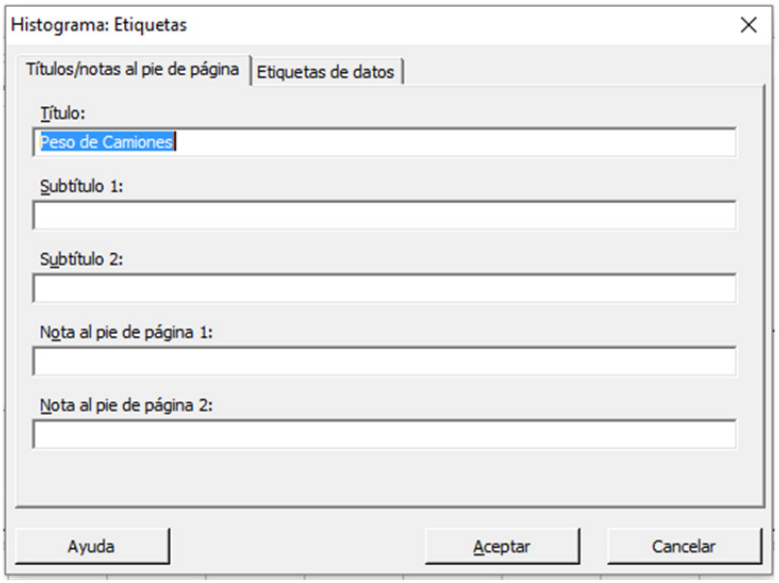

Histogramas
Simple
Con ajuste
Con grupos
Con ajuste y grupos
4.Seleccionar “Simple” y “Aceptar”
Ayuda
Aceptar
Cancelar

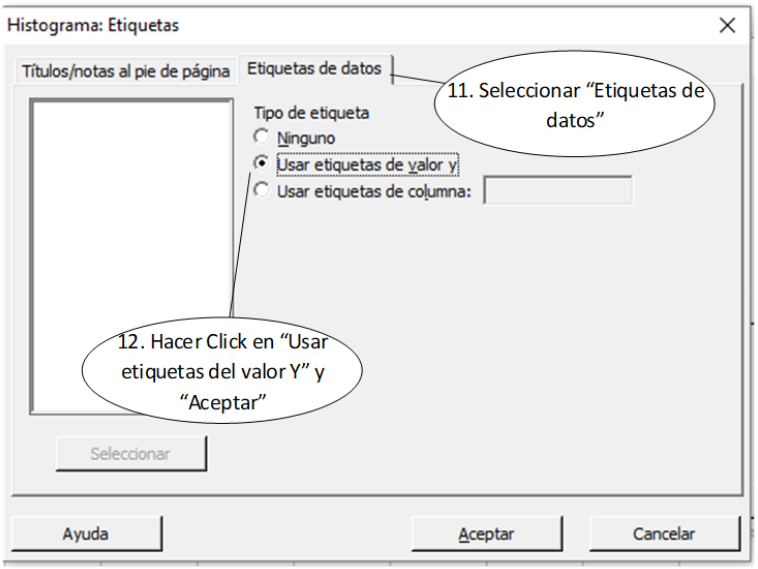

Histograma: Simple
C1 Peso de Camiones
Variables de gráficas:
'Peso de Camiones'
5. Hacer Click en el Recuadro
6. Seleccionar “Peso de Camiones”
Escala...
Múltiples gráficas...
Opciones de datos...
Seleccionar
Ayuda
Aceptar
Cancelar

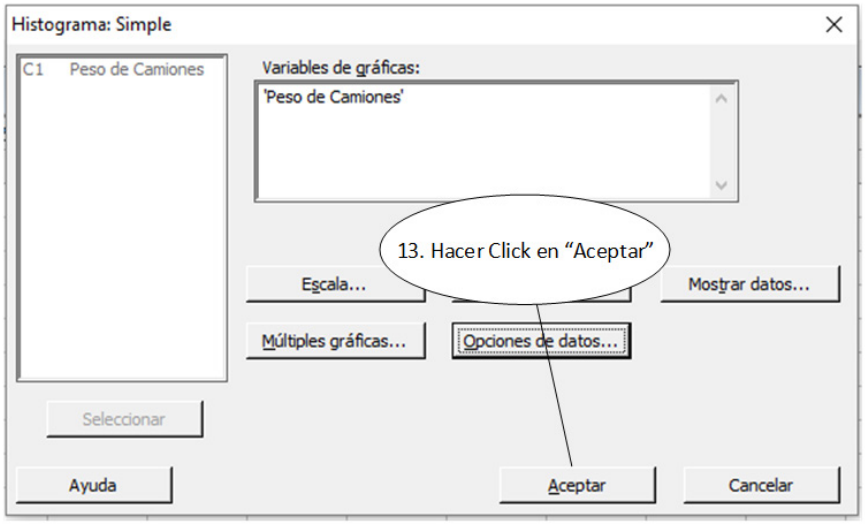

Histograma: Simple
C1 Peso de Camiones
Variables de gráficas:
'Peso de Camiones'
7. Hacer Click en "Escala"
Escala...
Etiquetas...
Mostrar datos...
Múltiples gráficas...
Opciones de datos...
Seleccionar
Ayuda
Aceptar
Cancelar


Histograma: Escala
Ejes y marcas
Tipo de escala Y
Líneas de cuadrícula
Líneas de referencia
Tipo de escala Y
Frecuencia
Porcentaje
Densidad
8. Seleccionar "Tipo de Escala Y"
9. Seleccionar "Porcentaje" y "Aceptar"
Ayuda
Aceptar
Cancelar

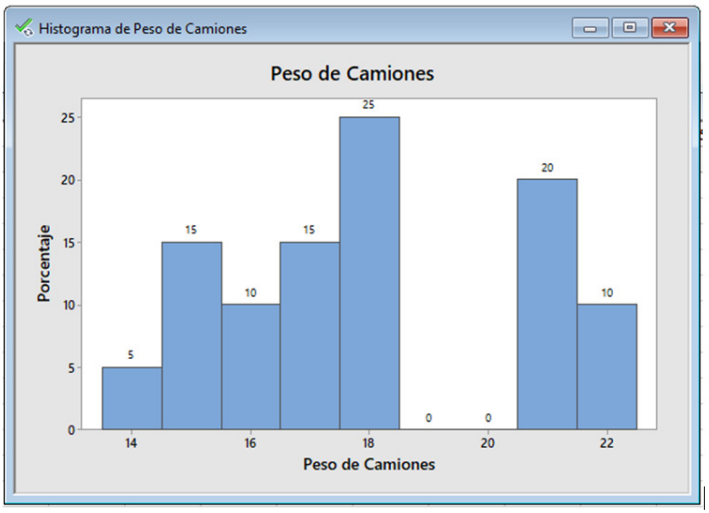

Histograma: Simple
C1 Peso de Camiones
Variables de gráficas:
'Peso de Camiones'
10. Seleccionar "Etiquetas"
Escala...
Etiquetas...
Mostrar datos...
Múltiples gráficas...
Opciones de datos...
Seleccionar
Ayuda
Aceptar
Cancelar


Histograma: Etiquetas
Títulos/notas al pie de página
Etiquetas de datos
Título:
Peso de Camiones
Subtítulo 1:
Subtítulo 2:
Nota al pie de página 1:
Nota al pie de página 2:
Ayuda
Aceptar
Cancelar





Como se observará a en el gráfico de MiniTab el histograma es diferente al hallado de la manera tradicional, esto se debe a que MiniTab emplea otro método para calcular el número de clases.



3.9 POLÍGONO DE FRECUENCIAS

Es un gráfico de líneas o gráfico de 90 grados para representar variables de intervalo o razón trazadas sobre el eje horizontal. Las frecuencias relativas de puntuación descritas por las alturas de puntos localizados sobre puntuaciones y enlazados por líneas rectas.

Los polígonos de frecuencia son especialmente útiles para comparar dos o más muestras, Por ejemplo, comparemos las distribuciones de evaluaciones de rendimiento de combustible para autos compactos contra los vehículos utilitarios de tracción en las cuatro ruedas (SUV),La tabla 3-3 de las distribuciones de frecuencia y frecuencia porcentual para ambos tipos de vehículos, Nótese que difieren los tamaños muestrales de los tipos de vehículos, Hay 106 modelos de autos compactos pero solo 68 modelos de los SUV, Si usamos frecuencias sin elaborar para construir los polígonos, el polígono para los autos compactos más numerosos hará empequeñecer el polígono para los SUV.

3.10 GRÁFICO DE FRECUENCIA ABSOLUTA

Una variación del Histograma es el Gráfico de Frecuencia Absoluta. Este gráfico representa el número de veces o frecuencia con la que aparece un valor por cada categoría. Se construyen las barras con base de igual amplitud y la altura de éstos toma el valor de la frecuencia absoluta de la categoría que representa.

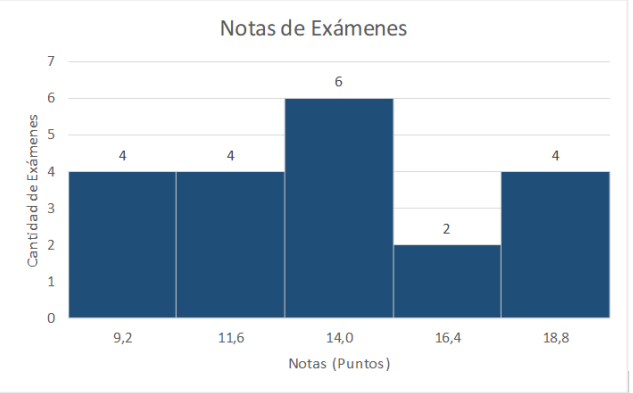
Notas de Exámenes (Unidades)									
12,00	12,00	17,00	17,50	8,00	18,00	15,00	13,00	9,00	14,00
20,00	19,00	12,00	15,00	10,00	11,00	10,00	13,00	14,00	18,00

Elemento	Valor	Fórmula
N	20,0 exámenes	Número de valores
Valor Mínimo	8,0 unidades	Valor Mínimo
Valor Máximo	20,0 unidades	Valor Máximo
Rango	12,0 unidades	Valor Máximo – Valor Mínimo
Número de Clases	5,0 clases	A criterio del lector*
Amplitud	2,4 unidades	Rango / Cantidad de Clases

* El número de clases puede ser elegido empleado la Ley de Sturges (1 + 3,33*log(N)) o a criterio del lector.

$Límite\ Inferior = Valor\ Mínimo$
$Límite\ Superior = Límite\ Inferior + Amplitud$
$Marca\ de\ Clase = \frac{(Límite\ Inferior + Limite\ Superior)}{2}$
$Notas = Notas\ con\ puntaje\ entre\ el\ Límite\ Inferior\ y\ el\ Límite\ Superior$
$Frecuencia\ Relativa = \frac{Cantidad\ de\ Exámenes\ por\ Clase}{Total\ de\ Exámenes}\ \text{x}\ 100\%$

Notas de Exámenes (Puntos)					
Número de Clase	Límite Inferior	Límite Superior	Marca de Clase	Cantidad de Camiones	Frecuencia Relativa
1	8,0	10,4	9,2	4	20%
2	10,4	12,8	11,6	4	20%
3	12,8	15,2	14,0	6	30%
4	15,2	17,6	16,4	2	10%
5	17,6	20,0	18,8	4	20%
Total				20	100%



3.11 GRÁFICO DE FRECUENCIA ABSOLUTA ACUMULADA

Este gráfico tiene forma de escalera, indica en el eje de las ordenadas, la frecuencia del valor del eje de las abscisas.

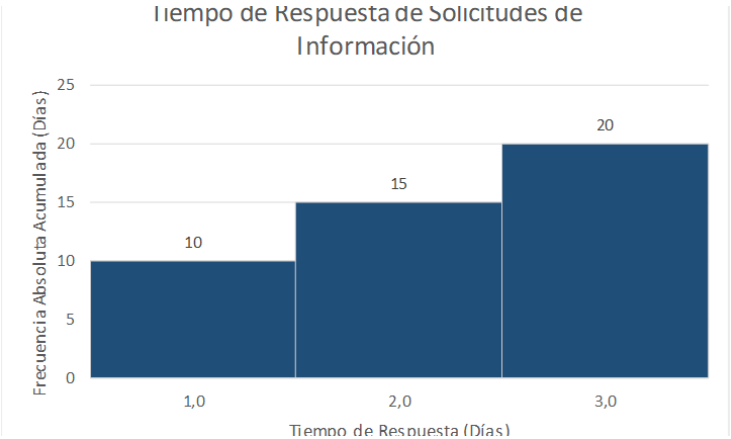
Tiempo de Respuesta de Solicitudes de Información (En Días)									
1	2	3	4	5	1	2	3	4	5
2	3	2	1	3	2	3	1	4	2

Elemento	Valor	Fórmula
N	20 muestras	Número de valores
Valor Mínimo	1,0 días	Valor Mínimo
Valor Máximo	5,0 días	Valor Máximo
Rango	4,0 días	Valor Máximo – Valor Mínimo
Cantidad de Clases	3,0 clases	A Criterio del Lector
Amplitud	1,3 unidades	Rango / Cantidad de Clases

* El número de clases puede ser elegido empleado la Ley de Sturges (1 + 3,33*log(N)) o a criterio del lector.

$Límite\ Inferior = Valor\ Mínimo$
$Límite\ Superior = Límite\ Inferior + Amplitud$
$Marca\ de\ Clase = \frac{(Límite\ Inferior + Limite\ Superior)}{2}$
$Notas = Notas\ con\ puntaje\ entre\ el\ Límite\ Inferior\ y\ el\ Límite\ Superior$
$Frecuencia\ Absoluta = Suma\ del\ Valor\ de\ la\ Frecuencia\ de\ cada\ Clase$
$Frecuencia\ Relativa = \frac{Cantidad\ de\ Exámenes\ por\ Clase}{Total\ de\ Exámenes}\ \text{x}\ 100\%$

Tiempo de Respuesta de Solicitudes de Información (En Días)						
Número de Clase	**Límite Inferior**	**Límite Superior**	**Marca de Clase**	**Días**	**Frecuencia Acumulada Absoluta**	**Frecuencia Relativa**
1	1,0	2,3	1,7	10	10 = (10)	50%
2	2,3	3,7	3,0	5	15 = (10 + 5)	25%
3	3,7	5,0	4,4	5	20 = (10 +5 +5)	25%
Total				20	Total	100%



Interpretación

En el gráfico se muestra que han existido 10 casos donde el tiempo de respuesta ha sido de hasta 1 día, en 15 casos el tiempo de respuesta ha sido de hasta 2 días y en 20 casos el total de casos el tiempo de respuesta ha sido de hasta 3 días. Los 20 casos representan el total de la muestra.

3.12 DIAGRAMA DE PARETO

También se le denomina distribución ABC. Este gráfico representa los valores organizados de mayor a menor y separados por barras. Se fundamente en el principio de Pareto o del 80/20, el cual indica que existen pocos elementos vitales y muchos triviales. Este gráfico permite asignar orden de prioridades o discernir las causas más importantes de un problema determinado.

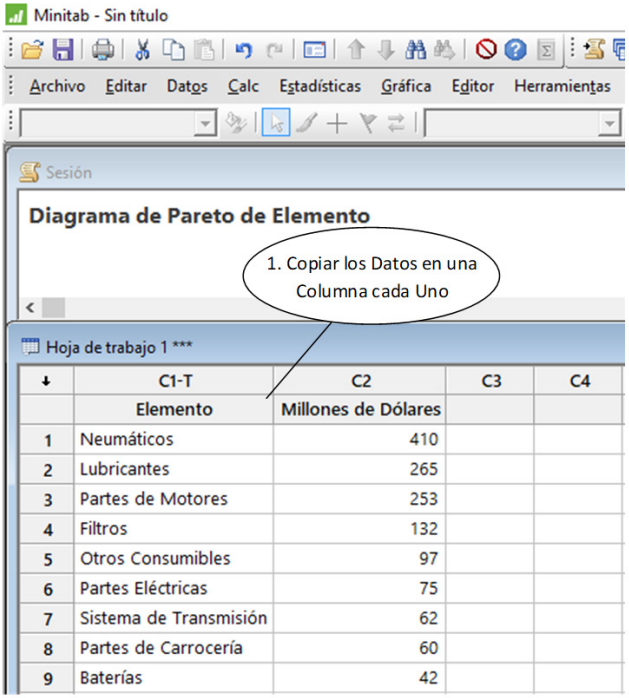
Importación de Suministros (FOB USD$)	
Suministro	**Millones de USD$**
Neumáticos	410
Lubricantes	265
Partes de Motores	253
Filtros	132
Otros Consumibles	97
Partes Eléctricas	75
Sistema de Transmisión	62
Partes de Carrocería	60
Baterías	42
Sistemas de Frenos	38
Elementos de Cuidado	36
Láminas Polarizadas	30
Accesorios Menores	26
Sensores	22
Pantallas / Monitores	17
Láminas de Protección	13

Forma Tradicional

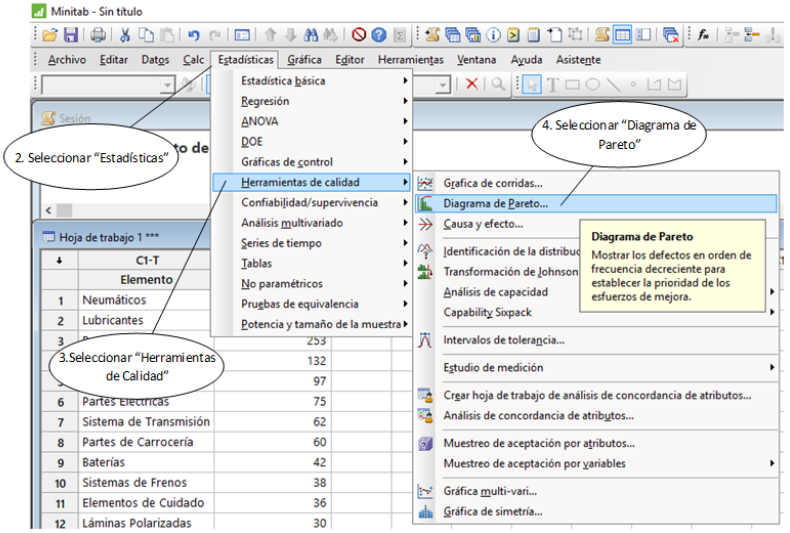
$Frecuecia\ Relativa = \frac{Millones\ de\ USD\$\ por\ cada\ Elemento}{Total\ de\ Millones\ de\ USD\$}$
$Frecuencia\ Acumulada = Frecuencia\ Relativa + Frecuencia\ Acumulada\ Anterior$
$Clasificación\ ABC$: $A\ hasta\ el\ 79.9\%\ de\ la\ Frecuencia\ Acumulada$ $B\ del\ 80.0\%\ hasta\ el\ 94{,}9\%\ de\ la\ Frecuencia\ Acumulada$ $C\ del\ 95.0\%\ hasta\ el\ 100.0\%\ de\ la\ Freuencia\ Acumulada.$

Importación de Suministros (FOB USD$)				
Suministro	**Millones de USD$**	**Frecuencia Relativa**	**Frecuencia Acumulada**	**Clasificación ABC**
Neumáticos	410	26%	26%	A
Lubricantes	265	17%	43%	A
Partes de Motores	253	16%	59%	A
Filtros	132	8%	67%	A
Otros Consumibles	97	6%	73%	A
Partes Eléctricas	75	5%	78%	A
Sistema de Transmisión	62	4%	82%	B
Partes de Carrocería	60	4%	86%	B
Baterías	42	3%	88%	B
Sistemas de Frenos	38	2%	91%	B
Elementos de Cuidado	36	2%	93%	B
Láminas Polarizadas	30	2%	95%	C
Accesorios Menores	26	2%	97%	C
Sensores	22	1%	98%	C
Pantallas / Monitores	17	1%	99%	C
Láminas de Protección	13	1%	100%	C
Total	1 578	100%		

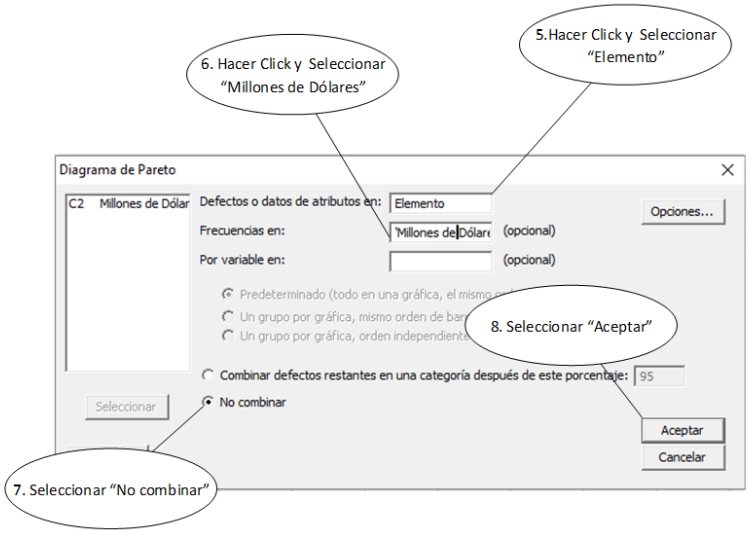
Aplicando MiniTab

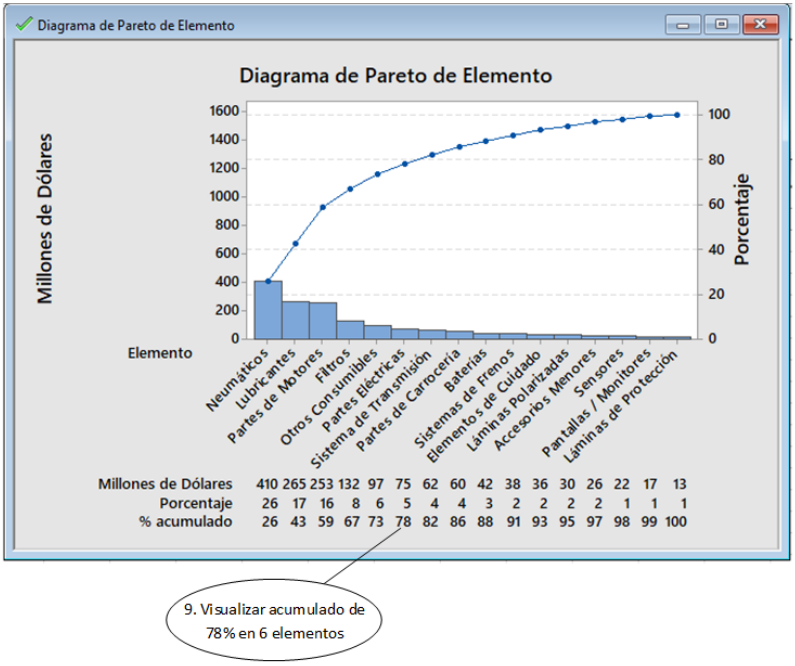


	C1-T	C2	C3	C4
	Elemento	Millones de Dólares		
1	Neumáticos	410		
2	Lubricantes	265		
3	Partes de Motores	253		
4	Filtros	132		
5	Otros Consumibles	97		
6	Partes Eléctricas	75		
7	Sistema de Transmisión	62		
8	Partes de Carrocería	60		
9	Baterías	42		



	C1-T	
	Elemento	
1	Neumáticos	
2	Lubricantes	
3		253
4		132
5		97
6	Partes Eléctricas	75
7	Sistema de Transmisión	62
8	Partes de Carrocería	60
9	Baterías	42
10	Sistemas de Frenos	38
11	Elementos de Cuidado	36
12	Láminas Polarizadas	30





9. Visualizar acumulado de 78% en 6 elementos

3.13 GRÁFICO DE CAJAS

También se le llama de caja y bigote o boxplot, es un gráfico que permite representar mediante cuartiles, la distribución de datos de una muestra. Se compone de un rectángulo o "caja" y dos brazos en la parte superior e inferior de ésta "bigotes". Permite visualizar los datos atípicos así como la simetría de la distribución y dispersión de los datos. Para elaborar un gráfico de barras se requieren los valores mínimos y máximos y los tres cuartiles. Se debe encontrar la mediana, representada por el segmento vertical que divide a la caja y después el primer y tercer cuartil.

3.14 Preguntas y respuestas de repaso

1. La calificación de los hoteles de una ciudad (en estrellas) se muestra a continuación:

2	1	1	4	5	2	1	3	5	4
1	4	4	1	2	1	1	2	4	2
2	5	2	2	1	5	4	1	2	1
3	1	4	4	2	3	1	5	1	3

a) Construya la tabla de distribución de frecuencias.

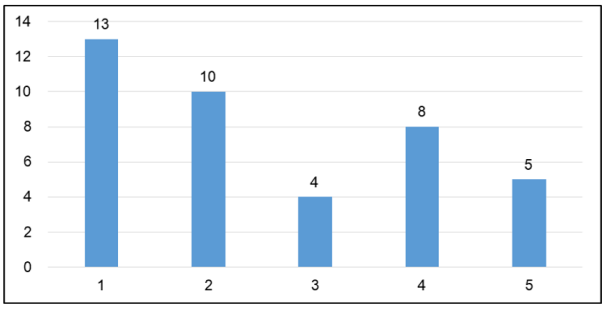
b) Dibuje el diagrama de barras.

Solución

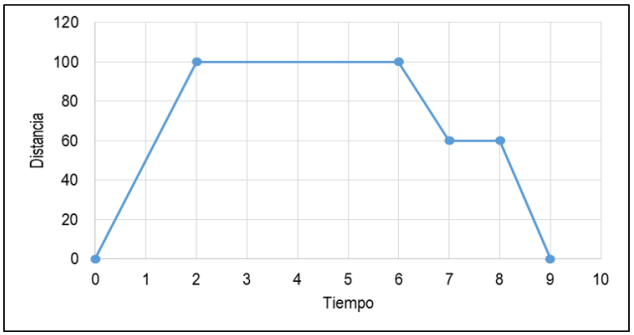
a)

x_i	f_i	F_i
1	13	13
2	10	23
3	4	27
4	8	35
5	5	40

b)



2. Los alumnos de cuarto grado salieron de paseo al Parque de las Leyendas, llevados por el bus del colegio. El siguiente gráfico muestra el tiempo que tomó dicho paseo (en horas) y la distancia recorrida (en km.):



a) ¿A cuántos km. del colegio se encuentra el zoológico?

b) ¿Cuánto tiempo estuvieron los alumnos en el zoológico?

c) ¿Hubo alguna parada en el camino?

d) ¿Cuánto tiempo duró el paseo?

Solución:

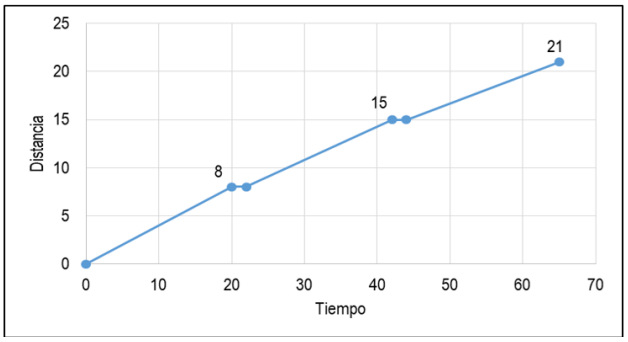
a) El zoológico se encuentra a 100 km.

b) Estuvieron 4 horas (de t = 2 a t = 6).

c) En la ida no, pero en la vuelta sí (1 hora).

d) El paseo duró 9 horas.

3. El tiempo (en minutos) y la distancia (en horas) que recorrió un atleta en la última Media Maratón de Lima se describe en el siguiente gráfico:



a) ¿Cuánto tiempo demoró en llegar a la meta?

b) El atleta hizo 2 paradas para rehidratarse. ¿Cuánto tiempo demoró en total? ¿A qué distancia se encuentran ambos puntos de la partida?

c) ¿A qué velocidad corrió en los primeros 20 minutos?

Solución:

a) El atleta llegó a la meta en 65 minutos, aproximadamente.

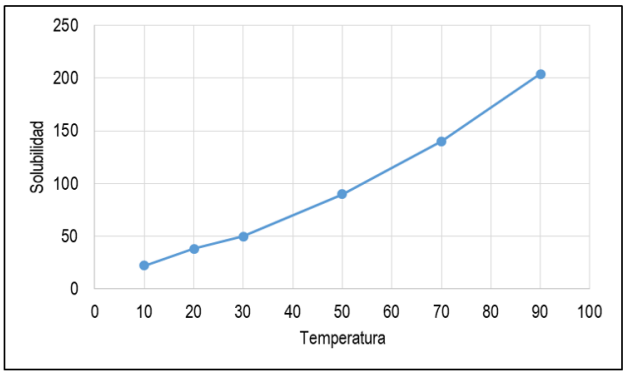
b) El atleta demoró alrededor de 2 minutos en cada punto de rehidratación, es decir, 4 minutos.

El primer punto se encuentra a 8 km. y el segundo a 15 km. de la partida.

c) El atleta recorrió 8 km. en 20 minutos.

$$Velocidad = \frac{8\ km}{20\ min \frac{1\ h}{60\ min}} = 24\ km/h$$

4. La solubilidad del Na2SO4 (en g. / 100 g. de agua) a diferentes temperaturas (en °C) se describe en el siguiente gráfico:



a) ¿Cuál es la solubilidad a 60 °C?

b) ¿Cuál es la variable independiente y dependiente?

c) A mayor temperatura, ¿la solubilidad aumenta o disminuye?

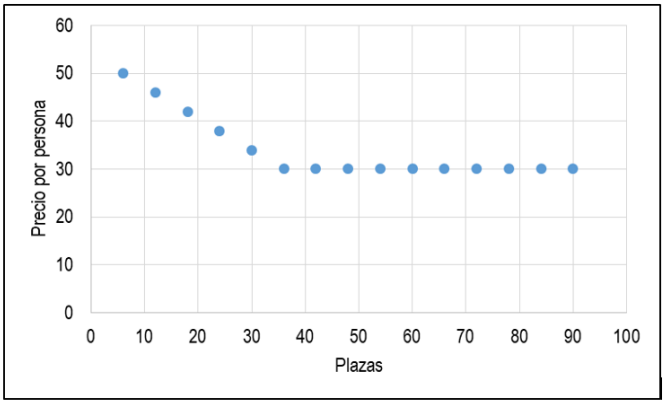
Solución:

a) La solubilidad es de 110 g / 100 g. de agua, aproximadamente.

b) La variable independiente es la temperatura y, la dependiente, la solubilidad.

c) La solubilidad aumenta.

5. Se va a organizar un reencuentro de promoción en un restaurante y el precio va a depender de las personas que confirmen su asistencia. Además, el restaurante tiene capacidad para 90 personas y, como condición, admite solo grupos de 6 personas.



Dicho escenario se describe en el siguiente gráfico:

a) ¿Qué significados tienen los puntos (18,42) y (60,30)?

b) ¿Por qué se dibuja la gráfica entre 6 y 90? ¿Se puede continuar?

c) ¿Es una función continua o discontinua?

d) ¿Por qué no se unen los puntos?

Solución

a) (18,42): Si se ocupan 18 plazas, cada persona pagará 42 soles.

(60,30): Si se ocupan 60 plazas, cada persona pagará 30 soles.

b) Porque el mínimo de plazas es 6 y el máximo 90. No se puede continuar, dado que excedería el máximo.

c) Es discontinua.

d) Porque solo se admiten grupos de 6 personas.

CAPÍTULO IV

ESTADÍSTICA DESCRIPTIVA

4.1 INTRODUCCIÓN

El capítulo IV denominado *Estadística descriptiva*, desarrolla los siguientes temas: La estadística descriptiva, llamada también promedios o medidas de tendencia central, la media, la media aritmética ponderada, debilidades de la media, la mediana, debilidades de la mediana y la moda. Se presenta también un cuadro resumen.

4.2 LA ESTADÍSTICA DESCRIPTIVA LLAMADA TAMBIÉN PROMEDIOS O MEDIDAS DE TENDENCIA CENTRAL

La estadística descriptiva, llamada promedios o medidas de tendencia central, es la técnica matemática que obtiene, organiza, presenta y describe un conjunto de datos con el propósito de facilitar su uso generalmente con el apoyo de tablas, medidas numéricas o gráficas. Además, calcula parámetros estadísticos como las medidas de centralización y de dispersión que describen el conjunto estudiado (Cervantes, 2016, p. 12).

En la estadística descriptiva, solo resumimos y describimos los datos recolectados. Por ejemplo, al analizar a los estudiantes de sexto año de la carrera de Derecho y Ciencia Jurídica en de la Universidad Nacional Mayor de San Marcos en el año 2018 y encontramos que el 45% usa ternos color azul. Así el 45% es un estadístico descriptivo, sin embargo, no pretendemos sugerir que el 45% de los estudiantes de la carrera profesional de Derecho y Ciencia Política en Perú, o ni siquiera en las otras Facultades Académicas utilizan ternos de color azul. Es decir solo se describe el dato que se registró.

Así, la información estadística que aparece, en las revistas de ciencia jurídica, informes jurídicos, libros, tesis con trabajos de campo y demás publicaciones, consiste en datos resumidos y presentados en forma clara para el investigador jurídico. Estos resúmenes de datos, que pueden ser tabulados, gráficos o numéricos se llaman estadísticas descriptivas.

Así la abreviación y exposición de los aspectos más importantes de un conjunto de datos se llama estadística descriptiva, la cual incluye la presentación de datos en formas de tablas, gráficos, cálculos de indicadores numéricos de centralidad y variabilidad. Así, a la estadística descriptiva puede ser definida también como el método que incluyen la recolección, presentación y caracterización de un conjunto de datos con el fin de describir adecuadamente sus diversas características. Estos métodos pueden ser aplicados tanto a los datos.

La estadística descriptiva tiene como función el manejo de los datos recopilados en cuanto se refiere a su ordenación y presentación, para poner en evidencia ciertas características en la forma que sea más objetiva y útil.

En este sentido los datos organizados den una distribución de frecuencias destacan sus características más esenciales. Como marcas de clases, centro, forma de distribución (asimétrica o simétrica), etc. Sin embargo, los indicadores que describen a los datos en forma más precisa deben calcularse. Estos indicadores que resumen los datos en número denominados medidas descriptivas se refieren a la centralización, o la dispersión o variabilidad, a la asimetría y a la curtosis. También son métodos numéricos para describir los datos, indicadores conocidos como, medidas de posición relativa que describen la posición de una observación relativa a las demás observaciones de la distribución, estos son los percentiles y los valores estandarizados. (Córdova 2009, p. 37)

Una población o universo objeto de una investigación aplicada a las ciencias jurídicas puede ser finita si sus elementos se pueden contar. Por ejemplo, el número de alumnos del tercer año de la Facultad de Derecho y Ciencia Jurídica de la Universidad Nacional de Trujillo.

Una población o universo es infinita cuando el número es demasiado grande y el investigador jurídico no puede someter a medición cada uno de ellos. Cuando se miden cualitativamente las características de una población, resultan categorías que necesariamente tienen que ser exhaustivas, es decir, que se puede clasificar a toda la población y también deben de ser mutuamente excluyentes, toda vez, que un mismo elemento no puede pertenecer simultáneamente a dos o más categorías. Por ejemplo, edad de una persona de 25 o 30 años.

Una muestra debe cumplir ciertas condiciones, de aquí el concepto de muestra aleatoria que es aquella obtenida de modo que cada elemento de la población tiene una oportunidad igual e independiente de ser elegido. La investigación estadística aplicada a la ciencia jurídica es toda operación orientada a la recopilación de información sobre una población.

En palabras de Mode (1967, p. 80) muchas inferencias estadísticas relativas a universos deben hacerse a partir de nuestras aleatorias. El primer paso en la obtención de estas inferencias es la descripción de las características numéricas de la muestra. La descripción usualmente implica promedio que resumen características numéricas de la muestra. Un promedio es un valor típico o representativo; es un número único que se emplea para remplazar un conjunto de números. Existen muchos tipos diferentes de promedios, cada uno de ellos con sus propiedades particulares propias. La mediana o valor central de un grupo ordenado, la moda o valor más frecuente y la media aritmética son tres de las medidas más útiles de tendencia central.

Los valores relativos de los tres estadígrafos de tendencia central nos informan acerca de la forma de una distribución de puntuaciones.

4.3 LA MEDIA

Existen tres medidas de tendencia central o promedios utilizadas en la investigación jurídica: la media aritmética, la mediana y la moda. La media es la más común de las

dos y se define como la suma de las marcas dividida pro el número total de los casos comprendidos. Para indicar la media se utiliza por convención el símbolo X. aunque a veces se emplee también la letra M.

La media aritmética es un promedio razonablemente estable. No es afectada hondamente por algunos pocos valores moderadamente pequeños o moderadamente grandes y esta estabilidad aumenta con la frecuencia total N. Sin embargo, uno o más valores extremos pueden algunas veces afectar grandemente su valor y reducir su utilidad. La media aritmética es una medida estadística muy conveniente, debido a su estabilidad general. Tiene infinidad sus usos en campos diversos. En meteorología, para obtener la temperatura media o la precipitación pluvial media; en medicina, para calcular la duración media de una enfermedad, en antropología, para estimar ciertas características de un grupo de seres humanos, en economía, para celebrar salarios medios. Precios, números, índices (Murray y Larry, 2006, p.62); en derecho, para calcular la duración media de un proceso judicial.

Según Murray y Larry (2006, p.62), la media aritmética, o brevemente la media, de un conjunto de N número se denota así: $\bar{X}$ (que se lee "X barra") y está definida como:

$$\bar{X} = \frac{X_1+X_2+X_3+ \dots +X_N}{N} = \frac{\sum_{j=1}^{N} X_j}{N} \dots (1)$$

Ejemplo: La media aritmética de los números de 8, 3, 5,12 y10 es:

$$\bar{X} = \frac{8+3+5+12+10}{5} = \frac{38}{5} = 7.6$$

Si los números X_1, X_2, …. , X_3 se presentan f_1, f_2, … , f_k veces, respectivamente (es decir, se presentan con frecuencia f_1, f_2, … , f_k), su media aritmética es:

$$\bar{X} = \frac{f_1X_1+f_2X_2+ \dots +f_kX_k}{f_1+f_2+ \dots + f_k} = \frac{\sum_{j=1}^{N} f_jX_j}{\sum_{j=1}^{N} f_j} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} \dots (2)$$

Donde $N= \sum f$, es la suma de las frecuencias (es decir, la cantidad total de los casos)

Ejemplo: Si, 5,8, 6 y 2 se presentan con frecuencias 3, 2, 4 y 1, respectivamente, su media aritmética es:

$$\bar{X} = \frac{(3)(5)+(2)(8)+(4)(6)+(1)(2)}{3+2+4+1} = \frac{15+16+24+2}{10} = 5.7$$

4.3.1 La media aritmética ponderada

Para Murray y Larry (2006, p.62), algunas veces, a los números X1,X2, … , XK se les asignan a ciertos factores de ponderación (o pesos) w1, w2, …, wk, que dependen del significado o importancia que se les asigne a estos números. En este caso, a

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \ ... \ + w_kX_k}{w_1 + w_2 + \ ... + \ w_k} = \frac{\sum wX}{\sum w}$$

Se le llama media aritmética ponderada. Obsérvese la semejanza con la ecuación(2), la cual se puede considerar como una media aritmética ponderada con pesos frecuencia f1, f2, … , fk.

Ejemplo: Si en una clase, al examen final se le da el triple de valor que a los exámenes parciales y un estudiante obtiene 95 en el examen final, y 70 y 90 en los dos exámenes parciales, su puntuación media es:

$$\bar{X} = \frac{(3)(95) + (1)(70) + (1)(90)}{3 + 1 + 1} = \frac{415}{5} = 83$$

4.3.2 Debilidades de la media aritmética

Cuando se reporta un estadístico de tendencia central tendemos a suponer que su valor es representativo de puntuaciones típicas en la parte central de una distribución. En ocasiones, sin embargo, cuando se informa la media puede conducir a errores al respecto. Este es el caso porque el cálculo de la media puede inflarse o desinflarse debido a puntuaciones valores extremos. (Ritchey, 2006).

Se presenta el siguiente caso: cantidad de dinero en efectivo que llevan 10 estudiantes al azar y se calcula su media.

4.4 LA MEDIANA

La mediana es una puntuación posicional, la puntuación central en una distribución ordenada. Es igual al 50% del percentil.

En datos sin tabular: los datos se ordenan de menor a mayor y se ubica el valor central. Si hay dos valores centrales, entonces se promedian.

En datos tabulados:

$$Md = L_i + c\ (\frac{\frac{n}{2} - N_i - 1}{n_i})$$

La mediana se encuentra dentro de la clase (categoría) que contiene a la posición **n/2**. Donde $\mathbf{L_i}$ es el límite inferior de esta clase, c es la amplitud de esta clase, $\mathbf{N_i - 1}$ es la frecuencia acumulada anterior a esta clase y $\mathbf{n_i}$ es la frecuencia absoluta.

Cuando una distribución está sesgada, la mediana es el estadístico a elegir porque su valor caerá entre la media y la moda y así, minimizar el error.

Ejemplo: la mediana de los números 3,4,5,6,8,8,8 y 10 es 6.

4.4.1 Debilidades de la mediana

Se tiene el siguiente caso: muestra de 5 ingresos mensuales de 5 familias elegidas al azar.

4.5 LA MODA

La moda es la puntuación o categoría que ocurre con más frecuencia en una distribución. La moda puede verse como puntuación o categoría más popular, pero no debe confundirse a la moda con "la mayoría de las puntuaciones".

La moda es fácil de ubicar en tablas y gráficos. Al identificar la moda debes tener cuidado en recordar que es una puntuación (X), no una frecuencia (f).

La moda es la menos útil de las medidas de tendencia central por sí misma, porque tiene un alcance informativo limitado, esto es, nos dice poco. La moda es insensible a los valores de puntuaciones de distribución e insensible al tamaño muestral. Dos distribuciones de puntuaciones deben tener formas radicalmente diferentes, pero tener la misma boda.

Ejemplo: La edad de los niños de una clase son 7, 7 ,8, 6, 8, 7, 8, 5 y 7 años, la moda seria 7 años.

Resumen

Estadística de Tendencia Central	Definición	Fortaleza y Aplicación	Debilidad Potencial
Media	Promedio aritmético del total de los datos	Abierto a operaciones matemáticas y útil con variables de tipo intervalo	Su cálculo se distorsiona por valores extremos
Mediana	Valor central de una distribución ordenada	Mayormente utilizada cuando distribución está sesgada	Insensibles a los valores de X pero sensible a los cambios de tamaño de la muestra
Moda	Dato repetido con mayor frecuencia.	Mayormente utilizado en conteos donde el porcentaje en la repetición mayor es alto	Insensible a la distribución de los valores de X

4.6 PREGUNTAS Y RESPUESTAS DE REPASO

1. En el curso de Estadística, la distribución de las notas del examen final se dio de la siguiente manera:

Nota	f_i
[0,4[	4
[4,8[	7
[8,12[	11
[12,16[	12
[16,20]	5

a) Hallar la media de la distribución estadística.

Solución

a)

Nota	x_i	f_i	$x_i * f_i$
[0,4[	2	4	8
[4,8[	6	7	42
[8,12[	10	11	110
[12,16[	14	12	168
[16,20]	18	5	90
		39	418

Entonces:

$$\bar{x} = \frac{418}{39} = 10.72$$

2. Los resultados de lanzar un dado 170 veces muestran lo siguiente:

x_i	f_i
1	24
2	a
3	29
4	31
5	27
6	b

a) Hallar a y b, sabiendo que la media es de 3.6.

Solución

a)

x_i	f_i	x_i*f_i
1	24	24
2	A	2a
3	29	87
4	31	124
5	27	135
6	B	6b

Se sabe:

$$24 + a + 29 + 31 + 27 + b = 170$$

$$a + b = 59$$

$$\frac{24 + 2a + 87 + 124 + 135 + 6b}{170} = 3.6$$

$$2a + 6b = 242$$

Resolviendo:

$$a = 28, b = 31$$

3. En un aula de clase, la cantidad de personas que conforman la familia de cada alumno (padre, madre y hermanos) se distribuye de la siguiente manera:

x_i	f_i
3	7
4	11
5	8
6	5
7	2
8	1

a) Hallar la media aritmética.

b) Hallar la moda.

c) Hallar la mediana.

Solución

a)

Se sabe:

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

Donde:

$$N = \sum_{i=1}^{n} f_i = 34$$

Entonces:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N} = \frac{157}{34} = 4.6176$$

b)

El valor que más abunda es f2 = 11. Por lo tanto, la moda es 4.

c)

x_i	f_i	f_i
3	7	7
4	11	18
5	8	26
6	5	31
7	2	33
8	1	34

$$\frac{N}{2} = \frac{34}{2} = 17$$

Se busca, en las frecuencias acumuladas (Fi), el valor inmediatamente superior a 17. Por lo tanto, la mediana es 4.

4. La cantidad de cursos que llevan unos alumnos en la universidad se distribuyen de la siguiente manera:

x_i	f_i	F_i	h_i
1	5		0.0625
2	7		
3		23	0.1375
4	14		0.1750
5	18		
6		67	
7	6	73	
8			

a) Completar la tabla.

b) Calcular la media aritmética.

c) Calcular la moda.

d) Calcular la mediana.

Solución

a)

En la 1ra fila: $$F_1 = f_1 = 5$$

$$N = \frac{f_1}{h_1} = \frac{5}{0.0625} = 80$$

En la 2da fila:

$$F_2 = F_1 + f_2 = 5 + 7 = 12$$

$$h_2 = \frac{f_2}{N} = \frac{7}{80} = 0.0875$$

En la 3ra fila:

$$f_3 = F_3 - F_2 = 23 - 12 = 11$$

En la 4ta fila:

$$F_4 = F_3 + f_4 = 23 + 14 = 37$$

En la 5ta fila:

$$F_5 = F_4 + f_5 = 37 + 18 = 55$$

$$h_5 = \frac{f_5}{N} = \frac{18}{80} = 0.2250$$

En la 6ta fila:

$$f_6 = F_6 - F_5 = 67 - 55 = 12$$

$$h_6 = \frac{f_6}{N} = \frac{12}{80} = 0.1500$$

En la 7ma fila:

$$h_7 = \frac{f_7}{N} = \frac{6}{80} = 0.0750$$

En la 8va fila:

$$F_8 = N = 80$$

$$f_8 = F_8 - F_7 = 80 - 73 = 7$$

$$h_8 = \frac{f_8}{N} = \frac{7}{80} = 0.0875$$

Finalmente:

x_i	f_i	F_i	h_i
1	5	5	0.0625
2	7	12	0.0875
3	11	23	0.1375
4	14	37	0.1750
5	18	55	0.2250
6	12	67	0.1500
7	6	73	0.0750
8	7	80	0.0875

b)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N} = \frac{368}{80} = 4.6$$

c)

El valor que más abunda es $f_5 = 18$. Por lo tanto, la moda es 5.

d)

$$\frac{N}{2} = \frac{80}{2} = 40$$

Se busca, en las frecuencias acumuladas (F_i), el valor inmediatamente superior a 40. Por lo tanto, la mediana es 5.

5. Las edades de los integrantes de varias familias se distribuye de la siguiente manera:

Edad	x_i	f_i	F_i
[0,10[	5	5	5
[10,20[	15	9	14
[20,30[	25	12	26
[30,40[	35	13	39
[40,50]	45	6	45

a) Calcular la media aritmética.

b) Calcular la moda.

c) Calcular la mediana.

Solución

a)

Se sabe:

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

Donde:

$$N = \sum_{i=1}^{n} f_i = 45$$

Entonces:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N} = \frac{1{,}185}{45} = 26.33$$

b)

El mayor valor de frecuencia absoluta es f4 = 13. Debido a que se trata de variables discretas agrupadas, se aplica lo siguiente:

$$Mo = L_i + c_i \left(\frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \right)$$

Donde:

Li: Límite inferior de la clase donde se encuentra el mayor valor de frecuencia absoluta.

ci: Amplitud del intervalo donde se encuentra el mayor valor de frecuencia absoluta.

f_{i-1}: Frecuencia absoluta anterior.

f_{i+1}: Frecuencia absoluta posterior.

Reemplazando:

$$Mo = 30 + 10 \left(\frac{13 - 12}{(13 - 12) + (13 - 6)} \right) = 31.25$$

c)

$$\frac{N}{2} = \frac{45}{2} = 22.5$$

Se busca, en las frecuencias acumuladas (Fi), el valor inmediatamente superior a 22.5. Debido a que se trata de variables discretas agrupadas, se aplica lo siguiente:

$$Me = L_i + c_i \left(\frac{\frac{N}{2} - F_{i-1}}{f_i} \right)$$

Donde:

L_i: Límite inferior de la clase donde se encuentra el valor inmediatamente superior de la frecuencia absoluta acumulada.

c_i: Amplitud del intervalo donde se encuentra el valor inmediatamente superior de la frecuencia absoluta acumulada.

F_{i-1}: Frecuencia absoluta acumulada anterior.

Reemplazando:

$$Me = 20 + 10 \left(\frac{22.5 - 14}{12} \right) = 27.08$$

CAPÍTULO V

MEDIDAS DE DISPERSIÓN

5.1. INTRODUCCIÓN

El capítulo V denominado *Medidas de dispersión,* desarrolla los siguientes temas: El rango, la varianza y la desviación estándar y la desviación estándar y la distribución normal.

Ahora bien, las medidas de tendencia central tienen como objetivo resumir los datos en un valor representativo, las medidas de dispersión nos dice hasta qué punto estas medidas de tendencia central son representativas como síntesis de la información.

Las medidas de dispersión tiene como función a determinar los próximos o alejados que están los datos de la muestra de un punto central. Estas medidas de dispersión indican el grado de variabilidad que hay en la muestra y la representatividad de dicho punto central.

Para Veliz (2000, p.49), "(…) dos grupos diferentes de datos pueden tener iguales medidas de tendencia central; sin embargo, las características de su distribución pueden ser diferentes. Un grupo de datos puede tener mayor o menor dispersión que el otro con respecto dela medida central. Precisamente, para interpretar mejor los datos se construyen medidas de dispersión o estadígrafos de dispersión. Las medidas de dispersión ilustran sobre la manera como varían los datos observados alrededor de una medida de tendencia central, indican como están concentrados los datos alrededor del parámetro de centralización, permite comparar una información con otra y ayudan a verificar si determinadas medidas de tendencia central son o no significativas. Por ejemplo, cuando la dispersión es muy grande la media aritmética no tiene mucha significación; sin embargo, si la dispersión es baja, la media adquiere significación. Entre las medidas de dispersión están el recorrido o rango, la varianza, la desviación estándar, el coeficiente de variación, los cuartiles, etc."

Explica Quispe (2010, p. 99) que se llaman medidas de dispersión al grado en que los datos numéricos tienden a extenderse alrededor de un valor medio o central. Existen algunas medidas de dispersión que son muy usadas en estadística que permiten corregir una serie de valores numéricos en estudios de gran importancia.

En esa misma lógica enseña Ritchey (2008, pp. 136-137), que "la dispersión es la forma como se distribuyen las puntuaciones de una variable de intervalo/razón de menor a mayor y la forma de la diseminación entre éstas. Así la dispersión explica la forma como se dispersan las puntuaciones de una variable de intervalo/razón de menor a mayor y la forma de la distribución de estas. Existe un número infinito de posibles formas de distribución para una variable con una media dada. Todas las puntuaciones podrían agruparse alrededor de la media con la clara forma de una curva de campana, aunque

la curva podría ser de diferentes tamaños según el tamaño de la muestra, o bien, las puntuaciones podrían estar ligeramente o muy sesgadas hacia un lado. Además, una sola variable puede tener dispersiones muy diferentes de una población a otra. Por ejemplo, el ingreso familiar anual de residentes en Estados Unidos varía desde cero hasta decenas de millones de dólares, mientras que el ingreso familiar de los pobres que viven en proyectos habitacionales va de cero a unos pocos de miles de dólares."

Las variables de intervalo tienen características de una unidad numérica de medición definida, es decir, identifican las diferencias en monto, cantidad, grado o distancia y se les asigna puntuaciones numéricas.

5.2 RANGO

El rango de un conjunto de datos, es de uso muy limitado. Así, es la diferencia entre el dato mayor y menor. Según Ritchey (2008, 138), "el rango es una expresión de como las puntuaciones de una variable de intervalo/ razón se distribuyen de menor a mayor, es decir, es la distancia entre las puntuaciones más elevada y el valor más bajo de una muestra. Se calcula como la diferencia entre las puntuaciones máximas y la mínima de una muestra, más el valor de la unidad de redondeo. El valor de la unidad de redondeo (1, por ejemplo, si las puntuaciones se redondean al número entero más cercano, 0.1 si las puntuaciones se redondean al décimo más cercano, y así sucesivamente) se suma para considerar el límite real inferior de la puntuación más baja y el límite real superior de la puntuación más alta."

Cálculo del rango de una variable x de intervalo/razón

Enseña Ritchey (2008, p.138) cuatro pasos a seguir:

1. Ordenar las puntuaciones de la distribución de menor a mayor
2. Identifica las puntuaciones mínima y máxima
3. Identifica el valor de la unidad de redondeo
4. Calcular el rango:

   Rango = (Puntuación máxima – puntuación mínima) + valor de la unidad de redondeo.

   Por ejemplo: En el aula 310 de la Escuela Académico Profesional de Ciencia Jurídica de la Universidad Nacional Mayor de San Marcos se encuentran reunidos 7 alumnos de Ciencia Jurídica, se pide calcular la edad promedio de dichos alumnos.

   Por tanto X= Edad

   La siguiente distribución:

   21, 23, 46, 27, 20, 21,25

   Paso 1: Ordenar las puntuaciones:

   20, 21, 21, 23, 25, 27, 46

Paso 2: Identificar la puntuación máxima y la mínima.

Max: 46
Min: 20

Paso 3: Identificar la unidad de redondeo

Como la variable X= EDAD, la unidad de redondeo es 1

Paso 4: Calcular el rango

Rango = (46-20)+ 1 = 27 años

Sin embargo, existen situaciones donde se pueden reportar errores en el rango:

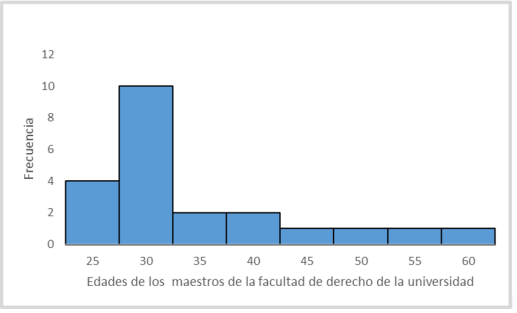
Puesto que el rango utiliza las puntuaciones más extremas de una distribución, un valor aislado inflará enormemente su cálculo. Esto sucedió para las siete edades indicadas anteriormente. Los 46 años hicieron que el rango pareciera estar extendido por encima de los 27 años. Reportar esto daría la impresión de que la muestra tiene un número considerable de sujetos de 30 y 50 años.

Un reporte más exacto sería, que el estudiante de derecho de 46 años sea eliminado, por tanto las edades tendrían un rango de 8 años (27-20+1= 8 años); al omitir la edad de 46 es una forma razonable de ajustar el rango.

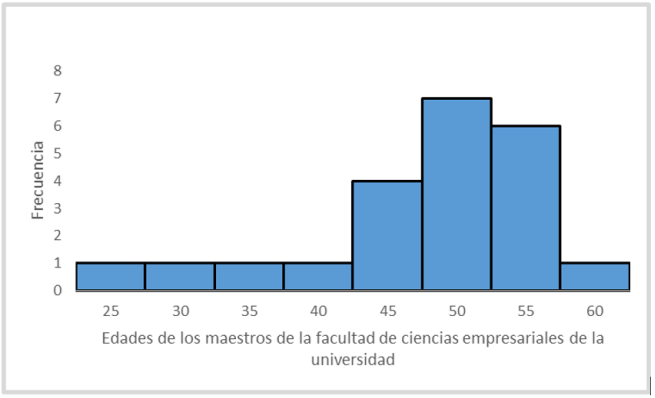
Otra de las limitaciones que presenta el rango es su estrecho alcance informativo, en cuanto a la forma de la distribución entre las puntuaciones extremas. Por ejemplo en la figura 1 presentamos la comparación entre las edades de los maestros de la Facultad de Derecho y Ciencia Jurídica de la Universidad Nacional Mayor de San Marcos y la Facultad de Ciencias Empresariales de la Universidad Nacional Mayor de San Marcos, las dos distribuciones descritas en la figura 1 tiene el mismo rango, lo que sugiere formas similares, pero de hecho sus formas son radicalmente diferentes.

Por último, hay poco que puede hacerse matemáticamente con el rango. En suma, el rango tiene utilidad limitada, en especial cuando se reporta solo.

Figura 1- Comparación de dos distribuciones con formas diferentes que tiene el mismo rango.



Elaboración propia



Elaboración propia

5.3 DESVIACIÓN ESTÁNDAR

Para Mode (1967, p. 87) la desviación estándar es la medida de variabilidad más importante y la que se usa con mayor frecuencia. Un valor relativamente pequeño de la desviación estándar implica concentración alrededor de la media aritmética, un valor relativamente grande, gran dispersión alrededor de la media aritmética. Una razón poderosa de su utilidad se debe al hecho de que las sumas de cuadrados se prestan fácilmente a manipulaciones algebraicas simples y producen relaciones interesantes y útiles. Una suma de valores absolutos. La desviación estándar constituye una unidad estadística conveniente para ser empleada en la construcción de otras medidas y para comparaciones entre ellas. En otras palabras, muchas medidas se expresan en términos de la desviación estándar como unidad.

Así, nos ilustra Ritchey (2008, p.140), que "la desviación estándar es un estadístico de dispersión. La desviación estándar es la raíz cuadrada de la varianza. En contraste, la desviación estándar describe la forma en que las puntuaciones de una variable de intervalo/ razón se dispersan a lo largo de la distribución en relación con la puntuación media. Así, la media es un estadístico de tendencia central y como tal proporciona un punto de enfoque que se centra "dentro" de la distribución. Observar la dispersión a partir de la media con su desviación estándar es como mirar desde el centro de la cancha, el centro de atención está en la distancia del centro de la cancha a otros puntos en cualquier dirección. Al igual que la media, la desviación estándar es muy apropiada con variables de intervalo/ razón. Para una variable de intervalo/ razón, la desviación estándar se calcula determinando qué tan alejada está cada puntuación de la media, es decir, cuánto se desvía de la media. En este sentido, la desviación estándar es una derivada (o producto) de la media, y las dos medidas siempre se reportan juntas. La desviación estándar, como una medida sumaria de todas las puntuaciones de una distribución, nos dice con qué amplitud se agrupan las puntuaciones alrededor de la media. También es útil junto con la curva normal."

Cálculo de la desviación estándar, vamos a seguir lo desarrollado por Ritchey (2008, pp. 140-141):

$$S_x = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

Donde:

S_x=desviación estándar para la variable X de intervalo/razón

Se analizará brevemente como se obtiene el cálculo de la desviación estándar.

Paso 1: Identifica las especificaciones

Comenzamos por identificar la información dada.

X=variable de intervalo/razón

n=tamaño muestral, y una distribución de puntuaciones en bruto para X.

Paso 2: Calcular la media

Calculamos la media porque la desviación estándar está diseñada para medir la dispersión alrededor de la media.

$$\bar{X} = \frac{\sum X}{n}$$

Paso 3: Calcular las puntuaciones de desviación

A continuación determinamos qué tan alejada está la puntuación de cada individuo respecto a la media. La diferencia entre una puntuación y su media se llama puntuación de desviación, es decir, cuanto difiere o se "desvía" de la media una puntuación individual:

$X\text{-}\bar{X}$=Puntuación de desviación para un valor de X

Paso 4: Sumar las puntuaciones de desviación

El siguiente paso para calcular la desviación estándar es sumar las puntuaciones de desviación. Esta suma siempre será igual a cero.

$\sum(X\text{-}\bar{X}) = 0$=suma de las puntuaciones de desviación

Paso 5: Elevar al cuadrado las puntuaciones de desviación y suma los cuadrados

La dispersión de una variable a menudo se compara para dos o más muestras.

Variación= $\sum(X - \bar{X})^2$

Divide la suma de cuadrados entre n-1 para ajustar el tamaño y el error de la muestra: pensamiento proporcional

En resumen, dividimos la variación (suma de cuadrados) entre *n-1* para compensar tanto los efectos del tamaño del tamaño muestral de la suma como el error de muestreo. El resultado se llama varianza, y su simbología es S_{X^2}.

$$S_{x^2} = \frac{\sum(X - \bar{X})^2}{n - 1} = Varianza$$

5.4 VARIANZA

La varianza es la variación promedio de las puntuaciones en una distribución. Medida estadística que mide la dispersión de los valores respecto a un valor central (media); es decir, es la media de la suma de cuadrados. Debe ser siempre positiva.

Cálculo de la varianza

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

La varianza es perfectamente aceptable para cálculos, pero no se interpreta de manera directa porque las unidades de medida están elevadas al cuadrado.

Calcular la desviación estándar

Para generar una medida de dispersión se necesita de un último paso. Sacar la raíz cuadrada de la varianza para obtener la desviación estándar. El resultado es la desviación estándar.

$$S_x = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}} = \sqrt{S_{x^2}}$$

En resumen, los elementos de la ecuación de la desviación estándar son las puntuaciones de desviación, la suma de cuadrados o variación y la varianza.

Por Ejemplo:

Los alumnos de la Escuela Profesional de Ciencia Jurídica de la Universidad Nacional Mayor del tercer ciclo desean participar en el campeonato interno entre escuelas, por ello han seleccionado a sus 12 mejores jugadores de fútbol con sus respectivos pesos. Se pide calcular la desviación estándar.

Paso 1: Identifica la variable *X*

X=peso de jugadores de fútbol (kg)

Jugador	Peso (kg)
1	75
2	78
3	80
4	80
5	64
6	75
7	75
8	69
9	72
10	81
11	65
12	71
n=12	$\sum$X=885 kg

Paso 2: Calcular la media

$$\overline{X} = \frac{\Sigma X}{n}$$

$$\bar{X} = \frac{885}{12} = 73.75\ kg$$

Paso 3: Calcular las puntuaciones de desviación

Puntuación de desviación jugador 1=X-$\bar{X}$=75-73.75=1.25 kg

Puntuación de desviación jugador 2=X-$\bar{X}$=78-73.75=4.25 kg

Jugador	Peso (kg)	X- $\bar{X}$
1	75	1.25
2	78	4.25
3	80	6.25
4	80	6.25
5	64	-9.75
6	75	1.25
7	75	1.25
8	69	-4.75
9	72	-1.75
10	81	7.25
11	65	-8.75
12	71	-2.75
n=12	$\sum$X=885 kg	$\sum$(X-$\bar{X}$)=0 kg

Paso 4: Sumar las puntuaciones de desviación

$$\sum(X - \bar{X}) = 0$$

Paso 5: Elevar al cuadrado las puntuaciones de desviación y suma los cuadrados.

Jugador	Peso (kg)	$X- \bar{X}$	$(X- \bar{X})^2$
1	75	1.25	1.5625
2	78	4.25	18.0625
3	80	6.25	39.0625
4	80	6.25	39.0625
5	64	-9.75	95.0625
6	75	1.25	1.5625
7	75	1.25	1.5625
8	69	-4.75	22.5625
9	72	-1.75	3.0625
10	81	7.25	52.5625
11	65	-8.75	76.5625
12	71	-2.75	7.5625
n=12	$\sum$X=885 kg	$\sum(X-\bar{X})$=0 kg	$\sum(X-\bar{X})^2$ =32.57 kg

Peso de jugadores de fútbol del tercer ciclo de la Escuela Profesional de Ciencia Jurídica de la Universidad Nacional Mayor de San Marcos

Paso 6: Calcular la varianza

$$S_{X^2} = \frac{\sum(X - \bar{X})^2}{n - 1}$$

$$S_{X^2} = \frac{32.57}{11} = 5.70\ kg$$

Paso 7: Calcular la desviación estándar

En el caso del peso del equipo de fútbol de los alumnos del tercer ciclo de la Escuela Profesional de Ciencia Jurídica la desviación estándar sería 5.70 kg.

$$S_x = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}} = \sqrt{S_{x^2}}$$

$$\sqrt{1\ 391.45} = 37.30\ libras$$

5.4.1 Limitaciones de la desviación estándar

Ritchey (2008, p.145), enseña que como la desviación estándar se calcula a partir de la media, al igual que ésta se infla por los valores extremos. Estos generan puntuaciones con grandes desviaciones. Cuando se elevan al cuadrado, estas puntuaciones, ya sean positivas o negativas, producen un alto resultado positivo e inflado. Así, la desviación estándar puede ser muy confusa cuando se reporta para una distribución sesgada, en la que pocas puntuaciones se extienden en una dirección.

¿Por qué se llama desviación "estándar"?

La desviación estándar recibe ese nombre por el hecho que proporciona una unidad de medida común para comparar variables con unidades observadas de medida muy diferentes.

Por ejemplo María y Eduardo solicitan una beca con base a su desempeño en los exámenes de admisión de la universidad.

María contestó la prueba académica de la Universidad Nacional Mayor de San Marcos y obtuvo 26 puntos.

Eduardo hizo lo propio con la prueba de admisión de la y obtuvo 900 puntos en la Universidad Nacional de Ingeniería.

Los dos resultados de las pruebas tiene unidades de medida muy diferentes: los puntos de la prueba de la Universidad Nacional Mayor de San Marcos van de cero a 36; y los de la prueba de la Universidad Nacional de Ingeniería, de 200 a 1600.

Con los siguientes estadísticos, encontramos que, en comparación con otros aspirantes que contestan las pruebas, María obtuvo la más alta.

X=*puntuación de la prueba SM*

$\bar{X}$=*22 puntos SM*

S_X=*2 puntos SM*

Y=*puntuación de la prueba UNI*

$\bar{Y}$=*1000 puntos UNI*

S_Y=*100 puntos UNI*

La puntuación de SM de 26 que obtuvo María tiene una desviación estándar de 2 arriba de la media de aquellos que toman la prueba SM, es decir, su puntuación está 4 puntos SM, esto es, 2 por 2 desviaciones estándar sobre el promedio de 22. La puntuación de Eduardo es de 1 desviación estándar debajo de la media de

aquellos que contrastan la prueba UNI , es decir, su puntuación está 100 puntos UNI, 1 desviación estándar abajo del promedio de 1 000. Sin lugar a dudamos podemos otorgarle la beca a María.

$$Z_X = \frac{X - \bar{X}}{S_X} \rightarrow \quad Z_X = \frac{26 - 22}{2} \rightarrow Z_X = 2.00 \ SD.$$

La puntuación Z de Mary es: 2.00 SD.

Puntuaciones estandarizadas (puntuaciones Z)

Hay tres formas de expresar el valor de cualquier puntuación de una variable de intervalo/razón:

a. Puntuación en bruto: Aquellas que expresamos en sus unidades de medidas observadas, originales.
b. Puntuación de desviación: Lo que expresamos como una desviación de la media
c. Puntuación estandarizada o puntuación Z: Expresamos su puntuación como un número de desviación estándar de le media de la puntuación

Cálculo de puntuaciones estandarizadas

$$Z_X = \frac{X - \bar{X}}{S_X}$$

Donde:

Z_X=Puntuación estandarizada para un valor de x

X=Variable de intervalo razón

$\bar{X}$=Media de X

S_X=Desviación estándar de X

Si hacemos que la puntuación *X=SM con $\bar{X}$=22 puntos,SM y S_X=2 puntos SM,la puntuación Z de Mary es:*

$$Z_X = \frac{X - \bar{X}}{S_X} = \frac{26 - 22}{2} = 2.00 \ SD.$$

Donde SD significa "desviación estándar". Por tanto una puntuación Z es la distancia de una puntuación X hacia la media dividida entre la desviación estándar de las distancias.

5.4.2 La desviación estándar y la distribución normal

La desviación estándar es una herramienta estadística tan valiosa que es una parte matemática de la curva normal. Cuando se sigue la curva desde su centro en cualquier dirección, la curva cambia de forma para aproximarse al eje *X*. Desde el pico, el punto en el que la curva empieza a desplazarse hacia fuera es 1 desviación estándar desde la media. Ese punto recibe el nombre de punto de inflexión de la curva (*Ver figura 2*).

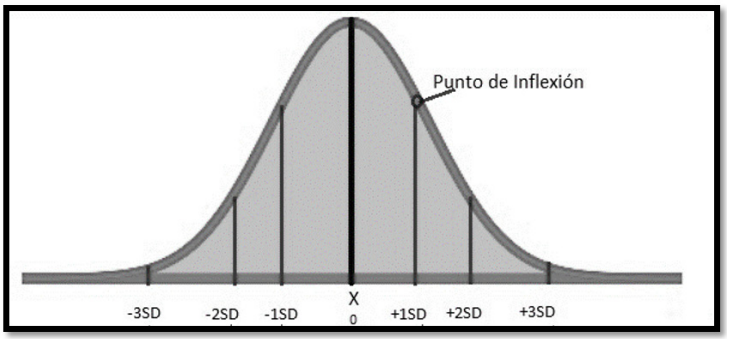


Figura 2. Relación de entre desviación estándar y la curva normal

Elaboración propia.

Uno de los rasgos más importantes del fenómeno de normalidad, que ocurre naturalmente, es que ofrece predicciones precisas sobre cuantas puntuaciones de una población cae dentro de cualquier rango de puntuaciones (*Ver figura 3*).

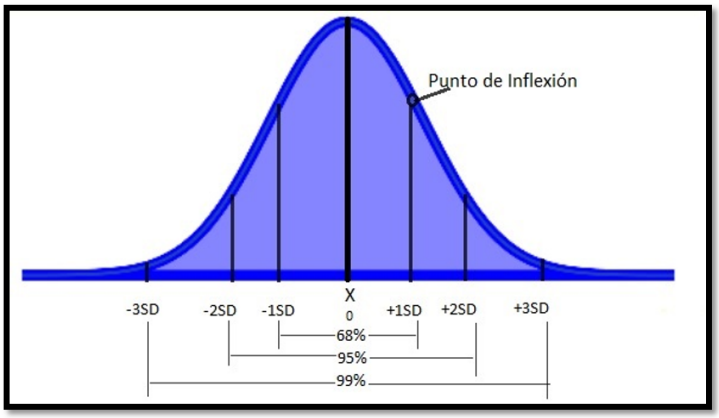
a. 50% de las puntaciones caen arriba de la media y 50% caen debajo.

b. Todas las puntuaciones (99.7%) caen no más de tres desviaciones estándar de la media en ambas direcciones.

c. Alrededor del 95% de las puntuaciones caen no más de dos desviaciones estándar de la media en ambas direcciones.

d. Alrededor del 68% de las puntuaciones de una variable normalmente distribuida caen a no más de 1 desviación estándar de la media en ambas direcciones.

Figura 3. Uso de la curva normal para estimar la distribución



Elaboración propia

Si la desviación estándar es mayor que la media, la distribución de puntuaciones no puede tener forma normal. Es probable que un histograma de la variable deje ver un sesgo o una distribución de puntuaciones de forma extraña.

5.5 PREGUNTAS DE REPASO

1. ¿Cuáles son las medidas de dispersión?
2. ¿Defina el rango?
3. ¿Defina la desviación estándar?
4. ¿Defina la variancia?

5.6 RESPUESTAS

1. Rango, desviación estándar y varianza
2. El rango es la puntuación máxima – puntuación mínima+ el valor de la unidad de redondeo.
3. La desviación estándar es la medida de variabilidad más importante y la que se usa con mayor frecuencia. Un valor relativamente pequeño de la desviación estándar implica concentración alrededor de la media aritmética, un valor relativamente grande, gran dispersión alrededor de la media aritmética.
4. La varianza es la variación promedio de las puntuaciones en una distribución. Medida estadística que mide la dispersión de los valores respecto a un valor central (media); es decir, es la media de la suma de cuadrados. Debe ser siempre positiva.

CAPÍTULO VI

LA TEORÍA DE LAS PROBABILIDADES

6.1 INTRODUCCIÓN

El capítulo V denominado *La teoría de las probabilidade*s, presenta los siguientes temas: la toma de decisiones frente a situaciones problemáticas, experimento aleatorio, espacio muestral, eventos, operaciones con eventos, probabilidad, probabilidad condicional, principios de multiplicación, eventos independientes, probabilidad total, teorema de Bayes, variables aleatorias , variables aleatorias discretas, esperanza matemática.

6.2 LA TOMA DE DECISIONES FRENTE A SITUACIONES PROBLEMÁTICAS

En muchos aspectos de la vida cotidiana necesitamos tomar decisiones frente a situaciones problemática.

Por ejemplo, a un candidato a la presidencia no le llamará mucho la atención si una persona vota por él, pero si le gustaría saber el porcentaje de personas que votarán por él.

Para un técnico tendrá sentido conocer cuál es la probabilidad de que un mecanismo diseñado por él funcione más de 500 horas en lugar de saber el tiempo total de duración.

Esto sucede debido a que se tiene que contar con fundamentos de experimentos para decidir correctamente.

En el transcurso del desarrollo del capítulo iremos conociendo los fundamentos de las probabilidades y sus diversas aplicaciones.

Por lo que, al finalizar, el estudiante estará en la capacidad de construir el espacio muestral que corresponde a un experimento aleatorio en problemas contextualizados, definir y realizar operaciones entre eventos en problemas contextualizados y calcular probabilidades simples y condicionales de eventos en juegos de azar, en problemas de producción, control de calidad o bajo otro contexto.

6.3 EXPERIMENTO ALEATORIO (ε):

Es un experimento que se puede repetir indefinidamente, sin la necesidad de cambiar sus condiciones. Los resultados de estos experimentos no se pueden predecir con exactitud antes de realizarlos, pero si se pueden describir sus resultados posibles.

6.4 ESPACIO MUESTRAL (Ω)

Es el conjunto de todos los resultados posibles de un experimento aleatorio. A veces se le denomina espacio muestral Ω asociado al experimento Ɛ.

A continuación, se mostrará diversos ejemplos para la compresión de los conceptos:

Ɛ1: Lanzar una moneda y observar la cara superior.

Ω(Ɛ1): {Cara, Sello}

Ɛ2: Lanzar un dado y observar el número que aparece.

Ω(Ɛ2): {1, 2, 3, 4, 5, 6}

Ɛ3: Extraer un artículo de una línea de producción y comprobar si es defectuoso o no.

Ω(Ɛ3): {Defectuoso, No Defectuoso}.

6.5 EVENTOS

Si hemos definido al espacio muestral como todos los posibles resultados de un experimento aleatorio, por lo que vendría el universo o conjunto universal. Por lo tanto, el evento es cualquier subconjunto de un espacio muestral y se denotan por A, B, C, etc.

Igualmente, el espacio muestral Ω y el evento nulo Φ (evento imposible) son considerados eventos.

Al espacio muestral (Ω) se le denomina evento seguro, porque siempre ocurre, es imposible que no pueda ocurrir.

Se dice que un evento ocurre cuando al menos uno de los resultados que lo conforman ocurre.

6.6 OPERACIONES CON EVENTOS:

Mostraremos las diferentes opciones de operaciones que se pueden dar con los eventos dentro de un espacio muestral.

Sean:

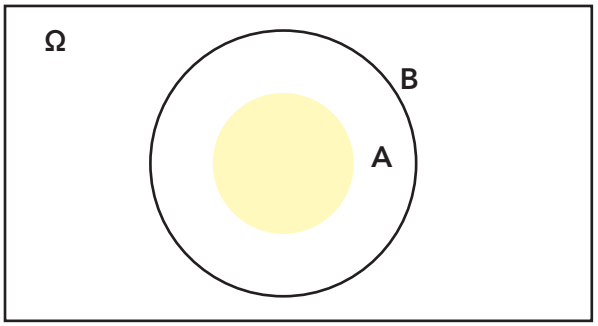
A, B: Eventos

Ω: Espacio Muestral

Tenemos que:

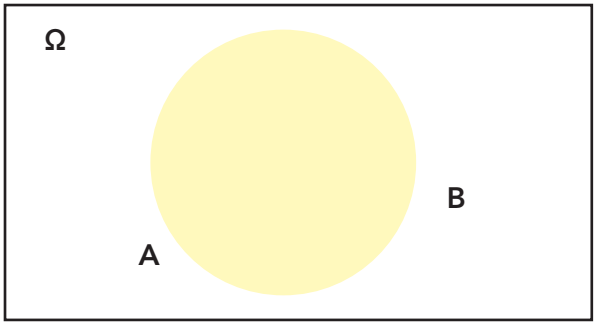
a. Si A se está incluido dentro de B, ósea si los resultados del evento A ocurrirán en el evento B. Se denota de la siguiente manera: $A \subset B$.

Gráficamente:



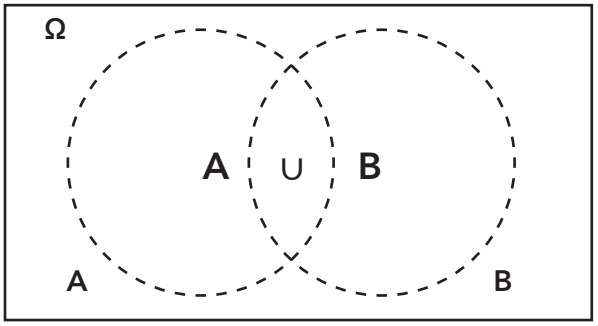
b. Si A es igual que B, ósea si los resultados del evento A coinciden en el evento B. Se denota de la siguiente manera: A = B y por lo tanto se cumple que: $A \subset B$ y $B \subset A$.

Gráficamente:



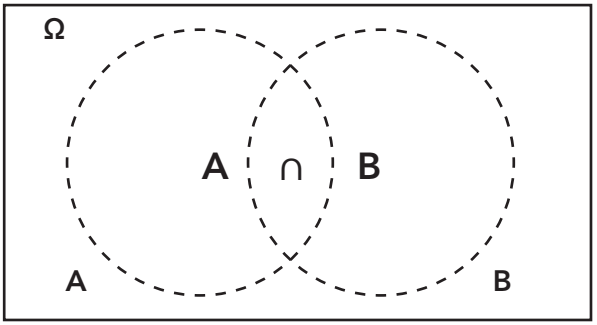
c. Si queremos unir los resultados del evento A y B, comúnmente llamado como la unión de eventos. Se denota como $A \cup B$.

Gráficamente:



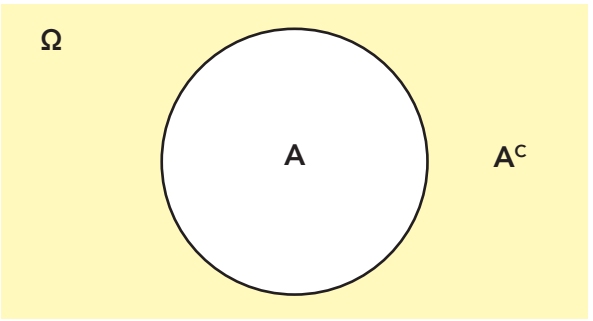
d. Si deseamos realizar una intersección entre los resultados del evento A y B, comúnmente llamado como la intersección de eventos. Se denota como A ∩ B.

Gráficamente:



e. Si tenemos un evento A dentro de un espacio muestral, el complementa de dicho evento es todo lo que no le pertenece. Se denota como: AC

Gráficamente:

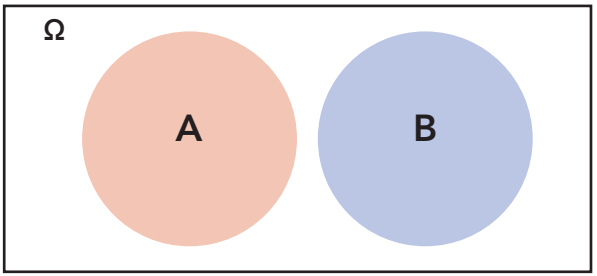


Adicionalmente se tiene algunas operaciones adicionales que cuentan con condiciones propias, la cuales son:

a. Eventos Mutuamente Excluyentes:

Sean A y B eventos de un espacio muestral, donde A ∩ B = Φ.

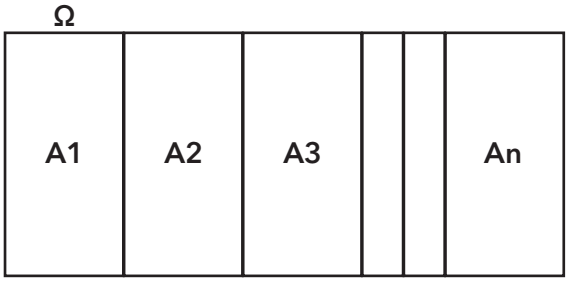
Gráficamente:



b. Eventos Colectivamente Exhaustivos:

Sean A1, A2, A3,…An eventos de un espacio muestral donde A1 ∪ A2 ∪ A3---∪ An = Ω.

Gráficamente:



Ejemplos:

Sea el experimento:

Σ: Lanzar dos dados y observar los números que aparecen en las caras superiores.

Tenemos que el espacio muestral será:

Ω (Σ):

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

Sea el evento:

A = Los números en las caras superiores son iguales.

A = {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}

Por lo tanto, N(A): 6

1. Se tiene un experimento aleatorio que consiste en lanzar dos monedas a la vez. Obtenga el espacio muestral.

   Primera moneda: {Cara, Sello}

   Segunda moneda: {Cara, Sello}

   Entonces:

   Ω = {(CC), (CS), (SC), (SS)}

2. Se tiene el experimento aleatorio que consiste en lanzar tres monedas. Obtenga el espacio muestral.

   Primera moneda: {Cara, Sello}

Segunda moneda: {Cara, Sello}

Tercera moneda: {Cara, Sello}

Entonces:

Ω = {(CCC), (CCS), (CSC), (CSS), (SCC), (SCS), (SSC), (SSS)}

3. Se tiene el experimento aleatorio que consiste en lanzar una moneda y un dado. Obtenga el espacio muestral.

   Solución:

   N(Ω): 2 x 6 = 12

   Podemos trabajarlo con una tabla de doble entrada:

Moneda	Dado					
	1	2	3	4	5	6
Cara	C,1	C,2	C,3	C,4	C,5	C,6
Sello	S,1	S,2	S,3	S,4	S,5	S,6

   Entonces:

   Ω = {(C1), (C2), (C3), (C4), (C5), (C6), (S1), (S2), (S3), (S4), (S5), (S6)}

6.7 PROBABILIDAD:

Si N(A) representa el número de elementos del evento A y N(Ω), el número de elementos del espacio muestral, entonces la probabilidad del evento A esta dada por:

P(A): N(A) / N(Ω)

Las probabilidades cuentan con axiomas, las cuales son:

- Primer Axioma: La probabilidad de un evento A ∈ Ω, solo puede tomar valores entre cero y uno (0 ≤ P(A) ≥ 1).
- Segundo Axioma: La probabilidad del espacio muestral es igual a la unidad. Esto es, P(Ω) = 1.
- Tercer Axioma: Para un número finito de k eventos mutuamente excluyentes, A1, ..., Ak que pertenecen a un Ω. La probabilidad de la unión de estos eventos está dada por:

$$P(A1) + P(A2) + P(A3) \ldots. + P(Ak) = P(\Omega)$$

Propiedades de Probabilidad:

Se tiene que:

1. Si Φ es el evento imposible o nulo, entonces $P(\Phi) = 0$.

Para cada evento $A \in \Omega$, se tiene:

$$P(A) = 1 - P(Ac)$$

2. Si A y B $\in \Omega$, tales que $A \subset B$, entonces $P(A) \leq P(B)$.

3. Si A y B $\in \Omega$, eventos cualesquiera, entonces:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \text{ --- Principio de Adición}$$

Ejemplo:

A continuación, se muestra a 50 profesionales de una empresa agrupados según especialidad por sexo.

Especialidad	Sexo		Totales
	Masculino (M)	Femenino (F)	
Ingeniería Industrial (I)	15	4	19
Ingeniería Informática (S)	12	8	20
Ingeniería Civil (C)	8	3	11
Totales	35	15	50

La empresa desea seleccionar un profesional para ocupar un cargo directivo, calcule:

a. La probabilidad de que sea de la especialidad de ingeniería industrial.

$$P(I) = 19/50 = 0.38$$

b. La probabilidad de que sea mujer y de ingeniería civil.

$$P(M \cap C) = 3/50 = 0.06$$

c. La probabilidad de que sea de la especialidad de ingeniería civil o ingeniería industrial.

$$P(C \cup I) = 30/50 = 0.60$$

d) La probabilidad de que sea hombre o Ingeniero informático.

$$P(H \cup S) = 43/50 = 0.86$$

6.8 PROBABILIDAD CONDICIONAL:

En muchas ocasiones se pide calcular la probabilidad de que ocurra un evento sabiendo que otro evento ha ocurrido.

Definición

Para dos eventos cualquiera A y B con P(B)>0, la probabilidad condicional de A dado que ocurrió B está definida por:

P (A/B) = P (A ∩ B) / P(B), si P(B) > 0

Propiedades:

Si $A \cap B = \emptyset$, entonces, $P\ (A/B) = 0$

Si $A \subset B$, entonces, $P\ (A/B) = P(A)/\ P(B)$

Si $B \subset A$, entonces, $P(A/B) = P(B)\ /P(B) = 1$

Ejemplos:

1. Un abogado de una empresa manufacturera recibió un estudio sobre la pureza del aire, para esto se seleccionó 300 muestras de aire que las clasificó de acuerdo con la presencia de dos moléculas raras. Los resultados se muestran en la siguiente tabla:

Moléculas de las muestras de aire		Molécula 1		Totales
		SI	NO	
Molécula 2	SI	27	20	47
	NO	41	212	253
Totales		68	232	300

Si se selecciona una muestra de aire al azar, calcule.

a. La probabilidad de que la muestra presente una molécula tipo 1.

Por lo tanto, se declaran los siguientes eventos:

A: Muestra presenta la molécula tipo 1.

B: Muestra presenta la molécula tipo 2.

P(A): 68/300 = 0.2267

b. La probabilidad de que la muestra no presente una molécula tipo 2.

P(B) = 1 – P(B) = 1 – 47/300 = 1- 0.1566 = 0.8433

c. La probabilidad de que la muestra presente una molécula tipo 1 dado que muestra una molécula tipo 2.

Por lo tanto, se aplica el concepto de probabilidad condicional.

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

$= (27/300) / (47/300) = 0.5745$

d. Si se verificó que había presencia de la molécula tipo 1, calcule la probabilidad de que se encuentre la molécula tipo 2.

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

$= (27/300) / (68/300) = 0.3971$

e. Si se verificó que no había presencia de la molécula tipo 1, calcule la probabilidad de que se encuentre la molécula tipo 2.

$P(B/A^c) = P(B \cap A^c) / P(A^c)$

$= (20/300) / (232/300) = 0.0862$

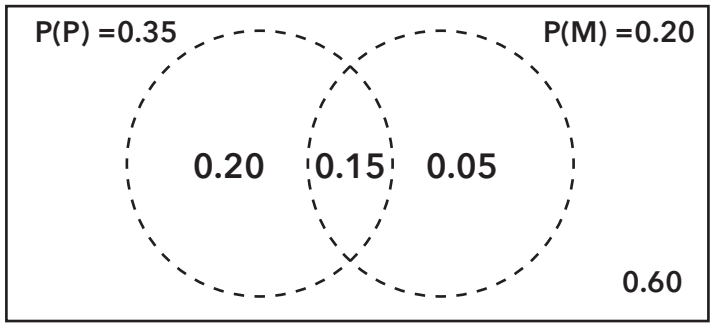
2. A un estudio de abogados llega las muestras de agua para detectar plomo y mercurio en el campamento minero Toromocho para una demanda constitucional. El 40% de las muestras presentan niveles tóxicos de plomo o mercurio, el 35% de plomo y el 15% de ambos metales.

Si se selecciona una muestra al azar, calcule la probabilidad de que:

Se tiene que:

P: Muestra tiene nivel toxico de Plomo.

M: Muestra tiene nivel toxico de Mercurio.



a. Tenga niveles tóxicos solamente de plomo.

   Respuesta: 0.20

b. No tenga niveles tóxicos de plomo y no tenga niveles de tóxicos de mercurio.

   Respuesta: 0.60

c. Tenga niveles tóxicos de mercurio dado que tiene niveles tóxicos de plomo.

   $P(M/P) = P(M \cap P) / P(P) = 0.15/0.35 = 0.4286$

d. Tenga niveles tóxicos de plomo si tiene niveles tóxicos de mercurio.

   $P(P/M) = P(P \cap M) / P(M) = 0.15/0.20 = 0.75$

6.9 PRINCIPIO DE MULTIPLICACIÓN

Para dos eventos A y B cualesquiera, se cumple:

$P(A \cap B) = P(B) \times P(A/B)$

Ejemplos

1. La probabilidad de que la batería de un avión sujeta a altas temperaturas dentro del comportamiento del motor reciba una corriente de carga mayor que la normal es 0.9; mientras que, la probabilidad de que la batería quede expuesta a altas temperaturas es 0.07.

Calcule la probabilidad de que la batería experimente tanto una corriente mayor que la normal como una temperatura alta.

Solución:

Tenemos que:

A: Batería experimenta una corriente de carga mayor que la normal.

B: Batería está expuesta a altas temperaturas.

$P(A/B) = 0.9$ $P(B) = 0.07$

Entonces, la probabilidad de que la batería experimente tanto una corriente de carga alta como una temperatura alta se calcula de la siguiente manera:

$P(A \cap B) = P(B) \times P(A/B) = 0.07 \times 0.9 = 0.063$

2. En una fábrica de botellas de plástico hay un lote que contiene 250 botellas de los cuales 25 son defectuosos. En un proceso de control de calidad, las botellas son seleccionadas una después de la otra para ver si son defectuosas.

a. ¿Cuál es la probabilidad de que las dos botellas seleccionadas sean defectuosas?

Sean los eventos:

A: Primer botella es defectuosa.

B: Segunda botella es defectuosa.

La probabilidad de seleccionar una preforma defectuosa es: P(A) = 25/250

La probabilidad que se seleccione otra preforma defectuosa dado que ya se seleccionó una defectuosa en la primera toma es: P(B/A) = 24/249

Por el principio de multiplicación, se tiene:

$P(A \cap B) = P(A) \times P(B/A) = 25/250 \times 24/249 = 600/62250 = 0.0096$

a. Si se selecciona otra botella, ¿cuál es la probabilidad de que las tres preformas sean defectuosas?

Sean los eventos:

A: Primer botella es defectuosa.

B: Segunda botella es defectuosa.

C: Tercera botella defectuosa.

$P(A \cap B \cap C) = P(A) \times P(B/A) \times P(C/A \cap B)$

$= 25/250 \times 24/249 \times 23/248 = 13800/15438000 = 0.000089$

6.10 EVENTOS INDEPENDIENTES:

Se dice que dos eventos A y B son independientes entre sí cuando la ocurrencia de uno de los eventos no afecta en nada la probabilidad de ocurrencia del otro.

Definición

Si dos eventos A y B son independientes, se cumple:

$P(A/B) = P(A)$, si $P(B) > 0$… (1)

Por otra parte, de la definición de probabilidad condicional, se tiene:

$P(A/B) = P(A \cap B) / P(B)$ … (2)

Luego, si A y B son eventos independientes, se cumple:

$P(A \cap B) = P(A) \times P(B)$ …. De 1 y 2

En otro caso se dice que los eventos A y B son dependientes.

Ejemplos:

1. En una empresa agroindustrial, el abogado corporativo recibió el control de calidad a un lote de producción de espárragos, para ello tomó una muestra de 100 unidades. En la muestra encontró que 40 espárragos cumplían con los estándares de calidad. Entre los que cumplían con los estándares de calidad, 18 eran espárragos blancos; mientras que, entre los que no cumplían los estándares de calidad, 45 eran morados.

¿Son los eventos "cumplen con los controles de calidad" y "tipo de esparrago morado" independientes?

Sean:

	Tipo de Esparrago		**Totales**
	Blanco	**Morado**	**NO**
Cumple con el control de calidad	18	22	**40**
No cumple con el control de calidad	15	45	**60**
Totales	**33**	**67**	**100**

C: Espárrago elegido cumple con los estándares de calidad

M: Esparrago elegido es de tipo morado

P (C) = 40/100 = 0.4

P (M) = 33/100 = 0.33

La probabilidad condicional de que cumplan con los estándares de calidad, dado que es de tipo blanco se obtiene de la siguiente manera:

$P(C/M) = P(C \cap M) / P(M) = (22/100) / (67/100) = 22/67 = 0.3284$

Como P (C/M) = 0.3284 es diferente a P(C) = 0.40, se concluye que los eventos C y M no son independientes.

2. Un motor de auto de carreras tiene cuatro pistones que funcionan de manera independiente, cuyas probabilidades de falla son, 0.06, 0.14, 0.25, 0.32 respectivamente.

Probabilidad de que falle:

P(A) = 0.06, P(B)= 0.14, P(C)= 0.25, P(D) = 0.32

Probabilidad de que no falle:

P(Ac) = 0.94, P(Bc) = 0.86, P(Cc) = 0.75, P(Dc) = 0.68

a) ¿Cuál es la probabilidad de que falle solo uno de los pistones?

P (A, Bc, Cc, Dc) = 0.06 x 0.86 x 0.75 x 0.68 = 0.026316

P (Ac, B, Cc, Dc) = 0.94 x 0.14 x 0.75 x 0.68 = 0.067116

P (Ac, Bc, C, Dc) = 0.94 x 0.86 x 0.25 x 0.68 = 0.137428

P (Ac, Bc, Cc, D) = 0.94 x 0.86 x 0.75 x 0.32 = 0.194016

P (falle solo uno de los pistones) = 0.424876

b) ¿Cuál es la probabilidad de que fallen solo dos de los pistones?

P (A, B, Cc, Dc) = 0.06 x 0.14 x 0.75 x 0.68 = 0.004284

P (Ac, B, C, Dc) = 0.94 x 0.14 x 0.25 x 0.68 = 0.0022372

P (Ac, Bc, C, D) = 0.94 x 0.86 x 0.25 x 0.32 = 0.064672

P (A, Bc, Cc, D) = 0.06 x 0.86 x 0.75 x 0.32 = 0.012384

P (Ac, Bc, C, Dc) = 0.06 x 0.86 x 0.25 x 0.68 = 0.008772

P (Ac, B, Cc, D) = 0.94 x 0.14 x 0.75 x 0.32 = 0.031584

P (falle solo dos de los pistones) = 0.144068

a) ¿Cuál es la probabilidad de que fallen a lo más dos de los pistones?

P (ningún pistón falle) = P (Ac, Bc, Cc, Dc) = 0.94 x 0.86 x 0.75 x 0.68 = 0.412284

P (X ≤ 2): P(X=0) + P(X=1) + P(X=2)

P (X ≤ 2): = 0.412284 + 0.424876 + 0.144068 = 0.981228

b) ¿Cuál es la probabilidad de que falle al menos uno de los pistones?

P (X ≥ 1) = P (X =1) + P (X=2) + P (X=3) + P (X=4)

También:

P (X ≥ 1) = 1 – P (X=0)

P (X ≥ 1) = 1 – 0.412284 = 0.587716

6.11 PROBABILIDAD TOTAL

El teorema nos permite calcular la probabilidad de un suceso a partir de probabilidades condicionadas.

Por lo tanto, si {Ei} es una colección de "k" eventos mutuamente excluyentes, donde P(Ei) > 0 y

$$P(E1) + P(E2) + P(E3) + \ldots\ldots + P(Ek) = 1$$

Entonces, para todo evento A, se tiene:

$$P(A) = \sum P(Ei) \times P(A/Ei)$$

6.12 TEOREMA DE BAYES

Es utilizado para calcular la probabilidad de un evento, contando con la información de antemano sobre el evento.

Podemos calcular la probabilidad de un suceso A, sabiendo además que ese A cumple cierta característica que condiciona su probabilidad. El teorema de Bayes entiende la probabilidad de forma inversa al teorema de la probabilidad total. El teorema de la probabilidad total hace inferencia sobre un suceso B, a partir de los resultados de los sucesos A. Por su parte, Bayes calcula la probabilidad de A condicionado a B

Por lo tanto, si {Ei} es una colección de "k" eventos mutuamente excluyentes donde P(Ei) > 0, entonces para todo evento A con P(A) > 0.

$$P(Er/A) = \frac{P(Er) \times P(A/Er)}{\sum P(Ei) \times P(A/Ei)}, \text{ donde r: } 1,2,\ldots k$$

Ejemplos:

Un comerciante adquiere sacos de arroz de tres molineras nacionales (M1, M2 y M3) en los siguientes porcentajes: el 20% de los sacos proviene de M1, el 30% de los sacos proviene de M2 y el resto de los sacos proviene de M3. Además, el 30% de los sacos que recibe de M1, el 80% de los sacos que recibe de M2 y el 70% de los sacos que recibe de M3 son de calidad excepcional.

Sean los eventos:

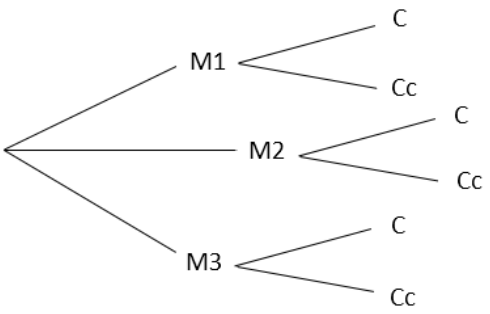
M1: El saco de arroz proviene de la molinera M1. → P(M1) = 0.2

M2: El saco de arroz proviene de la molinera M2. → P(M2) = 0.3

M3: El saco de arroz proviene de la molinera M3. → P(M3) = 0.5

C: El saco de arroz es de calidad excepcional.

Graficamos el diagrama del árbol:



a. Si un saco de arroz es escogido aleatoriamente. ¿Cuál es la probabilidad de que sea de calidad excepcional?

P(C) = P(M1) X P (C/M1) + P(M2) X P (C/M2) + P(M3) X P(C/M3)

P(C) = (0.2) * (0.3) + (0.3) * (0.8) + (0.5) * (0.7)

P(C) = 0.65

b. Si un saco de arroz es de calidad excepcional, ¿cuál es la probabilidad de que provenga de la molinera M2?

$$P(M2/C) = \frac{P(M2) * P(C/M2)}{P(C)} = \frac{(0.3) * (0.8)}{0.65} = 0.3692$$

1. En un bloque, el 70% de alumnos son mujeres, de ellas el 10% son fumadoras. De los varones, son fumadores el 20 %.

Sean los eventos:

M: Mujeres

H: Hombres

F: Fumado

a. ¿Cuál es la probabilidad de que un alumno tomado al azar sea fumador?

P(F): P(M) * P(F/M) + P(H) * P(F/H)

P(F): 0.7 *0.1 + 0.3 * 0.2 = 0.13

Se elige un individuo al azar y resulta fumador, ¿Cuál es la probabilidad de que sea un hombre?

$$P(H/F) = \frac{P(H) * P(F/H)}{P(H)*P(F/H) + P(M)* P(F/M)}$$

$$P(H/F) = (0.3 * 0.2) / 0.13 = 0.46$$

Dato: El primer caso de los ejemplos pertenece a Probabilidad Total ya que como nos dice el concepto, nos permite calcular la probabilidad a partir de probabilidades ya dadas. Mientras que el segundo caso de los ejemplos nos determina Teorema de Bayes, se calcula la probabilidad a partir de una condición ya dada.

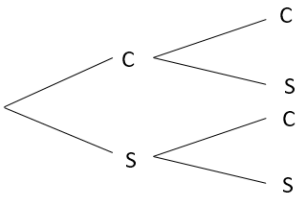
6.13 VARIABLES ALEATORIAS

Una Variable Aleatoria es una función que asigna un número a cada elemento en el espacio muestral.

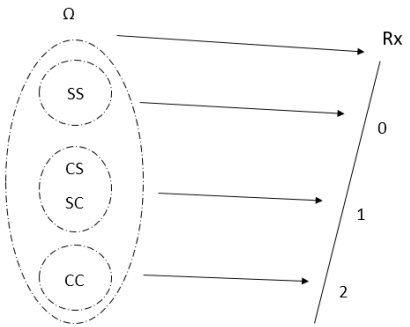
Por lo tanto, si una variable aleatoria es una función X: Ω → Rx entones Rx no es un Φ, tal que a cada elemento que pertenece al espacio muestral se le asocia un número real X ∈ Rx.

Daremos un ejemplo para entender la definición:

> Sea el experimento: Lanzar una moneda dos veces y sea la variable aleatoria X = Número de caras que se obtiene. Obtenga el espacio muestral y el recorrido de X.



Por lo tanto:



El espacio muestral es Ω = {CC, CS, SC, SS}

El recorrido de X está dado por Rx: {0,1,2}

Distribución de Probabilidad

Sea X: $\Omega \rightarrow Rx$ una variable aleatoria que toma los valores x1, x2, … Entonces P(Xi) es la distribución de probabilidad de la variable X, si a cada valor Xi se le asigna su respectiva probabilidad de ocurrencia, es decir:

$P(Xi) = P(X=xi) = P(w \in \Omega / X(w) = Xi)$

Función de Distribución

La función de distribución de una variable aleatoria X, denotada por F, es una función definida por:

$F(a) = P(X \leq a) = P(w \in \Omega / X(w) \leq a)$, para cualquier $a \in R$.

Es decir, F (x) es la probabilidad de que la variable aleatoria X tome valores menores o iguales que a. Además, F(x) está definida en todo R.

Propiedades:

La función de distribución es no decreciente.

Toda función de distribución es continua por la derecha.

6.14 VARIABLES ALEATORIAS DISCRETAS

Una variable aleatoria X: $\Omega \rightarrow Rx$ es discreta si su recorrido Rx es un conjunto contable finito o infinito numerable de números reales.

Función de Probabilidad:

La función de probabilidad conocida también como función de cuantía está definida de la siguiente manera:

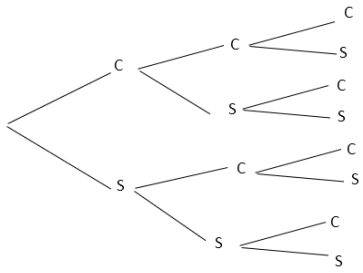
P(x) = P(X =x), si $x \in Rx$, mientras que 0 si x no pertenece a Rx.

Ejemplo:

Sea el experimento:

E1: Lanzar una moneda tres veces

Obtenga la distribución de probabilidad de la variable aleatoria: X = Número de caras que se obtiene, considere que ambos resultados son igualmente probables.

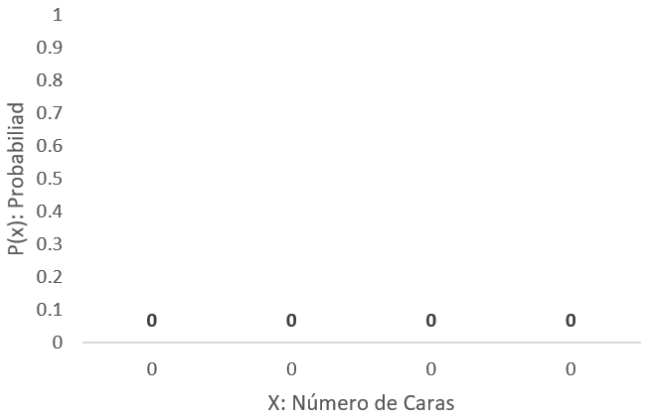


Obtenemos la distribución de probabilidad en base al espacio muestral:

Ω = {(CCC), (CCS), (CSC), (CSS), (SCC), (SCS), (SSC), (SSS)}

X	0	1	2	3
P(X)	1/8	3/8	3/8	1/8

Representaremos gráficamente:



Suponga que una compañía de cosméticos planea elaborar un nuevo perfume. El gerente de producto ha estimado las siguientes probabilidades subjetivas para las ventas del primer año (en millones de botellas):

X	0	1	2	3	4	5	6	7	8
P(X)	0.05	0.15	0.20	0.20	0.15	0.10	0.05	0.05	0.05

Encuentre las siguientes probabilidades:

$P(X < 3) = P(X=0) + P(X=1) + P(X=2) = 0.05 + 0.15 + 0.20 = 0.40$

$P(X > 5) = P(X=6) + P(X=7) + P(X=8) = 0.05 + 0.05 + 0.05 = 0.15$

$P(2 \leq X \leq 4) = P(X=2) + P(X=3) + P(X=4) = 0.20 + 0.20 + 0.15 = 0.55$

6.15 ESPERANZA MATEMÁTICA

Sea: $\Omega \rightarrow Rx$ una variable aleatoria discreta y p(x) su probabilidad por cada valor de x.

Se define la esperanza matemática o valor esperado de X, denotado por: E(X), está definida de la siguiente manera:

$E(X) = \sum x*p(x)$

Propiedades:

Si la variable aleatoria X toma un valor constante k, entonces: E(X) = k.

Si a y b son constantes y X una variable aleatoria cuya E(X) existe, entonces: E (a*X + b) = a*E(X) + b

6.15.1.- Varianza:

Sea: $\Omega \rightarrow Rx$ una variable aleatoria discreta y sea $\mu = E(x)$.

La varianza de X, denotado por V(X), está definida de la siguiente manera:

$V(X) = \sum (x - \mu)2 * p(x)$

Otra forma de obtener la varianza:

V(X) = E (X2) – E(X)2

Propiedades:

Si la variable aleatoria X toma un valor constante, entonces: V(X) = 0.

Si a y b son constantes y X una variable aleatoria cuya V(X) existe, entonces:

V (a*X + b) = a *V(X)

Ejemplo:

Calcule la esperanza y la varianza de la variable aleatoria. del ejemplo anterior.

X	0	1	2	3
P(X)	1/8	3/8	3/8	1/8

Para hallar la esperanza matemática o media, multiplicamos cada valor de x con su probabilidad dada para luego sumar los totales:

E(X): 0*(1/8) + 1*(3/8) + 2*(3/8) + 3*(1/8) = 1.5 caras

E (X 2): 0*0*(1/8) + 1*1*(3/8) + 2*2*(3/8) + 3*3*(1/8) = 3 caras

Por lo tanto: V(X) = 3- (1.5 *1.5) = 0.75

Existen distribuciones discretas especiales las cuales se presentan usualmente en el quehacer diario y son los más estudiados, entre los cuales tenemos:

6.16 PREGUNTAS Y RESPUESTAS DE REPASO

1. Un jugador del fútbol peruano ha tenido una buena temporada y existen rumores sobre su traspaso a un equipo extranjero. Se definen los siguientes eventos:

   A: El jugador es pretendido por PSV Eindhoven (Holanda).

   B: El jugador es pretendido por Valencia (España).

   C: El jugador es pretendido por Boca Juniors (Argentina).

   D: El jugador es pretendido por Santos (Brasil)

Describa, en término de los eventos antes mencionados, lo siguiente:

a) X: El jugador es pretendido por alguno de los equipos.

b) Y: El jugador es pretendido Valencia y no por los equipos americanos.

c) El significado del evento A^C∩B^C∩C^C∩D^C

Solución

a) X={A∪B∪C∪D}

b) Y={B∩C^C∩D^C }

c) Dicho evento significa que el jugador no es pretendido por ninguno de los equipos.

2. La lista de los deportistas peruanos calificados para los Juegos Panamericanos Lima 2019, dividida por deporte y género, es la siguiente:

Deporte	Género	
	Femenino	**Masculino**
Atletismo	11	16
Judo	4	9
Karate	10	7
Tabla	5	9
Tae Kwon Do	5	8
Vóley	14	0
Otros	19	36

Una empresa privada busca patrocinar a uno de estos deportistas durante la competición, por lo que lo elegirá al azar.

a. Calcule la probabilidad de que el deportista seleccionado sea hombre o atleta.
b. Calcule la probabilidad de que el deportista seleccionado no sea mujer ni karateca.
c. Si se sabe que el deportista seleccionado es hombre, cuál es la probabilidad de que sea judoka.
d. Si se selecciona al azar a otro deportista para patrocinarlo, ¿cuál es la probabilidad de que ambos sean surfistas?

Solución

a)

Se define:

H: Probabilidad de que el deportista seleccionado sea hombre.

A: Probabilidad de que el deportista seleccionado sea atleta.

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

Donde:

$$P(H) = \frac{85}{153} = 0.5556$$

$$P(A) = \frac{27}{153} = 0.1765$$

$$P(H \cap A) = P(H|A) * P(A) = \frac{16}{27}(0.1765) = 0.1046$$

Reemplazando:

$$P(H \cup A) = 0.5556 + 0.1765 - 0.1046 = 0.6275$$

b)

Se define:

M: Probabilidad de que el deportista seleccionado sea mujer.

K: Probabilidad de que el deportista seleccionado sea karateca.

$$P(M^C \cap K^C) = \frac{16 + 9 + 9 + 8 + 0 + 36}{153} = 0.5098$$

c)

Se define:

H: Probabilidad de que el deportista seleccionado sea hombre.

J: Probabilidad de que el deportista seleccionado sea judoka.

$$P(J|H) = \frac{P(J \cap H)}{P(H)}$$

Donde:

$$P(H) = \frac{85}{153} = 0.5556$$

$$P(J \cap H) = \frac{9}{85}(0.5556) = 0.0588$$

Reemplazando:

$$P(J|H) = \frac{0.0588}{0.5556} = 0.1059$$

d)

Se define:

S: Probabilidad de que el deportista seleccionado sea surfista.

$$P(S_1) = \frac{14}{153} = 0.0915$$

Para seleccionar a otro surfista:

$$P(S_2) = \frac{14 - 1}{153 - 1} = 0.0855$$

Entonces:

$$P(S_1) * P(S_2) = 0.0915 * 0.0855 = 0.0078$$

3. Con la finalidad de evaluar la calidad de un lote de 75 válvulas que se requieren en una planta industrial, el almacenero selecciona una muestra aleatoria de 5 piezas del lote. Si en la muestra hay, como máximo, una pieza defectuosa, el almacenero recibe la mercadería. De lo contrario, realiza una inspección completa del lote.

   a. Suponga que ha llegado un lote con 11 válvulas defectuosas y 64 en buen estado. Calcule la probabilidad de que el almacenero decida recibirlo (a priori, el almacenero no sabe cuántos defectuosos hay en el lote).

   b. Suponga que el almacenero debe revisar un lote con 16 válvulas defectuosas y 59 en buen estado. Calcule la probabilidad de que solo la última pieza seleccionada de la muestra sea defectuosa.

Solución

a)

$$P(Recibir) = \frac{C_{11,0} * C_{64,5}}{C_{75,5}} + \frac{C_{11,1} * C_{64,4}}{C_{75,5}} = 0.4418 + 0.4049 = 0.8467$$

b)

Se define:

X: Probabilidad de que la última pieza seleccionada sea defectuosa.

$$P(X) = \frac{59}{75} * \frac{58}{74} * \frac{57}{73} * \frac{56}{72} * \frac{16}{71} = 0.0844$$

4. Se tiene 3 urnas: A con 7 bolas rojas y 4 verdes, B con 5 bolas rojas y 6 verdes, y C con 3 bolas rojas y 5 verdes. Se escoge una urna al azar y se extrae una bola.

   a. Si la bola ha sido verde, ¿cuál es la probabilidad de haber sido extraída de la urna A?

Solución

a)

Utilizando el Teorema de Bayes:

$$P(A|V) = \frac{P(A) * P(V|A)}{P(A) * P(V|A) + P(B) * P(V|B) + P(C) * P(V|C)}$$

$$P(A|V) = \frac{\frac{1}{3} * \frac{4}{11}}{\frac{1}{3} * \frac{4}{11} + \frac{1}{3} * \frac{6}{11} + \frac{1}{3} * \frac{5}{8}} = 0.2370$$

5. Tres máquinas de coser (A, B y C) realizan el 38%, 35% y 27%, respectivamente, del total de las piezas producidas en una planta. De igual manera, los porcentajes de producción defectuosa de estas máquinas son del 4%, 5% y 6%.
   a. Si selecciona una pieza al azar, calcule la probabilidad de que sea defectuosa.
   b. Si selecciona una pieza al azar y resulta ser defectuosa, ¿cuál es la probabilidad de que haya sido producida por la máquina B?
   c. ¿Qué máquina tiene la mayor probabilidad de haber producido dicha pieza defectuosa?

Solución:

a)

Se define:

D: Probabilidad de que la pieza seleccionada sea defectuosa.

Por la propiedad de probabilidad total, se sabe:

$$P(D) = P(A) * P(D|A) + P(B) * P(D|B) + P(C) * P(D|C)$$
$$P(D) = 0.38 * 0.04 + 0.35 * 0.05 + 0.27 * 0.06 = 0.0489$$

b)

Utilizando el Teorema de Bayes:

$$P(B|D) = \frac{P(B) * P(D|B)}{P(A) * P(D|A) + P(B) * P(D|B) + P(C) * P(D|C)}$$
$$P(B|D) = \frac{0.35(0.05)}{0.38(0.04) + 0.35(0.05) + 0.27(0.06)} = 0.3579$$

c)

Utilizando el Teorema de Bayes:

$$P(A|D) = \frac{0.38(0.04)}{0.38(0.04) + 0.35(0.05) + 0.27(0.06)} = 0.3108$$
$$P(C|D) = \frac{0.27(0.06)}{0.38(0.04) + 0.35(0.05) + 0.27(0.06)} = 0.3313$$

Por lo tanto, la máquina con mayor probabilidad de haber producido es B.

CAPÍTULO VII

DISTRIBUCIONES

7.1 INTRODUCCIÓN

El capítulo VII, denominada Distribuciones, presenta: la distribución binominal y la distribución Poisson. La distribución binomial es una distribución discreta más útiles. Su área de aplicación es muy amplia. Por su parte, la distribución de Poisson es una distribución de probabilidad discreta, esta se manifiesta a partir de una frecuencia de ocurrencia media.

7.2 DISTRIBUCIÓN BINOMIAL

Es una de las distribuciones discretas más útiles. Su área de aplicación incluye, las ventas, la inspección de calidad de productos, medicina, investigación de mercados, etc.

A continuación, se presentan algunos experimentos que aceptan un modelo Binomial:

Lanzar una moneda quince veces y observar el número de caras que aparecen.

El número de artículos defectuosos producido por una máquina.

Observar el número de partículas extrañas en diferentes muestras de aire.

Características:

El experimento consiste en n ensayos idénticos e independientes.

En cada ensayo, existen solo dos posibles resultados, a uno se le denomina éxito (E) y al otros Fracaso (F).

La probabilidad de éxito p es la misma en cada ensayo.

Por lo tanto:

Una variable aleatoria discreta X sigue una distribución Binomial con parámetros n y p y se representa por X ~ B (n, p) si su función de probabilidad es la siguiente:

$$P(X=x): C_x^n * p^x * (1-p)^{n-x}$$

Donde: x: 0,1,2, --- n

n: número de ensayos

p: probabilidad de éxito en cada ensayo

Medidas de resumen:

Esperanza matemática: E(X): n*p

Varianza: V(X): n*p*(1-p)

Ejemplo:

En una fábrica manufacturera, el jefe de control de calidad observó que el 12% de las partículas en cada muestra de aire son extrañas. Si usted selecciona una partícula de cada una de 20 muestras de aire.

Sea la variable:

X= Número de partículas extrañas de un total de 20

Parámetros:

n: 20 y p: 0.12

Se pide:

Calcule la probabilidad que se encuentren a lo más dos partículas extrañas.

P (X ≤ 2): P (X = 0) + P (X = 1) + P (X = 2)

Por lo tanto:

$$P(X=0): C(20,0) * 0.12^{0} * (1-0.12)^{20-0} = 0.0776$$

$$P(X=1): C(20,1) * 0.12^{1} * (1-0.12)^{20-1} = 0.2115$$

$$P(X=2): C(20,2) * 0.12^{2} * (1-0.12)^{20-2} = 0.2740$$

Entonces: P(X≤2) = 0.5631

La probabilidad que se encuentre por lo menos tres partículas extrañas es:

P (X≥3) = 1 – P (X≤2)

P (X≥3) = 1 – 0.5631 = 0.4369

Calcule el valor esperado y la varianza de las partículas extrañas en el aire es:

E(X): n*p = 20 *0.12 = 2.4

V(X): n*p*(1-p) = 20*0.12*(1-0.12) = 2.112

7.3 DISTRIBUCIÓN POISSON

Esta distribución se aplica cuando nos encontramos en situaciones en que interesa determinar que un evento aislado ocurra un número específico de veces en un intervalo de tiempo o espacio.

A continuación, se presentan algunos experimentos que aceptan un modelo Poisson:

Número de autos que llegan a un taller automotriz en un lapso específico.

Número de impulsos electrónicos errados transmitidos durante espacio de tiempo específico.

Número de llamadas telefónicas que ingresan a un conmutador por minuto.

Número de defectos de una tela por m2.

Número de bacterias por cm2 de un cultivo.

Cantidad de átomos que se desintegran en sustancia radioactiva.

Número de accidentes automovilísticos en un cruce específico durante una semana.

Por lo tanto:

Una variable aleatoria X sigue una distribución Poisson con parámetro si su función de densidad de probabilidad está dada por:

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

Dónde:

λ = Número promedio de ocurrencias del evento de interés.

Medidas de resumen:

Esperanza matemática: E(X): λ

Varianza: V(X): λ

Ejemplo:

En el departamento de mantenimiento informático de una empresa se reciben en promedio dos llamadas solicitando servicio por día.

Sea la variable:

X = Número de llamadas solicitando servicio de mantenimiento informático en un día.

Parámetro λ =2 → 1 dia

Se pide:

a) Calcule la probabilidad que se reciban exactamente tres llamadas en un día cualquiera.

$P(X = 3) = (e^{-2} * 2^{3}) / 3! = 0.1804$

b) Calcule la probabilidad que se reciban a lo más dos llamadas en un día cualquiera.

$P(X \leq 2)$: $P(X = 0) + P(X = 1) + P(X = 2)$

$P(X \leq 2)$: $(e^{-2} * 2^{0}) / 0! + (e^{-2} * 2^{1}) / 1! + (e^{-2} * 2^{2}) / 2! = 0.6767$

7.4 PREGUNTAS Y RESPUESTAS DE REPASO

1. Un doctor afirma que la aspirina causa efectos secundarios a 7 de cada 100 pacientes. Para contrastar ello, elige al azar a 20 pacientes, a quienes pide que se tomen una pastilla.
   a. ¿Cuál es la probabilidad de que ningún paciente tenga efectos secundarios?
   b. ¿Cuál es la probabilidad de que, al menos, dos tengan efectos secundarios?
   c. ¿Cuál es el número medio de pacientes que espera el doctor que sufran efectos secundarios, si elige a 300 pacientes al azar?

Solución

a)

$$p = \frac{7}{100} = 0.07 \Rightarrow q = 0.93$$

Se sabe:

$$P(X = k) = C_{n,k}(p^k)(q^{n-k})$$

Entonces:

$$P(X = 0) = C_{20,0}(0.07^0)(0.93^{20-0}) = 0.2342$$

b)

$$P(X \geq 2) = 1 - P(X < 2) = 1 - \big(P(X = 0) + P(X = 1)\big)$$

$$P(X \geq 2) = 1 - \Big(C_{20,0}(0.07^0)(0.93^{20-0}) + C_{20,1}(0.07^1)(0.93^{20-1})\Big)$$

$$P(X \geq 2) = 1 - (0.2342 + 0.3526) = 0.4131$$

c)

$$\mu = np = 300(0.07) = 21$$

2. En una planta de café instantáneo, el área de calidad realizó una inspección e identificó 0.07 productos defectuosos, en promedio, por minuto.
   a. Determine la probabilidad de identificar un defectuoso en 5 minutos.
   b. Determine la probabilidad de identificar, al menos, tres defectuosos en 10 minutos.
   c. Determine la probabilidad de identificar, como máximo, un defectuoso en 15 minutos.

Solución

a)
Para 5 minutos:

$$\lambda = 0.07 * 5 = 0.35$$

Se sabe:

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

Entonces:

$$P(X = 1) = \frac{(e^{-0.35})(0.35^1)}{1!} = 0.2466$$

b)
Para 10 minut

$$\lambda = 0.07 * 10 = 0.70$$

Entonces:

$$P(X \geq 3) = 1 - P(X < 3) = 1 - \big(P(X = 0) + P(X = 1) + P(X = 2)\big)$$

$$P(X \geq 3) = 1 - \left(\frac{(e^{-0.70})(0.70^0)}{0!} + \frac{(e^{-0.70})(0.70^1)}{1!} + \frac{(e^{-0.70})(0.70^2)}{2!}\right)$$

$$P(X \geq 3) = 1 - (0.4966 + 0.3476 + 0.1217) = 0.0341$$

c)
Para 15 minutos:

$$\lambda = 0.07 * 15 = 1.05$$

Entonces:

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

$$P(X \leq 1) = \frac{(e^{-1.05})(1.05^0)}{0!} + \frac{(e^{-1.05})(1.05^1)}{1!}$$

$$P(X \leq 1) = 0.3499 + 0.3674 = 0.7174$$

3. En una bodega, la cual atiende todos los días, los ingresos diarios se modelan con una variable aleatoria normal con una media de S/ 274 y desviación estándar de S/ 19.

   a) Calcule la probabilidad de que los ingresos diarios superen los S/ 300.

   b) La dueña del restaurante quiere comprar una nueva congeladora, la cual cuesta S/ 1,650 y ha decidido ahorrar el 5% de los ingresos diarios para poder comprarla. ¿Cuál es la probabilidad de que, en 120 días, la dueña logre ahorrar lo suficiente para adquirir la congeladora? Suponer que el costo no variará durante el tiempo.

Solución

a)
Sea X: Ingreso diario de la bodega (en soles), donde X ~ N(274,19).

$$P(X > 300) = 1 - P(X \leq 300)$$

Estandarizando:

$$P(X > 300) = 1 - P\left(Z \leq \frac{300 - 274}{19}\right) = 1 - \varphi(1.368) = 0.0856$$

b)
Sea Y: Ahorro diario (en soles)

Donde:

$$E(Y) = 0.05\big(E(X)\big) = 0.05(274) = 13.7$$

$$Var(Y) = Var(0.05X) = 0.05^2 Var(X)$$

$$Var(Y) = 0.05^2(19)^2 = 0.9025$$

Sea A: Ahorro en 120 días (en soles), donde A ~ N(1,644,108.3).

$$P(A > 1{,}650) = 1 - P(A \leq 1{,}650)$$

Estandarizando:

$$P(A > 1{,}650) = 1 - P\left(Z \leq \frac{1{,}650 - 1{,}644}{\sqrt{108.3}}\right) = 1 - \varphi(0.5766) = 0.2821$$

4. La probabilidad de que una dieta aumente el rendimiento físico es de 0.12. Si 15 personas han empezado con dicha dieta, calcular la probabilidad de que aumenten su rendimiento en los siguientes casos:

   a) En 3 personas.

   b) En ninguna persona.

   c) En menos de 5 personas.

   d) En más de 3 personas.

   e) Entre 2 y 4 personas.

Solución

a)

Se sabe: p = 0.12, q = 0.88

$$p = 0.12 \Rightarrow q = 0.88$$

$$P(X = k) = C_{n,k}\left(p^k\right)\left(q^{n-k}\right)$$

Entonces:

$$P(X = 3) = C_{15,3}(0.12^3)(0.88^{15-3}) = 0.1696$$

b)

$$P(X = 0) = C_{15,0}(0.12^0)(0.88^{15-0}) = 0.1470$$

c)

$$P(X < 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$P(X < 5) = C_{15,0}(0.12^0)(0.88^{15-0}) + C_{15,1}(0.12^1)(0.88^{15-1}) + C_{15,2}(0.12^2)(0.88^{15-2}) + C_{15,3}(0.12^3)(0.88^{15-3}) + C_{15,4}(0.12^4)(0.88^{15-4})$$

$$P(X < 5) = 0.1470 + 0.3006 + 0.2870 + 0.1696 + 0.0694 = 0.9735$$

d)

$$P(X > 3) = 1 - P(X \leq 3)$$

$$= 1 - \left(P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\right)$$

$$P(X > 3) = 1 - \left(C_{15,0}(0.12^0)(0.88^{15-0}) + C_{15,1}(0.12^1)(0.88^{15-1}) + C_{15,2}(0.12^2)(0.88^{15-2}) + C_{15,3}(0.12^3)(0.88^{15-3})\right)$$

$$P(X > 3) = 1 - (0.1470 + 0.3006 + 0.2870 + 0.1696) = 0.0959$$

e)

$$P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4)$$

$$P(2 \leq X \leq 4) = C_{15,2}(0.12^2)(0.88^{15-2}) + C_{15,3}(0.12^3)(0.88^{15-3}) + C_{15,4}(0.12^4)(0.88^{15-4})$$

$$P(2 \leq X \leq 4) = 0.2870 + 0.1696 + 0.0694 = 0.5259$$

5. Las familias ingresan a un zoológico de acuerdo a un proceso Poisson con una media de 5 personas cada 7 minutos, donde el precio de la entrada, para niños y adultos, es de S/ 5.

   a) Calcule la probabilidad de que en 3 minutos ingresen menos de 4 personas.

   b) Si los visitantes pueden ingresar al museo desde las 8 a.m. hasta las 3 p.m., ¿cuál es el valor esperado de los ingresos diarios por la venta de entradas?

Solución

a)

Para 3 minutos:

$$\lambda = \frac{5}{7} * 3 = 2.1429$$

Se sabe:

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

Entonces:

$$P(X < 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$P(X < 4) = \frac{(e^{-2.1429})(2.1429^0)}{0!} + \frac{(e^{-2.1429})(2.1429^1)}{1!} + \frac{(e^{-2.1429})(2.1429^2)}{2!} \frac{(e^{-2.1429})(2.1429^3)}{3!}$$

$$P(X < 4) = 0.1173 + 0.2514 + 0.2694 + 0.1924 = 0.8305$$

b)

Para 7 horas (420 minutos):

$$\lambda = \frac{5}{7} * 420 = 300$$

$$E(X) = 300(5) = 1{,}500\ soles$$

CAPÍTULO VIII

LA MUESTRA EN LA INVESTIGACIÓN JURÍDICA

8.1 INTRODUCCIÓN

El capítulo VIII denominado La muestra en la investigación jurídica, desarrolla los siguientes temas: población, muestra, individuo, tipos de muestras, determinar el tamaño de una muestra en una investigación jurídica, cálculo del tamaño de la muestra desconociendo el tamaño de la población, cálculo del tamaño de la muestra conociendo el tamaño de la población.

8.2 POBLACIÓN

Es el conjunto de todos los elementos que son objeto del estudio estadístico. También se le considera como un conjunto de medidas. Si la característica observada ha sido medida, recibe el nombre de variable continua, si solamente se hace el recuento se le denomina atributo o puede ser una variable discreta.

Para Martínez (2016, p.658):

> Considerar la población como un conjunto de unidades o elementos, debe entenderse como un grupo de personas, familias, establecimientos, manzanas barrios, objetos, etc. Pero en realidad es un conjunto de medidas obtenidas de las características estudiadas. Por ejemplo, al considerar un grupo de personas de acuerdo a la finalidad de la investigación, se podrá estudiar su nivel educativo, sexo, ocupación, edad, preferencias, hábitos o cualquier otra característica. Para un conjunto de personas que presentan las mismas características, se buscará información cuantificada, algunas veces mediante el recuento de unidades. En ese caso particular, número de personas con determinado nivel educativo; número de hombres y/o mujeres; cuantos de ellas se encuentran trabajando, etc. En otros casos la característica será medida (estatura, pesos, ingresos, etc).

8.3 MUESTRA

Es un subconjunto, extraído de la población (mediante técnicas de muestreo), cuyo estudio sirve para inferir características de toda la población.

8.4 INDIVIDUO

Es cada uno de los elementos que forman la población o la muestra.

8.5 TIPOS DE MUESTRAS

Los criterios de clasificación pueden dividirse en dos grandes grupos: métodos de muestreo probabilísticos y métodos de muestreo no probabilísticos.

8.5.1 El muestreo probabilístico

El método de muestreo probabilísticos es aquel que se basa en el principio de equiprobabilidad. Es decir, aquellos en los que todos los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra y, consiguientemente, todas las posibles muestras de tamaño n tienen la misma probabilidad de ser seleccionadas.

El método de muestreo probabilístico permite la representatividad de la muestra extraída y son, por tanto, los más recomendables.

Los métodos de muestreo probabilísticos son los siguientes tipos:

8.5.2 Muestreo aleatorio simple

Se asigna un número a cada individuo de la población y a través de algún medio mecánico (bolas dentro de una bolsa, tablas de números aleatorios, números aleatorios generados con una calculadora u ordenador, etc. se eligen tantos sujetos como sea necesario para completar el tamaño de muestra requerido. Este procedimiento, tiene poca utilidad cuando la población es muy grande.

8.5.3 Muestreo aleatorio sistemático

Es un tipo de muestreo provistico, motivo por el cual requiere tener un control preciso de la muestra seleccionada de individuos seleccionables conjuntamente con la probabilidad de que sean seleccionados, es decir, consiste en escoger un individuo inicial de forma aleatoria entre la población y luego seleccionar para la muestra a cada enésimo individuo disponible en el marco de la muestra.

El muestreo sistemático es un proceso simple y sólo requiere la elección de un individuo al azar. El resto del proceso es rápido. Los resultados que se obtiene son representativos de la población, de forma similar al muestreo aleatorio simple, siempre y cuando no haya algún factor intrínseco en la forma en que los individuos están listados que haga que se reproduzcan ciertas características poblacionales cada cierto número de individuos.

8.5.4 El muestreo aleatorio estratificado

Es un tipo de muestreo probabilístico que se utiliza cuando en la población se puede 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. Este tipo de muestreo presenta básicamente dos ventajas: 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.

La distribución de la muestra en función de los diferentes estratos se denomina afijación, y puede ser de tres tipos:

Afijación simple: a cada estrato le corresponde igual número de unidades de la muestra.

Afijación proporcional: la distribución se realiza de acuerdo con el tamaño de la población en cada estrato.

Afijación óptima: se tiene en cuenta la previsible dispersión de los resultados, de modo que se considera la proporción y la desviación típica.

8.5.5 El muestreo aleatorio por conglomerados

El conglomerado más utilizado en la investigación es un conglomerado geográfico. Por ejemplo, un investigador desea estudiar el rendimiento académico de los estudiantes de Derecho en Per. Puede dividir a toda la población (población de Perú) en diferentes conglomerados (ciudades del Perú).

8.5.6 El muestreo no probabilístico

En el muestreo probabilístico se seleccionan a los sujetos siguiendo determinados criterios procurando, en la medida de lo posible, que la muestra sea representativa. En algunas circunstancias el muestreo probabilístico permite resolver los problemas de representatividad, cuando el investigador jurídico conoce la población, por ejemplo los estudios de caso-control, por parte del Órgano de control de la magistratura, donde los casos no son seleccionados aleatoriamente de la población. Entre los métodos de muestreo no probabilísticos más utilizados en investigación encontramos

Las aproximaciones no probabilísticas son utilizadas cuando el investigador no dispone del marco muestral para la población en estudio o cuando simplemente, no se considera necesario o adecuado el uso de un procedimiento probabilístico.

Entre los métodos de muestreo no probabilísticos más utilizados en investigación encontramos

8.5.7 El muestreo por cuotas

Enseña Del Val Cid (2007, p. 39), que el muestreo por cuotas encuentra su fundamento en que, si se conocieran todas las características de la población y sus proporciones correspondientes, sería posible organizar una muestra que estuviera "cuotificada" en todas sus dimensiones, de tal manera que fuera representativa de la población, sin

necesidad de extracción aleatoria individual de sus unidades.

Por ejemplo, si se pretendiera abordar la influencia del nivel de estudios en los hábitos de lectura de una población universitaria, se deberían establecer cuotas a partir de los distintos niveles de estudio que se pueden alcanzar. Las cuotas más habituales en la investigación social vienen definidas por el sexo, la edad, la educación, la etnia, la religión y el nivel socioeconómico, pues la mayoría de los marcos o bases muestrales recogen estas variables.

En el muestreo por cuotas, se suelen dar instrucciones generales a los entrevistadores como: buscar dos hombres y tres mujeres en una manzana específica y asegurarse que cuatro tengan más de 25 años y uno menor de 25 años. Por otro lado, en el muestreo probabilístico a los entrevistadores se les proporcionan nombres o direcciones ya seleccionadas aleatoriamente sin subjetividad humana. (Pérez: 2018).

8.5.8 Muestreo intencional o de conveniencia

De acuerdo a Question Pro (2020):

> El muestreo intencional o de conveniencia es una técnica de muestreo no probabilístico y no aleatorio es utilizada para crear muestras de acuerdo a la facilidad de acceso, la disponibilidad de las personas de formar parte de la muestra, en un intervalo de tiempo dado o cualquier otra especificación práctica de un elemento particular.
>
> Los investigadores utilizan técnicas de muestreo en situaciones en las que hay grandes poblaciones para ser evaluadas, ya que, en la mayoría de los casos, es casi imposible realizar pruebas a toda una población.
>
> Este muestreo por conveniencia es la técnica de muestreo que se utiliza de manera más común, ya que es extremadamente rápida, sencilla, económica y, además, los miembros suelen estar accesibles para ser parte de la muestra.
>
> Esta técnica de muestreo es utilizada cuando no existen criterios que deban considerarse para que una persona pueda ser parte de la muestra. Cada individuo de la población puede ser un participante y es elegible para ser parte de la muestra. Estos participantes dependen de la cercanía al investigador. Es muy frecuente su utilización en sondeos preelectorales de zonas que en anteriores votaciones han marcado tendencias de voto. También puede ser que el investigador seleccione de manera intencional los individuos de la población. El caso más frecuente de este procedimiento es utilizar como muestra los individuos que se tiene fácil acceso (los profesores de la Facultad de Derecho utilizan con mucha frecuencia a sus propios estudiantes)

8.6 DETERMINACIÓN DEL TAMAÑO DE UNA MUESTRA EN UNA INVESTIGACIÓN JURÍDICA

Determinar el tamaño de la muestra que se va a seleccionar es un paso importante en el estudio de la investigación jurídica social, se debe justificar de acuerdo al planteamiento del problema, la población, los objetivos y la hipótesis de la investigación.

El tamaño de la muestra dependerá de las decisiones estadísticas y no estadísticas, que puedan incluir por ejemplo la disponibilidad de los recursos, el presupuesto o el equipo que estará en el trabajo de campo.

Para calcular el tamaño de la muestra es necesario determinar varios aspectos:

El tamaño de la población

Una población es una agrupación bien definida de objetos o individuos que tienen características similares.

Hablamos de dos tipos de población:

- Población objetivo, que suele tiene diversas características y también es conocida como la población teórica.
- La población accesible, es la población sobre la que los investigadores aplicaran sus conclusiones.
- El margen de error, es un estadígrafo que expresa la cantidad de error de muestreo aleatorio en los resultados de una encuesta, es decir, es la medida estadística del número de veces de cada 100 que se espera que los resultados se encuentren dentro de un rango específico.
- El nivel de confianza, son los intervalos aleatorios que se usan para acotar un valor con una determinada probabilidad alta. Por ejemplo, un intervalo de confianza de 95% significa que los resultados de una acción probablemente cubrirán las expectativas el 95% de las veces
- La desviación estándar, es un índice numérico de la dispersión de un conjunto de datos (o población). Mientras mayor es la desviación estándar, mayor es la dispersión de la población.

8.7 CÁLCULO DEL TAMAÑO DE LA MUESTRA DESCONOCIENDO EL TAMAÑO DE LA POBLACIÓN

La población infinita, está conformada por un indeterminado número de unidades, tal es el número de libros jurídicos producidos en el mundo. El comportamiento de una población demasiado grande, aun siendo finita, es considerado como una población infinita al calcular el tamaño de la muestra, por ejemplo, el número de artículos jurídicos producidos en América.

Al determinar el tamaño de la muestra en una población finita demasiada extensa el resultado no varía en lo más mínimo al establecido por la población finita.

La fórmula para calcular el tamaño de muestra cuando se desconoce el tamaño de la población es la siguiente:

$$n = \frac{Z_a^{\ 2} \times p \times q}{d^2}$$

donde:

- $Z_{\square}^{\ 2} = 1.96^2$ (ya que la seguridad es del 95%)
- p = proporción esperada (en este caso 5% = 0.05)
- q = 1 – p (en este caso 1 – 0.05 = 0.95)
- d = precisión (en este caso deseamos un 3%)

$$n = \frac{1.96^2 * 0.05 * 0.95}{0.03^2} = 203$$

8.8 CÁLCULO DEL TAMAÑO DE LA MUESTRA CONOCIENDO EL TAMAÑO DE LA POBLACIÓN

La población finita es aquella constituida por un determinado de elementos y unidades y en la mayoría de casos, considerada como relativa pequeña.

Si la población es finita, es decir conocemos el total de la población y deseásemos saber cuántos del total tendremos que estudiar la respuesta seria:

$$n = \frac{N * Z_{\alpha}^2 p * q}{d^2 * (N-1) + Z_{\alpha}^2 * p * q}$$

Donde:

- N = Total de la población
- $Z_{\square}^{\ 2} = 1.96^2$ (si la seguridad es del 95%) Nivel de confianza
- p = proporción esperada o probabilidad de éxito (en este caso 5% = 0.05)
- q = 1 – p probabilidad de fracaso (en este caso 1-0.05 = 0.95)
- d = precisión (en este caso deseamos un 3%). Error máximo admisible.

¿A cuántas personas tendríamos que estudiar de una población de 15.000 miembros del Colegio de Abogados del Callao para conocer su conocimiento del derecho arbitral?

Seguridad = 95%; Precisión = 3%; proporción esperada = asumamos que puede ser próxima al 5%; si no tuviese ninguna idea de dicha proporción utilizaríamos el valor p = 0.5 (50%) que maximiza el tamaño muestral.

$$n = \frac{15.000 * 1.96^2 * 0.05 * 0.95}{0.03^2 (15.000 - 1) + 1.96^2 * 0.05 * 0.95} = 200$$

Según diferentes seguridades el coeficiente de $Z_{\square}$ varía, así:

- Si la seguridad $Z_{\square}$ fuese del 90% el coeficiente sería 1.645
- Si la seguridad $Z_{\square}$ fuese del 95% el coeficiente sería 1.96
- Si la seguridad $Z_{\square}$ fuese del 97.5% el coeficiente sería 2.24
- Si la seguridad $Z_{\square}$ fuese del 99% el coeficiente sería 2.576

1. El tiempo en que los alumnos de un colegio terminan sus exámenes sigue una distribución normal con desviación estándar de 8.5 minutos. Para estimar la media del tiempo en que terminan, se quiere calcular un intervalo de confianza que tenga una amplitud menor o igual a 5 minutos, con un nivel de confianza del 99 %.

   a. Determine cuál es el tamaño mínimo de la muestra que es necesario observar.

   Solución:

   a)

   La longitud del intervalo de confianza es la siguiente:

$$Longitud = \left(\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

   Se pide:

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < 5$$

$$2(2.576)\frac{8.5}{\sqrt{n}} < 5$$

$$n > 76.6994 \approx 77\ alumnos$$

2. Para estimar la proporción de familias en Lima que cuentan con internet en casa, se quiere utilizar una muestra aleatoria de valor n.

   a. Calcular el valor mínimo de n para garantizar que, a un nivel de confianza del 95 %, el error en la estimación sea menor que 0.05.

Solución

a)

Dado que se desconoce la proporción, se ha de tomar el caso más desfavorable

$$\hat{p} = \hat{q} = 0.5.$$

El error en la estimación es el siguiente:

$$e = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Se pide:

$$e = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < 0.05$$

$$n > (1.96)^2\frac{(0.5)(0.5)}{(0.05)^2} = 384.1459 \approx 385\ familias$$

3. Una empresa que produce celulares sabe que la vida media de estos equipos sigue una distribución normal con media de 30 meses y desviación estándar de 5 meses.

   a. Determine el mínimo tamaño muestral que garantiza, con un nivel confianza del 98%, que la vida media de los celulares en dicha muestra se encuentre entre 28 y 32 meses.

Solución

a)

$$Longitud = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 4\ meses$$

Se pide:

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < 4$$

$$2(2.326)\frac{5}{\sqrt{n}} < 4$$

$$n > 33.8243 \approx 34\ celulares$$

4. Se desea obtener la media de una variable aleatoria que se distribuye normalmente con una desviación estándar de 3.85. Para ello, se toma una muestra de 75 elementos, donde se obtiene una media de 43.5.
   a. ¿Con qué nivel de confianza se puede afirmar que la media de la población está entre 42.3 y 44.2?
   b. Si la desviación estándar de la población fuera 3.5, ¿qué tamaño mínimo debería tener la muestra con la cual se estimó la media poblacional para un nivel de confianza del 99 %, con un error admisible que no supere el valor de 0.75?

Solución

a)

$$Longitud = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 44.2 - 42.3$$

$$2z_{\alpha/2}\frac{3.85}{\sqrt{75}} = 1.9$$

$$z_{\alpha/2} = 2.1369$$

Luego:

$$1 - \frac{\alpha}{2} = 0.9837 \Rightarrow \alpha = 0.0326$$

Por lo tanto, el nivel de confianza es de 96.74%.

b)

Para $\alpha = 0.01$ y $\sigma = 3.5$, se tiene:

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < 0.75$$

$$n > (2.576)^2\frac{(3.5)^2}{(0.75)^2} = 144.4933 \approx 145\ elementos$$

5. El área de calidad de una empresa que produce gaseosas está estudiando el peso de sus presentaciones familiares, las cuales tienen una media de 3.5 kg. y una desviación estándar de 0.2 kg. Para ello, se tomó una muestra de 15 botellas y se especifica la probabilidad de error tipo I como $\alpha = 0.05$.

   a. Si el peso real promedio de las botellas es de 3.35 kg., halle la probabilidad de que la prueba detecte que no se cumple el estándar especificado.

   b. Si es importante que la potencia de la prueba sea de, al menos, 0.9 en caso la media verdadera sea de 3.35 kg., halle el tamaño de muestra requerido. Use $\alpha = 0.05$.

Solución

Se sabe: $\mu=3.5$, $\sigma=0.2$

a)

El estadístico de la prueba es el siguiente:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 3.5}{0.2/\sqrt{15}}$$

Donde:

$$Z_{0.025} \leq \frac{\bar{x} - 3.5}{0.2/\sqrt{15}} \leq Z_{0.975}$$

$$3.399 \leq \bar{x} \leq 3.601$$

Para $\mu = 3.35$:

$$H_0: \mu = 3.35, H_1: \mu \neq 3.35$$

$$\alpha = P(Rechazar\ H_0 \mid H_0\ es\ verdadera) = 1 - P(Aceptar\ H_0 \mid H_0\ es\ verdadera)$$

Estandarizando:

$$\alpha = 1 - P\left(\frac{3.399 - 3.5}{0.2/\sqrt{15}} \leq Z \leq \frac{3.601 - 3.5}{0.2/\sqrt{15}}\right)$$

$$\alpha = 1 - P(0.945 \leq Z \leq 4.865)$$

$$\alpha = 1 - (\varphi(4.865) - \varphi(0.945))$$

$$\alpha = 1 - (1 - 0.8276) = 0.8276$$

b)

La potencia debe ser de, al menos, 0.9 ($\beta \leq 0.1$). Estandarizando:

$$P\left(\frac{\left(-1.96\left(\frac{0.2}{\sqrt{n}}\right)+3.5\right)-3.35}{0.2/\sqrt{n}} \leq Z \leq \frac{\left(1.96\left(\frac{0.2}{\sqrt{n}}\right)+3.5\right)-3.35}{0.2/\sqrt{n}}\right) \leq 0.1$$

$$P\left(-1.96+0.75\sqrt{n} \leq Z \leq 1.96+0.75\sqrt{n}\right) \leq 0.1$$

$$\varphi\left(1.96+0.75\sqrt{n}\right)-\varphi\left(-1.96+0.75\sqrt{n}\right) \leq 0.1$$

$$1-\varphi\left(-1.96+0.75\sqrt{n}\right) \leq 0.1$$

$$0.9 \leq \varphi\left(-1.96+0.75\sqrt{n}\right)$$

Entonces:

$$-1.96+0.75\sqrt{n} \geq 1.282$$

$$n \geq 18.6799 \approx 19\ botella$$

8.9 PREGUNTAS Y RESPUESTAS DE REPASO

1. El tiempo en que los alumnos de un colegio terminan sus exámenes sigue una distribución normal con desviación estándar de 8.5 minutos. Para estimar la media del tiempo en que terminan, se quiere calcular un intervalo de confianza que tenga una amplitud menor o igual a 5 minutos, con un nivel de confianza del 99 %.

   a. Determine cuál es el tamaño mínimo de la muestra que es necesario observar.

Solución

a)

La longitud del intervalo de confianza es la siguiente:

$$Longitud = \left(\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)-\left(\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Se pide:

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}<5$$

$$2(2.576)\frac{8.5}{\sqrt{n}}<5$$

$$n>76.6994 \approx 77\ alumnos$$

2. Para estimar la proporción de familias en Lima que cuentan con internet en casa, se quiere utilizar una muestra aleatoria de valor n.

   a. Calcular el valor mínimo de n para garantizar que, a un nivel de confianza del 95 %, el error en la estimación sea menor que 0.05.

Solución

a)

Dado que se desconoce la proporción, se ha de tomar el caso más desfavorable $\hat{p}=\hat{q}=0.5$.

El error en la estimación es el siguiente:

$$e = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$e = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < 0.05$$

$$n > (1.96)^2\frac{(0.5)(0.5)}{(0.05)^2} = 384.1459 \approx 385\ familias$$

3. Una empresa que produce celulares sabe que la vida media de estos equipos sigue una distribución normal con media de 30 meses y desviación estándar de 5 meses.

   a. Determine el mínimo tamaño muestral que garantiza, con un nivel confianza del 98%, que la vida media de los celulares en dicha muestra se encuentre entre 28 y 32 meses.

Solución

a)

$$Longitud = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 4\ meses$$

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < 4$$

$$2(2.326)\frac{5}{\sqrt{n}} < 4$$

$$n > 33.8243 \approx 34\ celulares$$

4. Se desea obtener la media de una variable aleatoria que se distribuye normalmente con una desviación estándar de 3.85. Para ello, se toma una muestra de 75 elementos, donde se obtiene una media de 43.5.
   a. ¿Con qué nivel de confianza se puede afirmar que la media de la población está entre 42.3 y 44.2?
   b. Si la desviación estándar de la población fuera 3.5, ¿qué tamaño mínimo debería tener la muestra con la cual se estimó la media poblacional para un nivel de confianza del 99 %, con un error admisible que no supere el valor de 0.75?

Solución

a)

$$Longitud = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 44.2 - 42.3$$

$$2z_{\alpha/2}\frac{3.85}{\sqrt{75}} = 1.9$$

$$z_{\alpha/2} = 2.1369$$

$$1 - \frac{\alpha}{2} = 0.9837 \Rightarrow \alpha = 0.0326$$

Por lo tanto, el nivel de confianza es de 96.74%.

b)

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < 0.75$$

$$n > (2.576)^2\frac{(3.5)^2}{(0.75)^2} = 144.4933 \approx 145\ elementos$$

5. El área de calidad de una empresa que produce gaseosas está estudiando el peso de sus presentaciones familiares, las cuales tienen una media de 3.5 kg. y una desviación estándar de 0.2 kg. Para ello, se tomó una muestra de 15 botellas y se especifica la probabilidad de error tipo I como $\alpha = 0.05$.
   a. Si el peso real promedio de las botellas es de 3.35 kg., halle la probabilidad de que la prueba detecte que no se cumple el estándar especificado.
   b. Si es importante que la potencia de la prueba sea de, al menos, 0.9 en caso la media verdadera sea de 3.35 kg., halle el tamaño de muestra requerido. Use $\alpha = 0.05$.

Solución

Se sabe: μ=3.5, σ=0.2

a)

El estadístico de la prueba es el siguiente:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 3.5}{0.2/\sqrt{15}}$$

Donde:

$$Z_{0.025} \leq \frac{\bar{x} - 3.5}{0.2/\sqrt{15}} \leq Z_{0.975}$$

$$3.399 \leq \bar{x} \leq 3.601$$

Para μ = 3.35:

$$H_0: \mu = 3.35, H_1: \mu \neq 3.35$$

$$\alpha = P(Rechazar\ H_0 \mid H_0\ es\ verdadera)$$

$$= 1 - P(Aceptar\ H_0 \mid H_0\ es\ verdadera)$$

Estandarizando:

$$\alpha = 1 - P\left(\frac{3.399 - 3.5}{0.2/\sqrt{15}} \leq Z \leq \frac{3.601 - 3.5}{0.2/\sqrt{15}}\right)$$

$$\alpha = 1 - P(0.945 \leq Z \leq 4.865)$$

$$\alpha = 1 - (\varphi(4.865) - \varphi(0.945))$$

$$\alpha = 1 - (1 - 0.8276) = 0.8276$$

b)

La potencia debe ser de, al menos, 0.9 (β ≤ 0.1). Estandarizando:

$$P\left(\frac{\left(-1.96\left(\frac{0.2}{\sqrt{n}}\right) + 3.5\right) - 3.35}{0.2/\sqrt{n}} \leq Z \leq \frac{\left(1.96\left(\frac{0.2}{\sqrt{n}}\right) + 3.5\right) - 3.35}{0.2/\sqrt{n}}\right) \leq 0.1$$

$$P\left(-1.96 + 0.75\sqrt{n} \leq Z \leq 1.96 + 0.75\sqrt{n}\right) \leq 0.1$$

$$\varphi\left(1.96 + 0.75\sqrt{n}\right) - \varphi\left(-1.96 + 0.75\sqrt{n}\right) \leq 0.1$$

$$1 - \varphi\left(-1.96 + 0.75\sqrt{n}\right) \leq 0.1$$

$$0.9 \leq \varphi\left(-1.96 + 0.75\sqrt{n}\right)$$

Entonces:

$$-1.96 + 0.75\sqrt{n} \geq 1.282$$

$$n \geq 18.6799 \approx 19\ botellas$$

CAPÍTULO IX

PRUEBA DE HIPÓTESIS

9.1. INTRODUCCIÓN

La prueba de hipótesis es utilizada para corroborar posibles teorías que se generan a partir de información obtenida de datos por medio de la observación, experiencias personales, experiencias de terceros, etc. Estas teorías son un conjunto de ideas organizadas e interrelacionadas de manera lógica que explican un evento o fenómeno de interés y permite probar la solidez de estas ideas contra hechos observados. (Ritchey 2008, p. 267)

El modo de evaluar si estos hechos son válidos es a través de la prueba de hipótesis.

El capítulo IX denominado Prueba de hipótesis, estudia los siguientes temas: hipótesis, inferencia estadística, pasos para la inferencia estadística, prueba de hipótesis curva normal, prueba de hipótesis T-student, prueba de hipótesis Chi cuadrado.

9.2. HIPÓTESIS

La hipótesis es una predicción sobre la relación entre dos variables, una variable dependiente cuya diferencia entre sus medidas estará relacionada a diferencia de las medidas de la variable independiente.

La hipótesis simple es cualquier hipótesis estadística para la cual el valor del parámetro se especifique. Por ejemplo, el valor de la media de edades es de 30 años. Por otro lado, la hipótesis compuesta es cuando el valor del parámetro no se especifica. Por ejemplo, el valor de la media de edades es mayor a 30 años.

Durante el desarrollo de nuestra vida realizamos diferentes predicciones, algunas relacionadas a los resultados de un partido de fútbol, otras relacionadas a los resultados de un juicio con relación a los alegatos de las partes involucradas y las pruebas presentadas, el tiempo estimado de vida de la población, entre otras.

Ejemplo: ¿Qué ruta debo tomar hoy?

Hipótesis: La ruta (variable independiente) que tome definirá el tiempo (variable dependiente) que demoraré en llegar a mí destino.

Observación: Tomar alguna ruta y ver cronometrar el tiempo que demorará.

9.3. INFERENCIA ESTADÍSTICA

La inferencia estadística es generar conclusiones de una población a partir de datos y estadísticos obtenidos de una muestra. La inferencia contiene un grado de error medible en términos estadísticos.

9.4. PASOS PARA LA INFERENCIA ESTADÍSTICA

9.4.1. Formular H0 y HA

La Hipótesis Nula o H0 es la hipótesis que inicialmente será tomada como verdadera, esta hipótesis será comprobada experimentalmente. En tal sentido, la hipótesis nula será la principal hipótesis por probar.

El resultado del experimento determinará si la hipótesis es tomada como verdadera o será rechazada.

La Hipótesis Nula establece que no existe asociación o diferencia significativa entre la exposición de interés y el resultado.

La Hipótesis Alternativa o HA es la hipótesis contraria a la hipótesis nula H0.

La hipótesis alternativa se acepta en caso de que la hipótesis nula sea rechazada.

9.4.2. Distribución muestral

La distribución muestral es una proyección de resultados muestrales que es probable que ocurran en el muestreo repetido cuando H0 es cierta. Una distribución muestral consiste en un listado de resultados muestrales posibles y una definición de la probabilidad de cada uno.

Se empleará tendrá que evaluar en función del tipo de población, tipo de muestra y variable qué distribución a emplear. En el presente material se explicarán las tres pruebas de hipótesis más empleadas, las cuales son

Z – Normal	Tamaño de muestra $n \geq 30$. Variables de intervalo o razón
T – Student	Tamaño de muestra $n < 30$ Variables de intervalo o razón
X2 – Chi Cuadrado	Variables nominales y ordinales

9.4.3 Nivel de significancia (α), la dirección de la prueba y valores críticos de la prueba.

- **Nivel de significancia (α)**

Mediante el nivel de significancia (α) o error tipo I se establece la confiabilidad del estudio y determinará el punto de rechazo de la prueba de hipótesis. Por convención, se recomienda trabajar con el nivel de significancia de 5%. Así mismo, es aceptable trabajar con niveles entre 1% y 10%. Los niveles mayores restarán confiabilidad al estudio que se desea realizar.

- **Error del tipo I (α) o error del tipo II (β):**

Al realizar un estudio existen cuatro posibles acciones que determinarán si la decisión tomada es correcta o no.

- **La primera es rechazar la H0 cuando la H0 es verdadera.**
  Decisión incorrecta conocida como Error Tipo I.
- **La segunda es rechazar la H0 cuando la H0 es falsa.**
  Decisión correcta.
- **La tercera es aceptar la H0 cuando la H0 es verdadera.**
  Decisión correcta.
- **La cuarta es aceptar la H0 cuando la H0 es falsa.**
  Conocida como Error Tipo II.

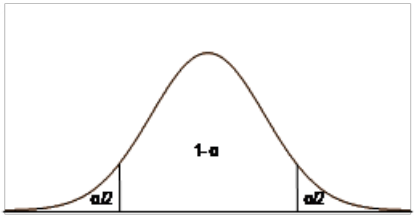
	Estado Real	
Decisión Real	**H_0 es verdadera**	**H_0 es falsa**
Rechazar H_0	Error Tipo I	
Aceptar H_0		Error Tipo II

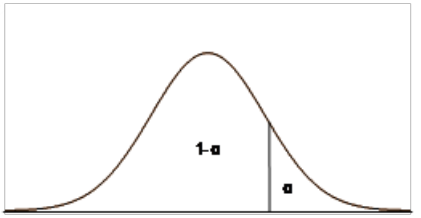
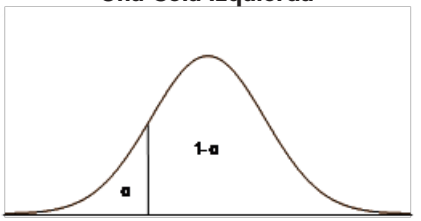
En otras palabras, el Error Tipo I es la probabilidad de rechazar la H0 cuando esta es verdadera, conocida como nivel de significancia. El Error Tipo II es la probabilidad de aceptar la H0 cuando H0 es falsa, conocida como potencia de la prueba.

El Error Tipo II se puede disminuir incrementando el tamaño de la muestra.

- **Dirección de la prueba**

La dirección de la prueba se establece en función del planteamiento la hipótesis alternativa, existiendo tres tipos.

H0: ux = u H1: ux ≠ u	**Dos Colas**  Para una prueba de dos colas con un nivel de significancia α = 0.05, la cola de la izquierda contendrá la primera región de rechazo correspondiente a $\alpha/2$ y la cola de la derecha, la segunda región de rechazo igual a $\alpha/2$.

H0: u = ux H1: u > ux	**Una Cola Derecha**  Para una prueba de una cola con un nivel de significancia α = 0.05, la cola de la derecha contendrá la región de rechazo igual a α
H0: u = ux H1: u < ux	**Una Cola Izquierda**  Para una prueba de una cola con un nivel de significancia α = 0.05, la cola de la izquierda contendrá la región de rechazo igual a α.

9.4.4 Valores críticos de la prueba e interpretación de resultados

Los valores críticos de la prueba se establecen según el nivel de significancia (α). Los valores críticos marcan la región de rechazo o no rechazo de la prueba.

4. Calcular los estadísticos de prueba de la muestra y el valor p.

Los estadísticos de prueba de la muestra son aquellos valores hallados mediante una fórmula estadística que mide la posibilidad de ocurrencia del efecto a medir, los estadísticos a emplear en el presente capítulo son tres, estadístico Z (Normal), t (T - Student) y X^2 (Chi Cuadrado). El valor p es la probabilidad que la diferencia que existe entre los valores a estudiar sea producto del azar.

El valor de rechazo o no rechazo por medio del valor p se da en función al nivel de significancia (α):

El valor p representa una región crítica que se interpreta como la probabilidad de cometer el error tipo I.

Si p < α; la probabilidad que las diferencias en las medidas del estudio se deban al azar es baja, por lo tanto, la probabilidad que las diferencias sean reales es alta. Es decir, la prueba es estadísticamente significativa.

Decisión: Rechazar la H_0 y no rechazar la hipótesis H_A

Si p > α; la probabilidad que las diferencias en las medidas del estudio se deban al azar es alta, por lo tanto, la probabilidad que las diferencias sean reales es baja. Es decir, la prueba no es estadísticamente significativa.

Decisión: No rechazar la H0 y rechazar la hipótesis HA.

Después de tomar la decisión de rechazo o no rechazo se debe interpretar qué significa el resultado, para comunicar la interpretación se debe considerar quién o quienes serán el público objetivo.

9.5. PRUEBA DE HIPÓTESIS CURVA NORMAL

Es un procedimiento estadístico que permite rechazar o no rechazar una afirmación sobre un suceso.

La curva normal se aplica en variables de intervalo o razón, cuando se cuenta con un tamaño de muestra mayor o igual a 30 datos donde se cumple que:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N\,(0,1)$$

9.5.1 Formular H0 y HA

Se desea estudiar el tiempo de viaje de los estudiantes de ciencia jurídica desde la Pontificia Universidad Católica del Perú a sus centros de labores. Para ello se ha realizado una encuesta a 35 estudiantes obteniendo la siguiente información:

Tabla 1: Tiempo de viaje

Tiempo de viaje (minutos)						
51	40	48	72	58	73	83
52	61	49	61	42	47	53
64	56	46	70	41	71	52
40	37	48	66	58	39	44
51	63	44	59	65	53	53

El investigador cree que el tiempo de viaje es superior a 60 minutos. Por ende, se procede a formular la hipótesis nula y la hipótesis alternativa.

H_0: u = El tiempo de viaje es igual a 60 minutos

H_A: u ≥ El tiempo de viaje es mayor o igual a 60 minutos

9.5.2 Distribución muestral

Para el presente caso, al ser una variable de razón y al contar con 35 datos se puede asumir que sigue una distribución normal considerando que:

μ_0 = 54,57 minutos

$$Z = \frac{\bar{X} - 54{,}57}{\sigma / \sqrt{n}} \sim N(0,1)$$

9.5.3. Nivel de significancia (α), la dirección de la prueba y valores críticos de la prueba.

- **Nivel de significancia (α)**

Para el presente caso se define un nivel de significancia α= 0.05.

- **Dirección de la prueba**

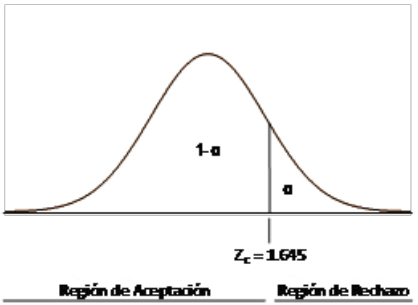
Al establecer la hipótesis alternativa de la siguiente forma:

HA: u ≥ El tiempo de viaje es mayor o igual a 60 minutos

Se define una prueba de una cola con una cola derecha.

9.5.4 Valores críticos de la prueba e interpretación de resultados

Estableceremos los valores críticos o el valor crítico de la prueba. Para ello debemos recordar que definimos un nivel de significancia α = 0.05.



Para ello buscamos el valor Z_c dentro de la Tabla de la Distribución Normal Estándar en el Anexo C.

Donde identificamos que para el área (1 – 0.05 = 0.95) se tiene un valor igual a 1.645.

Partiendo de la H0 definimos que:

$$\alpha > \mathrm{P}\left[\text{ Rechazar } H_0 \middle| H_0 \textit{ es correcta}\right]$$

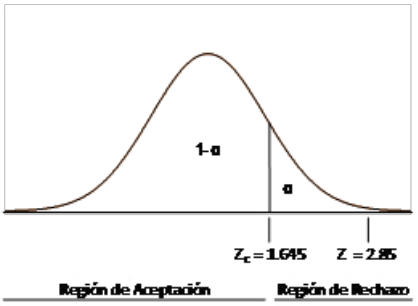
$$\alpha > \mathrm{P}\left[\bar{X} < 60\right]$$

$$0.05 > P\left[\frac{\bar{X} - \mathrm{u}}{\sigma\sqrt{n}} < \frac{60 - 54.57}{11.27\sqrt{35}}\right]$$

$$0.05 > P\left[Z < 2.85\right]$$

$$0.05 > 1 - P\left[Z > 2.85\right]$$

Se procede a buscar dentro de la Tabla de la Distribución Normal Estándar a qué área corresponde para el valor 2.85 identificando el valor de 0.9978. Donde el valor de 1-0.9978 se conoce como el valor p.



$$.05 > 1 - 0.9978$$

$$0.05 > 0.0022$$

Por ende, podemos concluir que:

$$\alpha > valor\ p$$

y se procede a rechazar la H0 y aceptar la HA. Es decir, con un nivel de significancia de 0.05 el tiempo de viaje de los estudiantes de las ciencias jurídicas desde la Pontificia Universidad Católica del Perú a sus centros de labores es mayor a 60 minutos.

9.6. PRUEBA DE HIPÓTESIS T - STUDENT

Es un procedimiento estadístico que permite rechazar o no rechazar una afirmación sobre un suceso, es decir, determinar si un parámetro de una población es igual al un valor referente especificado.

La curva T de Student se aplica en variables intervalo o razón, cuando se cuenta con un tamaño de muestra menor a 30 datos donde se cumple que:

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t\,(1-\alpha, g)$$

Como conocemos, cuando se cuenta con un mayor número de datos el error en la prueba estadística se reduce, en la distribución t, al trabajar con una cantidad menor de 30 datos se presenta un error mayor que se traduce en un aplanamiento de la curva, el aplanamiento depende de la cantidad de datos que pueden ser parametrados con los grados de libertad.

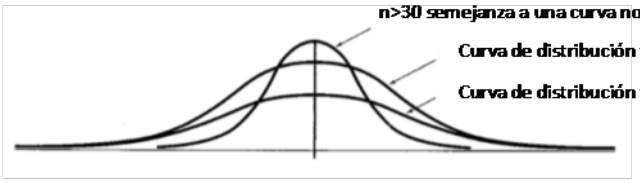
$$gl = n - 1$$

Donde:

gl: grados de libertad

n: tamaño de la muestra

Cuando la cantidad de datos se va incrementando la distribución t empieza a tomar la forma de la distribución normal.



9.6.1 Formular H_0 y H_A

Se realizó un estudio sobre la utilización del agua en un distrito de Lima, para ello se consideró una muestra de 28 casas. El número de litros de agua que utilizaron durante el muestreo fue el siguiente y sigue una distribución aproximadamente normal.

Consumo de Agua (litros/día)						
180	169	200	177	210	203	171
182	172	169	199	182	188	173
184	175	217	199	205	173	158
186	180	220	214	187	171	176

El investigador cree que el consumo diario de agua es menor a 160 litros por día. Por ende, se procede a formular la hipótesis nula y la hipótesis alternativa.

H_0: u = El consumo diario de agua es igual a 180 litros.

H_A: u < El consumo diario de agua es menor a 180 litros.

9.6.2 Distribución muestral

Para el presente caso, al ser una variable de razón y al contar con 28 datos se puede asumir que sigue una distribución normal considerando que:

μ_0 = 186,43 litros

$$t = \frac{\bar{X} - 186{,}43}{\frac{s}{\sqrt{n}}} \sim t\,(0{,}95\,;27)$$

9.6.3. Nivel de significancia (α), la dirección de la prueba y valores críticos de la prueba.

- **Nivel de significancia (α)**

Para el presente caso se define un nivel de significancia α= 0.05.

- **Dirección de la prueba**

Al establecer la hipótesis alternativa de la siguiente forma:

H_A: u < El consumo diario de agua es menor a 180 litros.

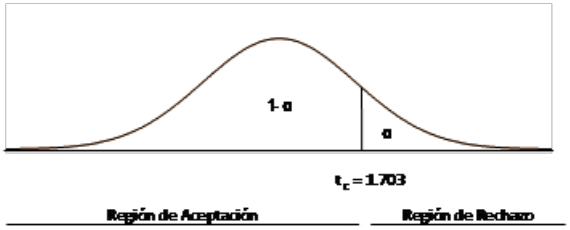
Se define una prueba de una cola con una cola izquierda.

9.6.4 Valores críticos de la prueba e interpretación de resultados

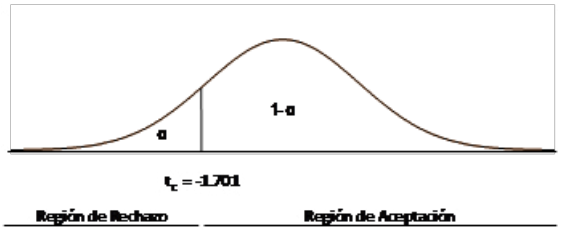
Estableceremos los valores críticos o el valor crítico de la prueba.

Al buscar el valor t_c dentro de la Tabla de la Distribución t - Student en el Anexo D para 27 grados de libertad y nivel de significancia 0.05 identificamos que solo existe

la gráfica para pruebas con cola a la derecha como se muestra en la siguiente imagen.



Es por ello, que debemos inferir que si para el t_c = 1.703 se ha acumulado el 95% de probabilidades, entonces para el t_c = -1,703 se ha acumulado el 5% de probabilidades.



Partiendo de la H_0 definimos que:

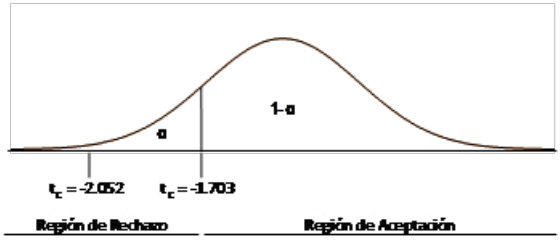
$$\alpha > \mathrm{P}\,[\,\text{Rechazar } H_0 \mid H_0 \textit{ es correcta}\,]$$

$$\alpha > \mathrm{P}\,[\,\bar{X} > 160\,]$$

$$0.05 > P\left[\frac{\bar{X} - \mathrm{u}}{\frac{s}{\sqrt{n}}} > \frac{180 - 186.43}{\frac{16.40}{\sqrt{28}}}\right]$$

$$0.05 > P\,[\,t > -2.05]$$

Se procede a buscar dentro de la Tabla de la Distribución t - Stundet a qué área corresponde el estadístico t = -2,05. Recordemos que la tabla solo contiene valor positivo, es decir, si para t = 2,05 se tiene un área bajo de la curva de 97,5%, para t = -2,05 se tendrá un área de 2,5%. Donde el valor de 2,5% o 0,025 se conoce como el valor p.



0.05>0,025

Por ende, podemos concluir que:

α>valor p

y se procede a rechazar la H0 y aceptar la HA. Es decir, con un nivel de significancia de 0.05 el consumo diario de agua en los hogares es menor a 180 litros.

9.7. PRUEBA DE HIPÓTESIS X^2 – CHI CUADRADO

Es un procedimiento estadístico que permite rechazar o no rechazar la relación entre dos variables que pueden contar con dos o más categorías.

La curva X^2 – Chi Cuadrado se aplica a variables nominales u ordinales, donde se cumple que:

$$X^2 = \sum \frac{(fo - fe)^2}{fe}$$

Donde:

fo: Frecuencia observada

fe: Frecuencia esperada

Como conocemos, cuando se cuenta con un mayor número de datos el error en la prueba estadística se reduce, en la distribución X^2, al trabajar con una cantidad menor de 30 datos se presenta un error mayor que se traduce en una concentración de datos con un sesgo hacia la izquierda, esto depende de la cantidad de datos que pueden ser parametrados con los grados de libertad.

$$gl = (c - 1)(f - 1)$$

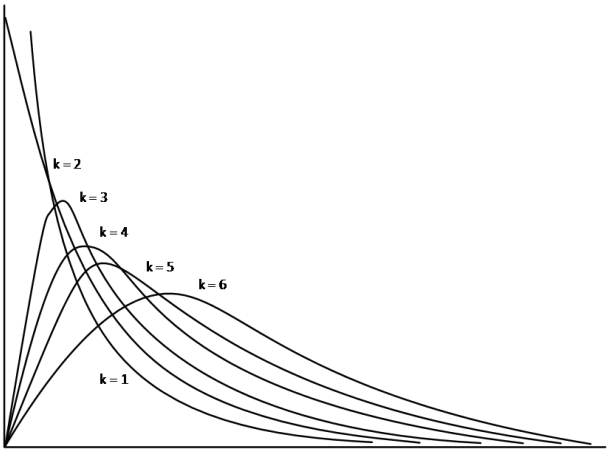
Donde:

gl: grados de libertad

c: cantidad de columnas

f: cantidad de filas

Cuando la cantidad de datos se va incrementando la distribución $\mathbf{X^2}$ empieza a tomar la forma de la distribución normal debido al teorema del límite central.



Tablas cruzadas de frecuencias

Para estudiar la relación que existe entre las variables de estudio se emplea las tablas cruzadas de frecuencias, donde se comparan dos variables nominales u ordinales al mismo tiempo. Estas tablas permiten probar la hipótesis de relación entre las variables.

Ejemplo:

En un estudio descriptivo requiere identificar si existe relación entre la asistencia a clases y la relación de género. Para ello, se recopiló información sobre la asistencia de alumnos a las clases de Derecho Constitucional donde se identificó lo siguiente:

Tabla 1: Asistencia a Clases de Derecho Constitucional

Asistencia a Clases	Varones	Mujeres	Total margina
Asistieron	132	121	253
No Asistieron	214	162	376
Total marginal	346	283	Gran total (629)

Los totales que agrupan la cantidad de varones o la cantidad de mujeres se conoce como totales marginales (346 varones o 319 mujeres), de igual manera, los totales que agrupan la cantidad de personas que asistieron o no asistieron a clases también se conoce como totales marginales (253 personas asistieron o 412 personas no asistieron). Por otro lado, la suma de la cantidad de varones y mujeres o la suma de la cantidad de personas que asistieron o no asistieron se conoce como gran total (665 personas).

Porcentaje de participación: Filas y columnas.

Después de recopilar la información, haciendo uso de la tabla de cruzada frecuencias se procede a realizar el cálculo de porcentajes.

Porcentaje de fila (% de fila):

Varones del total de asistentes: 132 / 253 x 100% = 52%

Mujeres del total de asistentes: 121 / 253 x 100% = 48%

Varones del total de no asistentes: 214 / 376 x 100% = 57%

Mujeres del total de no asistentes: 162 / 376 x 100% = 43%

Porcentaje de columna (% de columna):

Asistentes del total de varones: 132 / 346 x 100% = 38%

No Asistentes del total de varones: 214 / 346 x 100% = 62%

Asistentes del total de varones: 121 / 283 x 100% = 43%

No Asistentes del total de mujeres: 162 / 283 x 100% = 57%

Asistencia a Clases	Varones	Mujeres	Total
Asistieron	132	121	**253**
No Asistieron	214	162	**376**
Total	**346**	**283**	**629**

% de fila	
52%	48%
57%	43%

38%	43%
62%	57%
% de columnas	

De estos datos podemos obtener valiosa información descriptiva. Como por ejemplo que tanto en los varones como en las mujeres más del 57% no asistieron a clases y entre ambos, los varones son los que tienen mayores inasistencias con un 62%.

9.7.1 Formular H_0 y H_A

La prueba de hipótesis en el caso de la X^2 permite enunciar la relación o no relación entre una variable u otra.

Ejemplo:

Se desea comprobar si existe relación entre la procedencia del colegio público o privada y las notas finales obtenidas durante el primer semestre en la universidad. Por tal motivo, se recopiló información de la cantidad de estudiantes que obtuvieron

notas insuficientes, suficiente, notable y sobresaliente.

Tabla 1 – Notas obtenidas por colegio de procedencia

Colegio de Procedencia	Insuficiente	Bien	Notable	Sobresaliente
Privado	12	28	34	18
Público	60	64	34	6

El investigador cree que el rendimiento de los alumnos tiene relación con los centros educativos de procedencia.

Por ende, se plantean las siguientes hipótesis. Considerando que en la prueba X^2 solo se puede establecer la relación o no relación. Y el sentido solo puede ser direccional ya que se toman valores al cuadrado que finalmente siempre son positivos.

H_0: $X^2 = 0$; Entendiendo que no existe relación entre el colegio de procedencia y la nota obtenida.

H_A: $X^2 > 0$; Entendiendo, que existe relación entre el colegio de procedencia y la nota obtenida.

9.7.2 Distribución muestral

Para el presente caso, al ser una variable ordinal (puntaje obtenido) y nominal (colegio de procedencia) se puede asumir que sigue una distribución X^2.

9.7.3 Frecuencia Observada y Cálculo de frecuencias esperadas

En primer lugar, agrupamos los datos de la tabla inicial en los totales marginales y el gran total.

La frecuencia observada es la frecuencia o la cantidad de veces que se ha obtenido el dato durante el muestreo. Como ejemplo podemos decir, que la frecuencia observada de los estudiantes de colegios públicos con calificación insuficiente es de 12, y la frecuencia observada de los estudiantes de colegios privados con calificación sobresaliente es de 6.

Colegio de Precedencia	Insuficiente	Bien	Notable	Sobresaliente	Total Marginal
Público	12	28	34	18	92
Privado	60	64	34	6	164
Total Marginal	72	92	68	24	256

Seguidamente obtenemos la frecuencia esperada con la siguiente fórmula, la cual permitirá comprender la cantidad de veces que ocurra un evento.

$$Frecuencia\ Esperada = \frac{Total\ Marginal\ Columna\ x\ Total\ Marginal\ Fila}{Gran\ Total}$$

Frecuencia Esperada= (Total Marginal Columna x Total Marginal Fila)/(Gran Total)

Combinación	Total Marginal Columna	Total Marginal Final	Gran Total	Frecuencia Esperada
Público - Insuficiente	72	92	256	25,88
Privado - Insuficiente	72	164	256	46,13
Público - Bien	92	92	256	33,06
Privado - Bien	92	164	256	58,94
Público - Notable	68	92	256	24,44
Privado - Notable	68	164	256	43,56
Público - Sobresaliente	24	92	256	8,63
Privado - Sobresaliente	24	164	256	15,38
Total				256,00

De la tabla XX podemos decir, que de un total de 256 casos se puede esperar que existan 25,88 casos donde el puntaje obtenido por un alumno que proviene de un colegio pública sea insuficiente.

A modo de resumen decimos que para la combinación colegio público con calificación insuficiente la frecuencia observada es de 12 y la frecuencia esperada es de 25. Cuando se presentan grandes diferencias entre la frecuencia esperada y la frecuencia observada es más probable que se acepte la hipótesis alternativa. Recordemos que, al ser una muestra, está sujeta al error de muestreo normal esperado, ya que en un primer muestreo obtuvimos la cantidad observada de 12, habiendo la posibilidad que en el siguiente muestreo este pueda fluctuar. Es por ello que, los resultados de la prueba de hipótesis dependerán de la calidad de la muestra única que obtengamos.

Por otro lado, debemos emplear la prueba chi cuadrada X2 cuando la frecuencia esperada de cada casilla cruzada en la tabla cruzada es de al menos 5.

9.7.4 Grados de Libertad

Los grados de libertad se definen como:

$$gl = (c - 1)(f - 1)$$

Donde:

gl: grados de libertad

c: cantidad de columnas

f: cantidad de filas

Por consiguiente, para el ejemplo definimos 3 grados de libertad.

gl = (4 - 1) x (2 - 1) = 3 x 1 = 3

9.7.5 Nivel de significancia (α), la dirección de la prueba y valores críticos de la prueba.

- **Nivel de significancia (α)**

  Para el presente caso se define un nivel de significancia α= 0.05.

- **Dirección de la prueba**

  Al establecer la hipótesis alternativa de la siguiente forma:

  HA: X2 > 0; Entendiendo, que existe relación entre el colegio de procedencia y la nota obtenida.

  Se define una prueba de una cola con una cola a la derecha.

9.7.6 Valores críticos de la prueba e interpretación de resultados

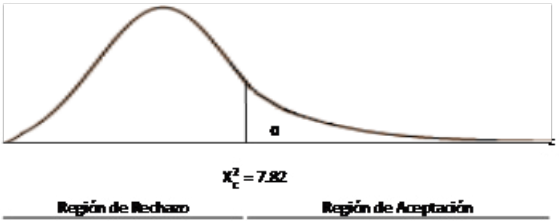
Estableceremos los valores críticos o el valor crítico de la prueba.

$$X^2 = \sum \frac{(fo - fe)^2}{fe}$$

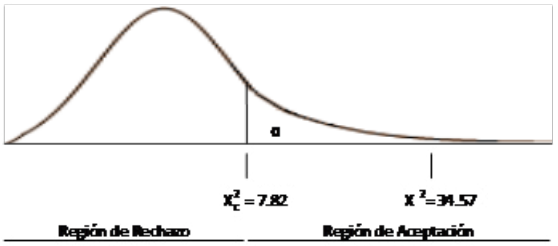
Realizando los cálculos correspondientes tenemos que:

Casilla o combinación	F. Observada (*fo*)	F. Esperada (*fe*)	$(fo-fe)^2$	$\frac{(fo-fe)^2}{fe}$
Publico - Insuficiente	12	25,88	192,52	7,44
Privado - Insuficiente	60	46,13	192,52	4,17
Público - Bien	28	33,06	25,63	0,78
Privado - Bien	64	58,94	25,63	0,43
Público - Notable	34	24,44	91,44	3,74
Privado - Notable	34	43,56	91,44	2,10
Público - Sobresaliente	18	8,63	87,89	10,19
Privado - Sobresaliente	6	15,38	87,89	5,72
Estadístico chi cuadrada				34,57

Al buscar el valor X^2c dentro de la Tabla de la Distribución X^2 Chi cuadrada Anexo D para 3 grados de libertad y nivel de significancia 0.05 obtenemos que el valor crítico es de 7.82.



De la tabla xxx anterior podemos identificar que el estadístico chi cuadrada tenía un valor de 34.57 que pertenece a la región de rechazo.



Partiendo de la H0 definimos que:

$$\alpha > \mathrm{P}\,[\,\text{Rechazar } H_0 |\, H_0 \textit{ es correcta}]$$

$$\alpha > \mathrm{P}\,[\,X^2 < 0\,]$$

$$0.05 > 1 - P\,[\,X^2 > 34.57]$$

$$0.05 > 1 - 0.999999$$

$$0.05 > 0.000001$$

Se procede a buscar dentro de la Tabla de la Distribución X2 a qué área corresponde el estadístico X2 = 34,57, y por medio de extrapolación se obtiene que el área bajo la curva es de 99,9999%. Donde el valor de 1-0.999999=0.000001 se conoce como el valor p.

Por ende, podemos concluir que:

$$\alpha > valor\ p$$

y se procede a rechazar la H0 y aceptar la HA. Es decir, con un nivel de significancia de 0.05 el colegio de precedencia si influye en la calificación y que la diferencia observada no se debe al azar.

9.8 PREGUNTAS Y RESPUESTA DE REPASO

1. El gerente de una cadena de joyerías cree que las balanzas utilizadas en sus tiendas están dando valores mayores al peso real, lo cual afecta a su negocio. Por ello, solicitó que se pese la misma piedra de cuarzo (de 3 kg.) en cada una de las balanzas de sus 13 tiendas.

   Asumiendo normalidad, se pide lo siguiente:

   a. Plantee la hipótesis del caso.

   b. El gerente concluye que las balanzas utilizadas en sus tiendas están mal calibradas, tal que el promedio de los pesos supera un cierto valor C. Halle dicho valor, de manera que el nivel de significación de la prueba sea de $\alpha = 0.05$.

   c. ¿Qué podría concluir el gerente si el promedio de los pesos es de 3,08 kg., con una desviación estándar de 0.15 kg., usando el mismo nivel de significación de la pregunta anterior?

Solución

a)

Sea X: Peso medido por la balanza de cada local de la joyería, donde X ~ N(μ,σ).

$$H_0: \mu = 3, H_1: \mu > 3$$

b)

El gerente plantea la siguiente región crítica:

$$R.C: \bar{X} > C$$

Para el nivel de significación planteado:

$$\alpha = P(Rechazar\ H_0 \mid H_0\ es\ verdadera)$$

$$0.05 = P(\bar{X} > C \mid \mu = 3)$$

$$0.05 = P\left(T_0 > \frac{C-3}{S/\sqrt{n}}\right) = 1 - P\left(T_0 \leq \frac{C-3}{S/\sqrt{n}}\right)$$

Donde:

$$T_0 = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{1-\alpha}(n-1)$$

Entonces:

$$\frac{C-3}{S/\sqrt{13}} = t_{0,95}(12) = 1.7823$$

$$C = 3 + 1.7823(S)$$

c)

Para S = 0.15:

$$C = 3 + 1.7823(0.15) = 3.0741$$

Debido a que $\overline{X}$=3.08>3.0741, la región crítica se satisface, por lo que se rechazará H0 y el gerente podrá asegurar, con una probabilidad de equivocarse del 5%, que las balanzas que se utilizan en sus tiendas están dando valores mayores al peso real.

4. Una empresa que produce focos ahorradores afirma que éstos, en promedio, duran al menos 8,000 horas. Las pruebas con 65 focos dan como resultado una duración media de 7,950 horas, con una desviación estándar de 250 horas.

   Asumiendo normalidad, se pide lo siguiente:

   a. Comprobar si hay evidencia suficiente para rechazar la afirmación de la empresa, a un nivel de significación del 5%.

   b. ¿Cuál es el p-valor?

Solución

a)

Sea X: Duración de los focos en horas, donde X ~ N(μ,σ).

$$H_0: \mu \geq 8{,}000, H_1: \mu < 8{,}000$$

En el muestreo de una población normal con varianza desconocida, con muestras grandes (n > 30), se tiene:

$$\bar{x} \sim N\left(\mu, \frac{s_x}{\sqrt{n}}\right) = N\left(8{,}000, \frac{250}{\sqrt{65}}\right) = N(8{,}000, 31.0087)$$

Para el nivel de significación planteado:

$$\alpha = P(Rechazar\ H_0 \mid H_0\ es\ verdadera)$$

$$0.05 = P(\bar{x} < k \mid \mu \geq 8{,}000)$$

$$0.05 = P\left(\frac{\bar{x} - 8{,}000}{31.0087} < \frac{k - 8{,}000}{31.0087}\right) = P\left(z < \frac{k - 8{,}000}{31.0087}\right)$$

Entonces:

$$\frac{k - 8{,}000}{31.0087} = -1.645$$

$$k = 7{,}948.9953$$

Debido a que $\bar{x}$=7,950>7,948.9953, se acepta H_0, por lo que se acepta la afirmación de la empresa, con un nivel de confianza del 95%.

b)

El p-valor (αp) es el menor nivel de significación para el que se rechaza H_0.

$$\alpha_p = P(Rechazar\ el\ estadístico\ muestral \mid H_0\ es\ verdadera)$$

$$\alpha_p = P(\bar{x} < 7{,}950 \mid N(8{,}000, 31.0087))$$

$$\alpha_p = P\left(\frac{\bar{x} - 8{,}000}{31.0087} < \frac{7{,}950 - 8{,}000}{31.0087}\right) = P\left(z < \frac{7{,}950 - 8{,}000}{31.0087}\right)$$

$$\alpha_p = (z < -1.612) = 0.0534$$

Debido a que αp > α = 0.05, se acepta H0, por lo que se acepta que los focos ahorradores tienen una duración de al menos 8,000 horas, con un nivel de confianza del 95%.

3. Para determinar la tasa de empleabilidad en las principales ciudades del país, una empresa encuestadora le realizó la pregunta a una muestra aleatoria de 150 personas en las ciudades de Lima, Arequipa y Trujillo, obteniendo los siguientes datos:

Empleo	Ciudad		
	Lima	Arequipa	Trujillo
Tiene	135	133	111
No tiene	15	17	39

   a. ¿Es la proporción real de empleabilidad la misma para las 3 ciudades? Use α = 0.05. Sugerencia: Efectúe pruebas de hipótesis para diferencia de proporciones de dos en dos ciudades.

   b. Si las proporciones reales no son iguales, establezca la jerarquía de los parámetros (de ser posible) con un nivel de confianza del 95%.

Solución

a)

Se sabe:

$$\bar{p}_1 = \frac{135}{150} = 0.9$$

$$\bar{p}_2 = \frac{133}{150} = 0.8867$$

$$\bar{p}_3 = \frac{111}{150} = 0.74$$

Para α = 0.05:

$$z_{1-\alpha/2} = z_{0.975} = 1.96$$

Entonces:

$$R.C:]-\infty, -1.96[\cup]1.96, \infty[$$

Lima vs. Arequipa:

$$H_0: p_1 = p_2, H_1: p_1 \neq p_2$$

$$\hat{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} = \frac{150(0.9) + 150(0.8867)}{150 + 150} = 0.8933$$

$$z_c = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.9 - 0.8867}{\sqrt{(0.8933)(0.1067)\left(\frac{1}{150} + \frac{1}{150}\right)}} = 0.3741$$

Como z_c está fuera del R.C, se acepta H_0.

Lima vs. Trujillo:

$$H_0: p_1 = p_3, H_1: p_1 \neq p_3$$

$$\hat{p} = \frac{n_1\bar{p}_1 + n_3\bar{p}_3}{n_1 + n_3} = \frac{150(0.9) + 150(0.74)}{150 + 150} = 0.82$$

$$z_c = \frac{\bar{p}_1 - \bar{p}_3}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_3}\right)}} = \frac{0.9 - 0.74}{\sqrt{(0.82)(0.18)\left(\frac{1}{150} + \frac{1}{150}\right)}} = 3.6067$$

Como z_c está dentro del R.C, se rechaza H_0.

Arequipa vs. Trujillo:

$$H_0: p_2 = p_3, H_1: p_2 \neq p_3$$

$$\hat{p} = \frac{n_2\bar{p}_2 + n_3\bar{p}_3}{n_2 + n_3} = \frac{150(0.8867) + 150(0.74)}{150 + 150} = 0.8133$$

$$z_c = \frac{\bar{p}_2 - \bar{p}_3}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_2} + \frac{1}{n_3}\right)}} = \frac{0.8867 - 0.74}{\sqrt{(0.8133)(0.1867)\left(\frac{1}{150} + \frac{1}{150}\right)}} = 3.2598$$

Como z_c está dentro del R.C, se rechaza H_0.

En conclusión: $p_1 = p_2 \neq p_3$.

b)

$$m = C_2^3 = 3$$

$$\alpha_0 = \frac{\alpha}{m} = \frac{0.05}{3} = 0.0167$$

$$z_{1-\alpha_0/2} = z_{0.9917} = 2.394$$

Los intervalos de confianza se definen de la siguiente manera:

Lima vs. Arequipa:

$$\left[(\bar{p}_1 - \bar{p}_2) - z_{1-\frac{\alpha_0}{2}}\sqrt{\frac{\bar{p}_1\bar{q}_1 + \bar{p}_2\bar{q}_2 + 2\bar{p}_1\bar{p}_2}{n_1 + n_2 + n_3}}, (\bar{p}_1 - \bar{p}_2) + z_{1-\alpha_0/2}\sqrt{\frac{\bar{p}_1\bar{q}_1 + \bar{p}_2\bar{q}_2 + 2\bar{p}_1\bar{p}_2}{n_1 + n_2 + n_3}}\right]$$

$$\left[(0.9 - 0.8867) - 2.394\sqrt{\frac{(0.9)(0.1) + (0.8867)(0.1133) + 2(0.9)(0.8867)}{150 + 150 + 150}}, (0.9 - 0.8867) + 2.394\sqrt{\frac{(0.9)(0.1) + (0.8867)(0.1133) + 2(0.9)(0.8867)}{150 + 150 + 150}}\right]$$

$$[-0.1375, 0.1642]$$

Como 0 está dentro del intervalo, se puede afirmar, con un nivel de confianza del 95%, que $p_1 = p_2$.

Lima vs. Trujillo:

$$\left[(\bar{p}_1 - \bar{p}_3) - z_{1-\frac{\alpha_0}{2}}\sqrt{\frac{\bar{p}_1\bar{q}_1 + \bar{p}_3\bar{q}_3 + 2\bar{p}_1\bar{p}_3}{n_1 + n_2 + n_3}}, (\bar{p}_1 - \bar{p}_3) + z_{1-\alpha_0/2}\sqrt{\frac{\bar{p}_1\bar{q}_1 + \bar{p}_3\bar{q}_3 + 2\bar{p}_1\bar{p}_3}{n_1 + n_2 + n_3}}\right]$$

$$\left[(0.9 - 0.74) - 2.394\sqrt{\frac{(0.9)(0.1) + (0.74)(0.26) + 2(0.9)(0.74)}{150 + 150 + 150}}, (0.9 - 0.74) + 2.394\sqrt{\frac{(0.9)(0.1) + (0.74)(0.26) + 2(0.9)(0.74)}{150 + 150 + 150}}\right]$$

$$[0.0166, 0.3034]$$

Como 0 está fuera del intervalo, se puede afirmar, con un nivel de confianza del 95%, que $p_1 > p_3$.

Arequipa vs. Trujillo:

$$\left[(\bar{p}_2 - \bar{p}_3) - z_{1-\frac{\alpha_0}{2}}\sqrt{\frac{\bar{p}_2\bar{q}_2 + \bar{p}_3\bar{q}_3 + 2\bar{p}_2\bar{p}_3}{n_1 + n_2 + n_3}}, (\bar{p}_2 - \bar{p}_3) + z_{1-\alpha_0/2}\sqrt{\frac{\bar{p}_2\bar{q}_2 + \bar{p}_3\bar{q}_3 + 2\bar{p}_2\bar{p}_3}{n_1 + n_2 + n_3}}\right]$$

$$\left[(0.8867 - 0.74) - 2.394\sqrt{\frac{(0.8867)(0.1133) + (0.74)(0.26) + 2(0.8867)(0.74)}{150 + 150 + 150}}, (0.8867 - 0.74) - 2.394\sqrt{\frac{(0.8867)(0.1133) + (0.74)(0.26) + 2(0.8867)(0.74)}{150 + 150 + 150}}\right]$$

$$[0.0037, 0.2896]$$

Como 0 está fuera del intervalo, se puede afirmar, con un nivel de confianza del 95%, que $p_2 > p_3$.

En conclusión: $p_1 = p_2 > p_3$.

4. En una empresa metalmecánica se compraron tuercas de diámetro aleatorio, con una distribución normal de μ = 32 mm. y σ = 1.1 mm., para que sean utilizadas con los tornillos, cuyos diámetros tienen una distribución normal con media μ y σ = 0.8 mm.

   a. ¿Cuál debe ser el valor de μ para que la probabilidad de que una tuerca no entre en un tornillo sea de 0,05?

   b. Suponga que el personal que utiliza ambos elementos sospecha que los diámetros de éstos no están cumpliendo la especificación dada. Si se desea detectar una desviación de la especificación de 1.3 mm., con una probabilidad de, al menos, 0.99 y un nivel de significación α = 0.05, ¿cuál sería el tamaño de la muestra que se debería tomar, a fin de aclarar sus sospechas?

   Solución

   a)

   Sean:

   X: Diámetro de una tuerca, donde X ~ N(32,1.1).

   Y: Diámetro de un tornillo, donde X ~ N(μ,0.8).

   Donde:

$$0.05 = P(X > Y) = P(X - Y > 0) = 1 - P(X - Y \leq 0)$$

   Se sabe:

$$X - Y \sim N\left(\mu_x - \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2}\right) = N(32 - \mu, \sqrt{1.1^2 + 0.8^2})$$

$$X - Y \sim N(32 - \mu, 1.3601)$$

   Entonces:

$$0.95 = P\left(Z \leq \frac{\mu - 32}{1.3601}\right)$$

$$\frac{\mu - 32}{1.3601} = 1.645 \Rightarrow \mu = 34.2372$$

   b)

$$H_0: \mu_x = 32, H_1: \mu_x \neq 32$$

La potencia debe ser de, al menos, 0.99 ($\beta \leq 0.01$), a fin de detectar una desviación δ = 1.3 mm. A un nivel de significación α = 0.05, se tiene:

$$F_z\left(1.96 - \frac{\sqrt{n}}{1.1}\right) - F_z\left(-1.96 - \frac{\sqrt{n}}{1.1}\right) \leq 0.01$$

Dado que el segundo término de la fórmula anterior es pequeño (< 0.025), se considerará solo el primero, de tal manera que:

$$1.96 - \frac{\sqrt{n}}{1.1} \leq -2.326$$

$$n \geq 22.2307 \approx 23$$

Reemplazando:

$$F_z\left(1.96 - \frac{\sqrt{23}}{1.1}\right) - F_z\left(-1.96 - \frac{\sqrt{23}}{1.1}\right) = 0.0001 \leq 0.01$$

Con ello, el personal podría utilizar una muestra de, al menos, 23 tuercas.

5. Una empresa quiere saber si es que un curso de comunicación efectiva aumentaría sus niveles de ventas. Por tal motivo, seleccionó a sus 12 asesores para que tomen el curso. Al mes siguiente de que éste haya terminado, la cantidad de ventas realizadas por cada asesor fueron las siguientes:

Antes del curso	Después del curso
6	12
13	12
10	9
12	8
13	11
7	12
9	10
7	8
11	9
6	13
11	12
14	7

a. ¿Se podría decir que este curso origina diferente variabilidad en la cantidad de ventas? Use $\alpha = 0.05$.

b. ¿Se podría afirmar que el curso afecta positivamente en el nivel de ventas? Use $\alpha = 0.05$.

Solución

a)

Sean:

X: Cantidad de ventas antes del curso, donde X ~ N(μ1, σ1).

Y: Cantidad de ventas después del curso, donde X ~ N(μ2, σ2).

Al nivel de significación planteado:

$$H_0: \sigma_1^2 = \sigma_2^2, H_1: \sigma_1^2 \neq \sigma_2^2$$

$$F_0 = \frac{S_1^2}{S_2^2} = \frac{(2.8749)^2}{(2.0057)^2} = 2.0546$$

Se rechaza H_0 si se satisface la región crítica.

$$R.C: F_0 < F_{0.025}(11,11) = 0.2879, F_0 > F_{0.975}(11,11) = 3.4737$$

Dado que F_0 está fuera de la región crítica, se acepta H_0. Es decir, ambos escenarios originan variabilidades similares.

b)

Al nivel de significación planteado:

$$H_0: \mu_1 = \mu_2, H_1: \mu_1 < \mu_2$$

Donde σ_1^2 y σ_2^2 son desconocidos y, de la pregunta anterior, se puede asumir que $\sigma_1^2 = \sigma_2^2$.

Se rechaza H_0 si se satisface la región crítica.

$$R.C: T_0 < -t_{1-\alpha}(n_1 + n_2 - 2) = -t_{0.95}(22) = -1.7171$$

Donde:

$$T_0 = \frac{X - Y}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(12 - 1)(2.8749)^2 + (12 - 1)(2.0057)^2}{12 + 12 - 2}} = 2.4787$$

Reemplazando:

$$T_0 = \frac{9.9167 - 10.25}{2.4787\sqrt{\frac{1}{12} + \frac{1}{12}}} = -0.3294$$

Dado que T_0 está fuera de la región crítica, se acepta H0. Es decir, no se puede garantizar que el curso afecte positivamente al nivel de ventas.

CAPÍTULO X

INTERVALOS DE CONFIANZA

10.1. INTRODUCCIÓN

La estimación de parámetros mediante los intervalos de confianza es parte de la estadística inferencial, donde los resultados derivados de las muestras nos dan conclusiones acerca de las características de la población.

Tenemos dos tipos de estimaciones para estimar parámetros de la población: Puntuales y las de intervalo. Una estimación puntual es un valor de un solo estadístico de la muestra, mientras la estimación de intervalo de confianza es un rango de números, llamado intervalo, construido alrededor de la estimación puntual.

En este capítulo X denominado Intervalos de confianza, se abordaran los siguientes temas: intervalos de confianza de una media poblacional, interpretación apropiada de los intervalos de confianza, intervalo de confianza de una proporción poblacional calculado a partir de una muestra grande y la selección de un tamaño de la muestra para elecciones, encuestas, y estudios de investigación

10.2. INTERVALOS DE CONFIANZA DE UNA MEDIA POBLACIONAL (IC)

Es un rango de valores posibles de un parámetro, el cual es expresado con un grado de confianza específico. Para ello tenemos la siguiente formula:

$$IC\ de\ 95\%\ de\ \mu_x = \bar{X} \pm (Z_\propto)(S_{\bar{x}})$$

Donde:

$IC\ de\ \mu_x : Intervalo\ de\ confianza\ de\ una\ media\ poblacional.$

$\bar{X} : Media\ Muestral.$

$Z_\propto :\ Puntuacion\ Z\ critica\ que\ corresponde\ al\ nivel\ estipulado\ de\ significancia\ y\ confianza.$

$S_{\bar{x}}$: Error Estandar Estimado de la Media.

Nivel de Confianza

Es el grado de confianza calculado, que un procedimiento estadístico realizado con datos muestrales, producirá un resultado correcto para la población muestreada. Se muestra en la siguiente formula (resaltado en negro):

$$IC\ de\ \mathbf{95\%}\ de\ \mu_x = \bar{X}\ \pm (Z_\propto)(S_{\bar{x}})$$
$$IC\ de\ \mathbf{99\%}\ de\ \mu_x = \bar{X}\ \pm (Z_\propto)(S_{\bar{x}})$$

Donde:

$95\% : Nivel\ de\ confianza$

$IC\ de\ \mu_x : Intervalo\ de\ confianza\ de\ una\ media\ poblacional.$

$\bar{X} : Media\ Muestral.$

$Z_\propto :\ Puntuacion\ Z\ critica\ que\ corresponde\ al\ nivel\ estipulado\ de\ significancia\ y\ confianza$

$S_{\bar{x}}$: Error Estandar Estimado de la Media.

Nivel de significancia α

Llamado también error esperado, se encuentra relacionado con el nivel de confianza, donde el máximo nivel de confianza seria 100%. Entonces el Nivel de significancia será:

$$NS(Error\ esperado) = \alpha = 100\% - NC$$

Donde:

$\alpha : Nivel\ de\ significancia\ o\ Error\ esperado$

$NC : Nivel\ de\ confianza, valores\ mas\ usados\ son\ 95\%\ o\ 99\%.$

Error Estándar (estimado) de un Intervalo de Confianza de una Media Poblacional

Como la media y la desviación estándar son desconocidas, entonces utilizamos la desviación estándar de la media para estimar el error estándar de la media.

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$

Donde:

$S_{\bar{x}} : Error\ estandar\ estimado\ de\ medias\ para\ una\ variable\ de\ intervalo$
$/razon\ \ X.$

$S_x : Desviacion\ estandar\ de\ una\ muestra.$

$n : Tamaño\ de\ la\ muestra.$

Selección de la Puntuación Z crítica (Zα)

Es una puntuación Z, que corresponde a los niveles de confianza y significancia elegidos.

Al tener el nivel de confianza podemos utilizar la tabla de distribución normal estándar (anexo 1), entonces podemos representar a la puntuación Z crítica en el siguiente gráfico.

Estos valores de la puntuación Z crítica corresponden a una distribución normal y a una cantidad de muestras n > 30.

Calculo del término de error de un intervalo de confianza de una media poblacional (cuando n >30)

El término de error es la multiplicación del error estándar y la puntuación Z , tal como lo vemos en la siguiente formula.

$$Termino\ del\ error = (Z_{\propto})(S_{\bar{x}})$$

Donde:

α : *Nivel de significancia o Error esperado*

$Z_{\propto}$: *Puntuacion Z critica que corresponde a los niveles estipulados de significancia y confianza.*

$S_{\bar{x}}$: *Error estandar estimado de un intervalo de confianza de la media*

Calculo del Intervalo de Confianza de una Media Poblacional (cuando n>30)

Este intervalo de confianza es una media muestral más y menos un término de error, otra manera de expresarlo sería mediante intervalos que consiste en un límite inferior y un límite superior.

$$(100\%-\propto) IC\ de\ 95\%\ de\ \mu_x = \bar{X} \pm (Z_{\propto})(S_{\bar{x}})$$

Donde:

$\propto : Nivel\ de\ significancia\ (o\ error\ esperado, expresado\ en\ \%)$

$100-\propto : Nivel\ de\ confianza$

$IC\ de\ \mu_x : Intervalo\ de\ confianza\ de\ una\ media\ poblacional.$

$\bar{X} : Media\ Muestral.$

$Z_{\propto} :\ Puntuacion\ Z\ critica\ que\ corresponde\ al\ nivel\ estipulado\ de\ sign\ y\ conf.$

$S_{\bar{x}} :$ Error Estandar Estimado de la Media.

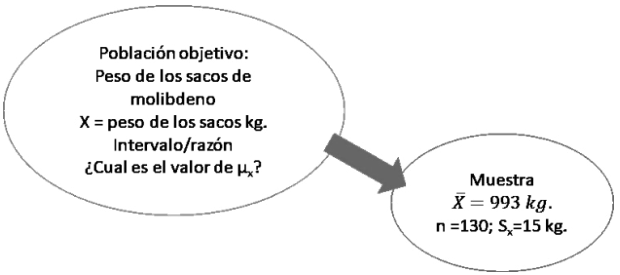
Cuando Calcular un Intervalo de Confianza de una Media Poblacional (cuando n>30)

- Cuando se requiera estimar un parámetro poblacional.
- Cuando la variable de interés (X) es de nivel de medición de intervalo/razón. Por lo tanto, debemos de proporcionar una estimación del intervalo del valor de un parámetro de la población μx
- Cuando se trabaja con una muestra única y representativa.
- Cuando el tamaño de la muestra es mayor que 30.

Ejemplo de Aplicación para la Estimación de Parámetros Mediante Intervalos de Confianza

En el área de llenado (Packing) la empresa Minera MinaCobre, se necesita obtener el peso promedio de la producción embolsada de Molibdeno. Se toman una muestra de 130 bolsas empacadas. De estas muestras se obtuvieron una media de 993 kg. Y una desviación estándar de 15 kg. Se pide calcular el intervalo de confianza de 95% para el peso promedio de las bolsas empacadas.

Paso 1. Realizar la pregunta de investigación: Con un rango especifico de Kg. ¿Cuál es el parámetro μx peso del saco de Molibdeno medio de la producción?



Paso 2. Calculo del error estándar, puntuación Z crítica y termino de error.

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} = \frac{15}{\sqrt{130}} = 1.32\ kg.$$

Para un nivel de confianza de 95%, Zα = 1.96

$$Termino\ de\ error = Z_{\alpha})(S_{\bar{x}}) = (1.96)(1.32) = 2.58\ kg.$$

Paso 3. Calculo de los Límites de confianza inferior (LCI) y Límites de confianza superior (LCS).

$$IC\ de\ 95\%\ de\ \mu_x = \bar{X} \pm (1.96)(S_{\bar{x}})$$

$$media\ muestral \pm termino\ de\ error$$

$$993\ kg. \pm (1.96)(1.32) = 993\ kg. \pm 2.58$$

$$LCI = 993kg. - 2.58kg. = 990.42kg.$$

$$LCS = 993\ kg. + 2.58kg. = 995.58kg.$$

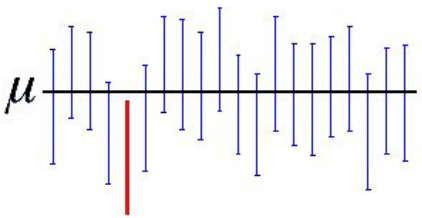
Paso 4. Interpretación en lenguaje común: Estoy 95% seguro de que el peso medio de los sacos llenados de molibdeno están entre 990.42 kg. Y 995.58 kg.

Paso 5. Interpretación estadística ilustrando la noción de confianza en el procedimiento: si se realizaran los mismos procedimientos muestrales y estadísticos 100 veces, 95 veces el parámetro poblacional μx estará comprendido en los intervalos calculados y 5 veces no. Por lo tanto, tengo una confianza de 95% que este intervalo de confianza individual que calcule incluye al parámetro real.

10.3 INTERPRETACIÓN APROPIADA DE LOS INTERVALOS DE CONFIANZA

Al decir que "estoy 95% seguro" estamos expresando confianza en nuestro método empleado para calcular nuestro parámetro. Para la interpretación estadística seria: con los mismos procedimientos muestrales y estadísticos se realizan 100 veces, 95 veces la media poblacional real μx estará comprendida en los intervalos calculados.

Grafico tasa de éxito de un intervalo de confianza de 95% al proporcionar una estimación del intervalo que comprende el valor real del parámetro poblacional.



Una mal interpretación común que se maneja en la interpretación de estos intervalos es: Según el ejemplo desarrollado, declaramos "estoy 95% seguro de que el peso medio de los sacos de molibdeno esta entre 990.42 kg y 995.58 kg., No estamos diciendo que 95% de los sacos de molibdeno pesan entre dichas cifras.

10.4. INTERVALO DE CONFIANZA DE UNA PROPORCIÓN POBLACIONAL CALCULADO A PARTIR DE UNA MUESTRA GRANDE

Utilizado para las variables nominal/ordinal, estos intervalos de confianza proporcionan una estimación de la proporción de una población que cae en la categoría de éxito de la variable. Para ello tenemos la siguiente ecuación.

$$(100\%-\propto) IC\ de\ P_u = P_s \pm (Z_\propto)(Sp_s)$$

$$= proporcion\ muestral \pm termino\ del\ error$$

Donde:

$P = P[categoria\ de\ exito] de\ una\ variable\ dominal/ordinal$

$\propto = nivel\ de\ significancia\ (o\ error\ esperado)$

$100-\propto : Nivel\ de\ confianza$

$IC\ de\ P_u = se\ lee\ "el\ Intervalo\ de\ confianza\ de\ una\ proporcion\ poblacional.$

$P_s = Proporcion\ Muestral.$

$Z_\propto :\ Puntuacion\ Z\ critica\ que\ corresponde\ al\ nivel\ estipulado\ de\ sign\ y\ conf.$

S_{ps} : Error Estandar Estimado de un intervalo de confianza de una proporcion.

Cuando calcular un intervalo de confianza de una proporción de la población (para una variable nominal/ordinal)

Cuando debemos proporcionar una estimación de un intervalo del valor de un parámetro de la población, Pu, donde P = p[de la categoría de éxito] de una variable nominal/ordinal.

Cuando tenemos una sola muestra representativa de la población.

Cuando el tamaño de la muestra (n) es lo suficientemente grande que (pmenor)(n)>=5, resultado en una distribución muestral que es normal.

Calculo del Error Estándar de un Intervalo de Confianza de una Proporción de la Población (para variable nominal/ordinal)

El error estándar es calculado a base de los datos muestrales, teniendo la siguiente formula.

$$Sp_S = \sqrt{\frac{P_S Q_S}{n}}$$

Donde:

Sp_s : $Error\ estandar\ estimado\ de\ proporciones\ para\ una\ variable\ de\ nominal\ con\ P = p[categoria\ de\ exito]$

P_s : $p[de\ la\ categoria\ de\ exito\ en\ la\ muestra]$

Q_s : $p[de\ la\ categoria\ de\ fracaso\ en\ la\ muestra] = 1 - P$

n : $Tamaño\ de\ la\ muestra.$

Cálculo del Término del Error de un Intervalo de Confianza de una Proporción de la Población

$$Termino\ del\ error = (Z_\propto)(Sp_s)$$

Donde:

α : $Nivel\ de\ significancia\ o\ Error\ esperado$

$Z_\propto$: $Puntuacion\ Z\ critica\ que\ corresponde\ a\ los\ niveles\ estipulados\ de\ significancia\ y\ confianza.$

Sp_s : $Error\ estandar\ estimado\ de\ proporciones\ para\ una\ variable\ nominal/ordinal\ donde\ P = p[de\ la\ categoria\ de\ exito]$

Cálculos de Intervalos de Confianza de 95% y 99% de una Proporción de la Población cuando (pmenor)(n)>=5

$$IC\ de\ 95\%\ de\ P_u = P_s \pm (1.96)(Sp_s)$$

$$IC\ de\ 99\%\ de\ P_u = P_s \pm (2.58)(Sp_s)$$

Donde:

$P = p[de\ la\ categoria\ de\ exito]de\ una\ variable\ nominal/ordinal$

$IC\ 95\%\ de\ P_u$: $intervalo\ de\ confianza\ de\ 95\%\ de\ una\ proporcion\ de\ una\ pobl.$

$IC\ 99\%\ de\ P_u$: $intervalo\ de\ confianza\ de\ 95\%\ de\ la\ poblacion$

$P_s\ proporcion\ de\ la\ muestra$

Sp_s : $Error\ estandar\ estimado\ de\ un\ inervalo\ de\ confianza\ de\ una\ proporc.$

Ejemplo de Aplicación para la Estimación de Parámetros Mediante Intervalos de Confianza de una Proporción de la Población

La empresa de lavadoras marca SpeedWash, quiere saber la proporción de amas de casa que prefieren usar su marca en la ciudad Nuevo Horizonte. El gerente general afirma que por lo menos el 30% de las amas de casa prefieren dicho producto. Una muestra de 400 amas de casa indico que 95 usarían dicha marca. Calcular el intervalo de confianza del 95% e indicar si el gerente tiene la razón.

Paso 1. Pregunta de investigación: con una seguridad de 95%, ¿podemos concluir que el gerente tenía la razón?, es decir que por lo menos el 33% de las amas de casa prefiere la marca de la empresa.

$$P = p[de\ los\ que\ prefieren\ la\ marca\ SpeedWash]$$

$$Q = p[de\ los\ que\ prefieren\ otra\ marca]$$

P=p[de los que prefieren la marca SpeedWash]

Q=p[de los que prefieren otra marca]

Muestra:

n = 450 encuestados

de encuestados que prefieren la marca de la empresa = 80

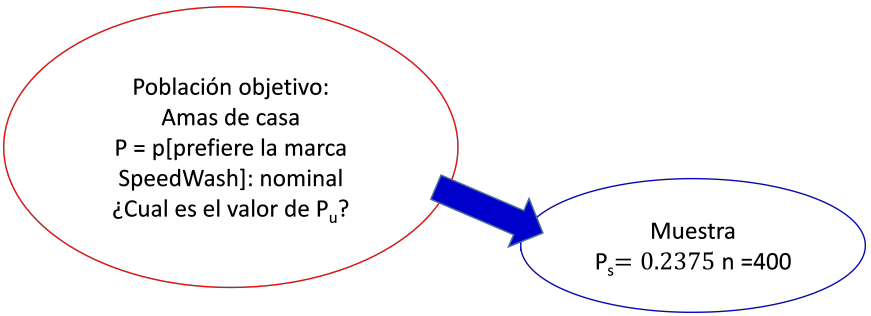
$$P_s = \frac{\#\ \text{pref. la SpeedWash}}{\#\ \text{pref. otra marca}} = \frac{95}{400} = 0.2375$$

$$Q_s = 1 - P_s = 1 - 0.2375 = 0.7625$$

Verificamos si n es lo suficientemente grande si (pmenor)(n)>5

$$(p_{menor})(n) = (0.2375)(400) = 95, entonces\ 95 > 5$$

Por lo tanto podemos continuar con los cálculos.



Paso 2. Calculo del error estándar y termino de error.

$$Sp_S = \sqrt{\frac{P_S Q_S}{n}} = \sqrt{\frac{(0.2375)(0.7625)}{400}} = 0.0213$$

Para una seguridad de 95%, Z_α=1.96

$$Termino\ del\ error = (Z_\alpha)(Sp_s) = (1.96)(0.0213) = 0.0417$$

Paso 3. Calculo de Límite de confianza inferior (LCI) y Límite de confianza superior (LCS)

$$IC\ de\ 95\%\ de\ P_u = P_s \pm (1.96)(Sp_s)$$

$$= 0.2375 \pm (1.96)(0.0213)$$

$$= 0.2375 \pm 0.0417$$

$$proporcion\ de\ la\ muestra \pm termino\ del\ error$$

$$LCI = 0.2375 - 0.0417 = 19.58\%$$

$$LCS = 0.2375 + 0.0417 = 27.92\%$$

Paso 4. Interpretación en lenguaje común: Estoy 95% seguro de que el porcentaje de amas de casa de la ciudad Nuevo Horizonte, tienen una preferencia entre 19.58% y 27.92% por la marca de lavadoras SpeedWash. Entonces podemos decir que el gerente de la empresa está equivocado al pensar que por lo menos el 30% de las amas de casa preferían la marca SpeedWash.

Paso 5. Interpretación estadística ilustrando la noción de "seguridad en el procedimiento": "si se realizan los mismos procedimientos de muestreo y estadístico 100 veces, 95 veces el parámetro poblacional real, Pu, estará comprendido en los intervalos calculados y 5 veces no. Por lo tanto, "estoy 95% seguro de que el intervalo de confianza individual que calcule incluye al parámetro real".

10.5 SELECCIÓN DE UN TAMAÑO DE LA MUESTRA PARA ELECCIONES, ENCUESTAS, Y ESTUDIOS DE INVESTIGACIÓN

El tamaño de la muestra es un componente importante en el tamaño del error estándar, en estas ecuaciones lo ubicamos en el denominador (n), mientras más grande sea "n" mejor será la muestra y producirá un error estándar pequeño.

$$Termino\ de\ error = (Z_\alpha)(Sp_S) = (Z_\alpha)\sqrt{\frac{P_S Q_S}{n}}$$

Entonces de esta ecuación podemos despejar el tamaño de la muestra "n";

$$n = \frac{(P_S Q_S)(Z_\alpha)^2}{\text{termino del error}^2}$$

Donde:

$n = tamaño\ muestral\ necesario$

$Z_\alpha = puntuacion\ Z\ que\ correponde\ al\ nivel\ de\ confianza\ y$

$significancia\ estipulados\ (por\ ejemplo, Z_\alpha = 1.96\ para\ NC = 95\%)$

$P_s =: p[de\ la\ categoria\ de\ exito\ en\ la\ muestra]$

$Q_s =: p[de\ la\ categoria\ de\ fracaso\ en\ la\ muestra] = 1 - P$

$termino\ del\ error = precision\ deseada\ en\ los\ resultados\ que\ se\ reportaran$

Ejemplo de Cálculo del Tamaño Muestral para el Intervalo de Confianza de una Proporción Poblacional de Variable nominal/ordinal

En las elecciones de una segunda vuelta, quedan 2 candidatos favoritos, y se quiere saber con una seguridad de 95% si es probable que el candidato A gane las elecciones. Se pide determinar el tamaño de la muestra. Los datos proporcionados son: termino de error ±3%, al no conocerse las proporciones a favor de candidato A ni B, se asume como valor razonable de 0.5 para cada posibilidad.

Entonces tendríamos lo siguiente:

$$n = \frac{(P_S Q_S)(Z_\alpha)^2}{\text{termino del error}^2} = \frac{(0.5)(0.5)(1.96)^2}{0.03^2} = 1067\ encuestas$$

Una consideración para reducir el tamaño de la muestra seria aumentar el término de error, por decir si estamos dispuesta a aceptar un mayor error, por ejemplo de ±5%, el tamaño de la muestra bajaría a 384 encuestas.

10.6 PREGUNTAS Y RESPUESTAS DE REPASO

1. El peso (en kg.) de las personas que viven en Lima Metropolitana sigue una distribución N(μ,18.5). Se han tomado los pesos de 20 personas seleccionadas aleatoriamente, y los resultados fueron los siguientes:

71.0	105.6	61.9	75.9	95.0
113.8	51.8	73.3	108.7	89.7
98.4	65.0	82.0	62.6	117.6
58.6	70.4	119.6	97.3	64.3

a. Obtener los intervalos de confianza al 90%, 95% y 99% para la media poblacional.

b. Determinar cuál sería el tamaño muestral necesario para conseguir, con un 95% de confianza, un intervalo de longitud igual a 2.5 kg.

c. Suponiendo ahora que σ es desconocida, calcular los intervalos de confianza para la media al 90%, 95% y 99%.

Solución

a)

Se sabe: $\bar{x}$=84.1250,σ=18.5

El intervalo de confianza para la media poblacional μ de varianza conocida se define de la siguiente manera:

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Al N.C = 90% (α = 0.1):

$$\left[84.1250 - 1.645\frac{18.5}{\sqrt{20}}, 84.1250 + 1.645\frac{18.5}{\sqrt{20}}\right] = [77.3207, 90.9293]$$

Al N.C = 95% (α = 0.05):

$$\left[84.1250 - 1.96\frac{18.5}{\sqrt{20}}, 84.1250 + 1.96\frac{18.5}{\sqrt{20}}\right] = [76.0172, 92.2328]$$

Al N.C = 99% (α = 0.01):

$$\left[84.1250 - 2.576\frac{18.5}{\sqrt{20}}, 84.1250 + 2.576\frac{18.5}{\sqrt{20}}\right] = [73.4695, 94.7805]$$

b)

$$Longitud = \left(\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{2z_{\alpha/2}\sigma}{Longitud}\right)^2$$

Reemplazando:

$$n = \left(\frac{2(1.96)(18.5)}{2.5}\right)^2 \approx 841\ personas$$

c)

Debido a que la muestra es pequeña (n ≤ 30), el intervalo de confianza se define de la siguiente manera:

$$\left[\bar{x} - t_{\frac{\alpha}{2}}(n-1)\frac{s_x}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}(n-1)\frac{s_x}{\sqrt{n}}\right]$$

Donde:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{8{,}813.9575}{20-1}} = 21.5382$$

Al N.C = 90% (α = 0.1):

$$\left[84.1250 - 1.7291\frac{21.5382}{\sqrt{20}}, 84.1250 + 1.7291\frac{21.5382}{\sqrt{20}}\right] = [75.7974, 92.4526]$$

Al N.C = 95% (α = 0.05):

$$\left[84.1250 - 2.093\frac{21.5382}{\sqrt{20}}, 84.1250 + 2.093\frac{21.5382}{\sqrt{20}}\right] = [74.0448, 94.2052]$$

Al N.C = 99% (α = 0.01):

$$\left[84.1250 - 2.8609\frac{21.5382}{\sqrt{20}}, 84.1250 + 2.8609\frac{21.5382}{\sqrt{20}}\right] = [70.3465, 97.9035]$$

2. Una muestra aleatoria extraída de una población normal de varianza 85, presenta una media muestral de 174.25. Con una muestra de tamaño 150, se pide lo siguiente:
   a. Calcular un intervalo de confianza al 95% para la media poblacional.
   b. Calcular un intervalo de confianza al 90% para la media poblacional.
   c. Comparar ambos intervalos, desde el punto de vista de la información que generan.
   d. Si se quiere tener una confianza al 95% de que su estimación se encuentra a una distancia de 0.63 unidades de la verdadera media poblacional, ¿cuántas observaciones adicionales deberían tomarse?

Solución

Se sabe: $\bar{x}$ =174.25,σ=9.2195

a)

Al N.C = 95% (α = 0.05):

$$\left[174.25 - 1.96\frac{9.2195}{\sqrt{150}}, 174.25 + 1.96\frac{9.2195}{\sqrt{150}}\right] = [172.7746,175.7254]$$

b)

Al N.C = 90% (α = 0.1):

$$\left[174.25 - 1.645\frac{9.2195}{\sqrt{150}}, 174.25 + 1.645\frac{9.2195}{\sqrt{150}}\right] = [173.0118,175.4882]$$

c)

Al calcular la longitud de cada intervalo de confianza, se tiene lo siguiente:

$$L_{95\%} = 175.7254 - 172.7746 = 2.951$$

$$L_{90\%} = 175.4882 - 173.0118 = 2.476$$

Se observa que el segundo intervalo de confianza es de menor longitud, por lo que podría parecer más preciso. Sin embargo, su nivel de confianza es también menor.

d)

El error absoluto (e) del intervalo de confianza es el siguiente:

$$e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{Z_{\alpha/2}\sigma}{e}\right)^2$$

Reemplazando:

$$n = \left(\frac{(1.96)(9.2195)}{0.63}\right)^2 \approx 823\ observaciones$$

Por lo tanto, se debería tomar una muestra adicional de 673 observaciones.

3. La cantidad de personas que visitaron Machu Picchu durante el verano, medida a través de una muestra aleatoria de 20 días, fueron las siguientes:

4,103	3,726	3,716	3,863	3,428
3,780	3,608	3,886	3,574	4,050
3,620	4,077	3,497	3,886	3,636
4,168	3,671	3,726	3,607	3,963

Suponga que los niveles de asistencia siguen una distribución normal, con desviación típica muestral de 207.92.

a. ¿Se podría afirmar, con un 95% de confianza, que la asistencia diaria promedio a Machu Picchu es de 3,750 personas?

b. El Ministerio de Cultura indica que la asistencia media es constante y que la dispersión sería de unas 150 personas. ¿Se puede probar esta afirmación con los datos disponibles, a un 95% de confianza?

Solución

a)

Se sabe: $\bar{x}$=3,779.25

El intervalo de confianza para la media μ de una distribución normal de varianza poblacional desconocida (n ≤ 30), se define de la siguiente manera:

$$\left[\bar{x} - t_{\frac{\alpha}{2}}(n-1)\frac{s_x}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}(n-1)\frac{s_x}{\sqrt{n}}\right]$$

Donde:

$$s_x = \sqrt{\frac{n\sigma_x^2}{n-1}}$$

Reemplazando:

$$s_x = \sqrt{\frac{20(207.92)^2}{20-1}} = 213.3214$$

Al N.C = 95% (α = 0.05):

$$\left[3{,}779.25 - 2.093\frac{213.3214}{\sqrt{20}}, 3{,}779.25 + 2.093\frac{213.3214}{\sqrt{20}}\right] = [3{,}679.4125, 3{,}879.0875]$$

Debido a que 3,750 se encuentra dentro del intervalo de confianza, se puede afirmar, con un 95% de confianza, que la asistencia diaria promedio es de 3,750 personas.

b)

El intervalo de confianza para la desviación σ de una distribución normal es la siguiente:

$$\left[\sqrt{\frac{(n-1)s_x^2}{\chi^2_{\alpha/2}(n-1)}}, \sqrt{\frac{(n-1)s_x^2}{\chi^2_{1-\alpha/2}(n-1)}}\right]$$

Donde:

$$\chi^2_{0.025}(19) = 32.8523$$
$$\chi^2_{0.975}(19) = 8.9065$$

Reemplazando:

$$\left[\sqrt{\frac{(20-1)(213.3214)^2}{32.8523}}, \sqrt{\frac{(20-1)(213.3214)^2}{8.9065}}\right] = [162.2289, 311.5712]$$

Debido a que 150 no se encuentra dentro del intervalo de confianza, no se puede afirmar, con un 95% de confianza, que la dispersión de la asistencia sea de 150 personas.

4. En Estados Unidos, los dirigentes del partido demócrata afirman que la intención de voto del partido republicano para las elecciones del 2020, en California, es la misma que en Nueva York. Para ello, se realizó una encuesta a 350 personas en California, donde 116 mostraron su apoyo al partido republicano, y a otras 350 personas en Nueva York, donde 131 eligieron al partido republicano.

   a. Construir un intervalo de confianza al 90% para la proporción de personas que votarían por el partido republicano en California.

   b. ¿A cuántas personas habría que encuestar para obtener un margen de error o error de estimación de ± 2,5%, al nivel de confianza anterior?

   c. Construir un intervalo de confianza al 90% para la diferencia de proporciones en la estimación del voto del partido republicano en los dos estados. ¿Se puede afirmar que los dirigentes del partido demócrata tienen razón?

Solución

Se sabe:

$$\hat{p}_1 = \frac{116}{350} = 0.3314$$

$$\hat{q}_1 = 1 - \hat{p}_1 = 0.6686$$

$$\hat{p}_2 = \frac{131}{350} = 0.3743$$

$$\hat{q}_2 = 1 - \hat{p}_2 = 0.6257$$

a)

Como el tamaño de la muestra es grande (n1 = 350), se puede utilizar la aproximación normal, donde el intervalo de confianza se define de la siguiente manera:

$$\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Al N.C = 90% (α = 0.1):

$$\left[0.3314 - 1.645\sqrt{\frac{(0.3314)(0.6686)}{350}}, 0.3314 + 1.645\sqrt{\frac{(0.3314)(0.6686)}{350}}\right]$$

$$[0.2900, 0.3728]$$

b)

El error de estimación (e) del intervalo de confianza es el siguiente:

$$e = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(\hat{q})}{n}}$$

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2(\hat{p})(\hat{q})}{e^2}$$

El caso más desfavorable se dará cuando

$$\hat{p} = \hat{q} = 0.5$$

Reemplazando:

$$n = \frac{(1.645)^2(0.5)(0.5)}{0.025^2} \approx 1{,}082\ personas$$

c)

El intervalo de confianza para la diferencia de parámetros poblacionales (p1 - p2) de dos distribuciones binomiales, con tamaños de muestra grandes (n1 = n2 = 350), se puede determinar mediante aproximación normal, donde el intervalo sería el siguiente:

$$\left[(\hat{p}_1-\hat{p}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},(\hat{p}_1-\hat{p}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

Reemplazando:

$$\left[(0.3314-0.3743)-1.645\sqrt{\frac{(0.3314)(0.6686)}{350}+\frac{(0.3743)(0.6257)}{350}},(0.3314-0.3743)\right.$$
$$\left.+1.645\sqrt{\frac{(0.3314)(0.6686)}{350}+\frac{(0.3743)(0.6257)}{350}}\right]$$
$$[-0.1022,0.0165]$$

Debido a que 0 se encuentra dentro del intervalo de confianza, no existe diferencia significativa entre la intención de voto del partido republicano en ambos estados, por lo que los dirigentes del partido demócrata tienen razón, a un nivel de confianza del 90%.

5. Un equipo de nutricionistas está interesado en ver si una dieta basada en frutos secos reduce el nivel de glucosa en la sangre. Para ello, toma una muestra de 15 pacientes y determina sus niveles de glucosa antes y después del tratamiento. Los datos obtenidos, expresados en mg/dL, fueron los siguientes:

Antes	Después
89	103
118	85
90	115
105	131
128	102
134	117
129	138
125	83
108	89
111	96
95	104
106	93
131	115
94	103
118	119

a. Construya un intervalo de confianza al 95% para la diferencia del nivel medio de glucosa en la sangre antes y después del tratamiento.

Solución

a)

Debido a que la muestra es pequeña (n ≤ 30), el intervalo de confianza se define de la siguiente manera:

$$\left[\bar{d} - t_{\frac{\alpha}{2}}(n-1)\frac{s_d}{\sqrt{n}}, \bar{d} + t_{\frac{\alpha}{2}}(n-1)\frac{s_d}{\sqrt{n}}\right]$$

Donde:

$$d_i = x_{antes} - x_{después}$$

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{88}{15} = 5.8667$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}} = \sqrt{\frac{6{,}053.7333}{15-1}} = 20.7945$$

Al N.C = 95% (α = 0.05):

$$\left[5.8667 - 2.1448\frac{20.7945}{\sqrt{15}}, 5.8667 + 2.1448\frac{20.7945}{\sqrt{15}}\right]$$

$$= [-5.6489, 17.3823]$$

Debido a que 0 se encuentra dentro del intervalo de confianza, no existe diferencia significativa en la diferencia del nivel medio de glucosa en la sangre antes y después del tratamiento, a un nivel de confianza del 95%.

CAPÍTULO XI

CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE

11.1. INTRODUCCIÓN

En la investigación estadística, al igual que en los eventos cotidianos que vivimos, afrontamos con gran frecuencia escenarios que tratamos de relacionar y explicar por medio de variables que suponemos tienen una dependencia o relación. Es así como, desde el comienzo de nuestra historia, tenemos la necesidad de explicar estos hechos y para ello, recurrimos a las matemáticas, las cuales han constituido y constituyen un mecanismo clave para brindar mayor precisión a los cálculos y por ende, una mejor predicción.

Es sumamente importante analizar dentro del estudio las variables sobre las cuales se tiene una ligera sospecha de la existencia de una relación, en algunos escenarios, la relación será irrebatible y en otros, en menor medida; esto depende mucho de las experiencias, información o conocimientos de los investigadores. Debe quedar completamente claro que dentro del presente capítulo no se explica causalidad sino, solamente, correlación.

En este capítulo, denominado Correlación y regresión lineal simple, se tratan los siguientes puntos: modelo de regresión lineal simple, diagrama de esparcimiento y método de los mínimos cuadrados, Interpretación de la pendiente de regresión b, correlación lineal, coeficiente de correlación lineal, fórmulas alternativas para el cálculo de r y coeficiente de determinación.

11.2. MODELO DE REGRESIÓN LINEAL SIMPLE

El modelo de regresión lineal simple se emplea la relación lineal entre una variable independiente "Y" explicada mediante un variable dependiente "X", en otras palabras, que existe una relación entre las variables X e Y la cual se expresa de la siguiente manera:

Y=f(X); comprendida como "Y en función de X"

Para elaborar el análisis debemos contar inicialmente con los datos sobre los valores para X e Y, pues con estos datos podremos estructurar el modelo matemático probabilístico. De la ecuación se desprende que cada valor observado de Y nace a partir de un valor de X, adicionando un valor constante y un error que no podrá ser explicado por modelo (salvo que guarde una correlación perfecta, en este caso el error es igual a cero).

$$Y = \alpha + \beta X + \varepsilon$$

Donde:

α: ordenada de origen que toma un valor constante

β: pendiente de la regresión (inclinación)

ε: error o residuo del modelo

Seguidamente, en el ejercicio de la estimación podemos decir que para cada valor de X se cuenta con un valor esperado de Y*; por ende, el error se expresa como la diferencia entre el valor real contra el valor estimado.

$$\varepsilon = Y - Y^*$$

Y considerando el principio de linealidad tenemos que el valor esperado para ε es igual a cero, pues todos los puntos esperados se superponen sobre la línea de regresión, quedando finalmente la expresión con la que trabajaremos, la forma para obtener los parámetros "a" y "b" serán explicados en el punto 8.2:

$$Y^* = a + bX$$

Otros principios que acompañan a la regresión simple son de:

Independencia: Las variables X e Y son variables independientes, por tal motivo se asume que los errores son independientes estadísticamente.

Igualdad de varianzas: Bajo la suposición que las varianzas $\sigma^2_{Y_i}$ de Y_i en cada X_i son iguales a la varianza común σ^2 conocida como varianza de la regresión.

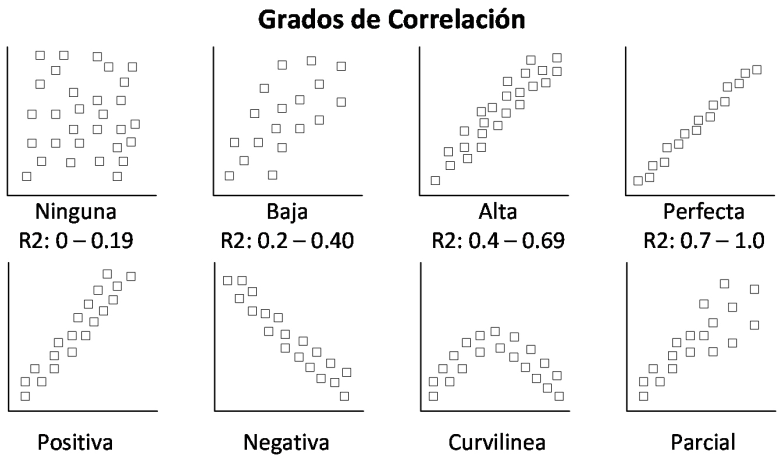
Normalidad: Bajo la suposición que cada variable aleatoria Y_i sigue una distribución normal, partido de esta idea se puede suponer que ε_i sigue una distribución normal.

11.3. DIAGRAMA DE ESPARCIMIENTO Y MÉTODO DE LOS MÍNIMOS CUADRADOS

DIAGRAMA DE ESPARCIMIENTO

El diagrama de esparcimiento se emplea para representar las "n" observaciones bidimensionales agrupándolas en pares de variables (X;Y). Los valores obtenidos se trasladan gráficamente mediante puntos a un plano cartesiano.

Esta representación otorgará al investigador una idea general sobre el comportamiento de los datos como se muestra en las siguientes imágenes.



MÉTODO DE LOS MÍNIMOS CUADRADOS

El método de los mínimos cuadros se construye a partir de variables bidimensionales (X;Y) para hallar una correspondencia entre la nube de puntos de los "n" datos observados. Se busca una función matemática que represente estadísticamente la relación entre variables, la representación con menor error o sesgo se considerará como una mejor representación. Es por ello que la mejor curva es la cual minimiza la suma de los cuadrados de las desviaciones entre los puntos de correspondencia Yi y Y*.

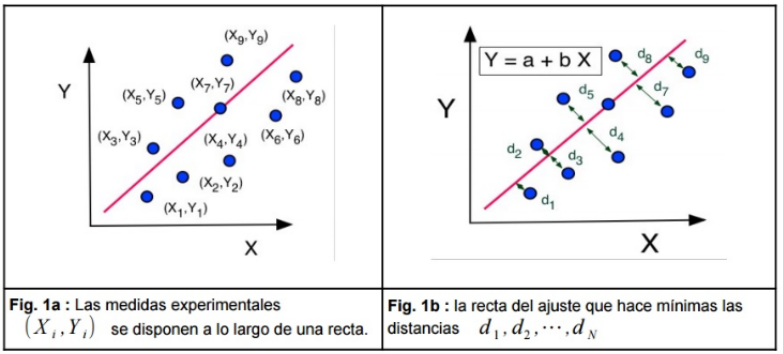


Fig. 1a : Las medidas experimentales (X_i, Y_i) se disponen a lo largo de una recta.

Fig. 1b : la recta del ajuste que hace mínimas las distancias $d_1, d_2, \cdots, d_N$

Se identifican los puntos y se trazan las distancias a la recta con la menor distancia.

Donde se busca minimizar:

$$\phi = \sum \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - Y_i^*)^2 = \text{Mínimo}$$

Donde:

Y_i: Valor observado

Y^*_i: Valor estimado a partir de la ecuación de regresión $Y^* = a + bX$

n: número de pares observados (X;Y)

Para $Y^*=f(X)$ es la ecuación definida para expresar la recta de regresión donde se minimiza el error por el método de los mínimos cuadrados.

Ahora, reemplazamos la sumatoria de Y* por a + bX se tiene:

$$Min\ \phi = \sum_{i=1}^{n} (Y_i - a - bX)^2$$

Para calcular el mínimo se procede a derivar ϕ respecto a los parámetros "a" y "b", teniendo un caso de derivación parcial.

Las derivaciones parciales se igual a cero obteniendo:

$$\frac{d\phi}{da} = 0 \ \ y \ \ \frac{d\phi}{db} = 0$$

Resultado dos ecuaciones:

$$\frac{d\phi}{da} = 2\sum (Y - a - \ bX)(-1) = 0$$

(2)

$$\frac{d\phi}{db} = 2\sum (Y - a - \ bX)(-X) = 0$$

Donde:

(1)

$$\sum (Y - a - \ bX) = 0$$

(2)

$$\sum (Y - a - \ bX)(X) = 0$$

Y por medio de propiedades de sumatoria se obtiene dos ecuaciones con dos incógnitas, las cuales son "a" y "b". Las cuales deberán ser reemplazadas en la ecuación de regresión lineal

$$\sum Y = an + b \sum X$$

$$\sum XY = a \sum X + b \sum X^2$$

Recordemos que los valores X y Y son los obtenidos de las observaciones realizadas por el investigador. El valor "n" es el límite superior de la sumatoria.

Finalmente realizando el reemplazo de los valores correspondiente se logra la ecuación de estimación:

$$Y^* = a + bX$$

11.4 INTERPRETACIÓN DE LA PENDIENTE DE REGRESIÓN B.

Como mencionamos anteriormente, en la estimación de la recta de regresión simple el parámetro "a" indica el cruce de la recta con la ordenada en el origen.

Por otro lado, el coeficiente b es la pendiente de la recta de regresión y se entiende como la razón de cambio promedio. Es decir, cuál es el cambio promedio en Y cuando X cambia en una unidad de medición. Por ejemplo, el valor de X cambia en "z" unidades, entonces el valor de Y cambiará en "z" unidades multiplicadas por "b".

De la misma manera, el parámetro "b" indica la tendencia del modelo de regresión.

Si b > 0; entonces la tendencia lineal es creciente, por ello a medida que se incrementa X se tiende a incrementar Y

Si b < 0; entonces la tendencia línea es decreciente, por ello a medida que se reduce X se tiende a reducir Y.

Si b = 0; no existe regresión muestra, puesto a medida que se incrementa o disminuye X el valor de Y permanece estacionario.

11.5 CORRELACIÓN LINEAL

La correlación indica el grado de afinidad o asociación entre las variables estudiadas; así mismo, la correlación también explica el grado de bondad de ajuste de la recta de regresión. En términos generales, la correlación denota la interdependencia entre las variables.

Cuando las variables se encuentran asociadas, la ecuación de la función de regresión

explica el comportamiento de la variable dependiente en función de la variable independiente.

Se debe considerar que existe la correlación simple al hallar la dependencia entre dos variables, correlación línea cuando se habla de la recta de regresión y correlación no lineal si la función no corresponde a la ecuación de una recta.

11.6 COEFICIENTE DE CORRELACIÓN LINEAL

El grado de afinidad o asociación entre las variables relacionadas se conoce como coeficiente de correlación rectilínea. Se denota como "r" y se expresa de la siguiente forma:

$$r = \sqrt{\frac{S_{y*}^2}{S_y^2}}$$

Donde:

$S_{y^*}^2$: **Varianza explicada.** Es aquella varianza del total de Y que es explicada por la recta de regresión.

S_y^2 : **Varianza total.** Es la varianza total de los valores observados de Y.

Estas dos varianzas se definen como:

$$S_{y*}^2 = \frac{\sum(Y^* - \bar{Y})^2}{n} \; y \; S_y^2 = \frac{\sum(Y - \bar{Y})^2}{n}$$

Reemplazando se obtiene:

$$r = \sqrt{\frac{S_{y*}^2}{S_y^2}} = \sqrt{\frac{\sum(Y^* - \bar{Y})^2}{\sum(Y - \bar{Y})^2}}$$

Es importante saber que la varianza total se forma como la suma de la varianza explicada más la varianza no explicada.

Varianza Total = Varianza Explicada + Varianza No Explicada

$$\frac{\sum(Y - \bar{Y})^2}{n} = \frac{\sum(Y^* - \bar{Y})^2}{n} + \frac{\sum(Y - Y^*)^2}{n}$$

$$S_y^2 = S_{y*}^2 + S_{yx}^2$$

Considerando al Coeficiente de Correlación

$$r^2 = \frac{S_{y*}^2}{S_y^2}$$

Como

$$S_{y*}^2 = S_y^2 - S_{yx}^2$$

Reemplazando se obtiene:

$$r = \sqrt{\frac{S_y^2 - S_{yx}^2}{S_y^2}} = \sqrt{1 - \frac{S_{yx}^2}{S_y^2}}$$

Que define la expresión para calcular el Coeficiente de Correlación, en la fórmula se sabe que:

$$S_{yx}^2 = \frac{\sum Y^2 - a\sum Y - \sum XY}{n}$$

$$S_y^2 = \frac{\sum Y^2}{n} - \left(\frac{\sum \bar{Y}}{n}\right)^2$$

Ambos pueden calcularse de manera independiente y por ende, ser reemplazadas para realizar el cálculo de "r" o Coeficiente de Correlación el cual se encuentra entre el rango y su signo es el mismo de la pendiente de la recta de regresión "b".

$$-1 \leq r \leq 1$$

De los valores calculados para el "r" se pueden interpretar de la siguiente manera:

- Si r > 0; se presenta una correlación directa positiva.
- Si r = +1; se presenta una correlación perfecta positiva.
- Si r < 0; se presenta una correlación inversa negativa.
- Si r = -1; se presenta una correlación perfecta negativa.
- Si r2 = 1; se presenta una correlación rectilínea.
- Si r = 0; los datos no guardan correlación.

La significancia del resultado obtenido en "r" está dado según:

$0.00 \leq r < 0.20$ presenta una correlación no significativa.

$0.20 \leq r < 0.40$ presenta una correlación baja.

$0.40 \leq r < 0.70$ existe una correlación significativa.

$0.70 \leq r < 1.00$ existe un algo grado de correlación.

Estos rangos nos son definitivos, pues debe considerarse la materia en investigación, así como los datos y el tamaño de la muestra.

11.7 FÓRMULAS ALTERNATIVAS PARA EL CÁLCULO DE R

Adicionalmente al método inicial propuesto se desarrollan dos formas alternativas de calcular el coeficiente de correlación.

11.7.1 Fórmula de Pearson

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (Y)^2]}}$$

11.7.2 Fórmula de Covarianza

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

Donde:

$$Cov\,(X;Y) = \frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)$$

$$S_Y^2 = \frac{\sum Y^2}{n} - \bar{Y}^2 \qquad S_Y^2 = \frac{\sum X^2}{n} - \bar{X}^2$$

Recordando que $\bar{X}$ e $\bar{Y}$ son la media de los valores de X e Y.

11.8. COEFICIENTE DE DETERMINACIÓN

El coeficiente de determinación R^2 es un estadístico usado en el contexto de un modelo estadístico cuyo principal propósito es predecir posibles o futuros resultados. Este coeficiente determina el nivel de explicación de los resultados. El coeficiente de

determinación "R^2"es obtiene al elevar al cuadrado el coeficiente de correlación "r"; sus valores pueden variar entre 0 y 1.

$$R^2=r^2$$

Si obtenemos un R^2 con un valor de 0.86 nos indicaría que los resultados se explican en un 86% mediante el modelo empleado y no se explican en un 14%.

11.9. PREGUNTAS Y RESPUESTA DE REPASO

1. 1. En un estudio llevado a cabo en Estados Unidos, se realizaron exámenes médicos a 10 pacientes, a fin de ver si es que existe una relación entre los niveles de colesterol y glucosa en sangre. Los datos obtenidos se muestran en la siguiente tabla:

Nivel de colesterol (mmol/L)	Nivel de glucosa (mmol/L)
2.86	6.56
2.58	5.71
4.23	6.90
4.32	7.27
4.97	7.91
4.51	5.83
5.72	6.92
3.21	6.28
2.68	4.72
4.82	6.77

   a. Construya un diagrama de dispersión para estos datos.

   b. ¿Existe evidencia de relación lineal entre los niveles de colesterol y glucosa?

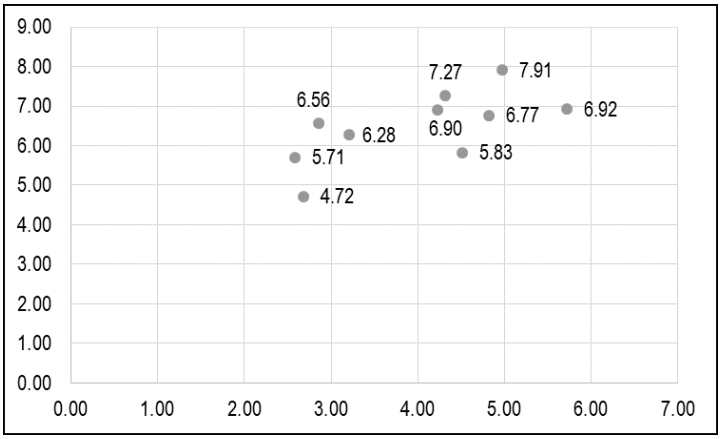
   c. Estime el nivel de glucosa cuando el nivel de colesterol es de 3.85 mmol/L.

   d. ¿Cuál es el porcentaje de variación del nivel de glucosa explicado por el nivel de colesterol?

Solución

a)

Los valores del nivel de colesterol se identifican en el eje de abscisas x, mientras que el nivel de glucosa en el eje de ordenadas y.



El diagrama de dispersión no muestra una clara tendencia lineal.

b)

Se sabe:

$$\bar{x} = 3.9900,\ S_x = 1.0888,\ \bar{y} = 6.4870,\ S_y = 0.8995,\ \sum_{i=1}^{10} x_i y_i = 264.7150$$

El coeficiente de correlación lineal se define de la siguiente manera:

$$r_{x,y} = \frac{Cov(x,y)}{S_x S_y}$$

Donde:

$$Cov(x,y) = \frac{\sum_{i=1}^{10} x_i y_i}{n} - \bar{x}\bar{y} = \frac{264.7150}{10} - (3.9900)(6.4870) = 0.5884$$

Entonces:

$$r_{x,y} = \frac{0.5884}{(1.0888)(0.8995)} = 0.6008$$

Debido a que el valor del coeficiente de correlación lineal de Pearson es 0.6008, no existe una evidencia fuerte de que un modelo lineal pueda ser bueno para predecir la relación entre el nivel de colesterol y el nivel de glucosa. En todo caso, se tiene una tendencia a relacionarse de manera directamente proporcional.

c)

Para efectuar una estimación es necesario haber decidido antes el modelo que se va a ajustar a los datos experimentales. Si tuviera la forma de una línea recta, se puede asumir lo siguiente:

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\mathrm{x}}$$

Donde:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x^2} = \frac{264.7150 - 10(3.9900)(6.4870)}{(10-1)(1.0888)^2} = 0.5515$$

$$\hat{\beta}_0 = 6.4870 - (0.5515)(3.9900) = 4.2865$$

Entonces:

$$\hat{y}_x = 4.2865 + 0.5515\mathrm{x}$$

Para x = 3.85:

$$\hat{y}_x = 4.2865 + 0.5515(3.85) = 6.4098$$

d)

El indicador que se define como porcentaje de variación de nivel de glucosa explicado por el nivel de colesterol es el coeficiente de determinación R^2.

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} = \frac{\sum_{i=1}^{10}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{10}(y_i - \bar{y})^2} = \frac{3.2448}{7.2820} = 0.4456$$

El porcentaje de variación de nivel de glucosa explicado por el nivel de colesterol es de 0.4456, lo cual es bastante bajo. Debería buscarse un modelo que haga mejores predicciones, dado que el modelo lineal simple es deficiente.

2. 2. Los datos de la siguiente tabla relacionan la solubilidad del cloruro de bario Y con la temperatura del agua (X en °C). A la temperatura indicada X, Y gramos de cloruro de bario se disuelven en 100 gramos de agua, obteniendo:

X	Y
0	31.6
10	33.3
20	35.7
30	38.2
40	40.7
50	43.6
60	46.4
70	49.4
80	52.4
100	58.8

a. Usando el método de mínimos cuadrados, obtenga los estimadores de los parámetros.

b. ¿Qué puede decir de la calidad del ajuste obtenido?

c. ¿Cree que, a mayor temperatura, existe mayor solubilidad del cloruro de bario?

d. Determine el error estándar de estimación con el modelo ajustado y obtenga un intervalo de longitud con dos errores estándar de estimación, centrado en la estimación de la solubilidad del cloruro de bario cuando la temperatura es de 47 °C.

Solución

a)

Se sabe:

$$\bar{x} = 46,\ S_x = 32.0416,\ \bar{y} = 43.01,\ S_y = 8.7974,\ \sum_{i=1}^{10} x_i y_i = 22{,}315$$

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\mathrm{x}}$$

Donde:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x^2} = \frac{22{,}315 - 10(46)(43.01)}{(10-1)(32.0416)^2} = 0.2739$$

$$\hat{\beta}_0 = 43.01 - (0.2739)(46) = 30.4128$$

Entonces:

$$\hat{y}_x = 30.4128 + 0.2739\mathrm{x}$$

b)

Puesto que el ajuste del modelo es lineal, el coeficiente de correlación entre X e Y es el siguiente:

$$r_{x,y} = \frac{Cov(x,y)}{S_x S_y}$$

Donde:

$$Cov(x,y) = \frac{\sum_{i=1}^{10} x_i y_i}{n} - \bar{x}\bar{y} = \frac{22{,}315}{10} - (46)(43.01) = 253.04$$

Entonces:

$$r_{x,y} = \frac{253.04}{(32.0416)(8.7974)} = 0.8977$$

El valor obtenido es cercano a 1 y, por tanto, este modelo es bueno para estimar los valores de Y sobre la base de los valores de X.

c)

La covarianza permite averiguar si las variables X e Y tienen una relación directa. Del inciso anterior:

$$Cov(x,y) = 253.04 > 0$$

Puesto que la covarianza es positiva, la relación entre X e Y es directa. A mayor temperatura, existe mayor solubilidad del cloruro de bario.

d)

Para x = 47 °C:

$$\hat{y}_x = 30.4128 + 0.2739(47) = 43.2839$$

Como el modelo es lineal, el error estándar de estimación se define de la siguiente manera:

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{3.5918}{10-2}} = 0.6701$$

Para determinar el intervalo centrado en la estimación de y cuando x = 47 °C, con una longitud de dos errores estándar, se procede como sigue:

$$[\hat{y}_x - S_{yx}, \hat{y}_x + S_{yx}] = [43.2839 - 0.6701, 43.2839 + 0.6701] = [42.6138, 43.9539]$$

3. 3. Una empresa del sector retail decide investigar si es que existe alguna relación entre los ingresos generados por compras con tarjetas de débito y crédito. Para ello, decidió tomar una muestra de 15 transacciones realizadas (en soles) en el último mes con ambos tipos de tarjeta y se obtuvo los siguientes datos:

Débito	Crédito
236.82	828.65
203.12	483.32
170.62	340.98
145.24	396.25
152.27	320.26
237.76	701.12
198.19	447.45
245.06	555.87
153.74	478.68
208.45	613.39
205.03	509.46
221.76	429.28
153.62	368.12
243.50	834.23
240.13	841.89

a. ¿Es bueno el modelo lineal para estimar las compras a crédito en base a las de débito? Justifique su respuesta con un indicador estadístico y escriba el modelo ajustado con los parámetros estimados.

b. Determinar un intervalo centrado en la estimación de las compras a crédito para una compra de S/ 252.56 con una tarjeta de débito y una longitud de 3 errores estándar de estimación.

c. ¿Cuál es el porcentaje de variación de las compras a crédito explicada por las compras a débito, a través de este modelo lineal ajustado?

Solución

a)

La calidad del modelo se puede calcular con el coeficiente de regresión lineal.

$\bar{x} = 201.0207$, $S_x = 37.1419$, $\bar{y} = 543.2633$, $S_y = 181.1268$, $\sum_{i=1}^{15} x_i y_i = 1{,}713{,}461.8785$

$$r_{x,y} = \frac{Cov(x,y)}{S_x S_y}$$

Donde:

$$Cov(x,y) = \frac{\sum_{i=1}^{15} x_i y_i}{n} - \bar{x}\bar{y} = \frac{1{,}713{,}461.8785}{15} - (201.0207)(543.2633) = 5{,}023.6345$$

Entonces:

$$r_{x,y} = \frac{5{,}023.6345}{(37.1419)(181.1268)} = 0.7467$$

Este coeficiente indica que el modelo lineal es relativamente bueno para estimar las compras a crédito en base a las de débito.

Luego:

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\mathrm{x}}$$

Donde:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{15} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x^2} = \frac{1{,}713{,}461.8785 - 15(201.0207)(543.2633)}{(15-1)(37.1419)^2} = 3.9017$$

$$\hat{\beta}_0 = 543.2633 - (3.9017)(201.0207) = -241.0582$$

Entonces:

$$\hat{y}_x = -241.0582 + 3.9017\mathrm{x}$$

b)

Para x = 252.56:

$$\hat{y}_x = -241.0582 + 3.9017(252.56) = 744.3541$$

Como el modelo es lineal, el error estándar de estimación se define de la siguiente manera:

$$S_{yx} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{165{,}286.4404}{15-2}} = 112.7579$$

Para determinar el intervalo centrado en la estimación de y cuando x = 252.56, con una longitud de 3 errores estándar, se procede como sigue:

$$\left[\hat{y}_x - 1.5S_{yx}, \hat{y}_x + 1.5S_{yx}\right] = [744.3541 - (1.5)(112.7579), 744.3541 + (1.5)(112.7579)]$$

$$[575.2173{,}913.4910]$$

c)

El indicador que se define como porcentaje de variación de nivel de compras con tarjetas de crédito explicada por las compras con tarjetas de débito es el coeficiente de determinación R^2.

$$R^2 = r_{x,y}^2 = (0.7467)^2 = 0.5576$$

El porcentaje de variación de nivel de compras con tarjetas de crédito explicada por las compras con tarjetas de débito es de 0.5576, lo cual es bajo. Debería buscarse un modelo que haga mejores predicciones, dado que el modelo lineal simple es deficiente.

4. Una empresa pesquera ha realizado un estudio acerca de la cantidad de peces que una flota recolecta en función de la temperatura del mar. En 12 días, considerando la temperatura promedio, se pescaron los siguientes volúmenes Y (en kg.):

Temperatura (°C)	Volumen
-4	203.54
-3	198.85
0	182.73
2	176.59
3	157.97
4	136.25
5	131.68
5	126.91
7	119.85
8	107.54
9	95.67
10	91.36

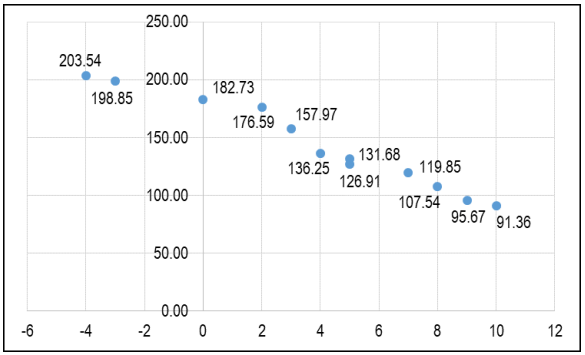
A la flota le interesa saber si mañana su volumen de pesca será de, por lo menos, 60 kg., a fin de que le sea rentable salir al mar.

a. Haga su diagrama de dispersión y ajuste la recta de mínimos cuadrados.

b. Si la empresa pesquera pronostica para mañana una temperatura media de 11 °C, ¿recomendaría o no, con un nivel de confianza de 95%, que la flota salga al mar? ¿Por qué?

Solución

a)



El diagrama de dispersión muestra una tendencia lineal inversa.

Se sabe:

$$\bar{x} = 3.8333,\ S_x = 4.4890,\ \bar{y} = 144.0783,\ S_y = 39.1198,\ \sum_{i=1}^{12} x_i y_i = 4{,}728.23$$

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\mathrm{x}}$$

Donde:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{12} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x^2} = \frac{4{,}728.23 - 12(3.8333)(144.0783)}{(12-1)(4.4890)^2} = -8.5686$$

$$\hat{\beta}_0 = 144.0783 - (-8.5686)(3.8333) = 176.9246$$

Entonces:

$$\hat{y}_x = 176.9246 - 8.5686\mathrm{x}$$

b)

Para x = 11:

$$\hat{y}_x = 176.9246 - 8.5686(11) = 82.67$$

Si bien, a esta temperatura, se está por encima de las 52 toneladas requeridas, no se puede garantizar que el volumen de pesca para este día específico lo supere. En tal sentido, para tomar la decisión, se debe hallar un intervalo de predicción al 95%.

$$\left[\hat{y}_x - t_{1-\frac{\alpha}{2}}(n-2)S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{y}_x + t_{1-\frac{\alpha}{2}}(n-2)S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]$$

Donde:

$$t_{0.975}(10) = 2.228$$

$$S_e = \sqrt{\frac{n-1}{n-2}\left(S_y^2 - \hat{\beta}_1^2 S_x^2\right)} = \sqrt{\frac{11(39.1198^2 - (-8.5686)^2(4.4890)^2)}{10}} = 7.4766$$

Reemplazando:

$$\left[82.67 - 2.228(7.4766)\sqrt{1+\frac{1}{12}+\frac{(11-3.8333)^2}{221.6667}}, 82.67 + 2.228(7.4766)\sqrt{1+\frac{1}{12}+\frac{(11-3.8333)^2}{221.6667}}\right]$$

$$[63.5664, 101.7736]$$

Dado que el intervalo supera el volumen de 60 kg. requerido, sí se recomendaría salir al mar.

5. Los datos de la producción del melón (X en toneladas) y el precio pagado al productor (Y en soles / tonelada) en las regiones del país que se cultiva, fueron los siguientes:

Región	Producción	Precio
Arequipa	1,555.70	575.59
Ica	4,380.12	673.11
La Libertad	433.00	964.40
Lambayeque	942.00	544.49
Lima	3,753.00	696.37
Loreto	3,698.00	528.29
Piura	501.00	1,081.96
Tacna	305.00	1,152.01
Tumbes	3,111.45	590.00
Ucayali	3,668.96	686.78

a. Ajuste la recta de regresión por el método de mínimos cuadrados.

b. Calcule la varianza residual (S_2^R).

c. Calcula un intervalo de confianza al 95% para la pendiente de la recta de regresión obtenida.

d. Contraste la hipótesis de que el precio pagado al productor depende linealmente de la producción de melón, usando un nivel de significación $\alpha = 0.05$.

Solución

a)

Se sabe:

$$\bar{x} = 2{,}234.8228\ S_x = 1{,}632.4075,\ \bar{y} = 749.2998,\ S_y = 230.45,\ \sum_{i=1}^{10} x_i y_i = 14{,}590{,}291.7922$$

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\mathrm{x}}$$

Donde:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x^2} = \frac{14{,}590{,}291.7922 - 10(2{,}234.8228)(749.2998)}{(10-1)(1{,}632.4075)^2} = -0.0899$$

$$\hat{\beta}_0 = 749.2998 - (-0.0899)(2{,}234.8228) = 950.1337$$

Entonces:

$$\hat{y}_x = 950.1337 - 0.0899\mathrm{x}$$

b)

Usando la recta de regresión, se calculan los residuos ei.

$$S_R^2 = \frac{\sum_{i=1}^{10} e_i^2}{n-2} = \frac{284{,}283.5766}{10-2} = 35{,}535.4471$$

c)

El intervalo de confianza al 100 (1-α)% para β_1 se define de la siguiente manera:

$$\left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{S_R^2}{(n-1)S_x^2}}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{S_R^2}{(n-1)S_x^2}}\right]$$

Donde:

$$t_{1-\frac{\alpha}{2}}(n-2) = t_{0.975}(8) = 2.306$$

Reemplazando:

$$\left[-0.0899 - 2.306\sqrt{\frac{35{,}535.4471}{(10-1)(1{,}632.4075)^2}}, -0.0899 + 2.306\sqrt{\frac{35{,}535.4471}{(10-1)(1{,}632.4075)^2}}\right]$$

$$[-0.1786, -0.0011]$$

d)

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$

De la pregunta anterior:

$$-0.1786 \leq \beta_1 \leq -0.0011$$

Debido a que el intervalo no contiene a 0, se rechaza H0, por lo que la producción de melón contribuye significativamente en él.

CAPÍTULO XII

REGRESIÓN LINEAL MÚLTIPLE

12.1. INTRODUCCIÓN

El presente capítulo denominado "Regresión lineal múltiple" se presentan los siguientes temas: modelo de regresión lineal múltiple, estimación del modelo de regresión, análisis de los coeficientes de regresión y coeficiente de determinación múltiple ajustado

12.2. MODELO DE REGRESIÓN LINEAL MÚLTIPLE

El análisis de regresión lineal múltiple es una ampliación de la regresión simple para escenarios que implican dos o más variables independientes X1, X2, X3, X4 … Xk (k > 2), las cuales se relacionan con una variable dependiente "Y" formulado a través de un modelo estadístico. En otras palabras, es la explicación de una variable por medio de dos o más variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon$$

Donde:

$\beta_0, \beta_1, \beta_k$ son parámetros desconocidos. (La estimación de estos parámetros será visto más adelante)

ε es el término que define de error, es una variable aleatoria, trabajando bajo el supuesto que sigue una distribución normal con media o valor esperado $E(\varepsilon) = 0$ y varianza σ^2.

Es importante conocer, que el modelo estadístico de la regresión línea múltiple es equivalente al modelo matemático de la misma:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

Este modelo puede ser representado gráfico hasta con dos variables, es decir, cuando K toma un valor igual a 2. Sin embargo, cuando K toma valores mayores a 2, ya no es posible su representación tradicional debido a que formaría parte de un hiperplano.

Si el estudiante desea comprender la interacción entre el valor estimado E(Y) y una o dos variables, se recomienda hacer un análisis mediante un gráfico de dispersión para E(Y) y la variable X_1, X_2, o X_k que se quiera representar.

Se debe tener en claro que los coeficientes de regresión β_i de X_i indican el promedio del cambio de Y que guarda correspondencia con un incremento unitario en X_i cuando las demás variables X permanecen constantes.

En el presente material didáctico se abordar el modelo de regresión de efectos fijos, es decir, que las variables X no son aleatorias.

En este sentido, el objetivo es analizar un modelo de regresión líneal múltiple que intenta explicar el comportamiento de la variable aleatoria Y (variable de intervalo o razón), por medio de la aplicación de información proporcionada por una muestra aleatoria de tamaño n, expresada a través de las variables matemáticas, $X_1, X_2, X_3, \ldots, X_k, Y_i$ donde $i=1,2,\ldots,n$ y $n<k$.

El análisis de regresión lineal múltiple es una técnica muy útil empleada en diversas disciplinas, investigaciones y/o estudios complementarios. Con la aplicación de softwares de cómputo la labor de su cálculo se vuelve más sencilla, además que permite operar con una mayor cantidad de datos y variables independientes.

Dicho modelo estadístico en función de la muestra de variables aleatorias es denotado mediante la expresión:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \, , i = 1,2,\ldots,n$$

Los supuestos del análisis de regresión múltiple se ajustan los mismos supuestos del análisis de regresión lineal simple. Suponiendo que los residuos $\varepsilon_i = Y_i - \mu_{Y_i}$, son variables aleatorias, donde cada una de ella presenta una media igual a cero y varianza común σ^2, conocido como homocedásticidad.

Asimismo, se supone que los residuos $\varepsilon_i = Y_i - \mu_{Y_i}$, tienen distribución normal, entendido estadísticamente como normalidad.

Por otro lado debemos suponer que las variables $X_1, X_2, X_3, \ldots, X_k$ son variables independientes y cuando este supuesto no se cumple, se dice que el modelo presenta multicolinearidad.

12.3. ESTIMACIÓN DEL MODELO DE REGRESIÓN

Como primer objetivo de un estudio de regresión es estimar el modelo de regresión.

Para una población:

$$\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

Para una muestra:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k$$

Donde, $\hat{Y}$ es la estimación de μ_Y y $b_0, b_1, \ldots, b_k$ (denotados también por $\hat{\beta}_j$) son las estimaciones de los parámetros β_j con $j=1,2,\ldots,k$.

Los coeficientes de regresión muestral $b_0, b_1, \ldots, b_k$ son calculados mediante el método

de mínimos cuadrados a los datos de un muestra aleatoria de tamaño n, cuyos valores observados los expresamos por: $(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)$, $i = 1,2, \dots, n$ y $n < k$.

En este sentido, y_i es el valor de la variable dependiente Y para los valores $x_{1i}, x_{2i}, \dots, x_{ki}$ de las k variables independientes.

Para cada $i=1,2,\dots,n$ los datos de la muestra satisfacen la ecuación de regresión muestral:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k + e_i$$

Donde, las diferencias $e_i = y_i - \hat{y}_i$, son denominados errores o residuos.

El **método de mínimos cuadrados** tiene como finalidad determinar los coeficientes $b_0, b_1, \dots, b_k$ de manera que generen mínima la suma de los cuadrados de los residuos expresada mediante:

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2$$

Este requisito se cumple, según el teorema de Gauss – Markow resolviendo el sistema de las k + 1 ecuaciones normales.

$$\begin{aligned}
nb_0 + b_1 \sum x_1 + b_2 \sum x_2 + \cdots + b_k \sum x_k &= \sum y \\
b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \dots + b_k \sum x_1 x_k &= \sum x_1 y \\
b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 \dots + b_k \sum x_2 x_k &= \sum x_2 y \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots & \qquad\qquad \vdots \\
b_0 \sum x_k + b_1 \sum x_k x_1 + b_2 \sum x_k x_2 \dots + b_k \sum x_k^2 &= \sum x_k y
\end{aligned}$$

Donde, $\sum x_j = \sum_{i=1}^{n} x_{ji}$, $\sum x_j y = \sum_{i=1}^{n} x_{ji}\, y_i$, para $j = 1,2,\dots,k$.

12.4. ANÁLISIS DE LOS COEFICIENTES DE REGRESIÓN

Los coeficientes de regresión del modelo estimado de la regresión múltiple pueden ser analizados de la siguiente forma:

Los coeficientes de regresión indican la tendencia de b_i para cada X_i indican, la tendencia. Cuando b_i es positivo la tendencia es creciente, es decir, cuando se incrementan los

valores de X_i se incrementan los valores de Y. La constante b_0 es la ordenada en el origen.

12.5 COEFICIENTE DE DETERMINACIÓN MÚLTIPLE AJUSTADO

El coeficiente de determinación R^2 crece a medida que el número de variables independientes del modelo de regresión se incrementa. Para evitar este sesgo se aplica el coeficiente ajustado.

El coeficiente de determinación múltiple ajustado o corregido de la estimación del modelo de regresión es:

$$R_A^2 = 1 - \frac{MCE}{MCT} = 1 - \frac{SCE/(n-k-1)}{SCT/(n-1)}$$

Este coeficiente nos permite comparar descriptivamente dos o más modelos de regresión con diferentes números de variables independientes.

12.6 PREGUNTAS Y RESPUESTAS DE REPASO

1. Un científico busca demostrar que el rendimiento de una reacción química (Y en %) depende de la concentración de su reactivo (x) y de la temperatura de la operación (z en °C). Para estudiar ello, se registraron los siguientes datos:

Y	x	z
81	0.90	133
85	0.85	146
88	0.80	159
90	0.75	172
91	0.70	180
90	0.65	148
92	0.60	151
93	0.55	172
94	0.50	177
95	0.45	175

   a. Ajuste los datos a un modelo de regresión múltiple y analice la correlación.

   b. Realice el contraste de significación del modelo. Use = 0.05. ¿Se podría decir que este modelo es mejor que uno de regresión lineal simple, tomando solo en cuenta la temperatura?

   c. ¿En cuánto estimaría el rendimiento medio de una reacción química en la cual se utiliza una concentración de 0,69 a una temperatura de 150 °C?

Solución

a) En cualquier modelo de regresión lineal múltiple con dos variables independientes, su matriz y el vector columna de la variable dependiente tiene la siguiente forma:

$$X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \dots & \dots & \dots \\ 1 & x_n & z_n \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}$$

Luego, el sistema $(X^tX)B = X^tY$ produce el siguiente sistema de ecuaciones normales:

$$n\beta_0 + \sum_{i=1}^{n} x_i\beta_1 + \sum_{i=1}^{n} z_i\beta_2 = \sum_{i=1}^{n} Y_i$$

$$\sum_{i=1}^{n} x_i\,\beta_0 + \sum_{i=1}^{n} x_i^2\beta_1 + \sum_{i=1}^{n} x_i z_i\beta_2 = \sum_{i=1}^{n} x_i Y_i$$

$$\sum_{i=1}^{n} z_i\,\beta_0 + \sum_{i=1}^{n} x_i z_i\beta_1 + \sum_{i=1}^{n} z_i^2\beta_2 = \sum_{i=1}^{n} z_i Y_i$$

En este caso: n = 10.

$$\sum_{i=1}^{10} x_i = 6.75$$

$$\sum_{i=1}^{10} x_i^2 = 4.7625$$

$$\sum_{i=1}^{10} z_i = 1{,}613$$

$$\sum_{i=1}^{10} z_i^2 = 262{,}513$$

$$\sum_{i=1}^{10} Y_i = 899$$

$$\sum_{i=1}^{10} Y_i^2 = 80{,}985$$

Reemplazando:

$$10\beta_0 + 6.75\beta_1 + 1{,}613\beta_2 = 899$$

$$6.75\beta_0 + 4.7625\beta_1 + 1{,}074.65\beta_2 = 601.35$$

$$1{,}613\beta_0 + 1{,}074.65\beta_1 + 262{,}513\beta_2 = 145{,}506$$

La solución de este sistema nos provee de las estimaciones de mínimos cuadrados y el plano de regresión:

$$\hat{\beta}_0 = 89.2684$$

$$\hat{\beta}_1 = -20.4240$$

$$\hat{\beta}_2 = 0.0894$$

$$\hat{y}_{(x,z)} = 89.2684 - 20.4240\mathrm{x} + 0.0894\mathrm{z}$$

Para medir el grado de ajuste de los datos a un modelo lineal, se halla el coeficiente de determinación y el error estándar de estimación.

$$SCT = (n-1)S_Y^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = 164.9$$

$$SCR = \hat{\beta}_1\left(\sum_{i=1}^{n} x_i Y_i - 10\bar{x}\bar{Y}\right) + \hat{\beta}_2\left(\sum_{i=1}^{n} z_i Y_i - 10\bar{z}\bar{Y}\right) = 156.2722$$

$$R^2 = \frac{SCR}{SCT} = 0.9477$$

Debido a que el R2 tiende a 1, el ajuste es muy bueno.

$$S_e = \sqrt{\frac{SCT - SCR}{n-k-1}} = \sqrt{\frac{8.6278}{10-2-1}} = 1.1102$$

Mientras más pequeño sea S_e, mejor ajuste tendrán los datos al modelo, por lo que indicaría un muy buen ajuste lineal.

b) Para la significación del modelo, se contrasta a nivel $\alpha = 0.05$.

$$H_0: \beta_1 = \beta_2 = 0, H_1: \beta_j \neq 0 \; \exists \; j \in \{1,2\}$$

La tabla ANOVA es la siguiente:

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Medias cuadráticas	F_0
Regresión	156.2722	2	78.1361	63.3940
Error	8.6278	7	1.2325	
Total	164.9	9		

Como $F_0 = 63.3940 > F_{0.95}(2,7) = 4.74$, se rechaza H_0 y el modelo lineal dado es aceptable.

Para ver si el modelo de regresión múltiple es mejor que el lineal indicado, se debe probar que la concentración contribuye significativamente en la estimación del valor medio de Y. Es decir, contrastar a nivel $\alpha = 0.05$.

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$

Se rechaza H_0 si:

$$\left| T_{0j} = \frac{\hat{\beta}_j}{S_e \sqrt{c_{j+1,j+1}}} \right| > t_{1-\alpha/2}(n-k-1)$$

Donde: $j = 1$ y $c_{j+1,j+1}$ es la entrada j+1, j+1 de la matriz $(X^tX)^{-1}$. Para determinar esta entrada, se define X:

$$X = \begin{bmatrix} 1 & 0.90 & 133 \\ 1 & 0.85 & 146 \\ 1 & 0.80 & 159 \\ 1 & 0.75 & 172 \\ 1 & 0.70 & 180 \\ 1 & 0.65 & 148 \\ 1 & 0.60 & 151 \\ 1 & 0.55 & 172 \\ 1 & 0.50 & 177 \\ 1 & 0.45 & 175 \end{bmatrix}$$

$$(X^t X) = \begin{bmatrix} 10 & 6.75 & 1{,}613 \\ 6.75 & 4.7625 & 1{,}074.65 \\ 1{,}613 & 1{,}074.65 & 262{,}513 \end{bmatrix}$$

$$(X^t X)^{-1} = \begin{bmatrix} 33.7739 & -13.6563 & -0.1516 \\ -13.6563 & 8.2751 & 0.0500 \\ -0.1516 & 0.0500 & 0.0007 \end{bmatrix}$$

Reemplazando:

$$\left| T_{01} = \frac{-20.4240}{1.1102\sqrt{8.2751}} = -6.3952 \right| > t_{0.975}(7) = 2.365$$

Por lo tanto, se rechaza H_0. Es decir, la concentración sí contribuye con información significativa en la predicción de Y, por lo que es mejor un modelo de regresión lineal múltiple que el simple propuesto.

c) Como el modelo múltiple es mejor, se tiene:

$$\hat{y}_{(0.69,150)} = 89.2684 - 20.4240(0.69) + 0.0894(150) = 88.5836$$

2. Tomando en cuenta la última Encuesta Nacional de Hogares, el INEI realizó un estudio sobre la posible relación entre las siguientes variables:

Y: Gastos mensuales (en miles de soles).

X_1: Ingreso mensual del hogar (en miles de soles).

X_2: Tamaño de la familia.

En una muestra de 15 hogares elegidos al azar, se obtuvieron los siguientes datos:

Y	X1	X2
7.5	8.5	8
7.2	8.4	7
7.0	8.0	7
6.5	7.8	6
6.2	7.3	6
6.3	7.5	6
5.5	7.0	5
5.4	6.5	5
5.2	6.0	5
4.8	5.8	4
4.7	5.5	4
4.5	5.4	4
4.2	4.5	4
3.4	3.5	3
3.0	3.0	3

a. Establezca el sistema de ecuaciones normales de la regresión de mínimos cuadrados.

b. Halle la ecuación de regresión múltiple muestral (estimada).

c. Estime el gasto mensual para un hogar compuesto de 7 personas, cuyo ingreso mensual es de S/ 6,800.

d. Interprete los coeficientes de regresión.

e. ¿Cuál de las 2 variables independientes contribuye más a la predicción de los gastos mensuales?

Solución

a)

$$n\beta_0 + \sum_{i=1}^{n} X_{1i}\beta_1 + \sum_{i=1}^{n} X_{2i}\beta_2 = \sum_{i=1}^{n} Y_i$$

$$\sum_{i=1}^{n} X_{1i}\,\beta_0 + \sum_{i=1}^{n} X_{1i}^2\beta_1 + \sum_{i=1}^{n} X_{1i}X_{2i}\beta_2 = \sum_{i=1}^{n} X_{1i}Y_i$$

$$\sum_{i=1}^{n} X_{2i}\,\beta_0 + \sum_{i=1}^{n} X_{1i}X_{2i}\beta_1 + \sum_{i=1}^{n} X_{2i}^2\beta_2 = \sum_{i=1}^{n} X_{2i}Y_i$$

En este caso: n = 15.

$$\sum_{i=1}^{15} X_{1i} = 94.7$$

$$\sum_{i=1}^{15} X_{1i}^2 = 638.99$$

$$\sum_{i=1}^{15} X_{2i} = 77$$

$$\sum_{i=1}^{15} X_{2i}^2 = 427$$

$$\sum_{i=1}^{15} X_{1i}X_{2i} = 520.2$$

$$\sum_{i=1}^{15} Y_i = 81.4$$

$$\sum_{i=1}^{15} Y_i^2 = 467.5$$

Reemplazando:

$$15\beta_0 + 94.7\beta_1 + 77\beta_2 = 81.4$$
$$94.7\beta_0 + 638.99\beta_1 + 520.2\beta_2 = 546.03$$
$$77\beta_0 + 520.2\beta_1 + 427\beta_2 = 445.9$$

b)

$$\hat{\beta}_0 = 0.5342$$
$$\hat{\beta}_1 = 0.4435$$
$$\hat{\beta}_2 = 0.4076$$
$$\hat{y}_{(X_1,X_2)} = 0.5342 + 0.4435X_1 + 0.4076X_2$$

c)

$$\hat{y}_{(6.8,7)} = 0.5342 + 0.4435(6.8) + 0.4076(7) = 6.4033 \text{ miles de soles}$$

d)

$\hat{\beta}_0$=0.5342. Es el monto de gastos mensuales, en caso $X_1 = X_2 = 0$.

$\hat{\beta}_1$=0.4435. Significa que, por cada aumento de 1 unidad del ingreso mensual familiar (X_1), el gasto mensual total aumenta en 0.4435 miles de soles, manteniendo X_2 constante.

$\hat{\beta}_2$=0.4076. Significa que, por cada aumento de 1 unidad del tamaño de la familia (X_2), el gasto mensual total aumenta en 0.4076 miles de soles, manteniendo X_1 constante.

e)

Dado que las variables no tienen las mismas unidades, se utiliza el coeficiente β, a fin de estandarizar.

$$\beta_1 = \frac{S_{X_1}\hat{\beta}_1}{S_Y}$$
$$\beta_2 = \frac{S_{X_2}\hat{\beta}_2}{S_Y}$$
$$S_{X_1}^2 = \frac{\sum_{i=1}^{15} X_{1i}^2 - 15(\bar{X}_1)^2}{15-1} = \frac{638.99 - 15(6.3133)^2}{14} = 2.9370$$

$$S_{X_1} = 1.7138$$

$$S_{X_2}^2 = \frac{\sum_{i=1}^{15} X_{2i}^2 - 15(\bar{X}_2)^2}{15-1} = \frac{427 - 15(5.1333)^2}{14} = 2.2667$$

$$S_{X_2} = 1.5055$$

$$S_Y^2 = \frac{\sum_{i=1}^{15} Y^2 - 15(\bar{Y})^2}{15-1} = \frac{467.5 - 15(5.4267)^2}{14} = 1.8407$$

$$S_Y = 1.3567$$

Entonces:

$$\beta_1 = \frac{(1.7138)(0.4435)}{1.3567} = 0.5603$$

$$\beta_2 = \frac{(1.5055)(0.4076)}{1.3567} = 0.4523$$

Debido a que β_1 es mayor a β_2, se puede decir que un cambio en la variable X_1 es más significativo que en X_2, es decir, los ingresos mensuales del hogar contribuyen más a la predicción de los gastos mensuales que el tamaño de la familia.

3. El jefe de ventas de una empresa que se dedica al comercio de café tostado a través de una cadena nacional de tiendas por departamento está interesado en estudiar la relación que tienen el precio de su producto y la publicidad con las ventas. Para ello, registró las ventas anuales Y (en miles de soles) que su empresa generó a diferentes precios (X_1 en soles) y proporciones X_2 de gastos en publicidad en cada una de las 25 regiones del país, respecto a lo gastado el año pasado. Al realizar un análisis de regresión múltiple, obtuvo la siguiente información:

Variable	Media	Desviación estándar
y	54.2389	8.5622
x_1	9.6253	1.2579
x_2	13.4795	2.3568

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Medias cuadráticas	F_0
Regresión	852.9253	2	426.4627	48.5473
Error	193.2584	22	8.7845	
Total	1,046.1837	24		

Los elementos de entradas 2,2 y 3,3 de $(X^tX)^{-1}$, siendo X la matriz de variables independientes, fueron c22 = 25.6472 y c33 = 0.8563, y el plano de mínimos cuadrados ajustado fue el siguiente:

$$\hat{y}_{x_1,x_2} = 25.2893 - 42.5632x_1 + 6.2389x_2$$

a. A un nivel de significación $\alpha = 0.05$, ¿qué le dice la tabla ANOVA anterior?

b. Halle el coeficiente de determinación R2 e indique si es que el ajuste de los datos al modelo lineal hallado es bueno.

c. ¿Contribuyen, a un nivel de significación $\alpha = 0.05$, las dos variables independientes con información significativa en la estimación de las ventas anuales? ¿Cuál de las dos variables da mayor contribución y por qué?

Solución

a)

$$H_0: \beta_1 = \beta_2 = 0, H_1: \beta_j \neq 0 \; \exists \; j \in \{1,2\}$$

Se rechaza H_0 si:

$$F_0 > F_{1-\alpha}(k, n-k-1) = F_{0.95}(2,22) = 3.44$$

Por lo tanto, se rechaza H_0 y el modelo es válido.

b)

$$R^2 = \frac{SCR}{SCT} = \frac{1{,}042.917}{1{,}520.743} = 0.8153$$

Debido a que el R2 tiende a 1, el ajuste es relativamente bueno.

c)

$$S_e = \sqrt{\frac{SCT - SCR}{n - k - 1}} = \sqrt{\frac{193.2584}{24 - 2 - 1}} = 2.9639$$

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$

$$\hat{\beta}_1 = -42.5632$$

Se rechaza H0 si:

$$\left| T_{0j} = \frac{\hat{\beta}_j}{S_e \sqrt{c_{j+1,j+1}}} \right| > t_{1-\alpha/2}(n - k - 1)$$

Donde: j = 1 y $c_{j+1,j+1}$ es la entrada j+1, j+1 de la matriz $(X^tX)^{-1}$.

$$\left| T_{01} = \frac{-42.5632}{2.9639\sqrt{25.6472}} = -2.8357 \right| > t_{0.975}(22) = 2.074$$

Por lo tanto, se rechaza H0 y la variable X1 es significativa.

$$H_0: \beta_2 = 0, H_1: \beta_2 \neq 0$$

$$\hat{\beta}_2 = 6.2389$$

Se rechaza H0 si:

$$\left| T_{0j} = \frac{\hat{\beta}_j}{S_e \sqrt{c_{j+1,j+1}}} \right| > t_{1-\alpha/2}(n - k - 1)$$

Donde: j = 2 y $c_{j+1,j+1}$ es la entrada j+1, j+1 de la matriz $(X^tX)^{-1}$.

$$\left| T_{02} = \frac{6.2389}{2.9639\sqrt{0.8563}} = 2.2748 \right| > t_{0.975}(22) = 2.074$$

Por lo tanto, se rechaza H_0 y la variable X_2 es significativa.

Entonces, ambas variables son significativas en la estimación de las ventas anuales. Para determinar cuál de ambas da mayor contribución, se utiliza el coeficiente β, dado que éstas no tienen las mismas unidades.

$$\beta_1 = \frac{S_{X_1}\hat{\beta}_1}{S_Y} = \frac{(1.2579)(-42.5632)}{8.5622} = -6.2531$$

$$\beta_2 = \frac{S_{X_2}\hat{\beta}_2}{S_Y} = \frac{(2.3568)(6.2389)}{8.5622} = 1.7173$$

Debido a que β_2 es mayor a β_1, se puede decir que un cambio en la variable X_2 es más significativo que en X_1, es decir, los gastos en publicidad contribuyen más a las ventas anuales que los precios.

4. Una empresa de consumo masivo decidió evaluar la productividad de los 30 operarios que trabajan en planta, tomando en cuenta las siguientes variables:

   Y: Productividad en el trabajo (calificado de 0 a 20).

   X_1: Horas semanales de trabajo.

   X_2: Cantidad de productos elaborados por semana.

   X_3: Años de experiencia.

   Con los datos de la muestra, se obtuvo los siguientes resultados:

$$\hat{y} = 1.5693 + 2.3645X_1 + 1.1397X_2 + 0.5683X_3$$

   $R^2 = 0.89$, $S_Y = 4.2531$.

   Entradas c_{11}, c_{22}, c_{33} de la matriz $(X^tX)^{-1}$: 0.4269, 0.0852, 0.0514, respectivamente.

   a. ¿Considera, a nivel de significación $\alpha = 0.05$, que el modelo estimado es válido para predecir Y en términos de X_1, X_2, y X_3?

   b. ¿Cuál o cuáles de las variables independientes no contribuyen a la predicción del comportamiento hacia el trabajo? Use prueba de hipótesis con $\alpha = 0.05$.

   c. Escriba el modelo (solo la forma) que resulta de eliminar las variables que no contribuyen al modelo general. Si con este modelo se obtuvo SCE = 60.2563, ¿considera que el modelo resultante se ajusta adecuadamente a los datos de la muestra? ¿Por qué?

Solución

a)

Se sabe: n = 30, k = 3.

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0, H_1: \beta_j \neq 0 \; \exists \; j \in \{1,2,3\}$$

$$SCT = (n-1)S_Y^2 = 29(4.2531)^2 = 524.5769$$

$$R^2 = \frac{SCR}{SCT} \Rightarrow 0.89 = \frac{SCR}{524.5769}$$

$$SCR = 466.8735$$

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Medias cuadráticas	F0
Regresión	466.8735	3	155.6245	70.1212
Error	57.7035	26	2.2194	
Total	524.5769	29		

Se rechaza H0 si:

$$F_0 > F_{1-\alpha}(k, n-k-1) = F_{0.95}(3,26) = 2.98$$

Por lo tanto, se rechaza H0 y el modelo es válido.

b)

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$

$$\hat{\beta}_1 = 2.3645$$

Se rechaza H0 si:

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{ES} > t_{1-\alpha/2}(n-k-1)$$

$$ES = \sqrt{MCE(c_{11})} = \sqrt{2.2194(0.4269)} = 0.9734$$

$$t_0 = \frac{2.3645 - 0}{0.9734} = 2.4292 > t_{0.975}(26) = 2.0555$$

Por lo tanto, se rechaza H0 y la variable X1 es significativa.

$$H_0: \beta_2 = 0, H_1: \beta_2 \neq 0$$

$$\hat{\beta}_2 = 1.1397$$

Se rechaza H0 si:

$$t_0 = \frac{\hat{\beta}_2 - \beta_2}{ES} > t_{1-\alpha/2}(n-k-1)$$

$$ES = \sqrt{MCE(c_{22})} = \sqrt{2.2194(0.0852)} = 0.4348$$

$$t_0 = \frac{1.1397 - 0}{0.4348} = 2.6209 > t_{0.975}(26) = 2.0555$$

Por lo tanto, se rechaza H0 y la variable X2 es significativa.

$$H_0: \beta_3 = 0, H_1: \beta_3 \neq 0$$

$$\hat{\beta}_3 = 0.5683$$

Se rechaza H0 si:

$$t_0 = \frac{\hat{\beta}_3 - \beta_3}{ES} > t_{1-\alpha/2}(n-k-1)$$

$$ES = \sqrt{MCE(c_{33})} = \sqrt{2.2194(0.0514)} = 0.3378$$

$$t_0 = \frac{0.5683 - 0}{0.3378} = 1.6826 > t_{0.975}(26) = 2.0555$$

Por lo tanto, se acepta H0 y la variable X3 no es significativa.

c)

Quitando la variable X3, el modelo sería el siguiente:

$$\hat{y} = 2.3645X_1 + 1.1397X_2$$

Para comparar de manera descriptiva dos o más modelos con desigual número de variables independientes, se utiliza el coeficiente de determinación ajustado.

$$R_A^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1} = 1 - \frac{(1-0.89)(30-1)}{30-3-1} = 0.8773$$

Debido a que el R^2_A tiende a 1, se concluye que el modelo resultante se ajusta a los datos de la muestra.

5. Una empresa de comercio electrónico está estudiando el sistema de reparto de sus productos. Su objetivo es predecir el tiempo de entrega Y (en horas). El responsable del estudio concluyó que los dos factores más importantes que determinan este tiempo son el número de unidades de producto que se entregan (X1), así como la máxima distancia que debe recorrer el distribuidor (X2 en km.). Además, presentó la siguiente información para 25 repartos elegidos al azar:

$$\hat{y} = 3.5647 + 1.9523X_1 + 1.4752X_2$$

$$\sum_{i=1}^{n} Y_i = 538.4698$$

$$\sum_{i=1}^{n} Y_i^2 = 21{,}356.8321$$

$$MCE = 32.3426$$

La matriz inversa $(X^tX)^{-1}$, donde X es la matriz de variables independientes, es la siguiente:

$$(X^tX)^{-1} = \begin{bmatrix} 3.4358 & -0.1065 & -0.3285 \\ -0.1065 & 0.0256 & 0.0268 \\ -0.3285 & 0.0268 & 0.0124 \end{bmatrix}$$

a. ¿Coincide su análisis con el del especialista? Realice primero una prueba de hipótesis global y luego individual de los coeficientes de regresión, al nivel de significación $\alpha = 0.05$.

b. b) Estime, en un intervalo de confianza de 95%, el tiempo medio de servicio que se requerirá para satisfacer un pedido de 20 unidades de producto que se ubica a 42 km. de distancia.

Solución

a)

Se sabe: n = 25, k = 2.

$$H_0: \beta_1 = \beta_2 = 0, H_1: \beta_j \neq 0 \; \exists \; j \in \{1,2\}$$

$$SCT = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} = 21{,}356.8321 - \frac{(538.4698)^2}{25}$$

$$SCT = 9{,}758.8431$$

$$SCE = MCE(n-k-1) = 32.3426(22) = 711.5372$$

$$SCR = SCT - SCE = 9{,}758.8431 - 711.5372 = 9{,}047.3059$$

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Medias cuadráticas	F0
Regresión	9,047.3059	2	4,523.6529	139.8667
Error	711.5372	22	32.3426	
Total	9,758.8431	24		

Se rechaza H0 si:

$$F_0 > F_{1-\alpha}(k, n-k-1) = F_{0.95}(2,22) = 3.44$$

Por lo tanto, se rechaza H0 y el modelo es válido.

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$

$$\hat{\beta}_1 = 1.9523$$

Se rechaza H0 si:

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{ES} > t_{1-\alpha/2}(n-k-1)$$

$$ES = \sqrt{MCE(c_{11})} = \sqrt{32.3426(0.0256)} = 0.9099$$

$$t_0 = \frac{1.9523 - 0}{0.9099} = 2.1456 > t_{0.975}(22) = 2.0739$$

Por lo tanto, se rechaza H_0 y la variable X_1 es significativa.

$$H_0: \beta_2 = 0, H_1: \beta_2 \neq 0$$
$$\hat{\beta}_2 = 1.4752$$

Se rechaza H_0 si:

$$t_0 = \frac{\hat{\beta}_2 - \beta_2}{ES} > t_{1-\alpha/2}(n-k-1)$$
$$ES = \sqrt{MCE(c_{22})} = \sqrt{32.3426(0.0124)} = 0.6333$$
$$t_0 = \frac{1.4752 - 0}{0.6333} = 2.3294 > t_{0.975}(22) = 2.0739$$

Por lo tanto, se rechaza H_0 y la variable X_2 es significativa.

Finalmente, luego de aplicar las pruebas de hipótesis global e individual, se está de acuerdo con el responsable, debido a que el modelo planteado es válido.

b)

El intervalo de confianza al 100 (1-α)% para el valor medio de Y, dado $\vec{x} = (x_1, x_2, \ldots, x_k)^t$ se define de la siguiente manera:

$$\left[\hat{y}_{\vec{x}} - t_{1-\frac{\alpha}{2}}(n-k-1)S_e\sqrt{a^t(X^tX)^{-1}a}, \hat{y}_{\vec{x}} + t_{1-\frac{\alpha}{2}}(n-k-1)S_e\sqrt{a^t(X^tX)^{-1}a}\right]$$

Donde Se es el error estándar de estimación y a representa el vector columna $a = (1, x_1, x_2, \ldots, x_k)^t$

Se sabe:

$$\vec{x} = (20{,}42)^t$$
$$a = (1{,}20{,}42)^t$$

Entonces:

$$\hat{y}_{\vec{x}} = 3.5647 + 1.9523X_1 + 1.4752X_2$$

$$\hat{y}_{\vec{x}} = 3.5647 + 1.9523(20) + 1.4752(42) = 104.5691$$

$$t_{1-\alpha/2}(n-k-1) = t_{0.975}(22) = 2.0739$$

$$S_e = \sqrt{\frac{SCT - SCR}{n-k-1}} == \sqrt{\frac{711.5372}{22}} = 5.6871$$

$$\sqrt{a^t(X^tX)^{-1}a} = \sqrt{\begin{bmatrix}1 & 20 & 42\end{bmatrix}\begin{bmatrix}3.4358 & -0.1065 & -0.3285\\ -0.1065 & 0.0256 & 0.0268\\ -0.3285 & 0.0268 & 0.0124\end{bmatrix}\begin{bmatrix}1\\20\\42\end{bmatrix}}$$

$$\sqrt{a^t(X^tX)^{-1}a} = \sqrt{\begin{bmatrix}-12.4912 & 1.5311 & 0.7283\end{bmatrix}\begin{bmatrix}1\\20\\42\end{bmatrix}} = \sqrt{48.7194} = 6.9799$$

Reemplazando:

$$[104.5691 - 2.0739(5.6871)(6.9799), 104.5691 + 2.0739(5.6871)(6.9799)]$$

$$[22.2451, 186.8931]$$

REFERENCIAS BIBLIOGRÁFICAS

Ander-Egg, E. (1987). *Técnicas de investigación social*. Editorial Magisterio del Río de la Plata.

Anduiza, E. et al. (2000). *Metodología de la Ciencia Política*, Fondo editorial de la Universidad de Lima.

Blanco, M. y Villapando, P. (2012). "Nociones científicas del Protocolo de Investigación", *Metodología para investigaciones de alto impacto en las ciencias sociales,* Universidad Autónoma De Nuevo León.

Bolívar, R. (2001). "La jurídica como ciencia". *Revista de Estudios Políticos,* núm. 28, sexta época, septiembre-diciembre. Universidad Nacional Autónoma de México.

Buendía, et al. L. (1998). *Métodos de investigación en psicopedagogía*. McGraw-Hill.

Bunge, M. (1972). *La ciencia, su método y su filosofía*. Ediciones Siglo Veinte.

Cazau, P. (2006). Introducción a la investigación en ciencias sociales. Amauta 21.

Centro Universitario Interamericano. (2019). Investigación explicativa.

http://metodologiainter.weebly.com/uploads/1/9/2/6/19268119/investigacin_correlacional.pdf.

Cervantes, F. (2016). *Estadística Descriptiva y Probabilidad*. Facultad de Estudios Superiores Cuautitlán. Universidad Nacional Autónoma de México.

Córdova, M. (2009). *Estadística descriptiva e inferencial,* Librería Moshera.

Chulla, E. y Agulló, M. (2012). Como se hace un trabajo de investigación jurídica, Catarata.

Gómez, F. (2013). "Qué es la Ciencia Política", *Criterio Libre Jurídico*. Volumen 10 núm. 1.

Del Val Cid, C. (2007)."El muestreo, métodos y aplicaciones", *La investigación social del turismo, Perspectivas y Aplicaciones*. Jesús Gutiérrez Brito Coordinador. Paraninfo.

Gestipolis. (2019).Tipos de estudio y métodos de investigación

https://nodo.ugto.mx/wp-content/uploads/2016/05/Tipos-de-estudio-y-metodos-de-investigaci%C3%B3n.pdf.

Hernández, R. et.al. (2014). *Metodología de la investigación*. Mac Graw Hill.

Mac Donald, A. (1972). *Elementos para un análisis cuantitativo en sociología*. Pontificia Universidad Católica del Perú.

Martínez, C. (2016). *Estadística y Muestreo*. Ecoe ediciones.

Mills, Ch. W. (1959). *La imaginación sociológica*. Fondo de Cultura Económica.

Mitac, M. (2014). *Tópicos de estadística descriptiva y probabilidad*, San Marcos.

Mode, E. (1990). Elementos de probabilidad y estadística. Editorial Reverte.

Murray, S. y Larry, S. (2006). Estadística. McGraw – Hill, México.

Question Pro. ¿Qué es el muestreo por conveniencia? *https://www.questionpro.com/blog/es/muestreoporconveniencia/muestreopor20conveniencia20es,pr20un*

Quispe, U. (2010). *Fundamentos de estadística básica*, Editorial San Marcos.

Reynaga, J. El método estadístico, en *http://www.cobatab.edu.mx/descargasgrales/academico2011*.

Ritchey, F.J. (2008).*The statistical imagination statics for the social sciences*, The Mac Graw-Hill, editions.

Rojas, M. (2015). "Tipos de Investigación científica: Una simplificación de la complicada incoherente nomenclatura y clasificación". *Revista Electrónica de Veterinaria,* vol. 16, núm. 1.

Sánchez, L. (2018). Investigación correlacional e investigación explicativa.

https://luiserveychavez.files.wordpress.com/2008/11/investigacion-correlacional.

Sierra, R. (2009). Técnicas de investigación social. Teoría y ejercicios, Paraninfo.

Strauss, A. y Corbin, J. (2002). Bases de la investigación cualitativa. Técnicas y procedimientos para desarrollar la teoría fundamentada. Editorial Universidad de Antioquía.

Unesco. (2011). Métodos de investigación cuantitativa para el planeamiento de la educación. Instituto Internacional de Planeamiento de la Educación.

Universidad de Jaén. (2018). *Estudios correlaciónales*.

www4.ujaen.es/eramirez/Descargas/tema5. .

Universidad Nacional Autónoma de México, Métodos de investigación.

http://www.psicol.unam.mx/Investigacion2/pdf/metodos.pdf.

Vallejo, M. (2002). "El diseño de investigación: una breve revisión metodológica". *Revista mexicana de cardiología,* México. núm. 1.

Veliz, C. (2000). *Estadística*. *Aplicaciones*. Editorial San Marcos.

Francisco Carruitero Lecca

Abogado, Magister y Doctor en Derecho (Revalidada) por la Pontificia Universidad Católica del Perú, Máster en Teoría de las Organizaciones por la Universidad de Burdeos Francia. Doctor en Derecho por la Universidad de Castilla La Mancha España. Profesor invitado en el Doctorado en Derecho en la Universidad Privada Antenor Orrego y Profesor Ordinario de la Universidad Nacional Mayor de San Marcos.

Tula Benites Vásquez

Abogada por la Universidad Nacional de Trujillo. Doctora en Derecho Constitucional por la Universidad Privada Antenor Orrego, Especialista en Justicia Constitucional y Tutela Jurisdiccional, Diplomada en Perfeccionamiento en Alta Formación en Justicia Constitucional por la Universidad de Pisa Italia, Graduada del Programa de Gobernabilidad y Gerencia Política por George Washington University –PUCP. Docente Ordinario de la Universidad Privada Antenor Orrego.

Angel Hospinal Alvarez

Ingeniero Industrial por la Pontificia Universidad Católica del Perú. Estudios de Maestría en Gestión Económica Empresarial Integrada por la Universidad Mayor de San Marcos, Especialista en Mejora Estadística de Procesos, Lean Six Sigma y Sistemas de Gestión. Consultor en entidades públicas y privadas.

UPAO | FONDO EDITORIAL