ESTADÍSTICA
PARA INGENIERÍA

Ciencia Latina Internacional
CID editorial
ALAC
REDIB
Crossref
Redilat
Red de Investigadores Latinoamericanos

Marzo 2022 – CID - Centro de Investigación y Desarrollo
Copyright © CID - Centro de Investigación y Desarrollo
Copyright del texto © 2022 de Autores
libros.ciencialatina.org
editorial@ciencialatina.org
Atención por WhatsApp al +52 22 2690 3834

Datos Técnicos de Publicación Internacional
Título: ESTADISTICA PARA INGENIERIA. Autor: Arreguín Samano, Moisés Co-Autores: Carrillo Espinosa, Guillermo; Hernández Allauca, Andrea Damaris y; Sampayo Maldonado, Salvador Editor: CID - Centro de Investigación y Desarrollo Diseño de tapa: CID - Centro de Investigación y Desarrollo Corrección de Estilo: CID - Centro de Investigación y Desarrollo Formato: PDF Páginas: 231 p. Tamaño: Sobre C5 162 x 229 mm Requisitos de sistema: Adobe Acrobat Reader Modo de acceso: World Wide Web Incluir: Bibliografía ISBN: 978-99925-13-05-7 DOI: https://doi.org/10.37811/cli_w743

1ª. Edición. Año 2022. Editorial CID - Centro de Investigación y Desarrollo.

El contenido del libro y sus datos en su forma, corrección y fiabilidad son responsabilidad exclusiva de los autores. Permite la descarga de la obra y compartir siempre que los créditos se atribuyan a los autores, pero sin la posibilidad de cambiarlo de cualquier forma o utilizarlo con fines comerciales.

Prohibida su reproducción por cualquier medio.
Distribución gratuita.

Título: ESTADÍSTICA PARA INGENIERÍA

Autor:

Arreguín Samano, Moisés

Ph. D. Profesor UEB. Provincia Bolívar, Ecuador.

Co-Autores:

Carrillo Espinosa, Guillermo.

Maestro de Ciencias y Profesor de División de Ciencias Forestales (DICIFO), Universidad Autónoma Chapingo (UACh), estado de México, México.

Hernández Allauca, Andrea Damaris.

Ph. D. Profesora ESPOCH, Provincia Chimborazo, Ecuador. (C. Ph. D. en Matemática)

Sampayo Maldonado, Salvador

Doctor en Ciencias Forestales e Investigador de UNAM, Campus Iztacala, estado de México.

Estudiantes colaboradores del libro (UEB):

Número	Apellidos	Nombre(s)	Cédula de identidad
1	Álava Sánchez	Andrés Abdón	1205962127
2	Bayas Chimborazo	Jessenia Selena	1501092991
3	Carvajal Culqui	Johana Estefania	O202206355
4	Del Pozo Guaquipana	Edgar Vinicio	O202298915
5	Granizo Tuaza	Bryan Patricio	O250147873
6	Guerrero Zurita	Jheyson Mesias	O202303319
7	Maliza Villa	Tupac Curicamac	1850323591
8	Manobanda Baño	Mesias Bladimir	O250186228
9	Ochoa Rea	Mirian Liliana	1726928250
10	Ortiz Chacha	Silvia Mercedes	O202232591
11	Poaquiza Caiza	Klever Goevanny	1600698490
12	Quinatoa Quinatoa	Ángel Gustavo	O202380994

Resumen:

La estadística es el arte y la ciencia que reúne datos, analiza, presenta e interpreta mediante el uso de modelos matemáticos. Tiene uso en múltiples disciplinas científicas, como ingeniería forestal, agronomía, agro negocios, economía, procesamiento de alimentos, ciencias de la salud, marketing, elecciones políticas, estudios de mercado, control de calidad, inventarios, padrón de cobro de impuestos, muestreo, decisiones empresariales, operaciones logísticas de transporte, mercantiles, servicios y productos, cuánto producir con base en variables estocásticas-no estocásticas, dónde producir, a qué precio ofertar, a qué precio comprar materia prima, desde pequeños negocios a un gran número de compañías, organizaciones y corporaciones multinacionales para que las decisiones sean significativamente mejores respecto a tomadas con base en su propio criterio. Para un investigador que recién inicia, el procesamiento de recuento, enumeración, conteo o cuantificación de grandes volúmenes productivos, cosechas agrícolas, listas de productos u otros grandes conjuntos de datos están íntimamente relacionados, mediante técnicas estadísticas, con actividades soporíferas de organización de bases, tabulación de datos, presentación gráfica, estimación de cantidades "representativas", procesos inductivos, procesos deductivos, uso indiscriminado de software, posible pago de licencias, manejo complejo, material bibliográfico, procesamiento automático de datos, criterios interpretativos incomprensibles o transcripción mecánica de resultados ([1]). Con base en esto, Estadística para Ingeniería con Ofimática presenta terminología básica estadística, demostraciones matemáticas superiores, probabilidad con cálculo en espacios muestrales, análisis combinatorio, distribución de probabilidades, función de probabilidades con cálculo de una variable aleatoria discreta, continua, espacios muestrales, esperanza matemática, representación gráfica probabilística, distribución muestral, variables, medidas de tendencia central, medidas de dispersión, variabilidad, muestreo con tipos de distribuciones, muestreo probabilístico, muestreo no probabilístico, hipótesis con pruebas paramétricas y pruebas no paramétricas. Con el uso de información del Censo Nacional Agropecuario (CNA, 2000),

[1] Fuente: Capa, S., H. 2015

Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC, 2002 a 2018) del Instituto Nacional de Estadística y Censos (INEC) del Ecuador a nivel Nacional, Regional, Provincial y Cantonal de cultivos permanentes (banano, cacao, café, caña de azúcar para biocombustible, caña de azúcar tallo fresco, naranja, palma africana, plátano y tomate de árbol), transitorios (arroz, cebada, fréjol, maíz duro seco, maíz suave choclo, papa, trigo y yuca), producción animal (asnal, caballar, caprino, mular, pollos-gallinas en campo, pollos-gallinas en planteles avícolas, porcino y ovino), pastos (cultivados y naturales), montes-bosques, software con licencia (Microsoft Excel, Statistical Analysis System -SAS- y Wolfram System Modeler –Wolfram-) y open office (R, R-Studio y Python-Jupyter) esta obra presenta y desarrolla ejercicios teórico-prácticos con diferentes niveles de rigurosidad científica, interpretación con criterios informales o nocionales y formales de manera confiable, práctica y de uso común en estadística. El objetivo de Estadística para Ingeniería con Ofimática es contribuir a superar deficiencias de estudiantes, investigadores y personas involucradas en el manejo de bases de datos mediante la sensibilidad y experiencia de profesionales del área; por lo tanto, esta obra interactúa amplia y profundamente con Regresión Lineal con Ofimática y Diseño Experimental Simplificado para Ingeniería con Ofimática, que paralelamente usan software con licencia y open office mencionados para desarrollo de ejercicios teóricos-prácticos complementarios.

ÍNDICE

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

La estadística es una rama de la matemática que usa, según ([2]), un gran conjunto de datos provenientes de una muestra poblacional, recursos naturales e industriales con el fin de hacer inferencias basadas en cálculo de probabilidades; es decir, es el arte y la ciencia que reúne datos, analizar, presenta e interpreta mediante el uso de modelos matemáticos, entendidos como la abstracción simplificada de una realidad más compleja en que siempre existirá una cierta discrepancia entre lo observado y lo previsto por estos ([3]). Tiene uso en múltiples disciplinas científicas, como ingeniería forestal, agronomía, agro negocios, economía, procesamiento de alimentos, ciencias de la salud, marketing, elecciones políticas, estudios de mercado, control de calidad, inventarios, padrón de cobro de impuestos, muestreo, entre otras.

Para un investigador que recién inicia, el procesamiento de recuento, enumeración, conteo o cuantificación de grandes volúmenes productivos, cosechas agrícolas, listas de productos, censos poblacionales u otros grandes conjuntos de datos están íntimamente relacionados, mediante técnicas estadísticas, con actividades soporíferas de organización de bases, tabulación de datos, presentación gráfica, estimación de cantidades "representativas", procesos inductivos, procesos deductivos, uso indiscriminado de software, posible pago de licencias, manejo complejo, material bibliográfico o tutoriales abstrusos, procesamiento automático de datos, criterios interpretativos incomprensibles y transcripción mecánica de resultados ([4]).

Con base en esto, el libro "Estadística para Ingeniería con Ofimática" presenta y desarrolla ejercicios teóricos-prácticos, interpretación con criterios informales o nocionales y formales respecto a estadística. Sus capítulos abordan conocimiento progresivo mediante bases de estadística, probabilidad (presenta, antecedentes, cálculo en espacios muestrales, análisis combinatorio, distribución de probabilidades, multidimensionales, independencia, condicionalidad y teorema de Bayes), variables, variabilidad y función de probabilidades (caracteriza y calcula una variable aleatoria discreta, continua, espacios muestrales,

[2] Fuente RAE (2018): dle.rae.es/srv/fetch?id=GjpDTiC
[3] Fuente: Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014.
[4] Fuente: Capa, S., H. 2015

esperanza matemática, representación gráfica probabilística, distribución muestral), muestreo (resume, define y clasifica sus tipos de distribuciones, infiere los tipos de muestreo probabilístico y no probabilístico e hipótesis (pruebas paramétricas y pruebas no paramétricas).

Finalmente, este libro pretende contribuir a superar deficiencias de estudiantes, investigadores y personas involucradas en manejo de bases de datos a través de la sensibilidad y experiencia de profesionales del área, pues su objetivo es aportar conocimiento en manejo de datos estadísticos de manera confiable, práctica y de uso común en diferentes disciplinas mediante el uso de ofimática actual, diversa, práctica, con licencia pagada o versión libre (Microsoft Excel, Calculadora TI – Nspire CAS de Texas Instrument, R Project for Statistical Computing –R i386 3.3.3 y R Studio –, Statistical Analysis System –SAS-, Wolfram System Modeler y Matrix Laboratory –MatLab-), con diferentes niveles de rigurosidad científica y mediante ejercicios prácticos que incluyan herramientas informáticas de uso simple hasta programación avanzada para que el usuario tengan certeza sobre salidas o resultados antes de aceptarlos inmediatamente.

2. BASES DE ESTADÍSTICA

2.1. INTRODUCCIÓN

2.1.1. Qué es Estadística.

Cuando coloquialmente se habla de estadística, se suele pensar en una relación de datos numéricos presentada de forma ordenada y sistemática ([5]). Sólo cuando se adentra en un mundo más específico como es el campo de la investigación de las Ciencias Sociales: Medicina, Biología, Psicología se percibe que la Estadística no sólo es algo más, sino que se convierte en la única herramienta que permite dar luz, obtener resultados y por tanto beneficios, en cualquier tipo de estudio, cuyos movimientos y relaciones, por su variabilidad intrínseca, no puedan ser abordadas desde la perspectiva de las leyes determistas.

Definiciones:

- Ciencia que estudia cómo debe emplearse la información y cómo dar una guía de acción en situaciones prácticas que entrañan incertidumbre ([6]).
- Se ocupa de los métodos y procedimientos para recoger, clasificar, resumir, hallar regularidades y analizar los datos, siempre y cuando la variabilidad e incertidumbre sea una causa intrínseca de los mismos; así como de realizar inferencias a partir de ellos, con la finalidad de ayudar a la toma de decisiones y en su caso formular predicciones ([7])).
- La estadística es una ciencia que se interesa por la recogida, presentación y resumen de datos y la obtención de información a partir de ellos con el propósito de estudiar posibles relaciones entre variables de interés para el hombre ([8]).
- (Sinónimo de dato): este se refiere al concepto popular de la estadística, es equivalente a datos con algún ordenamiento coherente y sistemático ([9]).
- Estudia el comportamiento de los fenómenos naturales y sociales en todas las ciencias. Esto se debe a la particularidad de ofrecer técnicas y métodos precisos para obtener información y analizarla ([10]).

[5] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014.
[6] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014.
[7] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014.
[8] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[9] Aragón, S. L. G. 2016.
[10] Aragón, S. L. G. 2016.

- "Rama de las matemáticas que estudia la recolección, análisis, interpretación y presentación de masas de información numérica" (Webster's New Collegiate Dictionary citado por ([11]).
- "Estadística es la rama del método científico que estudia los datos obtenidos por contar o medir las propiedades de poblaciones" (Stuart y Ord, 1991 citados por ([12])).
- Se "ocupa esencialmente de procedimientos para analizar información, en especial aquella que en algún sentido vago tenga un carácter aleatorio" (Rice, 1995 citado por ([13])).
- El arte y la ciencia de reunir datos, analizarlos, presentarlos e interpretarlos" ([14]).

Estadística se define como el arte y la ciencia de reunir datos, analizarlos, presentarlos e interpretarlos mediante el uso de modelos matemáticos (econometría). Especialmente en los negocios y en la economía, la información obtenida proporciona a directivos, administradores y personas que deben tomar decisiones una mejor comprensión del negocio o entorno económico, permitiéndoles así tomar mejores decisiones con base en mejor información.

2.1.2. Estadística y Agricultura

Los trabajos que estudian la relación entre Agricultura y Estadística datan el inicio de su revisión histórica a finales del siglo XVIII. El artículo de Crete de Palluel (1788) se menciona como la primera referencia en la que se cita un experimento en agricultura, diseñado en parte según los principios estadísticos que serían formalizados mucho después. En la historia de la mencionada relación, se pueden distinguir, a grandes rasgos, tres periodos claramente diferenciados:

- **Desde finales del siglo XVIII hasta el comienzo de la primera guerra mundial**: William Sealy Gosset (1876-1937), "Student", realiza ensayos de variedades de cebada, para la industria cervecera, utilizando métodos empíricos, Student estudia la distribución t, (caracterizada más tarde analíticamente por Fisher), en su intento por superar las

[11] Anderson, C. D; Sweeney, D.J; Williams, T. A; Camm, J. D. y Cochran, J. J. 2016.
[12] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[13] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[14] Anderson, C. D; Sweeney, D.J; Williams, T. A; Camm, J. D. y Cochran, J. J. 2016

limitaciones de la teoría existente hasta el momento (grandes muestras), cuando se trate con muestras de pequeña.

- **El periodo entre las dos guerras mundiales**: En Rothamsted entre los años 1919 y 1933, Ronald Aylmer Fisher, establece las bases de la matemática estadística, los fundamentos de las técnicas modernas de diseño, análisis de experimentos y desarrolla un gran número de métodos originales corno respuesta a los requerimientos de los investigadores que le consultan.
- **La época actual, posterior a la segunda gran guerra**: El periodo de la postguerra supone una época de consolidación y unificación de ideas anteriores, destacándose los libros de Frank Yates (1949) sobre muestreo, Finney (1952) sobre bioensayo, y William Gemmell Cochran-Gertrude Mary Cox (1950) sobre diseño de experimentos.

2.1.3. Investigación Científica o Convencional

Son investigaciones realizadas en campos experimentales de universidades o centros de investigación. En su mayoría se trata de investigaciones básicas, en laboratorios o invernaderos, donde las condiciones son conocidas y manipulables. Para poder calificar una investigación científica como tal, se debe seguir los pasos del método científico. A través de los años la investigación ha evolucionado, con el afán de garantizar objetividad y replicabilidad, se han creado normas, diseñado sistemas cada vez más complejos para situaciones y condiciones específicas, que tienen su validez sólo bajo esas circunstancias. Cada una de las fórmulas requiere de su sustento teórico y tiene sus limitaciones, que muchas veces son ignorados por investigadores menos capacitados y versados ([15]).

En lo que refiere a la estadística, que en la investigación convencional juega un rol mucho más importante que en la participativa, no hay que olvidar que la estadística y todas sus fórmulas se han creado para poder demostrar diferencias que a simple vista no se ven o son distorsionadas por la subjetividad del ojo. Muchos sobrestiman la estadística, sin tener en cuenta que con fórmulas mal usadas se pueden distorsionar ensayos. Cada resultado puede ser manipulado y presentado con la estadística a manera de ver del autor. Depende de la sinceridad de éste en no desfigurar los resultados.

[15] Ortiz, P. J. 2013

2.1.4. Método científico y Estadística

De acuerdo con ([16]), el método científico se basa en la toma y análisis de datos; sin embargo, los datos y las conclusiones obtenidas aplicando metodología estadística ejercen una profunda influencia en casi todos los campos de la actividad científica y humana. La estadística invade cada vez más cualquier investigación relativa a la ciencia La aplicación de la estadística en el método científico busca incrementar la credibilidad y confiabilidad de las investigaciones, pero no garantiza que en todos los casos el método estadístico aplicado haya sido correctamente utilizada o, peor aún, que sea válido. ¿Por qué debe preocupar la aplicación incorrecta de métodos estadísticos en un trabajo científico o en un informe técnico?:

- Las conclusiones pueden ser incorrectas.
- No todos los lectores están en condiciones de detectar el error y, esto, genera un importante "ruido" en la bibliografía científica.

El estudio de la Estadística y el modo de pensamiento capacita a la persona para evaluar objetiva y efectivamente si la información que recibe (tablas, gráficos, porcentajes, tasas, etc.) es relevante y adecuada. Por supuesto, la interpretación de cualquier problema requiere no sólo de conocimientos metodológicos, sino también de un profundo conocimiento del tema. Aun cuando una persona no esté interesada en especializarse en estadística, un entrenamiento básico en el teme permite una mejor comprensión de la información cuantitativa y cualitativa. Las áreas en que puede dividirse la Estadística son:

a) **Diseño** (planteamiento y desarrollo de investigaciones). Consiste en definir como se desarrollará la investigación para dar respuesta a las preguntas que motivaron la misma. Un estudio bien diseñado resulta simple de analizar y las conclusiones suelen ser obvias. Asimismo, un experimento pobremente diseñado o con datos inapropiadamente recolectados o registrados puede ser incapaz de dar respuesta a las preguntas que motivaron la investigación, más allá de lo sofisticado que sea el análisis estadístico.

b) **Descripción** (resumen y exploración de datos). Los métodos de la Estadística Descriptiva o Análisis Exploratorio de Datos ayudan a presentar los datos de modo que

[16] Ortiz, P. J. 2013

sobresalga su estructura. Existe la forma gráfica y resumirlos en uno o dos números que pretenden caracterizar el conjunto con la menor distorsión o pérdida de información posible. No obstante, datos erróneos o inesperados serán procesados de modo inapropiado y ni usted ni el pc se darán cuenta a menos que realice previamente un análisis exploratorio de los datos.

c) **Inferencia estadística** hace referencia a un conjunto de métodos que permiten hacer predicciones acerca de características de un fenómeno sobre la base de información parcial acerca del mismo.

Los métodos de la inferencia permiten proponer el valor de una cantidad desconocida (estimación) o decidir entre dos teorías contrapuestas cuál de ellas explica mejor los datos observados (test o prueba de hipótesis). El fin último de cualquier estudio es aprender sobre las poblaciones, pero es usualmente necesario y, más práctico, estudiar solo una muestra de cada una de las mismas.

De acuerdo con ([17]), "los datos muestrales deben reunirse de forma adecuada, como en un proceso de selección aleatoria y si los datos muestrales no se reúnen de forma adecuada, resultaría inútil que ningún método estadístico podría salvarlos".

2.1.5. Población, Muestra, Parámetro, Estimador, Estimada y Estadístico

Individuos o elementos. Personas u objetos que contienen cierta información que se desea estudiar.

Población. Es la colección completa de todos los elementos individuos, objetos o medidas de interés, como puntuaciones, personas, mediciones, etc., a estudiar. Se dice que la colección es completa, pues incluye a todos los sujetos que estudiarán ([18]) y ([19])). Ejemplo: Los estudiantes inscritos en la UEB, los alumnos de la Escuela de Ingeniería Agroindustrial o todos los estudiantes de la Facultad de Ciencias Agropecuarias, Recursos Naturales y del Ambiente.

Muestra. Es un subconjunto, una porción o una parte de miembros seleccionados de una población de interés. Ejemplo: Un sondeo de Gallup preguntó a 1087 adultos "¿consume

[17] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[18] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[19] Anderson, C. D; Sweeney, D.J; Williams, T. A; Camm, J. D. y Cochran, J. J. 2016

bebidas alcohólicas como licor, vino o cerveza o es abstemio?". Las 1087 personas de la encuesta constituyen una muestra, mientras que la población consiste en el conjunto de 202,682, 345 personas adultas ([20]).

Censo. Es la colección de datos de cada uno de los miembros de la población. Ejemplo: Cada 10 años los gobiernos intentan obtener datos de cada ciudadano, pero no logra su objetivo, pues es imposible localizar a cada uno de ellos ([21]). Según ([22]), cuando existen datos para toda la población (censo) no es necesario usar métodos de estadística inferencial, pues es posible calcular exactamente los parámetros de interés.

Parámetro. Es una medición numérica que describe algunas características de una población. Ejemplo: Cuando Lincoln fue elegido presidente por primera vez recibió el 39.82% de 1´865,908 votantes. Si se supone que el conjunto de todos esos votos es la población por considerar, entonces el 39.82 % es un parámetro, no un estadístico ([23]).

Estimador. Estadístico utilizado para estimar un parámetro cuya función de densidad se llama función de densidad muestral ([24]).

Estimada o Valor Estimado. Cuando las variables aleatorias observables son reemplazadas por valores muestrales observados se generan valores estimados de los parámetros. Es importante tener presente que los valores estimados en particular no pueden ser evaluados como "buenos" ni como "malos" ([25]).

Estadístico. Es una medición numérica que describe algunas características de una muestra. Ejemplo: Con base en una muestra de 877 ejecutivos encuestados, se encontró que el 45 % de ellos no contrataría a alguien con un error ortográfico en su solicitud de empleo. Esta cifra del 45 % es un estadístico, pues está basada en una muestra y no en la población completa de todos los ejecutivos ([26])

Datos. Son las observaciones recolectadas, como mediciones, géneros, respuestas de encuesta, entre otras. *Datos cuantitativos* consisten en números que representan conteos o

[20] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[21] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[22] Ortiz, P. J. 2013
[23] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[24] Oteyza, E; Lam, E; Hernández, C. y Carrillo, A. 2015
[25] Oteyza, E; Lam, E; Hernández, C. y Carrillo, A. 2015
[26] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.

mediciones. Ejemplo: Los pesos de las supermodelos, estaturas de personas, cantidad de automóviles que poseen. *Datos cualitativos*, categóricos o de atributo, se dividen en diferentes categorías que se distinguen por alguna característica no numérica. Ejemplo: colores de ojos verde, café, azul, marrón, etc., género (masculino o femenino) de atletas profesionales, etcétera ([27]).

2.1.6. División de la Estadística

Según ([28]), para aplicar métodos estadísticos a la información disponible, es necesario tener presente los tipos de problemas que esta ciencia resuelve.

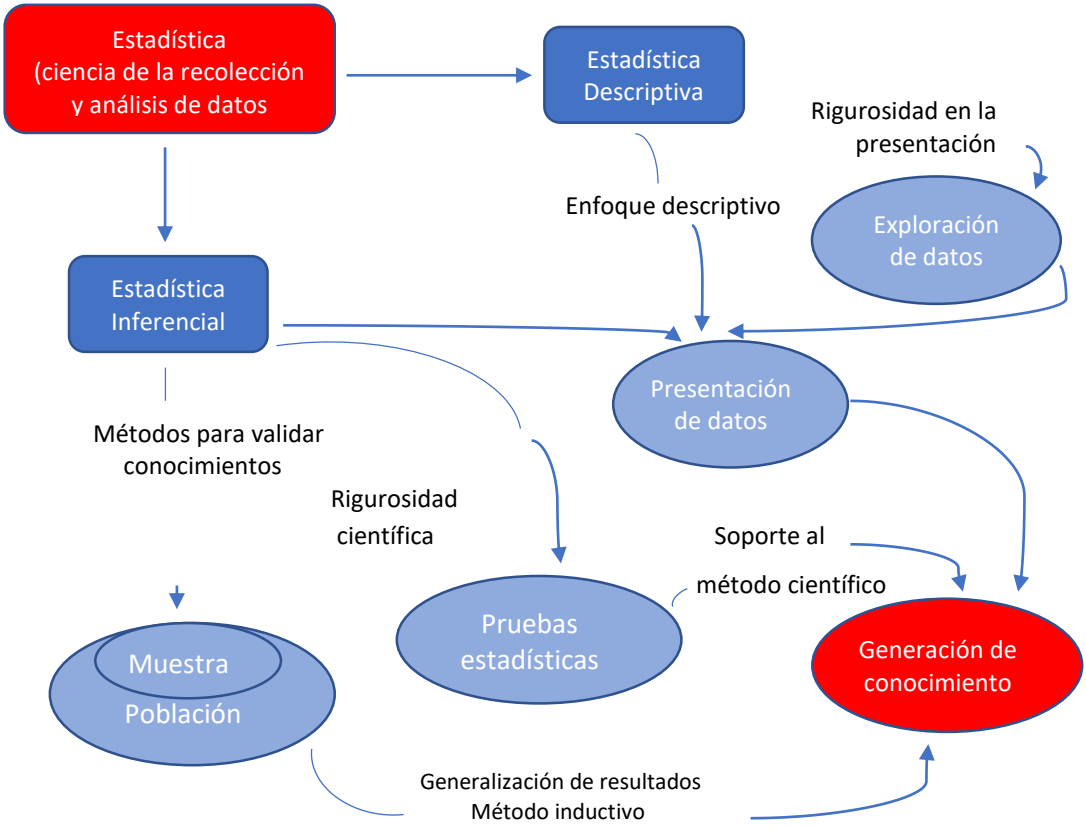

Figura 1. Mapa conceptual de Estadística

[27] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[28] Capa, S.,H. 2015

Estadística descriptiva. Son técnicas para recopilar, organizar, procesar y presentar datos obtenidos en muestras. El primer problema que, históricamente, aborda la Estadística es la descripción de datos. Suponga que se toman ciertas mediciones, que pueden ser gastos de alimentación en familias, producción de máquinas de un taller o preferencias en un grupo votante. El objetivo de esta división es describir, analizar y representar un grupo de datos utilizando métodos numéricos, gráficos que resumen y presentan la información contenida en ellos.

Estadística inferencial. Es frecuente que, por razones técnicas o económicas, no sea posible estudiar elementos de una población. Por ejemplo, para determinar la opinión de una población ante elecciones sólo se investiga a un grupo pequeño, pues es imposible consultar a todas las personas en capacidad de votar. Análogamente, se acude a una muestra para estudiar la rentabilidad de un proceso de fabricación o determinar el nivel de ocupación de la población. Es decir, apoyándose en el cálculo de probabilidades y a partir de datos muéstrales, efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos llamado población. Es un procedimiento aplicado para decidir si un proceso industrial funciona o no, según ciertas especificaciones, estudiar relación entre consumo de tabaco con cáncer, juzgar la demanda potencial de un producto, mediante un estudio de mercado, orientar estrategias electorales de un partido político, interpretar una prueba de inteligencia, etcétera.

Diseño de experimentos. Es cualquier proceso o estudio en que se realiza una recolección de datos donde el investigador, usualmente, tiene control sobre algunas condiciones bajo que el experimento tiene lugar. Por lo tanto, su objeto es planificar recogida de datos tal que queden garantizadas las suposiciones que requieren los métodos estadísticos que a utilizar en el análisis de los mismos. Por ejemplo: el desarrollo de un nuevo medicamento, preparación de una nueva aleación de acero para uso en automóviles u otras investigaciones, es necesario realizar experimentos para comparar su efectividad con otros previos.

Parámetro. Función definida sobre los valores numéricos de características medibles de una población

Población Finita. Como es el caso del número de personas que llegan al servicio de urgencia de un hospital en un día.

Población Infinita. Si por ejemplo estudiamos el mecanismo aleatorio que describe la secuencia de caras y cruces obtenida en el lanzamiento repetido de una moneda al aire.

Caracteres. Propiedades, rasgos o cualidades de los elementos de la población. Estos caracteres pueden dividirse en cualitativos y cuantitativos.

Modalidades. Diferentes situaciones posibles de un carácter. Las modalidades deben ser a la vez exhaustivas y mutuamente excluyentes —cada elemento posee una y sólo una de las modalidades posibles.

Clases. Conjunto de una o más modalidades en el que se verifica que cada modalidad pertenece a una y sólo una de las clases.

2.1.7. Campos de aplicación. La estadística tiene su aplicación en: Ciencias de la Salud ([29])**,** Procesamiento de Alimentos ([30])**,** Marketing ([31])**,** Elecciones políticas**.** ([32])**,** Estudios de mercado**. (**([33]) [34]**).**

2.1.8. Pensamiento crítico

El éxito en el curso de estadística por lo regular requiere de más sentido común que destreza matemática, aunque Voltaire advirtió "el sentido común no es muy común". Actualmente, con la ayuda de pc y calculadoras no es necesario dominar algoritmos complejos de operaciones matemáticas. En su lugar, se enfoca en interpretación de datos y resultados.

El estadístico Benjamín Disraeli dijo "Hay tres clases de mentiras: mentiras, viles mentiras y estadísticas". Aunque, "las cifras no mientan, los mentirosos calculan las cifras". Andrew Lang dijo "Algunas personas utilizan la estadística como un borracho usa los postes de alumbrado: como apoyo más que como iluminación". El autor Franklin P. Jones escribió "la estadística puede usarse para sustentar cualquier cosa, en especial a los estadísticos". En

[29] Ortiz, P. J. 2013
[30] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[31] Anderson, C. D; Sweeney, D.J; Williams, T. A; Camm, J. D. y Cochran, J. J. 2016
[32] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[33] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015.
[34] Un conjunto de ejercicios de identificación de población de interés, meta inferencial y cómo emprendería la recolección de una muestra se ubicar en la carpeta "y sub carpeta".

Esar´s Comic Dictionary se encuentra la definición que un estadístico es “un especialista que reúne pensamientos y luego los conduce al extravío”.

Estas afirmaciones se refieren a ejemplos donde los métodos estadísticos se usaron de forma incorrecta, de manera que resultaron engañosos en última instancia. Hay dos fuentes de este engaño:

1) El intento malintencionado por parte de personas deshonestas.

2) Los errores de descuido cometidos por personas que no conocen nada mejor.

3) La falta de capacitación, actualización, interés por mejorar sus conocimientos, técnicas, uso de herramientas, aplicarlos a la vida profesional real, al servicio del ser humano, sin intereses pretenciosos de dimensiones ambiciosas, cuyo objetivo sea servir a la ciencia y a la humanidad.

También, se debe tener en cuenta, como ciudadanos responsables y empleados profesionales valiosos, la fuente de información, calidad, cantidad y una habilidad básica para distinguir entre conclusiones estadísticas que parecen ser válidas de las que son gravemente defectuosas. Asimismo, en aplicaciones reales y con significado se debe ser cuidadoso para interpretar correctamente los resultados de métodos estadísticos válidos.

Muestra de respuesta voluntaria (muestra autoseleccionada). Es aquella donde los sujetos deciden ser incluidos como elementos de la muestra por sí mismos. Ejemplos: se puede aplicar una encuesta por internet anuncios en periódico, radio o tv, u otros medios, donde los individuos por sí mismos deciden si participan o no. Puede darse el caso que las muestras no sean representativas de toda la población e incluso pueden adolecer de una carencia importante. Con muestras de respuesta voluntaria como éstas, sólo es posible llegar a conclusiones válidas acerca del grupo específico que decide participar, pero sería una práctica incorrecta común establecer conclusiones acerca de una población más grande.

Muestras pequeñas. Las conclusiones no deben basarse en muestras que son sumamente pequeñas. Ejemplo: El Children´s Defense Fund reportó que, de los estudiantes de escuela secundaria suspendidos en una región, el 67% fueron suspendidos al menos tres veces. ¡Pero esta cifra está basada en una muestra de sólo tres estudiantes! Incluso, una muestra puede parecer relativamente grande (como una encuesta de “2000 adultos estadounidenses seleccionados al azar”), pero si se obtienen conclusiones acerca de los

subgrupos, estas estarían basadas en muestras demasiado pequeñas. Si es importante una muestra que sea suficientemente grande, también lo es tener datos muestrales que se recolecten de una forma adecuada, como la elección aleatoria. Aun las muestras grandes llegan a ser muestras erróneas.

Gráficas. Las gráficas, como barras y circulares, en ocasiones sirven para exagerar o disfraza la verdadera naturaleza de los datos. Ejemplo: las siguientes gráficas representan los mismos datos del Berau of Labor Statistics, aunque el inciso b está diseñado para exagerar la diferencia entre salarios semanales de hombres y mujeres. Al no iniciar el eje vertical en cero, la gráfica del inciso b tiene a producir una impresión subjetiva engañosa, que hace que los lectores incorrectamente crean que la diferencia es mucho peor de la realidad. Se debe analizar la información numérica dada en ella, para no engañarse por su forma general.

Mediana de ingresos semanales para edades de 16 a 24 años (en dólares).

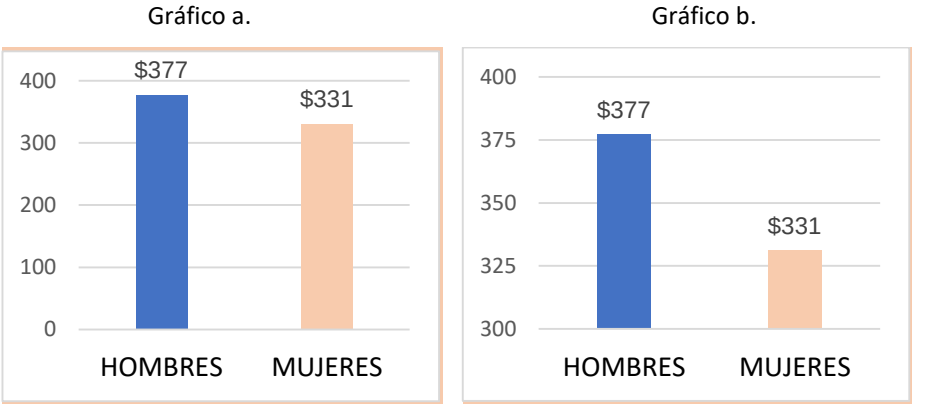

Figura 2. Gráfica de salarios hombres y mujeres del Berau of Labor Statistics ([35])

Pictogramas. Los dibujos de objetos, llamados pictogramas, también pueden resultar engañosos. Algunos objetos que se usan comúnmente para representar datos incluyen objetos tridimensionales, como bolsas de dinero, pilas de monedas, tanques militares (gastos militares), barriles (producción petrolera) y casas (construcción de casas). Al dibujar estos objetos, los artistas llegan a crear impresiones falsas que distorsionan las diferencias, sino que aumenta en un factor de cuatro. Ejemplo: si se duplica cada lado de un cuadrado, el área no tan sólo se duplica, sino que aumenta en un factor de cuatro. Si duplica cada lado de un

[35] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015

cubo, el volumen no se duplica simplemente, sino que se incrementa en un factor de ocho. Entonces, Si los impuestos se duplican durante una década, un artista podría representar las cantidades de impuestos con una bolsa de dinero para el primer año y otra de dinero dos veces más ancha, dos veces más alta y dos veces más profunda para el segundo año. En vez de parecer que los impuestos se duplican, parecerá que alimentaron en un factor de ocho y así el dibujo distorsionaría la verdad.

Figura 3 Representación de un pictograma (36)

Porcentajes. A veces se utilizan porcentajes engañosos o pocos claros. Si usted toma el 100% de alguna cantidad, está tomándolo todo (no debería requerir de un 110% de esfuerzo para que tenga sentido). Ejemplo:

a) Para encontrar el porcentaje de una cantidad, excluya el símbolo % y divida el valor del porcentaje entre 100, después multiplique por la cantidad. Este ejemplo muestra que el 6% de 1200 es 72: el 6% de 1200 $respuestas$=(6/100)∗1200=72.

[36] Fuente: Wackerly, D. D; Mendenhall, W. y Scheaffer, R. L. 2010

b) Para convertir de una fracción a un porcentaje divida el denominador entre el numerador para obtener un número decimal equivalente y después multiplíquelo por 100 y agregue el símbolo %: $3/4=0.75 \Rightarrow 0.75*100=75\%$.

c) Para convertir de un número decimal a un porcentaje, multiplíquelo por 100%: $0.234 \Rightarrow 0.234*100=23.4\%$.

d) Para convertir de un porcentaje a un número decimal, elimine el símbolo % y divida entre 100: $85\%=85/100=0.85$.

Preguntas predispuestas. Existen muchos aspectos que afectan las preguntas de una encuesta. Éstas llegan a estar "cargadas" o redactadas intencionalmente de manera que propicien una respuesta deseada. Ejemplos:

a) 97% sí: "¿Debe el presidente utilizar su poder de veto para eliminar los desperdicios?".

b) 57% sí: "¿Debe el presidente utilizar su poder de veto o no?"

c) Se preguntó a diferentes sujetos: "se gasta muy poco dinero en subsidios del Estado" y "se gasta muy poco dinero en asistencia a los pobres". Aun cuando el pobre es quien recibe el dinero del Estado, sólo el 19% estuvo de acuerdo cuando se usaron las palabras "subsidio del Estado", aunque el 63% estuvo de acuerdo con "asistencia a los pobres".

Orden de las preguntas. En ocasiones las preguntas de una encuesta se cargan de forma no intencional, en virtud de factores como el orden de los reactivos que se someten a consideración. Ejemplo:

a) "¿Cree usted que el tránsito vehicular contribuye a la contaminación del aire más o menos que la industria?" y "¿Cree ud que la industria contribuye a la contaminación del aire más o menos que el tránsito vehicular?". Cuando se presentó primero el tránsito, el 45% culpó al tránsito y el 27% culpó a la industria. Cuando la industria se presentó primero, el 24% culpó al tránsito y el 57% culpó a la industria.

Rechazo. Cuando se invita a las personas a contestar una encuesta, algunas se niegan con firmeza a responder. La tasa de rechazo ha crecido en años recientes, en parte porque muchos vendedores persistentes de empresas de telemercadeo buscan vender bienes o servicios comenzando con una introducción de ventas que suena como si fuera parte de una encuesta de opinión. Wheeler Michael, autor de Lies, Damn Lien and Statistics, indica "las

personas que se niegan a hablar con los entrevistadores parecen ser diferentes de quienes no lo hacen. Algunas quizás tengan miedo a extraños y otras sean celosas de su privacidad, pero su negativa a hablar demuestra que su visión del mundo circundante es marcadamente diferente de aquellas otras personas que permiten a los entrevistadores entrar en sus hogares".

Correlación y causalidad. El término correlación indica que dos variables están altamente relacionadas (como riqueza y coeficiente intelectual), aunque la correlación no implica causalidad (una asociación estadística entre dos variables, no se puede concluir que una de las variables es la causa de la otra o que la afecta directamente). En los medios de comunicación es bastante común reportar una correlación recién encontrada con una redacción que indica o implica directamente que una de las variables es causa de la otra.

Números precisos. "En la actualidad existen 103´215,027 hogares en Estados Unidos". Puesto que esta cantidad es muy precisa, mucha gente considera que es exacta. En este caso, ese número es un estimado y se considera mejor decir que el número de hogares es alrededor de 103 millones.

Estudios para el propio beneficio. Algunas veces los estudios reciben patrocinio de grupos con intereses específicos que buscan promover. Ejemplo: Kiwi Brands, fabricante de abrillantador de calzado, encargó un estudio en que se suscitó una declaración impresa en algunos periódicos "de acuerdo con una encuesta nacional realizada a 250 empleadores profesionales, la razón más común del fracaso de un solicitante de trabajo del sexo masculino al dar una buena impresión, fue llevar los zapatos desaseados". En los últimos años ha generado preocupación creciente la práctica de las compañías farmacéuticas de financiar a doctores que realizan experimentos clínicos y reportan sus resultados en revistas de prestigio, como Journal of America Medical Association.

Imágenes parciales. "El 90% de todos nuestros automóviles, vendidos en el país en los últimos 10 años continúa circulando". Millones de consumidores escucharon ese anunció comercial, pero ese 90% fueron vendidos en los últimos 3 años. La afirmación era técnicamente correcta, aunque muy engañosa al presentar resultados completos.

Distorsiones deliberadas. Cynthia Crossen publicó, en el libro Tainted Truth, resultados que mostraban que, entre las compañías de renta de automóviles, Avis fue la

ganadora en una encuesta realizada a personas que utilizaban ese servicio. Cuando Hertz solicitó información detallada acerca de la encuesta, las respuestas originales de ésta desaparecieron y el coordinador de encuestas renunció. Hertz demandó a Avis (por publicidad falsa basada en la encuesta) y a la revista: Al final las compañías llegaron a un acuerdo.

2.1.9. Variable aleatoria

Función con valores numéricos definida sobre un espacio muestral. En otras palabras, es una variable aleatoria si el valor que asume es un suceso numérico aleatorio. Ejemplo: observar el número de defectos en un mueble o el registro de aprovechamiento de un estudiante en particular. Tipos de variables[37]:

- **Discreta**. Es una variable que sólo puede asumir un conjunto numerable de valores. Es decir, que si se puede enumerar es discreta. Ejemplos: número de tornillos defectuosos en una muestra de 10 unidades extraídas de una producción industrial. Número de casas rurales que cuentan con servicio eléctrico en una región. El número de fallas de un aeroplano en un periodo de tiempo. El número de personas que esperan una consulta en un consultorio médico.
- **Continua**. Es una variable que puede asumir el número infinitamente grande de valores correspondientes a los puntos sobre un intervalo en una línea recta. Es decir, la palabra continua significa que procede sin interrupción y proporciona la clave para identificar a las variables aleatorias continuas. Una característica importante de esta variable es que si las mediciones u observaciones tienen un conjunto de valores que forman puntos sobre una línea sin interrupción o espacio entre ellos. Ejemplos: Estatura de una persona. Tiempo de vida de una célula humana. Cantidad de azúcar en una naranja. Tiempo requerido para completar una operación de montaje en un proceso de fabricación

[37] Un conjunto de ejercicios se encuentra en la carpeta “”

2.2. DATOS

2.2.1. Características

Según ([38]), datos son hechos/informaciones y cifras que se recogen, analizan y resumen para su presentación e interpretación. A todos los datos reunidos para un determinado estudio se les llama conjunto de datos para el estudio. Ejemplo:

Variable: característica de los elementos de interés Bolsa de Valores: N (Bolsa de Nueva York) y NQ

Variables: Ticker (identifica la acción en la lista de la bolsa); Posición en Business Week (Fortaleza de empresa del 1-500); Precio por acción ($): precio de cierre al 28/02/2005 y Ganancia por acción ($) ganancias/acción en últimos 12 meses,

TABLA 1 conjunto DE DATOS DE 25 EMPRESAS S&P 500

Empresa	Bolsa de valores	Denominación abreviada Ticker	Posición en Business Week	Precio por acción ($)	Ganancia por acción ($)
Abbott Laborarories	N	ABT	90	46	2.02
Altria Group	N	MO	148	66	4.57
Apollo Group	NQ	APOL	174	74	0.90
Bank of New York	N	BK	305	30	1.85
Bristol-Myers Squibb	N	BMY	346	26	1.21
Cincinnati Financial	NQ	CINF	161	45	2.73
Comcast	NQ	CMCSA	296	32	0.43
Deere	N	DE	36	71	5.77
eBay	NQ	EBAY	19	43	0.57
Federated Dept. Stores	N	FD	353	56	3.86
Hasbro	N	HAS	373	21	0.96
IBM	N	IBM	216	93	4.94
International Paper	N	IP	370	37	0.98
Knight-Riddeer	N	KRI	397	66	4.13
Manor Care	N	HCR	285	34	1.90
Medtronic	N	MDT	53	52	1.79
National Senuconducior	N	NSM	155	20	1.03
Novellus Systems	NQ	NVLS	386	30	1.06

Elemento: entidades de las que se obtienen los datos

Observación: conjunto de mediciones obtenidas para un determinado elemento. P. ej. IBM es N, IBM, 216,93 y 4.98

[38] Anderson, C. D; Sweeney, D.J; Williams, T. A; Camm, J. D. y Cochran, J. J. 2016

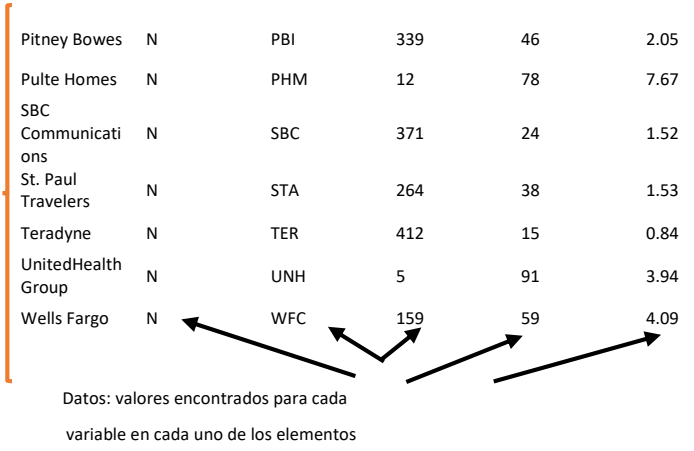

Figura 4 Representación tabular de datos con sus variables ([39])

La tabla anterior muestra un conjunto de datos donde estas empresas representan 76% de la capitalización de mercado de todas las acciones de Estados Unidos. Las acciones de S&P 500 son estrechamente observadas por los inversionistas y por los analistas de Wall Street.

2.2.2. Tipos

Elementos son las entidades de las que se obtienen los datos. En el conjunto de datos de la tabla 1.1, cada acción de una empresa es un elemento; los nombres de los elementos aparecen en la primera columna. Como se tienen 25 acciones, el conjunto de datos contiene 25 elementos.

Una variable es una característica de los elementos que es de interés. El conjunto de datos de la tabla anterior contiene las cinco variables siguientes: Bolsa de valores (mercado bursátil). Dónde se comercializa (cotiza) la acción: N (Bolsa de Nueva York) y NQ (Mercado Nacional Nasdaq). Ticker (denominación abreviada): Abreviación usada para identificar la acción en la lista de la bolsa. Posición en BusinessWeek: Número del 1 al 500 que indica la fortaleza de la empresa. Precio por acción ($): El precio de cierre (28 de febrero de 2005). Ganancia por acción ($): las ganancias por acción en los últimos 12 meses.

2.2.3. Escalas de medición

Nominal. Cuando el dato de una variable es una etiqueta o un nombre que identifica un atributo de un elemento. Tal que, se usa un código o una etiqueta no numérica. Ejemplo:

[39] Fuente: Anderson, C. D; Sweeney, D.J; Williams, T. A; Camm, J. D. y Cochran, J. J. 2016

variable bolsa de valores (mercado bursátil) es nominal porque N y NQ son etiquetas que se usan para indicar dónde cotiza la acción de la empresa. Aunque, puede ser para facilitar la recolección de los datos y para guardarlos en una base de datos en una computadora puede emplearse un código numérico en el que 1 denote la Bolsa de Nueva York y 2 el Mercado Nacional Nasdaq, pero sigue siendo escala nominal.

Ordinal. Si los datos muestran las propiedades de los datos nominales y además tiene sentido el orden o jerarquía de los datos. Tienen las propiedades de los datos nominales, pero además pueden ser ordenados o jerarquizados en relación con la calidad del servicio. Ejemplo: una empresa Lechera (Rey Leche) envía a sus clientes cuestionarios para obtener información sobre su servicio de reparación. Cada cliente evalúa el servicio de reparación como excelente, bueno o malo. Un dato excelente indica el mejor servicio, seguido por bueno y, por último, malo.

Intervalo. Los datos de intervalo siempre son numéricos. Incluso, una escala de medición para una variable es de este tipo. Ejemplo: Las calificaciones obtenidas por tres alumnos en la prueba de matemáticas con 620, 550 y 470, pueden ser ordenadas en orden de mejor a peor.

De razón. Esta escala requiere que se tenga el valor cero para indicar que en este punto no existe la variable, pues si los datos tienen todas las propiedades de datos de intervalo y la proporción entre dos valores tiene significado. Ejemplo: considere el costo de un automóvil. El valor cero para el costo indica que el automóvil no cuesta, que es gratis. Además, si se compara el costo de un automóvil de $30 000, con el costo de otro automóvil, $15 000, la propiedad de razón muestra que $\$30{,}000/\$15{,}000 = 2$: el primer automóvil cuesta el doble del costo del segundo.

2.2.4. Valores atípicos, inusual o extremo

Es un conjunto de datos, una observación lejana, en valor, del resto de datos; es decir, un dato inusualmente grande o pequeño respecto al resto. Puede ser el resultado de un error en una medición, en cuyo caso distorsiona la interpretación de datos al tener una influencia excesiva respecto a cálculos a partir de la muestra. Si el valor atípico es un resultado genuino es importante, pues podría indicar un comportamiento extremo del proceso de estudio. Con

base en esto, todos los valores atípicos serán examinados cuidadosamente antes de realizar un análisis formal y no se eliminarán sin una justificación previa.

2.2.5. Representación gráfica

Es una forma eficiente de conocer el comportamiento de un conjunto de datos, pues permite una descripción rápida y fácil de comprender. Su importancia es tal que todo análisis estadístico debe ser acompañado de esta forma.

2.2.5.1. Puntos

Es una manera de resumir datos cuantitativos, en que cada observación se representa mediante un punto sobre una recta numérica. Si se tuviera un resumen de muchos datos, cada punto puede representar un número fijo de individuos. Esta representación aprecia una localización general de sus observaciones, su dispersión y presencia de observaciones inusuales, atípicos o extremos. Se recomienda usarlo cuando se representa un máximo de 20 observaciones individuales, caso contrario será difícil distinguirlos. Se puede combinar sus representaciones de dos o más conjuntos de datos sobre un mismo gráfico con una forma sencilla de interpretación. Por ejemplo: triángulos, círculos, cuadrados, rectángulos u otra forma en vez de puntos.

Al crear este tipo de gráfico se determinará el rango de observaciones, cómo se representará y, también, fijar una escala apropiada que permita una buena representación de sus datos. En datos nominales u ordinales, un diagrama de puntos es similar a uno de barras, reemplazadas por un conjunto de puntos. En datos continuos, esta gráfica es similar a un histograma, rectángulos son sustituidos por puntos.

2.2.5.2. Tallo y hojas

El diagrama de puntos tiene algunas desventajas, como regresar de puntos a tallos y puede ser confuso si tiene gran cantidad de datos. Es conveniente usar otras herramientas gráficas. El diagrama de tallos y hojas es una técnica semi gráfica que, usada para ilustrar las principales características de datos, como localización, dispersión y simetría. Tiene la ventaja de presentar valores de datos y, por su forma, se puede usar en conjunto de datos hasta de 100 elementos.

Ejemplo:

08	19	17	01	07	09	05	16
13	15	04	02	00	04	01	12

1) Los datos se clasifican consideran las decenas tal que se consideran dos grupos. Uno comienza con 0 y el otro en 1 formando el tallo verticalmente:

0
1

2) Para cada elemento se anota el segundo dígito a la derecha de barra vertical, que construyen las hojas:

0	8	1	7	9	5	4	2	0	4	1
1	9	7	6	3	5	2				

3) Se ordena en forma ascendente:

0	0	1	1	2	4	4	5	7	8	9
1	2	3	5	6	7	9				

4) Se crean dos categorías en forma ascendente en cada decena, los dígitos de unidades del 0-4 forman el primero y 5-9 forman el segundo grupo:

0	0	1	1	2	4	4
	5	7	8	9		
1	2	3				
	5	6	7	9		

En casos en que la base de datos consta de más de dos cifras se escoge los rangos par agrupaciones que se harán. Después, mediante una coma se separan, llenadas las hojas:

33	55	79	106	188	47	118	248
47	58	82	113	208	60	88	

Con base en estos datos, se puede construir dos diagramas de tallo y hojas:

1) Primer diagrama:

Categoría	**Dígito**									
0-100	0	33	47	47	55	58	60	79	82	88
100-200	1	06	13	18	88					
200-300	2	08	48							

O, también:

Categoría	Dígito						
0-50	0	33	47	47			
50-100	0	55	58	60	79	82	88
100-150	1	06	13	18			
150-200	1	88					
200-250	2	08	48				
250-300	2						

Sin embargo, diagramas múltiples se pueden usar para comparar dos conjuntos de datos. Por consiguiente, se coloca un tallo común y hojas de un conjunto se sitúan a la izquierda y hojas del segundo conjunto a la derecha del tallo, respectivamente:

Dígito										
			44	1	33	47	47			
		57	79	1	55	58	60	79	82	88
01	23	34	42	2	06	13	18			
		06	78	2	88					
			33	3	08	48				
			5	3						
				4						

Los datos izquierdos están más agrupados en respecto a valores bajos, con rango mayor y fuerte asimetría, mientras que el conjunto derecho es asimétrico y con menor dispersión. Finalmente, estos diagramas se emplean para representar datos con decimales:

0.80	0.46	1.23	1.15	2.23	1.89	0.95	1.02	2.06	0.61	0.52	1.94

Su diagrama ascendente es:

Categoría	Dígito									
0-1.0	0.	46	52	61	80	95				
1.0-2.0	1.	02	15	23	89	94				
2.0-3.0	2.	06	23							

2.2.5.3. Sectores circulares o gráfica de pastel

Una alternativa para la representación de frecuencias relativas de un conjunto de categorías es la utilización de gráficos de sectores. En este caso a cada categoría se le asigna un sector representa su frecuencia. Una limitante para este tipo de representaciones es el número de categorías, ya que cuando éstas son muchas, la lectura del gráfico no es simple.

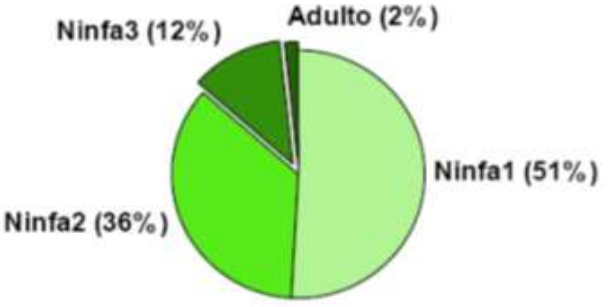

Figura 5 Representación gráfica de sectores circulares ([40])

2.2.5.4. Histograma

Es un conjunto de rectángulos, cada uno representa un intervalo de agrupación. Sus bases son iguales al intervalo de clase empleado en la distribución de frecuencias y alturas son proporcionales a la frecuencia absoluta n_i o relativa f_i de clase. Es apropiado para datos continuos, medidos con una misma escala y se emplea cuando es laborioso hacer un diagrama de tallo y hojas. Puede ayudar a revelar observaciones atípicas y brecha alguna entre datos ([41]).

Otra forma alternativa de presentar los resultados de la Tabla 1.4 es mediante el clásico histograma. La Figura siguiente presenta el histograma de frecuencias relativas y el polígono correspondientes al peso de las larvas del estadio 1. Lo más destacable que puede observarse es la marcada asimetría de la distribución; en comparación con la representación en box-plot es más difícil identificar los percentiles.

[40] Fuente: Galindo, E. 2006
[41] Capa, S.,H. a. 2015

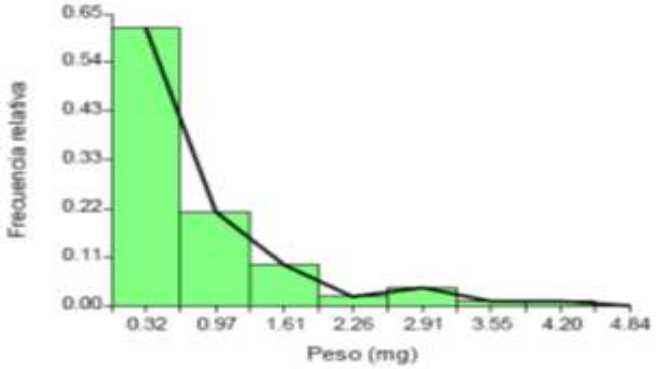

Figura 6 Representación gráfica de histograma (42)

2.2.5.5. Polígono de frecuencias

Es un gráfico obtenido de unir segmentos de recta con puntos que tienen proporcionalmente como abscisa a la marca de clase y como ordenada la frecuencia respectiva. Se cierra en ambos extremos en marcas adyacentes con frecuencia cero (43).

Cuando se estudia la asociación entre 2 variables (por ejemplo X e Y) es muy útil hacer un diagrama de dispersión. Este es un gráfico en el que cada observación está representada en el plano XY por un punto cuyas coordenadas están dadas por los valores registrados en ambas variables. En algunos casos un diagrama de dispersión puede ser modificado incluyendo segmentos de recta que unen los puntos del plano según un orden dado por el eje de abscisas. Como ejemplo, supóngase que se evalúa el número de callos obtenidos en cultivos de 200 anteras sometidas a un número creciente de días de frío.

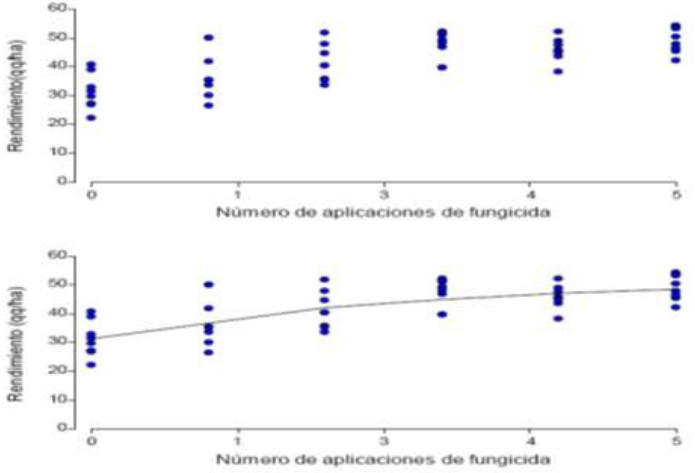

Figura 7. Representación gráfica de polígono de frecuencias (44)

42 Fuente: Galindo, E. 2006
43 Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014
44 Fuente: Galindo, E. 2006

Días	0	1	2	3	4	5	6
# Callos	150	50	45	40	38	27	15

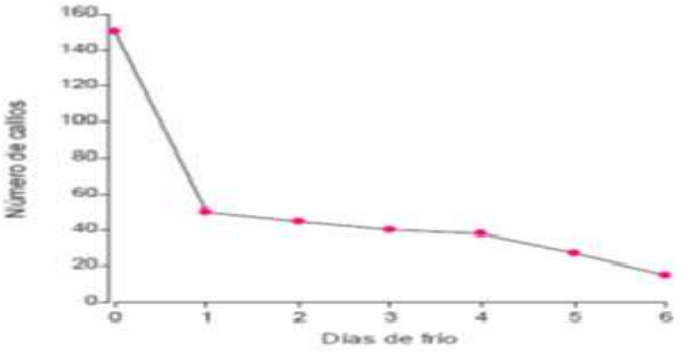

Figura 8. Representación gráfica de polígono de frecuencias con líneas ([45])

2.2.5.6. Ojiva

Es un polígono de frecuencias acumuladas o, en otras palabras, en abscisas se colocan los límites superiores de cada intervalo de clase y en ordenadas la frecuencia acumulada (absoluta o relativa) de la clase. Es útil en cálculos del número o porcentaje de observaciones correspondientes a un intervalo determinado de la variable y estimar percentiles de distribución de datos ([46]).

2.2.5.7. Balanza

Los gráficos antes mencionados no requieren realizar cálculos de medidas estadísticas. En consecuencia, este tipo de gráfico sí requiere estas y, por ello, son más útiles al analizar. Fue introducido en año 2000 como una herramienta que muestra, en un mismo gráfico, la forma de datos, su valor central y variabilidad al representar su promedio, mínimo, máximo y desviación estándar de sus datos. Los pasos de su cálculo son:

- Se estima el promedio, desviación estándar, mínimo y máximo de un conjunto de datos a analizar.
- En una recta se ubican sus valores promedio, mínimo y máximo. Los segmentos que unen el promedio con su mínimo se llaman brazos de balanza.
- Sobre la recta se ubican dos puntos –uno a la izquierda y el otro a la derecha de su valor promedio-, a una distancia igual a la desviación estándar.

[45] Fuente: Galindo, E. 2006
[46] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014

- Bajo el valor promedio se dibuja un triángulo:

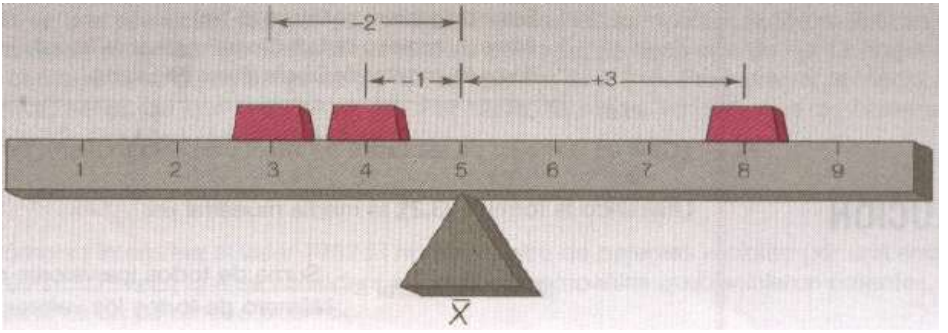

Figura 9 Representación gráfica de balanza ([47])
Este diagrama se interpreta:

- Si datos son simétricos, su valor promedio se sitúa en su centro.
- Si datos están agrupados en torno al centro, los brazos de balanza serán cortos. Caso contrario, si sus datos están dispersos en torno al centro, los brazos de la balanza serán largos.
- Si uno de los brazos de la balanza es mucho más largo que el otro indica que sus datos son asimétricos y existe posible presencia de valores atípicos en sus observaciones.

Puede ser útil combinar, en mismo gráfico, con un diagrama de puntos para visualizar la manera en que se distribuyen sus observaciones.

Ejemplo:

5	5	5	5	10	10	2	20	27	35
39	55	55	60	60	60	68	75	90	90

Con base en su análisis, estos datos presentan valores $\text{mín} = 5$, $\text{máx} = 90$, $\bar{x} = 90$ y $s = 29.3$. Por lo tanto, $\bar{x} - s = 39.7 - 29.3 = 10.4$ y $\bar{x} + s = 39.7 + 29.3 = 69.0$. Entonces, si se hace el gráfico su promedio no se ubica en el centro del rango; por lo que, se deduce que los datos son asimétricos y sus brazos de balanza no tienen igual longitud, que denota la posible presencia de valores atípicos en extremo derecho ([48]).

[47] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004
[48] Capa, S.,H. a. 2015

2.2.5.8. Box-Plot o caja con bigotes

Fue introducido en 1977 por John Wilder Tukey como una herramienta que muestra, en un gráfico, la forma de sus datos, su valor central y su variabilidad al presentar la mediana, cuartiles, rango Inter cuartil y rango de observaciones. Fundamentalmente, es útil para examinar la simetría de datos, la presencia de valores atípicos y comparar dos conjuntos de muchos datos. Las secuencias para construirla son:

1) Sobre una línea horizontal se ubican la mediana, cuartiles inferior-superior, datos mínimo y máximo.

2) Se construye una caja angosta que unen Q_1 y Q_3. Enseguida, se divide esta caja en dos partes mediante una línea que pase por Q_2.

3) Finalmente, se trazan vallas, dos rectas, desde cada extremo de la caja hacia el valor mínimo y el valor máximo de datos.

Estos gráficos tienen por objeto presentar sintéticamente los aspectos más importantes de una distribución de frecuencias. Aunque el box-plot es una representación apropiada para la distribución de frecuencias muestrales, a veces el tamaño de la muestra es pequeño y los cuantiles muestrales que de ella se obtienen no son confiables desde el punto de vista estadístico y en consecuencia la construcción del box-plot, que requiere de estas medidas, puede no ser buena. Ejemplo: Se toman muestras aleatorias de tamaño $n = 100$ de cada uno de tres estadios larvales de una especie de polilla forestal. Cada individuo es pesado y los resultados se presentan en tabla siguiente:

Cuadro 1. Base datos de 100 larvas por estadio de polilla forestal [49]

Peso (mg) de 100 larvas de cada estadio de una polilla forestal								
Estadio 1			**Estadio 2**			**Estadio 3**		
0.47	2.87	0.06	2.40	4.85	3.09	22.47	7.96	10.03
0.05	0.24	0.63	3.48	4.46	9.22	3.63	11.19	4.54
0.25	0.00	0.86	3.69	10.67	5.28	8.17	15.34	10.88
1.43	0.00	0.00	5.35	1.75	2.25	9.82	5.14	4.68
0.49	0.28	0.04	3.01	0.92	2.19	7.59	11.01	5.32
4.52	0.39	0.00	1.98	1.46	3.97	8.33	7.48	14.40
2.92	1.06	0.47	1.88	4.51	4.15	12.49	10.19	10.83
0.14	0.11	0.12	12.47	2.35	2.81	7.74	10.95	5.54
1.76	1.00	0.07	11.24	5.47	3.75	23.73	12.87	9.75
0.18	0.01	2.94	5.43	4.07	0.73	6.79	13.67	6.51
0.69	0.37	0.92	7.29	14.67	2.59	8.28	7.56	9.93
0.00	0.56	0.03	3.88	1.40	3.83	6.46	9.12	9.10
0.20	1.20	0.01	4.19	5.07	2.92	11.99	10.93	11.80
0.75	0.40	0.05	3.34	3.43	6.40	14.52	22.87	15.05
3.02	3.77	0.76	11.69	9.01	5.50	18.25	4.57	12.49
0.29	0.28	0.39	2.98	6.09	7.22	13.62	11.30	5.48
1.68	0.46	1.06	1.36	5.31	5.60	8.74	8.56	6.68
0.37	0.31	0.84	2.97	9.54	4.29	8.53	3.93	10.45
0.06	0.84	0.12	1.93	7.55	4.68	9.61	23.12	11.35
0.72	0.91	0.51	3.84	8.33	2.32	2.83	5.44	9.58
0.09	0.23	1.87	2.33	2.89	3.93	13.69	14.41	5.56
0.10	0.06	0.75	3.02	4.64	5.11	10.83	2.63	8.52
0.69	0.27	0.03	5.02	9.59	3.03	8.10	6.52	7.73
0.00	1.87	1.80	6.25	7.13	3.46	9.49	17.35	7.02
0.77	1.26	0.56	9.29	3.29	2.05	3.16	10.24	5.56
0.10	0.82	0.85	2.83	7.16	1.67	10.64	12.34	16.14
0.14	0.00	0.05	6.31	0.35	4.45	5.13	6.81	10.95
0.90	0.00	0.05	1.61	2.81	3.47	10.18	4.17	5.22
0.00	1.57	0.53	5.89	9.33	5.76	4.18	8.38	11.05
1.25	0.04	0.02	6.49	3.01	1.75	6.04	4.87	20.70
2.50	0.36	0.01	8.35	6.65	1.97	17.87	5.46	10.24
2.05	0.01	0.04	4.22	6.44	9.47	5.97	10.45	7.97
1.82	0.20		2.95	5.94		5.18	17.90	
1.76	0.00		2.61	5.43		10.19	3.44	

[49] Fuente: Galindo, E. 2006

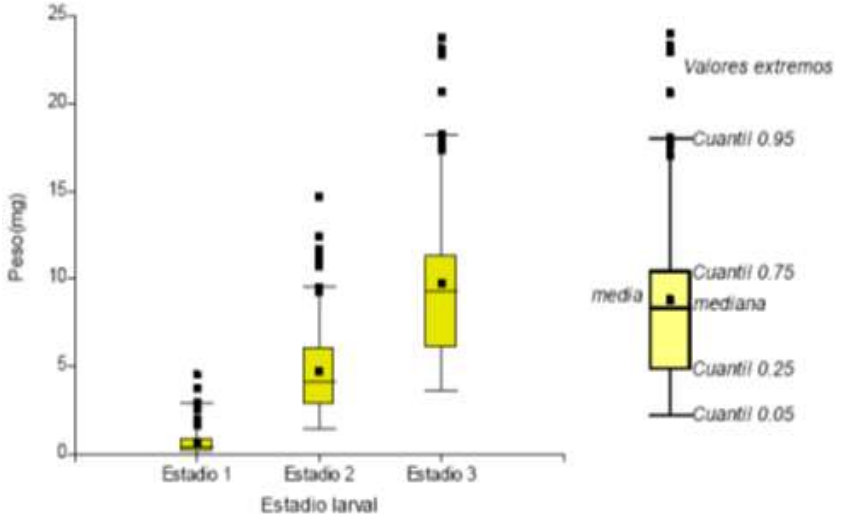

Figura 10. Representación gráfica de Box-Plot o caja con bigotes (50)

Todo conjunto de datos presenta ciertas características que permiten, en una primera aproximación, deducir el comportamiento del proceso del que fueron obtenido. Las tres principales características son localización, dispersión y simetría.

2.2.5.9. Localización

Fue ideada por Sir Ronald Aylmer Fisher en 1922. Es la posición relativa que ellos presentan. En general, se mide por el valor que tiene el punto medio del conjunto de datos. Un ejemplo es la medición de estatura de un grupo de personas, sus mediciones se localizarán según edades y porte de las personas, se supone que sus estaturas se pueden caracterizar mediante un valor promedio.

2.2.5.10. Dispersión

Este concepto fue ideado por Francis Galton (1886) y Wilhelm Lexis (1887). Parte de la idea que los valores muestrales no son iguales tal que su variación se llama dispersión. Al medirla se desea detectar el grado de diseminación de valores individuales alrededor del centro de las observaciones. En procesos de manufactura o medición, una alta precisión está asociada con una baja dispersión.

[50] Fuente: Galindo, E. 2006

2.2.5.11. Densidad de puntos (´Dot − Plot´)

En algunas circunstancias no sólo se quiere tener una imagen de los aspectos generales de la distribución sino, también, una visualización de los valores efectivamente observados. En estos casos el dot-plot, puede ser la representación más satisfactoria. El procedimiento de construcción es simple y consiste en dibujar un punto por cada uno de los valores observados en la muestra, ubicados según una escala (recta real) que se pone como referencia. Cuando hay más de una observación con el mismo valor, ésta se representa con otro punto ubicado en posición contigua al anterior y, sucesivamente, con el resto de las observaciones repetidas. Ejemplo: La siguiente tabla presenta los resultados observados del número de plántulas de malezas por m^2 para dos muestras de tamaño $\mathrm{n} = 20$.

Cuadro 2. Número de plántulas de malezas ([51])

Número de plántulas de malezas por $\boldsymbol{m^2}$

Potrero 1	5	9	5	4	8	4	7	4	7	5
	3	4	7	5	1	4	5	8	3	5
Potrero 2	1	2	3	1	4	4	5	6	1	2
	1	1	1	2	3	1	6	2	5	3

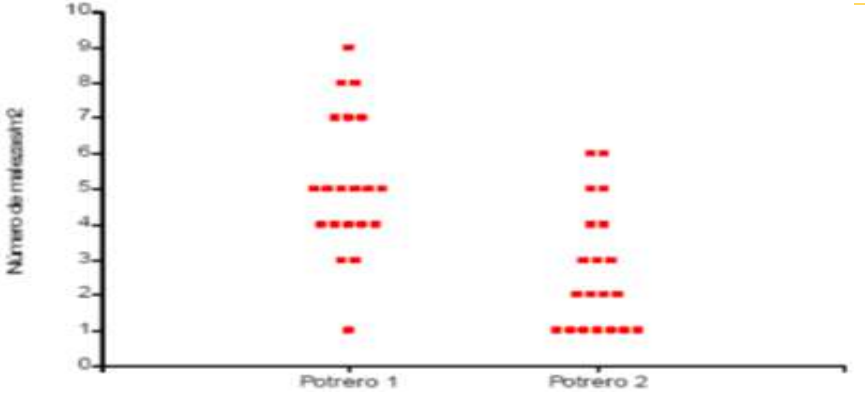

Figura 11. Representación gráfica de Densidad de puntos (´Dot − Plot´) ([52])

2.2.5.12. Coeficiente de asimetría

Un conjunto de datos es simétrico cuando sus valores están distribuidos de igual manera por encima y debajo de su punto medio. Los datos simétricos son fáciles de

[51] Fuente: Galindo, E. 2006
[52] Fuente: Galindo, E. 2006

interpretar, pues los datos que están por encima y debajo de su punto medio son considerados con un mismo criterio, permiten una fácil detección de valores atípicos y admiten comparación en conjuntos de datos similares, en términos de dispersión.

La asimetría en un conjunto de datos es el agrupamiento que presentan a un lado de su centro. Sus valores situados a un lado de la mitad de datos tienden a estar más alejados que sus valores ubicados en el otro lado.

2.2.5.13. Apuntamiento o curtosis

El coeficiente de apuntamiento o curtosis de una variable sirve para medir el grado de concentración de valores que toma en torno a su media. Se elige como referencia una variable con distribución normal tal que su coeficiente de apuntamiento de esta última es cero.

$$A_p = \frac{\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{n}\right)}{s^4} - 3$$

Con base en esto, una variable puede ser:

- **Leptocúrtica** si $A_p > 0$: es más apuntada que la normal. Valores que toma la variable están muy concentrados en torno a su media y existen pocos valores extremos.
- **Mesocúrtica** si $A_p = 0$: es tan apuntada como la normal.
- **Platicúrtica** si $A_p < 0$: es menos apuntada que la normal. Existen muchos valores extremos, las colas de la variable son muy pesadas.

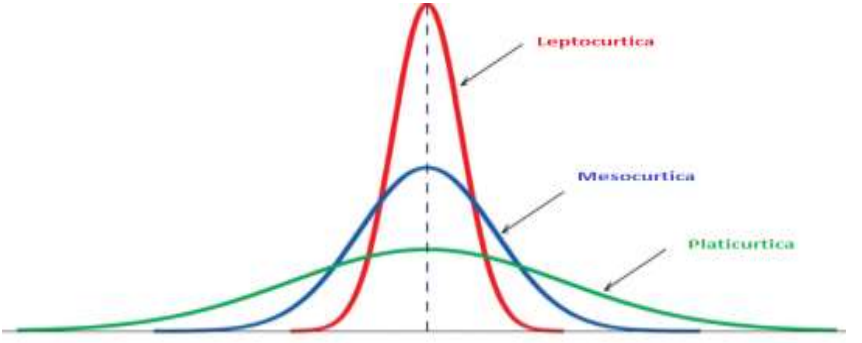

Figura 12. Representación gráfica de apuntamiento o curtosis ([53])

Ejemplo: Calcule coeficientes de simetría y apuntamiento de sueldos de diez personas que ganan su salario en dólares americanos.

[53] Fuente: www.google.com/search?q=Apuntamiento+o+curtosis&client=firefox-b&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjFmZX6tOjdAhVBh-AKHbMPBqQQ_AUICigB&biw=1366&bih=635#imgrc=aH6euGYmWmzLjM:

170	172	168	165	173	178	180	165	167	172

Se conoce que $\bar{x} = 171$ y $s = 5.1$. Asimismo:

$$A_p = \frac{\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n}\right)}{s^3}$$

$$= \frac{\left(\frac{(170-171)^3 + (172-171)^3 + (168-171)^3 + (165-171)^3 + \cdots + (172-171)^3}{10}\right)}{(5.1)^3}$$

$$= 0.421$$

$$A_p = \frac{\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{n}\right)}{s^4} - 3$$

$$= \frac{\left(\frac{(170-171)^4 + (172-171)^4 + (168-171)^4 + (165-171)^4 + \cdots + (172-171)^4}{10}\right)}{(5.1)^3}$$

$$= -1.239$$

Se concluye que los datos son asimétrico, simétrica a su derecha. También, son platicúrticos con posible presencia de valores atípicos ([54])

2.3. MEDIDAS DE TENDENCIA CENTRAL DE DATOS NO AGRUPADOS Y AGRUPADOS

2.3.1. Media aritmética o promedio ([55])

Esta media es una medida de tendencia central que se utiliza ampliamente. Tiene varias propiedades ([56]):

2.3.1.1. Media muestral

- Todo conjunto de datos de nivel de intervalo tiene un valor medio. Por ejemplo, los datos de nivel de intervalo comprenden datos de edades, ingresos y pesos, siendo constante la distancia entre los números.
- Para evaluar la media se consideran todos los valores.
- Un conjunto de datos sólo tiene una media, siendo valor único.
- La media es una medida muy útil para comparar dos o más poblaciones. Por ejemplo, puede emplearse para comparar el trabajo en la producción de los operarios del primer turno

[54] Capa, S., H. a. 2015

[55] Su teorema demuestra que si $\{a_n\}_{n\geq a}$ es una sucesión convergente, pues $\lim_{n\to\infty}(a_n) = \lim_{n\to\infty}\left(\frac{a_1+a_2+a_3+a_4+\cdots+a_n}{n}\right)$.

[56] Aragón, S. L. G. 2016

de una planta de transmisiones Chevrolet con el desempeño laboral de los operarios del segundo, tercer, cuarto o más turnos.

- La media aritmética es la única medida de tendencia central donde la suma de las desviaciones de cada valor, respecto de la media, siempre es igual a 0. De forma simbólica esto es:

$\sum(X - \overline{X}) = 0$. Entonces, como ejemplo, la media de 3, 8 y 4 es 5: $\sum(X - \overline{X}) = (3 - 5) + (8 - 5) + (4 - 5) = -2 + 3 - 1 = 0$

- La media puede considerarse como un punto de equilibrio de un conjunto de datos. Ejemplo, supóngase que se tiene una barra rectangular larga, marcada con los números $1, 2, 3, \ldots, n$, espaciados uniformemente sobre la barra. Se colocan tres lingotes de oro de igual peso sobre la barra en los números 3, 4 y 8. El punto de equilibrio queda fijado en 5, que es la media de los tres números. Las desviaciones hacia debajo de la media (-3) son iguales a las de arriba (+3):

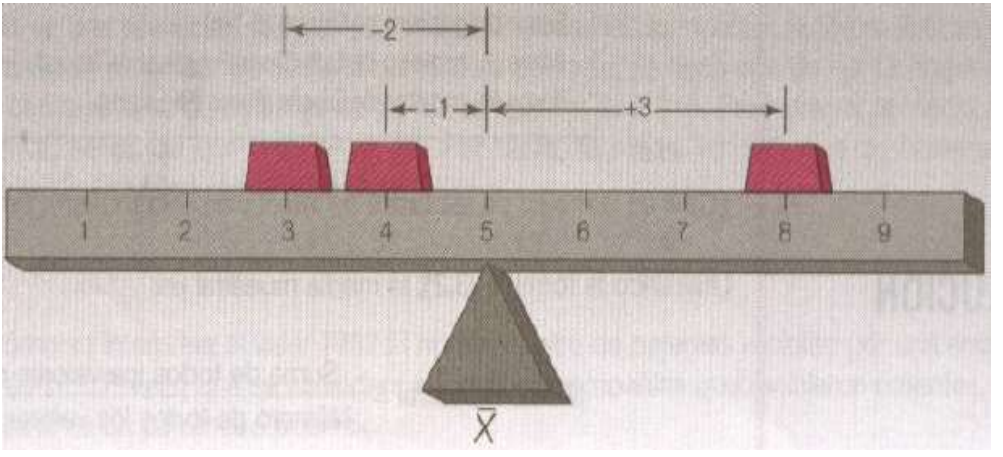

Figura 13. Representación gráfica de media muestral [57]

- La media es indebidamente afectada en forma notable por valores muy grandes o muy pequeños, pues para su cálculo se utiliza el valor de cada elemento de una muestra o una población. Ejemplo: supóngase que los ingresos anuales en dólares de un pequeño grupo de corredores de acciones de Merrill Lynch son 62,900, 61,600, 62,500, 60,800 y 1.2 millones. El ingreso medio sería de 289,560 USD, pero resulta obvio que no es representativo de este

[57] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

grupo porque todos, excepto un corredor (el de 1.2 millones de USD), tienen un ingreso en el intervalo de 60-63,000 USD. En consecuencia, el ingreso atípico afecta indebidamente a la media.

- No se puede determinar la media para datos con un extremo abierto. La media también es inadecuada si hay una clase de extremos abiertos en el caso de datos agrupados en una distribución de frecuencias. Si una distribución tiene una clase de extremo abierto de "100 000 y más", si hay 10 personas de esa clase, en realidad no se sabe si sus ingresos se aproximan a 100 000, 500 000 o 16 millones de USD. Como no se tiene información acerca de sus ingresos, no es posible determinar la media aritmética del ingreso para esta distribución de extremo abierto.

Según ([58]), con frecuencia se seleccionar una muestra de la población, pues se quiere evaluar algo acerca de una característica específica de tal población. Por ejemplo, un departamento de control de calidad necesita tener la seguridad que el diámetro exterior de cojines de bolas que están produciendo es aceptable. Entonces, podría seleccionarse una muestra de cinco cojines y estimar su diámetro promedio de todos los cojines producidos. Para datos a granel (no agrupados), la media es la suma de todos los valores, dividida entre el número total de los mismos:

$$\text{Media Muestral} = \frac{\text{Suma de todos los valores de la muestra}}{\text{Número de valores en la muestra}}$$

$$\overline{X}\ ([59]) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Cualquier característica medible de una muestra se denomina Dato Estadístico o Estimador (característica de una muestra). La media de una población es un estimador.

Ejemplo: La empresa Merrill Lynch Global se especializa en obligaciones a largo plazo de países extranjeros. Interesa saber la tasa de interés de estas obligaciones. Una muestra aleatoria de seis bonos reveló lo siguiente ([60]). ¿Esta información es una muestra o una población? y ¿Cuál es la media de tasas de interés de obligación es a largo plazo?:

[58] Aragón, S. L. G. 2016

[59] $\overline{X}$ simboliza la media muestral. Se lee "X con barra, X supra línea o media muestral". La letra n designa al número total de valores de la muestra

[60] Aragón, S. L. G. 2016

Cuadro 3. Tasa de interés (%) de obligaciones a largo plazo de empresa Merril Lynch Global ([61])

Artículo	Tasa de interés (%)
Bonos del gobierno de Australia	9.50
Bonos del gobierno de Bélgica	7.25
Bonos del gobierno de Canadá	6.50
Bonos del gobierno de Francia (B-Tan)	4.75
Bonos del gobierno de Italia (Buoni Poliennali de Tesora)	12.00
Bonos del gobierno de España (Bonos del Estado)	8.30

La media muestral se obtiene mediante la fórmula ([62]):

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{9.50 + 7.25 + 6.50 + \cdots + 8.30}{6} = 8.05\ \%$$

2.3.1.2. Media Poblacional

Según ([63]), la media se obtiene mediante la división de los datos o valores de una población entre el número total de datos. Para estimar la media de una población se utiliza la siguiente formula:

$$\text{Media Poblacional} = \frac{\text{Suma de todos los valores de la población}}{\text{Número de valores en la población}}$$

En vez de expresar con palabras las instrucciones completas para calcular la media poblacional, o cualquier otra media, es más conveniente usar símbolos matemáticos:

$$\mu\ ([64]) = \frac{\sum_{i=1}^{N} X_i}{N}$$

Ejemplo: Hay 12 empresas de fabricantes de autos en EUA. Enseguida se presenta el número de patentes otorgadas el año pasado por el gobierno de USA a cada negociación. ¿Esta información es una muestra o una población? y ¿Cuál es el número medio de patentes otorgadas?:

[61] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004
[62] Un conjunto de ejercicios se encuentra en carpeta "Media Muestral".
[63] Aragón, S. L. G. 2016
[64] μ es media de población (letra griega "mu" minúscula), N es número total de elementos de la población, X representa cualquier valor en particular, Σ indica la operación de suma (letra griega "sigma" mayúscula) y ΣX sumatoria de todos los valores X.

Cuadro 4. Número de patentes otorgadas a 12 fabricantes de autos de EUA ([65])

Empresa	N_o de patentes otorgadas
General Motors (GM)	511
Nissan	385
Daimier Chrysler	275
Toyota	257
Honda	249
Ford	234
Mazda	210
Chrysler	97
Porsche	50
Mitsubishi	36
Volvo	23
BMW	13

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{511 + 385 + 275 + 257 + \cdots + 13}{12} = \frac{2{,}340}{12} = 195$$

Para evaluar la media aritmética de datos organizados en una distribución de frecuencias, se considera que las observaciones en cada clase están representadas por el punto medio de clase. La media de una muestra de datos organizados en una distribución de frecuencias se estima así:

$$\overline{X}\ ([66]) = \frac{\sum fX}{n}$$

Sin embargo, la media de datos agrupados en una distribución de frecuencias puede ser diferente de la media de datos reales. El hecho de agrupar datos produce una pérdida de información.

Ejemplo ([67]):

- Las operaciones necesarias para calcular la media aritmética de datos agrupados en una distribución de frecuencias se mostrarán con base en precios de venta de los vehículos mostrados como datos en la siguiente tabla. Determine la media aritmética del precio de venta de los vehículos.

[65] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

[66] $\overline{X}$ es media aritmética, X es valor central o punto medio de clase, f es frecuencia de cada clase, fX frecuencia de cada clase multiplicada por el punto medio de la clase, Σ fX suma de esos productos y n número total de frecuencias

[67] Éste ejemplo calculado en Microsoft Excel se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubica en carpeta "Media Muestral"

Cuadro 5. Frecuencia de precio de ventas de vehículos ([68])

Precio de venta (miles de USD)	Frecuencia
12 hasta 15	8
15 hasta 18	23
18 hasta 21	17
21 hasta 24	18
24 hasta 27	8
27 hasta 30	4
30 hasta 33	2
Total	**80**

Los ingresos netos de una muestra de grandes importadores de antigüedades se organizaron en la siguiente tabla:

a) ¿Cómo se llama a este tipo de tabla?

Evalúe la media aritmética del ingreso neto basándose en la distribución.

2.3.1.3. Media aritmética ponderada

La media ponderada es un caso especial de la media común (media aritmética). Se presenta cuando hay varias observaciones con un mismo valor, que puede ocurrir si los datos se han agrupado en una distribución de frecuencias. En general, la media aritmética ponderada de un conjunto de números designados por X_1, X_2, X_3,…, X_n con las ponderaciones o "pesos" correspondientes W_1, W_2, W_3,…, W_n se calcula así:

$$\overline{X}_W = \frac{W_1X_1 + W_2X_2 + W_3X_3 + \cdots + W_nX_n}{W_1 + W_2 + W_3 + \cdots + W_n} = \frac{\sum_{i=1}^{n}(WX)}{\sum_{i=1}^{n} W}$$

Ejemplo ([69]):

a) En un restaurante se venden refrescos medianos, grandes y extragrandes. Sus precios son 0.90, 1.25 y 1.50 USD, respectivamente. De los últimos 10 refrescos que se vendieron 3 eran medianos, 4 grandes y 3 extragrandes.

b) La constructora Carter Construction Co paga a sus empleados 6.50, 7.50 o 8.50 USD/Hr. Hay 26 empleados contratados por hora: 14 reciben la tarifa de 6.50 USD, 10 reciben

[68] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

[69] Estos ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubica en carpeta "Media aritmética ponderada"

7.50 USD y 2 reciben 8.50 USD. ¿Cuál es la media de la tarifa por hora que se paga a los 26 trabajadores?

2.3.1.4. Media armónica H

La media armónica H de un conjunto de N números X_1, X_2, X_3,…, X_N es el recíproco de la media aritmética de los recíprocos de los números:

$$H = \frac{1}{\frac{1}{N}\sum_{j=1}^{N}\frac{1}{X_j}} = \frac{N}{\sum\frac{1}{X}}$$

Aunque en la práctica es más fácil:

$$\frac{1}{H} = \frac{\sum\frac{1}{X}}{N} = \frac{1}{N}\sum\frac{1}{X}$$

Ejemplo[70]: Estime la media armónica H de los números 2, 4 y 8.

$$H = \frac{N}{\sum\frac{1}{X}} = \frac{3}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}} = \frac{3}{\frac{7}{8}} = 3.4$$

2.3.1.5. Media geométrica G ([71])

La media geométrica es útil para encontrar el promedio de porcentajes, razones, índices o tasas de crecimiento. Por ejemplo, en negocios y economía, pues frecuentemente interesa determinar el cambio porcentual en ventas, sueldos o cifras económicas, como el Producto Interno Bruto (PIB). Tiene varias propiedades:

- La media geométrica será menor o igual que la media aritmética; es decir, nunca será mayor que la media aritmética.
- Todos los datos deben ser positivos para determinar la media geométrica.
- Un segundo uso de la media geométrica es encontrar aumentos porcentuales promedio en un intervalo de tiempo:

[70] Este ejemplo calculado en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubica en carpeta "Media armónica H"

[71] Su teorema demuestra que si $\{a_n\}_{n\geq a}$ es una sucesión convergente tal que $\lim_{n\to\infty}(a_n) = \lim_{n\to\infty}\left(\sqrt[n]{(a_1 * a_2 * a_3 * a_4 * \dots * a_n)}\right)$. Por ejemplo, si se estima $\lim_{n\to\infty}\left(\sqrt[n]{\left(\frac{4}{15} * \frac{7}{30} * \frac{12}{55} * \dots * \frac{n^2+3}{5n^2+10}\right)}\right) \Rightarrow \lim_{n\to\infty}(a_n) = \lim_{n\to\infty}\left(\frac{n^2+3}{5n^2+10}\right) \Rightarrow f'_{(x)} = \frac{2n+0}{10+0}$; $f''_{(x)} = \frac{2+0}{2*5+0} = \frac{1}{5}$ o, también, $\lim_{n\to\infty}(a_n) = \lim_{n\to\infty}\left(\frac{\frac{n^2}{n^2}+\frac{3}{n^2}}{\frac{5n^2}{n^2}+\frac{10}{n^2}}\right) = \lim_{n\to\infty}\left(\frac{1+\frac{3}{n^2}}{5+\frac{10}{n^2}}\right) \Rightarrow \lim_{n\to\infty}\left(\frac{1+\frac{3}{(\infty)^2}}{5+\frac{10}{(\infty)^2}}\right) = \lim_{n\to\infty}\left(\frac{1+0}{5+0}\right) = \left(\frac{1}{5}\right)$.

$$\mathrm{MG}_{(\text{Aumento \% Promedio en un Periodo Determinado})} = \sqrt[n]{\frac{\text{Valor al final del periodo}}{\text{Valor al inicio del periodo}}} - 1$$

- La media geométrica de un conjunto de n números positivos se define como la raíz n-ésima del producto de los n valores:

$$\text{Media Geométrica (MG)} = \sqrt[n]{(X_1)(X_2)\dots(X_n)}$$

Ejemplos ([72]):

- Un Ingeniero Agroindustrial recibe un aumento de sueldo del 5% este año y recibirá uno de 15% el siguiente año. El aumento porcentual promedio es 9.886 y no de 10.0%. El sueldo mensual del Ingeniero es de $3,000 USD. Explique el porqué. Verifíquelo.
- Las ganancias obtenidas por la Constructora Atkins en cuatro proyectos fueron 3, 2, 4 y 6%. ¿Cuál es la media geométrica de la ganancia? Explique el porqué. Verifíquelo.
- Un Ingeniero ganó $30,000 USD en 2005 y $50,000 USD en 2015. ¿Cuál es la tasa de aumento anual en el periodo?
- Supóngase que la población en Haarlan, Alaska, en 2000 era de 2 personas y en 2010 eran 22. ¿Cuál fue la tasa del incremento porcentual anual promedio para el periodo?
- Encuentre por método convencional y método de Log: Media geométrica y Media aritmética de los números 3, 5, 6, 6, 7, 10 y 12. Se supone que los números son exactos.

2.3.2. Mediana

Para datos que contienen uno o dos valores muy grandes o muy pequeños, la media aritmética puede ser no representativa. Es el valor que corresponde al punto medio de los valores después de ordenarlos de menor a mayor o mayor a menor; tal que 50% de las observaciones son menores y 50% de las mismas son mayores que la Mediana. Tiene varias propiedades:

- Es única. A semejanza de la media, sólo existe una mediana para un conjunto de datos.
- No se ve afectada por valores extremadamente grandes o muy pequeños. Por lo tanto, es una media valiosa de tendencia central cuando se presenta esta clase de valores.

[72] Éstos ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubican en carpeta "Media geométrica G"

- Puede calcularse para una distribución de frecuencias con una clase de extremo abierto.
- Puede calcularse para datos de nivel de razón, intervalo y ordinal, excepto para nivel nominal (por ejemplo, se puede ordenar por intervalos de menor a mayor, como en un estudio de mercado de una barra de chocolate las personas la consideran: “excelente”, “muy bien”, “bien”, "regular” y “mal”. Por lo tanto, la mediana es “bien”).

Ejemplo:

a) Una persona desea adquirir un condominio en Palm Aire. Su agente de bienes raíces le indicó que el precio promedio de las unidades disponibles en este momento es de $110,000 USD. Si tuviera un presupuesto máximo entre $60,000-70,000 USD podría pensar que está fuera de sus posibilidades. ¿De todas formas querría considerar lo anterior? Sin embargo, al verificar los precios individuales de los condominios podría cambiar de idea. Los precios son $60,000, $65,000, $70,000, $80,000 y un Penthouse muy lujoso cuesta $275,000 USD. La media aritmética del precio es $110,000 según indicó el agente de bienes raíces, pero un valor ($275,000 USD) está haciendo que la media aritmética se incline hacia arriba, por lo que es un promedio no representativo. Parecería que un precio entre $65,000-$75,000 USD es un precio más típico o representativo: Enseguida se muestran los rendimientos anuales totales de cinco años, de las seis acciones con mejor desempeño de fondos comunes de inversión con crecimiento dinámico ¿Cuál es la ganancia mediana anual?

Cuadro 6. Cálculo de mediana de rendimiento total anual (%) de condominio en Palm Aire ([73])

Nombre del fondo	Rendimiento Total Anual (%)
PBHG Growth	28.5
Dean Witter Developing Growth	17.2
AIM Aggressive Grown	25.4
Twentieth Century Giftrust	28.6
Robertson Stevens Emerging Growth	22.6
Seligman Frontier A	21.0

[73] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

Cuadro 7. Cálculo de mediana de rendimiento total anual (%) del fondo [74]

Nombre del fondo	Rendimiento Total Anual (%)	
PBHG Growth	28.5	
Dean Witter Developing Growth	17.2	
AIM Aggressive Grown	25.4	$\frac{48.0}{2} = 24.0\ \%$
Twentieth Century Giftrust	28.6	
Robertson Stevens Emerging Growth	22.6	
Seligman Frontier A	21.0	

Recuerde que la mediana, de datos agrupados, se define como el valor abajo del cual se encuentra la mitad de los valores y arriba del mismo la otra mitad. Dado que los datos sin agrupar se han organizado en una distribución de frecuencias, parte de la información ya no es indispensable. Como resultado, no es posible determinar la media exacta. Sin embargo, puede estimarse:

1) Localizando la clase en que se encuentra la mediana.

2) Realizando interpolaciones dentro de esta clase para obtener dicho valor. La razón de este enfoque es que se supone que los elementos de la clase en que se encuentra la mediana están espaciados de manera uniforme en toda la clase.

Su fórmula es:

$$\text{Mediana } ([75]) = \underline{\text{L}} + \left[\left(\frac{\text{n}}{2} - \text{FA}\right)(\text{i})\right]$$

La mediana sólo se basa en las frecuencias y los límites de la clase que la contienen. Las clases de extremo abierto que se presentan en los extremos rara vez se necesitan. En consecuencia, se podrá determinar la mediana de una distribución de frecuencias con una clase de extremos abiertos. La media aritmética de una distribución de frecuencias con una clase de extremo abierto no puede evaluarse de forma exacta, a menos que se estimen los puntos medios de las clases de ese tipo. Además, también se puede determinar la mediana si se tienen frecuencias porcentuales en lugar de frecuencias absolutas. Esto se debe a que la mediana el valor con un 50% de distribución por arriba y 50% por debajo de ella, así que

[74] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

[75] L es límite inferior de la clase que contiene a la mediana, n es el número total de frecuencias, f es frecuencia de la clase que contiene la mediana, FA es número acumulado de frecuencias en todas las clases que proceden a la clase que contiene a la mediana e i es amplitud (anchura) de la clase en que encuentra la mediana.

no depende de los conteos reales. Los porcentajes se consideran como sustitutos de las frecuencias verdaderas. En cierto sentido, son frecuencias absolutas cuyo total es 100.00.

Ejemplos ([76]):

b) Los datos que incluyen los precios de venta de los vehículos en la agencia Whitner Pontiac se utilizarán para mostrar el procedimiento a seguir para calcular la mediana. Las frecuencias acumuladas en la columna de la derecha se utilizarán en breve. ¿Cuál es la mediana del precio de venta de los vehículos nuevos en esta agencia?

Cuadro 8. Precios, número vendidos y frecuencia acumulada de precios de venta de vehículos en agencia Whitner Pontiac ([77])

Precio de venta (Miles de USD)	Número vendido (*f*)	Frecuencia acumulada (*FA*)
12 hasta 15	8	8
15 hasta 18	23	31
18 hasta 21	17	48
21 hasta 24	18	66
24 hasta 27	8	74
27 hasta 30	4	78
30 hasta 33	2	80
Total	**80**	**385**

2.3.3. Moda

Es el valor de la observación que aparece con más frecuencia. La moda es especialmente útil para describir los niveles de mediciones nominales y ordinales. Tiene varias propiedades:

- Se puede determinar la moda para datos de los niveles: nominal, ordinal, intervalo y razón.
- Tiene la ventaja de no verse afectada por valores extremadamente altos o bajos.
- Al igual que la mediana, puede usarse como medida de tendencia central en distribuciones con clases de extremo abierto.

[76] Este ejemplo calculado en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubican en carpeta "Mediana"
[77] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

➢ En muchos conjuntos de datos no hay valor modal porque ningún valor aparece más de una vez. Ejemplo: 19, 21, 23, 20 y 18. Cada valor es diferente, podría argumentarse que cada valor es modal.

➢ En ciertos conjuntos de datos hay más de una moda (Bimodal o dos modas). Ejemplo: 22, 26, 27, 27, 31, 35 y 35. Entonces, 27 y 35 son edades modales.

La moda de un conjunto de números es el valor que se presenta con más frecuencia; es decir, es el valor más frecuente Puede no haber moda, pero también puede no ser única. Para datos agrupados en una distribución de frecuencias, es posible aproximar la moda usando el punto medio de la clase que contiene el mayor número de frecuencias de clase. Dos valores pueden presentarse un número elevado de veces. Entonces se dice que la distribución es bimodal. Si el conjunto de datos tiene más de dos valores modales, la distribución se denomina multimodal. En estos casos probablemente no se consideraría ninguna de las modas como representativa del valor central de los datos.

Ejemplos ([78]):

a) Nivel Nominal. Una compañía ha desarrollado cinco aceites para baño. Enseguida se muestran los resultados de un estudio de mercado diseñado para descubrir la preferencia de los consumidores.

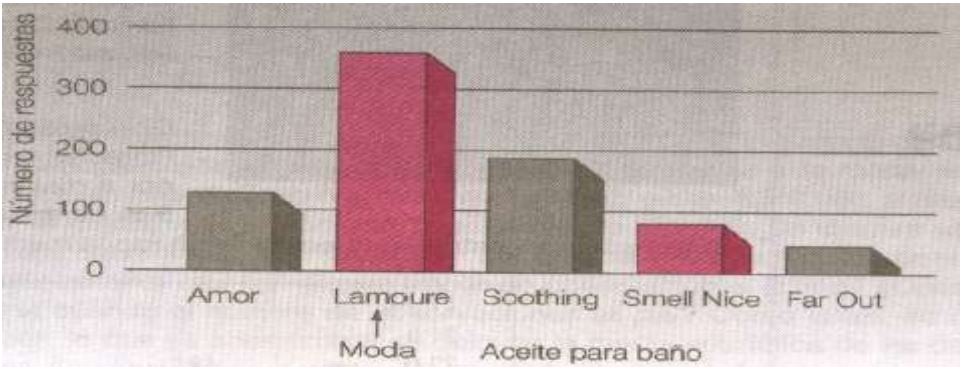

Figura 14. Representación de Moda de venta de cinco aceites para baño ([79])

Se muestran los sueldos anuales en dólares americanos de gerentes de control de calidad en algunos estados. ¿Cuál es su valor modal?

[78] Estos ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubica en carpeta "Moda"
[79] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

Cuadro 9. Sueldos anuales en USD por estados de EUA ([80])

Estado	Sueldo (USD)
Arizona	35,000
California	49,100
Colorado	60,000
Florida	60,000
Idaho	40,000
Illinois	58,000
Lousiana	60,000
Maryland	60,000
Massachusetts	40,000
New Jersey	65,000
Ohio	50,000
Tennessee	60,000
Texas	71,400
West Virginia	60,000
Wyoming	55,000

b) Las ventas netas de una muestra de pequeñas plantas de estampado se organizaron en la siguiente distribución de frecuencias porcentuales. ¿Cuál es la mediana y moda estimadas de las ventas netas?

Cuadro 10. Ventas netas de una muestra de pequeñas plantas de estampado ([81])

Ventas neta (Millones de USD)	Porcentaje total
1 hasta 4	13
4 hasta 7	14
7 hasta 10	40
10 hasta 13	23
13 y superior	10

c) En una distribución bimodal supóngase que las edades de una muestra de trabajadores son 22, 27, 30, 30, 30, 30, 34, 58, 60, 60, 60, 60 y 65.

d) Una muestra de la producción diaria de transmisores/receptores de comunicación marca Scott Electronics se organizó en la siguiente distribución. Calcule la mediana y la moda de la producción.

[80] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004
[81] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

Cuadro 11. Muestra de producción diario de transmisores/receptores ([82])

Producción diaria	Frecuencia
80 hasta 90	5
90 hasta 100	9
100 hasta 110	20
110 hasta 120	8
120 hasta 130	6
130 hasta 140	2

2.3.4. Relación Empírica entre Media, Mediana y Moda

En las curvas de frecuencias unimodales que son ligeramente sesgadas (asimétrica), se tiene la relación empírica:

$$\text{Media} - \text{Moda} = 3(\text{Media} - \text{Mediana})$$

Las posiciones relativas de la Media, Mediana y Moda en curvas de frecuencias sesgadas a la derecha o izquierda se muestran respectivamente:

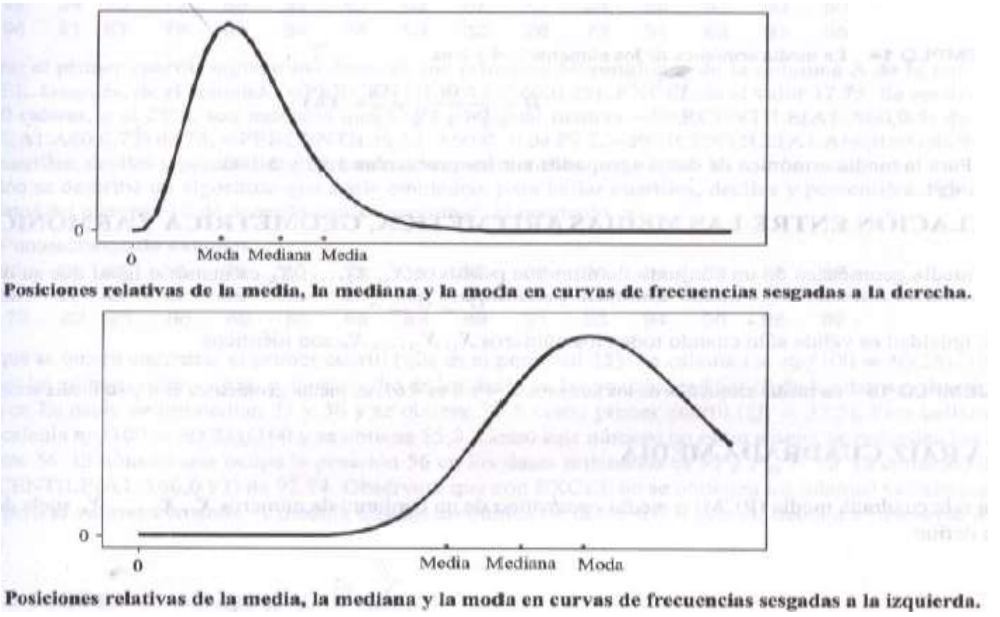

Figura 15. Representación empírica entre media, mediana y moda ([83])

Asimismo, en las curvas simétricas, la Media, Mediana y Moda coinciden.

2.3.5. Relación Empírica entre las Medias Aritmética, Geométrica y Armónica

Según Murray y Larry (2009), la media geométrica de un conjunto de números positivos X_1, X_2, X_3,…, X_N es menor o igual que su media aritmética, pero mayor o igual que su media

[82] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004
[83] Fuente: Murray R. S. y Larry J. S. 2009

armónica: $H \leq G \leq \overline{X}$. Tal que, la igualdad es válida sólo cuando los números X_1, X_2, X_3,…, X_N son idénticos.

Ejemplo: La media aritmética de los números 2, 4 y 8 es 4.67. Su media geométrica es 4 y su media armónica es 3.43.

2.3.6. Conjunto Cuantiles

En un conjunto de datos ordenados conforme a su magnitud, el valor medio (media aritmética de dos valores de en medio), que divide al conjunto en dos partes iguales es la mediana. Continuando con esta idea se puede pensar en aquellos valores que dividen al conjunto de datos en cuatro partes iguales. Estos valores, denotados por Q_1, Q_2 y Q_3 son el primero, segundo y tercer cuartiles, respectivamente. Tal que, el valor Q_2 coincide con la Mediana.

De igual manera, los valores que dividen al conjunto en diez partes iguales son deciles y se denotan D_1, D_2, D_3,…, D_9. Los valores que dividen al conjunto en 100 partes iguales son llamados percentiles, se denotan por P_1, P_2, P_3,…, P_{99}. Donde, el quinto decil y el percentil 50 coinciden con la Mediana. Los percentiles 25 y 75 coinciden con el primer y tercer cuartiles, respectivamente. A los cuartiles, deciles, percentiles y otros valores dividiendo al conjunto de datos en partes iguales se llama Conjunto Cuantiles.

Ejemplos ([84]):

a) A continuación, se presenta un conjunto de puntuaciones. Obtenga Q_1, Q_2, Q_3 y P_{99}.

b) Describa el algoritmo que suele emplearse para hallar cuartiles, deciles y percentiles.

2.4. MEDIDAS DE DISPERSIÓN DE DATOS NO AGRUPADOS Y AGRUPADOS

2.4.1. Dispersión o variación

Un promedio, como media o medina, solamente localiza el centro de los datos. Sin embargo, un promedio nada indica acerca de la diseminación de datos. El grado de dispersión de datos numéricos respecto a un valor promedio se llama dispersión o variación de datos. Existen varias medidas de dispersión (variación). Las más usadas son rango, desviación media, rango semi intercuartil, rango percentil 10-90 y desviación estándar. La desviación media, la varianza y la desviación estándar se basan en desviaciones respecto a la media.

[84] Estos ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta “Ejemplos” y, también, un conjunto de ejercicios se ubica en carpeta “Conjunto cuantiles”

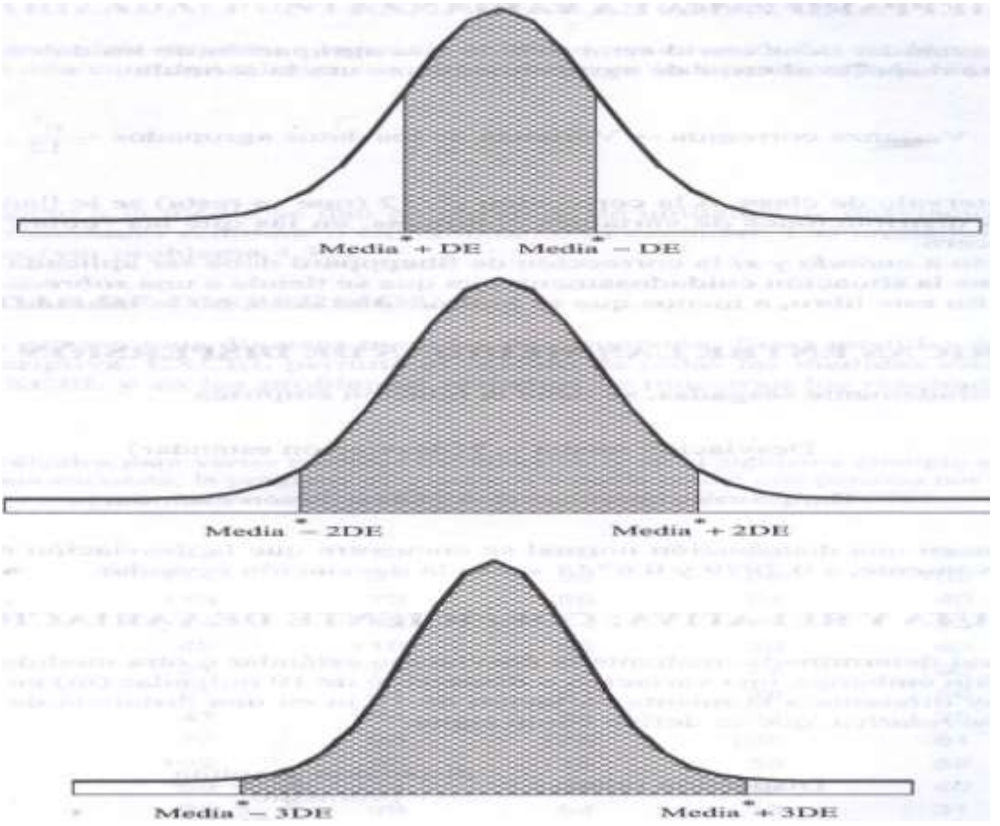

Figura 16. Representación de dispersión o variación de datos ([85])

Sin embargo, para las distribuciones moderadamente sesgadas, se tiene la siguiente relación empírica:

$$\text{Desviación media} = \frac{4}{5}(\text{Desviación estándar})$$

$$\text{Rango semi} - \text{intercuartil o desviación cuartil} = \frac{2}{3}(\text{Desviación estándar})$$

2.4.2. Rango o amplitud intercuartil (RIC o RIQ)

El rango o amplitud intercuartil (RIC) se obtiene como diferencia entre los cuartiles 1^o y 3^o: $\text{RIC} = \text{Q}_3 - \text{Q}_1$. Una variante del mismo, conocido como amplitud semi-intercuartil es:

$$\text{Q}_{(\text{RSIC})} = \frac{(\text{Q}_3 - \text{Q}_1)}{2}$$

[85] Fuente: www.google.com/search?client=firefox-b&biw=1366&bih=635&tbm=isch&sa=1&ei=9Xa0W-ekIMLk_AbGpq2wBA&q=Promedio+y+desviaciones+est%C3%A1ndar&oq=Promedio+y+desviaciones+est%C3%A1ndar&gs_l=img.3...57413.58192.0.59027.8.6.0.0.0.0.0.0..0.0....0...1c.1.64.img..8.0.0....0.X7Y_c-_uWhI#imgrc=H1Ja2zxMOOZAIM:

Ambos índices tienen como ventaja respecto al Rango que no se ven afectados por la existencia de valores atípicos en la variable, pues no se obtienen a partir de los dos valores más extremos de la variable, sino a partir de dos valores más centrado como son el Q_3 y Q_1.

En una distribución normal se encuentra que la desviación media y el rango semi-intercuartil son iguales, respectivamente, a 0.7979 y 0.6745 vez la desviación estándar. Sin embargo, un valor pequeño en una medida de dispersión indica que los datos se acumulan estrechamente, por ejemplo, alrededor de la media aritmética. En consecuencia, el valor medio se considera representativo de los datos. Por el contrario, una medida de dispersión grande indica que la media no es confiable. Una segunda razón para estudiar la dispersión de un conjunto de datos es comparar la dispersión en dos o más distribuciones.

Ejemplos ([86]):

a) Si una guía geográfica informa que el cauce de un río tiene en promedio 3 pies de profundidad, ¿Usted lo cruzaría sin tener información adicional? Probablemente no. Desearía saber algo acerca de la variación de la profundidad. ¿Es la profundidad máxima del río 3. 25 y la mínima 2.75 pies, respectivamente? Si es el caso, probablemente decidirá cruzar. ¿Qué ocurriría si se entera que la profundidad del río varía de 0.50 a 5.5 pies? Su decisión probablemente sería no atravesarlo. Antes de decidir si cruza o no el río, usted requiere información acerca de la profundidad típica y la variación en la profundidad de este.

b) Se han organizado los datos de 100 empleados de Struthers & Wells Inc., una compañía fabricante de acero, en un histograma basado en el número de años que han sido empleados de la misma. La media es 4.9 años, pero la variabilidad de datos va desde 6 meses a 16.8 años. El valor medio, 4.9 años, no es muy representativo de todos los empleados.

[86] Estos ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubica en carpeta "RIC o RIQ"

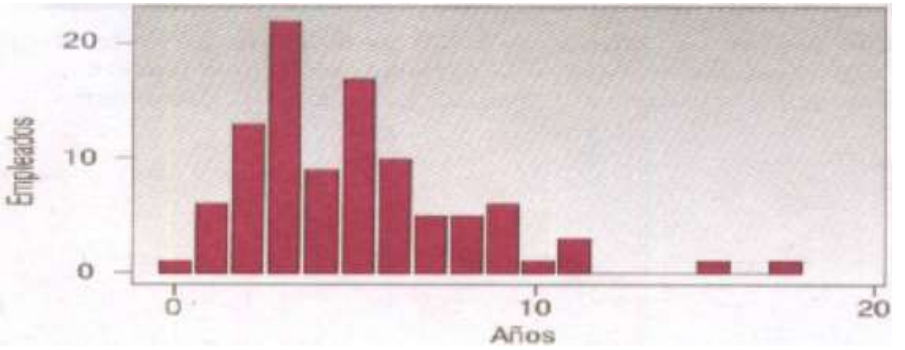

Figura 17. Histograma de años de servicio en Struthers & Well, Inc. ([87])

c) Supóngase que la nueva computadora i9 se ensambla en Baton Rouge y, también, en Tucson, Arizona. La media aritmética de la producción diaria en la planta de Baton Rouge es 50 y, también, en la de Tucson es igual. Con base en ambos valores medios se podría concluir que las distribuciones de las producciones diarias son idénticas. No obstante, los registros de producción de 9 días en las dos plantas revelan que esta conclusión es incorrecta. La producción de Baton Rouge varía de 48 a 52 ensambles/día, pero la producción en Tucson es más errática, pues varía de 40 a 60 ensambles/día.

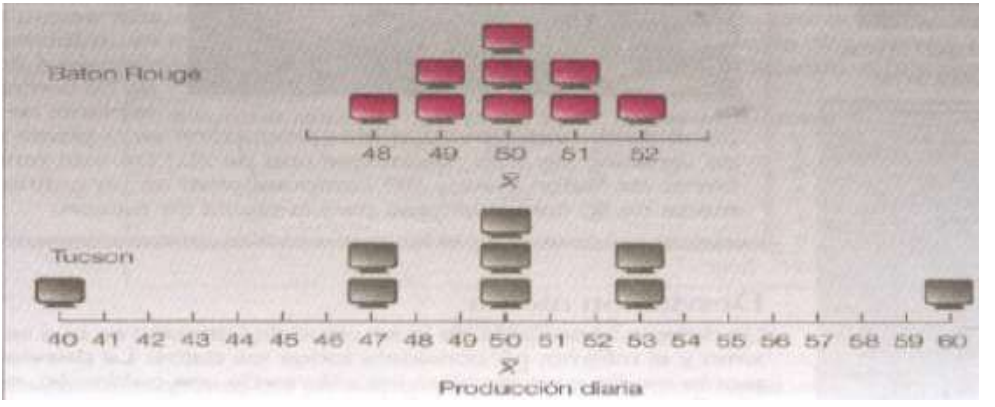

Figura 18. Producción diaria de computadoras en plantas ([88])

2.4.3. Rango (Range)

El rango de un conjunto de números es la diferencia entre el número mayor y el número menor del conjunto. La medida de dispersión más sencilla es la amplitud de rango (Amplitud de variación = valor más grande − Valor más pequeño), pues en inglés se

[87] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004
[88] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

utiliza el término range para designar la amplitud de variación o "rango". Asimismo, el término Rank, si equivale a rango, pero con el significado de jerarquía o grado.

Desventajas:

- No utiliza todas las observaciones (sólo dos).
- Se puede ver muy afectada por alguna observación extrema (dato atípico).
- El rango aumenta con el número de observaciones o, bien se queda igual. En cualquier caso, nunca disminuye.

Ejemplos ([89]):

a) Se tienen los siguientes costos de producción de yogurt: 50, 60, 80 y 120. El rango de los costos es igual a $120 - 50 = 70$. El 100% de los costos se encuentra distribuido a una distancia de 70 nuevos USD.

b) El rango del conjunto 2, 3, 3, 5, 5, 5, 8, 10, 12 es $12 - 2 = 10$.

c) Algunas veces el rango se da mediante el número menor y el número mayor. En el caso anterior, simplemente se indica de 2 a 12 o $2 - 12$.

d) La amplitud de variación en la producción diaria de computadoras en la planta Baton Rouge es 4 computadoras (52−48=4). La amplitud de variación de la producción diaria en la planta Tucson es 20 computadoras (60−40=20). Se puede concluir que: 1) Hay menos dispersión en la producción diaria de la planta Baton Rouge que en la de Tucson y 2) la producción en la planta Baton Rouge se acumula más cerca de la media 50, que la producción de la planta Tucson. Entonces, la producción media en la planta Baton Rouge es un promedio más representativo que la media de 50 unidades para la planta Tucson.

2.4.4. Desviación media $(\mathbf{D_m})$

Un defecto importante de la amplitud de variación o rango es que se basa sólo en dos valores (máximo y mínimo), pues no considera todos los datos. Caso contrario, la desviación media sí lo hace, mide el monto medio en que varían los valores de una población o muestra respecto a su media. Desviación media es el promedio aritmético de valores absolutos de las desviaciones con respecto a la media aritmética. Es importante mencionar que esta desviación no considera los signos de las desviaciones respecto a la media, pues caso

[89] Éstos ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubican en carpeta "Range"

contrario las desviaciones + y - se compensarían y la desviación media siempre sería = 0, por lo que sería un valor estadístico inútil y es difícil trabajar con valores absolutos, así que no se usa frecuentemente. Como se consideran desviaciones absolutas. Entonces:

$$D_m = \frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|$$

Ventajas de la Desviación Promedio:

- Utiliza en su cálculo todos los valores en la muestra.
- Es fácil de comprender, pues representa el promedio en que los valores se desvían con respecto a la media. No obstante, su principal desventaja es el uso de valores absolutos. En consecuencia, la desviación media no se usa con la misma frecuencia que las otras medidas de dispersión, como es el caso de la desviación estándar.

La desviación media, desviación media absoluta o desviación promedio de un conjunto de N números $X_1,\ X_2, \dots, X_N$ se abrevia DM y se define como:

$$\text{Desviación Media (DM)}^{90} = \frac{\sum_{j=1}^{N}|X_j - \overline{X}|}{N} = \frac{\sum|X - \overline{X}|}{N} = |X_j - \overline{X}|$$

Ejemplos[91]:

a) Suponga que desea hallar la desviación media absoluta de los siguientes ingresos por ventas mensuales durante un semestre: 100, 200, 300, 400, 500 y 700.

b) Encuentre la desviación media del conjunto 2, 3, 6, 8 y 11 ($\overline{X} = 6$ y $\text{DM} = 2.8$).

c) El número de pacientes atendidos en sala de urgencias del Hospital IESS, Quito, Ecuador para una muestra de 5 días al año fue: 103, 97, 101, 106 y 103. Determine e interprete la desviación media.

Si $X_1,\ X_2, \dots, X_k$ se presentan con frecuencias $f_1, f_2, \dots, f_k$, respectivamente. La desviación media puede expresarse de la siguiente manera:

[90] $\overline{X}$ es media aritmética de los números y $|X_j - \overline{X}|$ es valor absoluto de la desviación de X_j respecto de $\overline{X}$ (valor absoluto de un número es el número sin signo). El valor absoluto de un número se indica por medio de dos barras verticales colocadas a los lados del número, así $|-4| = 4,\ |+3| = 3,\ |6| = 6$ y $|-0.84| = 0.84$

[91] Estos ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubican en carpeta "D_m"

$$DM^{92} = \frac{\sum_{j=1}^{K} f_j \left|X_j - \overline{X}\right|}{N} = \frac{\sum f|X - \overline{X}|}{N} = \left|X_j - \overline{X}\right|$$

En ocasiones, la desviación media se define en términos de las desviaciones absolutas respecto de la mediana o de otro promedio y no respecto a la media. Una propiedad interesante de la suma $\sum_{j=1}^{N}\left|X_j - a\right|$ es que es mínima cuando a es la mediana (la desviación media absoluta con respecto a la mediana es un mínimo).

2.4.5. Varianza

Cuando se establecieron los sistemas de medición fue necesario establecer una "referencia", medir por qué hay diferencias y, con ellas, la variabilidad en un conjunto de datos, pues no hay diferencias en un solo dato y, tampoco, existe variabilidad. La principal utilidad de la varianza es que sirve para tomar decisiones con información incompleta (muestra) conociendo la probabilidad que la decisión sea acertada, como si un elemento no pertenece al mismo grupo que otros (experimento), un valor es mejor que aquel otro (experimentos) y si el valor real se encuentra entre dos límites (muestreo).

2.4.5.1. Poblacional $(\boldsymbol{\sigma^2})$ y Muestral $(\mathbf{s^2})$

La varianza de un conjunto de datos poblacionales se define como el cuadrado de la desviación estándar y, por lo tanto, corresponde al valor σ^2 de las ecuaciones $\sigma = \sqrt{\frac{\sum_{j=1}^{N}(X_j - \overline{X})^2}{N}} = \sqrt{\frac{\sum(X-\overline{X})^2}{N}} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{(X - \overline{X})^2}$ y $\sigma = \sqrt{\frac{\sum_{j=1}^{K} f_j (X_j - \overline{X})^2}{N}} = \sqrt{\frac{\sum f(X-\overline{X})^2}{N}} = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{(X - \overline{X})^2}$. Asimismo, la varianza poblacional de datos no agrupados (tabulados en una distribución de frecuencias) se obtiene $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$. Cuando es necesario distinguir la desviación estándar de una población de la desviación estándar de una muestra obtenida de esa población, se suele emplear σ y s, respectivamente. En consecuencia, σ^2 y s^2 representan la varianzas poblacional y muestral, correspondientemente.

En el último caso, la conversión de la varianza poblacional a varianza muestral no es tan directa, pues debe hacerse una ligera modificación en el denominador. En lugar de

[92] $N = \sum_{j=1}^{K} f_j = \sum f$. Esta fórmula es útil para datos agrupados, donde las X_j representan las marcas de clase y las f_j las frecuencias correspondientes. Asimismo, es más apropiado usar el término desviación media absoluta en vez de desviación media

introducir n (número en la muestra) en vez de N (número en población), el denominador se hace igual a $n-1$. La varianza muestral se estima con:

$$s^2 = \frac{\sum(X - \overline{X})^2}{n-1} \text{ o } s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$$

El cambio en el denominador se debe a que n tiende a subestimar la varianza de la población, σ^2. El uso de $n-1$ en el denominador proporciona la corrección adecuada para esta tendencia.

Ejemplos[93]:

a) Las edades de los pacientes del pabellón de aislados en el Hospital IESS de Quito son 38, 26, 13, 41 y 22 años. ¿Cuál es la varianza de esa población?

b) La oficina en Filadelfia de la empresa Price Waterhouse Coopers LLP contrató a cinco pasantes de contabilidad este año. Sus sueldos mensuales iniciales fueron (USD): $2,536, $2,173, $2,448, $2,121 y $2,622 USD. Estime:

1. Calcule la media de la población.
2. Determine la varianza.
3. Obtenga la desviación estándar poblacional.
4. La oficina de Pittsburgh contrató a 6 pasantes. Su sueldo mensual promedio fue de $2,550 USD y la desviación estándar $250. Compare ambos grupos.

c) Los salarios por hora en una muestra de operarios de medio tiempo en la empresa Fruit Packers, Inc., son (USD): $2, $10, $6, $8 y $9. ¿Cuál es la varianza muestral?

2.4.5.2. Corrección de Sheppard para varianza

El cálculo de la desviación estándar tiene cierto error debido a la agrupación de los datos en clases (error de agrupamiento). Para hacer un ajuste respecto al error de agrupamiento se usa la fórmula:

$$\text{Varianza corregida} = \text{Varianza de datos agrupados} - \frac{c^2_{(\text{Tamaño del intervalo de clase})}}{12}$$

Entonces, a la corrección $c^2/12$ (se resta) se le llama Corrección de Sheppard. Esta corrección se usa para distribuciones de variables continuas, en las que las "colas", en ambas

[93] Estos ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubican en carpeta "Varianza Poblacional y Muestral"

direcciones, se aproximan gradualmente a cero. No obstante, hay discrepancia a cuándo y si la corrección de Sheppard debe ser aplicada. Desde luego no debe aplicarse antes de que se examine la situación cuidadosamente, pues se tiene a una sobrecorrección, sólo se sustituye un error por otro.

2.4.6. Desviación estándar

2.4.6.1. Poblacional $(\boldsymbol{\sigma})$ y Muestral $(\mathbf{s})$

Propiedades de la Desviación Estándar:

A. Se puede definir como: $\sigma = \sqrt{\frac{\sum_{j=1}^{N}(X_j - a)^2}{N}}$. En el caso de la Desviación Estándar Poblacional es $\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$.

Donde a es un promedio cualquiera además de la media aritmética. De todas las desviaciones estándar, la mínima es aquella en la que $a = \overline{X}$, pues "en un conjunto de números X_j, la suma de los cuadrados de desviaciones respecto a un número a es un mínimo si y sólo sí $a = \overline{X}$". Demostración:

Probar que $w^2 + pw + q$, donde p y q con constantes dadas, es mínimo si y sólo sí $w = -\frac{1}{2}p$. Entonces, $w^2 + pw + q = (w + \frac{1}{2}p)^2 - \frac{1}{4}p^2 + q$. Como $-\frac{1}{4}p^2 + q$ es constante, esta expresión tiene su mínimo valor si y sólo sí $w + \frac{1}{2}p = 0$ $(w = -\frac{1}{2}p)$. Empleando el anterior inciso probar que $\frac{\sum_{j=1}^{N}(X_j - a)^2}{N}$ o $\frac{\sum(X-a)^2}{N}$ es mínimo si y sólo sí $a = \overline{X}$. Entonces, $\frac{\sum(X-a)^2}{N} = \frac{\sum(X^2 - 2aX + a^2)}{N} = \frac{\sum X^2 - 2a\sum X + Na^2}{N} = a^2 - 2a\frac{\sum X}{N} + \frac{\sum X^2}{N}$. Comparando esta última expresión con $(w^2 + pw + q)$, se tiene: $w = a$, $p = -2\frac{\sum X}{N}$ y $q = \frac{\sum X^2}{N}$. Por lo tanto, la expresión tiene un mínimo en $a = -\frac{1}{2}p = \frac{\sum X}{N} = \overline{X}$.

B. En las distribuciones normales, moderadamente sesgadas, se encuentra que los siguientes porcentajes se satisfacen de manera aproximada:

a) 68.27% de los casos está comprendido entre $\bar{x} \pm s$ (una desviación estándar a cada lado de la media).

b) 95.45% de los casos está comprendido entre $\bar{x} \pm 2s$ (dos desviaciones estándar a cada lado de la media).

c) 99.73% de los casos está comprendido entre $\bar{x} \pm 3s$ (tres desviaciones estándar a cada lado de la media).

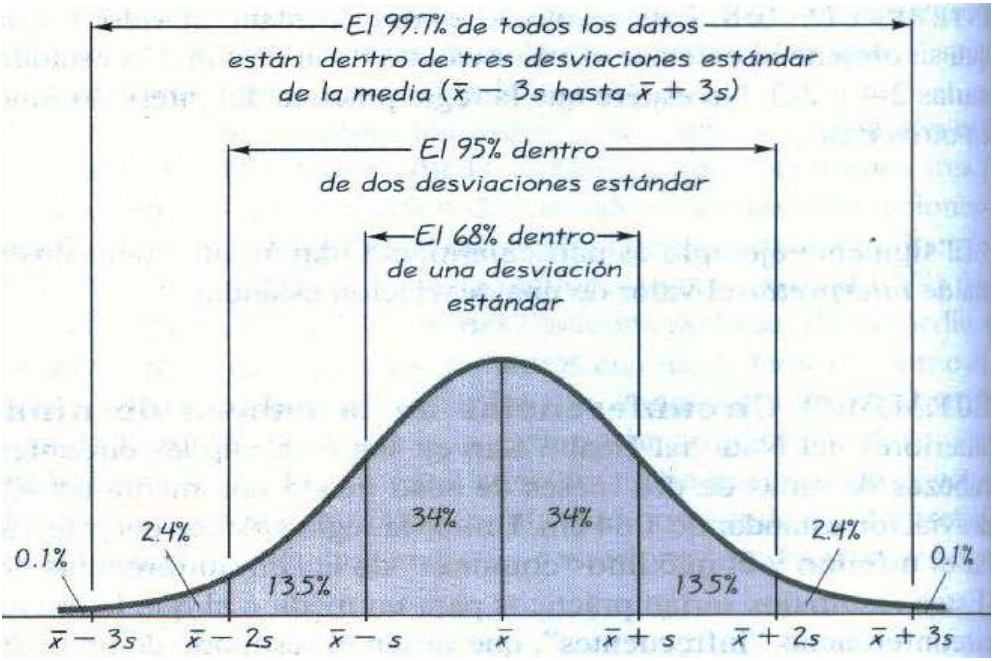

Figura 19. Nivel de probabilidad por número de desviaciones estándar muestrales([94])

C. Supóngase que dos conjuntos que constan de N_1 y N_2 números (dos distribuciones de frecuencia con frecuencias totales N_1 y N_2) tienen varianzas σ_1^2 y σ_2^2, respectivamente, así como una misma media $\bar{X}$. Entonces, la varianza combinada o conjunta de los dos conjuntos o de las dos distribuciones de frecuencia está dada por:

$$\sigma^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2}$$

Obsérvese que ésta es una media aritmética ponderada de las varianzas. Esta fórmula puede generalizarse a tres o más conjuntos.

D. El matemático ruso Pafnuti Lvóvich Chebyshev (1821-1894) desarrolló un teorema que permite determinar la proporción mínima de los valores que se encuentran dentro de un

[94] Fuente: www.google.com/search?client=firefox-b&biw=1366&bih=635&tbm=isch&sa=1&ei=9Xa0W-ekIMLk_AbGpq2wBA&q=Promedio+y+desviaciones+est%C3%A1ndar&oq=Promedio+y+desviaciones+est%C3%A1ndar&gs_l=img.3...57413.58192.0.59027.8.6.0.0.0.0.0.0..0.0....0...1c.1.64.img..8.0.0....0.X7Y_c-_uWhI#imgrc=fe44E-gbjywjyM:

número específico de desviaciones estándar con respecto a la media. “Para un conjunto de observaciones (muestra o población), la proporción mínima de valores que se encuentran dentro de k desviaciones estándar desde la media es por lo menos $1-\frac{1}{k^2}$, donde k es una constante mayor que 1”. Entonces, el teorema de Chebyshev establece que para $k>1$, por lo menos $\left(1-\left(\frac{1}{k^2}\right)\right)*100\%$ de la distribución de probabilidad de cualquier variable está a no más de k desviaciones estándar de la media. En particular, para $k=2$, por lo menos $\left(1-\left(\frac{1}{2^2}\right)\right)*100\%$ o bien 75% de los datos está en el intervalo $\overline{X}\pm 2s$, para $k=3$, por lo menos $\left(1-\left(\frac{1}{3^2}\right)\right)*100\%$ u 89% de los datos está en el intervalo $\overline{X}\pm 3s$ y para $k=4$, por lo menos $\left(1-\left(\frac{1}{4^2}\right)\right)*100\%$ o 93.75% de los datos está en el intervalo $\overline{X}\pm 4s$.

La desviación estándar de un conjunto de N números $X_1,\ X_2,\ldots,X_N$ se denota como σ y está definida por $\sigma=\sqrt{\frac{\sum_{j=1}^{N}(X_j-\overline{X})^2}{N}}=\sqrt{\frac{\sum(X-\overline{X})^2}{N}}=\sqrt{\frac{\sum x^2}{N}}=\sqrt{(X-\overline{X})^2}$. Donde x representa la desviación de cada uno de los números X_j respecto a la media $\overline{X}$. Por lo tanto, s es la raíz cuadrada de la media (RCM) de las desviaciones de la media o, como suele llamársele algunas veces, la desviación raíz-media-cuadrado. Si $X_1,\ X_2,\ldots,X_N$ se presentan con frecuencias $f_1,f_2,\ldots,f_k$, respectivamente, la desviación estándar se puede expresar como:

$$\sigma=\sqrt{\frac{\sum_{j=1}^{K}f_j\left(X_j-\overline{X}\right)^2}{N}}=\sqrt{\frac{\sum f(X-\overline{X})^2}{N}}=\sqrt{\frac{\sum fx^2}{N}}=\sqrt{(X-\overline{X})^2}$$

Donde $N=\sum_{j=1}^{K}f_j=\sum f$. Esta fórmula es útil para datos agrupados. Algunas veces la desviación estándar de una muestra se define usando como el denominador, en las ecuaciones $\sigma=\sqrt{\frac{\sum_{j=1}^{N}(X_j-\overline{X})^2}{N}}=\sqrt{\frac{\sum(X-\overline{X})^2}{N}}=\sqrt{\frac{\sum x^2}{N}}=\sqrt{(X-\overline{X})^2}$ y $\sigma=\sqrt{\frac{\sum_{j=1}^{K}f_j(X_j-\overline{X})^2}{N}}=\sqrt{\frac{\sum f(X-\overline{X})^2}{N}}=\sqrt{\frac{\sum fx^2}{N}}=\sqrt{(X-\overline{X})^2}$, $(N-1)$ en vez de N. Esto se debe a que el valor que así se obtiene es una mejor aproximación a una desviación estándar de la población de la que se ha tomado la muestra. Con valores grandes de $N(N>30)$, prácticamente no hay diferencia

entre las dos definiciones. Cuando se necesita una estimación mejor, ésta siempre se puede obtener multiplicando por $\sqrt{\frac{N}{(N-1)}}$ la desviación estándar obtenida según con la primera definición. El Método Abreviado para el Cálculo de la Desviación Estándar es[95] $\sigma = \sqrt{\frac{\sum_{j=1}^{N}(X_j-\overline{X})^2}{N}} = \sqrt{\frac{\sum(X-\overline{X})^2}{N}} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{(X-\overline{X})^2}$ y $\sigma = \sqrt{\frac{\sum_{j=1}^{K} f_j(X_j-\overline{X})^2}{N}} = \sqrt{\frac{\sum f(X-\overline{X})^2}{N}} = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{(X-\overline{X})^2}$ se pueden expresar, respectivamente, mediante las siguientes fórmulas:

$$\sigma = \sqrt{\frac{\sum_{j=1}^{N} X_j^2}{N} - \left(\frac{\sum_{j=1}^{N} X_j}{N}\right)^2} = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \overline{X}^2}$$

$$\sigma = \sqrt{\frac{\sum_{j=1}^{K} f_j X_j^2}{N} - \left(\frac{\sum_{j=1}^{K} f_j X_j}{N}\right)^2} = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\overline{X^2} - \overline{X}^2}$$

La Desviación Estándar Muestral se utiliza como un estimador de la desviación estándar poblacional. Es la raíz cuadrada de la varianza muestral. En caso de datos no agrupados se estima como sigue:

$$s = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}} \text{ o } s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}} \text{ (Formula directa)}$$

Si los datos que interesan están en forma agrupada (distribución de frecuencias), la desviación estándar muestral puede aproximarse al sustituir $\sum X^2$ por $\sum fX^2$ y $\sum X$ por $\sum fX$:

$$s = \sqrt{\frac{\sum fX^2 - \frac{(\sum fX)^2}{n}}{n-1}}$$

[95] Donde $\overline{X^2}$ representa la media de los cuadrados de los diversos valores de X, en tanto que $\overline{X}^2$ denota el cuadrado de la media de los diversos valores de X. Existen ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubica en carpeta "Desviación Estándar Poblacional – Muestral"

2.4.7. Coeficiente de variación $(\mathbf{CV}\ 96)$

La variación o dispersión real determinada mediante la desviación estándar u otra medida de dispersión se le conoce como dispersión absoluta. Sin embargo, una variación o dispersión de 10" (in o pulgadas) en una distancia de 1000 ft (pies) tiene un significado muy diferente a la misma variación de 10 in en 20 ft. Este efecto se puede medir mediante la dispersión relativa, definida como:

$$\text{Dispersión relativa} = \frac{\text{Dispersión absoluta}}{\text{Promedio}}$$

Si la dispersión absoluta es la desviación estándar (s) y el promedio es la media $(\overline{\mathrm{X}})$; entonces, la dispersión relativa es el Coeficiente de Variación o Coeficiente de Dispersión (V):

$$\text{Coeficiente de Variación Poblacional } (\mathrm{CV_P}) = \frac{\sigma}{\mu} * 100(\%)$$

$$\text{Coeficiente de Variación Muestral } (\mathrm{CV_m}) = \frac{\mathrm{s}}{\overline{\mathrm{X}}} * 100(\%)$$

Es importante mencionar que, el coeficiente de variación es independiente de las unidades que se empleen y, por esto, es útil cuando se trata de comparar distribuciones en las que las unidades son diferentes. Una desventaja del coeficiente de variación es que no es útil cuando el valor $\overline{\mathrm{X}}$ es cercano a 0.

2.4.8. Relaciones empíricas entre medias de dispersión

Un promedio, como media o medina, solamente localiza el centro de los datos. Sin embargo, un promedio nada indica acerca de la diseminación de datos. Para las distribuciones moderadamente sesgadas, se tiene la siguiente relación empírica:

$$\text{Desviación media} = \frac{4}{5}(\text{Desviación estándar})$$

$$\text{Rango semi} - \text{intercuartil o desviación cuartil} = \frac{2}{3}(\text{Desviación estándar})$$

[96] Existen ejemplos calculados en Microsoft Excel, R, SAS, Wolfram, Matlab y Texas Instrument se encuentra en carpeta "Ejemplos" y, también, un conjunto de ejercicios se ubica en carpeta "CV"

3. PROBABILIDAD

3.1. HISTORIA

En la naturaleza y en la vida cotidiana se presentan fenómenos con resultados determinados de forma anticipada mediante la aplicación de ciertas leyes o fórmulas. Por ejemplo: resultados de mediciones geométricas, cálculos financieros o ciertos procesos físicos. Asimismo, existen fenómenos con resultados puede no ser anticipado con certeza, sino que existe una probabilidad que un cierto resultado se dé. Por lo tanto, nadie puede dar un resultado certero con anticipación a eventos considerados tal que, si se da una respuesta, existe incertidumbre en su resultado. Con base en esto, la teoría de probabilidades fue desarrollada con el fin de dar una explicación matemática a resultados que aparecen relacionados con el azar.

En casi todas las culturas antiguas es posible hallar referencias que indican el estudio de fenómenos aleatorios fue importante, como dados, presencia de lluvia, clima, entre otras. Según ([97]), en épocas del renacimiento, hubo un abandono progresivo de explicaciones teológicas que condujo a una reconsideración de experimentos de resultado incierto y, con ello, matemáticos italianos del siglo XVI interpretaron resultados de experimentos aleatorios simples. Por ejemplo, Gerolamo Cardano o Girolamo Cardano en 1526 estableció, por condiciones de simetría, la equiprobabilidad de aparición de caras de un dado. También, Galileo Galilei publicó en un tratado llamado Considerazoine sopra il giuoco dei dadi por qué es más difícil obtener valor 9 tirando 3 dados que obtener 10, pues de las 216 combinaciones posibles equiprobables, 25 conducen a 9 y 27 al número 10.

El desarrollo del análisis matemático de juegos de azar se produjo en siglos XVI y XVII. Algunos autores consideran su origen en el cálculo de probabilidades la resolución del problema de puntos en correspondencia entre Blaise Pascal y Pierre de Fermat. En siglo XVIII, el cálculo de probabilidades se extendió a problemas físicos y seguros marítimos. El factor principal de su desarrollo fue el conjunto de problemas de astronomía y física que surgieron ligados a la constatación empírica de la teoría de Isaac Newton.

[97] RAE (2018: dle.rae.es/srv/fetch?id=ZVAt4Ig): Ciencia que trata de Dios, de sus atributos y perfecciones. Del griego θεος o theos que significa "dios" y λογος o logos que expresa "estudio" o "razonamiento", en consecuencia, significa el estudio de dios y de hechos relacionados con él.

Pierre-Simón Laplace introdujo la primera definición explícita de probabilidad y desarrolló la ley normal como modelo para describir la variabilidad de errores de medida. También hubo importantes contribuciones de matemáticos, como Adrien-Marie Legendre y Johann Carl Friedrich Gauss para hacer predicciones del comportamiento de ciertos fenómenos. Durante el siglo XIX, matemáticos y astrónomos continuaron ampliando la teoría, tal que a mediados del mismo existían herramientas que permitieron su consolidación como una rama científica. Sin embargo, su aplicación se restringía a física y astronomía.

Una descripción axiomática de probabilidad fue dada por Andréi Nikoláyevich Kolmogórov que contribuyo como la base de la moderna teoría. Se consiguió elaborar modelos complejos y aplicar las probabilidades a muchas ciencias y campos de la vida. Actualmente, su empleo en las ciencias naturales, sociales, ingeniería, cálculo actuarial o economía creció ampliamente.

3.2. ANÁLISIS COMBINATORIO

El factorial es de un entero positivo n, la factorial de n o n factorial se define en principio como el producto de todos los números enteros positivos desde 1; es decir, los números naturales hasta n ([98]). En términos más sencillos, el factorial de un número entero positivo n es el producto expresado como $n! = n * (n-1) * \dots * 2 * 1$, con $0! = 1$. Considere un conjunto finito compuesto por n elementos diferentes, como $\{a_1, a_2, a_3, a_4, \dots, a_n\}$. Desea formar una colección constituida por k elementos $(k \leq n)$. El número de subconjuntos depende de si los conjuntos son ordenados o no. Las colecciones ordenadas se llaman variaciones (a cada uno de los arreglos ordenados de k elementos, tomados de otro de n elementos, $k \leq n$, tal que estos arreglos difieren en algún elemento u orden de colocación) y las no ordenadas combinaciones (a cada uno de los subconjuntos de k elementos, tomados de otro n elementos , $k \leq n$, tal que sin tener en cuenta el orden de los mismos, no puede haber dos combinaciones con igual elementos).

[98] Se puede simbolizar $n! = \prod_{k=1}^{n} k$ y, también, es posible definirlo mediante la relación de recurrencia $n! = \begin{cases} 1, & \text{si } n = 0; \\ (n-1)! * n, & \text{si } n > 0; \end{cases}$

El número de variaciones de k elementos que pueden combinarse a partir de un conjunto de n elementos es $V_n^k = \frac{n!}{(n-k)!}$, mientras que el número de combinaciones de k elementos que pueden obtenerse de un conjunto de n elementos es $C_{n\ (\text{Coeficiente Binomial})}^k = \frac{n!}{k!*(n-k)!}$.

Ejemplo: Halle el número de variaciones y combinaciones de dos elementos que pueden obtenerse a partir del conjunto $\{a_1, a_2, a_3\}$. Entonces, $n = 3, k = 2$ tal que se puede formar $V_3^2 = \frac{3!}{(3-2)!} = 6$ variaciones: (a_1, a_2), (a_2, a_1), (a_2, a_3), (a_3, a_1), (a_2, a_3) y (a_3, a_2). Por otra parte, se pueden formar $C_{3\ (\text{Coeficiente Binomial})}^2 = \frac{3!}{2!*(3-2)!} = 3$ combinaciones: (a_1, a_2), (a_1, a_3) y (a_2, a_3).

La permutación de n elementos es cada una de las variaciones de n elementos distintos. Su número de permutaciones se calcula mediante $P_n = n!$. **Por ejemplo**: encuentre las permutaciones que se forman con base en el conjunto $\{a_1, a_2, a_3\}$. Sea $P_3 = 3! = 6$ permutaciones: (a_1, a_2, a_3), (a_1, a_3, a_2), (a_3, a_1, a_2), (a_3, a_2, a_1), (a_2, a_3, a_1) y (a_2, a_1, a_3). Si se considera la permutación con repetición de k elementos obtenidos a partir de un conjunto de n elementos es un arreglo de k elementos ordenados en que los elementos se repiten arbitrariamente y, por lo tanto, se consideran en arreglos múltiples del mismo conjunto. Se calcula mediante $P_n^k = n^k$. **Por ejemplo**: con elementos del conjunto $A = (a_1, a_2, a_3)$, ¿cuántas permutaciones con repetición, de dos elementos, se pueden formar? Se forman parejas considerando dos veces el conjunto A tal que se tiene $n = 3$ y $k = 2$. Existe un total de $n^k = 3^2 = 9$ permutaciones con repetición: (a_1, a_1), (a_1, a_2), (a_1, a_3), (a_2, a_1), (a_2, a_2), (a_2, a_3), (a_3, a_1), (a_3, a_2) y (a_3, a_3).

Considere dos conjuntos con m y n elementos: $A = \{a_1, a_2, a_3, a_4, \dots, a_m\}$ y $D = \{d_1, d_2, d_3, d_4, \dots, d_n\}$. Las parejas con m elementos de A y n elementos de D es posible formar $m * n$ parejas (a_i, d_k) que contengan un elemento de cada conjunto. **Por ejemplo**: en una fábrica de calzado se confeccionan 4 modelos de zapatos para damas, en 6 tamaños diferentes. Entonces, se puede fabricar $4 * 6 = 24$ diferentes tipos de zapatos. Este concepto se generaliza mediante arreglos múltiples, considera conjuntos $A = \{a_1, a_2, a_3, a_4, \dots, a_m\}$ de m y $D = \{d_1, d_2, d_3, d_4, \dots, d_n\}$ de n elementos, respectivamente, hasta $O =$

$\{o_1, o_2, o_3, o_4, \dots, o_s\}$ de s elementos. Por lo tanto, es posible formar $m * n * \dots * s$ arreglos $(a_i, d_j, \dots, o_r)$ que contiene un elemento de cada conjunto. Sin embargo, existe otra manera es considerar un procedimiento A a ejecutarse de m formas, un procedimiento D de n maneras, hasta un procedimiento O tal que se puede efectuar $m * n * \dots * s$ modos diferentes. **Por ejemplo**: se clasifica un grupo de estudiantes universitarios según su sexo (masculino o femenino), estado civil (soltero, casado o divorciado) y carrera que estudian (ingeniería agroindustrial, alimentos, agronomía, mecánica agrícola, irrigación, matemática, civil y electrónica). Por lo tanto, existe un total de $2 * 3 * 7 = 42$ clasificaciones diferentes.

3.3. EVENTOS Y ESPACIOS MUESTRALES

Evento (ω) es cualquier resultado posible de un experimento u otras situaciones que involucre incertidumbre.

Espacio Muestral es la colección de todos los elementos elementales, denotado por $\Omega = \{\omega/\omega\}$ es un evento elemental. Entonces, un evento no es más que un subconjunto del espacio muestral Ω. El concepto de espacio muestral fue introducido por Galileo para resolver el problema de porqué en lanzamiento de 3 dados, "10" y "11" aparece más frecuentes que "9" y "12" ([99]).

3.4. ELEMENTOS BÁSICOS

3.4.1. Definiciones.

Las teorías matemáticas, específicamente relacionadas con fenómenos económicos o naturales, se construyen generalmente a partir de conceptos intuitivos, claros, para que puedan aplicarse en las primeras formulaciones teóricas, pero no lo suficientemente rigurosos tal que reciben objeciones cuando logran cierto desarrollo. La etapa siguiente es revisar los fundamentos con el fin de elaborar una construcción axiomática. En vez de iniciar con una formulación axiomática, es mejor comenzar con definiciones tal vez no muy exactas y con ejemplos simples, substanciales, para comprender luego el verdadero sentido de los axiomas ([100]). Con base en esto, se presentan las siguientes definiciones:

- Probabilidad provee una descripción cuantitativa de posibilidad de ocurrencia de un evento particular y puede pensar que es su frecuencia relativa, en una serie

[99] Capa, S. H. 2015 b
[100] Capa, S. H. 2015 a

larga de repeticiones de una prueba, en que uno de los resultados es el evento de interés. Probabilidad de Laplace es la razón entre el número de casos favorables y número total de casos posibles (eventos elementales), siempre que todos tengan la misma probabilidad.

Ejemplos:

1) Se considera el lanzamiento de un dado, cuyo $\Omega = \{1,2,3,4,5,6\}$. El "evento A es sale 5" o $A = \{5\}$. Si el dado está bien construido y su lanzamiento se hace completamente al azar, no hay nada que obligue a creer que una de las caras tenga que salir de preferencia a las demás; por lo tanto, las seis caras son igualmente po1sibles:

$$P(A) = \frac{\text{Número de caos favorables}}{\text{Número de casos posibles}} = \frac{1}{6}$$

Si en vez de un dado cúbico, se supone que el mismo se alarga según una de sus dimensiones hasta formar un prisma recto de base cuadrada y altura mayor que lados de la base, las seis caras dejan de ser "igualmente probables". Es más probable que salga una cara lateral que una base y esta probabilidad se incrementa al aumentar la altura del prisma.

2) Se considera una urna con 10 bolas blancas y 5 rojas. Si se saca una bola al azar, ¿cuál es la probabilidad que salga roja? $\Omega = \left\{ \begin{matrix} 1,2,3,4,5,6\ldots 10 \\ 10 \text{ bolas blancas} \end{matrix} \quad \begin{matrix} 1,2,3,,\ldots 5 \\ 5 \text{ bolas rojas} \end{matrix} \right\}>$

$$P(A) = \frac{\text{Número de caos favorables}}{\text{Número de casos posibles}} = \frac{\cancel{5}}{(\cancel{5} * 3)} = \frac{1}{3}$$

Este ejemplo supone que todas las bolas son igualmente posibles. Si hubiera algunas con más probabilidad de salir que otras, como si tuvieran diferentes tamaños o pesos, la probabilidad sería diferente. Sin embargo, existen casos, como los anteriores, en que el cumplimiento o no de "igualmente probables" es obvio, mientras que otras veces este hecho pasa inadvertido y requiere mucha atención, así como "buen sentido" para evidenciarlo.

Con base en lo anterior, aun cuando un evento A no sea elemental, se puede aplicar la definición anterior, teniendo cuidado de contar cada evento elemental tantas veces como posibilidades verificativas A existan.

3) Se lanza una moneda 2 veces, ¿Cuál es la probabilidad que se obtenga cara por lo menos una vez? Podría creerse que los casos posibles son: 2 caras, una vez cara-una vez sello –sol- y 2 veces sello –sol-. Si se cuentan 2 casos favorables y 3 casos posibles, entonces es

equivalente a $\frac{2}{3}$. Aunque, se tiene $\Omega = \{(C, C), (C, S), (S, C), (S, S)\}$ tal que evento A: "sale al menos una cara" es $A = \{(C, C), (C, S), (S, C)\}$ y $P(A) = \frac{3}{4}$.

Laplace supone que el número de casos favorables y, por tanto, casos posibles es finito. Entonces, la probabilidad de un evento es siempre un $\mathbb{R}$ en $[0, 1]$ donde la probabilidad 0 indica que no hay ningún caso favorable; es decir, el suceso es "improbable". La probabilidad 1 indica que el número de casos favorables es igual al número de casos posibles o, en otras palabras, el suceso es "seguro" ([101]).

- Experimento aleatorio es un experimento en que no se conoce con certeza su resultado, como lanzamiento de una moneda dos ocasiones.
- Expacio muestral es un conjunto de todos los resultados posibles de un experimento aleatorio y se representa mediante $\Omega_{(\text{Letra griega omega})}$; por ejemplo: en lanzamiento de un dado $\Omega = \{1,2,3,4,5,6\}$ o, en lanzamiento de una moneda, $\Omega = \{\text{aguila} - \text{sello} -, \text{sol} - \text{cara} -\}$. De manera formal, la probabilidad de un evento A se define como una funcion que cumple: A_1. Para $\forall$ A: $0 \leq Pr(A) \leq 1$, A_2. $Pr(\Omega) = 1$ y A_3. Si A y B son incompatibles su $Pr(A \cup B) = Pr(A) + Pr(B)$. Entonces, se cumple la relación $Pr(A \cup B)_{(\text{Fórmula de probabilidad para unión})} = Pr(A) + Pr(B) - Pr(A \cap B)$. **Ejemplos**:

1) Dados eventos A, B y C del espacio muestral Ω. Exprese mediante operaciones entre conjuntos eventos:

a) Tan sólo ocurre A: Puede ocurrir A, simultáneamente no ocurre B ni C. Es decir, $x \in A \cap B^C \cap C^C$.

Si ocurre A, no ocurre B: Si no ocurre B entonces sucede B^C o, también, "si ocurre A, también sucede B^C" tal que $x \in A \subset B^C$.

b) Por lo menos uno de los dos eventos sucede: Suceden $(A - B)$ o $(A - C)$ o $(B - C)$ o $(A - B - C)$ aunque este último está contenido en los 3 primeros. El resultado es $x \in (A \cap B) \cup (A \cap C) \cup (B \cap C)$

[101] Capa, S. H. 2015 b

2) Demuestre que ([102]):

a) $\Pr(A^C) = 1 - \Pr(A)$: Sea $\Omega = A \cup A^C$, con A y A^C disjuntos ([103]) tal que por A_3. Si A y B son incompatibles su $\Pr(A \cup B) = \Pr(A) + \Pr(B)$, $\Pr(\Omega) = \Pr(A) + \Pr(A^C)$ tal que, por A_2. $\Pr(\Omega) = 1$. Por lo tanto, se obtiene $1 = \Pr(A) + \Pr(A^C)$ siendo el resultado inmediato.

b) Si $A \subset B$ entonces $\Pr(A) \leq \Pr(B)$: Donde $B = A \cup (A^C \cap B)$ incompatibles. Por A_3. Si A y B son incompatibles su $\Pr(A \cup B) = \Pr(A) + \Pr(B)$, entonces $\Pr(B) = \Pr(A) + \Pr(A^C \cap B)$ tal que por A_1. Para $\forall A: 0 \leq \Pr(A^C \cap B) \leq \Pr(A)$ y $\Pr(A) \leq \Pr(B)$.

De igual forma, existen definiciones importantes:

1. Dos eventos son "igualmente probables" si $\Pr(A) = \Pr(B)$.
2. Evento "A es más probable que B" si $\Pr(A) > \Pr(B)$.
3. Evento cierto o seguro (Ω) aparece en realización de un experimento, su probabilidad es 1.
4. Evento imposible $(\emptyset)$ es aquel que jamás puede suceder, su probabilidad es 0.
5. Sucede A o B: $A \cup B$.
6. Ocurre A y sucede B: $A \cap B$.
7. No sucede A: $\overline{A} = A^C = \frac{\Omega}{A}$.
8. Eventos incompatibles $(A \cap B = \emptyset)$.

➢ Evento o suceso es todo subconjunto de Ω tal que un evento A es elemental si es un subconjunto unitario.

3.4.2. Axiomas

Se conoce por probabilidad de un evento A a un $\mathbb{R}$ $P(A)$ tal que su funcion real P cumple con:

i) $P(A) \geq 0, \forall A \subseteq \Omega$.

ii) Si $A_1, A_2, A_3, A_4, \ldots, A_n$ son sucesos incompatibles, en relación dos a dos, implica que $P(\bigcup_i A_i) = \sum_i P(A_i)$.

iii) $P(\Omega) = 1$.

[102] Capa, S. H. 2015 a

[103] El conjunto Ø se denomina **conjunto vacío** y se entiende que es el conjunto que no contiene elementos. Una afirmación equivalente sería decir que $\mathbb{E}$ y $\mathbb{S}$ son **disjuntos**; es decir, si no tienen ningún elemento en común. Equivalentemente, dos conjuntos son disjuntos si su intersección es vacía. Por ejemplo, {1, 2, 3} y {a, b, c}.

Las consecuencias de estos axiomas es que definen la probabilidad y mediante aplicación de simples relaciones de teoría de conjuntos, se puede deducir las siguientes propiedades sin recurrir a su demostración:

1. $\Pr(\emptyset) = 0$.
2. $\Pr(A^C) = 1 - \Pr(A)$.
3. Si $A \subset B$ entonces $\Pr(A) \leq \Pr(B)$ y $\Pr((B|A) = \Pr(B) - \Pr(A)$.
4. Si $A \subset \Omega$ entonces $\Pr(A) \leq 1$.
5. Para 2 eventos A y B en Ω: $(A \cup B) = \Pr(A) + \Pr(B) - (A \cap B)$.

Enseguida se presenta como estos axiomas y sus correspondencias se relacionan con la definición clásica de probabilidad. Suponga que Ω se compone de un número finito de eventos elementales, como $A_1, A_2, A_3, A_4, \ldots, A_n$ incompatibles en relación dos a dos. Es decir, $\Omega = A_1 \cup A_2 \cup A_3 \cup A_4 \cup \ldots \cup A_n, A_i \cap A_j \neq \emptyset \; \forall \; i \neq j$. Para axiomas ii) y iii):

$$\Pr(\Omega) = P(A_1) + P(A_2) + P(A_3) + P(A_4) + \cdots + P(A_n) = 1$$

Además, supone que $\forall \, A_i$ tienen la misma probabilidad o, por definición clásica, "igualmente probables", se tiene $P(A_1) = P(A_2) = P(A_3) = P(A_4) = \cdots = P(A_n)$; por lo que, $P_i = \frac{1}{n}$. Sea un evento A_j formado por unión de m eventos (A_i) o, en otras palabras, $A = A_{i1} \cup A_{i2} \cup A_{i3} \cup A_{i4} \cup \ldots \cup A_{in}, A_{im}$. La probabilidad de A será:

$$\Pr(A) = mP(A_i) = \frac{m}{n}$$

Con esta ecuación se obtiene el resultado de definición clásica.

Ejemplos:

A. Se lanzan dos dados al azar y se requiere hallar la probabilidad que la suma de los puntos sea igual a 10.

La solución es que el espacio muestral es un conjunto de 36 pares ordenados (i, j) tal que i puede tomar valores de 1, 2, 3, 4, 5 y 6 del primer dado, mientras que j toma iguales valores respecto al segundo. Se supone que estos dados no están cargados; es decir, son legales. La probabilidad que cada cara salga y, paralelamente, cada par (i, j) es la misma, cualquiera que sea el par $(\frac{1}{36})$ y considerando que la suma de probabilidades es 1. El problema, hallar la probabilidad que el suceso $A = \{(i, j); i + j = 10\}$ tal que la probabilidad del suceso formado por pares $(4, 6), (5, 5)$ y $(6, 4)$. Existen 3 casos favorales, con igual

probabilidad, por axioma "si $A_1, A_2, A_3, A_4, \dots, A_n$ son sucesos incompatibles, en relación dos a dos, implica que $P(\cup_i A_i) = \sum_i P(A_i)$" equivale a $P(A) = \frac{3*1}{3*12} = \frac{1}{12}$.

B. En reunión de x personas $(x > 1)$, ¿cuál es la probabilidad que, por lo menos dos de ellas, cumplan años el mismo día, aunque no el mismo número de años?

No se considera la posibilidad que alguien haya nacido un día específico, como 21 de mayo, tal que supone que el año tiene 365 días. El espacio muestral se compone de posibles conjuntos de x fechas con 365^x elementos con igual probabilidad $(\frac{1}{365^x})$. Se busca la probabilidad que ningún par de personas cumpla años un día. La primera persona tiene 365 posibilidades para nacer, la segunda, no habiendo nacido el mismo día que la primera, tiene 364 posibilidades, la tercera tiene 363 posibilidades, la cuarta tiene 362, sucesivamente hasta la última persona, con $365 - (x - 1)$ posibilidades. Con base en esto, el evento opuesto consta de $365 * 364 * 363 * 362 * \dots * [365 - (x - 1)]$ elementos y su probabilidad es igual a este número dividido por 365^x. Entonces, la probabilidad buscada es $P_r = 1 - \left[\frac{365*364*363*362*\dots*(365-(x-1))}{365^x}\right]$. Estimando algunos posibles resultados se obtiene:

Cuadro 12. Probabilidad estimada por número de personas

Número de personas	Probabilidad estimada P_r
5	0.027
10	0.117
20	0.411
23	0.507
30	0.706
40	0.890
60	0.027

Se considera el número $x = 23$, pues la probabilidad es casi $\frac{1}{2}$. Además, si las personas son 60 o más, la probabilidad es superior a 99%; es decir, existe casi certeza que por lo menos dos personas cumplan años el mismo día.

3.4.3. Probabilidad condicional

Sea B un evento tal que $P(B) \neq 0$. Donde probabilidad condicional de un evento A, dado B, es $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Entonces, $P(A \cap B) = P(A|B) * P(B)$. De forma análoga, dado un suceso A y si $P(A) > 0$ se tiene probabilidad condicional

$P(B|A) = \frac{P(B \cap A)}{P(A)} \Leftrightarrow P(A \cap B) = P(B|A) * P(A)$. Para interpretar la definición de probabilidad condicional se considera el lanzamiento de un dado, cuyos eventos $A = \{2\}$ y $A = \{2,4,6\}$, ¿cómo se modifica la probabilidad a priori $P(A)$, si tiene información que ha sucedido B, salió un número par? El razonamiento directo indica que existen 3 casos posibles, uno favorable, tal que $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3}$. Sin embargo, si se aplica la definición equivale a $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$. Entonces, la probabilidad condicional será la probabilidad modificada con información que sucedió B, probabilidad a posteriori de A. Con base en esto, dos eventos A y B son independientes si $P(A|B) = P(A)$ tal que equivale a decir que dos eventos son independientes si $P(A \cap B) = P(A) * P(B)$ e implica que $P(B|A) = P(B)$.

Ejemplos:

A) Una urna contiene 10 bolas rojas y 5 blancas. Se extraen dos unidades y se desea conocer la probabilidad que las dos sean rojas si las extracciones se hacen con y sin reposición.

Se considera eventos A "primera bola roja" y B "segunda bola roja". Se quiere calcular $P(A \cap B)$. En caso con reposición, ambas extracciones son independientes e igual condiciones $P_{(\text{con reposición})} = P(A \cap B) = P(A) * P(B) = \left(\frac{5}{15}\right) * \left(\frac{5}{15}\right) = \left(\frac{1}{9}\right)$.

B) No existe independencia en resultados de ambas extracciones, se obtiene:

$$P_{(\text{sin reposición})} = P(A \cap B) = P\big((B|A)\big) * P(A) = \left(\frac{4}{14}\right) * \left(\frac{5}{15}\right) = \left(\frac{2}{21}\right)$$

La definición de independencia se generaliza a más de dos eventos mediante Fórmula de Bayes: varios eventos $A_1, A_2, A_3, A_4, \dots, A_n$ son independientes si se comprueba $P(A_{i1} \cap A_{i2} \cap A_{i3} \cap A_{i4} \dots A_{ik}) = P(A_{i1})P(A_{i2})P(A_{i3})P(A_{i4}) \dots P(A_{ik})$ para $k = 1, 2,3,4 \dots, n$ donde $(i_1, i_2, i_3, i_4, \dots, i_k)$ son combinación de cualquiera de n números $1, 2,3,4 \dots, n$. Por ejemplo, para que 3 eventos A, B y C sean independientes se cumplirá $P(A \cap B) = P(A) * P(B)$, $P(A \cap C) = P(A) * P(C)$, $P(B \cap C) = P(B) * P(C)$ y $P(A \cap B \cap C) = P(A) * P(B) * P(C)$ tal que las cuatro condiciones son necesarias, pues esta última no se deducirá de las tres primeras.

3.5. PROBABILIDADES EN ESPACIOS MUESTRALES

3.5.1. Finito

Si se considera el evento $A = \{\omega_1, \omega_2, \omega_3, \omega_4, .., \omega_k\}$ su probabilidad está complemente determinada si se conoce los valores en cada elemento $P_r(\{\omega_1\}), P_r(\{\omega_2\}), P_r(\{\omega_3\}), P_r(\{\omega_4\}), .., P_r(\{\omega_k\})$. Por lo tanto, $P_r(A) = \sum_{i=1}^{k} P_r(\{\omega_i\})$. Un caso particular en importancia es cuando todas las $P_r(\{\omega\})$ son iguales. Si se asigna $Card(A)$ el número k de elementos del $\{A\}$ y $Card(\Omega)$ en número N de elementos de espacio muestral; entonces $P_r(A) = \left(\frac{\text{Casos favorables A}}{\text{Casos posibles}}\right) = \left(\frac{Card(A)}{Card(\Omega)}\right) = \left(\frac{k}{N}\right)$. Es decir, la probabilidad de un evento aleatorio A es igual a relación entre número de eventos elementales favorables (cuando A sucede) y el número total de eventos elementales del espacio muestral. Esta definición es aplicable en área agroindustrial con experimentos sencillos.

Ejemplos:

1) En laboratorio de microbiología existen dos libros de $\text{Metodología Microbiológica}$ y 3 libros de $\text{Microbiología General}$. Al azar se toma un libro y, después, un segundo. Halle la probabilidad que un libro de $\text{Microbiología General}$ sea seleccionado a) la primera ocasión, b) ambas ocasiones.

Con base en espacio muestral $\Omega = \{MM_1, MM_2, MG_1, MG_2, MG_3\}$ tal que sea A el evento de escoger un MG o $A = \{MG_1, MG_2, MG_3\}$. Por lo tanto, $P_r(A) = \left(\frac{\text{Casos favorables A}}{\text{Casos posibles}}\right) = \left(\frac{Card(A)}{Card(\Omega)}\right) = \left(\frac{3}{5}\right)$. Finalmente, el evento en que se seleccione dos veces un MG tal que la primera elección es un MG, entonces se tienen 3 casos aceptables. La segunda elección implica que sea un MG tal que hay dos casos admisibles. En consecuencia, el número de casos favorables es $3 * 2 = 6$. El número de casos posibles respecto al número total de parejas sin repetir es $5 * 4 = 20$. Por lo tanto, la probabilidad estimada será $p = \left(\frac{3*\cancel{2}}{10*\cancel{2}}\right) = \left(\frac{3}{10}\right) = 0.30$.

2) Entre 100 frascos de encurtidos de una caja se halla un frasco buscado respecto a características visuales, como coloración. De la caja aleatoriamente se extrae 10 unidades. Estime la probabilidad que entre estas unidades se halle el frasco de encurtido deseable.

Según esta información, Ω está formado por conjuntos de 10 unidades, que pueden formarse a partir de 100. $Card(\Omega) = \binom{10}{100} = C_{100}^{10}$ tal que el número de resultados aceptables es la

manera en cómo pueden seleccionar 9 frascos de los 99 restantes ($\mathrm{Card}(\mathrm{A}) = \binom{9}{99} = \mathrm{C}_{99}^{9}$).

Por lo tanto, $\mathrm{P_r}(\mathrm{A}) = \left(\frac{\binom{9}{99}}{\binom{10}{100}}\right) = \left(\frac{\mathrm{C}_{99}^{9}}{\mathrm{C}_{100}^{10}}\right) = \left(\frac{1}{10}\right)$.

3.5.2. Infinito numerable

Si $\mathrm{A} = \{\omega_1, \omega_2, \omega_3, \omega_4, .., \omega_\mathrm{n}\}$ es un espacio muestral infinito numerable tal que $\sum_{\mathrm{i}=1}^{\infty} \mathrm{P_r}(\{\omega_\mathrm{i}\}) = 1$. Si A es un evento de A, su probabilidad se estima mediante $\mathrm{P_r}(\mathrm{A}) = \sum_{\omega_\mathrm{i} \in \mathrm{A}} \mathrm{P_r}(\{\omega_\mathrm{i}\})$. En el cálculo de probabilidades se usan series numéricas infinitas.

Ejemplo:

1) En una muestra de control de calidad de bebidas gaseosas de 30 unidades, existen 18 unidades que cumplen con NTE ([104]) del INEN ([105]) concentración respecto a 15 mg de sodio, 16 unidades cumplen con el parámetro de la norma de 25 gr de carbohidratos totales y 6 unidades que no cumplen con ninguna de las anteriores. Se elige al azar una unidad del producto gaseoso elaborado. Estime que probabilidad existe de que la bebida cumpla con los dos parámetros asignados, sabiendo que cumple con la concentración de 15 mg de sodio, cuál es la probabilidad de que cumpla con los requerimientos para los carbohidratos totales y ¿son independientes los sucesos de cumplimiento de esta NTE-INEN para la bebida gaseosa?

Con base en esta información, sean sucesos E, cumpla respecto a concentración de 15 mg de sodio y C, cumpla respecto a concentración de 25 gr de carbohidratos totales, organizando los datos en una tabla de doble entrada tal que P(E∩C) = $\frac{10}{30} = \frac{1}{3}$, P(C/E) = $\frac{P(C \cap E)}{P(E)} = \frac{10}{18} = \frac{5}{9}$, son independientes sí P(E ∩ C) = P(E).P(C) tal que P(E) = $\frac{18}{30} = \frac{3}{5}$, P(C) = $\frac{16}{30} = \frac{8}{15}$, P(E).P(C) = $\frac{3}{5} \times \frac{8}{15}$ = $\frac{24}{75} = \frac{8}{25}$ y $\frac{1}{3}$ = P(E ∩ C) ≠ P(E).P(C) =$\frac{8}{25}$ no son independientes.

3.5.3. Continuo

Suponga que tiene una figura plana Ω tal que en ella se ubica una figura A. Sobre figura Ω se marca un punto a al azar. La probabilidad que el punto se localice en A es proporcional al área de la figura y no de su forma o posición. Entonces, esta probabilidad es $\mathrm{P_r}(\mathrm{A}) = \left(\frac{\text{Área A}}{\text{Área } \Omega}\right)$.

[104] Norma Técnica Ecuatoriana (NTE): apps.normalizacion.gob.ec/descarga/

[105] Instituto Ecuatoriano de Normalización (INEN): www.normalizacion.gob.ec/

Por lo tanto, si A es un evento de espacio muestral continuo Ω tal que su unidad de medida (espacio, peso, masa, distancia, amplitud-magnitud-módulo-norma, volumen, tiempo, etcétera).

Ejemplo:

1) Según la teoría de "Bolas abiertas y conjuntos abiertos" se trazan dos n-bola abierta de radio r y centro a $(B(a,r))$ tal que sus circunferencias concéntricas de radios 5 y 10 Cm se trazan. Estime la probabilidad que un punto marcado aleatoriamente en circunferencia mayor se ubique en área de "Corona Circular" ([106])([107]) y ([108]).

El área del círculo mayor es $S = 10^2\pi\,Cm^2$ tal que el área de "Corona Circular" es equivalente a la diferencia entre ambas áreas $(T = 10^2\pi - 5^2\pi)Cm^2 = 75\pi\,Cm^2$. Por lo tanto, $P_r(A) = \left(\frac{\text{Área A}}{\text{Área }\Omega}\right) = \left(\frac{T}{S}\right) = \left(\frac{75\cancel{\pi\,Cm^2}}{10^2\cancel{\pi\,Cm^2}}\right) = 0.75.$

3.6. INDEPENDENCIA Y CONDICIONALIDAD

En la teoría de probabilidades una teoría muy útil es la independencia de eventos, que significa si la ocurrencia de uno de los eventos no da información sobre si otro evento sucederá; es decir, los eventos no influyen uno sobre otro ó son independientes. En consecuencia, dos eventos A y B se denominan independientes si la probabilidad que ambos sucedan es igual al producto de las probabilidades eventos individuales: $P_r(A \cap B) = P_r(A) * P_r(B)$. Por lo tanto, esta expresión se extiende a cualquier número de eventos ([109])([110])).

3.7. TEOREMA DE THOMAS BAYES

Teorema de Bayes. Thomas Bayes, Londres, Inglaterra, 1702 - Tunbridge Wells, 1761, fue un matemático británico y ministro presbiteriano. Su obra más conocida es el Teorema de Bayes que establece que sean $A_1,\ A_2,\ A_3,\ A_4, \ldots,\ A_k$, eventos que forman una partición (recubrimiento en el que los subconjuntos que pertenecen a una misma familia son disjuntos,

---

[106] Una corona circular, también llamada anillo, es la región entre dos círculos concéntricos. Su área equivale a la diferencia de áreas de estos dos círculos concéntricos calculada mediante $A_{(\text{Área})} = \pi\left(R^2_{(\text{Radio de círculo mayor})} - r^2_{(\text{Radio de círculo menor})}\right)$.

[107] Capa, S. H. 2015 a

[108] Fuente: Apostol, T. M. 1985.

[109] Si A y B son independientes, se puede demostrar que sus complementos son independientes tal que se cumple $P_r(A \cap B^C) = P_r(A) * P_r(B^C)$, $P_r(A^C \cap B) = P_r(A^C) * P_r(B)$ y $P_r(A^C \cap B^C) = P_r(A^C) * P_r(B^C)$. Es importante no confundir conceptos de eventos independientes con mutuamente incompatibles (disjuntos).

[110] Capa, S. H. 2015 a

su intersección, de a pares, es vacía. El recubrimiento, por su parte, hace referencia a una colección de subconjuntos A de un conjunto X: es decir, la colección de dichos subconjuntos es un recubrimiento de X) de un espacio muestral M. Sea B un evento en M. Suponga que $P(A_1)$, $P(A_2)$, $P(A_3)$, $P(A_4), \dots, P(A_k)$, $P(B/A_1)$, $P(B/A_2), P(B/A_3)$, $P(B/A_4), \dots, P(B/A_K)$ son probabilidades conocidas. Entonces:

$$P(A_i/B)_{(\text{Probabilidad condicional de } A_i \text{ dado B})} = \frac{P(B/A_i)P(A_i)}{\sum_{i=1}^{k} P(B/A_i)P(A_i)}; i = 1,2,3,4, \dots, k$$

Donde:

- $P(A_i)$: Probabilidad a priori.
- $P(B/A_i)$: Probabilidad condicional.
- $\sum_{i=1}^{k} P(B/A_i)P(A_i) = P(B_i)$: $Probabilidad\ total$.
- $P(A_i/B)$: Probabilidad a posteriori (en un evento aleatorio es la probabilidad condicional que es asignada después de que la evidencia es tomada en cuenta)

Regla de Bayes. El concepto de probabilidad condicional da lugar a ramificaciones muy discutidas en las inferencias obtenidas usando el cálculo de probabilidades. Estas dificultades provienen de la aplicación del llamado Teorema de Bayes, que es una consecuencia simple de la definición de probabilidad condiciones, pues:

$$P(A/B)_{(\text{Probabilidad condicional de A dado B})} = \frac{P(AB)}{P(B)}$$

De donde $P(A \cap B) = P(A/B)P(B)$ y, también, $P(B/A) = \frac{P(AB)}{P(A)}$

Por lo que $P(A \cap B) = P(B/A)P(A)$. Tal que $P(A/B)P(B) = P(B/A)P(A)$ o $P(A/B) = \frac{P(B/A)P(A)}{P(B)}$.

Complementariamente, la Fórmula de Bayes sostiene que si $A_1, A_2, A_3, A_4, \dots, A_n$ son eventos incompatibles dos a dos y cuya unión es todo el espacio muestral Ω; es decir, $\Omega = U_i A_1$, los A_i forman una partición Ω. Sea B un evento tal que $P(B) > 0$. Se supone que se conocen tanto las probabilidades condicionales $P(B|A_i)$, as{i como probabilidades $P(A_i)$. El problema de Bayes consiste en estimar, con datos, las probabilidades $P(A_k|B)$, $k = 1,2,3,4, . . , n$. Se sabe que:

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(B|A_k)P(A_k)}{P(B)}$$

Además, se tiene:

$$\mathrm{P(B)} = \mathrm{P(B \cap \Omega)} = P\left[B \cap \left(\bigcup_{i=1}^{n} (\mathrm{B \cap A_i})\right)\right] = \mathrm{P}\left(\bigcup_{i=1}^{n} (\mathrm{B \cap A_i})\right) = \sum_{i=1}^{n} \mathrm{P(B \cap A_i)}$$

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

Por lo tanto, se puede concluir que la Probabilidad de Causas se estima mediante la ecuación:

$$\mathrm{P(A_k|B)} = \frac{\mathrm{P(A_k \cap B)}}{\sum_{i=1}^{n} \mathrm{P(A_i \cap B)}} = \frac{\mathrm{P(B|A_k)P(A_k)}}{\sum_{i=1}^{n} \mathrm{(B|A_k)\, P(A_i)}}$$

Esta ecuación resuelve el problema en que un evento B puede producirse como consecuencia de cualquiera de sucesos $\mathrm{A_i}$ y conociendo que B se ha producido tal que se desea averiguar la probabilidad que haya sido por causa $\mathrm{A_i}$.

Ejemplo:

1) Una empresa tiene 3 fábricas A_1, A_2 y A_3 que produce ciertas piezas, como A_1 30% de piezas con un 2% defectuosas, A_2 produce 25% de piezas con 1% defectuosas y A_3 produce 45% de piezas con 3% defectuosas. Se elige al azar una pieza que resulta defectuosa tal que se desea conocer la probabilidad que provenga de fábricas A_1, A_2 y A_3, respectivamente.

Si B, una pieza defectuosa, se tiene que $\mathrm{P(A_1)} = 0.30$, $\mathrm{P(A_2)} = 0.25$, $\mathrm{P(A_3)} = 0.45$, $\mathrm{P(B|A_1)} = 0.02$, $\mathrm{P(B|A_2)} = 0.01$ y $\mathrm{P(B|A_3)} = 0.03$. Las probabilidades $\mathrm{P(A_k|B)}$ que la pieza defectuosa provenga de máquina $\mathrm{A_k}$ es:

$$\mathrm{P(A_1|B)} = \frac{\mathrm{P(B|A_1)P(A_1)}}{\sum_{i=1}^{n} \mathrm{(B|A_k)\, P(A_i)}} = \left[\frac{(0.02) * (0.30)}{(0.02 * 0.3) + (0.01 * 0.25) + (0.03 * 0.45)}\right] = \mathbf{0.273}$$

$$\mathrm{P(A_2|B)} = \frac{\mathrm{P(B|A_2)P(A_2)}}{\sum_{i=1}^{n} \mathrm{(B|A_k)\, P(A_i)}} = \left[\frac{(0.01) * (0.25)}{(0.02 * 0.3) + (0.01 * 0.25) + (0.03 * 0.45)}\right] = \mathbf{0.114}$$

$$\mathrm{P(A_3|B)} = \frac{\mathrm{P(B|A_3)P(A_3)}}{\sum_{i=1}^{n} \mathrm{(B|A_k)\, P(A_i)}} = \left[\frac{(0.03) * (0.45)}{(0.02 * 0.3) + (0.01 * 0.25) + (0.03 * 0.45)}\right] = \mathbf{0.614}$$

3.8. DISTRIBUCIONES DE PROBABILIDAD

3.8.1. Uniforme discreta

Se dice que una variable aleatoria X tiene Distribución Uniforme Discreta si toma n valores, denotados por $\mathrm{X_1, X_2, X_3, X_4, \ldots, X_n}$, con la misma probabilidad:

$$\mathrm{P(X_i)} = \mathrm{P_r(X = k)} = \frac{1}{n};\ \mathrm{k} = 1,2,3,4, \ldots, \mathrm{n}$$

Donde la esperanza matemática es:

$$E(X) = \sum_X XP(X = k) = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{n+1}{2} = \overline{X}$$

Mientras que:

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n^2-1}{12}$$

Ejemplos:

1) Si $X =$ número que aparece cuando se landa un dado, se tiene:

$P(X_i) = P_r(X = X_i) = \frac{1}{6}; k = 1,2,3,4, \dots ,6$

$$E(X) = \sum_X XP(X = k) = \frac{1}{6}\sum_{i=1}^{n} X_i = \frac{1}{6} * (1+2+3+4+5+6) = \frac{7 * \cancel{3}}{2 * \cancel{3}} = \frac{7}{2}$$

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{6}\left[\left(1-\frac{7}{2}\right)^2 + \left(2-\frac{7}{2}\right)^2 + \left(3-\frac{7}{2}\right)^2 + \left(4-\frac{7}{2}\right)^2 + \cdots \left(6-\frac{7}{2}\right)^2\right]$$

$$= \frac{(6)^2-1}{2} = \frac{35 * \cancel{3}}{2 * \cancel{3}} = \frac{35}{2}$$

2) Una máquina registra en minutos completos la diferencia de tiempo en el paso por la banda que transporta cartones Tetra Pak de x marca de leche. Sabe que la diferencia máxima puede ser de 9 minutos. Suponga que las llegadas son aleatorias, calcule el tiempo que se esperaría entre dos llegadas consecutivas, su varianza y desviación estándar.

$$E(X) = \sum_X XP(X = k) = \frac{1}{2}\sum_{i=1}^{n} X_i = \frac{1}{2} * (9+1) = \frac{5 * \cancel{2}}{\cancel{2}} = 5$$

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(9)^2-1}{12} = \frac{20 * \cancel{4}}{3 * \cancel{4}} = \frac{20}{3} = 6.67 \Longrightarrow \sigma = 2.58$$

3.8.2. Hipergeométrica

La Distribución Hipergeométrica de parámetros k, n y $N(HG(k, n, N)$ surge en situaciones en donde el modelo aproximado de probabilidad se corresponde con muestreo sin reemplazamiento de una población dicotómica (Éxito y Fracaso) finita. Concretamente, las suposiciones que llevan a considerar esta distribución son:

- La población o conjunto donde deba hacerse el muestreo consta de N individuos o elementos a seleccionar.

- Cada individuo puede ser caracterizado como un éxito (E) o fracaso (F).
- Se selecciona una muestra de n individuos de entre k individuos marcados como éxito y $N - k$ restantes marcados como fracaso.
- Hay selección equi probable en cada paso.

Una variable aleatoria X tiene una Distribución Hipergeométrica si, para algunos enteros positivos k, n y N, tal que X es número de individuos de un total de n con cierta característica (éxito) si en N hay un total de k. Entonces, $P(X = x) = \begin{cases} \frac{\binom{k}{x}\binom{N-k}{n-k}}{\binom{N}{n}} \text{ si } \max\{0, n - (n-k)\} \leq x \leq \min(n,k) \end{cases}$. En otro caso, la Distribución Hipergeométrica se aproxima a Distribución Binomial, pues cuando N es grande $(N > 10\,n)$ supone $X \in Bi(n,p)$ con $p = \frac{k}{N}$. La Distribución Hipergeométrica de parámetros k, n y N presenta una media μ, $V(X) = npq\left(\frac{N-n}{N-1}\right)$ y $\sigma = \sqrt{npq\left(\frac{N-n}{N-1}\right)}$.

En esta distribución es importante recordar algunos aspectos básicos de la teoría de ecuaciones diferenciales ordinarias de segundo orden, con el fin de presentar la ecuación diferencial Hipergeométrica de Gauss y, posteriormente, sus lımites confluentes $(^{111})$. Su ecuación es $p(x,n) = \frac{{}_aC_x^*\ {}_{N-a}C_{n-x}}{{}_NC_n}$. Donde $p(x,n)$ es probabilidad de obtener X objetos defectuosos de entre n seleccionados, ${}_aC_x^*\ {}_{N-a}C_{n-x}$ son muestras de n objetos en donde hay X que son defectuosos y $n - X$ buenos y ${}_NC_n = \delta$ son todas las muestras posibles de seleccionar de n objetos tomadas de entre N objetos en total o, también, llamado espacio muestral.

Ejemplos:

1) En una fábrica, en una determinada máquina hay 2 artículos defectuosos por cada 2000 envases para atún, el departamento de control de calidad observa 150 envases y rechaza el lote si el número de defectuosos es mayor que 1. Cuál será la probabilidad de que el lote sea rechazado.

[111] Ortiz, P. J. 2013

Sea X el número de defectuosos de un total de $n = 150$ si en $N = 2000$ hay un total de $k = 2 \in HG(k = 2, n = 150, N = 2000)$. Se pide $P(X > 1) = 1 - [P(X = 0) + P(X = 1)] = 1 - \left[\frac{\binom{2}{0}\binom{200-2}{150}}{\binom{2000}{150}} + \frac{\binom{2}{1}\binom{2000-2}{150-1}}{\binom{2000}{150}}\right]$. Sin embargo, como sus cálculos son complicados, se va a aproximar la variable aleatoria X por binomial $Bi(n = 150, p = \frac{2}{2000} = 0{,}001)$ tal que $P(X > 1) = 1 - [P(X = 0) + P(X = 1)] = 1 - \left[\binom{150}{0}0{,}001°(1 - 001)^{150-0} + \binom{150}{1}0.001^1(1 - 0.001)^1(1 - 0.001)^{150-1}\right] \approx 0.442$.

2) Hallar la probabilidad que una selección de frutas contenga 2 de 3 de las mejores cajas si en 5 productores elaboran un extracto, cuya calidad varía entre productores. Si elige 3 productores al azar, estime la calidad físicoquímica del estrato.

Con base en esta información, se tiene que $N = 5$, $m = 3$, $n = 3$ y $k = 2$. Además, $P(X = x) = \left\{\frac{\binom{k}{x}\binom{N-k}{n-k}}{\binom{N}{n}}\right.$, $P(X = 2) = \left\{\frac{\binom{3}{2}\binom{5-3}{3-2}}{\binom{5}{2}}\right.$, $P(X = 2) = \left\{\frac{\binom{3}{2}\binom{2}{1}}{\binom{5}{3}}\right.$, $P(X = 2) = \frac{3C_2 * 2C_1}{5C_3}$ y $P(X = 2) = 0{,}6$.

3.8.3. Bernoulli y binomial

Un ensayo de Bernoulli es un experimento aleatorio que tiene sólo 2 resultados posibles, denotados por éxito (E) y fracaso (F), cuya suma es 1. Es decir, tiene un papel importante en investigaciones en que sólo se tienen dos posibles resultados mutuamente excluyentes. Ocurren con probabilidades p, $q = 1 - p$ y, por ende, $p + q = 1$. La variable X tiene una distribución de probabilidades de Bernoulli si $X =$ número de éxitos en un ensayo Bernoulli ($X \sim B(p)$); es decir, X toma valores 1 o 0 con probabilidades p y q, respectivamente.

Donde la esperanza matemática es:

$$E(X) = 0 * q + 1 * p = p$$

Mientras que:

$$V(X) = (0 - p)^2 q + (1 - p)^2 p = q(p^2 + pq) = q(p^2 + p - p^2) = pq$$

Ejemplos:

1) Un lanzamiento de una moneda sólo tiene dos posibles resultados tal que $X =$ número de caras tal que $X \sim B(\frac{1}{2})$, $E(X) = 0 * \frac{1}{2} + 1 * \frac{1}{2} = \frac{1}{2}$ y $V(X) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$. Sin embargo, si la moneda está "cargada" p y q serán $\neq \frac{1}{2}$.

2) El lanzamiento de un dado sólo tiene seis posibles resultados tal que $X =$ número de ocasiones que sale 4 tal que $X \sim B(\frac{1}{6})$, $E(X) = 0 * \frac{1}{6} + 1 * \frac{1}{6} = \frac{1}{6}$ y $V(X) = \left(\frac{1}{6}\right)\left(\frac{5}{6}\right) = \frac{5}{36}$. Sin embargo, si el dado está "cargado" p y q serán $\neq \frac{1}{6}$.

La distribución Binomial es una ley de distribución de una variable aleatoria discreta X que describe el número k de éxitos en una sucesión de n pruebas de Bernoulli independientes, en cada una de las cuales la probabilidad de éxitos es igual a p. Además, considera un experimento aleatorio con n ensayos repetidos tal que son independientes, cada uno tiene sólo 2 resultados y probabilidad constante de éxito por ensayo. Con base en esto, una distribución Bernoulli es caso particular de una distribución Binomial con $n = 1$.

$P(X) = P(X = x) = C_n^k p^X q^{n-X}$; $k = 0,1,2,3,4, \dots, n$ tal que $C_n^k = \frac{n!}{X!(n-X)!}$, $\sum_{i=0}^{n} P(X) = 1$. Se tiene $\sum_{i=0}^{n} P(X = x) = \sum_{i=0}^{n} C_n^k p^X q^{n-X} = (p+q)^n = [p + (1-p)]^n = 1^n = 1$. Este resultado se obtiene con base en Binomio de Newton, que se complementa con el Número Factorial, cuya expresión $n! = n * (n-1)(n-2)(n-3)(n-4) \dots 2 * 1$ y las combinaciones $C_n^k = \frac{n!}{X!(n-X)!}$ tal que $\forall a, b \in \mathbb{R}$ y $n \in \mathbb{N}$: $(a+b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b^1 + \binom{n}{2}a^{n-2}b^2 + \binom{n}{3}a^{n-3}b^3 + \binom{n}{4}a^{n-4}b^4 + \cdots + \binom{n}{n}a^0 b^n = \sum_{i=1}^{n}\binom{n}{x}a^{n-x}b^x$.

Donde la esperanza matemática es:

$$E(X) = \sum_{i=0}^{n} x C_n^k p^X q^{n-X} = x\left(\frac{n!}{X!\,(n-X)!}\right) p^X q^{n-X} = np \sum_{j=0}^{n-1} \left(\frac{(n-1)!}{j!\,(n-1-j)!}\right) p^j q^{(n-1)-(x-1)}$$

$$= np(p+q)^{n-1} = np$$

Mientras que:

$$V(X) = V\left(\sum_{i=1}^{n} P(X_i)\right) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} pq = npq$$

Esta distribución tiene una amplia aplicación en teoría de muestreo cuando se puede contestar a una pregunta únicamente con dos opciones (sí o no).

Ejemplos:

1) Una empresa de cárnicos genera 5% de productos embutidos defectuosos. Se toma un lote de 10 unidades para calcular las probabilidades de hallar 0 defectuosos, una unidad defectuosa, todas defectuosas y al menos una unidad defectuosa.

Se considera un evento E un "evento resulta defectuoso", con probabilidad $\mathrm{p} = 0.05$ $(\mathrm{P(E)} = 0.05)$ y X el "número de elementos defectuosos en lote" tal que $\mathrm{X} \sim \mathrm{B}(10,0.05)$. Con base en esto, $\mathrm{P(X} = 0) = (10 * 0.05^0)(0.95^{10}) = 0.95^{10} = 0.599$, $\mathrm{P(X} = 1) = (10 * 0.05^1)(0.95^9) = 10 * 0.05 * 0.95^9 = 0.315$, $\mathrm{P(X} = 10) = (1 * 0.05^{10})(0.95^0) = 1 * 0.05^{10} \approx 0.0$ y $\mathrm{P(X} > 1) = \mathrm{P(X} = 1) + \mathrm{P(X} = 2) + \mathrm{P(X} = 3) + \mathrm{P(X} = 4) + \cdots + \mathrm{P(X} = 10)$; es decir, en el último de los casos, $\mathrm{P(X} > 1) = 1 - \mathrm{P(X} \leq 1) = 1 - 0.60 - 0.32 = 0.08$.

2) Una fábrica artesanal de lácteos elabora lotes experimentales de 4 unidades de quesos tipo mozzarella $(^{112})$ en que la mitad de la producción presenta unidades con alto grado de deshidratación. Con base en esta información, forme la ley de la variable aleatoria que cuantifica el número de unidades con alta deshidratación, estime la probabilidad que uno de estos lotes exista una unidad más con este problema y cuántos quesos defectuosos se esperaría por lote.

Se sabe que $\mathrm{p}\left(\frac{1}{2}\right)$, el número de quesos por lote es $\mathrm{n} = 4$, la variable aleatoria X representa el "número de quesos con alto grado de deshidratación" tal que sigue una distribución binomial $\mathrm{X} \sim \mathrm{Bin}(4, 0.5)$. En consecuencia, la probabilidad que exista 0 unidades con alto grado de deshidratación en un lote experimental es $\mathrm{P(X} = 0) = \mathrm{C}_4^0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{4-0} = \frac{1}{16} = 0.063$; si existe una unidad defectuosa, su probabilidad será $\mathrm{P(X} = 1) = \mathrm{C}_4^1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{4-1} = \frac{1}{4} = 0.250$; en caso de haber 2 unidades defectuosas, el cálculo de su probabilidad es $\mathrm{P(X} = 2) = \mathrm{C}_4^2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{4-2} = \frac{3}{8} = 0.375$; si existen 3 unidades con este problema, su probabilidad será $\mathrm{P(X} = 3) = \mathrm{C}_4^3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{4-3} = \frac{1}{4} = 0.250$ y, también, en caso

[112] Wikipedia (2018, es.wikipedia.org/wiki/Mozzarella): La mozzarella (mozarela, muzarela, muzarella, musarela2) del italiano mozzare "cortar" o de su variante regional muzzare, es un tipo de queso originario de la cocina italiana. Existe una variante de este queso en Dinamarca, pero la tradición italiana es más antigua.

que halla 4 unidades con alto grado de deshidratación, su probabilidad será $P(X = 4) = C_4^4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^{4-4} = \frac{1}{16} = 0.063$. Por lo tanto:

Cuadro 13. Probabilidades de unidades

Resumen de probabilidades de unidades con alto grado de deshidratación					
$X_{(Unidades)}$	0	1	2	3	4
$P_{(Probabilidad)}$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

Por otro lado, el cálculo de $P(X > 1) = P(X = 2) + P(X = 3) + P(X = 4) = 1 - P(X \leq 1) = 1 - \left[\frac{1}{16} + \frac{1}{4}\right] = \frac{11}{16} = 0.688$. Por último, el número de unidades con alto grado de deshidratación se estima mediante

$$E(X) = np \sum_{j=0}^{n-1} \left(\frac{(n-1)!}{j!\,(n-1-j)!} \right) p^j q^{(n-1)-(x-1)} = np = (2 * \cancel{2}) \left(\frac{1}{\cancel{2}}\right) = 2.00$$

3.8.4. Geométrica y binomial negativa

Considere una secuencia de pruebas Bernoulli, con probabilidad de éxito p, pero en vez de contar el número de éxitos, interesa saber el número de intentos hasta obtener el primer éxito. Una sucesión de pruebas de este tipo forma un Experimento Geométrico. Una variable aleatoria discreta X que puede tomar un número infinito de valores $1,2,3,4,\ldots,n$, se dice que sigue una ley Distribución Geométrica de parámetro $p(0 < p < 1)$, si la probabilidad que X tome valor k es $P_r(X = k) = p(1-p)^{k-1}$; $k = 1,2,3,4,\ldots,n$. A esta variable aleatoria es denotada como $X \sim g(p)$. Su esperanza y varianza son iguales a $E(X) = \left(\frac{1}{p}\right)$ y $V(X) = \left(\frac{1-p}{p^2}\right)$. La distribución geométrica se aplica a investigaciones de mercado y muestreo, su objetivo es conocer cuántas compras se hacen en una promoción para obtener un premio.

Ejemplo:

1) Una promoción de marcas de papas ecuatorianas incluye, en cada una de sus bolsas o fundas, una de las figuras coleccionables de una película de recién estreno en cines a nivel nacional. Si un consumidor estima que existe igual número de figuras de cada personaje en la promoción, ¿Cuántas bolsas o fundas de papas espera comprar para coleccionar las tres figuras de la película?

En la primera compra se obtiene una figura coleccionable que no se tenía previamente; por lo tanto, $E(X_1) = 1$. En la segunda compra tiene una probabilidad $p_2 = \left(\frac{2}{3}\right)$ de conseguir una nueva figura coleccionable tal que el número de compras que se harán es $E(X_2) = \left(\frac{1}{p_2}\right) = \left(\frac{1}{\frac{2}{3}}\right) = \left(\frac{3}{2}\right)$. Una vez coleccionadas dos figuras de la película, la probabilidad de hallar la figura faltante será $p_3 = \left(\frac{1}{3}\right)$ y el número de compras que se hará para obtenerla es $E(X_2) = \left(\frac{1}{p_3}\right) = \left(\frac{1}{\frac{1}{3}}\right) = \left(\frac{3}{1}\right)$. En consecuencia, el número total de compras a realizar es $E(X) = E(X_1) + E(X_2) + E(X_3) = 1 + \frac{3}{2} + \frac{3}{1} = 5.50 \approx 6.0$. Por lo tanto, el número de bolsas o fundas de papas a comprar será al menos 6 para obtener la colección de figuras completa de la película.

A la ley de Distribución Binomial Negativa se llama Distribución de Pascal, pues tiene las mismas aplicaciones que la Ley Geométrica. Si se generaliza el concepto de Ley Geométrica e interesa el número de pruebas de Bernoulli obligatorias para obtener r éxitos. Entonces, una variable aleatoria X que puede tomar un número infinito de valores $r, r+1, r+2, r+3, r+4+, \dots, +r+n$ se dice que sigue una ley de Distribución Binomial Negativa de parámetros $(r+p)(r \geq 1, 0 < p < 1)$, si la probabilidad que X tome valor k es $P_r(X = k) = (C_{k-1}^{r-1})(p^r)(1-p)^{k-r};\ k = r, r+1, r+2, r+3, r+4+, \dots, +r+n$. El parámetro r es el número de éxtiso que se desea obtener tal que p es probabilidad de obtener éxtio. A esta variable aleatoria se denota por $X \sim BN(r, p)$. En consecuencia, $E(X) = \left(\frac{r}{p}\right)$ y $V(X) = r\left(\frac{1-p}{p^2}\right)$.

2) Una máquina, dañada, envasa latas de conserva de duraznos en almíbar ([113]) por unidad de forma independiente. Se considera que 5% del producto envasado es defectuoso. Si la máquina se detiene en el tercer producto defectuoso. ¿Cuál es el número de latas producidas hasta detenerse?, ¿cuál es la probabilidad que la máquina se detenga en la 9na lata buena? y ¿cuál que se detenga sin producir lata alguna?

[113] Wikipedia (2018, es.wikipedia.org/wiki/Almíbar): Del árabe clásico maybah, es un jarabe a base de una disolución sobresaturada de agua y azúcar; es decir, la denominación de almíbar se aplica a la solución acuosa de azúcar, en caliente, destinada a líquido de cobertura o a confecciones de confitería y repostería.

Si se define variable aleatoria X como el número de latas producidas hasta que haya 3 defectuosas tal que $X \sim BN(3, 0.05)$. Entonces, se calcula la esperanza $E(X) = \left(\frac{r}{p}\right) = \left(\frac{3}{0.05}\right) = 60$ por lo que se espera producir 60 unidades de latas hasta que se detenga la máquina. Se calcula $P_r(X = 9) = (C_{9-1}^{3-1})(0.05^3)(1 - 0.05)^{9-3} = 0.003$. Finalmente, la probabilidad que ninguna lata producida sea aceptable se estima mediante, que indica que las 3 primeras fueron defectuosas $(k = 3)$, $P_r(X = 3) = (C_{9-1}^{3-1})(0.05^3)(1 - 0.05)^{3-3} = 0.00013$

3.8.5. Poisson

Realmente se trata de un caso particular de aplicación del Teorema Central de Límite forma $Linderberg - Lévy$ ([114]), cuya particularidad consiste en

---

[114] Es considerado un caso particular de forma Lyapounov, primera demostración rigurosa hecha en 1901 respecto a un Teorema Central del Límite válida para Distribuciones Binomiales, pues las premisas previas son más restrictivas: dada una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$ independientes de manera que las variables tendrán por medias y varianzas $E(X_i) = \mu_i$ y $D^2(X_i) = \sigma_i^2$ tal que se tendrá la sucesión $\eta_n = \left[\frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}\right]$ que converge a una distribución $N[0,1]$. La forma $Linderberg - Lévy$ es un caso particular de forma Lyapounov, pues las premisas previas son más restrictivas tal que dada una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$ independientes y con misma distribución tal que las variables tendrán igual media y varianza: $E(X_i) = \mu$ y $D^2(X_i) = \sigma^2$se tendrá $\eta_n = \left[\frac{\sum_{i=1}^{n} X_i^{-n*\mu}}{\sigma\sqrt{n}}\right]$ que converge a una distribución $N[0,1]$, donde $\sum_{i=1}^{n} X_i \longrightarrow N[n * \mu; \sigma\sqrt{n}]$ tal que si se suma un gran número de variables aleatorias independientes e igualmente distribuidas, con la misma media y varianza. Esta suma se distribuye normalmente con media n veces la media común y desviación típica raíz cuadrada de n veces la desviación típica común. Para demostrar este teorema se prueba que la $F.G.M.$ de η_n tiende a $F.G.M.$ de distribución normal reducida cuando $n \longrightarrow \infty$: $Lím_{\varphi_\eta} = e^{\left(\frac{t^2}{2}\right)}$ que considera nuevas variables tal que $W_i = X_i - \mu$ para $i = 1,2,3,4,\ldots,n$ tal que $\eta_n = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma\sqrt{n}}\right) = \sum_{i=1}^{n}\left(\frac{W_i}{\sigma\sqrt{n}}\right)$, como todas las X son estocásticamente independientes e idénticamente distribuidas, las W_i serán independientes y, por lo tanto, tendrán igual distribución tal que la $F.G.M.$ de η_n será $\varphi_{\eta_n}(t) = \prod_{i=1}^{n}\left(\varphi_{W_i}\left(\frac{t}{\sigma\sqrt{n}}\right)\right) = \left[\varphi_{W_i}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$ por lo que primero se obtiene $\varphi_{W_i}\left(\frac{t}{\sigma\sqrt{n}}\right) = E\left[\frac{tW_i}{e\sigma\sqrt{n}}\right]$ tal que $\varphi_{W_i}\left(\frac{t}{\sigma\sqrt{n}}\right) = E\left[1 + \frac{t}{\sigma\sqrt{n}}W_i + \frac{t^2}{2en^2}W_i^2 + \phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right] = 1 + \frac{t}{\sigma\sqrt{n}}E[W_i] + \frac{t^2}{2n\sigma^2}E[W_i^2] + E\left[\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]$ dado que $E[W_i] = 0$ y $E[W_i^2] = \sigma^2$ se tendrá que $\varphi_{\eta_n}(t) = \left[1 + \frac{t^2}{2n} + E\left[\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]\right]^n$ si se toman límites cuando $n \longrightarrow \infty$ la función ϕ es un infinitésimo de orden superior a $\frac{t^2}{2n}$ y, por lo tanto, $Lím_{\varphi_\eta}(t) = \lim_{n\to\infty}\left[1 + \frac{t^2}{2n} + E\left[\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]\right]^n = \lim_{n\to\infty}\left(1 + \frac{t^2}{2n}\right)^n = e^{\left(\frac{t^2}{2}\right)}$ que es la expresión de $F.G.M.$ de la Normal $(0,1)$ tal que esta demostración es sólo válida para el caso en que $\exists\ F.G.M.$, sino fuera así se usa análogamente las funciones características. Del propio Teorema del Límite en forma $Lindeberg - Lévy$ se infiere, denominado su versión media. Así, dada una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}$ independientes y con igual distribución de manera que las variables tendrán igual media y varianza $E(X_i) = \mu$ y $D^2(X_i) = \sigma^2$ tal que se tiene la sucesión $\omega_i = \sum_{i=1}^{n}\left(\frac{X_i}{n}\right)$; es decir, la media de la sucesión y dado que se conoce por $Lindeberg - Lévy$ que la sucesión η_n se define como $\eta_n = \left[\frac{\sum_{i=1}^{n} X_i^{-n*\mu}}{\sigma\sqrt{n}}\right]$ converge en distribución a una distribución $N[0,1]$ tal que la nueva sucesión ω se tendrá que la sucesión η_n es definida como $\eta_n = \left[\frac{\sum_{i=1}^{n}\omega_i^{-\mu}}{\frac{\sigma}{\sqrt{n}}}\right]$ converge en distribución a una distribución $N[0,1]$, donde $\omega_i = \sum_{i=1}^{n}\left(\frac{X_i}{n}\right) \xrightarrow{d} N\left[\mu, \frac{\sigma}{\sqrt{n}}\right]$; es decir, la media aritmética de un gran número de variables aleatorias

que las variables aleatorias que forman la sucesión son o se distribuyen según una Poisson de parámetro λ. El hecho de tratarla radica en su utilidad y practicidad. Dada una sucesión de variables aleatorias $\{X_n\}_{n=1}^{\infty}\infty$ donde $X_i \longrightarrow \text{Poisson}\ (\lambda)$, por lo que la media común λ y su varianza común. En Teorema Central del Límite se tiene la sucesión η_n es definida como $\eta_n = \left[\frac{\sum_{i=1}^{n} \omega_i^{-\mu}}{\frac{\sigma}{\sqrt{n}}}\right]$ converge en distribución a una distribución $N[0,1]$. Dado que la distribución de Poisson cumple el Teorema de Adición para parámetro λ, se tiene que $\sum_{i=1}^{n}(X_i) = J \longrightarrow \varrho[\lambda_i = n\lambda]$ tal que se conoce que $\mu_i = \lambda_i = n\lambda = n\mu$ y la desviación típica sería $\sigma_j = \sqrt{\lambda_i} = \sqrt{n\lambda} = \sigma\sqrt{n}$ por lo que $\eta_n = \left[\frac{J-n\mu}{\sigma\sqrt{n}}\right] = \left[\frac{J-\lambda_j}{\sqrt{\lambda_j}}\right] \xrightarrow{d} N[0,1]$ tal que se deduce que una distribución de Poisson cuando $\lambda \longrightarrow \infty$ converge a una $N[0,1]$ con media λ y desviación típica $\sqrt{\lambda}$.

Una variable aleatoria discreta X que puede tomar un número infinito de valores $0,1,2,3,4,\ldots,n$ sigue una ley de Poisson de parámetro $\lambda(\lambda > 0)$ si la probabilidad que X tome el valor k es $P_r(X = k) = \left(\frac{e^{-\lambda}\lambda^k}{k!}\right); k = 0,1,2,3,4\ldots,n$ a esta variable aleatoria se le nota $X\sim P(\lambda)$. Su esperanza y su varianza son, respectivamente, iguales a $E(X) = \lambda$ y $V(X) = \lambda$.

La distribución Poisson se aplica a sucesos que se presentan en el tiempo o en espacio, tal como número de accidentes de tráfico, número de llamadas telefónicas a una central, número de goles que marca un equipo en un partido, número de bacterias en una placa, etcétera. El significado del parámetro λ es promedio del aparecimiento del evento en n pruebas. Esta ley de probabilidad es una buena aproximación a la Binomial cuando n es relativamente grande $(n \geq 30)$ y p pequeño $(p \leq 0.05)$, sabiendo $\lambda = np$. Puesto que muchas aplicaciones de esta distribución dependen del tiempo, es conveniente ponerla de forma $P_r(X = k) = \left(\frac{e^{-\lambda t}(\lambda t)^k}{k!}\right); k = 0,1,2,3,4\ldots,n$ que se interpreta como la probabilidad que sucedan exactamente k evetos en un intervalo de tiempo fijo de duración t. Para la distribución de Poisson también existe una fórmula de recurrencia para el cálculo de probabilidades dada por $P_0 = e^{-\lambda}$ y $P_k = (P_{k-1})\left(\frac{\lambda}{k}\right); k = 0,1,2,3,4\ldots,n$.

independientes e igualmente distribuidas, con igual media y varianza se distribuirá normalmente con media común y desviación típica, desviación típica común dividida por $\sqrt{n}$.

Ejemplos:

1) El promedio de llamadas que pasan en un centro de atención al cliente de una empresa de refrescos o colas durante un minuto es igual a dos. Estime la probabilidad que en tres minutos se hagan 4 llamadas, menos de 4 llamadas y al menos 4 llamadas.

Es necesario usar la segunda forma de Ley de Poisson con $\lambda = 2$ y $t = 3$ tal que $P_r(X = k) = \left(\frac{e^{-\lambda t}(\lambda t)^k}{k!}\right)$. La probabilidad que en 2 minutos se hagan 4 llamadas es $P_r(X = 4) = \left(\frac{e^{-2*3}(2*3)^4}{4!}\right) = \frac{e^{-6}(6)^4}{(4*3*2*1)} = 0.134$. Además, la probabilidad buscada $P_r(X < 4) = P_r(X = 3) + P_r(X = 2) + P_r(X = 1) + P_r(X = 0) = \left(\frac{e^{-2*3}(2*3)^3}{3!}\right) + \left(\frac{e^{-2*3}(2*3)^2}{2!}\right) + \left(\frac{e^{-2*3}(2*3)^1}{1!}\right) + \left(\frac{e^{-2*3}(2*3)^0}{0!}\right)$. Los eventos "se hicieron menos de 4 llamadas" y "se hicieron al menos de 4 llamadas" son complementarios, por eso, su probabilidad es $P_r(X \geq 4) = 1 - P_r(X < 4) = 1 - 0.151 = 0.849$.

2) El gerente de una fábrica procesadora de lácteos concibe adquirir una nueva máquina elaboradora de yogurt entre tipos A y B. Por día de funcionamiento, el número de reparaciones X que requiere la máquina A es una variable aleatoria de Poisson cuya media es $0.1\ t$, siendo t el tiempo de funcionamiento diario en horas. El número de reparaciones diarias Y de máquina B es una variable aleatoria de Poisson con media $0.12\ t$. El costo diario operativo de máquina A es $P_A(t) = 10\ t + 30\ X^2$ y para máquina $P_B(t) = 8\ t + 30\ Y^2$, ¿Cuál de las máquinas da el menor costo esperado, si un día de trabajo consiste en a) 10 Hr y b) 20 Hr?

El costo esperado para máquina A es $E(C_A[t]) = E(10\ t + 30\ X^2) = 10\ t + 30\ E(X^2) = 10\ t + 30\ \left[Var(X) + \left(E(X^2)\right)^2\right] = 10\ t + 30\ [0.1\ t + (0.12\ t)^2] = 13\ t + 0.3\ t^2$. También, $E(C_B[t]) = E(8\ t + 30\ Y^2) = 8\ t + 30\ E(Y^2) = 8\ t + 30\ \left[Var(Y) + \left(E(Y^2)\right)^2\right] = 8\ t + 30\ [0.12\ t + (0.12\ t)^2] = 11.6\ t + 0.432\ t^2$. Además, $E(C_A[10]) = 13(10) + 0.3(10)^2 = 160$ y $E(C_B[10]) = 11.6(10) + 0.432(10)^2 = 159.20$ tal que si $t = 10$, el menor costo será de máquina B con 159.20. Por último, $E(C_A[20]) = 13(20) + 0.3(20)^2 = 380$ y $E(C_B[20]) = 11.6(10) + 0.432(20)^2 = 404.80$ tal que si $t = 20$, el menor costo será de máquina A con 404.80.

3.8.6. Uniforme

La distribución de probabilidad de una variable aleatoria continua X se llama uniforme si en el intervalo $[a, b]$ la función de de densidad es constante e igual a $f(x) = \begin{cases} \frac{1}{b-a}, & \text{si } x \in [a, b] \\ 0, & \text{si } x \notin [a, b] \end{cases}$. A esta variable aleatoria se la nota como $X \cup [a, b]$ tal que $\begin{cases} 0 \text{ si } x < a \\ \frac{x-a}{b-a} \text{ si } a \leq x \leq b \\ 1, \text{ si } x > \end{cases}$.

La esperanza y la varianza son, respectivamente $E(X) = \frac{a+b}{2}$ y $Var(X) = \frac{(b-a)^2}{12}$. Esta distribución es el análogo continuo de la distribución uniforme discreta, que asigna igual probabilidad a cada resultado de un experimento. Tiene amplia aplicación en problemas de simulación estadística y en fenómenos que presentan regularidad en su aparecimiento, pero que no es posible usar variables discretas, como cuando dependen del tiempo. También, el error originado por el redondeo de un número se describe satisfactoriamente mediante una distribución uniforme en el intervalo $\left[-\frac{1}{2}, \frac{1}{1}\right]$.

Ejemplos:

1) Una variable aleatoria X tiene distribución uniforme sobre $[-2,3]$. a) calcular $Pr(X = 1), Pr(X<1.3), Pr(|X| < 1.5)$ y b) Hallar el valor de t tal que $Pr(X > t) = \frac{1}{3}$.

Con base en esta información, se sabe que $f(x) = 0$ si $x \notin [-2,3]$. Se tienen los límites $a = -2, b = 3$; por lo que, $f(x) = 1/(3 - (-2)) = \frac{1}{5}$, en $[-2,3]$. La función de densidad es $f(x) = \begin{cases} \frac{1}{5}, & \text{si } x \in [-2,3] \\ 0, & \text{si } x \notin [-2,3] \end{cases}$. Tal que, $Pr(X = 1) = \int_1^1 f(x)dx = 0$, pues X es una variable aleatoria continua. Por lo tanto, $Pr(X<1.3) = \int_{-\infty}^{1.3} f(x)dx = \int_{-2}^{1.3} \frac{1}{5}dx = 0,66$ y $Pr(|X| < 1.5) = Pr(-1.5<X<1.5) = \int_{-1.5}^{1.5} f(x)dx = \int_{-1.5}^{1.5} \frac{1}{5}dx = \frac{3}{5} = 0,6$. Finalmente, $Pr(X > t)$ es $Pr(X > t) = \int_t^{\infty} f(x)dx = \int_t^3 \frac{1}{5}dx + \int_3^{-\infty} 0dx = \left[\frac{x}{5}\right]_t^3 = \frac{3-t}{5}$. Entonces, $\frac{3-t}{5} = \frac{1}{3}$, con lo cual $t = \frac{4}{3}$.

2) Dos empleados de una empresa agroindustrial, A y B, se encontrán en una parada de bus entre 9:00 y 10:00 Hr. Cada uno espera un máximo de 10 minutos. ¿Cuál es la probabilidad de que no se encuentren, si A llegará a las 9:30 en punto ([115])?

Se sabe que la variable aleatoria X que describe el tiempo de llegada de A puede tomar cualquier valor entre las 9:00 y 10:00 Hr ó entre 0 y 60 minutos. En consecuencia, $X \sim U[a, b]$ y, por ende, su función de densidad será $f(x) = \begin{cases} \frac{1}{60}, & \text{si } 0 \leq t \leq 60; \\ 0, & \text{caso contrario} \end{cases}$. Puesto que B llegará a 9:30 ó 30 minutos después de 9:00 y esperará máximo 10 minutos, A no se encontrará con B si llega entre 9:00 y 9:20 o si llega después de 9:40 a.m. Finalmente, la probabilidad de que no se encuentren es $\Pr[(0 < 20) o (40 < X \leq 60)] = \int_0^{20} \frac{1}{60} dt + \int_{40}^{60} \frac{1}{60} dt = \frac{1}{3} + \frac{1}{3} = 0{,}667.$

3.8.7. Exponencial

Se tiene X con una distribución exponencial de parámetros λ, denotada por $X \sim E(\lambda)$ si $f(X) = \begin{cases} \lambda e^{-\lambda x} & , si\ x \geq 0 \\ 0 & , si\ x < 0 \end{cases} (\lambda > 0, fijo)$. También, se verifica que a) $\int_{-\infty}^{\infty} (\lambda e^{-\lambda x})\, dx = \lambda \int_0^{\infty} \left(\frac{-e^{-\lambda x}}{\lambda}\right)\Big|_0^{\infty} = 1$, b) $F(x) = \int_0^x (\lambda e^{-\lambda x})\, dt = 1 - e^{-\lambda x}$, c) $E(x) = \int_0^{\infty} x f(x)\, dx = \frac{1}{\lambda}$ y d) $V(x) = E(X^2) - E^2(X) = \int_0^{\infty} x^2 f(x)\, dx - \left(\frac{1}{\lambda}\right)^2 = \left(\frac{1}{\lambda^2}\right)$.

Ejemplo:

1) Un Ingeniero Agroindustrial supone que el tiempo de vida útil (T), en días, de una marca de queso artesanal en estado fresco sigue una distribución exponencial $E(0.01)$. Calcule a) el tiempo promedio de vida, b) probabilidad que el queso dure 20 Hr, c) probabilidad que el queso dure menos de 10 Hr, d) probabilidad que el queso dure al menos 50 Hr y e (probabilidad que el queso dure entre 20 y 100 Hr.

Con base en la información, se estima que $\lambda = 0.01$ tal que $E(T) = \left(\frac{1}{\lambda}\right) = \left(\frac{1}{0.01}\right) = 100$ Hr. Además, $P(T = 20) = 0$ tal que T es variable aleatoria continua. También, $P(T \leq 10) = \int_0^{10} [0.01 * (e)^{-0.01x}]\, dx = 1 - (e)^{-0.01*10} = 1 - (e)^{-0.1} = 0.095.$

Simultáneamente, $P(T \leq 50) = \int_{50}^{\infty} [0.01 * (e)^{-0.01x}]\, dx = (e)^{-0.01*50} = 0.61$ y, por último,

[115] Capa, S. H. 2015 a

$P(20 \leq T \leq 100) = F(100) - F(20) = [(1 - (e)^{-0.01*100}) - (1 - (e)^{-0.01*20})] = [(e)^{-0.20} - (e)^{-1.00}] = 0.45.$

3.8.8. Normal

Según ([116]), la distribución normal, también conocida como de Gauss, es la distribución más utilizada en la estadística. Constituye un buen modelo para muchas, aunque no para todas las poblaciones continuas. Parte de esto último se debe al teorema del límite central.

La distribución de probabilidad continua que más se utiliza es la distribución normal, con la conocida forma de campana que estudiamos en relación con la regla empírica. Su función de densidad normal es: Se dice que una variable Y tiene una distribución normal de probabilidad si y sólo si, para $\sigma > 0$ y $-\infty < \mu < \infty$, la función de densidad de Y es $f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(-y-\mu)^2/(2\sigma^2)}$, $-\infty < \gamma < \infty$. La función normal contiene dos parámetros μ y σ ([117]). Tal que, si Y es una variable aleatoria normalmente distribuida con parámetros μ y σ, entonces:

$E(Y) = \mu$ y $V(Y) = \sigma^2$.

Sin embargo, según ([118]), Si X es una variable aleatoria cuya función de densidad de probabilidad es normal con media μ y varianza σ^2, se expresa como $X \sim N(\mu, \sigma^2)$. Según ([119]), los resultados contenidos en el este Teorema implican que el parámetro μ localiza el centro de la distribución y que σ mide su dispersión. Una gráfica de una función de densidad normal se muestra así:

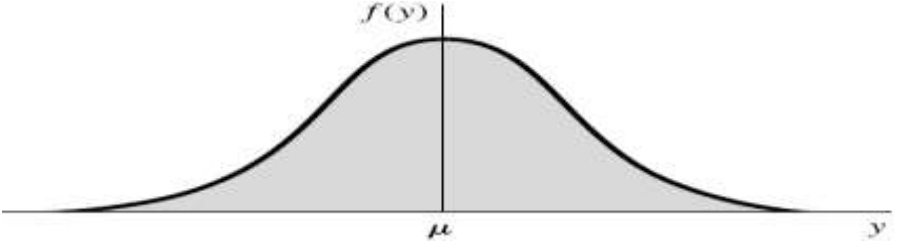

Figura 20. Distribución normal o de Gauss (120)

[116] Oteyza, E; Lam, E; Hernández, C. y Carrillo, A. 2015
[117] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015
[118] Oteyza, E; Lam, E; Hernández, C. y Carrillo, A. 2015
[119] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015
[120] Fuente: Navidi, W. 2006

Las áreas bajo la función de densidad normal correspondientes a $P(a \leq Y \leq b)$ requieren la evaluación de la integral $\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(-y-\mu)^2/(2\sigma^2)} dy$.

Aun cuando hay un número infinito de distribuciones normales (μ puede tomar cualquier valor finito, en tanto que σ puede tomar cualquier valor finito positivo), sólo necesitamos la tabla 4 para calcular áreas bajo densidades normales. La función de densidad normal es simétrica alrededor del valor μ, de modo que las áreas tienen que ser tabuladas en sólo un lado de la media. Las áreas tabuladas están a la derecha de los puntos z, donde z es la distancia desde la media, medida en desviaciones estándar ([121]).

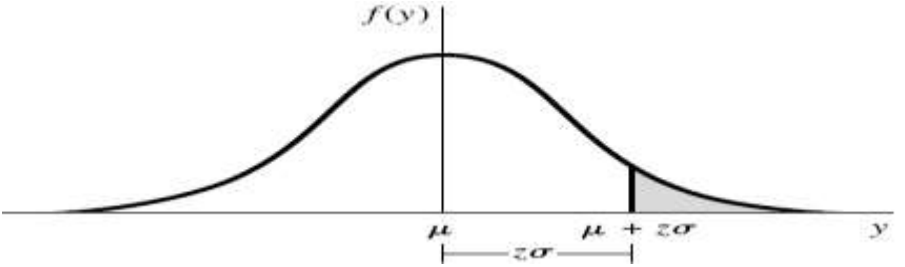

Figura 21. Distribución normal o de Gauss según número de desviaciones estándar ([122])

Muchos de los fenómenos observados en el mundo real tienen distribuciones de frecuencia relativas que se pueden modelar en forma adecuada con una distribución de probabilidad normal. Por tanto, en muchos problemas prácticos es razonable suponer que las variables aleatorias observables en una muestra aleatoria, $Y_1, Y_2 \ldots, Y_n$, son independientes con la misma función de densidad normal (Wackerly, Mendenhall y Scheaffer, 2010).

Teorema. Sea $Y_1, Y_2 \ldots, Y_n$, una muestra aleatoria de tamaño n de una distribución normal con media μ y varianza σ^2. Entonces: $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Se distribuye normalmente con media $\mu_{\overline{Y}} = \mu$ y varianza $\sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$ (Wackerly, Mendenhall y Scheaffer, 2010). Como $Y_1, Y_2, \ldots, Y_n$ es una muestra aleatoria de una distribución normal con media μ y varianza σ^2, Y_i, $i = 1, 2, \ldots, n$ son variables independientes distribuidas normalmente, con $E(Y_i) =$

[121] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015
[122] Fuente: Wackerly, D. D; Mendenhall, W. y Scheaffer, R. L. 2010

μ y $V(Y_i) = \sigma^2$. Además: $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}(Y_1) + \frac{1}{n}(Y_2) + \cdots + \frac{1}{n}(Y_n) = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$. Donde: $a_i = \frac{1}{n}$, $i = 1, 2, \ldots, n$ ([123]).

Así, $\overline{Y}$es una distribución lineal de $Y_1, Y_2, \ldots, Y_n$ y se puede aplicar, según ([124]), el teorema 6. 3 para concluir que $\overline{Y}$ está distribuida normalmente con:

$$E(\overline{Y}) = E\left[\frac{1}{n}(Y_1) + \cdots + \frac{1}{n}(Y_n)\right] = \frac{1}{n}(\mu) + \cdots + \frac{1}{n}(\mu) = \mu$$

$$V(\overline{Y}) = V\left[\frac{1}{n}(Y_1) + \cdots + \frac{1}{n}(Y_n)\right] = \frac{1}{n^2}(\sigma^2) + \cdots + \frac{1}{n^2}(\sigma^2) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

Tal que, la distribución muestral de $\overline{Y}$es normal con media $\mu_{\overline{Y}} = \mu$ y varianza $\sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$. De acuerdo con el anterior teorema, se deduce que $Z = \frac{\overline{Y} - \mu_{\overline{Y}}}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\overline{Y} - \mu}{\sigma}\right)$ tiene una distribución normal estándar.

Características de la Distribución de Probabilidad Normal:

- La curva normal es acampanada y presenta un solo presenta un solo pico en el centro de la distribución. La media aritmética, la mediana y la moda de la distribución son iguales y están localizadas en el pico. Tal que, la mitad del área bajo la curva se encuentra por arriba de este punto central y, la otra mitad, por debajo.

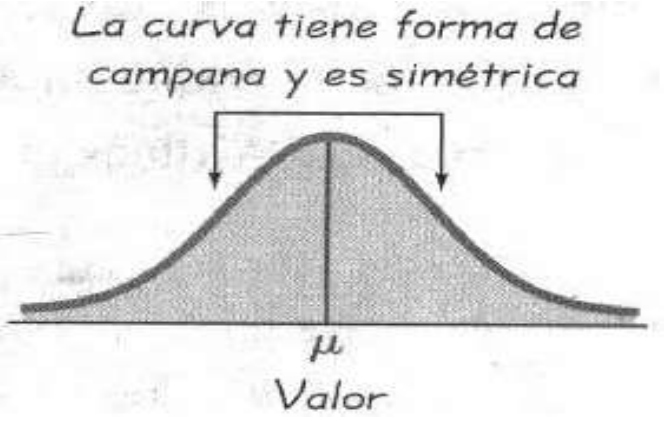

Figura 22. Distribución normal o de Gauss ([125])

[123] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015
[124] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015
[125] Fuente: www.google.com/search?client=firefox-b&biw=1366&bih=635&tbm=isch&sa=1&ei=g320W9DBFIy5ggfgk56ACQ&q=Forma+sim%C3%A9trica+de+campana

➢ La distribución de probabilidad normal es simétrica alrededor de μ con respecto a su media. Si se corta la curva normal verticalmente en este valor central, ambas mitades serán como imágenes en el espejo.

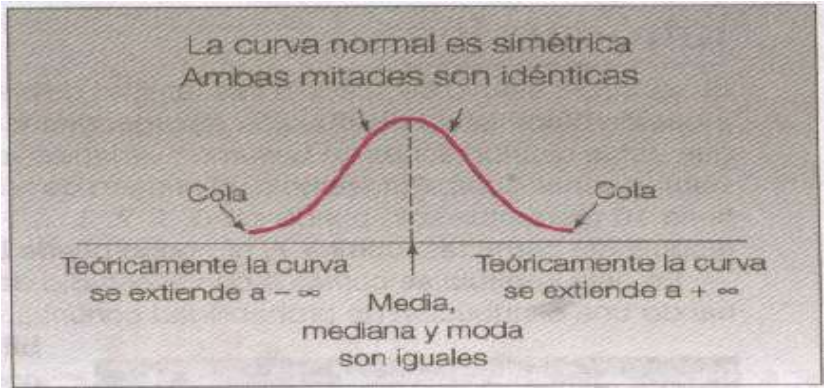

Figura 23. Características de distribución normal o de Gauss ([126])

➢ La curva normal decrece uniformemente en ambas direcciones a partir del valor central. Es asintótica, la curva ser acerca cada vez más al eje X, pero en realidad nunca llega a tocarlo. Es decir, los puntos extremos de la curva se extienden indefinidamente en ambas direcciones.

➢ El área total bajo la curva deber ser igual a 1.

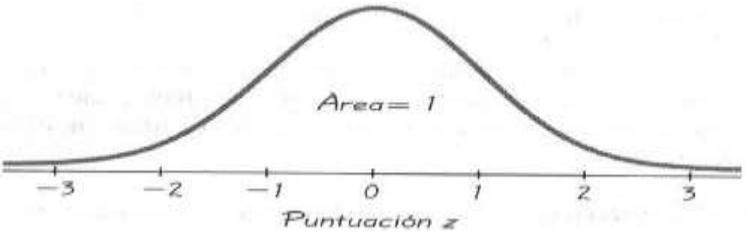

Figura 24. Área de distribución normal o de Gauss según desviaciones estándar ([127])

➢ Cada punto de la curva debe tener una altura vertical igual o mayor que 0 (No puede estar por debajo del eje x).

+de+gauss&oq=Forma+sim%C3%A9trica+de+campana+de+gauss&gs_l=img.3...255692.260478.0.260872.35.15.0.0.0.0.0..0.0....0...1c.1.64.img..35.0.0....0.PWB3RrUUtLM#imgrc=qyc-T2i07ZgRRM:

[126] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

[127] Fuente: www.google.com/search?client=firefox-b&biw=1366&bih=635&tbm=isch&sa=1&ei=kX60W_3xOYjl_QaZ56qoAg&q=%C3%80rea+bajo+la+campana+de+gauss&oq=%C3%80rea+bajo+la+campana+de+gauss&gs_l=img.3...176694.180877.0.181326.17.11.0.0.0.0.0.0..0.0....0...1c.1.64.img..17.0.0....0.7AyAIz4_4uw#imgrc=Gm6WcPY3MyxIBM:

➢ Toda población normal se caracteriza por:

Aproximadamente 68% de la población se encuentra en intervalo $\mu \pm \sigma$, Aproximadamente 95% de la población se encuentra en intervalo $\mu \pm 2\sigma$ y Aproximadamente 99.7% de la población se encuentra en intervalo $\mu \pm 3\sigma$.

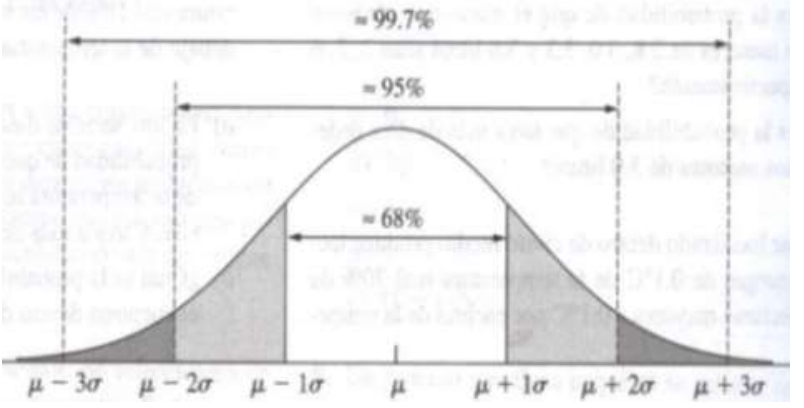

Figura 25. Probabilidad de distribución normal o de Gauss según desviaciones estándar poblacionales ([128]

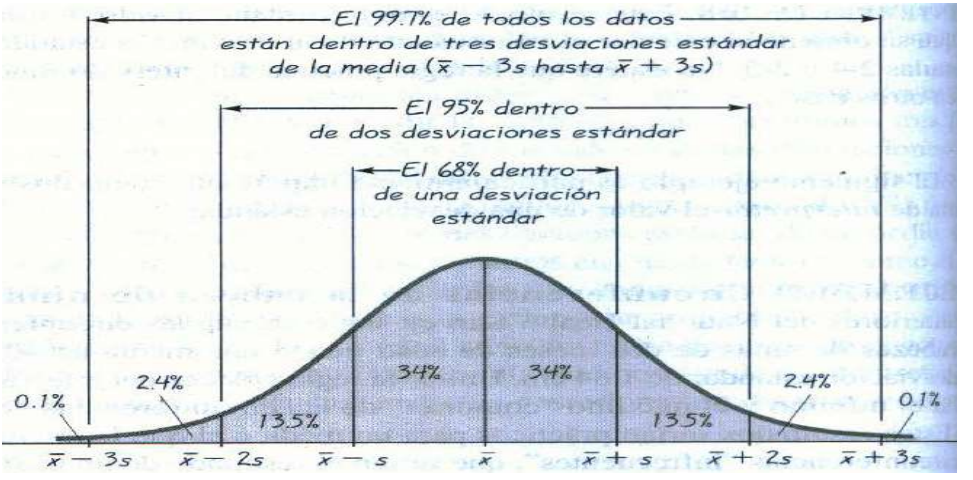

Figura 26. Probabilidad de distribución normal o de Gauss según desviaciones estándar poblacionales ([129])

[128] Fuente: Navidi, W. 2006
[129] Fuente: Lind, D. A; Marchal, W. G. y Mason, R. D. 2004

➢ Un miembro de la familia de distribuciones normales Z tiene una media de cero ($\mu = 0$) y una desviación estándar de 1 ($\sigma = 1$) ([130])

([131]) afirman hay una familia de distribuciones normales. Cada distribución puede tener una media (μ) o desviación estándar (σ). En consecuencia, el número de distribuciones es ilimitado. Cualquier distribución normal puede convertirse en "distribución normal estándar" restando la media a cada observación y dividiendo entre la desviación estándar.

Valor Z es la diferencia entre un valor elegido, denotado por X y la media μ, dividida entre la desviación estándar, σ. Por lo tanto, un valor Z es la distancia a la media, medida en unidades de desviación estándar.

$$\text{Valor Normal Estándar: } \mathbf{Z} = \frac{\mathbf{X} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

Donde:
X es valor de cualquier media y observación específica

μ media de la distribución

σ desviación estándar de la distribución

Ejemplos ([132])**:**
1) Las calificaciones para un examen de admisión a una universidad están normalmente distribuidas con media de 75 y desviación estándar 10. ¿Qué fracción de las calificaciones se encuentra entre 80 y 90? ([133]).

Sol. Recuerde que z es la distancia desde la media de una distribución normal expresada en unidades de desviación estándar. Entonces: $\mathrm{z} = \frac{\mathrm{y}-\mu}{\sigma}$. Por lo tanto, la fracción deseada de la población está dada por el área entre $\mathrm{z}_1 = \frac{80-75}{10} = .5$ y $\mathrm{z}_2 = \frac{90-75}{10} = 1.5$. En consecuencia:

[130] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014
[131] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014
[132] Un conjunto de ejercicios se encuentran en carpeta "Distribución Normal".
[133] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015

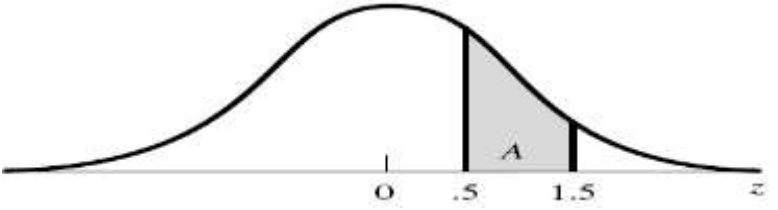

Figura 27. Cálculo de un intervalo de probabilidad [134]

Si $A = A(.5) - A(1.5) = .3085 - .0668 = .2417$. En otras palabras, si se localiza el intervalo de interés, (80, 90), en el eje horizontal inferior marcado como Y. Las calificaciones z correspondientes se dan en el eje horizontal superior y es evidente que el área sombreada da $P(80 < Y < 90) = P(0.5 < Z < 1.5) = .2417$:

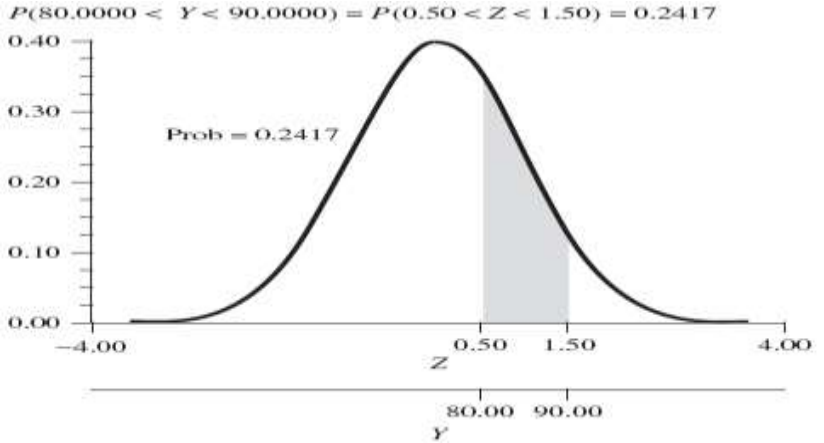

Figura 28. Cálculo de un intervalo de probabilidad [135]

2) Denote con Z una variable aleatoria normal con media 0 y desviación estándar 1. Encuentre [136]:

a) $\mathrm{P}(\mathrm{Z} > 2)$. **Sol.** Como $\mu = 0$ y $\sigma = 1$. Implique que $\mathrm{z} = 2$ desviaciones estándar arriba del valor medio (media). En la tabla 4 se ubica este valor (2.0). Ésta área denotada por (z), por lo que $\mathrm{A}(2.0) = .0228$. Entonces, $\mathrm{P}(\mathrm{Z} > 2) = .0228$.

b) $\mathrm{P}(-2 \leq \mathrm{Z} \leq 2)$. **Sol.** En el inciso a se encontró que $\mathrm{A}_1 = \mathrm{A}(2.0) = .0228$. Como la función de densidad es simétrica alrededor de la media $\mu = 0$, se deduce por lógica que $\mathrm{A}_2 = \mathrm{A}_1 = .0228$. Por lo tanto, $\mathrm{P}(-2 \leq \mathrm{Z} \leq 2) = 1 - \mathrm{A}_1 - \mathrm{A}_2 = 1 - 2(.0228) = .9544$.

[134] Fuente: Wackerly, D. D; Mendenhall, W. y Scheaffer, R. L. 2010
[135] Fuente: Wackerly, D. D; Mendenhall, W. y Scheaffer, R. L. 2010.
[136] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015

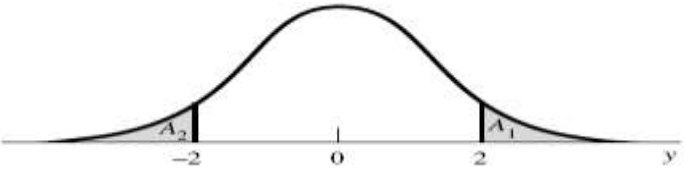

Figura 29. Cálculo de un intervalo de probabilidad ([137])

c) $P(0 \leq Z \leq 1.73)$. **Sol.** Como $P(Z > 0) = A(0) = .5$, se tiene que $P(0 \leq Z \leq 1.73) = .5 - A(1.73)$. Tal que si éste último valor se obtiene de la tabla 4 se tiene que $A(1.73) = .0418$. En consecuencia, $P(0 \leq Z \leq 1.73) = .5 - .0418 = .4582$.

d) Una máquina embotelladora puede ser regulada para que descargue un promedio de μ onzas por botella. Se ha observado que la cantidad de líquido dosificado por la máquina está distribuida normalmente con $\sigma = 1.0$ onza. Una muestra de $n = 9$ botellas se selecciona aleatoriamente de la producción de la máquina en un día determinado (todas embotelladas con el mismo ajuste de la máquina) y las onzas de contenido líquido se miden para cada una. Determine la probabilidad de que la media muestral se encuentre a no más de 0.3 onza de la verdadera media μ para el ajuste seleccionado de la máquina y ¿Cuántas observaciones deben estar incluidas en la muestra, si deseamos que Y se encuentre a no más de 0.3 onza de μ con probabilidad de 0.95? ([138]).

Sea $Y_1, Y_2 \ldots, Y_9$, denota el contenido en onzas de las botellas que se van a observar, entonces sabemos que las Y_i están distribuidas normalmente con media μ y varianza $\sigma^2 = 1$ y para $i = 1, 2, , \ldots, 9$. Con base en el anterior teorema, $\bar{Y}$ posee una distribución muestral normal con media $\mu_{\bar{Y}} = \mu$ y varianza $\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n} = \frac{1}{9}$. Deseamos hallar:

$$P(|\bar{Y} - \mu| \leq 0.3) = P[-0.3 \leq (\bar{Y} - \mu) \leq 0.3] = P\left(-\frac{0.3}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{0.3}{\frac{\sigma}{\sqrt{n}}}\right)$$

Como $\frac{(\bar{Y} - \mu_{\bar{Y}})}{\sigma_{\bar{Y}}} = \frac{(\bar{Y} - \mu)}{\frac{\sigma}{\sqrt{n}}}$ tiene una distribución normal estándar, se deduce que:

$$P(|\bar{Y} - \mu| \leq 0.3) = P\left(-\frac{0.3}{\frac{1}{\sqrt{9}}} \leq Z \leq \frac{0.3}{\frac{1}{\sqrt{9}}}\right) = P(-0.9 \leq Z \leq 0.9)$$

[137] Fuente: Wackerly, D. D; Mendenhall, W. y Scheaffer, R. L. 2010
[138] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015

Usando la Tabla 4, Áreas de Curvas Normal:

$$\mathrm{P}(-0.9 \leq \mathrm{Z} \leq 0.9) = 1 - 2\mathrm{P}(\mathrm{Z} > 0.90) = 1 - 2(.1841) = 0.6318$$

Por consiguiente, la probabilidad es sólo 0.6318 de que la media muestral se encuentre a no más de 0.3 onza de la verdadera media poblacional. Ahora buscamos $\mathrm{P}(|\overline{\mathrm{Y}} - \mu| \leq 0.3) = \mathrm{P}[-0.3 \leq (\overline{\mathrm{Y}} - \mu) \leq 0.3] = 0.95$. Si se divide cada término de la desigualdad entre $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$ (recuerde que $\sigma = 1$) se tiene:

$$\mathrm{P}\left(-\frac{0.3}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\overline{\mathrm{Y}} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{0.3}{\frac{\sigma}{\sqrt{n}}}\right) = \mathrm{P}\left(-0.3\sqrt{n} \leq \mathrm{Z} \leq 0.3\sqrt{n}\right) = 0.95$$

Usando la tabla de Probabilidad de Distribución Normal Estandarizada, Áreas de Curvas Normal:

$$\mathrm{P}\left(-0.3\sqrt{n} \leq \mathrm{Z} \leq 0.3\sqrt{n}\right) \Longrightarrow (-1.96 \leq \mathrm{Z} \leq 1.96) = 0.95 \text{ (Prob. numérica de curva)}$$

Esto indica: $0.3\sqrt{n} = 1.96$ o bien que es equivalente a $n = \left(\frac{1.96}{0.3}\right)^2 = 42.68$. En la práctica, es imposible tomar una muestra de tamaño 42.68. Por lo tanto, una muestra de tamaño 42 es insuficientemente grande para llegar al objetivo y, en consecuencia, si $n = 43$, $P(|\bar{Y} - \mu| \leq 0.3)$es ligeramente mayor que 0.95 (95 % de probabilidad).

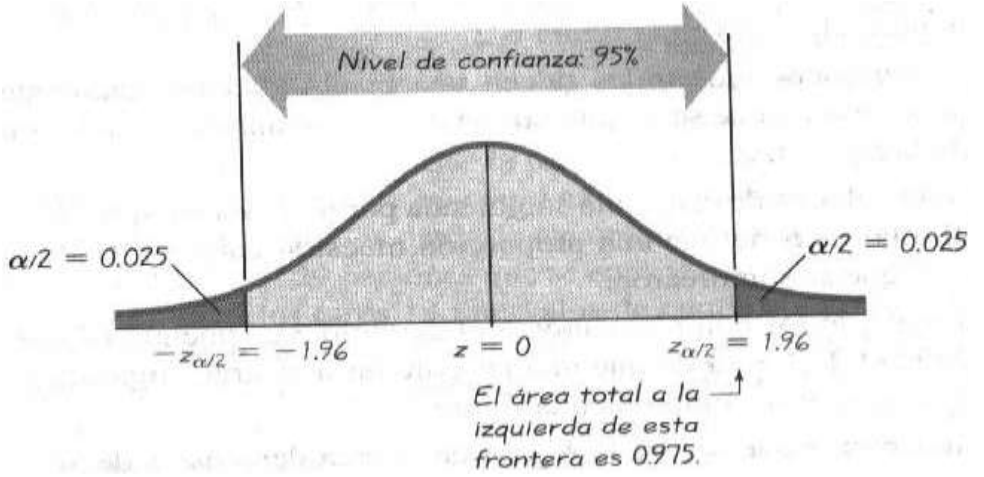

Figura 30. Intervalo de 95% de probabilidad ([139])

[139] Fuente: www.google.com/search?client=firefox-b&biw=1366&bih=635&tbm=isch&sa=1&ei=TH-0W5vkCdG-ggfV0ZfQDQ&q=Nivel+de+confiabilidad+estad%C3%ADstica+de+95%25&oq=Nivel+de+confiabilidad+estad%C3%A

Esto indica: $0.3\sqrt{n} = 1.96$ o bien que es equivalente a $n = \left(\frac{1.96}{0.3}\right)^2 = 42.68$. En la práctica, es imposible tomar una muestra de tamaño 42.68. Por lo tanto, una muestra de tamaño 42 es insuficientemente grande para llegar al objetivo y, en consecuencia, si $n = 43$, $P(|\overline{Y} - \mu| \leq 0.3)$es ligeramente mayor que 0.95 (95 % de probabilidad).

3.8.9. Teorema de límite central

También es conocido como Teorema Central de Límite (TCL), si $X_1, X_2, X_3, X_4, \ldots, X_n$ es una muestra de X tal que $E(X) = \mu$ y $V(X) = \sigma^2$ con μ y σ^2 finitos. Entonces, $z = \frac{(\sqrt{n})(\overline{X}-\mu)}{\sigma} \sim N(0,1)$ asintóticamente.

3.9. DISTRIBUCIONES MULTIDIMENSIONALES

3.9.1. Variables aleatorias bidimensionales $(^{140})$

Función con valores numéricos definida sobre un espacio muestral. En otras palabras, es una variable aleatoria si el valor que asume es un suceso numérico aleatorio. Ejemplo: observar el número de defectos en un mueble o el registro de aprovechamiento de un estudiante en particular. Tipos de variables:

3.9.1.1. Discretas

Es una variable que sólo puede asumir un conjunto numerable de valores. Es decir, que si se puede enumerar es discreta. Ejemplos: número de tornillos defectuosos en una muestra de 10 unidades extraídas de una producción industrial. Número de casas rurales que cuentan con servicio eléctrico en una región. El número de fallas de un aeroplano en un periodo de tiempo. El número de personas que esperan una consulta en un consultorio médico.

3.9.1.2. Continuas

Es una variable que puede asumir el número infinitamente grande de valores correspondientes a los puntos sobre un intervalo en una línea recta. Es decir, la palabra continua significa que procede sin interrupción y proporciona la clave para identificar a las variables aleatorias continuas. Una característica importante de esta variable es que si las

Dstica+de+95%25&gs_l=img.3...534927.542472.0.542713.41.33.0.0.0.0.258.3871.0j18j5.23.0....0...1c.1.64.img..18.16.2791...0j0i67k1j0i5i30k1j0i24k1j0i30k1.0.XinQWnr39ZM#imgrc=W83iuXZzpjloPM:

[140] Un conjunto de ejercicios se encuentra en carpeta "Variables aleatorias bidimensionales"

mediciones u observaciones tienen un conjunto de valores que forman puntos sobre una línea sin interrupción o espacio entre ellos. Ejemplos: Estatura de una persona. Tiempo de vida de una célula humana. Cantidad de azúcar en una naranja. Tiempo requerido para completar una operación de montaje en un proceso de fabricación.

3.9.2. Distribución condicionada

Sean a y b dos números reales cualquiera, X y Y dos variables aleatorias tal que la $P_r(Y \leq b) \neq 0$. La probabilidad condicionada que $X \leq a$ dado que $Y \leq b$, se representa mediante $P_r(X \leq a \,|Y \leq b)$, se define mediante $P_r(X \leq a \,|Y \leq b) = \left(\frac{P_r(X \leq a, Y \leq b)}{P_r(Y \leq b)}\right)$. Además, en variables aleatorias discretas, la probabilidad condicionada de x, para un valor fijo de variable y esta dada por $P_r(X = x \,|Y = y) = \left(\frac{P_r(X=a, Y=b)}{P_r(Y=y)}\right)$. También, para variables aleatorias continuas, la función de densidad condicionada de x, para un valor fijo de variable y se calcula por $f(x \,|y) = \left(\frac{f(x,y)}{fY(y)}\right)$ como $fY(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_{-\infty}^{\infty} f(x \,|y)\, fY(y) dx$ tal que $f(x \,|y) = \left(\frac{f(X \,|y)\, fY(y) dx}{\int_{-\infty}^{\infty} f(X \,|y)\, fY(y) dx}\right)$ que puede interpretarse como Teorema de Bayes para funciones de densidad ([141]).

3.9.3. Esperanza y covarianza de variable aleatoria bidimensional

Al igual que en caso de variables aleatorias unidimensionales, en bidimensionales es posible calcular la esperanza y varianza, previa transformación de variables. Sea (X, Y) un vector aleatorio bidimensional y $g(x, y)$ una funcion real tal que $\begin{matrix} g: \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ (x, y) & \longrightarrow & g(x, y) \end{matrix}$ por lo que si (X, Y) es un vector aleatorio discreto, su funcion de probabilidad es $f(x, y)$ tal que $E(g(x, y)) = \sum_{x_i} \sum_{y_i} g(x_i, y_i) * f(x_i, y_i) = \sum_{x_i} \sum_{y_i} g(x_i, y_i) * p_{ij}$. Análogamente, si (X, Y) es un vector aleatorio continuo con funcion de densidad conjunta $f(x, y)$ tal que $E(g(x, y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dy\, dx$ por lo que si X y Y son independientes se entiende que $E[g(X)h(Y)] = E[g(X)] * E[h(Y)]$.

Por otro lado, estas variables tienen una medida estadística llamada Covarianza, que permite estimar la relación entre variables aleatorias X y Y. Su representación matemática es

[141] Capa, S.,H. a. 2015

$\mathrm{Cov}(X,Y) = E[(X - E(X)) * (Y - E(Y))]$ o, equivalente a, $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$. Sus propiedades son que $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$, $\mathrm{Cov}(X,X) = V(X)$, si a_1 con a_2 son constantes positivas entonces $\mathrm{Cov}(a_1X_1 + a_2X_2, Y) = a_1\mathrm{Cov}(X_1,Y) + a_2\mathrm{Cov}(X_2,Y)$, si variables aleatorias son independientes, su covarianza es cero debido a que $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$ y $|\mathrm{Cov}(X,Y)| \leq \sqrt{V(X)V(Y)}$. De forma análoga, se deduce una expresión para varianza de suma de dos variables aleatoria cualquiera como $V(X+Y) = E(X+Y)^2 - [E(X+Y)]^2 = E(X^2) + E(Y^2) + 2E(XY) - [E^2(X) + E^2(Y) + 2E(X)E(Y)] = [E(X^2) - E^2(Y)] + [E(Y^2) - E^2(Y)] + 2[E(XY) - E(X)E(Y)] = V(X) + V(Y) + 2\mathrm{Cov}(X,Y)$. Además, la varianza del producto de dos variables aleatorias X y Y será $V(X+Y) = V(X) + V(Y) + 2\mathrm{Cov}(X,Y)$ y, finalmente, $V(XY) = \mu_Y^2 V(X) + \mu_X^2(X)V(Y) + 2\mu_X\mu_Y\mathrm{Cov}(X,Y) + 2\mu_X E[(X-\mu_X) * (Y-\mu_Y)^2] + 2\mu_Y E[(Y-\mu_Y) * (X-\mu_X)^2] + 2E[(X-\mu_X)^2 * (Y-\mu_Y)^2] - [\mathrm{Cov}(X,Y)]^2$ donde $\mu_X = E(X)$ y, también, $\mu_Y = E(Y)$.

Con base en esto, el Coeficiente de Correlación se define como una medida de dependencia entre variables aleatorias X y Y. Es estimado mediante $\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)V(Y)}}$. Sus propiedades son $\rho(X,Y) = \rho(Y,X)$, su valor fluctúa entre ± 1 $(-1 \leq \rho(X,Y) \leq 1)$, si $Y = b + mx$ (linealmente) tal que a y m son constantes por lo que $|\rho(X,Y)| = 1$, si $|\rho(X,Y)| = 1 \Longrightarrow \exists$ dependencia lineal entre variables X y Y y, por último, si X y Y son variables aleatorias independientes implica que $\rho(X,Y) = 0$ ([142]).

3.9.4. Variables aleatorias multidimensionales

Sobre un mismo espacio muestral Ω están definidas las variables aleatorias $x_1, x_2, x_3, x_4, \ldots, x_n$ talque $Z = (x_1, x_2, x_3, x_4, \ldots, x_n)$ es un vector aleatorio o una variable aleatoria $n - \mathrm{dimensional}$. Además, si $x_1, x_2, x_3, x_4, \ldots, x_n$ son variables aleatorias discretas, el vector aleatorio Z es discreto y su función probabilística es $fz = (x_1, x_2, x_3, x_4, \ldots, x_n) = P_r(X_1 = x_1, x_2, x_3, x_4, \ldots, X_n = x_n)$ tal que si $x_1, x_2, x_3, x_4, \ldots, X_n$ son variables aleatorias continuas, el vector aleatorio Z es continuo y probabilidad de evento $(x_1, x_2, x_3, x_4, \ldots, X_n = x_n) \epsilon E \subset \mathbb{R}^n$, estimado por $P_r[(x_1, x_2, x_3, x_4, \ldots, X_n = x_n) \epsilon E] =$

[142] Capa, S.,H. a. 2015

$\int \dots \int_E fz(x_1, x_2, x_3, x_4, \dots, x_n)\, dx_1, dx_2, dx_3, dx_4, \dots, dx_n$ la función es llamada densidad conjunta de $x_1, x_2, x_3, x_4, \dots, X_n$.

La función de distribución de variable aleatoria multivariante Z está definida por $Fz(x_1, x_2, x_3, x_4, \dots, x_n) = P_r(X_1 \leq x_1, x_2, x_3, x_4, \dots, X_n \leq x_n)$. Las variables aleatoria $x_1, x_2, x_3, x_4, \dots, X_n$ son independientes si $Fz(x_1, x_2, x_3, x_4, \dots, x_n) = Fx_1(x_1), Fx_2(x_2), Fx_3(x_3), Fx_4(x_4), \dots, Fx_n(x_n)$ o, equivalente a, $fz(x_1, x_2, x_3, x_4, \dots, x_n) = fx_1(x_1), fx_2(x_2), fx_3(x_3), fx_4(x_4), \dots, fx_n(x_n)$. Sea $g: \mathbb{R}^{\mathbf{n}} \to \mathbb{R}, E[g(x_1, x_2, x_3, x_4, \dots, x_n)]$, según ley de Z es discreta y por $E[g(x_1, x_2, x_3, x_4, \dots, x_n)] = \sum_{x_1} \dots \sum_{x_n} g(x_1, x_2, x_3, x_4, \dots, x_n) fz(x_1, x_2, x_3, x_4, \dots, x_n)$ tal que cuando Z es discreta y por $E[g(x_1, x_2, x_3, x_4, \dots, x_n)] =$

$\int \dots \int_{\mathbb{R}^{\mathbf{n}}} g(x_1, x_2, x_3, x_4, \dots, x_n) fz(x_1, x_2, x_3, x_4, \dots, x_n) dx_1, dx_2, dx_3, dx_4, \dots, dx_n$ *c*ando Z es continua. Donde, $Cov(X_r, X_m) = E[(X_r - E(X_r))(X_m - E(X_m))] = E(X_r, X_m) - E(X_r)E(X_m) = \sigma_{rm}$. Por último, el coeficiente de correlación se estima mediante $p(X_r, X_m) = \frac{Cov(X_r, X_m)}{\sqrt{Var(X_r) * Var(X_m)}}$ tal que al vector aleatorio $Z = (x_1, x_2, x_3, x_4, \dots, x_n)$ se asocia su matriz de varianza − covarianza o, sino, matriz covarianza definida:

H

$$= \begin{pmatrix} \mathbf{Var(X_1) = Cov(X_1, X_1)} & Cov(X_1, X_2) & Cov(X_1, X_3) & Cov(X_1, X_4) & \dots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & \mathbf{Var(X_2) = Cov(X_2, X_2)} & Cov(X_2, X_3) & Cov(X_2, X_4) & \dots & Cov(X_2, X_n) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & \mathbf{Var(X_3) = Cov(X_3, X_3)} & Cov(X_3, X_4) & \dots & Cov(X_3, X_n) \\ Cov(X_4, X_1) & Cov(X_4, X_2) & Cov(X_4, X_3) & \mathbf{Var(X_4) = Cov(X_4, X_4)} & \dots & Cov(X_4, X_n) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X_m, X_1) & Cov(X_m, X_2) & Cov(X_m, X_3) & Cov(X_m, X_4) & \dots & \mathbf{Var(X_n) = Cov(X_m, X_n)} \end{pmatrix}$$

Por propiedades de covarianza, la matriz H es simétrica tal que la relación que da varianza de suma de variables aleatorias es $Var(x_1, x_2, x_3, x_4, \dots, x_n) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \leq i < j \leq n} Cov(X_i, X_j)$ ([143]).

3.9.5. Distribuciones importantes

3.9.5.1. Multinomial

El vector aleatorio $k - dimensional\ X = (x_1, x_2, x_3, x_4, \dots, x_k)$ sigue una distribución multinomial de parámetros $(n; p_1, p_2, p_3, p_4, \dots, p_k)$ tal que $0 < p_i < 1$, $\sum_{i=1}^{k} p_i = 1 \rightarrow$

[143] Fuente: Galindo, E. 2006

$P_r(X = N) = P_r(X_1 = n_1, n_2, n_3, n_3, \dots, X_k = n_k) = \frac{n!}{n_1! \dots n_k!} * \left(p_1^{n_1}, p_2^{n_2}, p_3^{n_3}, p_4^{n_4}, \dots p_k^{n_k}\right)$ para $N = (n_1, n_2, n_3, n_3, \dots, n_k)$ talque $\sum_{i=1}^{k} n_i = n$. La esperanza, varianza y covarianza son iguales a $E = n(p_1, p_2, p_3, p_4, \dots, p_k), Var(X_i) = np_i(1 - p_i); i = 1,2,3,4, \dots, k$ y $Cov(X_i, X_j) = -np_ip_j; i \neq j$. Esta distribución es la generalización multivariante de la distribución binomial. La distribución marginal de cada componente X_i es binomial de parámetros (n_i, p_i) y cualquier distribución condicionada es multinomial. Esta ley de distribución tiene su aplicación en análisis estadístico de datos cualitativos.

Ejemplo:

1) En Banco Pichincha, Ecuador, opera tarjetas de crédito, registrando sus posibles causas de renovación. Se estima que 60% son pérdidas, 25% vencimiento de uso y 15% deterioro. En consecuencia, un día de la semana se recibió 28 solicitudes de su renovación. Evalúe probabilidad que 15 renovaciones sean por pérdidas, 7 por vencimiento y 6 por deterioro.

Con base en esta información, X_1 es número de renovaciones por pérdida, X_2 número de renovaciones por vencimiento y X_3 número de renovaciones por deterioro tal que probabilidad para $n_1 = 15, n_2 = 7$ y $n_3 = 6$ donde $n_1, +n_2 + n_3 = 28$; po lo tanto, $P_r(X_1 = 15, X_2 = 7 \text{ y } X_3 = 6) = \frac{28!}{15!*7!*6!} * \left(0.15^6 * 0.25^7 * 0.6^{15}\right) = 0.021.$

3.9.5.2. Uniforme

Un vector aleatorio $X = (x_1, x_2, x_3, x_4, \dots, x_n)$ tiene distribución uniforme $S = [a_1, b_2] * \dots * [a_n, b_n] \epsilon \mathbb{R}^{\mathbf{n}}$ si función de densidad probabilística $f(x_1, x_2, x_3, x_4, \dots, x_n)$ es $f(x_1, x_2, x_3, x_4, \dots, x_n) = \begin{cases} \frac{1}{\prod_{i=1}^{n}(b_i - a_i)}, & \text{si } x \epsilon S; \\ 0, & \text{si } x \notin S; \end{cases}$. Esta distribución es análoga multivariante de distribución uniforme. Las distribuciones marginales de variables aleatorias $X_i (i = 1,2,3,4, \dots, n)$ son uniformes con densidad $fX_i(x) = \begin{cases} \frac{1}{b_i - a_i}, & \text{si } x \epsilon [a_i, b_i]; \\ 0, & \text{si } x \notin [a_i, b_i]; \end{cases}$.

3.9.5.3. Normal bivariante

Sea un vector aleatorio $Z = (X, Y)$ que tiene distribución normal bivariante no degenerada si su función de densidad conjunta es $f(x, y) = \frac{1}{\left(2*\pi*\sigma_1*\sigma_2*\sqrt{1-\rho^2}\right)} * e^{\left\{-\frac{1}{\left(2*(1-\rho^2)\right)}*\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(2\rho)*(x-\mu_1)*(x-\mu_2)}{(\sigma_1*\sigma_2)} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}}$, donde se sabe que ρ es coeficiente de correlación entre variables aleatorias X y Y. Las leyes marginales de X e Y son $X \sim N(\mu_1, \sigma_1^2)$ y

$Y \sim N(\mu_2, \sigma_2^2)$. Finalmente, existe una gran cantidad de variables aleatorias multidimensionales y su función de densidad es ([144]):

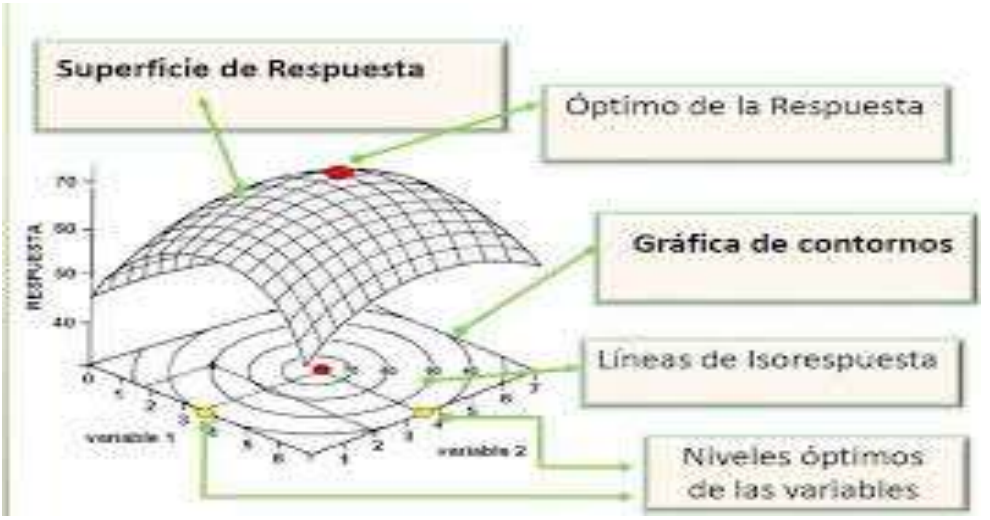

Figura 31. Superficie de respuesta ([145])

[144] La Metodología de Superficies de Respuesta (RSM) es un conjunto de técnicas matemáticas utilizadas en el tratamiento de problemas en los que una respuesta de interés está influida por varios factores de carácter cuantitativo. Su propósito inicial es diseñar un experimento que proporcione valores razonables de la variable respuesta y, a continuación, determinar el modelo matemático que mejor se ajusta a los datos obtenidos. El objetivo final es establecer los valores de los factores que optimizan el valor de la variable respuesta. Su metodología tiene dos etapas distintas, modelamiento y desplazamiento, que son repetidas tantas veces cuantas fueran necesarias, con el objetivo de alcanzar una región optima de la superficie investigada. El modelamiento, generalmente es hecho ajustándose a modelos simples (en general, lineares o cuadráticos) una respuesta obtenidas con planeamientos factoriales o con planeamientos factoriales ampliados. El desplazamiento se da siempre a lo largo del camino de máxima inclinación de un determinado modelo que es una trayectoria en que la respuesta varía de forma más pronunciada. Si el valor real esperado (η) que toma la variable de interés está influido por niveles de k factores cuantitativos $X_1, X_2, X_3, X_4, \dots, X_k$ tal que indica que alguna función $X_1, X_2, X_3, X_4, \dots, X_k$ continua en X_i para $\forall\, i = 1,2,3,4, \dots, k$ queestima valor η para alguna combinación de niveles $\eta = f(X_1, X_2, X_3, X_4, \dots, X_k)$ tal que su variable respuesta puede escribirse como $Y = \eta = f(X_1, X_2, X_3, X_4, \dots, X_k) + \varepsilon$; por lo tanto, la relación existente entre η y k factores cuantitativos se representa mediante una hipersuperficie, subconjunto de un espacio euclídeo $(k + 1)$- dimensional, llamada Superficie de Respuesta.

[145] Fuente: www.google.com/search?q=Superficie+de+respuesta&client=firefox-b&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiygYrY6ujdAhWCwFkKHfaEBvwQ_AUICigB&biw=1366&bih=635#imgrc=xqmjwVdMboFdpM:

4. VARIABLES, VARIABILIDAD Y FUNCIÓN DE PROBABILIDADES

4.1. VARIABLE ALEATORIA

Una variable aleatoria es una descripción numérica de resultados de un experimento. Es la cantidad resultante de un experimento, pero debido al azar, puede tomar valores diferentes. En general, una variable aleatoria es una función que a cada resultado posible de un experimento aleatorio le asocia un número real o, en otras palabras, es una función definida sobre un espacio muestral. Ejemplo: observar el número de defectos en un mueble o el registro de aprovechamiento de un estudiante en particular.

Ejemplos:

a) $\mathrm{X}_1 =$ Número de hijos en una familia.

b) X_2= Número de Panes de Pascua vendidos en una semana.

c) $\mathrm{X}_3 =$ Tirar un dado es un experimento, pues se puede presentar cualquiera de los 6 resultados posibles.

d) $X_4 =$ Si se cuenta el número de empleados ausentes de su turno de trabajo el lunes, el número puede ser 0, 1, 2, 3… El número de inasistencias es la variable aleatoria.

e) $\mathrm{X}_5 =$ Si se pesa un lingote de acero, el resultado (libras) puede ser 2,500, 2,500.1, 2,500.13 y, así sucesivamente, dependiendo de la precisión de la báscula. El peso es la variable aleatoria.

f) $\mathrm{X}_6 =$ Si se tiran al aire dos monetas y se cuenta el número de caras, el mismo puede ser 0, 1 o 2. Puesto que el número de caras se debe al azar, dicho número de caras es variable aleatoria.

g) $\mathrm{X}_7 =$ Otras variables podrían ser: número de lámparas defectuosas producidas durante 1 semana, estaturas de integrantes de un equipo de basquetbol femenil, cantidad de corredores en un maratón y número diario de automovilistas que cometieron infracción por manejar bajo la influencia del alcohol.

h) $\mathrm{X}_8 =$ En un experimento se aplico insecticida a tres larvas de un insecto y, al cabo de cierto tiempo, se observó los insectos vivos (v) y muertos (m). En este caso el espacio muestral es:

$$\mathrm{M} = \{\mathrm{vvv},\ \mathrm{mvv},\ \mathrm{vmv},\ \mathrm{vvm},\ \mathrm{vmm},\ \mathrm{mvm},\ \mathrm{mmv},\ \mathrm{mmm}\}$$

Supóngase que sólo se está interesado en el número de insectos muertos. Entonces, los resultados pueden representarse por los números 0, 1, 2 y 3. Si se considera estos números como valores tomados por una variable X. Entonces, esa variable toma valores según resultados de un experimento aleatorio y, por ende, se puede pensar en X como una *Variable Aleatoria*, que en este caso representa el número de insectos muertos de un total de tres. En este experimento X toma el valor 0 si ocurre el evento $\{\mathrm{vvv}\}$, valor 1 si ocurre cualquier de tres eventos $\{\mathrm{mvv}\}$, $\{\mathrm{vmv}\}$ o $\{\mathrm{vvm}\}$, valor 2 si $\{\mathrm{vmm}\}$, $\{\mathrm{mvm}\}$ o $\{\mathrm{mmv}\}$ y, finalmente, el valor 3 en caso de que $\{\mathrm{mmm}\}$**.** Entonces, si a los 8 puntos ordenados del especio M en el experimento de los insectos se les representa por $\beta_1, \beta_2, \beta_3, \beta_4 \ldots \beta_8$, se tiene que la variable aleatoria X, aplicada a cada punto del espacio muestral que es su dominio, da:

$$\mathrm{X}(\beta_1) = 0 \qquad \mathrm{X}(\beta_5) = 2$$
$$\mathrm{X}(\beta_2) = 1 \qquad \mathrm{X}(\beta_6) = 2$$
$$\mathrm{X}(\beta_3) = 1 \qquad \mathrm{X}(\beta_7) = 2$$
$$\mathrm{X}(\beta_4) = 1 \qquad \mathrm{X}(\beta_8) = 3$$

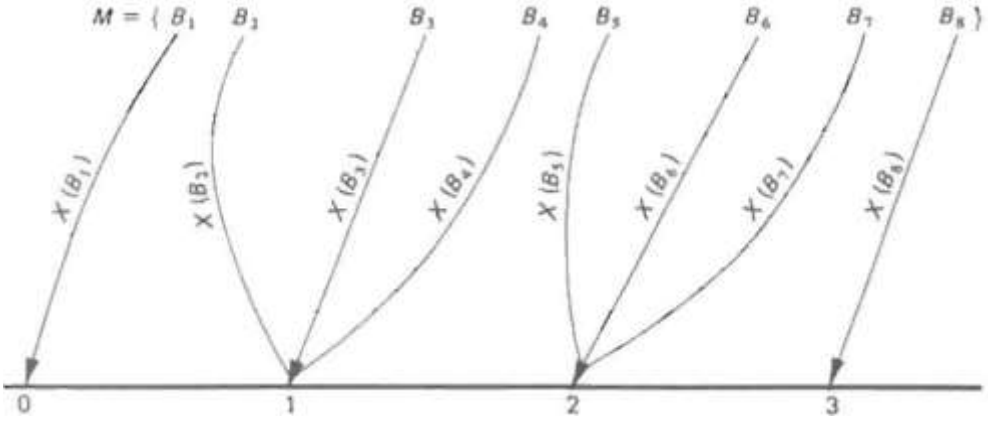

Figura 32. Espacio muestral combinatorio (146)

Algunos experimentos dan como resultado que son cuantitativos ($ USD, peso corporal o número de hijos) y, en cambio, otros dan como resultado que son cualitativos (color o preferencia religiosa). Existen dos tipos de variable aleatoria:

[146] Fuente: Infante, G. S. y Zárate de L., G. P. 1984.

4.1.1. Discreta

Es aquella variable que puede tomar cuando más un número finito o infinito numerable o contable (numerable) de valores; es decir, sólo puede asumir un conjunto numerable de valores. En otras palabras, si se puede enumerar es discreta. Aunque, en la inmensa mayoría de las situaciones prácticas, representan conteos de alguna característica, por lo que en ocasiones sus distribuciones (probabilidades de tomar valores específicos) reciben el nombre de distribuciones de conteos.

Ejemplos: Número de hijos en una familia, número de panes de pascua vendidos en una semana, puntuaciones otorgadas por jueces a aspectos técnicos y forma en patinaje sobre hielo (7.2, 8.9 o 9.7), número de tornillos defectuosos en una muestra de 10 unidades extraídas de una producción industrial, número de casas rurales que cuentan con servicio eléctrico en una región, número de fallas de un aeroplano en un periodo de tiempo, número de personas que esperan una consulta en un consultorio médico. Los primeros valores son discretos porque existe una distancia entre las calificaciones, como 8.3 y 8.4, así que no podría ser la calificación 8.34 o 8.347.

4.1.2. Continua

Es aquella variable que puede tomar un número infinito no numerable de valores o puede asumir el número infinitamente grande de valores correspondientes a los puntos sobre un intervalo en una línea recta. Es decir, la palabra continua significa que procede sin interrupción y proporciona la clave para identificar a las variables aleatorias continuas. Una característica importante de esta variable es que si las mediciones u observaciones tienen un conjunto de valores que forman puntos sobre una línea sin interrupción o espacio entre ellos. Ejemplos: Peso de una persona, tiempo en llegar al centro de trabajo, ancho de una habitación, altura de una persona, presión de un neumático de automóvil, distancia (millas) entre Ciudades de Atlanta-Los Ángeles podría ser 2,254, 2,254.1, 2,254.162 y sucesivamente, dependiendo de la precisión del dispositivo de medición o la presión de un neumático (Libras por pulgada cuadrada o Psi) podría ser 28, 28.6, 28.62, 28.624, etcétera. Esto es debido a que puede tomar un valor de una cantidad infinitamente grande de valores dentro de ciertas limitaciones.

4.1.2.1. Función de Densidad

4.1.3. Espacios muestrales

Los espacios muestrales se clasifican en discretos y continuos. Una clasificación similar puede hacerse para variables aleatorias. En consecuencia, una Variable Aleatoria es Discreta si puede tomar cuando más un número infinito o infinito numerable o contable (numerable) de valores y es Continua si puede tomar cualquier valor infinito no numerable de valores en un intervalo dado. Si se organiza un conjunto de valores posibles de una variable aleatoria discreta, en una distribución de probabilidades, por lógica la distribución se denomina Distribución de Probabilidad Discreta.

4.1.4. Distribución

La distribución de probabilidades para una variable aleatoria describe cómo se distribuyen las probabilidades entre todos los valores de la variable aleatoria. La distribución de probabilidades se define por una función de probabilidades denotadas por f(X), que provee probabilidades a cada valor de la variable aleatoria.

Ejemplos:

a) En el espacio muestral anterior, los eventos $\{\beta_1\}$ a $\{\beta_8\}$ forman una partición de M. Suponga que a estos eventos les asocia las siguientes probabilidades:

$$\mathrm{P}(\{\beta_1\}) = \frac{1}{32};\ \mathrm{P}(\{\beta_2\}) = \mathrm{P}(\{\beta_3\}) = \mathrm{P}(\{\beta_4\}) = \frac{3}{32}$$

$$\mathrm{P}(\{\beta_5\}) = \mathrm{P}(\{\beta_6\}) = \mathrm{P}(\{\beta_7\}) = \frac{5}{32};\ \mathrm{P}(\{\beta_8\}) = \frac{7}{32}$$

A cada evento del espacio se le a asociado un número, que es el valor que toma una variable denotada por X. Así, por ejemplo, con el evento $\{\beta_1\}$ se asocia con el valor 0, tal que X toma el valor 0 con la misma probabilidad asignada a $\{\beta_1\}$:

$$\mathbf{P(X = 0)} = \mathrm{P}(\{\beta_1\}) = \mathbf{\frac{1}{32}}$$

$$\mathbf{P(X = 1)} = \mathrm{P}(\{\beta_2\}) \cup (\{\beta_3\}) \cup (\{\beta_4\}) = \mathrm{P}(\{\beta_2\}) + \mathrm{P}(\{\beta_3\}) + \mathrm{P}(\{\beta_4\}) = \frac{3}{32} + \frac{3}{32} + \frac{3}{32} = \mathbf{\frac{9}{32}}$$

$$\mathbf{P(X = 2)} = \mathrm{P}(\{\beta_5\}) \cup (\{\beta_6\}) \cup (\{\beta_7\}) = \mathrm{P}(\{\beta_5\}) + \mathrm{P}(\{\beta_6\}) + \mathrm{P}(\{\beta_7\}) = \frac{5}{32} + \frac{5}{32} + \frac{5}{32} = \mathbf{\frac{15}{32}}$$

$$\mathbf{P(X = 3)} = \mathrm{P}(\{\beta_8\}) = \mathbf{\frac{7}{32}}$$

Lo anterior se puede resumir de la siguiente forma respecto a Distribución de Probabilidades para el número de insectos muertos:

Cuadro 14. Distribución de probabilidades para insectos ([147])

Eventos	$\{\boldsymbol{\beta_1}\}$	$\{\boldsymbol{\beta_2}\},\{\boldsymbol{\beta_3}\},\{\boldsymbol{\beta_4}\}$	$\{\boldsymbol{\beta_5}\},\{\boldsymbol{\beta_6}\},\{\boldsymbol{\beta_7}\}$	$\{\boldsymbol{\beta_8}\}$
Valor de X	0	1	2	3
Probabilidad de X	$\frac{1}{32}$	$\frac{9}{32}$	$\frac{15}{32}$	$\frac{7}{32}$

La tabla anterior proporciona la distribución de probabilidades la variable aleatoria X definida sobre la partición $\{\beta_1\},\dots,\{\beta_8\}$ en el espacio muestral M. El hecho de que la información del experimento pueda resumirse al considerar únicamente los valores de X conduce a la noción de función de probabilidades de una variable aleatoria. Debido a que existen diferencias importantes en la presentación del concepto cuando la variable aleatoria es discreta o continua, en adelante se referirá por separado a estos casos.

4.1.5. Media, varianza y desviación estándar de una distribución de probabilidad

La media indica la ubicación central de los datos, mientras que la varianza describe su dispersión. De manera semejante, una distribución de probabilidad se resume indicando su media y su varianza. La media de una distribución de probabilidad se denota con la letra griega mu minúscula (μ) y la desviación estándar con la letra griega sigma minúscula (σ).

Media ($\boldsymbol{\mu}$). La media es un valor típico que sirve para representar una distribución de probabilidad. También es el valor promedio, a largo plazo, de la variable aleatoria. A la media de una distribución de probabilidad se le conoce como su "valor esperado". Esta media es un promedio ponderado en que los valores posibles se ponderan mediante sus probabilidades correspondientes de ocurrencia.

4.1.6. Valor Esperado o esperanza matemática

Sea X una variable con función probabilidad $f(X)$, sea $Y = g(X)$ una función real de la variable X. Entonces, el valor esperado o esperanza matemática de $g(X)$ se define como:

$$E(Y) = E[g(X)] = \sum_{X \in R_X} g(X_i) f(X_i)$$

[147] Fuente: Infante, G. S. y Zárate de L., G. P. 1984.

Ejemplos [148]:

a) Un juego consiste en tirar dos dados. Si la suma de sus caras es ≥ 11 se ganan \$200 USD. Si esta suma está comprendida entre 5 y 10 inclusive se ganan \$150 USD y, para cualquier otro resultado, no se gana nada. ¿Cuál es el valor esperado del premio?

El espacio muestral para el problema es $S = \{(1,1),\ (1,2),\ (1,3),\ \ldots,(1,6),\ (2,1),\ (2,2),\ (2,3),\ldots,(2,6),\ (3,1),\ (3,2),\ (3,3),\ldots,\ (3,6),\ \ldots,\ (6,6)\}$ Esto arroja 36 puntos muestrales. Todos estos tienen la misma probabilidad $\frac{1}{36}$. Se define la variable aleatoria X: suma de las dos caras. El rango de la variable X es $R_X = \{2,\ 3,\ 4,\ldots,7,7,\ 3,\ 4,\ 5,\ldots,\ 8,\ 4,\ 5,\ 6,\ldots,\ 12\}$. La distribución de probabilidad de la variable X es la siguiente:

Cuadro 15. Espacio muestral de eventos de un premio [149]

Rango de variable X (R_X)	Espacio Muestral de Eventos (S)	$f(X)$
2	$\{(1,1)\}$	$1/36$
3	$\{(1,2),\ (2,1)\}$	$2/36$
4	$\{(1,3),\ (2,2),\ (3,1)\}$	$3/36$
5	$\{(1,4),\ (2,3),\ (3,2),\ (4,1)\}$	$4/36$
6	$\{(1,5),\ (2,4),\ (3,3),\ (4,2),\ (5,1)\}$	$5/36$
7	$\{(1,6),\ (2,5),\ (3,4),\ (4,3),\ (5,2),\ (6,1)\}$	$6/36$
8	$\{(2,6),\ (3,5),\ (4,4),\ (5,3),\ (2,2)\}$	$5/36$
9	$\{(3,6),\ (4,5),\ (5,4),\ (6,3)\}$	$4/36$
10	$\{(4,6),\ (5,5),\ (6,4)\}$	$3/36$
11	$\{(5,6),\ (6,5)\}$	$2/36$
12	$\{(6,6)\}$	$1/36$

[148] Un conjunto de ejercicios se encuentra en carpeta "Valor esperado o Esperanza Matemática"
[149] Fuente: Infante, G. S. y Zárate de L., G. P. 1984.

Si se define la función premio como $g(X)$ entonces la distribución de probabilidad de la función premio es:

Cuadro 16. Rango de una variable de la función de un premio ([150])

Rango de variable X (R_X)	Función Premio $g(X)$	$f(g(X))$
2	0	$1/36$
3		$2/36$
4		$3/36$
5	150	$4/36$
6		$5/36$
7		$6/36$
8		$5/36$
9		$4/36$
10		$3/36$
11	200	$2/36$
12		$1/36$

Por lo tanto, el valor esperado del premio es:

$$\mathrm{E(Y)} = \mathrm{E[g(X)]} = \sum_{\mathrm{X}\in \mathrm{R_X}} \mathrm{g(X_i)f(X_i)}$$

$$= \left[\left(0 * \frac{1}{36}\right) + \left(0 * \frac{2}{36}\right) + \left(0 * \frac{3}{36}\right)\right]$$

$$+ \left[\left(150 * \frac{4}{36}\right) + \left(150 * \frac{5}{36}\right) + \left(150 * \frac{6}{36}\right) + \cdots + \left(150 * \frac{3}{36}\right)\right]$$

$$+ \left[\left(200 * \frac{2}{36}\right) + \left(200 * \frac{1}{36}\right)\right] = \$\mathbf{129.17}\ \boldsymbol{USD}$$

4.1.7. Función de probabilidades

Una Variable Aleatoria es Discreta si puede tomar cuando más un número infinito o infinito numerable o contable (numerable) de valores. En la inmensa mayoría de las situaciones prácticas, las variables aleatorias discretas representan conteos de alguna característica, por lo que en ocasiones sus distribuciones (probabilidades de tomar valores específicos) reciben el nombre de "Distribuciones de Conteos". Cuando el número de valores

[150] Fuente: Infante, G. S. y Zárate de L., G. P. 1984.

que puede tomar la variable aleatoria es grande, resulta impráctico presentar todos los valores posibles en una tabla y, por ello, es más conveniente describir su comportamiento probabilístico mediante una ecuación. Se adoptará la convención de denotar por X a una variable aleatoria y por x los valores que puede tomar. Tal que si se tiene una ecuación que genera las probabilidades correspondientes a dichos valores se escribe:

$$\mathrm{f_X(x) = P(X = x).\ Tal\ que\ para\ x = 2\ \Rightarrow f_X(2) = P(X = 2)}$$

La función de probabilidades de una variable aleatoria discreta X es el conjunto de pares ordenados $\{\mathrm{x},\ \mathrm{f_X(x) = P(X = x)}\}$ donde x es cada uno de los valores que puede tomar la variable aleatoria X y $f_X(x)$ la probabilidad asociada con el valor particular x. Puesto que el caso de una variable aleatoria discreta los valores posibles de X tienen correspondencia con una partición del espacio muestral, es claro que deben cumplirse las siguientes propiedades para $x,\ f_X(x)$:

- $\mathrm{x},\ \mathrm{f_X(x) \geq 0},\ \text{para todo valor x de X.}$
- $\sum \mathrm{f_X(x) = 1}$

Donde la suma es para todos los valores x que puede tomar la variable aleatoria X. Asimismo, la función de probabilidades de una variable aleatoria discreta X debe cumplir las siguientes condiciones:

- $\mathrm{f_X(x) \geq 0},\ \text{para todo valor x de X.}$
- $\sum \mathrm{f_X(x) = 1}$

La media de una distribución de probabilidad discreta es:

$$\mu^{151} = \sum [\mathrm{x\ P(X)}]$$

La Varianza y Desviación Estándar de una distribución de probabilidad discreta es:

$$\sigma^2 = \sum (\mathrm{X} - \mu)^2 \mathrm{P(X)}$$

4.1.8. Representación gráfica de funciones de probabilidad

A menudo es conveniente representar gráficamente las funciones de probabilidad. Las dos formas más usadas para su representación gráfica son diagramas de puntos e histogramas de probabilidades.

[151] $\mathrm{P(X)}$ es probabilidad de cada valor que puede tomar la variable aleatoria X. En otras palabras, se multiplica cada valor de X por su respectiva probabilidad de ocurrencia y, luego, se suman estos productos

Ejemplos[152]:

a) Sea X una variable aleatoria con la siguiente función de probabilidades:

x	-2	-1	0	1	2	3	4
$f_X(x)$	0.1	0.1	0.3	0.2	0.1	0.1	0.1

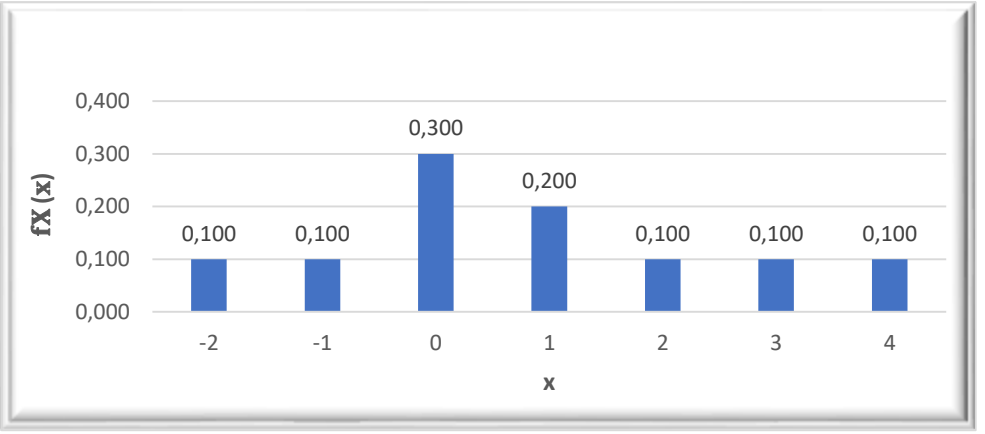

Figura 33. Representación gráfica de funciones de probabilidad ([153])

La forma más común de construir un histograma de probabilidades es elegir rectángulos de base unitaria, cuyo centro es cada uno de los valores que puede tomar la variable aleatoria X y cuya altura es $\mathrm{fX(x)}$ para el valor en cuestión y el área total del histograma es uno. Esta asociación de las probabilidades con áreas es esencial para presentar distribuciones de variables aleatorias de tipo continuo.

Función de Probabilidades de una Variable Aleatoria Continua. Una variable aleatoria continua es aquella variable cuyos valores pueden ser cualquier número real dentro de un intervalo cualquiera.

Función de Densidad. Una función se dice que es de densidad si cumple las dos propiedades:

- $\mathrm{f(x)} \geq 0$
- $\int_{-\infty}^{+\infty} \mathrm{f(x)dx} = 1$

[152] La desviación estándar (σ) se determina tomando la raíz cuadrada de σ^2 $(\sigma = \sqrt{\sigma^2})$. Ejemplos y un conjunto de ejercicios se encuentran en carpeta "Función de probabilidades"

[153] Fuente: Elaboración propia (2018) con base en datos de Infante, G. S. y Zárate de L., G. P. 1984

Además, verifica que dado $a < b$, se tiene que:

$$\mathrm{P}[\mathrm{a} \leq \mathrm{X} \leq \mathrm{b}] = \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\mathrm{dx}$$

La probabilidad entre dos valores dados a y b, es el área bajo la curva entre los puntos a y b.

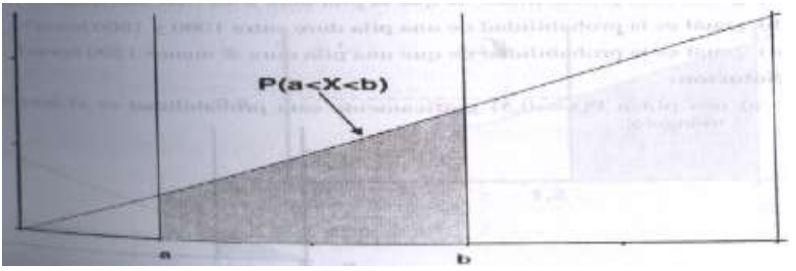

Figura 34. Representación gráfica de valores a y b [154]

Por ser f una función integrable, la probabilidad de un punto es nula:

$$\mathrm{P}[\mathrm{X} = \mathrm{a}] = \mathrm{P}[\mathrm{a} \leq \mathrm{X} \leq \mathrm{a}] = \int_{\mathrm{a}}^{\mathrm{a}} \mathrm{f}(\mathrm{x})\mathrm{dx} = 0$$

Al calcular la probabilidad de un intervalo no afectará nada el que esté abierto o cerrado por cualquiera de sus extremos, pues estos son puntos y, por tanto, de probabilidad nula:

$$\mathrm{P}[\mathrm{a} \leq \mathrm{X} \leq \mathrm{b}] = \mathrm{P}[\mathrm{a} < \mathrm{X} \leq \mathrm{b}] = \mathrm{P}[\mathrm{a} \leq \mathrm{X} < \mathrm{b}] = \mathrm{P}[\mathrm{a} < \mathrm{X} < \mathrm{b}]$$

La función de distribución de la Variable Aleatoria Continua, F, se define como:

$$\mathrm{F}: \Re \Rightarrow [0,1]$$

$$\mathrm{x} \Rightarrow \mathrm{F}(\mathrm{x}) = \mathrm{P}[\mathrm{X} \leq \mathrm{x}] = \int_{-\infty}^{\infty} \mathrm{f}(\mathrm{t})\mathrm{dt}$$

Ejemplos:

a) La vida útil de horas de cierta marca de pilas tiene la siguiente función de densidad:

- ¿Cuál es la probabilidad que la pile dure a lo más 500 Hr?
- $\mathrm{P}(\mathrm{X} \leq 0.5) = \frac{(0.5*0.25)}{2} = \frac{(500*2.5)}{2*100} = 0.0625\ (6.25\ \%)$

[154] Fuente: Daza P. G. F. 2006

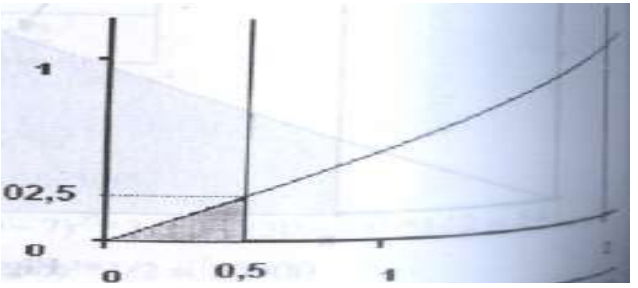

Figura 35. Representación gráfica de probabilidad que la pila dure a lo más 500 Hr ([155])

- ¿Cuál es la probabilidad que una pila dure entre 1000 y 1,800 Hr?
- $$P(1 \leq X \leq 1.8) = (0.8 * 0.5) + \frac{(0.8*0.4)}{2} = \frac{\left((800*0.5)+\frac{(800*0.4)}{2}\right)}{10} = 0.56\ (56\ \%)$$

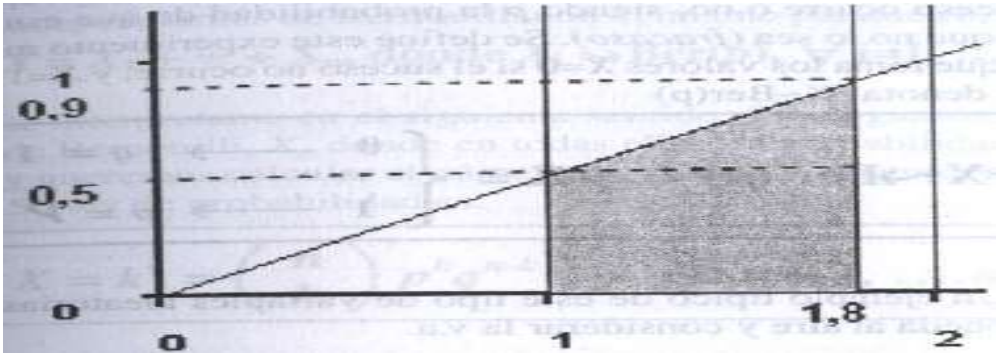

Figura 36. Representación gráfica de probabilidad que pila dure entre 1000 y 1,800 Hr ([156])

- ¿Cuál es la probabilidad que una pila dure al menos 1,200 Hr?
- $$P(X \geq 1.2) = 1 - 1.2 * \frac{(0.6)}{2} = \left(2 - 1.2 * \left(\frac{0.6}{2}\right)\right) - 1 = 0.64\ (64\ \%)$$

La primera dificultad que se enfrenta es que no se puede construir una tabla con valores posibles de la variable, pues el número de valores que puede tomar es infinito. En consecuencia, la única forma que se tiene de caracterizar la distribución de probabilidades es mediante una ecuación. La distribución de probabilidades de una variable aleatoria continua puede visualizarse gráficamente como una "forma límite" del histograma de frecuencias relativas.

[155] Fuente: Daza P. G. F. 2006
[156] Fuente: Daza P. G. F. 2006

Ejemplo:

a) En el último histograma se puede dar respuesta a algunas preguntas que no es posible en el histograma de frecuencias relativas para duración de 50 pistas de disco con intervalos de clase en un minuto; por ejemplo, la frecuencia relativa de pistas con duración entre 2 y 2.5 minutos es 0.212 (21.2 %). Puede observarse también que si se construye un polígono de frecuencias tiene ángulos más "suaves" que el que pueda construirse en el primer caso. Si se continúa incrementando el número de observaciones se puede disminuir simultáneamente la anchura de los intervalos. Sin embargo, este proceso no puede continuar indefinidamente, pues la precisión de las mediciones es limitada.

Puesto que se ha interpretado la probabilidad como la forma estabilizada de frecuencia relativa, la "forma límite" de la siguiente figura representa la distribución de probabilidades de la característica X en estudio. Tal que, el área bajo $f_X(x)$, delimitada por esta y dos líneas verticales levantadas sobre los puntos a y b $(a < b)$ es la probabilidad que la variable aleatoria X tome un valor entre a y b.

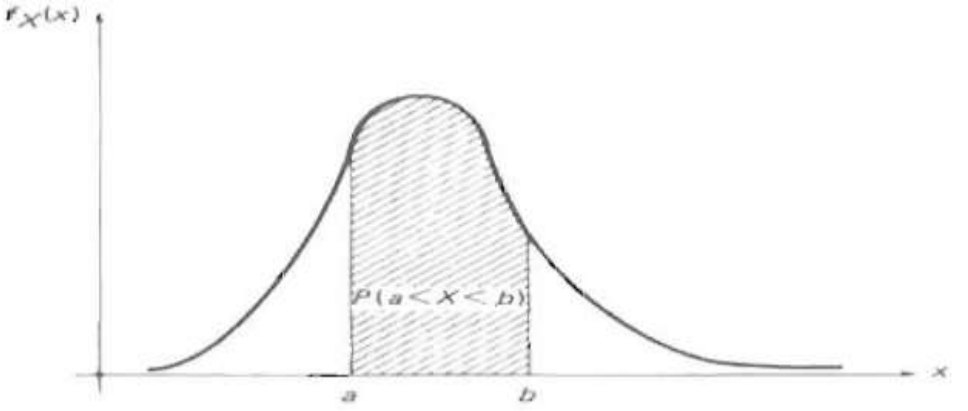

Figura 37. Distribución de probabilidades de la característica X en estudio [157]

Con base en lo anterior, se establecen las propiedades que debe cumplir una función para ser una función de densidad de probabilidades de una variable aleatoria continua X, denotada por $f_X(x)$:

[157] Fuente: www.google.com/search?client=firefox-b&biw=1366&bih=635&tbm=isch&sa=1&ei=bIG0W_PhC4uIggfgvZvwDg&q=Probabilidad+bajo+campana+de+gauss&oq=Probabilidad+bajo+campana+de+gauss&gs_l=img.3...1251139.1259662.0.1260058.56.25.0.0.0.0.642.642.5-1.1.0....0...1c.1.64.img..56.0.0.0...0.fecuth8BXlA#imgrc=sCXCeyClejQp4M:

- La curva siempre se encuentra sobre el eje de las abcisas (eje x): $f_X(x) \geq 0$.
- El área total bajo $f_X(x)$es $= 1$.
- El área delimitada por dos líneas verticales levantadas sobre los puntos a y b $(a < b)$, así como la curva, es la probabilidad que X tome un valor entre a y b.

Una diferencia fundamental entre las distribuciones de variables aleatorias discretas y continuas es que, en el caso de discretas, la probabilidad total se encuentra distribuida en un número finito o infinito contable (numerable) de puntos. Como consecuencia, si la variable puede tomar un valor k, puede asignarse un valor a $P(X = k)$ con la condición de que la suma de probabilidades para todos los puntos sea 1. Si se trata de una variable aleatoria continua, ésta puede tomar cualquier valor en un intervalo dado y si se asigna una probabilidad, por pequeña que sea, a cada uno de estos valores, la probabilidad total no sumará 1. En conclusión, para una variable aleatoria continua, la probabilidad asociada con un punto cualquiera es cero $P(X = k) = 0$ para cualquier valor que tome la variable. Entonces, si X es continua $f_X(x) \neq P(X = k)$. Aunque, se puede interpretar diciendo que, dada la precisión de los instrumentos de medición, no se puede distinguir entre las pistas con duración de 1.8 minutos y las que tienen 1.8001 o 1.7999. Entonces, en lo que se está interesado es la probabilidad del intervalo $\{1.7999,\ 1.8001\}$.

Aunque en general, es necesario evaluar una integral para calcular probabilidades en una variable aleatoria continua, en este caso el problema se simplifica, puesto que la probabilidad de un intervalo cualquiera $[a, b]$ es el área de un rectángulo con base $b - a$ y altura 1. Se puede calcular:

1. $P\left(X \leq \frac{1}{2}\right) = P\left(0 \leq X \leq \frac{1}{2}\right) = \left(\frac{1}{2}\right)(1) = \frac{1}{2}$
2. $P\left(X \leq \frac{1}{3}\right) = P\left(0 \leq X \leq \frac{1}{3}\right) = P\left(X < \frac{1}{3}\right) = \left(\frac{1}{3}\right)(1) = \frac{1}{3}$
3. $P\left(\frac{1}{3} < X \leq 1\right) = \left(\frac{2}{3}\right)(1) = \frac{2}{3}$
4. $P(0.2 \leq X < 0.8) = (0.6)(1) = 0.6$
5. $P(X \geq 0.4) = P(0.4 \leq X \leq 1) = (0.6)(1) = 0.6$
6. $P(X \geq 1) = 0$
7. $P(X \leq 0) = 0$

Suponga que una variedad de maíz puede producir plantas con 0, 1, 2 y 3 mazorcas. Se sabe además que las probabilidades que se tengan plantas con esos números de mazorcas son 0.1, 0.7, 0.1 y 0.1, respectivamente. Haciendo una tabla sería:

x	0	1	2	3	**Σ**
$f_X(x)$	0.1	0.7	0.1	0.1	**1.0**

Con esta función de probabilidades:

$$P(X \leq 1) = P(X = 0) + P(X = 1) = f_X(0) + f_X(1) = 0.1 + 0.7 = 0.8$$

En cambio:

$$P(X < 1) = P(X - 0) = f_X(0) = 0.1$$

Entonces:

$$P(X \leq 1) \neq P(X < 1)$$

Debido a que la variable aleatoria discreta del ejemplo tiene una masa probabilística de 0.7 concentrada en $X = 1$.

4.1.9. Distribución Muestral

Distribución Muestral es una función de probabilidad asociada a un estimador o valor estadístico, que se genera con todas las muestras de tamaño n que se pueden extraer de una población. Por ejemplo $\overline{X}$, $\overline{X}_1 - \overline{X}_2$, S^2, etc. son estimadores que tienen cada distribución de probabilidad particular, así se tendrá la distribución de la media muestral, distribución de la diferencia de medias muestrales, distribución de varianza muestral, etcétera.

4.1.9.1. Media

La distribución muestral de la media es la distribución de las medias de todas las muestras del mismo tamaño seleccionadas aleatoriamente en una población. Sea $X \sim N(\mu, \sigma^2)$, si se extrae una muestra aleatoria de tamaño n de una población entonces la media muestral $\overline{X}$ se distribuye así:

A. $X \sim N(\mu, \sigma_{\bar{x}}^2)$ es decir $\mu_{\bar{x}} = \mu$ y $\sigma_{\bar{x}}^2 = \left\{\frac{\sigma^2}{n}\right\}$ si la población es infinita o es finita con tamaño N, selección con o sin reemplazo y $\frac{n}{N} < 0.05$.

B. $\sigma_{\bar{x}}^2 = \left\{\frac{\sigma^2}{n} * \frac{N-n}{N-1}\right\}$ si la población es finita con tamaño N y la selección es sin reemplazo con $\frac{n}{N} \geq 0.05$.

C. Tal que $\sigma_{\bar{x}}$ se denomina error estándar poblacional de la media muestral. En caso de que la desviación estándar poblacional, σ, sea desconocida se le estimará mediante la desviación estándar muestral S o, en este caso, error estándar muestral de la media muestral $(S_{\bar{x}})$.

D. Se sabe que $Z = \frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, si la población es infinita o selección con reemplazo o si $\frac{n}{N} < 0.05$ y $Z = \frac{\bar{X}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}*\frac{N-n}{N-1}}$ si la población es finita y $\frac{n}{N} \geq 0.05$.

4.1.9.2. Diferencia de Medias Muestrales $(^{158})$

Si se extraen dos muestras aleatorias de tamaño n_1 y n_2 de manera independiente de dos poblaciones cuyas distribuciones de probabilidad son $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$. Entonces, las diferencias de medias muestrales de ambas muestras $\bar{X}_1 - \bar{X}_2$ tienen la siguiente distribución de probabilidad:

$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2,\ \sigma^2_{\bar{x}_1-\bar{x}_2}$; es decir, $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$ y $\sigma^2_{\bar{x}_1-\bar{x}_2} = \left\{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}\right\}$

Si ambas poblaciones son infinitas o finitas con tamaño N_1 y N_2, selecciones con reemplazo o sin reemplazo y $\frac{n_1}{N_1} < 0.05, \frac{n_2}{N_2} < 0.05$.

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \left\{\frac{\sigma_1^2}{n_1} * \frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2} * \frac{N_2 - n_2}{N_2 - 1}\right\}$$

Si las poblaciones son finitas con tamaño N_1 y N_2, selecciones sin reemplazo y $\frac{n_1}{N_1} \geq 0.05, \frac{n_2}{N_2} \geq 0.05$.

4.1.10. Distribución de funciones

4.1.10.1. Distribución Chi Cuadrado ($\chi 2$) (Bondad de ajuste, independencia y homogeneidad)

Los resultados obtenidos de las muestras no siempre coinciden exactamente con resultados teóricos esperados según las reglas de probabilidad. Por ejemplo, aunque de acuerdo con las consideraciones teóricas en 100 lanzamientos de una moneda se esperarían 50 caras y 50 cruces, es raro que se obtengan exactamente estos resultados.

Ejemplo:

[158] Ejemplos y un conjunto de ejercicios se encuentran en carpeta "Diferencias de medias muestrales"

a) En una muestra determinada se observa la ocurrencia de un conjunto de eventos $E_1, E_2, E_3, E_4, \ldots, E_k$ con frecuencias observadas $fo_1, fo_2, fo_3, fo_4, \ldots, fo_k$ y que, según reglas de la probabilidad, se esperaría que estos eventos ocurrieran con frecuencias esperadas o teóricas $fe_1, fe_2, fe_3, fe_4, \ldots, fe_k$.

Eventos	$\mathbf{E_1}$	$\mathbf{E_2}$	$\mathbf{E_3}$	$\mathbf{E_4}$	...	$\mathbf{E_k}$
Frecuencias observadas	fo_1	fo_2	fo_3	fo_4	...	fo_k
Frecuencias esperadas	fe_1	fe_2	fe_3	fe_4	...	fe_5

Una medida de discrepancia entre frecuencias observadas y frecuencias esperadas la proporciona el estadístico $\chi 2$:

$$\chi 2 = \frac{(fo_1 - fe_1)^2}{fe_1} + \frac{(fo_2 - fe_2)^2}{fe_2} + \frac{(fo_3 - fe_3)^2}{fe_3} + \frac{(fo_4 - fe_4)^2}{fe_4} + \cdots + \frac{(fo_k - fe_k)^2}{fe_k} = \sum_{j=1}^{k} \frac{(fo_j - fe_j)^2}{fe_j} \quad \text{Ecuación a}$$

Una expresión equivalente a esta última ecuación es:

$$\chi 2 = \sum \frac{fo_j^2}{fe_j} - N \quad \text{Ecuación b}$$

Tal que, si la frecuencia total es N: $\sum fo_j = \sum fe_j = N$ Ecuación c.

Si $\chi 2 = 0$, las frecuencias observadas y frecuencias teóricas coinciden exactamente; en tato, que si $\chi 2 > 0$, la coincidencia no es exacta. Donde, en cuanto mayor sea el valor $\chi 2$, mayor la discrepancia entre frecuencias observadas y esperadas. Una variable aleatoria X tiene una distribución Chi o Ji Cuadrado si su función de densidad de probabilidad está dada por:

$$f(x) = \left\{ \begin{matrix} \frac{1}{2^{v/2}\boldsymbol{\Gamma}(v)} X^{v/2-1} e^{x/2}, \; X \geq 0 \\ 0, \; X < 0 \end{matrix} \right\}$$

ó

$$f(X^2) = \frac{1}{2^{v/2}\boldsymbol{\Gamma}(v/2)} (X^2)^{(v/2)-1} e^{-X^2/2} \quad \text{para } X^2 > 0$$

$$\Gamma_{(\text{Letra griega gamma})}(a) = (a-1)! \text{ llamada Función Gamma de a}$$

Donde v es número de grados de libertad. No obstante, la distribución Chi Cuadrado es un caso particular de la distribución Gamma, cuya función de densidad está dada por:

$$f(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{\Gamma(k)} \; t \geq 0$$

Donde $\mathrm{G(k)}$ es función gamma de k. Los valores correspondientes de los parámetros 1 y k son $1 = 1/2$ y $\mathrm{k} = \mathrm{n}/2$. Tal que, el valor esperado y la varianza de la distribución Chi Cuadrado están dados por $\mathrm{E}\ (\mathrm{X}) = \mathrm{n}, \mathrm{V(X)} = 2\mathrm{n}$. Finalmente, el gráfico de una distribución Chi Cuadrado con pocos grados de libertad presenta un gran sesgo.

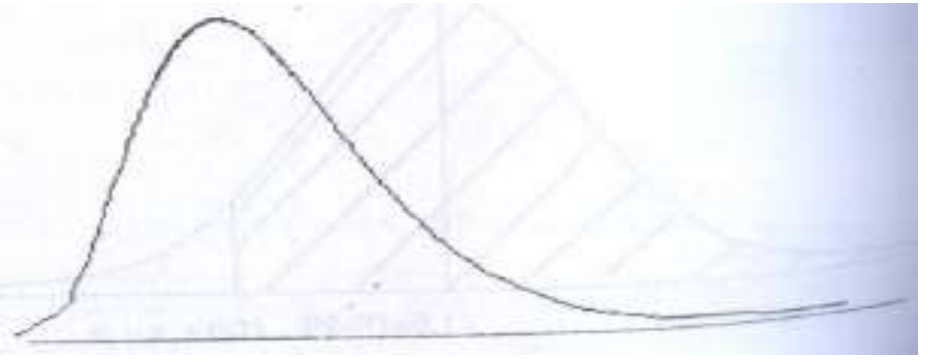

Figura 38. Distribución Chi Cuadrado con pocos grados de libertad ([159])

La distribución muestral de $\boldsymbol{\chi 2}$ se puede aproximar con bastante exactitud mediante la distribución Chi o Ji Cuadrada:

$$\mathrm{Y} = \mathrm{Y_0}(\chi 2)^{1/2(\mathrm{v}-2)}\mathrm{e}^{-1/2\boldsymbol{\chi 2}} = \mathrm{Y_0}\mathrm{X}^{\mathrm{v}-2}\mathrm{e}^{-1/2\boldsymbol{\chi 2}}$$

Si $\mathrm{fe} \geq 5$. La aproximación mejora cuanto mayor sean estos valores. El número de grados de libertad (v) es:

1. $v = \mathrm{k} - 1$ si las frecuencias esperadas pueden calcularse sin tener que estimar parámetros poblacionales a partir de estadísticos muestrales, pues conociendo $k - 1$ se las frecuencias esperadas, queda determinada la frecuencia restante.
2. $v = \mathrm{k} - 1 - \mathrm{m}$ si las frecuencias esperadas sólo pueden calcularse estimando m parámetros poblacionales a partir de estadísticos muestrales.

Propiedades de las Distribuciones Chi o Ji-Cuadrada:

1. Los valores $\chi 2\ \mathrm{son}\ \geq 0$.
2. La forma de una distribución $\chi 2$ depende de los grados de libertad. Por lo tanto, hay un número infinito de distribuciones $\chi 2$.
3. El área bajo la curva Chi o Ji-Cuadrada, sobre el eje horizontal, es 1.
4. Presenta distribución normal Z.

[159] Fuente: Daza P. G. F. 2006

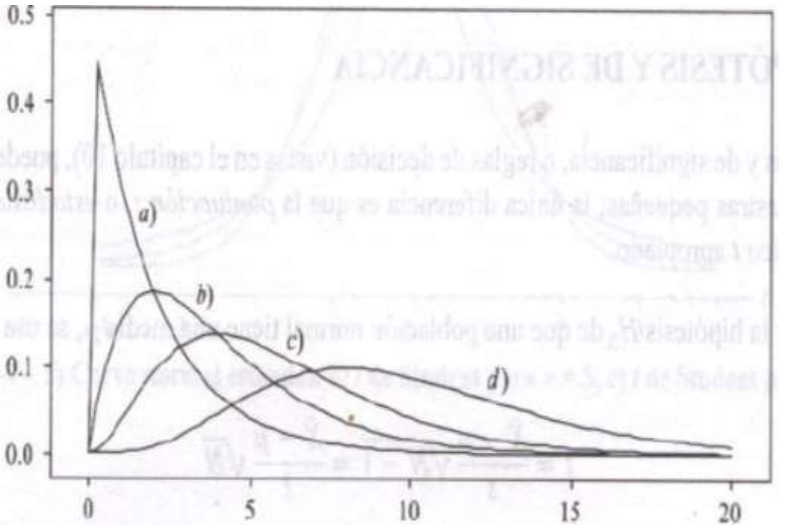

Figura 39. Distribución Chi Cuadrado con a) 2, b) 4, c) 6 y d) 10 grados de libertad [160]

Teoremas:

- Si Z es una variable aleatoria con distribución normal $(0,\ 1)$; entonces Z^2 tiene una distribución $\chi 2$ con un grado de libertad $(v = 1)$.
- Si $\mathrm{X_1},\ \mathrm{X_2},\ \mathrm{X_3},\ \mathrm{X_4}, \ldots,\ \mathrm{X_n}$ es un conjunto de n variables independientes e idénticamente distribuidos normalmente, $\mathrm{N}(0,\ 1)$, entonces tiene una distribución Chi Cuadrado con n grados de libertad.
- Si $\mathrm{X_1},\ \mathrm{X_2},\ \mathrm{X_3},\ \mathrm{X_4}, \ldots,\ \mathrm{X_k}$ es un conjunto de k variables independientes con distribuciones Chi Cuadrado con $\mathrm{v_1},\ \mathrm{v_2},\ \mathrm{v_3},\ \mathrm{v_4}, \ldots,\ \mathrm{v_k}$ grados de libertad, respectivamente. Entonces, la variable aleatoria $\mathrm{X} = \mathrm{X_1} + \mathrm{X_2} + \mathrm{X_3} + \mathrm{X_4} + \cdots + \mathrm{X_k}$ tiene una distribución Chi Cuadrado con $\mathrm{m} = \mathrm{v_1},\ \mathrm{v_2},\ \mathrm{v_3},\ \mathrm{v_4}, \ldots,\ \mathrm{v_k}$ grados de libertad.
- Si $\overline{\mathrm{X}}$ y S^2 son media y varianza de una muestra tomada de una población normal con media μ y varianza σ^2 entonces:

a) $\overline{\mathrm{X}}$ y S^2 son independientes**.**

b) La variable aleatoria $\frac{(\mathrm{n}-1)\mathrm{S}^2}{\sigma^2}$ tiene una distribución Chi Cuadrado con $\mathrm{n} - 1$ grados de libertad, pues si considera la siguiente suma $\sum_{i=1}^{n}(X_i - \mu)^2$ puede representarse de la siguiente forma, sumando y restando $\bar{X}$:

[160] Fuente: Daza P. G. F. 2006

c) $\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2=\sum_{i=1}^{n}[(X_i-\bar{X})^2+2(X_i-\bar{X})(\bar{X}-\mu)+(\bar{X}-\mu)^2]=\sum_{i=1}^{n}[(X_i-\bar{X})^2+n(\bar{X}-\mu)^2]$

Dado que $\sum_{i=1}^{n}(X_i-\bar{X})=0$. Si se divide ambas expresiones por σ^2:

$$\sum_{i=1}^{n}\left[\frac{X_i-\mu}{\sigma}\right]^2=\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{\sigma^2}+n\left(\frac{\bar{X}-\mu}{\sigma}\right)$$

Tal que si se multiplica y divide la primera parte de la expresión derecha por $(n-1)$:

$$\sum_{i=1}^{n}\left[\frac{X_i-\mu}{\sigma}\right]^2=\frac{(n-1)S^2}{\sigma^2}+\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$$

Tal que, $\sum_{i=1}^{n}\left[\frac{X_i-\mu}{\sigma}\right]^2$ sigue una distribución Chi Cuadrado con n grados de libertad y el término $\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2$ sigue una igual, pero con un grado de libertad. En consecuencia, la expresión $\frac{(n-1)S^2}{\sigma^2}$ sigue la misma distribución, pero con $n-1$ grados de libertad.

En la práctica, las frecuencias esperadas se calculan basándose en la hipótesis H_0. Si de acuerdo con esta hipótesis el valor calculado para $\chi 2$, mediante ecuaciones a, b y c, es mayor a algún valor crítico ($X^2_{95\ o\ \alpha=0.05}$ o $X^2_{99\ o\ \alpha=0.01}$), se concluye que las frecuencias observadas difieren en forma significativa de las frecuencias esperadas y no se acepta o se rechaza H_0 al correspondiente nivel de significancia; sino es así, no se rechaza H_0. A este procedimiento se le conoce como Prueba Chi o Ji Cuadrada de hipótesis o de significancia

(Si X_c^2 $> X^2_{95\ o\ \alpha=0.05}$ o $X^2_{99\ o\ \alpha=0.01}$: No se acepta H_0. Por lo tanto, las frecuencias observadas difieren significativamente)

(Si i X_c^2 $< X^2_{95\ o\ \alpha=0.05}$ o $X^2_{99\ o\ \alpha=0.01}$: No se rechaza H_0. Por lo tanto, las frecuencias observadas no difieren significativamente)

Es necesario notar que hay que tener desconfianza de aquellas circunstancias en las que $\chi 2$ tenga un valor demasiado cercano a cero, pues es raro que exista una coincidencia tan buena entre frecuencias observadas y frecuencias esperadas. Para examinar tales situaciones se determina si el valor obtenido para $\chi 2$ es $< X^2_{95\ o\ \alpha=0.05}$ o $X^2_{99\ o\ \alpha=0.01}$, en cuyo caso se decide que a los niveles de significancia 0.05 o 0.01 la coincidencia es demasiado buena. La prueba chi o ji cuadrada puede emplearse para determinar qué tan bien se ajustan una distribución teórica (distribución normal o binominal, por ejemplo) a una distribución empírica (obtenida a partir de datos muestrales).

Ejemplo:

a) Un par de dados se lanzan 500 veces (números observados) y las sumas de las caras que caen hacia arriba son:

Suma	2	3	4	5	6	7	8	9	10	11	12
Observada	15	35	49	58	65	76	72	60	35	29	6

Los números esperados, si el dado no está cargado, se determinan a partir de la distribución de x:

x	2	3	4	5	6	7	8	9	10	11	12
P(x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Tablas de contingencia:

Eventos	$\mathbf{E_1}$	$\mathbf{E_2}$	$\mathbf{E_3}$	$\mathbf{E_4}$	…	$\mathbf{E_k}$
Frecuencias observadas	fo_1	fo_2	fo_3	fo_4	…	fo_k
Frecuencias esperadas	fe_1	fe_2	fe_3	fe_4	…	fe_k

Las frecuencias observadas ocupan un solo renglón se les llama tablas de clasificación en un solo sentido. Como el número de columnas es k, se les llama tablas de 1*k (leídas como "1 por k"). Entonces, se obtienen tablas de clasificación en dos sentidos o tablas h*k, en que las frecuencias observadas ocupan h renglones y k columnas. A estas tablas se les llama "tablas de contingencia". Asimismo, en una tabla de contingencia h*k, para cada frecuencia observada hay una frecuencia esperada (o teórica), que se calcula basándose en algunas hipótesis y sujetándose a las reglas de probabilidad. A las frecuencias que ocupan las celdas de una tabla de contingencia se les llama frecuencias de celda. Al total de las frecuencias de un renglón o de una columna se la llama frecuencia marginal. Para investigar el grado de coincidencia entre frecuencias observadas y esperadas se estima con el estadístico:

$$\chi 2 = \sum \frac{fo_j^2}{fe_j} - N$$

Tal que, la suma se realiza sobre todas las celdas de la tabla de contingencia y donde los símbolos fo_j y fe_j representas las frecuencias observada y esperada en la celda j, respetivamente. Asimismo, contiene hk términos. La suma de todas las frecuencias observadas, que se denota como, es igual a todas las sumas de las frecuencias esperadas (Ecuación b). Las frecuencias esperadas se establecen basándose en la hipótesis nula (H_0) de que se trate. Una de la hipótesis más empleada es que las dos clasificaciones son

independientes una de otra. Las tablas de contingencia pueden extenderse a dimensiones mayores. Así, se pueden tener, por ejemplo, tablas h*k*1, en las que hay tres clasificaciones.

Cuando a datos discretos se aplican fórmulas para datos continuos es necesario hacer una corrección por continuidad. Para el empleo de la distribución Chi o Ji Cuadrada hay una corrección que consiste en reescribir la ecuación a la Corrección de Yates:

$$X^2_{\text{Corregida}} = \frac{(|fo_1 - fe_1| - 0.5)^2}{fe_1} + \frac{(|fo_2 - fe_2| - 0.5)^2}{fe_2} + \frac{(|fo_3 - fe_3| - 0.5)^2}{fe_3} + \frac{(|fo_4 - fe_4| - 0.5)^2}{fe_4} + \cdots + \frac{(|fo_k - fe_k| - 0.5)^2}{fe_k}$$

En general, esta corrección sólo se hace cuando el número de grados de libertad (v=1). Cuando se tienen muestras grandes, se tiene prácticamente el mismo resultado que con χ2 no corregida, pero cerda de los valores críticos pueden surgir dificultades. Cuando se tienen muestras pequeñas, donde cada una de las frecuencias esperadas está entre 5 y 10, quizá sea mejor comparar ambos valores χ2 (corregido y no corregido). Si ambos valores conducen a la misma conclusión respecto a la hipótesis, por ejemplo, a 0.05, es raro que se encuentren dificultades. Si ambos valores conducen a conclusiones diferentes se puede recurrir a aumentar el tamaño de la muestra o, si es posible, se pueden usar métodos de probabilidad en que se use la distribución multinomial.

Como se hace en la distribución normal y distribución t pueden definirse límites de confianza 95%, 99% u otros límites empleando la tabla de distribución X^2. De esta manera puede estimarse la desviación estándar poblacional σ en términos de desviación estándar muestral dentro de determinados límites de confianza. Por ejemplo, si $X^2_{0.025}$ y $X^2_{0.975}$ son valores de X^2(valores críticos), tales que 2.5% del área se encuentra repartida en ambas colas de la distribución, entonces el intervalo de confianza de 95% es:

$$X^2_{0.025} < \frac{NS^2}{\sigma^2} < X^2_{0.975}$$

De donde se ve que puede estimarse que σ se encuentra en intervalo:

$$\frac{s\sqrt{N}}{X^2_{0.975}} < \sigma < \frac{s\sqrt{N}}{X^2_{0.025}}$$

Con 95% de confianza. De manera similar, se pueden encontrar otros intervalos de confianza. Los valores $X^2_{0.025}$ y $X^2_{0.975}$ representan los percentiles 2.5 y 97.5, respectivamente.

Si se tienen valores grandes de $v(v \geq 30)$, se puede usar el hecho que $\left(\sqrt{2X^2} - \sqrt{2v-1}\right)$ se aproxima mucho a una distribución normal o media 0 y desviación estándar 1. Por lo tanto, las tablas para la distribución normal pueden emplearse cuando $v(v \geq 30)$. Si X_p^2 y z_p son los percentiles p de la distribución Chi o Ji Cuadrada y de la distribución normal, respectivamente se tiene:

$$X_p^2 = \frac{1}{2}\left(z_p + \sqrt{2v-1}\right)^2$$ [161]

4.1.10.2. Distribución T de Student

William Sealy Gossett, 11 de junio de 1876–16 de octubre de 1937, fue un estadístico, mejor conocido por su sobrenombre literario "Student". Gossett publicó el error probable de una media y casi todos sus artículos usando el pseudónimo Student en la publicación Biometrika creada por Pearson. Sin embargo, fue Ronald Aylmer Fisher quien apreció la importancia de los trabajos de Gossett sobre muestras pequeñas, tras recibir correspondencia de Gossett en la que le decía le envío una copia de las Tablas de Student. Fisher creyó que Gossett había efectuado una "revolución lógica". Irónicamente la estadística t por la que Gossett es famoso fue realmente creación de Fisher. La estadística de Gossett era z = t/√(n - 1). Fisher introdujo la forma t debido a que se ajustaba a su teoría de grados de libertad. Fisher es responsable también de la aplicación de la distribución t a la regresión. La distribución t-Student se construye como un cociente entre una normal y la raíz de una X^2 independientes (Bioestadística, 2009. Pp. 155). De modo preciso, se llamará distribución t-Student con n grados de libertad, t_n a la de una variable aleatoria T: $t = \frac{\overline{X}-\mu}{s}\sqrt{N-1} = \frac{\overline{X}-\mu}{\hat{s}\sqrt{N}}$ siendo análogo (semejante) al estadístico $z = \frac{\overline{X}-\mu}{\sigma/\sqrt{N}}$. Si se consideran muestras de tamaño N extraídas de una población normal (o aproximadamente normal) cuya media es μ y si para cada muestra se calcula t, usando la media muestral $\overline{X}$ y la desviación estándar muestral s o $\hat{s}$ se obtiene la distribución muestral t:

[161] Ejemplos y un conjunto de ejercicios se encuentran en carpeta "$\chi 2$"

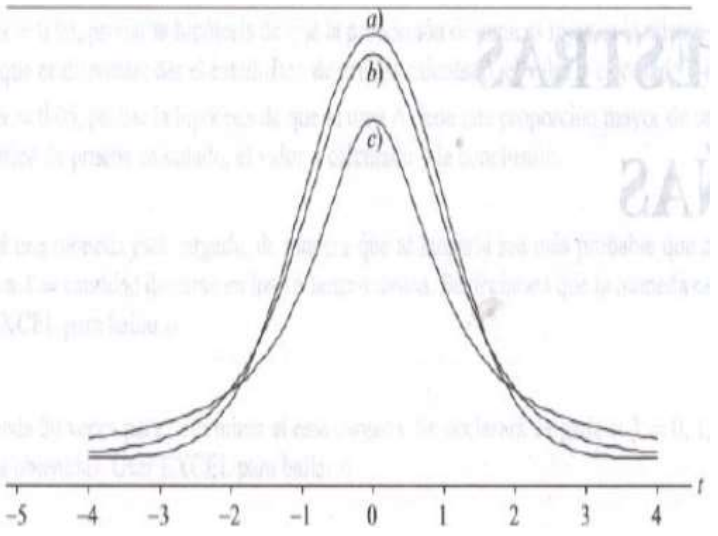

Figura 40. a) Curva normal estándar, b) t de Student con $\mathbf{v = 5}$ y c) t de Student con $\mathbf{v = 1}$ [162]

O, en otras palabras, sean Y y Z dos variables aleatorias independientes. Y con una distribución Chi cuadrado con v grados de libertad y Z con una distribución normal estándar $(0,1)$:

$$t = \frac{Z}{\sqrt{Y/v}} = \frac{\frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{(n-1)}}} = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$$

Donde la frecuencia de t está dada por:

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v \Gamma(v/2)}}\left(1 - \frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)} \quad -\infty < t < +\infty$$

La distribución t-Student está dada por:

$$Y = \frac{Y_0}{\left(1 + \frac{t^2}{N-1}\right)^{N/2}} = \frac{Y_0}{\left(1 + \frac{t^2}{v}\right)^{(v+1)/2}}$$

[162] Fuente: Daza P. G. F. 2006

Donde Y_0 es una constante que depende de N, tal que el área total bajo la curva sea 1 y donde a la constante $v = (N - 1)$ se le conoce como número de grados de libertad (v letra griega nu). Si los valores v o de N son grandes $(N \geq 30)$, la curva se aproxima a la normal estándar:

$$\mathrm{Y} = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\left(\frac{1}{2}\right)\mathrm{t}^2}$$

La distribución t de Student tiene propiedades parecidas a $\mathrm{N}(0,1)$:

a) El valor esperado es cero $\mathrm{E}(\mathrm{T}) = 0$.

b) Distribución simétrica con respecto a 0.

c) La varianza de t está dada por $\mathrm{V}(\mathrm{t}) = \mathrm{v}/(\mathrm{v} - 2)\ \ \mathrm{v} > 2$.

d) La varianza de t es ligeramente mayor que 1.0; es decir, es ligeramente mayor que la distribución normal estandarizada.

e) A medida que aumenta v, la dispersión de la curva t correspondiente disminuye. Es decir, un número alto de grados de libertad se puede aproximar la distribución t-Student por la normal: $\mathrm{t\_n}\ \ (\mathrm{n} \longrightarrow \infty) \longrightarrow\ \mathrm{N}(0,1)$.

f) A medida que $\mathrm{v} \longrightarrow \infty$, la secuencia de curvas de t se aproxima a la curva normal estándar.

g) Es de media cero y simétrica con respecto a la misma.

h) Es algo más dispersa que la normal, pero la varianza decrece hasta 1 cuando el número de grados de libertad aumenta.

Como en las distribuciones normales, se pueden definir intervalos de confianza de 95 %, 99 % u otros intervalos usando la tabla de la distribución t. Tal que puede estimarse la media poblacional μ dentro de determinados límites de confianza. Por ejemplo, si $-\mathrm{t}_{0.975}$ y $\mathrm{t}_{0.975}$ son valores de t para los cuales 2.5% del área se encuentra repartida en cada una de las colas de la distribución t; entonces, el intervalo de confianza para t de 95% es:

$$-\mathrm{t}_{0.975} < \frac{\overline{\mathrm{X}} - \mu}{\mathrm{s}} \sqrt{\mathrm{N} - 1} < \mathrm{t}_{0.975}$$

A partir de esto se puede estimar que μ se encuentra en el intervalo:

$$\overline{\mathrm{X}} - \mathrm{t}_{0.975} \frac{\mathrm{s}}{\sqrt{\mathrm{N} - 1}} \left(\mathrm{L}\right) < \mu < \overline{\mathrm{X}} + \mathrm{t}_{0.975} \frac{\mathrm{s}}{\sqrt{\mathrm{N} - 1}} \left(\mathrm{L}\right) \therefore \overline{\mathrm{X}} \pm \mathrm{t}_{\mathrm{c\ (calculada)}} \frac{\mathrm{s}}{\sqrt{\mathrm{N} - 1}}$$

Donde, los valores $\pm \mathrm{t}_{\mathrm{c\ (calculada)}}$ son llamados Valores críticos o Coeficientes de Confianza, dependen del nivel de confianza deseado y del tamaño del a muestra. Con una

confianza de 95% (probabilidad de 0.95). $t_{0.975}$ representa el valor percentil de 97.5 y $t_{0.25} = -t_{0.975}$ indica el valor del percentil 2.5. Se supone que la muestra se toma de una población normal. Esta suposición se puede verificar empleando la prueba para normalidad de Kolmogorov-Smirnov.

Las pruebas de hipótesis y de significancia o reglas de decisión pueden extenderse fácilmente a problemas con muestras pequeñas.

Media. Para probar la hipótesis H_0 que una población normal tiene una media μ se usa la puntuación t (estadístico t):

$$\mathrm{t} = \frac{\overline{\mathrm{X}} - \mu}{\mathrm{s}}\sqrt{\mathrm{N}-1} = \frac{\overline{\mathrm{X}} - \mu}{\hat{\mathrm{s}}}\sqrt{\mathrm{N}}$$

Donde $\bar{X}$ es media de una muestra de tamaño N. Esto es análogo (semejante) a usar puntuación z:

$$\mathrm{z} = \frac{\overline{\mathrm{X}} - \mu}{\sigma/\mathrm{N}}$$

Para una N grande, talque que si usa $\hat{\mathrm{s}} = \sqrt{\mathrm{N}(\mathrm{N}-1)}$ en lugar de σ. La diferencia es que mientras z está distribuida normalmente, t sigue una distribución de Student. A medida que N aumenta, estas distribuciones tienden a coincidir.

- Diferencias entre medias. Suponga que de poblaciones normales cuyas desviaciones estándar son iguales ($\sigma_1 = \sigma_2$) se toman dos muestras aleatorias de tamaños N_1 y N_2. Suponga, además, que las medias de estas dos muestras son $\bar{X}_1$ y $\bar{X}_2$, así como sus desviaciones estándar son s_1 y s_2, respectivamente.

Para probar la H_0 que las muestras provienen de una misma población ($\mu_1 = \mu_2$ y también $\sigma_1 = \sigma_2$) se usa la puntuación t dada por:

$$\mathrm{t} = \frac{\overline{\mathrm{X}}_1 - \overline{\mathrm{X}}_2}{\sigma\sqrt{1/\mathrm{N}_1 + 1/\mathrm{N}_2}} \text{ donde } \sigma = \sqrt{\frac{\mathrm{N}_1\mathrm{S}_1^2 + \mathrm{N}_2\mathrm{S}_2^2}{\mathrm{N}_1 + \mathrm{N}_2 - 2}}$$

Esta distribución t tiene una distribución de Student con $v = \mathrm{N}_1 + \mathrm{N}_2 - 2$ grados de libertad. El uso de la ecuación anterior se hace plausible al hacer $\sigma_1 = \sigma_2 = \sigma$ en la puntuación z, así como usar la media ponderada como estimación de σ^2:

$$\frac{(N_1-1)\hat{s}_1^2+(N_2-1)\hat{s}_2^2}{(N_1-1)+(N_2-1)}=\frac{N_1S_1^2+N_2S_2^2}{N_1+N_2-2}$$

Donde $\hat{s}_1^2$ y $\hat{s}_2^2$ son estimadores insesgados de σ_1^2 y σ_2^2[163].

4.1.10.3. Distribución F de Fisher

En algunas aplicaciones es importante conocer la distribución muestral de la diferencia entre las medias $(\overline{X}_1-\overline{X}_2)$ de dos muestras. De igual manera, algunas veces se necesita la distribución muestral de la diferencia entre varianzas $(S_1^2-S_2^2)$. Sin embargo, resulta que esta distribución es bastante complicada. En consecuencia, se considera el estadístico S_1^2/S_2^2, pues un cociente grande o pequeño indica una gran diferencia, en tanto que un cociente cercano a 1 indica una diferencia pequeña. En este caso se puede encontrar una distribución muestral llamada Distribución F, en honor a Ronald Aylmer Fisher, Londres, 17 de Febrero de 1890–Adelaida, 29 de Julio de 1962.

Teorema. Si U y W son dos variables aleatorias independientes, cada una con distribución Chi Cuadrado con $v_1\ y\ v_2$ grados de libertad, respectivamente. Entonces la distribución de la siguiente variable aleatoria:

$$F=\frac{\frac{U}{v_1}}{\frac{U}{v_2}}$$

Está dada por:

$$f(X)=\frac{\frac{\Gamma\left[\frac{(v_1+v_2)}{2}\right]\left(\frac{v_1}{v_2}\right)^{v_1}}{2X^{(n/2)-1}}}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)\left(\frac{1+v_1X}{v_2}\right)^{\frac{(v_1+v_2)}{2}}}$$

Se denomina "distribución F con v_1 (v_1 grados de libertad en numerador) y v_2 (v_2 grados de libertad en denominador) grados de libertad" y con la forma gráfica siguiente:

[163] Ejemplos y un conjunto de ejercicios se encuentran en carpeta "t − Student"

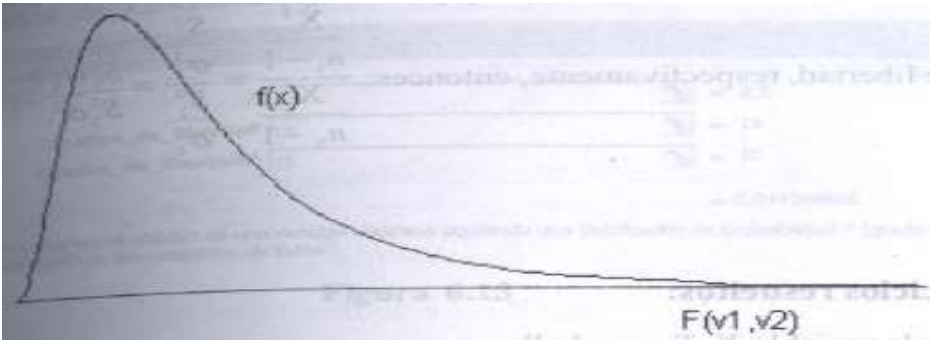

Figura 41. "Distribución F con v_1 ($v_1$grados de libertad en numerador) y v_2 (v_2 grados de libertad en denominador) grados de libertad [164]

En consecuencia, la media y varianza de la Distribución F son:

$\mu = \left(\frac{v_1}{v_2}\right) - 2 \quad$ para $v_2 > 2$.

$$\sigma^2 = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \quad \text{para } v_2 > 4$$

Propiedad:

$Fv_1, v_2, P = \frac{1}{Fv_2, v_1 . 1-P}$ si se invierte la definición de la distribución F.

Sean:

$$X_1^2 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \text{ y } X_2^2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$$

Variables aleatorias con distribuciones Chi Cuadrado con $n_1 - 1$ y $n_2 - 1$ grados de libertad, respectivamente, entonces:

$$\frac{\frac{X_1^2}{n_1 - 1}}{\frac{X_2^2}{n_2 - 1}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F(n_1 - 1, n_2 - 1)$$

Otra forma de interpretarlo es suponiendo que se tienen dos muestras, 1 y 2, de tamaños N_1 y N_2, obtenidas de dos poblaciones normales, o casi normales, cuyas varianzas son σ_1^2 y σ_2^2. Sea el estadístico:

[164] Fuente: Daza P. G. F. 2006

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} = \frac{N_1 S_1^2/(N_1 - 1)\sigma_1^2}{N_2 S_2^2/(N_2 - 1)\sigma_2^2}$$

Donde:

$\hat{S}_1^2 = \frac{N_1 S_1^2}{N_1 - 1}$ y $\hat{S}_2^2 = \frac{N_2 S_2^2}{N_2 - 1}$

Entonces a la distribución muestral de F se le llama Distribución F de Fisher o, simplemente, Distribución F, con $v_1 = N_1 - 1$ y $v_2 = N_2 - 1$ grados de libertad. Esta distribución está dada por:

$$Y = \frac{C_{(\text{Constante dependiente de } v_1 \text{ y } v_2)} F^{(v_1/2)-1}}{(v_1 F + v_2)^{(v_1+v_2)/2}}$$

Tal que, el área total bajo la curva es 1. Esta curva tiene una forma similar a la siguiente; aunque puede variar notablemente según $v_1 + v_2$:

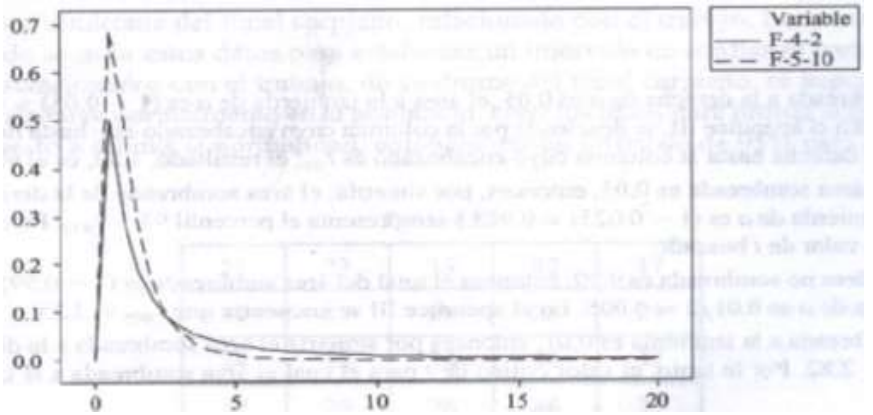

Figura 42. "Distribución F con 4 y 2 grados de libertad, línea punteada es Distribución F con 5 y 10 grados de libertad ([165])

Los valores percentiles de F para las áreas en la cola derecha son 0.05 y 0.01, respectivamente, denotados por $F_{0.95}$ y $F_{0.99}$, representados en niveles de significancia del 5% y 1%, usado para determinar si la varianza S_1^2 es significativamente mayor que la varianza S_2^2.

[165] Fuente: Daza P. G. F. 2006

El número de grados de libertad de un estadístico, generalmente denotado por v, se define como la cantidad N de observaciones en la muestra (tamaño muestral) menos la cantidad k de parámetros poblacionales que tengan que estimarse a partir de las observaciones muestrales. En símbolo es esto es $\mathrm{v} = \mathrm{N} - \mathrm{k}$. Para calcular un estadístico, por ejemplo (1) y (8), es necesario emplear observaciones obtenidas de una muestra y también ciertos parámetros poblacionales. Si estos parámetros no se conocen, es necesario estimarlos a partir de la muestra. En el caso del estadístico (1), la cantidad de observaciones independientes en la muestra es N y a partir de ella se se calculan $\bar{X}$ y S. Sin embargo, como se necesita estimar μ, $\mathrm{k} = 1$ y, por lo tanto, $v = \mathrm{N} - 1$. En el caso del estadístico (8), la cantidad de observaciones independientes en la muestra es N y a partir de ella se se calculan S. Sin embargo, como se necesita estimar σ, $\mathrm{k} = 1$ y, por lo tanto, $v = \mathrm{N} - 1$ ([166]).

Valor p (p value) es, según Ramírez (2015), la "Probabilidad de un valor más grande que el que se calculó" (probabilidad de un valor del estadístico más grande que el valor calculado) y, también, es la forma en que los paquetes estadísticos, como SAS, evitan la necesidad de ver los valores de tablas, de F o T o Ji. Si el valor p es 0.05 o menor es equivalente a que tu valor calculado sería mayor al de tablas, de F o t o Ji-Chi Cuadrado, al 5% y se dice que la prueba resultó significativa. Si el valor p es 0.01 o menor es equivalente a que tu valor calculado sería mayor al de tablas, de F o t o Ji-Chi, al 1% y se dice que la prueba resultó altamente significativa. Si el valor p es mayor que 0.05 es una forma de decir que la prueba resultó no significativa, es decir que "No se tienen elementos para Rechazar Ho" ([167]). Los valores-p correspondientes a los estadísticos z, t, X^2 y F que se usan en las pruebas de hipótesis:

$$\text{Valor} - \mathrm{p} = 1 - \text{Probabilidad acumulada}$$

En otras palabras, el valor p es la probabilidad de observar un valor muestral tan extremo, o más extremo, que el valor observado dado que H_0 es verdadera o dicho de otra manera valor-ρ es una probabilidad que aporta una medida de una evidencia suministrada por la muestra contra la hipótesis nula, pues valores-p pequeños indican una evidencia mayor

[166] Ejemplos y un conjunto de ejercicios se encuentran en carpeta "F de Fisher"

[167] Fuente: conv. pers. Ing. M. C. José Artemio Cadena Meneses Ph. D. Profesor Investigador del Departamento de Zootecnia. Universidad Autónoma Chapingo. E-mail: cadena@correo.chapingo.mx

contra la hipótesis nula y, por ello, el valor-p se usa para determinar si la hipótesis nula debe ser rechazada. Regla de decisión: ρ<nivel de significancia no se acepta H_0, pero si ρ >nivel de significancia no se rechaza H_0. Entonces, si $\rho = 0.0001$ indica que hay poca probabilidad que H_0 sea verdadera, pero si $\rho = 0.2033$ indica que Ho no se rechaza y hay poca probabilidad que ésta sea falsa. Interpretación del peso de evidencia contra H_0:

- Si $\rho < 0.10$ tiene alguna evidencia que H_0 no es verdadera
- Si $\rho < 0.05$ tiene fuerte evidencia que H_0 no es verdadera
- Si $\rho < 0.01$ tiene muy fuerte evidencia que H_0 no es verdadera
- Si $\rho < 0.001$ tiene una evidencia extremadamente fuerte que H_0 o no es verdadera

5. MUESTREO

5.1. DISTRIBUCIONES

5.1.1. Historia

La Estadística actual es el resultado de la fusión de dos disciplinas que evolucionaron de forma independiente hasta converger en Siglo XIX. La primera corresponde al Cálculo de Probabilidades mientras que la segunda a Estadística, que estudia descripción de datos y tiene raíces más antiguas. La integración de ambas líneas de pensamiento dio lugar a cómo obtener conclusiones de investigación empírica mediante modelos matemáticos. Los inicios de Estadística se hallan en antiguo Egipto, cuyos faraones recopilaron hacia el año 350 a.c., datos relativos a población y riqueza del país. Según Heródoto, este registro se hizo con la finalidad de aplicarlo a construcción de pirámides.

Los chinos efectuaron censos hace más de 40 siglos. Los griegos hicieron censos tributarios, sociales y militares. La investigación histórica revela que se hicieron 69 censos para estimar impuestos, determinar derechos de voto y ponderar la potencia guerrera. Sin embargo, los romanos emplearon los recursos de la Estadística de la mejor manera. Cada 5 años realizaban un censo poblacional y sus funcionarios públicos tenían la obligación de anotar los censos de nacimientos, defunciones y matrimonios, sin olvidar recuentos periódicos de recursos agropecuarios de imperios conquistados. Durante los mil años siguientes a la caída el imperio romano se hizo muy pocas investigaciones estadísticas. El primer intento de aplicar un razonamiento propiamente estadístico, en sentido actual del término, a datos demográficos es debido a John Graunt, en 1662, quien se planteó el problema de estimar la población inglesa de la época.

Godofredo Achenwall, Profesor de la Universidad de Gotinga, acuñó en 1760 la palabra Estadística, que extrajo del término italiano statista –Estadista-. Creía que los datos de la nueva ciencia sería el aliado más eficaz de gobernantes conscientes.

Durante el siglo XVIII y la mayor parte del XIX, la Estadística evolucionó como ciencia separada del Cálculo de Probabilidades. Una contribución importante al desarrollo de la Estadística es debida a Lambert Adolphe Jacques Quetelet (1846), sostuvo la importancia del Cálculo de Probabilidades para estudio de datos humanos. Lambert Adolphe Jacques

Quetelet demostró que la estatura de reclutas de un regimiento seguía una ley probabilística e introdujo el concepto "hombre medio".

A finales de siglo XIX, Sir Francis Galton ideó el método conocido como Correlación, que tenía por objeto medir la influencia relativa de factores. Sus investigaciones se dirigieron a aplicar métodos cuantitativos en estudio de herencia humana. La importancia del aporte de Galton radico no sólo en nuevo enfoque que introdujo en problemas de Estadística, sino en su influencia directa sobre Walter Frank Raphael Weldon, Karl Pearson y Francis Ysidro Edgeworth, etcétera. Además, fundó el primer departamento de Estadística.

Sin embargo, quien más influyó en Estadística moderna fue Ronald Aylmer Fisher, 1890-1962. Fisher se interesó primero por la Eugenesia ([168]), que condujo siguiendo los pasos de Galton a la investigación Estadística. Sus trabajos culminaron con la publicación del libro "Statistical Methods for Researchs Workers". En esta obra aparece el cuerpo metodológico básico de Estadística actual.

A partir de 1950, inicia la época moderna de Estadística. Un aspecto diferencial respecto a periodos anteriores es la aparición de computadores, que revolucionaron la metodología Estadística abriendo posibilidades para construcción de modelos complejos.

Actualmente, la Estadística es una disciplina que actúa como un nexo entre modelos matemáticos y fenómenos reales. Un modelo es una abstracción simplificada de una realidad más compleja y siempre existirá discrepancia entre lo observado y lo previsto por el modelo. La Estadística proporciona una metodología para evaluar y juzgar estas discrepancias entre realidad y teoría ([169]).

[168] definicion.de/eugenesia/: Disciplina que busca aplicar las leyes biológicas de la herencia para perfeccionar la especie humana. La eugenesia supone una intervención en rasgos hereditarios para ayudar al nacimiento de personas más sanas e inteligentes.
[169] Capa, S. H. b. 2015

5.1.2. Clases de distribución

5.1.2.1. Media

5.1.2.1.1. Media cuando varianza es desconocida

En apartados anteriores estudiamos el comportamiento de la media muestral y vimos que ésta dependía tanto del valor de la media poblacional, como de la varianza poblacional, parece lógico pensar que si nuestro interés radica en inferir comportamientos de la población partiendo de la muestra parece ilógico pensar que conozcamos la varianza. De ahí la importancia de establecer una distribución para la media muestral que la relacione únicamente con la poblacional, lo que hará que conocida la muestral concreta podamos aventurar el comportamiento de la poblacional. En consecuencia, $\bar{x} \Rightarrow N\left[\mu \frac{\sigma}{\sqrt{n}}\right]$ tal que $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \Rightarrow N[0;1]$. Sin necesidad de demostrar, $\frac{nS^2}{\sigma^2} \Rightarrow X^2 \text{con } (n-1)\text{gl}$ tal que $\sqrt{\frac{nS^2}{\frac{\sigma^2}{n-1}}}$. Se sabe que $t_n = \frac{N[0;1]}{\sqrt{\frac{X_n^2}{n}}} \longrightarrow \frac{\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{nS^2}{\frac{\sigma^2}{n-1}}}} \Rightarrow t_{n-1}$. Simplificando, se tiene que $\frac{\bar{x}-\mu}{S}\sqrt{n-1} \Rightarrow t_{n-1}$. Esta expresión relaciona ambas medias y la varianza muestral con una distribución conocida. Por último, se supone que, por Teorema de Límite Central, siempre se tiene convergencia hacia ley normal, pero no siempre sucede, como varianza poblacional desconocida.

Ley de distribución t de Student. Su gráfico de función densidad de ley t tiene forma semejante a la normal, simétrico respecto a 0 y asintótico respecto al eje real. Sus valores probabilísticos son tabulados, que dependen del grado de libertad, pues la ley de probabilidad t cambia si n varía tal que cuando n aumenta, su distribución t se aproxima a la normal.

Ley de distribución $\bar{x}$. Se obtiene una muestra $X_1, X_2, X_3, X_4, \dots, X_n$ de una población que sigue una ley de distribución normal $N(\mu, \sigma^2)$ tal que σ^2 es desconocida. Se cumple que la variable aleatoria $T = \frac{(\bar{X}-\mu)}{\frac{s}{\sqrt{n}}}$ sigue una ley $t - \text{Student}$, con $(n-1)_{(\text{Grados de libertad})}$ tal que $P_r(\bar{X} \leq t) = P_r\left(T \leq \frac{(t-\mu)}{\left(\frac{s}{\sqrt{n}}\right)}\right)$.

5.1.2.1.2. Diferencia de dos medias con varianzas conocidas

Se dispone de dos poblaciones que tienen medias μ_1 y μ_2 con varianzas σ_1^2 y σ_2^2, respectivamente. Si $\overline{X}_1$ y $\overline{X}_2$ medias muestrales de dos muestras aleatorias independientes de tamaños n_1 y n_2, seleccionadas de las poblaciones 1 y 2, respectivamente. Entonces, $\overline{X}_1$ y $\overline{X}_2$ cumplen que $\mu_{\overline{X}_1-\overline{X}_2} = E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$; $\sigma^2_{\overline{X}_1-\overline{X}_2} = Var(\overline{X}_1 - \overline{X}_2) = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$; para n_1 y n_2 suficientemente grandes, la variable aleatoria $Z = \frac{[(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)]}{\sqrt{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}}$ sigue aproximadamente una ley normal estándar; es decir,

$$P_r(\overline{X}_1 - \overline{X}_2 \leq t) \approx P_r\left(Z \leq \frac{(t-(\mu_1-\mu_2))}{\sqrt{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}}\right) = \Phi\left(\frac{(t-(\mu_1-\mu_2))}{\sqrt{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}}\right)$$ tal que para la mayoría de aplicaciones se tiene una aproximación buena si n_1 y $n_2 \geq 25$.

Ejemplo:

1) Una marca de bebidas alimenticias tiene dos embazadoras que procesan el mismo producto. Su rendimiento medio promedio de bebidas será de 40 g por galón, con desviación estándar de 5 g por galón. La marca compara directamente rendimientos de bebidas, seleccionando muestras aleatorias en ambas embazadoras. Se toan muestras de tamaño 30, se controla embazado promedio por unidad. Estime probabilidad que la diferencia entre promedio sea menor a 2 g por galón.

Con base en esto, $\mu_1 = \mu_2 = 40$, $\sigma_1 = \sigma_2 = 5$, $n_1 = n_2 = 30$; por lo tanto, $P_r(\overline{X}_1 - \overline{X}_2 \leq 2) \approx P_r\left(-\frac{(2-(40-40))}{\sqrt{\left(\frac{5^2}{30}+\frac{5^2}{30}\right)}} \leq Z \leq \frac{(2-(40-40))}{\sqrt{\left(\frac{5^2}{30}+\frac{5^2}{30}\right)}}\right) = \Phi(1.55) - \Phi(-1.55) = 0.939 - 0.061 = 0.879$.

5.1.2.1.3. Diferencia de dos medias con varianzas desconocidas

Se dispone de dos poblaciones que siguen una ley normal, pues población 1 sigue una ley de distribución $N(\mu_1, \sigma_1^2)$ mientras que población 2 sigue una $N(\mu_2, \sigma_2^2)$. Si $\overline{X}_1$ y $\overline{X}_2$ son dos medias muestrales aleatorias e independientes con tamaño n_1 y n_2, seleccionadas de poblaciones 1 y 2.

Caso 1. Sus varianzas poblaciones son iguales $\sigma_1^2 = \sigma_2^2 = \sigma^2$, su variable aleatoria será $T = \frac{[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]}{s* \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ sigue una ley de distribución t con $(n_1 + n_2 - 2)_{(\text{Grados de libertad})}$, donde $s^2 = \left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}\right)$.

Caso 2. Sus varianzas poblaciones son diferentes $\sigma_1^2 \neq \sigma_2^2$, su variable aleatoria será $T = \frac{[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]}{s* \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$, sigue una ley de distribución t con $g_{(\text{Grados de libertad})}$,

donde $g = \left(\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)} \right)} \right)$ cuando $g \notin \mathbb{N}$ se redondea al entero más cercano.

5.1.2.2. Proporción

Se tiene una muestra aleatoria $X_1, X_2, X_3, X_4, \ldots, X_n$ proveniente de una población que sigue una ley de Bernoulli, $\text{Ber}(p)$ se define $Y = \sum_{i=1}^{n} X_i$ tal que $X_i = 1$ con probabilidad p y $X_i = 0$ con probabilidad $q = 1 - p$; $i = 1,2,3,4, \ldots, n$. Donde, Y cuenta número de éxitos en n intentos mientras que l proporción de éxitos en muestra es $\hat{p} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$. Variable aleatoria Y tiene distribución binomial de parámetros (n, p); por lo que, $\mu_Y = np$, con $\sigma_Y^2 = npq$ tal que cumple: 1) $E(\hat{p}) = \frac{E(Y)}{n} = p$; 2) $\text{Var}(\hat{p}) = \frac{\text{Var}(Y)}{n^2} = \frac{pq}{n}$ y 3) $\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$ sigue aproximadamente una ley normal estándar, por $\text{Teorema de Límite Central}$; es decir, $P_r(\hat{p} \leq t) \approx P_r\left(Z \leq \frac{(t-p)}{\sqrt{\frac{pq}{n}}}\right) = \Phi\left(\frac{(t-p)}{\sqrt{\frac{pq}{n}}}\right)$ tal que Z es variable aleatoria normal estándar.

Ejemplo:

1) Un proceso de elaboración de queso de soya, el 20% de productos tiene algún defecto. Fueron seleccionados 100 unidades como muestra, contando el número de productos con algún defecto. Estime probabilidad que proporción de defectuosos se halle entre 15% y 29%.

Con base en esta información, $p = 0.20$, $\frac{pq}{n} = \frac{(0.20*0.80)}{100} = \frac{(0.16)}{100} = 0.0016$ tal que se desea evaluar la probabilidad $P_r(0.15 \leq \hat{p} \leq 0.29) \Rightarrow P_r(0.15 \leq \hat{p} \leq 0.29) = P_r(\hat{p} \leq 0.29) - P_r(\hat{p} \leq 0.15) \approx \Phi\left(\frac{(0.29-0.20)}{\sqrt{\frac{0.16}{100}}}\right) - \Phi\left(\frac{(0.15-0.20)}{\sqrt{\frac{0.16}{100}}}\right) = \Phi(2.25) - \Phi(-1.25) = 0.988 - 106 = 0.882$

5.1.2.3. Varianza

Según Teorema de Límite Central, la convergencia de ley de $\overline{X}$ está asegurada e independiente de ley que siguen las observaciones. En otros estadísticos, se requieren hipótesis adicionales para asegurar su convergencia hacia una ley de distribución, como varianza muestral.

Ley de distribución χ^2. Sea $X_1, X_2, X_3, X_4, \ldots, X_n$ n variables aleatorias independientes que siguen distribución normal estándar, variable aleatoria definida como $T = \sum_{i=1}^{n} X_i^2$ tiene una distribución χ^2 con $n_{(\text{Grados de libertad})}$ denominada $\chi^2(n)$. Su función de densidad está da por

$$f(x) = \begin{cases} \frac{X^{\frac{(n-2)}{2}} * e^{-\left(\frac{x}{2}\right)}}{X^{\left(\frac{n}{2}\right)} * \Gamma * \left(\frac{n}{2}\right)}, & \text{si } x \geq 0; \\ 0, & \text{si } x < 0; \end{cases}$$

Ley de distribución S^2. Suponga que existe una muestra $X_1, X_2, X_3, X_4, \ldots, X_n$ de una población que sigue una ley normal $N(\mu, \sigma^2)$ tal que a partir de la muestra se estima varianza muestral $S^2 = \frac{\sum_{i=1}^{n}(X_1 - \overline{X})^2}{n-1}$ tal que 1) $E(S^2) = \sigma^2$, 2) $Var(S^2) = \frac{2*\sigma^4}{n-1}$, 3) $\frac{(n-1)*S^2}{\sigma^2}$ sigue una ley normal $\chi^2(n-1)$.

5.1.2.4. Diferencia de dos proporciones

Se dispone, de dos poblaciones independientes, que siguen distribuciones de Bernoulli de parámetros p_1 y p_2, respectivamente. De la primera población se escoge una muestra tamaño $n_1: X_1, X_2, X_3, X_4 \ldots . X_{n1}$ y de la segunda una muestra con tamaño $n_2: Y_1, Y_2, Y_3, Y_4 \ldots . Y_{n2}$ tal que se construyen las proporciones muestrales de éxitos mediante $\hat{p}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$, $\hat{p}_2 = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}$ tal que $\hat{p}_1 - \hat{p}_2$ cumple 1) $\mu_{\hat{p}_1 - \hat{p}_2} = E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$, 2) $\sigma^2_{\mu_{\hat{p}_1 - \hat{p}_2}} = Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} - \frac{p_2(1-p_2)}{n_2}$, 3) Para suficientemente grandes n_1 y n_2,

variable aleatoria $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$ que sigue aproximadamente una ley normal estándar, por Teorema de Límite Central tal que

$$P_r(\hat{p}_1 - \hat{p}_2 \leq t) \approx P_r\left(Z \leq \frac{(t-(p_1-p_2))}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}\right) = \Phi\left(\frac{(t-(p_1-p_2))}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}\right).$$

Ejemplo:

1) Una consultoría especializada en encuestas sostiene que 30% de mujeres y 20% de hombres están a favor que permanezca en el mercado una leche deslactosada. Si realiza un monitoreo aleatorio a 150 consumidores de cada género, ¿con qué probabilidad la diferencia entre proporciones muestrales de mujeres y hombres es, en valor absoluto, menor a 0.19? Con base en esta información, $p_h = 0.30$, $p_m = 0.20$, $n_h = n_m = 150$. Además, $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{0.30*(1-0.30)}{150} + \frac{0.20*(1-0.20)}{150} = 0.002$, buscando la probabilidad $P_r(|\hat{p}_h - \hat{p}_m| < 0.19) \Rightarrow$ $P_r(|\hat{p}_h - \hat{p}_m| < 0.19) = P_r(-0.19 < \hat{p}_h - \hat{p}_m < 0.19) = \Phi\left(\frac{(0.19-(0.30-0.20))}{\sqrt{0.002}}\right) -$ $\Phi\left(-\frac{(0.19-(0.30-0.20))}{\sqrt{0.002}}\right) = \Phi(1.81) - \Phi(-1.81) = 0.965 - 0.035 = 0.929.$

5.1.2.5. Razón de dos varianzas

Ley de distribución F. Si X_1 y X_2 dos variables aleatorias independientes con distribución χ^2 con n_1 y n_2 grados de libertad, respectivamente tal que la variable aleatoria $V = \left[\frac{\left(\frac{X_1}{n_1}\right)}{\left(\frac{X_2}{n_2}\right)}\right]$ sigue distribución $F_{(\text{Snedecor o Fisher})}$ con $(n_1, n_2)_{(\text{Grados de libertad})}$ denotados como $F(n_1, n_2)$. Su función de densidad es $f(x) = \begin{cases} \frac{\Gamma*\left(\frac{n_1+n_2}{2}\right)*n_1^{\left(\frac{n_1}{2}\right)}*n_2^{\left(\frac{n_2}{2}\right)}}{\Gamma*\left(\frac{n_1}{2}\right)*\Gamma*\left(\frac{n_2}{2}\right)} * \left(x^{\left(\frac{n_1}{2-1}*(n_2+n_1 x)^{-\left(\frac{n_1+n_2}{2}\right)}\right)}\right), & \text{si } x > 0; \\ 0, \quad \text{si } x \leq 0; \end{cases}$

$0, \quad \text{si } x \leq 0;$

Donde, la esperanza y varianza son $E(V) = \left(\frac{n_2}{n_2-2}\right), \text{si } n_2 > 2$ y $Var(V) = \frac{2n_2^2*(n_1+n_2-2)}{n_1*(n_2-2)^2*(n_2-4)}, \text{si } n_2 > 4$. Finalmente, para lectura de valores % de extremo inferior de tabla de ley F se usa el cosiente $F_{(1-\alpha)(n_1,n_2)} = \frac{1}{F_{(\alpha)*(n_1,n_2)}}$.

Ley de distribución $\left(\frac{S_1^2}{S_2^2}\right)$.

Se dispone de dos poblaciones con una ley de distribución normal. La población primera tiene una ley $N(\mu_1, \sigma_1^2)$ mientras que la segunda sigue una ley $N(\mu_2, \sigma_2^2)$. Entonces, si S_1^2 y S_2^2 son varianzas aleatorias muestrales independientes de tamaños n_1 y n_2 tal que su variable aleatoria es

$$F_{(n_1-1,n_2-1)(\text{Grados de libertad})} = \left[\frac{\left(\frac{S_1^2}{\sigma_1^2}\right)}{\left(\frac{S_2^2}{\sigma_2^2}\right)}\right]; \quad \text{por} \quad \text{consiguiente,} \quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \Rightarrow F = \left(\frac{S_1^2}{S_2^2}\right) \sim F(n_1 - 1, n_2 - 1).$$

Ejemplos:

1) Estime el valor x tal que $P_r(V > x) = 0.05$ donde $V \sim F(6,7)$. En tabla F con nivel de significancia $\alpha = 0.05$, valores $n_1 = 6$ y $n_2 = 9$ el valor de $F_{(\text{Tablas})}$. Finalmente, estime el valor de x tal que $P_r(V < x) = 0.05$, $V \sim F(6,7)$. Considere valores $n_1 = 6$, $n_2 = 9$ tal que si $P_r(V < x) = 0.05$ implica que $P_r(V > x) = 0.95$.

Con base en esto, $x = F_{(0.05;6,9)} = 3.37$ y, también, $F_{(0.95;6,9)} = \frac{1}{F_{(\alpha=0.05)*(6,9)}} = 0.244$.

2) Una empresa de jugos naturales tiene dos plantas procesadoras que producen el mismo jugo. La calidad-cantidad de jugo deberá presentar igual media y desviación estándar. Esta empresa suele comparar estos elementos mediante muestras al azar en ambas procesadoras. Entonces, se toman 30 muestras estimando sus valores dichos. Estime la probabilidad que la desviación estándar de una muestra sea al menos 1.5 veces mayor que la segunda.

Con base en esta información, $P_r\left(\frac{S_1}{S_2} \geq 1.5\right) = P_r\left(\frac{S_1^2}{S_2^2} \geq 2.25\right)$ tal que $F = \left(\frac{S_1^2}{S_2^2}\right) \sim F(30 - 1, 30 - 1)$ entonces $0.01 < P_r\left(\frac{S_1^2}{S_2^2} \geq 2.25\right) < 0.025$; por lo tanto, su posibilidad exacta es $P_r\left(\frac{S_1^2}{S_2^2} \geq 2.25\right) = 0.016$ ([170]).

[170] Capa, S. H. b. 2015

5.2. TIPOS DE MUESTREO

5.2.1. Introducción

Recursos Forestales [171]:

- Flora
- Fauna
- Agua
- Suelo

Inventario Forestal. Los inventarios forestales suelen considerarse como sinónimos de estimaciones de calidad, cantidad, condiciones y distribución de árboles de un bosque y algunas características propias de la zona donde crecen. En otras palabras: cuántos hay, cómo están y dónde se ubican los recursos forestales. Su finalidad es planear su óptimo manejo de manera constante, sostenida, sustentable.

Clasificación de Inventarios Forestales según el tema [172]:

➢ **Botánicos**. Inventarios que se realizan sobre diferentes condiciones de vegetación forestal y en que se evalúa sus diferentes niveles de clasificación.

➢ **Faunísticos**. Son los inventarios que se realizan en diferentes condiciones de vegetación forestal y en que se evalúan únicamente los recursos faunísticos.

➢ **Edafológicos**. Son los inventarios que se realizan en diferentes condiciones de vegetación forestal y en que se evalúan únicamente exclusivamente los suelos.

➢ **Hidrológicos**. Son los inventarios que se realizan en diferentes condiciones de vegetación forestal y en que se evalúan únicamente exclusivamente el agua.

➢ **Integrales**. Se efectúan sobre las diferentes condiciones de vegetación forestal y en que se evalúan todos los factores interrelacionados en la misma, como agua, suelo, vegetación, fauna, etcétera.

Clasificación de Inventarios Forestales según Tipo de Influencia:

❖ **Nacionales**. Su área de influencia es muy amplia y su nivel de estudio es más generalizado.

❖ **Provinciales**. Su área de influencia es muy amplia, pero es semi-detallado.

[171] Carrillo, E. G. 2005
[172] Carrillo, E. G. 2010

❖ **Regionales**. Su área de influencia es más específica y, por tratarse de una zona o región específica, su nivel de estudio es detallado.

Clasificación de Inventarios Forestales según Tipo de Periodicidad:

Convencionales o Temporales. Se utilizan sitios temporales en el muestreo, generalmente de dimensiones fijas y la medición de sus variables son únicas. La intensidad de muestreo es mayor y por sus características de evaluar las condiciones del bosque en el momento de hacer las mediciones y por no requerir un cotejo su nivel de precisión es mayor.

Continuos o Periódicos. Los sitios utilizados en el muestreo son para levantamiento de la información son permanentes, dimensiones fijas por lo general (1000, 400 o 80 m^2), bajo un patrón sistemático, con diferentes intensidades de muestreo en cada uno de ellos, las remediciones de sus variables son periódicas y, en consecuencia, un registro detallado del arbolado presente en el área de estudio es necesario para su control en el tiempo. Además, por su finalidad de evaluar la dinámica del cambio de bosque con el paso del tiempo se requiere de un mayor número de observaciones tomadas con mayor laboriosidad y, con ello, un aumento en costos, utilidad y precisión en resultados. El Inventario Forestal Continuo (IFC) permite, mediante sus remediciones, precisar el crecimiento en diámetro y altura de todos los árboles localizados en las unidades de muestreo usadas en el levantamiento de la información.

5.2.1.1. Definición

Es una colección de objetos definidos y distinguibles cuya única propiedad indispensable es que sean identificados como pertenecientes a dicho conjunto, a cada uno de los objetos que lo constituyen se le llama elemento. Cabe mencionar que en la mayoría de los casos que involucran las técnicas de muestreo los objetos suelen ser de la misma naturaleza, o al menos muy semejante. Las técnicas de muestreo se aplican directamente a conjuntos de objetos con valores medidos en escalas continuas o discretas.

Los métodos de muestreo e inventarios con validez estadística se vuelto importantes para generar estimadas confiables y científicamente defendibles. Los inventarios son la base para la planeación de proyectos, manejo o administración y toma de decisiones estratégicas, de manera que se generen bases de datos confiables. Sin embargo, un censo o conteo completo de recursos sería demasiado costoso y tardado, el muestreo se hace

imprescindible. Además, no se cuenta con la totalidad de la información existente sino solamente con una fracción de ella (muestra).

Es decir, la información sobre cantidades y calidades de un recurso para tomar una decisión pueden ser obtenidas mediante una evaluación exhaustiva cuantificando o calificando todo el recurso. Sin embargo, en la mayoría de las circunstancias no es posible o conveniente hacer la evaluación exhaustiva, tal que sólo se hace en una parte del recurso (muestra), esperando que las determinaciones hechas también pertenezcan a la totalidad.

5.2.1.2. Razones para preferir el muestreo

a) Las técnicas que forman la estructura de las metodologías propias de un inventario forestal, temporal o convencional no captan los cambios en la dinámica del bosque originados por los volúmenes maderables que se extraen o eliminan por causas naturales, acción del hombre y volúmenes de la regeneración que se integran a la masa forestal.

b) La enumeración o medición completa puede ser imposible. Ejemplos: Determinar la cantidad exacta de madera en un bosque podría costar varias veces su valor.

c) Con frecuencia el muestreo proporciona la información esencial a un costo inferior al del conteo completo. En especial, para poblaciones grandes, los datos colectados por muestreo son más confiables.

d) La inferencia estadística es muy importante y su entendimiento apropiado es crucial para discutir el papel del muestreo en el proceso de las inferencias. La inferencia científica se convierte en inferencia estadística cuando la conexión entre el desconocido "estado de la naturaleza" y la información observada se expresa en términos probabilísticos.

e) La estadística proporciona métodos sobre cómo hacer tales análisis. Los métodos estadísticos se usan para predecir y explicar fenómenos, lo que con frecuencia es una tarea desafiante. Cramer (1946) citado por ([173]) resume el papel de la inferencia estadística en tres funciones:

1. **Descripción**. Es la reducción de los conjuntos de datos en un grupo de números tan pequeño como sea posible, como la media, la varianza, la asimetría de una distribución, etc.

[173] Schreuder, Ernst y Ramírez. 2006

Esto nos permite describir una población tan concisa y brevemente como sea posible y puede permitir la comparación.

2. **Análisis**. Es el resumen de lo datos para un propósito u objetivo particular. Por ejemplo: ¿cuáles son las estimadas de ciertas características de la población?, ¿cierta muestra proviene de una población dada?, ¿dadas dos muestras, provienen de la misma población o no? La estadística proporciona métodos sobre cómo hacer tales análisis.

3. **Predicción**. Los métodos estadísticos se usan para predecir y explicar fenómenos.

5.2.1.3. Análisis teórico de estimadores

Algunos parámetros y estimadores incluyen en su definición la suma de varios valores o datos. Si se simboliza por y_i a cualquiera de estos datos, como el i-ésimo de ellos y se tienen n datos, la suma de éstos datos se simboliza empleando operador suma (Σ):

$$y_1 + y_2 + y_3 + \cdots y_n = \sum_{i=1}^{n} y_i$$

O

$$y_1^2 + y_2^2 + y_3^2 + \cdots y_n^2 = \sum_{i=1}^{n} y_i^2$$

$$\sum_{i=1}^{5} y_i = (y_1 + y_2) + (y_3 + y_4 + y_5) \qquad \text{donde } y_i = \text{cualquier valor}$$

$$= \sum_{i=1}^{2} y_i + \sum_{i=3}^{5} y_i$$

5.2.1.4. Valor esperado o esperanza matemática $E(Y)$

Se llama valor esperado o esperanza matemática es la media de datos de una población, que es simplemente el promedio ponderado de posibles valores cuando se usan las probabilidades como facto de ponderación.

- **Variables Continuas** (variable aleatoria que puede tomar cualquier valor en un intervalo, p. ej. Cantidad de azúcar en una naranja, estatura de una persona, etc.):

$$E[Y] = \int_a^b y\, f(y)\, \delta(y)$$

Donde a y b son límites superior e inferior del rango de la variable aleatoria Y. $f(y)$ es la función de densidad de probabilidad.

➢ **Variable Discreta** (si se puede enumerar es discreta, p. ej. Número de casas rurales, número de personas en espera):

$$\mathrm{E[Y]} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{y_i P(y_i)}$$

Donde $\mathrm{P(y_i)}$ es la probabilidad que ocurra el valor $\mathrm{y_i}$.

Ejemplo: Una empresa necesita saber la ganancia promedio que obtendrá si vende un nuevo tipo de computadora. Si la probabilidad que una persona adquiera el nuevo tipo de computadora a un costo de $18,000 es 0.4 y la probabilidad que adquiera el modelo ya existente a un costo de $10,000 es 0.6 ¿Cuál sería la ganancia esperada?

$$\mathbf{E[Y]} = (18{,}000 * 0.4) + (10{,}000 * 0.6) = \mathbf{13{,}200}$$

5.2.1.5. Parámetros

Sobre el conjunto población se pueden definir funciones muy diversas como el valor más pequeño, el más grande, el que ocupa la posición central una vez que han sido ordenados ascendente o descendentemente, la suma de todos ellos después de elevarlos al cuadrado, el valor que se repite el mayor número de veces y muchos otros más, todas esas funciones son parámetros. Los parámetros suelen ser representados por letras griegas, como μ , τ , σ, en tanto que los estimadores generalmente se simbolizan con otros caracteres específicos.

Las funciones que se pueden proponer como parámetros, también se pueden definir para el conjunto muestra y aun otras funciones adicionales, entonces reciben el nombre de estimadores, pues a cada parámetro pueden corresponder uno o más estimadores. También existe un número infinito de estimadores, pero solo algunos tienen interés práctico.

Conclusión: Parámetro es una función que describe el total o una parte de la población, usualmente en forma numérica y estimador es una función de datos disponibles mediante la muestra que se usa para estimar los parámetros.

5.2.1.6. Población

La cantidad total de un recurso se denomina población y una parte de ese total constituye una muestra. Así, las mediciones se hacen en la muestra y se espera que los valores obtenidos correspondan a la población. Los valores de interés de la población se denominarán Parámetros y sus correspondientes en la muestra serán Estimadores. En consecuencia, es necesario que estudiantes conozcan este tipo de técnicas que les ayudarán

a medir o cuantificar ciertos fenómenos y justificar técnicamente estas aseveraciones en su especialidad -Forestal, Agroindustrial, Agronomía, Veterinaria, Zootécnica, Fitotecnia, Medicina, Parasitología Agrícola, Economía, Finanzas, Sociología, Agroecología, Administración, entre otras-.

5.2.1.7. Muestra

5.2.1.7.1. Unidad muestral

5.2.1.7.2. Marco muestral

5.2.1.7.3. Ventajas

5.2.1.7.4. Desventajas

5.2.2. Probabilístico

5.2.2.1. Muestreo simple aleatorio (MSA), muestreo simple al azar (MSA), muestreo completamente aleatorio (MCA) o muestreo irrestricto al azar (MIA)

Se denomina muestreo aleatorio simple o completamente al azar al diseño que habiendo decidido que el tamaño de la muestra será de n unidades de muestreo (o simplemente de tamaño n), le asigna la misma probabilidad de ser la elegida a cada una de todas las muestras posibles de ese tamaño. Es decir, cualquiera de las muestras distintas que podemos obtener de la población tendrá la misma probabilidad de ser elegida. Además, ésta es la propuesta más simple de muestreo probabilístico. Todas las unidades muestrales tienen una probabilidad de selección de n/N y cada conjunto de dos unidades tiene la probabilidad conjunta de selección de $\frac{n(n-1)}{(N-1)}$. Aunque esto puede parecer difícil implementar porque hay $\frac{N!}{n!(N-n)!}$muestras posibles si el muestreo es sin reemplazo (todas las n unidades en la muestra son diferentes).

El MSA no es difícil de implementar si está disponible una lista de las unidades en la población ([174]). Lo único que debemos asegurar es que la selección de una unidad no esté influenciada por las otras unidades, sea o no que estén incluidas en la muestra. Por ejemplo, se puede asignar a cada unidad un número distinto, desde 1 hasta *N* y elegir n números aleatorios distintos entre 1 y *N*. También tiene la ventaja de que al tener todas las unidades la misma probabilidad de selección, las técnicas de análisis aplicables son fáciles de aplicar y

[174] Schreuder, Ernst y Ramírez. 2006

la estimación es inmediata y entendible, esto es, al estimar la media μ o el total Y de una población. El estimador insesgado del total es: $\widehat{Y} = \frac{N\sum_{i=1}^{n} y_i}{n} = N\bar{y}$. Con un tamaño de muestra n, e y_i el valor de la variable de interés para la unidad de muestreo i, la varianza del estimador del total es:

$$V(\widehat{Y}) = \frac{N^2(N-n)}{Nn}\frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{(N-1)} = \frac{N^2(N-n)}{N}\frac{S^2}{n} = N^2(1-f)\frac{S^2}{n}$$

Un estimador insesgado de la varianza del estimador es:

$$v(\widehat{Y}) = \frac{N^2(N-n)}{Nn}\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n-1)} = \frac{N^2(N-n)}{N}\frac{s^2}{n}$$

5.2.2.1.1. Modalidades

✓ **Muestreo Aleatorio Simple con Reemplazo**. En el muestreo con reemplazo si el tamaño de la muestra es n y el de la población es N, existen N^n muestras diferentes y el procedimiento de selección consiste en seleccionar una unidad que tiene la posibilidad de ser incluida nuevamente en la muestra. Esta opción genera fórmulas de estimación más fáciles, pero en la práctica tiene poco sentido medir más de una ocasión la misma unidad muestral, salvo en diseños específicos u otros más elaborados en los que las complicaciones teóricas sugieren simplificar los supuestos en que se sustenta su análisis.

✓ **Muestreo Aleatorio Simple sin Reemplazo**. En el muestreo sin reemplazo se pueden construir tantas muestras diferentes como combinaciones se pueden hacer de *N* elementos de tamaño n ($N\ C_n$) que se calcula:

$${}_N C_N = \frac{N!}{n!\,(N-n)!}$$

Obtención de una Muestra Aleatoria Simple. Para hacer la selección de las unidades muestrales que constituirán la muestra se parte del marco muestral. A cada unidad se le asigna una etiqueta que la identifique, por ejemplo, números consecutivos y la selección se puede llevar a cabo usando una tabla de números aleatorios, la mayoría de los textos sobre muestreo tienen tablas de números aleatorios regularmente de 10 000 dígitos. Por ejemplo: si la población tiene entre 10 y 100 unidades se requieren dos dígitos para representarla (00 hasta 99), si son entre 100 y 1000 unidades se requiere tres dígitos (000 hasta 999) y así sucesivamente. Entrando sin ningún orden a la tabla se eligen los números aleatorios ubicados consecutivamente, tantos como sean necesarios para representar el total de

números muestrales. Si el número elegido es mayor que el número total de unidades de marco muestral no se toma en cuenta, si es igual o menor entonces la unidad con esa etiqueta se incluirá en la muestra. El proceso sigue análogamente con el siguiente número ubicado en cualquier dirección de la tabla hasta completar el número de unidades elegidas que deben constituir la muestra **n**.

Estimación de la Media y del Total Poblacional.

Al evaluar variables cuantitativas, los parámetros que con mayor frecuencia interesa estimar son media (μ_Y) o el total (τ_Y) de la variable "Y" para toda la población. Estos parámetros tienen las siguientes definiciones:

Media de la Población $\mu_Y = \mu = \frac{\sum_{i=1}^{N} y_i}{N}$

Total de la Población $\tau_Y = \tau\mu = \sum_{i=1}^{N} y_i = N\mu$

Naturalmente al no tener acceso a todas las *N* muestrales, de donde proviene cada y_i, se hace necesario definir estimadores sobre los datos que proporcionan las mediciones que se hacen en las unidades de muestreo incluidas en la muestra. Las expresiones siguientes se denominan estimadores y, una vez que se ejecutan usando los datos de una muestra específica, los valores que se obtienen se denominan estimadas. Estos estimadores de μ y τ son:

Estimador de la Media Poblacional (μ): $\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Estimador del Total Poblacional (τ): $\hat{\tau} = N\bar{y}$. $\hat{\tau}$ es un estimador insesgado del total de la población.

Estimación de la Varianza de Población.

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

Al igual que $\hat{\mu}$ y $\hat{\tau}$, σ^2 tiene su estimador que se obtiene con la muestra ($S^2_{(\text{varianza estimada})}$): $S_Y^2 = S^2_{(\text{varianza estimada})} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

Estimador de la Varianza de la Media ($\mu_{\bar{y}} = \mu$). Tal que $\sigma_{\bar{Y}}^2 = \frac{\sigma_{\bar{y}}^2}{n}\frac{N-n}{N}$. Al conocer los parámetros incluidos en estas expresiones se recurre a usar sus estimadores $\hat{\mu}_{\bar{y}} = \hat{\mu} = \bar{y}$. $S_{\bar{y}}^2 = \frac{S^2_{(\text{varianza estimada})}}{n}\left[\frac{N-n}{N}\right]$.

Estimador de la Varianza del Total $(\mu_{\hat{\tau}} = \tau = N\mu)$**.** Tal que $\sigma^2_{\hat{\tau}} = N^2 \frac{\sigma^2_{\bar{y}}}{n}\left[\frac{N-n}{N}\right]$. Sin embargo, al no conocer los parámetros incluidos en estas expresiones se recurre a usar sus estimadores $\hat{\mu}_{\hat{\tau}} = N\hat{\mu} = N\bar{y}$. Entonces $S^2_{\hat{\tau}} = N^2 \frac{S^2_{\bar{y}}}{n}\left[\frac{N-n}{N}\right]$.

Estimación del Intervalo de Confianza de la Media μ y del total τ de la Población

$$\bar{y} \pm \left[\left(t_{[n-1,(\alpha/2)]}\right) * S_{\bar{y}}\right].\ \text{Donde}\ S_{\bar{y}} = \sqrt{\frac{S^2_Y\,(S^2_{(\text{varianza estimada})})}{n}}\ \textbf{Media Poblacional}$$

$$\hat{\tau} = N\bar{y} \pm \left[\left(t_{[n-1,(\alpha/2)]}\right) * S_{\hat{\tau}}\right].\ \text{Donde}\ S_{\hat{\tau}} = \sqrt{N^2 \frac{S^2}{n}\left[\frac{N-n}{N}\right]}$$

$$= N\sqrt{\frac{S^2}{n}\left[\frac{N-n}{N}\right]}\ \textbf{Total Poblacional}$$

Intervalo de Confianza para Estimación del Total Poblacional

$$\hat{\tau} = N\bar{y} \pm \left[\left(t_{[n-1,(\alpha/2)]}\right) * S_{\hat{\tau}}\right].\ \text{Donde}\ S_{\hat{\tau}} = N\sqrt{\frac{S^2_{\bar{y}}}{n}\left[\frac{N-n}{N}\right]}$$

Tamaño muestral de estimar μ con tamaño de error $\mathfrak{b}$

$$n = \frac{t^2_{(n-1,\alpha/2)} N S^2_{\bar{y}} (S^2_{(\text{varianza estimada})})}{\mathfrak{b}^2_m (N-1) + S^2_{\bar{y}} (S^2_{(\text{varianza estimada})}) t^2_{(n-1,\alpha/2)}}$$

$$\mathfrak{b}_{(tamaño\ de\ error)} = t\sqrt{\frac{S^2_{y\,(estimador\ o\ S^2)}}{n} - \left[\frac{N-n}{N}\right]} \Rightarrow \mathfrak{b}^2 = t^2\left[\frac{S^2_y}{n} - \left[\frac{N-n}{N}\right]\right]$$

Dónde: N es número de unidades muestrales de la población; $S^2_{\bar{y}}$ varianza estimada de la población de interés y $\mathfrak{b}_m$ es tamaño del error de estimación de la media que se acepta. En estricto sentido, para estimar n se requiere un valor de S^2 que se puede lograr tomando una muestra previa y verificar el valor correcto de n (circularidad).

Los estimadores tienen propiedades estadísticas, en cambio las estimadas son realizaciones de los estimadores, los estimadores son variables aleatorias que tienen propiedades estadísticas derivadas de la probabilidad. Enseguida se muestran **dos propiedades deseables de estimadores de media y del total de la población:**

✓ **Insesgamiento:** Un estimador insesgado es aquel que en un número muy grande de estimaciones, tiene un promedio que difiere muy poco del valor del parámetro ($\text{sesgo}(\Theta) = \mathrm{E}(\widehat{\Theta}) - \Theta = 0$).

✓ **Consistencia:** Esta propiedad indica que cuando el tamaño de la muestra es igual al tamaño de la población ($\mathrm{n} = \mathrm{N}$), el estimador es igual al parámetro. La demostración resulta casi evidente por las definiciones de $\bar{y}$, μ,$\hat{\tau}$ y τ. Esta propiedad indica que cuando $\mathrm{n} \to \mathrm{N} \Rightarrow \widehat{\Theta} \to \Theta$. Interpretación: $\bar{y}$ es un estimador consistente de la media poblacional μ y $\hat{\tau}$ lo es del total poblacional τ.

Ejemplo ([175])**:** Se desea hacer un inventario sobre una plantación de coníferas de 20 años de edad, en una superficie de 100 Ha y su objetivo es estimar el volumen medio por Ha de los árboles cuyo diámetro normal sea mayor de 15 Cm. Para esto, se usarán sitios cuadrados de 0.1 Ha (1000 m^2), que se dibujaran en un mapa del área de estudio, se controlaran con números de 1 a 1000 de una tabla de números aleatorios, se seleccionarán 25 sitios de tres dígitos para identificar a las unidades de muestreo que habrán de incluirse en la premuestra. El procedimiento de selección será sin reemplazo

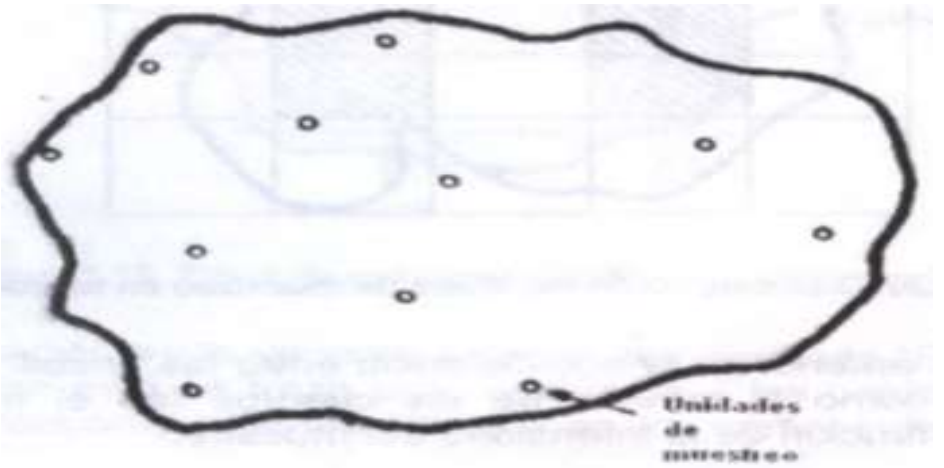

Figura 43. Distribución de unidades de muestreo en MSA ([176])

[175] La resolución de éste ejemplo se encuentra en carpeta “Tipos de Muestreo” y sub carpeta "MSA” o “MIA”. Fuente: Carrillo, E. G. 2005

[176] Fuente: Carrillo, E. G. 2010

Datos:

N = 100 Ha.

n (número sitios total premuestra) = 25 sitios de 0.1 Ha.

$$\bar{y} = \text{Vol. medio por Ha a estimar}$$

En el siguiente cuadro se presentan los datos de volumen (m^3/sitio) con base en la información dasométrica obtenida en pre-muestreo:

Cuadro 17. Información dasométrica de pre-muestreo

Sitio	Sitio	Sitio	Sitio	Sitio
18	16	20	18	17
16	16	18	16	15
18	22	13	19	18
16	15	15	16	18
17	17	15	18	15

5.2.2.2. Muestro aleatorio estratificado (MAE)

Este plan de muestreo trata de hacer aún más precisas las estimaciones que podemos obtener con un diseño básico de muestreo como el aleatorio simple.

Definición: Es el que divide la población en N individuos en E subpoblaciones o *estratos* con respecto a criterios que puedan ser importantes en la investigación. Los estratos contienen $N_1, N_2, \ldots N_E$ unidades tal que:

$$N = \sum_{h=i}^{E} N_h$$

En cada uno de estos estratos o subpoblaciones se hace un muestreo aleatorio simple con muestras de tamaño n_h, por lo que su tamaño es:

$$n = \sum_{h=i}^{E} n_h$$

Muestra aleatoria estratificada. Considera categorías típicas diferentes entre sí (estratos) que poseen gran homogeneidad entre unidades muestrales respecto a alguna característica. Por ejemplo, según la profesión, el municipio de residencia, el género, el estado civil, etc. Lo que se pretende con este tipo de muestreo es asegurarse de que todos los estratos de interés estarán representados adecuadamente en la muestra y que estos no

presentarán traslapes. Cada estrato funciona independientemente, pudiendo aplicarse dentro de ellos un muestreo aleatorio simple, para elegir los elementos concretos que formarán parte de la muestra.

Características de MAE ([177]):

Se usa si la población es muy heterogénea y las consideraciones de costo limitan el tamaño muestral. En MSA sería imposible estimaciones suficientemente precisas y sería demasiado costoso. La población se divide en subpoblaciones llamadas estratos de acuerdo a alguna semejanza a fin de reducir considerablemente la variación entre las mediciones en cada estrato, donde los elementos en cada uno de éstos no se traslapan y en su conjunto constituyen a toda la población. Una vez dividida la población se realiza la selección de una muestra aleatoria irrestricta para cada estrato, lo que nos permite la estimación separada de parámetros poblacionales dentro de cada uno de ellos.

¿Cuándo se usa el MAE?:

- Si se desea cierta precisión en algún estrato y cada uno se considera una población.
- Se obtiene un límite para el error de estimación más pequeño, esto es cierto cuando las mediciones en estrado son homogéneas.
- Reducir costo/observación de encuesta mediante estratificación de elementos de población en grupos convenientes y fácilmente diferenciables.
- Estratificación puede tener mayor precisión en estimaciones de características de total poblacional.

Ventajas del Muestreo Aleatorio Estratificado (MAE) respecto a Muestreo Aleatorio Simple (MAS):

- Generalmente, el estimador de μ presenta menor varianza.
- El costo por muestrear y analizar es menor, pues en vez de tomar observaciones sobre toda la población, sólo toma observaciones sobre estrato, que son más pequeños.
- Al final del análisis, se tiene estimaciones sobre total poblacional y estratos individuales.

[177] Arana Ovalle, R. I. 2003

¿Cómo seleccionar una muestra? Va a ser diferente por estrato, pues cada uno tiene características y costos de medición diferentes, por lo que el número de unidades será diferente. La muestra debe ser mayor si tiene mayor número de unidades, es muy variable internamente en características de sus elementos o es más barato en el estrato. Por el contrario, va a ser menor el tamaño muestral si su costo es elevado. Antes de seleccionar una muestra considere el tamaño del error de estimación y de acuerdo a esto seleccionar el tamaño muestral.

¿Cómo delimitar los estratos? En algunos casos los estratos están ya implícitos pues se conoce el comportamiento con base antiguos registros, o a características fenotípicas; también podría ser con base en la experiencia o simplemente a la naturaleza de los resultados que deseamos obtener.

¿Con base en qué se delimitará los estratos? Una primera aproximación sería en el caso cuantitativo, el construirlos dado un interés particular, pues muchas veces al momento del diseño de la evaluación, se conoce los rangos que se gustaría analizar para obtener las estimaciones. Pero también se puede dar el caso en el que se tiene el rango de salida de los datos y algunas frecuencias en categorías generales de variable de interés o de alguna variable altamente correlacionada ("Método Acumulativo de Raíz Cuadrada de Frecuencia"). Notación de escritura en Muestreo Aleatorio Estratificado **(MAE):**

- **E** número de estratos en población.
- **N** total de unidades muestrales en población.
- $\mathbf{N_h}$ número total de unidades en estrato h.
- $\mathbf{n_h}$ número de unidades muestrales en estrato h.
- $\mathbf{y_{hi}}$ valor obtenido de i-ésima unidad en estrato h.
- $\mathbf{W_h = N_h/N}$ ponderación de estrato (tamaño relativo del estrato)
- $\mathbf{\int_h = n_h/N_h}$ fracción de muestreo para estrato h.
- $\mathbf{y_{st}}$ valores medios por cada estrato $\mathrm{i} = 1,\ 2,\ 3,\ \ldots,\ \mathrm{E}$
- $\mathbf{\bar{y}_h = \frac{\sum_{i=1}^{n_h} \bar{y}_{hi}}{n_h}}$ media de estrato h.
- $\mathbf{S_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}}$ varianza de estrato h
- $\mathrm{t}_{(n_{\text{No. efectivo gl}},\ \alpha/2)}$ Valor T-Sudent con N_o. efectivo de gl y $\alpha/2$

Tal que:

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \text{ Media del estrato h}$$

$$S_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1} \text{ Varianza del estrato h}$$

Estimación de Media Poblacional (μ_{st}): $\bar{y}_{st} = \frac{\sum_{h=1}^{E} N_h \bar{y}_h}{N} = \sum_{h=1}^{E} W_h \bar{y}_h$. Cada estrato es independiente y las $\bar{y}_h$con h=1, 2, 3,… E son independientes. La varianza $\bar{y}_{st}$es la suma de varianzas de medias de cada estrato y es insesgado.

Estimación de varianza de $\bar{y}_{st}$: $\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\left[N_1^2 \hat{V}(\bar{y}_1) + N_2^2 \hat{V}(\bar{y}_2) + \cdots N_E^2 \hat{V}(\bar{y}_E)\right]$

$$= \frac{1}{N^2}\left[N_1^2 \left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{S_1^2}{n_1}\right) + \cdots N_E^2 \left(\frac{N_E - n_E}{N_E}\right)\left(\frac{S_E^2}{n_E}\right)\right] = \frac{1}{N^2}\sum_{h=1}^{E} N_h^2 \left(\frac{N_h - n_h}{N_h}\right)\left(\frac{S_h^2}{n_h}\right)$$

$$= \sum_{h=1}^{E} \frac{N_h^2}{N^2}\left(\frac{N_h - n_h}{N_h}\right)\left(\frac{S_h^2}{n_h}\right) = \sum_{h=1}^{E} W_h^2 \left(\frac{N_h - n_h}{N_h}\right)\left(\frac{S_h^2}{n_h}\right) = \sum_{h=1}^{E} W_h^2 \, S_{\bar{y}_h}^2$$

Enseguida se muestra el método aproximado de asignación de número efectivo de grados de libertad a $S^2(\bar{y}_{st})$(Satterthwaite, 1946):

$$n_e = \frac{\left(\sum_{h=1}^{E} g_h S_h^2\right)^2}{\sum_{h=1}^{E} \frac{g_h^2 S_h^4}{n_h - 1}} \text{ Donde: } g_h = N_h\left(\frac{N_h - n_h}{N_h}\right)$$

Intervalo de confianza para estimador:

$$\bar{y}_{st} \pm t\sqrt{\hat{V}(\bar{y}_{st})} = \bar{y}_{st} \pm t_{(n_e, \alpha/2)} \sqrt{\frac{1}{N^2}\sum_{h=1}^{E} N_h^2 \left(\frac{N_h - n_h}{N_h}\right)\left(\frac{S_h^2}{n_h}\right)} = \bar{y}_{st} \pm t_{(n_e, \alpha/2)} \sqrt{\frac{1}{N^2}\sum_{h=1}^{E} W_h^2 \, S_{\bar{y}_h}^2}$$

Estimador del Total Poblacional:

$$\hat{\tau}_{st} = N\bar{y}_{st} = N_1\bar{y}_1 + \cdots N_E\bar{y}_E = \sum_{h=1}^{E} N\bar{y}_{st}$$

Varianza Estimada del Estimador del Total Poblacional:

$$\hat{V}(N\bar{y}_{st}) = N^2\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{E} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\left(\frac{S_i^2}{n_i}\right)$$

Intervalo de confianza: $N\bar{y}_{st} \pm t_{(n_e, \alpha/2)}\sqrt{N_h^2 \left(\frac{N_h - n_h}{N_h}\right)\left(\frac{S_h^2}{n_h}\right)}$

Tamaño de Muestra Aproximado que se requiere para estimar μ y τ:

$$n = \frac{\frac{\sum_{h=1}^{E} N_h^2 S_h^2}{W_h}}{N^2 B_M^2 / \left(t_{(\alpha, n)}\right)^2 + \sum_{h=1}^{E} N_h S_h^2}$$

Donde: $W_h = N_h/N$ y B_MTamaño de error que desea acepta en estimar en μ

$$n = \frac{\frac{\sum_{h=1}^{E} N_h^2 S_h^2}{W_h}}{B_M^2 / \left(t_{(\alpha, n)}\right)^2 + \sum_{h=1}^{E} N_h S_h^2}$$

B_TTamaño de error que desea acepta en estimar en τ

Ejemplo ([178])**:** Es necesario muestrear un bosque de 800 Ha para la estimación del volumen medio de madera por hectárea expresado en metros cúbicos. Mediante fotos aéreas, la superficie se dividió en 3 estratos: Teca, Fernán Sánchez y Cedros. Se conocen los límites y la extensión total de cada tipo de bosque. Se seleccionaron al azar y sin reemplazo en cada estrato diez sitios de un décimo de hectárea cada uno. Así las observaciones se dividieron de la siguiente manera:

Cuadro 18. Información dasométrica de 3 estratos ([179])

No	Estrato	Tamaño	Observaciones	Media muestral ($\bar{y}_i$)
1	Teca (*Tectona grandis)*	3 200 sitios	16 14 17 18 17 22 14 19 20 16 ∑ = **173**	$\bar{y}_1 = \frac{173}{10} = 17.3\ m^3/sitio$
2	Fernán Sánchez (*Triplaris cumingiana*)	1 400 sitios	15 18 23 20 22 16 22 25 24 24 ∑ = **209**	$\bar{y}_2 = \frac{209}{10} = 20.9\ m^3/sitio$
3	Cedro (*Cedrela sp)*	3 400 sitios	12 15 9 6 5 8 8 7 6 10 ∑ = **86**	$\bar{y}_3 = \frac{86}{10} = 8.6\ m^3/sitio$
		$\sum = \mathbf{8\ 000}\ \boldsymbol{sit}$		

[178] La resolución de este ejemplo se encuentra en carpeta “Tipos de Muestreo” y sub carpeta "MAE".
[179] Carrillo, E. G. 2005

5.2.2.3. Muestro por Conglomerados en Una Etapa (MCUE)

Características de MCUE:

- Muestreo por conglomerados es ampliamente usado cuando el costo de muestrear unidades primarias es despreciable en relación con el censo de unidades secundarias.
- La selección primaria de elementos que estarán en la muestra sigue el mismo procedimiento que en el muestreo simple aleatorio, por lo que los estimadores de la media μ y el total τ se obtienen de manera similar.
- Los datos que nos proporciona el muestreo por conglomerados nos permiten obtener estimaciones a diferentes niveles de la población.

Muestra por Conglomerados. Se le denomina así a la muestra obtenida aleatoriamente (de la misma forma que en el muestreo simple aleatorio) y a las unidades obtenidas les llamaremos conglomerados, los cuales son grupos o colecciones de elementos sobre los que se hará la medición o revisión de la característica de interés (en un bosque se podría elegir sitios de cierta superficie como conglomerados). Además, si desea hacer una selección aleatoria de elementos se debe contar con el marco de muestreo, para después hacer el sorteo, pero representa un costo que se incrementa al tomar mediciones que se encuentran separadas entre sí por una gran distancia física. En el muestreo por conglomerados, este costo se reduce, pues al levantar la información de elementos contiguos o muy cercanos entre sí se evita el gasto de traslado.

Cuadro 19. Comparación Muestreo por Conglomerados vs Estratificado [180]

Comparación Muestreo por Conglomerados vs Estratificado	
Muestreo Estratificado	**Muestreo por Conglomerados**
1. Generalmente nos da **más precisión en relación con muestreo simple aleatorio.**	1. Generalmente nos da **menos precisión en relación con muestreo simple aleatorio.**
2. Para una mayor precisión los estratos deben contener elementos que sean lo más homogéneo posible entre ellos.	2. Para una mayor precisión los conglomerados deben contener elementos que sean lo más heterogéneo posible entre ellos.
3. Para una mayor precisión la diferencia entre estratos debe ser considerable.	3. Para una mayor precisión los conglomerados deben ser muy similares.
4. La varianza de la estimación de la media depende de la variabilidad de los valores dentro del estrato.	4. La varianza de la estimación de la media depende de la variabilidad que existe entre las medias de los conglomerados.

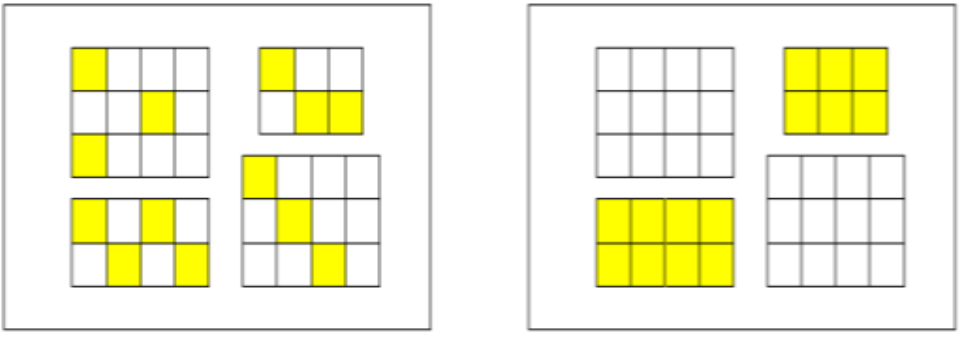

Figura 44. Comparación gráfica de muestreo estratificado vs conglomerados [181]

Notación:

[180] Fuente: Arana Ovalle, R. I. 2003

[181] Fuente: Arana Ovalle, R. I. 2003

❖ **UNIDADES PRIMARIAS**:

✓ $\boldsymbol{N}$ Número de Conglomerados en la Población

✓ $\boldsymbol{n}$ Número de Conglomerados seleccionados de una muestra simple aleatoria

✓ $\boldsymbol{y_i} = \sum_{j=1}^{\boldsymbol{M_i}} \boldsymbol{y_{ij}}$ Total en unidad Primaria o Conglomerado i

✓ $\boldsymbol{\tau} = \sum_{i=1}^{\boldsymbol{N}} \boldsymbol{y_i} = \sum_{i=1}^{\boldsymbol{N}} \sum_{j=1}^{\boldsymbol{M_i}} \boldsymbol{y_{ij}}$ Total, de la población

❖ **UNIDADES SECUNDARIAS**:

✓ $\boldsymbol{M_i}$ Número de Elementos en Conglomerado $(i = 1,\ 2,\ 3, \dots,\ N)$

✓ $\boldsymbol{M}$ Número de elementos en la población $(M = \sum_{i=1}^{N} M_i)$

✓ $\boldsymbol{\bar{M}}$ Tamaño Promedio del Conglomerado en la Muestra

✓ $\boldsymbol{y_i}$ Total del Conglomerado i.

✓ $\boldsymbol{y_{ij}}$ j-ésima observación en i-ésimo Conglomerado

✓ $\boldsymbol{\bar{y}_{iD}} = \sum_{j=1}^{\boldsymbol{M_i}} \frac{\boldsymbol{y_{ij}}}{\boldsymbol{M_i}} = \frac{\boldsymbol{y_i}}{\boldsymbol{M_i}}$ Media de Población en Unidad Primaria i

Elementos a estimar en MCUE:

➢ **Estimador de Media Poblacional:**

$$\bar{y}_c = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$$

➢ **Varianza Estimada de** $\bar{y}_c$:

$$\hat{V}(\bar{y}_c) = \left[\frac{N - n}{Nn\bar{M}^2}\right]\left[\frac{\sum_{i=1}^{n}(y_i - \bar{y}_c M_i)^2}{n - 1}\right]$$

➢ **Intervalo de Confianza** $\bar{y}_c$**:**

$$\bar{y}_c \pm t_{(n-1,\ \alpha/2)}\sqrt{\hat{V}(\bar{y}_c)}$$

➢ **Estimador del Total Poblacional**.

$$\hat{\tau}_c = M\bar{y}_c = \sum_{i=1}^{N} M_i * \left[\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}\right]$$

➢ **Varianza Estimada de** $\hat{\tau}_c$:

$$\hat{V}(\hat{\tau}_c) = N^2\left[\frac{N - n}{Nn}\right]\frac{\sum_{i=1}^{n}(y_i - \bar{y}_c M_i)^2}{n - 1}$$

➢ **Varianza del Estimador del Total Poblacional**:

$$\sigma^2_{(\hat{\tau}_c)} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_c M_i)^2}{n - 1}$$

- **Intervalo de Confianza para $\hat{\tau}_c$:**

$$\hat{\tau}_c \pm t_{(n-1,\,\alpha/2)}\sqrt{\hat{\mathbb{V}}(\hat{\tau}_c)}$$

Ejemplo ([182])**:** El dueño de una plantación forestal necesita estimar el volumen de madera en m^3 que tiene su terreno, lo que ha pensado es hacer un muestreo por conglomerados, para esto divide la plantación en 600 sitios de los cuales muestrea todos los elementos de 60 de ellos, en este caso nuestras unidades primarias (conglomerados) son los sitios y las unidades secundarias son los árboles.

Cuadro 20. Información de muestreo por conglomerados de plantación en 600 sitios (183)

n	M_i	y_i	$(y_i - \bar{y}_c M_i)^2$
1	508	1709	39054
2	302	1075	3418
3	693	3087	236235
4	598	1729	265644
5	459	1497	50947
6	695	2725	13580
7	476	2143	127080
8	675	2945	169413
9	432	1355	70957
10	576	2267	11059
11	657	2724	66644
12	650	2537	9492
13	667	3284	609373
14	598	2370	15774
15	548	2026	945
16	657	1987	229292
17	508	1479	182859
18	499	1668	41960
19	549	2163	10506
20	543	2463	180641

[182] La resolución de este ejemplo se encuentra en carpeta “Tipos de Muestreo” y sub carpeta "MCUE"
[183] Fuente: Arana Ovalle, R. I. 2003

n	M_i	y_i	$(y_i - \bar{y}_e M_i)^2$
21	558	2240	21235
22	598	2005	57316
23	532	2057	3637
24	599	2562	98496
25	607	1853	180783
26	609	2698	169998
27	640	3066	440842
28	659	1948	275994
29	589	1942	72161
30	674	2413	13607
31	508	1870	1341
32	302	987	21451
33	693	3258	431702
34	598	2700	207565
35	459	1750	745
36	583	2007	32800
37	476	1231	308600
38	675	2701	28089
39	432	1669	2268
40	567	1904	50202

n	M_i	y_i	$(y_i - \bar{y}_c M_i)^2$
41	657	1722	553305
42	653	2653	40872
43	667	3092	346477
44	608	2153	16625
45	548	1883	30188
46	657	1650	665603
47	506	2266	134606
48	499	2478	366217
49	449	2151	216987
50	543	1851	34962
51	558	1309	616663
52	598	1881	132064
53	532	2324	107128
54	599	2766	268158
55	607	2142	18546
56	609	1968	100928
57	640	1842	313645
58	659	2862	151048
59	589	1951	67407
60	674	2447	6831
Σ	34,500	129,485	8'941,964

5.2.2.4. Muestro por Conglomerados en Dos Etapas (MCDE)

Características de MCDE:

- El muestreo por conglomerados en dos etapas es en esencia muy parecido al de una etapa y también busca facilitar el manejo de los datos para reducir el costo de operación.
- Las organizaciones privadas o de gobierno desean resultados confiables. Al diseñar una encuesta por conglomerados se debe resolver: a) Precisión global necesaria, b) Número de unidades primarias a seleccionar, c) Número de unidades secundarias a seleccionar por cada unidad primaria y d) Los valores a buscar son n y el de todas m_i. La mejor selección de estos valores depende de las fuentes de variación (dentro de los conglomerados y la que existente entre ellos).
- Proporciona más ideas que podrán ayudar a decidir el diseño más adecuado para medir el fenómeno en cuestión.
- La diferencia entre el muestreo por conglomerados de una y dos etapas, radica en la forma de seleccionar las unidades secundarias.

- Presenta dos principales ventajas: a) No se tiene que hacer el proceso de aleatorización a cada elemento de la población, lo cual puede ahorrarnos una tarea muy complicada y b) Si se trata de unidades que se encuentran geográficamente separadas puede ahorrar costos en transportación para la toma de la muestra.
- Este muestreo es muy útil cuando se trata de poblaciones con muchos elementos, como municipios, unidades habitacionales; en la industria puede resultar muy útil cuando necesitamos muestrear cientos de unidades que vienen empacadas en cajas para validar su calidad o cuando se trata de productos que tienen varios componentes.
- El muestreo por conglomerados en dos etapas es usado cuando resulta menos costoso hacer un censo de unidades secundarias contiguas que hacer un muestro aleatorio sobre toda la población.
- Este muestreo presenta la ventaja de hacer más económico el costo de muestrear, pues en lugar de hacer un censo en cada unidad primaria o conglomerado, se toma una muestra aleatoria. La selección de unidades primarias se hace como en el muestreo aleatorio simple por lo que la media μ y el total τ se obtienen de la misma manera que en el muestreo en una etapa.
- Si las medias entre conglomerados varían mucho unas de otras y sus mediciones son homogéneas, entonces se selecciona muchos conglomerados de pocos elementos, pero si las mediciones varían de manera considerable entre ellas y las medias entre conglomerados son homogéneas se muestrea pocos conglomerados con muchas mediciones en cada uno de ellos.

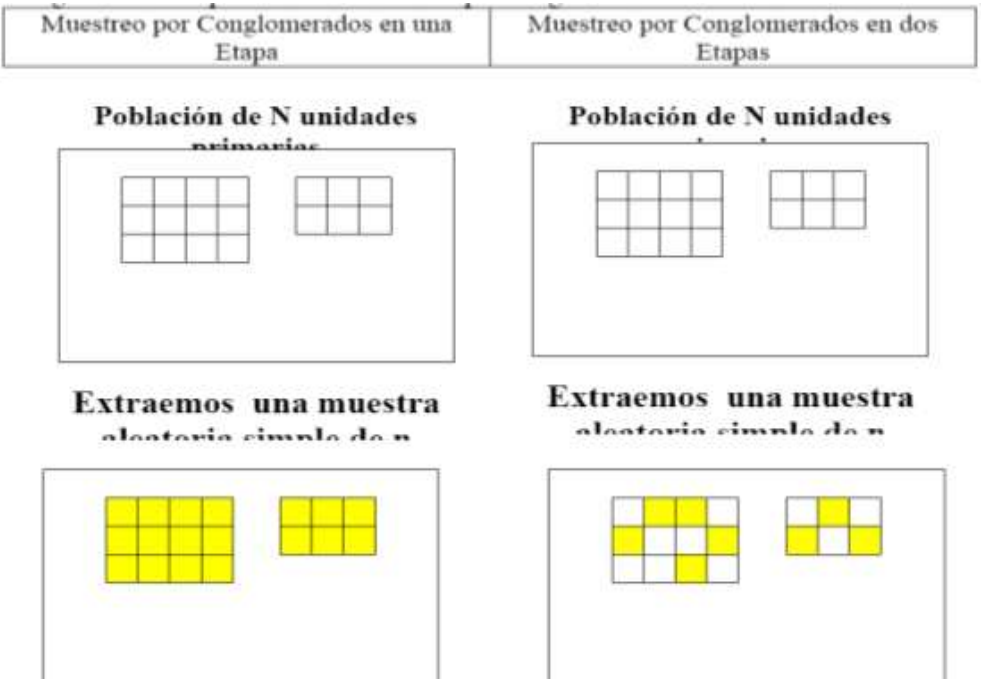

Figura 45. Comparación de muestreos por conglomerados ([184])

Notación para Muestreo por Conglomerados en Dos Etapas (MCDE)**:**

UNIDADES PRIMARIAS:

- ✓ **N** Número de Conglomerados o Unidades Primarias en la Población
- ✓ **n** Número de Conglomerados seleccionados de una muestra simple aleatoria

UNIDADES SECUNDARIAS:

- ✓ $\mathbf{M_i}$ Número de unidades secundarias en Conglomerado $(\mathrm{i} = 1,\ 2,\ 3, \ldots,\ \mathrm{N})$
- ✓ $\mathbf{m_i}$ Número de unidades secundarias seleccionadas en una muestra aleatoria del conglomerado i
- ✓ **M** Número de unidades secundarias en la población $\left(\mathrm{M} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{M_i}\right)$
- ✓ $\overline{\mathbf{M}}$ Tamaño del Conglomerado Promedio en la Población $\left(\overline{\mathrm{M}} = \frac{\mathrm{M}}{\mathrm{N}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{M_i}}{\mathrm{N}}\right)$
- ✓ $\bar{\mathbf{y}}_{\mathbf{i}} = \frac{\mathbf{1}}{\mathbf{m_i}} \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{m_i}} \mathbf{y_{ij}}$ Media muestral para el i-ésimo conglomerado
- ✓ $\mathbf{y_{ij}}$ j-ésima unidad secundaria en i-ésimo Conglomerado

[184] Fuente: Arana Ovalle, R. I. 2003

Elementos a estimar en MCDE:

Total:

$$\tau = \frac{N}{n}\sum_{i=1}^{n} y_i$$

Estimador de Media Poblacional:

$$\bar{y}_{2c} = \left[\frac{N}{M}\right]\left[\frac{\sum_{i=1}^{n} M_i \bar{y}_i}{n}\right]$$

Estimador de Varianza de $\bar{y}_{2c}$:

$\hat{V}(\bar{y}_{2c}) = \left[\frac{N-n}{N}\right]\left[\frac{1}{n\bar{M}^2}\right] s_b^2 + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{n} M_i^2 \left[\frac{M_i - m_i}{M_i}\right]\left[\frac{s_i^2}{m_i}\right]$ Donde:

$$s_b^2 = \frac{\sum_{i=1}^{n}(M_i\bar{y}_i - \bar{M}\hat{\mu})^2}{n-1} \quad s_i^2 = \frac{\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2}{m_i - 1} \text{ Tal que } i = 1,\ 2,\ 3,\ \ldots,\ n$$

Intervalo de Confianza para $\bar{y}_{2c}$:

$$\bar{y}_{2c} \pm t_{(n-1,\,\alpha/2)}\sqrt{\hat{V}(\bar{y}_{2c})}$$

Estimador del Total Poblacional:

$$\hat{\tau}_{2c} = M\bar{y}_{2c} = M\left[\frac{N}{M}\right]\left[\frac{\sum_{i=1}^{n} M_i\bar{y}_i}{n}\right] = N\left[\frac{\sum_{i=1}^{n} M_i\bar{y}_i}{n}\right]$$

Varianza Estimada de $\hat{\tau}_{2c}$:

$$\hat{V}(\hat{\tau}_{2c}) = M^2\hat{V}(\bar{y}_{2c}) = \left[\frac{N-n}{N}\right]\left[\frac{N^2}{n}\right]s_b^2 + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left[\frac{M_i - m_i}{M_i}\right]\left[\frac{s_i^2}{m_i}\right]$$

Intervalo de Confianza para $\hat{\tau}_{2c}$:

$$\hat{\tau}_{2c} \pm t_{(n-1,\,\alpha/2)}\sqrt{\hat{V}(\hat{\tau}_{2c})}$$

Es importante **considerar:**

- ✓ σ_b^2= Varianza entre las medias de conglomerados
- ✓ σ_w^2= Varianza entre elementos dentro de conglomerados
- ✓ $C = c_1 + nmc_2$Es costo total por muestrear. Donde: c_1 es costo de muestrear cada unidad primaria y, por ende, c_2 es costo de muestrar cada unidad secundaria. Bajo estos supuestos, el valor m que minimiza la varianza con un costo fijo está dado por la ecuación:

Tamaño de m aproximado requerido para estimar μ.

$\mathrm{m} = \sqrt{\frac{\sigma_w^2 c_1}{\sigma_b^2 c_2}}$ Donde: σ_w^2 es estimada por $s_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2$ y

$$\sigma_b^2 \text{ es estimada por } s_b^2 = \left[\frac{1}{n-1}\sum_{i=1}^{n}[\bar{y}_i - \hat{\mu}]\right] - \frac{s_w^2}{m}$$

Para conocer el número de unidades primarias que minimizarán la varianza, se usa la siguiente expresión:

Tamaño de n aproximado requerido para estimar μ:

$n = \frac{1}{\sigma_{2c}}\left[\sigma_b^2 + \frac{\sigma_w^2}{m}\right]$Donde σ_{2c}^2es estimada por $\hat{V}(\bar{y}_{2c})$

Ejemplo ([185])**:** Un centro de investigación desea saber la cantidad de maíz que produce una planta de una nueva variedad con la que están experimentando. Cuentan con 40 campos donde plantaron la nueva variedad en una melga (surcos) de experimentación que mide $1\ \mathrm{m} * 100\ \mathrm{m}$, tal que las semillas se plantaron a una distancia de $1\ m$ una de otra. Han decido que aplicarán un muestreo por conglomerados en dos etapas. En este caso se seleccionaron al azar 15 unidades primarias (melgas donde cultivan la nueva variedad de maíz) y, también, al azar las unidades secundarias sobre las que se hará la observación (plantas que están dentro de las melgas). Hay que tomar en cuenta que aunque se sembraron 8,000 semillas, no todas llegaron a desarrollarse y que en general se toma un 8% de mortandad por lo que $\mathrm{M} = 7{,}200$ plantas.

[185] La resolución de éste ejemplo se encuentra en carpeta “Tipos de Muestreo” y sub carpeta "MCDE"

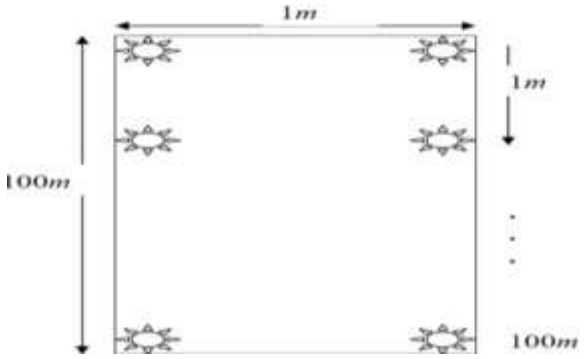

Figura 46. Comparación de muestreos por conglomerados [186]

Cuadro 21. Investigación de Productividad de una nueva variedad de Maíz [187]

Investigación de Productividad de una nueva variedad de Maíz							
n	M_i	m_i	$\bar{y}_i(Kg)$	$M_i\bar{y}_i$	S_i^2	$\sum_{i=1}^{n}(M_i\bar{y}_i - \bar{M}\hat{\mu})^2$	$\sum_{i=1}^{n} M_i^2\left[\frac{M_i - m_i}{M_i}\right]\left[\frac{S_i^2}{m_i}\right]$
1	194	16	0.590	114.46	0.00266	134.185	5.741
2	200	16	0.610	122.00	0.00329	358.579	7.567
3	175	14	0.533	93.28	0.00279	92.838	5.615
4	163	13	0.495	80.69	0.00299	494.829	5.624
5	181	14	0.522	94.48	0.00249	8.926	5.376
6	171	14	0.524	89.60	0.00269	179.483	5.158
7	197	16	0.605	119.19	0.00239	261.454	5.326
8	186	15	0.567	105.46	0.00259	5.912	5.492
9	175	14	0.533	93.28	0.00399	92.838	8.030
10	192	15	0.586	112.51	0.00318	90.033	7.205
11	174	14	0.533	92.74	0.00284	103.400	5.648
12	173	14	0.529	91.52	0.00332	132.842	6.523
13	185	15	0.567	104.90	0.00289	3.477	6.059
14	187	15	0.571	106.78	0.00239	15.121	5.125
15	197	16	0.600	118.20	0.00254	231.996	5.661
Σ	**2750**	**221**		**1539.07**		**2205.913**	**90.149**
Promedio	**183.33**						

[186] Fuente: Arana Ovalle, R. I. 2003
[187] Fuente: Arana Ovalle, R. I. 2003

5.2.2.5. Muestro Sistemático (MS)

Características del MS:

- El muestreo sistemático puede ser una excelente alternativa para sustituir al simple aleatorio y, algunas veces, es más preciso, pero depende de las características de la población a analizar, por lo que es necesario conocer algo sobre la estructura de la población.
- El muestreo sistemático generalmente resulta más simple y barato al momento de seleccionar la muestra.
- El muestreo sistemático es preferible cuando la población está ordenada, pues tiene la seguridad de recorrer todos los elementos de la población y tener una muestra representativa de esta, pero si la muestra es aleatoria los resultados son equivalentes al muestreo simple aleatorio.
- Debe tener cuidado al momento de tener una población periódica debido a que puede ocurrir que, al elegir el tamaño de nuestra k, las unidades muestrales siempre caigan en un lugar del ciclo y dejen la otra parte de este, por lo que la población no sería representada por la muestra.
- Es importante hacer notar que puede hacer un muestreo sistemático en un diseño estratificado, de razón o por conglomerados.
- Anteriores muestreos usan forma aleatoria en la selección de la muestra que implica un proceso complicado y costoso.
- El diseño de muestreo o de encuestas por muestreo sistemático, el cual es ampliamente utilizado pues representa una significativa reducción del proceso de selección de la muestra.
- Este diseño elimina la necesidad de desarrollar métodos de aleatorización elaborados ya que sólo requiere fijar un intervalo y de ahí recorrer la población seleccionando las unidades que se encuentren en el punto seleccionado del intervalo. Esto, evidentemente facilita el trabajo de campo en el muestreo y reduce sustancialmente los errores que se podrían cometer en caso de hacer uso de un procedimiento más elaborado.
- En este método se tiene la certeza de cubrir la totalidad de la población a analizar desde un inicio.

➢ De esta manera, el tiempo que consumirá y el costo de selección por unidad muestral será menor.

➢ El principal problema de las muestras extraídas de este tipo de población radica en extraer una muy sesgada, pues si elige un tamaño k demasiado pequeña que siempre cayera en el mismo lugar del intervalo y obtendría la caracterización de esa parte del ciclo y no de la población total. Si tomará un valor de k más grande que lograra romper el ciclo, los resultados serán más alentadores $(\mathrm{V}(\bar{y}_{sy}) \leq \mathrm{V}(\bar{y})$: varianza del MSA es mayor que varianza de Mestreo Sistemático)

Muestra Sistemática. Es una muestra que se obtiene seleccionando una unidad muestral por cada k unidades en una población de tamaño N. De esta manera, tomando el valor apropiado de k, se dice que se tiene una muestra de 1 en k. A este tipo de muestra se denota como y_{sY} (o media de muestra sistemática). Regularmente N es múltiplo de k y a cada conjunto de k unidades se le llama grupo (figura 1). Cabe señalar que existe el muestreo sistemático cuando N no es múltiplo de k.

Grupo	1	2	3	⋯	k
1	1	2	3	⋯	k
2	$k+1$	$k+2$	$k+3$	⋯	$2k$
3	$2k+1$	$2k+2$	$2k+3$	⋯	$3k$
⋮	⋮	⋮	⋮		⋮
j	$(j-1)k+1$	$(j-1)k+2$	$(j-1)k+3$	⋯	jk
⋮	⋮	⋮	⋮		⋮
n	$(n-1)k+1$	$(n-1)k+2$	$(n-1)k+3$	⋯	$nk=N$

Figura 47. Muestreo Sistemático [188]

Selección de una Muestra Sistemática. Primero se decide el tamaño del intervalo "1-en k" unidades, luego selecciona aleatoriamente una unidad que se encuentre dentro del intervalo de la primera hasta la k-ésima unidad y así seguirá tomando los múltiplos de k, hasta llegar a N. Para seleccionar una k adecuada para una muestra sistemática de n elementos en una población de N, k debe ser menor o igual que N/n; si no conociera a N, entonces determinará

[188] Fuente: Arana Ovalle, R. I. 2003.

un tamaño de muestra "n" aproximado para la encuesta y así estar en la posibilidad de obtener una k estimada. Sin embargo, es necesario tener una precisión dada desde el principio de estudio.

Población Aleatoria. Se llama así cuando se encuentre a las unidades muestrales ordenadas al azar dentro de la población. La muestra extraída de una población aleatoria debe conservar un coeficiente de correlación aproximadamente igual a cero $(\rho_{xy} \sim 0)$, es decir, que si se tiene una N grande, la varianza de y_{sY} es aproximadamente igual a la varianza de y, de esta forma el muestreo sistemático es equivalente al simple aleatorio. Sin embargo, este tipo de muestras suelen ser heterogéneas y generalmente con un coeficiente de correlación menor o igual a cero $(\rho_{xy} \leq 0)$. Si fuese el caso y tiene una N suficientemente grande se tiene que $V(\bar{y}_{sy}) \leq V(\bar{y})$. Es decir, la varianza del MSA es mayor que la Sistemática. Por lo tanto, una muestra sistemática ordenada proporciona más información que una muestra simple aleatoria por unidad de costo.

Estimación de la Media de Muestra Sistemática (μ):

$$\hat{\mu} = \bar{y}_{sY} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Varianza Estimada de $(\hat{\mu})$:

$$\hat{V}(\bar{y}_{sY}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

Intervalo de Confianza de $(\hat{\mu})$:

$$\bar{y}_{sY} \pm t_{(n-1,\, \alpha/2)} \sqrt{\frac{s^2}{n}\left[\frac{N-n}{N}\right]}$$

Así la varianza verdadera del estimador de la media de una muestra sistemática es:

$$V(\bar{y}_{sY}) = \frac{\sigma^2}{n}\left[1 + (n-1)\rho_{xy}\right]$$

Así pues, el muestreo sistemático estará muy ligado a este indicador ya que si ρ_{xy} es cercano a 1, quiere decir que los elementos están estrechamente relacionados y esto nos producirá una mayor varianza de la media que en el muestreo simple aleatorio, por lo que este último será el más indicado. En el caso contrario, si ρ_{xy} es cercano a cero, nuestra

estimación por muestreo sistemático es la más recomendada ya que en este caso la varianza es aproximadamente igual al muestreo simple aleatorio.

Ejemplo ([189])**:** El dueño de una compañía de transportes vendiendo su proveedor es de la calidad especificada a los 3 meses de haberlo suministrado a los vehículos; uno de los principales inconvenientes es que en ningún momento están todos los automotores en la central, por lo que hacer un muestreo simple aleatorio podría representar algunos problemas, por esto se ha decidido tomar un muestreo sistemático, seleccionando cada *k* vehículos según su llegada a la central, sabemos que la compañía cuenta con 1,200 vehículos e interesa muestrear 60 de ellos y a cada uno de ellos sacarle una muestra de aceite para medir sus grados Poise (que es la viscosidad) en el laboratorio, cabe mencionar que el método de medición es complicado pues se toman diferentes variables en cuenta. Asimismo, según especificaciones el aceite debe estar entre 0.5 y 0.7 grados Poise dependiendo del vehículo.

Se elige una muestra k:

$$k = \frac{N}{n} = \frac{1200}{60} = 20$$

Esto indica que se deben muestrear cada 20 elementos, eligiendo de manera aleatoria el primer elemento entre los primeros 20. Datos:

Cuadro 22. Muestreo Sistemático de 20 elementos ([190])

N$_o$. de Muestra	**Grados Poise**
Vehículo 2	0.5342
Vehículo 22	0.6340
Vehículo *k*-ésimo	0.6780
Vehículo 1,142	0.7128
Σ	33.9538
$\boldsymbol{\sigma^2}$	0.0935

5.2.2.6. Muestro Sistemático con Repeticiones ó Replicado (MSR)

Se usa cuando la población no es aleatoria y para estimar la varianza de la media, que usa el mismo principio que el Muestreo Sistemático Simple. Usa replicas, por lo que se recorre

[189] La resolución de éste ejemplo se encuentra en carpeta “Tipos de Muestreo” y sub carpeta "MS”
[190] Fuente: Arana Ovalle, R. I. 2003

la población tomando varias muestras sistemáticas al mismo tiempo, que tendrán un punto de inicio k diferente. Esto es:

- Se encuentra con una población de N elementos que se puede enumerar consecutivamente, de donde se selecciona una muestra de tamaño n.
- Se obtiene $k = {}^{N}/_{n}$, se selecciona un número aleatorio entre 1 y k. De ahí, se construye $k' = n_s * k$ será el nuevo tamaño de intervalo. Se muestrea elementos de 1 a k'. Tal que $n_s$es número de replicas a usar en el diseño, que usualmente son 10 para obtener estimaciones satisfactorias para la varianza. Observación: el valor de k'se construye tal que al final se tienen el mimo número de mediciones que se obtendría con una sola muestra de 1 en k¸ por lo que muestrear con réplicas no representa un costo mayor.
- Se selecciona $n_s$números aleatorios entre uno y k'que será puntos de inicio para cada una de las muestras, de ahí se recorre la población de k'en k'para cada una de éstas hasta llegar al último elemento N, en este momento se tendrá n^* elementos de cada réplica. Tal que, $n = n_s . n^*$(n es tamaño total de muestra), donde n representa el número de unidades muestrales incluidas en una muestra sistemática sin repeticiones.

Elementos a estimar en el MSR o Muestreo Replicado:

Estimación de la Media μ para Muestras Sistemáticas Replicadas:

$$\hat{\mu} = \bar{y}_{sY} = \sum_{i=1}^{n_s} \frac{\bar{y}_i}{n_s}$$

Varianza estimada de $\hat{\mu}$:

$$\hat{V}(\bar{y}_{sY}) = \left[\frac{N-n}{N}\right]\left[\frac{\sum_{i=1}^{n_s}(\bar{y}_i - \hat{\mu})^2}{n_s(n_s-1)}\right]$$

Intervalo de Confianza:

$$\bar{y}_{sY} \pm \left(t_{(n_{(\text{tamaño total de muestra})}-1,\, \alpha/2)}\sqrt{\hat{V}(\bar{y}_{sY})}\right)$$

Estimación del Total τ_{sY} para Muestras Sistemáticas Replicadas:

$$\hat{\tau}_{sY} = N\bar{y}_{sY} = N\sum_{1=1}^{n_s} \frac{\bar{y}_i}{n_s}$$

Varianza Estimada de τ_{sY}:

$$\hat{V}(\hat{\tau}_{sY}) = N^2\left[\frac{N-n}{N}\right]\frac{\sum_{i=1}^{n_s}[\bar{y}_i - \hat{\mu}]^2}{n_s(n_s-1)}$$

Intervalo de Confianza del Estimador del Total:

$$\hat{\tau}_{sY} \pm \left(t_{(n-1,\, \alpha/2)} \sqrt{\hat{V}(\hat{\tau}_{sY})} \right)$$

Ejemplo ([191])**:** Una empresa desea conocer la opinión de sus clientes acerca de sus servicios; para esto realiza una encuesta de opinión sobre los 1,000 clientes y cree suficiente muestrear a 70 de estos. Las respuestas de los clientes van de uno en uno hasta cinco donde $1 = \text{muy mal servicio}$ y $5 = \text{muy buen servicio}$. Es importante mencionar que cada cliente tiene un número consecutivo que lo identifica y la empresa cuenta con un listado, sobre el cual se seleccionará sistemáticamente. $k = \frac{1000}{70} = 14.2$ Aunque con $n_s = 10 \Rightarrow k' = 10 * 14 = 140$. Entonces, se selecciona aleatoriamente 10 números entre 1 y 140. Los resultados de muestreo se exponen, donde los números entre paréntesis representa el número de cliente seleccionado y la cantidad que enseguida aparece es la calificación dada a la empresa.

[191] La resolución de éste ejemplo se encuentra en carpeta “Tipos de Muestreo” y sub carpeta "MSR"

Cuadro 23. Base de datos de opinión de 70 clientes ([192])

1ra muestra		2da. muestra		3ra muestra		4ta muestra		5ta muestra		6ta muestra		7ma muestra		$\bar{y}_i$	$\sum_{i=1}^{n_s}(\bar{y}_i - \hat{\mu})^2$
Cliente	Calif.	Cliente	Calif.	Cliente	Calif.	Cliente	Calif.	Cliente	Calif.	Cliente	Calif.	Cliente	Calif.		
2	1	142	2	282	4	422	3	562	5	702	4	842	1	2.86	0.22
5	3	145	3	285	4	425	3	565	3	707	4	845	2	3.14	0.03
25	2	165	5	305	5	445	4	585	4	725	2	865	5	3.86	0.28
62	5	202	5	342	2	482	5	622	5	762	3	902	3	4.00	0.45
67	5	207	4	347	2	487	1	627	5	767	2	907	3	3.14	0.03
80	3	220	2	365	3	500	2	640	1	780	1	920	4	2.29	1.09
98	2	238	1	378	1	518	2	658	2	798	5	938	2	2.14	1.41
122	4	262	1	402	5	542	5	682	3	822	4	962	4	3.71	0.15
123	5	263	3	403	4	543	3	683	4	823	5	963	5	4.14	0.66
135	4	275	3	415	2	555	4	695	5	835	5	975	5	4.00	0.45
Σ														**33.29**	**4.78**

[192] Fuente: Arana Ovalle, R. I. 2003

5.2.2.7. Muestro de Razón, Regresión y Diferencia (MRRD) ([193])

5.2.3. No probabilístico

5.2.3.1. Accidental o bola de nieve

Es adecuado utilizar una muestra de bola de nieve cuando los miembros de una población son difíciles de localizar, como personas sin hogar, trabajadores migrantes o inmigrantes indocumentados. Una muestra de bola de nieve es aquella en que el investigador recopila datos sobre los pocos miembros de la población objetivo que puede localizar y, después, les pide que le proporcionen la información necesaria para localizar a otros miembros que conozcan de esa población. Se fundamenta en reclutar casos hasta que se completa el número de sujetos necesario para completar el tamaño de muestra deseado. Estos, se eligen de manera casual tal que quienes realizan el estudio eligen un lugar, a partir del cual reclutan los sujetos a estudio de la población que accidentalmente se encuentren a su disposición. Es similar al muestreo por conveniencia, excepto que intenta incluir a todos los sujetos accesibles como parte de la muestra. Por ejemplo: si un investigador quiere entrevistar a inmigrantes indocumentados de México, podría entrevistar a algunos indocumentados que conozca o pueda localizar, y luego dependerá de esos sujetos para que lo ayuden a localizar a más individuos indocumentados. Este proceso continúa hasta que el investigador tenga todas las entrevistas que necesita o hasta que se hayan agotados todos los contactos. Esta técnica es útil cuando se estudia un tema sensible en el que la gente podría no hablar abiertamente, o si hablar sobre los temas investigados podría poner en peligro su seguridad. Una recomendación de un amigo o conocido de que el investigador es confiable funciona para aumentar el tamaño de la muestra. Otro ejemplo, sería entre todos los sujetos con CA, seleccionar los primeros 50 incluibles que lleguen al servicio de urgencias del Hospital Regional de Temuco. **Ventajas**: El proceso en cadena permite que el investigador llegue a poblaciones que son difíciles de probar cuando se utilizan otros métodos de muestreo, el proceso es barato, simple y rentable, así como también necesita poca planificación y menos mano de obra que otras técnicas de muestreo. **Desventajas**: El investigador tiene poco control sobre el método de muestreo, pues los sujetos que el investigador puede obtener se

[193] Pérez López, C. 2005

basan principalmente en sujetos observados anteriormente, la representatividad de la muestra no está garantizada y el investigador no tiene ni idea de la verdadera distribución de la población ni de la muestra. **Ejemplo**: Una fundación desea captar más donadores, emplea un muestreo, bola de nieve, a partir de la selección de uno de sus donadores cautivos, para solicitarle nombres y direcciones de varios conocidos que sean susceptibles de ser donadores ([194]).

5.2.3.2. Intencional o de conveniencia

Este tipo de muestreo se caracteriza por un esfuerzo deliberado de obtener muestras "representativas" mediante la inclusión en la muestra de grupos supuestamente típicos. 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 directa e intencionadamente los individuos de la población. El caso más frecuente de este procedimiento el utilizar como muestra los individuos a los que se tiene fácil acceso, como profesores de universidad emplean con mucha frecuencia a sus propios alumnos. Es la muestra que está disponible en el tiempo o periodo de investigación. También, permite seleccionar casos característicos de una población limitando la muestra sólo a estos casos. Se utiliza en escenarios en las que la población es muy variable y consiguientemente la muestra es muy pequeña. Ejemplo: Todos los pacientes que asistan a una clínica en particular cierto día, semana, pueden ser requeridos para participar o, también, entre todos los sujetos con CA, seleccionar a aquellos que más convengan al equipo investigador, para conducir la investigación. **Ventajas:** Menos costoso, no requiere mucho tiempo, fácil de administrar, por lo general asegura alta tasa de participación y es posible generalización a sujetos similares. **Desventajas:** Difícil generalizar a otros sujetos, menos representativa de una población específica, los resultados dependen de las características únicas de la muestra, mayor probabilidad de error debido al investigador o influencia de sujetos (sesgos) y la muestra puede ser poco representativa de la población que se desea estudiar. **Ejemplo**: Una empresa comercial realiza entrevistas de puerta en puerta, eligiendo casas donde no existan perros, cercanas a la empresa, en planta baja u hogares con personas amables ([195]).

[194] Aragón, S. L. G. 2016
[195] Aragón, S. L. G. 2016

5.2.3.3. Discrecional o por expertos

Denominado también como muestreo de juicio, es una técnica utilizada por expertos para seleccionar especímenes, unidades o porciones representativas o típicas, según el criterio del experto; es decir, a criterio del investigador los elementos son elegidos sobre lo que él cree que pueden aportar al estudio y aplica bien para estudios de pre-prueba o prueba piloto para un instrumento. La idea se centra en que el investigador elige la muestra por que los considera los más representativos. Son utilizadas también en estudios exploratorios y en investigaciones de tipo cualitativo. Por ejemplo: la selección de un conjunto de especímenes con determinadas características, para un experimento de laboratorio, o la selección de determinadas semanas del año para llevar a cabo algunas auditorías. Es importante hacer notar que en este caso los criterios de selección pueden variar de experto a experto al determinar cuáles son las unidades de muestreo representativas de la población. **Ventajas**: Este método es sumamente fácil de aplicar, no es costoso y depende del conocimiento que tiene el investigador. **Desventajas**: No es tan preciso, pues depende del juicio del investigador ([196]).

5.2.3.4. Por cuotas

Denominado en ocasiones "accidental". Se asienta generalmente sobre la base de un buen conocimiento de los estratos de la población y/o de los individuos más "representativos" o "adecuados" para los fines de la investigación. Mantiene, por tanto, semejanzas con el muestreo aleatorio estratificado, pero no tiene el carácter de aleatoriedad de aquel. En este tipo de muestreo se fijan unas "cuotas" que consisten en un número de individuos que reúnen unas determinadas condiciones. Por ejemplo: 20 individuos de 25 a 40 años, de sexo femenino y residentes en Gijón. Una vez determinada la cuota se eligen los primeros que se encuentren que cumplan esas características. Este método se utiliza mucho en las encuestas de opinión. **Ventajas**: Es semejante al muestreo aleatorio estratificado y se basa en los individuos más representativos de la población. **Desventajas**: En algunos casos esta técnica no es totalmente representativa de la población y se debe saber que se han tenido en cuenta solo los rasgos seleccionados de la población para formar los subgrupos.

[196] Aragón, S. L. G. 2016

Ejemplo: En la Ciudad de México se desea conocer la preferencia de compra de un producto para arreglo personal, luego se elegirán personas que pasen por varios puntos fijos establecidos en norte, sur, este y oeste de la metrópoli, descartando niños y adultos mayores, pues se desea encuestar sólo personas con poder adquisitivo hasta llegar a la cota de elementos deseados en la muestra ([197])

[197] Aragón, S. L. G. 2016

6. HIPÓTESIS

6.1. QUÉ SON LAS HIPÓTESIS

La esencia de una prueba de hipótesis es probar si alguna relación expresada entre variables existe; éstas son las sospechas que el investigador supone por anticipado del problema en estudio, para dar a los hechos la oportunidad de demostrar o negar algo. Hipótesis es una expresión a manera de conjetura; es decir, una proposición tentativa en modo afirmativo acerca de la relación general o específica entre dos o más variables. En la formulación de cualquier hipótesis es conveniente observar los criterios: deben expresar relaciones entre variables y además, ser inferencias que permitan probar las relaciones establecidas.

El objetivo de una prueba estadística de hipótesis es determinar si un supuesto sobre alguna o algunas características de la población, está ampliamente respaldado por la información obtenida a través de datos muestrales. "una hipótesis estadística, denotada H_0, es un enunciado sobre la población. Su posibilidad de ser evaluada es con base en la información obtenida de una muestra aleatoria de la población" ([198]).

A la hipótesis nula se le caracterizó como la afirmación hecha sobre la población, tal afirmación puede tener dos resultados posibles y complementarios al probar su validez ([199]):

- $Hipótesis\ \mathrm{H}_0\ \ es\ cierta$ ($hipótesis\ \mathrm{H}_0\ \ está\ respaldada\ por\ los\ datos\ de\ la\ muestra$)
- $Hipótesis\ \mathrm{H}_0\ es\ falsa$ ($hipótesis\ \mathrm{H}_0\ \ no\ está\ ampliamente\ respaldada\ por\ los\ datos\ de\ la\ muestra$)

Al proceso en el cual se selecciona una de estas dos acciones se le llama **prueba estadística de hipótesis** ([200]).

6.1.1. Características de una hipótesis ([201]):

1. No debe contener palabras ambiguas ni términos valorativos. Los términos (variables) de la hipótesis tienen que ser comprensibles, precisos, lo más concretos posibles,

[198] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014
[199] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014
[200] Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014
[201] Vargas, A. D. 2006

claros por medio de definiciones conceptuales y operacionales. Una hipótesis debe expresar al menos una relación entre dos variables, una independiente, exógena o explicativa y una dependiente, endógena o explicada.

2. La relación entre variables propuesta por una hipótesis debe ser clara, verosímil (lógica) y acorde con fenómenos conocidos.

3. Los términos generales o abstractos deben ser operacionales. Los términos abstractos que no tienen referente empírico no deben ser considerados. Los términos de la hipótesis y la relación planteada entre ellos tienen que observarse y medirse; es decir, tener referentes en la realidad y estar libres de cualquier sesgo.

4. Cuando sea posible, la hipótesis debe formularse en términos cuantitativos.

5. La forma sintáctica debe ser la de una proposición simple.

6. La hipótesis causal o estadística debe considerar sólo dos variables.

7. Deberá excluir tautologías, repetición de una palabra o su equivalente en una frase.

8. Deberá evitar el uso de disyunciones ([202]).

9. Deberá ser doblemente pertinente tanto en su referencia al fenómeno real de investigación como en apoyo teórico que la sostiene. Las hipótesis deben referirse a una situación social real, coincidir con hechos conocidos, no estar en conflicto con leyes o principios establecidos.

10. Deberá referirse a aspectos de realidad que no han sido investigados.

11. Una característica de la hipótesis científica es su falibilidad, que implica que una vez comprobada puede perfeccionarse a través del tiempo.

12. Las hipótesis deben estar relacionadas con técnicas disponibles para probarlas. Deben ser medibles; es decir, la evaluación de hipótesis depende de la existencia de métodos para probarlas.

13. Las hipótesis serán la transformación directa de preguntas de investigación.

14. Las hipótesis sustituyen a objetivos y preguntas de investigación para guiar el estudio.

[202] Wikipedia (es.wikipedia.org/wiki/Disyunción_lógica): disyunción lógica (∨ {\displaystyle \lor } {\displaystyle \lor }) (en específico, una disyunción inclusiva) entre dos proposiciones es un conector lógico, cuyo valor de la verdad resulta en falso sólo si ambas proposiciones son falsas, y en cierto de cualquier otra forma.1 Existen diferentes contextos donde se utiliza la disyunción lógica.

15. Debe tener un nivel de generalidad y especificidad.

16. Deben dar respuesta parcial a lo investigado.

17. Deben ser lógicamente consistentes.

18. Hipótesis no es equivalente a generalización, es comprobable y surge de un número grande de observaciones e, incluso, en algunos casos se establece como conclusión de análisis.

19. Otras características que deben cumplir las hipótesis son enunciado verificable, grado de generalidad, formular de manera categórica, atinencia o ser asertivo ([203]), compatibilidad, simplicidad, plausibilidad o admisibilidad, poder predictivo o explicativo.

20. Consta de dos partes: una base o cimiento y un cuerpo o estructura.

21. Capacidad de inferir y hacer predicciones verificables, sugerir nuevas experiencias y formular otras hipótesis.

22. Pueda someterse al método experimental, que es prueba práctica.

23. Serán redactadas en términos claros, sencillos, de manera específica y formularse como aseveración.

6.1.2. Tipos de hipótesis ([204])

En toda situación en la que se desee probar la validez de una afirmación, la hipótesis nula se suele basar en la suposición de que la afirmación sea verdadera. Entonces, la hipótesis alternativa se formula de manera que rechazar H0 proporcione la evidencia estadística de que la suposición establecida es incorrecta. Resumen: En las pruebas de hipótesis para la media poblacional, μ_0 denota el valor hipotético y para la prueba de hipótesis hay que escoger una de las formas:

- **Pruebas de una cola** (izquierda)**:**

$$H_0: \mu \geq \mu_0$$
$$H_a: \mu < \mu_0$$

Ejemplo: La Federal Trade Commission, FTC, realiza periódicamente estudios estadísticos con objeto de comprobar las afirmaciones de los fabricantes acerca de sus productos. Por

[203] Diccionario de la Real Academia Española (dle.rae.es/?id=3yQsnyJ): Dicho de una persona que expresa su opinión de manera firme

[204] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015

ejemplo, en la etiqueta de una lata grande de Hilltop Coffee dice que la lata contiene 3 libras de café. La FTC sabe que el proceso de producción de Hilltop no permite llenar las latas con 3 libras exactas de café por lata, incluso si la media poblacional del peso de llenado de todas las latas es de 3 libras por lata. Sin embargo, mientras la media poblacional del peso de llenado sea por lo menos 3 libras por lata, los derechos del consumidor estarán protegidos. Por tanto, la FTC interpreta que la información de la etiqueta de una lata grande de café Hilltop tiene una media poblacional del peso de llenado de por lo menos 3 libras por lata. Si μ denota la media poblacional del peso de llenado, las hipótesis nula y alternativa son las siguientes:

$$H_0: \mu \geq 3$$
$$H_a: \mu < 3$$

Interpretación: Si los datos muestrales indican que H_0 no se puede rechazar, las evidencias estadísticas no conducirán a concluir que ha habido una violación en lo que se afirma en la etiqueta. Pero, si los datos muestrales indican que se puede rechazar H_0 se concluirá que la hipótesis alternativa $H_a: \mu \geq 3$ es verdadera. En este caso la conclusión de que hay falta de peso y un cargo por violación a lo que se establece en la etiqueta estará justificada.

- **Pruebas de dos colas:**

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

Ejemplo: La U.S. Golf Association, USGA, establece reglas que deben satisfacer los fabricantes de equipos de golf si quieren que sus productos se acepten en los eventos de USGA. MaxFlight emplea procesos de fabricación de alta tecnología para producir pelotas de golf que tienen una distancia media de recorrido de 295 yardas. Sin embargo, algunas veces el proceso se desajusta y se producen pelotas de golf que tienen una distancia media de recorrido diferente a 295 yardas. Cuando la distancia media es menor que 295 yardas, a la empresa le preocupa perder clientes porque las pelotas de golf no proporcionen la distancia anunciada. Cuando la distancia es mayor que 295 yardas, las pelotas de MaxFlight pueden ser rechazadas por la USGA por exceder los estándares respecto de distancia de vuelo y carrera. El programa de control de calidad de MaxFlight consiste en tomar muestras periódicas de 50 pelotas de golf y vigilar el proceso de fabricación. Con cada muestra se realiza una prueba de hipótesis para determinar si el proceso se ha desajustado. Para

elaborar las hipótesis nula y alternativa, se empieza por suponer que el proceso está funcionando correctamente; es decir, las pelotas de golf que se están produciendo alcanzan una distancia media de 295 yardas. Ésta es la suposición que establece en la hipótesis nula. La hipótesis alternativa es que la distancia media no es 295 yardas. Como el valor hipotético es $\mu_0 = 295$, las hipótesis nula y alternativa en el caso de la prueba de hipótesis de MaxFlight son las siguientes:

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

- **Pruebas de una cola** (derecha):

$$H_0: \mu \leq \mu_0$$

$$H_a: \mu > \mu_0$$

Ejemplo: John Morrell & Company, que se inició en Inglaterra en 1827, es considerado el fabricante de productos de carne más antiguo de Estados Unidos que ha funcionado con continuidad. Las investigaciones de mercado de Morrell proporcionan a los directivos información actualizada acerca de los diversos productos de la empresa y sobre su posición en relación con las otras marcas de productos similares. En esta prueba de comparación de los tres productos, se empleó una muestra de consumidores para que indicaran cómo calificaban a los productos en términos de sabor, apariencia, aroma y preferencia. Una de las cuestiones que se deseaba investigar era si el producto de Morrell era la elección de preferencia de más de 50% de la población de consumidores. Si p representa la proporción poblacional que prefiere el producto de Morrell, la prueba de hipótesis para la cuestión que se investiga es la siguiente:

$$H_0: p \leq 0.5$$

$$H_a: \mu > 0.50$$

En un estudio independiente se hizo una prueba de degustación empleando una muestra de 224 consumidores de Cincinnati, Milwaukee y Los Ángeles, 150 consumidores eligieron el producto de Morrell como el de su preferencia. A partir del procedimiento estadístico de prueba de hipótesis, se rechazó la hipótesis nula. Mediante el estudio se encontraron evidencias estadísticas que favorecían a Ha y se llegó a la conclusión de que el

producto de Morrell es preferido por más de 50% de la población de consumidores. La estimación puntual de la proporción poblacional es $\bar{p} = \frac{150}{224} = 0.67$

Interpretación: los datos muestrales sirvieron para hacer publicidad en una revista de alimentos en la cual se mostraba que, en una comparación del sabor de los tres productos, el producto de Morrell era "preferido en una relación 2 a 1".

6.1.3. ¿Cuál es la utilidad de las Hipótesis? ([205])

a) Son las herramientas de trabajo de la teoría, esto es, de las teorías se pueden deducir hipótesis.

b) Estas se pueden demostrarse; es decir, se puede establecer que son probablemente ciertas o probablemente falsas.

c) Son un instrumento poderoso para el progreso del conocimiento, porque ayudan a confirmar o negar una teoría en forma independiente de la opinión del investigador.

Pasos previos a la prueba de hipótesis:

a) **Postular modelos**. El establecimiento de modelos es la parte en la que se proponen las distribuciones y las ecuaciones de tipo estadístico que relacionan a las variables que intervienen en la evaluación de los supuestos teóricos; entre los más conocidos se pueden mencionar el binomial, el normal, el de regresión, etcétera.

b) **Recolectar información**. La recolección de información se puede realizar a través de algún método de muestreo o el diseño de algún experimento, con el fin de cuantificar las variables estudiadas, si aún no se dispone de los datos necesarios en las fuentes de información acreditadas existentes.

c) **Registro presentación de datos**. El registro y la presentación de los datos se hace mediante tablas y gráficas.

d) **Resumen de datos**. El resumen de la información se lleva a cabo a través de indicadores que caracterizan el comportamiento general de las variables estudiadas, como, por ejemplo: las medidas de tendencia central, dispersión, asimetría, etcétera.

[205] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015

e) **Estimación y predicción**. En la parte de estimación y predicción se obtienen los estadísticos de los parámetros de los modelos propuestos, con el fin de estimar las tendencias generales del fenómeno en estudio.

f) **Pruebas de hipótesis**. En esta parte se verifica si los supuestos teóricos sobre el fenómeno estudiado (parámetros del modelo) contrastan con las observaciones (modelo estimado), a través de ciertos procedimientos estadísticos aceptados.

Recuerde que a menudo uno de los objetivos de la estadística es hacer inferencias acerca de parámetros poblacionales desconocidos con base en información contenida en datos muestrales. Estas inferencias se interpretan de dos formas: como estimaciones de los parámetros respectivos o como pruebas de hipótesis acerca de sus valores ([206]). Por lo tanto, las pruebas de hipótesis se llevan a cabo en todos los campos en los que la teoría se pueda probar contra observación. Todas estas hipótesis pueden ser tema de verificación estadística mediante el uso de datos muestrales observados. Por lo general, de tiene una teoría; es decir, una hipótesis de investigación respecto a parámetros que se desea apoyar

6.2. PRUEBAS DE HIPÓTESIS PARAMÉTRICAS

6.2.1. Elementos de una prueba

Cuando se hace una prueba de hipótesis se empieza por hacer una suposición tentativa acerca del parámetro poblacional. A esta suposición tentativa se le llama hipótesis nula. Después se define otra hipótesis, llamada hipótesis alternativa, que dice lo contrario de lo que establece la hipótesis nula ([207]). Estos elementos son ([208]):

1) **Hipótesis nula** (H_0)

2) **Hipótesis alternativa** (H_a)

3) **Estadístico de prueba** (al igual que un estimador es una función de las mediciones muestrales)

4) **Región de rechazo** (estará denotada por RR, especifica los valores del estadístico de prueba para el cual la hipótesis nula ha de ser rechazada a favor de la hipótesis alternativa. Si, para una muestra particular, el valor calculado del estadístico de prueba cae en la región

[206] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015

[207] Anderson, C. D; Sweeney, D.J; Williams, T. A; Camm, J. D. y Cochran, J. J. 2016

[208] Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015

de rechazo RR, rechazamos la hipótesis nula H_0 y aceptamos la hipótesis alternativa Ha. Si el valor del estadístico de prueba no cae en la RR, aceptamos H_0).

6.2.2. De hipótesis

6.2.2.1. Media con varianza conocida [209]

Es uno de los casos más comunes en realización de pruebas de hipótesis. El parámetro θ que se desea probar es μ y su estimador $\hat{\theta}$ es $\overline{X}$. De manera sucinta, se exponen las pruebas estadísticas, bilateral y unilaterales cuando se desea probar que el valor de la media poblacional μ es igual al valor prefijado μ_0.

a) Prueba Bilateral para Media.

1. $H_0: \mu = \mu_0$
2. $H_a: \mu \neq \mu_0$
3. Estadístico de prueba $Z_{Calculada} = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
4. Región de rechazo $Z_{Calculada} < -Z_{\alpha/2}$ o $Z_{Calculada} > Z_{\alpha/2}$

Ejemplo: Una empresa farmacéutica establece un comprimido que tendrá un peso medio $\mu_0 = 0.5\ Gr$ y una desviación estándar $\sigma = 0.11\ Gr$. Se toma una muestra de 144 comprimidos de un lote de fármacos, cuyo peso promedio es $\overline{X} = 0.53\ Gr$:

a) Para un nivel de significación $\alpha = 0.01$, ¿el peso de comprimidos en el lote se diferencia del admisible por esta empresa?

La prueba es bilateral, $H_0: \mu = 0.5\ Gr$, H_a o $H_1: \mu \neq 0.5\ Gr$, estadístico de prueba $Z_{Calculada} = \frac{0.53 - 0.50}{\frac{0.11}{\sqrt{144}}} = 3.273$ y región de rechazo, con $\alpha = 0.01$, $Z_{Calculada\ (3.273)} < -Z_{\alpha/2\ (-2.570)}$ o $Z_{Calculada\ (3.273)} > Z_{\alpha/2(2.570)}$. En otras palabras, $|Z_{Calculada\ (3.273)}| > |Z_{\alpha/2\ (2.570)}| \Longrightarrow$ No se acepta H_0 y no se rechaza H_a o H_1 o, también, el valor de $Z_{Calculada\ (3.273)}$ cae en la región de rechazo, se descarta $H_0: \mu = 0.5\ Gr$ tal que la probabilidad de rechazar $H_0: \mu = 0.5\ Gr$, suponiendo que sea cierta es sólo $\alpha = 0.01$; por lo tanto, en al menos 99% de las ocasiones, la decisión es correcta.

b) Determine el $\rho - Value$ de la prueba

[209] Si el tamaño de la muestra es suficientemente grande ($n \geq 30$), se desconoce su varianza, es posible usar estas pruebas reemplazando σ por su estimador s sin pérdida de exactitud.

Si $Z_{\text{Calculada (3.273)}}$, su valor de probabilidad es $\Phi_{\text{(Letra girega mayúscula Phi)}}(3.273) = 0.9995 \Rightarrow 0.9995 + \frac{\alpha}{2} = 1$; por lo tanto, $\alpha = 2(1 - 0.9995) = 0.001$ e indica que se puede cometer un error en mil veces, que complementa y verifica la anterior decisión.

b) Prueba Unilateral para Media.

5. $H_0: \mu = \mu_0$

6. $H_a: \mu > \mu_0$ o $H_1: \mu < \mu_0$

7. Estadístico de prueba $Z_{\text{Calculada}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

8. Región de rechazo $Z_{\text{Calculada}} < -Z_{\alpha}$ $(H_1: \mu < \mu_0)$ o $Z_{\text{Calculada}} > Z_{\alpha}$

Ejemplo: Una empresa farmacéutica establece un comprimido que tendrá un peso medio $\mu_0 = 0.5\ \text{Gr}$ y una desviación estándar $\sigma = 0.11\ \text{Gr}$. Se toma una muestra de 144 comprimidos de un lote de fármacos, cuyo peso promedio es $\bar{X} = 0.53\ \text{Gr}$. Además, el peso máximo admisible para que el medicamento no sea tóxico $\mu_0 = 0.52\ \text{Gr}$:

a) Se desea saber si los comprimidos del lote son aptos para el consumo humano, a un nivel de significación $\alpha = 0.05$.

9. $H_0: \mu = 0.52\ \text{Gr}$

10. $H_a: \mu > 0.52\ \text{Gr}$

11. Estadístico de prueba $Z_{\text{Calculada}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{0.53 - 0.52}{\frac{0.11}{\sqrt{144}}} = 1.091$

12. Región de rechazo $Z_{\text{Calculada}(1.091)} < Z_{\alpha\ (1.65)}$

En otras palabras, $\left|Z_{\text{Calculada (1.091)}}\right| < \left|Z_{\alpha\ (1.65)}\right| \Longrightarrow$ No se rechaza H_0 y no se acepta H_a o H_1 o, también, el valor de $Z_{\text{Calculada (1.091)}}$ no cae en la región de rechazo, se descarta $H_a: \mu > 0.52\ \text{Gr}$ tal que no existe razón para descartar H_0, por lo que se podría asegurar que el medicamento es apto para consumo humano.

b) Determine el nivel de significación de prueba.

El nivel de significación de esta prueba es $1 - \Phi_{\text{(Letra girega mayúscula Phi)}}(1.091) = 1 - 0.862 = 0.138 \Rightarrow \rho - \text{Value}_{(0.138)} > \alpha_{(0.05)}$ no se rechaza H_0 ([210]).

[210] **Potencia de prueba**: La probabilidad de cometer error tipo II, denotado por β y definido como $\beta = \Pr(\text{error tipo II}) = \Pr(\text{Aceptar } H_0 | \mu_1) \Rightarrow H_1: \mu = \mu_1 \Rightarrow$ por Teorema del Límite Central, $Z_{\text{Calculada}} = \frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}}$ sigue una ley normal estándar. En este ejemplo, si calcula la potencia de la prueba si el verdadero valor es $\mu = 0.54\ \text{Gr}$.

6.2.2.2. Media con varianza desconocida

En este caso, no es posible aplicar el Teorema del Límite Central. Para que sea viable aplicar la prueba, es necesario que la muestra provenga de una población que sigue una ley normal, tal que el estadístico de prueba sigue una distribución t:

a) Prueba Bilateral para Media

13. $H_0: \mu = \mu_0$
14. $H_a: \mu \neq \mu_0$
15. Estadístico de prueba $t_{Calculada} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$
16. Región de rechazo $t_{Calculada} < -t_{(\alpha/2;\, n-1)}$ o $t_{Calculada} > t_{(\alpha/2;\, n-1)}$

Ejemplo: Según el Ministerio de Educación de Ecuador, el costo medio de lista de útiles escolares de educación básica es de 87.00 USD. Para verificarlo, un investigador tomo una muestra:

Costo (X_i)	68	75	93	101	123
Número (n_i)	2	3	4	6	5

Para un nivel de significación $\alpha = 0.05$, verifique la hipótesis que la máquina cumple con la especificación. Considere $\bar{X} = 97.7$ y $s = 18.728$.

1. $H_0: \mu = 87$
2. $H_a: \mu \neq 87$
3. Estadístico de prueba $t_{Calculada} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{97.7 - 87}{\frac{18.728}{\sqrt{20}}} = 2.555$
4. Región de rechazo $t_{Calculada\ (2.555)} > t_{(\alpha=0.05;\, n-1=19)=2.093}$

La región de rechazo es $t_{Calculada\ (2.555)} < -t_{(\alpha/2;\, n-1)=-2.093}$ o $t_{Calculada\ (2.555)} > t_{(\alpha/2;\, n-1)=2.093}$. En otras palabras, $|t_{Calculada\ (2.555)}| > |t_{tablas\ (2.093)}| \Longrightarrow$ No se acepta H_0 y no se rechaza H_a o H_1 o, también, el valor de $t_{Calculada\ (2.555)}$ se encuentra en región critica, pues $2.555 > 2.093 \Rightarrow$ no se acepta H_0; es decir, el precio medio de listas de útiles escolares es diferente al afirmado por el Ministerio de Educación de Ecuador.

Entonces, "no se rechaza H_0" si $Z_{Calculada}(\frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}}) \leq Z_{tablas\ (1.65)}$; es decir, $\frac{\bar{X} - 0.52}{\frac{0.11}{\sqrt{144}}} \leq 1.65 \Rightarrow \bar{X} \leq 0.535 \Rightarrow \Pr(\beta)$ o $\Pr(\text{Aceptar } H_0|\mu_1) = \Pr(\bar{X} \leq 0.535|\mu_1 = 0.54) = \Pr\left(\frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} \leq \frac{0.535 - 0.54}{\frac{0.11}{\sqrt{144}}}\right) = \Pr(Z \leq -0.531) = 0.298 \Rightarrow$ la potencia de la prueba es $\mathbf{Pot} = \mathbf{1} - \boldsymbol{\beta} = 0.702$.

b) Prueba Unilateral para Media

17. $H_0: \mu = \mu_0$

18. $H_a: \mu > \mu_0$ o $H_1: \mu < \mu_0$

19. Estadístico de prueba $t_{Calculada} = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$

20. Región de rechazo $t_{Calculada} < -t_{(\alpha;\, n-1)}$ $(H_1: \mu < \mu_0)$ o $t_{Calculada} > t_{(\alpha;\, n-1)}$

Ejemplo: Según previsiones del gobierno, la inflación para este año será 3.9%. Un profesional desconfiado, realizo una investigación por su cuenta y registro la variación de precios en 22 artículos que a su juicio tienen la mayor incidencia en la economía popular. Obtuvo una variación de 4.5% y una desviación estándar de 1.3%. Pruebe si la cifra de inflación del investigador será mayor que la del gobierno.

Se tiene que $n = 22$, $\overline{X} = 4.5$ y $s = 1.3$

1. $H_0: \mu = 3.9$

2. $H_a: \mu > 3.9$

3. Estadístico de prueba $t_{Calculada} = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{4.5 - 3.9}{\frac{1.3}{\sqrt{22}}} = 2.165$

4. Región de rechazo $t_{Calculada\ (2.165)} > t_{(\alpha=0.01;\, n-1=21)=2.831}$

La región de rechazo es $t_{Calculada\ (2.555)} < -t_{(\alpha/2;\, n-1)=-2.093}$ o $t_{Calculada\ (2.555)} > t_{(\alpha/2;\, n-1)=2.093}$.

6.2.2.3. Varianza

Para realizar una prueba de hipótesis sobre éste tipo, supone que las observaciones provienen de una distribución normal, para que el estadístico $\frac{(n-1)s^2}{\sigma^2}$ siga una distribución χ^2 con $(n-1)$ grados de libertad. Bajo éste supuesto, las pruebas de hipótesis son:

a) Prueba Bilateral para Varianza

21. $H_0: \sigma^2 = \sigma_0^2$

22. $H_a: \sigma^2 \neq \sigma_0^2$

23. Estadístico de prueba $\chi^2_{Calculada} = \frac{(n-1)s^2}{\sigma_0^2}$

24. Región de rechazo $\chi^2_{Calculada} < -\chi^2_{Tablas\ (1-\frac{\alpha}{2};\, n-1)}$ o $\chi^2_{Calculada} > \chi^2_{Tablas\ (1-\frac{\alpha}{2};\, n-1)}$

b) Prueba Unilateral para Varianza

25. $H_0: \sigma^2 = \sigma_0^2$

26. $H_1: \sigma^2 > \sigma_0^2$ o $H_1: \sigma^2 < \sigma_0^2$

27. Estadístico de prueba $\chi^2_{\text{Calculada}} = \frac{(n-1)s^2}{\sigma_0^2}$

28. Región de rechazo $\chi^2_{\text{Calculada}} < -\chi^2_{\text{Tablas }(1-\alpha;\, n-1)}$ $(H_1: \sigma^2 < \sigma_0^2)$o $\chi^2_{\text{Calculada}} > \chi^2_{\text{Tablas }(\alpha;\, n-1)}$

Ejemplo: Un fabricante de cables de cobre afirmó que su producto tiene una resistencia a la ruptura relativamente estable y se ubicaría en un rango de 40 Kg-Fuerza (KgF). Una muestra de 16 mediciones arroja una varianza de $s^2 = 195$.

a) ¿Hay evidencia suficiente para no aceptar la afirmación del fabricante?

El fabricante da el rango de variación de resistencia, se puede estimar la desviación estándar mediante la relación aproximada $\sigma \approx \frac{\text{Rango}}{4} \Rightarrow \sigma = 10\ \text{KgF}$. Con base en esto, la prueba es:

1. $H_0: \sigma^2 = \sigma_0^2 = (10)^2 = 100$
2. $H_1: \sigma^2 > 100$
3. Estadístico de prueba $\chi^2_{\text{Calculada}} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(16-1)*195}{100} = 29.25$
4. Región de rechazo. Para nivel de significación $\alpha = 0.05$, 15 grados de libertad, $\chi^2_{\text{Tablas} = 0.05}(15) = 25.00 \Rightarrow \left|\chi^2_{\text{Calculada }(29.25)}\right| > \left|\chi^2_{\text{Tablas }(25.00)}\right|$ No se acepta H_0 y no se rechaza H_1. Por lo tanto, la variación de mediciones excede las especificaciones del fabricante

b) Encuentre el $\rho - \text{Value}$ de la prueba.

Para hallar el nivel de significación aproximado de la prueba se examina, en tabla de distribución χ^2_{Tablas}, en renglón correspondiente a 15 grados de libertad. Se observa que el valor $\chi^2_{\text{Tablas }(\alpha\, =\, 0.025\, \Rightarrow\, 27.49)} < \chi^2_{\text{Calculada }(29.25)} < \chi^2_{\text{Tablas }(\alpha\, =\, 0.01\, \Rightarrow\, 30.58)}$. Por lo tanto, el nivel de significación observado es menor que 0.025 e implica que no se acepta H_0 para $\forall\ \alpha \geq 2.5\ \%$.

6.2.2.4. Proporción

Suponga que dispone de n observaciones provenientes de una población con distribución de Bernoulli y desea que el parámetro ρ es igual a un valor prefijado ρ_0. Recuerde que entre n observaciones hay y éxitos, la proporción se estima con $\hat{\rho} = \frac{y}{n}$. La realización de estas pruebas usará la aproximación de ley binomial mediante ley normal.

a) Prueba Bilateral para Varianza

29. $H_0: \rho = \rho_0$

30. $H_1: \rho \neq \rho_0$

31. Estadístico de prueba $Z_{Calculada} = \frac{\hat{\rho} - \rho_0}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho} - \rho_0}{\sqrt{\frac{\rho_0 \, q_0}{n}}} \Rightarrow \hat{\rho} = \frac{y}{n}$.

32. Región de rechazo $Z_{Calculada} < -Z_{\alpha/2}$ o $Z_{Calculada} > Z_{\alpha/2}$

b) Prueba Unilateral para Varianza

33. $H_0: \rho = \rho_0$

34. $H_1: \rho > \rho_0$ o $H_1: \rho < \rho_0$

35. Estadístico de prueba $Z_{Calculada} = \frac{\hat{\rho} - \rho_0}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho} - \rho_0}{\sqrt{\frac{\rho_0 \, q_0}{n}}} \Rightarrow \hat{\rho} = \frac{y}{n}$.

36. Región de rechazo $Z_{Calculada} < -Z_{\alpha}$, cuando $H_1: \rho < \rho_0$, o $Z_{Calculada} > Z_{\alpha}$

Ejemplo: Una empresa realizó una investigación de mercado para determinar el nivel de consumo de un refresco –cola-, consultando a 200 consumidores, en que 28 expresaron su preferencia por el producto. El fabricante, según sus ventas, estima que tiene el 10% del mercado de refrescos o colas.

a) ¿Son estos resultados de investigación consistentes con datos del fabricante?

Se tiene que $\rho_0 = 0.10$ y su estimador de proporción es $\hat{\rho} = \frac{y}{n} = \frac{28}{200} = 0.14$ tal que la prueba queda:

1. $H_0: \rho = 0.1$

2. $H_1: \rho \neq 0.1$

3. Estadístico de prueba $Z_{Calculada} = \frac{\hat{\rho} - \rho_0}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho} - \rho_0}{\sqrt{\frac{\rho_0 \, q_0}{n}}} = \frac{0.14 - 0.10}{\sqrt{\frac{0.10 * 0.90}{200}}} = 1.886.$

4. Región de rechazo. Al escoger $\alpha = 0.05$, $Z_{Tablas: \frac{0.05}{2} = 0.025} = 1.96 \Rightarrow Z_{Calculada\ (1.886)} < Z_{Tablas:1.96}$. $Z_{Calculada}$ no cae en la zona de rechazo o, igualmente, $|Z_{Calculada\ (1.886)}| < |Z_{Tablas:1.96}|$ no se rechaza H_0 y no se acepta H_a. Por lo tanto, no hay evidencia que la proporción de consumidores sea diferente del 10%.

b) Determine el nivel de significación de contraste

Sea $Z_{\text{Calculada }(1.886)}$, su valor de probabilidad es $\Phi(1.886) = 0.9706$. Al ser una prueba bilateral, cumple que $0.9706 + \frac{\alpha}{2} = 1 \Rightarrow \alpha = 0.0588$.

Cálculo de potencia de prueba

La probabilidad de cometer un error tipo II se denota como β y es definida como $\beta = P_r(\text{error tipo II}) = P_r(\text{Aceptar } H_0|\rho_1)$ tal que supone verdadera la hipótesis $H_1: \rho = \rho_1$. Entonces, variable $Z_{\text{Calculada}} = \frac{\hat{\rho} - \rho_1}{\sqrt{\frac{\rho_1(1-\rho_1)}{n}}}$ sigue una ley normal estándar.

Ejemplo: Con base en ejemplo anterior, calcule la potencia de prueba si el verdadero valor de la proporción es 0.12.

No se rechaza H_0 si $\left|\frac{\hat{\rho} - \rho_0}{\sqrt{\frac{\rho_0\, q_0}{n}}}\right| \leq 1.96$; es decir, $\left|\frac{\hat{\rho} - 0.10}{\sqrt{\frac{0.10*0.90}{200}}}\right| \leq 1.96 \Rightarrow 0.05842 \leq \hat{\rho} \leq 0.14159$ ∴, la probabilidad β se puede poner según:

$$P_r(\text{Aceptar } H_0|\rho_1) = P_r(0.05842 \leq \hat{\rho} \leq 0.14159|\rho_1 = 0.12)$$

$$= \Phi\left(\frac{0.14159 - 0.12}{\sqrt{\frac{0.12 * 0.88}{200}}}\right) - \Phi\left(\frac{0.05842 - 0.12}{\sqrt{\frac{0.12 * 0.88}{200}}}\right)$$

$$= \Phi(0.93958) - \Phi(2.6799) = 0.8226$$

Su potencia de prueba es $\mathbf{Pot = 1 - \beta} = 0.1774$. Un resumen de pruebas de hipótesis con muestra única es:

Cuadro 24. Cálculo de potencia de prueba por distribución ([211])

Concepto	H_0	H_a o H_1	Estadístico de prueba	Región de rechazo
Distribución general	$\mu = \mu_0$	$\mu \neq \mu_0$	$Z_{Calculada} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	$\lvert Z_{Calculada}\rvert > \lvert Z_{Tablas(\frac{\alpha}{2})}\rvert$
Varianza conocida (muestra grande)	$\mu \leq \mu_0$	$\mu > \mu_0$		$\lvert Z_{Calculada}\rvert > \lvert Z_{Tablas(\alpha)}\rvert$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$\lvert Z_{Calculada}\rvert < -\lvert Z_{Tablas(\alpha)}\rvert$
Distribución general	$\mu = \mu_0$	$\mu \neq \mu_0$	$t_{Calculada} = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$ $(n - 1)$ grados de libertad	$\lvert t_{Calculada}\rvert > \lvert t_{Tablas(\frac{\alpha}{2})}\rvert$
Varianza desconocida	$\mu \leq \mu_0$	$\mu > \mu_0$		$\lvert t_{Calculada}\rvert > \lvert t_{Tablas(\alpha)}\rvert$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$\lvert t_{Calculada}\rvert < -\lvert t_{Tablas(\alpha)}\rvert$
Distribución normal	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2_{Calculada} = \frac{(n-1)s^2}{\sigma_0^2}$ $(n - 1)$ grados de libertad	$\lvert \chi^2_{Calculada}\rvert < \lvert \chi^2_{Tablas\ (1-\frac{\alpha}{2})}\rvert$ $\lvert \chi^2_{Calculada}\rvert > \lvert \chi^2_{Tablas\ (\frac{\alpha}{2})}\rvert$
	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$\lvert \chi^2_{Calculada}\rvert < \lvert \chi^2_{Tablas\ (\alpha)}\rvert$
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$\lvert \chi^2_{Calculada}\rvert < \lvert \chi^2_{Tablas\ (1-\alpha)}\rvert$
Distribución	$\rho = \rho_0$	$\rho \neq \rho_0$	$Z_{Calculada} = \frac{\hat{\rho} - \rho_0}{\hat{\sigma}_{\hat{\rho}}} = \frac{\hat{\rho} - \rho_0}{\sqrt{\frac{\rho_0\, q_0}{n}}}$	$\lvert Z_{Calculada}\rvert > \lvert Z_{Tablas(\frac{\alpha}{2})}\rvert$
Binomial (muestra grande)	$\rho \leq \rho_0$	$\rho > \rho_0$		$\lvert Z_{Calculada}\rvert > \lvert Z_{Tablas(\alpha)}\rvert$
	$\rho \geq \rho_0$	$\rho < \rho_0$		$\lvert Z_{Calculada}\rvert < -\lvert Z_{Tablas(\alpha)}\rvert$

6.2.3. Diferencia entre dos medias

Al igual que en pruebas de hipótesis sobre la media, considerar caos en que es posible aplicar el Teorema del Límite Central es importante y, también, aquellos en que no es posible.

[211] Fuente: Galindo, E. 2006

Además, un tercer caso se tomará en cuenta, cuando las muestras provienen de una misma unidad muestral, mediante mediciones repetidas.

6.2.3.1. Varianzas supuestas conocidas

Suponga que dispone de dos poblaciones, nombradas P_1 y P_2, quiere probarse si la diferencia entredós medias poblacionales es igual a una cantidad D_0 o, en otras palabras, $H_0: \mu_1 - \mu_2 = D_0$, así como si desea probar un caso particular de igualdad de sus medias, $H_0: \mu_1 = \mu_2$. Se extrae de P_1 una muestra de tamaño n_1 y de P_2 n_1. Si sus varianzas poblacionales correspondientes son conocidas, las pruebas de hipótesis son:

a) Prueba Bilateral para Diferencia de dos Medias

37. $H_0: (\mu_1 - \mu_2) = D_0$
38. $H_1: (\mu_1 - \mu_2) \neq D_0$
39. Estadístico de prueba $Z_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.
40. Región de rechazo $Z_{Calculada} > Z_{\left(\frac{\alpha}{2}\right)}$ o $Z_{Calculada} < -Z_{\left(\frac{\alpha}{2}\right)}$

b) Prueba Unilateral para Diferencia de dos Medias

41. $H_0: (\mu_1 - \mu_2) = D_0$
42. $H_1: (\mu_1 - \mu_2) > D_0$ o $H_1: (\mu_1 - \mu_2) < D_0$
43. Estadístico de prueba $Z_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.
44. Región de rechazo $Z_{Calculada} > Z_{(\alpha)}$ o $Z_{Calculada} < -Z_{(\alpha)}$ $(H_1: (\mu_1 - \mu_2) < D_0)$

Ejemplo: Un inversionista tiene dos hoteles en la ciudad, uno al norte y el otro al sur. Sospecha que el consumo medio en restaurante del norte es menor en comparación del sur. Del primer local obtiene una muestra de 30.00 facturas y, con ello, un consumo medio de 59.00 USD. Del segundo local toma una muestra de 50.00 facturas, con un consumo de 63.00 USD. Las varianzas de consumos en ambos locales son conocidas e iguales a 60.00 y 80.00 USD, respectivamente.

a) Con un nivel de significación $\alpha = 0.05$, verifique lo escrito.

Local norte $\overline{X}_1 = 59.00$, $\sigma_1^2 = 60.00$ y $n_1 = 30.00$, mientras que local sur $\overline{X}_1 = 63.00$, $\sigma_1^2 = 80.00$ y $n_1 = 50.00$. Con base en la información anterior, es una prueba unilateral con $D_0 = 0$:

1. $H_0: \mu_1 = \mu_1$
2. $H_1: \mu_1 < \mu_2$
3. Estadístico de prueba $Z_{Calculada} = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(59-63) - 0}{\sqrt{\frac{60}{30} + \frac{80}{50}}} = -2.11.$
4. Región de rechazo. Al escoger $\alpha = 0.05$, $Z_{Tablas:\, 0.05} = 1.645 \Rightarrow |Z_{Calculada\ (211)}| > |Z_{Tablas:1.645}|$. Cae en la zona de rechazo o, igualmente, $|Z_{Calculada\ (211)}| > |Z_{Tablas:1.645}|$ no se acepta H_0 y no se rechaza H_a. Por lo tanto, el consumo en local del sur es mayor que el consumo en local del norte.

b) Determine significación de la prueba ([212]).

Como $Z_{Calculada} = \frac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(59-63) - 0}{\sqrt{\frac{60}{30} + \frac{80}{50}}} = -2.11 \Rightarrow \Phi(-2.11) = 0.0174;$ es decir, $\rho - Value = 0.0174$ y, en consecuencia, se tiene una fuerte evidencia que H_0 no es verdadera o, en otras palabras, se tiene una fuerte evidencia que H_1 es verdadera.

6.2.3.2. Varianzas desconocidas

6.2.3.2.1. Supuestas iguales

Se dispone de dos poblaciones y desea probar si la diferencia entre sus medias poblacionales, correspondientes son igual a D_0 o, en otras palabras, $H_0: (\mu_1 - \mu_2) = D_0$. Se dispone de dos muestras de tamaños n_1 y n_2. Si sus varianzas poblacionales son desconocidas, supuestas iguales. Entonces, se calcula el estimador ponderado σ^2 mediante:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} = \frac{\sum_{i=1}^{n_1}(X_{1i} - \overline{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \overline{X}_1)^2}{(n_1 + n_2 - 2)}$$ [213]

a) Prueba Bilateral para Diferencia de dos Medias

45. $H_0: (\mu_1 - \mu_2) = D_0$
46. $H_1: (\mu_1 - \mu_2) \neq D_0$

[212] Si el tamaño de muestras es suficientemente grande ($n_i \geq 30$) y se desconoce sus varianzas, se usa pruebas sustituyendo σ_1^2 y σ_1^2 por sus estimadores s_1^2 y s_1^2, respectivamente.
[213] s_1^2 y s_1^2 son varianzas muestrales 1 y 2, respectivamente.

47. Estadístico de prueba $t_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

48. Región de rechazo $t_{Calculada} > t_{\left(\frac{\alpha}{2};\, n_1+n_2-2\right)}$ o $t_{Calculada} < -t_{\left(\frac{\alpha}{2};\, n_1+n_2-2\right)}$

b) Prueba Unilateral para Diferencia de dos Medias

49. $H_0: (\mu_1 - \mu_2) = D_0$

50. $H_1: (\mu_1 - \mu_2) > D_0$ o $H_1: (\mu_1 - \mu_2) < D_0$

51. Estadístico de prueba $t_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

52. Región de rechazo $t_{Calculada} > t_{(\alpha;\, n_1+n_2-2)}$ o $t_{Calculada} < -t_{(\alpha;\, n_1+n_2-2)}$, cuando $H_1: (\mu_1 - \mu_2) < D_0$ [214].

Ejemplo: Un inversionista no sabe si invertir en bonos emitidos por un país A o un B. Para tomar una decisión, seleccionó dos muestras correspondientes a rendimientos de bonos emitidos por ambos países, obteniendo:

Cuadro 25. Bonos emitidos por países

Concepto	País								
	A						B		
Rendimiento (%)	x_i	12.3	12.5	12.8	13.0	13.5	y_i	12.2	12.3	13.0
Frecuencia	n_i	1	2	4	2	1	m_i	6	8	2

Con un nivel de significación $\alpha = 0.01$ verifique si los tiempos el rendimiento de bonos de ambos países es igual, asumiendo que éstos siguen una distribución normal $N(0,1)$ y homocedasticidad.

Cuadro 26. Tiempos de rendimiento de bonos de países

País	Medida		
	Tendencia Central	Dispersión	Tamaño muestral
A	$\bar{x} = 12.80$	$s_x^2 = 0.11$	$n_x = 10$
B	$\bar{y} = 12.35$	$s_y^2 = 0.07$	$n_y = 16$

$$s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)} = \frac{(10-1)0.11 + (16-1)0.07}{(10+16-2)} = \frac{1}{12}$$

Con base en esto, la prueba es bilateral o de dos colas:

[214] El supuesto de igualdad entre varianzas poblacionales es comprobado mediante la "prueba de hipótesis para razón entre dos varianzas".

1. $H_0: \mu_x = \mu_y$

2. $H_1: \mu_x \neq \mu_y$

3. Estadístico de prueba $t_{Calculada} = \frac{(\bar{x}-\bar{y})}{s\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} = \frac{(12.80-12.35)}{\sqrt{\frac{1}{12}}\sqrt{\frac{1}{10}+\frac{1}{16}}} = 3.870.$

4. Región de rechazo $t_{Calculada\,=\,3.870} > t_{\left(\frac{0.01}{2}=0.005;\ 10+16-2=24\right)\,=\,2.797}$.

$\Rightarrow |t_{Calculada\,=\,3.870}| > |t_{Tablas:2.797}|$. Cae en la zona de rechazo o, igualmente, $|t_{Calculada\,=\,3.870}| > |t_{Tablas:2.797}|$ no se acepta H_0 y no se rechaza H_a. Por lo tanto, los bonos de ambos países tienen rendimientos diferentes.

6.2.3.2.2. Supuestas distintas

Suponga que dispone de dos poblaciones y desea probar si la diferencia entre sus correspondientes medias poblacionales es $= D_0$ o, es decir, $H_0: (\mu_1 - \mu_2) = D_0$ tal que se admite que las poblaciones son normales, sus varianzas desconocidas y diferentes.

a) Prueba Bilateral para Diferencia de dos Medias

53. $H_0: (\mu_1 - \mu_2) = D_0$

54. $H_1: (\mu_1 - \mu_2) \neq D_0$

55. Estadístico de prueba $t_{Calculada} = \frac{(\bar{X}_1-\bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$.

56. Región de rechazo $t_{Calculada} > t\left(\frac{\alpha}{2};\ g=\left[\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right]\right)$[215].

b) Prueba Unilateral para Diferencia de dos Medias

57. $H_0: (\mu_1 - \mu_2) = D_0$

58. $H_1: (\mu_1 - \mu_2) > D_0$ o $H_1: (\mu_1 - \mu_2) < D_0$

59. Estadístico de prueba $t_{Calculada} = \frac{(\bar{X}_1-\bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$.

[215] $g \notin \mathbb{N}$, se redondea al número más cercano.

60. Región de rechazo $t_{Calculada} > t_{\left(\alpha;\, g=\left[\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right]\right)}$ o $t_{Calculada} > -t_{\left(\alpha;\, g=\left[\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right]\right)}$, cuando $H_1: (\mu_1 - \mu_2) < D_0$.

Ejemplo: Se desea conocer el efecto del frío extremo sobre la realización de operaciones manuales. Se eligió al azar 20 voluntarios, divididos en dos grupos de 10 personas. El primer grupo fue expuesto a una temperatura de 4 0_C, mientras que el segundo se mantuvo a temperatura ambiente. Se contó el número de veces que los voluntarios podrían abrir y cerrar la mano en un periodo de 15 segundos:

Cuadro 27. Efecto del frío extremo en operaciones manuales de dos grupos

Equipos expuestos	Número de veces en que pueden abrir-cerrar la mano en 15 segundos									
Temperatura ambiente	54	51	40	45	48	46	45	52	49	50
Frío	32	29	38	33	34	35	36	35	29	23

La hipótesis a probar es que al estar expuesto un grupo de personas reduce su capacidad de abrir y cerrar la mano (H_0) en más de 12 veces (D_0). Con base en un nivel de significación de $\alpha = 0.05$ haga sus estimaciones.

Cuadro 28. Efecto del frío extremo en operaciones manuales de dos grupos con diferentes medidas

Equipos expuestos	Medida		
	Tendencia Central	Dispersión	Tamaño muestral
Temperatura ambiente	$\bar{x} = 48.0$	$s_x^2 = 16.89$	$n_x = 10$
Frío	$\bar{y} = 32.4$	$s_y^2 = 19.16$	$n_y = 10$

1. $H_0: (\mu_1 - \mu_2) = 12$
2. $H_1: (\mu_1 - \mu_2) > 12$
3. Estadístico de prueba $t_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{(48.0-32.4) - 12}{\sqrt{\frac{16.89}{10}+\frac{19.16}{10}}} = 1.896.$

4. Región de rechazo $t_{Calculada\ (1.896)} > t\left(\alpha=0.05;\ g=\left[\frac{\left(\frac{16.89}{10}+\frac{19.16}{10}\right)^2}{\frac{\left(\frac{16.89}{10}\right)^2}{10-1}+\frac{\left(\frac{19.16}{10}\right)^2}{10-1}}\right]=17.9\approx 18\right)=1.734$

$\Rightarrow |t_{Calculada\ =\ 1.896}| > |t_{Tablas:\ 1.734}|$. Cae en la zona de rechazo o, igualmente, $|t_{Calculada\ =\ 1.896}| > |t_{Tablas:\ 1.734}|$ no se acepta H_0 y no se rechaza H_a. Por lo tanto, el frío reduce la capacidad de realizar operaciones manuales.

6.2.3.3. Varianzas conocida y desconocida

Suponga que dispone de dos poblaciones y desea probar si la diferencia entre sus correspondientes medias poblacionales es $= D_0$ o, en otras palabras, $H_0: (\mu_1 - \mu_2) = D_0$. Considere que las poblaciones son normales $N(0, 1)$, σ_1^2 es conocida y σ_2^2 no lo es.

a) Prueba Bilateral para Diferencia de dos Medias

61. $H_0: (\mu_1 - \mu_2) = D_0$
62. $H_1: (\mu_1 - \mu_2) \neq D_0$
63. Estadístico de prueba $t_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{s_2^2}{n_2}}}$.
64. Región de rechazo $t_{Calculada} > t\left(\frac{\alpha}{2};\ g=\left[\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right]\right)$[216].

b) Prueba Unilateral para Diferencia de dos Medias

65. $H_0: (\mu_1 - \mu_2) = D_0$
66. $H_1: (\mu_1 - \mu_2) > D_0$ o, también, $H_1: (\mu_1 - \mu_2) < D_0$
67. Estadístico de prueba $t_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{s_2^2}{n_2}}}$.

[216] $g \notin \mathbb{N}$, se redondea al número más cercano.

68. Región de rechazo $t_{Calculada} > t_{\left(\alpha;\, g=\left[\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right]\right)}$ o, sino, $t_{Calculada} > -t_{\left(\alpha;\, g=\left[\frac{\left(\frac{\sigma_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}\right]\right)}$, cuando $H_1: (\mu_1 - \mu_2) < D_0$.

6.2.3.4. Diferencia por parejas

Las pruebas de diferencia de medias anteriores se aplican cuando dos muestras son independientes, pero existen casos en que la información recogida no lo es (tomas muestrales de un individuo de forma repetida). Sea $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \dots, (x_n, y_n)$ una muestra aleatoria de pares de observaciones tal que (x_i, y_i) indican dos mediciones tomadas de la misma unidad muestral (antes y después de un tratamiento o fenómeno x). Desea conocer si la población cambió de forma mínima aceptable luego del tratamiento o fenómeno. Se emplea la prueba de diferencias por parejas.

Con base en esto, se construye una muestra aleatoria de diferencias $d_1, d_2, d_3, d_4 \dots, d_n \Rightarrow d_i = x_i - y_i$ donde $i = 1, 2, 3, 4, \dots, n$ con una distribución $N(\mu_d, \sigma_d^2) \Rightarrow$ sus estimadores son $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ y $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$.

a) Prueba Bilateral para Diferencia por Medias

69. $H_0: \mu_d = D_0$

70. $H_1: \mu_d \neq D_0$

71. Estadístico de prueba $t_{Calculada} = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}}$.

72. Región de rechazo $t_{Calculada} > t_{\left(\frac{\alpha}{2};\,(n-1)\right)}$ o, sino, $t_{Calculada} > -t_{\left(\frac{\alpha}{2};\,(n-1)\right)}$.

b) Prueba Unilateral para Diferencia por Medias

73. $H_0: \mu_d = D_0$

74. $H_1: \mu_d > D_0$ o, también, $H_1: \mu_d < D_0$

75. Estadístico de prueba $t_{Calculada} = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}}$.

76. Región de rechazo $t_{Calculada} > t_{(\alpha;\,(n-1))}$ o, sino, $t_{Calculada} > -t_{(\alpha;\,(n-1))}$.

Ejemplo: Una investigación clínica realiza mediciones de frecuencia cardiaca a 9 personas antes y después de un entrenamiento físico:

Cuadro 29. Mediciones de frecuencia cardiaca

Concepto	Medición a exposición de ejercicio físico								
Antes	81	85	82	82	90	83	73	93	92
Después	76	71	87	79	90	89	76	75	89

Con nivel de significancia $\alpha = 0.05$ y una distribución normal $N(0, 1)$, realice sus cálculos respecto si el acondicionamiento físico varió significativamente las frecuencias cardiacas de las personas.

El promedio y desviación estándar de diferencias de mediciones son $\bar{d} = \frac{1}{9}(29) = 3.22$, $\sqrt{s_d^2} = 8.21$ y $D_0 = 0$.

1. $H_0: \mu_d = 0$
2. $H_1: \mu_d \neq 0$
3. Estadístico de prueba $t_{Calculada} = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}} = \frac{3.22 - 0}{\frac{8.21}{\sqrt{9}}} = 1.177$.
4. Región de rechazo $t_{Calculada\ (1.177)} > t_{\left(\frac{0.05}{2};\,(9-1)=2.306\right)}$.

$\Rightarrow |t_{Calculada\ =\ 1.177}| < |t_{Tablas:\ 2.306}|$. No cae en la zona de rechazo o, igualmente, $|t_{Calculada\ =\ 1.177}| < |t_{Tablas:\ 2.306}|$ no se rechaza H_0 y no se acepta H_a. Por lo tanto, no existe variación mínima aceptable en frecuencia cardiaca de personas a quienes se tomó las mediciones.

6.2.4. De hipótesis para razón entre dos varianzas

Suponga desea probar igualdad de varianzas de dos poblaciones normalmente distribuidas, extraídas dos muestras independientes; es decir, $H_0: \sigma_1^2 = \sigma_2^2$.

a) Prueba Bilateral para Razón entre Dos Varianzas

77. $H_0: \sigma_1^2 = \sigma_2^2$
78. $H_1: \sigma_1^2 \neq \sigma_2^2$

79. Estadístico de prueba $F_{Calculada} = \frac{s_1^2}{s_2^2}$.

80. Región de rechazo $F_{Calculada} > F_{\left(\frac{\alpha}{2};\,(n_1-1;\,n_2-1)\right)}$.

b) Prueba Unilateral para Razón entre Dos Varianzas

81. $H_0: \sigma_1^2 = \sigma_2^2$

82. $H_1: \sigma_1^2 > \sigma_2^2$ o, también, $H_1: \sigma_1^2 < \sigma_2^2$

83. Estadístico de prueba $F_{Calculada} = \frac{s_1^2}{s_2^2}$ ([217]).

84. Región de rechazo $F_{Calculada} > F_{(\alpha;\,(n_1-1;\,n_2-1))}$.

Ejemplo: Un inversionista no sabe si invertir en bonos emitidos por países A o B. Para tomar una decisión, selección dos muestras, correspondientes a rendimiento de bonos emitidos por ambos países, obtuvo unas varianzas muestrales iguales a $s_x^2 = 0.11$ y $s_y^2 = 0.07$. Con un nivel de significación $\alpha = 0.05$, verifique si sus varianzas poblacionales en rendimientos de ambos bonos son iguales.

Con base en esta información, la prueba es bilateral o de dos colas.

1. $H_0: \sigma_x^2 = \sigma_y^2$

2. $H_1: \sigma_x^2 \neq \sigma_y^2$

3. Estadístico de prueba $F_{Calculada} = \frac{s_x^2}{s_y^2} = \frac{0.11}{0.07} = 1.571$.

4. Región de rechazo $\left|F_{Calculada\ (1.571)}\right| > \left|F_{\left(\frac{0.05}{2};\,(10-1;\,16-1)\right)=3.12}\right|$.

$\Rightarrow \left|F_{Calculada\ (1.571)}\right| < \left|F_{\left(\frac{0.05}{2};\,(10-1;\,16-1)\right)=3.12}\right|$. No cae en la zona de rechazo o, igualmente, $\left|F_{Calculada\ (1.571)}\right| < \left|F_{\left(\frac{0.05}{2};\,(10-1;\,16-1)\right)=3.12}\right|$ no se rechaza H_0 y no se acepta H_a. Por lo tanto, no existe diferencia entre las varianzas de bonos emitidos por ambos países o, en otras palabras, existe homocedasticidad respecto a los bonos emitidos por ambos países.

[217] $s_1^2 > s_2^2$; es decir, s_1^2 es la mayor varianza muestral.

6.2.5. Diferencia entre dos proporciones

Suponga selecciona dos muestras aleatoria e independientemente respecto a dos poblaciones binomiales, con tamaños n_1 y n_2 que son suficientemente altos para que sus distribuciones muestrales de $\hat{p}_1$ y $\hat{p}_2$ sean aproximadamente normales. Desea probar si la diferencia de proporciones muestrales es $= D_0$ tal que se deben tomar en cuenta los casos $D_0 = 0$ –igualdad proporcional- y $D_0 \neq 0$.

6.2.5.1. $\mathbf{p_1 - p_2}$ cuando $\mathbf{D_0 = 0}$

Se estiman las proporciones p_1 y p_2 a partir de muestras, $\hat{p}_1$ y $\hat{p}_2$ respectivamente. Después, se calcula su estimador ponderado de proporción p mediante $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

a) Prueba Bilateral para Diferencia entre dos Proporciones

85. $H_0: p_1 - p_2 = 0$ o, también, $H_0: p_1 = p_2 = p$.

86. $H_1: p_1 - p_2 \neq 0$ o, también, $H_0: p_1 \neq p_2 \neq p$.

87. Estadístico de prueba $Z_{Calculada} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$.

88. Región de rechazo $Z_{Calculada} > Z_{\left(\frac{\alpha}{2}\right)}$ o, también, $Z_{Calculada} < -Z_{\left(\frac{\alpha}{2}\right)}$.

b) Prueba Unilateral para Diferencia entre dos Proporciones

89. $H_0: p_1 - p_2 = 0$ o, también, $H_0: p_1 = p_2 = p$.

90. $H_1: p_1 > p_2$ o $H_1: p_1 < p_2$

91. Estadístico de prueba $Z_{Calculada} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$.

92. Región de rechazo $Z_{Calculada} > Z_{(\alpha)}$ o, también, $Z_{Calculada} < -Z_{(\alpha)}$.

Ejemplo: El dueño de un supermercado cree que el porcentaje de cheques protestados, con clientes han pagado sus cuentas, ha aumentado con respecto al año anterior. En una muestra correspondiente al primer trimestre del año pasado, encontró 5 cheques protestados de 80 cheques admitidos; con otra muestra de 68 cheques correspondientes al primer trimestre del presente año, el número de cheques protestados fue de 6. Con base en estos datos, ¿hay evidencia suficiente que indique un incremento en porcentaje de cheques protestados?

Sean p_1 y p_2 proporciones de cheques protestados en primer trimestre del año pasado y presente. Se estima las proporciones muéstrales.

$\hat{p}_1 = \frac{5}{80} = 0.063$, $\hat{p}_2 = \frac{6}{68} = 0.082$ y $\hat{p} = \frac{5+6}{80+68} = 0.074$.

Se desea estimar si existe un aumento en proporción, tal que H_a: $p_1 - p_2 > 0$, su prueba es:

1. H_0: $p_1 = p_2$.
2. H_1: $p_1 < p_2$
3. Estadístico de prueba $Z_{Calculada} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.063-0.082}{\sqrt{(0.074*(1-0.0743))\left(\frac{1}{80}+\frac{1}{68}\right)}} = -0.595.$
4. Región de rechazo $Z_{Calculada\ (-0.595)} > Z_{(\alpha_{(0.05)}=1.645)}$.

$\Rightarrow \left|Z_{Calculada\ (-0.595)}\right| < \left|Z_{(\alpha_{(0.05)}=1.645)}\right|$. Cae en la zona de rechazo o, igualmente, $\left|Z_{Calculada\ (-0.595)}\right| < \left|Z_{(\alpha_{(0.05)}=1.645)}\right|$ no se rechaza H_0 y no se acepta H_a. Por lo tanto, no existe evidencia de aumento en número de cheques protestados respecto al año anterior.

6.2.5.2. $p_1 - p_2$ cuando $D_0 \neq 0$

En caso de probar que dos proporciones son distintas y su diferencia es igual a un valor preestablecido (D_0) se tiene las pruebas:

a) Prueba Bilateral para Diferencia entre dos Proporciones

93. H_0: $(p_1 - p_2) = D_0$.
94. H_1: $(p_1 - p_2) \neq D_0$.
95. Estadístico de prueba $Z_{Calculada} = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\left(\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}\right)}}$.
96. Región de rechazo $Z_{Calculada} > Z_{\left(\frac{\alpha}{2}\right)}$ o, también, $Z_{Calculada} < -Z_{\left(\frac{\alpha}{2}\right)}$.

b) Prueba Unilateral para Diferencia entre dos Proporciones

97. H_0: $(p_2 - p_1) = D_0$.
98. H_1: $(p_2 - p_1) > D_0$ o H_1: $(p_2 - p_1) < D_0$
99. Estadístico de prueba $Z_{Calculada} = \frac{(\hat{p}_2 - \hat{p}_1) - D_0}{\sqrt{\left(\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}\right)}}$.

Región de rechazo $Z_{Calculada} > Z_{(\alpha)}$ o, también, $Z_{Calculada} < -Z_{(\alpha)}$.

Ejemplo: En un partido de soccer se permite sustituir al portero sólo para que detenga tiros de penaltis. Al definir su estratégica para un partido, su entrenador examina estadísticas individuales de porteros titular y suplente. En una muestra de registros de entrenamientos

del último mes, el titular obtuvo 128 detenciones de penaltis de 510 y el suplente 183 detenciones de penaltis de 480 tiros. Su entrenador decidirá sustituir al portero titular si el suplente ha detenido al menos 10% más de tiros que su titular. Con un nivel de significancia $\alpha = 0.05$, ¿cuál decisión debe tomar su entrenador?

Con base en esta información, $\hat{p}_1 = \frac{128}{510} = 0.251$, $\hat{p}_2 = \frac{183}{480} = 0.381$, prueba unilateral de cola derecha y $D_0 = 0.10$.

1. $H_0: (p_2 - p_1) = 0.10$.
2. $H_1: (p_2 - p_1) > 0.10$
3. Estadístico de prueba $Z_{Calculada} = \frac{(\hat{p}_2 - \hat{p}_1) - D_0}{\sqrt{\left(\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}\right)}} = \frac{(0.381 - 0.251) - 0.10}{\sqrt{\left(\frac{0.251 * 0.749}{510} + \frac{0.381 * 0.619}{480}\right)}} = 1.023$.
4. Región de rechazo $Z_{Calculada\ (1.023)} > Z_{\alpha_{(0.05)}=1.64}$.

$\Rightarrow \left|Z_{Calculada\ (1.023)}\right| < \left|Z_{(\alpha_{(0.05)}=1.645)}\right|$. No cae en la zona de rechazo o, igualmente, $\left|Z_{Calculada\ (1.023)}\right| < \left|Z_{(\alpha_{(0.05)}=1.645)}\right|$ no se rechaza H_0 y no se acepta H_a. Por lo tanto, no existe evidencia respecto a que la diferencia es ≥ 0.10 y, en consecuencia, su entrenador no deberá sustituir a su portero titular.

Cuadro 30. Resumen de prueba de hipótesis con dos muestras ([218])

Población	H_0	H_a o H_1	Estadístico de prueba	Región de rechazo
General con varianzas iguales	$\mu_1 - \mu_2 = D_0$ $\mu_1 - \mu_2 \leq D_0$ $\mu_1 - \mu_2 \geq D_0$	$\mu_1 - \mu_2 \neq D_0$ $\mu_1 - \mu_2 > D_0$ $\mu_1 - \mu_2 < D_0$	$t_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$\lvert Z_{Calculada}\rvert > \lvert Z_{Tablas(\frac{\alpha}{2})}\rvert$ $\lvert Z_{Calculada}\rvert > \lvert Z_{Tablas(\alpha)}\rvert$ $\lvert Z_{Calculada}\rvert < \lvert -Z_{Tablas(\alpha)}\rvert$
Normal, varianzas desconocidas supuestas iguales			$t_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $n_1 + n_2 - 2$ gl	$\lvert t_{Calculada}\rvert > \lvert t_{Tablas(\frac{\alpha}{2})}\rvert$ $\lvert t_{Calculada}\rvert > \lvert t_{Tablas(\alpha)}\rvert$ $\lvert t_{Calculada}\rvert < \lvert -t_{Tablas(\alpha)}\rvert$
Normal, varianzas desconocidas supuestas distintas			$t_{Calculada} = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ r gl	
Normal observaciones emparejadas	$\mu_D = \mu_{D_0}$ $\mu_D \leq \mu_{D_0}$ $\mu_D \geq \mu_{D_0}$	$\mu_D \neq \mu_{D_0}$ $\mu_D > \mu_{D_0}$ $\mu_D < \mu_{D_0}$	$t_{Calculada} = \frac{\bar{d} - \mu_{D_0}}{\frac{s_d}{\sqrt{n}}}$ $n - 1$ gl	
Normal	$\sigma_1^2 = \sigma_2^2$ $\sigma_1^2 \leq \sigma_2^2$ $\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$ $\sigma_1^2 > \sigma_2^2$ $\sigma_1^2 < \sigma_2^2$	$F_{Calculada} = \frac{s_1^2}{s_2^2}$ $\nu_1 = n_1 - 1;\ \nu_2 = n_2 - 1$	$\lvert F_{Calculada}\rvert > \lvert F_{Tablas(\frac{\alpha}{2})}\rvert$ $\lvert F_{Calculada}\rvert > \lvert F_{Tablas(\alpha)}\rvert$ $\lvert F_{Calculada}\rvert < \lvert -F_{Tablas(1-\alpha)}\rvert$

[218] Fuente: Galindo, E. 2006

Población	H_0	H_a o H_1	Estadístico de prueba	Región de rechazo
Binomial	$p_1 = p_2$ $p_1 \leq p_2$ $p_1 \geq p_2$	$p_1 \neq p_2$ $p_1 > p_2$ $p_1 < p_2$	$Z_{Calculada} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$	$\lvert Z_{Calculada}\rvert > \left\lvert Z_{Tablas(\frac{\alpha}{2})}\right\rvert$ $\lvert Z_{Calculada}\rvert > \lvert Z_{Tablas(\alpha)}\rvert$ $\lvert Z_{Calculada}\rvert < \lvert -Z_{Tablas(\alpha)}\rvert$
Binomial $(D_0 \neq 0)$	$p_1 - p_2 = D_0$ $p_1 - p_2 \leq D_0$ $p_1 - p_2 \geq D_0$	$p_1 - p_2 \neq D_0$ $p_1 - p_2 > D_0$ $p_1 - p_2 < D_0$	$Z_{Calculada} = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\left(\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}\right)}}$	

6.3. PRUEBAS DE HIPÓTESIS NO PARAMÉTRICAS

Existen aplicaciones en ciencias e ingeniería en que no es posible conocer las distribuciones de poblaciones de las que se extraen muestras o datos se reportan como valores en escala ordinal. En casos así, se usan métodos alternativos equivalentes a parámetros, llamados Métodos No Paramétricos o Distribución Libre.

Usualmente, se usan pruebas no paramétricas cuando se trata de inferencias con muestras pequeñas y distribución poblacional desconocida, pues se puede usar el Teorema de Límite Central. La aplicación de métodos no paramétricos no requiere conocimientos matemáticos avanzados su trabajo será ordenar por rangos datos observados.

Si verifica condiciones exigidas en uso de prueba paramétrica implica usarla en vez de no paramétrica. Esto se debe a que, si usa el mismo nivel de significancia en ambas pruebas, la potencia de una no paramétrica es menor a una paramétrica. También, con métodos no paramétricos se pierde gran cantidad de información al calcular con rangos en lugar de sus valores originales.

Cuadro 31. Comparación de métodos no paramétricos con paramétricos

Métodos No Paramétricos respecto a Paramétricos	
Ventajas	**Desventajas**
1. Son fáciles de usar y entender.	1. Ocasionalmente, ignoran, desperdician o pierden información.
2. Eliminan necesidad de suposiciones restrictivas de pruebas paramétricas.	2. No son tan eficientes como las paramétricas.
3. Se pueden usar con muestras pequeñas.	3. Llevan mayor probabilidad de no rechazar H_0 falsa (Error tipo II)
4. Se pueden usar con datos cualitativos.	

Las pruebas no paramétricas se dividen en los grupos χ^2 de bondad de ajuste a una ley y Tablas de contingencia.

6.3.1. χ^2 de bondad de ajuste a una ley

Trata de procedimientos con el objetivo de determinar si un conjunto de observaciones sigue cierto esquema probabilístico. Estos comparan frecuencias observadas respecto a teóricas del modelo probabilístico con que se corrobora mediante un estadístico de prueba con ley χ^2.

6.3.1.1. Parámetros en un experimento multinomial

Si en la investigación se realizan n observaciones tal que n_1 están en la primera, n_2 en la segunda,…, n_k en la $k-$ ésima categoría tal que se cumple $n_1 + n_2 + n_3 + n_4 + \cdots + n_k = n$. Con base en esto, una investigación que presenta esta característica se llama Experimento Multinomial con las particularidades siguientes:

a) Experimento consta de n ensayos idénticos e independientes entre sí.

b) El resultado de cada ensayo se ubica en una y sólo una de las categorías.

c) La probabilidad que un ensayo se ubique en $i-$ ésima categoría es p_i, $i = 1, 2, 3, 4, \ldots, k$ y cumple $p_1 + p_2 + p_3 + p_4 + \cdots + p_k = 1$:

Cuadro 32. Nomenclatura de parámetros en un experimento multinomial

Concepto	Nomenclatura	Sumatoria
Categoría	$C_1\ C_2\ C_3\ C_4 \dots C_k$	Total
Frecuencia	$n_1\ n_2\ n_3\ n_4 \dots n_k$	n
Probabilidad	$p_1\ p_2\ p_3\ p_4 \dots p_k$	1

Con base en esto, interesa saber si el número de observaciones ubicadas en cada categoría se ajusta a un esquema de probabilidad dado o, en otras palabras, si las probabilidades de pertinencia por grupo tienen valores específicos $p_1 = p_{10}, p_2 = p_{20}, p_3 = p_{30}, p_4 = p_{40}, , \dots, p_k = p_{k0}$ tal que la prueba es:

1. $H_0: p_1 = p_{10}, p_2 = p_{20}, p_3 = p_{30}, p_4 = p_{40}, , \dots, p_k = p_{k0}.$
2. $H_1: p_1 \neq p_{10}, p_2 \neq p_{20}, p_3 \neq p_{30}, p_4 \neq p_{40}, , \dots, p_k \neq p_{i0}$
3. Estadístico de prueba $\chi^2_{Calculada} = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}.$
4. Región de rechazo $\chi^2_{Calculada} > \chi^2_{tablas\ (\alpha;\ k-1)}.$

Ejemplo: En un año se registró 100 nacimientos de gemelos en Hospital IESS, Quito. Según su sexo, su distribución es 2 varones 29 niños, 2 mujeres 38 niñas y 1 varón-1 mujer 33 niños. Considere que estos datos están distribuidos según ley trinomial de parámetros $(100, p_1, p_2, p_3)$ con $p_1, = p_2 = \frac{1}{4}, p_3 = \frac{1}{2}$ y un nivel de significancia $\alpha = 0.05.$

1. $H_0: p_1 = \frac{1}{4}, p_2 = \frac{1}{4}, p_3 = \frac{1}{2}.$
2. $H_1: p_1 \neq \frac{1}{4}, p_2 \neq \frac{1}{4}, p_3 \neq \frac{1}{2}$
3. Estadístico de prueba $\chi^2_{Calculada} = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \frac{\left(29 - \frac{1}{4} * 100\right)^2}{\left(\frac{1}{4} * 100\right)} + \frac{\left(38 - \frac{1}{4} * 100\right)^2}{\left(\frac{1}{4} * 100\right)} + \frac{\left(33 - \frac{1}{2} * 100\right)^2}{\left(\frac{1}{2} * 100\right)} = 11.260.$
4. Región de rechazo $\left|\chi^2_{Calculada(11.260)}\right| > \left|\chi^2_{tablas\ (\alpha=0.05;k-1=3-1)=5.991}\right|.$

$\Rightarrow \left|\chi^2_{Calculada(11.260)}\right| > \left|\chi^2_{tablas\ (\alpha=0.05;k-1=3-1)=5.991}\right|$. Cae en la zona de rechazo o, igualmente, $\left|\chi^2_{Calculada(11.260)}\right| > \left|\chi^2_{tablas\ (\alpha=0.05;k-1=3-1)=5.991}\right|$ no se acepta H_0 y no se rechaza H_a. Por lo tanto, el nacimiento de hermanos gemelos no sigue la ley establecida.

6.3.1.2. Bondad de ajuste a una ley

Suponga no puede asignar previamente probabilidades de pertinencia por grupo, sino será necesario estimar a partir de una ley de distribución teórica -Uniforme, Normal, Poisson u otra-, cuyos parámetros son conocidos o estimados con base en datos muestrales.

Se dispone de un conjunto de n observaciones provenientes de una ley de distribución dada, sea Y una variable aleatoria con ley señalada con valores $y_1 + y_2 + y_3 + y_4 + \cdots + y_i$ tal que $P_r(Y = y_i) = p_i$ tal que si Y sigue ley Poisson será $P_r(Y = y_i) = p_i = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$ tal que parámetro λ puede estar especificado previamente o será estimado. Por lo tanto, a partir de probabilidades teóricas se calcula frecuencia esperada, como $e_i = np_i$.

Cuando clase alguna tiene una frecuencia observada menor a 5 se agrupa en alguna otra adyacente y suma probabilidades respectivas. Luego de agruparlas, se obtienen k clases, se dispone de una tabla:

Cuadro 33. Agrupamiento por k clases

Grupo	Observación	Frecuencia	
		Observada	Esperada
1	x_1	n_1	$e_1 = np_1$
2	x_2	n_2	$e_2 = np_2$
3	x_3	n_3	$e_3 = np_3$
4	x_4	n_4	$e_4 = np_4$
⋮	⋮	⋮	⋮
k	x_k	n_k	$e_k = np_k$
Total		**n**	**n**

Con base en esto, el estadístico de prueba que comprueba si los datos siguen una ley específica es $\chi^2_{Calculada} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$, que sigue una distribución χ^2 con $[(k-1) - (\text{Número de parametros estimados})]$ grados de libertad. Si supone una ley de Poisson, con parámetro λ, el estadístico $\chi^2_{Calculada}$ sigue una distribución $\chi^2(k-2)$. La prueba de hipótesis será:

1. H_0: Datos siguen una ley $\mathcal{L}(p)$.
2. H_a: Datos no siguen una ley $\mathcal{L}(p)$.
3. Estadístico de prueba $\chi^2_{Calculada} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$ con k número de clases una vez agrupados sus datos.

4. Región de rechazo $\chi^2_{\text{Calculada}} > \chi^2_{\text{tablas }(\alpha;\, k-1-\ell)}$ con ℓ número de parámetros estimados con base en la muestra.

Ejemplos:

1) En una agencia bancaria hay 5 ventanillas con atención al cliente. Un día al azar, el gerente contabilizó el número de personas que atendía cada ventanilla. Los resultados son cajas 1, 2, 3, 4 y 5 atendieron, respectivamente, 34, 54, 39, 48 y 45 dando un total de 220 clientes atendidos. Con base en esta información, ¿existe preferencia por alguna de las cajas?

Con base en esta información, $p_i = P_r(Y = y_i) = \frac{1}{5}; i = 1, 2, 3, 4 \text{ o } 5$. Las frecuencias esperadas son $e_i = np_i = \left(220 * \frac{1}{5}\right) = 44;\ i = 1, 2, 3, 4 \text{ o } 5$.

1. H_0: Datos siguen una ley $\mathcal{L}(p)$ o no hay preferencia por alguna ventanilla.
2. H_a: Datos no siguen una ley $\mathcal{L}(p)$ o hay preferencia por alguna ventanilla.
3. Estadístico de prueba $\chi^2_{\text{Calculada}} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} = \frac{(34-44)^2}{44} + \frac{(54-44)^2}{44} + \frac{(39-44)^2}{44} + \frac{(48-44)^2}{44} + \frac{(45-44)^2}{44} = 5.50$ con k número de clases una vez agrupados sus datos.
4. Región de rechazo $\left|\chi^2_{\text{Calculada }(5.50)}\right| < \left|\chi^2_{\text{Tablas }(\alpha = 0.05;\, k-1 = 4:\, 9.49)}\right|$.

$\Rightarrow \left|\chi^2_{\text{Calculada }(5.50)}\right| < \left|\chi^2_{\text{tablas }(\alpha = 0.05;\, k-1 = 4:\, 9.49)}\right|$. No cae en la zona de rechazo o, igualmente, $\left|\chi^2_{\text{Calculada }(5.50)}\right| < \left|\chi^2_{\text{Tablas }(\alpha = 0.05;\, k-1 = 4:\, 9.49)}\right|$ no se rechaza H_0 y no se acepta H_a. Por lo tanto, la preferencia por alguna ventanilla no existe.

2) En una ensambladora de automóviles se registra el número de defectos o fallas por unidad en una muestra de 100 unidades que se inspeccionan en una semana, sus frecuencias son respecto a número de defectos $0, 1, 2, 3, 4$, y número de carros $64, 19, 8, 5, 4$, respectivamente. Contraste la hipótesis que estos datos se ajustan a una distribución Poisson.

$$\hat{\lambda} = \frac{[(0 * 64) + (1 * 19) + (2 * 8) + (3 * 5) + (4 * 3) + (5 * 1)]}{(64 + 19 + 8 + 5 + 4)} = 0.67$$

Según ley $\mathcal{P}_{(\text{Poisson})}(0.67)$, las probabilidades de pertinencia por grupo es $p_0 = P_r(X = 0) = \frac{e^{-067}(0.67)^0}{0!} = 0.512,\ p_1 = P_r(X = 1) = \frac{e^{-067}(0.67)^1}{1!} = 0.343,\ p_2 =$

$P_r(X=2)=\frac{e^{-067}(0.67)^2}{2!}=0.115,\ \ p_3=P_r(X=3)=\frac{e^{-067}(0.67)^3}{3!}=0.026$ y $p_4=P_r(X=4)=\frac{e^{-067}(0.67)^4}{4!}=0.004$. Además, las frecuencias esperadas son $e_0=(100*0.512)=51.2,\ \ e_1=(100*0.343)=34.3,\ \ e_2=(100*0.115)=11.5,\ \ e_3=(100*0.026)=2.6$ y $e_4=(100*0.004)=0.4$. Con base en esto, se estima estadístico de prueba $\chi^2_{Calculada}=\sum_{i=0}^{4}\frac{(n_i-e_i)^2}{e_i}=\frac{(64-52.1)^2}{52.1}+\frac{(19-34.3)^2}{34.3}+\frac{(8-11.5)^2}{11.5}+\frac{(5-2.6)^2}{2.6}+\frac{(4-0.4)^2}{0.4}=45.71.$

Entonces, la prueba estadística es:

1. H_0: Datos siguen una ley Poisson $\mathcal{P}(0.67)$.
2. H_a: Datos no siguen una ley Poisson $\mathcal{P}(0.67)$.
3. Estadístico de prueba $\chi^2_{Calculada}=\sum_{i=0}^{4}\frac{(n_i-e_i)^2}{e_i}=\frac{(64-52.1)^2}{52.1}+\frac{(19-34.3)^2}{34.3}+\frac{(8-11.5)^2}{11.5}+\frac{(5-2.6)^2}{2.6}+\frac{(4-0.4)^2}{0.4}=45.71.$
4. Región de rechazo $\left|\chi^2_{Calculada\ (45.71)}\right|>\left|\chi^2_{Tablas\ (\alpha=0.05;\ k-1-1=3:\ 7.81)}\right|$.

$\Rightarrow \left|\chi^2_{Calculada\ (45.71)}\right|>\left|\chi^2_{Tablas\ (\alpha=0.05;\ k-1-1=3:\ 7.81)}\right|$. Cae en la zona de rechazo o, igualmente, $\left|\chi^2_{Calculada\ (45.71)}\right|>\left|\chi^2_{Tablas\ (\alpha=0.05;\ k-1-1=3:\ 7.81)}\right|$ no se acepta H_0 y no se rechaza H_a. Por lo tanto, el número de defectos no sigue una ley Poisson $\mathcal{P}(0.67)$.

3) Se mide la duración de vida en anaquel de 200 frascos de mermeladas tal que la primera columna indica intervalos de tiempo (días) y la segunda el tiempo de vida del producto entre límites de intervalos.

Cuadro 34. Agrupamiento por k clases de duración de vida de anaquel de 200 frascos de mermeladas

Tiempo de vida en anaquel del producto (días)	Frecuencia (n_i)
$0-5$	133
$5-10$	45
$10-15$	15
$15-20$	4
$20-25$	2
$25-30$	1

Con un nivel de confiabilidad estadística 99% o nivel de significancia de 1%, verifique que la vida en anaquel del producto está distribuida según ley exponencial. Con base en esta información, H_0: Datos siguen una ley $\mathcal{L}(\varepsilon)$ tal que $\varepsilon(0.2)$,

H_a: Datos no siguen una ley $\mathcal{L}(\varepsilon)$ tal que $\varepsilon(0.2)$, estadístico de prueba $\chi^2_{\text{Calculada}} = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^{4} \frac{(n_i - e_i)^2}{e_i} = \frac{(133-126.42)^2}{126.42} + \frac{(45-46.52)^2}{46.52} + \frac{(15-17.10)^2}{17.10} + \frac{(7-9.48)^2}{9.48} = 1.30$, parámetro λ, que sigue una distribución exponencial, se estima mediante $\hat{\lambda} = \frac{1}{\bar{x}} = \bar{x}^{-1} = 0.2$, las probabilidades que la variable aleatoria tome valores en cada intervalo son estimadas mediante $p_1 = P_r(0 \leq x \leq 5) = e^{(-0.2*0)} - e^{(-0.2*5)} = 0.632$. Análogamente, $p_2 = 0.233$, $p_3 = 0.086$, $p_4 = 0.032$, $p_5 = 0.012$ y $p_2 = 0.004$ tal que con $e_i = np_i$ se estiman frecuencias esperadas:

Cuadro 35. Comparación de frecuencia observada y teórica de la vida de anaquel de 200 frascos de mermelada por ley exponencial

Número	Probabilidad de variable aleatoria $(\mathbf{p_i})$	Elementos muestrales $(\mathbf{n})$	Frecuencia observada $(\mathbf{n_i})$	Frecuencia teórica $(\mathbf{e_i})$	
1	0.632	200	133	126.42	126.42
2	0.233		45	46.52	46.52
3	0.086		15	17.10	17.10
4	0.032		4	6.30	9.48
5	0.012		2	2.32	
6	0.004		1	0.86	

Región de rechazo, como $\hat{\lambda} = 0.2$,

$\ell_{(\text{Número de parámetros estimados con base en la muestra})} = 1$, se examina la ley $\chi^2_{\text{Calculada}}$ con $(4 - 1 - 1) = 2$ grados de libertad y $\left|\chi^2_{\text{Calculada: } 1.30}\right| > \left|\chi^2_{\text{Tablas } (\alpha=0.01;\ k-1-\ell=4-1-1)=2):9.21}\right|$. Por lo tanto, con un nivel de confiabilidad estadística 99% o nivel de significancia de 1%, no se rechaza

H_0: Datos siguen una ley $\mathcal{L}(\varepsilon)$ y no se acepta H_a: Datos no siguen una ley $\mathcal{L}(\varepsilon)$ tal que $\varepsilon(0.2)$.

6.3.2. Tablas de contingencia

Cuando se tiene información de dos variables de tipo cualitativo se resume en una tabla de contingencia, que es una tabla de frecuencias de doble entrada en donde en las filas se ponen las modalidades de una variable y en columnas las modalidades de la otra, en celdas resultantes de cruce de las filas y columnas se coloca el número de elementos que presentan ambas modalidades.

Si se tiene información de N elementos acerca de las variables A y B tal que presentan r y c modalidades respectivamente. Su tabla de contingencia $r * c$, con r filas y c columnas, es:

Cuadro 36. Tabla de contingencia de variables A y B

Variable						
A	**B**					
	B_1	$\cdots$	B_i	$\cdots$	B_c	**Total**
A_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1c}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rc}	$n_{r.}$
Total	$\mathbf{n_{.1}}$	$\cdots$	$\mathbf{n_{.j}}$	$\cdots$	$\mathbf{n_{.c}}$	**N**

Tal que $n_{i.} = \sum_{j=1}^{c} n_i$ (total de frecuencia de fila i−ésima) y $n_{.j} = \sum_{j=1}^{c} n_i$ (total de frecuencia de columna j−ésima).

Ejemplo:

1) Los padres de familia que viven en una ciudad pueden clasificarse en dueños o arrendatarios de su casa y, según su nivel de instrucción, en primario, secundario y superior y, en consecuencia, se pueden clasificar en una tabla de contingencia $2 * 3$.

6.3.2.1. Independencia

Considérese las probabilidades $p_{i.}$, ubicada en la fila i; $p_{.j}$, ubicada en columna j; y p_{ij} es probabilidad de hallarse en celda (i, j). Además, su cumplirá que $\sum_{i=1}^{r} p_{i.} = 1$ y $\sum_{i=1}^{c} p_{.j} = 1$ pueden estimarse mediante $\widehat{p_{i.}} = \frac{n_{i.}}{N}, i = 1,2,3,4, \ldots . r$ y, también, $\widehat{p_{.j}} = \frac{n_j}{N}, j = 1,2,3,4, \ldots . r$. Bajo hipótesis de independencia entre filas y columnas, la frecuencia esperada en celda ubicada en la $i - ésima$ fila y $j - ésima$ columna es $e_{ij} = n\widehat{p_i}\,\widehat{p_j} = \frac{n_{i.} * n_j}{N}$. En consecuencia, el estadístico que permite probar su hipótesis de independencia es $\chi^2_{(Calculada)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$, que sigue aproximadamente una distribución normal χ^2 con $[(r-1)(c-1)]$ grados de libertad. Por lo tanto, su prueba para independencia es

H_0 (Hipótesis nula o negada: Variable A es independiente de variable B): $p_{i.} * p_{.j}$,

H_a (Hipótesis alterna o alternativa: Variable A y B no son independientes para ≥ 1 celda de tabla): $p_{ij} \neq p_{i.} * p_{.j}$, su estadístico de prueba es $\chi^2_{(Calculada)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$ y su región de rechazo será $\left|\chi^2_{(Calculada)}\right| > \left|\chi^2_{(Tablas:\ \propto;\ [(r-1)(c-1)]_{(Grados\ de\ libertad)})}\right|$ o $\left|\chi^2_{(Calculada)}\right| < \left|\chi^2_{(Tablas:\ \propto;\ [(r-1)(c-1)]_{(Grados\ de\ libertad)})}\right|$ ([219]).

Ejemplos:

1) En una investigación se desea conocer si existe relación entre consumo y origen de los carros que circulan por la ciudad de Quito.

Cuadro 37. Relación entre países de origen y consumo

Consumo	Origen			Total
	EEUU	Europa	Japón	
Bajo	76	56	70	202
Alto	160	14	9	183
Total	**236**	**70**	**79**	**385**

Con un nivel de confiabilidad estadística de 95% o nivel de significancia de 5%, se verifica si ambas variables están asociadas tal que, por un lado, las frecuencias esperadas (e_{ij}) son estimadas mediante $e_{11} = \frac{(n_{1.} * n_{.1})}{N} = \frac{(202*236)}{385} = 123.82$, $e_{12} = \frac{(n_{1.} * n_{.2})}{N} = \frac{(202*70)}{385} = 36.72$, $e_{13} = \frac{(n_{1.} * n_{.3})}{N} = \frac{(202*79)}{385} = 41.45$, $e_{21} = \frac{(n_{2.} * n_{.1})}{N} = \frac{(183*236)}{385} = 112.18$, $e_{22} = \frac{(n_{2.} * n_{.2})}{N} = \frac{(183*70)}{385} = 33.27$ y $e_{23} = \frac{(n_{2.} * n_{.3})}{N} = \frac{(183*79)}{385} = 37.55$, cuya tabla es:

Cuadro 38 Nivel de asociación entre países de origen y consumo

Consumo	Origen		
	EEUU	Europa	Japón
Bajo	123.82	36.72	41.45
Alto	112.18	33.27	37.55

[219] Valor absoluto es la distancia que existe entre un número y el 0 en la recta de números $\mathbb{R}$. Su función es $f(x) = |x|$, es a trozos y está definida por $|x| = \begin{cases} -x, & si\ < 0; \\ x, & si\ \geq 0; \end{cases}$ Algunas de sus propiedades son 1) $|-a| = a$, 2) $\sqrt{(a)^2} = |a|$, 3) $|a * b| = |a| * |b|$, 4) $\left|\frac{a}{b}\right| = \frac{|a|}{|b|}$, 5) $|a^2| = |a|^2$, 6) $|a + b| \leq |a| + |b|$, 7) $|a - b| \leq |a| + |b|$, 8) $|a| - |b| \leq |a - b|$, 9) $||a - b|| \leq |a - b|$, 10) $|a| < b \Leftrightarrow -b < a < b$, 11) $-|a| \leq a \leq |a|$, etcétera. Otra forma de demostrarlo es mediante $(-1)^2 = 1 \Rightarrow (-1)^2 - 1 = (-1 * -1) - 1 = -1(-1 + 1) \Longrightarrow (-1)^2 = (-1)^2+) - 1 + 1) = [((-1)^2 - 1) + 1] = 1$
(Douchet, J. y Zwahlen, B. 1983. Calcul diffèrentiel et intègral. 3 Fontions rèeles d´une variable réelle)

Con base es esto, su prueba para independencia es

H_0 (Hipótesis nula o negada: Origen del automóvil son independientes): $p_{i.} * p_{.j}$,

H_a (Hipótesis alterna o alternativa: Origen del carro y consumo están relacionados): $p_{ij} \neq p_{i.} * p_{.j}$, su estadístico de prueba es $\chi^2_{(\text{Calculada})} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(76-123.82)^2}{123.82} + \frac{(56-36.72)^2}{36.72} + \frac{(70-41.45)^2}{41.45} + \frac{(160-112.18)^2}{112.18} + \frac{(14-33.27)^2}{33.27} + \frac{(9-37.55)^2}{37.55} = 101.51$ y su región de rechazo será $\left|\chi^2_{(\text{Calculada:}101.51)}\right| > \left|\chi^2_{(\text{Tablas: } \propto=0.05;\, [(2-1)(3-1)]_{(\text{Grados de libertad})})=5.99}\right|$, por lo que no se acepta

H_0 (Hipótesis nula o negada: Origen del automóvil son independientes) y no se rechaza

H_a (Hipótesis alterna o alternativa: Origen del carro y consumo están relacionados).

En tabla de contingencia de $2 * 2$, se desarrollada una fórmula de cálculo de $\left|\chi^2_{(\text{Calculada})}\right|$ que no requiere la determinación de frecuencias esperadas tal que se dispone de una tabla de contingencia con dos variables, sólo pueden tomar dos valores:

Cuadro 39. Tabla de contingencia con dos variables

Variable			
A	**B**		**Total**
	B_1	B_2	
A_1	a	b	$a + b$
A_2	c	d	$c + d$
Total	$A + c$	$b + d$	n

Su estadístico de prueba se estima mediante $\chi^2_{(\text{Calculada})} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$, que sigue una $\chi^2_{(1 \text{ Grado de libertad})}$ tal que su prueba para independencia es

H_0 (Hipótesis nula o negada: Variable A es independiente de variable B): $p_{i.} * p_{.j}$,

H_a (Hipótesis alterna o alternativa: Variable A y B no son independientes para ≥ 1 celda de tabla): $p_{ij} \neq p_{i.} * p_{.j}$, su estadístico de prueba es $\chi^2_{(\text{Calculada})} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ y su región de rechazo será $\left|\chi^2_{(\text{Calculada})}\right| > \left|\chi^2_{(\text{Tablas: } \propto;\, [(r-1)(c-1)]_{(\text{Grados de libertad})})}\right|$ o $\left|\chi^2_{(\text{Calculada})}\right| < \left|\chi^2_{(\text{Tablas: } \propto;\, [(r-1)(c-1)]_{(\text{Grados de libertad})})}\right|$.

2) Se muestra la reacción por sexo de espectadores ante un comercial de televisión.

Cuadro 40. Reacción sexual por espectadores

Reacción	Sexo		Total
	Hombres	Mujeres	
Desfavorable	10	5	15
Favorable	3	7	10
Total	**13**	**12**	**25**

Con base en esta información,

H_0 (Hipótesis nula o negada: Reacción del comercial es independiente del sexo del espectador): $p_{i.} * p_{.j}$,

H_a (Hipótesis alterna o alternativa: Reacción del comercial no es independiente del sexo del espectador): $p_{ij} p_{i.} * p_{.j}$, su estadístico de prueba es $\chi^2_{(\text{Calculada})} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{25((10*7)-(5*3))^2}{(15)(10)(13)(12)} = 3.23$ y su región de rechazo será $\left|\chi^2_{(\text{Calculada}):3.23}\right| < \left|\chi^2_{(\text{Tablas: } \propto=0.05;\ 1\ (\text{Grados de libertad})):3.84}\right|$. Por lo tanto, no se rechaza

H_0 (Hipótesis nula o negada: Reacción del comercial es independiente del sexo del espectador) y no se acepta

H_a (Hipótesis alterna o alternativa: Reacción del comercial no es independiente del sexo del espectador)

6.3.2.2. Homogeneidad

Una muestra es homogénea si todas sus observaciones son generadas por el mismo modelo de distribución de probabilidad o pertenecen a una misma población. La metodología de prueba de independencia puede aplicarse para averiguar si existe homogeneidad entre dos o más muestras independientes. Su prueba estadística es

H_0 (Hipótesis nula o negada: Muestras provienen de una población−homogéneas−)

H_a (Hipótesis alterna o alternativa: Muestras no provienen de una población−heterogéneas), su estadístico de prueba es $\chi^2_{(\text{Calculada})} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$ y su región de rechazo será $\left|\chi^2_{(\text{Calculada})}\right| > \left|\chi^2_{(\text{Tablas: } \propto;\ [(r-1)(c-1)]_{(\text{Grados de libertad})})}\right|$ o $\left|\chi^2_{(\text{Calculada})}\right| < \left|\chi^2_{(\text{Tablas: } \propto;\ [(r-1)(c-1)]_{(\text{Grados de libertad})})}\right|$.

Ejemplos:

1) En la Facultad de Ingeniería, Universidad Nacional de Chimborazo (UNACH), se clasifica las notas obtenidas por sus estudiantes (bajas, medias y altas), luego do rendir el mismo examen de matemáticas. También, se registró el profesor que dictaba el curso.

Cuadro 41. Relación de calificaciones estudiantiles con dos profesores

Profesor	Calificación			Total
	Baja	Media	Alta	
A	12	23	7	42
B	25	17	4	46
Total	**37**	**40**	**11**	**88**

Por lo tanto, sus frecuencias esperadas son:

Cuadro 42. Frecuencias esperadas de calificaciones estudiantiles con dos profesores

Profesor	Calificación		
	Baja	Media	Alta
A	17.7	19.1	5.3
B	19.3	20.9	5.8

Con base en esto,

H_0 (Hipótesis nula o negada: Diferencias entre notas no son debidas al profesor del curso)

H_a (Hipótesis alterna o alternativa: Diferencias entre notas son debidas al profesor del curso), su estadístico de prueba es $\chi^2_{(\text{Calculada})} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(12-17.7)^2}{17.7} + \frac{(23-19.1)^2}{19.1} + \cdots + \frac{(4-5.8)^2}{5.8} = 6.12$ y su región de rechazo será $\left|\chi^2_{(\text{Calculada}):\, 6.12}\right| > \left|\chi^2_{(\text{Tablas: } \propto=0.05;\, [(2-1)(3-1)]_{(\text{Grados de libertad})}):\, 5.99}\right|$. Por lo tanto, no se acepta

H_0 (Hipótesis nula o negada: Diferencias entre notas no son debidas al profesor del curso) y no se rechaza

H_a (Hipótesis alterna o alternativa: Diferencias entre notas son debidas al profesor del curso).

7. FUENTES BIBLIOGRAFICAS

1. Anderson, C. D; Sweeney, D.J; Williams, T. A; Camm, J. D. y Cochran, J. J. 2016. Estadística para Administración y Economía. 12a. Edición. Editorial CENPAGE Learning Editores S. A. de C. V. México, D. F.
2. Apostol, T. M. 1985. Calculus. Cálculo con funciones de varias variables y álgebra lineal, con aplicaciones a las ecuaciones diferenciales y a las probabilidades. Vol. II. Ed. Reverté, S.A. 297-300 Pp..
3. Aragón, S. L. G. 2016. Estadística en el área de Ciencias Sociales y Administrativas. Alfaomega Grupo Editor, S. A. de C. V. México.
4. Arana Ovalle, R. I. 2003. Métodos de Muestreo. Tesis de Licenciatura en Estadística. División de Ciencias Forestales. Universidad Autónoma Chapingo.
5. Capa, S. H. b. 2015. Investigación por muestreo. Fundamentos y Aplicaciones. Escuela Politécnica Nacional (EPN).
6. Capa, S., H. a. 2015. Probabilidad y Estadística. Editorial EPN
7. Carrillo, E. G. 2005. Notas de muestreo forestal. División de ciencias Forestales. Universidad Autónoma Chapingo
8. Carrillo, E. G. 2010. Epidometría. División de ciencias Forestales. Universidad Autónoma Chapingo
9. Daza P. G. F. 2006. Estadística Aplicada con Excel
10. Galindo, E. 2006. Estadística. Métodos y Aplicaciones. PROCIENCIA EDITORES
11. Infante, G. S. y Zárate de L., G. P. 1984. Métodos Estadísticos. Un enfoque interdisciplinario
12. Levine, D. M; Krehbel, T. C. y Berenson, M. L. 2014. Estadística para administración. Editorial PEARSON EDUCACIÓN. México.
13. Lind, D. A; Marchal, W. G. y Mason, R. D. 2004. Estadística para Administración y Economía
14. Mendehall, III Jr; Beaver, R. J. y Barbara, M. B. 2015. Introducción a probabilidad
15. Murray R. S. y Larry J. S. 2009. Estadística
16. Navidi, W. 2006. Estadística para Ingenieros y Científicos. Ed. Mc Graw-Hill

17. Ortiz, P. J. 2013. Principios de estadística aplicada. Editorial Ediciones de la U. Bogota, Colombia.
18. Oteyza, E; Lam, E; Hernández, C. y Carrillo, A. 2015. Probabilidad y Estadística. Editorial PEARSON EDUCACIÓN. México.
19. Pérez López, C. 2005. Muestreo estadístico. Conceptos y problemas resueltos. Universidad Complutense de Madrid. Instituto de Estudios Fiscales. Editorial PEARSON EDUCACIÓN. Madrid.
20. Schreuder, Ernst y Ramírez. 2006. Statistical techniques for sampling and monitoring natural resources. División de Ciencias Forestales. Universidad Autónoma Chapingo
21. Vargas, A. D. 2006. Manejo instrumental del concepto de hipótesis en diseño de un proyecto de investigación. Área de soporte técnico a procesos de investigación e innovación
22. Wackerly, D. D; Mendenhall, W. y Scheaffer, R. L. 2010. Estadística Matemática con Aplicaciones

Marzo 2022 - CID - Centro de Investigación y Desarrollo
Copyright © - CID - Centro de Investigación y Desarrollo
Copyright del texto © 2022 de Autores
Formato: PDF
Paginas: 231
Tamaño: Sobre C5
Requisitos de sistema: Adobe Acrobat Reader
Modo de acceso: World Wide Web
ISBN: 978-99925-13-05-7
DOI: https://doi.org/10.37811/cli_w743
libros.ciencialatina.org
editorial@ciencialatina.org
Atención por WhatsApp al +52 1 999 102 8914