Prontuario para la materia de bioestadística

Ernesto Bravo Núñez

Prontuario para la materia de bioestadística

Universidad Autónoma de la Ciudad de México

M en C. Juan Carlos Aguilar Franco
Rector

Dra. María Elizbeth Álvarez Sánchez
Coordinadora Académica

Lic. Jorge Luis Rubio Hernández
Coordinador de Difusión Cultural y Extensión Universitaria

Equipo de la Biblioteca del Estudiante

Ángeles Godínez Guevara
Responsable

Ana Beatriz Alonso Osorio
Ana Lina Graciano Franco
Daniel Valentín Cruz
Florina Piña Cancino
María del Pilar Aparicio Romero
Sergio Javier Cortés Becerril

Prontuario para la materia de bioestadística

Ernesto Bravo Núñez

UACM
Universidad Autónoma
de la Ciudad de México
NADA HUMANO ME ES AJENO

Biblioteca
BE
del
Estudiante

FICHA CATALOGÁFICA E-S/N

Bravo Núñez, Ernesto

Prontuario para la materia de bioestadística / Ernesto Bravo Núñez. -- Primera edición. -- Ciudad de México : Universidad Autónoma de la Ciudad de México, 2024.

173 páginas : gráficas, tablas ; 21 cm.

Bibliografía: página 173.

ISBN: En trámite.

1. Estadística - Libros de texto universitarios. 2. Biometría - Libros de texto universitarios. 3. Distribución (Teoría de probabilidades). 4. Prueba de hipótesis estadística. I. Título.

LC QH 323.5 Dewey 570.15195

Prontuario para la materia de bioestadística

primera edición, 2025

García Diego 168, col. Doctores,

alc. Cuauhtémoc, c. p. 06720, México, D F

ISBN: 978-607-2615-66-3

https://www.uacm.edu.mx/Organizacion/CoordinacionAcademica/Biblioteca_Estudiante

Hecho e impreso en México

Indice

Introducción

El prontuario para la materia de Bioestadística se ubica en el tercer semestre de las licenciaturas en Promoción de la Salud, Ciencias Ambientales y Protección Civil y Gestión de Riesgos de la Universidad Autónoma de la Ciudad de México, para las que es obligatoria.

El objetivo académico es que los estudiantes de la materia tengan disponible un texto de consulta rápida en el que puedan revisar la aplicación e interpretación de algunas herramientas sumamente utilizadas en Bioestadística, no solamente el libro guía y las explicaciones de la clase del profesor de la materia; así como también a profesionistas que ya tengan conocimientos de Estadística o Bioestadística.

El prontuario no reemplaza el estudio de la Bioestadística en libros, en un curso o la asesoría del profesor. Es un complemento para la enseñanza-aprendizaje que contribuye a facilitar el aprendizaje de los estudiantes y la enseñanza de los profesores.

Los libros de consulta que se encuentran al final del texto, son de nivel universitario, abarcan la teoría, herramientas y ejemplos, cuyo dominio depende del estudio que realicen los estudiantes. El presente texto, en su extensión y profundidad académica, se localiza entre el libro de texto y el formulario: es un prontuario con ejemplos de cada herramienta estadística aplicada en la Bioestadística, útil para cada una de las licenciaturas mencionadas. Las herramientas son expuestas de manera detallada para su cálculo y aplicación. El Prontuario está vinculado a sus intereses de aprender, certificar la materia y tener un ejercicio capacitado acorde con los de profesionistas en ciernes.

El presente documento no incluye teoría y no es un formulario: es un prontuario de Bioestadística en donde se encontrarán ejemplos nacionales en los que se aplican las herramientas bioestadísticas y se interpretan.

El prontuario abarca: léxico estadístico general, pruebas paramétricas y no paramétricas, bibliografía; y se recomienda el *software Statdisk,* elaborado para la enseñanza por M. F. Triola, autor del libro: *Elementary Statistics,* y al que se puede acceder desde su plataforma con la dirección **statdisk.com**.

1.- Léxico estadístico

La bioestadística es una rama aplicada de la estadística, por lo que se utilizan los mismos conceptos que en esta área de las matemáticas, así como algunos conceptos generales para el manejo de datos; los términos utilizados en el programa de la materia de bioestadística se podrían reducir a los siguientes:

ALEATORIEDAD. Cualquier proceso cuyo resultado está influido por el azar (por una causa desconocida), el que es el conjunto de situaciones complejas desconocidas por el ser humano en la que él mismo no interviene o no puede intervenir.

Ejemplo 1. Al lanzar un dado honesto (que no está cargado para favorecer a ninguno de los seis resultados posibles; esto es, que cada uno de los seis lados tiene la misma probabilidad de ser el resultado), se desconoce cuál de los seis números será el ganador. Entonces se dice que es un proceso aleatorio, como muchos otros, cuyo resultado es debido al azar y no está sujeto por causa alguna a ser siempre el mismo (determinístico).

BASES DE DATOS. Es el conjunto de datos cualitativos o cuantitativos que se registran o miden a cada una de las unidades de estudio. En el programa Excel las columnas codificadas con letras corresponden a las variables en tanto que los renglones son los registros de todas las variables para una unidad de estudio.

Ejemplo 2. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, la base de datos en el programa Excel incluiría en las columnas las variables de cada sismo, como pueden ser: número consecutivo, coordenadas geográficas, magnitud en escala de Richter, fecha, nombre del epicentro, hora en la que ocurrió, año de ocurrencia, profundidad de ocurrencia, etc. En tanto que, en los renglones, estaría para cada sismo la información que le corresponde según la variable en turno de ser llenada.

Ejemplo 3. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, la base de datos en el programa Excel incluiría en las columnas las variables de cada estudiante, como pueden ser, número consecutivo, nombre del estudiante, sexo, nombre de la escuela, curso predilecto, años estudiados, etc. En tanto que, en los renglones, estaría para cada estudiante la información que le corresponde según la variable en turno de ser llenada.

CALIDAD DE LAS BASES DE DATOS. Una base de datos no debe tener errores de los registros o el nombre de las variables para que pueda ser utilizada con plena confianza en el análisis estadístico que se requiera. Si tiene errores de «dedo» y/o registro en al menos una unidad de estudio y no es corregido, serán arrastrados a los análisis a los que la base de datos sea sujetada; esta situación puede conducir a errores en las conclusiones y en consecuencia a la ejecución de acciones equivocadas. Una base de datos debe garantizar la fiabilidad de la información, por lo que antes de darla por terminada se debe asegurar que no contiene equivocaciones.

CENSO. Consiste en que cada una de las unidades de estudio que forman parte de la población objetivo son registrados o medidos en el desarrollo del proyecto, sin dejar de incluir a ninguno independientemente del tamaño de la población objetivo. Al realizar un censo no se conoce el error estadístico, el costo y tiempo son mayores que en el caso de una muestra. En estadística es común trabajar con muestras en las que se conoce el nivel de confianza y el tamaño del error.

Ejemplo 4. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, el censo queda determinado por la seguridad de que todas y cada una de las unidades de estudio, esto es, cada sismo ocurrido que integra a la población objetivo es registrado, sin faltar alguno. Sin embargo, conforme aumenta la cantidad de unidades de estudio en la población objetivo, la probabilidad de que alguno no sea registrado también aumenta.

Ejemplo 5. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, el censo queda determinado por la seguridad de que todas y cada una de las unidades de estudio, esto es, cada estudiante que integra la población objetivo es medido o registrado, sin faltar alguno. Sin embargo, conforme aumenta la cantidad de unidades de estudio en la población objetivo, la probabilidad de que alguno no sea registrado también aumenta.

CRITERIOS DE ELIMINACIÓN. Conjunto de afirmaciones utilizadas para eliminar, de un estudio experimental ya en ejecución, las unidades de estudio (fenómenos, objetos o personas) que dejan de cumplir con los Criterios de Inclusión o que por cualquier otra causa pudieran alterarlo. Estas afirmaciones son complementarias a los criterios de inclusión y heredadas por la muestra.

Ejemplo 6. Si en un estudio de seguimiento de una dieta para adelgazar, los sujetos a experimentación están comprometidos a apegarse a la dieta, pero, uno (a) de los (as) será eliminada (o) en el caso de que deje de seguir la dieta, aceptada por él o ella en el tiempo que dure el experimento o por alguna causa que pudiera alterar los resultados del experimento.

CRITERIOS DE EXCLUSIÓN. Son el conjunto de afirmaciones utilizadas para no incluir en la Población Objetivo a los fenómenos, objetos o personas. Estas afirmaciones son complementarias a los criterios de inclusión y heredadas por la muestra.

Ejemplo 7. En el caso de estudiar todos y cada uno de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, los criterios de exclusión son todos y cada uno de los sismos que no ocurrieron en dicha alcaldía, en ese año o menores o iguales a una magnitud de 4.

Ejemplo 8. En el caso de todos y cada uno de los estudiantes que estudiaron la licenciatura en Salud, en la ciudad de México en el año 2008, los criterios de exclusión son los estudiantes que no estudiaron la licenciatura en Salud a nivel universitario, que no lo hicieron en la Ciudad de México y los que estudiaron la licenciatura en Salud en un año diferente al 2008.

CRITERIOS DE INCLUSIÓN. Son el conjunto de afirmaciones utilizadas para delimitar a la población objetivo que incluye fenómenos, objetos o personas. Estas afirmaciones son heredadas por la muestra. Los criterios de inclusión deben cumplirse simultáneamente por

cada una de las unidades de estudio. A mayor cantidad de criterios de inclusión menor será el tamaño de la población objetivo.

Ejemplo 9. En el caso de estudiar todos y cada uno de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, los criterios de inclusión son los sismos, la alcaldía de Coyoacán, el año en que ocurrieron y la magnitud con que se presentaron superior a una magnitud 4 en la escala de Richter.

Ejemplo 10. En el caso de todos y cada uno de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, los criterios de inclusión son los estudiantes que estudiaron la licenciatura en Salud, en la Ciudad de México y en el año 2008.

DATOS CONTINUOS. Son aquellos que sí tiene sentido subdividir, incluso infinitamente, entre dos enteros consecutivos. Corresponden a mediciones con algún artefacto o registro. Se presentan en las variables de intervalo y en las de razón.

Ejemplo 11. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, un dato continuo podría ser la distancia en kilómetros de un epicentro a otro (1.45 km, 2.87 km, 4.97 km...) o la magnitud de acuerdo con la escala de Richter (magnitud 3.6, magnitud 4.7, magnitud 5.2...) etc.

Ejemplo 12. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, un dato continuo es el tiempo de transportación, en horas y minutos, de su casa a la escuela donde estudia (0.45 hrs., 1.30 hrs., 2.34 hrs...) etc.

DATOS CUANTITATIVOS. Son características que pueden ser medidas o contadas. Corresponden a las variables de intervalo o de razón.

Ejemplo 13. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, un dato cuantitativo es la cantidad de sismos ocurridos N = 3,765 o n = 349, la magnitud de los sismos en la escala de Richter, etc.

Ejemplo 14. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, un dato cuantitativo es la cantidad de estudiantes N = 11,040 o n = 570, la cantidad de horas de asistencia a clase, etc.

DATOS DISCRETOS. Son datos cuantitativos que no deben subdividirse entre dos enteros consecutivos ya que fraccionarlos no tiene sentido. Corresponden a conteos. Se pueden presentar en todo tipo de variables.

Ejemplo 15. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, un dato discreto es la cantidad de sismos ocurridos por epicentros, etc.

Ejemplo 16. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, un dato discreto es la cantidad de estudiantes mujeres, la cantidad de estudiantes hombres, etc.

DESCRIPCIÓN DE UNA MUESTRA. La descripción de la muestra es todo análisis estadístico de las unidades de estudio que la conforman. La mejor forma de efectuarla, aunque cuando no se requiera la inferencia de la población objetivo, es a partir de una muestra significativa y representativa, y en caso de que se trate de una muestra no probabilística (a juicio), no debe hacerse inferencia hacia la población objetivo, pues se desconoce la confianza, el

error aceptable y si las unidades de estudio tuvieron la misma probabilidad, por lo tanto, hacer inferencia a partir de una muestra no probabilística, conlleva de manera implícita, la posibilidad de equivocarse en las determinaciones sobre la población objetivo.

ESTADÍSTICOS. Se localizan en cualquier tipo de muestra; son una estimación que puede ser insesgada de los parámetros de la población objetivo y se representan con letras arábigas: media ($\bar{x}$), varianza (s^2), desviación estándar (s) o por medio de letras arábigas minúsculas como el tamaño de la muestra (n).

Ejemplo 17. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, los estadísticos de la muestra que estiman a los parámetros de la población objetivo son, media de = 0.96 sismos/día, varianza $s^2 = 0.9$ (sismos/día)^2 y desviación estándar de s = 0.03 sismos/día.

Ejemplo 18. En la medición o registro de los estudiantes de la licenciatura en Salud en la Ciudad de México en el año 2008, los estadísticos de la muestra son, una media de asistencia a clases de = 4.8 asistencia a clase/mes, varianza de $s^2 = 8$ (asistencia a clase/mes)^2 y una desviación estándar de s = 2.83 asistencia a clase/mes.

INFERENCIA. Esta característica consiste en que, a partir de una muestra significativa y representativa los resultados de medir y/o registrar los datos de las unidades de estudio pueden ser aplicados a la totalidad de la población objetivo, con la característica de saber la confianza y el tamaño del error aceptable.

Ejemplo 19. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, se determina una muestra significativa y representativa de la población objetivo, y los resultados obtenidos de dicha muestra son aplicados a la totalidad de la población objetivo, esto es, si el total de sismos es de N = 3,765 el tamaño de la muestra para medir el promedio del parámetro requerido con una confianza del 95%, un error aceptable de 1.0 unidad sobre el parámetro requerido y una desviación estándar poblacional = 10, el tamaño de la muestra sería de n = 349 o n = 385 unidades de estudio (sismo) si se desconoce el total de unidades de estudio (sismos) de la población objetivo. A esta muestra de unidades de estudio (sismos) se les toman los datos para el proyecto de investigación. Los resultados obtenidos en la muestra se aplican al total de N = 3,765 sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a la magnitud 4 en la escala de Richter.

Ejemplo 20. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, se determina una muestra significativa y representativa de la población objetivo y los resultados obtenidos de dicha muestra son aplicados a la totalidad de la población objetivo, esto es, si el total de estudiantes fuera de N = 11,040 para determinar una proporción de mujeres con una confianza del 95% y un error aceptable de 0.04 sobre el parámetro de estudio, el tamaño de la muestra es n = 570 o n = 601 unidades de estudio (estudiante de la licenciatura en Salud) si se desconoce el total de unidades de estudio. En esta muestra de unidades de estudio se registra la proporción de mujeres. Los resultados obtenidos en la muestra son aplicados a los N = 11,040 estudiantes de la licenciatura en Salud en la Ciudad de México en el año 2008.

MUESTRA. Es el subconjunto de la población objetivo al cual se le medirá o registrará la o las variables de interés para el proyecto de investigación. Hay dos tipos de muestras, las probabilísticas y las no probabilísticas o a juicio.

Ejemplo 21. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, sería un subconjunto o fracción de todos los sismos (unidades de estudio) que forman la población objetivo.

Ejemplo 22. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, sería un subconjunto o fracción de todos los(as) estudiantes (unidades de estudio) que forman la población objetivo.

MUESTRA NO PROBABILÍSTICA O A JUICIO. De la población objetivo se determina un tamaño de la muestra por decisión del investigador, sin utilizar una ecuación específica, no toma en cuenta la confianza estadística y el tamaño aceptable de error.

Ejemplo 23. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, sería un subconjunto o fracción determinado por decisión, esto es, si el total de sismos es de N = 3,765 o cualquier otra cantidad que integra la población objetivo, el investigador toma la decisión de que solamente se registrarán n = 100 sismos (unidades de estudio) u otra cantidad del total que forman la población objetivo para realizar el estudio del parámetro requerido.

Ejemplo 24. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, sería un subconjunto o fracción determinado por decisión, esto es, si el total de estudiantes fuerade N = 11,040 o cualquier otra cantidad que integre a la población objetivo, el investigador toma la decisión de que solamente se registrarán y/o medirán n = 120 estudiantes u otra cantidad del total que forman la población objetivo para realizar el estudio.

MUESTRA PROBABILÍSTICA. De la población objetivo se determina un tamaño de la muestra, a través de una ecuación específica en la que toman parte la confianza estadística y el tamaño del error aceptable.

Ejemplo 25. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, una muestra probabilística de la población objetivo, N = 3, 765 el tamaño de la muestra para medir el promedio del parámetro requerido con una confianza del 95%, un error aceptable de 1.0 sobre el parámetro requerido y una desviación estándar = 10, el tamaño de la muestra sería de n = 349 o n = 385 unidades de estudio (sismo) si se desconoce el total de unidades de estudio (sismos) de la población objetivo.

Ejemplo 26. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, una muestra probabilística de la población objetivo N = 11,040 para determinar una proporción de mujeres con una confianza del 95% y un error aceptable de 0.04 sobre el parámetro de estudio, el tamaño de la muestra es de n = 570 o n = 601 unidades de estudio (estudiante de la licenciatura en Salud) si se desconoce el total de unidades de estudio.

MUESTRA SIGNIFICATIVA Y REPRESENTATIVA. De la población objetivo se determina una muestra probabilística, y las unidades de estudio de la población objetivo que la integran tienen la misma probabilidad de ser parte de esa muestra, obteniendo de esta manera una muestra significativa y representativa.

PARÁM**ETROS**. En una ecuación son valores constantes para un conjunto de datos, como pueden ser las constantes del modelo de regresión lineal b_o (ordenada al origen)

y b_1 (pendiente) entre otros. En una población objetivo son todas aquellas características estadísticas cuyo cálculo repetido e inmediato no cambian de valor y se representan con letras griegas: media (μ), la varianza (^{2}) y la desviación estándar ( ) o por medio de mayúsculas arábigas como el tamaño de la población objetivo (N) o minúsculas como la proporción (p).

Ejemplo 27. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, los parámetros de la población objetivo podrían ser N = 3,765 sismos en el año con una media anual de μ= 5.4 grados Richter, varianza de ^{2} = 4 (grados Richter)^2 y desviación estándar de = 2 grados Richter.

Ejemplo 28. En la medición o registro de los estudiantes de la licenciatura en Salud en la Ciudad de México en el año 2008, los parámetros de la población objetivo podrían ser N = 11,040 estudiantes en el año, una media de asistencia a clase de μ= 3.6 asistencia a clase/mes, varianza de ^{2} = 9 (asistencia a clase/mes)^2 y una desviación estándar de = 3 asistencia a clase/mes.

POBLACIÓN. La población o universo se refiere a un conjunto de elementos del que se quiere saber algún aspecto. La población puede estar constituida por objetos, fenómenos o personas.

Ejemplo 29. Todos y cada uno de los sismos ocurridos. La población o universo solamente queda definida por el fenómeno de interés principal, que son los sismos.

Ejemplo 30. Todos y cada uno de los estudiantes que estudiaron la licenciatura en Salud. De igual forma solamente se incluye el fenómeno de interés y que en este caso se refiere a las personas que estudiaron la licenciatura en Salud.

POBLACIÓN FINITA. Si se conoce la cantidad total de una población (N), por el hecho de conocerla se dice que es una población finita.

Ejemplo 31. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, si la cantidad total de sismos N es conocida es una población finita.

Ejemplo 32. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, si la cantidad total de estudiantes, N es conocida es una población finita.

POBLACIÓN INFINITA. Si no se conoce la cantidad total de una población, por el hecho de no conocerla se dice que es una población infinita, situación equivalente a cuando una población es muy grande, como la cantidad de bacterias de una especie en el almacenamiento de agua en alguna localidad, la cantidad de estrellas en el cielo nocturno, etc. Poblaciones de este tipo serán tratadas como poblaciones infinitas.

Ejemplo 33. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, si la cantidad total de sismos N no es conocida basta para considerarla población infinita.

Ejemplo 34. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, si la cantidad total de estudiantes N no es conocida basta para considerarla población infinita.

POBLACIÓN OBJETIVO. Es el conjunto de objetos, fenómenos o personas que comparten la característica de interés y, está delimitada por los criterios de inclusión.

Ejemplo 35. Si son todos y cada uno de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, la población objetivo queda delimitada por los criterios de inclusión que son los sismos, la alcaldía de Coyoacán, el año en que ocurrieron y la magnitud a registrar.

Ejemplo 36. Si son todos y cada uno de los estudiantes que estudiaron la licenciatura en Salud a nivel universitario en la Ciudad de México en el año 2008, la población objetivo queda delimitada por los criterios de inclusión que son los estudiantes que estudiaron la licenciatura en Salud, en la Ciudad de México, a nivel universitario y en el año 2008.

VARIABLES. Corresponden a las características de la población objetivo que pueden cambiar de valor entre las unidades de estudio o en una misma a través del tiempo.

Ejemplo 37. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, una variable es la magnitud del sismo medido en escala de Richter.

Ejemplo 38. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, una variable de las unidades de estudio, puede ser la asistencia a clase durante el año.

VARIABLES DE INTERVALO. La característica definitoria de este tipo de variable es que el valor 0 no implica la desaparición de la variable y existe una forma cuantitativa de medirle la cuantía. Las variables de intervalo son cuantitativas y pueden ser discretas o continuas.

Ejemplo 39. La variable tiempo es una variable de intervalo, ya que al momento de iniciar el conteo de los años transcurridos para una persona o sea su edad, que en nuestra cultura es al momento de nacer, no implica que el tiempo (edad) ha desaparecido. En nuestra cultura tenemos fijado el año 0 al momento de nacer, lo que no implica que la variable edad haya desaparecido, ya que no se consideran los nueve meses del embarazo de la madre.

VARIABLES NOMINALES. Es un tipo de variable que se refiere a características cualitativas de las unidades de estudio. Es la variable que proporciona la menor cantidad estadística de información.

Ejemplo 40. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, una variable nominal puede ser el nombre de la localidad en la que se presentó cada uno de los sismos.

Ejemplo 41. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, una variable nominal pueden ser los nombres de los(as) estudiantes.

VARIABLES ORDINALES. Es una variable cualitativa que se divide en valores apreciativos de magnitud de mayor a menor con la cantidad de categorías que se requieran, pero que no es posible determinar cuantitativamente la magnitud de cada una. El investigador crea las categorías que forman la variable ordinal, pudiendo codificarlas con números, pero generalmente no es factible realizar cálculos estadísticos puesto que no proporcionan nada útil. Sin embargo, existe una excepción, la escala de Likert, que establece cinco categorías cualitativas numeradas, conforme se tenga o no acuerdo, satisfacción u otro concepto que pueda dividirse en categorías descendentes, con una proposición cualitativa. Con ellas sí se pueden realizar cálculos estadísticos. La escala puede ser: 5 =totalmente de acuerdo, totalmente satisfecho... El número 5 indica el máximo grado obtenido, ya

sean los planteados u otros y a partir del cual la magnitud de la categoría disminuye, así tenemos, 4 = de acuerdo, satisfecho..., 3 = no sabe, esta categoría señala el cambio total de una declaración cualitativa a otra, 2 = en desacuerdo, insatisfecho... y 1 = totalmente en desacuerdo, totalmente insatisfecho.... El número uno indica el mínimo grado respecto del valor 5.

Ejemplo 42. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, una variable ordinal puede ser la opinión de las personas sobre la intensidad del sismo descrita como: intensidad muy fuerte, intensidad fuerte, intensidad media, intensidad baja o intensidad muy baja.

Ejemplo 43. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México, en el año 2008, una variable ordinal puede ser la opinión de las unidades de estudio que tengan del desempeño docente en cierta materia del profesor(a), así las categorías podrían ser: muy buen docente, buen docente, mal docente y muy mal docente.

VARIABLES DE RAZÓN. En este tipo de variable la característica definitoria es que el 0 de valor sí implica la desaparición de la variable y existe una forma cuantitativa de medirle la cuantía. Las variables de razón son cuantitativas y pueden ser discretas o continuas.

Ejemplo 44. En el registro de los sismos ocurridos en la alcaldía de Coyoacán en el año 2000, mayores a una magnitud de 4 en la escala de Richter, una variable de razón es la magnitud del sismo medido por una escala especializada como la de Richter, esta escala considera por lo menos una cifra decimal.

Ejemplo 45. En la medición o registro de los estudiantes que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, una variable de razón puede ser la circunferencia de cintura (CCI) en centímetros (cm) de los o las estudiantes.

2.- Base de datos

La base de datos es parte del análisis estadístico, es indispensable que sea confiable, lo que se consigue al satisfacer, al menos, que:

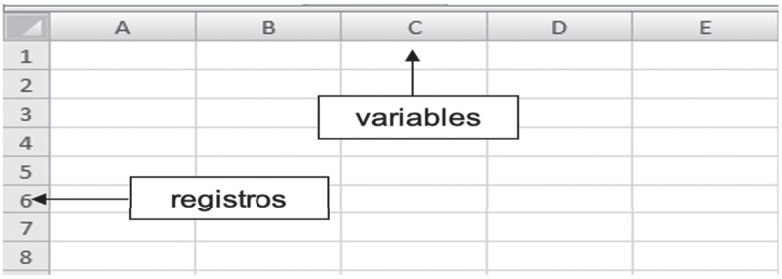
1. No contenga ningún tipo de error introducido al momento de su realización.
2. La información obtenida, en la etapa de recolección de información, sea congruente con los objetivos del trabajo.
3. Los objetivos del trabajo sean congruentes con el desarrollo del trabajo.
4. El contenido del instrumento para la obtención de información (hoja de registro, hoja de mediciones, cuestionario, etc.) sea congruente con los objetivos e hipótesis, si la hay.
5. La información a recolectar sea precisa y exacta.
6. El rigor aplicado sea muy alto, por parte del personal encargado de la obtención de la información.
7. El rigor aplicado sea muy alto por parte del control y elaboración de la base de datos.

En caso de una base de datos no confiable (dudosa), los análisis estadísticos, las gráficas, presentaciones y las conclusiones a las que se llegue, derivadas de dicha base, al menos serán sospechosas, si no falsas. El apoyarse en computadoras o gráficas, el análisis estadístico no mejora la calidad de datos erróneos.

El Excel es el programa de hoja de cálculo más común, en él se puede realizar una base de datos de muy buena calidad independientemente del área o ámbito del proyecto de trabajo.

En la hoja de cálculo, en la parte superior las columnas están codificadas alfabéticamente y en extremo izquierdo están numerados progresivamente los renglones.

En las columnas codificadas alfabéticamente estarían las variables provenientes de la información a recabar, en tanto que, en los renglones numerados progresivamente estarían los sujetos u objetos de los cuales se determinan todas las variables. En consecuencia, una base de datos consta, en este caso, de variables (columnas) y registros (renglones). Las unidades de estudio están registradas en los renglones.



La figura anterior muestra una fracción de la hoja de cálculo de Excel

Ejemplo 46. En el registro de sismos ocurridos en la Ciudad de México en el 2016 de 1 grado Richter o mayor, la base de datos podría ser de localidad, año, magnitud y hora local. Una fracción de la base de datos puede ser:

variables

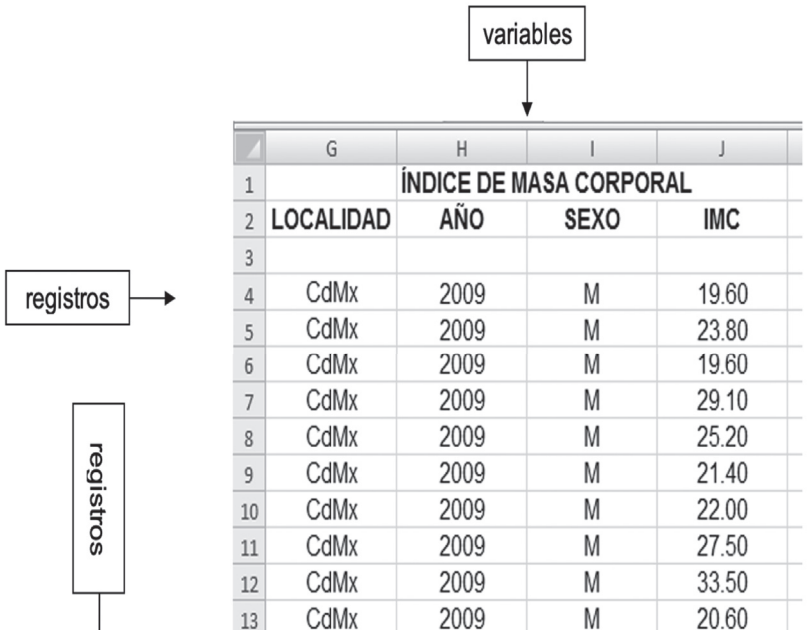
registros

registros

	A	B	C	D
1	SISMOS			
2	LOCALIDAD	AÑO	MAGNITUD	HORA LOCAL
3				
4	CdMx	2016	3.60	23.29
5	CdMx	2016	3.30	22.52
6	CdMx	2016	3.30	22.47
7	CdMx	2016	3.80	22.26
8	CdMx	2016	3.40	22.14
9	CdMx	2016	3.50	21.57
10	CdMx	2016	3.70	21.27
11	CdMx	2016	3.20	21.13
12	CdMx	2016	3.50	20.41
13	CdMx	2016	3.60	20.40

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

Ejemplo 47. En el registro del índice de masa corporal (IMC) para mujeres (M) de la Ciudad de México en 2009, la base de datos podría ser localidad, año, sexo (M) e IMC. Una fracción de la base de datos puede ser:



	G	H	I	J
1		ÍNDICE DE MASA CORPORAL		
2	LOCALIDAD	AÑO	SEXO	IMC
3				
4	CdMx	2009	M	19.60
5	CdMx	2009	M	23.80
6	CdMx	2009	M	19.60
7	CdMx	2009	M	29.10
8	CdMx	2009	M	25.20
9	CdMx	2009	M	21.40
10	CdMx	2009	M	22.00
11	CdMx	2009	M	27.50
12	CdMx	2009	M	33.50
13	CdMx	2009	M	20.60

Fuente: Datos modificados del software estadístico statdisk 13.

3.- Tamaño de muestra

Cuando se quiere hacer inferencia de una Muestra a una Población o Población Objetivo (estos términos se emplearán indistintamente) es necesario que la muestra sea probabilística, para lo que es inevitable calcular el tamaño de esta con cierto nivel de confianza y margen de error aceptable.

Es posible determinar el tamaño de la muestra para la proporción o la media, poblacional si esta es finita o infinita. El nivel de confianza es una probabilidad que desea el investigador y comúnmente es 0.90, 0.95 o 0.99 que depende de que tan estricto se deseé hacer el cálculo. Sin embargo, puede ser cualquier valor entre 0 y 1. Al especificar la probabilidad el valor únicamente puede ser en el intervalo mencionado, en cambio, al referirse al nivel de confianza la probabilidad se multiplica por 100 que resulta en 90%, 95% y 99%.

Tamaño de Muestra para una Proporción

Al determinar la proporción (*p*) de una Población, se tiene que varía entre 0 y 1. Es importante estimarla a partir de un estudio previo o uno de una muestra preliminar, tan grande como sea posible. Sin embargo, existe la posibilidad de que no sea posible la estimación previa de *p*, por lo que se modifica la ecuación. La muestra siempre es un entero, si no lo es se redondea al entero superior más cercano.

Tamaño de muestra con proporción poblacional estimada y población infinita

Donde:

n = tamaño de muestra

Z_α= obtenida de las tablas de probabilidad normal

$\hat{p}$ = estimación de la proporción poblacional $\hat{p} = \frac{x}{n}$

$\hat{q} = 1 - \hat{p}$

$E = Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$ la n es de la muestra de la que se obtuvo $\hat{p}$

Nivel de Confianza (%)		Valor
99	0.005	2.575
95	0.025	1.960
90	0.050	1.645

Tamaño de muestra con proporción poblacional estimada y población finita

$$n = \frac{N\hat{p}\hat{q}\left[Z_{\alpha/2}\right]^2}{\left[Z_{\alpha/2}\right]^2\hat{p}\hat{q} \ + \ (N-1)E^2}$$

Donde:

n = tamaño de muestra

N = tamaño de la población

Z_α= obtenida de las tablas de probabilidad normal

$\hat{p}$ = estimación de la proporción poblacional (p)

$\hat{q} \ = 1 - \hat{p}$

$E = Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}}$ (n es de la muestra en que se estimó)

Tamaño de muestra con proporción poblacional desconocida y población infinita

$$n = \frac{\left[Z_{\alpha/2}\right]^2 \cdot 0.25}{E^2}$$

Donde:

n = tamaño de muestra

Z_α= obtenida de las tablas de probabilidad normal

E = valor aceptable del error 0 < E <1

Tamaño de muestra con proporción poblacional desconocida y población finita

Donde:

n = tamaño de muestra

N = tamaño de la población

σ^2 = varianza poblacional

Z_α= obtenida de las tablas de probabilidad normal

E = error aceptable en las mismas unidades de la media investigada

Cálculo del Tamaño de Muestra para una Proporción Poblacional Desconocida y Población Infinita

Ejemplo 48, En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, el tamaño de muestra para determinar la proporción de sismos superiores a 3 grados Richter con una confianza del 95% y un error de *E* = 0.02 es:

$$n = \frac{[Z_{\alpha/2}]^2 \cdot 0.25}{E^2} = \frac{[1.96]^2 \cdot 0.25}{0.02^2} = 2{,}401$$

Entonces, el tamaño de muestra para determinar la proporción de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, superiores a 3 grados Richter con una confianza del 95% y un error de *E* = 0.02 es de *n* = 2,401 sismos.

Ejemplo 49. En el registro de estudiantes que estaban inscritos en la licenciatura en Salud en la Ciudad de México en el año 2008, el tamaño de la muestra para determinar la proporción de mujeres, con una confianza del 99% y un error de *E* = 0.05 es:

$$n = \frac{[Z_{\alpha/2}]^2 \cdot 0.25}{E^2} = \frac{[2.575]^2 \cdot 0.25}{0.05^2} = 664$$

Entonces, el tamaño de muestra para determinar la proporción poblacional de las mujeres inscritas en la licenciatura en Salud en la Ciudad de México en el año 2008, con una confianza del 99% y un error de *E* = 0.05 es de *n* = 664 estudiantes.

Tamaño de Muestra para Determinar la Media Poblacional

Al determinar la media (μ) de una Población, es importante conocer la desviación estándar poblacional (σ).

Tamaño de muestra en la que σ es conocida y la población es infinita

$$n = \left[\frac{\sigma Z_{\alpha/2}}{E}\right]^2$$

Donde:

n = tamaño de muestra

σ = desviación estándar poblacional

Z_α= obtenida de las tablas de probabilidad normal

E = error aceptable en las mismas unidades de la media investigada

Tamaño de muestra en la que σ es conocida y la población es finita

$$n = \frac{N\sigma^2\left[Z_{\alpha/2}\right]^2}{(N-1)E^2 + \sigma^2\left[Z_{\alpha/2}\right]^2}$$

Donde:

n = tamaño de muestra

N = tamaño de la población

σ^2 = varianza poblacional

Z_α= obtenida de las tablas de probabilidad normal

E = error aceptable en las mismas unidades de la media investigada

Cuando no sea factible contar con σ, se reemplaza en las ecuaciones para población infinita o finita, por con n - 1 grados de libertad, y se busca en las tablas de las probabilidades de esa distribución.

Cálculo del Tamaño de Muestra para Determinar la Media Poblacional, Desviación Estándar Poblacional Conocida y Población Infinita

Ejemplo 50. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, el tamaño de muestra para determinar la media poblacional de los sismos superiores a 3 grados Richter con grados Richter, una confianza del 95% y un error aceptable de E = 0.03.

$$n = \left[\frac{\sigma Z_{\alpha/2}}{E}\right]^2 = \left[\frac{0.25 \cdot 1.96}{0.03}\right]^2 = 266.77778$$

El resultado no es un entero, por lo cual se redondea al entero inmediato superior.

n = 267 sismos

Entonces, el tamaño de muestra para determinar la media poblacional de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, superiores a 3 grados Richter con grados Richter, una confianza del 95% y un error aceptable de E = 0.03 grados Richter es n = 267 sismos.

Ejemplo 51. En el registro de estudiantes que estaban inscritos en la licenciatura en Salud en la Ciudad de México en el año 2008, el tamaño de la muestra para determinar la media de la circunferencia de cintura con cm, una confianza de 99% y un error aceptable de E = 5.00 cm es:

$$n = \left[\frac{\sigma Z_{\alpha/2}}{E}\right]^2 = \left[\frac{15.40 \cdot 2.575}{5}\right]^2 = \left[\frac{39.655}{5}\right]^2 = 62.900761$$

El resultado no es un entero por lo cual se redondea al entero inmediato superior.

$$n = 63 \text{ estudiantes}$$

Entonces, el tamaño de muestra para determinar la media de la circunferencia de cintura en los estudiantes inscritos en la licenciatura en Salud en la Ciudad de México, en el año 2008 con cm, una confianza del 99% y un error aceptable de $E = 5.00$ cm es de $n = 63$ estudiantes.

4.- Tablas de frecuencias no agrupadas

La tabla de frecuencia no agrupada es para variables cualitativas: nominal u ordinal. Los conteos de cada una de las categorías de las variables son datos discretos y queda conformada, de izquierda a derecha, por las siguientes columnas:

Cada una de las categorías registradas.

Frecuencia absoluta (F_i) de cada una de las categorías: F_i = total del conteo de cada evento en cada categoría. La sumatoria debe ser igual al tamaño de la muestra:

Donde:

n = tamaño de la muestra

$\sum$ = sumatoria de todas y cada una de las F_i

i = contador que va de i =1 a i = n

x_i = cada dato de la muestra

Frecuencia Relativa (Fr_i) de cada categoría: $Fr_i = Fi/n$. Esta variable lleva dos dígitos a la derecha del punto. La suma de las Fr_i debe ser 1. Sí la diferencia es de milésimas o menos es correcto.

Frecuencia Relativa Porcentual ($\%Fr_i$) de cada categoría: $\%Fr_i = Fr_i * 100$. La suma de $\%Fr_i$ debe ser 100 . Sí la diferencia es de milésimas o menos es correcto.

Frecuencia Relativa Porcentual Acumulada ($A\%Fr_i$) de cada categoría: $A\%Fr_i = \%Fr_i + \%Fr_i+1$. El total 100% se localiza en la última categoría, ya que se calcula el porcentaje acumulado hasta cada categoría.

Tabla de Frecuencia No Agrupada para Variable Nominal

Ejemplo 52. Al registrar los sismos ocurridos en las alcaldías de la Ciudad de México en el año 2017, de magnitud 1 o mayor, la tabla de frecuencias no agrupadas, para una variable nominal, puede estar constituida por la cantidad de sismos que se presentaron en cada alcaldía.

Sismos Ocurridos en Alcaldías de la Ciudad de México en 2017				
Alcaldías por orden alfabético	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$
Álvaro Obregón	1	0.04	4	4
Benito Juárez	3	0.12	12	16
Coyoacán	11	0.44	44	60
Cuajimalpa de Morelos	3	0.12	12	72

Sismos Ocurridos en Alcaldías de la Ciudad de México en 2017				
Alcaldías por orden alfabético	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$
La Magdalena Contreras	2	0.08	8	80
Tlalpan	5	0.20	20	100
Totales	25	1	100	

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

Ejemplo 53. Al realizar un inventario de medicinas en una farmacia de la Ciudad de México en febrero del 2005, se podría tener el siguiente registro.

Medicamentos en la Bodega de una Farmacia de la Ciudad de México en 2005				
Medicamentos por orden alfabético	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$
Aspirina	66	0.23	23	23
Captotril	24	0.08	8	31
Cymbalta	38	0.13	13	44
Diclofenaco	15	0.05	5	49
Lop	26	0.09	9	58
Loratadina	48	0.17	17	75
Metformina	16	0.06	6	81
Odivitor	32	0.11	11	92
Trayenta duo	24	0.08	8	100
Totales	289	1	100	

Tabla de Frecuencia No Agrupada para Variable Ordinal

Ejemplo 54. Al registrar los sismos ocurridos en las alcaldías de la Ciudad de México en el año 2017, de magnitud 1 o mayor, la tabla de frecuencias no agrupadas, para una variable ordinal, puede estar constituida por el tipo de opinión de las personas, relativa a la intensidad de los sismos ocurridos.

Opinión Sobre la Intensidad de los Sismos Ocurridos en Alcaldías de la Ciudad de México en 2017				
Opinión sobre la intensidad del sismo	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$
Muy fuerte	2	0.08	8	8
Fuerte	4	0.16	16	24
Medianamente fuerte	12	0.48	48	72
Poco fuerte	4	0.16	16	88
Nada fuerte	3	0.12	12	100
Totales	25	1	100	

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

Ejemplo 55. En el caso de la opinión sobre la efectividad del ejercicio personal para tener buena salud en la Ciudad de México en 2009, se podrían tener los siguientes resultados:

Opinión Sobre la Efectividad del Ejercicio para una Buena Salud en la Ciudad de México en 2009				
Opinión sobre el ejercicio para tener buena salud	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$
Muy efectivo	66	0.23	23	23
Efectivo	96	0.33	33	56
Medianamente efectivo	38	0.13	13	69
poco efectivo	31	0.11	11	80
nada efectivo	58	0.20	20	100
Totales	289	1	100	

5.- Tablas de frecuencias agrupadas

La tabla de frecuencia agrupada es para variables cuantitativas: de razón o de intervalo. Por lo general se usan cuando n 30. Se espera tener cuando menos 5 intervalos de clase. Los conteos para cada intervalo de clase son datos discretos. Los intervalos están delimitados por su límite real inferior y el límite superior por datos continuos. El intervalo de clase menor y el mayor deben contener a los valores menor y mayor de los datos, respectivamente, esto es, la tabla debe incluir a todos los datos.

Para la construcción de la tabla de datos agrupados, primero se debe determinar el número de intervalos de clase, la amplitud de cada uno de ellos y sus límites reales inferior y superior.

Determinación del número de intervalos (K)

Regla de Sturges $K = 1+ 3.322(\log_{10} n)$

Donde:
K = Número de intervalos de clase
n = tamaño de la muestra

En caso de que K tenga fracciones, se redondea al entero superior inmediato.

Determinación de la amplitud de cada intervalo de clase (A_i)

$A_i = R/ k$

Donde:
R (rango) = Valor mayor de la muestra - Valor menor de la muestra

Determinación de los límites inferiores (LI) y superiores (LS)

El límite inferior de la tabla es el valor menor de todos los datos y una vez seleccionado se calculan todos los límites inferiores (LI), de acuerdo con la amplitud redondeada al entero superior inmediato si la cifra decimal es 0.5, o al dígito a la derecha del punto que indica las decimales, según el tipo de los datos que se tengan. El intervalo de clase (IC) es la distancia entre un límite inferior (LI) y el LI inmediato superior.

A continuaciónse calculan los límites superiores (LS). El LS de la clase 1 se calcula restando la unidad de redondeo al límite inferior (LI) del intervalo 2, una vez realizado se calculan los demás límites superiores, sumándole a cada uno la amplitud del intervalo ya calculada.

La unidad de redondeo es determinada de acuerdo con los datos originales.

Sí son enteros = 1
Sí llegan hasta decimales = 0.1
Sí llegan hasta centésimas = 0.01
Sí llegan a milésimas = 0.001

Y así consecutivamente dependiendo cuántas cifras significativas tienen los datos originales y cada intervalo de clase debe cumplir con ser excluyente, exhaustivo y la frecuencia 1.

Determinación de los límites reales inferiores (LRI) y superiores (LRS)

El cálculo de los límites reales inferiores (LRI), se hace restando la mitad de la unidad de redondeo a cada uno de los límites inferiores.

El cálculo de los límites reales superiores (LRS), se hace sumando la otra mitad de la unidad de redondeo a cada uno de los límites superiores.

En consecuencia, la tabla de datos agrupados queda conformada, de izquierda a derecha, al menos por las siguientes columnas:

Número consecutivo de los intervalos de clase.

Cada uno de los intervalos de clase (límite inferior real y límite superior real) hasta llegar a K. Cada uno tiene la misma amplitud (Ai). Se ordenan del menor al mayor y cada uno debe ser excluyente (no debe haber duda de a que intervalo de clase pertenece cada uno), exhaustivo (debe contener todos los datos cuya magnitud este entre el límite inferior y el límite superior) y su frecuencia absoluta debe ser mayor a cero.

Frecuencia absoluta (F_i) de cada uno de los intervalos de clase: F_i = total del conteo de cada evento en cada intervalo de clase (*i*). La sumatoria debe ser igual al tamaño de la muestra.

$$n = \sum_{i=1}^{n} Fi.$$

Donde:
n = tamaño de la muestra
$\sum$ = sumatoria de todas y cada una de las F_i
= contador que va de =1 a = n
x_i = cada dato de la muestra

Frecuencia Relativa (Fr_i) de cada uno de los intervalos de clase: $Fr_i = F_i/n$. Esta variable lleva dos o más valores a la derecha del punto. La suma de las Fr_i debe ser 1.

Frecuencia Relativa Porcentual ($\%Fr_i$) de los intervalos de clase: $\%Fr_i = Fr_i * 100$. La suma de $\%Fr_i$ debe ser 100.

Frecuencia Relativa Porcentual Acumulada ($A\%Fr_i$) de los intervalos de clase: $A\%Fr_i = \%Fr_i + \%Fr_i+1$. El total 100% se localiza en el último intervalo de clase, ya que se calcula el porcentaje acumulado hasta cada categoría.

Marca de Clase (MC) = (Límite Superior Real + Límite Inferior Real)/2.

Elaboración de la Tabla de Frecuencia Agrupada

Ejemplo 56. En el caso del registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019 de magnitud 1 o más en la escala de Richter, la tabla de frecuencias agrupadas para una variable de razón que en este caso es la magnitud de los sismos. Magnitud de 89 sismos registrados:

No.	Magnitud
1	1.0
2	1.0
3	1.3
4	1.3
5	1.3
6	1.4
7	1.5
8	1.5
9	1.5
10	1.5
11	1.5
12	1.5
13	1.7
14	1.7
15	1.7
16	1.7
17	1.7
18	1.7
19	1.7
20	1.8
21	1.8
22	1.8
23	1.8
24	1.8
25	1.8
26	1.9
27	1.9
28	1.9
29	1.9
30	1.9

No.	Magnitud
31	2.0
32	2.0
33	2.0
34	2.0
35	2.0
36	2.0
37	2.1
38	2.1
39	2.1
40	2.1
41	2.1
42	2.1
43	2.1
44	2.2
45	2.2
46	2.2
47	2.2
48	2.2
49	2.2
50	2.2
51	2.3
52	2.3
53	2.3
54	2.3
55	2.3
56	2.4
57	2.4
58	2.4
59	2.4
60	2.4

No.	Magnitud
61	2.4
62	2.4
63	2.4
64	2.4
65	2.5
66	2.5
67	2.5
68	2.5
69	2.5
70	2.5
71	2.6
72	2.6
73	2.6
74	2.6
75	2.7
76	2.7
77	2.7
78	2.7
79	2.7
80	2.9
81	2.9
82	3.0
83	3.3
84	3.3
85	3.5
86	3.6
87	4.0
88	4.1
89	4.2

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

Determinación del número de intervalos

Regla de Sturges $K = 1 + 3.322(\log_{10} 89) = 7.475874 \quad 8$

Determinación de la amplitud de cada intervalo de clase (A_i)

$A_i = R/ k = (4.2 - 1)/8 = 0.40$

Determinación de los límites inferiores (LI) y superiores (LS)

IC	LI	LS
1	1.0	1.30
2	1.4	1.70
3	1.8	2.10
4	2.2	2.50
5	2.6	2.90
6	3.0	3.30
7	3.4	3.70
8	3.8	4.10
9	4.2	4.50

Se agrega un intervalo de clase para incluir el sismo cuya magnitud es la mayor, y como los datos originales llegan a décimas la unidad de redondeo es = 0.1.

Determinación de los límites reales inferiores (LRI) y superiores (LRS)

Resta de 1/2 de la unidad de redondeo = 0.05 a cada límite inferior

Suma de 1/2 de la unidad de redondeo = 0.05 a cada límite superior

IC	LS	LRS
1	1.30	1.35
2	1.70	1.75
3	2.10	2.15
4	2.50	2.55
5	2.90	2.95
6	3.30	3.35
7	3.70	3.75
8	4.10	4.15
9	4.50	4.55

Entonces, la tabla de frecuencia agrupada queda:

MAGNITUD DE LOS SISMOS OCURRIDOS EN LA CIUDAD DE MÉXICO DE 2016 A 2019							
IC	LRI	LRS	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$	MC
1	0.95	1.35	5	0.0562	5.62	5.62	1.15
2	1.35	1.75	14	0.1573	15.73	21.35	1.55
3	1.75	2.15	24	0.2697	26.97	48.31	1.95
4	2.15	2.55	27	0.3034	30.34	78.65	2.35
5	2.55	2.95	11	0.1236	12.36	91.01	2.75
6	2.95	3.35	3	0.0337	3.37	94.38	3.15
7	3.35	3.75	2	0.0225	2.25	96.63	3.55
8	3.75	4.15	2	0.0225	2.25	98.88	3.95
9	4.15	4.55	1	0.0112	1.12	**100**	4.35
		Total	89	1	100		

Ejemplo 57. En el caso del registro de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, la tabla de frecuencias agrupadas para una variable de razón, que en este caso son las CCI.

Datos de la circunferencia de cintura (CCI) de 43 mujeres:

No.	CCI
1	66.70
2	67.20
3	67.70
4	67.80
5	68.60
6	68.70
7	69.10
8	70.00
9	72.70
10	72.90
11	73.60
12	74.50
13	74.50
14	74.50
15	75.40

No.	CCI
16	75.50
17	75.90
18	78.80
19	79.50
20	81.40
21	82.50
22	82.60
23	85.00
24	85.70
25	85.70
26	91.10
27	92.80
28	92.80
29	93.00
30	94.00

No.	CCI
31	95.50
32	98.00
33	99.30
34	99.40
35	100.70
36	104.70
37	105.50
38	105.50
39	112.80
40	115.30
41	118.90
42	126.00
43	126.50

Fuente: Datos modificados del software estadístico statdisk 13.

Determinación del número de intervalos

Regla de Sturges K = 1+ 3.322($\log_{10}$ 43) = 6.42638221 7

***Determinación de la amplitud de cada intervalo de clase (*Ai)**

Ai = R/ k = (126.5 - 66.7)/7 = 8.54285714 9

Determinación de los límites inferiores (LI) y superiores (LS)

IC	LI	LS
1	66.70	75.60
2	75.70	84.60
3	84.70	93.60
4	93.70	102.60
5	102.70	111.60
6	111.70	120.60
7	120.70	129.60

Como los datos originales llegan a décimas la unidad de redondeo es = 0.1

Determinación de los límites reales inferiores (LRI) y superiores (LRS)

Resta de 1/2 de la unidad de redondeo = 0.05 a cada límite inferior

IC	LI	LRI
1	66.70	66.65
2	75.70	75.65
3	84.70	84.65
4	93.70	93.65
5	102.70	102.65
6	111.70	111.65
7	120.70	120.65

Suma de 1/2 de la unidad de redondeo = 0.05 a cada límite superior

IC	LS	LRS
1	75.60	75.65
2	84.60	84.65
3	93.60	93.65
4	102.60	102.65
5	111.60	111.65
6	120.60	120.65
7	129.60	129.65

Entonces, la tabla de frecuencia agrupada queda:

CIRCUNFERENCIA DE CINTURA DE MUJERES QUE ESTUDIARON LA LICENCIATURA EN SALUD EN LA CIUDAD DE MÉXICO EN 2008							
IC	LRI	LRS	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$	MC
1	66.65	75.65	16	0.37	37	37	71.15
2	75.65	84.65	6	0.14	14	51	80.15
3	84.65	93.65	7	0.16	16	67	89.15
4	93.65	102.65	6	0.14	14	81	98.15
5	102.65	111.65	3	0.07	7	88	107.15
6	111.65	120.65	3	0.07	7	95	116.15
7	120.65	129.65	2	0.05	5	**100**	125.15
		Total	43	1	100		

6.- Gráficas

Las gráficas de columnas son para variables cualitativas: nominal y ordinal. El histograma, el polígono de frecuencias y la ojiva para variables cuantitativas de razón o de intervalo ya sean continuas o discretas.

El eje horizontal para variables cualitativas no es numérico, estarían cada uno de los nombres de las categorías correspondientes. En las variables cuantitativas, el eje horizontal si es numérico y se le llamará “X”.

En la gráfica de columnas el eje “Y” puede ser una de las siguientes tres opciones, las Frecuencias Absolutas (F_i), las Frecuencias relativas (Fr_i) o la Frecuencia Relativa Porcentual ($\%Fr_i$). En el eje horizontal estarían los nombres de las categorías correspondientes.

Frecuencia absoluta (F_i) de cada una de las categorías: F_i = Total del conteo de cada evento en cada categoría.

Frecuencia Relativa (Fr_i) de cada categoría: $Fr_i = F_i/n$.

Frecuencia Relativa Porcentual ($\%Fr_i$) de cada categoría: $\%Fr_i = Fr_i * 100$.

En el histograma también son válidas las tres opciones anteriores para el eje “Y”. Sin embargo, para mayor facilidad de interpretación y comparación con otra muestra se utiliza el $\%Fr_i$, especialmente si el tamaño de las muestras es diferente. En cambio, en el eje “X” estarían los Intervalos de Clase (IC).

En el polígono de frecuencias también son válidas las opciones mencionadas, pero el área bajo la curva cambia según cada una de ellas, al utilizar en el eje “Y” F_i el área encerrada por el polígono es n, si se usa $\%Fr_i$ el área encerrada equivale al 100% y con Fr_i es 1. Sin embargo, para su construcción a las Marcas de Clase (MC) existentes en la tabla de frecuencias agrupadas, se agrega, antes de la primera MC, una MC restándole a la misma la amplitud del Intervalo de Clase y a la última MC se le suma la amplitud del Intervalo de Clase. Al efectuar los ajustes indicados, el Polígono de Frecuencias queda cerrado en el eje “X”.

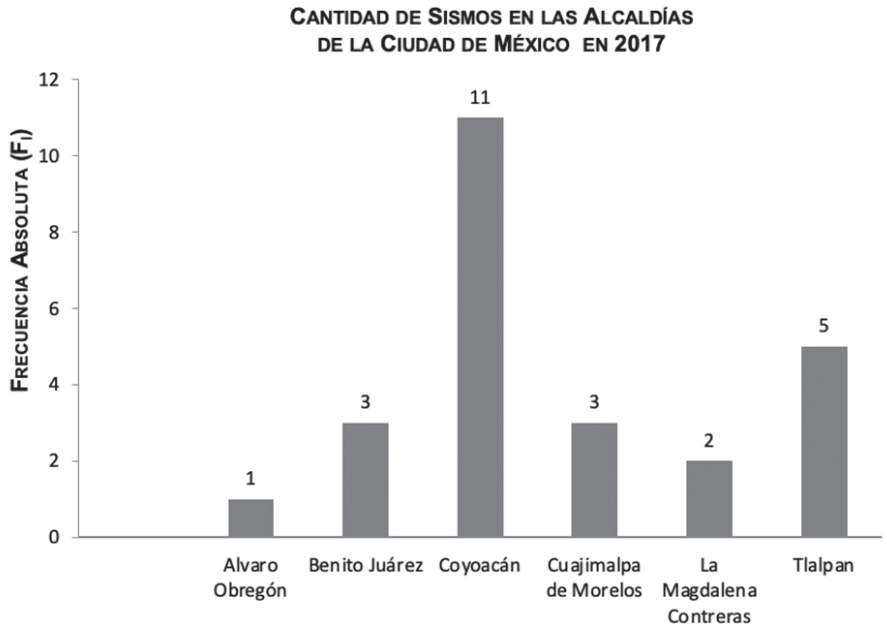
La gráfica de ojiva, que es acumulativa de las frecuencias, cualquiera de las tres opciones mencionadas sería válida para el eje “Y”. En tanto que en el eje “X” estarían los Límites Reales Inferiores (LRI).

Gráfica de Columnas para Variable Nominal

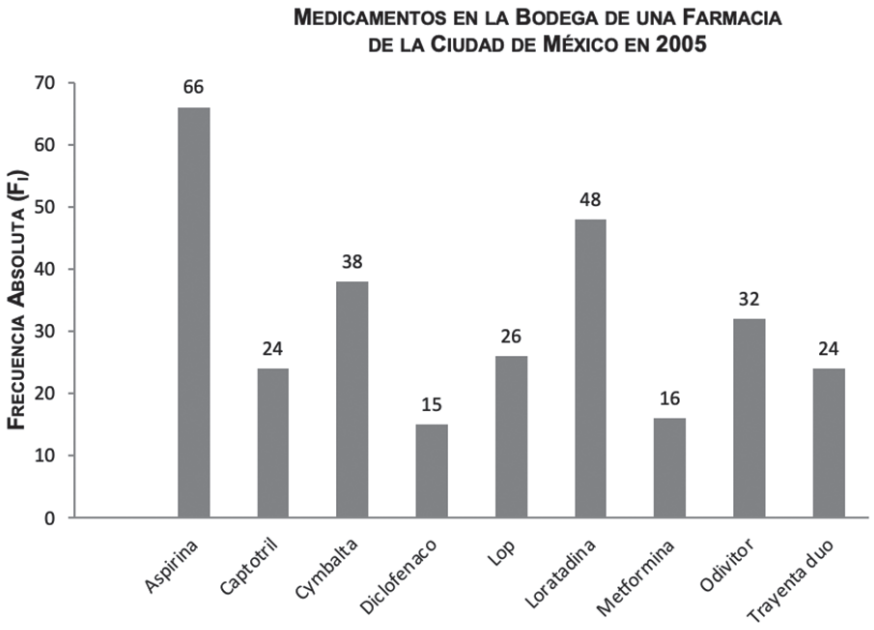
Ejemplo 58: La tabla de datos no agrupados proviene del ejemplo 52:

Sismos Ocurridos en Alcaldías de la Ciudad de México en 2017				
Alcaldías por orden alfabético	**F_i**	**Fr_i**	**$\%Fr_i$**	**$A\%Fr_i$**
Álvaro Obregón	1	0.04	4	4
Benito Juárez	3	0.12	12	16
Coyoacán	11	0.44	44	60
Cuajimalpa de Morelos	3	0.12	12	72
La Magdalena Contreras	2	0.08	8	80
Tlalpan	5	0.20	20	**100**
Totales	**25**	**1**	**100**	

Gráfica nominal para la Frecuencia Absoluta (F_i) de la tabla de frecuencias no agrupadas del ejemplo 52:



Ejemplo 59. La tabla de datos no agrupados proviene del ejemplo 53:



Medicamentos en la Bodega de una Farmacia de la Ciudad de México en 2005				
Medicamentos por orden alfabético	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$
Aspirina	66	0.23	23	23
Captotril	24	0.08	8	31
Cymbalta	38	0.13	13	44
Diclofenaco	15	0.05	5	49
Lop	26	0.09	9	58
Loratadina	48	0.17	17	75
Metformina	16	0.06	6	81
Odivitor	32	0.11	11	92
Trayenta duo	24	0.08	8	**100**
Totales	289	1	100	

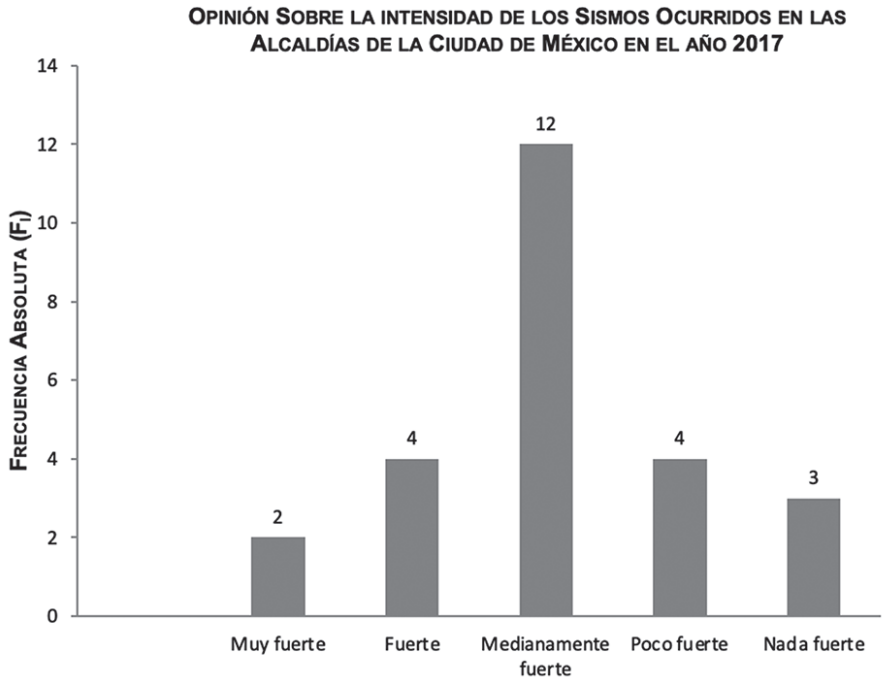
Gráfica nominal para la Frecuencia Absoluta (F_i) de la tabla de frecuencias no agrupadas del ejemplo 53:

Gráfica de Columnas para Variable Ordinal

Ejemplo 60. La tabla de datos no agrupados proviene del ejemplo 54:

Opinión Sobre la Intensidad de los Sismos Ocurridos en Alcaldías de la Ciudad de México en 2017				
Opinión sobre la intensidad del sismo	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$
Muy fuerte	2	0.08	8	8
Fuerte	4	0.16	16	24
Medianamente fuerte	12	0.48	48	72
Poco fuerte	4	0.16	16	88
Nada fuerte	3	0.12	12	**100**
Totales	25	1	100	

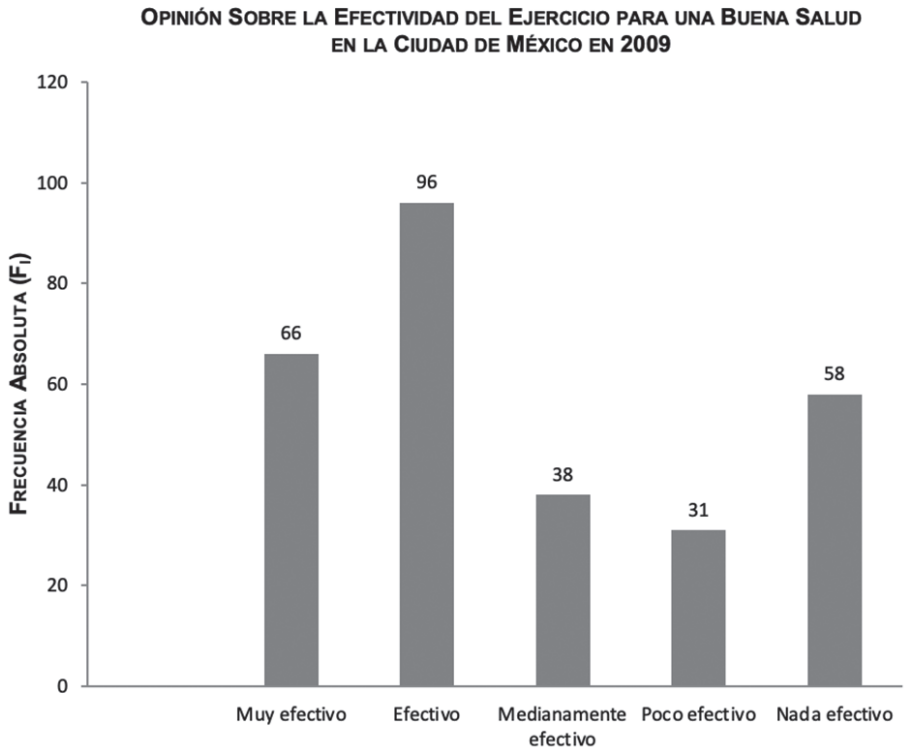
Gráfica ordinal para la Frecuencia Absoluta (F_I) de la tabla de frecuencias no agrupadas del ejemplo 54:



Ejemplo 61. La tabla de datos no agrupados proviene del ejemplo 55:

Opinión Sobre la Efectividad del Ejercicio para una Buena Salud en la Ciudad de México en 2009				
Opinión sobre el ejercicio para tener buena salud	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$
Muy efectivo	66	0.23	23	23
Efectivo	96	0.33	33	56
Medianamente efectivo	38	0.13	13	69
poco efectivo	31	0.11	11	80
nada efectivo	58	0.20	20	**100**
Totales	289	1	100	

Gráfica ordinal para la Frecuencia Absoluta (F_i) de la tabla de frecuencias no agrupadas del ejemplo 55:

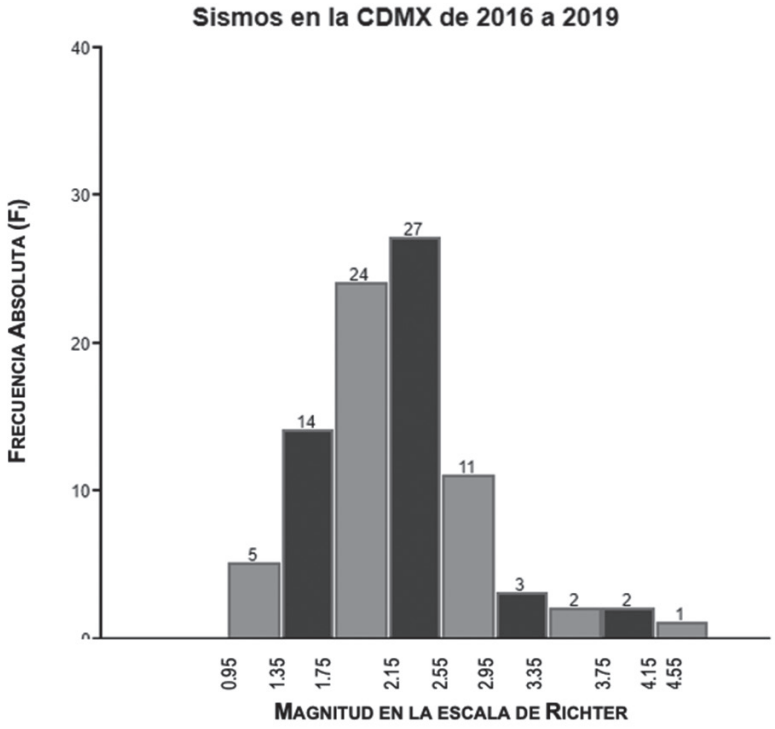


Histograma para una Variable de Razón

Ejemplo 62. La tabla de datos agrupados proviene del ejemplo 56:

Magnitud de los Sismos Ocurridos en la Ciudad de México de 2016 a 2019							
IC	LRI	LRS	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$	MC
1	0.95	1.35	5	0.0562	5.62	5.62	1.15
2	1.35	1.75	14	0.1573	15.73	21.35	1.55
3	1.75	2.15	24	0.2697	26.97	48.31	1.95
4	2.15	2.55	27	0.3034	30.34	78.65	2.35
5	2.55	2.95	11	0.1236	12.36	91.01	2.75
6	2.95	3.35	3	0.0337	3.37	94.38	3.15
7	3.35	3.75	2	0.0225	2.25	96.63	3.55
8	3.75	4.15	2	0.0225	2.25	98.88	3.95
9	4.15	4.55	1	0.0112	1.12	**100**	4.35
		Total	89	1	100		

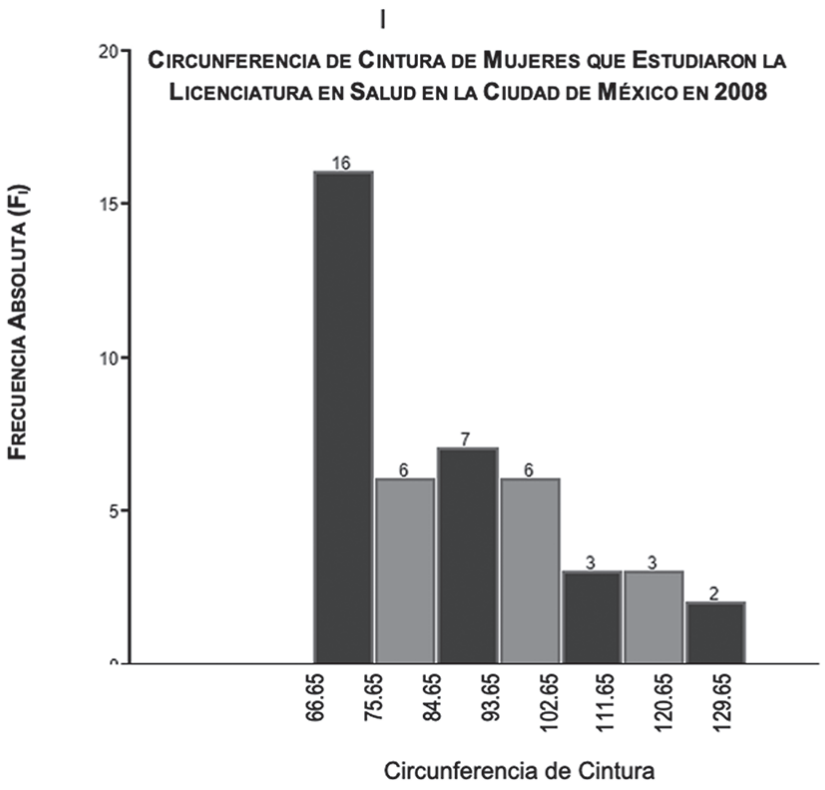
Histograma para la Frecuencia Absoluta (F_i) de la tabla de frecuencias agrupadas del ejemplo 56:



Ejemplo 63. La tabla de datos agrupados proviene del ejemplo 57:

Circunferencia de Cintura de Mujeres que Estudiaron la Licenciatura en Salud en la Ciudad de México en 2008							
IC	LRI	LRS	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$	MC
1	66.65	75.65	16	0.37	37	37	71.15
2	75.65	84.65	6	0.14	14	51	80.15
3	84.65	93.65	7	0.16	16	67	89.15
4	93.65	102.65	6	0.14	14	81	98.15
5	102.65	111.65	3	0.07	7	88	107.15
6	111.65	120.65	3	0.07	7	95	116.15
7	120.65	129.65	2	0.05	5	**100**	125.15
		Total	43	1	100		

Histograma para la Frecuencia Absoluta (F_i) de la tabla de frecuencias agrupadas del ejemplo 57:

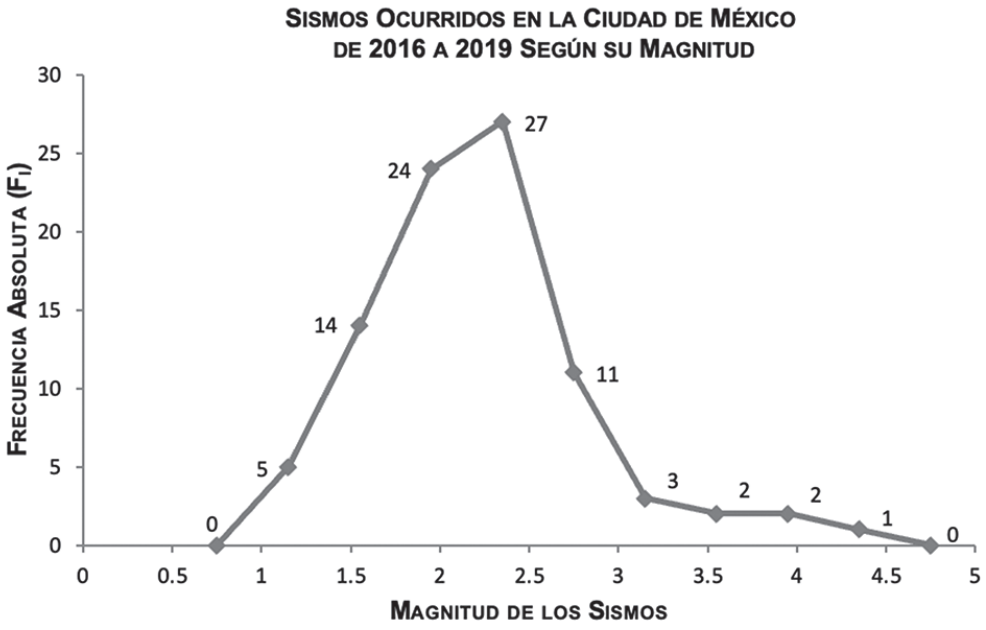


Polígono de Frecuencias para una Variable de Razón

Ejemplo 64. La tabla de datos agrupados proviene del ejemplo 56:

Magnitud de los Sismos Ocurridos en la Ciudad de México de 2016 a 2019							
IC	LRI	LRS	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$	MC
1	0.95	1.35	5	0.0562	5.62	5.62	1.15
2	1.35	1.75	14	0.1573	15.73	21.35	1.55
3	1.75	2.15	24	0.2697	26.97	48.31	1.95
4	2.15	2.55	27	0.3034	30.34	78.65	2.35
5	2.55	2.95	11	0.1236	12.36	91.01	2.75
6	2.95	3.35	3	0.0337	3.37	94.38	3.15
7	3.35	3.75	2	0.0225	2.25	96.63	3.55
8	3.75	4.15	2	0.0225	2.25	98.88	3.95
9	4.15	4.55	1	0.0112	1.12	**100**	4.35
		Total	89	1	100		

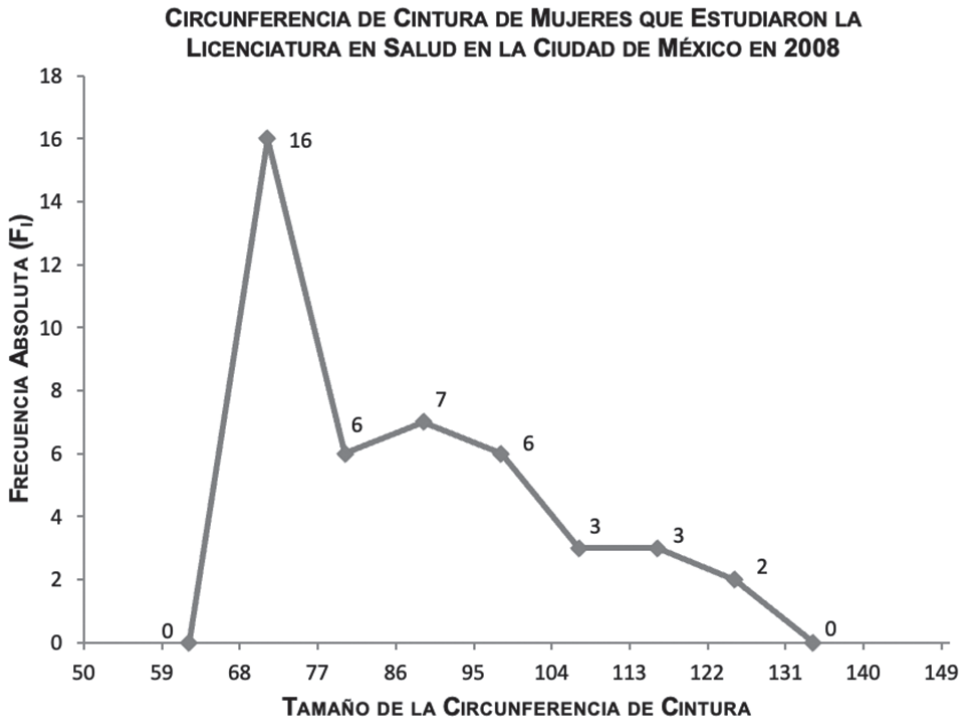
El Polígono de frecuencias para la Frecuencia Absoluta (F_i) de la tabla de frecuencias agrupadas del ejemplo 56:



Ejemplo 65. La tabla de datos agrupados proviene del ejemplo 57:

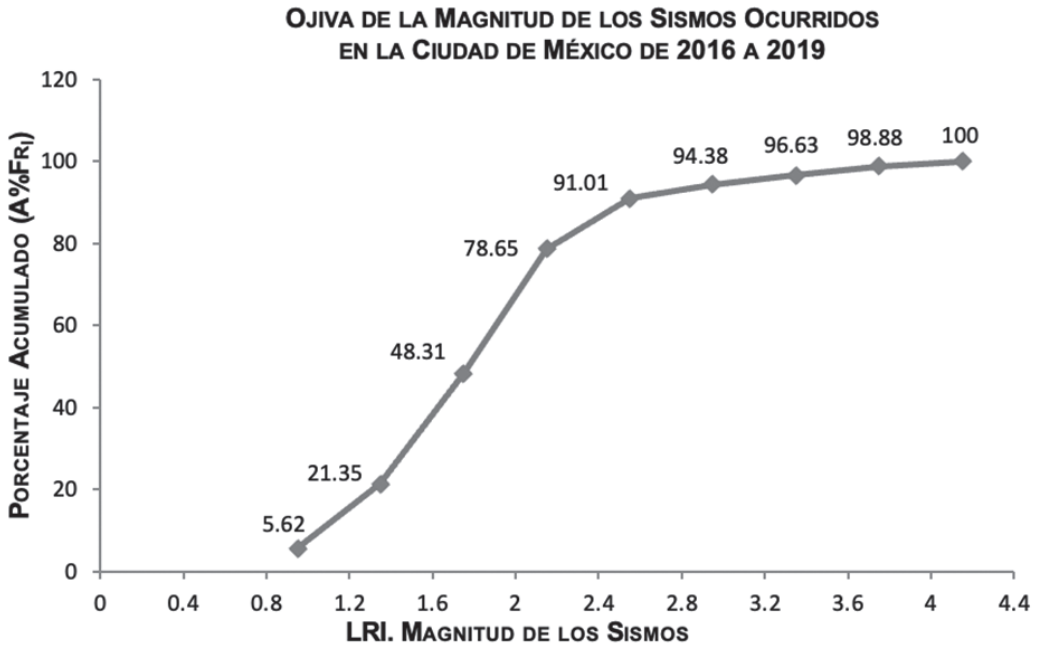
Circunferencia de Cintura de Mujeres que Estudiaron la Licenciatura en Salud en la Ciudad de México en 2008							
IC	LRI	LRS	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$	MC
1	66.65	75.65	16	0.37	37	37	71.15
2	75.65	84.65	6	0.14	14	51	80.15
3	84.65	93.65	7	0.16	16	67	89.15
4	93.65	102.65	6	0.14	14	81	98.15
5	102.65	111.65	3	0.07	7	88	107.15
6	111.65	120.65	3	0.07	7	95	116.15
7	120.65	129.65	2	0.05	5	**100**	125.15
		Total	43	1	100		

El Polígono de frecuencias para la Frecuencia Absoluta (F_i) de la tabla de frecuencias agrupadas del ejemplo 57:



Ojiva para una Variable de Razón

Ejemplo 66. La tabla de datos agrupados proviene del ejemplo 56:



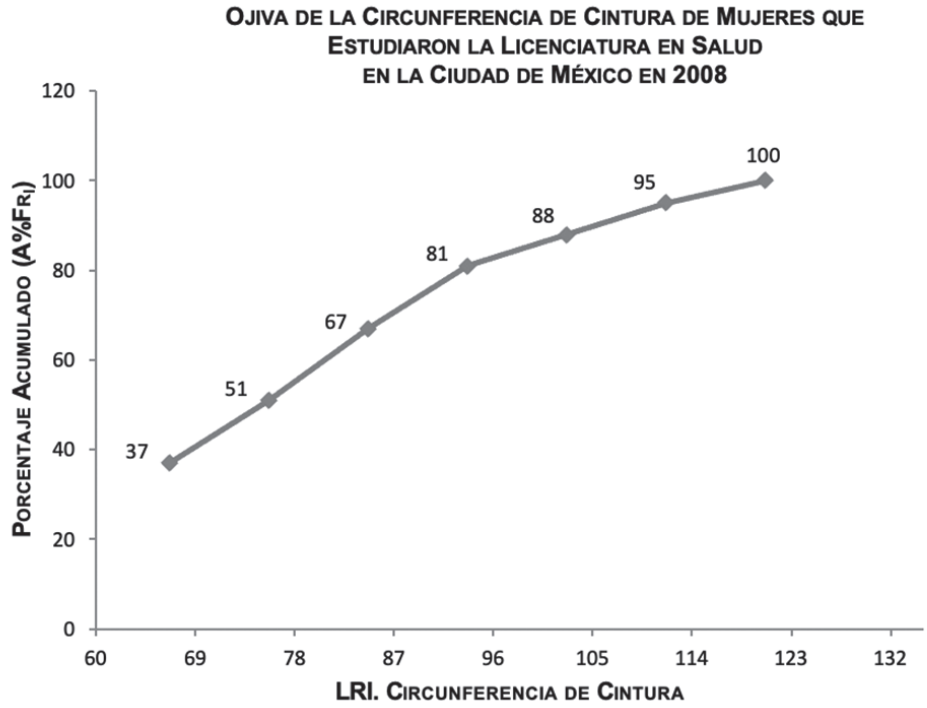
Magnitud de los Sismos Ocurridos en la Ciudad de México de 2016 a 2019							
IC	LRI	LRS	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$	MC
1	0.95	1.35	5	0.0562	5.62	5.62	1.15
2	1.35	1.75	14	0.1573	15.73	21.35	1.55
3	1.75	2.15	24	0.2697	26.97	48.31	1.95
4	2.15	2.55	27	0.3034	30.34	78.65	2.35
5	2.55	2.95	11	0.1236	12.36	91.01	2.75
6	2.95	3.35	3	0.0337	3.37	94.38	3.15
7	3.35	3.75	2	0.0225	2.25	96.63	3.55
8	3.75	4.15	2	0.0225	2.25	98.88	3.95
9	4.15	4.55	1	0.0112	1.12	**100**	4.35
		Total	89	1	100		

La ojiva para la frecuencia Acumulada del Porcentaje de la Frecuencia Relativa ($A\%Fr_i$) de la tabla de frecuencias agrupadas del ejemplo 56:

Ejemplo 67. La tabla de datos agrupados proviene del ejemplo 57:

Circunferencia de Cintura de Mujeres que Estudiaron la Licenciatura en Salud en la Ciudad de México en 2008							
IC	LRI	LRS	F_i	Fr_i	$\%Fr_i$	$A\%Fr_i$	MC
1	66.65	75.65	16	0.37	37	37	71.15
2	75.65	84.65	6	0.14	14	51	80.15
3	84.65	93.65	7	0.16	16	67	89.15
4	93.65	102.65	6	0.14	14	81	98.15
5	102.65	111.65	3	0.07	7	88	107.15
6	111.65	120.65	3	0.07	7	95	116.15
7	120.65	129.65	2	0.05	5	**100**	125.15
		Total	43	1	100		

La ojiva para la frecuencia Acumulada del Porcentaje de la Frecuencia Relativa ($A\%Fr_i$) de la tabla de frecuencias agrupadas del ejemplo 57:



7.- Medidas de tendencia central

Estas medidas corresponden a variables cuantitativas discretas o continuas. Representan un único valor para el tamaño de la Población o Población Objetivo (N), en la que son parámetros, pero en la muestra (n) son estadísticos. La más usada es la media aritmética o promedio. Hay diversas medidas de tendencia central de las que se mostrarán la media aritmética, la mediana y la moda. Esta última, puede no existir en un conjunto de datos, haber una o más de una. Sin embargo, en cada caso es importante determinar cuál de ellas es la que representa mejor a los datos y/o es más útil al trabajo.

Media Aritmética

La media consta de varios tipos como la media acotada, la media geométrica, la media sopesada, la media aritmética o promedio, etc. Hay una expresión para el tamaño de la Población o Población Objetivo que tiene N elementos y otra para el de la muestra con n componentes. En este caso Población y Población Objetivo se usarán indistintamente.

La media aritmética, solamente es correcto calcularla para variables cuantitativas discretas o de razón. Es sensible a valores extremos, superiores o inferiores. En el primer caso queda sesgada hacia un valor mayor y en el segundo hacia un valor menor. Un valor extremo se considera aquel cuyo valor es muy diferente (mayor o menor) a la *nube* de datos encontrados. Tiene las mismas unidades de medición que los datos originales, incluye en su cálculo a todos y cada uno de los valores poblacionales o muestrales, y siempre existe, a diferencia de la moda que puede no existir. Si se trata de una población con distribución normal es el valor que tiene la mayor probabilidad.

En la población la expresión es:

Donde:

μ = media aritmética poblacional
$\sum$ = señala la sumatoria de cada uno de los datos (X)
N = tamaño de la población
i = contador que va de =1 a *i* = N
X_i = cada dato de la población

En la muestra la expresión es:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Donde:

$\bar{x}$ = media aritmética muestral

$\sum$ = señala la sumatoria de cada uno de los datos (x)

n = tamaño de la muestra

i = contador que va de i =1 a i = n

x_i = cada dato de la muestra

Cálculo de la Media Aritmética

Ejemplo 68. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor en la escala de Richter, para una variable de razón que en este caso es la magnitud de los sismos, la media aritmética es:

Magnitud de 89 sismos registrados:

No.	Magnitud
1	1.0
2	1.0
3	1.3
4	1.3
5	1.3
6	1.4
7	1.5
8	1.5
9	1.5
10	1.5
11	1.5
12	1.5
13	1.7
14	1.7
15	1.7
16	1.7
17	1.7
18	1.7
19	1.7
20	1.8
21	1.8
22	1.8
23	1.8
24	1.8
25	1.8
26	1.9
27	1.9
28	1.9
29	1.9
30	1.9
31	2.0
32	2.0
33	2.0
34	2.0
35	2.0
36	2.0
37	2.1
38	2.1
39	2.1
40	2.1
41	2.1
42	2.1
43	2.1
44	2.2
45	2.2
46	2.2
47	2.2
48	2.2
49	2.2
50	2.2
51	2.3
52	2.3
53	2.3
54	2.3
55	2.3
56	2.4
57	2.4
58	2.4
59	2.4
60	2.4
61	2.4
62	2.4
63	2.4
64	2.4
65	2.5
66	2.5
67	2.5
68	2.5
69	2.5
70	2.5
71	2.6
72	2.6
73	2.6
74	2.6
75	2.7
76	2.7
77	2.7
78	2.7
79	2.7
80	2.9
81	2.9
82	3.0
83	3.3
84	3.3
85	3.5
86	3.6
87	4.0
88	4.1
89	4.2

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1 + 1 + 1.3 + \cdots + 4.2}{89} = 2.21797753$$

La media aritmética obtenida se redondea a una cifra, más a la derecha del punto, que los datos originales. En este caso, alcanzan las decimales, por lo que será redondeada a la centésima:

$$\bar{x} = 2.22 \text{ grados Richter}$$

Entre los años 2016 y 2019, en la ciudad de México, hubo una media aritmética de la magnitud de 1 o mayor en la escala de Richter, de 2.22 grados Richter.

Ejemplo 69. En el registro de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, la media aritmética es:

Datos de la circunferencia de cintura (CCI) de 43 mujeres:

No.	CCI
1	66.70
2	67.20
3	67.70
4	67.80
5	68.60
6	68.70
7	69.10
8	70.00
9	72.70
10	72.90
11	73.60
12	74.50
13	74.50
14	74.50
15	75.40

No.	CCI
16	75.50
17	75.90
18	78.80
19	79.50
20	81.40
21	82.50
22	82.60
23	85.00
24	85.70
25	85.70
26	91.10
27	92.80
28	92.80
29	93.00

No.	CCI
30	94.00
31	95.50
32	98.00
33	99.30
34	99.40
35	100.70
36	104.70
37	105.50
38	105.50
39	112.80
40	115.30
41	118.90
42	126.00
43	126.50

Fuente: Datos modificados del software estadístico statdisk 13.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{66.70 + 67.20 + 67.70 + \cdots + 126.5}{43} = 87.1697674$$

La media aritmética obtenida se redondea a una cifra más a la derecha del punto que los datos originales. En este caso, alcanzan las decimales, por lo que será redondeada a la centésima:

$$\bar{x} = 87.17 \text{ cm}$$

El resultado obtenido es que la media aritmética de la circunferencia de cintura de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el 2008 fue de = 87.17 cm.

Mediana

Hay una expresión para el tamaño de la Población o Población Objetivo que tiene N elementos y otra para el de la muestra con n componentes. También, en este caso, Población y Población Objetivo se usarán indistintamente.

La mediana poblacional (Me) y la muestral (me) dividen a la mitad los datos de variables cuantitativas, discretas o continuas, una mitad abarca al 50% de datos menor a ella y la otra es el 50% mayor y depende de si el tamaño de la población (N) o de la muestra (n) es par o impar, el primer paso para calcular la mediana es ordenar los datos de menor a mayor. La mediana siempre está a la mitad de los datos.

La mediana solamente se puede calcular correctamente para variables cuantitativas discretas o de razón. A diferencia de la media aritmética no es sensible a valores extremos, superiores o inferiores. Tiene las mismas unidades de medición que los datos originales, solamente utiliza el dato central cuyo valor es la mediana poblacional o muestral. Siempre existe, a diferencia de la moda que puede no existir.

En la población la expresión es:

Si N es impar la mediana es:

$$Me = X_{\frac{N+1}{2}}$$

Donde:
Me = mediana poblacional es el valor del dato central
X = dato que ocupa la posición (N+1)/2

Si N par es:

$$Me = \frac{X_{\frac{N}{2}} + X_{\left(\frac{N}{2}\right)+1}}{2}$$

Donde:
Me = mediana poblacional es el valor promedio de los dos datos centrales
X = dato que ocupa la posición N/2
X = dato que ocupa la posición (N/2) +1

En la muestra la expresión es:

Si n es impar es:

$$me = x_{\frac{n+1}{2}}$$

Donde:
me = mediana muestral, es el valor del dato central
x = dato que ocupa la posición (n+1)/2

Si n es par es:

$$me = \frac{x_{\frac{n}{2}} + x_{\left(\frac{n}{2}\right)+1}}{2}$$

Donde:
me = mediana muestral, es el valor promedio de los dos datos centrales
x = dato que ocupa la posición n/2
x = dato que ocupa la posición (n/2) +1

Cálculo de la Mediana

Ejemplo 70. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor en la escala de Richter, provenientes del ejemplo 68, la mediana es:
El total n = 89 por lo tanto es impar; los datos se ordenan de menor a mayor y entonces la mediana de la magnitud de los sismos es:

$$me = x_{\frac{n+1}{2}} = x_{\frac{89+1}{2}} = x_{45} = \text{ 2.2 grados Richter}$$

La mediana obtenida no se redondea ya que es el valor del dato central.

El resultado es que entre los años 2016 y 2019, en la ciudad de México, la mediana de la magnitud de los sismos de 1 o mayor en la escala de Richter, fue de me = 2.2 grados Richter.

Ejemplo 71. En el registro de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, provenientes del ejemplo 69 que son 43, se incluyó en los datos de circunferencia un dato más para utilizar el cálculo de la mediana cuando n es par, por lo cual la mediana es:
El total n = 44 por lo tanto es par; los datos se ordenan de menor a mayor y entonces la mediana de la circunferencia de cintura es:

$$me = \frac{x_{\frac{n}{2}} + x_{\left(\frac{n}{2}\right)+1}}{2} = \frac{x_{\frac{44}{2}} + x_{\left(\frac{44}{2}\right)+1}}{2} = \frac{82.6 + 85.0}{2} = 83.80 \text{ cm}$$

La mediana obtenida se redondea a una cifra más, a la derecha del punto, que los datos originales. Sin embargo, en este caso no se hace necesario.

El resultado obtenido es que la mediana de la circunferencia de cintura de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México, en el 2008 fue de me = 83.80 cm.

Moda

La Moda (M_o) tanto en la población como en la muestra es el valor más frecuente, se determina calculando el dato cuya Frecuencia Absoluta (F_i) es la mayor. Es la única que puede no existir o haber una o más modas. Si hay una sola moda se dice que los datos son unimodales, si hay dos, bimodales y si hay tres o más se les conoce como polimodales. Puede utilizarse en cualquier tipo de variable. No es afectada por valores extremos, solamente toma en cuenta al valor más frecuente o a los igualmente frecuentes. No es afectada por valores extremos, utiliza un solo valor en su cálculo y es poco usada.

Determinación de la Moda

Ejemplo 72. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, provenientes del ejemplo 68, la moda es:
La Frecuencia Absoluta (F_i) mayor es F_i = 9 y corresponde a la magnitud de 2.4 grados Richter.

El resultado es que entre los años 2016 y 2019, en la ciudad de México, la moda de la magnitud de los sismos fue de 2.4 grados Richter. Los datos son unimodales.

Ejemplo 73. En el registro de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, provenientes del ejemplo 69 la moda es:
La Frecuencia Absoluta (F_i) mayor es F_i = 3 y corresponde al valor de 74.5 cm de circunferencia de cintura. Los datos son unimodales.

El resultado obtenido es que la moda de la circunferencia de cintura de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el 2008 fue de 74.5 cm.

8.- Medidas de variablidad

Estas medidas corresponden a variables cuantitativas discretas o continuas. Representan un único valor para el tamaño de la Población o Población Objetivo (N), en la que son parámetros, pero en la muestra (n) son estadísticos. Es muy usada la desviación estándar, sin embargo, su cálculo implica determinar la varianza; en cada caso es importante determinar cuál de ellas es la que representa mejor a los datos. Hay diversas medidas de variabilidad de las que se mostrarán la varianza, la desviación estándar y el rango.

Varianza

Esta media de variabilidad en el caso de la Población o Población Objetivo (se usarán indistintamente) y de la muestra constan de diferentes expresiones. La primera consta de N elementos y la segunda con n - 1.

La varianza,solo se puede calcular correctamente para variables cuantitativas discretas o de razón. Tiene las unidades de medición de los datos originales, pero al cuadrado. En su cálculo incluye a todas las distancias, al cuadrado, de cada dato respecto de la media aritmética poblacional o muestral. Siempre es positiva, su menor valor es 0 y el mayor es abierto. La raíz cuadrada de la varianza poblacional o muestral es la desviación estándar.

En la población la expresión es:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

Donde:

^{2}= varianza poblacional. Es la distancia promedio, en la población, de todos los datos a la media aritmética.

$\sum$ = sumatoria de todas las distancias al cuadrado de cada dato a la media aritmética. $(xi - \mu)^2$.

N = tamaño de la población.

i= contador que va de *i* =1 a *i* = N

Xi = cada dato de la población

En la muestra la expresión es:

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

Donde:

s^2= varianza muestral. Es la distancia promedio, en la muestra, de todos los datos a la media aritmética.

$\sum$ = sumatoria de todas las distancias al cuadrado de cada dato a la media aritmética. $(x_i - \bar{x})^2$.

n = tamaño de la muestra.

i = contador que va de i =1 a i = n

x_i = cada dato de la muestra

Cálculo de la Varianza

Ejemplo 74. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, la varianza muestral es:

Magnitud de 89 sismos registrados:

No.	Magnitud
1	1.0
2	1.0
3	1.3
4	1.3
5	1.3
6	1.4
7	1.5
8	1.5
9	1.5
10	1.5
11	1.5
12	1.5
13	1.7
14	1.7
15	1.7
16	1.7
17	1.7
18	1.7
19	1.7
20	1.8
21	1.8
22	1.8
23	1.8
24	1.8
25	1.8
26	1.9
27	1.9
28	1.9
29	1.9
30	1.9
31	2.0
32	2.0
33	2.0
34	2.0
35	2.0
36	2.0
37	2.1
38	2.1
39	2.1
40	2.1
41	2.1
42	2.1
43	2.1
44	2.2
45	2.2
46	2.2
47	2.2
48	2.2
49	2.2
50	2.2
51	2.3
52	2.3
53	2.3
54	2.3
55	2.3
56	2.4
57	2.4
58	2.4
59	2.4
60	2.4
61	2.4
62	2.4
63	2.4
64	2.4
65	2.5
66	2.5
67	2.5
68	2.5
69	2.5
70	2.5
71	2.6
72	2.6
73	2.6
74	2.6
75	2.7
76	2.7
77	2.7
78	2.7
79	2.7
80	2.9
81	2.9
82	3.0
83	3.3
84	3.3
85	3.5
86	3.6
87	4.0
88	4.1
89	4.2

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(1-2.22)^2 + (1-2.22)^2 + (1.3-2.22)^2 + \cdots + (4.2-2.22)^2}{89-1} = 0.383082227$$

La varianza muestral obtenida se redondea a una cifra más a la derecha del punto que los datos originales. En este caso, alcanzan las decimales, por lo que será redondeada a la centésima:

$$s^2 = 0.38 \text{ (grados Richter)}^2$$

El resultado es que entre los años 2016 y 2019, en la ciudad de México, hubo una varianza muestral de la magnitud de 1 o mayor en la escala de Richter de s_2 = de 0.38 (grados Richter)^2.

Ejemplo 75. En el caso del registro de la circunferencia cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, la varianza muestral es:

Datos de la circunferencia de cintura (CCI) de 43 mujeres:

No.	CCI
1	66.70
2	67.20
3	67.70
4	67.80
5	68.60
6	68.70
7	69.10
8	70.00
9	72.70
10	72.90
11	73.60
12	74.50
13	74.50
14	74.50
15	75.40

No.	CCI
16	75.50
17	75.90
18	78.80
19	79.50
20	81.40
21	82.50
22	82.60
23	85.00
24	85.70
25	85.70
26	91.10
27	92.80
28	92.80
29	93.00
30	94.00

No.	CCI
31	95.50
32	98.00
33	99.30
34	99.40
35	100.70
36	104.70
37	105.50
38	105.50
39	112.80
40	115.30
41	118.90
42	126.00
43	126.50

Fuente: Datos modificados del software estadístico statdisk 13.

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

$$= \frac{(66.70 - 87.17)^2 + (67.20 - 87.17)^2 + \cdots + (126.5 - 87.17)^2}{43 - 1} =$$

$$=283.0621595$$

La varianza obtenida se redondea a una cifra más a la derecha del punto que los datos originales. En este caso, alcanzan las decimales, por lo que será redondeada a la centésima:

$$s^2=283.06 \text{ (cm)}^2$$

El resultado obtenido es que la varianza de la circunferencia de cintura de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el 2008 fue de $s^2 = 283.06 \text{ (cm)}^2$.

Desviación Estándar

En esta media de variabilidad la Población o Población Objetivo (se usarán indistintamente) y la muestra constan de diferentes expresiones. La primera con N elementos y la segunda con n - 1.

La desviación estándar solamente se puede calcular correctamente para variables cuantitativas discretas o de razón. Tiene las mismas unidades de medición que los datos originales, a diferencia de la varianza las unidades no son al cuadrado. Incluye en su cálculo a todas las distancias de cada dato respecto de la media aritmética poblacional o muestral. Siempre es positiva, su valor menor es 0 y el mayor es abierto.

En la población la expresión es:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Donde:

σ = desviación estándar poblacional. La distancia promedio, en la población, de todos los datos a la media aritmética.

$\sum$ = sumatoria de todas las distancias al cuadrado de cada dato a la media aritmética. $(x_i - \mu)^2$.

N = tamaño de la población.

i = contador que va de i =1 a i = N

X_i = cada dato de la población

En la muestra la expresión es:

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

Donde:

s = desviación estándar muestral. Es la distancia promedio, en la muestra, de todos los datos respecto a la media aritmética.

$\sum$ = sumatoria de todas las distancias al cuadrado de cada dato a la media aritmética. $(x_i - \bar{x})^2$.

n = tamaño de la muestra.

i = contador que va de i =1 a i = n

x_i = cada dato de la muestra

Cálculo de la Desviación Estándar

Ejemplo 76. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, en los sismos del ejemplo 74, la desviación estándar es:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(1-2.22)^2 + (1-2.22)^2 + (1.3-2.22)^2 + \cdots + (4.2-2.22)^2}{89-1}} =$$

$$=0.\ 61893637$$

La desviación estándar muestral obtenida se redondea a una cifra más, a la derecha del punto, que los datos originales. En este caso, alcanzan las decimales, por lo que será redondeada a la centésima:

$$s = 0.62 \text{ grados Richter}$$

El resultado es que entre los años 2016 y 2019 en la ciudad de México, hubo una desviación estándar muestral de la magnitud de 1 o mayor en la escala de Richter de s = 0.19 grados Richter.

Ejemplo 77. En el registro de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, para los datos del ejemplo 75 la desviación estándar muestral es:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(66.70-87.17)^2 + (67.20-87.17)^2 + (67.70-87.17)^2 + \cdots + (126.5-87.17)^2}{43-1}} =$$

$$=16.82445124$$

La desviación estándar obtenida se redondea a una cifra más, a la derecha del punto, que los datos originales. En este caso, alcanzan las decimales, por lo que será redondeada a la centésima:

$$s = 16.82 \text{ cm}$$

El resultado obtenido es que la desviación estándar de la circunferencia de cintura de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México, en el 2008 fue s = 16.82 cm.

Rango

Esta medida de variabilidad está entre las menos utilizadas, su uso correcto solamente es en variables cuantitativas, discretas o continuas. Nada más toma en cuenta al valor mayor y al valor menor de los datos. Es muy fácil de calcular y puede ser la primera aproximación a la dispersión de los datos, mientras más grande la dispersión es mayor y viceversa. Es la misma expresión tanto para la población como en la muestra.

Únicamente ocupa dos valores de todos los datos, el valor mayor y el valor menor; es fácil de calcular; los datos extremos lo afectan en el tamaño de su intervalo.

En la población o la muestra la expresión es:

$$R = \text{Valor Mayor} - \text{Valor Menor}$$

Donde :

R = rango de los datos
Valor Mayor = valor más grande de la Población o la Muestra
Valor menor = valor más pequeño de la Población o la Muestra

Cálculo del Rango

Ejemplo 78. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, para los datos del ejemplo 74 el Rango muestral es:

$$R = \text{Valor Mayor} - \text{Valor Menor} = 4.2 - 1 = 3.2$$

El rango obtenido no se redondea, ya que se utilizan los datos originales, así mismo, el valor resultante tiene las unidades de medición de esos elementos, por lo que el resultado corresponde a:

$$R = 3.2 \text{ grados Richter}$$

En consecuencia, el resultado es que entre los años 2016 y 2019, en la ciudad de México, hubo un rango muestral de la magnitud de 1 o mayor en la escala de Richter de R = 3.2 grados Richter.

Ejemplo 79. En la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, para los datos del ejemplo 75 el Rango muestral es:

$$R = \text{Valor Mayor} - \text{Valor Menor} = 126.5 - 66.7 = 59.8$$

El rango obtenido no se redondea, ya que se utilizan los datos originales, así mismo, el valor resultante tiene las unidades de medición de esos elementos, por lo que el resultado corresponde a:

$$R = 59.8 \text{ cm}$$

El resultado es que el rango de la CCI de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el 2008 fue R = 59.8 cm.

Coeficiente de Variación

Esta medida de variabilidad es muy usada en la comparación de variables que podrían no ser las mismas. Es un porcentaje de variación de cada Población o Población Objetivo (se usarán de manera indistinta ambos conceptos) o de la muestra, por lo tanto la comparación es de la variación dentro cada una. Aprovecha la determinación de la media aritmética y de la desviación estándar. La forma de cálculo difiere en la Población o Población Objetivo y en la Muestra

En la población la expresión es:

$$C.V. = (/\mu)100$$

Donde:
C. V. = coeficiente de variación en porcentaje
= desviación estándar poblacional
μ = media aritmética poblacional

En la muestra la expresión es:

$$C.V. = (s/)100$$

Donde:
C. V. = coeficiente de variación en porcentaje
s = desviación estándar muestral
$\overline{x}$= media aritmética muestral

Cálculo del Coeficiente de Variación

Ejemplo 80. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, el Coeficiente de Variación (C.V.) muestral para los datos del ejemplo 74 es:

$$C.V. = (s/\overline{x})100 = (0.62/2.22)*100 = 27.927927927$$

El Coeficiente de Variación (C.V.) se redondea a una cifra más, a la derecha del punto que los datos originales. En este caso, alcanzan las decimales, por lo que será redondeado a la centésima:

$$C.V. = 27.93\%$$

El resultado es que entre los años 2016 y 2019, en la ciudad de México, hubo un coeficiente de variación de la magnitud de 1 o mayor en la escala de Richter, el coeficiente de variación es C.V. = 27.93%.

Ejemplo 81. En el registro de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, para los datos del ejemplo 75 el Coeficiente de Variación (C.V.) es:

$$C.V. = (s/)100 = (16.82/87.17)*100 = 19.295629230$$

El Coeficiente de Variación (C.V.) se redondea a una cifra más, a la derecha del punto, que los datos originales. En este caso, alcanzan las decimales, por lo que será redondeado a la centésima:

$$C.V. = 19.30\ \%$$

El resultado obtenido es que el coeficiente de variación de la circunferencia de cintura de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el 2008 fue C.V. = 19.30%.

Comparación de los Coeficientes de Variación

Los coeficientes de variación de la magnitud de los sismos (C.V. = 27.93%) y de la circunferencia de cintura (C.V. = 19.30%) pueden ser comparables, ya que ambos están en porcentaje, en consecuencia, el de la segunda es menor que el de la primera, por lo cual, la circunferencia de cintura de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el 2008 fue menor que el coeficiente de variación de la magnitud de los sismos entre los años 2016 y 2019, en la ciudad de México.

9.- Medidas de localización

Estas herramientas permiten una localización más precisa de los valores dentro de la Población o Población Objetivo (estos términos se utilizarán de manera indistinta) o la Muestra. Dividen en porcentajes iguales al conjunto de datos, pues los dividen desde los valores menores a los mayores. Cualquier conjunto de datos puede ser fragmentado en sectores de igual porcentaje, sin embargo, los más comunes son los cuartiles y los deciles. Los cuartiles (Q) dividen en cuatro partes porcentualmente iguales a los datos, en tanto que los deciles (D) lo hacen en diez.

Cuartiles

Los cuartiles representan cada uno de ellos el 25% de los datos. Consta de tres cuartiles: $Q_{1,}$ Q_2 y Q_3. Ordenados los datos de menor a mayor valor, entonces se tiene que:

En la población o la muestra los cuartiles son:

Q_1 25% de los valores menores o iguales a Q_1.

Q_2 50% de los valores menores o iguales a Q_2 .

Q_3 75% de los valores menores o iguales a Q_3.

Cada cuartil corresponde a un Percentil (*P*), que dividen a los datos en 100 secciones en la que en cada una hay 1% de los datos. Los cuartiles corresponden a los siguientes percentiles: $Q_1 = P_{25}$; $Q_2 = P_{50}$ y $Q_3 = P_{75}$

Entre cada una de las secciones (valor menor a Q_1; Q_1 a Q_2 y Q_2 a Q_3) se encontraría el 25% de todos los datos de la Población o de la Muestra, cuya sumatoria abarca al 100%. No hay acuerdo entre diversos autores y tampoco entre diferentes programas computacionales en la mejor forma de cálculo, sin embargo, la ecuación más empleada que es a partir de los percentiles correspondientes es *P*(n+1), y será utilizada en los ejemplos que están a continuación.

Cálculo de Cuartiles

Ejemplo 82. En el caso del registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, los cuartiles son:

Magnitud de 89 sismos registrados:

No.	Magnitud
1	1.0
2	1.0
3	1.3
4	1.3
5	1.3
6	1.4
7	1.5
8	1.5
9	1.5
10	1.5
11	1.5
12	1.5
13	1.7
14	1.7
15	1.7
16	1.7
17	1.7
18	1.7
19	1.7
20	1.8
21	1.8
22	1.8
23	1.8
24	1.8
25	1.8
26	1.9
27	1.9
28	1.9
29	1.9
30	1.9

No.	Magnitud
31	2.0
32	2.0
33	2.0
34	2.0
35	2.0
36	2.0
37	2.1
38	2.1
39	2.1
40	2.1
41	2.1
42	2.1
43	2.1
44	2.2
45	2.2
46	2.2
47	2.2
48	2.2
49	2.2
50	2.2
51	2.3
52	2.3
53	2.3
54	2.3
55	2.3
56	2.4
57	2.4
58	2.4
59	2.4
60	2.4

No.	Magnitud
61	2.4
62	2.4
63	2.4
64	2.4
65	2.5
66	2.5
67	2.5
68	2.5
69	2.5
70	2.5
71	2.6
72	2.6
73	2.6
74	2.6
75	2.7
76	2.7
77	2.7
78	2.7
79	2.7
80	2.9
81	2.9
82	3.0
83	3.3
84	3.3
85	3.5
86	3.6
87	4.0
88	4.1
89	4.2

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

Para el cálculo de los cuartiles se utilizará la ecuación mencionada y los datos de los sismos ya están ordenados del menor al mayor, así que:

$Q_1 = 0.25(89+1) = 0.25(90) = x_{22.5}$ para la separación de los datos en cuartiles el dato del cuartil siempre tendrá que ser un entero por lo que $Q_1 = x_{23}$

$Q_2 = 0.5(89+1) = 0.5(90) = x_{45}$

$Q_3 = 0.75(89+1) = 0.75(90) = x_{67.5} = x_{68}$, por las mismas razones que Q_1, $Q_3 = x_{68}$

Los cuartiles de la magnitud de los sismos ocurridos en la Ciudad de México en los años del

2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, son:

$$Q_1 = x_{23} = 1.8 \text{ grados Richter}$$
$$Q_2 = x_{45} = 2.2 \text{ grados Richter}$$
$$Q_3 = x_{68} = 2.5 \text{ grados Richter}$$

Ejemplo 83. En el registro de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, los cuartiles serían para los datos de 43 mujeres:

No.	CCI
1	66.70
2	67.20
3	67.70
4	67.80
5	68.60
6	68.70
7	69.10
8	70.00
9	72.70
10	72.90
11	73.60
12	74.50
13	74.50
14	74.50
15	75.40

No.	CCI
16	75.50
17	75.90
18	78.80
19	79.50
20	81.40
21	82.50
22	82.60
23	85.00
24	85.70
25	85.70
26	91.10
27	92.80
28	92.80
29	93.00
30	94.00

No.	CCI
31	95.50
32	98.00
33	99.30
34	99.40
35	100.70
36	104.70
37	105.50
38	105.50
39	112.80
40	115.30
41	118.90
42	126.00
43	126.50

Fuente: Datos modificados del software estadístico statdisk 13.

Para el cálculo de los cuartiles se utilizará la ecuación mencionada y los datos de las CCI ya están ordenados del menor al mayor, así que:

$Q_1 = 0.25(43+1) = 0.25(44) = x_{11}$ para la separación de los datos en cuartiles el dato del cuartil siempre tendrá que ser un entero por lo que $Q_1 = x_{11}$.

$Q_2 = 0.5(43+1) = 0.5(44) = x_{22}$

$Q_3 = 0.75(43+1) = 0.75(44) = x_{33}$

Los cuartiles de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México, en el año 2008, son:

$$Q_1 = x_{11} = 73.60 \text{ cm}$$
$$Q_2 = x_{22} = 82.60$$
$$Q_3 = x_{33} = 99.30 \text{ cm}$$

Deciles

En la separación en deciles, cada uno representa el 10% de los datos. Consta de nueve deciles: D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8 y D_9. La suma de los deciles equivale al 100% de los datos. Al igual que en los cuartiles los valores están ordenados de menor a mayor, entonces se tiene que:

En la población o la muestra los deciles son:

D_1 10% de los valores menores o iguales a D_1
D_2 20% de los valores menores o iguales a D_2
D_3 30% de los valores menores o iguales a D_3
D_4 40% de los valores menores o iguales a D_4
D_5 50% de los valores menores o iguales a D_5
D_6 60% de los valores menores o iguales a D_6
D_7 70% de los valores menores o iguales a D_7
D_8 80% de los valores menores o iguales a D_8
D_9 90% de los valores menores o iguales a D_9

Cada decil corresponde a un Percentil (*P*), que dividen a los datos en 100 secciones en la que cada una hay 1% de los datos. Los deciles corresponden a los siguientes percentiles: $D_1 = P_{10}$'; $D_2 = P_{20}$; $D_3 = P_{30}$; $D_4 = P_{40}$; $D_5 = P_{50}$; $D_6 = P_{60}$; $D_7 = P_{70}$; $D_8 = P_{80}$ y $D_9 = P_{90}$.

Entre cada una de las secciones (valor menor a D_1; D_1 a D_2; D_2 a D_3, etc.) se encontraría el 10% de todos los datos de la Población o de la Muestra, cuya sumatoria abarca al 100%. No hay acuerdo entre diversos autores y tampoco entre diferentes programas computacionales en la mejor forma de cálculo, sin embargo, la ecuación más empleada que es a partir de los percentiles correspondientes es *P*(n+1), y será utilizada en los ejemplos que están a continuación.

Cálculo de los Deciles

Ejemplo 84. En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, los deciles para los datos del ejemplo 82 son:

$$D_1 = 0.10(89+1) = 0.10(90) = x_9$$
$$D_2 = 0.20(89+1) = 0.20(90) = x_{18}$$
$$D_3 = 0.30(89+1) = 0.30(90) = x_{27}$$
$$D_4 = 0.40(89+1) = 0.40(90) = x_{36}$$
$$D_5 = 0.50(89+1) = 0.50(90) = x_{45}$$
$$D_6 = 0.60(89+1) = 0.60(90) = x_{54}$$
$$D_7 = 0.70(89+1) = 0.70(90) = x_{63}$$
$$D_8 = 0.80(89+1) = 0.80(90) = x_{72}$$
$$D_9 = 0.90(89+1) = 0.90(90) = x_{81}$$

Los deciles de la magnitud de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor, en la escala de Richter, serían:

$D_1 = x_9 = 1.5$ grados Richter
$D_2 = x_{18} = 1.7$ grados Richter
$D_3 = x_{27} = 1.9$ grados Richter
$D_4 = x_{36} = 2.0$ grados Richter
$D_5 = x_{45} = 2.2$ grados Richter
$D_6 = x_{54} = 2.3$ grados Richter
$D_7 = x_{63} = 2.4$ grados Richter
$D_8 = x_{72} = 2.6$ grados Richter
$D_9 = x_{81} = 2.9$ grados Richter

Ejemplo 85. En el registro de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, los deciles para los datos del ejemplo 83 serían:

$D_1 = 0.10(43+1) = 0.10(44) = x_{4.4}$ para la separación de los datos en deciles el dato del decil tendrá que ser un entero por lo que, se redondea al siguiente entero inmediato superior $D_1 = x_5$
$D_2 = 0.20(43+1) = 0.20(44) = x_{8.80} = x_9$
$D_3 = 0.30(43+1) = 0.30(44) = x_{13.20} = x_{14}$
$D_4 = 0.40(43+1) = 0.40(44) = x_{17.60} = x_{18}$
$D_5 = 0.50(43+1) = 0.50(44) = x_{22.00} = x_{22}$
$D_6 = 0.60(43+1) = 0.60(44) = x_{26.40} = x_{27}$
$D_7 = 0.70(43+1) = 0.70(44) = x_{30.80} = x_{31}$

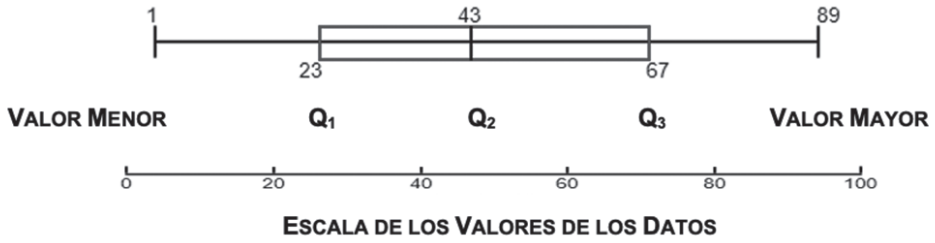
$$D_8 = 0.80(43+1) = 0.80(44) = x_{35.20} = x_{36}$$
$$D_9 = 0.90(43+1) = 0.90(44) = x_{39.60} = x_{40}$$

Los deciles de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México, en el año 2008, serían:

$$D_1 = x_5 = 68.60 \text{ cm}$$
$$D_2 = x_9 = 72.70 \text{ cm}$$
$$D_3 = x_{14} = 74.50 \text{ cm}$$
$$D_4 = x_{18} = 78.80 \text{ cm}$$
$$D_5 = x_{22} = 82.60 \text{ cm}$$
$$D_6 = x_{27} = 92.80 \text{ cm}$$
$$D_7 = x_{31} = 95.50 \text{ cm}$$
$$D_8 = x_{36} = 104.70 \text{ cm}$$
$$D_9 = x_{40} = 115.30 \text{ cm}$$

Gráfica de Caja

La gráfica de caja (boxplot) a partir de los cuartiles muestra, en el eje inferior, la escala de valores de los datos asociados con su valor. Es la misma gráfica ya sea que se trate de la Población, Población Objetivo o Muestra. Señala en una caja el valor de los cuartiles Q_1, Q_2 y Q_3. Además, incluye al valor menor y al mayor.

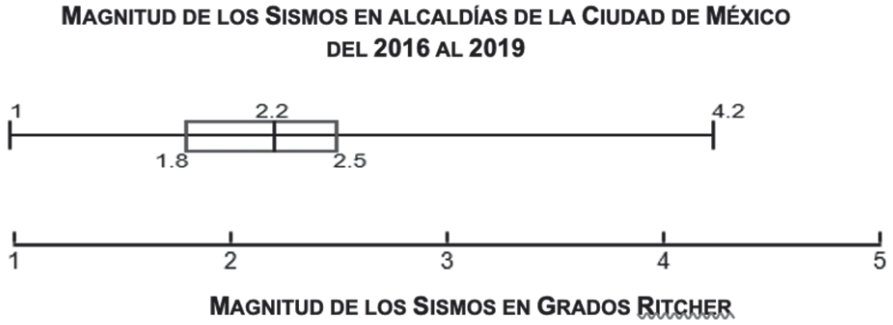


Fuente: Modificado del software estadístico statdisk 13.

Es útil para revisar el comportamiento de los datos y sus valores, así como para señalar el sesgo de la distribución de los datos. Entre el Valor Menor y Q_1 existe el 25% de los datos, la misma cantidad porcentual está entre Q_1 y Q_2; Q_2 y Q_3 e igualmente entre Q_3 y el Valor Mayor, cuya sumatoria es igual al 100% de los datos. Además puede ser comparable con la o las gráficas de caja de otros datos siempre y cuando tengan las mismas unidades de medición o registro.

Construcción de la Gráfica de Caja

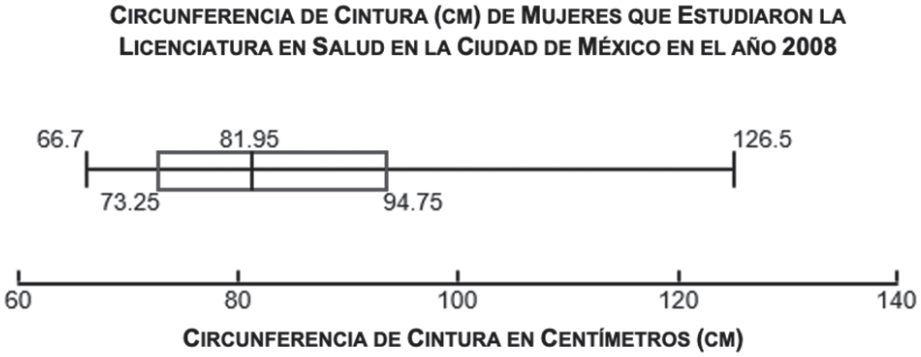
Ejemplo 86. En el caso del registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor en la escala de Richter, la gráfica de caja sería para los datos del ejemplo 82:



Fuente: Modificado del software estadístico statdisk 13.

La gráfica de caja muestra que la diferencia entre el valor menor y Q_1 la diferencia es de 0.8 grados Richter para el 25% de los datos, en cambio, es mayor de Q_3 al valor mayor que es de 1.7 grados Richter e igualmente representa 25% de los terremotos, por lo que hay un sesgo de la distribución hacia las magnitudes mayores. Los sismos menores al Q_1 son más semejantes en cuanto a su magnitud que los mayores a Q_3. Entre Q_1 y Q_2 la desigualdad es de 0.4 grados Richter mayor de Q_2 a Q_3 que alcanza 0.3 grados Richter. Sin embargo, de Q_1 a Q_3 es de 0.7 grados Richter. Podría decirse que, si las circunstancias del 2016 al 2019 son semejantes en años posteriores, al menos el 50% de los sismos en alcaldías de la Ciudad de México serán de magnitud igual o mayor a 2.2 grados Richter y el otro 50% de los sismos tendrán una igual o menor.

Ejemplo 87. En el registro de la circunferencia de cintura en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, la gráfica de caja sería para los datos del ejemplo 83 sería:



Fuente: Modificado del software estadístico statdisk 13.

La gráfica de caja muestra que la diferencia entre el valor menor y Q_1 es de 6.55 cm para el 25% de los datos, en cambio, es mayor de Q_3 al valor mayor que es de 31.75 cm e igualmente representa 25% de las mujeres a las que se les midió la circunferencia de cintura, por lo que hay un sesgo de la distribución hacia las circunferencias de cinturas mayores. Las circunferencias de cinturas menores al Q_1 son más semejantes en cuanto a su tamaño que los mayores a Q_3. Entre Q_1 y Q_2 la desigualdad es de 8.70 cm menor que de Q_2 a Q_3 que es de 12.80 cm. Sin embargo, de Q_1 a Q_3 es de 21.50 cm. Podría decirse que si las circunstancias del 2008 son semejantes en años posteriores, al menos el 50% de las mujeres que estudien la licenciatura en Salud en la Ciudad de México será superior o igual a 81.95 cm y el otro 50% tendrán una igual o menor

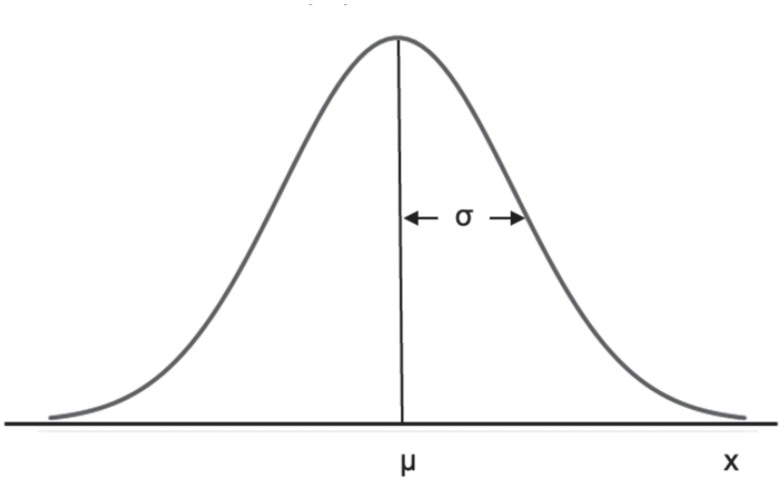
10.- Distribuciones de probabilidad

En la Estadística existen diversas distribuciones de probabilidad que son de utilidad en el estudio de variados fenómenos, cada una tiene una tabla de distribuciones utilizada de acuerdo con el análisis determinado a los datos producto de la investigación. Entre ellas se encuentran las utilizadas en el curso: Normal Estándar o Estandarizada, *t* de Student y la Distribución F (también es mencionada en variadas fuentes como Distribución F de Fisher, Distribución F de Snedecor o Distribución F de Fisher-Snedecor). En cualquier buen libro de Estadística, Bioestadística o internet, se encuentran las tablas de las probabilidades correspondientes y no son reproducidas en el presente documento por rebasar sus objetivos.

La más utilizada en el curso y tal vez en la mayoría de las investigaciones es la Distribución Normal Estándar, derivada de la distribución Normal. La distribución *t* de Student se utiliza en lugar de la Distribución Normal cuando así se requiera. En el curso la distribución F es aplicada preliminarmente en las pruebas de hipótesis para saber estadísticamente si las desviaciones estándar poblacionales ( ) de dos muestras son o no iguales al momento de realizar una inferencia sobre dos medias (μ).

Distribución Normal

La Distribución Normal, Campana de Gauss o simplemente Distribución de Campana, es para datos cuantitativos continuos; muestra cómo se distribuyen los valores en una gráfica cuyo eje “X” son los valores poblacionales (X), con media aritmética (μ) y desviación estándar ( ). Es una distribución simétrica que abarca desde los valores más pequeños a los más grandes. En el eje “Y” puede estar la frecuencia absoluta (F_i) de X o bien la Probabilidad de “X” (P_x).



Fuente: Modificado de https://www.get-digital-help.com/es/c%C3%B3mo-graficar-una-distribuci%C3%B3n-normal/

Muchas pruebas estadísticas paramétricas se basan en la Distribución Normal, siempre y cuando la población la tenga. Debido a su simetría bilateral la media aritmética (μ) = Mediana (M_e) = Moda (M_o).

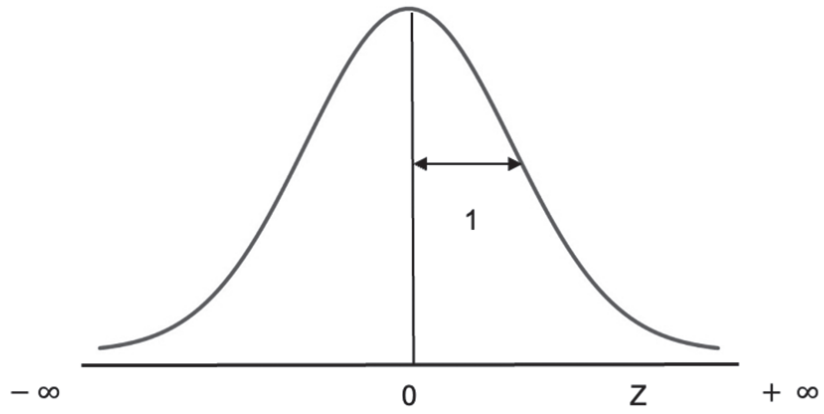
Distribución Normal Estándar

Cada población de datos cuantitativos continuos tiene unidades de medición diferentes, que se evitan al estandarizarlos y forman una Distribución Normal Estándar, cuyo eje "Y" estaría la probabilidad de cada valor estandarizado (Z) los cuales estarían en el eje "X". La estandarización es para el eje "X" y se calcula para cada dato de acuerdo con la siguiente fórmula:

En la población la expresión es:

Donde:
Z_i = valor estandarizado del dato i de la población
X_i = valor del dato i de la población
μ = media aritmética de la población
= desviación estándar poblacional

El resultado es que, para cada valor X_i hay un valor estandarizado Z. Al considerar todos los valores teóricos la Distribución Normal Estándar resulta de la estandarización, se distribuye $N \sim (0,1)$ esto es, su distribución es como una normal con media 0 y desviación estándar 1. Z abarca desde: $-\infty$ *a*$+\infty$:



Fuente: Modificado de https://www.get-digital-help.com/es/c%C3%B3mo-graficar-una-distribuci%C3%B3n-normal/

Es muy útil para analizar cualquier variable cuantitativa continua cuyos datos originales sean de cualquier tipo de unidades, ya que al estandarizarlos se obtienen cifras sin unidades. Hereda las características de la distribución Normal pero cuyo eje de simetría bilateral es 0. Es importante notar que la curva de la gráfica no toca al eje donde está Z, ni a la izquierda ya que se prolonga hasta $-\infty$ ni a la derecha ya que se continua hasta $+\infty$ Muchas pruebas estadísticas la usan, siempre y cuando la población tenga una distribución normal, de no

ser así se utiliza la distribución de *t* de Student que depende de los grados de libertad, los que al aumentar la *t* se va asemejando a la normal, y si son en gran cantidad se puede usar la Distribución Normal.

En la muestra la expresión es:

$$Z_i = \frac{x_i - \bar{x}}{s}$$

Donde:

Z_i = valor estandarizado del dato i de la Muestra
x_i = valor del dato i de la Muestra
$\bar{x}$= media aritmética de la Muestra
s = desviación estándar Muestra

Cálculo de la Estandarización

Ejemplo 88. En el caso del registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor en la escala de Richter, los valores estandarizados se calculan a partir de:

Magnitud de 89 sismos registrados:

No.	Magnitud
1	1.00
2	1.00
3	1.30
4	1.30
5	1.30
6	1.40
7	1.50
8	1.50
9	1.50
10	1.50
11	1.50
12	1.50
13	1.70
14	1.70
15	1.70
16	1.70
17	1.70
18	1.70

No.	Magnitud
19	1.70
20	1.80
21	1.80
22	1.80
23	1.80
24	1.80
25	1.80
26	1.90
27	1.90
28	1.90
29	1.90
30	1.90
31	2.00
32	2.00
33	2.00
34	2.00
35	2.00
36	2.00

No.	Magnitud
37	2.10
38	2.10
39	2.10
40	2.10
41	2.10
42	2.10
43	2.10
44	2.20
45	2.20
46	2.20
47	2.20
48	2.20
49	2.20
50	2.20
51	2.30
52	2.30
53	2.30
54	2.30

No.	Magnitud
55	2.30
56	2.40
57	2.40
58	2.40
59	2.40
60	2.40
61	2.40
62	2.40
63	2.40
64	2.40
65	2.50
66	2.50

No.	Magnitud
67	2.50
68	2.50
69	2.50
70	2.50
71	2.60
72	2.60
73	2.60
74	2.60
75	2.70
76	2.70
77	2.70
78	2.70

No.	Magnitud
79	2.70
80	2.90
81	2.90
82	3.00
83	3.30
84	3.30
85	3.50
86	3.60
87	4.00
88	4.10
89	4.20

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

$\bar{x}$= 2.2179 grados Richter; s = 0.6189 grados Richter

Estandarización de los datos de la Muestra de la magnitud de los sismos:

$$Z_i = \frac{x_i - \bar{x}}{s}$$

Al sustituir para el primer valor de x_i = 1 grado Richter:

$$Z_i = \frac{1 - 2.2179}{0.6189} = -1.9679$$

La ecuación se aplica a las 89 magnitudes de los sismos, con lo que se obtiene:

No.	Magnitud	Z
1	1	-1.9679
2	1	-1.9679
3	1.3	-1.4832
4	1.3	-1.4832
5	1.3	-1.4832
6	1.4	-1.3216
7	1.5	-1.1600

No.	Magnitud	Z
8	1.5	-1.1600
9	1.5	-1.1600
10	1.5	-1.1600
11	1.5	-1.1600
12	1.5	-1.1600
13	1.7	-0.8369
14	1.7	-0.8369

No.	Magnitud	Z
15	1.7	-0.8369
16	1.7	-0.8369
17	1.7	-0.8369
18	1.7	-0.8369
19	1.7	-0.8369
20	1.8	-0.6753
21	1.8	-0.6753
22	1.8	-0.6753
23	1.8	-0.6753
24	1.8	-0.6753
25	1.8	-0.6753
26	1.9	-0.5137
27	1.9	-0.5137
28	1.9	-0.5137
29	1.9	-0.5137
30	1.9	-0.5137
31	2	-0.3522
32	2	-0.3522
33	2	-0.3522
34	2	-0.3522
35	2	-0.3522
36	2	-0.3522
37	2.1	-0.1906
38	2.1	-0.1906
39	2.1	-0.1906
40	2.1	-0.1906
41	2.1	-0.1906
42	2.1	-0.1906
43	2.1	-0.1906
44	2.2	-0.0290
45	2.2	-0.0290
46	2.2	-0.0290
47	2.2	-0.0290
48	2.2	-0.0290
49	2.2	-0.0290
50	2.2	-0.0290
51	2.3	0.1325
52	2.3	0.1325
53	2.3	0.1325
54	2.3	0.1325
55	2.3	0.1325
56	2.4	0.2941
57	2.4	0.2941
58	2.4	0.2941
59	2.4	0.2941
60	2.4	0.2941
61	2.4	0.2941
62	2.4	0.2941
63	2.4	0.2941
64	2.4	0.2941
65	2.5	0.4557
66	2.5	0.4557
67	2.5	0.4557
68	2.5	0.4557
69	2.5	0.4557
70	2.5	0.4557
71	2.6	0.6172
72	2.6	0.6172
73	2.6	0.6172
74	2.6	0.6172
75	2.7	0.7788
76	2.7	0.7788
77	2.7	0.7788
78	2.7	0.7788
79	2.7	0.7788
80	2.9	1.1019
81	2.9	1.1019
82	3	1.2635
83	3.3	1.7482
84	3.3	1.7482
85	3.5	2.0713
86	3.6	2.2329
87	4	2.8792
88	4.1	3.0407
89	4.2	3.2023

Ejemplo 89. En el de la circunferencia de cintura (CCI) en centímetros (cm) de mujeres que estudiaron la licenciatura en Salud en la Ciudad de México en el año 2008, los valores estandarizados se calculan a partir de:

Datos de la circunferencia de cintura (CCI) de 43 mujeres:

No.	CCI
1	66.70
2	67.20
3	67.70
4	67.80
5	68.60
6	68.70
7	69.10
8	70.00
9	72.70
10	72.90
11	73.60
12	74.50
13	74.50
14	74.50
15	75.40

No.	CCI
16	75.50
17	75.90
18	78.80
19	79.50
20	81.40
21	82.50
22	82.60
23	85.00
24	85.70
25	85.70
26	91.10
27	92.80
28	92.80
29	93.00

No.	CCI
30	94.00
31	95.50
32	98.00
33	99.30
34	99.40
35	100.70
36	104.70
37	105.50
38	105.50
39	112.80
40	115.30
41	118.90
42	126.00
43	126.50

Fuente: Datos modificados del software estadístico statdisk 13.

$\bar{x}$= 123.80 cm; s = 16.82 cm

Estandarización de los datos de la Muestra de la circunferencia de cintura de mujeres:

$$Z_i = \frac{x_i - \bar{x}}{s}$$

Al sustituir para el primer valor de xi = 66.70 cm:

$$Z_i = \frac{66.70 - 123.80}{16.82} = -3.3939$$

La ecuación se aplica a las 43 mediciones de CCI con lo que se obtiene:

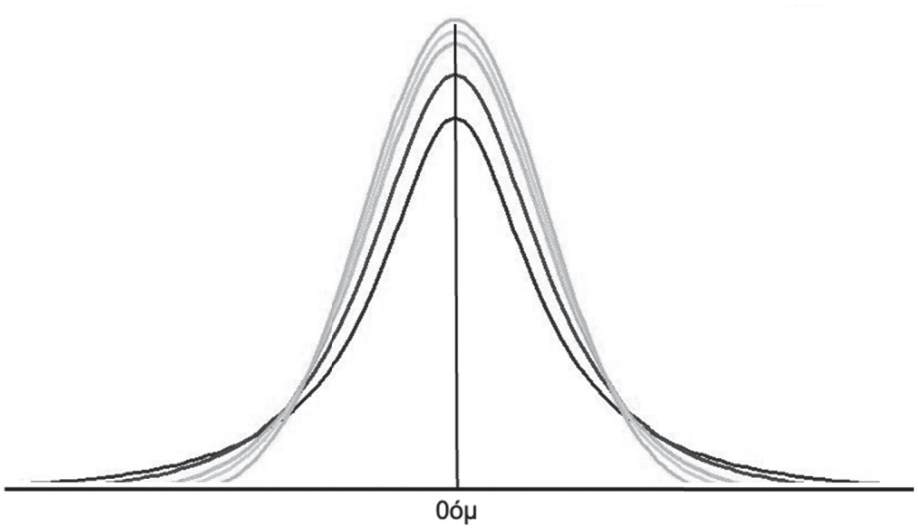
No.	CCI	Z
1	66.70	-3.3939
2	67.20	-3.3642
3	67.70	-3.3344
4	67.80	-3.3285
5	68.60	-3.2809
6	68.70	-3.2750
7	69.10	-3.2512
8	70.00	-3.1977
9	72.70	-3.0372
10	72.90	-3.0254
11	73.60	-2.9838
12	74.50	-2.9303
13	74.50	-2.9303
14	74.50	-2.9303
15	75.40	-2.8768
16	75.50	-2.8708
17	75.90	-2.8470
18	78.80	-2.6747
19	79.50	-2.6331
20	81.40	-2.5201
21	82.50	-2.4548
22	82.60	-2.4488

No.	CCI	Z
23	85.00	-2.3062
24	85.70	-2.2646
25	85.70	-2.2646
26	91.10	-1.9436
27	92.80	-1.8426
28	92.80	-1.8426
29	93.00	-1.8307
30	94.00	-1.7712
31	95.50	-1.6821
32	98.00	-1.5335
33	99.30	-1.4562
34	99.40	-1.4503
35	100.70	-1.3730
36	104.70	-1.1353
37	105.50	-1.0877
38	105.50	-1.0877
39	112.80	-0.6538
40	115.30	-0.5052
41	118.90	-0.2912
42	126.00	0.1308
43	126.50	0.1605

Distribución t de Student

Esta distribución se utiliza cuando no es factible la Distribución Normal, siempre y cuando la población se distribuya como una normal y el tamaño de la muestra sea pequeña ($n < 30$). Depende de los grados de libertad (g.l.) que a su vez derivan del tamaño de la Muestra (n); para cada uno de ellos hay una distribución *t* de Student: g.l. = n -1. Su forma es muy semejante a la Distribución Normal. Igual que la Distribución Normal es simétrica con su eje de simetría bilateral en la media aritmética (μ) o en 0 si los datos están estandarizados.

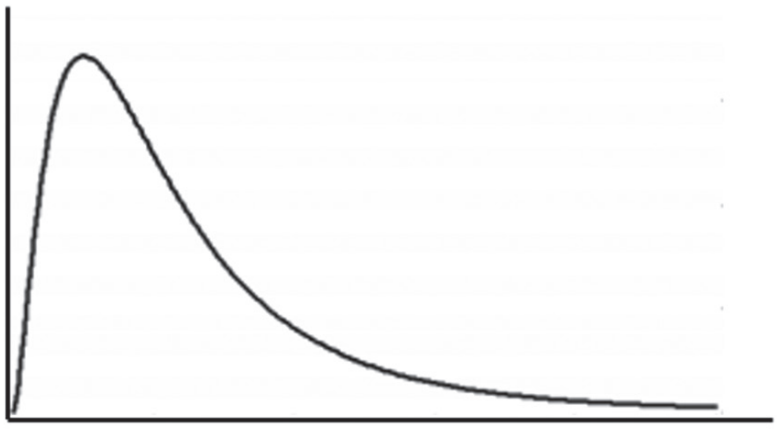
Las curvas que se muestran en la figura siguiente representan diferentes distribuciones de *t* de Student, cada una para cierto número de grados de libertad, los que al incrementarse la desviación estándar aumenta y la altura es mayor.



Fuente: Modificado de https://es.wikipedia.org/w/index.php?title=Distribuci%C3%B3n_t_de_Student&oldid=119488972

Distribución F

Es una distribución que, entre otras cosas, se utiliza para evaluar si las varianzas (^{2}) de dos poblaciones ($_1^2$y $_2^2$) son o no iguales, durante el proceso de una prueba de hipótesis. Está determinada por los grados de libertad del cociente de varianzas: grados de libertad del numerador (n_1-1) y grados de libertad del denominador (n_2-1). No es simétrica y los valores de la distribución F no pueden ser negativos.



Fuente: Modificado de http://www.ivia.gva.es/documents/16186 2582/162430134/capitulo+9.pdf/e31fdd7c-2bc5-4bcd-a477-5afc03441f59

11.- Pruebas de hipótesis

En las investigaciones puede ser necesario analizar proporciones (*p*), medias (μ) o desviaciones estándar ( ); determinar estadísticamente la diferencia de ellas contra un valor especificado o la diferencia entre dos de ellas (p_1 y p_2; μ_1 y μ_2; $_1$ y $_2$). Hay diversos tipos de pruebas de hipótesis, tanto en la Estadística Paramétrica como en la Estadística no Paramétrica. Las hipótesis estadísticas **SIEMPRE** van en pares, hipótesis nula (H_0) e hipótesis alternativa (H_1 o H_a) y siempre se efectúan para los parámetros de la población.

Al menos hay seis equivocaciones en la utilización e interpretación de las pruebas de hipótesis: 1) evaluar el resultado solamente en cuanto a la significación estadística pero no la contextual, 2) realizar en un mismo estudio varias pruebas de hipótesis acerca de las mismas variables y tomar el Error tipo I de manera individual, 3) interpretar aisladamente los resultados sin tomar en cuenta la existencia de otros estudios relativos al caso, 4) pensar que no rechazar la hipótesis nula (H_0) representa una prueba de su veracidad, 5) aplicar los resultados a una o unas poblaciones diferentes a la que se les realizó la prueba de hipótesis y 6) realizar las pruebas de hipótesis en muestras no aleatorias.

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Determinación del Valor P

En la determinación del *valor P* [no confundir con la proporción de la población (*p*)] para tomar una decisión sobre si se rechaza o no H_0 con los siguientes criterios: a) Si la prueba es de cola izquierda, es la probabilidad a la izquierda del estadístico de prueba. b) Si es de cola derecha, es la probabilidad a la derecha del estadístico de prueba. Si es de dos colas, existen dos posibilidades: c_1) Si el resultado del estadístico de la prueba de hipótesis está a la izquierda del eje de simetría de la Distribución Normal Estándar, entonces será dos veces la probabilidad del estadístico de prueba a la izquierda y c_2) Si el resultado del estadístico de la prueba de hipótesis está a la derecha del eje de simetría de la Distribución Normal Estándar entonces será dos veces la probabilidad del estadístico de prueba a la derecha. La determinación del *valor P* puede obtenerse al realizar la prueba de hipótesis con el programa Statdisk o cualquier otro programa computacional profesional.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Si el *valor P*< , H_0 se rechaza, o si el valor absoluto del estadístico de prueba es mayor que el valor absoluto del valor crítico de tablas, H_0 se rechaza, pero si el *valor P* , H_0 no se rechaza, o si el valor absoluto del estadístico de prueba es menor que el valor absoluto del valor crítico de tablas H_0 no se rechaza.

Prueba de Hipótesis para la Diferencia Entre Dos Proporciones

H_0: $p_1 = p_2$

H_1: $p_1 < p_2$

$p_1 > p_2$

$p_1 \quad p_2$

Donde:

p_1 = proporción de la población 1

p_2 = proporción de la población 2

H_o = hipótesis nula lleva el signo de igual (=).

H_1 = hipótesis alternativa. Es la que se desea probar. El signo es uno de los siguientes <, >ó

Conforme el signo de la H_1 o H_a la prueba de hipótesis será de cola derecha (>), cola izquierda (<) o de dos colas ( ).

Ambas muestras cumplen con:

1.-ser muestras independientes y,

2.- n_1p_1 5 y n_1q_1 5 ; n_2p_2 5 y n_1q_1 5. Si *p* y *q* son desconocidas se les reemplaza con los valores apareados de $\bar{p}$ y $\bar{q}$:

$$p = \frac{x_1}{n_1}$$

$$q = 1 - p_1$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Donde:

p 1 proporción de éxitos

p 2 proporción de fracasos

$\bar{p}$ = proporción de éxitos apareada

x_1 = número de éxitos en la muestra n_1

x_2 = número de éxitos en la muestra n_2.

n_1 = tamaño de la muestra 1

n_2 = tamaño de la muestra 2

$\overline{q} = 1 - \overline{p}$

Estadístico de prueba:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

Donde:

Donde:

Z = probabilidad normal estándar de Z

$\hat{p}_1 = x_1/n_1$

$\hat{p}_2 = x_2/n_2$

$\bar{p}$ = proporción de éxitos apareada

x_1 = número de éxitos en la muestra n_1

x_2 = número de éxitos en la muestra n_2

n_1 = tamaño de la muestra 1

n_2 = tamaño de la muestra 2

$\bar{q} = 1 - \bar{p}$

Prueba de Hipótesis para la Diferencia Entre Dos Medias

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 < \mu_2$

$\mu_1 > \mu_2$

$\mu_1 \quad \mu_2$

Donde:

μ_1 = media aritmética de la población 1

μ_2 = media aritmética de la población 2

H_o = hipótesis nula lleva el signo de igual (=)

H_1 = hipótesis alternativa. La que se desea probar. El signo es uno de los siguientes <, >ó

Conforme el signo de la H_1 la prueba de hipótesis será de cola derecha (>), cola izquierda (<) o de dos colas ( ).

Conforme la prueba de *F* si $\sigma_2^1 \neq \sigma_2^2 \neq$ y ambas muestras cumplen con:

1.-ser independientes o,

2.- ser aleatorias simples y,

3.- ambas muestras son grandes, $n_1 > 30$ y $n_2 > 30$

Estadístico de prueba:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Donde:

t = estadístico de prueba de distribución t

$\bar{x}_1$ = media aritmética de la muestra 1

$\bar{x}_2$ = media aritmética de la muestra 2

s_1^2 = varianza de la muestra 1

s_2^2 = varianza de la muestra 2

n_1 = tamaño de la muestra 1

n_2 = tamaño de la muestra 2

Grados de libertad:

$$g.l. = \frac{(A+B)^2}{\frac{A^2}{n_1-1} + \frac{B^2}{n_2-1}}$$

Donde:

g.l. = grados de libertad

A = s_1^2/n_1

B = s_2^2/n_2

Conforme la prueba de *F* si $\sigma_2^1 = \sigma_2^2$ y ambas muestras cumplen con:

1. ser independientes o,
2. ser aleatorias simples y,
3. ambas muestras son grandes, $n_1 > 30$ y $n_2 > 30$

Estadístico de prueba:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Donde:

t = estadístico de prueba de distribución *t*
$\bar{x}_1$= media aritmética de la muestra 1
$\bar{x}_2$ = media aritmética de la muestra 2
s_p^2= promedio ponderado de la varianza de la muestra 1 y la muestra 2
n_1 = tamaño de la muestra 1
n_2 = tamaño de la muestra 2

Varianza agrupada de la muestra 1 con la muestra 2:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Donde:

s_p^2 = promedio ponderado de la varianza de la muestra 1 y la 2

n_1 = tamaño de la muestra 1

n_2 = tamaño de la muestra 2

s_1^2 = varianza de la muestra 1

s_2^2 = varianza de la muestra 2

Grados de libertad:

$$g.l = n_1 + n_2 - 2$$

Prueba de F para determinar si σ_2^1 y σ_2^2son iguales o no

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

Donde:

σ_1^2 = varianza de la población 1

σ_2^2 = varianza de la población 2

H_o = hipótesis nula lleva el signo de igual (=).

H_1 = hipótesis alternativa. Es la que se desea probar. En este caso el signo es ≠ por lo que la prueba es de dos colas.

Estadístico de prueba:

$$F = \frac{s_1^2}{s_2^2}$$

Donde:

F = estadístico de prueba que se distribuye como una *F*

s_1^2 = varianza mayor de s_1^2 y s_2^2

s_2^2 = varianza menor de s_1^2 y s_2^2

Cálculo de la Prueba de Hipótesis para la Diferencia de Dos Medias

Ejemplo 89. En los sismos ocurridos en la Ciudad de México y en la ciudad de Toluca en el 2016 de 1 grado Richter o mayor, la prueba de hipótesis podría ser de la diferencia de las medias de la magnitud.

Magnitud de 85 sismos registrados en la Ciudad de México y la ciudad de Toluca:

No.	Magnitud
Ciudad de México	
1	1.00
2	1.00
3	1.30
4	1.30
5	1.30
6	1.40
7	1.50
8	1.50
9	1.50
10	1.50
11	1.50
12	1.50
13	1.70

No.	Magnitud
14	1.70
15	1.70
16	1.70
17	1.70
18	1.70
19	1.70
20	1.80
21	1.80
22	1.80
23	1.80
24	1.80
25	1.80
26	1.90
27	1.90

No.	Magnitud
28	1.90
29	1.90
30	1.90
31	2.00
32	2.00
33	2.00
34	2.00
35	2.00
36	2.00
37	2.10
38	2.10
39	2.10
40	2.10
41	2.10

No.	Magnitud
42	2.10
43	2.10
44	2.20
45	2.20
Ciudad de Toluca	
1	2.20
2	2.20
3	2.20
4	2.20
5	2.20
6	2.30
7	2.30
8	2.30
9	2.30
10	2.30

No.	Magnitud
11	2.40
12	2.40
13	2.40
14	2.40
15	2.40
16	2.40
17	2.40
18	2.40
19	2.40
20	2.50
21	2.50
22	2.50
23	2.50
24	2.50
25	2.50

No.	Magnitud
26	2.60
27	2.60
28	2.60
29	2.60
30	2.70
31	2.70
32	2.70
33	2.70
34	2.70
35	2.90
36	2.90
37	3.00
38	3.30
39	3.30
40	3.50

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

Ciudad de México: $\bar{x}_1$= 1.77 grados Richter; s_1=0.30 grados Richter; n_1=45 sismos

Toluca: $\bar{x}_2$= 2.55 grados Richter; s_2=0.31 grados Richter; n_2=40 sismos

$$H_0: \mu_1 = \mu 2$$

$$H_1: \mu_1 \ \mu_2$$

Nivel de significancia: = 0.05, como es una prueba de dos colas ( ) se tiene en cada cola /2 = 0.025.

Regla de Decisión: si el *valor P*< , H_0 se rechaza, pero si *P*> , H_0 no se rechaza.

Prueba *F* para determinar si $^{1}_{2}$y $^{2}_{2}$y son o no iguales:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$F = \frac{s_1^2}{s_2^2} = \frac{(0.31)^2}{(0.30)^2} = \frac{0.0961}{0.09} = 1.0678$$

Como es de dos colas ( ) /2 = 0.05/2 = 0.025 y,

$F_{/2 = 0.025,\ g.l.= 40\text{-}1/45\text{-}1} = F_{/2 = 0.025,\ g.l.} = 39/44 = 1.8752$

De acuerdo con la Regla de Decisión como el *valor P* = 0.8290 = 0.05 la H_o no se rechaza.

Retomando la prueba de hipótesis relativa a las medias.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \quad \mu_2$$

Como $\frac{1}{2} = \frac{2}{2}$ la varianza será la ponderada:

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{(44)0.09+(39)0.0961}{44+39} = \frac{(44)0.09+(39)0.0961}{83} =$$

$$= \frac{3.96+3.75}{83} = \frac{7.71}{83} = 0.093$$

Por lo tanto, el valor del estadístico de prueba es:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}} = \frac{1.77-2.55}{\sqrt{\frac{0.093}{45}+\frac{0.093}{40}}} = \frac{-0.78}{\sqrt{0.0021+0.0023}} = \frac{-0.78}{\sqrt{0.0044}}\frac{-0.78}{0.0663} = -11.7647$$

Grados de libertad:

$$g.l= 45+40-2=83$$

En la tabla de probabilidades de *t* de Student se determina $t_{/2 = 0.025,\ g.l. = (45+40)-2} = t_{/2 = 0.025,\ g.l.} = 83 \pm 1.990$

Conclusión estadística:de acuerdo con la Regla de Decisión, como el *valor P* = 0.0000 < = 0.05, H_0 se rechaza, por lo tanto, $\mu_1 \quad \mu_2$.

Conclusión contextual: la media de la intensidad de los sismos de 1 o más grados Richter, entre la Ciudad de México y la ciudad de Toluca sí es diferente para el año 2016, con una media para la segunda de μ= 2.55 grados Richter, en tanto que la primera tuvo una media de μ = 1.77 grados Richter. Lo anterior, para una = 0.05 n_1 = 45 y n_2 = 40.

Ejemplo 90. En el caso del índice de masa corporal (IMC) para mujeres y hombres de la Ciudad de México en 2009, la prueba de hipótesis sería la diferencia entre los IMC de mujeres y hombres.

IMC de 80 mujeres y hombres de la ciudad de México:

No.	IMC Mujeres	No.	IMC Hombres
1	19.60	1	23.80
2	23.80	2	23.20
3	19.60	3	24.60
4	29.10	4	26.20
5	25.20	5	23.50
6	21.40	6	24.50
7	22.00	7	21.50
8	27.50	8	31.40
9	33.50	9	26.40
10	20.60	10	22.70
11	29.90	11	27.80
12	17.70	12	28.10
13	24.00	13	25.20
14	28.90	14	23.30
15	37.70	15	31.90
16	18.30	16	33.10
17	19.80	17	33.20
18	29.80	18	26.70
19	29.70	19	26.60
20	31.70	20	19.90

No.	IMC Mujeres	No.	IMC Hombres
21	23.80	21	27.10
22	44.90	22	23.40
23	19.20	23	27.00
24	28.70	24	21.60
25	28.50	25	30.90
26	19.30	26	28.30
27	31.00	27	25.50
28	25.10	28	24.60
29	22.80	29	23.80
30	30.90	30	27.40
31	26.50	31	28.70
32	21.20	32	26.20
33	40.60	33	26.40
34	21.90	34	32.10
35	26.00	35	19.60
36	23.50	36	20.70
37	22.80	37	26.30
38	20.70	38	26.90
39	20.50	39	25.60
40	21.90	40	24.20

Fuente: Datos modificados del statdisk 13

IMC Mujeres: $\bar{x}_1$=25.74 imc; s_1=6.166 imc; n_1=40 mujeres

IMC Hombres: $\bar{x}_2$=25.997 imc; s_2=3.43 imc; n_2=40 hombres

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \quad \mu_2$$

Nivel de significancia: = 0.05, como es una prueba de dos colas ( ) se tiene en cada cola /2 = 0.025.

Regla de Decisión: si el *valor P*< , H_0 se rechaza, pero si *P* , H_0 no se rechaza.

Prueba *F* para determinar si $^2_1$ y $^2_2$ son o no iguales:

$$H0: \ ^2_1 = \ ^2_2$$

$$H1: \ ^2_1 \quad ^2_2$$

$$F = \frac{s_1^2}{s_2^2} = \frac{(6.166)^2}{(3.43)^2} = \frac{38.014}{11.77} = 3.23$$

Como es de dos colas ( ) /2 = 0.05/2 = 0.025 y, $F_{/2 = 0.025,\ g.l.}$= 40-1/40-1 = $F_{/2 = 0.025,\ g.l.}$ = 39/39 = 1.8752

De acuerdo con la Regla de Decisión como el *valor P* = 0.0004 < = 0.05 la H_0 se rechaza, por lo tanto $^{2}_{1}$ $^{2}_{2}$.

Retomando la prueba de hipótesis relativa a las medias.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \ \mu_2$$

Como $^{2}_{1}$ $^{2}_{2}$, la varianza será la de cada muestra, por lo tanto, el valor del estadístico de prueba es:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{25.74 - 25.997}{\sqrt{\frac{(6.166)^2}{40} + \frac{(3.43)^2}{40}}} = \frac{-0.257}{\sqrt{0.950 + 0.294}} = \frac{-0.257}{\sqrt{1.244}} = \frac{-0.257}{1.115} =$$

$$= -0.2305$$

Grados de libertad:

$$g.l. = \frac{(A+B)^2}{\frac{A^2}{n_1-1} + \frac{B^2}{n_2-1}} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_{2-1}}} = \frac{\left(\frac{38.014}{40} + \frac{11.77}{40}\right)^2}{\frac{\left(\frac{38.014}{40}\right)^2}{40-1} + \frac{\left(\frac{11.77}{40}\right)^2}{40-1}} =$$

$$= \frac{(0.9504 + 0.2943)^2}{\frac{(0.9504)^2}{39} + \frac{(0.2943)^2}{39}} = \frac{1.5493}{\frac{0.9033}{39} + \frac{0.0866}{39}} = \frac{1.5493}{0.0254} = 60.9961$$

En la tabla de probabilidades de *t* de Student se determina $t_{/2 = 0.025,\ g.l. = 61}$ ± 2.0

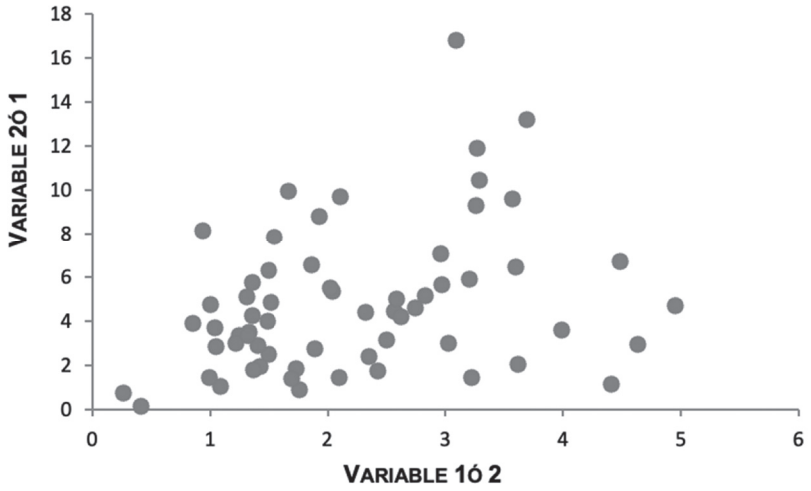
Conclusión estadística: de acuerdo con la Regla de Decisión, como el *valorP* = 0.8186 = 0.05 la H_0 no se rechaza.

Conclusión contextual: la media del índice de masa corporal (IMC) de mujeres y hombres de la Ciudad de México en 2009, puede ser igual en ambos sexos debido a que no se rechaza el planteamiento de que así sea. Esta es una de dos posibilidades, la otra es que realmente no sean iguales. Lo anterior para = 0.05 y n_1 = 40 y n_2 = 40.

12.- Correlación lineal simple

La correlación es la relación de cambio en una variable en relación con otra, puede aplicarse a cualquier tipo de variable cuantitativa. En este caso es la correlación lineal simple, lo que significa que es la relación rectilínea de dos variables. También existe correlación no lineal, lineal múltiple y no rectilínea múltiple. Al examinar la correlación entre un par de variables se encuentra en primer término el visual, que abarca: a) diagrama de correlación o de dispersión, b) la determinación genérica de si hay o no correlación entre las variables y su dirección y c) el grado de compactación de la nube de puntos.

El diagrama de correlación o dispersión es una gráfica en la que cada variable cuantitativa ocupa uno de los ejes sin importar cuál de ellos. Una de ellas tiene sus valores en el eje “X” y la otra en el “Y”. En la parte del plano en la que se localizan se tendría un punto, por lo que habría tantos puntos como datos se graficaran. Al Conjunto de puntos se le denomina “nube de puntos”.



La figura anterior es el diagrama de dispersión o de correlación para las variables 1 y 2, muestra gran cantidad de puntos, cada uno indica un dato con coordenadas “X” y “Y” con sus unidades de medición o registro. El conjunto de puntos forma la nube de puntos que se estudia de manera integral y no punto por punto, al ser de esa forma se aprecia que la “nube de puntos” tiene valores menores a la izquierda del diagrama y superiores a la derecha del mismo. El grado de compactación de la nube de puntos no cuenta con una escala consensada, queda a criterio del investigador.

El diagrama de correlación o dispersión es una exploración del comportamiento de los datos. Seguidamente se calcula del coeficiente de correlación simple de Pearson (r) que determina analíticamente la dirección y la fuerza de la correlación entre las variables. Solamente es útil al tratarse de un fenómeno lineal de dos variables. El valor de r es independiente, de la variable que esté en “X” o en “Y”, es el mismo resultado si las variables 1 y 2 se intercambian y la 1 queda en “Y” y la 2 en “X”.

Los errores más generalizados al interpretar los resultados de la correlación son: a) concluir que si hay correlación hay una relación de causa efecto entre las variables. Correlación no es sinónimo de causa efecto, puede haber correlación, pero no causa efecto, pero si hay causa efecto sí hay correlación, b) la utilización de medias en lugar de los datos para determinar el coeficiente de correlación y c) puede ser que la relación entre las variables exista pero no sea lineal y que al determinarla como si fuera lineal no sea significativa.

La expresión de la correlación lineal simple difiere entre la población y la muestra, ya que en la primera se utiliza la desviación estándar poblacional y en la segunda la desviación estándar muestral. No hay que olvidar que la correlación es una variable cuantitativa continua y los términos de Población y Población objetivo se usan de manera equivalente.

En la población la expresión de es:

$$\rho = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

Donde:

ρ = correlación entre las variables "x" y "y" en la población

σ_{XY} = variación conjunta de "x" y "y" o productos cruzados de "x" y "y"

σ_X^2 = varianza de la variable "x"

σ_Y^2 = varianza de la variable "y"

En la muestra la expresión de r es:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}$$

Donde:

r = coeficiente de correlación de Pearson

x_i =valores de la variable "x"

$\bar{x}$ = media aritmética de la variable "x"

y_i =valores de la variable "y"

$\bar{y}$ = media aritmética de la variable "y"

n = tamaño de la muestra

s_x = desviación estándar de la variable "x"

s_y = desviación estándar de la variable "y"

Valores Posibles del Coeficiente de Correlación de Pearson (r)

El coeficiente de Pearson está acotado en sus valores inferior y superior pero no deja de ser una variable cuantitativa continua, por lo que puede tener cualquier valor de -1 a +1 pasando por el 0.

Cuando el valor es - 1 la correlación es negativa (al aumentar el valor de una variable la otra decrece) y perfecta (los puntos se alinean formando una recta); sí es +1 la correlación será positiva (al aumentar el valor de una variable la otra también crece) y perfecta (los puntos se alinean formando una recta), en cambio, si el resultado es 0 la correlación no existe. El signo del coeficiente indica si la correlación es positiva o negativa, el valor numérico del coeficiente señala la fuerza de la correlación entre las variables.

El obtener algún valor para el coeficiente de correlación de Pearson (r) no implica que sea significativo, por lo que se debe realizar una prueba de hipótesis para determinar si el valor de r es o no significativo.

Prueba de Hipótesis para la Significancia de la Correlación

H_0: = 0 (no hay correlación lineal)

H_1: 0 (sí hay correlación lineal)

Nivel de Significancia:

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Estadístico de Prueba (Coeficiente de Correlación de Pearson)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}$$

Regla de Decisión

Si el *valor P*< que , H_0 se rechaza (si hay correlación lineal) pero si el *valor P*> , H_0 no se rechaza (no hay correlación lineal).

Determinación de la Correlación y su Significancia en una Muestra

Ejemplo 91. En el registro de sismos ocurridos en la Ciudad de México en el 2016 de 1 grado Richter o mayor, el análisis de correlación podría ser entre las intensidades en grados Richter y la hora de su ocurrencia.

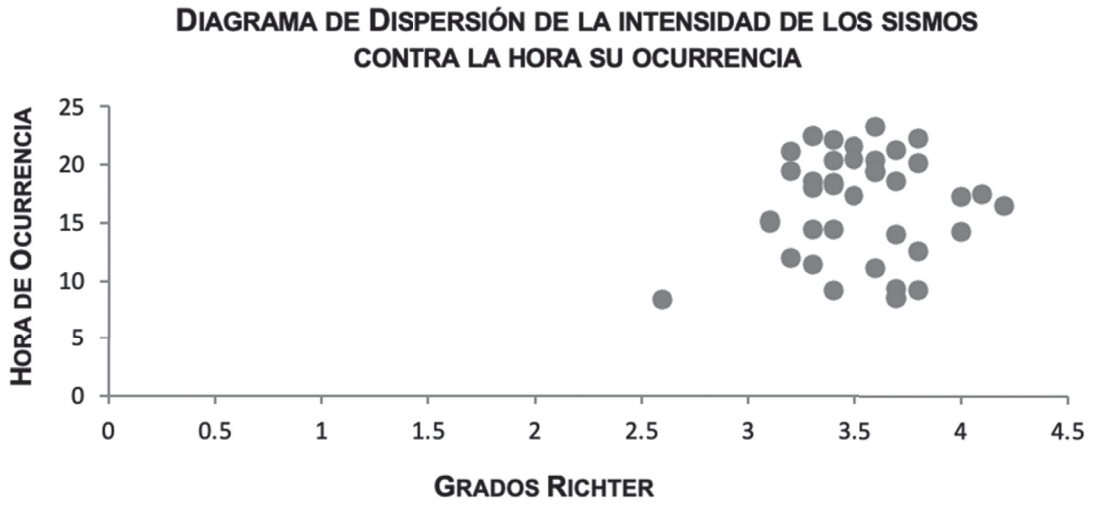
Magnitud y hora local en la que ocurrieron 40 sismos en la Ciudad de México:

No.	Magnitud	Hora local
1	3.60	23.29
2	3.30	22.52
3	3.30	22.47
4	3.80	22.26
5	3.40	22.14
6	3.50	21.57
7	3.70	21.27
8	3.20	21.13
9	3.50	20.41
10	3.60	20.40
11	3.40	20.34
12	3.80	20.07
13	3.60	19.46
14	3.20	19.44
15	3.60	19.30
16	3.30	18.55
17	3.70	18.54
18	3.40	18.41
19	3.40	18.20
20	3.30	18.05

No.	Magnitud	Hora local
21	4.10	17.43
22	3.50	17.28
23	4.00	17.19
24	4.20	16.42
25	3.10	15.17
26	3.10	15.00
27	3.40	14.47
28	3.30	14.40
29	4.00	14.19
30	3.70	14.02
31	3.80	12.49
32	3.20	12.00
33	3.30	11.44
34	3.60	11.08
35	3.70	9.23
36	3.40	9.19
37	3.80	9.12
38	3.70	8.50
39	3.70	8.45
40	2.60	8.35

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

Diagrama de dispersión:



La nube de puntos muestra que a excepción de un sismo (el que está más a la izquierda) los demás se compactan sin indicar claramente la dirección en que lo hacen.

Cálculo del coeficiente de correlación de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y} =$$

$$= \frac{\sum_{i=1}^{40}[(3.60 - 3.52) * (23.29 - 16.581)] + \ldots + [(2.60 - 3.52) * (8.35 - 16.581)]}{40 * 0.307346 * 4.603299} =$$

$$= 0.0326076$$

El coeficiente de correlación de Pearson para la intensidad de los sismos y la hora en la que ocurrieron tiene un valor de r = 0.0326076. El signo indica que la correlación es positiva y que la fuerza de la correlación entre las variables es cercana a 0 ó sea mucho muy débil.

Ahora bien, la prueba de hipótesis para determinar la significancia del coeficiente de correlación de Pearson es:

$H_{0:}$ = 0 (no hay correlación lineal)

$H_{1:}$ 0 (sí hay correlación lineal)

Estadístico de prueba: r = 0.0326076

Valor crítico de tablas: $r_{=0.05,\ n=4}$) ± 0.3120061

Conclusión estadística:de acuerdo con la Regla de Decisión, como el *valor P* = 0.84268 >

= 0.05, H_0 no se rechaza, por lo tanto, no hay correlación lineal simple entre la intensidad en grados Richter de los sismos y la hora en la que ocurrieron.

Conclusión contextual: no hay correlación lineal simple entre la intensidad en grados Richter de los sismos ocurridos en la Ciudad de México en el 2016 y la hora en la que ocurrieron ya que r = 0.0326076 no es significativo para = 0.05 y n = 40.

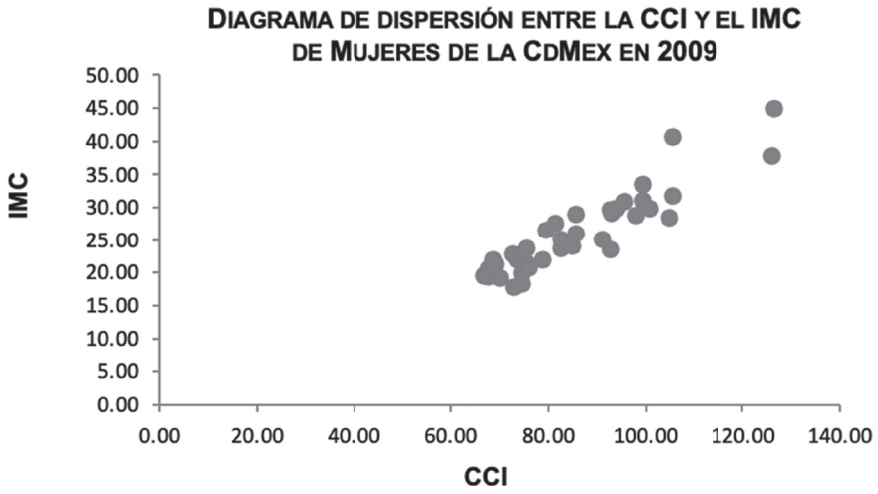
Ejemplo 92. En el índice de masa corporal (IMC) y la circunferencia de cintura (CCI) en mujeres de la Ciudad de México en 2009, el análisis de correlación podría ser entre ambas variables.

IMC y CCI de 40 mujeres de la Ciudad de México:

No.	CCI	IMC	No.	CCI	IMC
1	67.20	19.60	21	75.50	23.80
2	82.50	23.80	22	126.50	44.90
3	66.70	19.60	23	70.00	19.20
4	93.00	29.10	24	98.00	28.70
5	82.60	25.20	25	104.70	28.50
6	75.40	21.40	26	67.80	19.30
7	73.60	22.00	27	99.30	31.00
8	81.40	27.50	28	91.10	25.10
9	99.40	33.50	29	74.50	22.80
10	67.70	20.60	30	95.50	30.90
11	100.70	29.90	31	79.50	26.50
12	72.90	17.70	32	69.10	21.20
13	85.00	24.00	33	105.50	40.60
14	85.70	28.90	34	78.80	21.90
15	126.00	37.70	35	85.70	26.00
16	74.50	18.30	36	92.80	23.50
17	74.50	19.80	37	72.70	22.80
18	94.00	29.80	38	75.90	20.70
19	92.80	29.70	39	68.60	20.50
20	105.50	31.70	40	68.70	21.90

Fuente: Datos modificados del statdisk 13

Diagrama de dispersión



La “nube de puntos” muestra una gran compactación y al aumentar una variable la otra también lo hace, por lo que la correlación es positiva.

Cálculo del coeficiente de correlación de Pearson

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y} =$$

$$= \frac{\sum_{i=1}^{40}[(67.20 - 85.3) * (19.6 - 25.74)] + \cdots + [(68.7 - 85.3) * (21.9 - 25.74)]}{40 * 15.401355 * 6.16557024} =$$

$$= 0.91806664$$

El coeficiente de correlación de Pearson entre la CCI y el IMC es de r = 0.91806664. El signo indica que la correlación es positiva y que la fuerza de la correlación entre las variables es cercana a 1 ó sea mucho muy fuerte.

Ahora bien, la prueba de hipótesis para determinar la significancia del coeficiente de correlación de Pearson es:

$H_{0:}$ = 0 (no hay correlación lineal)

$H_{1:}$ 0 (sí hay correlación lineal)

Estadístico de prueba: r = 0.91806664

Valor crítico de tablas: $r_{=0.05,\ n=40} \pm 0.3120061$

Conclusión estadística:de acuerdo con la Regla de Decisión, como el *valor P* = 0.00000 < = 0.05, H_0 se rechaza, por lo tanto Sí hay correlación lineal simple entre la CCI y el IMC.

Conclusión contextual: sí hay correlación lineal simple entre la circunferencia de cintura (CCI) y el índice de masa corporal (IMC) de mujeres de la Ciudad de México en 2009, ya que r = 0.91806664 es significativo para = 0.05 y n = 40.

13.- Regresión lineal simple

Este tipo de regresión se refiere a la recta y al uso de dos variables. Se aplica a variables cuantitativas continuas para modelar su comportamiento y predecir el valor estimado (y) de “y” a partir de un valor de “x”, pero dentro del intervalo de los valores de “x” de los datos originales, los utilizados para determinar las constantes del modelo (estadístico) muestral, ya que se desconoce su comportamiento fuera de ese rango.

Únicamente si la correlación entre ambas variables es significativa (ver sección de correlación lineal simple), se procede a la regresión, por lo que primeramente se efectúa la prueba de hipótesis de correlación y rechazar H_0 (no hay correlación entre las variables). El modelo se ajusta a los datos pero nunca los datos al modelo, ya que esto último sería un craso error, en todo caso, si el modelo de la recta no ajusta a los datos, hay que recurrir a un modelo diferente, pero nunca a modificar los datos originales. No se debe usar una ecuación de regresión vieja para analizar datos actuales y tampoco hay que usarla para aplicarla a una población diferente de la que provienen los datos.

En la regresión es muy importante decidir qué variable será "x" (variable independiente) y cuál «y» (variable dependiente). La decisión la debe tomar el investigador. Sin embargo, puede ocurrir que esté muy clara la variable independiente y cuál la dependiente.

Modelo general de la recta de regresión:

$$Y= \beta_0+\beta_1 X+\varepsilon$$

Donde:
Y = variable "y" a estimar o predecir
β_0 = ordenada al origen
β_1 = pendiente de la recta
X = variable predictora de "y"
ε = error: variable aleatoria, correspondiente a la variación en "y" no explicable por la relación entre las variables "x" e "y".

Estadístico muestral:
Como el llamado error (ε) es una variable aleatoria independiente el valor esperado de «y» es $\beta_0+\beta_1 X$, pero las constantes muestrales son β_0 y β_1.

$$y_i= b_0+b_1x_i$$

Donde:
$\hat{y}_i$ = $\hat{y}_i$ de la variable estimada
b_0 = ordenada al origen
b_1 = pendiente de la recta
x_i = variable predictora "x"

Determinación de las constantes del modelo muestral:

Ordenada al origen

$$b_0 = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

Pendiente

$$b_1 = \frac{n(\sum_{i=1}^{n} x_i\, y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

Donde:

x_i = valores originales de la variable "x" o variable independiente

y_i = valores originales de la variable "y" o variable dependiente

Suma de cuadrados y tipos de variaciones:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Donde:

$\sum_{i=1}^{n} (y_i - \bar{y})^2$ = variación total

$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ = variación explicada

$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ = variación sin explicar

y_i = valor original de y_i

$\hat{y}_i$ = valor estimado de y_i

$\bar{y}$ = media aritmética de la variable dependiente "y"

Coeficiente de determinación (R^2):

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = (r)^2$$

R^2 = coeficiente de determinación

$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ = variación explicada

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ = variación total

r = coeficiente de correlación de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}$$

Donde:

r = coeficiente de correlación de Pearson

x_i =valores de la variable "x"

$\bar{x}$ = media aritmética de la variable "x"

y_i =valores de la variable "y"

$\bar{y}$ = media aritmética de la variable "y"

n = tamaño de la muestra

s_x = desviación estándar de la variable "x"

s_y = desviación estándar de la variable "y"

Valores Posibles de R^2 y r:

Los valores del coeficiente de determinación R^2 son de 0 a 1 y los del coeficiente de correlación de Pearson r de -1 a +1.

Capacidad del modelo de explicar la variación de «y» por la relación de "x" y "y":

$$\%R^2 = R^2\ 100$$

Se considera que los mejores modelos son los que en la ecuación anterior tienen un valor 75%.

Determinación de la Regresión Lineal Simple

Ejemplo 93. En el registro de sismos ocurridos en la Ciudad de México en el 2016 de 1 grado Richter o mayor, el análisis de regresión podría ser entre las intensidades en grados Richter y la hora de su ocurrencia.

Magnitud y hora local en la que ocurrieron 40 sismos en la Ciudad de México:

No.	Magnitud	Hora local
1	3.60	23.29
2	3.30	22.52
3	3.30	22.47
4	3.80	22.26
5	3.40	22.14
6	3.50	21.57
7	3.70	21.27
8	3.20	21.13
9	3.50	20.41
10	3.60	20.40
11	3.40	20.34
12	3.80	20.07
13	3.60	19.46
14	3.20	19.44
15	3.60	19.30
16	3.30	18.55
17	3.70	18.54
18	3.40	18.41
19	3.40	18.20
20	3.30	18.05

No.	Magnitud	Hora local
21	4.10	17.43
22	3.50	17.28
23	4.00	17.19
24	4.20	16.42
25	3.10	15.17
26	3.10	15.00
27	3.40	14.47
28	3.30	14.40
29	4.00	14.19
30	3.70	14.02
31	3.80	12.49
32	3.20	12.00
33	3.30	11.44
34	3.60	11.08
35	3.70	9.23
36	3.40	9.19
37	3.80	9.12
38	3.70	8.50
39	3.70	8.45
40	2.60	8.35

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

Cálculo del coeficiente de correlación de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y} =$$

$$= \frac{\sum_{i=1}^{40}[(3.60 - 3.52) * (23.29 - 16.581)] + \ldots + [(2.60 - 3.52) * (8.35 - 16.581)]}{40 * 0.307346 * 4.603299} =$$

$$= 0.0326076$$

El coeficiente de correlación de Pearson para la intensidad de los sismos y la hora en la que ocurrieron tiene un valor de r = 0.0326076.

Ahora se procede a realizar la prueba de hipótesis para determinar la significancia del coeficiente de correlación de Pearson obtenido:

$H_{0:}$ = 0 (no hay correlación lineal)

H_1: 0 (sí hay correlación lineal)

Estadístico de prueba: r = 0.0326076

Valor crítico de tablas: $r_{\alpha=0.05,\ n=40} \pm 0.3120061$

Conclusión estadística:de acuerdo con la Regla de Decisión, como el *valor P* = 0.84268 > = 0.05, H_0 no se rechaza, por lo tanto no hay correlación lineal simple entre las variables en estudio.

Conclusión contextual: no hay correlación lineal simple entre la intensidad en grados Richter de los sismos ocurridos en la Ciudad de México en el 2016 y la hora en la que ocurrieron, ya que r = 0.0326076 no es significativo para $r_{= 0.05 \text{ y } n = 40}$. En consecuencia, no es adecuado realizar el análisis de regresión entre esas variables.

Incluso, al determinar la **capacidad del modelo de explicar la variación de "y" por la relación de "x" y "y"** se encuentra que:

$$\%R^2 = R^2\ 100 = (r)^2\ 100 = (0.0326076)^2\ 100 = (0.0010632556)100 = 0.10632556$$

Además de que H_0 no es rechazada, el modelo únicamente explicaría el 0.10632556% de la variación de «y» debida a la relación lineal de "x" e "y", lo que indica un modelo inadecuado.

Ejemplo 94. En el registro del índice de masa corporal (IMC) y la circunferencia de cintura (CCI) en mujeres de la Ciudad de México en 2009, el análisis de regresión podría ser entre ambas variables.

IMC y CCI de 40 mujeres de la Ciudad de México:

No.	CCI	IMC
1	66.70	19.60
2	67.20	19.60
3	67.70	20.60
4	67.80	19.30
5	68.60	20.50
6	68.70	21.90
7	69.10	21.20
8	70.00	19.20
9	72.70	22.80
10	72.90	17.70
11	73.60	22.00
12	74.50	18.30
13	74.50	19.80
14	74.50	22.80

No.	CCI	IMC
15	75.40	21.40
16	75.50	23.80
17	75.90	20.70
18	78.80	21.90
19	79.50	26.50
20	81.40	27.50
21	82.50	23.80
22	82.60	25.20
23	85.00	24.00
24	85.70	28.90
25	85.70	26.00
26	91.10	25.10
27	92.80	29.70
28	92.80	23.50

No.	CCI	IMC
29	93.00	29.10
30	94.00	29.80
31	95.50	30.90
32	98.00	28.70
33	99.30	31.00
34	99.40	33.50
35	100.70	29.90
36	104.70	28.50
37	105.50	31.70
38	105.50	40.60
39	126.00	37.70
40	126.50	44.90

Fuente: Datos modificados del software estadístico statdisk 13.

Cálculo del coeficiente de correlación de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y} =$$

$$= \frac{\sum_{i=1}^{40}[(67.20 - 85.30) * (19.60 - 25.74)] + \ldots + [(68.70 - 85.30) * (21.90 - 25.74)]}{40 * 15.401355 * 6.16557024} =$$

$$= 0.91806664$$

El coeficiente de correlación de Pearson para la correlación entre la CCI y el IMC tiene un valor de r = 0.91806664.

Ahora se procede a realizar la prueba de hipótesis para determinar la significancia del coeficiente de correlación de Pearson obtenido:

H_0: = 0 (no hay correlación lineal)

H_1: 0 (sí hay correlación lineal)

Estadístico de prueba: r = 0.91806664

Valor crítico de tablas: $\mathbf{r}_{=0.05,\ n = 40} \pm 0.3120061$

Conclusión estadística:de acuerdo con la Regla de Decisión, como el *valor P* = 0.00000 < = 0.05, H0 se rechaza, por lo tanto, sí hay correlación lineal simple entre las variables en estudio.

Conclusión contextual: sí hay correlación lineal simple entre la circunferencia de cintura (CCI) y el índice de masa corporal (IMC) de mujeres de la Ciudad de México en 2009, ya que r = 0.91806664 es significativo para = 0.05 y n = 40. En consecuencia, es adecuado realizar el análisis de regresión entre esas variables.

Incluso, al determinar la **capacidad del modelo de explicar la variación de "y" por la relación de "x" e "y"** se encuentra que:

$$\%R^2 = R^2\ 100 = (r)^2 100 = (0.91806664)^2\ 100 = (0.8428463555)100 =$$

$$= 84.28463555$$

Además de que H0 es rechazada, el modelo explicaría el 84.28463555% de la variación de "y" debida a la relación lineal de "x" e "y", lo que indica un muy buen modelo. Solamente no explica el 15.71536445%.

Entonces la variable independiente (x) sería la CCI que es muy fácil de determinar, a diferencia del Índice de Masa Corporal que requiere de más información y cálculos por lo que sería la variable dependiente (y).

Entonces, los parámetros del modelo muestral son:

Ordenada al origen

$$b_0 = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} =$$

$$\frac{1{,}029.60 * 298{,}471.91 - 3{,}401.30 * 90{,}949.40}{40 * 298{,}471.91 - 11{,}568{,}841.69} =$$

$$\frac{307{,}306{,}679 - 309{,}346{,}194}{11{,}938{,}876.40 - 11{,}568{,}841.69} = -5.511687$$

Pendiente

$$b_1 = \frac{n(\sum_{i=1}^{n} x_i\, y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} =$$

$$= \frac{40 * 90{,}949.40 - 3{,}401.3 * 1{,}029.60}{40 * 298{,}471.91 - 11{,}568{,}841.69} = 0.3675264$$

En consecuencia, el modelo muestral de la relación entre la CCI y el IMC es:

$$\hat{y}_i = -5.511687 + 0.3675264 x_i$$

En caso de que se realizaran predicciones, solamente serán válidas dentro del intervalo dado por el valor mínimo y el valor máximo de los datos originales del CCI, que son 66.70 cm y 126.5 cm respectivamente (ver tabla de datos originales, corresponden a los valores mínimo y máximo), que son con los que se determinaron los parámetros del modelo. En valores de CCI menores a 66.70 cm o mayores a 126.50 cm no es adecuado, puesto que se desconoce el comportamiento de las variables.

14.- Razones y tasas

Las razones y las tasas son formas prácticas de reportar, comparar información, fortalecer la toma de decisiones, cantidad de insumos requeridos o existentes, aspectos que pueden ser requeridos en una epidemia o en la gestión de un desastre. Se aplican a conteos (datos cuantitativos discretos).

Tasa

Las tasas, se utilizan para referirse a la ocurrencia de algún evento. El numerador del cociente de una tasa forma parte de su denominador. El propósito del multiplicador *k* (base), es evitar que los resultados abarquen números muy pequeños, facilitar su comprensión. El valor elegido para*k*, depende de la magnitud del numerador y la del denominador, de la posible comparación con otras tasas o de las necesidades correspondientes. La tasa refleja la cantidad de algún evento referida a un número k, y se expresa:

$$Tasa = \left(\frac{a}{a+b}\right)k$$

Donde:
a = frecuencia de ocurrencia del evento 1
b = frecuencia de ocurrencia del evento 2
a + b = frecuencia total de los eventos 1 y 2
k = base: 10, 100, 1, 000, 10, 000, o algún otro múltiplo de 10

Cálculo de una Tasa

Ejemplo 95. ¿Cuál fue la tasa de mujeres en la alcaldía de Benito Juárez en el primer semestre del 2015 por cada 1,000 habitantes?
Total de hombres 193,007; total de mujeres 224,409, (Anuario Estadístico y Geográfico de la Ciudad de México 2017 en www.datatur.sectur.gob.mx/ITxEF_Docs/CDMX_ANUARIO_PDF.pdf)

$$Tasa = \left(\frac{a}{a+b}\right)k = \left(\frac{Mujeres}{Mujeres + Hombres}\right)1{,}000 =$$

$$= \left(\frac{224{,}409}{224{,}409 + 193{,}007}\right)1{,}000 = 5{,}37.61 \approx 538\ Mujeres$$

En conclusión, hubo casi 538 mujeres por cada 1,000 habitantes de la alcaldía de Benito Juárez. Lo anterior indica que las mujeres eran más del 50%.

Ejemplo 96. ¿Cuál fue la tasa de hombres, en la alcaldía de Iztapalapa en el primer semestre del 2015 por cada 10,000 habitantes?
Total de hombres 878,365; Total de mujeres 949,503 (Anuario Estadístico y Geográfico de la Ciudad de México 2017 en www.datatur.sectur.gob.mx/ITxEF_Docs/CDMX_ANUARIO_PDF.pdf)

$$Tasa = \left(\frac{a}{a+b}\right)k = \left(\frac{Hombres}{Mujeres + Hombres}\right)10{,}000 =$$

$$= \left(\frac{878{,}365}{949{,}503 + 878{,}365}\right)10{,}000 = 4{,}805.41 \approx 4{,}800\ Hombres$$

En conclusión, hubo casi 4,800 hombres por cada 10,000 habitantes de la alcaldía de Iztapalapa. Lo anterior indica que menos del 50% eran hombres.

Razón

En una razón, la frecuencia del evento ubicada en el numerador del cociente no forma parte del denominador, o sea es independiente. Es posible calcular la razón del número de personas que requieren atención médica con relación al número de médicos existentes o la razón de cantidad de personas afectadas por una inundación en relación a la cantidad de personas no afectadas por la inundación. La cantidad de camas en la unidad de cuidados intensivos con relación a la cantidad de personas que requieren cuidados intensivos.

$$Razón = \left(\frac{c}{d}\right)k$$

Donde:
- c = frecuencia de ocurrencia del evento 1
- d = frecuencia de ocurrencia del evento 2
- k = base. Los valores utilizados con mayor frecuencia son 1 y 100.

Cálculo de una Razón

Ejemplo 97. ¿Cuál fue la razón de mujeres por vivienda en el municipio de Jojutla en el estado de Morelos?
Total de mujeres 28,685; total de viviendas en Jojutla 20,839 (https://mexico.pueblosamerica. com/l/munest/morelos/jojutla)

$$Razón = \left(\frac{c}{d}\right) k = \left(\frac{Mujeres}{Viviendas}\right) 1 = \left(\frac{28{,}685}{20{,}839}\right) 1 = 1.38\ Mujeres/Vivienda$$

ó

$$Razón = \left(\frac{c}{d}\right) k = \left(\frac{Mujeres}{Vivienda}\right) 100 = \left(\frac{28{,}685}{20{,}839}\right) 100 =$$

$$= 138\ Mujeres\ por\ cada\ 100\ Casas$$

Sin embargo, en caso de que fueran viviendas por mujer se tendría:

$$Razón = \left(\frac{c}{d}\right) k = \left(\frac{Viviendas}{Mujeres}\right) 1 = \left(\frac{20{,}839}{28{,}685}\right) 1 = 0.73\ de\ Vivienda/Mujer$$

ó

$$Razón = \left(\frac{c}{d}\right) k = \left(\frac{Viviendas}{Mujeres}\right) 100 = \left(\frac{20{,}839}{28{,}685}\right) 100 =$$

$$= 73\ Viviendas\ por\ cada\ 100\ Mujeres$$

Ejemplo 98. ¿Cuál fue la razón del número de habitantes por bombero en la Ciudad de México en el 2015?

Total habitantes CdMx 4,687,003 (Anuario Estadístico y Geográfico de la Ciudad de México 2017 en www.datatur.sectur.gob.mx/ITxEF_Docs/CDMX_ANUARIO_PDF.pdf) y total de bomberos 2015, 1,900 (https://es.wikipedia.org/wiki/Heroico_Cuerpo_de_Bomberos_de_la_Ciudad_de_M%C3%A9xico).

$$Razón = \left(\frac{c}{d}\right) k = \left(\frac{Población\ CdMex}{Bomberos}\right) 1 = \left(\frac{4{,}687{,}003}{1{,}900}\right) 1 =$$

$$= 2{,}466.84 \approx 2{,}477\ Habitantes/Bombero$$

ó

$$Razón = \left(\frac{c}{d}\right) k = \left(\frac{Población\ CdMex}{Bomberos}\right) 100 = \left(\frac{4{,}687{,}003}{1{,}900}\right) 100 =$$

$$= 246{,}684.37 \approx 246{,}685\ Habitantes\ por\ cada\ 100\ Bomberos$$

Sin embargo, en caso de que fueran bomberos por habitante se tendría:

$$Razón = \left(\frac{c}{d}\right)k = \left(\frac{Bomberos}{Población\ CdMex}\right)1 = \left(\frac{1{,}900}{4{,}687{,}003}\right)1 =$$

$$= 0.00040538\ de\ Bombero/Habitante$$

ó

$$Razón = \left(\frac{c}{d}\right)k = \left(\frac{Bomberos}{Población\ CdMex}\right)100 = \left(\frac{1{,}900}{4{,}687{,}003}\right)100 =$$

$$= 0.041\ Bomberos\ por\ cada\ 100\ Habitantes$$

15.- Números índice

Los números índice se calculan con variables cuantitativas y sirven para hacer comparaciones en el valor de una de ellas, en dos o más situaciones diferentes, en relación al tiempo (*t*) o entre zonas geográficas (z); la más antigua o primaria es llamada base (*0*) la referencia para el periodo o zona geográfica de comparación (*t* o *z*).

Los podemos emplear en diferentes áreas del conocimiento, como una forma de visualizar los cambios generados en alguna variable de interés, ya sea a través del tiempo o de diferentes zonas geográficas. Los números índice enfocados a la comparación temporal o espacial pueden aportar información que de otra manera podría quedar "oculta" de diversos procesos.

Es imprescindible que sea la misma variable en el tiempo o zona geográfica base y la que servirá para la comparación pasado el tiempo (t) o en una zona geográfica (z) diferente, la medición de esta debe ser en las mismas unidades, para que el cociente entre ellas sea adimensional y pueda convertirse en porcentaje.

La comparación se realiza entre momentos diferentes del tiempo o zonas geográficas. La cantidad base lleva el subíndice *0* y el valor de comparación *t (*tiempo*)* o *z* (zona geográfica) y en ambos casos *x* es el valor de la variable en estudio. El cociente se multiplica por 100 ya que el numero índice se expresa en porcentaje. De manera general se tiene que:

$$I_{t\,ó\,z} = \frac{x_{t\,ó\,z}}{x_0}\,100$$

Donde:

$I_{t\,ó\,z}$= número índice de cambio de la variable *x* del tiempo o zona *0* al tiempo *t* o zona de comparación *z*, expresado en porcentaje

$X_{t\,ó\,z}$= valor de la variable al tiempo *t* o zona *z* diferente al valor base

X_0 = valor base de la variable en el tiempo *t* o zona *z*

El número índice puede tomar valores tales que, , si es menor indica que en relación con el tiempo *t* o zona *z* hubo una reducción en el valor de la variable, si es mayor señala un incremento y si es igual a 100 que no existe cambio en el valor de la variable.

La clasificación de los números índices se efectúa por la naturaleza de las variables a estudiar. Pueden ser simples o complejos, y en el caso de los complejos, por la importancia relativa de cada componente en el conjunto.

a. Los números índices simples, son los que estudian, de una sola variable de interés, el desenvolvimiento de un proceso.

b. Los números índices complejos sin ponderar, se utilizan cuando en el desenvolvimiento de varias variables todas son consideradas igualmente importantes para el proceso en estudio, entonces se les otorga el mismo pesorelativo.

c. Los númerosíndicescomplejosponderados, se usan cuando en el desenvolvimiento de varias variables cada una representa un peso relativo diferente, por lo que a cada una se les asigna un valor de ponderación.

En términos generales los números índice deben satisfacer los criterios (propiedades) de existencia, identidad, inversión, circularidad y proporcionalidad. Todas estas propiedades son cumplidas por los números índices simples, sin embargo, no es así para los índices complejos. Ahora bien, los números índice pueden ser vinculados en serie (la referencia base es fija) o en cadena (referencia variable).

Cálculo del Número Índice Simple

Ejemplo 99.- ¿Cómo se ha comportado porcentualmente del año 2017 al 2018 la cantidad de sismos de magnitud 2 y 3 en la alcaldía de Coyoacán de la Ciudad de México?

ALCALDIA DE COYOACÁN	
AÑO	**MAGNITUD**
2017	2.4
2017	2.6
2017	2.4
2017	2.4
2018	2.5
2018	2.1

Fuente: Datos modificados de http://www2.ssn.unam.mx8080/catalogo

En el año 2017 hubo 4 sismos y en el 2018, 2 sismos. La base es el año 2017. Entonces el número índice será:

$$I_{2018} = \frac{x_{2018}}{x_0}\ 100 = \frac{2}{4}\ 100 = 50\%$$

En consecuencia, la cantidad de sismos de magnitud 2 y 3 en 2018 es el 50% de los ocurridos en el 2017 en la alcaldía de Coyoacán.

Ejemplo 100.- ¿Cuál es el cambio porcentual de los precios para 2018 de un preservativo de cierta marca en la alcaldía de Tlalpan y en la de Gustavo A. Madero (GAM), si en la primera el valor es de $7.50 y en la segunda de $8.5, y la base para la comparación es la segunda alcaldía?

$$I_{GAM} = \frac{x_{GAM}}{x_0} 100 = \frac{8.50}{7.50} 100 = 113.4\%$$

El preservativo de la marca estudiada costaba 113.4% en la alcaldía de la GAM comparada con la alcaldía de Tlalpan, a la que le corresponde el 100%. En consecuencia, en la GAM el costo era 13.4% (113.4% -100% = 13.4%) mayor que en la de Tlalpan. Por lo tanto, en el Norte de la CDMX el preservativo era más caro que en el Sur.

16.- Prueba para independencia u homogeneidad

Esta prueba es para datos categóricos en que los conteos son la forma de cuantificarlos. Trabaja con dos variables diferentes presentadas en lo que se conoce como tablas de contingencia de 2 entradas; las Frecuencias Esperadas son menores; los renglones corresponden a una variable y las columnas a la otra. Comúnmente son utilizadas en el análisis de encuestas.

TABLA DE CONTINGENCIA*				
	Variable B			
	Categoría 1	Categoría 2	Categoría 3	Total Renglón
Variable A Categoría 1	F_{11}	F_{12}	F_{13}	$F_{11} + F_{12} + F_{13}$
Variable A Categoría 2	F_{21}	F_{22}	F_{23}	$F_{21} + F_{22} + F_{23}$
Total Columna	$F_{11} + F_{21}$	$F_{12} + F_{22}$	$F_{13} + F_{23}$	$F_{11} + ... + F_{23}$
				Gran Total

*El número de categorías depende de las necesidades del estudio.

La última celda de la tabla es n_{jm}. En este caso es $\mathbf{F_{23}}$

Prueba de Independencia

Se pone a prueba la hipótesis nula (H_0) de que la variable localizada en los renglones es independiente de la ubicada en las columnas o lo que es igual a que entre ambas no hay asociación, se aplica a una sola muestra.

No se requiere que la población de la que provienen los datos tenga algún tipo de distribución particular; la frecuencia esperada (E) debe ser por lo menos de 5, esto no se aplica a ninguna frecuencia observada.

H_0: las variables son independientes

H_1: las variables son dependientes

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia, establecer la regla de decisión y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1,

es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Si el *valor P*< , ó bien, $\chi^2_{calculada} > \chi^2_{,g.l.=1}$, H_0 se rechaza. Es una prueba de cola derecha.

Estadístico de Prueba

$$\chi^2 = \sum_{celda\ 11}^{celda\ n_{jm}} \frac{(O-E)^2}{E}$$

Donde:
^{2} = ji- cuadrada
O = frecuencias observadas
E = frecuencias esperadas
n = números de la última celda

Frecuencia Esperada (E):

$$E = \frac{(total\ renglón)(total\ Columna)}{(Gran\ Total)}$$

Grados de Libertad:

g.l. = (r - 1)(c - 1)

Donde:
r = renglones
c = columnas

Cálculo de la Prueba de Independencia

Ejemplo 101.- La opinión de los habitantes de la CdMx sobre la veracidad de las noticias en la radio relativas a las consecuencias del sismo ocurrido en septiembre del 2017 en dicha ciudad, es independiente de su sexo biológico. Ejemplo 101.- La opinión de los habitantes de la CdMx sobre la veracidad de las noticias en la radio relativas a las consecuencias del sismo ocurrido en septiembre del 2017 en dicha ciudad, es independiente de su sexo biológico.

TABLA DE CONTINGENCIA				
Sexo biológico	Estaciones de Radio (ER)			
	ER1	ER2	ER3	Total
Mujeres	107	65	258	430
Hombres	378	129	130	637
Total Columna y Gran Total	485	194	388	1067

H_0: la opinión sobre la veracidad de diferentes estaciones de radio es independiente del sexo biológico de la persona

H_1: la opinión sobre la veracidad de diferentes estaciones de radio es dependiente del sexo biológico de la persona

Nivel de significancia: = 0.05

Regla de decisión: si el *valor P*< , ó bien, si $\chi^2_{calculada} > \chi^2_{0.05,g.l.=1}$, H0 se rechaza

Cálculo de las Frecuencias Esperadas *E*:

Todas las columnas con el renglón 1, para obtener E_{11}, E_{12} y E_{13}. Posteriormente todas las columnas con el renglón 2, para obtener E_{21}, E_{22} y E_{23}. Es necesario conservar las cifras a la derecha del punto. Solamente se detallan para cada uno de los renglones, las 2 primeras frecuencias esperadas, pero en este caso debe haber 6 ya que son 2 renglones y 3 columnas.

Renglón 1:

$$E_{11} = \frac{(total\ renglón_1)(total\ Columna_1)}{(Gran\ Total)} = \frac{430(485)}{1067} = \frac{208550}{1067} = 195.45$$

$$E_{12} = \frac{(total\ renglón_1)(total\ Columna_2)}{(Gran\ Total)} = \frac{430(194)}{1067} = \frac{83420}{1067} = 78.18$$

Renglón 2:

$$E_{21} = \frac{(total\ renglón_2)(total\ Columna_1)}{(Gran\ Total)} = \frac{637(485)}{1067} = \frac{308945}{1067} = 289.54$$

$$E_{22} = \frac{(total\ renglón_2)(total\ Columna_2)}{(Gran\ Total)} = \frac{637(194)}{1067} = \frac{123578}{1067} = 115.81$$

TABLA DE CONTINGENCIA CON FRECUENCIAS ESPERADAS				
Sexo biológico	Estaciones de Radio (ER)			
	ER1	ER2	ER3	Total
Mujeres	107	65	258	430
Frecuencia esperada	195.45	78.18	156.36	
Hombres	378	129	130	637
Frecuencia esperada	289.54	115.81	231.63	
Total Columna y Gran Total	485	194	388	1067

Cálculo del estadístico de prueba:

$$\chi^2 = \sum_{celda\ 11}^{celda\ 23} \frac{(O-E)^2}{E} = \frac{(107-195.45)^2}{195.45} + \cdots + \frac{(130-231.63)^2}{231.63} =$$

=40.02+…+44.59= 181.41

Conclusión Estadística: como el *valor P= 0.000* **< α = 0.05, ó bien, =181.41 > $\chi^2_{calculada}$ =181.41 >$\chi^2_{0.05,g.l.=1}$ = 3.841, H_0 se rechaza.**

Conclusión Contextual: La opinión de las personas sobre las consecuencias del sismo de septiembre de 2017 en la CdMx, depende de si es hombre o mujer.

Ejemplo 102.- ¿El Índice de Masa Corporal (IMC) es independiente del sexo biológico de la persona?

TABLA DE CONTINGENCIA					
Sexo biológico	IMC				
	Desnutrido	Bien Nutrido	Sobrepeso	Obesidad	Total
Mujeres	257	1294	665	1179	3395
Hombres	850	987	1244	998	4079
Total Columna y Gran Total	1107	2281	1909	2177	7474

H_0: el IMC es independiente de sexo biológico de la persona

H_1: el IMC es dependiente del sexo biológico de la persona

Nivel de significancia: = 0.05

Regla de decisión: si el *valor P*< , ó bien, si $^2_{calculada} > {}^2_{0.05,g.l.=1}$, H_0 se rechaza

Cálculo de las Frecuencias Esperadas *E*:

Todas las columnas con el renglón 1, para obtener E_{11}, E_{12}, E_{13} y E_{14}. Posteriormente todas las columnas con el renglón 2, para obtener E_{21}, E_{22}, E_{23} y E_{24} . Es necesario conservar las cifras a la derecha del punto. Solamente se detallan para cada uno de los renglones, las 2 primeras frecuencias esperadas, pero en este caso debe haber 8 ya que son 2 renglones y 4 columnas.

Renglón 1:

$$E_{11} = \frac{(total\ renglón_1)(total\ Columna_1)}{(Gran\ Total)} = \frac{3395(1107)}{7474} = \frac{3758265}{7474} = 502.84$$

$$E_{12} = \frac{(total\ renglón_1)(total\ Columna_2)}{(Gran\ Total)} = \frac{3395(2281)}{7474} = \frac{7743995}{7474} = 1036.12$$

Renglón 2:

$$E_{21} = \frac{(total\ renglón_2)(total\ Columna_1)}{(Gran\ Total)} = \frac{4079(1107)}{7474} = \frac{4515453}{7474} = 604.15$$

$$E_{22} = \frac{(total\ renglón_2)(total\ Columna_2)}{(Gran\ Total)} = \frac{4079(2281)}{7474} = \frac{9304199}{7474} = 1244.87$$

TABLA DE CONTINGENCIA CON FRECUENCIAS ESPERADAS					
Sexo biológico	IMC				
	Desnutrido	Bien Nutrido	Sobrepeso	Obesidad	Total
Mujeres	257	1294	665	1179	3395
Frecuencia esperada	502.84	1036.12	867.14	988.88	
Hombres	850	987	1244	998	4079
Frecuencia esperada	604.15	1244.87	1041.85	1188.11	
Total Columna y Gran Total	1107	2281	1909	2177	7474

Cálculo del estadístico de prueba:

$$\chi^2 = \sum_{celda\ 11}^{celda\ 24} \frac{(O-E)^2}{E} = \frac{(257-502.84)^2}{502.84} + \cdots + \frac{(998-1188.11)^2}{1188.11} =$$

$$= 120.19 + \cdots + 30.41 = 491.12$$

Conclusión Estadística: como el *valor P= 0.000*< = 0.05, ó bien $^2_{calculada}$=491.12> $^2_{0.05,g.l.=1}$, H_0 se rechaza.

Conclusión Contextual: El IMC de las personas estudiadas es dependiente de si su sexo biológico es hombre o mujer.

Prueba de Homogeneidad

En la prueba de homogeneidad se utiliza el mismo procedimiento que en la prueba de independencia, sin embargo, en este caso, la Hipótesis nula (H_0) es sobre las proporciones de las diversas categorías de las variables en estudio y se tienen varias muestras independientes de diferentes poblaciones.

Cálculo de la Prueba de Homogeneidad

Ejemplo 103.- La opinión de los habitantes de las alcaldías estudiadas de la CdMx, es homogénea sobre si el sismo de septiembre del 2017 fue fuerte o débil.

TABLA DE CONTINGENCIA			
Opinión sobre la Intensidad del Sismo	Alcaldía de Ocurrencia del Sismo		
	Cuauhtémoc	Magdalena Contreras	Total
Opinión de que el Sismo fue Fuerte	490	90	580
Opinión de que el Sismo fue Débil	149	376	525
Total Columna y Gran Total	639	466	1105

H_0: Las proporciones de las opiniones (fuerte/débil) son iguales sin importar la alcaldía

H_a: Las proporciones de las opiniones (fuerte/débil) son diferentes en las alcaldías

Nivel de significancia: = 0.05

Regla de decisión: si el *valor P*< , ó bien, $^2_{calculada}$> $^2_{0.05,g.l.=1}$, H_0 se rechaza

Cálculo de las Frecuencias Esperadas *E*:

Todas las columnas con el renglón 1, para obtener E_{11} y E_{12}; posteriormente todas las columnas con el renglón 2, para obtener E_{21} y E_{22} . Es necesario conservar las cifras a la derecha del punto. Se detalla el cálculo para cada uno de los renglones y las 2 columnas, por lo que las frecuencias esperadas son 4.

Renglón 1:

$$E_{11} = \frac{(total\ renglón_1)(total\ Columna_1)}{(Gran\ Total)} = \frac{580(639)}{1105} = \frac{370620}{1105} = 335.4$$

$$E_{12} = \frac{(total\ renglón_1)(total\ Columna_2)}{(Gran\ Total)} = \frac{580(466)}{1105} = \frac{270280}{1105} = 244.59$$

Renglón 2:

$$E_{21} = \frac{(total\ renglón_2)(total\ Columna_1)}{(Gran\ Total)} = \frac{525(639)}{1105} = \frac{335475}{1105} = 303.59$$

$$E_{22} = \frac{(total\ renglón_2)(total\ Columna_2)}{(Gran\ Total)} = \frac{525(466)}{1105} = \frac{244650}{1105} = 221.40$$

TABLA DE CONTINGENCIA CON FRECUENCIAS ESPERADAS			
Opinión sobre la Intensidad del Sismo	Alcaldía de Ocurrencia del Sismo		
	Cuauhtémoc	Magdalena Contreras	Total
Opinión de que el Sismo fue Fuerte	490	90	580
Frecuencia esperada	335.40	244.59	
Opinión de que el Sismo fue Débil	149	376	525
Frecuencia esperada	303.59	221.40	
Total Columna y Gran Total	639	466	1105

Cálculo del estadístico de prueba:

$$\chi^2 = \sum_{celda\ 11}^{celda\ 24} \frac{(O-E)^2}{E} = \frac{(490-335.40)^2}{335.40} + \cdots + \frac{(376-221.40)^2}{221.40} =$$

$$= 71.26 + \cdots + 107.95 = 355.62$$

Conclusión Estadística: como el *valor P= 0.000*< = 0.05, ó bien, $^{2}_{calculada}$=355.62>$\chi^2_{0.05,g.l.=1}$ =3.841 H_0 se rechaza.

Conclusión Contextual: Las opiniones de las personas depende de en qué alcaldía se haga la pregunta sobre si el sismo del 2017 fue fuerte o débil.

Ejemplo 104.- Las proporciones de la presencia de enfermedades en dos comunidades son iguales al momento del estudio. Se registró el número de personas que padecen la enfermedad 1 o la 2.

TABLA DE CONTINGENCIA			
Comunidad			
	Enfermedad 1	Enfermedad 2	Total
Comunidad 1	155	190	345
Comunidad 2	160	187	347
Total Columna y Gran Total	315	377	692

H_0: La proporción de las enfermedades 1 y 2 son iguales en ambas comunidades

H_a: La proporción de las enfermedades 1 y 2 no son iguales en ambas comunidades

Nivel de significancia: = 0.05

Regla de decisión: si el *valor P*< , ó bien, $\chi^2_{calculada}$>$\chi^2_{0.05,g.l.}$=1, H_0 se rechaza

Cálculo de las Frecuencias Esperadas *E*:

Todas las columnas con el renglón 1, para obtener E_{11} y E_{12}; posteriormente todas las columnas con el renglón 2, para obtener E_{21} y E_{22} . Es necesario conservar las cifras a la derecha del punto. Se detalla el cálculo para cada uno de los renglones y las 2 columnas, por lo que las frecuencias esperadas son 4.

Renglón 1:

$$E_{11} = \frac{(total\ renglón_1)(total\ Columna_1)}{(Gran\ Total)} = \frac{345(315)}{692} = \frac{108675}{692} = 157.04$$

$$E_{12} = \frac{(total\ renglón_1)(total\ Columna_2)}{(Gran\ Total)} = \frac{345(377)}{692} = \frac{130065}{692} = 187.95$$

Renglón 2:

$$E_{21} = \frac{(total\ renglón_2)(total\ Columna_1)}{(Gran\ Total)} = \frac{347(315)}{692} = \frac{109305}{692} = 157.95$$

$$E_{22} = \frac{(total\ renglón_2)(total\ Columna_2)}{(Gran\ Total)} = \frac{347(377)}{692} = \frac{130819}{692} = 189.04$$

TABLA DE CONTINGENCIA CON FRECUENCIAS ESPERADAS			
Comunidad	Enfermedad (E)		
	E1	E2	Total
Comunidad 1	155	190	345
Frecuencia esperada	157.04	187.95	
Comunidad 2	160	187	347
Frecuencia esperada	157.95	189.04	
Total Columna y Gran Total	315	377	692

Cálculo del estadístico de prueba:

$$\chi^2 = \sum_{celda\ 11}^{celda\ 22} \frac{(O-E)^2}{E} = \frac{(155-157.04)^2}{157.04} + \cdots + \frac{(187-189.04)^2}{189.04} =$$

$$= 0.02 + \cdots + 0.02 = 0.08$$

Conclusión Estadística: Como el *valor P= 0.7549*> = 0.05, ó bien, $\chi^2_{calculada}=0.08<\chi^2_{0.05,g.l.=1}=3.841$, H_0 no se rechaza.

Conclusión Contextual: Las proporciones de las enfermedades E1 y E2 son las mismas en ambas comunidades.

17.- Prueba del signo

Prueba no paramétrica, basada en las frecuencias de los signos positivos y negativos en los que el valor cero es descartado para determinar si son significativamente diferentes. Puede utilizarse en aseveraciones para probar datos apareados, datos nominales o relativas a la mediana de una sola población. Los datos se seleccionan aleatoriamente y no se requiere que cumplan con cierta distribución. En este procedimiento se contrastan las hipótesis.

Hipótesis:

Aseveraciones para datos apareados:

H_0: mediana = 0 no hay diferencia.

H_1: mediana 0 si existe diferencia (la mediana de las diferencias no es 0)

Aseveraciones para datos nominales

H_0: p = 0.5 no hay diferencia en las proporciones

H_1: p 0.5 si existe diferencia en las proporciones

Aseveraciones sobre la mediana de una sola población (Me)

H_0: Me = valor propuesto

H_1: Me < valor propuesto en H_0

Sin embargo esta hipótesis puede ser de dos colas.

El caso presentado es la hipótesis alternativa (H_1 ó H_a) de diferencia ( ) y no se incluyen los casos de mayor que (>) o menor que (<) en los dos primeros casos y de diferencia ( ) o mayor (>) que en el tercer caso, con la finalidad de facilitar el seguimiento, evitar confusiones y errores de interpretación.

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia, establecer la regla de decisión y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de

cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

Regla de decisión: en consecuencia el valor crítico para z se obtiene de la tabla de probabilidades de la distribución normal del escogido, si la prueba es de una cola o /2 si es de dos colas. Para n > 25, si $z_{calculada}$ z ó $z_{calculada}$ z /2 , ó bien, si el *valor P*< , H_0 se rechaza. Lo que también se aplica a 25, donde x = menor frecuencia de signos Valor Crítico de la tabla para la prueba del signo, H_0 se rechaza.

Estadístico de Prueba

n = número total de signos positivos más los negativos, sin contar los ceros.

Si n 25 el estadístico de prueba será el número de veces que ocurre el signo menos frecuente (x).

Si n > 25 el estadístico de prueba será:

$$z = \frac{(x + 0.5) - \left(\frac{n}{2}\right)}{\frac{\sqrt{n}}{2}}$$

Donde:

z = valores *z*
x = número de veces que ocurre el signo menos frecuente
n = total de signos positivos más el total de los signos negativos

Cálculo de la Prueba del Signo

Ejemplo 105.- La mediana de la magnitud de los sismos registrados en la CdMx de 2016 a 2019 en la escala de Richter es de 2 grados.

Magnitud de 89 sismos registrados:

No.	Magnitud	No.	Magnitud	No.	Magnitud	No.	Magnitud
1	1.0	7	1.5	14	1.7	21	1.8
2	1.0	8	1.5	15	1.7	22	1.8
3	1.3	9	1.5	16	1.7	23	1.8
4	1.3	10	1.5	17	1.7	24	1.8
5	1.3	11	1.5	18	1.7	25	1.8
6	1.4	12	1.5	19	1.7	26	1.9
		13	1.7	20	1.8	27	1.9

No.	Magnitud	No.	Magnitud	No.	Magnitud	No.	Magnitud
28	1.9	44	1.9	60	2.0	75	2.7
29	1.9	45	1.9	61	2.0	76	2.7
30	1.9	46	1.9	62	2.0	77	2.7
31	1.9	47	1.9	63	2.4	78	2.7
32	1.9	48	1.9	64	2.4	79	2.7
33	1.9	49	1.9	65	2.5	80	2.9
34	1.9	50	1.9	66	2.5	81	2.9
35	1.9	51	1.9	67	2.5	82	3.0
36	1.9	52	1.9	68	2.5	83	3.3
37	1.9	53	2.0	69	2.5	84	3.3
38	1.9	54	2.0	70	2.5	85	3.5
39	1.9	55	2.0	71	2.6	86	3.6
40	1.9	56	2.0	72	2.6	87	4.0
41	1.9	57	2.0	73	2.6	88	4.1
42	1.9	58	2.0	74	2.6	89	4.2
43	1.9	59	2.0				

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

H_0: Me = 2

H_1: Me < 2

Significancia: = 0.05

Número de valores mayores a 2 (signos positivos) = 27

Número de valores menores a 2 (signos negativos) = 52

Número de valores iguales a 2 (valores 0) = 10 se descartan

Regla de decisión: si el valor $- z_{calculada} > - z_{=0.05}$, ó bien, si el *valor P*< , H_0 se rechaza.

Estadístico de prueba, como n = 52 + 27 = 79 > 25 el estadístico de prueba será:

$$z = \frac{(x + 0.5) - \left(\frac{n}{2}\right)}{\frac{\sqrt{n}}{2}}$$

Cálculo del estadístico de prueba:

$$z = \frac{(27 + 0.5) - \left(\frac{79}{2}\right)}{\frac{\sqrt{79}}{2}} =$$

$$= \frac{27.5 - 39.5}{\frac{\sqrt{79}}{2}} = \frac{-12}{4.44409721} = -2.70021096$$

Conclusión estadística: $z_{calculada} = -2.70021096 > z_{=0.05} = -1.645$, H_0 se rechaza.

Conclusión contextual: La mediana de la magnitud de los sismos registrados en la CdMx entre 2016 y 2019 es menor a 2 grados Richter.

Ejemplo 106.- Una Nutrióloga desarrolló una dieta para bajar de peso a mujeres obesas cuyo Índice de Masa Corporal (IMC) era de 25 a 34.9. Al experimento se sometieron 25 mujeres durante 1 mes; los resultados obtenidos se presentan en la tabla que está a continuación y la pregunta es que si la dieta contribuye a bajar o no el IMC.

IMC de 25 mujeres al inicio y al final del experimento 30 días después

No.	IMC INICIAL	IMC FINAL	Signo
1	25.10	19.00	+
2	25.20	19.20	+
3	26.00	25.00	+
4	26.50	26.00	+
5	27.50	26.00	+
6	28.50	20.00	+
7	28.70	20.10	+
8	28.90	20.20	+
9	29.10	25.20	+
10	29.70	25.30	+
11	29.80	24.80	+
12	29.90	26.90	+
13	30.90	30.80	+
14	31.00	31.20	-
15	31.70	31.70	0
16	33.50	33.70	-
17	34.00	33.00	+
18	34.00	33.00	+
19	34.00	28.00	+
20	34.50	29.00	+
21	34.50	30.60	+
22	34.50	31.30	+

No.	IMC INICIAL	IMC FINAL	Signo
23	34.50	31.90	+
24	34.80	34.90	-
25	34.80	34.50	+

Fuente: Datos sin signo modificados del software estadístico statdisk 13.

H_0: Me = 0 No existe diferencia entre las medianas

H_1: Me 0 Si hay diferencia entre las medianas

Significancia: = 0.05

Número de signos positivos (bajaron de IMC) = 21

Número de signos negativos (subieron de IMC) = 3

Número de valores 0 = 1 se descarta

Regla de decisión: si el valor obtenido de los signos menos frecuentes (x) Valor Crítico de las tablas para la prueba del signo, H_0 se rechaza.

Estadístico de prueba, como n = 21 + 3 = 24 25 el estadístico de prueba será x = número menor de signos obtenidos = 3.

Cálculo del estadístico de prueba: x = 3

Conclusión estadística: como x = 3 <*Valor Crítico* $_{/2 = 0.05/2, n=24}$ = 6, H_0 se rechaza.

Conclusión contextual: la dieta experimental sí redujo el IMC de las mujeres en un mes.

18.- Prueba exacta de Fisher

Se utiliza en tablas de contingencia de 2x2 para variables cualitativas categóricas dicotómicas, en muestras pequeñas menor a 20 o cuando una o varias de las frecuencias esperadas es menor o igual a 5. Prueba la independencia o dependencia de las variables y totales de renglón y columna fijos.

H_0: las variables son independientes

H_1: las variables son dependientes

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia y seguidamente establecer la regla de decisión. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Si el *valor P* del contraste < , H_0 se rechaza.

Determinación del *Valor P*

TABLA DE CONTINGENCIA 2x2			
	Variable 2 nivel 1	Variable 2 nivel 2	Total
Variable 1 nivel 1	A	B	$a + b = n_1$
Variable 1 nivel 2	C	D	$c + d = n_2$
	$a + c = n_3$	$d + d = n_4$	$a + b + c + d = n$

Probabilidad de obtener los datos observados:

$$P = \frac{n_1!\, n_2!\, n_3!\, n_4!}{n!\, a!\, b!\, c!\, d!}$$

Para calcular el *valor P* del contraste, hay 2 formas, 1.- el *valor P* = suma de la probabilidad de los datos observados más las probabilidades menores o iguales a la probabilidad observada de los datos posibles. 2.- el *valor P* = suma de la probabilidad de los datos observados más las probabilidades más favorables de los datos posibles a la hipótesis alternativa, que los datos observados, esto es los valores más extremos. Lo anterior es válido para una prueba de hipótesis de una sola cola. Sin embargo, si se trata de dos colas la forma 2 de calcular el *valor P* se multiplica por dos.

Cálculo de la Prueba Exacta de Fisher

Ejemplo 107.- A 10 (una muestra de 3 mujeres y 7 hombres) personas de la CdMx se les preguntó, a finales del 2017, su opinión sobre la intensidad del sismo del 19 de septiembre, si había sido fuerte o débil.

H_0: las opiniones son independientes del sexo de las personas

H_1: las opiniones son dependientes del sexo de las personas

Significancia: = 0.05

Regla de decisión: si el *valorP* del contraste < , H_0 se rechaza

Tabla 1.- datos observados

SEXO	FUERTE	DÉBIL	TOTAL
MUJER	2	1	3
HOMBRE	4	3	7
TOTAL	6	4	10

Tabla 2.- datos posibles 1

SEXO	FUERTE	DÉBIL	TOTAL
MUJER	1	2	3
HOMBRE	5	2	7
TOTAL	6	4	10

Tabla 3.- datos posibles 2

SEXO	FUERTE	DÉBIL	TOTAL
MUJER	0	3	3
HOMBRE	6	1	7
TOTAL	6	4	10

Tabla 4.- datos posibles 3

SEXO	FUERTE	DÉBIL	TOTAL
MUJER	3	0	3
HOMBRE	3	4	7
TOTAL	6	4	10

Tabla resumen:

	a	b	c	d	n_1	n_2	n_3	n_4	n	Valor P
Observados	2	1	4	3	3	7	6	4	10	0.500
Posibles 1	1	2	5	2	3	7	6	4	10	0.300
Posibles 2	0	3	6	1	3	7	6	4	10	0.033
Posibles 3	3	0	3	4	3	7	6	4	10	0.167

Obtención del valor P del contraste:

Forma 1 cálculo *valor P* del contraste:

Valor P = 0.500+0.300+0.033+0.167 = 1.000

Forma 2 cálculo *valor P* del contraste:

Valor P = 0.500+0.033 = 0.533

Conclusión estadística: como *valor P* = 1 ó 0.533 > = 0.05, H_0 no se rechaza.

Conclusión contextual: la opinión de si el sismo de septiembre del 2017 fue fuerte o débil no depende del sexo de la persona entrevistada.

Ejemplo 107.- Se realizó un estudio sobre la eficacia del preservativo y a 12 mujeres, se les preguntó si cuando resultaron embarazadas su pareja había utilizado preservativo. Ellas no usaron preservativo ni método que estimulara la concepción.

H_0: los embarazos son independientes del uso de preservativo

H_1: los embarazos dependen del uso del preservativo

Significancia: = 0.05

Regla de decisión: si el *valor P* del contraste < , H_0 se rechaza

Tabla 1.- datos observados

	CON PRESERVATIVO	**SIN PRESERVATIVO**	**TOTAL**
EMBARAZO	0	7	7
NO EMBARAZO	5	0	5
TOTAL	5	7	12

Tabla 2.- datos posibles 1

	CON PRESERVATIVO	**SIN PRESERVATIVO**	**TOTAL**
EMBARAZO	1	6	7
NO EMBARAZO	4	1	5
TOTAL	5	7	12

Tabla 3.- datos posibles 2

	CON PRESERVATIVO	SIN PRESERVATIVO	TOTAL
EMBARAZO	2	5	7
NO EMBARAZO	3	2	5
TOTAL	5	7	12

Tabla 4.- datos posibles 3

	CON PRESERVATIVO	SIN PRESERVATIVO	TOTAL
EMBARAZO	3	4	7
NO EMBARAZO	2	3	5
TOTAL	5	7	12

Tabla 5.- datos posibles 4

	CON PRESERVATIVO	SIN PRESERVATIVO	TOTAL
EMBARAZO	4	3	7
NO EMBARAZO	1	4	5
TOTAL	5	7	12

Tabla 6.- datos posibles 5

	CON PRESERVATIVO	SIN PRESERVATIVO	TOTAL
EMBARAZO	5	2	7
NO EMBARAZO	0	5	5
TOTAL	5	7	12

Tabla resumen:

	a	b	c	d	n_1	n_2	n_3	n_4	n	Valor *P*
Observados	0	7	5	0	7	5	5	7	12	0.0013
Posibles 1	1	6	4	1	7	5	5	7	12	0.0442
Posibles 2	2	5	3	2	7	5	5	7	12	0.2652
Posibles 3	3	4	2	3	7	5	5	7	12	0.4419
Posibles 4	4	3	1	4	7	5	5	7	12	0.2210
Posibles 5	5	2	0	5	7	5	5	7	12	0.02656

Obtención del *valor P* del contraste:

Forma 1 cálculo *valor P* del contraste:

Valor *P* = 0.0013

Forma 2 cálculo *valor P* del contraste:

Valor *P* = 0.0013

Conclusión estadística: como *valor P* = 0.0013 < = 0.05, H_0 se rechaza.

Conclusión contextual: el embarazo es dependiente de si la pareja de la mujer utiliza preservativo.

19.- Prueba de McNemar

Es para tablas de contingencia de 2x2 de variables nominales categóricas dicotómicas con datos apareados o muestras dependientes, que son aquellos en los que, si los de una muestra se pueden utilizar para determinar los de la otra muestra. Utiliza conteos de frecuencias de datos apareados (no independientes) para probar las hipótesis siguientes:

H_0: las frecuencias de las categorías diferentes ocurren en la misma proporción

H_1: las frecuencias de las categorías diferentes no ocurren en la misma proporción

La tabla de contingencia sería:

Tabla de Contingencia con Conteo de Frecuencias para Datos Apareados			
Efectiva		Variable X	
		No efectiva	
	Efectiva	a	b
Variable Y			
	No efectiva	c	d

Al aplicar la prueba de McNemar es necesario verificar que los datos muestrales sean apareados o conteos de frecuencias para una variable nominal dicotómica: Así mismo cada dato (frecuencia) que sea posible clasificarla de dos maneras: 1) según la categoría a la que pertenece (Variable X ó Variable Y) y 2) según otra categoría con dos valores posibles (Efectiva, No Efectiva; Positiva, Negativa). En tablas como la anterior las frecuencias deben ser tales que b + c 10. La prueba de McNemar únicamente se basa en los pares de datos discordantes, los que indican que hubo un cambio por el uso de la variable X ó Y, esto es, de Efectiva a No efectiva y de No Efectiva a Efectiva lo cual corresponde a las literales b y c.

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia, establecer la regla de decisión y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Si el *valor P* < , ó bien, si $\chi^2_{calculada} > \chi^2_{\alpha,g.l.=1}$, H_0 se rechaza. La prueba de McNemar es de cola derecha.

Estadístico de Prueba

Al poner a prueba las Hipótesis mencionadas, la prueba de McNemar solamente utiliza las frecuencias discordantes lo que implica seleccionar las de b y c, entonces la ecuación del estadístico de prueba queda:

$$\chi^2 = \frac{(|\, b - c \,| - 1)^2}{b + c}$$

Donde:

^{2} = ji - cuadrada

b = frecuencias de la variable X de No Efectiva a Efectiva en la variable Y

c = frecuencias de la variable X de Efectiva a No efectiva en la variable Y

Grados de Libertad:

g.l. = 1

Cálculo de la Prueba de McNemar

Ejemplo 109.- A diversas personas de la CDMX se les impartió un curso sobre qué hacer en caso de sismos y después de uno se le preguntó a la misma gente, si el curso lo había hecho sentir seguro:

Tabla de Contingencia con Conteo de Frecuencias para Datos Apareados			
Seguro		Antes del curso	
		No seguro	
	Seguro	20	90
Después del curso			
	No seguro	5	34

H_0: las proporciones de las categorías antes y después del curso son iguales

H_1: las proporciones de las categorías antes y después del curso no son iguales

Significancia: = 0.05

Regla de decisión: si $^2_{calculada} > {}^2_{0.05,g.l.=1}$, ó si el *valor P* < = 0.05, H_0 se rechaza

Cálculo del estadístico de prueba:

$$\chi^2 = \frac{(|\, b - c \,| - 1)^2}{b + c} = \frac{(|\, 90 - 5 \,| - 1)^2}{90 + 5} =$$

$$= \frac{-84^2}{95} = \frac{7056}{95} = 74.274$$

Conclusión estadística: como $\chi^2_{calculada}$=74.274 > $\chi^2_{0.05,g.l.=1}$ =3.841, ó bien, el *valor P = 0.000* < = 0.05, H_0 se rechaza.

Conclusión contextual: Las proporciones no son iguales, indican que el curso si tiene efecto sobre las personas

Ejemplo 110.- Se desarrolló una pomada para la limpieza facial alternativa a la tradicional, para adolescentes con problemas médicos de erupciones en la cara, pero se requiere saber si el nuevo método es igual o no al de siempre. Se realizó un experimento con varios adolescentes, diariamente se limpiaron la mitad de la cara con el nuevo sistema y la otra mitad con el sistema de siempre.

Tabla de Contingencia con Conteo de Frecuencias para Datos Apareados			
Limpia		Método Experimental	
		No limpia	
	Limpia	8	21
Método Tradicional			
	No limpia	23	6

H_0: las proporciones de las categorías de los dos sistemas de limpieza facial son iguales

H_1: las proporciones de las categorías de los dos sistemas de limpieza facial no son iguales

Significancia: = 0.05

Regla de decisión: si $^2_{calculada} > {}^2_{0.05,g.l.=1}$, ó si el *valor P* < = 0.05, H_0 se rechaza

Cálculo del estadístico de prueba:

$$\chi^2 = \frac{(|\, b - c \,| - 1)^2}{b + c} = \frac{(|\, 21 - 23 \,| - 1)^2}{21 + 23} =$$

$$= \frac{1^2}{44} = \frac{1}{44} = 0.023$$

Conclusión estadística: como <ó bien, el *valor P = 0.880* > = 0.05, H_0 no se rechaza.
Conclusión contextual: Las proporciones son iguales, señalan que el método experimental no es mejor para la limpieza facial que el método tradicional.

20.- Prueba de Wilcoxon

Es una prueba no paramétrica, tiene dos variaciones, una para datos apareados y otra, que aquí se detalla es la prueba de la suma de rangos de Wilcoxon para datos independientes, utilizada para dos grupos de datos muestrales independientes, no relacionados, asociados o apareados e intenta probar que la mediana de ambas poblaciones es igual. Así mismo, ambas muestras pueden ser de distribución libre. Las hipótesis serían:

H_0: las dos muestras provienen de poblaciones con medianas iguales

H_1: las dos muestras provienen de poblaciones con medianas diferentes

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia, establecer la regla de decisión y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Si el *valor P* < , ó bien, si $z_{calculada} > z_{/2}$, H_0 se rechaza. Es una prueba de dos colas.

Estadístico de Prueba

El tamaño de cada muestra debe ser n > 10 datos de lo contrario hay que utilizar tablas especiales. Ambas poblaciones pueden ser de libre distribución:

a. ambas muestras se mezclan formando una sola.
b. ordenar la muestra resultante de "a" de menor a mayor.
c. en la muestra resultante del inciso "b" cada valor en reemplazado por un rango.
d. el rango, en la muestra resultante de "b", es progresivo de menor a mayor. El rango menor es 1.

e. si hay datos con el mismo valor (empates) en la muestra "d" se les asigna el promedio de los rangos implicados.

f. se separan las dos muestras implicadas en "e" y se calcula la suma de los rangos de cada una, por lo que se obtiene R1 = suma de los rangos de la muestra 1 y R2 = suma de los rangos de la muestra 2.

g. se calcula el estadístico de prueba "z".

Donde:

R_1 = suma de rangos de la muestra con tamaño n_1

$$\mu_R = \frac{n_1(n_1+n_2+1)}{2}$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$$

n_1 = tamaño de la muestra 1 a partir de la que se calcula la suma de rangos

n_2 = tamaño de la muestra 2 a partir de la que se calcula la suma de rangos

Cálculo de la Prueba de Wilcoxon

Ejemplo 111.- En los sismos ocurridos en la Ciudad de México y en la de Toluca en el 2016 de 1 grado Richter o mayor, se podría investigar si las poblaciones de ambos grupos de sismos tienen o no el mismo valor para la mediana.

Magnitud de 26 sismos.

No.	Magnitud
Ciudad de México	
1	1.8
2	1.5
3	2
4	2.6
5	2.2
6	3.5
7	1.0
8	2.9
9	2.1

No.	Magnitud
10	2.4
11	1.3
12	2.2
Ciudad de Toluca	
13	2.3
14	2.3
15	1.8
16	3.3
17	2.6
18	1.7

No.	Magnitud
19	2.2
20	2.1
21	1.5
22	1.4
23	1.8
24	1.0
25	1.8
26	1.9

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

H_0: las dos muestras provienen de poblaciones con medianas iguales

H_1: las dos muestras provienen de poblaciones con medianas diferentes

Significancia: = 0.05

Regla de decisión:si el *valor P* < , ó bien, $z_{calculada} > z_{/2}$, H_0 se rechaza.

Rango definitivo para cada sismo:

No.	Magnitud	Rango Inicial	Rango Definitivo
Ciudad de México			
1	1.8	9	9.5
2	1.5	5	5.5
3	2	13	13
4	2.6	22	22.5
5	2.2	16	17
6	3.5	26	26
7	1	2	1.5
8	2.9	24	24
9	2.1	15	14.5
10	2.4	21	21
11	1.3	3	3
12	2.2	17	17
Ciudad de Toluca			
13	2.3	19	19.5
14	2.3	20	19.5
15	1.8	10	9.5
16	3.3	25	25
17	2.6	23	22.5
18	1.7	7	7
19	2.2	18	17
20	2.1	14	14.5
21	1.5	6	5.5
22	1.4	4	4
23	1.8	11	9.5
24	1	1	1.5
25	1.8	8	9.5
26	1.9	12	12

Cálculo del estadístico de prueba:

Ciudad de México: $n_1 = 12$, $R_1 = 174.50$

Ciudad de Toluca: $n_2 = 14$, $R_2 = 176.50$

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{12(12 + 14 + 1)}{2} = 162$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{12 * 14 * (12 + 14 + 1)}{12}} = 19.44$$

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{174.50 - 162}{19.44} = 0.64$$

Conclusión estadística: $z_{calculada} = 0.64 < z_{/2= 0.05/2} \pm 1.96$, H_0 no se rechaza.

Conclusión contextual: hay evidencia para asegurar que las dos poblaciones de sismos tienen el mismo valor de la mediana.

Ejemplo 112.- Se sospecha que los hombres y las mujeres no tienen la misma mediana en cuanto al nivel de colesterol.

IMC de 27 hombres y mujeres

Colesterol	
No.	**Mujeres**
1	264
2	181
3	267
4	384
5	98
6	62
7	126
8	89
9	531
10	130

Colesterol	
11	175
12	44
13	8
14	112
No.	**Hombres**
15	522
16	127
17	740
18	49
19	230

Colesterol	
20	316
21	590
22	466
23	121
24	578
25	78
26	265
27	250

Fuente: Datos modificados del software estadístico statdisk 13.

H_0: las dos muestras provienen de poblaciones con medianas iguales

H_1: las dos muestras provienen de poblaciones con medianas diferentes

Significancia: = 0.05

Regla de decisión: si el *valor P* < , ó bien, $z_{calculada} > z_{/2}$, H0 se rechaza.

Rangos para el colesterol de hombres o mujeres:

No.	Colesterol	Rango
Mujeres		
1	264	17
2	181	14
3	267	19
4	384	21
5	98	7
6	62	4
7	126	10
8	89	6
9	531	24
10	130	12
11	175	13
12	44	2
13	8	1
14	112	8
Hombres		
15	522	23
16	127	11
17	740	27
18	49	3
19	230	15
20	316	20
21	590	26
22	466	22
23	121	9
24	578	25
25	78	5
26	265	18
27	250	16

Cálculo del estadístico de prueba:

Mujeres: $n_1 = 14, R_1 = 158$

Hombres: $n_2 = 13, R_2 = 220$

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{14(14 + 13 + 1)}{2} = 196$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{14 * 13 * (14 + 13 + 1)}{12}} = 20.61$$

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{158 - 196}{20.61} = -1.84$$

Conclusión estadística: $z_{calculada} = -1.84 < z_{/2} = 0.05/2 = \pm 1.96$, H_0 no se rechaza.

Conclusión contextual: en las poblaciones de mujeres y hombres la mediana del colesterol tiene el mismo valor.

21.- Prueba de Kruskal-Wallis

Es una prueba no paramétrica, también llamada prueba *H*, para al menos de tres muestras o mayor número de ellas pero independientes; utiliza los rangos de los datos de las muestras pueden ser de distribución libre. Las hipótesis serían:

H_0: las muestras provienen de poblaciones con medianas iguales

H_1: las muestras provienen de poblaciones con medianas diferentes

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia, establecer la regla de decisión y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Si el *valorP* < ó bien $H_{calculada} > H_{crítica} = {}^2_{,g.l.=k-1}$, H_0 se rechaza. Es una prueba de cola derecha.

Estadístico de Prueba

El tamaño de cada muestra debe ser n 5 para usar la aproximación ${}^2_{,g.l.=k-1}$, donde k = total de muestras. Si las muestras tienen menos de 5 observaciones hay que utilizar tablas especiales. Las tres o más poblaciones pueden ser de libre distribución.

a. las muestras se mezclan formando una sola.
b. ordenar la muestra resultante de "a" de menor a mayor.
c. en la muestra resultante del inciso "b" cada valor en reemplazado por un rango.
d. el rango, en la muestra resultante de "b", es progresivo de menor a mayor. El rango menor es 1.
e. si hay datos con el mismo valor (empates) en la muestra «d» se les asigna el promedio de los rangos implicados.
f. se separan las muestras implicadas en "e" y se calcula la suma de los rangos de cada

una, por lo que se obtiene R_1 = suma de los rangos de la muestra 1, R_2 = suma de los rangos de la muestra 2 ... R_k = suma de los rangos de la k-ésima muestra.

g. se calcula el estadístico de prueba *H*.

$$H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ \ldots + \frac{R_k^2}{n_k}\right) - \ 3(N+1)$$

Donde:

H = valor de la prueba de Kruskal-Wallis

N = número total de observaciones en todas las muestras combinadas

R_1^2 = suma de los rangos de la muestra 1 al cuadrado

R_2^2 = suma de los rangos de la muestra 2 al cuadrado

R_k^2 = suma de los rangos de la k-ésima muestra

n_1 = número de observaciones de la muestra 1

n_2 = número de observaciones de la muestra 2

n_k = número de observaciones de la k-ésima muestra

Cálculo de la Prueba de Kruskal-Wallis

Ejemplo 113.- En los sismos ocurridos en 5 alcaldías de la CdMx en el 2016 de 1 grado Richter o mayor, se podría investigar si las poblaciones de los grupos de sismos tienen o no la misma mediana.

Magnitud de 25 sismos.

No.	Magnitud
Cuauhtémoc	
1	2
2	2.6
3	1.3
4	1.8
5	2.1
Iztapalapa	
6	2.2
7	2.2
8	2.2

No.	Magnitud
9	1
10	2.1
Gustavo A. Madero	
11	2.3
12	1.8
13	2.3
14	1
15	1.8
Tláhuac	
16	2.6

No.	Magnitud
17	3.3
18	2.9
19	1.4
20	1.8
Venustiano Carranza	
21	1.5
22	1.7
23	3.5
24	2.4
25	1.5

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

H_0: las 5 muestras provienen de poblaciones con medianas iguales

H_1: las 5 muestras provienen de poblaciones con medianas diferentes

Significancia: = 0.05

Regla de decisión: si el *valorP* < , o bien, $H_{calculada} > H_{crítica} = {}^{2}{}_{,g.l.=5-1}$, H_0 se rechaza.

Rango definitivo para cada sismo:

No.	Magnitud	Rango Inicial	Rango Definitivo
Cuauhtémoc			
1	2	12	12
2	2.6	21	21.5
3	1.3	3	3
4	1.8	8	9.5
5	2.1	13	13.5
Iztapalapa			
6	2.2	15	16
7	2.2	17	16
8	2.2	16	16
9	1	1	1.5
10	2.1	14	13.5
Gustavo A. Madero			
11	2.3	18	18.5
12	1.8	10	9.5
13	2.3	19	18.5
14	1	2	1.5
15	1.8	11	9.5

No.	Magnitud	Rango Inicial	Rango Definitivo
Tláhuac			
16	2.6	22	21.5
17	3.3	24	24
18	2.9	23	23
19	1.4	4	4
20	1.8	9	9.5
Venustiano Carranza			
21	1.5	6	5.5
22	1.7	7	7
23	3.5	25	25
24	2.4	20	20
25	1.5	5	5.5

Alcaldía de Cuauhtémoc: $n_1 = 5$, $R_1 = 59.5$

Alcaldía de Iztapalapa: $n_2 = 5$, $R_2 = 63$

Alcaldía de Gustavo A. Madero: $n_3 = 5$, $R_3 = 57.5$

Alcaldía de Tláhuac: $n_4 = 5$, $R_4 = 82$

Alcaldía de Venustiano Carranza: $n_5 = 5$, $R_5 = 63$

Cálculo del estadístico de prueba:

$$H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ldots + \frac{R_k^2}{n_k}\right) - 3(N+1) =$$

$$= \frac{12}{25(25+1)}\left(\frac{59.5^2}{5}+\frac{63^2}{5}+\frac{57.5^2}{5}+\frac{82^2}{5}+\frac{63^2}{5}\right)-3(25+1)=$$

$$= 0.01846154(708.05+793.80+661.25+1344.80+793.80)-78=$$

$$= 79.416-78= 1.416$$

Conclusión estadística: como $H_{calculada}$ =1.416 > $H_{crítica}$ = $^{2}_{0.05,g.l.5-1}$ = 9.488ó bien, *valor P = 0.8414>* = 0.05, H_0 no se rechaza.

Conclusión contextual: hay evidencia para asegurar que las 5 poblaciones de sismos tienen el mismo valor de la mediana.

Ejemplo 114.- Se sospecha que 4 poblaciones de mujeres de diferentes alcaldías tienen la misma mediana en cuanto al tamaño de la cintura en centímetros.

No.	Cintura
Tlalpan	
1	78.80
2	75.40
3	104.70
4	75.50
5	74.50
6	95.50
7	74.50
8	67.80
9	72.90
10	81.40
Milpa Alta	
11	75.90
12	126.50
13	74.50
14	93.00
15	126.00
16	67.20
17	98.00
18	105.50
19	91.10
20	68.60

No.	Cintura
Coyoacán	
21	82.50
22	79.50
23	82.60
24	85.70
25	92.80
26	72.70
27	94.00
28	73.60
29	69.10
30	85.70
Álvaro Obregón	
31	99.30
32	100.70
33	85.00
34	105.50
35	68.70
36	66.70
37	92.80
38	67.70
39	70.00
40	99.40

Fuente: Datos modificados del software estadístico statdisk 13.

H_0: las cuatro muestras provienen de poblaciones con medianas iguales

H_a: las cuatro muestras provienen de poblaciones con medianas diferentes

Significancia: = 0.05

Regla de decisión: si el *valor P* < , o bien, $H_{calculada} > H_{crítica} = {}_{2,g.l.=4-1}$, H_0 se rechaza.

Rangos definitivos para la cintura en centímetros para mujeres:

No.	Cintura	Rango Inicial	Rango Definitivo
Tlalpan			
1	78.80	18	18
2	75.40	15	15
3	104.70	36	36
4	75.50	16	16
5	74.50	13	13
6	95.50	31	31
7	74.50	14	13
8	67.80	4	4
9	72.90	10	10
10	81.40	20	20
Milpa Alta			
11	75.90	17	17
12	126.50	40	40
13	74.50	12	13
14	93.00	29	29
15	126.00	39	39
16	67.20	2	2
17	98.00	32	32
18	105.50	37	37.5
19	91.10	26	26
20	68.60	5	5

No.	Cintura	Rango Inicial	Rango Definitivo
Coyoacán			
21	82.50	21	21
22	79.50	19	19
23	82.60	22	22
24	85.70	24	24.5
25	92.80	28	27.5
26	72.70	9	9
27	94.00	30	30
28	73.60	11	11
29	69.10	7	7
30	85.70	25	24.5
Álvaro Obregón			
31	99.30	33	33
32	100.70	35	35
33	85.00	23	23
34	105.50	38	37.5
35	68.70	6	6
36	66.70	1	1
37	92.80	27	27.5
38	67.70	3	3
39	70.00	8	8
40	99.40	34	34

Alcaldía Tlalpan: $n_1 = 10$, $R_1 = 176$

Alcaldía Milpa Alta: $n_2 = 10$, $R_2 = 240.5$

Alcaldía Coyoacán: $n_3 = 10$, $R_3 = 195.5$

Alcaldía Álvaro **Obregón:** $n_4 = 10$, $R_4 = 208$

Cálculo del estadístico de prueba:

$$H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ \ldots + \frac{R_k^2}{n_k}\right) - \ 3(N+1) =$$

$$= \frac{12}{40(40+1)}\left(\frac{176^2}{10} + \frac{240.5^2}{10} + \frac{195.5^2}{10} + \frac{208^2}{10}\right) - \ 3(40+1) =$$

$$= \ 0.007317073(3097.60 + 5784.025 + 3822.025 + 4326.40) - 123 =$$

$$= \ 0.007317073(17030.05) - 123 = 1.61011904$$

Conclusión estadística: como $H_{calculada}$ = 1.610 < $H_{crítica}$ = $^{2}_{0.05,g.l.=4-1}$ = 7.815ó bien, *valor P = 0.6571>* = 0.05, H_0 no se rechaza.

Conclusión contextual: hay evidencia suficiente para asegurar que las medianas de las cuatro poblaciones son iguales.

22.- Prueba de correlación de rangos de Spearman

Es una prueba no paramétrica para datos apareados, para probar la existencia o carencia de asociación entre dos variables, utiliza rangos para calcular el coeficiente de correlación de rangos = r_s (el subíndice s indica que se trata de la letra inicial de Spearman). Las hipótesis se realizan en relación con el coeficiente de correlación de rangos a nivel poblacional (p_s) para dos muestras.

H_0: $p_s = 0$ no hay correlación entre las variables

H_1: p_s 0 sí hay correlación entre ambas variables

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia, establecer la regla de decisión y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Si el *valor P* < , ó bien, si $r_{scalculada} > r_{s,}$, H_0 se rechaza. Es una prueba de dos colas. La determinación del valor crítico ($r_{s,}$) depende de si n (número de pares de datos muestrales) 30 se consultan las tablas del coeficiente de correlación de rangos de Spearman, pero si n > 30 los valores críticos de r_s se calculan según:

$$r_s = \frac{\pm z}{\sqrt{n-1}}$$

Donde:

r_s = valores críticos

z = valores de la distribución normal estándar de acuerdo con

n = número de pares de datos muestrales

Estadístico de Prueba

Antes de calcular el estadístico de prueba, es necesario checar si las muestras se componen de rangos o no. Si los datos de las muestras no son rangos hay que convertirlos. Hay que acomodar los rangos de las dos muestras de la siguiente manera:

a. ordenar la muestra guía 1 del valor menor al mayor.

b. a cada uno de los valores de la muestra resultante de "a" se le asigna un rango progresivo, el menor es el número 1.

c. las categorías se ordenan según el rango que les corresponde en "a". Tabla 1.

d. se acomodan los rangos de la otra muestra de acuerdo con cada una de las categorías. Los rangos no necesariamente quedan en orden progresivo en esta otra muestra y se calculan las diferencias. Tabla 1.

Tabla 1.- *Arreglo correcto de las muestras*

Categorías	Muestra Guía (MG rangos)	La otra muestra (OM)	Diferencia (*d*)
A	1	rango de A en OM	MG 1 - rango de A en OM
B	2	rango de B en OM	MG 2 - rango de B en OM
C	3	rango de C en OM	MG 3 - rango de C en OM
D	4	rango de D en OM	MG 4 - rango de D en OM
E	5	rango de E en OM	MG 5 - rango de E en OM
...	**...**	**...**	**...**
k-ésima categoría	k-ésimo rango	rango de la k-ésima categoría en OM	MG k-ésimo rango - rango de la k-ésima categoría en OM

Ahora bien, del estadístico de prueba se tienen dos tipos: 1) sin empates del rango dentro de cada variable se utiliza:

$$r_s = 1 - \frac{6\sum_{i=1}^{rm} d^2}{n(n^2 - 1)}$$

Donde:
r_s = coeficiente de correlación de rangos de Spearman
n = número de pares de datos muestrales
rm = máximo rango
d^2 = diferencia del rango mayor al rango menor de los datos pareados

2) con empates de los rangos en cualquiera de las dos variables se utiliza:

$$r_s = \frac{n\sum_{i=1}^{rm} xy - (\sum_{i=1}^{rm} x)(\sum_{i=1}^{rm} y)}{\sqrt{n(\sum_{i=1}^{rm} x^2) - (\sum_{i=1}^{rm} x)^2}\sqrt{n(\sum_{i=1}^{rm} y^2) - (\sum_{i=1}^{rm} y)^2}}$$

Donde:

r_s = coeficiente de correlación de rangos de Spearman
n = número de pares de datos muestrales
rm = máximo rango
x = rangos de una muestra
y = rangos de la otra muestra
xy = producto de los rangos apareados

Cálculo de la Prueba de Correlación de Rangos de Spearman

Ejemplo 115.- Se lesolicitó al mismo número de mujeres y de hombres que habitan por lo menos desde hace 20 años en la CDMX que por consenso de cada grupo, según su sexo, acomoden en orden creciente de rango 7 alcaldías según su opinión de donde se sienten más los sismos.
Al seguir los pasos de la "a" a la "e", anotados más arriba, aplicados al consenso de la opinión de mujeres y hombres se tiene que:

Alcaldía	Rango Mujeres	Rango Hombres
Tlalpan	1	2
Xochimilco	2	1
Álvaro Obregón	3	3
Benito Juárez	4	4
Cuauhtémoc	5	7
Tláhuac	6	5
Iztapalapa	7	6

H_0: p_s = 0 no hay correlación entre el consenso de las mujeres y el de los hombres

H_1: p_s 0 sí hay correlación entre el consenso de las mujeres y el de los hombres

Significancia: = 0.05

Regla de decisión: si el *valor P* < , ó bien, $r_{s\ calculada} > r_s$, , H_0 se rechaza

Cálculo del estadístico de prueba:

No.	Alcaldía	Rango Mujeres	Rango Hombres	D	d^2
1	Tlalpan	1	2	1	1
2	Xochimilco	2	1	1	1
3	Álvaro Obregón	3	3	0	0
4	Benito Juárez	4	4	0	0
5	Cuauhtémoc	5	7	2	4
6	Tláhuac	6	5	1	1
7	Iztapalapa	7	6	1	1
Total					8

$$r_s = 1 - \frac{6\sum_{i=1}^{rm} d^2}{n(n^2 - 1)} = 1 - \frac{6 * 8}{7 * 48} = 1 - \frac{48}{336} = 1 - 0.143 = 0.857$$

Conclusión estadística: como $r_{s\ calculada}$ =0.857 > $r_{s,\ 0.05}$ = 0.786 , H_0 se rechaza.

Conclusión contextual: hay evidencia para asegurar que las opiniones consensuadas de mujeres y hombres están correlacionadas, ya que al aumentar la de las mujeres también aumenta la de los hombres.

Ejemplo 116.- Se le solicitó al mismo número de mujeres y de hombres que por consenso de cada grupo, según su sexo, acomoden en orden creciente de agrado a 7 pueblos originarios de la alcaldía de Tláhuac.
Al seguir los pasos de la "a" a la "e", anotados más arriba, aplicados al consenso de la opinión de mujeres y hombres se tiene que:

Alcaldía Tláhuac	Rango Mujeres	Rango Hombres
San Andrés Mixquic	1	3
San Francisco Tlaltenco	2	2
Santa Catarina Yecahuízotl	3	6
San Juan Ixtayopan	4	4
Santiago Zapotitlán	5	1
San Nicolás Tetelco	6	7
San Pedro Tláhuac	7	5

H_0: p_s = 0 no hay correlación entre el consenso de las mujeres y el de los hombres

H_1: p_s 0 sí hay correlación entre el consenso de las mujeres y el de los hombres

Significancia: = 0.05

Regla de decisión:si el *valorP* < , ó bien, $r_{s\ calculada} > r_{s,}$, H_0 se rechaza.

Cálculo del estadístico de prueba:

Diferencias entre los rangos (d) y d^2:

No.	Alcaldía Tláhuac	Rango Mujeres	Rango Hombres	d	d^2
1	San Andrés Mixquic	1	3	2	4
2	San Francisco Tlaltenco	2	2	0	0
3	Santa Catarina Yecahuízotl	3	6	3	9
4	San Juan Ixtayopan	4	4	0	0
5	Santiago Zapotitlán	5	1	4	16
6	San Nicolás Tetelco	6	7	1	1
7	San Pedro Tláhuac	7	5	2	4
Total					34

$$r_s = 1 - \frac{6\sum_{i=1}^{rm} d^2}{n(n^2 - 1)} = 1 - \frac{6 * 34}{7 * 48} = 1 - \frac{204}{336} = 1 - 0.607 = 0.393$$

Conclusión estadística: como $r_{s\ calculada}$ =0.393 < $r_{s,\ 0.05}$ = 0.786 H_0 no se rechaza.

Conclusión contextual: hay evidencia para asegurar que, las opiniones consensuadas de mujeres y hombres sobre su agrado de los pueblos estudiados de la alcaldía de Tláhuac no están correlacionados.

23.- Prueba de bondad de ajuste

Es una prueba no paramétrica para una muestra de la que se desea probar que una distribución de frecuencias observada se ajusta a una que se propone. Hay dos vertientes, 1.- con frecuencias desiguales y 2 .- frecuencias iguales, que es la que se detallará. Se utiliza en experimentos multinomiales, pues tiene más de dos categorías a diferencia del binomial con solamente dos. Sin embargo, para que sea considerado de ese tipo es necesario que el número de ensayos sea fijo e independientes; que sus resultados sean clasificados en una de las categorías y que la probabilidad de cada una sea la misma en cada uno de los ensayos. Las hipótesis a prueba son:

$$H_{0:}\, p_1 = p_2 = p_3 \ldots p_k$$

H_1: al menos una probabilidad es diferente a las otras

Donde:

k = número de categorías

p = probabilidad de la categoría en turno

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia, establecer la regla de decisión y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Si el *valor P* < , ó bien, si $^2_{calculada} > {}^2_{,g.l.=k-1}$, H_0 se rechaza. Es una prueba de cola derecha.

Estadístico de Prueba

La muestra debe cumplir con que los datos hayan sido seleccionados al azar; consisten de frecuencias para cada categoría; la frecuencia esperada (*E*) al menos es de 5 y no necesariamente son enteros, sin embargo, la observada si son enteros.

La forma de determinar las frecuencias esperadas, ya que en H_0 todas las frecuencias son iguales, es:

$$E = n/k$$

Donde:
n = suma de todas las frecuencias

Entonces el estadístico de prueba es:

$$\chi^2 = \sum_{k=1}^{k} \frac{(O-E)^2}{E}$$

Donde:
χ^2 = ji-cuadrada
O = frecuencia observada
E = frecuencia esperada
k = número total de categorías

Cálculo de la Prueba de Bondad de Ajuste

Ejemplo 117.- En el registro de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 o mayor en la escala de Richter se desea saber si las probabilidades de las magnitudes son iguales o no

Magnitud de 89 sismos registrados:

No.	Magnitud	No.	Magnitud	No.	Magnitud	No.	Magnitud
1	2.7	22	1.5	43	2.4	64	2.9
2	1.9	23	2.3	44	1.7	65	2.5
3	2.1	24	3.3	45	2.4	66	1.9
4	1.8	25	3.0	46	1.8	67	1.7
5	2.4	26	2.5	47	1.7	68	1.7
6	4.1	27	2.0	48	2.3	69	2.4
7	2.0	28	1.7	49	1.9	70	4.2
8	2.2	29	1.5	50	2.6	71	2.0
9	2.2	30	1.5	51	1.8	72	1.0
10	2.3	31	1.7	52	1.9	73	2.7
11	2.5	32	2.2	53	2.4	74	2.5
12	1.8	33	1.9	54	2.4	75	3.5
13	2.6	34	2.1	55	2.4	76	2.4
14	4.0	35	1.3	56	2.2	77	1.4
15	2.1	36	2.4	57	2.6	78	2.7
16	2.3	37	1.3	58	1.5	79	2.1
17	2.2	38	2.0	59	1.8	80	2.7
18	2.2	39	1.5	60	2.7	81	2.6
19	2.9	40	2.2	61	2.3	82	2.0
20	2.1	41	1.3	62	1.5	83	2.5
21	2.5	42	1.8	63	2.1	84	1.7

No.	Magnitud
85	3.6
86	3.3
87	2.0
88	2.1
89	1.0

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

H_0: $p_1 = p_2 = p_3 = p_4$

H_1: al menos una de las cuatro probabilidades es diferente a las otras

Significancia: = 0.05

Regla de decisión:si el *valor P* < , ó bien, $\chi^2_{calculada} > \chi^2_{\alpha,g.l.=k-1}$, H_0 se rechaza.

Frecuencias observadas

Magnitudes	Categoría	Frecuencia observada
1 a 1.9	1	30
2 a 2.9	2	51
3 a 3.9	3	5
4	4	3

Frecuencias esperadas:

E = n/k = 89/4 = 22.25

$$\chi^2 = \sum_{k=1}^{4} \frac{(O-E)^2}{E} = \frac{(30-22.25)^2}{22.25} + \cdots + \frac{(3-22.25)^2}{22.25} =$$

=2.699+ +16.654=69.8764

Conclusión estadística: como el *valor P= 0.0000* < = 0.05, ó bien, $\chi^2_{calculada}$= 69.8764 >$\chi^2_{0.05,g.l.=4-1}$= 7.815, H_0 se rechaza.

Conclusión contextual: hay evidencia para asegurar, que al menos una de las probabilidades de las categorías de la magnitud de los sismos es diferente a las demás.

Ejemplo 118.- De acuerdo con el Índice de Masa Corporal (IMC) de mujeres, se desea saber si las categorías tienen la misma probabilidad.

IMC de 40 mujeres:

No.	IMC
1	40.6
2	29.8
3	33.5
4	22.8
5	22.8
6	28.5
7	18.3
8	24.0
9	26.5
10	21.9
11	29.7
12	31.0
13	17.7
14	30.9
15	19.6
16	20.6
17	44.9
18	20.7
19	37.7
20	26.0

No.	IMC
21	19.2
22	27.5
23	23.5
24	29.1
25	23.8
26	31.7
27	29.9
28	28.7
29	25.1
30	21.4
31	21.2
32	20.5
33	19.8
34	19.6
35	28.9
36	21.9
37	22.0
38	25.2
39	23.8
40	19.3

Fuente: Datos modificados del software estadístico statdisk 13.

$$H_0: p_1 = p_2 = p_3 = p_4$$

H_1: al menos una de las cuatro probabilidades es diferente a las otras

Significancia: = 0.05

Regla de decisión: si el *valor P* < , ó bien, $^2_{calculada} > {}^2_{,g.l.=k-1}$, H_0 se rechaza.

Frecuencias observadas:

Clasificación del peso	Categoría	Frecuencia observada
Bajo Peso	1	2
Normal	2	19
Sobrepeso	3	12

Clasificación del peso	Categoría	Frecuencia observada
Obesidad	4	7

Frecuencias esperadas:

$$E = n/k = 40/4 = 10$$

$$\chi^2 = \sum_{k=1}^{4} \frac{(O-E)^2}{E} = \frac{(2-10)^2}{10} + \cdots + \frac{(7-10)^2}{10} =$$

$$= 6.4 + \cdots + 0.9 = 15.8$$

Conclusión estadística: como el *valorP= 0.0012* < = 0.05, ó bien, $^{2}_{alculada} = 15.8 > {}^{2}_{0.05,g.l.=4-1} = 7.815$, H_0 se rechaza.

Conclusión contextual: hay evidencia para asegurar que al menos una de las probabilidades de las categorías del IMC es diferente de las demás.

24.- PRUEBA DE RACHAS

Es una prueba no paramétrica para detectar aleatoriedad en una muestra y hay dos opciones, 1.- muestra chica o 2.- muestra grande. Somete a corroboración si los datos de la muestra que tienen cierta secuencia, es aleatoria o no. Se basa en datos muestrales que tienen dos características y analiza las rachas de cada una, esto es, la secuencia de cada una. Por definición, una racha es la secuencia continua de datos con una sola característica, y precedida o seguida por una de la otra característica o por ningún dato. La aleatoriedad se rechaza siempre que el número de rachas sea muy bajo o alto. La prueba se basa en el orden, no en la frecuencia. Las hipótesis son:

H_0: la secuencia de las rachas es aleatoria

H_1: la secuencia de las rachas no es aleatoria

Nivel de Significancia

Inmediatamente después de establecer las Hipótesis a probar **SIEMPRE** hay que determinar el nivel de significancia, establecer la regla de decisión y seguidamente escoger el estadístico de prueba. En cualquier prueba de Hipótesis el nivel de significancia ( ) está entre 0 y 1, es una variable cuantitativa continua. Los valores más empleados según la gravedad de cometer el Error Tipo I son = 0.10, = 0.05, = 0.01, y es muy común usar el valor = 0.05. El investigador escoge el valor correspondiente. Hay que tomar en cuenta que = Error Tipo I, significa no rechazar H_0 cuando es falsa.

Regla de Decisión

SIEMPRE se establece una vez determinado el valor de significancia y se aplica a los resultados del estadístico de prueba. Se tienen dos posibilidades, ya sea para muestras chicas o grandes. Si n_1 20 y n_2 20 y = 0.05, y *G* al valor más chico de la tabla para la prueba de rachas ó*G* al número más grande de la tabla para la prueba de rachas, H_0 se rechaza. Pero si $n_1 > 20$ ó $n_2 > 20$ ó 0.05, y $z_{calculada} > z_{/2}$, ó bien, *valor P*< , H_0 se rechaza. Es una prueba de dos colas.

Estadístico de Prueba

1.- Si n_1 y n_2 20 y = 0.05 el estadístico de prueba sería G = número de rachas:

$$G = RC1 + RC2$$

Donde:

G = número total de rachas
RC1 = número total de rachas de la característica 1
RC2 = número total de rachas de la característica 2

2.- Si n_1 ó n_2 > 20 ó 0.05, el estadístico de prueba sería:

$$z = \frac{G - \mu_G}{\sigma_G}$$

$$\mu_G = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$\sigma_G = \sqrt{\frac{(2n_1n_2)(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

Donde:

z = valor observado
G = número de rachas
μ_G = media del número de rachas (G)
$_G$ = desviación estándar del número de rachas(G)
n_1 = número de datos iguales en la muestra
n_2 = número de datos iguales en la muestra pero diferentes a n_1

Cálculo de la Prueba de Rachas

Ejemplo 119.- En el registro de algunos de los sismos ocurridos en la Ciudad de México en los años del 2016 al 2019, de magnitud 1 y < 2 , así como 2 y <3 en la escala de Richter, se desea saber si las magnitudes son o no aleatorias:

Magnitud de 69 sismos registrados:

No.	Magnitud
1	2.2
2	1.7
3	2.2
4	2.1
5	2.1
6	2.5
7	2.4
8	2.7
9	1.0

No.	Magnitud
10	1.9
11	1.9
12	2.2
13	2.5
14	1.7
15	2.4
16	1.4
17	1.5
18	2.4

No.	Magnitud
19	2.5
20	2.2
21	1.3
22	2.2
23	2.2
24	2.1
25	1.8
26	1.8
27	2.2

No.	Magnitud	No.	Magnitud	No.	Magnitud
28	1.3	42	1.8	56	2.6
29	1.8	43	2.4	57	1.8
30	2.7	44	2.1	58	2.3
31	2.5	45	2.6	59	2.5
32	1.5	46	2.6	60	2.6
33	2.3	47	2.3	61	2.7
34	1.7	48	2.4	62	1.9
35	2.3	49	1.8	63	2.1
36	1.7	50	2.7	64	1.5
37	2.4	51	2.4	65	2.9
38	1.7	52	2.5	66	1.7
39	2.4	53	1.0	67	1.5
40	2.7	54	1.7	68	1.9
41	2.0	55	1.9	69	2.0

Fuente: Datos modificados de http://www2.ssn.unam.mx:8080/catalogo/

H_0: la secuencia de las rachas de magnitud 1 y 2 es aleatoria

H_1: la secuencia de las rachas de magnitud 1 y 2 no es aleatoria

Significancia: = 0.05

Regla de decisión: si $z_{calculada} > z_{/2}$, ó bien, *valor P*< , H_0 se rechaza.

Determinación de rachas (G), n_1 y n_2:

RC1 = magnitud 1 y < 2 escala de Richter

RC2 = magnitud 2 y <3 escala de Richter

G = RC1+ RC2= 18 + 19 = 35

n_1 = 27 y n_2= 42. Como n_1> 20 y n_2> 20 el estadístico de prueba es z:

$$\mu_G = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2*27*42}{27+42} + 1 = \frac{2268}{69} + 1 = 33.87$$

$$\sigma_G = \sqrt{\frac{(2n_1n_2)(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}} = \sqrt{\frac{2268(2268-27-42)}{4761(68)}} = \sqrt{\frac{4987332}{323748}} =$$

$$=3.92491804$$

$$z = \frac{G-\mu_G}{\sigma_G} = \frac{35-33.87}{3.92} = \frac{1.13}{3.92} = 0.288$$

3.92491804

Conclusión estadística: como $z_{calculada} = 0.288$ $z_{0.05/2} < \pm 1.96$, H_0 no se rechaza.

Conclusión contextual: hay evidencia para asegurar que las rachas de los sismos son aleatorias.

Ejemplo 120.- De acuerdo con el sexo biológico (M = mujer y H= hombre) se desea saber si las rachas son o no aleatorias.

Muestra de 34 personas:

No.	SEXO
1	H
2	M
3	H
4	H
5	M
6	H
7	H
8	H
9	M
10	M
11	M
12	M
13	M
14	M
15	H
16	H
17	M

No.	SEXO
18	H
19	M
20	H
21	H
22	H
23	M
24	M
25	H
26	M
27	M
28	M
29	M
30	H
31	M
32	M
33	H
34	M

Fuente: Datos modificados del software estadístico statdisk 13.

H_0: la secuencia de las rachas de Mujer y Hombre es aleatoria

H_1: la secuencia de las rachas de Mujer y Hombre no es aleatoria

Significancia: = 0.05

Regla de decisión: si *G*< que el número menor de $G_{0.05,19,15}$ y G > que el número mayor de G0.05,19,15, H_0 se rechaza.

Determinación de rachas (G), n_1 y n_2:

RM = Rachas mujer

RH = Rachas hombre

$G = RM + RH = 9 + 9 = 18$

$n_1 = 19$ y $n_2 = 15$. Como n_1 20 y n_2 20 el estadístico de prueba es *G*:

Conclusión estadística: como $G = 18$ es mayor que el número menor de $G_{0.05,19,15} = 11$, pero menor que el mayor número de $G_{0.05,19,15} = 24$, H_0 no se rechaza.

Conclusión contextual: hay evidencia para asegurar que las rachas de la secuencia de Mujeres y Hombres son aleatorias.

25.- Bibliografía y programa selecto

Álvarez, R. C., 2007. *Estadística Aplicada a las Ciencias de la Salud*. Díaz de Santos, Madrid.

Anderson, D. R., D. J. Sweeney, Th. A. Williams, J. D. Camm y J. J. Cochran, 2016. *Estadística para Negocios y Economía*, 12a ed. CENGAGE Learning, México.

Best, J., 2001. *Dammed Lies and Statistics*. University of California Press, Berkeley.

Bhattacharyya, G. K. y R. A. Johnson, 1977. *Statistical Concepts and Methods*. John Wiley and Sons, Nueva York.

Campbell, S. 1974. *Flaws and Fallacies in Statistical Thinking*. Prentice Hall, Englenwood Cliffs.

Daniel, W. W., 1996. *Bioestadística, Base para el Análisis de las Ciencias de la Salud*. Editorial Limusa, México.

Hollander, M. y F. Proschan, 1984. The *Statistical Exorcist: Dispelling Statistics Anxiety*. Marcel Dekker, Nueva York.

Huff, D., 2011. *Cómo mentir con estadística*. Crítica, México.

Kotz, S. y D. Stroup, 1983. *Educated Guessing-How to Cope in an Uncertain World*. Marcel Dekker, Nueva York.

Marques de Cantú, M. J., 1991. *Probabilidad y Estadística para Ciencias Químico-Biológicas*. Preedición. McGraw-Hill, México.

Mendenhall, III, W. R. J. Beaver y B. M. Beaver, 2015. *Introducción a la Probabilidad y Estadística*. Cengage Learning, México.

Mendenhall, III, W. R. J. Beaver y B. M. Beaver, 2015. *Probabilidad y Estadística para las Ciencias Sociales del Comportamiento y la Salud*. Cengage Learning, México.

Richtey, F. J., 2008, *Estadística para las Ciencias Sociales*. McGraw-Hill, México.

Snedecor, G. W. y W. Cochran, 1981. *Métodos Estadísticos*. C.E.C.S.A., México.

Statdisk 13, 2020, plataforma https://www.statdisk.com/accounts/login/?next=/

Statdisk 13, recuperado de https://www.statdisk.com/accounts/login/?next=/

Triola, M. F., 2006. *Elementary Statistics*. Pearson Education/Adison Wesley, Boston.

Prontuario para la materia de bioestadística
se terminó de imprimir endiciembre de 2025,
en el taller de impresión de la
Universidad Autónoma de la Ciudad de México,
San Lorenzo, 290, Col. Del Valle,
Alcaldía Benito Juárez, C. P. 03100,
Ciudad de México con un tiraje de 500 ejemplares.
Cuidado de la edición: Ángeles Godínez Guevara
Diseño editorial: Sergio Cortés Becerril

El presente material es de utilidad para las licenciaturas de Promoción de la Salud, Ciencias Ambientales y Protección Civil y Gestión de Riesgos de la UACM, así como para cualquier profesionista que tenga conocimientos básicos de Estadística y Bioestadística, y se elaboró con el objetivo de facilitar a las y los estudiantes el acceso a un manual complementario al curso obligatorio de Bioestadística.

En su extensión y profundidad académica, este libro se localiza entre el libro de texto y el formulario: un prontuario de consulta rápida, en el que se puede revisar la interpretación y aplicación de las herramientas más utilizadas en Bioestadística. Las tablas, medidas y pruebas son expuestas de manera detallada para su cálculo y aplicación, cada una con ejemplos del contexto nacional, y ejercicios de refuerzo.

El contenido aborda pruebas paramétricas y no paramétricas, así como léxico estadístico general y otros contenidos de utilidad básica para la labor estadística. En la bibliografía se han incluido textos universitarios de consulta, que abarcan a profundidad la teoría Bioestadística, y cuyo dominio depende del estudio que realicen las y los estudiantes.

Raúl Ernesto de Guadalupe Bravo Núñez, es biólogo por Facultad de Ciencias de la UNAM. Estudió los posgrados de Estadística Aplicada en el IIMAS y de Manejo de Recursos Naturales, Ecología y Aprovechamiento de los Recursos Naturales en el ITESM campus Guaymas. Tiene 40 años como docente e investigador en diferentes universidades públicas y privadas del país. Desde su ingreso a la UACM en 2006 hasta la fecha, es profesor de tiempo completo, impartiendo el curso de Bioestadística en las licenciaturas de Promoción de la Salud y de Protección Civil y Gestión de Riesgos del Colegio de Ciencias y Humanidades, del cual fue Coordinador de 2017 a 2019.

UACM
Universidad Autónoma de la Ciudad de México
NADA HUMANO ME ES AJENO

Biblioteca BE del Estudiante

9 786072 615663