ANTOLOGIA

BIOESTADISTICA

LICENCIATURA EN ENFERMERIA

CUARTO CUATRIMESTRE

Marco Estratégico de Referencia

ANTECEDENTES HISTORICOS

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

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

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

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

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

Nuestra Universidad inició sus actividades el 18 de agosto del 2004 en las instalaciones de la 4ª avenida oriente sur no. 24, con la licenciatura en Puericultura, contando con dos grupos de

cuarenta alumnos cada uno. En el año 2005 nos trasladamos a nuestras propias instalaciones en la carretera Comitán – Tzimol km. 57 donde actualmente se encuentra el campus Comitán y el Corporativo UDS, este último, es el encargado de estandarizar y controlar todos los procesos operativos y Educativos de los diferentes Campus, Sedes y Centros de Enlace Educativo, así como de crear los diferentes planes estratégicos de expansión de la marca a nivel nacional e internacional.

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

MISIÓN

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

VISIÓN

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

VALORES

- Disciplina
- Honestidad
- Equidad
- Libertad

ESCUDO

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

ESLOGAN

"Mi Universidad"

ALBORES

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

BIOESTADISTICA

Objetivo de la materia:

Al finalizar la asignatura el alumno conocerá las definiciones básicas de la estadística descriptiva y su aplicación en el campo de la enfermería, manejará adecuadamente las técnicas de estimación y decisión en la realización de estudios de corte epidemiológico y en las investigaciones propias de la disciplina.

INDICE

UNIDAD 3. DISTRIBUCIONES DE PROBABILIDAD

UNIDAD 4. DEMOGRAFIA

UNIDAD I: ESTADÍSTICA DESCRIPTIVA

1.1 La estadística en enfermería.

En las ciencias de la salud, la estadística tiene una gran importancia ya que posee numerosas ventajas, por ejemplo, nos puede ayudar a conocer las problemáticas presentes en una comunidad, los factores de riesgo o predisposición a ciertas patologías y puede ser muy útil a la hora de buscar una respuesta a esta o al tratar de educar para evitarlas en futuras ocasiones.

Como los objetos de estudio de las ciencias de la vida son muy variados, la Bioestadística ha debido ampliar su campo para, de esta manera, incluir cualquier modelo cuantitativo, no solamente estadístico y que entonces pueda ser empleado para responder a las necesidades oportunas.

La principal ventaja del pensamiento estadístico interviniendo en las ciencias de la vida es que no solo resuelve, sino que también comprende una compleja metodología para dar respuesta a las hipótesis, además de agilizar la cuestión de organización del sistema de investigación, desde el diseño general, el de muestreo, el control de la calidad de información y la presentación de los resultados.

En Salud Pública la estadística permite analizar situaciones en las que los *componentes aleatorios* contribuyen de forma importante en la variabilidad de los datos obtenidos. En salud pública los componentes aleatorios se deben, entre otros aspectos, al conocimiento o a la imposibilidad de medir algunos determinantes de los estados de salud y enfermedad, así como a la variabilidad en las respuestas por los pacientes, similares entre sí, que son sometidos al mismo tratamiento.

La extensión de los conocimientos y aptitudes de carácter estadístico que necesitan adquirir los profesionales de la salud pública son importantes, porque el conocimiento de los principios y métodos estadísticos y la competencia en su aplicación se necesitan para el ejercicio eficaz de la salud pública y, adicionalmente, para la comprensión e interpretación de los datos sanitarios.

1.1.1 Introducción histórica.

El primer médico que utilizó métodos matemáticos para cuantificar variables de pacientes y sus enfermedades fue el francés Pierre Charles-Alexandre Louis (1787-1872). La primera aplicación del Método numérico (que es como tituló a su obra y llamó a su método) en su clásico estudio de la tuberculosis, que influyó en toda una generación de estudiantes. Sus discípulos, a su vez, reforzaron la nueva ciencia de la epidemiología con base en el método estadístico. En las recomendaciones de Louis para evaluar diferentes métodos de tratamiento están las bases de los ensayos clínicos que se hicieron un siglo después. En Francia Louis René Villermé (1782-1863) y en Inglaterra William Farr (1807-1883) que había estudiado estadística médica con Louis hicieron los primeros mapas epidemiológicos usando métodos cuantitativos y análisis epidemiológicos. Francis Galton (1822-1911), basado en el darwinismo social, fundó la biometría estadística.

Pierre Simón Laplace (1749-1827), astrónomo y matemático francés, publicó en 1812 un tratado sobre la teoría analítica de las probabilidades, Théorie analytique des probabilités, sugiriendo que tal análisis podría ser una herramienta valiosa para resolver problemas médicos.

Los primeros intentos de hacer coincidir las matemáticas de la teoría estadística con los conceptos emergentes de la infección bacteriana tuvieron lugar a comienzos del siglo XX. Tres diferentes problemas cuantitativos fueron estudiados por otros tantos autores. William Heaton Hamer (1862-1936) propuso un modelo temporal discreto en un intento de explicar la ocurrencia regular de las epidemias de sarampión; John Brownlee (1868-1927), primer director del British Research Council, luchó durante veinte años con problemas de cuantificación de la infectividad epidemiológica, y Ronald Ross (1857-1932) exploró la aplicación matemática de la teoría de las probabilidades con la finalidad de determinar la relación entre el número de mosquitos y la incidencia de malaria en situaciones endémicas y epidémicas. Pero el cambio más radical en la dirección de la epidemiología se debe a Austin Bradford Hill (1897-1991) con el ensayo clínico aleatorizado y, en colaboración con Richard Doll (n. 1912), el épico trabajo que correlacionó el tabaco y el cáncer de pulmón.

Los primeros trabajos bioestadísticos en enfermería los realizó, a mediados del siglo XIX la enfermera inglesa Florence Nightingale. Durante la guerra de Crimea, Florence Nightingale observó que eran mucho más numerosas las bajas producidas en el hospital que en el frente. Por lo tanto, recopiló información y dedujo que la causa de la elevada tasa de mortalidad se debía a la precariedad higiénica existente. Así, gracias a sus análisis estadísticos, se comenzó a tomar conciencia de la importancia y la necesidad de unas buenas condiciones higiénicas en los hospitales.

1.2 La estadística como herramienta de trabajo en enfermería.

El análisis y las técnicas estadísticas son un componente esencial en toda investigación biomédica, y la utilización de las técnicas estadísticas ha evolucionado considerablemente en los últimos años en las áreas de la investigación de ciencias de la salud. No hay duda de que tanto la actividad investigadora como los profesionales de la salud necesitan métodos estadísticos para el análisis de sus observaciones debido al crecimiento incesantemente de los mismos.

El empleo de técnicas estadísticas más específicas en investigación ha ido en aumento en las últimas décadas, motivado por la inclusión de la bioestadística en el currículo de los profesionales de la salud y por la inclusión de perfiles expertos en metodología en los equipos de investigación. Los análisis estadísticos empleados en un estudio dependen en gran medida del tipo de estudio, del objetivo que se pretende abordar y del tamaño de la muestra, así como del grado de conocimiento por parte de los investigadores de las técnicas estadísticas y del software para su implementación.

Es por ello que la estadística juega un papel fundamental en la investigación en ciencias de la salud, y a través de un equipo multidisciplinar que engloba a profesionales del ámbito sanitario, académico y perfiles expertos en metodología estadística se obtienen investigaciones de mayor calidad.

Esta disciplina es usada en diversos campos de la medicina y la salud pública, como la epidemiología, nutrición y salud ambiental. Asimismo, sus métodos son aplicados en estudios relacionados con la ecología y la genómica.

Algunas de las aportaciones más importantes de la bioestadística se han dado en el estudio de las enfermedades. A raíz de los datos arrojados por esta disciplina se ha logrado un mejor entendimiento de la propagación de ciertas enfermedades y las características de males crónicos como el cáncer y el sida. Además, ha contribuido enormemente al desarrollo de nuevos fármacos.

Sin lugar a dudas, el pensamiento estadístico ha permitido establecer un sistema organizado de investigación, desde el diseño de la misma, el muestreo, el control de calidad, el análisis y la presentación de la información. De ese modo, ha permitido resolver y optimizar la metodología para dar respuesta a las diversas hipótesis que se manejan en el mundo de las ciencias de la vida.

1.3 Descripción de una variable estadística.

Cuando hablamos de variable estadística estamos hablando de una cualidad que, generalmente adopta forma numérica. Por ejemplo, la altura de Juan es de 180 centímetros. La variable estadística es la altura y está medida en centímetros. También podríamos, por ejemplo, decir que el beneficio de una empresa ha sido de 22.300 dólares el último año. En este caso, la variable sería el beneficio y estaría medido en dólares. Ambas variables son del tipo cuantitativo (se expresan con un número).

Claro que no todas las variables estadísticas son iguales y, por supuesto, no todas se pueden (en principio) expresar en forma de número. Así, otra variable que podríamos encontrarnos es el color de ojos de una persona. Por ejemplo, Juan tiene los ojos verdes y Andrés los tiene azules. La variable sería el color de ojos y sería una variable cualitativa. Es decir, no se expresa con número.

1.3.1. Definiciones básicas.

Variable estadística: Una variable estadística es una característica de una muestra o población de datos que puede adoptar diferentes valores.

Aunque hay decenas de tipos de variables estadísticas, por norma general podemos encontrarnos dos tipos de variables:

Variable cuantitativa: Son variables que se expresan numéricamente.

- Variable continua: Toman un valor infinito de valores entre un intervalo de datos. Por ejemplo, el tiempo que tarda un corredor en completar los 100 metros lisos.
- Variable discreta: Toman un valor finito de valores entre un intervalo de datos. Ejemplo: Número de helados vendidos.

Variable cualitativa: Son variables que se expresan, por norma general, en palabras.

- Variable ordinal: Expresa diferentes niveles y orden. Por ejemplo, primero, segundo, tercero, etc.
- Variable nominal: Expresa un nombre claramente diferenciado. Por ejemplo, el color de ojos puede ser azul, negro, castaño, verde, etc.

Además, cada una de estas variables podría tener más subtipos, ya que tenemos variables de tipo económico, categóricas, dicotómicas, dependientes, independientes. Es decir, como ya hemos dicho, muchos tipos de variables estadísticas. Por ejemplo, podríamos tener una variable estadística de tipo cuantitativo, discreta y dependiente.

Adicionalmente, también debemos aclarar que el hecho que las variables cualitativas se expresen con nombre no quiere decir que no puedan ser parte de un modelo matemático. Así pues, podríamos crear una variable cuantitativa a partir de una variable cualitativa. Por ejemplo, para el color de ojos podríamos asignar un 1 si tiene los ojos azules, un 2 si tiene los ojos verdes y un 3 si tiene los ojos marrones. O, en otros casos, podríamos también convertir variables dicotómicas que indica SI o NO, en 1 o 0.

1.4 Representaciones gráficas.

Una gráfica o una representación gráfica o un gráfico, es un tipo de representación de datos, generalmente cuantitativos, mediante recursos visuales (líneas, vectores, superficies o símbolos), para que se manifieste visualmente la relación matemática o correlación estadística que guardan entre sí. También es el nombre de un conjunto de puntos que se plasman en coordenadas cartesianas y sirven para analizar el comportamiento de un proceso o un conjunto de elementos o signos que permiten la interpretación de un fenómeno. La representación gráfica permite establecer valores que

no se han obtenido experimentalmente sino mediante la interpolación (lectura entre puntos) y la extrapolación (valores fuera del intervalo experimental).

Tipos de representaciones gráficas

Cuando se muestran los datos estadísticos a través de representaciones gráficas, se ha de adaptar el contenido a la información visual que se pretende transmitir. Para ello, se barajan múltiples formas de representación:

- **Diagramas de barras**: muestran los valores de las frecuencias absolutas sobre un sistema de ejes cartesianos, cuando la variable es discreta o cualitativa.
- **Histogramas**: formas especiales de diagramas de barras para distribuciones cuantitativas continuas.
- **Polígonos de frecuencias**: formados por líneas poligonales abiertas sobre un sistema de ejes cartesianos.
- **Gráficos de sectores**: circulares o de tarta, dividen un círculo en porciones proporcionales según el valor de las frecuencias relativas.
- **Pictogramas**: o representaciones visuales figurativas. En realidad, son diagramas de barras en los que las barras se sustituyen con dibujos alusivos a la variable.
- **Cartogramas**: expresiones gráficas a modo de mapa.
- **Pirámides de población**: para clasificaciones de grupos de población por sexo y edad.

Diagramas de barras e histogramas

Los diagramas de barras se usan para representar gráficamente series estadísticas de valores en un sistema de ejes cartesianos, de manera que en las abscisas se indica el valor de la variable estadística y en las ordenadas se señala su frecuencia absoluta.

Estos gráficos se usan en representación de caracteres cualitativos y cuantitativos discretos. En variables cuantitativas continuas, se emplea una variante de los mismos llamada histograma

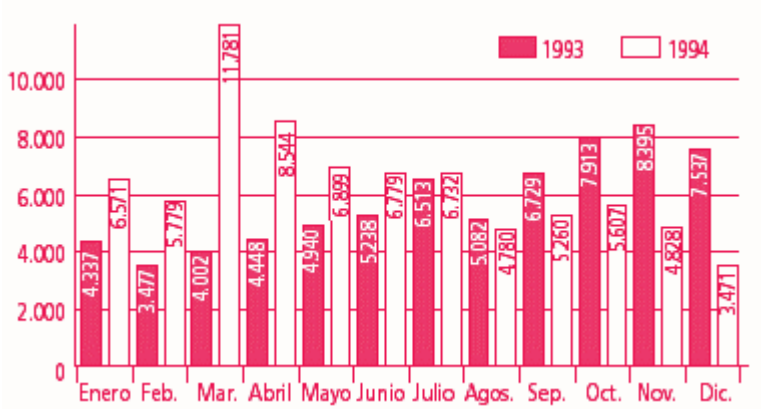


Figura 1. Diagrama de barras.

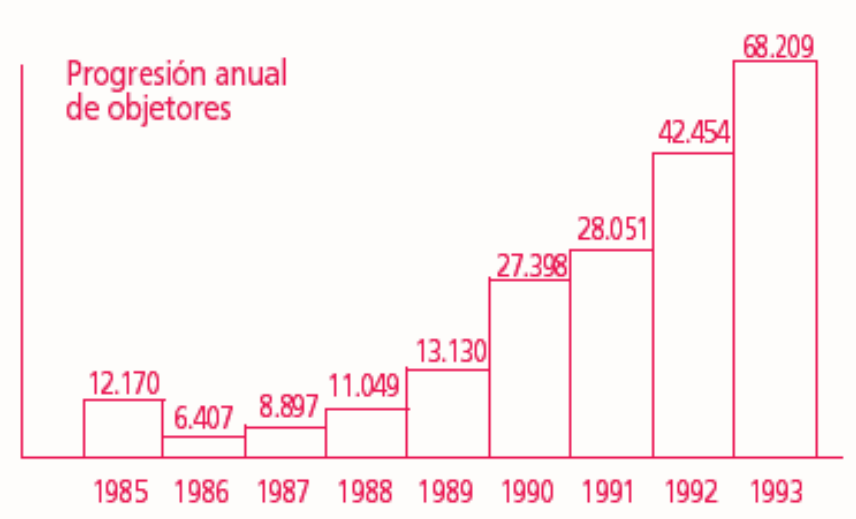
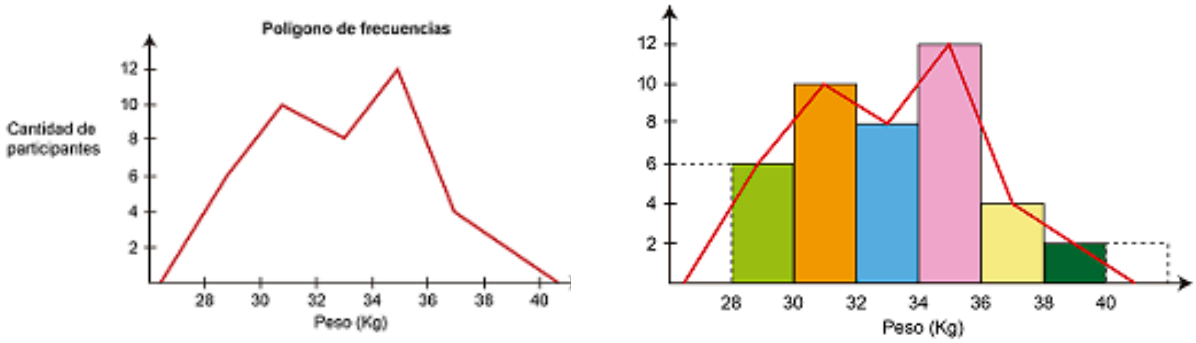


Figura 2. Histograma.

Polígonos de frecuencias

Esta gráfica se usa para representar los puntos medios de clase en una distribución de frecuencias. Para construir polígonos de frecuencias, se trazan las frecuencias absolutas o relativas de los valores de la variable en un sistema de ejes cartesianos y se unen los puntos resultantes mediante trazos rectos. Con ello se obtiene una forma de línea poligonal abierta.

Los polígonos de frecuencias se utilizan preferentemente en la presentación de caracteres cuantitativos, y tienen especial interés cuando se indican frecuencias acumulativas. Se usan en la expresión de fenómenos que varían con el tiempo, como la densidad de población, el precio o la temperatura.



Gráficos de sectores

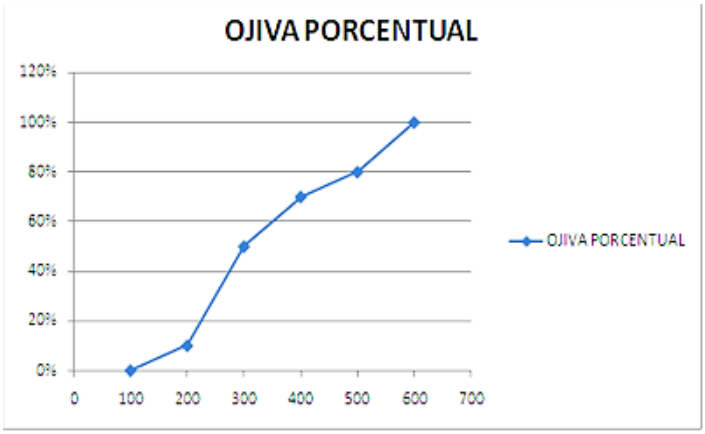
En los diagramas de sectores, también llamados circulares o de tarta, se muestra el valor de la frecuencia de la variable señalada como un sector circular dentro de un círculo completo. Por ello, resultan útiles particularmente para mostrar comparaciones entre datos, sobre todo en forma de frecuencias relativas de las variables expresadas en forma de porcentaje.

Pictogramas y cartogramas

Para aligerar la presentación de datos estadísticos, con frecuencia se recurre a imágenes pictóricas representativas del valor de las variables. Dos formas comunes de expresión gráfica de los datos son:

- **Los pictogramas**, que muestran diagramas figurativos con figuras o motivos que aluden a la distribución estadística analizada (por ejemplo, una imagen antropomórfica para indicar tamaños, alturas u otros).
- **Los cartogramas**, basados en mapas geográficos que utilizan distintas tramas, colores o intensidades para remarcar las diferencias entre los datos.

La ojiva: Esta gráfica consiste en la representación de las frecuencias acumuladas de una distribución de frecuencias. Puede construirse de dos maneras diferentes; sobre la base "menor que" o sobre la base "o más". Puede determinar el valor de la mediana de la distribución.



1.5 Representación numérica.

La tabla de frecuencias (o distribución de frecuencias) es una tabla que muestra la distribución de los datos mediante sus frecuencias. Se utiliza para variables cuantitativas o cualitativas ordinales.

La tabla de frecuencias es una herramienta que permite ordenar los datos de manera que se presentan numéricamente las características de la distribución de un conjunto de datos o muestra.

X_i	Frecuencia absoluta (n_i)	Frecuencia absoluta acumulada (N_i)	Frecuencia relativa (fi = ni/N)	Frecuencia relativa acumulada (Fi=Ni/N)
1	7	7	0,06	0,06
2	19	26	0,15	0,21
3	25	51	0,20	0,41
4	12	63	0,10	0,50
5	23	86	0,18	0,69
6	15	101	0,12	0,81
7	8	109	0,06	0,87
8	16	125	0,13	1,00
Total	**125**	**125**	**1**	**1**

Construcción de la tabla de frecuencias

Cabe distinguir entre:

- Tabla de frecuencias con datos no agrupados.
- Tabla de frecuencias con datos agrupados.

Construcción de una tabla de frecuencias con datos no agrupados

1. En la primera columna se ordenan de menor a mayor los diferentes valores que tiene la variable en el conjunto de datos.
2. En las siguientes columnas (segunda y tercera) se ponen las frecuencias absolutas y las frecuencias absolutas acumuladas.
3. Las columnas cuarta y quinta contienen las frecuencias relativas y las frecuencias relativas acumuladas.
4. Adicionalmente (opcional) se pueden incluir dos columnas (sexta y séptima), representando la frecuencia relativa y la frecuencia relativa acumulada como tanto por cien. Estos porcentajes se obtienen multiplicando las dos frecuencias por cien.

Construcción de una tabla de frecuencias con datos agrupados

Se emplea cuando hay un número alto de datos. Estos se agrupan en intervalos o clases para facilitar su tabulación y análisis. Está indicado para representarlos en un histograma.

Como en el caso anterior, se utiliza tanto para variables cuantitativas como en variables cualitativas ordinales.

Los pasos iniciales para formar una tabla de frecuencias con datos agrupados están encaminados a determinar el número de intervalos y definirlos (siempre que no se conozcan de antemano). Los pasos son:

1. Obtener el rango *R* de los datos. Es la diferencia entre el dato mayor y el menor del conjunto de valores que toma la variable a tabular. Se llama también amplitud total.

$$R = X_{máx} - X_{mín}$$

2. Fijar cuántos intervalos o clases se desea. Se tiende a que el número de clases sea impar y que esté entre 5 y 15. Hay dos maneras de hacerlo:
   - A criterio del investigador.
   - Mediante el método de Sturges, que emplea la siguiente fórmula:

$$n_{int} = 1 + 3{,}322 \cdot \log N$$

Donde n_{int} es el número de intervalos, el logaritmo es natural o base 10 y N es el número total de datos. El resultado se redondea al número entero más próximo.

3. Determinar la amplitud del intervalo o clase *I*:
   Es el resultado de dividir el rango *R* o amplitud total por el número de clases o intervalos n_{int} que se han fijado:

$$I = \frac{R}{n_{int}}$$

El valor obtenido en esta división no tiene porqué ser un número entero. En ese caso, se redondearía al valor entero más próximo. Los dos redondeos, el que se haya podido hacer en el número de intervalos n_{int} y el de la amplitud del intervalo *I* modificarán el valor de la amplitud total o rango, apareciendo un nuevo valor ajustado, con los valores definitivos, repartiendo la diferencia entre *R'* y *R* entre los dos extremos del rango:

$$R' = I \cdot n_{int}$$

4. Formar los diferentes intervalos o clases, partiendo del valor mínimo del nuevo rango *R'*. Cada intervalo tendrá unos extremos *a* y *b* separados por la amplitud de clase o intervalo *I*. En variables continuas, normalmente los intervalos son cerrados por la izquierda y abiertos por la derecha, [*a, b*) en el que *b* no pertenece a este intervalo, sino que es el valor mínimo del intervalo siguiente. En variables discretas ordinales o en variables continuas en los que el procedimiento de medición no pueda apreciar más allá de un valor entero, los intervalos o clases serán cerrados por los extremos [*a, b*].

5. Cada intervalo está representado por la llamada marca de clase. Es la media entre sus extremos.

$$c_i = \frac{a_i + b_i}{2}$$

Representará a los valores del intervalo o clase en los cálculos a partir de la tabla.

6. A partir de la columna de las clases, se formarán las columnas de las frecuencias, que son las que se describen a continuación y que son comunes para las tablas de datos no agrupados como en las de datos agrupados.

Tipos de frecuencias

Existen cuatro tipos de frecuencias:

Frecuencia absoluta

La frecuencia absoluta (n_i) de un valor X_i es el número de veces que el valor está en el conjunto (X_1, X_2,…, X_N). La suma de las frecuencias absolutas de todos los elementos diferentes del conjunto debe ser el número total de sujetos *N*. Si el conjunto tiene *k* números (o categorías) diferentes, entonces:

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_k = N$$

Frecuencia absoluta acumulada

La frecuencia absoluta acumulada(N_i) de un valor X_i del conjunto (X_1, X_2,…, X_N) es la suma de las frecuencias absolutas de los valores menores o iguales a X_i, es decir:

$$N_i = n_1 + n_2 + \ldots + n_i$$

Frecuencia relativa

La frecuencia relativa (f_i) de un valor X_i es la proporción de valores iguales a X_i en el conjunto de datos (X_1, X_2,…, X_N). Es decir, la frecuencia relativa es la frecuencia absoluta dividida por el número total de elementos *N*:

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

siendo $(X_1, X_2, ..., X_N)$ el conjunto de datos
y n_i el total de valores igual a X_i

Las frecuencias relativas son valores entre 0 y 1, $0 \leq f_i \leq 1$. La suma de las frecuencias relativas de todos los sujetos da 1. Si se multiplica la frecuencia relativa por cien se obtiene el porcentaje (tanto por cien %).

Frecuencia relativa acumulada

Definimos la frecuencia relativa acumulada (F_i) de un valor X_i como la proporción de valores iguales o menores a X_i en el conjunto de datos (X_1, X_2,…, X_N). Es decir, la frecuencia relativa acumulada es la frecuencia absoluta acumulada dividida por el número total de sujetos *N*:

$$F_i = \frac{N_i}{N}$$

siendo $(X_1, X_2, ..., X_N)$ el conjunto de datos
y N_i el total de valores igual o menor a X_i

La frecuencia relativa acumulada de cada valor siempre es mayor que la frecuencia relativa. De hecho, la frecuencia relativa acumulada de un elemento es la suma de las frecuencias relativas de los elementos menores o iguales a él, es decir:

$$F_i = f_1 + f_2 + \ldots + f_i$$

Ejercicio 1

Un profesor tiene la lista de las notas en matemáticas de 30 alumnos de su clase. Las notas son las siguientes:

NOTAS EN MATEMÁTICAS DE 30 ALUMNOS									
6	10	5	5	4	4	6	6	5	4
6	7	7	5	6	3	6	7	9	5
6	5	7	3	8	8	4	7	8	9

1) Frecuencia absoluta

Se realiza el recuento de la variable que se estudia (notas) para ver el número de veces que aparece cada nota. Una vez realizado el recuento, se representan las frecuencias absolutas de cada una de las notas (n_i). Las frecuencias son: $n_1(3) = 2$, $n_2(4) = 4$, $n_3(5) = 6$, $n_4(6) = 7$, $n_5(7) = 5$, $n_6(8) = 3$, $n_7(9) = 2$ y $n_8(10) = 1$.

X_i	Frecuencia absoluta (n_i)
3	2
4	4
5	6
6	7
7	5
8	3
9	2
10	1
Total	30

2) Frecuencia absoluta acumulada

Se calculan las frecuencias absolutas acumuladas (N_i) como la suma de las frecuencias absolutas de los valores menores o iguales a X_i:

X_i	Frecuencia absoluta (n_i)	Frecuencia absoluta acumulada (N_i)
3	2	2
4	4	6
5	6	12
6	7	19
7	5	24
8	3	27
9	2	29
10	1	30
Total	30	30

3) Frecuencia relativa

Se calcula la frecuencia relativa de cada elemento como la división de la frecuencia absoluta entre el total de elementos *N*=30.

- **$f_1(3)$** = $n_1(3)$ /N = 2/30 = **0,07**
- **$f_2(4)$** = $n_2(4)$ /N = 4/30 = **0,13**
- **$f_3(5)$** = $n_3(5)$ /N = 6/30 = **0,20**
- **$f_4(6)$** = $n_4(6)$ /N = 7/30 = **0,23**
- **$f_5(7)$** = $n_5(7)$ /N = 5/30 = **0,17**
- **$f_6(8)$** = $n_6(8)$ /N = 3/30 = **0,10**
- **$f_7(9)$** = $n_7(9)$ /N = 2/30 = **0,07**
- **$f_8(10)$** = $n_8(10)$ /N = 1/30 = **0,03**

X_i	Frecuencia absoluta (n_i)	Frecuencia relativa (fi = ni/N)	Frecuencia relativa (fi = ni/N) en %
3	2	0,07	7%
4	4	0,13	13%
5	6	0,20	20%
6	7	0,23	23%
7	5	0,17	17%
8	3	0,10	10%
9	2	0,07	7%
10	1	0,03	3%
Total	**30**	**1**	**100%**

Se pueden calcular las frecuencias relativas en porcentaje (%) multiplicándolas por 100.

4) Frecuencia relativa acumulada

Para obtener la frecuencia relativa acumulada se divide la frecuencia absoluta acumulada entre el número total de elementos (*N*=30). Esto da el tanto por uno de elementos iguales o menores al elemento que se estudia.

Las frecuencias relativas acumuladas son las siguientes:

$F_1(3) = f_1(3) = \mathbf{0,07}$

$F_2(4) = f_1(3) + f_2(4) = 0,07 + 0,13 = \mathbf{0,20}$

$F_3(5) = f_1(3) + f_2(4) + f_3(5) = 0,07 + 0,13 + 0,20 = \mathbf{0,40}$

$F_4(6) = f_1(3) + f_2(4) + f_3(5) + f_4(6) = 0,07 + 0,13 + 0,20 + 0,23 = \mathbf{0,63}$

$F_5(7) = f_1(3) + f_2(4) + f_3(5) + f_4(6) + f_5(7) = 0,07 + 0,13 + 0,20 + 0,23 + 0,17 = \mathbf{0,80}$

$F_6(8) = f_1(3) + f_2(4) + f_3(5) + f_4(6) + f_5(7) + f_6(8)$

$= 0,07 + 0,13 + 0,20 + 0,23 + 0,17 + 0,10 = \mathbf{0,90}$

$F_7(9) = f_1(3) + f_2(4) + f_3(5) + f_4(6) + f_5(7) + f_6(8) + f_7(9)$

$= 0,07 + 0,13 + 0,20 + 0,23 + 0,17 + 0,10 + 0,07 = \mathbf{0,97}$

$F_8(10) = f_1(3) + f_2(4) + f_3(5) + f_4(6) + f_5(7) + f_6(8) + f_7(9) + f_8(10)$

$= 0,07 + 0,13 + 0,20 + 0,23 + 0,17 + 0,10 + 0,07 + 0,03 = \mathbf{1,00}$

X_i	Frecuencia absoluta (n_i)	Frecuencia relativa ($f_i = n_i/N$)	Frecuencia relativa acumulada ($F_i = N_i/N$)	Frecuencia relativa acumulada (Fi=Ni/N) en %
3	2	0,07	0,07	7%
4	4	0,13	0,20	20%
5	6	0,20	0,40	40%
6	7	0,23	0,63	63%
7	5	0,17	0,80	80%
8	3	0,10	0,90	90%
9	2	0,07	0,97	97%
10	1	0,03	1,00	100%
Total	**30**	**1**	**1**	**100%**

Se pueden calcular las frecuencias relativas acumuladas en porcentaje (%) multiplicándolas por 100.

5) Tabla de frecuencias

Una vez se han calculado todas las frecuencias, se construye la tabla de frecuencias. La tabla es la siguiente:

X_i	Frecuencia absoluta (n_i)	Frecuencia absoluta acumulada (N_i)	Frecuencia relativa (fi = ni/N)	Frecuencia relativa acumulada (Fi=Ni/N)
3	2	2	0,07	0,07
4	4	6	0,13	0,20
5	6	12	0,20	0,40
6	7	19	0,23	0,63
7	5	24	0,17	0,80
8	3	27	0,10	0,90
9	2	29	0,07	0,97
10	1	30	0,03	1,00
Total	**30**	**30**	**1**	**1**

Adicionalmente, se pueden incluir dos columnas con los porcentajes de las frecuencias relativas y frecuencias relativas acumuladas. Se obtiene la siguiente tabla:

X_i	Frecuencia absoluta (n_i)	Frecuencia absoluta acumulada (N_i)	Frecuencia relativa (fi = ni/N)	Frecuencia relativa acumulada (Fi=Ni/N)	Frecuencia relativa (fi = ni/N) en %	Frecuencia relativa acumulada (Fi=Ni/N) en %
3	2	2	0,07	0,07	7%	7%
4	4	6	0,13	0,20	13%	20%
5	6	12	0,20	0,40	20%	40%
6	7	19	0,23	0,63	23%	63%
7	5	24	0,17	0,80	17%	80%
8	3	27	0,10	0,90	10%	90%
9	2	29	0,07	0,97	7%	97%
10	1	30	0,03	1,00	3%	100%
Total	**30**	**30**	**1**	**1**	**100%**	**100%**

Ejercicio 2

Los datos de lluvia caída en un día en 60 ciudades de una región, medida en l/m², han sido:

LLUVIA EN 60 LOCALIDADES					
23,2	17,6	15,7	16,2	19,9	3,4
4,2	16,6	8,8	23,6	4,5	9,5
23,8	17,0	13,2	5,8	12,2	26,4
24,0	10,1	14,7	21,2	17,7	7,7
2,8	18,2	18,0	23,0	19,0	15,0
15,2	18,3	26,2	5,1	14,8	11,7
3,4	22,1	17,2	23,4	19,8	19,4
22,4	20,6	2,2	9,8	21,8	3,9
22,8	20,9	25,7	18,9	20,2	7,2
25,5	16,0	21,0	11,2	25,4	22,4

Elaborar la correspondiente tabla de frecuencias con datos agrupados en clases:

Solución:

La amplitud total o rango se obtendrá de los valores extremos, que se han localizado entre los datos de la tabla inicial de precipitaciones:

$$R = X_{máx} - X_{mín} = 26{,}4 - 2{,}2 = 24{,}2$$

Fijamos el número de clases o intervalos por el método de Sturges, sabiendo que el total de observaciones son 60:

$$n_{int} = 1 + 3{,}322 \cdot \log N =$$
$$= 1 + 3{,}322 \cdot 1{,}778 = 6{,}9 \cong \mathbf{7}$$

Redondeamos el número de clases, n_{int}, a 7.

Se calcula la amplitud de cada clase, dividiendo el rango por el número de clases:

$$I = \frac{R}{n_{int}} = \frac{24{,}2}{7} = 3{,}45 \cong \mathbf{4}$$

Se redondea la amplitud de cada clase o intervalo a 4.

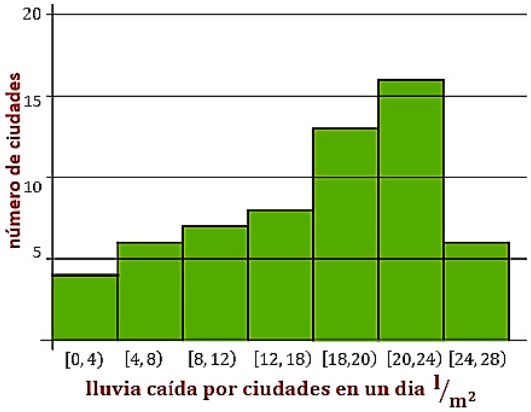
Ahora hay que recalcular un nuevo rango o amplitud total *R'*:

$$R' = I \cdot n_{int} = 4 \cdot 7 = \mathbf{28}$$

El exceso lo repartimos entre los extremos, que quedarán entre 0 y 28. Se puede ya construir la tabla de frecuencias buscada, poniendo antes en la segunda columna la marca de clase c_i, que será el punto medio entre los extremos de cada intervalo. El proceso para completar las columnas de las diferentes frecuencias es el mismo que en el ejercicio anterior. La tabla queda así:

l/m^2	marca de clase c_i	Frecuencia absoluta (n_i)	*Frecuencia* absoluta acumulada (N_i)	Frecuencia relativa $(f_i = n_i/N)$	Frecuencia relativa acumulada $(F_i = f_i/N)$
[0,4)	2	4	4	0,07	0,07
[4, 8)	6	6	10	0,10	0,17
[8, 12)	10	7	17	0,12	0,28
[12, 16)	14	8	25	0,13	0,42
[16, 20)	18	13	38	0,22	0,63
[20, 24)	22	16	54	0,27	0,90
[24, 28)	26	6	60	0,10	1,00
Total		**60**		**1**	

La tabla la vemos representada en este histograma:



1.6 Características de posición, dispersión y forma.

Medidas de posición

Las medidas de posición son indicadores estadísticos que permiten resumir los datos en uno solo, o dividir su distribución en intervalos del mismo tamaño.

Las medidas de posición se suelen dividir en dos grandes grupos: la de tendencia no central y las centrales. Las medidas de posición no centrales son los cuantiles. Estos realizan una serie de divisiones iguales en la distribución ordenada de los datos. De esta forma, reflejan los valores superiores, medios e inferiores. Los más habituales son:

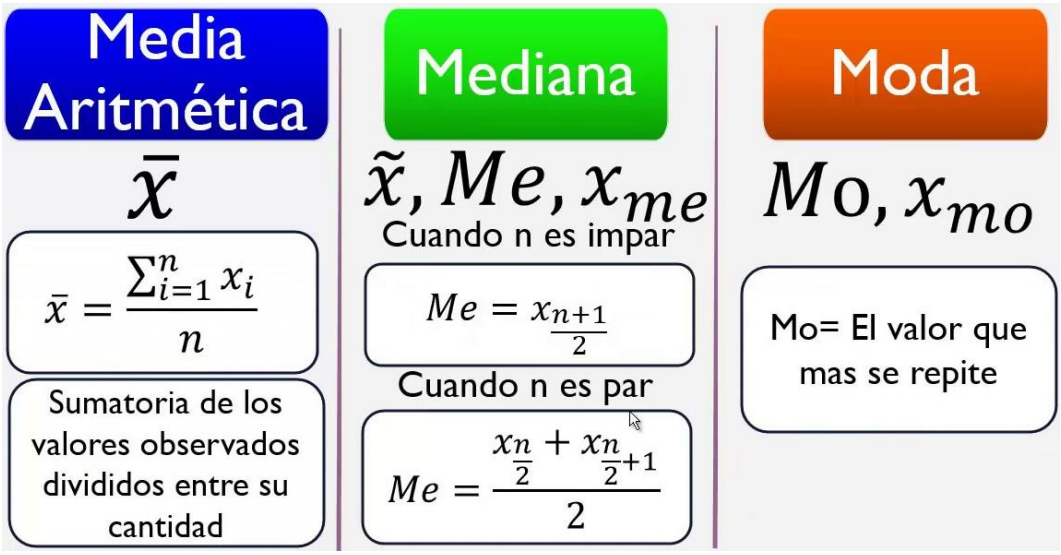
- **El cuartil:** Es uno de los más utilizados y divide la distribución en cuatro partes iguales. Así, existen tres cuartiles. Los valores inferiores de la distribución se sitúan por debajo del primero (Q1). La mitad o mediana son los valores menores iguales al cuartil dos (Q2) y los superiores son representados por el cuartil tres (Q3).
- **El quintil:** En este caso, divide la distribución en cinco partes. Por tanto, hay cuatro quintiles. Además, no existe ningún valor que divida la distribución en dos partes iguales. Es menos frecuente que el anterior.
- **El decil:** Estamos ante un cuantil que divide los datos en diez partes iguales. Existen nueve deciles, de D1 a D9. El D5 se corresponde con la mediana. Por su lado, los valores superiores e inferiores (equivalentes a los diferentes cuartiles) se sitúan en puntos intermedios entre estos.
- **El percentil:** Por último, este cuantil divide la distribución en cien partes. Hay 99 percentiles. Tiene, a su vez, una equivalencia con los deciles y cuartiles.

Cuartiles	Deciles	Percentiles
$\frac{k \cdot N}{4}$	$\frac{k \cdot N}{10}$	$\frac{k \cdot N}{100}$

Medidas de posición central

Estas nos permiten resumir la distribución de los datos en un solo valor central, alrededor del cual se sitúan; mientras que las segundas dividen la distribución en partes iguales.

- **La media aritmética, geométrica o armónica:** Son tres medidas centrales que nos indican un promedio ponderado de los datos. La primera es la más utilizada y la más conocida de las tres. La geométrica se aplica en series que muestran crecimientos porcentuales. Por su parte, la armónica es útil en el análisis de inversiones en bolsa.
- **La mediana:** En este caso, esta es la medida de posición central más reconocible. Divide la distribución en dos partes iguales. De esta forma, expresa el valor mediano, que no medio. Es muy útil en variables como los ingresos o salarios, a la vez que está muy relacionada con la media y algunos de los cuantiles vistos.
- **La moda:** Estamos ante una medida central de los valores más frecuentes. Por tanto, la moda nos informa sobre aquellos que se repiten en más ocasiones.



Medidas de dispersión

Las medidas de dispersión, o de variabilidad, expresan cómo se distribuyen los datos en torno a alguna de las medidas de centralización definidas antes, y son un complemento a estas últimas para describir más fielmente un conjunto de datos.

Varianza: La Varianza es una medida de dispersión que se utiliza para representar la variabilidad de un conjunto de datos respecto de la media aritmética de los mismo. Así, se calcula como la suma de los residuos elevados al cuadrado y divididos entre el total de observaciones.

Desviación estándar: La desviación estándar o desviación típica es una medida que ofrece información sobre la dispersión media de una variable. La desviación estándar es siempre mayor o igual que cero. Se obtiene al sacar la raíz cuadrada a la varianza.

$$\text{para la población } \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2 n_i}{N}$$

$$\text{para la muestra } S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2 n_i}{N-1}$$

Medidas de forma

Las medidas de forma son aquellas que nos muestran si una distribución de frecuencia tiene características especiales como simetría, asimetría, nivel de concentración de datos y nivel de apuntamiento que la clasifiquen en un tipo particular de distribución.

Para analizar estos aspectos recurriremos a dos tipos de medida:

- Coeficiente de asimetría de Fischer.
- Coeficiente de curtosis a apuntamiento de Fisher.

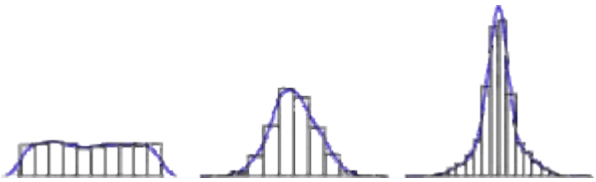
Coeficiente de asimetría de Fisher

Una distribución es simétrica cuando al trazar una vertical, en el diagrama de barras o histograma de una variable, según sea esta discreta o continua, por el valor de la media, esta vertical se transforma en eje de simetría y entonces decimos que la distribución es simétrica. En caso contrario, dicha distribución será asimétrica o diremos que presenta asimetría.

La asimetría puede ser de dos tipos:

- Asimétrica por la derecha.
- Asimétrica por la izquierda.

Coeficiente de curtosis o apuntamiento de Fisher



La otra medida de forma que vamos a considerar es el apuntamiento, al igual que con la simetría hemos de tomar una referencia para ver si la distribución de los datos es apuntada o no. La referencia citada es la distribución normal, y distinguiremos tres casos:

- Leptocúrtica, si la distribución es más picuda que la normal,
- Mesocúrtica, si la distribución es igual a la normal, y
- Platicúrtica, si la distribución es más aplastada que la normal.

1.7 Descripción numérica de una variable estadística bidimensional.

En numerosas ocasiones interesa estudiar simultáneamente dos (o más) caracteres de una población. En el caso de dos (o más) variables estudiadas conjuntamente se habla de variable bidimensional (multidimensional); si se trata de dos caracteres cualitativos, de par de atributos. Si de una cierta población se estudian dos caracteres simultáneamente se obtienen dos series de datos.

Variable estadística bidimensional es el conjunto de pares de valores de dos caracteres o variables estadísticas unidimensionales X e Y sobre una misma población.

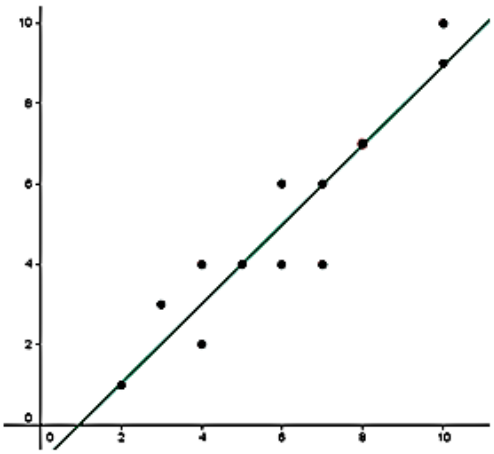
La variable estadística bidimensional se representa por el símbolo (X, Y) y cada uno de los individuos de la población viene caracterizado por la pareja (xi , yi), en el cual xi representa los datos, valores o marcas de clase x1, x2, ..., xn de la variable X; e yi representa los datos, valores o marcas de clase y1, y2, ..., ym de la variable Y.

Se denominan distribuciones bidimensionales a las tablas estadísticas bidimensionales formadas por todas las frecuencias absolutas de todos los posibles valores de la variable estadística bidimensional (X, Y). Las tablas estadísticas bidimensionales pueden ser: Simples y de doble entrada.

Ejemplo

Las notas de 12 alumnos de una clase en Matemáticas y Física son las siguientes:

Matemáticas	Física
2	1
3	3
4	2
4	4
5	4
6	4
6	6
7	4
7	6
8	7
10	9
10	10



1.8 Distribuciones marginales y condicionadas.

En teoría de probabilidades, la distribución marginal es la distribución de probabilidad de un subconjunto de variables aleatorias de un conjunto de variables aleatorias. La distribución marginal proporciona la probabilidad de un subconjunto de valores del conjunto sin necesidad de conocer los valores de las otras variables. Esto contrasta con la distribución condicional, que proporciona probabilidades contingentes sobre el valor conocido de otras variables.

El término variable marginal se usa para referirse a una variable del subconjunto de retenido y cuyos valores pueden ser conocidos.1 La distribución de las variables marginales, la distribución marginal, se obtiene marginalizando sobre la distribución de variables descartadas y las variables descartadas se llaman a veces variables marginalizadas.

Partiendo de una distribución bidimensional de frencuencias (xi, yj; nij), se denomina distribución condicionada de la variable X a un valor dado yj de la variable Y a la

distribución unidimensional definida por el conjunto de valores tomados por X y de las frencuencias condicionadas de dichos valores de X a qué Y tome el valor yj.

Análogamente, se denomina distribución de la variable y condicionada a un valor dado xi de la variable X a la distribución unidimensional definida por el conjunto de valores tomados por Y y de las frecuencias de dichos valores de Y condicionadas a que X tome el valor xi.

La función de probabilidad marginal es usada para hallar las diferentes distribuciones de probabilidad estadística de las variables individuales, con esta función podemos asignar diferentes valores a las variables conjuntas sin tener que relacionarlas, por ello se amplía las probabilidades de cada una de las variables.

La distribución marginal de X es simplemente la función de probabilidad de x, pero la palabra marginal sirve para distinguirla de la distribución conjunta de X e Y.
Una distribución marginal nos da la idea de la forma como depende una probabilidad con respecto a una sola variable.

La distribución marginal de dos variables aleatorias se puede obtener a partir de su distribución conjunta.

Para una variable aleatoria se puede especificar probabilidades para dicha variable sin tener en cuenta los valores de cuales quiera otras variables aleatorias.

Por ejemplo:

Genero	Messenger	Whatsup	Total
Hombres	254	356	610
Mujeres	169	221	390
	423	577	1000

P (H): 610/1000= 0.61

P (M): 390/1000= 0.39

P (Messenger) = 423/1000= 0.423

P (WhatsApp) = 577/1000= 0.577

De cada distribución bidimensional se pueden deducir dos distribuciones marginales: una correspondiente a la variable x, y otra correspondiente a la variable y, como en el ejemplo anterior: una distribución para Hombres y otra para Mujeres (sus totales)

Es cuando nos interesa conocer la distribución de un componente por separado, sin tener en cuenta a el otro componente. Eso se denomina "marginar", y la distribución de la variable por separado se llama "distribución marginal".

1.9 Independencia e incorrelación

Dos variables estadísticas son estadísticamente independientes cuando el comportamiento estadístico de una de ellas no se ve afectado por los valores que toma la otra; esto es cuando las relativas de las distribuciones condicionadas no se ven afectadas por la condición, y coinciden en todos los casos con las frecuencias relativas marginales.

Esta definición puede hacerse más operativa, a través de la caracterización siguiente: Dos variables son estadísticamente independientes cuando para todos los pares de valores se cumple que la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales.

Se dice que dos variables X e Y son independientes estadísticamente cuando la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales en todos los casos, es decir:

$$\frac{n_{ij}}{n} = \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}$$ Para todo i, j

Si esto no se cumple para todos los valores se dice que hay dependencia estadística.

Ejemplo: el suceso estatura de los alumnos de una clase y el color del pelo son independientes: el que un alumno sea más o menos alto no va a influir en el color de su cabello, ni viceversa.

Para que dos sucesos sean independientes tienen que verificar al menos una de las siguientes condiciones:

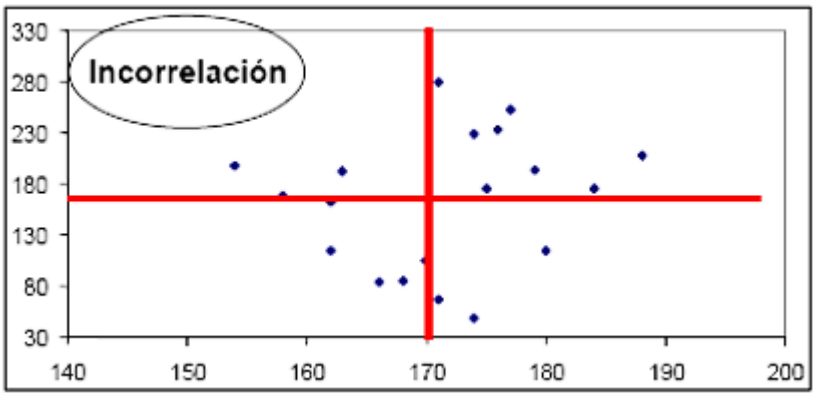
P (B/A) = P (B) es decir, que la probabilidad de que se dé el suceso B, condicionada a que previamente se haya dado el suceso A, es exactamente igual a la probabilidad de B.

Ejemplo: la probabilidad de que al tirar una moneda salga cara (suceso B), condicionada a que haga buen tiempo (suceso A), es igual a la propia probabilidad del suceso B.

P (A/B) = P (A) es decir, que la probabilidad de que se dé el suceso A, condicionada a que previamente se haya dado el suceso B, es exactamente igual a la probabilidad de A.

Incorrelación

Es el grado de dispersión entre los puntos de una variable, es decir, el cuándo los puntos no marchan en una misma dirección si no que están dispersos por todos lados, a diferencia de la correlación que es todo lo contrario.



Características numéricas

Los sistemas de numeración son conjuntos de dígitos usados para representar cantidades, así se tienen los sistemas de numeración decimal, binario, octal, hexadecimal, romano, etc.

Los cuatro primeros se caracterizan por tener una base (número de dígitos diferentes: diez, dos, ocho, dieciséis respectivamente) mientras que el sistema romano no posee base y resulta más complicado su manejo tanto con números, así como en las operaciones básicas.

Los sistemas de numeración que poseen una base tienen la característica de cumplir con la notación posicional, es decir, la posición de cada número le da un valor o peso, así el primer dígito de derecha a izquierda después del punto decimal, tiene un valor igual a b veces el valor del dígito, y así el dígito tiene en la posición n un valor igual a: (bn) * A dónde:

b = valor de la base del sistema

n = número del dígito o posición del mismo

A = dígito.

1.10 Regresión y correlación.

En forma más especifica el análisis de correlación y regresión comprende el análisis de los datos muéstrales para saber qué es y cómo se relacionan entre si dos o más variables en una población. El análisis de correlación produce un número que resume el grado de la correlación entre dos variables; y el análisis de regresión da lugar a una ecuación matemática que describe dicha relación.

El análisis de correlación generalmente resulta útil para un trabajo de exploración cuando un investigador o analista trata de determinar que variables son potenciales importantes, el interés radica básicamente en la fuerza de la relación. La correlación mide la fuerza de una entre variables; la regresión da lugar a una ecuación que describe dicha relación en términos matemáticos

Los datos necesarios para análisis de regresión y correlación provienen de observaciones de variables relacionadas.

Definiciones

En estadística, el análisis de la regresión es un proceso estadístico para estimar las relaciones entre variables. Incluye muchas técnicas para el modelado y análisis de diversas variables, cuando la atención se centra en la relación entre una variable dependiente y una o más variables independientes (o predictoras).

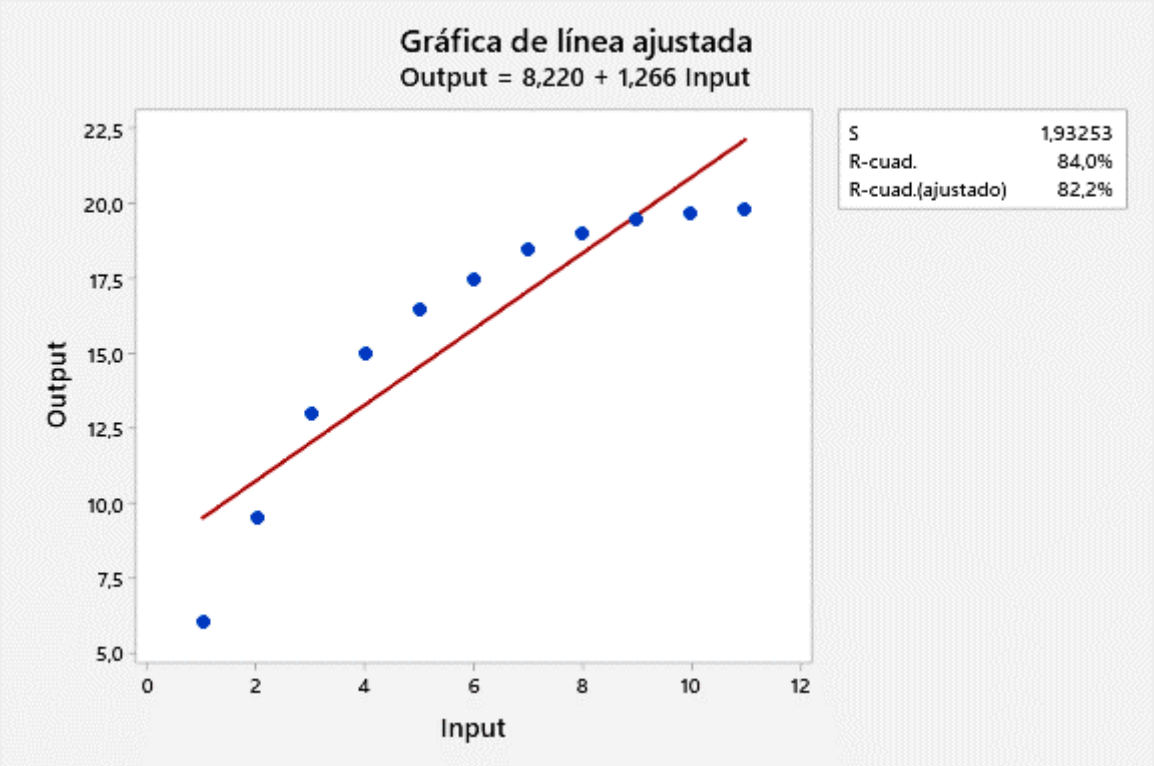
Más específicamente, el análisis de regresión ayuda a entender cómo el valor de la variable dependiente varía al cambiar el valor de una de las variables independientes, manteniendo el valor de las otras variables independientes fijas.

Más comúnmente, el análisis de regresión estima la esperanza condicional de la variable dependiente dadas las variables independientes - es decir, el valor promedio de la variable dependiente cuando se fijan las variables independientes. Con menor frecuencia, la atención se centra en un cuantil, u otro parámetro de localización de la distribución condicional de la variable dependiente dadas las variables independientes. En todos los casos, el objetivo de la estimación es una función de las variables independientes llamada la función de regresión.

En el análisis de regresión, también es de interés caracterizar la variación de la variable dependiente en torno a la función de regresión, la cual puede ser descrita por una distribución de probabilidad. Se denomina correlación al vínculo recíproco o correspondiente que existe entre dos o más elementos. El concepto se emplea de diferentes maneras de acuerdo al contexto. Correlación en el ámbito de las matemáticas y las estadísticas, la correlación alude a la proporcionalidad y la relación lineal que existe entre distintas variables. Si los valores de una variable se modifican de manera sistemática con respecto a los valores de otra, se dice que ambas variables se encuentran correlacionadas.

Curva de regresión y coeficiente de determinación.

La curva de regresión de Y sobre X visualiza como cambia la media de la variable Y de aquellos grupos de observaciones caracterizados por tener un mismo valor en la otra variable X. Es decir, como varía, por término medio, la variable Y en función de los valores de X. Por eso la variable Y recibe el nombre de variable dependiente y la variable X el de variable independiente.



Coeficiente de determinación

El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. Es también denominado R cuadrado y sirve para reflejar la bondad del ajuste de un modelo a la variable que se pretende explicar.

El coeficiente de determinación puede adquirir resultados que oscilan entre 0 y 1. Así, cuando adquiere resultados más cercanos a 1, mayor resultará el ajuste del modelo a la variable que se pretende aplicar para el caso en concreto. Por el contrario, cuando adquiere resultados que se acercan al valor 0, menor será el ajuste del modelo a la variable que se pretende aplicar y, justo por eso, resultará dicho modelo menos fiable.

El coeficiente de determinación es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, también llamado R cuadrado, refleja la bondad del ajuste de un modelo a la variable que pretender explicar.

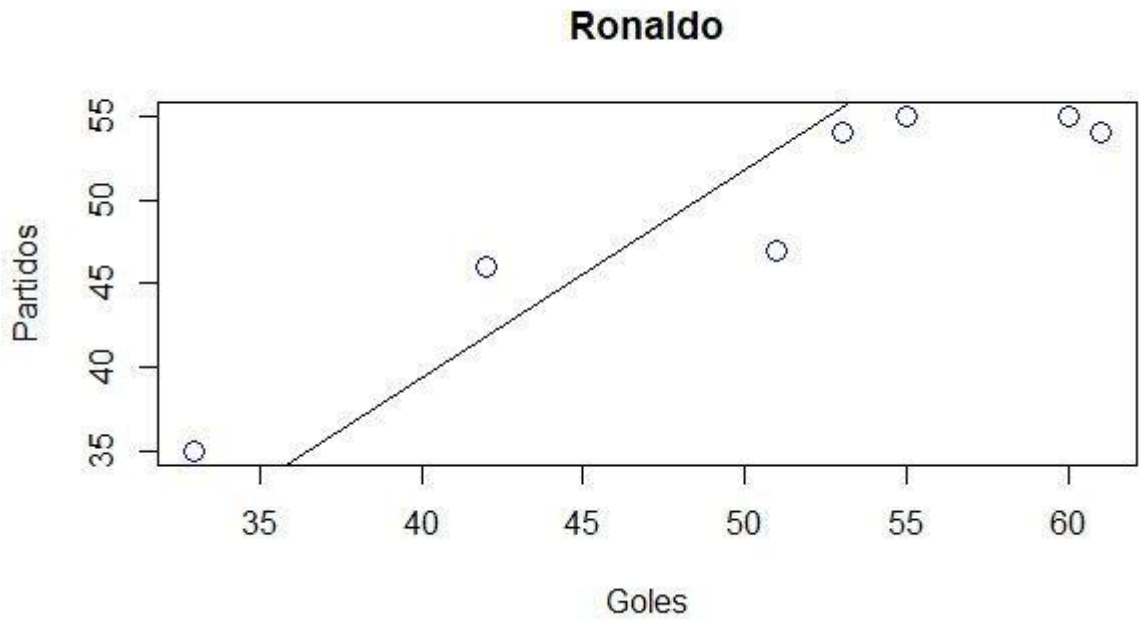
Es importante saber que el resultado del coeficiente de determinación oscila entre 0 y 1. Cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que

estamos intentando explicar. De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será.

$$R^2 = \frac{\sum_{t=1}^{T} (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{T} (Y_t - \bar{Y})^2}$$

Interpretación del coeficiente de determinación

Supongamos que queremos explicar la cantidad de goles que anota Cristiano Ronaldo según la cantidad de partidos que juega. Suponemos que, a mayor cantidad de partidos jugados, más goles meterá. Los datos pertenecen a las últimas 8 temporadas. De tal manera, tras extraer los datos, el modelo arroja la siguiente estimación:



Cómo podemos ver en el gráfico, la relación es positiva. A más partidos jugados, como es lógico, más goles anota en la temporada. El ajuste, según el cálculo del R cuadrado, es de 0,835. Esto quiere decir que es un modelo cuyas estimaciones se ajustan bastante bien a la variable real. Aunque técnicamente no sería correcto, podríamos decir algo así como que el modelo explica en un 83,5% a la variable real.

Regresión y correlación lineal.

Regresión Lineal

La regresión lineal simple comprende el intento de desarrollar una línea recta o ecuación matemática lineal que describe la reacción entre dos variables.

La regresión puede ser utilizada de diversas formas. Se emplean en situaciones en la que las dos variables miden aproximadamente lo mismo, pero en las que una variable es relativamente costosa, o, por el contrario, es poco interesante trabajar con ella, mientras que con la otra variable no ocurre lo mismo.

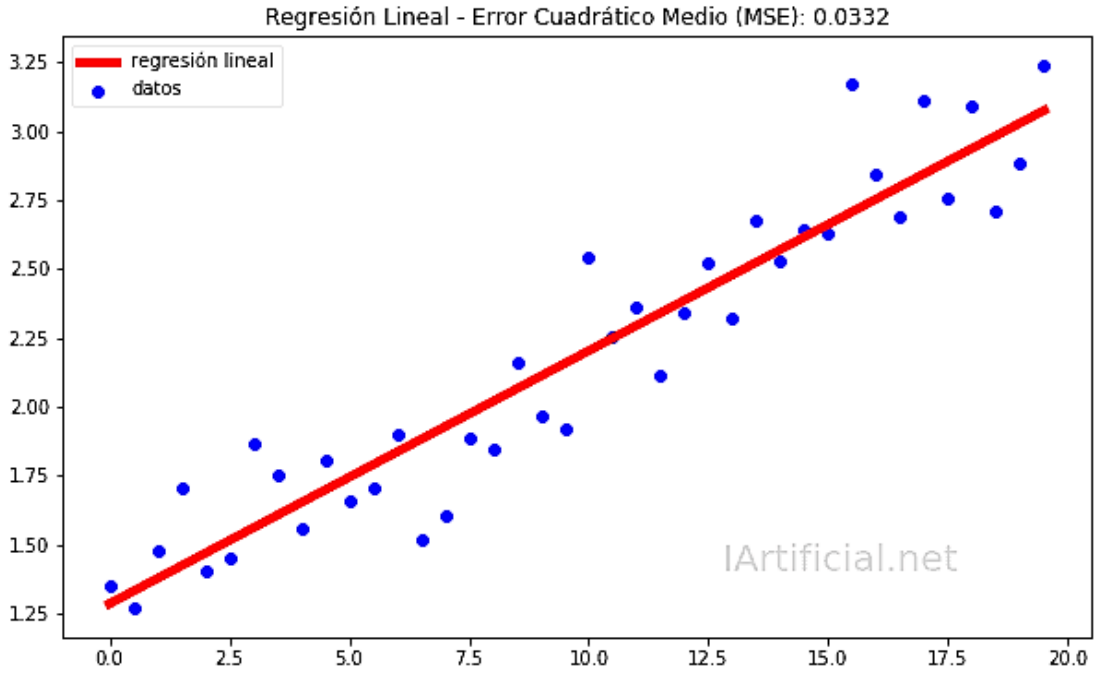
La finalidad de una ecuación de regresión seria estimar los valores de una variable con base en los valores conocidos de la otra.

Otra forma de emplear una ecuación de regresión es para explicar los valores de una variable en término de otra. Es decir, se puede intuir una relación de causa y efecto entre dos variables. El análisis de regresión únicamente indica qué relación matemática podría haber, de existir una. Ni con regresión ni con la correlación se pude establecer si una variable tiene "causa "ciertos valores de otra variable.

La fórmula para la regresión lineal con una sola variable *x* es:

$$y = Wx + b$$

Ejemplo de Regresión Lineal



Hemos usado una regresión lineal para encontrar los parámetros de la línea que minimiza el error de los datos que tenemos. El proceso de aprendizaje consiste en estimar los parámetros *w* y *b*. Así nos queda que, para estos datos, los mejores valores son:

W =0.0918

b =1.2859

Así que nos queda:

y = 0.0918x + 1.2859

Podemos usar este modelo de regresión lineal para estimar cuáles serán los resultados para otros valores de *x*. Por ejemplo, si queremos saber el resultado para *x* = 5, usaremos el modelo anterior y veremos que el resultado es 1.7449:

y = 0.0918 (5) + 1.2859 = 1.7449

Correlación Lineal

El coeficiente de correlación permite la medición de la correlación entre dos variables. Entre las ventajas por la que sobresale el coeficiente de correlación respecto a otras formas de medición de correlación, es la covarianza, los resultados del coeficiente de correlación son entre -1 y +1; y siendo su simpleza para comparar diferentes correlaciones de forma más directa y simple.

Si se analizan dos variables aleatorias X e Y relacionada con determinada población; el coeficiente de correlación de Pearson, la simbología con la letra Px, Y y se refiere a la expresión que permite calcular.

Teniendo en cuenta que:

- Es la covarianza de (X, Y)
- Es la desviación estándar de la variable X
- Es la desviación estándar de la variable Y

Se puede calcular de la siguiente forma, utilizando la fórmula:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Mientras que, adicionalmente se puede calcular el coeficiente sobre un estadístico muestral, reflejado en

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

Interpretación del valor del índice de correlación

Este varía en el intervalo [-1,1], estableciendo el signo el sentido de la relación, y la interpretación de cada resultado es el siguiente:

- Si r = 1: Correlación positiva perfecta. El índice refleja la dependencia total entre ambas dos variables, la que se denomina relación directa: cuando una de las variables aumenta, la otra variable aumenta en proporción constante.
- Si 0 < r < 1: Refleja que se da una correlación positiva.
- Si r = 0: En este caso no hay una relación lineal. Aunque no significa que las variables sean independientes, ya que puede haber relaciones no lineales entre ambas variables.
- Si -1 < r < 0: Indica que existe una correlación negativa.
- Si r = -1: Indica una correlación negativa perfecta y una dependencia total entre ambas variables lo que se conoce como "relación inversa", que es cuando una de las variables aumenta, la otra variable en cambio disminuye en proporción constante.

La correlación refleja la medida de asociación entre variables. Si se aplica en probabilidad y estadística, la correlación permite conocer la fuerza y dirección de la relación lineal que se dé entre dos variables aleatorias.

1.11 Otros tipos de regresión.

Regresión Múltiple: Este tipo se presenta cuando dos o más variables independientes influyen sobre una variable dependiente. Ejemplo: Y = f (x, w, z).

Análisis de Regresión Múltiple Dispone de una ecuación con dos variables independientes adicionales:

$$Y' = a' + b_1 x_1 + b_2 x_2$$

Se puede ampliar para cualquier número "m" de variables independientes:

$$Y' = a' + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_m x_m$$

Para poder resolver y obtener a, b_1 y b_2 en una ecuación de regresión múltiple el cálculo se presenta muy tediosa porque se tiene atender 3 ecuaciones que se generan por el método de mínimo de cuadrados:

$$\sum y = na + b_1 \sum x_1 + b_2 \sum x_2$$

$$\sum x_1 y = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$$

$$\sum x_2 y = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$

Para poder resolver se puede utilizar programas como AD+, SPSS y Minitab y Excel.

El error estándar de la regresión múltiple

Es una medida de dispersión la estimación se hace más precisa conforme el grado de dispersión alrededor del plano de regresión se hace más pequeño. Para medirla se utiliza la fórmula:

$$S_{xy} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - m - 1}}$$

Y : Valores observados en la muestra

$\hat{Y}$: Valores estimados a partir a partir de la ecuación de regresión

n : Número de datos

m : Número de variables independientes

El coeficiente de determinación múltiple

Mide la tasa porcentual de los cambios de Y que pueden ser explicados por x_1, x_2 y x_3 simultáneamente.

$$r^2 = \frac{SCregresiòn}{SCTotal}$$

Ejemplo de aplicación:

En la Facultad de Ingeniería de Sistemas se quiere entender los factores de aprendizaje de los alumnos que cursan la asignatura de PHP, para lo cual se escoge al azar una muestra de 15 alumnos y ellos registran notas promedios en las asignaturas de Algoritmos, Base de Datos y Programación como se muestran en el siguiente cuadro.

Alumno	PHP	Algoritmos	Base de Datos	Programación
1	13	15	15	13
2	13	14	13	12
3	13	16	13	14
4	15	20	14	16
5	16	18	18	17
6	15	16	17	15
7	12	13	15	11
8	13	16	14	15
9	13	15	14	13
10	13	14	13	10
11	11	12	12	10
12	14	16	11	14
13	15	17	16	15
14	15	19	14	16
15	15	13	15	10

Lo que buscamos es construir un modelo para determinar la dependencia que exista de aprendizaje reflejada en las notas de la asignatura de PHP, conociendo las notas de las asignaturas Algoritmos, Base de Datos y Programación. Se presentará la siguiente ecuación a resolver:

$$Y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Utilizando las fórmulas de las ecuaciones normales a los datos obtendremos los coeficientes de regresión o utilizando Regresión de Análisis de datos, en la Hoja de Cálculo de Excel podemos calcular también los coeficientes de regresión:

	Coeficientes	Error típico	Estadístico t	p-valor	Inferior 95%	Superior 95%
Intercepción	2.551	2.369	1.077	0.305	-2.663	7.766
Algoritmos	0.583	0.267	2.186	0.051	-0.004	1.169
Base de Datos	0.373	0.144	2.589	0.025	0.056	0.691
Programación	-0.242	0.270	-0.893	0.391	-0.837	0.354

Por lo tanto, podemos construir la ecuación de regresión que buscamos:

$$Y = 2.551 + 0.583x_1 + 0.373x_2 - 0.242x_3$$

El Error Estándar de Regresión Múltiple

Mediante esta medida de dispersión se hace más preciso el grado de dispersión alrededor del plano de regresión, se hace más pequeño. Para calcularla se utiliza la formula siguiente:

$$S_{yx} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-m-1}}$$

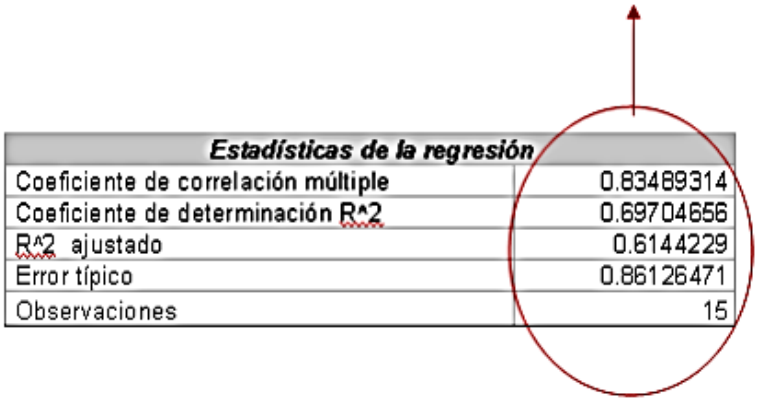
En los resultados de Excel se llama error típico y para explicar la relación del aprendizaje de PHP que se viene desarrollando es de 0.861

El coeficiente de determinación múltiple

Utilizaremos para determinar la tasa porcentual de Y para ser explicados las variables múltiples, utilizando la siguiente formula:

$$r^2 = \frac{SCregresiòn}{SCTotal}$$

$$r^2 = \frac{18.7737874}{26.9333333} = 0.69704656$$



Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.83489314
Coeficiente de determinación R^2	0.69704656
R^2 ajustado	0.6144229
Error típico	0.86126471
Observaciones	15

Conclusiones

El 69.70% del aprendizaje del Curso de PHP puede ser explicado mediante las notas obtenidas por las asignaturas de Algoritmos, Base de Datos y Programación.

1.12 Análisis de atributos

Su principal objetivo es el de evitar un error muy común consistente en tratar de encontrar la forma de mejorar un producto, servicio o proceso analizándolo como un todo. Muchas veces, la búsqueda de una idea global, salvadora, que mejore el todo, impide descubrir la característica específica que, por sí sola, podría producir el resultado deseado.

Características para las Gráficas de Control de Atributos

- Están basadas en decisiones de pasa/no pasa.
- Se pueden aplicar en casi cualquier operación donde se recolectan datos.
- Se utilizan en características de calidad que no pueden ser medidas o que son costosas o difíciles de medir. A diferencia de las gráficas de control de datos variables, las gráficas de datos atributos se pueden establecer para una característica de calidad o para muchas.

Tipos de Gráficas de Atributos:

- Defectivos
  – np - número de unidades no-conformes
  – p - proporción de unidades no-conformes
- Defectos
  – c - número de defectos
  – u - proporción de defectos

UNIDAD II: CALCULO DE PROBABILIDADES

En la vida cotidiana aparecen muchas situaciones en las que los resultados observados son diferentes, aunque las condiciones iniciales en las que se produce la experiencia sean las mismas. Por ejemplo, al lanzar una moneda unas veces resultará cara y otras, cruz. Estos fenómenos, denominados aleatorios, se ven afectados por la incertidumbre.

En el lenguaje habitual, frases como "probablemente...", "es poco probable que...", "hay muchas posibilidades de que..." hacen referencia a esta incertidumbre.

La teoría de la probabilidad pretende ser una herramienta para modelizar y tratar con situaciones de este tipo. Por otra parte, cuando aplicamos las técnicas estadísticas a la recogida, análisis e interpretación de los datos, la teoría de la probabilidad proporciona una base para evaluar la fiabilidad de las conclusiones alcanzadas y las inferencias realizadas.

El objetivo del Cálculo de Probabilidades es el estudio de métodos de análisis del comportamiento de fenómenos aleatorios.

2.1 La medida de probabilidad. Espacio Probabilístico

Para medir la incertidumbre existente en un experimento aleatorio1 dado, se parte de un espacio muestral M en el que se incluyen todos los posibles resultados individuales del experimento (sucesos elementales); es decir, el conjunto muestral es un conjunto exhaustivo (contiene todas las posibles ocurrencias) y mutuamente exclusivo (no pueden

darse dos ocurrencias a la vez). Una vez definido el espacio muestral, el objetivo consiste en asignar a todo suceso compuesto $A \subset M$ un número real que mida el grado de incertidumbre sobre su ocurrencia. Para obtener medidas con significado matemático claro y practico, se imponen ciertas propiedades intuitivas que definen una clase de medidas que se conocen como medidas de probabilidad.

Definición Medida de Probabilidad. Una función p que proyecta los subconjuntos $A \subset M$ en el intervalo [0, 1] se llama medida de probabilidad si satisface los siguientes axiomas:

Axioma 1: Un experimento se denomina aleatorio cuando puede dar resultados distintos al realizarse en las mismas condiciones (por ejemplo, lanzar un dado al aire y observar el número resultante).

Formalmente, una medida de probabilidad se define sobre una σ-algebra del espacio muestral, que es una colección de subconjuntos que es cerrada para los operadores de unión $A \cup B$ y complementario $A = M \setminus A$ (también para intersecciones $A \cap B = A \cup B$).

Sin embargo, optamos por una definición menos rigurosa y más intuitiva para introducir este concepto.

Axioma 2: Para cualquier sucesión infinita, A1, A2, . . ., de subconjuntos disjuntos de M, se cumple la igualdad. El Axioma 1 establece que, independientemente de nuestro grado de certeza, ocurrirá un elemento del espacio muestral M (es decir, el conjunto M es exhaustivo). El Axioma 2 es una fórmula de agregación que se usa para calcular la probabilidad de la unión de subconjuntos disjuntos. Establece que la incertidumbre de un cierto subconjunto es la suma de las incertidumbres de sus partes (disjuntas). Nótese que esta propiedad también se cumple para sucesiones finitas.

En general un espacio probabilístico está integrado por tres componentes. Primero, el conjunto (llamado espacio muestral) de los posibles resultados del experimento, llamados sucesos elementales. Segundo, por la colección de todos los sucesos aleatorios (no solo los elementales), que es una σ-álgebra sobre. El par es lo que se conoce como un espacio de medida. Por último, una medida de probabilidad o función de probabilidad, que asigna una probabilidad a todo suceso y que verifica los llamados axiomas de Kolmogórov.

2.2 Probabilidad condicionada.

Miraremos la forma en que cambia la probabilidad de un suceso A cuando se sabe que otro suceso B ha ocurrido.

A esta probabilidad se le denomina la probabilidad condicional del suceso A dado que el suceso B ha ocurrido.

La notación para esta probabilidad condicional es $P(A|B)$. Por conveniencia, esta notación se lee simplemente como la probabilidad condicional de A dado B.

Entonces, sean A y B dos sucesos cualesquiera de un mismo espacio muestral E, tales que $P(B) > 0$, así:

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

Ejemplo de probabilidad condicional

En un grupo de 100 estudiantes, 35 jóvenes juegan al fútbol y al baloncesto, mientras que 80 de los miembros practican fútbol. ¿Cuál es la probabilidad de que uno de los estudiantes que juega al fútbol, también juegue al baloncesto o básquet?

Como se puede advertir, en este caso conocemos dos datos: los estudiantes que juegan al fútbol y al baloncesto (35) y los estudiantes que juegan al fútbol (80).

Evento A: Que un estudiante juegue al baloncesto (*x*)
Evento B: Que un estudiante juegue al fútbol (*80*)
Evento A y B: Que un estudiante juegue al fútbol y al baloncesto (*35*)
P (A / B) = P (A∩B) / P (B)
P (A / B) = 35 / 80
P (A / B) = 0,4375
P (A / B) = 43,75%

Por lo tanto, esta probabilidad condicional indica que la probabilidad de que un estudiante juegue al baloncesto dado que también juega al fútbol es del 43,75%.

Probabilidad condicional para sucesos independientes

Dos sucesos, A y B, son independientes cuando la probabilidad de que suceda A no se ve afectada porque haya sucedido, o no, B.

Por ejemplo, Si tiramos dos veces una moneda, el segundo resultado que obtenemos no está influenciado por el primer resultado obtenido.

Si dos sucesos A y B son independientes, entonces $\mathbf{P(A \cap B) = P(A)P(B)}$.

Por tanto, si $P(B) \neq 0$, de la definición de probabilidad condicional resulta que:

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

En otras palabras, si dos sucesos A y B son independientes, entonces la probabilidad condicional de A cuando se sabe que B ha ocurrido es la misma que la probabilidad incondicional de A cuando no se dispone de información sobre B. El resultado recíproco también es cierto, si:

$$\mathbf{P(A|B) = P(A)}$$

entonces los sucesos A y B deben ser independientes.

Sucesos dependientes

Dos sucesos, A y B, son dependientes cuando la probabilidad de que suceda A se ve afectada porque haya sucedido, o no, B.

Dos sucesos A y B son dependientes si:

$$\mathbf{P(A|B) \neq P(A)}$$

2.3 Teoremas asociados.

El teorema de Bayes es utilizado para calcular la probabilidad de un suceso, teniendo información de antemano sobre ese suceso.

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

El teorema de Bayes ha sido muy cuestionado. Lo cual se ha debido, principalmente, a su mala aplicación. Ya que, mientras se cumplan los supuestos de sucesos disjuntos y exhaustivos, el teorema es totalmente válido.

Fórmula del teorema de Bayes

Para calcular la probabilidad tal como la definió Bayes en este tipo de sucesos, necesitamos una fórmula. La fórmula se define matemáticamente como:

$$P[A_n/B] = \frac{P[B/A_n] \cdot P[A_n]}{\sum P[B/A_i] \cdot P[A_i]}$$

Donde B es el suceso sobre el que tenemos información previa y A(n) son los distintos sucesos condicionados. En la parte del numerador tenemos la probabilidad condicionada, y en la parte de abajo la probabilidad total. En cualquier caso, aunque la fórmula parezca un poco abstracta, es muy sencilla. Para demostrarlo, utilizaremos un ejemplo en el que en lugar de A (1), A (2) y A (3), utilizaremos directamente A, B y C.

P(A) = 0,40 **P(D/A)** = 0,02

P(B) = 0,30 **P(D/B)** = 0,03

P(C) = 0,30 **P(D/C)** = 0,05

1. Si un envase ha sido fabricado por la fábrica de esta empresa en Estados Unidos ¿Cuál es la probabilidad de que sea defectuoso?

Se calcula la probabilidad total. Ya que, a partir los diferentes sucesos, calculamos la probabilidad de que sea defectuoso.

P(D) = [P(A) x P(D/A)] + [P(B) x P(D/B)] + [P(C) x P(D/C)] = [0,4 x 0,02] + [0,3 x 0,03] + [0,3 x 0,05] = 0,032

Expresado en porcentaje, diríamos que la probabilidad de que un envase fabricado por la fábrica de esta empresa en Estados Unidos sea defectuoso es del 3,2%.

2. Siguiendo con la pregunta anterior, si se adquiere un envase y este es defectuoso ¿Cuáles es la probabilidad de que haya sido fabricado por la máquina A? ¿Y por la máquina B? ¿Y por la máquina C?

Aquí se utiliza el teorema de Bayes. Tenemos información previa, es decir, sabemos que el envase es defectuoso. Claro que, sabiendo que es defectuoso, queremos saber cuál es la probabilidad de que se haya producido por una de las máquinas.

P(A/D) = [P(A) x P(D/A)] / P(D) = [0,40 x 0,02] / 0,032 = 0,25

P(B/D) = [P(B) x P(D/B)] / P(D) = [0,30 x 0,03] / 0,032 = 0,28

P(C/D) = [P(C) x P(D/C)] / P(D) = [0,30 x 0,05] / 0,032 = 0,47

Sabiendo que un envase es defectuoso, la probabilidad de que haya sido producido por la máquina A es del 25%, de que haya sido producido por la máquina B es del 28% y de que haya sido producido por la máquina C es del 47%.

2.4 Variable aleatoria.

Se llama variable aleatoria a toda función que asocia a cada elemento del espacio muestral E un número real.

Se utilizan letras mayúsculas $X, Y, \ldots$ para designar variables aleatorias, y las respectivas minúsculas $(x, y, \ldots)$ para designar valores concretos de las mismas.

Tipos de variable aleatoria

Dentro de las variables aleatorias existen, fundamentalmente, dos tipos. Su clasificación, depende del tipo de número que arroja la función matemática. Una variable aleatoria puede ser de dos tipos:

- **Variable aleatoria discreta:** Una variable aleatoria es discreta si los números a los que da lugar son números enteros. La forma de calcular las probabilidades de una variable aleatoria discreta es a través de la función de probabilidad.
- **Variable aleatoria continua:** Una variable aleatoria es continua en caso de que los números a los que dé lugar no sean números enteros. Es decir, tengan decimales. La probabilidad de que se dé un suceso determinado correspondiente a una variable aleatoria continua, viene establecida por la función de densidad.

Ejemplo de variable aleatoria

Una variable aleatoria bien podría ser la función de los resultados del lanzamiento de un dado. Es importante diferenciar aquí tres conceptos.

- **Dado:** No es la variable aleatoria. El dado es simplemente un objeto.
- **Lanzamiento de un dado:** No es la variable aleatoria. El lanzamiento de un dado es el experimento aleatorio.
- **Resultados del lanzamiento de un dado:** Sí es la variable aleatoria. Es la función que recoge los resultados del lanzamiento del dado. Un ejemplo de variable aleatoria podría ser: Que salga un número mayor que 2 al lanzar el dado.

X: Que salga mayor que 2 al lanzar el dado

Distribución de probabilidad: 1/3 no sale mayor que 2 y 2/3 si sale mayor que 2.

Es decir, la probabilidad se distribuye tal que la probabilidad de que salga un número menor o igual que 2 es de 1/3. Mientras, la probabilidad de que salga mayor que 2 es 2/3

Por tanto, nuestra variable aleatoria dependerá del resultado concreto del valor del dado. El tipo de variable al que estamos haciendo referencia es discreta. ¿Por qué lo sabemos? Porque cuando tiramos un dado solo podemos obtener 6 posibles resultados. Todos ellos, son números enteros. Concretamente, entre 1 y 6.

2.5 Concepto de variable aleatoria. Probabilidad inducida

Una variable es un símbolo que actúa en las funciones, las fórmulas, los algoritmos y las proposiciones de las matemáticas y la estadística. Según sus características, las variables se clasifican de distinto modo.

Variable aleatoria

Se denomina variable aleatoria (o estocástica) a la función que adjudica eventos posibles a números reales (cifras), cuyos valores se miden en experimentos de tipo aleatorio. Estos valores posibles representan los resultados de experimentos que todavía no se llevaron a cabo o cantidades inciertas. Cabe destacar que los experimentos aleatorios son aquellos que, desarrollados bajo las mismas condiciones, pueden ofrecer resultados diferentes. Arrojar una moneda al aire para ver si sale cara o ceca es un experimento de este tipo.

La variable aleatoria, en definitiva, permite ofrecer una descripción de la probabilidad de que se adoptan ciertos valores. No se sabe de manera precisa qué valor adoptará la variable cuando sea determinada o medida, pero sí se puede conocer cómo se distribuyen las probabilidades vinculadas a los valores posibles. En dicha distribución incide el azar.

Tal como hemos comentado, la definición formal de variable aleatoria impone una restricción matemática en la formulación vista hasta el momento. Definiremos una variable aleatoria como una aplicación de Ω en el conjunto de números reales, es decir, para todo número real x, el conjunto de resultados elementales tales que la variable aleatoria toma sobre ellos valores inferiores o iguales a x ha de ser un suceso sobre el cual podamos definir una probabilidad.

Dicha propiedad recibe el nombre de medibilidad y por tanto podríamos decir que una variable aleatoria es una función medible de Ω en los reales. Esta condición nos asegura que podremos calcular sin problemas, probabilidades sobre intervalos de la recta real a partir de las probabilidades de los sucesos correspondientes.

La expresión anterior se leería de la manera siguiente: La probabilidad de que la variable aleatoria tome valores inferiores o iguales a x es igual a la probabilidad del suceso formado por el conjunto de resultados elementales sobre los que el valor de la variable es menor o igual que x.

La probabilidad obtenida de esta manera se denomina probabilidad inducida. Se puede comprobar que, a partir de la condición requerida, se pueden obtener probabilidades sobre cualquier tipo de intervalo de la recta real.

2.6 Función de distribución.

En la teoría de la probabilidad y en estadística, la Función de Distribución Acumulada (FDA, designada también a veces simplemente como FD) o función de probabilidad acumulada asociada a una variable aleatoria real: X (mayúscula) sujeta a cierta ley de distribución de probabilidad, es una función matemática de la variable real: x (minúscula); que describe la probabilidad de que X tenga un valor menor o igual que x.

Intuitivamente, asumiendo la función f como la ley de distribución de probabilidad, la FDA sería la función con la recta real como dominio, con imagen del área hasta aquí de la función f, siendo aquí el valor x para la variable aleatoria real X.

La FDA asocia a cada valor x, la probabilidad del evento: "la variable X toma valores menores o iguales a x". El concepto de FDA puede generalizarse para modelar variables aleatorias multivariantes.

2.7 Variables aleatorias discretas y continuas

Una variable aleatoria es una función que asigna un valor numérico, al resultado de un experimento aleatorio. Una variable aleatoria puede ser discreta o continua.

- **Las variables aleatorias discretas** son aquellas que presentan un número contable de valores; por ejemplo, el número de personas que viven en una casa (3, 5 o 9).
- **Las variables aleatorias continuas** son aquellas que presentan un número incontable de valores; por ejemplo, el peso de las vacas en una granja (una vaca puede pesar 632.12 kg, otra puede pesar 583.12312 kg, otra 253.12012 kg, otra 198.0876 kg y nunca terminaríamos de enumerar todos los posibles valores). Como estas definiciones son muy difíciles de entender a simple vista, vamos a explicarlas a detalle.

Variable aleatoria

Una variable aleatoria es una función que asigna un valor numérico, al resultado de un experimento aleatorio. Recordemos que el resultado de un experimento aleatorio depende del azar. Veamos los ejemplos.

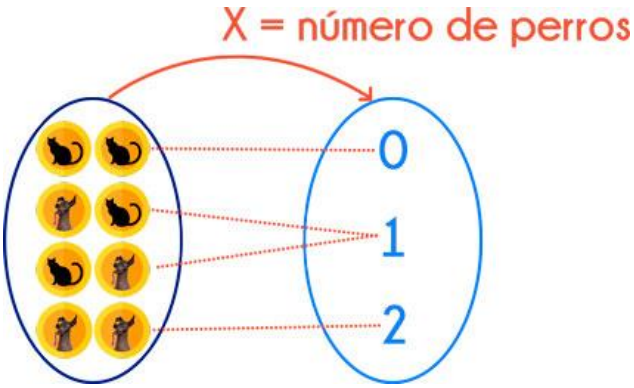
Ejemplo 1 de variable aleatoria

Tenemos una moneda que en sus caras tiene por un lado un gato y por el otro, un perro.

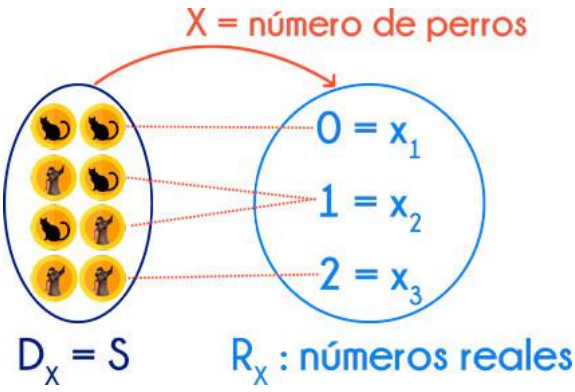
Vamos a realizar un experimento aleatorio que consiste en lanzar 2 monedas. Colocaremos los resultados en el siguiente gráfico:

Definimos nuestra variable aleatoria X:

- X = número de perros.



Ten en cuenta que la variable aleatoria siempre va con letras mayúsculas (en este caso X), mientras que los valores de su rango siempre con letras minúsculas (x_1 , x_2 , x_3).



Los valores del rango de esta variable aleatoria son:

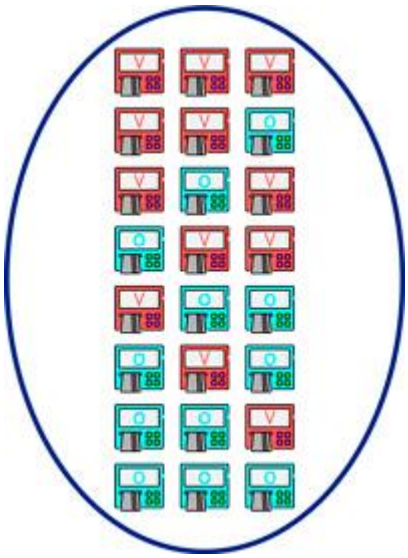
- $x_1 = 0$
- $x_2 = 1$
- $x_3 = 2$

En el dominio de la función tenemos el espacio muestral, es decir, todos los resultados posibles de nuestro experimento aleatorio. Mientras que el rango tenemos un conjunto de números reales.

Ejemplo 2 de variable aleatoria

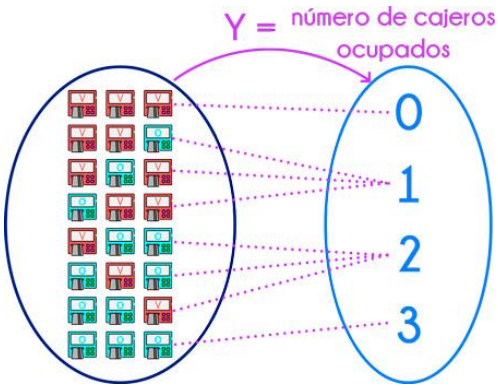
En un banco hay 3 cajeros automáticos. Vamos a realizar un experimento aleatorio que consiste en ir al banco a una hora al azar del día y ver qué cajeros están ocupados y qué cajeros están vacíos.

Colocamos en el siguiente gráfico los resultados, los cajeros vacíos (V) irán de color rojo y los ocupados (O) de color verde.

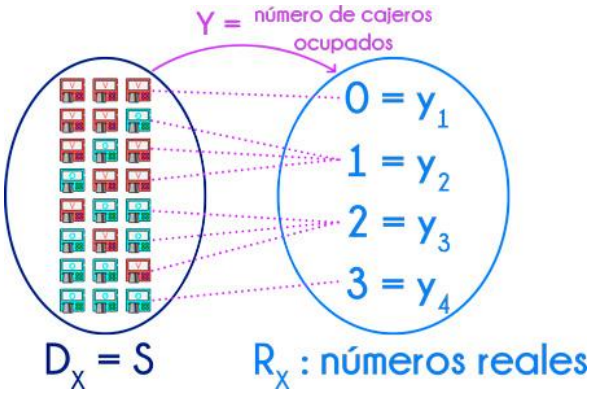


Definimos nuestra variable aleatoria Y:

- Y = número de cajeros automáticos ocupados.



Ten en cuenta que la variable aleatoria siempre va con letras mayúsculas (en este caso Y), mientras que los valores de su rango siempre con letras minúsculas (en este caso y_1 , y_2 , y_3, y_4).



Los valores del rango de esta variable aleatoria son:

- $y_1 = 0$
- $y_2 = 1$
- $y_3 = 2$
- $y_4 = 3$

Las variables aleatorias se clasifican en discretas o continuas en función de los valores numéricos que asumen. Veamos esto a detalle.

Variable aleatoria discreta

Una variable aleatoria discreta es aquella que puede asumir un número contable de valores.

Por ejemplo, si realizamos el experimento de salir a calle y seleccionar 10 personas al azar para un examen sorpresa de matemáticas, podemos definir la variable aleatoria A:
A = número de personas que aprobaron el examen.

Los valores que asume A (en su rango), van del 0 al 10 (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10). El rango lo expresaríamos de la siguiente manera:

$R_A = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

La variable aleatoria A asume un número contable de valores, por ello, es una variable aleatoria discreta.

Otro ejemplo, vamos a realizar el experimento de registrar los automóviles a una caseta de peaje. Podemos definir la variable aleatoria B:
B = número de vehículos que llegan durante el periodo de un día.

Los valores que asume V (en su rango), son 0, 1, 2, 3, 4, 5, …; así sean muchos vehículos los que llegan, siempre podremos contar la cantidad de valores que asume V. Por ello, la variable V es una variable aleatoria discreta.

Variable aleatoria continua

Es aquella que puede asumir un número incontable de valores. Por ejemplo: Si realizamos el experimento de ir a una granja y estudiamos las características de las vaquitas, podemos definir la variable aleatoria C:
C = peso de una vaca en la granja de Jorge (en kilogramos).

Alguna vaquita puede pesar 425.1872 kg; otra puede pesar 612.5874541 kg; otra puede pesar 545.897512121 kg. Si tomamos más vacas, podríamos tener más valores y nunca terminaríamos.

Se conoce que el becerro más pequeño tiene un peso de 30 kg, y la vaca más grande tiene un peso de 1000 kg.

Y así, tendríamos un número incontable de valores para el rango de esta variable. El rango de esta variable puede ser cualquier valor dentro del intervalo que va desde 30 kg hasta 1000 kg. Por ello, se trata de una variable aleatoria continua.

Otro ejemplo: Si vamos a una agencia del banco y registramos los datos de atención a los clientes, podemos definir la variable aleatoria D:

- D = tiempo de atención a los clientes del banco (en segundos).

Un cliente puede ser atentido en 24.123 S; otro cliente en 72.32142 S; otro en 51.123123 S. Si seguimos tomando más clientes, tendríamos más valores. Se conoce además que el tiempo mínimo de atención en ventanilla es de 1 segundo y el tiempo máximo es de 240 segundos.

Y así, tendríamos un número incontable de valores para el rango de esta variable. El rango de esta variable puede ser cualquier valor dentro del intervalo que va desde 1 s hasta 240 s. Por ello, se trata de una variable aleatoria continua.

En general, las variables aleatorias discretas representan datos que provienen del conteo del número de elementos, mientras que, las variables aleatorias continuas representan datos que provienen de mediciones, por ejemplo, tiempo, peso, longitud, etc.

2.8 Características de una variable

Las variables como entidades empíricas del problema de investigación presentan un conjunto de características significativas tales como:

- Están contenidas esencialmente en el título, el problema, el objetivo y las respectivas hipótesis de la investigación. En virtud de ello es que no se puede agregar nuevas variables de las que ya existen en los ítems mencionados.

- Son aspectos que cambian o adoptan distintos valores. Esto significa que las variables al ser medidas y observadas expresan diferencias entre los rasgos, cualidades y atributos de las unidades de análisis.
- Son enunciados que expresan rasgos característicos de los problemas medibles empíricamente. Estas variables en la práctica social pueden ser medidas y observadas con instrumentos convencionales, en mérito de que contienen rasgos, propiedades y cualidades.
- Son susceptibles de descomposición empírica. Dicho de otro término, que las variables pueden desagregarse en indicadores, índices, subíndices e ítems.

2.9 Esperanza de una variable aleatoria

En estadística la esperanza matemática (también llamada esperanza, valor esperado, media poblacional o media) de una variable aleatoria, es el número que formaliza la idea de valor medio de un fenómeno aleatorio.

Cuando la variable aleatoria es discreta, la esperanza es igual a la suma de la probabilidad de cada posible suceso aleatorio multiplicado por el valor de dicho suceso. Por lo tanto, representa la cantidad media que se "espera" como resultado de un experimento aleatorio cuando la probabilidad de cada suceso se mantiene constante y el experimento se repite un elevado número de veces.

Cabe decir que el valor que toma la esperanza matemática en algunos casos puede no ser "esperado" en el sentido más general de la palabra (el valor de la esperanza puede ser improbable o incluso imposible).

La esperanza matemática de una variable aleatoria es una característica numérica que proporciona una idea de la localización de la variable aleatoria sobre la recta real. Decimos que es un parámetro de centralización o de localización.

Su interpretación intuitiva o significado se corresponde con el valor medio teórico de los posibles valores que pueda tomar la variable aleatoria, o también con el centro de gravedad de los valores de la variable supuesto que cada valor tuviera una masa proporcional a la función de densidad en ellos.

La definición matemática de la esperanza en el caso de las variables aleatorias discretas se corresponde directamente con las interpretaciones proporcionadas en el párrafo anterior.

En caso de que el recorrido sea infinito la esperanza existe si la serie resultante es absolutamente convergente, condición que no siempre se cumple.

La definición se corresponde con un promedio ponderado según su probabilidad de los valores del recorrido y, por tanto, se corresponde con la idea de un valor medio teórico.

2.10 Momentos de una variable aleatoria

Cuando la distribución de probabilidad de una variable aleatoria no es conocida, diversas características de ella pueden proporcionar una descripción general de la misma.

Entre las distintas características de una distribución ocupan un importante lugar los momentos, entre los que cabe destacar los diferentes tipos que definimos a continuación:

- Momentos no centrados
- Momentos centrados en media

Los momentos centrados se calculan, como los no centrados, teniendo en cuenta la definición de esperanza de una función de una variable aleatoria.

La varianza de una variable, si existe, es el valor medio de las dispersiones cuadráticas de los valores de la variable respecto de su media. Por este motivo, tanto la varianza como su raíz cuadrada, σX, que se denomina desviación típica, se usan, como se verá posteriormente, como medidas de la dispersión de la variable.

2.11 Funciones asociadas a una variable aleatoria

Una función que asocia un número real, perfectamente definido, a cada punto muestral. A veces las variables aleatorias (v.a.) están ya implícitas en los puntos muéstrales.

La función que caracteriza las variables continuas es aquella función f positiva e integrable en los reales, tal que acumulada desde -∞ hasta un punto x, nos proporciona el valor de la función de distribución en x, F(x). Recibe el nombre de función de densidad de la variable aleatoria continua.

Las funciones de densidad discreta y continua tienen, por tanto, un significado análogo, ambas son las funciones que acumuladas (en forma de sumatorio en el caso discreto o en forma de integral en el caso continuo) dan como resultado la función de distribución.

La diferencia entre ambas, sin embargo, es notable. La función de densidad discreta toma valores positivos únicamente en los puntos del recorrido y se interpreta como la probabilidad de la que la variable tome ese valor f(x) = P (X = x).

La función de densidad continua toma valores en el conjunto de números reales y no se interpreta como una probabilidad. No está acotada por 1, puede tomar cualquier valor positivo. Es más, en una variable continua se cumple que probabilidades definidas sobre puntos concretos siempre son nulas.

UNIDAD III. DISTRIBUCIONES DE PROBABILIDAD

3.1 Modelos de distribución de probabilidad

MODELOS DISCRETOS

Los modelos discretos, son modelos de probabilidad de variable aleatoria discreta. Los más importante son los modelos de BERNOUILLI (especialmente "la distribución binomial") y la "distribución de Poisson".

Distribución Binomial.

El campo de variación de la variable es {0,1,2, 3..., n} y la función de cuantía es:

$$P(x) = \binom{n}{x} p^x q^{n-x}$$ para valores de x= 0,1,2, ...n siendo n Î N, p Î [0,1] y q=1-p

Si una variable aleatoria, X, sigue una distribución binomial de parámetros n y p se expresa como: X ~ B (n, p).

Situaciones que modeliza:

- Se realiza un número n de pruebas (separadas o separables).
- Cada prueba puede dar dos únicos resultados A y Ã
- La probabilidad de obtener un resultado A es p y la de obtener un resultado Ã es q, con q= 1-p, en todas las pruebas. Esto implica que las pruebas se realizan exactamente en las mismas condiciones. Si se trata de extracciones, (muestreo), las extracciones deberán ser con devolución (reemplazamiento) (M.A.S).

Distribución de Poisson

Formalmente: dada una variable aleatoria X con campo de variación

X Î {0,1, 2, ..., ¥}, es decir X Î **N** cuya función de cuantía sea:

$P(x) = \frac{e^{\lambda} \lambda^{x}}{x!}$ siendo l un parámetro positivo

diremos que X sigue una distribución de Poisson de parámetro l, X ~ P (l).

Situaciones que modeliza:

- Se observa la ocurrencia de hechos de cierto tipo durante un período de tiempo o a lo largo de un espacio, considerados unitarios
- El tiempo (o el espacio) pueden considerarse homogéneos, respecto al tipo de hechos estudiados, al menos durante el período experimental; es decir, que no hay razones para suponer que en ciertos momentos los hechos sean más probables que otros.
- La probabilidad de que se produzca un hecho en un intervalo infinitesimal es prácticamente proporcional a la amplitud del intervalo infinitesimal.

Distribución Hipergeométrica

Dada la siguiente situación:

- Una población constituida por N individuos en total.
- De los cuales Np individuos son del tipo A, y N q individuos son del tipo Ã.

De forma que la proporción de individuos A que hay en la población es p, y la proporción de individuos de tipo Ã, es q (p + q=1).

- Se realizan n (pruebas) extracciones sin reemplazamiento

De forma que la probabilidad de extraer un individuo A (Ã) en una de las extracciones depende de los resultados de las pruebas anteriores.

- Si consideramos la variable aleatoria X = n° de resultados A obtenidos en las n extracciones, X seguirá una distribución Hipergeométrica. X~H (N, n, p)

Puede comprobarse que la función de cuantía es, entonces:

$$P(x) = \frac{\binom{Np}{x}\binom{Nq}{(n-x)}}{\binom{N}{n}}$$

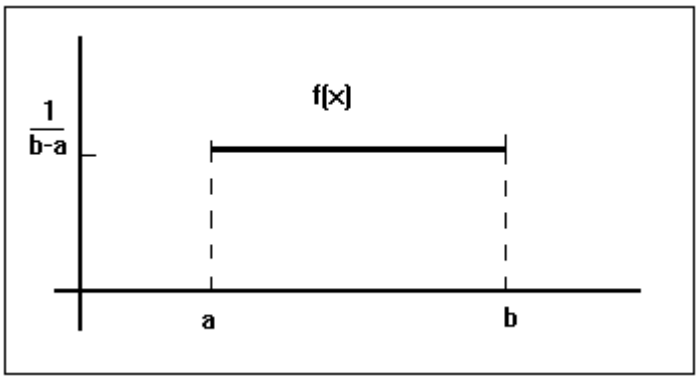
La distribución Hipergeométrica es semejante a la binomial, excepto en el hecho de que las pruebas no mantienen constantes las probabilidades de A y Ã.

MODELOS CONTINUOS

Distribución Uniforme (de V. Continua)

Dada una variable aleatoria continua, X, definida en el intervalo [a, b] de la recta real, diremos que X tiene una distribución uniforme en el intervalo [a, b] cuando su función de densidad sea: X ~ U ([a, b])

f(x)= 1/(b-a) para x Î [a, b].



De manera que la función de distribución resultará:

0 para x < a

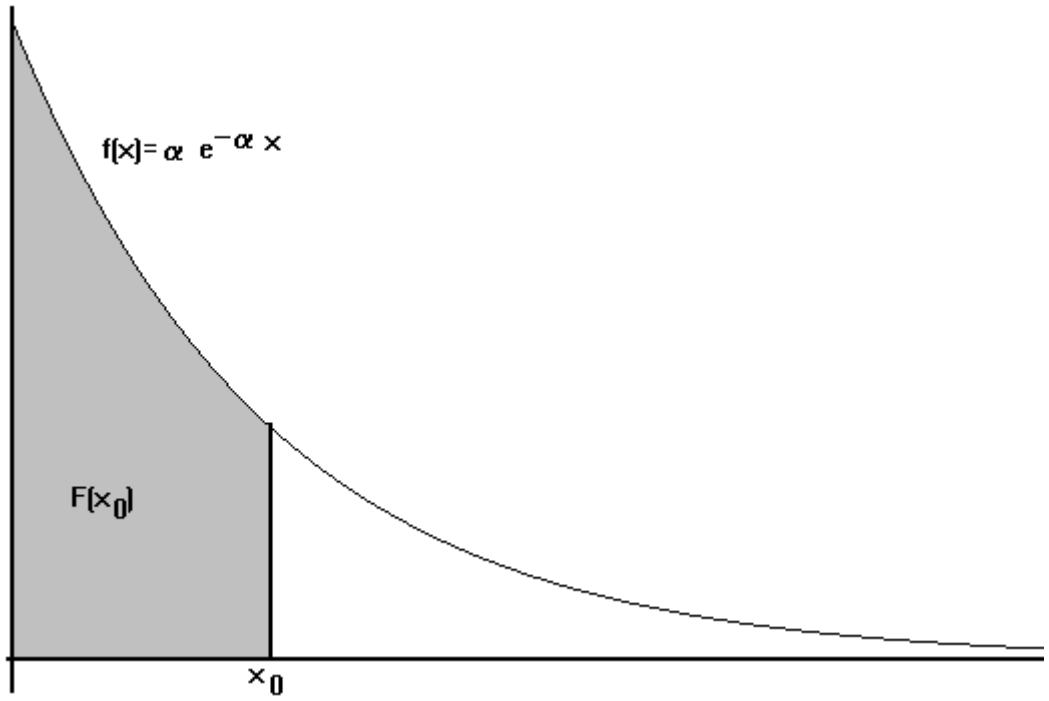
$$F(x) = \int_a^b \frac{dx}{b-a} = \frac{x-a}{b-a} \qquad \forall x \in [a,b]$$

1 para x ³ b

Distribución Exponencial

Dada una variable aleatoria continua, X, definida para valores reales positivos.

diremos que X tiene una distribución exponencial de parámetro a cuando su función de densidad sea: f(x) = a $e^{-a\,x}$ para x ³ 0 (siendo el parámetro a positivo)



La función de distribución será

$$F(x) = \begin{cases} = 0 \text{ para } x < 0 \\ = \int_0^x \alpha e^{-\alpha x} dx = \left[-e^{-\alpha x} \right]_0^x = 1 - e^{\alpha x} \quad para\ x > 0 \end{cases}$$

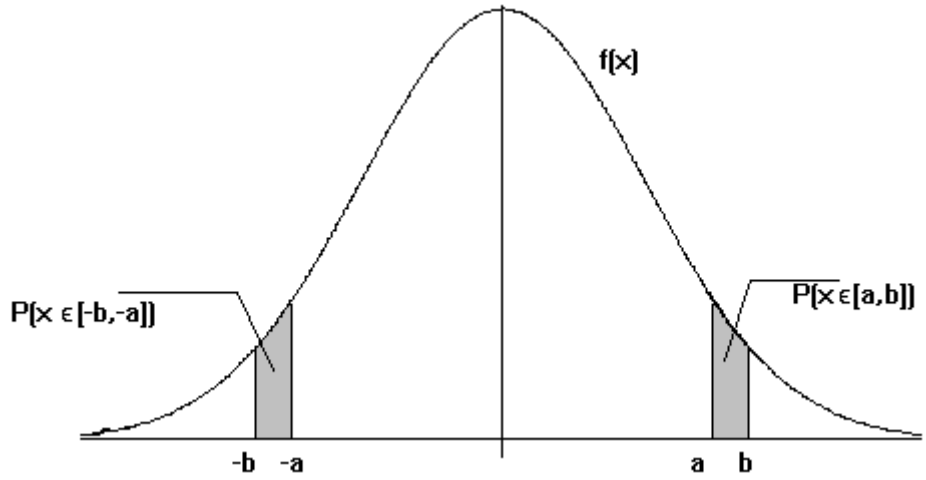
Distribución Normal

La distribución normal es la más importante de todas las distribuciones de probabilidad. Es una distribución de variable continua con campo de variación [-¥, ¥], que queda especificada a través de dos parámetros (que acaban siendo la media y la desviación típica de la distribución).

Una variable aleatoria continua, X, definida en [-¥, ¥] seguirá una distribución normal de parámetros m y s, (X ~ N (m; s)), si su función de densidad es:

$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$ para x Î [-¥, ¥]

cuya representación gráfica es:



Importancia de la distribución Normal.

a) Enorme número de fenómenos que puede modelizar: Casi todas las características cuantitativas de las poblaciones muy grades tienden a aproximar su distribución a una distribución normal.

b) Muchas de las demás distribuciones de uso frecuente, tienden a distribuirse según una Normal, bajo ciertas condiciones.

c) (En virtud del teorema central del límite). Todas aquellas variables que pueden considerarse causadas por un gran número de pequeños efectos (como pueden ser los errores de medida) tienden a distribuirse según una distribución normal.

3.2 Distribuciones Binomial y Poisson.

Distribución Binomial

Una distribución binomial es una distribución de probabilidad discreta que describe el número de éxitos al realizar n experimentos independientes entre sí, acerca de una variable aleatoria**.**

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

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

Propiedades de la distribución binomial

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

- En cada ensayo, experimento o prueba solo son posibles dos resultados (éxito o fracaso).
- La probabilidad del éxito ha de ser constante. Esta se representa mediante la letra p. La probabilidad de que salga cara al lanzar una moneda es 0,5 y esta es constante dado que la moneda no cambia en cada experimento y las probabilidades de sacar cara son constantes.
- La probabilidad de fracaso ha de ser también constate. Esta se representa mediante la letra q = 1 - P. Es importante fijarse que, mediante esa ecuación, sabiendo p o sabiendo q, podemos obtener la que nos falte.
- El resultado obtenido en cada experimento es independiente del anterior. Por lo tanto, lo que ocurra en cada experimento no afecta a los siguientes.
- Los sucesos son mutuamente excluyentes, es decir, no pueden ocurrir los 2 al mismo tiempo. No se puede ser hombre y mujer al mismo tiempo o que al lanzar una moneda salga cara y cruz al mismo tiempo.
- Los sucesos son colectivamente exhaustivos, es decir, al menos uno de los 2 ha de ocurrir. Si no se es hombre, se es mujer y, si se lanza una moneda, si no sale cara ha de salir cruz.

- La variable aleatoria que sigue una distribución binomial se suele representar como X~ (n, p), donde n representa el número de ensayos o experimentos y P la probabilidad de éxito.

Formula de la distribución binomial

La fórmula para calcular la distribución normal es:

$$P_{(x)} = \binom{n}{x} p^x q^{n-x}$$

Donde:

n = Número de ensayos/experimentos
x = Número de éxitos
p = Probabilidad de éxito
q = Probabilidad de fracaso (1-p)

Es importante resaltar que la expresión entre corchetes no es una expresión matricial, sino que es un resultado de una combinatoria sin repetición. Este se obtiene con la siguiente formula:

$$C_{n,x} = \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

El signo de exclamación en la expresión anterior representa el símbolo de factorial.

Ejemplo:

El número de llamadas telefónicas que llegan a una central telefónica es modelado frecuentemente como una variable aleatoria de Poisson. Asumiendo que en promedio hay 10 llamadas por hora. ¿Cuál es la probabilidad de que haya exactamente 5 llamadas en una hora? ¿Cuál es la probabilidad de que haya exactamente 5 llamadas en 30 minutos?

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

$$\lambda = 10$$

$$f(5) = \frac{10^5}{5!} e^{-10}$$

$$f(5) = 0.0378$$

Distribución de Poisson

La Distribución de Poisson se llama así en honor a Simeón Dennis Poisson (1781-1840), francés que desarrolló esta distribución basándose en estudios efectuados en la última parte de su vida.

La distribución de Poisson es una distribución de probabilidad discreta que se aplica a las ocurrencias de algún suceso durante un intervalo determinado. Nuestra variable aleatoria x representará el número de ocurrencias de un suceso en un intervalo determinado, el cual podrá ser tiempo, distancia, área, volumen o alguna otra unidad similar o derivada de éstas.

La probabilidad de nuestra variable aleatoria X viene dada por la siguiente expresión:

$$P(x) = \frac{\mu^x . e^{-\mu}}{x!}$$

donde:

- Nuestra variable aleatoria discreta puede tomar los valores: $x = 0, 1, 2, 3 \ldots$
- μ donde es la media del número de sucesos en el intervalo que estemos tomando, ya sea de tiempo, distancia, volumen, etc.
- "e" es una constante 2. 71..

La distribución de Poisson debe de cumplir los siguientes requisitos:

- La variable discreta x es el número de ocurrencias de un suceso durante un intervalo (esto es la propia definición que hemos dado anteriormente).
- Las ocurrencias deben ser aleatorias y no contener ningún vicio que favorezca unas ocurrencias en favor de otras.
- Las ocurrencias deben estar uniformemente distribuidas dentro del intervalo que se emplee.

La distribución de Poisson es particularmente importante ya que tiene muchos casos de uso. Podemos poner como ejemplos de uso: la disminución de una muestra radioactiva, la llegada de pasajeros de un aeropuerto o estación de trenes o autobuses, los usuarios que se conectan a una web determinada por hora.

Ejemplo:

Si un banco recibe en promedio 6 cheques sin fondo por día, ¿cuáles son las probabilidades de que reciba cuatro cheques sin fondo en un día dado?

Solución:

x = variable que nos define el número de cheques sin fondo que llegan al banco en un día cualquiera = 0, 1, 2, 3,…., etc., etc.

λ **=** 6 cheques sin fondo por día

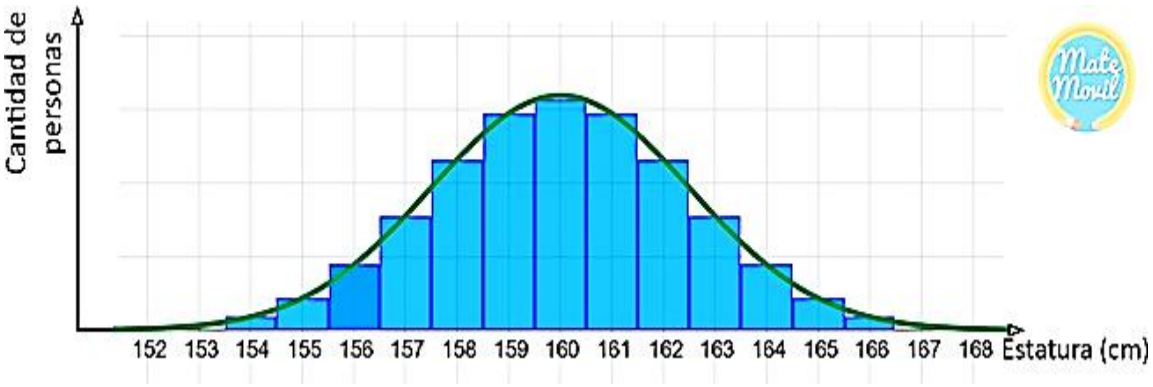
e = 2.718

$$p(x = 4, \lambda = 6) = \frac{(6)^4(2.718)^{-6}}{4!} = \frac{(1296)(0.00248)}{24} = 0.13392$$

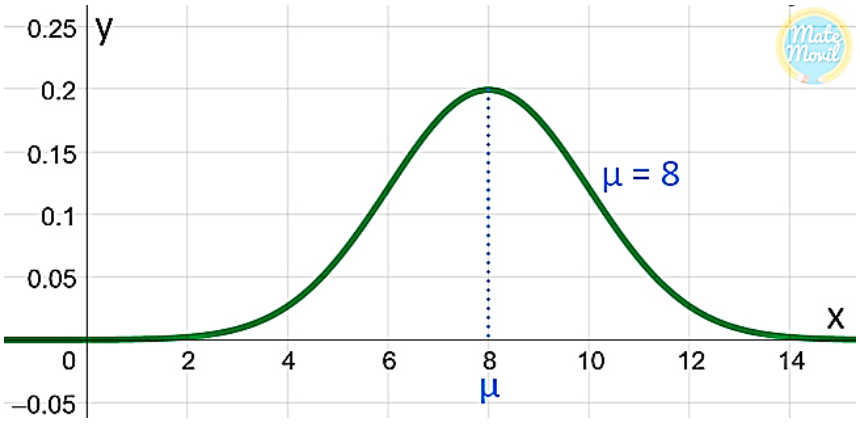
3.3 Distribución normal.

La distribución normal, distribución de Gauss o distribución gaussiana, es la distribución de probabilidad individual más importante. La distribución normal nos permite crear modelos de muchísimas variables y fenómenos, como, por ejemplo, la estatura de los habitantes de un país, la temperatura ambiental de una ciudad, los errores de medición y muchos otros fenómenos naturales, sociales y hasta psicológicos. Por ello, hoy vamos a revisar sus características y muchísimos problemas resueltos en 3 niveles de dificultad.

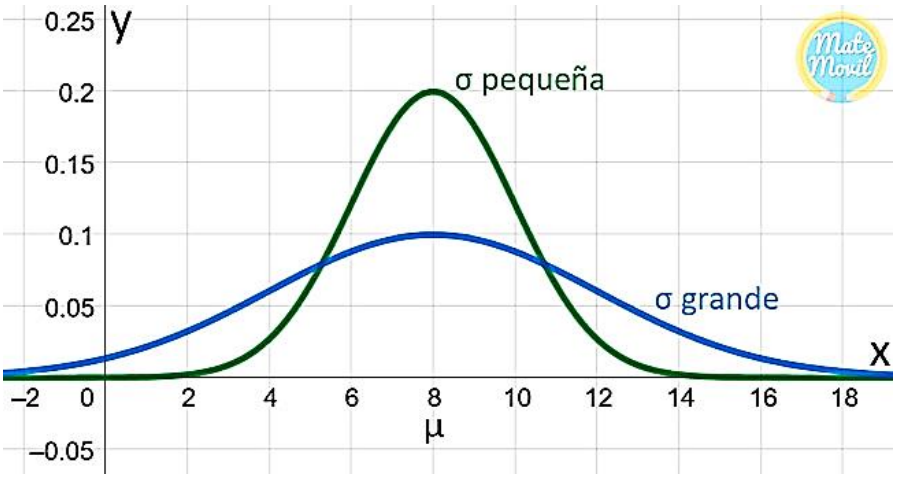
¿Qué pasaría si se realiza una encuesta en una ciudad a personas adultas consultando su estatura? A partir de los resultados obtenidos, se puede elaborar un histograma que tendría la siguiente forma:



Como vemos, el histograma tiene forma de campana, una característica importante de la distribución normal. Un parámetro muy importante es la media (μ) y siempre estará al centro de la curva con forma de campana. Por ejemplo, aquí tenemos la gráfica de una distribución normal con media igual a 8.



Además de la media, existe otro parámetro muy importante, se trata de la desviación estándar, representada con la letra griega **σ.** La desviación estándar es la medida de variabilidad más utilizada y nos indica que tan dispersos se encuentran los datos. Por ejemplo, aquí veremos dos curvas normales, una con desviación estándar pequeña, y otra con desviación estándar grande. Cuando la desviación estándar es pequeña, los datos tienen una dispersión baja y se agrupan alrededor de la media. En cambio, cuando la desviación estándar es alta, los datos tienen una dispersión alta y se alejan de la media.



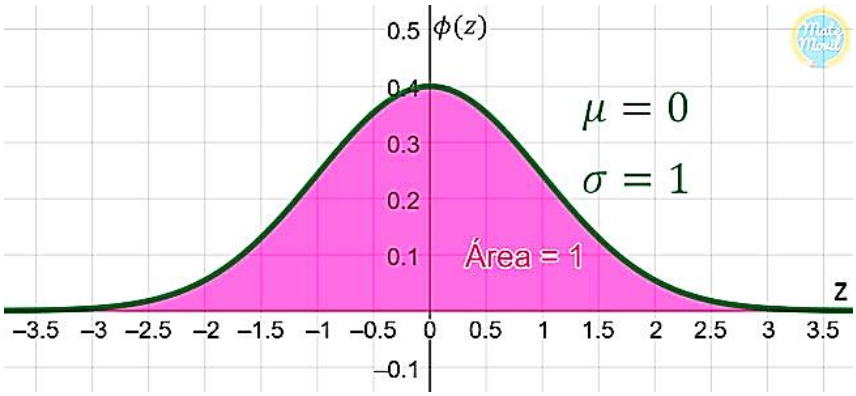
Características de la distribución normal

- Toma en cuenta la media(μ) y la desviación estándar(σ).
- El área bajo la curva es igual a 1.
- Es simétrica respecto al centro, o a la media.
- 50% de los valores son mayores que la media, y 50% de los valores son menores que la media.
- La media es igual a la mediana y a la moda.
- Tiene una asíntota en y = 0 (eje x).

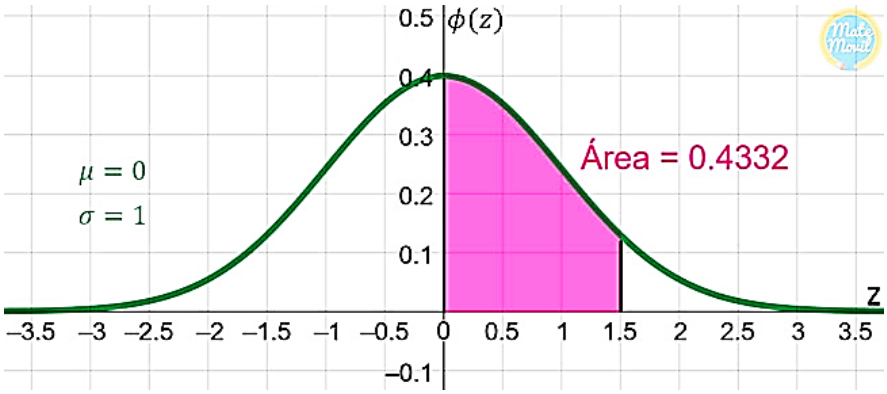
Para encontrar las probabilidades o cantidad de datos entre determinados valores de la variable, se calcula el área bajo la curva normal, que se encuentra en la tabla z o tabla de áreas bajo la curva normal estandarizada.

La distribución normal estándar

La distribución normal estándar, es aquella distribución normal que tiene una media igual a cero, y una desviación estándar igual a uno. Veamos la función densidad normal estandarizada, que trabaja con la variable estandarizada z en el eje horizontal:



Por ejemplo, si se desea encontrar la probabilidad de que la variable estandarizada z, tome un valor entre 0 y 1,50; hay que encontrar el área bajo la curva entre z = 0 y z = 1,50.



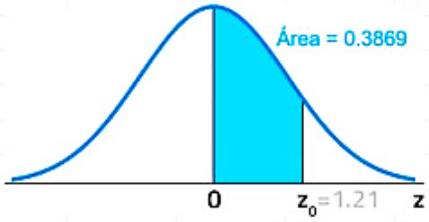
Para calcular el valor de esta área, se utiliza la tabla z y se busca el valor de 1,50:

Tabla de Distribución Normal Estandarizada

$$z = \frac{x - \mu}{\sigma}$$

$\mu = media$

$\sigma = desviación\ estándar$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

Como vemos, el valor del área bajo la curva es de 0,4332, y esa sería la probabilidad de que la variable estandarizada z, tome un valor comprendido entre 0 y 1,50.

¿Y si mi distribución normal no es estandarizada?

En la mayoría de problemas, cuando se analizan diferentes variables x, la distribución normal no tiene la forma estandarizada, es decir, la media no es cero y la desviación estándar no es uno. En esos casos, se convierten los valores de la variable (x) a z, es decir, se estandarizan los valores de la variable (x).

La fórmula de la variable estandarizada «z», la cual indica cuántas desviaciones estándar se aleja el valor x de la media, es la siguiente:

$$z = \frac{x - \mu}{\sigma}$$

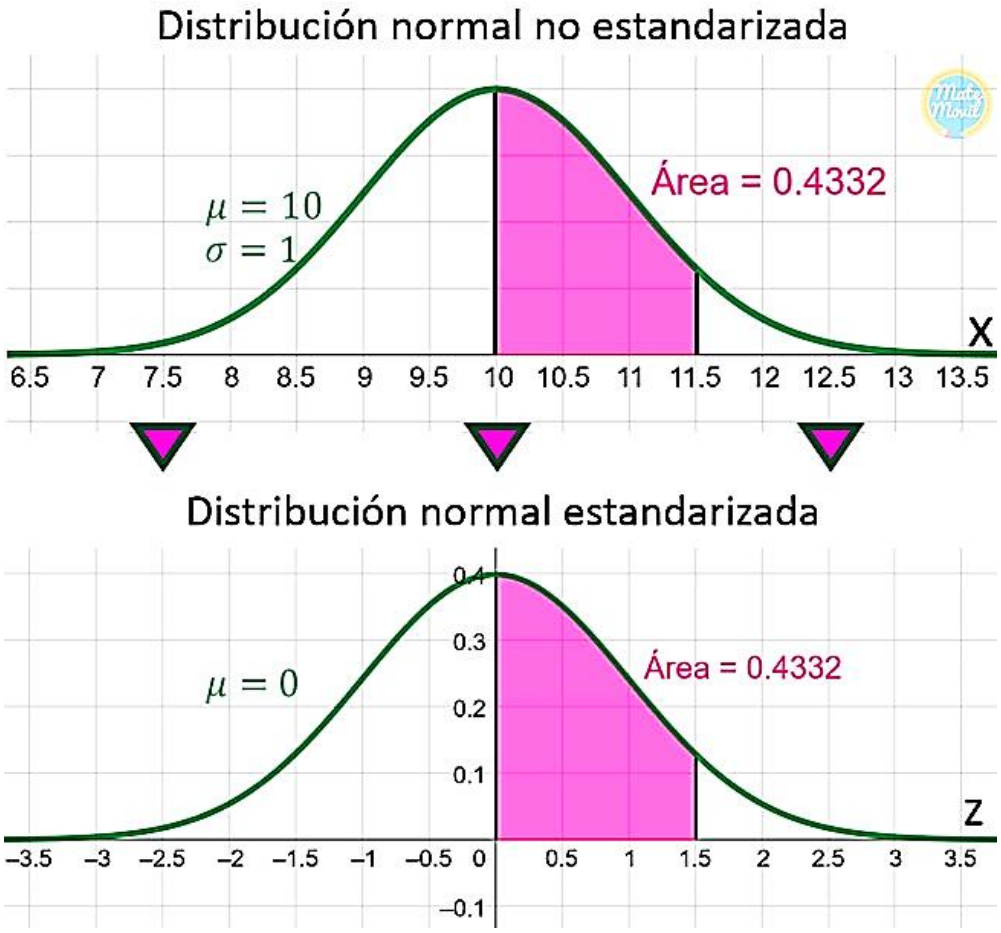
Y luego, con los valores de z, se utiliza la tabla y se calculan las áreas bajo la curva, porcentajes o probabilidades. En los videos que vienen líneas abajo, encontrarás muchos ejemplos.

Por ejemplo, si tenemos una variable aleatoria continua X con una distribución normal no estandarizada, con media igual a 10 y desviación estándar igual a 1, y el problema pide

calcular la probabilidad de que la variable X tome un valor entre 10 y 11,50, hay que estandarizar los valores de la variable X aplicando la fórmula de z:

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{11{,}50 - 10}{1} = \frac{1{,}50}{1} = 1{,}50$$



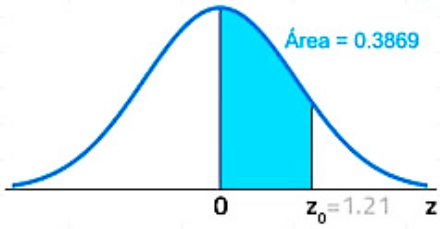
Y usando la tabla z, se calcula el área bajo la curva. Cuando z es igual a 1,50, el área bajo la curva es de 0,4332.

Tabla de Distribución Normal Estandarizada

$$z = \frac{x - \mu}{\sigma}$$

$\mu = media$

$\sigma = desviación\ estándar$

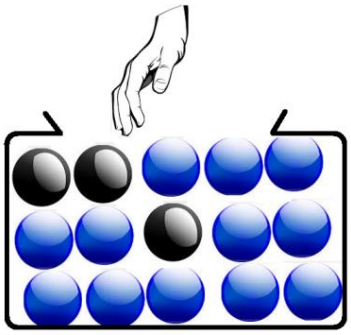


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

Podemos concluir que la probabilidad de que la variable estandarizada z, tome un valor comprendido entre 10 y 11.50 es de 0.4332.

3.4 Otras distribuciones discretas y continúas

Distribución Hipergeométrica



La distribución Hipergeométrica es especialmente útil en todos aquellos casos en los que se extraigan muestras o se realicen experiencias repetidas sin devolución del elemento extraído o sin retornar a la situación experimental inicial.

Es una distribución fundamental en el estudio de muestras pequeñas de poblaciones pequeñas y en el cálculo de probabilidades de juegos de azar. Tiene grandes aplicaciones en el control de calidad para procesos experimentales en los que no es posible retornar a

la situación de partida. Las consideraciones a tener en cuenta en una distribución Hipergeométrica:

- El proceso consta de "n" pruebas, separadas o separables de entre un conjunto de "N" pruebas posibles.
- Cada una de las pruebas puede dar únicamente dos resultados mutuamente excluyentes.
- El número de individuos que presentan la característica A (éxito) es "k".
- En la primera prueba las probabilidades son: P(A)= p y P(A)= q; con p+q=1.

En estas condiciones, se define la variable aleatoria X = "No. de éxitos obtenidos". La función de probabilidad de esta variable sería:

$$p(X=x)=\frac{\binom{k}{x}\cdot\binom{N-k}{n-x}}{\binom{N}{n}} \qquad \begin{array}{l} N=tamaño\ \ de\ \ población \\ K=n^{o}\ individuos\ que... \\ n=tamaño\ de\ la\ \ muestra \\ x=valor\ que\ toma\ \ la\ \ \text{var}iable \end{array}$$

Ejemplo:

De una urna que contiene seis bolas negras y nueve bolas rojas se extraen cinco bolas, ¿cuál es la probabilidad de obtener tres bolas rojas?

Sea X = "Número de bolas rojas al extraer cinco bolas de la urna". La probabilidad pedida es:

$$P(X=3)=\frac{\binom{6}{2}\binom{9}{3}}{\binom{15}{5}}=0'4196.$$

La distribución Gamma

Este modelo es una generalización del modelo Exponencial ya que, en ocasiones, se utiliza para modelar variables que describen el tiempo hasta que se produce p veces un determinado suceso.

Propiedades de la distribución Gamma

- Su esperanza es pα.
- Su varianza es pα2.
- La distribución Gamma (α, p = 1) es una distribución Exponencial de parámetro α. Es decir, el modelo Exponencial es un caso particular de la Gamma con p = 1.
- Dadas dos variables aleatorias con distribución Gamma y parámetro α común. Una consecuencia inmediata de esta propiedad es que, si tenemos k variables aleatorias con distribución Exponencial de parámetro α (común) e independientes, la suma de todas ellas seguirá una distribución G (α, k).

La distribución de Cauchy

Se trata de un modelo continuo Cuya integral nos proporciona la función de distribución.

Propiedades de la distribución de Cauchy

Se trata de un ejemplo de variable aleatoria que carece de esperanza (y, por tanto, también de varianza o cualquier otro momento).

Por tanto, la esperanza de una distribución de Cauchy no existe. Cabe señalar que la función de densidad es simétrica respecto al valor cero (que sería la mediana y la moda), pero al no existir la integral anterior, la esperanza no existe.

3.5 Muestreo aleatorio simple.

El muestreo aleatorio simple es un subconjunto de una muestra elegida de una población más grande. Cada individuo se elige al azar y por pura casualidad. En este tipo de muestreo cada individuo tiene la misma probabilidad de ser elegido en cualquier etapa del proceso.

Cabe mencionar que es diferente al muestreo sistemático. Una muestra aleatoria simple es una técnica de muestreo justa.

El muestreo aleatorio simple es un tipo de método de muestreo muy básico y puede ser fácilmente un componente de un método de muestreo más complejo. El principal atributo de este método de muestreo es que cada muestra tiene la misma posibilidad de ser elegida.

¿Quieres saber cómo realizar un muestreo aleatorio simple?

1.- Prepara una lista de todos los miembros de la población, posterior a esto marca a cada miembro con un número específico.

2.- De esta población, las muestras aleatorias se eligen de dos maneras: tablas de números aleatorios y con un software de generador de números aleatorios. Se recomienda (y comúnmente es lo que prefieren los investigadores) un software generador de números aleatorios, ya que los números de muestra se generan sin interferencia humana.

¿Quieres saber cómo realizar un muestreo aleatorio sin sesgo? Hay dos enfoques o métodos que se encargan de minimizar este punto:

Método de lotería:

El método de la lotería es uno de los métodos más antiguos y es definitivamente un ejemplo claro del mecanismo del muestreo aleatorio simple. En este método, cada miembro de la población debe estar numerado de manera sistemática y posterior a esto se escribe cada número en una hoja de papel por separado. Esos pedazos de papel se mezclan y se ponen en una caja y de esta de forma los números se extraen de manera aleatoria.

Uso de números aleatorios

El método de uso de números aleatorios es un método alternativo que también implica la numeración de la población. En este método, se utiliza una tabla similar a la de la siguiente imagen para aplicar la técnica:

20	17	42	01	72	33	94	55	89	65	58	60
74	49	04	27	56	49	11	63	77	79	90	31
94	70	49	49	05	74	64	00	26	07	23	00
22	15	78	49	74	26	50	94	13	90	08	14
93	29	12	20	26	87	66	98	37	53	82	62
45	04	77	48	87	72	66	91	42	98	17	26
44	91	99	08	72	97	33	58	12	08	91	12
16	23	91	95	97	87	52	49	40	37	21	46
04	50	65	37	99	98	74	98	93	99	78	30
32	70	17	05	79	63	50	26	54	30	01	88
03	64	59	55	85	96	49	46	61	89	33	79
62	49	00	67	28	94	19	65	13	44	78	39
61	00	95	85	86	60	64	17	47	67	87	59
89	03	90	40	10	05	18	43	97	37	68	97

Ejemplo de muestreo aleatorio simple

Una organización tiene 500 empleados. Y lo que queremos obtener es una lista de 100 de estos individuos.

Paso 1: Haz una lista de todos los empleados que trabajan para la organización (Como se mencionó anteriormente, hay 500 empleados en la organización, es por eso que la lista debe contener 500 nombres).

Paso 2: Asignar un número secuencial a cada uno de los 500 empleados (1, 2, 3 … n). Este es tu marco de muestreo (la lista de la cual extraerán tu muestra aleatoria simple).

Paso 3: Define cuál será el tamaño de tu muestra (en este caso es 100).

Paso 4: Utiliza un generador de números aleatorios para seleccionar la muestra, utilizando tu marco de muestreo (tamaño de la población) del paso dos y el tamaño de tu muestra, del paso 3. Por ejemplo, si tu tamaño de muestra es 100 y tu población es de 500, lo que debes hacer es generar 100 números entre 1 y 500.

Ventajas del muestreo aleatorio simple

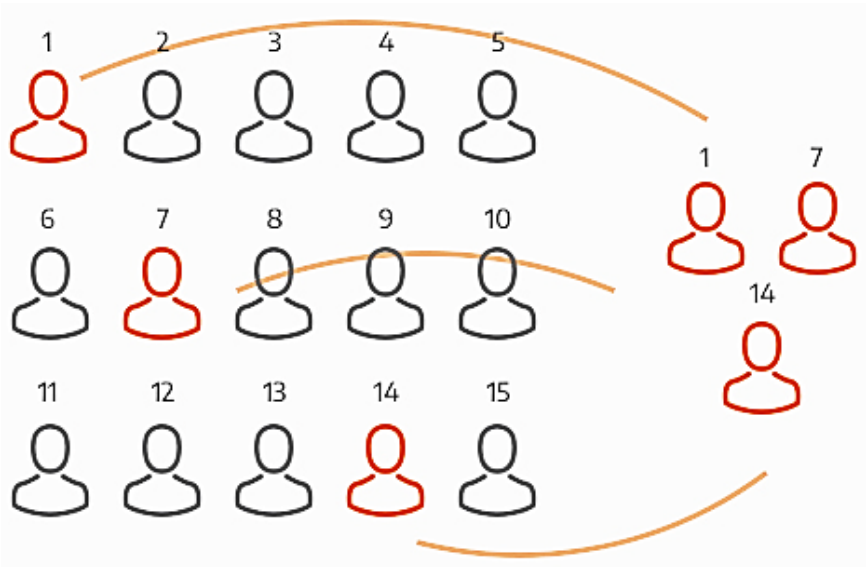
1.- Este es un método de muestreo justo, y si se aplica adecuadamente, ayuda a reducir cualquier sesgo involucrado con el muestreo en comparación con cualquier otro de los métodos de muestreo.

2.- Debido a que involucra un marco de muestra grande, generalmente es fácil seleccionar un tamaño de muestra más pequeño de una población más grande.

3.- La persona que realiza la investigación no necesita tener conocimiento previo de los datos que se recopilaran. Uno puede simplemente hacer una pregunta para reunir la información necesaria. La realidad es que en este caso no se necesita ser un experto en el tema.

4.- Este método de muestreo es un método básico de recopilación de datos. No se requieren conocimientos técnicos.

5.- Dado que el tamaño de la población es grande en este tipo de método de muestreo, no existe ninguna restricción en el tamaño de la muestra. De una población más grande se puede obtener con facilidad una muestra más pequeña.



3.6 Justificación del muestreo.

En vez de tomar un censo completo, los procedimientos de muestreo estadístico se han convertido en la herramienta preferida en la mayoría de las situaciones de investigación. Existen tres razones principales para extraer una muestra.

Antes que todo, por lo general, lleva demasiado tiempo realizar un censo completo. En segundo lugar, es demasiado costoso hacer un censo completo. Tercero, es demasiado molesto e ineficiente obtener un conteo completo de la población objeto.

Después que se han determinado las preguntas numéricas y categóricas más esenciales en la encuesta, el tamaño de muestra necesario se basará en la satisfacción de la pregunta con los requerimientos más rigurosos.

3.7 Función de Distribución empírica.

La función de distribución empírica es la función de distribución de la distribución empírica.

La función de distribución empírica es la función de $\mathbb{R}$ en $[0,1]$, que denotamos por $\widehat{F}$ y que toma los valores:

$$\widehat{F}(x) = \begin{cases} 0 & \text{para} \quad x < x_{(1)} \\ \vdots & \\ \frac{i}{n} & \text{para} \quad x_{(i)} \leq x < x_{(i+1)} \\ \vdots & \\ 1 & \text{para} \quad x \geq x_{(n)} \ . \end{cases}$$

En otras palabras, $\widehat{F}(x)$ es la proporción de los elementos de la muestra que son menores o iguales a x.

3.8 Estadísticos muéstrales. Distribuciones.

En estadística un estadístico (muestral) es una medida cuantitativa, derivada de un conjunto de datos de una muestra, con el objetivo de estimar o inferir características de una población o modelo estadístico.

Más formalmente un estadístico es una función medible T que, dada una muestra estadística de valores, les asigna un número, que sirve para estimar determinado parámetro de la distribución de la que procede la muestra. Así, por ejemplo, la media de los valores de una muestra (media muestral) sirve para estimar la media de la población de la que se ha extraído la misma; la varianza muestral podría usarse para estimar la varianza poblacional, etc. Esto se denomina como realizar una estimación puntual.

A partir de las muestras seleccionadas de una población pueden construirse variables aleatorias alternativas, de cuyo análisis se desprenden interesantes propiedades estadísticas. Las dos formas más comunes de estas variables corresponden a las distribuciones muéstrales de las medias y de las proporciones.

Dada una población constituida por un número n de elementos, cuya media aritmética es m y donde la desviación típica viene dada s, pueden formarse n2 muestras con reemplazamiento distintas, formadas por dos elementos de la población.

El estudio de determinadas características de una población se efectúa a través de diversas muestras que pueden extraerse de ella.

El muestreo puede hacerse con o sin reposición, y la población de partida puede ser infinita o finita. Una población finita en la que se efectúa muestreo con reposición puede considerarse infinita teóricamente. También, a efectos prácticos, una población muy grande puede considerarse como infinita. En todo nuestro estudio vamos a limitarnos a una población de partida infinita o a muestreo con reposición.

Consideremos todas las posibles muestras de tamaño n en una población. Para cada muestra podemos calcular un estadístico (media, desviación típica, proporción...) que variará de una a otra. Así obtenemos una distribución del estadístico que se llama distribución muestral.

Las dos medidas fundamentales de esta distribución son la media y la desviación típica, también denominada error típico.

Hay que hacer notar que si el tamaño de la muestra es lo suficientemente grande las distribuciones muéstrales son normales y en esto se basarán todos los resultados que alcancemos.

3.9 Estimación.

Estimar qué va a ocurrir respecto a algo (o qué está ocurriendo, o qué ocurrió), a pesar de ser un elemento muy claramente estadístico, está muy enraizado en nuestra cotidianidad. Dentro de ello, además hacemos estimaciones dentro de un intervalo de posibilidades. Por ejemplo: "creo que terminaré la tarea en unos 5-6 días". Lo que hacemos en el terreno del análisis de datos es aplicar matizaciones técnicas a este hábito. Vamos a dedicar este documento al concepto de estimación, comenzando con la estimación puntual. Después nos ocuparemos de desarrollar un modelo de estimación por intervalo donde identificaremos los elementos fundamentales, con su significado y símbolo. Y, por último, habrá que desarrollar cómo se calculan esos elementos.

La estimación puntual

Estimar puede tener dos significados interesantes. Significa querer e inferir. Desde luego, el primer significado es más trascendente. Pero no tiene ningún peso en la estadística, disciplina que no se ocupa de los asuntos del amor. El segundo significado es el importante aquí. Una estimación estadística es un proceso mediante el que establecemos qué valor debe tener un parámetro según deducciones que realizamos a partir de estadísticos. En otras palabras, estimar es establecer conclusiones sobre características poblacionales a partir de resultados muestrales.

Vamos a ver dos tipos de estimaciones: puntual y por intervalo. La segunda es la más natural. Y verás que forma parte habitual de nuestro imaginario como personas sin necesidad de una formación estadística. La primera, la estimación puntual, es la más sencilla y, por ese motivo, vamos a comenzar por ella. Ocurre, además, que la estimación

por intervalo surge, poco más o menos, de construir un intervalo de posibles valores alrededor de la estimación puntual.

Una estimación puntual consiste en establecer un valor concreto (es decir, un punto) para el parámetro. El valor que escogemos para decir "el parámetro que nos preocupa vale X" es el que suministra un estadístico concreto. Como ese estadístico sirve para hacer esa estimación, en lugar de estadístico suele llamársele estimador. Así, por ejemplo, utilizamos el estadístico "media aritmética de la muestra" como estimador del parámetro "media aritmética de la población". Esto significa: si quieres conocer cuál es el valor de la media en la población, estimaremos que es exactamente el mismo que en la muestra.

Ejemplos de estimaciones puntuales

Para obtener una estimación puntual se usa un estadístico que recibe el nombre de estimador o función de decisión. Algunos ejemplos de estadísticos son:

- La media muestral que sirve como estimación puntual de la media poblacional.

$$\bar{X} = \mu$$

- La desviación típica muestral que sirve de estimación para la desviación típica de la población.

$$S = 6$$

3.10 Propiedades de los estimadores

Las propiedades deseables de un estimador son las siguientes:

Sesgo: Se denomina sesgo de un estimador a la diferencia entre la esperanza (o valor esperado) del estimador y el verdadero valor del parámetro a estimar. Es deseable que un estimador sea insesgado o centrado, es decir, que su sesgo sea nulo por ser su esperanza igual al parámetro que se desea estimar.

Por ejemplo, si se desea estimar la media de una población, la media aritmética de la muestra es un estimador insesgado de la misma, ya que su esperanza (valor esperado) es igual a la media de la población.

Eficiencia: Un estimador es más eficiente o preciso que otro, si la varianza del primero es menor que la del segundo.

Convergencia: Para estudiar las características de un estimador no solo basta con saber el sesgo y la varianza, sino que además es útil hacer un análisis de su comportamiento y estabilidad en el largo plazo, esto es, su comportamiento asintótico. Cuando hablamos de estabilidad en largo plazo, se viene a la mente el concepto de convergencia. Luego, podemos construir sucesiones de estimadores y estudiar el fenómeno de la convergencia.

Consistencia: También llamada robustez, se utilizan cuando no es posible emplear estimadores de mínima varianza, el requisito mínimo deseable para un estimador es que a medida que el tamaño de la muestra crece, el valor del estimador tiende a ser el valor del parámetro, propiedad que se denomina consistencia.

3.11 Obtención de estimadores.

Método por Analogía. Consiste en aplicar la misma expresión formal del parámetro poblacional a la muestra , generalmente , estos estimadores son de cómoda operatividad , pero en ocasiones presentan sesgos y no resultan eficientes . Son recomendables , para muestras de tamaño grande al cumplir por ello propiedades asintóticas de consistencia.

Método de los momentos. Consiste en tomar como estimadores de los momentos de la población a los momentos de la muestra . Podríamos decir que es un caso particular del método de analogía. En términos operativos consiste en resolver el sistema de equivalencias entre unos adecuados momentos empíricos (muestrales) y teóricos (poblacionales).

Ejemplo:

conocemos que la media poblacional de una determinada variable x depende de un parámetro K que es el que realmente queremos conocer (estimar). Así

$\mu = 2K + 7$ por el método de los momentos tendríamos que

$\hat{\mu} = \bar{x} \xleftrightarrow{estimador} \hat{\mu}$ de donde $\hat{k} = (\bar{x} - 7) / 2$

Estimadores máximo - verosímiles. La verosimilitud consiste en otorgar a un estimador/estimación una determinada "credibilidad" una mayor apariencia de ser el cierto valor(estimación) o el cierto camino para conseguirlo(estimador).

En términos probabilísticos podríamos hablar de que la verosimilitud es la probabilidad de que ocurra o se dé una determinada muestra si es cierta la estimación que hemos efectuado o el estimador que hemos planteado.

Evidentemente, la máxima verosimilitud, será aquel estimador o estimación que nos arroja mayor credibilidad. En situación formal tendríamos:

Un estimador máximo-verosímil es el que se obtiene maximizando la función de verosimilitud (likelihood) de la muestra

$$L(x_1, x_2, \ldots\ldots x_n / \hat{\vartheta})$$

3.12 Estimación por intervalos de confianza

La estimación por intervalos consiste en establecer el intervalo de valores donde es más probable se encuentre el parámetro. La obtención del intervalo se basa en las siguientes consideraciones:

a) Si conocemos la distribución muestral del estimador podemos obtener las probabilidades de ocurrencia de los estadísticos muestrales.

b) Si conociéramos el valor del parámetro poblacional, podríamos establecer la probabilidad de que el estimador se halle dentro de los intervalos de la distribución muestral.

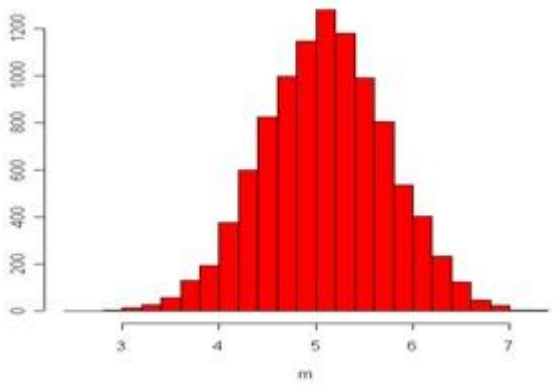
c) El problema es que el parámetro poblacional es desconocido, y por ello el intervalo se establece alrededor del estimador. Si repetimos el muestreo un gran número de veces y definimos un intervalo alrededor de cada valor del estadístico muestral, el parámetro se sitúa dentro de cada intervalo en un porcentaje conocido de ocasiones. Este intervalo es denominado "intervalo de confianza".

Ejemplo

Se generan 100000 muestras aleatorias (n=25) de una población que sigue la distribución Normal, y resulta:

	Población	Distribución muestral
Media	5.1	5.1
Desviación Típica	3.2	0.6

La distribución de las Medias muestrales aproxima al modelo Normal:



En consecuencia, el intervalo dentro del cual se halla el 95% de las Medias muestrales es

$$\mu_{\bar{X}} \pm 1.96\sigma_{\bar{X}} = 5.1 \pm (1.96)(0.6) = \begin{cases} 6.3 \\ 3.9 \end{cases}$$

(Nota: Los valores +-1.96 que multiplican la Desviación Típica de la distribución muestral son los valores cuya función de distribución es igual a 0.975 y 0.025 respectivamente y se pueden obtener en las tablas de la distribución Normal estandarizada o de funciones en aplicaciones informáticas como Excel). Seguidamente generamos una muestra de la población y obtenemos su Media, que es igual a 4.5. Si establecemos el intervalo alrededor de la Media muestral, el parámetro poblacional (5.1) está incluido dentro de sus límites:

$$\bar{X} \pm 1.96\sigma_{\bar{X}} = 4.5 \pm (1.96)(0.6) = \begin{cases} 5.7 \\ 3.3 \end{cases}$$

Ahora bien, la distancia de un punto A a un punto B es la misma que de B a A. Por esa razón, la distancia desde m a la Media muestral es la misma que va de la Media muestral a m. En consecuencia, si hacemos un muestreo con un número grande de muestras observamos que el 95% de las veces (aproximadamente) el valor de la Media de la población (m) se encuentra dentro del intervalo definido alrededor de cada uno de los valores de la Media muestral. El porcentaje de veces que el valor de m se halla dentro de alguno de los intervalos de confianza es del 95%, y es denominado nivel de confianza.
Si queremos establecer un intervalo de confianza en que él % de veces que m se halle dentro del intervalo sea igual al 99%, la expresión anterior es:

$$\bar{X} \pm 2.58\sigma_{\bar{X}}$$

(Obtenemos el valor +-2.58 que multiplica la Desviación Típica de la distribución muestral en las tablas de la distribución Normal estandarizada o de funciones en aplicaciones informáticas como Excel), y son los valores cuya función de probabilidad es igual a 0.995 y 0.005 respectivamente).

3.13 Contraste de hipótesis.

Una hipótesis estadística es una asunción relativa a una o varias poblaciones, que puede ser cierta o no. Las hipótesis estadísticas se pueden contrastar con la información

extraída de las muestras y tanto si se aceptan como si se rechazan se puede cometer un error.

La hipótesis formulada con intención de rechazarla se llama hipótesis nula y se representa por H_0. Rechazar H_0 implica aceptar una hipótesis alternativa (H_1).

La situación se puede esquematizar:

	H_0 cierta	H_0 falsa H_1 cierta
H_0 rechazada	Error tipo I (α)	Decisión correcta (*)
H_0 no rechazada	Decisión correcta	Error tipo II (β)

(*) Decisión correcta que se busca

a = p (rechazar H_0|H_0 cierta)

b = p (aceptar H_0|H_0 falsa)

Potencia =1-b = p (rechazar H_0|H_0 falsa)

Detalles a tener en cuenta

1. a y b están inversamente relacionadas.
2. Sólo pueden disminuirse las dos, aumentando n.

Los pasos necesarios para realizar un contraste relativo a un parámetro q son:

1. Establecer la hipótesis nula en términos de igualdad

$$H_0 : \theta = \theta_0$$

2. Establecer la hipótesis alternativa, que puede hacerse de tres maneras, dependiendo del interés del investigador

$$H_1 : \theta \neq \theta_0 \qquad \theta > \theta_0 \qquad \theta < \theta_0$$

en el primer caso se habla de contraste bilateral o de dos colas, y en los otros dos de lateral (derecho en el 2° caso, o izquierdo en el 3°) o una cola.

3. Elegir un nivel de significación: nivel crítico para a

4. Elegir un estadístico de contraste: estadístico cuya distribución muestral se conozca en H_0 y que esté relacionado con q y establecer, en base a dicha distribución, la región crítica: región en la que el estadístico tiene una probabilidad menor que a si H_0 fuera cierta y, en consecuencia, si el estadístico cayera en la misma, se rechazaría H_0.

Obsérvese que, de esta manera, se está más seguro cuando se rechaza una hipótesis que cuando no. Por eso se fija como H_0 lo que se quiere rechazar. Cuando no se rechaza, no se ha demostrado nada, simplemente no se ha podido rechazar. Por otro lado, la decisión se toma en base a la distribución muestral en H_0, por eso es necesario que tenga la igualdad.

5. Calcular el estadístico para una muestra aleatoria y compararlo con la región crítica, o equivalentemente, calcular el "valor p" del estadístico (probabilidad de obtener ese valor, u otro más alejado de la H_0, si H_0 fuera cierta) y compararlo con a.

Ejemplo:

Estamos estudiando el efecto del estrés sobre la presión arterial. Nuestra hipótesis es que la presión sistólica media en varones jóvenes estresados es mayor que 18 cm de Hg. Estudiamos una muestra de 36 sujetos y encontramos:

$$\overline{X} = 18{,}5 \qquad S = 3{,}6$$

1. Se trata de un contraste sobre medias. La hipótesis nula (lo que queremos rechazar) es:

$$H_0 : \mu = 18$$

2. la hipótesis alternativa

$$H_1 : \mu > 18$$

Es un contraste lateral derecho.

3. Fijamos "a priori" el nivel de significación en 0,05 (el habitual en Biología).

4. El estadístico para el contraste es

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

y la región crítica $T>t_a$

Si el contraste hubiera sido lateral izquierdo, la región crítica sería $T<t_{1-a}$ y si hubiera sido bilateral $T<t_{1-\ a/2}$ o $T>t_{\ a/2}$ En este ejemplo $t_{(35)0,05}$=1,69.

5. Calculamos el valor de t en la muestra

$$T = \frac{18,5 - 18}{3,6/\sqrt{36}} = 0,833$$

no está en la región crítica (no es mayor que 1,69), por tanto, no rechazamos H_0.

Una hipótesis estadística es una afirmación respecto a alguna característica de una población. Contrastar una hipótesis es comparar las predicciones con la realidad que observamos. Si dentro del margen de error que nos permitimos admitir, hay coincidencia, aceptaremos la hipótesis y en caso contrario la rechazaremos.

La hipótesis emitida se suele designar por Ho y se llama Hipótesis nula porque parte del supuesto que la diferencias entre el valor verdadero del parámetro y su valor hipotético es debida al azar, es decir no hay diferencia. La hipótesis contraria se designa por H1 y se llama Hipótesis alternativa

Los contrastes pueden ser unilaterales o bilaterales (también llamados de una o dos colas) según establezcamos las hipótesis, si las definimos en términos de igual y distinto estamos ante una hipótesis unilateral, si suponemos una dirección (en términos de mayor o menor) estamos ante uno unilateral.

Se trata pues, de extraer conclusiones a partir de una muestra aleatoria y significativa, que permitan aceptar o rechazar una hipótesis previamente emitida, sobre el valor de un parámetro desconocido de la población.

El método que seguiremos es el siguiente:

Enunciar la hipótesis

Elegir un nivel de significación a y construir la zona de aceptación, intervalo fuera del cual sólo se encuentran el a100% de los casos más raros. A la zona de rechazo la llamaremos región crítica, y su área es el nivel de significación.

Verificar la hipótesis extrayendo una muestra cuyo tamaño se ha decidido en el paso anterior y obteniendo de ella el correspondiente estadístico (media o proporción en nuestro caso).

Decidir. Si el valor calculado en la muestra cae dentro de la zona de aceptación se acepta la hipótesis y si no se rechaza.

Una hipótesis estadística es una asunción relativa a una o varias poblaciones, que puede ser cierta o no. Las hipótesis estadísticas se pueden contrastar con la información extraída de las muestras y tanto si se aceptan como si se rechazan se puede cometer un error.

La hipótesis formulada con intención de rechazarla se llama hipótesis nula y se representa por H0. Rechazar H0 implica aceptar una hipótesis alternativa (H1).

3.14 Construcción de Test de hipótesis.

Seis pasos básicos para configurar y realizar correctamente una prueba de hipótesis.

1. Especificar las hipótesis.
2. Elegir un nivel de significancia (también denominado alfa o α).
3. Determinar la potencia y el tamaño de la muestra para la prueba.
4. Recolectar los datos.
5. Comparar el valor p de la prueba con el nivel de significancia.
6. Decidir si rechazar o no rechazar la hipótesis nula.

Ejemplo:

Un gerente de ventas de libros universitarios afirma que en promedio sus representantes de ventas realiza 40 visitas a profesores por semana. Varios de estos representantes piensan que realizan un número de visitas promedio superior a 40. Una muestra tomada al azar durante 8 semanas reveló un promedio de 42 visitas semanales y una desviación estándar de 2 visitas. Utilice un nivel de confianza del 99% para aclarar esta cuestión.

Datos:

VP = 40

$\bar{x} = 42$

n = 8

$S = 2$

Nivel de confianza del 99%

Nivel de significación = (100%-99%)/2 = 0,5% = 0,005

$$t_{prueba} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

Solución:

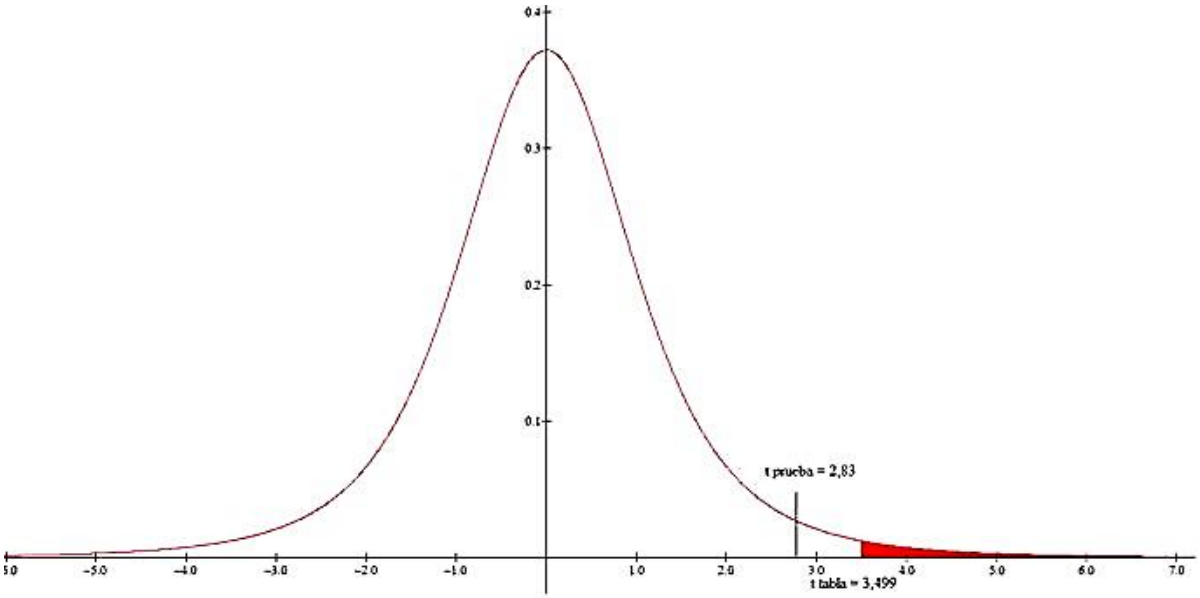
H0: VP = 40

H1: VP > 40

Grados de libertad: n-1 = 8-1 =7

a = 0,005 $\Longrightarrow t_{tabla} = 3{,}499$

$$t_{prueba} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{42 - 40}{\frac{2}{\sqrt{8}}} = \frac{2}{0{,}7071} = 2{,}83$$



H0 es aceptada, ya que t prueba (2,83) es menor que t tabla (3,499), por lo que no es acertado pensar que están realizando un número de visitas promedio superior a 40.

3.15 Contraste de hipótesis paramétricas.

Es la técnica estadística que se usa para estudiar si una determinada afirmación acerca de cierto parámetro poblacional es confirmada o invalidada por los datos de una muestra extraída de dicha población.

Ejemplo: ¿La selección del jurado es aleatoria? ¿p=0.5?

HIPÓTESIS NULA H_0	HIPÓTESIS ALTERNATIVA H_a
-Es la que se supone cierta, y debe aceptarse salvo prueba en contra	- Es la que niega la hipótesis nula
- Los datos muestrales pueden refutarla	-Los datos muestrales pueden mostrar evidencias a favor.
- No debe ser rechazada sin una gran evidencia en contra	-No debe ser aceptada sin una gran evidencia a favor

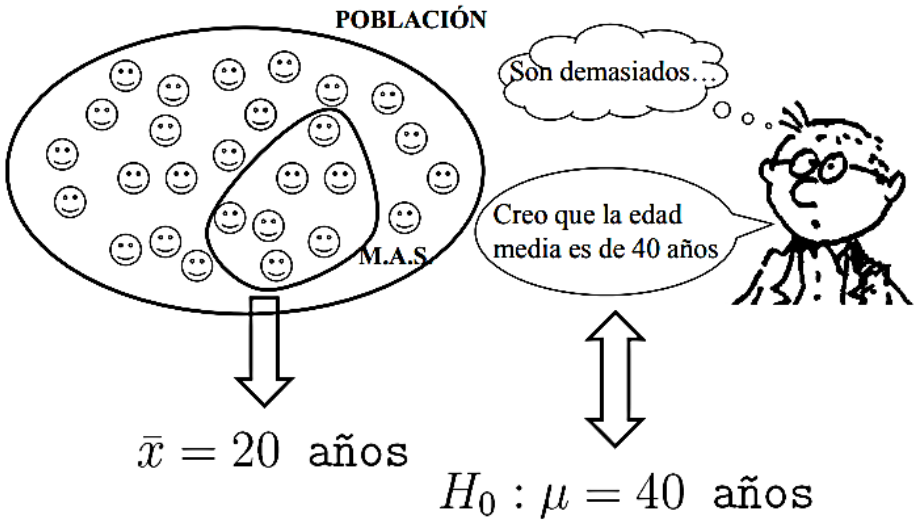
EJEMPLO: $\begin{cases} H_0 : p = 0.5 \\ H_a : p \neq 0.5 \end{cases}$

9

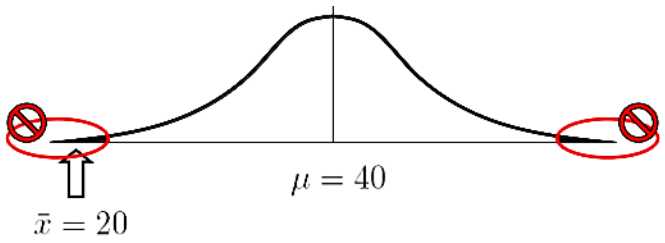
Razonamiento Básico del Contraste de Hipótesis

Localizar un suceso que sea muy improbable cuando la hipótesis nula se supone cierta; si, una vez extraída una muestra aleatoria acontece dicho suceso, o bien es que el azar nos ha jugado una mala pasada al elegir una muestra muy rara, o bien, como parece más razonable, la hipótesis nula es falsa.

Ejemplo: ¿Cuál es la edad media de la población?



Si suponemos que la hipótesis nula es cierta y normalidad en los datos, X = edad de un individuo de la población sigue una distribución normal de media 40…



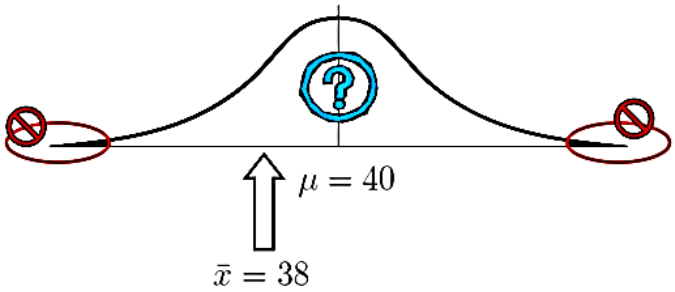
y por lo tanto, es muy improbable que la media muestral se encuentre en las zonas marcadas en rojo.

¡¡¡Sin embargo, ha ocurrido!!!

Rechazamos que H_0 sea cierta

12

Pero, ¿qué hubiéramos podido afirmar si hubiésemos obtenido una media muestral de 38 años?



- No hay evidencias contra la hipótesis nula

-No podemos rechazar la hipótesis nula (NI ACEPTARLA)

- El contraste no es significativo.

UNIDAD IV. DEMOGRAFIA

4.1 Test para poblaciones normales

Una población

En la inferencia sobre una variable numérica en una población, el objetivo principal de los test de hipótesis es contrastar el valor de alguna medida de posición (media o mediana), de dispersión (varianza) o de algún otro parámetro poblacional. Así, si se cuenta con información muestral sobre el número de horas diarias que un individuo está viendo la

televisión, trataremos de ver si podemos aceptar que el promedio de horas en la población toma un determinado valor.

En principio, si hacemos inferencia sobre un resumen de la variable podemos considerar que dicha variable sigue una distribución y plantear el problema desde la óptica paramétrica. Esa distribución puede ser cualquiera de las muchas distribuciones tipo (binomial, Poisson, normal, uniforme, gamma,). Ahora bien, por distintos motivos (variables, originales o transformadas, con distribución más o menos campaniforme, teorema del límite central, simplicidad en los procedimientos, ...), lo más habitual es la suposición de normalidad. Entonces, suponiendo que la variable *X* sigue una distribución normal de media μ y desviación σ, plantearemos contrastes de hipótesis sobre dichos parámetros.

En la inferencia sobre dos variables numéricas en una población se trata de ver si existe relación lineal entre las mismas a partir de la información muestral. Así, podemos tratar de analizar si existe relación entre la renta y el consumo de las familias y cuál es la intensidad y el sentido de la misma, concretándose la inferencia en el coeficiente de correlación ρ. Este tipo de análisis, denominado genéricamente de correlación, conduce de forma inmediata al análisis de regresión: si existe relación trataremos de encontrar la función que mejor exprese esta relación.

Dos poblaciones

En este caso, el objetivo fundamental es comparar la distribución de una variable cuantitativa *X* en las dos poblaciones (subpoblaciones) determinadas por las modalidades de una característica cualitativa dicotómica o, lo que es lo mismo, estudiar si la variable cuantitativa (variable respuesta) presenta diferencias significativas en cada uno de los dos niveles de la variable cualitativa (factor). Esta comparación se realizará a partir de la información parcial que proporcionan dos muestras. Denotaremos por X_1 y X_2 a la variable cuantitativa en cada una de las dos situaciones.

Si podemos aceptar normalidad, el objetivo general de comparar dos poblaciones se traduce en comparar las medias de la variable en cada una de ellas; aunque suele tener únicamente un interés instrumental, también se pueden comparar las varianzas.

Para poder realizar estas comparaciones se utilizan dos muestras provenientes de individuos diferentes (estudiar las posibles diferencias salariales en función del sexo a partir de una muestra de hombres y una muestra de mujeres); en este caso hablaremos de muestras independientes. A veces se puede utilizar una muestra con los mismos individuos para las dos situaciones de la variable (comparar la valoración de dos detergentes a partir de los datos que sobre uno y otro proporciona una única muestra de consumidores); en este caso hablaremos de muestras apareadas o relacionadas. Siempre que se puedan considerar muestras relacionadas, este procedimiento proporciona en principio mejores inferencias.

El problema de comparar medias se puede generalizar a más de dos poblaciones, los llamados procedimientos ANOVA.

Poblaciones normales	Una población	Una variable	**C. de H. sobre la media poblacional**
			C. de H. sobre la varianza poblacional
		Dos variables	**C. de H. sobre el coeficiente de correlación poblacional**
	Dos poblaciones	Muestras independientes	**C. de H. sobre la diferencia de medias poblacionales**
			C. de H. sobre el cociente de varianzas poblacionales
		Muestras relacionadas	**C. de H. sobre la diferencia de medias poblacionales**

4.2 Test para poblaciones binomiales y de Poisson

Nos encontramos con un modelo derivado de un proceso experimental puro, en el que se plantean las siguientes circunstancias.

Se realiza un número n de pruebas (separadas o separables).

Cada prueba puede dar dos únicos resultados A y Ã.

La probabilidad de obtener un resultado A es p y la de obtener un resultado Ã es q, con q= 1-p, en todas las pruebas. Esto implica que las pruebas se realizan exactamente en las mismas condiciones y son, por tanto, independientes en sus resultados. Si se trata de extracciones, (muestreo), las extracciones deberán ser con devolución

(reemplazamiento), o bien población grande (M.A.S). A este respecto hagamos una consideración: si el proceso consiste en extraer individuos de una población y observar si poseen cierta característica: el parámetro n será el número de extracciones (tamaño muestral) y el parámetro p la proporción de individuos de la población que poseen la característica en cuestión. Se ha comentado que para que la probabilidad, de que en cada extracción obtengamos un individuo poseedor de la característica sea constante en todas la pruebas es necesario que las proporciones poblacionales no cambien tras cada extracción es decir se reemplace cada individuo extraído .Sin embargo si la población es muy grande, aunque no reemplacemos los individuos extraídos las variaciones en las proporciones de la población restante serán muy pequeñas y, aunque de hecho las probabilidades de, obtener un éxito varíen tras cada prueba, esta variación será muy pequeña y podremos considerar que son constantes .

En estadística, el test binomial es un test exacto de la significación estadística de las desviaciones de una teóricamente distribución esperada de las observaciones en dos categorías.

Uso común

Un uso común del test binomial es en el caso donde la hipótesis nula es aquella en la que las dos categorías son igualmente probables de que ocurran (como el lanzamiento de una moneda). Las tablas están ampliamente disponibles para dar la importancia observada en el número de observaciones en las categorías para este caso. Sin embargo, como muestra el siguiente ejemplo, el test binomial no se limita a este caso.

Donde hay más de dos categorías, y un test exacto es requerido, el test multinomial, basado en la distribución multinomial, debe usarse en vez del test binomial.

Muestras grandes

Para muestras grandes como las del siguiente ejemplo, La distribución binominal es bien aproximada por convenientes distribuciones continuas, y éstos se utilizan como la base para las pruebas alternativas que son mucho más rápidas para computar, Prueba χ^2 de Pearson y el G-test. Sin embargo, para pequeñas muestras estas aproximaciones se descomponen, y no hay alternativa para el test binomial.

La distribución de Poisson se aplica a varios fenómenos discretos de la naturaleza (esto es, aquellos fenómenos que ocurren 0, 1, 2, 3... veces durante un periodo definido de tiempo o en un área determinada) cuando la probabilidad de ocurrencia del fenómeno es constante en el tiempo o el espacio.

Ejemplos de estos eventos que pueden ser modelados por la distribución de Poisson incluyen:

- El número de autos que pasan a través de un cierto punto en una ruta (suficientemente distantes de los semáforos) durante un periodo definido de tiempo.
- El número de errores de ortografía que uno comete al escribir una única página.
- El número de llamadas telefónicas en una central telefónica por minuto.
- El número de servidores web accedidos por minuto.
- El número de animales muertos encontrados por unidad de longitud de ruta.
- El número de mutaciones de determinada cadena de ADN después de cierta cantidad de radiación.
- El número de núcleos atómicos inestables que se han desintegrado en un determinado período.

4.3 Test basado en el estadístico Chi-cuadrado

Esta prueba puede utilizarse incluso con datos medibles en una escala nominal. La hipótesis nula de la prueba Chi-cuadrado postula una distribución de probabilidad totalmente especificada como el modelo matemático de la población que ha generado la muestra.

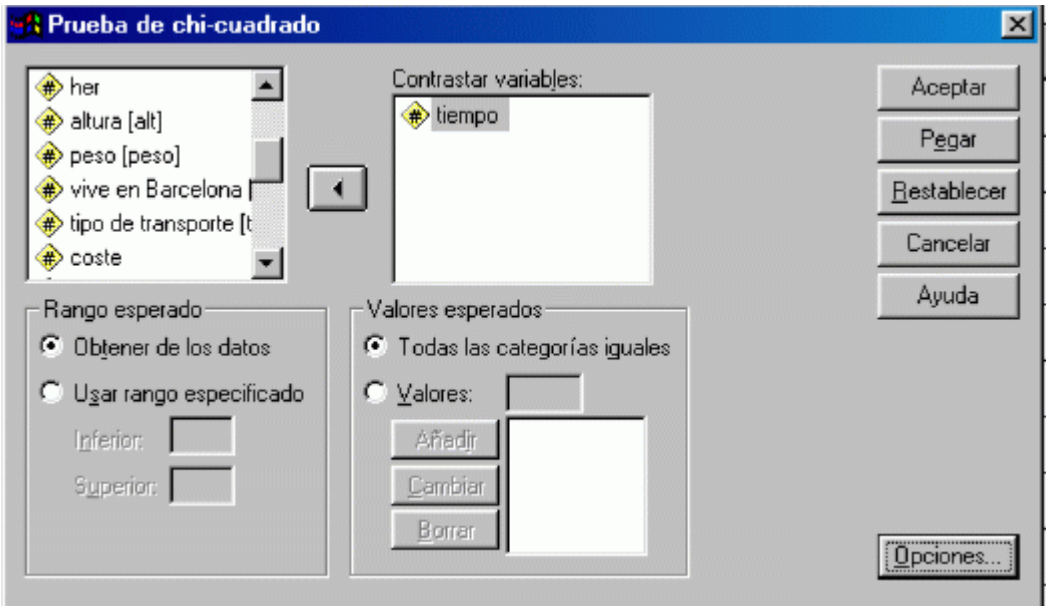
Para realizar este contraste se disponen los datos en una tabla de frecuencias. Para cada valor o intervalo de valores se indica la frecuencia absoluta observada o empírica (Oi). A continuación, y suponiendo que la hipótesis nula es cierta, se calculan para cada valor o intervalo de valores la frecuencia absoluta que cabría esperar o frecuencia esperada (Ei=n ·pi, donde n es el tamaño de la muestra y pi la probabilidad del i-ésimo valor o intervalo de valores según la hipótesis nula). El estadístico de prueba se basa en las diferencias entre la Oi y Ei y se define como:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Este estadístico tiene una distribución Chi-cuadrado con k-1 grados de libertad si n es suficientemente grande, es decir, si todas las frecuencias esperadas son mayores que 5. En la práctica se tolera un máximo del 20% de frecuencias inferiores a 5.

Si existe concordancia perfecta entre las frecuencias observadas y las esperadas el estadístico tomará un valor igual a 0; por el contrario, si existe una gran discrepancia entre estas frecuencias el estadístico tomará un valor grande y, en consecuencia, se rechazará la hipótesis nula. Así pues, la región crítica estará situada en el extremo superior de la distribución Chi-cuadrado con k-1 grados de libertad. Para realizar un contraste Chi-cuadrado la secuencia es:

- *Analizar*
- *Pruebas no paramétricas*
- *Chi-cuadrado*



En el cuadro de diálogo *Prueba chi-cuadrado* se indica la variable a analizar en *Contrastar variables*.

En *Valores esperados* se debe especificar la distribución teórica activando una de las dos alternativas. Por defecto está activada *Todas las categorías iguales* que recoge la hipótesis de que la distribución de la población es uniforme discreto. La opción *Valores* requiere especificar uno a uno los valores esperados de las frecuencias relativas o absolutas correspondientes a cada categoría, introduciéndolos en el mismo orden en el que se han definido las categorías.

El recuadro *Rango esperado* presenta dos opciones: por defecto está activada *Obtener de los datos* que realiza el análisis para todas las categorías o valores de la variable; la otra alternativa, *Usar rango especificado*, realiza el análisis sólo para un determinado rango de valores cuyos límites *Inferior* y *Superior* se deben especificar en los recuadros de texto correspondientes.

El cuadro de diálogo al que se accede con el botón *Opciones* ofrece la posibilidad de calcular los *Estadísticos Descriptivos* y/o los *Cuartiles*, así como seleccionar la forma en que se desea tratar los valores perdidos.

4.4 Test de bondad de ajuste.

La bondad de ajuste de un modelo estadístico describe lo bien que se ajusta un conjunto de observaciones. Las medidas de bondad en general resumen la discrepancia entre los valores observados y los valores esperados en el modelo de estudio.

Estas pruebas permiten verificar que la población de la cual proviene una muestra tiene una distribución especificada o supuesta.
Sea X: variable aleatoria poblacional $f_0(x)$ la distribución (o densidad) de probabilidad especificada o supuesta para X.
Se desea probar la hipótesis: Ho: f(x) = f0 (x)

En contraste con la hipótesis alterna: Ha: f(x) no= f0 (x) (negación de Ho)

Prueba de bondad de ajuste chi cuadrado x2

El procedimiento de la prueba requiere una muestra aleatoria de tamaño n proveniente de la población cuya distribución de probabilidad es desconocida.

Estas n observaciones se pueden distribuir en k intervalos de clases y pueden ser representadas en histogramas.

La prueba se puede utilizar tanto para distribuciones discretas como para distribuciones continuas.

La prueba se puede sintetizar en los siguientes pasos.

1. Se colocan los n datos históricos (muéstrales) en una tabla de frecuencia.

2. Se propone una distribución de probabilidad una distribución de probabilidad de acuerdo con la tabla de frecuencia o con la curva que muestre un histograma o polígono de frecuencia.

3. Con la distribución propuesta, se calcula la frecuencia esperada para cada uno de los intervalos (FEi) de la siguiente manera:

- Si la variable es continua se halla mediante la integración de la distribución propuesta y luego se multiplica por el número total de datos.

- Si la variable es continua se utiliza de modelo matemático de la distribución propuesta y se evalúan todas las categorías y luego se multiplica por el número total de datos.

4. Se calcula el estadístico de prueba

5. Si el estimador C es menor o igual al valor correspondiente X2 con m-k-1 grados de libertad (K= números de parámetros estimados de la distribución propuesta estimada por los estadísticos muéstrales) y a un nivel de confiabilidad de 1-α, entonces no se puede rechazar la hipótesis de que los datos siguen la distribución que se propuso.

4.5 Test de heterogeneidad

La heterogeneidad estadística es la presencia de diferencias entre los efectos calculados de la intervención, que son mayores que lo que es de esperar si se debieran solamente a las variaciones al azar (muestrales)

Cuando hablamos de heterogeneidad podemos distinguir dos aspectos: por un lado el relativo a las diferencias existentes entre los estudios en cuanto a características de los pacientes incluidos, la metodología utilizada, el tiempo de seguimiento, las dosis

empleadas, la localización geográfica, etc. y por otro lado el concepto de heterogeneidad estadística ,que únicamente cuantifica la variabilidad entre los resultados de los estudios, y que puede ser debida a las diferencias reales de planteamiento y ejecución entre los estudios incluidos, o a otras causas.

La elaboración de guías de práctica clínica se fundamenta cada vez más en los resultados de revisiones sistemáticas, en las que los meta-análisis constituyen una herramienta crucial, y puesto que la validez de la conclusión global de un meta-análisis depende en gran medida de la homogeneidad de los estudios incluidos, resulta de extraordinaria importancia disponer de algún parámetro que cuantifique la heterogeneidad.

Uno de los aspectos de la heterogeneidad, el relativo a las diferencias clínicas o biológicas entre estudios y a las diferencias de procedimientos, es en primer lugar un problema metodológico, ya que habrá que decidir si las diferencias entre los estudios, que siempre existen, permiten o no combinarlos, independientemente de los resultados que en ellos se haya obtenido, y por lo tanto es previo a la ejecución del meta-análisis. Mientras que la heterogeneidad estadística trata de cuantificar la variabilidad del resultado medido en los diferentes estudios con respecto al resultado global promedio, y determinar si dicha variabilidad es superior a la que sería esperable por puro azar.

La prueba estadística más ampliamente utilizada para verificar la posible existencia de heterogeneidad superior a la esperable por puro azar se denomina Q de Cochran, y se basa en calcular la suma de las desviaciones cuadráticas entre el resultado individual de cada estudio y el resultado global, ponderadas por el mismo peso con el que cada resultado interviene en el cálculo global.

Sin embargo, el empleo de esta prueba no está exento de problemas, ya que si el número de estudios es pequeño su capacidad para detectar heterogeneidad es muy baja (poca potencia de contraste), mientras que, por el contrario, cuando el meta-análisis combina gran número de estudios, el resultado puede ser estadísticamente significativo incluso cuando la magnitud de la heterogeneidad no sea de relevancia clínica. Esto no es nada nuevo: son los problemas inherentes a la metodología de las pruebas de contraste estadístico.

4.6 Test de homogeneidad

Se plantea el problema de la existencia de homogeneidad entre r poblaciones, para lo cual se realizan muestras independientes en cada una de ellas. Los datos muéstrales vienen clasificados en s clases y sus frecuencias absolutas se presentan en forma de una matriz r x s. siendo nij el número de observaciones en la i-ésima población pertenecientes a la j-ésima clase.

Se quiere contrastar la hipótesis nula de que las probabilidades asociadas a las s clases son iguales en las r poblaciones. Donde es el tamaño muestral para la i ésima población, es la frecuencia marginal de la j-ésima clase y es el tamaño muestral total. El estadístico L se distribuye como una con (r - 1) (s - 1) grados de libertad. El contraste se realiza con un nivel de significación del 5%.

Objetivo de la prueba: se utiliza cuando se tienen varias muestras independientes de n individuos que se clasifican respecto a una variable cualitativa y se desea conocer a partir de datos muéstrales, si provienen de la misma población (el objetivo es comparar diferentes muestras).

Es decir, en esta prueba se tienen varias muestras independientes correspondientes a las categorías de una de las variables y se clasifican las observaciones respecto a la otra variable. La prueba tiene la finalidad de conocer si la distribución de la variable estudiada difiere en las r" poblaciones subyacentes de las cuales se obtuvieron las muestras.

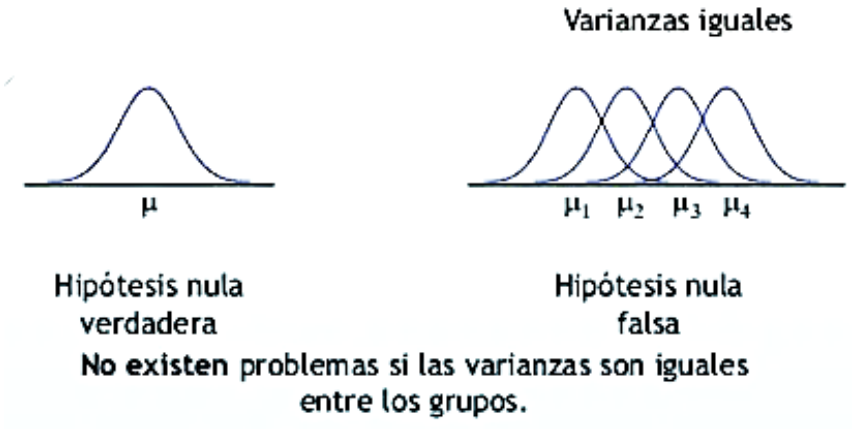
Limitaciones de la prueba (las mismas que para la prueba de Independencia):

-Se necesita que más del 20 % de los valores esperados estén por encima de 5 y que ninguna celda tenga valor esperado menor a 1.

-Si la tabla es de 2x2, todas las celdas deben tener valores esperados por encima de 5.

-En el caso de la tabla de 2x2 si existe una sola celda con valor esperado menor que 5, esto representaría un 25 % de las celdas con esa condición, por lo que se utilizaría la Prueba de las Probabilidades exactas de Fisher en lugar de la Prueba χ 2, ya que en éste caso no es posible agrupar categorías.

El estadígrafo de prueba y la regla de decisión son similares a los de la Prueba Ji cuadrado de independencia.



4.7 Tablas de Contingencia.

Una tabla de contingencia es una herramienta utilizada en la rama de la estadística, la cual consiste en crear al menos dos filas y dos columnas para representar datos categóricos en términos de conteos de frecuencia.

Esta herramienta, que también se conoce como tabla cruzada o como tabla de dos vías, tiene el objetivo de representar en un resumen, la relación entre diferentes variables categóricas.

Objetivos de una tabla de contingencia

La tabla permite medir la interacción entre dos variables para conocer una serie de información "oculta" de gran utilidad para comprender con mayor claridad los resultados de una investigación.

La tabla sólo mostrará los encuestados que respondieron ambas preguntas, lo que significa que las frecuencias mostradas pueden diferir de una tabla de frecuencias estándar.

El informe que ofrece también mostrará las Estadísticas Chi-cuadrado de Pearson, el cual representa el grado de correlación entre las variables que usan el chi-cuadrado, el valor p y el grado de libertad.

Los objetivos de la tabla de contingencia son los siguientes:

- Ordenar la información recolectada para un estudio cuando los datos se encuentran divididos de forma bidimensional, esto significa a que se relaciona con dos factores cualitativos.
- El otro objetivo de la tabla de contingencia es analizar si hay una relación entre las variables cualitativas, ya sean dependientes o independientes.

Ventajas de realizar una tabla de contingencia

Entre los principales beneficios de realizar una tabla de contingencia se encuentran los siguientes:

1. Facilita la lectura de los datos recolectados, ya que permite agruparlos cuando aún se encuentran sin procesar, lo que disminuye el margen de error al realizar un informe de investigación.
2. Gracias a la tabla de contingencia es posible realizar gráficas que permitan visualizar la información fácilmente para su comprensión.
3. A diferencia de otros métodos estadísticos de análisis de datos, la tabla de contingencia permite ahorrar tiempo durante la correlación de variables.
4. Las tablas ofrecen resultados claros y precisos que permiten tomar mejores decisiones y crear estrategias basadas en datos.

Cuando usar una tabla de contingencia

La tabla de contingencia generalmente se realiza en datos categóricos, es decir que se pueden dividir en grupos mutuamente excluyentes.

Un ejemplo de datos categóricos es la región de ventas de un producto. Típicamente, la región se puede dividir en categorías como área geográfica (norte, sur, noreste, oeste, etc.) o estado.

Es importante recordar que los datos categóricos no pueden pertenecer a más de una categoría.

Uno de los principales usos de una tabla de contingencia es analizar la relación que existe entre los datos, las cuales no son fáciles de identificar.

Importancia de hacer uso de una tabla de contingencia
La información sin procesar puede ser difícil de interpretar. Incluso para pequeños conjuntos de datos, es demasiado fácil obtener resultados incorrectos con solo mirar los datos. La tabla ofrece un método simple de agrupar variables, que minimiza el potencial de confusión o error al proporcionar resultados claros.

Además, una tabla puede ayudarnos a obtener grandes conocimientos de los datos sin procesar. Estas ideas no son fáciles de ver cuando los datos sin formato se organizan como una tabla.

Dado que la tabla de contingencia traza claramente las relaciones entre las preguntas categóricas, los investigadores pueden obtener información más profunda, que de otro modo se habría pasado por alto o habría tomado mucho tiempo descifrar de formas más complicadas de análisis estadístico.

La tabla facilita la interpretación de los datos, lo cual es beneficioso para los investigadores que tienen un conocimiento limitado del análisis estadístico. Las personas no necesitan programación estadística para correlacionar variables categóricas.

La claridad que ofrece una tabla ayuda a los profesionales a evaluar su trabajo actual y trazar estrategias futuras.

Ejemplo

Por ejemplo, se considera la distribución conjunta de dos variables y la correspondiente tabla de contingencia en una muestra de pacientes de un hospital. Se tiene la siguiente tabla donde se consideran el riesgo de ataque al corazón respecto a la toma de aspirinas:

- $X \equiv$ Se toma aspirina o placebo $(I = 2)$.
- $Y \equiv$ Se sufre ataque cardíaco o no $(J = 3)$.

	Mortal	No mortal	No ataque	Totales
Placebo	18	171	10845	11034
Aspirina	5	99	10933	11037

Como resumen de la información que presenta la tabla, de los 11034 enfermos que tomaron un placebo, 18 tuvieron un ataque al corazón, mientras que de los 11037 que tomaron aspirina, 5 tuvieron ataques al corazón. La distribución conjunta de dos variables categóricas determina su relación. Esta distribución también determina las distribuciones marginales y condicionales.

4.8 Demografía. Conceptos básicos

La Demografía es una ciencia social que estudia el volumen, crecimiento y características de un grupo de población humana en un periodo de tiempo determinado o a su evolución y se podría traducir como estudio de la población.

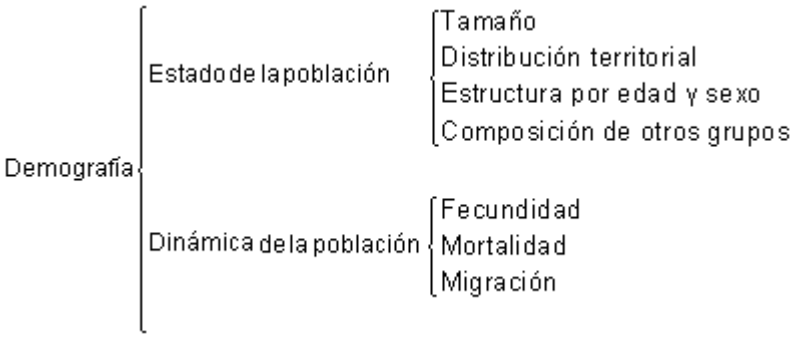
Se trata de estudios estadísticos relativos, por ejemplo, a la natalidad, mortalidad y la migración. Organismos oficiales se encargan de recoger este tipo de datos y se utilizan instrumentos como encuestas y padrones.

La Demografía es una ciencia que estudia las poblaciones humanas. No obstante, muchas otras ciencias tienen este mismo objetivo, entre otras: la Sociología, la Antropología, la Sicología, las Ciencias Políticas, la Economía, etc. De hecho, el objeto de estudio de todas las ciencias sociales es la población humana.

El objetivo de la Demografía consiste en estudiar los movimientos que se presentan en las poblaciones humanas. El término de población debe ser entendido como el conjunto de personas que se agrupan en cierto ámbito geográfico y está propenso a continuos

cambios. De esta manera, el área temática de la Demografía se concentra en el estado y la dinámica de estas poblaciones en el tiempo.

El estado de la población hace referencia a su tamaño, distribución territorial y estructura por edad, sexo, u otros subgrupos de interés. Mientras que la dinámica se enfoca en aquellos elementos que pueden provocar cambios en el estado a lo largo del tiempo. En este sentido, los componentes de mayor interés son la fecundidad, la mortalidad y la migración.



La demografía es la ciencia que se ocupa de estudiar la estructura, la evolución, las características y el tamaño de la población humana.

Sobre todo, la demografía es una ciencia social y sus estudios sobre la población humana pueden ser de forma comparativa y cuantitativa. La demografía se auxilia de la estadística y la utiliza como una herramienta fundamental para realizar sus estudios sobre los datos obtenidos y poder realizar las comparaciones necesarias.

Por otra parte, la población se forma por un conjunto de personas que comparten rasgos de tipo social, cultural, geográficos, políticos o de cualquier otro tipo. Lo que les permite mantener cierta homogeneidad y permanencia en el tiempo.

Tipos de demografía

La demografía puede ser de dos tipos: estática y dinámica.

1. Demografía estática

En efecto, se denomina demografía estática la que estudia problemas de población como sus características estructurales, el territorio y su dimensión en un momento determinado y definido.

Características estructurales: En realidad, en las características estructurales se pueden considerar variables como el lugar de nacimiento, la nacionalidad, la lengua que hablan, el nivel educativo, el nivel económico, la tasa de natalidad, la edad, el sexo, entre algunos de los más importantes.

El territorio: En cuanto al territorio se refiere a la relación conforme al lugar que ocupa la población. El cual podría ser un país, una comunidad, un departamento, un municipio, una provincia o una aldea.

La dimensión: Asimismo, la dimensión es el número de habitantes que residen en un espacio geográfico.

2. Demografía dinámica

Por otro lado, la demografía dinámica estudia la evolución de la población humana considerando aspectos como la edad, el sexo, la tasa de natalidad, fecundidad, la familia, la educación, la tasa de divorcios, tasa de mortalidad, el trabajo y las migraciones.

Variables más importantes que estudia la demografía

Las variables más importantes que estudia la demografía son:

1. Tasa de natalidad

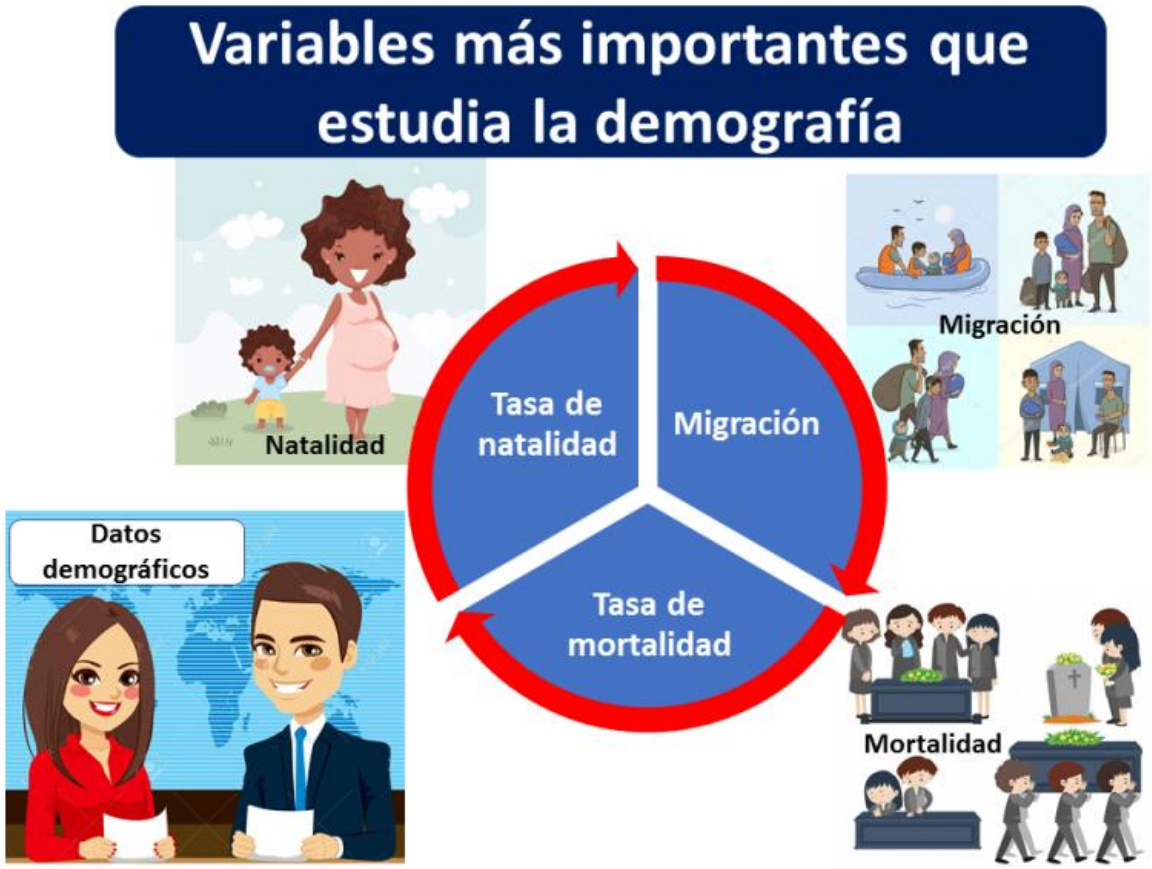
Con relación a la tasa de natalidad, esta mide la cantidad de nacimientos que se producen dentro de un grupo de población, considerando un periodo de tiempo determinado. Refleja el ritmo al que una población se reproduce y mantiene la supervivencia de la especie.

2. La migración

Ahora, la migración determina el grado de movilidad que experimenta una población, en este aspecto la demografía estudia todo movimiento o desplazamiento de población que se produce desde un lugar de origen a otro de destino.

3. Tasa de mortalidad

Por consiguiente, la tasa de mortalidad registra la frecuencia de fallecimientos dentro de un período determinado, considerando variables como la esperanza de vida y las causas de los fallecimientos.



Finalmente, los estudios demográficos son muy importantes para cualquier sociedad porque permiten conocer los cambios y la evolución que se producen en una población determinada, para poder obtener estos datos se utilizan especialmente los censos y los estudios estadísticos. El objetivo de tener los datos de las variables demográficas es que permite a los gobiernos planificar, diseñar e implementar políticas que se adapten a las necesidades particulares de una determinada población.

4.9 Modelos de crecimiento de poblaciones

El crecimiento poblacional se refiere al incremento del número de habitantes en un espacio y tiempo determinado, el cual se puede medir a través de una fórmula aritmética. También se puede emplear como sinónimo el término crecimiento demográfico.

Cuando se hace mención al crecimiento poblacional se puede hacer referencia a cualquier especie animal, sin embargo, se suele usar para referirse a los seres humanos, en especial cuando se realizan investigaciones acerca del crecimiento de la población.
Los datos que se obtienen de estos análisis son de gran importancia, tanto para los gobiernos de un país como, para las diferentes organizaciones internacionales.

Cabe resaltar que durante el siglo XX la población de seres humanos ha crecido y sigue creciendo en gran porcentaje, lo que ha generado preocupación, en especial por sus consecuencias sobre el uso y cuidado de los recursos naturales, entre otros.

Las zonas urbanas son las que presentan mayor crecimiento demográfico, así como, los países en vías de desarrollo. Por el contrario, el crecimiento poblacional es menor en los países desarrollados.

Por ejemplo, el crecimiento poblacional en México ha ido en aumento a lo largo de su historia, es el país con más habitantes hispanohablantes en América Latina. México tiene una población aproximada a 130 millones de habitantes, y se estima que seguirá creciendo gracias a diversos factores gracias a su continuo desarrollo político, económico y social.

Tipos de crecimiento poblacional

Existen dos tipos de crecimiento poblacional denominadas crecimiento exponencial y crecimiento logístico.

Crecimiento exponencial: presenta los datos con una curva en forma de J, y refleja cómo las poblaciones creen muy rápido y luego se detiene de manera repentina debido a diversos factores.

Crecimiento logístico: presenta los datos de crecimiento poblacional a través de una curva en forma de S (sigmoidea). Expone los datos de una población cuyo crecimiento tiene una etapa lenta, luego toma velocidad y crece y, finalmente decrece de manera gradual buscando un equilibrio.

Tasa de crecimiento poblacional

La tasa de crecimiento poblacional es un índice que se emplea tanto en las investigaciones demográficas como ecológicas con el fin de exponer cómo ha sido el incremento o disminución de la población de una especie en un lugar y tiempo específico.
Los resultados obtenidos se exponen, generalmente, en porcentajes y se emplean tanto para comparar con los análisis anteriores, como para realizar futuras aproximaciones.

Por otro lado, la medición de la tasa de crecimiento poblacional se ve afectada directamente tanto por cuatro importantes índices: natalidad, mortalidad, emigración e inmigración, las cuales que varían en el tiempo y por diversas circunstancias.

La fórmula para obtener los datos de la tasa de crecimiento poblacional se obtiene de la siguiente manera:

Tasa de crecimiento poblacional = (población final del período) – (población al principio del período) / población al principio del período.
Sin embargo, la ecuación que se suele emplear para expresar las variaciones del crecimiento poblacional durante un período y en porcentaje es la siguiente:

Porcentaje crecimiento = tasa / crecimiento x 100%
Ahora bien, si el resultado obtenido es un valor positivo, entonces quiere decir que el número de habitantes de un país o región ha aumentado.

En el caso contrario, si arroja un número negativo es porque ha disminuido el crecimiento poblacional. Pero, en caso de obtener un cero como resultado, eso quiere indicar que la población se encuentra equilibrada.

Factores que influyen en el crecimiento poblacional

Existen diversos factores que han influido en el crecimiento poblacional, entre ellos se pueden mencionar los siguientes.

- La elaboración y uso de herramientas que facilitaron diversos trabajos como la construcción de viviendas, el cultivo y la recolección de alimentos, entre otros.
- La actividad agrícola ha sido importante para el desarrollo humano ya que ha permitido la construcción de ciudades a su alrededor, la actividad comercial y el intercambio cultural.
- La Revolución industrial impactó el desarrollo humano de manera positiva tras alcanzar la posibilidad de mejorar la calidad de vida de los trabajadores, delimitar horas de trabajo, incrementar el número de empleos, así como el desarrollo tecnológico, entre otros.
- El desarrollo continuo en el área de la salud también ha sido un factor que ha incrementado el crecimiento poblacional al ofrecer mayores esperanzas de vida, la posibilidad de evita y prevenir enfermedades, entre otros.
- Mejoras en la calidad de vida, en términos generales, es decir, contar con un buen sistema de salud y de educación, posibilidad de encontrar empleo, estabilidad política, económica y social; entre otros, han sido factores que han elevado los números de habitantes en diversas poblaciones.

4.10 Fuentes históricas y naturales.

El estudio de la población ha alcanzado en los últimos años un gran desarrollo, y ha creado una ciencia propia: la demografía. Es fundamental saber el número de personas que habitan un determinado espacio, así como las principales características de sus habitantes. Para tomar decisiones sobre temas como el número de escuelas, hospitales, etc.

La demografía, es la ciencia que tiene como objetivo el estudio de las poblaciones humanas y que trata de su dimensión, estructura, evolución y características generales, considerados desde un punto de vista cuantitativo. Por tanto, la demografía estudia estadísticamente la estructura y la dinámica de la población y las leyes que rigen estos fenómenos.

Las principales fuentes de datos demográficos son los censos nacionales, el registro civil y, en algunos países, los muestreos a nivel nacional. Estas fuentes proporcionan el material de base para investigar las causas y las consecuencias de los cambios de población. La fuente más habitual es el censo de población, que contabiliza en un cierto momento todas las personas de un área dada, con sus datos personales y características sociales y económicas específicas. En el registro civil se produce la contabilización continua, por parte de las administraciones locales, de los nacimientos, fallecimientos, migraciones, matrimonios y divorcios. Su fiabilidad depende de lo veraces que sean los ciudadanos al proporcionar los datos. En el muestreo se utiliza una selección estadística representativa de la población total.

Se distinguen tres principales fuentes demográficas:

- Estadísticas Vitales
- Censo de población,
- Encuesta

Censos de Población y Vivienda

Los censos de población y vivienda es el proceso de recolección compilación, análisis, evaluación, publicación, y diseminación o difusión de los datos demográficos económicos, sociales, y culturales, pertenecientes a los habitantes de un territorio delimitado, en un momento determinado.

El censo es la principal fuente de información para conocer el tamaño de la población, su distribución geográfica en el territorio y por su contenido sobre las características de los habitantes de un país, constituye la base fundamental de datos del análisis demográfico

El objetivo del censo de población y vivienda es dar a conocer el número de localidades que hay en el territorio, la estructura de edad y sexo de la población, su lugar de nacimiento y residencia, así como su característica en materia educativa, ocupacional, y de servicios de vivienda, 50 social y económica

Características universales censales

Universalidad: Los censos deben incluir a todas las personas en la Área sin omisión ni duplicación.

Simultaneidad: Los datos deben referirse a un momento determinado es decir debe recoger la información al mismo tiempo en todo territorio.

Individualidad: La información debe referirse a unidades individuales, debe plasmar las características personales de cada uno de los ocupantes de la vivienda censada.

Territorio definido: El censo debe referirse a un territorio geográfico con fronteras definidas.

Compilación y publicación: los resultados deben ser compilados y publicados en tiempo razonable, una parte fundamental del trabajo censal es el ordenamiento, la sistematización y la publicación, por área geográfica de las variables demográficas, sociales, culturales y económicas básicas.

Normas internacionales censales: Debido que los datos censales incluyen información estrictamente privada de cada uno e los habitantes del país están protegidos por normas internacionales

Secreto estadístico: Los datos censales constituyen un secreto al que nadie puede acceder.

Auspicio del estado: Independientemente de que la agencia ejecutora sea gubernamental o no, el gobierno es el dueño, responsable y protector de la información recabada.

Periodicidad: Los censos deben llevarse periódicamente, se ha convenido internacionalmente que este periodo sea cada 10 años. esto permite establecer comparaciones internacionales estimar las tendencias de variables demográficas económicas y sociales.

Problemas de levantamiento censal: La complejidad del proceso censal, desde la planeación hasta su publicación implica un camino se presenta problemas que pueden reflejarse en los resultados censales.

4.11 Fenómenos Demográficos.

Se trata de tendencias o indicadores que provienen de la información demográfica; estudios estadísticos de las poblaciones humanas según su estado y distribución en un contexto particular, ya sea su posición geográfica, su evolución histórica o su contexto social.

Ellos nos brindan información clara y precisa sobre las cuestiones relativas a la población.

"Los indicadores demográficos son relaciones estadísticas referidas a algún tema en particular, por ejemplo, la natalidad, la mortalidad o la fecundidad de una población específica (de una ciudad, de una provincia, de un país, de una región, del mundo) y en un momento determinado".

Son importantes para conocer la situación socio-económica de la población. Mientras que las Tasas "son relaciones que se establecen entre un grupo de la población y la población total o entre dos sub-grupos de esa población (por ejemplo, relación entre nacimientos y mujeres en edad fértil)."

Ambos permiten analizar una población a través del tiempo, comparar diferentes poblaciones en el mismo momento histórico o establecer líneas de acción política a partir de su análisis, como puede ser políticas sanitarias, de vivienda, educativas o de empleo. A pesar de que existe una gran cantidad de indicadores demográficos, los más utilizados son las tasas de natalidad, mortalidad, mortalidad infantil, fecundidad, migración e incidencia, el índice de escolaridad, además del PBI (Producto Bruto Interno), el PBI per cápita o por persona, la esperanza de vida y el IDH (Índice de Desarrollo Humano). También se usan otros indicadores del estado socio-económico de una población, como el consumo diario de calorías por habitante.

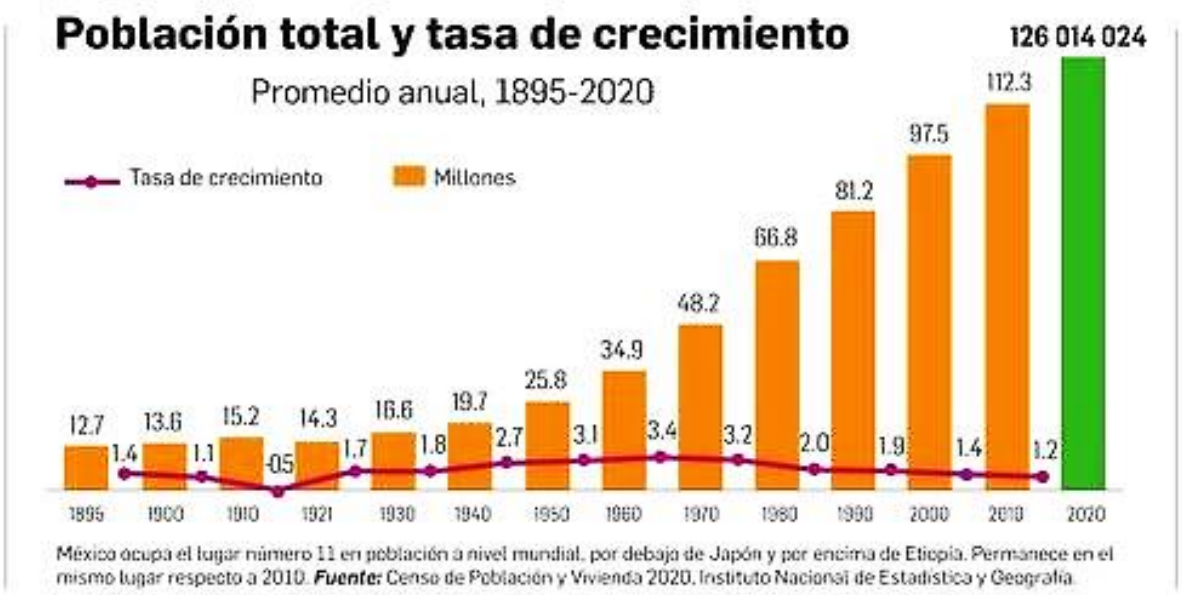
Tasa bruta de natalidad: Es el cociente entre el número de nacimientos ocurridos durante un período determinado, generalmente un año calendario, y la población media del período.

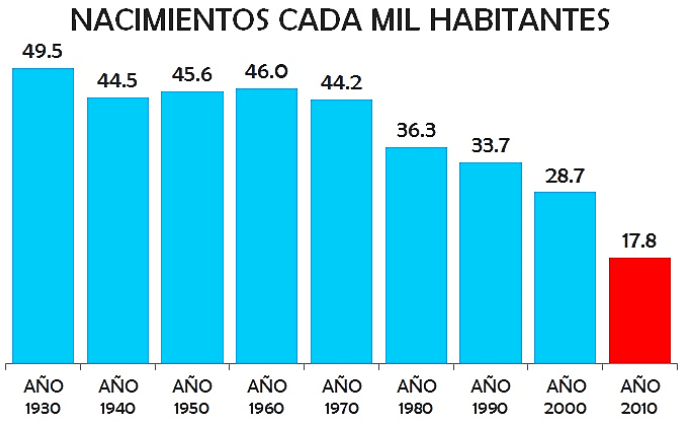
Tasa Global de fecundidad: Es el número de hijos que en promedio tendría cada mujer de una cohorte hipotética de mujeres que durante su vida fértil tuvieran sus hijos de acuerdo a las tasas de fecundidad por edad del período en estudio y no estuvieran expuestas al riesgo de mortalidad desde el nacimiento hasta el término de su período fértil.

Tasa de fecundidad por edad de la madre (por mil): Se define como el total de nacimientos de madres de una determinada edad, a lo largo del año, por cada 1.000 mujeres de dicho colectivo poblacional.

TFG: tasa de fertilidad general: es el cociente entre Nacidos Vivos ocurridos durante un período determinado y la población de mujeres entre 15 y 49 años.

Tasa bruta de reproducción: Se interpreta como el número de hijas promedio que tendría cada mujer de una cohorte hipotética, que cumpliera condiciones similares a las expresadas en la Tasa Global de Fecundidad. Dicha tasa se calcula multiplicando la tasa global de fecundidad por la proporción que representan los nacimientos femeninos respecto al total de nacimientos.





Referencias bibliográficas

Articulo (SD). Distribución Hipergeométrica. 22/05/2021, de Proyecto Descartes Sitio web: https://proyectodescartes.org/iCartesiLibri/materiales_didacticos/EstadisticaProbabilidadInferencia/VAdiscreta/4_1DistribucionHipergeometrica/index.html

Aula Fácil. (2019). Independencia de sucesos. 13/08/2021, de Aula Fácil Sitio web: https://www.aulafacil.com/cursos/estadisticas/gratis/independencia-de-sucesos-l11238

Arrondo, V. (2020). Regresión y correlación. 13/08/2021, de Sites Sitio web: https://www.ugr.es/~jsalinas/apuntes/C5.pdf

Alfaro, M. (2018). Función de distribución empírica. 13/08/2021, de Membres Sitio web: https://membres-ljk.imag.fr/Bernard.Ycart/emel/cours/sd/node6.html

Álvarez, H. (s.f.). Demografía y Fuentes Demográficas. 13/08/2021, de Sites Sitio web: https://sites.google.com/site/geografiaterceranoenm509/demografia-y-fuentes-demograficas-1

Berlín, F. (2018). Función de Distribución Empírica. 13/08/2021, de Humboldt Sitio web: https://wikis.hu-berlin.de/mmint/Basics:_Empirical_Distribution_Function/es

Carrillo, S. (2019). Representación gráfica de datos. 13/08/2021, de Hiru Sitio web: https://www.hiru.eus/es/matematicas/representacion-grafica-de-datos-estadisticos

Conexión Esan. (2015). ¿Cuál es la importancia de la bioestadística? 13/08/2021, de Conexión Esan Sitio web: https://www.esan.edu.pe/apuntes-empresariales/2015/10/auditoria-en-salud-cual-es-la-importancia-de-la-bioestadistica/

Guyatt. (2018). Contrastes de hipótesis. 13/08/2021, de CMAJ Sitio web: http://www.hrc.es/bioest/Introducion_ch.html