

HERRAMIENTAS DE ESTADÍSTICA

Dr. José Antonio Barreno Benítez



Universidad
Internacional
de Valencia



Universidad
Internacional
de Valencia

Este material es de uso exclusivo para los alumnos de la Universidad Internacional de Valencia. No está permitida la reproducción total o parcial de su contenido ni su tratamiento por cualquier método por aquellas personas que no acrediten su relación con la Universidad Internacional de Valencia, sin autorización expresa de la misma.

Edita

Universidad Internacional de Valencia

Herramientas de estadística

6 ECTS

Dr. José Antonio Barreno Benítez

Leyendas

-  Enlace de interés
-  Ejemplo
-  Importante

abc Los términos resaltados a lo largo del contenido en color **naranja** se recogen en el apartado **GLOSARIO**.

CAPÍTULO 1. ESTADÍSTICA DESCRIPTIVA	7
1.1. Medidas de centralización	8
1.1.1. Media	8
1.1.2. Mediana	9
1.1.3. Moda	9
1.2. Medidas de dispersión	10
1.2.1. Varianza	10
1.2.2. Desviación típica	10
1.2.3. Covarianza	11
1.3. Representación de datos	12
1.3.1. Histograma	13
1.3.2. Diagrama de sectores	14
1.3.3. Gráfico temporal	14
1.4. Medidas de forma	15
1.4.1. Coeficiente de asimetría de Fisher	16
1.4.2. Coeficiente de asimetría de Pearson	16
1.4.3. Coeficiente de asimetría de Bowley	17
1.4.4. Curtosis	18
CAPÍTULO 2. INTRODUCCIÓN A LA PROBABILIDAD	22
2.1. Regla de Laplace y definición axiomática de probabilidad	23
2.2. Probabilidad condicionada y teorema de Bayes	26
2.3. Variables aleatorias	28
2.3.1. Variables aleatorias discretas	29
2.3.2. Variables aleatorias continuas	31
2.3.3. Esperanza y varianza de una variable aleatoria	32
2.3.4. Covarianza entre dos variables aleatorias	34
2.4. Distribuciones	34
2.4.1. Funciones de densidad y distribución	35
2.4.2. Ejemplos de distribuciones comúnmente empleadas	37
2.5. Teorema del límite central y ley de grandes números	52

CAPÍTULO 3. INFERENCIA ESTADÍSTICA	54
3.1. Estimación de parámetros	55
3.1.1. Límites de variables aleatorias	57
3.1.2. Método de los momentos	57
3.1.3. Método de la máxima verosimilitud	58
3.2. Intervalos de confianza	59
3.2.1. Intervalos de confianza para la media	60
3.2.2. Intervalos de confianza para la varianza	62
3.2.3. Intervalos de confianza para la diferencia entre dos medias	63
3.2.4. Intervalo de confianza para el cociente entre dos varianzas	65
3.3. Contraste de hipótesis	66
3.3.1. Contraste de una media	69
3.3.2. Contraste de una varianza	71
3.3.3. Contraste de dos medias	72
3.3.4. Contraste de dos varianzas	75
3.3.5. Contraste del parámetro de una Binomial	76
3.3.6. Contraste de dos porcentajes	76
3.3.7. Test chi-cuadrado de bondad de ajuste	77
3.4. Introducción a la estadística bayesiana	79
3.4.1. Inferencia bayesiana	81
CAPÍTULO 4. AJUSTE Y REGRESIÓN	87
4.1. Ajuste lineal	89
4.1.1. Método de mínimos cuadrados	90
4.1.2. Coeficiente de correlación de Pearson	94
4.1.3. Intervalos de confianza	96
4.1.4. Predicciones	99
4.2. Ajustes no lineales	101
4.2.1. Ajuste parabólico	101
4.2.2. Ajuste hiperbólico	104
4.2.3. Ajuste potencial	106
4.2.4. Ajuste exponencial	108
GLOSARIO	112
BIBLIOGRAFÍA	117

The background of the page features a warm-toned photograph of hands working on a laptop. A semi-transparent circular graphic is overlaid on the left side, containing the chapter title. The entire page is decorated with a network of white dots and lines, and several white bar charts of varying heights are scattered across the right side.

Capítulo 1

Estadística descriptiva

Objetivos

- Conocer y trabajar conceptos básicos en la estadística descriptiva.
- Adquirir habilidades para conseguir una tabla de frecuencias y un diagrama de barras y saber interpretarlos.
- Relacionar la información contenida en una tabla con el tipo de variable que se está tabulando.
- Manejar diferentes maneras de presentar información, incluyendo el cálculo de algunos indicadores estadísticos.

Introducción

Los antecedentes históricos de la **estadística** se encuentran en los censos, que consisten en observaciones periódicas y sistemáticas sobre datos de la población para fines de finanzas y guerras, realizados años antes de Cristo.

Actualmente, la estadística no se limita únicamente a la recolección de datos, sino a la recopilación, organización y análisis de estos con un objetivo determinado.

La estadística se suele utilizar con dos significados diferentes:

1. Como colección de datos numéricos. Este significado es el más común de la palabra “estadística”. Se entiende que dichos datos numéricos tienen que estar presentados de manera sistemática y ordenada. Cualquier información numérica no tiene por qué constituir una estadística; para que lo sea, los datos deben ser coherentes, estar establecidos de manera sistemática y seguir un criterio de ordenación.

Existen muchas situaciones en las que se hace necesario organizar una multitud de datos para que se encuentren disponibles y se pueda extraer toda la información; en otras situaciones hay que hacer un resumen de ese conjunto de datos para tomar decisiones.

2. Como ciencia. La estadística como ciencia estudia el comportamiento de los fenómenos de masas. Al igual que todas las ciencias, estudia las características generales de un grupo y obvia las peculiaridades de cada elemento.

Por ejemplo, al estudiar el sexo de los nacimientos, se comienza el estudio tomando un amplio grupo de nacimientos y obteniendo después la proporción de varones. Es muy común encontrarse con fenómenos cuyo resultado es muy complicado de predecir; por ejemplo, no se puede dar una tabla con las personas que fallecerán a cierta edad.

Así pues, el objetivo de la estadística es calcular las regularidades que se producen en los fenómenos de masa.

También utiliza datos que se han obtenido a partir de una pequeña parte del grupo para predecir qué va a suceder con el grupo entero.

La estadística deductiva o descriptiva trata del recuento, clasificación y ordenación de los datos que se obtienen a través de las observaciones. Se construyen tablas y a partir de ellas se representan gráficamente los datos, para así simplificar la complejidad de aquellos que intervienen en la distribución. Así mismo, se calculan parámetros estadísticos que caracterizan la distribución. No se utiliza el cálculo de probabilidades y únicamente se limita a realizar deducciones a partir de los datos y parámetros obtenidos.

1.1. Medidas de centralización

Las medidas de centralización son aquellas que representan los valores centrales o típicos de un conjunto de datos, es decir, las encargadas de medir en torno a qué valores se sitúa un conjunto de datos.

Las medidas de centralización o de tendencia central que más se utilizan son: **media**, mediana y moda.

1.1.1. Media

La medida de centralización más utilizada y conocida es el promedio aritmético o **media aritmética**. Para expresarla se usan las mismas unidades que en los datos originales.

La media se puede describir como el hecho de repartir de forma equitativa la sumatoria de todos los datos entre los individuos que conforman la muestra.

Para calcular una media aritmética, se deben sumar todos los valores de los que se dispone y a continuación dividir esta suma entre el número de valores que se tienen.

Si a todos los valores que se tienen se les denota con la letra x , si se tienen 5 datos, se tendrá x_1, x_2 y así hasta x_5 . Si se tienen n datos, se tendrá x_1, x_2, \dots hasta x_n . De este modo, la fórmula sería:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n}$$

Si se está tratando con los datos en tablas de **frecuencia**, donde los valores de x_1, x_2, \dots, x_k no representan un dato en sí, sino los k posibles valores de los individuos, y las n_1, n_2, \dots, n_k son el número de individuos que toma cada valor; la fórmula sería:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \dots + x_k n_k}{n} = \sum_{i=1}^k \frac{n_i x_i}{n}$$

1.1.2. Mediana

La **mediana** es el valor que queda en el centro de un conjunto de datos si se ordenan dichos valores en orden de magnitud, es decir, de menor a mayor (o de mayor a menor). Dicho de otro modo, es el valor que tiene igual número de datos tanto a su lado derecho como a su izquierdo. Esta medida adquiere mucha relevancia si los valores están muy próximos en torno al mínimo o al máximo.

Puede ocurrir que el número de observaciones sea par; entonces, la mediana es equivalente a la media de los dos valores que se encuentran en el centro.

Por ejemplo, en una muestra donde se tienen los valores 3, 4, 6, 12, la mediana es $(4+6)/2 = 5$.

La mediana también se puede encontrar como el cuantil 50, aunque los **cuantiles** puedan ser de cualquier valor desde 0 hasta 100; de este modo, los deciles son los valores que dejan en la recta real a la izquierda al 10 % (llamado “primer decil”), el 20 % (segundo decil), el 30 %, etc., y así hasta el 100 %. De manera similar ocurre con los percentiles (de 1 % en 1 %) y con los cuantiles (de 25 % en 25 %).

1.1.3. Moda

La **moda** de un conjunto de valores se define como el valor más repetido en ese conjunto. En el caso de que los valores sean continuos, es decir, de que un dato pueda tomar infinitos valores dentro de un determinado rango (por ejemplo, la altura de una persona medida en metros con 3 cifras decimales), la moda no se puede calcular, a no ser que los datos sufran una **discretización**, es decir, que cada dato se transforme en un valor con un intervalo. En esos casos, la moda se hallaría realizando una interpolación dentro del intervalo que tuviese mayor densidad de frecuencia.

Para el caso en el que se estudien variables de tipo continuo, una vez que se identifica el intervalo modal, ciertos analistas toman la marca de clase de ese intervalo como valor para la moda, si bien es muy común hacer un cálculo de la moda mediante la siguiente fórmula:

$$M_0 = L_{i-1} + c \cdot \frac{n_{i+1}}{n_{i-1} + n_{i+1}}$$

Donde c corresponde a la amplitud y $[L_{i-1}, L_i)$ es el intervalo modal.

La moda no implica que sea un valor único: puede darse el caso de que haya varios valores de la variable con la frecuencia más alta. En este caso se dice que se trata de una distribución bimodal, trimodal, etc.

1.2. Medidas de dispersión

Las medidas de centralización en general no aportan suficiente información para una descripción apropiada de los datos, ya que no tienen en cuenta la concentración o **dispersión** de dichos datos; así pues, es necesario hacer uso de diferentes medidas que proporcionen el grado de **variabilidad** de los datos.

Estas medidas son útiles para hacer comparaciones significativas entre grupos de observaciones. Al medir cómo de **dispersos** están los valores de una **variable** respecto de una de sus medidas de posición, lo que se mide es el grado de representatividad que tiene esa medida de posición en el conjunto de los datos que se quieren resumir.

Por tanto, las medidas de dispersión miden cuánto se suelen alejar los datos de los valores centrales.

1.2.1. Varianza

Si se tiene una variable estadística discreta x , la cual tiene una distribución de frecuencias relativas definida por $\{(x_i, f_i)\}$ y media \bar{x} , se definirá la **varianza** de la variable estadística X , y se denotará como $v(x)$ o σ_x^2 , el promedio de los cuadrados de las desviaciones de los valores de la variable a su media, es decir:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Simplificando la ecuación, se obtiene:

$$\sigma_x^2 = \frac{1}{N} \sum_{i=1}^k n_i x_i^2 - \bar{x}^2$$

Cuando se tienen variables estadísticas discretas o continuas en las que los valores están recogidos en modo de intervalos, se determina de manera análoga sin más que cambiar los valores x_i por las marcas de clase que corresponden a cada intervalo de clase.

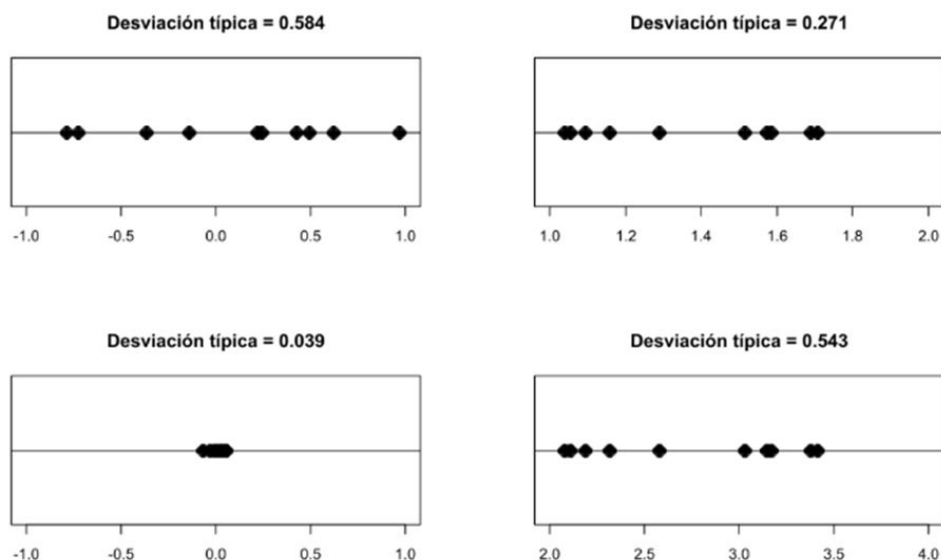
Dicho de otro modo, para calcular la varianza se resta a cada dato la media aritmética, el valor que se obtiene se eleva al cuadrado y se repite para cada dato tantas veces como observaciones o datos se tengan. Una vez realizado, se hace el sumatorio y se efectúa la división entre el número de observaciones.

1.2.2. Desviación típica

A través de la varianza se puede calcular la **desviación típica**, que se representa de forma general como σ . Para calcularla simplemente hay que hacer la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\text{Varianza}}$$

Tanto la desviación típica como la varianza son muy empleadas en estadística, ya que son capaces de simplificar rápidamente la manera en la que se reparten los valores en un conjunto. Lo ideal sería que estas medidas fuesen lo más pequeñas posible (es decir, con datos muy próximos entre sí), para que de ese modo las medidas de centralización describan mucho mejor los datos, que es el objetivo que persigue la estadística descriptiva.



En los dibujos de la izquierda se pueden observar diferencias entre los datos dispersos y los no dispersos; sin embargo, en los dibujos de la derecha los datos están igual de dispersos, pero como desviaciones diferentes. ¿Por qué ocurre esto?

Uno de los problemas que surgen con la desviación típica es su escala; es decir, una variable X puede estar más dispersa que otra variable Y , pero cuando se toman valores más pequeños, la desviación típica es también más pequeña. Por ejemplo, la desviación típica de una muestra de pesos de personas es siempre menor si la tomamos en kg que si lo hacemos en g (aunque la dispersión sea la misma).

Para solucionar este problema, cuando se comparan dos o más variables se emplea el **coeficiente de variación**, que es la división de la desviación típica entre la media aritmética.

Lo que se persigue con el coeficiente de variación es tener una medida normalizada de la dispersión de un conjunto de valores o muestras. A veces, cuando la media es negativa, el coeficiente de variación no es tan informativo como cuando todos los datos son positivos; por ello, en ocasiones, el valor de la media aritmética se expresa como valor absoluto.

1.2.3. Covarianza

La **covarianza** de una muestra bidimensional se define como la media aritmética de los resultados de los productos de las desviaciones de cada variable respecto a sus medias. Es decir:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Otra manera de expresar la covarianza es:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}$$

La covarianza determina el sentido de la correlación entre variables; de este modo, si el valor de la covarianza es positivo, la correlación es directa; y si el valor es negativo, la correlación es inversa.

Por ejemplo:

Las notas de 9 alumnos de una clase en inglés y francés son:

INGLÉS	2	5	8	6	4	4	7	10	5
FRANCÉS	4	4	5	7	9	5	8	8	6

Calcular la covarianza de la distribución.

x_i	y_i	$x_i \cdot y_i$
2	4	8
5	4	20
8	5	40
6	7	42
4	9	36
4	5	20
7	8	56
10	8	80
5	6	30
51	56	332

Después de haber tabulado los datos, hay que calcular las medias aritméticas:

$$\bar{x} = 51/9 = 5.67$$

$$\bar{y} = 56/9 = 6.2$$

$$\text{Covarianza} = (332/9) - 5.67 \cdot 6.2 = 1.74$$

1.3. Representación de datos

Cuando se hacen análisis estadísticos, es muy común usar representaciones visuales de manera complementaria a las tablas de frecuencia. Con dichas representaciones, que se adaptan al objetivo informativo que se persiga, los datos de los análisis se transmiten de manera directa, rápida y entendible para un gran grupo de personas.

Siempre que se muestran datos estadísticos a través de gráficos, debe adaptarse el contenido a la información visual que se quiere transmitir, para lo cual existen multitud de representaciones:

- Diagramas de barras. En ellos se muestra el valor de la **frecuencia absoluta** sobre ejes cartesianos, cuando la variable es **cualitativa** o discreta.
- Histogramas. Son formas características de diagramas de barras que se emplean en distribuciones cuantitativas continuas.
- Polígonos de frecuencias. Se trata de líneas poligonales abiertas sobre un sistema de ejes x e y .
- Gráficos de sectores. En ellos se divide un círculo proporcionalmente según el valor que tengan las frecuencias relativas.

- Pictogramas, también denominados representaciones visuales figurativas. Son como los diagramas de barras, pero en ellos se sustituyen las barras por dibujos que aluden a la variable que se estudia.
- Cartogramas. Se trata de expresiones gráficas a modo de mapa.
- Pirámides de población. Se emplean para clasificar grupos de población por edad y sexo.

1.3.1. Histograma

Los histogramas son una modificación de los diagramas de barras, que se usan en la representación gráfica de series estadísticas de valores en un eje de coordenadas, de modo que en el eje de abscisas se cuantifica el valor de la variable estadística y en el eje de ordenadas se indica su frecuencia absoluta.

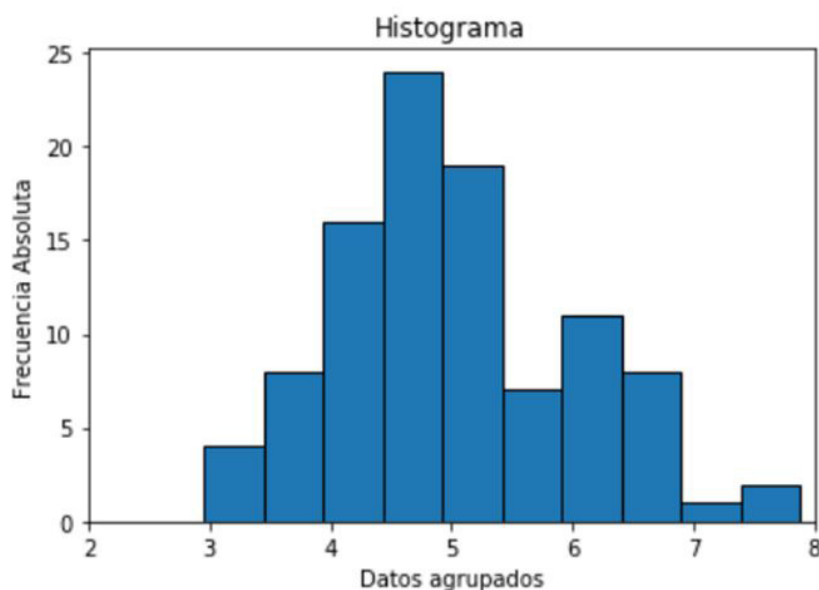
En el caso de los histogramas, se colocan en el eje X los intervalos en que la variable toma sus valores, y sobre cada intervalo se construyen barras rectangulares cuya base será la amplitud del intervalo, y cuya altura representará la frecuencia absoluta de cada intervalo.

En los histogramas, las amplitudes de los intervalos de clase pueden ser de dos tipos:

- Constantes.
- Variables.

Si estas amplitudes de los intervalos son constantes, como la base de los rectángulos que se construyen es siempre de la misma magnitud, puede tomarse esa base como unidad de medida de la variable. Así, las áreas de los rectángulos son equivalentes a las alturas. Cabe destacar que, en este caso, el área total cubierta por la figura es equivalente al número total de observaciones.

Sin embargo, cuando se trata de intervalos con amplitud distinta, las frecuencias no pueden tomarse como alturas. La representación correcta de la gráfica conlleva el cálculo de las alturas. En este caso, la altura se calculará como la frecuencia dividida entre la amplitud del intervalo.



1.3.2. Diagrama de sectores

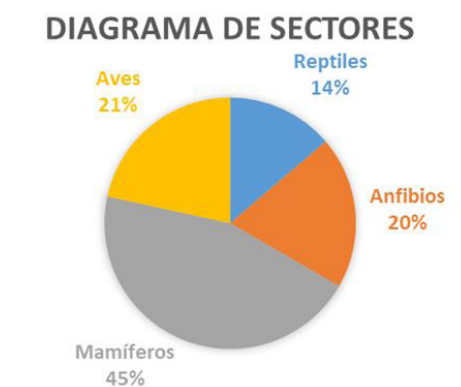
Los diagramas de sectores se emplean para hacer comparaciones entre las diferentes modalidades de un carácter. Para construirlo, simplemente hay que trazar un círculo y asignar a cada modalidad un sector de este cuyo ángulo sea proporcional a su **frecuencia relativa**; para ello, se establece que los 360° del círculo corresponden al sumatorio de todas las frecuencias, y de ahí se saca la proporcionalidad.

Por ejemplo:

En una determinada zona de muestreo en un parque natural, se anota el número de especies distintas encontradas de reptiles, anfibios, mamíferos y aves, de modo que se obtienen los siguientes datos:

- Reptiles: 7
- Anfibios: 10
- Mamíferos: 23
- Aves: 11

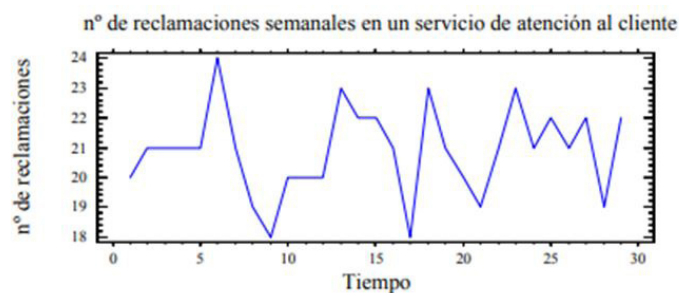
El correspondiente diagrama de sectores sería:



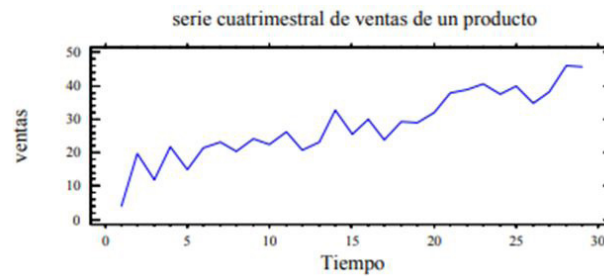
1.3.3. Gráfico temporal

Para construir un gráfico temporal, se sitúan los valores de la serie en el eje Y (o eje de ordenadas) y los instantes temporales en el eje X (o eje de abscisas). Este tipo de gráfico es de gran utilidad para ver el comportamiento de una serie temporal.

- Ejemplo A:



- Ejemplo B:



Las series temporales pueden clasificarse como estacionarias y no estacionarias.

- Series temporales estacionarias.

Se dice que una serie es estacionaria cuando es estable, esto es, cuando la variabilidad y la media son constantes a lo largo del tiempo. Esto se ve de forma gráfica cuando los valores de la serie oscilan alrededor de una media constante y la variabilidad en relación con esa media también es constante en el tiempo. Así pues, se trata de una serie estable a lo largo del tiempo, sin que se aprecien disminuciones ni aumentos sistemáticos de sus valores.

En el ejemplo A se representa una serie estacionaria discreta. La serie es estable alrededor de un valor central: en promedio, se reciben aproximadamente 21 reclamaciones a la semana.

- Series temporales no estacionarias.

Se trata de series en las que la variabilidad y/o la media cambian a lo largo del tiempo. Las variaciones en la media determinan una tendencia a aumentar o a disminuir a largo plazo, así que la serie no oscila alrededor de un valor constante.

El ejemplo B presenta una marcada tendencia creciente, aunque hay grandes oscilaciones con relación a dicha tendencia de crecimiento lineal.

1.4. Medidas de forma

Las medidas de forma son aquellas en las que se muestra si una distribución de frecuencia posee características especiales, como por ejemplo simetría o asimetría, nivel de apuntamiento o de concentración de datos, que la clasificarán en un modo característico de distribución.

Las medidas de asimetría son indicadores que permiten establecer el nivel de asimetría (o simetría) que tienen los datos de una distribución sin necesidad de hacer una representación gráfica de ellos. Para medir la simetría de una distribución de frecuencias se tomará como eje de asimetría la recta paralela al eje Y (o eje de ordenadas) que pasa por la media aritmética de la muestra.

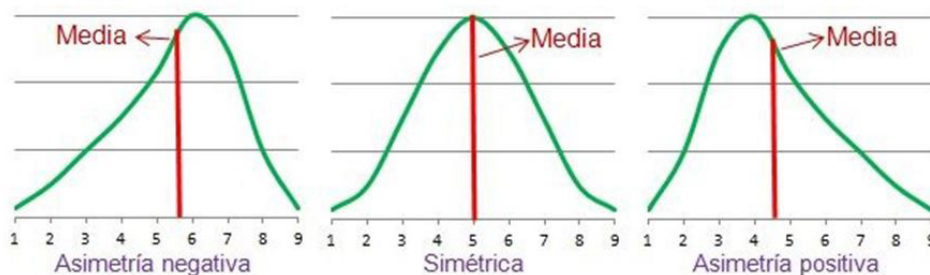
Cuando la distribución es simétrica, se puede decir que hay igual número de valores a la izquierda que a la derecha de la media aritmética, es decir, igual número de desviaciones con signo negativo que con signo positivo. Se necesita, pues, calcular las distancias entre la media aritmética y los datos, pero conservando los signos, y de ese modo comprobar si las distancias entre los valores inferiores y la media coinciden con las distancias entre los valores superiores y la media. Así, el momento de orden 3 respecto a la media aritmética constituye una medida de asimetría, es decir:

$$m_3 = \frac{\sum_{i=1}^N (x_i - \bar{x})^3 \cdot n_i}{N}$$

Si $m_3 = 0$, la distribución es simétrica.

Si $m_3 > 0$, la distribución es a derechas o asimétrica positiva.

Si $m_3 < 0$, la distribución es a izquierdas o asimétrica negativa.



Las unidades de m_3 son iguales que las de la variable, pero elevadas al cubo, y se obtienen a partir de los momentos con respecto al origen usando la siguiente ecuación:

$$m_3 = a_3 - 3a_2a_1 + 2a_1^3$$

1.4.1. Coeficiente de asimetría de Fisher

El coeficiente de asimetría de Fisher se basa en cómo se desvían los valores que se observan respecto de la media aritmética. La interpretación de los resultados que proporciona este coeficiente es equivalente a la del primer coeficiente de Pearson.

La fórmula para calcularlo es:

$$A_s = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{ns^3}$$

En el caso de que los datos estén tabulados, la fórmula sería:

$$A_s = \frac{\sum_{i=1}^k (y_i - \bar{x})^3 n_i}{ns^3}$$

La interpretación que se le da al coeficiente de asimetría de Fisher es que los valores que son negativos, es decir, menores que 0, indican asimetría negativa; y al contrario, si los valores son mayores que 0, mostrarán asimetría positiva. En cambio, si los valores son 0 o muy próximos a 0, presentan simetría.

1.4.2. Coeficiente de asimetría de Pearson

Cuando la gráfica de la distribución tiene forma de campana, es decir, cuando es campaniforme y moderadamente asimétrica, se utiliza el coeficiente de asimetría de Pearson para analizarla. Este coeficiente se define como el resultado de dividir la diferencia entre la media y la moda entre la desviación típica:

$$A_p = \frac{\bar{X} - M_o}{S}$$

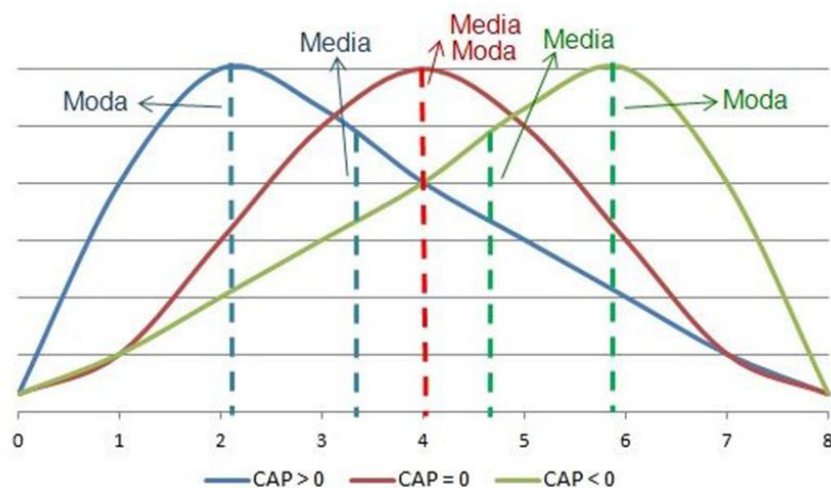
Según datos empíricos, se puede hacer una aproximación de esta medida con la siguiente expresión:

$$A_p \approx \frac{3(\bar{X} - M_o)}{S}$$

El coeficiente de asimetría de Pearson se basa en que si una distribución en forma de campana de Gauss es simétrica, se debe cumplir que su moda, mediana y media son coincidentes; y en este caso, A_p se hará 0.

En el caso de que la distribución sea asimétrica positiva, la media estará por encima de la moda y A_p será mayor que 0 (positiva); sin embargo, cuando la distribución sea asimétrica negativa, ocurrirá todo lo contrario.

Este procedimiento únicamente se empleará en distribuciones poco asimétricas y unimodales.



1.4.3. Coeficiente de asimetría de Bowley

El coeficiente de asimetría de Bowley, CA_B , toma los cuartiles como referencia para así averiguar si la distribución es o no simétrica. En la explicación de este coeficiente, hay que suponer que el modo en que se comporta la distribución en los extremos es análogo. Siendo el conjunto $X = (x_1, x_2, \dots, x_N)$, el coeficiente de asimetría de Bowley se define así:

$$CA_B = \frac{Q_3 + Q_1 - 2M_e(X)}{Q_3 - Q_1}$$

Siendo Q_3 el tercer cuartil, Q_1 el primer cuartil y $M_e(X)$ la mediana.

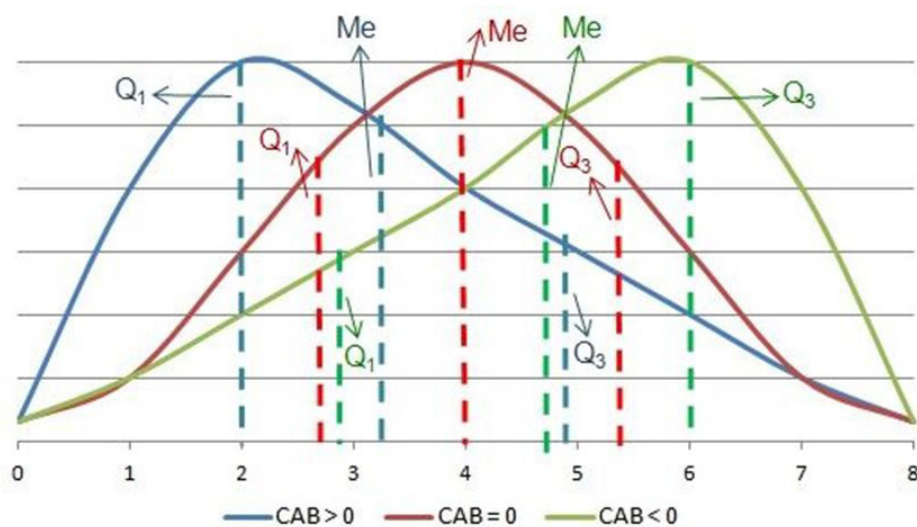
Esta fórmula tiene el siguiente origen:

$$CA_B = \frac{(Q_3 - M_e(X)) - (M_e(X) - Q_1)}{Q_3 - Q_1}$$

Hay que recordar que la M_e (mediana) es equivalente a Q_2 (segundo cuartil), por lo que el coeficiente de asimetría de Bowley se podría escribir también así:

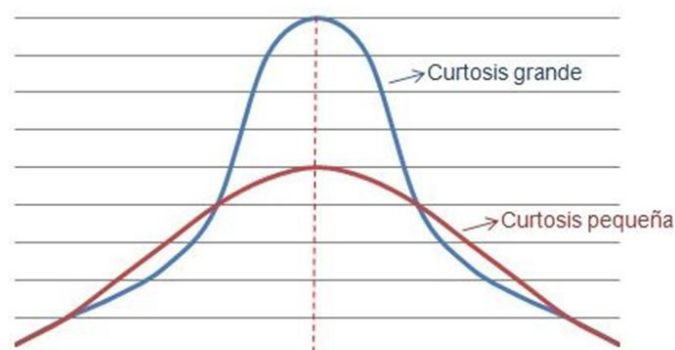
$$CA_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

- La distribución tiene una asimetría negativa si $CA_B < 0$, ya que la separación entre la mediana y el primer cuartil (Q_1) es mayor que entre aquella y el tercer cuartil (Q_3).
- La distribución es simétrica si $CA_B = 0$, ya que la mediana se encuentra a la misma distancia del primer cuartil que del tercero.
- La distribución tiene una asimetría positiva si $CA_B > 0$, ya que la separación entre la mediana y el tercer cuartil es mayor que entre aquella y el primero.



1.4.4. Curtosis

El apuntamiento o curtosis es una medida de forma que indica cuán achatada o escarpada está una distribución o curva.

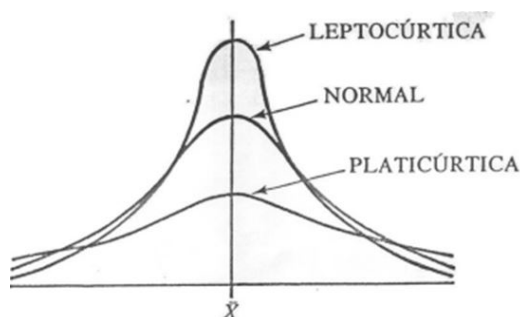


En la simetría existe un referente natural que en la curtosis de una distribución de frecuencias no se tiene, sino que se sustenta en la comparación con una distribución de referencia, más concretamente, en la distribución normal o campana de Gauss. Como consecuencia, su obtención únicamente tendrá sentido en las variables cuya distribución de frecuencias sea parecida a la de la curva normal.

El apuntamiento manifiesta el grado en que una distribución acumula casos en sus colas en comparación con aquellos que se acumulan en las colas de una distribución normal en las que la dispersión es equivalente. Como ocurre con la asimetría, se distinguen 3 categorías de apuntamiento:

- Distribución platicúrtica. Señala que en las colas hay más casos acumulados que en las colas de una distribución normal. Apuntamiento negativo.
- Distribución leptocúrtica. Indica que en las colas hay un menor número de casos acumulados que en las colas de una distribución normal (lo contrario a la distribución platicúrtica). Apuntamiento positivo.
- Distribución mesocúrtica. Ocurre lo mismo que en la distribución normal. Apuntamiento normal.

La representación gráfica de las distribuciones sería la siguiente:



La curtosis o apuntamiento se mide haciendo un promedio de la cuarta potencia de la diferencia entre cada elemento del conjunto y la media, dividido entre la desviación típica elevada a la cuarta potencia.

El coeficiente de curtosis para un conjunto $X = (x_1, x_2, \dots, x_N)$ se define así:

$$g_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 \cdot n_i}{NS_x^4}$$

Donde n_i es la frecuencia absoluta de x_i .

Como se trata de la curtosis de una distribución normal, en la fórmula hay que restar 3; de este modo, la curtosis tendrá un valor de 0 para la normal, que será la que se tome como referencia.

Cuando los datos están agrupados, o formando intervalos, la fórmula del coeficiente de curtosis se transforma del siguiente modo:

$$\text{Exceso de curtosis} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 \cdot n_i}{NS_x^4} - 3$$

Resumen

Siempre que se estudia de forma cuantitativa una población, ya sea de personas en una determinada ciudad, alumnos mayores de 16 años con nacionalidad española o bolas de cierto color dentro de una bolsa, se debe aportar una descripción de lo que se pretende estudiar. Para que esta sea concreta, es necesario determinar un modo de resumir toda la información que se nos presenta; es decir, sobre los elementos de la población (personas, alumnos mayores de 16 años, bolas de cierto color, etc.), que se denominan “individuos”, hay que diferenciar los rasgos que se quieren estudiar (**probabilidad** del 0 al 10 de que la persona vaya a votar, aprobados en un examen de matemáticas, existencia de algún defecto en la bola roja, etc.), algo que los estadísticos denominan “variables”, y obtener de este modo el valor que dichas variables adquieren para cada individuo (probabilidad de 7.5, 18 alumnos aprobados, bolas con rugosidades). Así, la información que se obtiene en la medición de variables se resume a través de la estadística descriptiva.

Términos como “media”, “desviación típica”, “varianza”, etc., se refieren a medidas que pueden ser de dispersión o de centralización. En esta ocasión, hemos estudiado qué significado tienen realmente y cómo de importantes pueden llegar a ser.

Aunque se pueden estudiar datos sobre cualquier cosa, distinguimos entre datos numéricos y datos no numéricos.

Los datos numéricos son aquellos que representan una magnitud que puede ser medida de forma numérica, como por ejemplo la altura en centímetros o el peso en kilogramos. De este modo, los datos numéricos pueden ser discretos (cuando se pueden enumerar los distintos valores de una variable) o continuos (cuando no se pueden enumerar):

- Número de días lluviosos al año: discreto (porque se sabe que el valor será un número de la secuencia 1, 2, 3..., 365).
- Número de goles que marca un jugador de fútbol en una temporada: discreto (aunque no se sepa dónde acaba la secuencia porque no existe un máximo de goles que puede marcar un jugador, se sabe que será cero o un número natural: 0, 1, 2, 3, etc.).
- La proporción de bolas rojas en una bolsa con tres bolas: discreto (se sabe que será uno de estos tres valores: 0.333, 0.666 y 1).
- La altura de un adulto en centímetros: continuo (una persona nunca mide 168 cm justos, ya que la altura siempre será un valor con infinitos decimales: 168,44958, o bien 168,84295...; que se redondee siempre a una cifra sin decimales es una cuestión de aproximación; es decir, existen infinitos valores posibles).
- Tiempo en minutos que se tarda en desplazarse desde un punto A hasta un punto B: continuo (un periodo de tiempo no puede ser un número exacto, por lo que habrá que usar una aproximación numérica para poder representarlo).

Los datos no numéricos son los que representan un atributo, una cualidad de algún elemento que no se puede medir con números; por ejemplo, si se coloca una fruta en un plato y se pregunta de qué fruta se trata, no se puede responder con un número, y si se hiciera (por haber asignado a cada fruta un valor numérico), dicho número no representaría ninguna magnitud.

Se distingue entre datos no numéricos ordinales, cuyos valores posibles son categorías en las que hay cierta jerarquía, como por ejemplo malo < bueno; y nominales, entre los que esa jerarquía no existe.

Todos estos datos suelen ser representados en lo que se denomina “tablas de frecuencias”.

Si se tiene un conjunto de k valores diferentes, la tabla de frecuencias de ese conjunto sería:

x	n	f	N
x_1	n_1	f_1	N_1
x_2	n_2	f_2	N_2
...
x_i	n_i	f_i	N_i
...
x_n	n_n	f_n	N_n
Suma	N	1	

Puede verse que:

- x_i son los valores que se dan en el conjunto de datos.
- n_i son las frecuencias absolutas, que representan el número de veces que se repite cada valor en el conjunto. La suma es el total de datos que se tienen en el conjunto.
- f_i son las frecuencias relativas; se obtienen al dividir la frecuencia absoluta de cada valor entre N (que es la suma de las frecuencias absolutas).
- N_i son las frecuencias acumuladas; se obtienen al sumar, por cada valor i , las frecuencias relativas de ese y de todos los valores que vienen detrás: $N_i = n_1 + n_2 + \dots + n_i$.



Capítulo 2

Introducción a la probabilidad

Objetivos

- Entender el concepto de experimento aleatorio.
- Valorar la probabilidad y sus aplicaciones.
- Calcular probabilidades.
- Comprender el concepto de probabilidad condicionada y aplicar con soltura los teoremas de la probabilidad total y de Bayes.

Introducción

Si se quieren extender los resultados que se obtienen al hacer un estudio descriptivo de las distintas variables estadísticas a poblaciones que no son observadas completamente, se hace necesario utilizar el concepto de modelo probabilístico.

De este modo, se introduce, en primera instancia, el concepto de probabilidad como idealización del concepto de frecuencia relativa.

Pero trabajar con los resultados cualitativos de un **experimento aleatorio** introduce ciertas complicaciones, por lo que es de gran utilidad cuantificar los resultados cualitativos del experimento aleatorio, o lo que es lo mismo, asignar un valor numérico a cada suceso del espacio muestral correspondiente al experimento aleatorio considerado. Esta relación entre los sucesos del espacio muestral y el valor numérico que se les asigna la establecemos mediante la **variable aleatoria**.

Al hacer un cálculo de probabilidades, se estudia el grado de error o fiabilidad de las conclusiones que se obtienen mediante inferencia estadística.

La probabilidad cuantifica o mide el grado de incertidumbre que se tiene sobre el resultado de un experimento aleatorio.

2.1. Regla de Laplace y definición axiomática de probabilidad

A lo largo del siglo XX, Andrei Kolmogorov (matemático de Rusia) definió por primera vez el concepto de probabilidad, que actualmente se sigue usando.

Si se realiza un experimento determinado, el cual está definido por un espacio muestral Ω , se definirá la probabilidad como la función que relaciona a cada uno de los sucesos A una probabilidad determinada $P(A)$, la cual cumple los siguientes axiomas:

1. La probabilidad de cualquier suceso A es 0 o positiva. Dicho de otro modo, $P(A) \geq 0$. En este caso, la probabilidad mide cuán difícil es que ocurra el suceso A ; de este modo, cuanto menor sea la probabilidad, más complicado es que suceda.
2. Un suceso seguro tiene una probabilidad de 1. Dicho de otro modo, $P(\Omega) = 1$. En este caso, la probabilidad siempre será mayor que 0 pero menor que 1. Una probabilidad de 0 significa que no existe ninguna probabilidad de que ocurra (es decir, el suceso es imposible) y, por el contrario, una probabilidad de 1 significa que el suceso siempre ocurre (suceso seguro).
3. La probabilidad de la unión de un conjunto cualquiera de sucesos incompatibles se define como el sumatorio de las probabilidades de cada suceso por separado. Así pues, si se tienen los sucesos A , B , C , y se trata de sucesos incompatibles dos a dos, la probabilidad será:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

Hay que aclarar que en matemáticas, un axioma se define como el resultado que será aceptado sin que sea necesaria una demostración. Esto es lo que se conoce comúnmente como definición axiomática de la probabilidad, ya que esta se define como una función que cumplirá los 3 axiomas que se han descrito.

Regla de Laplace

La regla de Laplace es muy importante, ya que permite hacer un cálculo de la probabilidad de un suceso, siempre y cuando los sucesos elementales sean equiprobables, o lo que es lo mismo, que todos los resultados que se obtengan tengan una probabilidad igual.

En estas condiciones, la probabilidad de cualquier suceso A se obtendrá al dividir el número de resultados que forman el suceso A entre el número total de resultados posibles.

Si se determina que los sucesos de A son los casos favorables, entonces se podrá escribir la regla de Laplace de este modo:

$$P(A) = \frac{\text{casos favorables a A}}{\text{casos posibles}}$$

No obstante, como esta regla únicamente funciona si todos los casos son equiprobables, no se puede hacer un razonamiento como el siguiente:

Se quiere calcular la probabilidad del suceso A (ser arrollado por un tren). Únicamente hay 2 posibles casos: ser arrollado y no serlo. Si suponemos que el caso favorable es "ser arrollado", la probabilidad es: $P(A) = \frac{1}{2}$.

Esto significa que cada vez que se saliese a la calle, se tendría el 50 % de probabilidad de ser arrollado. En este caso, el error radica en que ambos sucesos no tienen igual probabilidad, lo cual hace pensar que si se desea solucionar un problema con la regla de Laplace, es preciso tener cuidado si se quiere considerar que los posibles resultados están desordenados, ya que podría darse el caso de que no todos fuesen igual de probables.

Un ejemplo no válido sería:

Si una familia tiene 2 hijos, y se supone que la probabilidad de ser mujer es la misma que la de ser hombre, ¿cuál es la probabilidad de que los dos hijos sean del mismo sexo?

Se podría argumentar que nuestro espacio muestral sería:

$$\Omega = \{(M, M), (H, H), (H, M)\}$$

Es decir, "dos mujeres", "dos hombres" y "un hombre y una mujer"; entonces, los casos favorables (mismo sexo) serían 2 y los casos posibles serían 3, por lo que la probabilidad sería $\frac{2}{3}$.

En este caso, el error está en que los tres casos no son equiprobables, ya que para que cada hijo sea de un sexo, puede ocurrir que el primer hijo sea una mujer y el segundo un hombre, o que el primero sea un hombre y el segundo una mujer.

En conclusión, cuando se quiera aplicar la regla de Laplace, será necesario considerar el espacio muestral ordenado para así evitar errores.

Para aplicar correctamente esta regla en un ejemplo similar al anterior:

Si una familia tiene 2 hijos, y se supone que la probabilidad de ser mujer es la misma que la de ser hombre, ¿cuál es la probabilidad de que los dos hijos sean del mismo sexo?

En este caso, el espacio muestral es:

$$\Omega = \{(M, M), (M, H), (H, M), (H, H)\}$$

Los casos favorables son (H, H) y (M, M), con lo que la probabilidad será de $\frac{2}{4} = \frac{1}{2}$; es decir, de 0.5 (o del 50 %).

Principales propiedades de la probabilidad:

- $P(A) + P(\bar{A}) = 1$

Esto quiere decir que la suma de las probabilidades de los sucesos que son complementarios es igual a 1. Esta propiedad se utiliza mucho para averiguar la probabilidad del suceso complementario:

$$P(\bar{A}) = 1 - P(A)$$

Esto es así porque, en primer lugar, A y \bar{A} son incompatibles y, por otra parte, $A \cup \bar{A} = \Omega$, ya que lo uno es lo contrapuesto de lo otro. Esta es otra forma de explicar lo que ya se expuso anteriormente: que el suceso $A \cup \bar{A}$ es un suceso seguro, por lo que el segundo axioma, $P(A \cup \bar{A}) = 1$, o lo que es lo mismo, que ocurrirá siempre. Así pues, por el axioma 3, se tendrá:

$$P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

$$P(A \cup \bar{A}) = P(\Omega) = 1$$

Por lo tanto, se puede deducir que:

$$P(A) + P(\bar{A}) = 1$$

Esta propiedad, que es bastante útil, se puede generalizar de la siguiente manera: si se tienen 3 o más sucesos que sean incompatibles dos a dos, y tales que su unión sea el espacio muestral completo, o lo que es lo mismo, A, B, C incompatibles dos a dos, de manera que:

$$A \cup B \cup C = \Omega$$

De esa forma, por los axiomas 2 y 3 se cumple que:

$$P(A) + P(B) + P(C) = 1$$

En este caso se puede decir que A, B y C formarán un sistema completo de sucesos. Cabe destacar que siempre que se exprese Ω como un conjunto de sucesos elementales, lo que se está dando en realidad es un sistema completo de sucesos.

La consecuencia de esto es que $P(\emptyset) = 0$, lo cual quiere decir que la probabilidad de que un suceso sea imposible es nula, ya que, como se sabe que el suceso contrario al **suceso imposible** será un suceso seguro, de esa manera se puede cambiar esto en la igualdad de la propiedad $P(\emptyset) + P(\Omega) = 1$. Por tanto, por el axioma 2 de la probabilidad $P(\Omega) = 1$, se tendrá que $P(\emptyset) + 1 = 1$; luego $P(\emptyset) = 0$.

- Si $A \subset B$, se cumple que $P(A) \leq P(B)$.

El término “si $A \subset B$ ” se lee así: “si el suceso A se encuentra dentro del suceso B ”, o lo que es lo mismo, si todos los posibles resultados que se cumplen en el suceso A también se cumplen en el B .

Dicha propiedad es muy lógica y se entiende muy bien con el ejemplo de lanzar un dado: si se quiere enfrentar la probabilidad de A (que salga un 3) con B (que salga impar), entonces la probabilidad de A tiene que ser menor o igual que la de B , ya que si sale un 3, también sale un número impar. Es decir, al cumplirse A , también se está cumpliendo B ; por lo tanto, que se diese A debería ser más complicado que se cumpliera B . Es decir:

$$P(A) \leq P(B)$$

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

Esta ecuación se da como consecuencia de una propiedad de la teoría de los conjuntos, que dice que cuando se tienen dos conjuntos A y B, la unión de ambos se expresa así:

$$A \cup B = (A - B) \cup (A \cap B) \cup (B - A)$$

Que son incompatibles dos a dos y, como consecuencia, por el tercer axioma, se tiene:

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

Según la teoría de conjuntos, $A = (A - B) \cup (A \cap B)$, que se trata de dos **sucesos** incompatibles y, por tanto, por el tercer axioma: $P(A) = P(A - B) + P(A \cap B)$, es decir:

$$P(A - B) = P(A) - P(A \cap B).$$

De forma análoga, se tiene que:

$$B = (B - A) \cup (B \cap A) = (B - A) \cup (A \cap B)$$

Por lo que:

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

Si se sustituyen estas propiedades en la igualdad, se obtiene:

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

2.2. Probabilidad condicionada y teorema de Bayes

Cuando se supone que dos sucesos están relacionados de modo que la probabilidad de que ocurra uno de ellos depende de que el otro haya ocurrido o no, estas serán las condiciones en las que se definirá la **probabilidad condicionada** del suceso A dado que este ocurre, denotado como $P(A | B)$, así:

$$P(A | B) = P(A \cap B) / P(B) \quad P(B) > 0$$

Únicamente se puede condicionar a sucesos con una probabilidad diferente de 0.

Si se tienen 2 sucesos A y B que sean independientes, entonces:

$$P(A | B) = P(A) \text{ y } P(B | A) = P(B)$$

Así pues, se tiene:

$$P(A | B) \geq 0$$

$$P(E | B) = 1$$

Si A_1, A_2, \dots son tales que $A_i \cap A_j = \emptyset$ si $i \neq j$, entonces:

$$P(\cup_{i=1, \infty} A_i | B) = \sum_{i=1, \infty} P(A_i | B)$$

Teorema de la probabilidad compuesta

Si se tienen n sucesos A_1, A_2, \dots, A_n con $P(A_i) > 0$ para $i = 1, 2, \dots, n$, se cumple que:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) P(A_2 | A_1) \dots P(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

Si se tienen **sucesos independientes**:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) P(A_2) \dots P(A_n)$$

Ejemplo:

Un experimento trata de hacer observaciones meteorológicas en un determinado día. Sea A el suceso observar lluvia y B el suceso observar cielo nublado.

A y B se encuentran relacionados. La probabilidad $P(A)$ de lluvia no es la misma que la probabilidad de lluvia dado que el cielo está nublado. Supongamos que un 10 % de todos los días son nublados y lluviosos y un 30 % de todos los días son nublados.

$$P(A | B) = 0,1/0,3 = 1/3$$

Entonces, la probabilidad de que se observe lluvia dado que se observa cielo nublado es de $1/3$.

Teorema de la probabilidad total

Sea E un espacio muestral de un experimento; se consideran n sucesos A_1, A_2, \dots, A_n de E de modo que sean **disjuntos** y que su unión sea el espacio muestral. Se dice entonces que dichos sucesos constituyen una partición el espacio E .

Si A_1, A_2, \dots, A_n son tales que $A_i \cap A_j = \emptyset$ si $i \neq j$, y además $U_i = 1, \dots, n$ $A_i = E$, entonces la probabilidad del suceso B vendrá dada por lo siguiente:

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

Teorema de Bayes

El teorema de Bayes proporciona una manera sencilla para el cálculo de la probabilidad condicional de cada suceso A_i dado B a partir de las probabilidades condicionales de B dado uno de los sucesos A_i y la probabilidad que tendrá cada A_i .

Si se supone que los sucesos A_1, A_2, \dots, A_k constituyen una partición del espacio E tal que la probabilidad $P(A_j) > 0$ para $j = 1, 2, \dots, k$ y que sea B cualquier suceso tal que $P(B) > 0$, entonces para $i = 1, 2, \dots, k$:

$$P(A_i | B) = \frac{P(A_i) P(B | A_i)}{\sum_{j=1}^k P(A_j) P(B | A_j)}$$

Ejemplo:

Para fabricar un lote grande de artículos semejantes se han utilizado 3 máquinas: M_1 , M_2 , y M_3 . Se supone que el 20 % de los productos han sido construidos por la máquina M_1 , el 30 % por la máquina M_2 y el 50 % por la máquina M_3 .

Los productos defectuosos que ha fabricado cada máquina han sido:

1 % para la máquina M_1 .

2 % para la máquina M_2 .

3 % para la máquina M_3 .

Por último, se supone que se elige al azar uno de los productos fabricados y se observa que es defectuoso. Se va a determinar la probabilidad de que ese producto haya sido elaborado por la máquina M_2 .

Solución:

Sea A_i el suceso de que el artículo que se selecciona haya sido fabricado por la máquina M_i ($i=1, 2, 3$) y sea B el suceso de que el artículo que se selecciona sea defectuoso. Hay que hallar la probabilidad condicional $P(A_2/B)$.

La probabilidad $P(A_i)$ de que el artículo seleccionado al azar lo haya producido la máquina M_i es:

$$P(A_1) = 0.2$$

$$P(A_2) = 0.3$$

$$P(A_3) = 0.5$$

Además, la probabilidad $P(B/A_i)$ de que un artículo producido por la máquina M_i sea defectuoso es:

$$P(B/A_1) = 0.01$$

$$P(B/A_2) = 0.02$$

$$P(B/A_3) = 0.03$$

Del teorema de Bayes resulta que:

$$P(A_2 | B) = \frac{P(A_2) P(B | A_2)}{\sum_{j=1}^3 P(A_j) P(B | A_j)} = \frac{0.3 \cdot 0.02}{0.2 \cdot 0.01 + 0.3 \cdot 0.02 + 0.5 \cdot 0.03} = 0.26$$

2.3. Variables aleatorias

En la estadística descriptiva, el estudio se centraba en la información que correspondía a una variable; se obtenía su distribución de frecuencias y un conjunto de medidas descriptivas tales como la media, la varianza, la mediana, la moda, etc., las cuales englobaban toda la información sobre la variable que se estaba estudiando, de manera que se podía comparar entre diferentes poblaciones empleando solo sus características descriptivas.

En el ámbito de la probabilidad se da una situación similar, pues toda la información de una variable aleatoria llamada X estará implícita en su función de probabilidad $P(x)$ o en su **función de densidad** $f(x)$, según se trate de una variable aleatoria discreta o continua, respectivamente. Puede darse el caso de tener que comparar entre dos o más distribuciones de probabilidad, lo cual será más sencillo de hacer si se comparan los valores característicos de las variables aleatorias correspondientes a esas distribuciones, que si se comparan directamente las diferentes distribuciones de probabilidad o las funciones de densidad correspondientes.

Una variable es aleatoria (v. a.) cuando el hecho de que tome un valor determinado está condicionado por el azar, como por ejemplo el número de miembros de un grupo familiar o la altura de una persona. Si se considera un experimento en el que el espacio muestral es el conjunto E , una variable aleatoria, llámese X , será una función que otorga un número real $X(s)$ a cada resultado posible $s \in E$.

De forma intuitiva, una variable aleatoria es una cantidad o medida que cambia con relación al resultado concreto ei que se observa cuando se realiza un experimento aleatorio.

2.3.1. Variables aleatorias discretas

Se dice que una variable aleatoria X es discreta cuando el conjunto de los valores que puede tomar son finitos o numerables. En este caso, la **distribución de probabilidad** queda totalmente especificada por la función de probabilidad $P(X = k)$, siendo k cualquier valor que pueda tomar la variable. Evidentemente, se tiene que:

$$\sum_k P(X = k) = 1$$

Función de probabilidad

Sea X una variable aleatoria discreta que tomará valores en un conjunto $S = \{x_1, x_2, \dots\}$ finito o infinito numerable, con probabilidades $p_1 = P(X = x_1)$, $p_2 = P(X = x_2)$...

La función de probabilidad de X o función de masa de X se definirá así:

$$P(X = x) = \begin{cases} p_i, & \text{si } x = x_i \in S \\ 0, & \text{si } x \notin S \end{cases}$$

$$P(X = x) = \begin{cases} p_i, & \text{si } x = x_i \in S, \\ 0, & \text{si } x \notin S. \end{cases}$$

Por ejemplo:

Variable discreta con un número finito de valores: se considera un experimento aleatorio que consiste en lanzar 3 veces una moneda; se definirá la variable aleatoria X = número de caras. En este experimento, el espacio muestral que se tiene es:

$$E = \{ccc, xxx, ccx, cxc, xcc, cxx, xc x, xxc\}$$

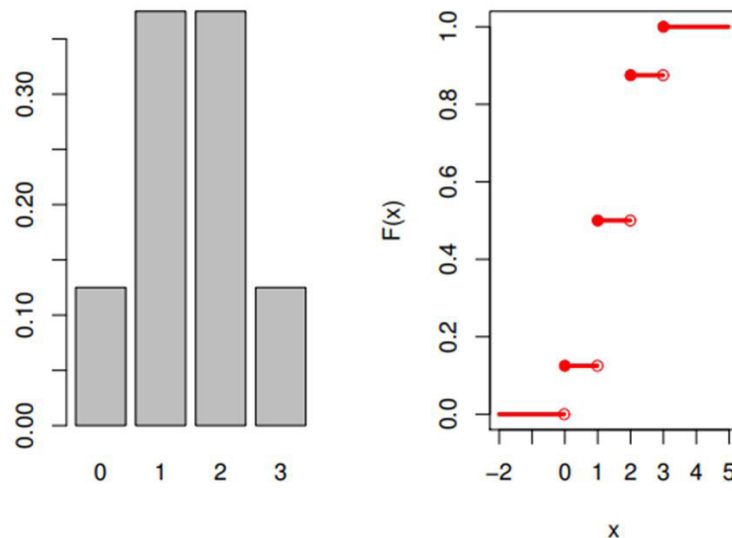
Los valores posibles de X en este experimento son $k = 0, 1, 2, 3$. Para cada k la probabilidad $P(X = k) = P(\{w \in E : X(w) = k\})$ puede obtenerse de manera fácil empleando la regla de Laplace, y se resume en la siguiente tabla:

k	0	1	2	3
$P(X = k)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

La **función de distribución** acumulativa de esta variable aleatoria es:

$$F_x(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{8} & 0 \leq x < 1 \\ \frac{4}{8} & 1 \leq x < 2 \\ \frac{7}{8} & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

Una representación gráfica de las funciones de probabilidad y de distribución acumulativa de dicha variable aleatoria es:



Propiedades de la función de probabilidad:

Si se tiene una variable aleatoria X discreta que tomará valores en el conjunto $S = \{x_1, x_2, \dots\}$ con probabilidades $p_1 = P(X = x_1)$, $p_2 = P(X = x_2)$, ...:

- $0 \leq P(X = x) \leq 1$, para todo $x \in \mathbb{R}$.
- $\sum_{x \in S} P(X = x) = \sum_i P(X = x_i) = \sum_i p_i = 1$
- $P(X \in A) = \sum_{x \in A} P(X = x)$
- $P(X \leq x) = P(X \in (-\infty, x]) = \sum_{i, x_i \leq x} P(X = x_i) = \sum_{i, x_i \leq x} p_i$
- $P(X > x) = 1 - P(X \leq x)$

Función de distribución

La función de probabilidad acumulada o función de distribución de una variable aleatoria X será una aplicación $F: \mathbb{R} \rightarrow [0, 1]$ en la que a cada valor de $x \in \mathbb{R}$ se le asigna la probabilidad $F(x) = P(X \leq x) = P(X \in (-\infty, x])$.

Propiedades de la función de distribución:

- $0 \leq F(x) \leq 1$, para todo $x \in \mathbb{R}$.
- $F(y) = 0$ para todo $y < \min S$. Por tanto, $F(-\infty) = 0$.

- $F(y) = 1$ para todo $y \geq \max S$. Por tanto, $F(\infty) = 1$.
- Si $x_1 \leq x_2$, entonces $F(x_1) \leq F(x_2)$, es decir, $F(x)$ es no decreciente.
- Para todo $a, b \in \mathbb{R}$.

$$P(a < X \leq b) = P(X \in (a, b]) = P(X \in (-\infty, b]) - P(X \in (-\infty, a]) = F(b) - F(a).$$

2.3.2. Variables aleatorias continuas

Las variables aleatorias continuas son aquellas cuya función de distribución acumulativa es continua. Se caracterizan por tomar valores en un intervalo o rango continuo sin que existan puntos en los que se acumule la probabilidad; es decir, si X es una variable continua, $P(X = x) = 0$ para cualquier valor $x \in \mathbb{R}$.

Un caso particular de variables aleatorias continuas es el de las que se conocen como absolutamente continuas, que se caracterizan por ser su función de distribución totalmente continua. Esto quiere decir que hay una función real f , que será integrable y positiva en el conjunto de números reales, de modo que la función de distribución acumulativa F se puede expresar así:

$$F(x) = \int_{-\infty}^x f(u) du$$

La función f se denomina “función de densidad de probabilidad de la variable aleatoria X ”. Dicho nombre no es al azar, ya que $f(x)$ admite una interpretación análoga a la del término físico de densidad. Como consecuencia de esta ecuación, se determina que:

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{P(X \leq x + \Delta x) - P(X \leq x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X \leq x + \Delta x)}{\Delta x}$$

Esto hace ver que $f(x)$ representa la cantidad de probabilidad en un entorno cercano de x , dividido entre la medida Δx de dicho entorno. Si se compara con un término físico, $P(x \leq X \leq x + \Delta x)$ se puede entender como la totalidad de la masa de la probabilidad que se concentra en un volumen Δx alrededor de x . La definición clásica de la densidad es la masa dividida entre el volumen, y esto justifica el nombre de la función f .

Del mismo modo, de la expresión anterior se deduce también que para un valor de Δx muy pequeño se tiene:

$$P[X \in (x, x + \Delta x)] \approx f(x) \Delta x$$

Esto quiere decir que la probabilidad de que la variable aleatoria X se encuentre dentro del intervalo pequeño que contenga a un valor x es prácticamente igual a $f(x)$ veces la amplitud de ese intervalo. Geométricamente, el término $f(x) \Delta x$ representa al área de un rectángulo cuya base es Δx y que tiene como altura $f(x)$.

Propiedades:

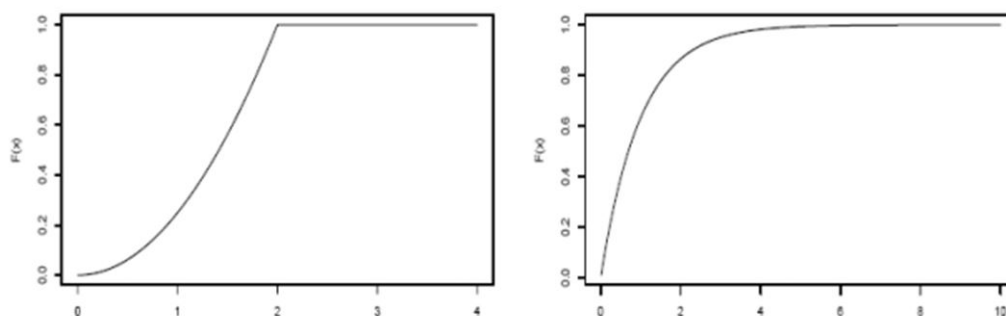
- $f(x) \geq 0$ para todo $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(x) dx = 1$, es decir, la totalidad del área de la función de densidad es 1.
- Para todo $a, b \in \mathbb{R}$, $P(a \leq X \leq b) = P(X \in [a, b]) = \int_a^b f(x) dx$ es el área que determinará la función de densidad de X sobre el intervalo $[a, b]$.
- Los intervalos $[a, b]$, (a, b) , $(a, b]$ y $[a, b)$ tienen todos el mismo valor de probabilidad.

Hay que tener en cuenta que la función de densidad tiene el mismo objetivo que la función de probabilidad para variables aleatorias discretas; sin embargo, para el caso continuo únicamente tendrá sentido el cálculo de probabilidades de intervalos, ya que $P(X = x) = 0$ para todos los valores de $x \in \mathbb{R}$.

Función de distribución

Dada una variable aleatoria continua X , la función de distribución se definirá como una función tal que $F(x) = P(X \leq x) = P(X \in (-\infty, x]) = \int_{-\infty}^x f(t)dt$, para todo $x \in \mathbb{R}$.

Como ocurre en el caso discreto, la función $F(x)$ da las probabilidades acumuladas hasta el punto $x \in \mathbb{R}$, pero en este caso se trata de una función continua. Algunos ejemplos son:



Propiedades:

- $0 \leq F(x) \leq 1$, para todo $x \in \mathbb{R}$.
- $F(-\infty) = 0$
- $F(\infty) = 1$
- Si $x_1 \leq x_2$, entonces $F(x_1) \leq F(x_2)$, es decir, $F(x)$ es no decreciente.
- Para todo $a, b \in \mathbb{R}$, $P(a \leq X \leq b) = F(b) - F(a)$.
- La función de densidad de X se calcula al derivar la función de distribución, o lo que es lo mismo, $f(x) = F'(x)$.

2.3.3. Esperanza y varianza de una variable aleatoria

La **esperanza matemática** se define como la generalización de la media aritmética que tiene toda la población, o lo que es lo mismo, la media de la variable aleatoria. Otra forma de denominarla es valor esperado, valor medio o esperanza matemática; y en todos los casos se representa por la letra μ .

Si se tiene una variable aleatoria discreta X , la cual será representada de forma general por una tabla de valores de x_i y probabilidades $p_i = P(X = x_i)$:

X	x_1	x_2	...	x_n
$P(X = x_i)$	p_1	p_2	...	p_n

Para calcular la esperanza, se hace la media aritmética de los valores, es decir, el sumatorio de los valores multiplicado por sus correspondientes probabilidades, que son equivalentes a sus frecuencias relativas:

$$\mu = E(X) = \sum_{i=1}^k x_i \cdot p_i$$

Si a una variable estadística se la representa por los valores que toma x_i , y las frecuencias relativas son $f_i = n_i / n$, entonces la media aritmética se podrá expresar así:

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

Es decir, el sumatorio de los valores por su frecuencia. Para el caso de las variables aleatorias, las frecuencias serán las probabilidades. Es por ello por lo que la esperanza es un valor medio esperado.

Para el caso de que X sea una variable aleatoria continua, la variable toma valores infinitos. En este caso, se incluye la función de densidad:

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Es preciso destacar que si la variable tiene un número concreto de valores (finito), el sumatorio que definirá la esperanza tendrá un número de sumandos finitos, y será siempre convergente. Por ello, la esperanza de una variable aleatoria con número finito de valores siempre existirá.

Sin embargo, si E_x es infinito, se hace necesaria la imposición de la condición de convergencia absoluta de la serie, porque, en caso contrario, diferentes reordenaciones de los sumandos darían diferentes valores de la esperanza.

Tanto para el caso continuo como para el caso discreto, la esperanza matemática tiene las siguientes propiedades:

- Si C es una constante, $E(C) = C$.
- $\forall a, b \in \mathbb{R}, E(aX + b) = aE(X) + b$.
- Si $g(X)$ es una función de X :
 - Si X es discreta, $E[g(X)] = \sum_{i=1}^{\infty} g(x_i) \cdot p_i$
 - Si X es continua, $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$
- Si X_1, X_2, \dots, X_n son variables aleatorias, entonces $E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i]$.

Desigualdad de Chebyshev

Este operador es muy utilizado para calcular la probabilidad cuando no se conoce la distribución de probabilidad de una variable aleatoria X .

Si se tiene una variable aleatoria con varianza y esperanza finitas, entonces se cumplirá que para todo $K \geq 1$:

$$P(|X - E(X)| \geq k) \leq \frac{\text{var}(X)}{k^2}$$

Equivalentemente, se tiene que:

$$P(|X - E(X)| < k) \geq 1 - \frac{\text{var}(X)}{k^2}$$

No obstante, la cota que estima la desigualdad de Chebyshev es muy amplia y únicamente deberá usarse cuando no se tenga la ley de la variable aleatoria X .

La varianza de una variable aleatoria X , la cual toma valores x_1, x_2, \dots, x_n con distribución de probabilidad, se define así:

Para el caso discreto:

$$V[X] = (x - E(X))^2 = \sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n x_i^2 P_i - \mu^2$$

Para el caso continuo:

$$V[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Propiedades de la varianza y la esperanza:

Sean X e Y dos variables aleatorias tales que $Y = aX + b$; entonces, siempre se cumplirá que:

- $E[Y] = a E[X] + b$
- $V[Y] = a^2 V[X]$

Si X e Y son variables independientes, entonces $E[XY] = E[X] E[Y]$.

2.3.4. Covarianza entre dos variables aleatorias

Si se tienen dos variables aleatorias X e Y , se definirá la covarianza entre las dos variables de la siguiente manera:

$$Cov = E[XY] - E[X] E[Y]$$

Donde el cálculo de las esperanzas dependerá de que las variables sean continuas o discretas. De dicha definición se deduce que cuando X e Y son independientes, el valor de la covarianza es 0:

$$E[X \cdot Y] = E[X] E[Y]$$

Entonces:

$$Cov = E[XY] - E[X] E[Y] = E[X] \cdot E[Y] - E[X] \cdot E[Y] = 0$$

2.4. Distribuciones

Como ya se ha visto anteriormente, el estudio de una variable aleatoria queda reducido al de la distribución de probabilidad, la cual es una función de conjunto.

Es cierto que para algunos tipos de variables aleatorias manejar estas funciones de conjunto puede parecer simple, por ejemplo, cuando se tienen variables con un determinado número de valores (finito); pero, generalmente, el estudio con estas funciones puede ser bastante complejo.

Para resolver este problema, se le asigna a cada distribución de probabilidad P_x una función de punto que la describa completamente y que se llamará “función de distribución de la variable aleatoria X ”.

Dicho de otro modo, si se tiene una variable aleatoria X que está definida sobre un espacio de probabilidad (Ω, A, P) con una distribución de probabilidad P_x , se denomina función de distribución de la variable a lo siguiente:

$$F_x : \mathbb{R} \rightarrow [0,1]$$

$$\mapsto F_x(x) = P_x((-\infty, x]) = P\{X \leq x\}$$

La función de distribución de una variable aleatoria X debe cumplir las siguientes características:

- Debe ser monótona no decreciente.
- Debe ser continua a la derecha.
- $\lim_{x \rightarrow +\infty} (F(x)) = 1$ y $\lim_{x \rightarrow -\infty} (F(x)) = 0$

2.4.1. Funciones de densidad y distribución

2.4.1.1. Caso discreto

Una variable aleatoria $X: (\Omega, A, P) \rightarrow (\mathbb{R}, B, P_x)$ será discreta si presenta valores en un conjunto numerable; esto quiere decir que: $\exists E_x \subset \mathbb{R}$ tal que:

$$P(X \in E_x) = 1$$

Donde E_x corresponde al conjunto de valores de la variable, y está claro que:

$$X \notin E_x, P(X = x) = 0$$

Función masa de probabilidad

Si se tiene una variable aleatoria discreta X con valores dentro del conjunto:

$$E_x = \{x_n, n \in \mathbb{N}\}$$

Su función masa de probabilidad viene definida así:

$$P_x: E_x \rightarrow [0, 1]$$

$$x_n \rightarrow P_x(x_n) = P(X = x_n) = P(\{x_n\})$$

Las propiedades que cumple la función masa de la probabilidad son:

- Es no negativa: $p_x(x) \geq 0, \forall x \in E_x$.
- La sumatoria de los datos da como resultado la unidad: $\sum_{x \in E_x} P_x(x) = 1$.
- La distribución de probabilidad de la variable estará determinada por la función masa de probabilidad de una variable aleatoria discreta.

$$\forall B \in \mathcal{B}, P_x(B) = P(X \in B) = \sum_{x \in B \cap E_x} P(X = x)$$

En consecuencia, su función de distribución:

$$\forall x \in \mathbb{R}, F_x(x) = \sum_{\substack{x_n \in E_x \\ x_n \leq x}} P(X = x_n) = \sum_{n=1}^{\infty} P(X = x_n) \mathbb{I}_{(x-x_n)}$$

Con lo siguiente:

$$\mathbb{I}_x(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Por todo ello se puede concluir que la función de distribución de una variable aleatoria discreta será una función en la que los números reales x_n serán los puntos de salto de la función, de modo que $P(\{X = x_n\}) > 0$, y tal probabilidad será la longitud del salto.

Teorema: toda colección numerable de números no negativos cuyo sumatorio sea 1 forma los valores que determinarán la función masa de probabilidad de alguna variable aleatoria discreta.

Dicha función será no decreciente, con límites 0 y 1 en ∞ y continua a la derecha. Así pues, con la aplicación del teorema de la correspondencia, habrá una variable aleatoria que tenga esa función como función de distribución y, por las características que muestra, esta variable será de tipo discreto.

2.4.1.2. Caso continuo

En la vida real se dan multitud de ejemplos prácticos en los que las variables de interés tienen valores en un determinado conjunto no numerable por puntos:

- Tiempo de caducidad de un producto determinado.
- Sumatorio de dos puntos que se eligen al azar en un determinado intervalo.

En esas situaciones, no es posible darle una probabilidad positiva a cada valor de la variable de forma que la sumatoria de esas probabilidades dé como resultado 1, como sucedía en el caso discreto; por lo tanto, el estudio de la probabilidad de este tipo de variables es distinto al de aquel de las variables discretas.

Una variable aleatoria $X: (\Omega, A, P) \rightarrow (\mathbb{R}, B, P_x)$ es continua cuando su función de distribución es derivable y continua en todo \mathbb{R} , y con derivada continua (puede no serlo en un conjunto pequeño de puntos). Esto es equivalente a que $\exists f_x: \mathbb{R} \rightarrow \mathbb{R}$ no negativa y que se puede integrar, de manera que:

$$F_X(x) = \int_{-\infty}^x f_x(t) dt, \forall x \in \mathbb{R}$$

La función f_x se define como función de densidad de la variable aleatoria X .

Propiedades de la función de densidad:

- La función de densidad de una variable aleatoria X será una función integrable, no negativa y $\int_{-\infty}^{\infty} f_x(t) dt = 1$.

Que sea no negativa e integrable forma parte de la definición, y:

$$\int_{-\infty}^{\infty} f_x(t) dt = \lim_{x \rightarrow +\infty} \int_{-\infty}^x f_x(t) dt = \lim_{x \rightarrow +\infty} F_X(x) = F_X(+\infty) = 1$$

- Si x_0 es un punto de continuidad de f_x , F_X será derivable en x_0 y $F'_X(x_0) = f_x(x_0)$.
- La función de densidad de una variable aleatoria continua determinará su función de distribución F_X y su distribución de probabilidad P_x .

Según la definición, se tiene:

$$F_X(x) = \int_{-\infty}^x f_x(t) dt, \forall x \in \mathbb{R}$$

Además, se puede probar que:

$$P_x(B) = P(X \in B) = \int_B f_x(x) dx$$

Teorema: si $f: \mathbb{R} \rightarrow \mathbb{R}$ es no negativa, existe su integral y además $\int_{-\infty}^{\infty} f_x(t) dt = 1$, f es la función de densidad de alguna variable aleatoria de tipo continuo.

2.4.2. Ejemplos de distribuciones comúnmente empleadas

2.4.2.1. Bernoulli

La distribución de Bernoulli consiste en un modelo teórico que se emplea para representar una variable aleatoria discreta, la cual únicamente puede resultar en dos sucesos que se excluyen mutuamente.

Dicho de otro modo, la distribución de **Bernoulli** se aplica a una variable aleatoria discreta, la cual únicamente tiene dos posibles sucesos:

- El resultado que se espera que ocurra, que se denominará “éxito”.
- El resultado distinto al esperado, que se denominará “no éxito” o “fracaso”.

Dada una variable aleatoria discreta Z cuya frecuencia puede aproximarse de forma satisfactoria a una distribución de Bernoulli con probabilidad p :

$$Z \sim b(p)$$

De forma general, se suele utilizar el término p para indicar la probabilidad de éxito de la variable aleatoria discreta Z ; entonces, los posibles resultados de la variable aleatoria Z son:

$$\text{Éxito: } P(Z = 1) = p$$

$$\text{Fracaso: } P(Z = 0) = (1 - p) = q$$

- Si la variable aleatoria Z resulta en el resultado que se había definido como “éxito” al principio del **experimento**, ($Z = 1$), entonces la probabilidad de que se obtenga dicho resultado es p .
- Si la variable Z resulta en un resultado diferente al que se había definido como “no éxito” al principio del experimento, ($Z = 0$), entonces la probabilidad de que se obtenga dicho resultado es $1 - p$.

Hay que destacar que el resultado “fracaso” o “no éxito” no quiere decir lo contrario de “éxito”; sino que se refiere a cualquier otro caso diferente al que representa “éxito”, siempre que se den dos o más posibilidades.

Por ejemplo, en el caso de que se tire un dado, si la condición de “éxito” se refiere a sacar un 3 en una tirada, la variable “no éxito” será sacar cualquier resultado distinto a 3.

El espacio muestral sería: $\{1, 2, 3, 4, 5, 6\}$.

Fórmula del parámetro p y la Regla de Laplace

Para la obtención del parámetro p se utilizará la regla de Laplace:

$$p = \frac{\text{casos probables}}{\text{casos posibles}}$$

- Casos posibles. Serán todos aquellos resultados posibles que se pueden obtener en un experimento.
- Casos probables. Serán aquellos resultados que se obtengan en el experimento de forma secuencial, es decir, que serán excluyentes (si ocurre uno, no pueden ocurrir los otros).

Propiedades:

- Media: $\mu = p$
- Varianza: $\text{VAR}[Z] = pq$

Ejemplo:

Un vendedor de seguros hace visitas a futuros clientes para vender seguros de coche. Se sabe que en el 60 % de las visitas tiene éxito y consigue vender un seguro. Se pide calcular la variable aleatoria asociada a ese experimento y obtener la media y la varianza.

Solución:

Vende seguro: $Z = 1$

No vende seguro: $Z = 0$

Media = 0.6

Varianza = 0.24

2.4.2.2. Binomial

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

Hay multitud de experimentos que pueden ser catalogados bajo este tipo de distribución de probabilidad; por ejemplo, si al tirar un dado 6 veces se cuentan los éxitos (sacar un 3) que se obtienen, la distribución de probabilidades se ajustará a una distribución binomial.

Es decir, la distribución binomial se entenderá como una serie de ensayos o pruebas en las que únicamente se pueden tener 2 resultados (éxito o fracaso), tomando el éxito como la variable aleatoria.

Este tipo de distribución tiene lugar cuando se repite muchas veces la prueba de Bernoulli. La probabilidad de éxito será igual en todas las repeticiones.

Dada una variable aleatoria X = número de éxitos en n pruebas, se dirá que sigue una distribución binomial de parámetros n y p , $X \sim B(n, p)$.

La variable puede tomar los valores $\{0, 1, 2, \dots, n\}$ y la función de probabilidad será:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

Donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, p es la probabilidad de éxito y $q = 1 - p$.

Propiedades:

- Media: $\mu = np$
- Varianza: $\text{VAR}[X] = npq$
- Propiedad reproductiva (cuando se tienen 2 variables): si X_1 y X_2 son dos variables aleatorias independientes distribuidas según una $B(n_1, p)$ y una $B(n_2, p)$ respectivamente, la variable aleatoria $X = X_1 + X_2$ se distribuirá según una $B(n_1 + n_2, p)$.

Características de la distribución binomial:

- En cada experimento, ensayo o prueba únicamente serán posibles 2 resultados (éxito o fracaso).
- La probabilidad de tener éxito debe ser constante. Dicha probabilidad se simboliza por la letra p . Por ejemplo, la probabilidad de que al tirar al aire una moneda salga cara es 0.5, y será constante, ya que la moneda no cambia en cada ensayo.
- La probabilidad de no tener éxito también deberá ser constante. Se representará por la letra q .
- El resultado que se obtenga de cada ensayo será independiente del resultado obtenido anteriormente, es decir, lo que ocurra en cada experimento no afectará a los experimentos posteriores.
- Los sucesos serán excluyentes mutuamente, es decir, no podrán darse los dos simultáneamente. Por ejemplo, si se lanza una moneda, no puede salir cara y cruz al mismo tiempo.
- Los sucesos serán colectivamente exhaustivos, es decir, al menos uno deberá ocurrir. Por ejemplo, si se lanza una moneda y no sale cara, deberá salir cruz.
- La variable aleatoria que sigue una distribución binomial suele representarse como $X \sim (n, p)$, donde n representa el número de experimentos o ensayos y p la probabilidad de tener éxito.

Ejemplo:

Se supone que un 80 % de las personas en todo el planeta vieron el partido de la final del mundial de fútbol. Una vez acabado el partido, 4 personas se reúnen para comentarlo. ¿Qué probabilidad existe de que 3 de ellas hayan visto el partido?

Las variables del experimento serán:

- $n = 4$: total de la muestra.
- $x = 3$: número de éxitos (haber visto el partido).
- $p = 0.8$: probabilidad de éxito.
- $q = 0.2$: probabilidad de fracaso.

Sustituyendo los valores en la fórmula, se obtiene:

$$P_{(3)} = \frac{4!}{3! (4-3)!} 0,8^3 0,2^{4-3}$$

El resultado final sería 0.4096; al multiplicarlo por 100 se obtiene una probabilidad del 40.96 % de que 3 de las 4 personas hayan visto la final del mundial de fútbol.

2.4.2.3. Poisson

La distribución de **Poisson** consiste en una distribución de la probabilidad discreta que se aplica a las ocurrencias de algún suceso durante un determinado intervalo. La variable aleatoria x representa el número de ocurrencias de un suceso en un determinado intervalo, el cual puede ser volumen, área, distancia, tiempo o alguna otra unidad similar o derivada de estas.

La siguiente expresión determinará la probabilidad de la variable aleatoria X :

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

Donde:

- x puede tomar valores $x = 0, 1, 2, 3, \dots$
- μ es la media del número de sucesos en el intervalo que se está tomando, ya sea de distancia, de tiempo, de volumen, etc. Es necesario entender que dicho valor es una media en sentido estadístico, y como tal se debe calcular mediante la expresión y nunca mediante proporcionalidad o regla de tres.
- Debe cumplirse la condición de normalización: $\sum_{x=0}^{\infty} P(x) = 1$.
- La desviación típica será: $\sigma = \sqrt{\mu}$.
- Cuando se realiza un experimento en el que se mide algún suceso y se obtiene un valor x , su error se determinará por la raíz de x : $x \pm \sqrt{x}$.

La distribución de Poisson debe cumplir una serie de condiciones:

- La variable discreta x será el número de ocurrencias de un suceso durante un intervalo.
- Las ocurrencias deberán ser aleatorias y no incluir ningún vicio que favorezca unas sobre otras.
- Las ocurrencias deberán estar repartidas de manera uniforme dentro del intervalo empleado.

La distribución de Poisson tiene gran importancia y se usa por ejemplo, en la disminución de una muestra radiactiva, en la llegada de los pasajeros de una estación de autobuses o trenes, o en los usuarios que se conectan a una determinada web por hora.

Distribución de Poisson como aproximación a la distribución binomial

A veces, la distribución de Poisson se emplea para aproximar la distribución binomial. Hay un consenso para que se pueda realizar esta aproximación, y es cuando se satisfacen las siguientes condiciones:

1. $n \geq 100$
2. $np \leq 10$

En el caso de que se haga la aproximación por cumplirse las dos condiciones, se necesitará el valor de μ , que se calculará a partir de la expresión:

$$\mu = np$$

Además de la que acabamos de ver, también se suele usar a menudo una aproximación de la distribución de Poisson a una distribución Gaussiana, aunque esta sea de probabilidad continua.

Ejemplo:

El promedio de llamadas de teléfono que se atienden en la centralita de una conocida empresa es de 36 llamadas por hora. ¿Qué probabilidad hay de que en un tiempo de 5 minutos se atiendan 5 llamadas? ¿Qué probabilidad hay de que se atiendan más de 2 llamadas?

Solución:

$$P[X = 5] = 0.1008$$

$$P[X > 2] = 0.5768$$

4.2.4.2. Uniforme

- Discreta

La distribución uniforme discreta consiste en la distribución a partes iguales de la masa de probabilidad entre un número finito de valores. Si una variable aleatoria X cuyos valores posibles son x_1, x_2, \dots, x_n tiene distribución uniforme discreta, se cumplirá:

$$P(X = x_1) = P(X = x_2) = \dots = P(X = x_n) = 1/n$$

Se supone que $x_1 \leq x_2 \leq \dots \leq x_n$; su función de distribución vendrá definida por lo siguiente:

$$F(x) = \begin{cases} 0, & \text{si } x < x_1 \\ 1/n, & \text{si } x_1 \leq x < x_2 \\ 2/n, & \text{si } x_2 \leq x < x_3 \\ \dots & \\ 1, & \text{si } x_n \leq x \end{cases}$$

Por otra parte, se tendrá que:

$$E[X] = 1/n \sum_{i=1}^n x_i \text{ y que } E[X^2] = 1/n \sum_{i=1}^n x_i^2$$

Sus características más importantes son:

- Valor de la media:

$$\mu = \frac{n+1}{2}$$

- Valor de la varianza:

$$V[X] = \frac{n^2+1}{12}$$

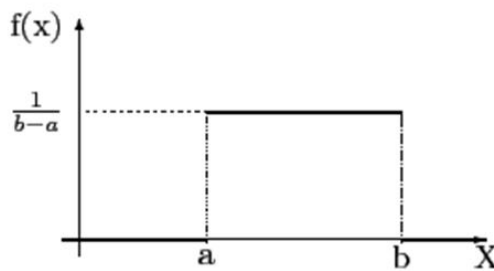
- Continua

La distribución uniforme continua se entiende como la extensión de la distribución uniforme discreta, es decir, como aquella que tomará con igual probabilidad valores dentro de dos conjuntos cualesquiera de la misma amplitud y que se encuentren dentro del intervalo de posibles valores de la variable.

La variable X tendrá una distribución rectangular o uniforme en el intervalo $[a, b]$ ($U[a, b]$) cuando su función de densidad venga definida de la siguiente manera:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{en el resto} \end{cases}$$

Su representación gráfica sería:



Sus características más importantes son:

- Valor de la media:

$$E[X] = \frac{a+b}{2}$$

- Valor de la varianza:

$$V[X] = \frac{(b-a)^2}{12}$$

- Valor de la función generatriz de momentos:

$$M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

Ejemplo:

El metro llega a cierta estación cada 15 minutos. ¿Cuál será la probabilidad de que una señora que llega a la parada en un momento determinado tenga que esperar el metro más de 5 minutos?

Solución:

Si se define X como el tiempo de espera, entonces $X \sim U(0, 15)$.

Lo primero que habrá que hacer es calcular la función de distribución:

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x/15 & \text{si } 0 \leq x \leq 15 \\ 1 & \text{si } x > 15 \end{cases}$$

La probabilidad vendrá definida por lo siguiente:

$$P(X > 5) = 1 - P(X \leq 5) = 1 - 5/15 = 0.67$$

2.4.2.5. Normal

Se denomina distribución normal al modelo teórico que es capaz de hacer una aproximación satisfactoria del valor de una variable aleatoria a una situación idílica.

Esta distribución se emplea a menudo en aplicaciones estadísticas. Su utilización se justifica por la normalidad con la que algunos fenómenos tienen tendencia a parecerse en su comportamiento a esta distribución.

Dicho de otra forma, la distribución normal adaptará una variable aleatoria a una función que dependerá de la desviación típica y de la media; es decir, la variable aleatoria y la función se representarán de la misma manera, pero con algunas diferencias.

Ciertas variables aleatorias continuas tienen una función de densidad que al representarlas gráficamente tienen forma de campana de Gauss.

En algunas ocasiones, al considerarse distribuciones binomiales tipo $B(n, p)$ para el mismo valor de p y con valores de n cada vez más altos, se observa que sus polígonos de frecuencia tienden a una curva de Gauss.

Se puede resumir que la importancia de la distribución normal se debe sobre todo a que hay un gran número de variables asociadas a fenómenos de la naturaleza que siguen este modelo:

- Caracteres morfológicos de individuos (animales, personas, plantas, etc.) de una especie: altura, peso, perímetros, diámetros, etc.
- Caracteres fisiológicos: efecto de la misma dosis de un fármaco, efecto de una misma cantidad de abono en una planta, etc.
- Caracteres sociológicos: consumo de alcohol en un mismo grupo de individuos, calificaciones en un examen, etc.
- Caracteres psicológicos: cociente intelectual, grado de adaptación al medio, etc.
- Errores al medir alguna magnitud.
- Valores estadísticos muestrales: la media.

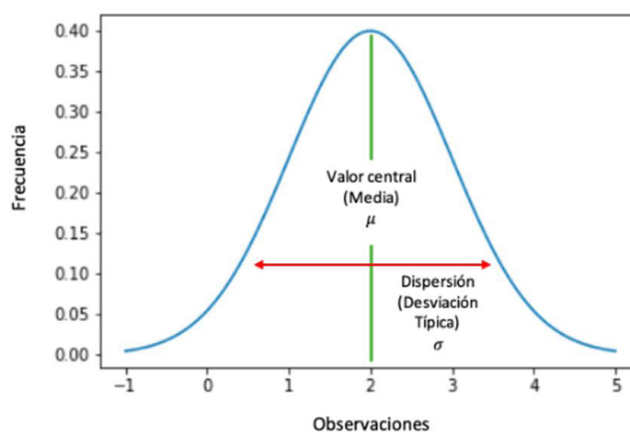
Y, en general, cualquier característica que se obtenga por medio de la suma de varios factores.

Si se tiene una variable aleatoria X , se dirá que la frecuencia de sus observaciones se puede aproximar de manera satisfactoria a una distribución normal tal que:

$$X \sim N(\mu, \sigma)$$

Los parámetros de la distribución serán la desviación típica (σ) y la media o valor central (μ).

La representación de la función de densidad de probabilidad de una variable aleatoria que sigue una distribución normal es la siguiente:



Se trata de una distribución simétrica, es decir, los resultados de la mediana, la media y la moda son equivalentes.

Además, se trata de una distribución unimodal. Los valores más frecuentes o que tendrán mayor probabilidad de darse se encuentran en torno a la media; a la vez, al alejarse de la media, la probabilidad de que aparezcan ciertos valores y su frecuencia disminuyen.

Distribución normal de media 0 y desviación típica 1 ($N(0, 1)$)

La distribución $N(\mu, \sigma)$ puede relacionarse con la distribución $N(0, 1)$ a través del siguiente proceso, que se denomina **estandarización** o tipificación:

$$X \sim N(\mu, \sigma) \rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

La distribución $N(0, 1)$ se define como normal estándar.

Aproximación de la binomial por la normal

En el caso de que $n > 30$, se puede aproximar la distribución binomial de parámetros n, p por la normal de media np , cuya desviación típica es $\sqrt{np(1-p)}$.

$$B(n, p) \approx N(np, \sqrt{np(1-p)})$$

Ejemplo:

A una potencia constante, un semiconductor láser tiene una vida que se distribuye con una media de 7000 horas y una desviación típica de 600 horas. ¿Qué probabilidad existe de que la vida del láser se encuentre entre 6280 y 7120 horas?

Solución:

$X \equiv$ horas de vida del semiconductor $\sim N(7000, 600)$

Tipificación:

$$Z = \frac{x - 7000}{600} \sim N(0, 1); \frac{6280 - 7000}{600} = -1.2; \frac{7120 - 7000}{600} = 0.2$$

$$P(6280 \leq X \leq 7120) = P(-1.2 \leq Z \leq 0.2) = P(Z \leq 0.2) - P(Z \leq -1.2) = 0.5793 - 0.1151 = 0.4642$$

2.4.2.6. t de Student

En la estadística y la probabilidad, la distribución t de Student o distribución t surge del problema de estimar la media de una población que generalmente se encuentra distribuida cuando el tamaño de la muestra es pequeño.

A la teoría de muestras pequeñas también se la conoce con el nombre de teoría exacta del muestreo, ya que, de igual modo, se puede utilizar con muestras aleatorias de mayor tamaño.

Es necesario tener en cuenta un concepto nuevo que será de utilidad para comprender la distribución t: la definición de grados de libertad.

Los grados de libertad se definen utilizando, a su vez, la definición de la varianza muestral:

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

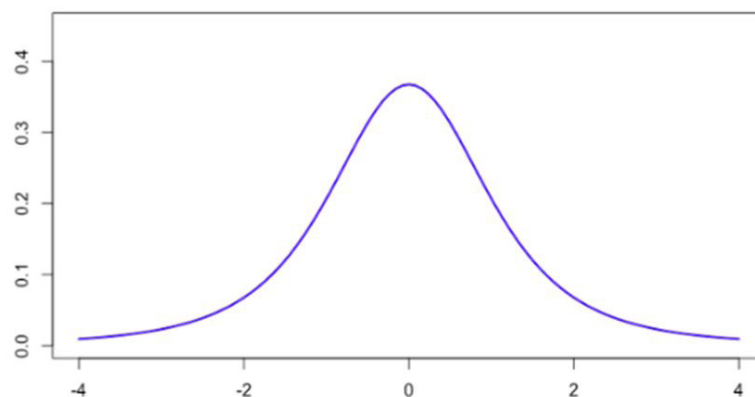
Esta fórmula se basa en $n - 1$ grados de libertad. Esta terminología surge del hecho de que si s^2 se basa en n cantidades $x_1 - \bar{x}$, $x_2 - \bar{x}$, ..., $x_n - \bar{x}$, estas sumarán 0; así que el valor restante se determinará a partir de los valores de cualquier $n - 1$. Por ejemplo, si $n = 4$ y $x_1 - \bar{x} = 8$; $x_2 - \bar{x} = -6$ y $x_4 - \bar{x} = -4$; entonces, de forma automática se sabrá que $x_3 - \bar{x} = 2$; de este modo, únicamente 3 de las 4 medidas de $x_i - \bar{x}$ estarán determinadas libremente, mientras que la otra deberá tomar el valor que haga que el resultado de la sumatoria sea 0. Por ello, solo hay 3 grados de libertad. Se puede determinar que:

$$\text{Grados de libertad} = \text{número de medidas} - 1$$

Una variable aleatoria se distribuirá según el **modelo de probabilidad** t o T de Student con k grados de libertad, siendo k un número entero y positivo, si su función de densidad está definida de la siguiente forma:

$$h_k(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-\frac{(k+1)}{2}} \quad -\infty < t < \infty, \text{ donde } \Gamma(p) = \int_0^\infty e^{-x} x^{p-1} dx$$

La representación gráfica de esta función de densidad es simétrica independientemente del valor que tome k , y tendrá un aspecto similar a la de una distribución normal:



Distribución t de Student con 3 grados de libertad.

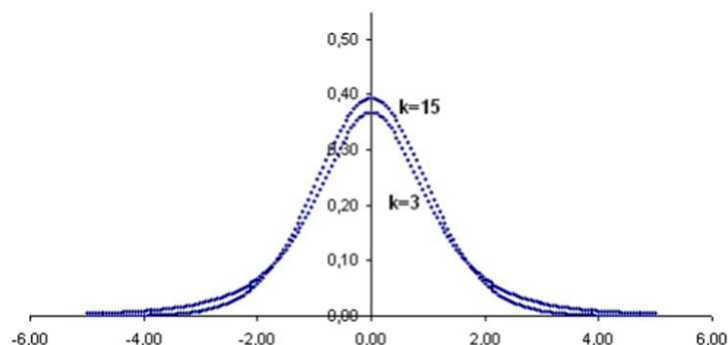
El resultado del valor medio y de la varianza se calculará de la siguiente forma:

$$E(T) = \mu = \int_{-\infty}^{\infty} t \cdot h_k(t) \cdot dt = \int_{-\infty}^{\infty} t \cdot \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-\frac{(k+1)}{2}} dt = \dots = 0$$

Si $k > 3$:

$$\text{Var}(T) = \sigma^2 = E((T - \mu)^2) = \int_{-\infty}^{\infty} (t - \mu)^2 \cdot h_k(t) \cdot dt = \int_{-\infty}^{\infty} t^2 \cdot \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-\frac{(k+1)}{2}} dt = \dots = 0$$

En la siguiente gráfica se representan varias distribuciones de t . Aparentemente, la distribución t es semejante a la distribución normal estándar: las dos son unimodales y simétricas, y el valor máximo en la ordenada se alcanzará en la media $\mu = 0$. Pero la distribución t de Student tendrá colas más amplias que la normal, es decir, la probabilidad de las colas será más alta que en la distribución normal. A medida que los grados de libertad aumentan, la forma límite de la distribución t de Student será la distribución normal estándar.



Propiedades de la distribución t :

- Cada curva t tendrá forma de campana con centro en 0.
- Cada curva t estará más dispersa que la curva normal estándar.
- Cuando el valor de k aumente, la dispersión de t disminuirá.
- Cuando k tienda a infinito, la secuencia de curvas de t se aproximará a la curva normal estándar.

La distribución t se emplea cuando:

- Se desea estimar la media de una población normalmente distribuida a partir de una muestra pequeña.
- El tamaño de la muestra es menor a 30 observaciones. (Desde un número de observaciones de 30 o más elementos, la distribución t se asemeja mucho a la distribución normal).
- Se desconoce la desviación típica de una población y debe estimarse a partir de las observaciones de la muestra.

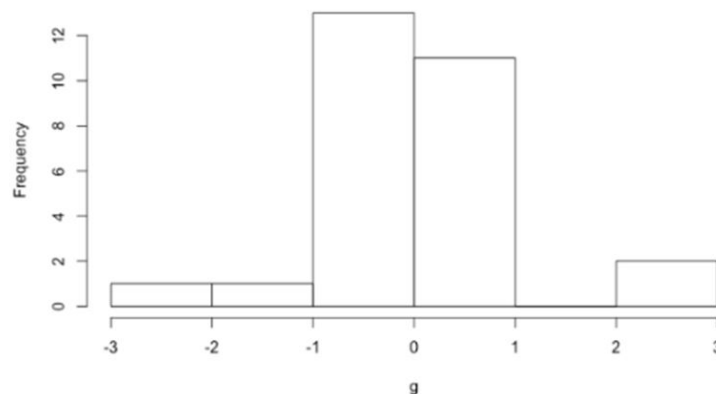
Ejemplo:

Se supone que hay 28 elementos de una variable aleatoria G que seguirá una distribución t de Student con 27 grados de libertad:

-0.66109321	0.87426449	-0.21001010	0.85784467	-2.48692810	0.17831511
-0.34501437	-0.95639484	0.35129492	2.27694001	2.36461405	-0.59664360
0.31923232	-0.15809322	-0.86069232	-0.47580231	0.65589350	-0.20863280
0.18857879	0.98738525	-0.08536776	0.85393439	-1.38541556	-0.80056519
0.66046656	-0.56274230	-0.45095378	0.13224412		

Como se trabajará con datos reales, la media, la mediana y la moda no siempre serán 0 ni exactamente iguales.

Al representar gráficamente en un **histograma** la frecuencia de cada observación de la variable G se obtiene:



Los motivos para pensar que la variable G sigue una distribución t de Student son:

- La distribución es simétrica, es decir, hay el mismo número de observaciones a la izquierda que a la derecha del valor central. Además, tanto la media como la mediana tienden a aproximarse al mismo valor. (La media es 0.016).
- Los datos con más probabilidad o frecuencia se encuentran alrededor del valor central.

2.4.2.7. Chi-cuadrado

La prueba de chi o Ji cuadrado (X^2) es la más conocida y posiblemente la más utilizada para analizar variables cualitativas, es decir, para la determinación de la existencia o no de la independencia entre dos variables. Se encuentra entre las pruebas correspondientes a la estadística descriptiva que se aplican al estudio de dos variables. Es preciso recordar que la estadística descriptiva extrae información sobre la muestra, mientras que la estadística inferencial extrae información sobre la población.

Con el estudio de la independencia se origina un método para comprobar si las frecuencias que se observan en cada categoría son o no compatibles con la independencia entre las variables.

Para la evaluación de la independencia entre variables es necesario calcular los valores que indicarían la independencia absoluta, que se denominan frecuencias esperadas, y se comparan con las frecuencias de la muestra.

Habitualmente, la hipótesis nula (H_0) indica que ambas variables son independientes; por el contrario, la hipótesis alternativa (H_1) indica que dichas variables tienen alguna relación o grado de asociación.

Pruebas de chi-cuadrado de ajuste o de independencia

Estas pruebas no son propias de la estadística paramétrica, ya que no se implantan suposiciones restrictivas respecto al tipo de variables que se admiten, ni respecto a su distribución de probabilidad, ni en los valores ni el conocimiento de sus parámetros. Se determinan en dos casos:

- Si se desea saber si una variable tiene una función de probabilidad determinada. La prueba se denomina chi-cuadrado de ajuste.
- Si se pretende estudiar si dos variables son independientes estadísticamente. En tal caso, se aplica la prueba de chi-cuadrado de independencia o de contingencia.

Chi-cuadrado de ajuste

En este caso, la hipótesis nula establece que una variable X tiene cierta distribución de probabilidad con ciertos valores de los parámetros. El tipo de distribución se determina según la definición propia de la variable, consideraciones teóricas al margen de esta y/o las evidencias que aportan datos que son anteriores al experimento actual.

En ocasiones, la definición del tipo de variable incluye los valores de sus parámetros o de parte de estos; si no es así, estos parámetros se determinarán a partir de la muestra de valores de la variable que se empleará para la realización de la prueba de ajuste.

Las hipótesis se definen así:

- Hipótesis nula. X tiene una distribución de probabilidad $f(x)$ de parámetros y_1, y_2, \dots, y_p .
- Hipótesis alternativa. X tiene otra distribución de probabilidad.

Cabe destacar que rechazar la hipótesis nula no significa que todos sus aspectos sean erróneos, sino solo el conjunto de ellos; por ejemplo, puede darse el caso de que el tipo de distribución sea válido, pero haya habido un error en los valores de los parámetros.

Es necesario tener una muestra de valores de la variable X . Si se trata de una variable discreta y tiene pocos valores posibles, se calculan las probabilidades de esos valores a través de sus frecuencias; si se trata de una variable continua (o discreta con un número grande de valores), se estima por intervalos.

La prueba se basa en comparar la serie de frecuencias absolutas que se observan de forma empírica para los valores de la variable (O_i) y las frecuencias absolutas teóricas correspondientes que se obtienen a partir de la función de probabilidad supuesta en la hipótesis nula.

Una vez que se han calculado las frecuencias absolutas de cada valor o de cada intervalo de valores, se obtendrá el sumatorio total de observaciones de la muestra (T) al sumar las frecuencias que se observan:

$$T = \sum_i O_i$$

Para el cálculo de las frecuencias que se esperan, se divide T en partes proporcionales a la probabilidad de cada suceso o grupo de ellos. Para esto, se determinan dichas probabilidades usando la función de probabilidad, de modo que cada término E_i estará definido por la siguiente expresión:

$$E_i = f(X_i) \cdot T$$

Así pues, para la prueba se dispone de los siguientes datos:

- Valor de la variable: $x_1, x_2, x_3, \dots, x_i, \dots, x_k$.
- Frecuencias que se observan: $O_1, O_2, O_3, \dots, O_i, \dots, O_k$.
- Frecuencias que se esperan: $E_1, E_2, E_3, \dots, E_i, \dots, E_k$.

Cuando la hipótesis nula es cierta, las diferencias entre los valores que se observan y los que se esperan son atribuibles al azar. En estos casos, se podría determinar un parámetro dependiente de ambos, el cual tendría una distribución ajustable a una chi-cuadrado:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \approx \chi^2$$

En cambio, si la hipótesis nula es falsa, los E_i ya no son los valores de las frecuencias que se esperan, por lo que las diferencias entre los valores observados y los esperados reflejan no solo el efecto del azar, sino también las diferencias entre los E_i y la serie auténtica de valores que se esperan. A consecuencia de esto, la diferencia entre los valores del numerador de la expresión anterior tenderá a ser mayor, y al ser un factor elevado al cuadrado, el sumatorio de cocientes será positivo y mayor de lo esperado para los valores de chi-cuadrado:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} \gg \chi^2$$

Así pues, dicho término será el estadístico de contraste de la prueba de hipótesis y la región crítica siempre se encontrará en la cola derecha de la distribución chi-cuadrado (dicha prueba siempre será de una sola cola).

Estadístico de contraste:

$$\chi^2 \approx \sum_i \frac{(O_i - E_i)^2}{E_i}$$

La hipótesis nula se aceptará si $\chi^2 < \chi^2_{1-\alpha, v}$, el percentil $1 - \alpha$ de la distribución chi-cuadrado con v grados de libertad.

Para calcular el número de grados de libertad de la variable chi-cuadrado se procede como se indica a continuación:

- En un principio, se tendrán tantos grados de libertad como parejas de frecuencia esperada y frecuencia observada haya.
- A esa cantidad se le deberá restar el número de restricciones lineales que se impongan a las frecuencias observadas, es decir, el número de datos que es necesario calcular de forma directa a partir de los datos que se observan para que de ese modo se establezcan los valores que se esperan. Dicho valor será como mínimo 1.

Una condición que se debe cumplir para poder llevar a cabo una prueba chi-cuadrado es que las frecuencias de las distintas clases deberán ser lo bastante altas para garantizar que cualquier pequeña desviación aleatoria en la muestra no tendrá gran repercusión sobre el resultado del estadístico de contraste.

Chi-cuadrado de independencia o contingencia

Esta prueba sirve para comprobar la independencia de frecuencias entre dos variables aleatorias X e Y .

Las hipótesis se definen así:

- Hipótesis nula. X e Y son independientes.
- Hipótesis alternativa. X e Y no son independientes.

X e Y son independientes si para cualquier pareja de valores x e y la probabilidad de que X tome el valor de x y la probabilidad de que Y tome el valor de y de forma simultánea es el producto de las probabilidades por separado:

X e Y son independientes $\leftrightarrow \forall x, y \ f(x, y) = f(x) \cdot f(y)$

X e Y son independientes $\leftrightarrow \forall x, y \ f(x, y) = f(x) \cdot f(y)$

Así pues, únicamente se necesitan algunas estimaciones de las funciones de probabilidad de las dos variables por separado y de la función de probabilidad conjunta.

Lo primero que se debe hacer es tomar una muestra de parejas de valores sobre la que se contará la frecuencia absoluta con la que aparece cada combinación (x_i, y_j) o de grupo de valores (i, j) (O_{ij}).

En la siguiente tabla se recogen estos datos, y se muestra la **estimación** de la función de probabilidad conjunta por el número total de datos (T):

X	Y	y_1	y_2	...	y_i	...	y_j	$F_i = \sum_j O_{ij}$
x_1		O_{11}	O_{12}	...	O_{1i}	...	O_{1j}	F_1
x_2		O_{21}	O_{22}	...	O_{2i}	...	O_{2j}	F_2
...	
x_i		O_{i1}	O_{i2}	...	O_{ii}	...	O_{ij}	F_i
...	
x_j		O_{j1}	O_{j2}	...	O_{ji}	...	O_{jj}	F_j
$C_j = \sum_i O_{ij}$		C_1	C_2	...	C_i	...	C_j	T

Para la obtención de las estimaciones de las funciones de probabilidad marginales, se deberán sumar por columnas y por filas los datos de las frecuencias conjuntas. Las sumas de filas (F_i) serán las veces que se obtiene un valor de X (x_i) en cualquier combinación con diferentes valores de Y , es decir, la estimación de la función de probabilidad de X por el número total de observaciones; de forma análoga, la sumatoria de columnas (C_j) será la estimación de la función de probabilidad de Y por el número total de observaciones.

El número total de observaciones puede obtenerse como la sumatoria de todas las frecuencias que se observan, o como la sumatoria de las sumas de columnas o de filas:

$$T = \sum_{ij} O_{ij} = \sum_i F_i = \sum_j C_j$$

Siempre que las variables sean independientes, se debe cumplir que:

$$\frac{O_{ij}}{T} = \frac{F_i}{T} = \frac{C_j}{T} = \frac{F_i \cdot C_j}{T^2} \quad \forall i, j$$

Evidentemente, no se prevé que dicha condición se cumpla por culpa de los errores que se cometen en el muestreo aleatorio, por lo que el inconveniente está en discernir entre las diferencias que se producen por el efecto del muestreo y las que son producidas por falta de independencia.

Si se tienen variables independientes, o lo que es lo mismo, si las frecuencias E_{ij} son en realidad los datos que se esperan de las frecuencias O_{ij} , se podrá determinar un factor que dependerá de las dos y tendrá una distribución chi-cuadrado:

$$\sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2$$

En el caso de que las variables sean dependientes, las diferencias entre las frecuencias esperadas y las observadas serán mayores que las que se atribuyen al azar y, al encontrarse elevadas al cuadrado, tenderán a ser mayores que el valor de la variable chi-cuadrado:

$$\sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \gg \chi^2$$

Este parámetro será el estadístico de la prueba de hipótesis, y la región crítica se encontrará en la cola derecha de la distribución chi-cuadrado. (Dicha prueba siempre será de una sola cola).

Estadístico de contraste:

$$\chi^2 \approx \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Se aceptará la hipótesis nula si $\chi^2 < \chi^2_{1-\alpha, \nu}$ el percentil $1 - \alpha$ de la distribución chi-cuadrado con ν grados de libertad.

Los grados de libertad de la chi-cuadrado que se emplea como contraste se calculan del siguiente modo:

- En principio, habrá tantos grados de libertad como combinaciones de valores x_i, y_j se tengan.
- Al número obtenido se le resta un término I , ya que para el cálculo de las frecuencias esperadas se necesitan calcular las I sumas de filas. Cuando se conocen las sumatorias de las filas, se obtiene el número total de observaciones sin que se pierda ningún grado de libertad.
- Por último, hay que determinar a partir de las frecuencias que se observan $J - 1$ de las sumas de columnas. Las restantes se obtienen al restar la suma de las anteriores del total de observaciones (T).

En definitiva, el número de grados de libertad de la prueba será el producto de las filas menos 1 por el número de columnas menos 1:

$$\nu = I \cdot J - I - (J - 1) = I \cdot J - I - J + 1 = (I - 1) \cdot (J - 1)$$

Como se ha visto, esta prueba no hace suposiciones sobre el tipo de distribución de ninguna variable implicada y solo emplea información contingente, es decir, de la muestra. Por esta razón se la denomina chi-cuadrado de contingencia.

2.4.2.8. F de Snedecor

Se trata de una distribución de probabilidad continua que se emplea en la **inferencia estadística**, generalmente en el contraste de la igualdad de varianzas de dos poblaciones y en el análisis de la varianza, la cual permite detectar la presencia o no de diferencias significativas entre muestras diferentes, y que será esencial en todos los casos en los que se investigue la importancia de un factor en la naturaleza y desarrollo de una característica.

También es conocida como distribución F o como distribución F de Fisher-Snedecor.

Una variable aleatoria de distribución F se construye a partir del siguiente cociente:

$$F = \frac{U_1/d_1}{U_2/d_2}$$

En el que los términos U_1 y U_2 serán independientes estadísticamente y seguirán una distribución chi-cuadrado con unos grados de libertad d_1 y d_2 respectivamente.

La función de densidad de una variable F será:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right) \left(\frac{m}{n}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} x^{\frac{m}{2}-1} \left(1 + \frac{mx}{n}\right)^{-\frac{m+n}{2}}$$

Con $x > 0$, siendo:

$$\Gamma(p) = \int_0^{\infty} x^{p-1} e^{-x} dx$$

La función gamma de Euler, con $p > 0$. Al final, la función de distribución viene dada por lo siguiente:

$$F(x) = \Pr\{X \leq x\} = \int_0^x f(t) dt$$

2.5. Teorema del límite central y ley de grandes números

Teorema central del límite

El teorema central del límite (TCL) es una teoría estadística que determina que, dada una muestra lo bastante grande de la población, la distribución de las medias muestrales seguirá una distribución normal. Además, a medida que el tamaño de la muestra se vaya incrementando, la media muestral se aproximará a la media de la población; por ello, mediante el TCL se puede definir la distribución de la media muestral de una población determinada con una varianza definida.

Propiedades principales del TCL:

- Si la muestra tiene un tamaño grande, la distribución de las medias muestrales seguirá una distribución normal. Se considerará un tamaño de muestra grande cuando sea superior a 30; en tal caso, la función de distribución de la media muestral será cercana a una normal. Esto se cumplirá siempre y no dependerá de la forma que tenga la distribución.
- La media muestral y la poblacional serán similares, es decir, la media de la distribución de las medias muestrales será la misma que la media de la población total.
- La varianza de la distribución de las medias muestrales estará definida por σ^2/n , es decir, la varianza de la población dividida entre el tamaño de muestra.

Tener una distribución de las medias muestrales similar a una normal es bastante útil, ya que la distribución normal es sencilla de aplicar en el contraste de hipótesis y la elaboración de intervalos de confianza. Además, el teorema central del límite permite hacer inferencias sobre la media de la población a partir de la media muestral, lo cual es muy útil cuando no hay muchos medios y no se pueden recoger datos de toda la población.

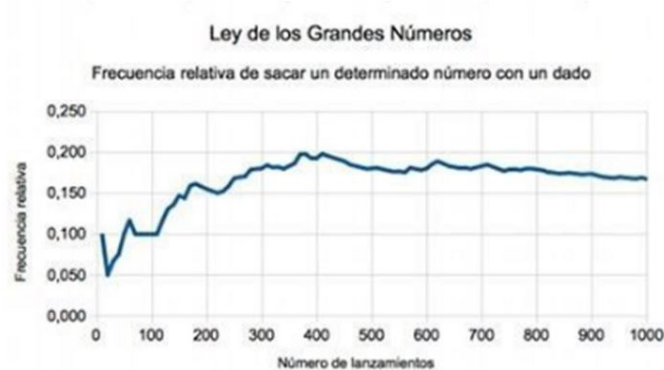
Ley de grandes números

La ley de los grandes números es un teorema fundamental de la probabilidad que anuncia que al repetir muchas veces (tendiendo al infinito) el mismo experimento, la frecuencia de que ocurra cierto evento tiende a ser constante.

Por ejemplo, se realiza el experimento de lanzar un dado y se considera el evento de que salga un 6. Como es sabido, la probabilidad de que salga el número 6 es $\frac{1}{6}$. Según la ley de grandes números, a medida que se aumenta el número de veces que se lanza el dado, la frecuencia con la que se repetirá el evento de que salga un 6 se acercará poco a poco a una constante, que tendrá un valor de 16.6 % o $\frac{1}{6}$.

Es posible que en los primeros lanzamientos la frecuencia con la que salga el 6 no sea del 16 %, sino otro porcentaje diferente; pero a medida que se hagan más lanzamientos (por ejemplo, 12 000) la frecuencia en que aparezca el 6 será próxima al 16.6 %.

En el siguiente gráfico se puede ver el ejemplo de lanzar un dado repetidas veces:



Se puede observar cómo se modifica la frecuencia relativa de sacar un número determinado.

Resumen

Los experimentos aleatorios se conciben de manera que los resultados del espacio muestral son cualitativos (por ejemplo, al tirar una moneda una vez, el espacio muestral es $\{c, x\}$), o cuantitativos (por ejemplo, la edad de una serie de estudiantes). El concepto de variable aleatoria proporciona un medio para relacionar cualquier resultado con una medida cuantitativa.

Se puede pensar que la población asociada a un experimento “se genera” al repetirlo un alto número de veces y al considerar el conjunto de mediciones asociadas. Nunca se llega a medir a todos los miembros de una población; sin embargo, se puede concebir la idea de hacerlo. Es posible obtener un subconjunto pequeño de estas mediciones (que, como se sabe, recibe el nombre de muestra) y, con la información contenida en esta muestra, describir o hacer inferencias acerca de la población.

La función de distribución de una variable aleatoria (o sus derivadas, la función de probabilidad en el caso discreto y la función de densidad en el caso continuo) es la herramienta que permite modelar la incertidumbre presente en los procesos de observación o experimentación.

La distribución normal o distribución de Gauss es, sin duda, la más importante y la de más aplicación de todas las distribuciones continuas. Es muy adecuada para describir la distribución de muchos conjuntos de datos que ocurren en la naturaleza, la industria y la navegación. Así pues, la distribución normal se considera apropiada para los siguientes conjuntos de datos:

- Los datos meteorológicos correspondientes a temperaturas, lluvias, etc.
- Las clasificaciones correspondientes a pruebas de aptitud.
- Las alturas de individuos de una edad y sexo dado.
- Las medidas físicas de productos manufacturados.
- La vida media de un tipo de lámparas con un voltaje dado, etc.



Capítulo 3

Inferencia estadística

Objetivos

- Conocer la inferencia y sus métodos.
- Entender el significado de la inferencia estadística.
- Calcular los parámetros de la distribución de medias o proporciones muestrales de tamaño n , extraídas de una población de media y varianza conocidas.
- Contrastar los resultados obtenidos a partir de muestras.

Introducción

La inferencia estadística es el conjunto de métodos que permiten inducir, a través de una **muestra** estadística, el comportamiento de una determinada población. Así pues, la inferencia estadística estudia cómo, a través de la aplicación de dichos métodos sobre los datos de una muestra, se pueden extraer conclusiones sobre los parámetros de la población de datos. De la misma manera, estudia también el grado de fiabilidad de los resultados extraídos del estudio.

Para entender este concepto es importante comprender tres conceptos:

- Inferencia. Inferir significa, literalmente, extraer juicios o conclusiones a partir de ciertos supuestos, sean estos generales o particulares.
- Población. Una población de datos es el conjunto total de datos que existen sobre una variable.
- Muestra estadística. Una muestra es una parte de la población de datos.

Si se tiene claro lo que significa el concepto de inferir, una de las dudas fundamentales recae en el hecho de elegir una muestra en lugar de una población.

Normalmente, en estadística se trabaja con muestras debido a la gran cantidad de datos que aporta una población. Por ejemplo, si se quieren sacar conclusiones, esto es, inferir los resultados de las elecciones generales, es imposible preguntar a toda la población del país. Para solventar ese problema se escoge una muestra variada y representativa, gracias a la cual se pueda extraer una estimación del resultado final. Escoger una muestra adecuada es tarea de las distintas técnicas de muestreo.

La otra gran rama de la estadística es la estadística descriptiva.

Los métodos y técnicas de la inferencia estadística se pueden dividir en dos tipos: métodos de estimación de parámetros y métodos de contraste de hipótesis.

- Métodos de estimación de parámetros. Se encargan de asignar un valor al parámetro o al conjunto de parámetros que caracterizan el campo sujeto a estudio. No obstante, al tratarse de una estimación, existe cierto margen de error. Para obtener estimaciones adaptadas a esta realidad, se crean intervalos de confianza.
- Métodos de contraste de hipótesis. Su objetivo es comprobar si una estimación se corresponde con los valores poblacionales. En todo contraste de hipótesis existen dos supuestos: la hipótesis nula (H_0) recoge la idea de que un valor tiene un valor predeterminado; si se rechaza la hipótesis nula, entonces se acepta la hipótesis alternativa (H_1).

3.1. Estimación de parámetros

En la inferencia estadística se asume que se tienen los datos de una muestra y que se pretende conocer cuáles serán las características (como por ejemplo la media, la mediana, la moda, la curtosis, etc.) no de dicha muestra, sino de la población a la que pertenece. A estos valores de dichas características a nivel poblacional se los denomina parámetros, y se representan mediante letras griegas.

Para conocer los valores de los parámetros se puede optar por recoger los datos para todos los elementos de la población (lo cual puede ser poco viable) o por realizar una estimación a partir de los datos de una muestra. Este camino es más habitual, aunque suponga un riesgo de error, ya que, en cuanto a la estimación, el valor que se recoja no tiene por qué coincidir con el valor verdadero de ese parámetro.

Pueden distinguirse dos grandes aproximaciones a la estimación de parámetros: la estimación puntual y la estimación por intervalos. La diferencia radica en que en el primer caso se proporciona una estimación que consiste en un valor concreto (valor puntual), mientras que en la estimación por intervalos se tiene un rango de valores (intervalo).

En el caso de disponer de los datos de una población para una variable X , la obtención de los parámetros que interesan es inmediata, solo hay que aplicar los índices estadísticos que correspondan para cada uno de los datos de la población.

Sin embargo, si se tienen datos de una muestra de esa población, para obtener la estimación de cualquiera de los parámetros hay que emplear un **estimador** del parámetro correspondiente, es decir, una función matemática que permita la obtención de una estimación del valor del parámetro a partir de los datos de la muestra.

La estimación de un parámetro se suele representar con un acento circunflejo sobre la letra del parámetro correspondiente; así, el parámetro σ_x simboliza el valor estimado de la desviación típica de la variable X en la población.

Para un parámetro determinado, se pueden considerar funciones matemáticas diferentes que ofrecen estimaciones del mismo parámetro; por ejemplo, las siguientes funciones pueden ser estimadores de la media (μ_x):

$$\hat{\mu}_x = \frac{\sum x_i^2}{n} \quad \hat{\mu}_x = \frac{\sum x_i}{n-2} \quad \hat{\mu}_x = \frac{\sum x_i}{n^2} \quad \hat{\mu}_x = \frac{\sqrt{\sum x_i^2}}{n} \quad \hat{\mu}_x = \frac{\sum x_i}{n} \quad \hat{\mu}_x = \frac{\sum x_i}{\sqrt{n}}$$

La función que se considera mejor estimador de un parámetro es aquella que cumpla con las siguientes propiedades:

- Ausencia de sesgo. Se dice que un estimador es insesgado cuando el promedio de las estimaciones que se obtienen en muestras diferentes es el valor del parámetro que se quiere estimar.
- Eficiencia. Esta propiedad se establece en términos comparativos, es decir, tendrá una eficiencia mayor el estimador que proporcione estimaciones del valor verdadero del parámetro con una variabilidad menor. Una de las maneras de evaluar la eficiencia de un estimador es obtener la desviación típica de sus estimaciones, lo que se conoce con el nombre de “error típico de estimación del estimador”. De este modo, será mejor el que proporcione un error típico de estimación menor.
- Consistencia. Un estimador es consistente cuando la probabilidad del valor que se estima coincide con la del parámetro y aumenta cuando el tamaño de la muestra crece.
- Suficiencia. El estimador es suficiente respecto a cierto parámetro si agota la información disponible en la muestra aprovechable para la estimación.

En general, los mejores estimadores de los parámetros correspondientes a los índices estadísticos son los propios índices estadísticos que se obtienen a partir de la muestra. Sin embargo, hay algunas excepciones, como son:

- El estimador del parámetro de la varianza (σ_x^2) no es el estadístico de la varianza (s_x^2), sino el de la cuasi-varianza ($s_x'^2$):

$$\hat{\sigma}_x^2 \rightarrow \frac{\sum (x_i - \bar{x})^2}{n-1} = s_x'^2$$

Esto se debe a que el índice estadístico de la varianza no cumple el requisito de ser insesgado del parámetro de la varianza; sin embargo, la cuasi-varianza sí (por eso a veces se la denomina varianza insesgada).

- El estimador del parámetro de la desviación estándar (σ_x) es el estadístico de la cuasi-desviación estándar (s_x'):

$$\hat{\sigma}_x \rightarrow s_x = \sqrt{s_x'^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

- El estimador del parámetro de la covarianza (σ_{xy}) es el estadístico de la cuasi-varianza (s'_{xy}):

$$\hat{\sigma}_{xy} \rightarrow s_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

3.1.1. Límites de variables aleatorias

Siempre que se realice un estudio, hay que medir las variables que caracterizan los resultados obtenidos. A estas variables se las denomina variables aleatorias.

Se dice que una variable es continua cuando puede tomar un valor cualquiera en un intervalo conocido; y es discreta si solo puede tomar algunos valores.

Las distribuciones de probabilidad de variables continuas se definen mediante una función denominada función de densidad o de probabilidad, la cual asocia valores de una variable aleatoria con sus probabilidades correspondientes.

La función de densidad de una variable aleatoria debe ser siempre positiva en todo el dominio, tomar valores entre cero y uno y permitir la obtención de la probabilidad de que un valor de la variable aleatoria esté entre dos puntos, siendo dicha probabilidad el área que se encuentra bajo la curva.

El TCL o teorema central del límite es uno de los instrumentos fundamentales de la estadística. Dicho teorema indica que si una muestra es lo suficientemente grande (se suele decir que cuando es mayor a 30), no importa cuánto vale la distribución de la media muestral, ya que seguirá una distribución normal. En otras palabras, para cualquier variable aleatoria, al extraer muestras de tamaño n (siendo n mayor a 30) y calcular los promedios muestrales, dichos promedios seguirán una distribución normal. A todo esto hay que añadir que la media será igual que la de la variable de interés, y la desviación estándar de la media muestral será, de forma aproximada, el error estándar.

3.1.2. Método de los momentos

Hay que recordar que si se tiene una variable X que es aleatoria, se definirá el momento de orden K en relación con el origen de la siguiente forma:

$$\mu_k = E[X^k] = \begin{cases} \sum_{x_i \in E} x_i^k P(X = x_i) & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{\infty} x^k f(x) dx & \text{si } X \text{ es continua} \end{cases}$$

Es necesario tener en cuenta que $\mu = \mu_1$ y $\sigma^2 = \mu_2 - \mu_1^2$. De la misma manera que la varianza y la esperanza se pueden escribir en función de los momentos, de forma general, si se tiene una variable aleatoria X , dependerá de unos valores que se desconocen $\theta_1, \theta_2, \dots, \theta_k$; en ocasiones se podrán definir estos parámetros en forma de alguna función de ciertos momentos de la variable, es decir, $\theta_j = g_j(\mu_1, \mu_2, \dots)$, $j = 1, 2, \dots, k$. El método de los momentos consiste en el cálculo de dichas funciones. Se calculan los momentos correspondientes a partir de sus muestras análogas:

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n X_i, \hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n X_i^2, \dots, \hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

Y finalmente se calculan los θ_j , a través de las anteriores funciones evaluadas en los momentos muestrales: $\hat{\theta}_j = g_j(\hat{\mu}_1, \hat{\mu}_2, \dots)$, $j = 1, 2, \dots, k$.

Dicho método se fundamenta en el hecho de que los momentos muestrales son estimadores insesgados de los momentos poblacionales. Como ya se sabe, al tomar una muestra aleatoria e ir aumentando su tamaño, su distribución empírica es cada vez más parecida a la distribución de probabilidad de la variable que se observa. Se puede intuir que los momentos muestrales serán cada vez más parecidos a los poblacionales según vaya aumentando el tamaño de la muestra.

3.1.3. Método de la máxima verosimilitud

Sea X una variable aleatoria en la que la distribución de probabilidad es una función de varios o de un solo parámetro desconocido $\theta_1, \theta_2, \dots, \theta_k$, y sea $f_\theta(x)$ su función de probabilidad o de densidad (dependiendo de si X es continua o discreta), siendo $\theta = (\theta_1, \theta_2, \dots, \theta_k)$. Se quiere calcular θ , y se supone que se dispone de una muestra aleatoria (X_1, X_2, \dots, X_n) que ha producido los valores (x_1, x_2, \dots, x_n) . Este método se basa en tomar como estimador de θ el valor que da una probabilidad mayor al grupo de valores que se observan. A través de dicho método, si la muestra aleatoria produce los datos (x_1, x_2, \dots, x_n) se debe a que es bastante probable que dichos valores se observen; por consiguiente, los valores que resultan veraces para θ serán los que hacen que sea muy probable que se observen los datos (x_1, x_2, \dots, x_n) , y el más verosímil es el que hace la probabilidad mayor.

Una definición más formal es la que se representa a continuación:

$$L(\theta) = L((\theta_1, \theta_2, \dots, \theta_k) | x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n | \theta = (\theta_1, \theta_2, \dots, \theta_k)) = f_\theta(x_1, x_2, \dots, x_n)$$

Esta función representa la densidad o probabilidad conjunta de las variables X_1, X_2, \dots, X_n en el punto (x_1, x_2, \dots, x_n) cuando el valor del parámetro es $\theta = (\theta_1, \theta_2, \dots, \theta_k)$. Como (X_1, X_2, \dots, X_n) es una muestra aleatoria, esto quiere decir que las X_i tendrán la misma distribución y serán independientes, y por ello su función de densidad o probabilidad conjunta es el resultado de multiplicar las funciones de densidad o probabilidad de cada variable. Así, se tiene:

Si X_1, X_2, \dots, X_n son variables discretas:

$$L(\theta) = f_\theta(x_1, x_2, \dots, x_n) = P_\theta(X_1 = x_1) P_\theta(X_2 = x_2) \dots P_\theta(X_n = x_n)$$

Siendo $f_\theta(x)$ la función de probabilidad o densidad de las X_i .

El estimador de máxima verosimilitud (denominado estimador MV) se define como el valor del parámetro $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ que maximiza la función:

$$\hat{\theta} = \arg \max L(\theta)$$

La mayoría de las veces, este valor se puede obtener derivando $L(\theta)$ respecto a cada θ_i , igualando a 0 y despejando las θ_i :

$$\frac{\partial}{\partial \theta_i} L(\theta_1, \theta_2, \dots, \theta_k) = 0, i = 1, 2, \dots, k$$

Hay que destacar que $L(\theta)$ es una multiplicación de n términos dependientes de θ , y el cálculo de su derivada es una operación bastante compleja. Por eso, para calcular el valor más alto de $L(\theta)$ se suele utilizar en su lugar la log-verosimilitud:

$$\ell(\theta) = \log(L(\theta)) = \begin{cases} \sum_{i=1}^n \log(P_\theta(X_i = x_i)) & \text{si las } X_i \text{ son discretas} \\ \sum_{i=1}^n \log(f_\theta(x_i)) & \text{si las } X_i \text{ son continuas} \end{cases}$$

Como el logaritmo es una función monótona, el máximo de $L(\theta)$ coincidirá con el máximo de su logaritmo $l(\theta)$, es decir:

$$\hat{\theta} = \arg \max L(\theta) = \arg \max l(\theta)$$

Siendo la derivada de $l(\theta)$ mucho más fácil de calcular.

Así pues, los estimadores de mayor verosimilitud se calcularán en la mayoría de las ocasiones resolviendo la siguiente ecuación:

$$\frac{\partial}{\partial \theta_i} l(\theta_1, \theta_2, \dots, \theta_k) = 0, i = 1, 2, \dots, k$$

Propiedades de los estimadores de máxima verosimilitud:

Suelen preferirse los estimadores de máxima verosimilitud a los estimadores que se obtienen por el método de los momentos (a veces, los estimadores que se obtienen por ambos métodos son coincidentes, aunque en general no ocurre así), ya que tienen propiedades más óptimas, como son:

- **Consistencia.** Los estimadores MV son consistentes, es decir, al aumentar el tamaño de una muestra, es más probable que el valor del estimador se aproxime cada vez más al valor del parámetro.
- **Eficiencia.** Cuando aumenta el tamaño de la muestra, los estimadores MV tienen un valor más bajo del error cuadrático medio de entre todos los posibles estimadores.
- **Normalidad asintótica.** Cuando aumenta el tamaño de la muestra, los estimadores MV tienden a una distribución normal.

3.2. Intervalos de confianza

La estimación puntual hace que a través de un factor se aproxime el resultado de un parámetro desconocido o una característica poblacional (el peso medio de los españoles, la intención de voto a un partido en unas elecciones municipales, el tiempo medio de resolución de un problema matemático, la cantidad de motocicletas, etc.), pero no indica qué error se comete en esa estimación.

En la práctica, lo razonable es añadir un intervalo, junto a la estimación puntual del parámetro, que calcule el margen de error de dicha estimación. El objetivo de la estimación por intervalos de confianza es la construcción de ese intervalo.

Un intervalo de confianza para un parámetro con un nivel de confianza $1 - \alpha$ ($0 < \alpha < 1$) es un intervalo cuyos extremos son aleatorios (L, U) y en el que, teniendo una probabilidad $1 - \alpha$, el parámetro esté contenido.

$$P(\text{parámetro} \in (L, U)) = 1 - \alpha$$

Los valores que más se usan del nivel de confianza $1 - \alpha$ pueden ser 0.99, 0.95 o 0.9 (se da en tanto por ciento, por lo que la confianza es del 99 %, el 95 % o el 90 %, respectivamente). Algunas veces también se suele emplear el concepto de niveles de significación para el valor α .

En una estimación que se hace por intervalos de confianza se parte de una muestra x_1, \dots, x_n . A partir de dichos valores se obtiene un intervalo numérico. Por ejemplo, se puede decir que, teniendo una confianza del 99 %, la proporción de voto a un grupo político "P" se encuentra entre el 27 % y el 30 %; o que, con una confianza del 92 %, la altura media está entre 1.82 y 1.87 metros.

Interpretación:

De cada una de las muestras también se puede obtener un intervalo de confianza. Entonces, con diferentes muestras, también se obtendrán intervalos distintos. A medida que vaya aumentando el número de intervalos que se han creado, el porcentaje de intervalos que contengan el valor verdadero del parámetro se aproximará al 100 (1 - α) %.

Por ejemplo, con un intervalo de confianza del 93 % se está asegurando que, si se toman 100 medidas, el valor verdadero del parámetro quedará incluido en el intervalo en aproximadamente 93 de todos los intervalos que se hayan tomado.

3.2.1. Intervalos de confianza para la media

Se va a estudiar la estimación de la media μ de una variable o población normal (peso medio, altura media, tiempo medio haciendo ejercicio, etc.).

Para la construcción matemática de un intervalo de confianza, se considerará una variable $X \sim N(\mu, \sigma)$, que representará la característica que se vaya a medir (altura, peso, etc.). Se supondrá que σ es conocida.

Se considerará una muestra aleatoria simple X_1, \dots, X_n de la variable X . Para un nivel de confianza $1 - \alpha$, se elegirá el denominado estadístico pivote:

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

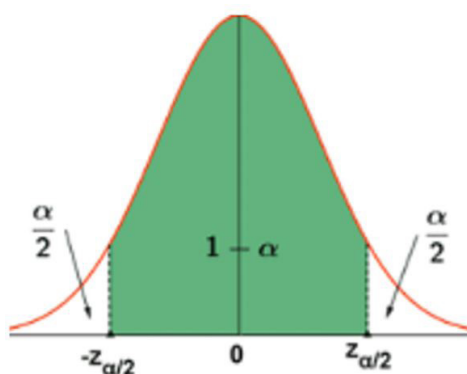
Un estadístico se define como una función de variables aleatorias y es, además, otra variable aleatoria. En este contexto, se verá cuál es la distribución que desarrolla esta variable T ("pivote" es un término que proviene de la nomenclatura que se utiliza en los test de hipótesis).

La media muestral verifica que:

$$\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Así pues, tipificando la variable (se resta la media y se divide entre la desviación típica) se obtiene la variable T , lo que significa que dicha variable tiene una distribución normal estándar [$N(0, 1)$].

Si se tiene en cuenta que $\alpha/2 = P(Z \geq z_{\alpha/2})$ (ver imagen), se sabe que:



Niveles de significación en una normal estandarizada.

$$1 - \alpha = P \left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right)$$

Se despeja el parámetro μ y se obtiene un intervalo de confianza para μ a un nivel de confianza de $1 - \alpha$ que es:

$$(L, U) = \left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

3.2.1.1. Varianza conocida

Sea X una variable aleatoria distribuida como $X \rightarrow N(\mu, \sigma)$. Si se emplea la media muestral (\bar{X}) como estimador, entonces:

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Tipificando y centrando el estimador, cambiando de escala y origen, se obtiene:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$$

Así pues, la probabilidad o el intervalo de confianza para el estimador "media" con una varianza conocida vendrá determinado por la siguiente ecuación:

$$P \left[-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right] = P \left[-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right] = P \left[-\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

Para obtener una media (μ) positiva hay que cambiar todos los signos:

$$P \left[\bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right] = (1 - \alpha)$$

Ordenando los términos, se obtiene:

$$P \left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right] = (1 - \alpha)$$

Con lo cual el intervalo sería:

$$\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

3.2.1.2. Varianza desconocida

- Con $n > 30$

Según el teorema central del límite, si se tiene un tamaño grande de muestra, se sabe que para cualquier distribución se puede hacer una aproximación o una distribución como una normal:

$$\bar{X} \rightarrow N\left(\mu, \frac{s}{\sqrt{n}}\right)$$

Siendo s la cuasi-desviación típica muestral; por lo tanto:

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \rightarrow N(0, 1)$$

Procediendo como en el caso anterior, se llega a que el intervalo de confianza es:

$$\left[\bar{X} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right]$$

- Con $n < 30$

En una población normal, en esas condiciones la variable aleatoria se distribuirá como una t-Student con $n - 1$ grados de libertad; de esta forma:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \rightarrow t_{n-1}$$

Entonces, el intervalo de confianza para un nivel $(1 - \alpha)$ será:

$$P\left[-t_{n-1; \alpha/2} < \frac{\bar{X} - \mu}{s/\sqrt{n}} < t_{n-1; \alpha/2}\right] = P\left[-t_{n-1; \alpha/2} \cdot \frac{s}{\sqrt{n}} < \bar{X} - \mu < t_{n-1; \alpha/2} \cdot \frac{s}{\sqrt{n}}\right] = 1 - \alpha$$

Si se continúa despejando de manera análoga a como se ha hecho en los casos anteriormente expuestos, se obtiene un nivel de confianza:

$$\left[\bar{X} - t_{n-1; \alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1; \alpha/2} \cdot \frac{s}{\sqrt{n}}\right]$$

3.2.2. Intervalos de confianza para la varianza

3.2.2.1. Media conocida

En poblaciones normales, la variable aleatoria $\frac{(n-1)s^2}{\sigma^2} \rightarrow \chi^2_{n-1}$ para un nivel de confianza de $(1 - \alpha) \%$ viene dada por lo siguiente:

$$P\left[\chi^2_{n-1; 1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1; \alpha/2}\right] = 1 - \alpha$$

Cuando se invierte y se despeja, queda:

$$P\left[\frac{1}{\chi^2_{n-1; 1-\alpha/2}} > \frac{\sigma^2}{(n-1)s^2} > \frac{1}{\chi^2_{n-1; \alpha/2}}\right] = P\left[\frac{(n-1)s^2}{\chi^2_{n-1; 1-\alpha/2}} > \sigma^2 > \frac{(n-1)s^2}{\chi^2_{n-1; \alpha/2}}\right] = 1 - \alpha$$

Con lo que el intervalo de confianza para la varianza es:

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1; \alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1; 1-\alpha/2}}\right]$$

3.2.2.2. Media desconocida

Para una población que se puede describir mediante una variable aleatoria X con distribución (μ, σ) , se establecerá el nivel de confianza $1 - \alpha$ ($0 < \alpha < 1$).

El estimador insesgado de σ^2 será:

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Se utilizará la siguiente aproximación:

$$\frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2_{n-1}$$

Hay que tener en cuenta que la distribución de χ^2_n no es simétrica, por lo que se tiene:

$$\chi^2_{n-1; \alpha/2} \neq \chi^2_{n-1; 1-\alpha/2}$$

El enunciado probabilístico quedaría como sigue:

$$P\left[\chi^2_{n-1; 1-\alpha/2} < \frac{(n-1)\hat{S}^2}{\sigma^2} < \chi^2_{n-1; \alpha/2}\right] = 1 - \alpha$$

Cuando se despeja de esta ecuación el valor de σ^2 :

$$P \left[\frac{(n-1) \hat{S}^2}{\chi^2_{n-1, \alpha/2}} < \sigma^2 < \frac{(n-1) \hat{S}^2}{\chi^2_{n-1, 1-\alpha/2}} \right] = 1 - \alpha$$

El intervalo probabilístico cuyos extremos son variables aleatorias será:

$$\left[\frac{(n-1) \hat{S}^2}{\chi^2_{n-1, \alpha/2}} ; \frac{(n-1) \hat{S}^2}{\chi^2_{n-1, 1-\alpha/2}} \right]$$

La probabilidad de que se tomen valores entre los que esté comprendido el valor verdadero de la varianza es $1 - \alpha$.

3.2.3. Intervalos de confianza para la diferencia entre dos medias

Sea $X_{11}, X_{12}, \dots, X_{1n}$ una muestra aleatoria de n_1 ensayos tomados de una primera población con un valor esperado μ_1 y una varianza de σ_{21} ; y sea $X_{21}, X_{22}, \dots, X_{2n}$ una muestra aleatoria de n_2 observaciones tomadas de la segunda población cuyo valor esperado es μ_2 y con una varianza de σ_{22} . Siendo \bar{X}_1 y \bar{X}_2 las medias muestrales, la estadística de $\bar{X}_1 - \bar{X}_2$ será un estimador puntual de $\mu_1 - \mu_2$, y si las dos poblaciones son normales tendrá una distribución normal, o cercana a normal, si se cumplen las condiciones del TLC (teorema del límite central); por lo que:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Si se desea calcular el intervalo de confianza para la diferencia de dos medias, hay que saber si las varianzas poblacionales son conocidas o no; y en el caso de que se trate de varianzas desconocidas, se deberá probar si son diferentes o iguales.

3.2.3.1. Varianzas conocidas

Si se tienen varianzas poblacionales conocidas pero que son diferentes, los pasos que se deben seguir para averiguar el intervalo de confianza son:

1. El estadístico que se emplee como estimador puntual de la diferencia de las dos medias $\mu_1 - \mu_2$ será $T = \bar{X}_1 - \bar{X}_2$.
2. La variable aleatoria asociada con el estimador será la **variable normal** estándar definida por lo siguiente:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

3. Para llevar a cabo el cálculo del intervalo de confianza, hay que tener en cuenta el nivel de confianza que se pretende considerar.

Teorema: si se tienen las medias de dos muestras aleatorias independientes $\bar{X}_1 - \bar{X}_2$, con un tamaño de muestra n_1 y n_2 , que han sido tomadas de poblaciones con una varianza conocida σ_1 y σ_2 , respectivamente, el intervalo de confianza para $\mu_1 - \mu_2$ será:

$$\bar{X}_1 - \bar{X}_2 - Z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + Z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3.2.3.2. Varianzas desconocidas, pero iguales entre sí

En el caso en que las varianzas sean desconocidas, se hace necesaria la realización previa de una prueba estadística para que se verifique si las varianzas serán iguales o distintas.

Para ello, hay que emplear la distribución F , ya sea a través del cálculo de probabilidad de que una muestra que se toma venga de dos poblaciones en las que las varianzas son iguales, o bien a través del uso de un intervalo de confianza para la relación de dos varianzas.

Al desconocerse las varianzas de la población, se utilizarán las varianzas de las muestras a modo de estimadores.

A continuación se describe el procedimiento que se debe seguir para calcular el intervalo de confianza para la diferencia de dos medias:

1. El estadístico que se emplee como estimador puntual de la diferencia de las dos medias $\mu_1 - \mu_2$ será $\bar{X}_1 - \bar{X}_2$.
2. La variable aleatoria asociada con el estimador será la variable que se define de la siguiente manera (se empleará t para el caso de muestras pequeñas):

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Donde S_p será el estimador combinado de las S^2 , que se define así:

$$S_p = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

En el cálculo del intervalo de confianza hay que considerar el nivel de confianza que se quiere determinar y los grados de libertad calculados:

$$\text{Grados de libertad} = n_1 + n_2 - 2$$

Al operar en la expresión anterior, se llegará al teorema siguiente, que definirá el intervalo de confianza para la diferencia de dos medias con varianzas desconocidas pero iguales:

Siendo $\bar{X}_1, \bar{X}_2, s_1^2, s_2^2$ las medias y varianzas de dos muestras aleatorias de tamaños n_1 y n_2 respectivamente, que se toman de dos poblaciones normales y que son independientes, cuyas varianzas son desconocidas pero iguales entre sí, entonces el intervalo de confianza para la diferencia de medias será:

$$\bar{X}_1 - \bar{X}_2 - t S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

3.2.3.3. Varianzas desconocidas, con tamaño de muestra grande

Para este caso, el estimador puntual de la diferencia de medias ($\mu_1 - \mu_2$) que se empleará será $\bar{X}_1 - \bar{X}_2$, que será suficiente.

La variable t será la variable aleatoria que estará asociada con el estimador, y se definirá de la siguiente forma:

$$z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

El intervalo de confianza estará determinado por un teorema que está basado en la distribución que tiene t con n grados de libertad. Dicho teorema dice:

Si $\bar{X}_1, \bar{X}_2, s_1^2, s_2^2$ son las medias y las varianzas de dos muestras aleatorias con tamaños n_1 y n_2 , que se han recogido de dos poblaciones normales y que son independientes con varianzas desconocidas y diferentes, entonces un intervalo de confianza para la diferencia de dos medias $\mu_1 - \mu_2$ será:

$$\bar{X}_1 - \bar{X}_2 - t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Para el cálculo de los grados de libertad se procederá de la siguiente manera:

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2 / n_1 - 1}{\left(\frac{S_2^2}{n_2} \right)^2 / n_2 - 1}}$$

Como el valor que se obtiene de v suele ser un número decimal, se suele redondear al entero menor más próximo; por ejemplo, si se obtiene un valor de 13.9, se deberá redondear a 13 (no a 14).

3.2.4. Intervalo de confianza para el cociente entre dos varianzas

Se considerarán dos poblaciones normales y que son independientes entre sí y cuyas varianzas son desconocidas, determinadas por σ_1^2, σ_2^2 respectivamente. Se pretende calcular un intervalo de nivel $1 - \alpha$ para el cociente de las dos varianzas σ_1^2 / σ_2^2

Se tomará una muestra aleatoria de tamaño n_1 de una de las dos poblaciones y una muestra de un tamaño n_2 de la otra población.

Si S_1^2, S_2^2 son las respectivas varianzas muestrales de cada población, se considerará el estadístico:

$$F = \frac{S_2^2 / \sigma_2^2}{S_1^2 / \sigma_1^2}$$

Hay que destacar que F contiene al parámetro de interés σ_1^2 / σ_2^2 , ya que F se define así:

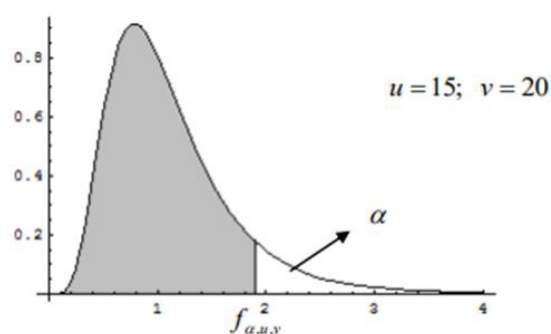
$$F = \frac{S_2^2 \cdot \sigma_1^2}{S_1^2 \cdot \sigma_2^2}$$

F tendrá una distribución denominada Fisher con $n_1 - 1$ y $n_2 - 1$ grados de libertad.

Siendo X una variable aleatoria continua, se dirá que tiene una distribución de Fisher con u grados de libertad en el numerador y v grados de libertad en el denominador si su función de probabilidad es de la siguiente forma:

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right) \left(\frac{u}{v}\right)^{\frac{u}{2}} x^{\frac{u}{2}-1}}{\Gamma\left(\frac{u}{2}\right) \Gamma\left(\frac{v}{2}\right) \left[\left(\frac{u}{v}\right)x + 1\right]^{\frac{u+v}{2}}} \quad 0 < x < \infty$$

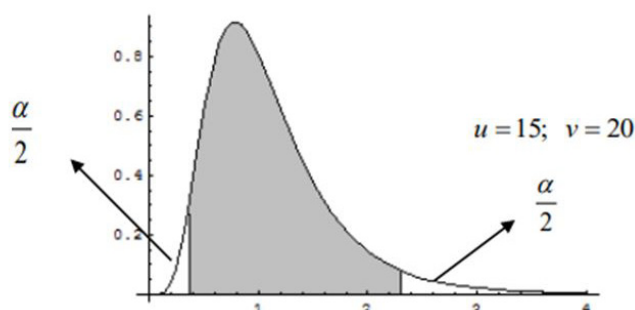
Gráficamente, la distribución de Fisher se asemeja a la gráfica de una chi-cuadrado, es decir, es asimétrica. Se anotará $f_{\alpha, u, v}$ al cuantil que tiene a su derecha un área de α bajo la curva de densidad:



Se plantea la siguiente ecuación: $P(a \leq F \leq b) = 1 - \alpha$, y se puede comprobar que la elección óptima de a y b es:

$$a = f_{1-\alpha/2, n_2-1, n_1-1}$$

$$b = f_{\alpha/2, n_2-1, n_1-1}$$



Así pues, se tendrá que:

$$P\left(f_{1-\alpha/2, n_2-1, n_1-1} \leq \frac{S_2^2}{S_1^2} \leq f_{\alpha/2, n_2-1, n_1-1}\right) = 1 - \alpha$$

Al despejar el cociente σ_1^2 / σ_2^2 , la expresión resultante será:

$$P\left(\frac{S_1^2}{S_2^2} f_{1-\alpha/2, n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} f_{\alpha/2, n_2-1, n_1-1}\right) = 1 - \alpha$$

Se puede concluir que el intervalo de confianza para el cociente entre dos varianzas conocidas de dos poblaciones normales e independientes será:

$$\left[\frac{S_1^2}{S_2^2} f_{1-\alpha/2, n_2-1, n_1-1}; \frac{S_1^2}{S_2^2} f_{\alpha/2, n_2-1, n_1-1} \right]$$

3.3. Contraste de hipótesis

De manera genérica, la investigación científica siempre ha intentado elegir la hipótesis más sencilla que explicara la realidad que se estuviese observando. La razón es que es más fácil descubrir las deficiencias y contrastar de manera empírica una hipótesis sencilla, lo cual implica que se aprenda de los datos de una manera más rápida y segura.

Dicho principio explica por qué varias investigaciones estadísticas tienen como objetivo comparar una hipótesis simplificada, como por ejemplo: una población es igual que otra que se tome como referencia (el efecto de un fármaco que acaba de salir al mercado es igual que el de otro que ya existía previamente, etc.), dos o más poblaciones son iguales entre sí, etc.

Una hipótesis se compara cuando se contrastan sus predicciones con la realidad: si son coincidentes, dentro de un margen de error que se admita, se mantendrá la hipótesis; por el contrario, si no son coincidentes, habrá que rechazarla.

Una hipótesis estadística se puede definir como una afirmación que se ha hecho sobre ciertas características de una determinada población (por ejemplo, que el tiempo de vida medio de una pila es X horas, que un fármaco calma el dolor, etc.). Un contraste o test de hipótesis es el procedimiento para aceptar o rechazar dicha afirmación o hipótesis.

Los test de hipótesis surgieron alrededor de 1925, con la publicación del libro de Ronald Fisher *Métodos estadísticos para investigadores*. En él se definían los contrastes de significación y se decía que su funcionamiento seguía el esquema del falsacionismo.

Test de significación (NHST)

Un test de significación consta de una hipótesis H_0 , denominada hipótesis nula, en la que se establece que el valor de un parámetro llamado θ es un valor concreto θ_0 :

$$H_0: \theta = \theta_0$$

La hipótesis de la que parte el investigador se definió con este nombre en honor a Fisher porque su significado era que no existía ningún cambio al emplear un fertilizante nuevo, con lo que su efecto era nulo. Es decir, si se quería demostrar que un fertilizante nuevo tenía efecto, se debía suponer que no lo tenía, y tratar de contradecir esta afirmación. Las siglas NHST son el acrónimo de *null hypothesis significance testing*.

Así pues, una vez que se ha delimitado la hipótesis nula que se quiere poner a prueba, la manera general de seguir consiste en elegir una muestra de esa población y estudiar si los resultados de la muestra elegida son coherentes con la afirmación que se ha realizado. En resumen, la evidencia que proporcione la muestra que se ha tomado deberá ser lo bastante sólida como para poder decidir si se acepta o no la hipótesis nula.

Para constatar la fuerza de la evidencia o la coherencia de los resultados, se deberá estudiar la diferencia entre lo que se observa en la muestra y lo que dice la hipótesis nula. Para esto hay que elegir el denominado estadístico pivote (estadístico T) del test y calcular su valor sobre los datos de la muestra que se observa (x_1, x_2, \dots, x_n), lo que se denota como $T(x_1, x_2, \dots, x_n)$. Como la distribución en el muestreo del estadístico pivote se debe conocer, se determinará la probabilidad de que el estadístico tenga un valor más extremo o un valor igual que el que se observa (x_1, x_2, \dots, x_n), suponiendo que la hipótesis nula sea cierta.

Simbólicamente, se escribirá:

$$P(T \geq T(x_1, x_2, \dots, x_n) / H_0)$$

Y ese número se denominará *p*-valor.

Cuando dicho *p*-valor es extremadamente pequeño (generalmente se toma como muy pequeño cuando es menor de 0.05), se dice que el test tiene un resultado significativo, ya que se puede descartar la hipótesis nula. En cualquier otro caso, el resultado del test no es significativo y no se puede rechazar dicha hipótesis.

De acuerdo con esto, se supone que, al trabajar con un nivel de significación que sea del 5 %, en promedio 5 de cada 100 veces que la hipótesis nula sea cierta se rechazará por azar. Es decir, cinco de cada cien veces que se rechace la hipótesis se estará cometiendo un error, ya que se estará admitiendo que es cierta cuando en realidad la prueba no justifica ni su certeza ni su falsedad.

La hipótesis nula únicamente se rechazará si, al observar una muestra como la que se da, la probabilidad es extremadamente baja. Es decir, la hipótesis nula tiene que ser rechazada si la muestra es muy distinta cuando la hipótesis nula es aceptada. Este razonamiento científico está basado en la disyunción lógica: "o bien ha ocurrido un suceso excepcional (muy improbable) o bien la hipótesis nula no es correcta".

La probabilidad de significación, también denominada p -valor, sirve como una suerte de evidencia en oposición a la hipótesis nula: cuanto más bajo sea, más evidencias habrá en su contra. Un valor muy bajo determina que la muestra que se observa se aleja mucho más de lo que se esperaría si solo se tuviese en cuenta el azar, las circunstancias del muestreo aleatorio. Por lo tanto, el investigador está frente a una hipótesis nula descartable o inverosímil.

Fisher describió los test de significación como un estudio para descartar la hipótesis nula, que de ninguna manera podía ser establecida o probada definitivamente. Dicho planteamiento refutacionista fue coherente con la corriente falsacionista. Fisher hizo una propuesta metodológica que era una especie de falsacionismo que se aplicaba a la estadística: "se trata de rechazar aquellas hipótesis para las cuales las observaciones sean relativamente inverosímiles".

La teoría de Neyman-Pearson

En esta teoría de Neyman-Pearson se empleaba el p -valor y el NHST de Fisher como parte de un proceso de decisión formal. Así, se planteó una elección verdadera entre dos hipótesis contrarias. El contraste de hipótesis se transformó en un método para decidir entre ambas hipótesis: la hipótesis alternativa H_1 y la hipótesis nula H_0 .

En cualquier contraste de hipótesis, el objetivo es rechazar o aceptar la hipótesis nula que se plantea (aceptando la hipótesis alternativa en caso de rechazar la hipótesis nula). No obstante, se pueden dar los siguientes casos:

- Se acepta la hipótesis nula siendo apta. Esta será una decisión correcta.
- Se rechaza la hipótesis nula siendo no apta. Esta será una situación correcta.
- Se rechaza la hipótesis nula siendo apta. En este caso se cometerá un error, también llamado error de tipo I. La probabilidad de que se cometa este error vendrá dada por el nivel de significación α que se fija en un principio.
- Se acepta la hipótesis nula siendo no apta. En este caso se cometerá un error, llamado error de tipo II. La probabilidad de que se cometa este error se representará por β , y a la probabilidad $1 - \beta$ se la denominará potencia del contraste, que medirá la probabilidad de que se rechace la hipótesis nula cuando sea falsa.

Pearson y Neyman justificaron que, en muchas ocasiones, cuando se ha fijado la probabilidad α de error de tipo I, es decir, cuando se ha acotado el tanto por ciento de veces que se tomará una decisión errónea al no aceptar la hipótesis nula cuando es verdadera, es posible utilizar y construir contrastes de potencia máxima, esto es, contrastes que minimizan el error de tipo II o la probabilidad de tipo β , por lo que maximizan la denominada potencia del test: su capacidad o sensibilidad para determinar que la hipótesis nula es errónea.

3.3.1. Contraste de una media

3.3.1.1. Varianza conocida

Si se considera una población normal, para hacer este contraste se tomará como estadístico la media muestral:

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Como su distribución es conocida, el estadístico de contraste vendrá dado por lo siguiente:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$$

Se pueden hacer 3 tipos de contrastes; se supondrá que la hipótesis nula es cierta, y se rechazará en los siguientes casos:

a. $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$

Se rechazará la hipótesis nula si:

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

b. $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$

Se rechazará la hipótesis nula si:

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha}$$

c. $H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$

Se rechazará la hipótesis nula si:

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha}$$

Para el caso contrario, se aceptará la hipótesis nula.

3.3.1.2. Varianza desconocida

- Con $n > 30$

Si se considera una población normal, para hacer este contraste se tomará como estadístico la media muestral:

$$\bar{X} \rightarrow N\left(\mu, \frac{s}{\sqrt{n}}\right)$$

Como su distribución es conocida, el estadístico de contraste vendrá dado por lo siguiente:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \rightarrow N(0, 1)$$

Se pueden hacer 3 tipos de contrastes; se supondrá que la hipótesis nula es cierta, y se rechazará en los siguientes casos:

a. $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$

Se rechazará la hipótesis nula si:

$$\left| \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right| > z_{\alpha/2}$$

b. $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$

Se rechazará la hipótesis nula si:

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > z_{\alpha}$$

c. $H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$

Se rechazará la hipótesis nula si:

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} < -z_{\alpha}$$

Para el caso contrario, se aceptará la hipótesis nula.

- Con $n < 30$

Si se considera una población normal, para hacer este contraste se tomará como estadístico la media muestral:

$$\bar{X} \rightarrow t_{n-1}$$

Como su distribución es conocida, el estadístico de contraste vendrá dado por lo siguiente:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \rightarrow t_{n-1}$$

Se pueden hacer 3 tipos de contrastes; se supondrá que la hipótesis nula es cierta, y se rechazará en los siguientes casos:

a. $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$

Se rechazará la hipótesis nula si:

$$\left| \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right| > t_{n-1; \alpha/2}$$

b. $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$

Se rechazará la hipótesis nula si:

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > t_{n-1; \alpha}$$

c. $H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$

Se rechazará la hipótesis nula si:

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} < -t_{n-1; \alpha}$$

Para el caso contrario, se aceptará la hipótesis nula.

3.3.2. Contraste de una varianza

3.3.2.1. Media conocida

Si se considera X_1, X_2, \dots, X_n una muestra aleatoria de una población $X \rightarrow \mathbb{N}(\mu, \sigma)$ y con una media conocida μ , el estadístico de contraste será:

$$\frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma_0^2}$$

Se supondrá que la hipótesis nula es cierta, y se rechazará en los siguientes casos:

a. $H_0: \sigma_2 = \sigma_0^2$ $H_1: \sigma_2 \neq \sigma_0^2$

Se rechazará la hipótesis nula si:

$$X^2 \leq X_{\alpha/2, n}^2 \text{ o bien, } X^2 \geq X_{1-\alpha/2, n}^2$$

b. $H_0: \sigma_2 = \sigma_0^2$ $H_1: \sigma_2 > \sigma_0^2$

Se rechazará la hipótesis nula si:

$$X^2 \geq X_{1-\alpha, n}^2$$

c. $H_0: \sigma_2 = \sigma_0^2$ $H_1: \sigma_2 < \sigma_0^2$

Se rechazará la hipótesis nula si:

$$X^2 \leq X_{\alpha, n}^2$$

3.3.2.2. Media desconocida

Si se considera una población normal, para hacer este contraste se tomará como estadístico la varianza muestral:

$$\frac{(n-1)s^2}{\sigma_2} \rightarrow \chi_{n-1}^2$$

La distribución del estadístico en este caso no es simétrica, por lo que se podrán estudiar 3 tipos de contrastes diferentes, pero se tendrá en cuenta esa no simetría. Se supondrá que la hipótesis nula es cierta, y se rechazará en los siguientes casos:

a. $H_0: \sigma_2 = \sigma_0^2$ $H_1: \sigma_2 \neq \sigma_0^2$

Se rechazará la hipótesis nula si:

$$\frac{(n-1)s^2}{\sigma_0^2} \notin (X_{\alpha/2}^2; X_{1-\alpha/2}^2)$$

b. $H_0: \sigma_2 \leq \sigma_0^2$ $H_1: \sigma_2 > \sigma_0^2$

Se rechazará la hipótesis nula si:

$$\frac{(n-1)s^2}{\sigma_0^2} < \chi_{\alpha}^2$$

$$c. H_0: \sigma_2 \geq \sigma_0^2 \quad H_1: \sigma_2 < \sigma_0^2$$

Se rechazará la hipótesis nula si:

$$\frac{(n-1)s^2}{\sigma_0^2} < \chi_{1-\alpha}^2$$

Para el caso contrario, se aceptará la hipótesis nula.

3.3.3. Contraste de dos medias

3.3.3.1. Varianzas conocidas

En este caso, se trata de hacer una comparación de dos muestras que se observan (en distintos sujetos) y que proceden de poblaciones equivalentes en relación con una determinada variable; por ejemplo, comparar las notas de una asignatura en concreto de los alumnos de una clase.

El contraste de medias en grupos independientes se divide en dos supuestos: normalidad y homocedasticidad. La normalidad implica que las muestras que se estudian provienen de una población normal; por su parte, la homocedasticidad quiere decir que las dos poblaciones normales tendrán la misma varianza. Cabe destacar que la condición de normalidad en el caso de muestras grandes (esto es, con $n \geq 30$) no es necesaria, y aunque respecto a la homocedasticidad las condiciones son más estrictas, tampoco resultan demasiado exigentes, por lo que los contrastes de medias son pruebas muy firmes. Aun así, si se quieren dar conclusiones rigurosas, cuando no se cumpla la condición de normalidad y la muestra sea pequeña, se podrá aplicar la prueba de Mann Whitney; y en el caso de que no se cumpla la condición de homocedasticidad, se trate o no de una muestra grande, la *t* de Welch.

Si se observan dos medias y se quiere saber si proceden de poblaciones iguales en relación con la variable que se está estudiando, se comienza partiendo de la hipótesis de que ambas poblaciones sí son iguales. El hecho de partir de la hipótesis de igualdad como hipótesis nula y no desde la desigualdad (hipótesis alternativa) se debe a que la hipótesis nula tiene un planteamiento muy sencillo, en contra de la hipótesis alternativa, que tiene bastantes más planteamientos.

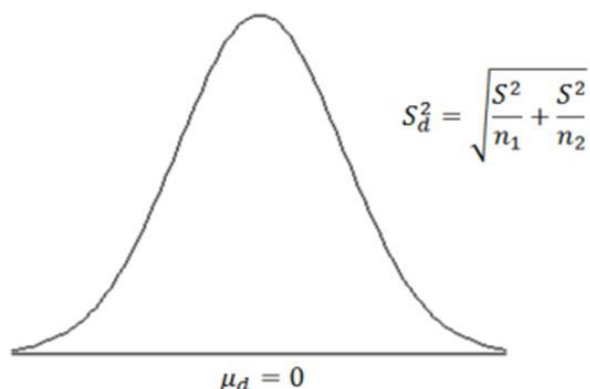
Dadas dos poblaciones iguales en varianzas y medias, si se toman pares de muestras de las dos poblaciones y se calculan las medias, se obtendrá una distribución muestral de diferencias de medidas, cuya media valdrá 0, debido a que ambas poblaciones tienen la misma media, y es de esperar que sus datos se encuentren alrededor del 0; respecto a la desviación típica, quedaría definida por lo siguiente:

$$\sigma_d^2 = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

Si se supone que las varianzas de las dos poblaciones son iguales, se podrá estimar la varianza σ^2 por S^2 , que será la varianza ponderada que se obtiene de las dos varianzas muestrales:

$$\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

Gráficamente, se tendrá:

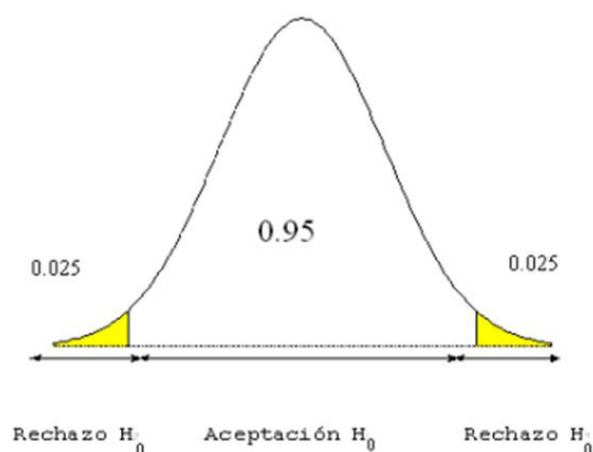


Las dos hipótesis posibles son:

- $H_0: \mu_1 = \mu_2$ (las dos medias que se observan proceden de poblaciones con medias iguales entre sí).
- $H_1: \mu_1 \neq \mu_2$ (las dos medias que se observan proceden de poblaciones con medias diferentes).

Al operar con el valor de $\alpha = 0.05$ (o $\alpha = 0.01$), se definen las dos zonas en la distribución muestral de diferencias de medias. Gráficamente, se diferencian dos zonas: la zona central, donde se encuentra el 95 % de los casos y que, además, es la zona de aceptación de la hipótesis nula; y la otra zona, que corresponde al 5 % de los casos, y que es la de la hipótesis alternativa, o zona donde se rechaza la hipótesis nula.

Cuando la diferencia de medias que se observa se encuentra en la zona de la hipótesis nula, se acepta dicha hipótesis con una probabilidad del 95 % de acierto; si, por el contrario, se encuentra en la zona de la hipótesis alternativa, resulta más exacto anunciar que las dos medias proceden de poblaciones distintas, aunque se asuma un riesgo del 5 % como máximo de equivocarse.



Simplemente se trata de conocer cuál es la región en la que se encuentra la diferencia de medias que se observan. Para esto, se tipificará la variable y se calculará el número de desviaciones tipo a que se encuentra esta diferencia de la media de la distribución muestral, que tendrá un valor de 0:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}}$$

Se situará dicha diferencia de medias en la distribución muestral de diferencias de medias que proceden de las mismas poblaciones y se calculará la probabilidad de que pertenezcan a la misma población. Si esa probabilidad es alta, es decir, mayor o igual al 5 %, se aceptará la hipótesis nula; en caso contrario, se rechazará.

3.3.3.2. Varianzas desconocidas, pero iguales entre sí

Para el estudio de contraste para la diferencia de dos medias, denominadas μ_1 y μ_2 , de poblaciones normales independientes, cuyas varianzas poblacionales son desconocidas pero iguales entre sí, es decir, $\sigma_1^2 = \sigma_2^2$, se definirá N_1 como el tamaño de la muestra de la población 1 y N_2 como el tamaño de la muestra de la población 2. Así pues, se tiene:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}}$$

La hipótesis nula, H_0 , será $\mu_1 - \mu_2 = \Delta_0$.

Existen 3 posibles casos de hipótesis alternativa, H_i :

- Si la hipótesis alternativa es $\mu_1 - \mu_2 < \Delta_0$, la región crítica vendrá dada por $(-\infty, t_\alpha)$
- Si la hipótesis alternativa es $\mu_1 - \mu_2 \neq \Delta_0$, la región crítica vendrá dada por $(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$.
- Si la hipótesis alternativa es $\mu_1 - \mu_2 > \Delta_0$, la región crítica vendrá dada por (t_α, ∞) .

3.3.3.3. Varianzas desconocidas, con tamaño de muestra grande

Para el estudio de contraste para la diferencia de dos medias, denominadas μ_1 y μ_2 , cuyas varianzas poblacionales σ_1^2 y σ_2^2 son desconocidas para muestras aleatorias independientes y con tamaño de muestra grande, se definirá N_1 como el tamaño de la muestra de la población 1 y N_2 como el tamaño de la muestra de la población 2. Así pues, se tiene:

$$Z \approx \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \sim N(0, 1)$$

La hipótesis nula, H_0 , será $\mu_1 - \mu_2 = \Delta_0$.

Existen 3 posibles casos de hipótesis alternativa, H_1 :

- Si la hipótesis alternativa es $\mu_1 - \mu_2 < \Delta_0$, la región crítica vendrá dada por $(-\infty, z_\alpha)$.
- Si la hipótesis alternativa es $\mu_1 - \mu_2 \neq \Delta_0$, la región crítica vendrá dada por $(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$.
- Si la hipótesis alternativa es $\mu_1 - \mu_2 > \Delta_0$, la región crítica vendrá dada por (z_α, ∞) .

3.3.4. Contraste de dos varianzas

En esta situación hay que distinguir dos casos: que las medias sean conocidas o que sean desconocidas.

- Con medias conocidas:

Si se consideran X_1, X_2, \dots, X_{n_1} , una muestra aleatoria de una población $X \rightarrow N(\mu_1, \sigma_1)$, y Y_1, Y_2, \dots, Y_{n_2} , una muestra aleatoria de una población $Y \rightarrow N(\mu_2, \sigma_2)$, y se supone que ambas poblaciones son independientes, el estadístico de contraste será:

$$F = \frac{\sum_{i=1}^{n_1} (x_i - \mu_1)^2 / n_1}{\sum_{i=1}^{n_2} (y_i - \mu_2)^2 / n_2}$$

Se supondrá que la hipótesis nula es cierta, y se rechazará en los siguientes casos:

a. $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$

Se rechazará la hipótesis nula si:

$$F \leq \frac{1}{f_{1-\alpha/2, n_2, n_1}} \text{ o bien } F \geq f_{1-\alpha/2, n_1, n_2}$$

b. $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$

Se rechazará la hipótesis nula si:

$$F \geq f_{1-\alpha/2, n_1, n_2}$$

c. $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$

Se rechazará la hipótesis nula si:

$$F \leq \frac{1}{f_{1-\alpha/2, n_2, n_1}}$$

- Con medias desconocidas:

Si se consideran X_1, X_2, \dots, X_{n_1} , una muestra aleatoria de una población $X \rightarrow N(\mu_1, \sigma_1)$ y Y_1, Y_2, \dots, Y_{n_2} , una muestra aleatoria de una población $Y \rightarrow N(\mu_2, \sigma_2)$ y se supone que ambas poblaciones son independientes, el estadístico de contraste será:

$$F = \frac{S_1^2}{S_2^2}$$

Se supondrá que la hipótesis nula es cierta, y se rechazará en los siguientes casos:

– $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$

Se rechazará la hipótesis nula si:

$$F \leq \frac{1}{f_{1-\alpha/2, n_2-1, n_1-1}} \text{ o bien } F \geq f_{1-\alpha/2, n_1-1, n_2-1}$$

– $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$

Se rechazará la hipótesis nula si:

$$F \geq f_{1-\alpha/2, n_1-1, n_2-1}$$

$$- H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 < \sigma_2^2$$

Se rechazará la hipótesis nula si:

$$F \leq \frac{1}{f_{1-\alpha/2, n_2-1, n_1-1}}$$

3.3.5. Contraste del parámetro de una Binomial

Se supone que X es una variable aleatoria con una distribución de probabilidad binomial con parámetro 1 y π tal que $X \rightarrow B(1, \pi)$, de la cual se extraerá una muestra aleatoria X_1, X_2, \dots, X_n de tamaño n . Se llamará p a la proporción poblacional. Se pretende comprobar si el parámetro π puede ser igual a un valor π_0 , es decir, se quiere resolver uno de los siguientes contrastes:

- Contraste bilateral:

$$H_0: \pi = \pi_0 \quad H_1: \pi \neq \pi_0$$

- Contrastes unilaterales:

$$\begin{cases} H_0: \pi \geq \pi_0 \\ H_1: \pi < \pi_0 \end{cases} \quad \begin{cases} H_0: \pi \leq \pi_0 \\ H_1: \pi > \pi_0 \end{cases}$$

El contraste de hipótesis para el parámetro p (proporción de éxitos) de una distribución binomial está basado en la distribución del estadístico muestral π para un tamaño de muestra n , que debe ser suficientemente grande.

La proporción de éxitos de la muestra de una distribución binomial se simboliza por \hat{p} y la expresión del estadístico de contraste para el parámetro p es:

$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Dicha expresión seguirá una distribución normal de media 0 y desviación típica 1 bajo la hipótesis nula.

Las regiones de no rechazo para las distintas hipótesis alternativas son:

- Si $H_1: \pi \neq \pi_0$, la región de no rechazo es $(-Z_{\alpha/2}, Z_{\alpha/2})$.
- Si $H_1: \pi > \pi_0$, la región de no rechazo es $(-\infty, Z_{\alpha})$.
- Si $H_1: \pi < \pi_0$, la región de no rechazo es $(-Z_{\alpha}, \infty)$.

3.3.6. Contraste de dos porcentajes

Si se considera X_1, X_2, \dots, X_{n1} , una muestra aleatoria sencilla de $X \rightarrow \text{Bernoulli}(p_1)$, y Y_1, Y_2, \dots, Y_{n2} , una muestra aleatoria sencilla de $Y \rightarrow \text{Bernoulli}(p_2)$, y se supone que ambas poblaciones serán independientes, el estadístico de contraste será:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}$$

Donde:

$$\frac{n_1 \hat{p}_2 + n_2 \hat{p}_2}{n_1 + n_2}$$

Se supondrá que la hipótesis nula es cierta, y se rechazará en los siguientes casos:

a. $H_0: p_1 = p_2$ $H_1: p_1 \neq p_2$

Se rechazará la hipótesis nula si:

$$Z \leq z_{\alpha/2}, \text{ o bien } Z \geq z_{1-\alpha/2}$$

b. $H_0: p_1 = p_2$ $H_1: p_1 > p_2$

Se rechazará la hipótesis nula si:

$$Z \geq z_{1-\alpha}$$

c. $H_0: p_1 = p_2$ $H_1: p_1 < p_2$

Se rechazará la hipótesis nula si:

$$Z \leq z_{\alpha}$$

3.3.7. Test chi-cuadrado de bondad de ajuste

Esta prueba se emplea para decidir si un conjunto de datos se ajusta a una distribución determinada.

Se considera una muestra aleatoria de tamaño n de la distribución de una variable aleatoria X , que se encuentra dividida en k clases exhaustivas y que no son compatibles, y N_i , donde $i = 1, 2, \dots, k$ el número de observaciones en la i -ésima clase. Se establece que la hipótesis nula H_0 es:

$$H_0: F(x) = F_0(x)$$

Donde el modelo de probabilidad que se propone $F_0(x)$ está especificado de manera completa en relación con todos los parámetros.

De este modo, se puede calcular la probabilidad p_i de obtener una observación en la i -ésima clase bajo esa hipótesis nula. Además, se cumple que:

$$\sum_i p_i = 1$$

Siendo n_i la realización de N_i para $i = 1, 2, \dots, k$, de modo que:

$$\sum_i n_i = n$$

La probabilidad de que se obtengan n_i observaciones de manera exacta en la i -ésima clase será:

$$p_i^{n_i} \text{ para } i = 1, 2, \dots, k$$

Como existen k categorías excluyentes mutuamente con probabilidades p_1, p_2, \dots, p_k , se puede deducir que bajo la H_0 la probabilidad de la muestra agrupada va a ser igual a la función de probabilidad de una determinada distribución multinomial:

$$p(x_1, x_2, \dots, x_{k-1}; n, p_1, p_2, \dots, p_{k-1}) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} \cdot p_2^{x_2} \dots p_k^{x_k}$$

Donde $x_k = 1 - x_1 - \dots - x_{k-1}$; $p_k = 1 - p_1 - \dots - p_{k-1}$.

En la deducción de una prueba estadística para la hipótesis nula, H_0 , se va a considerar el valor de $k = 2$. Este será el caso de la distribución binomial, así:

$$x = n_1$$

$$p = p_1$$

$$n - x = n_2$$

$$1 - p = p_2$$

Sea Y la variable aleatoria estandarizada, esta estará definida por lo siguiente:

$$Y = \frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

Para valores de n muy grandes, la variable aleatoria se distribuye según una $N(0, 1)$; además, se sabe que al elevar al cuadrado una variable aleatoria $N(0, 1)$, se distribuirá según una chi-cuadrado con un grado de libertad; entonces se puede determinar que el estadístico:

$$\begin{aligned} \frac{(N_1 - np_1)^2}{np_1(1-p_1)} &= \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_1 - np_1)^2}{np_2} = \frac{(N_1 - np_1)^2}{np_1} + \frac{(n - N_2 - n(1-p_2))^2}{np_2} = \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} = \\ &= \sum_{i=1}^2 \frac{(N_i - np_i)^2}{np_i} \rightarrow \chi_{2-1}^2 \end{aligned}$$

Siguiendo este razonamiento, se puede llegar a demostrar que para $k \geq 2$ categorías diferentes:

$$\sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \rightarrow \chi_{k-1}^2$$

Hay que destacar que N_i será la frecuencia que se observa en la i -ésima clase y np_i será la que se espera según H_0 .

A esta estadística se la conoce con el nombre de prueba de bondad de ajuste chi-cuadrada de Pearson.

Cuando exista una coincidencia perfecta entre la frecuencia esperada y la observada, el estadístico será 0. Si la diferencia entre ambas frecuencias es grande, el estadístico también será grande. Por ello se deduce que para un determinado valor del error de tipo I, la región crítica va a estar en la parte superior de la distribución chi-cuadrado y los grados de libertad serán $k - 1$.

Una de las ventajas que presenta la prueba de bondad de ajuste chi-cuadrado es que cuando se tengan valores grandes de población, la distribución límite de chi-cuadrado de la estadística será independiente de la forma que tenga la distribución $F_0(x)$ que se proponga en la hipótesis nula. Por este motivo, se emplea también la prueba de bondad en distribuciones de probabilidad en las que $F_0(x)$ sea continua; sin embargo, se debe insistir en que dicha prueba es discreta, es decir, se comparan frecuencias observadas o esperadas para un número concreto de categorías.

De acuerdo con esto, en el caso de que $F_0(x)$ sea continua, la prueba no comparará las frecuencias observadas de manera independiente con la función de densidad que se propone en la hipótesis nula, sino que la comparación se llevará a cabo con la aproximación de la distribución continua bajo la hipótesis nula con un número concreto de intervalos de clase.

Aun así, esta prueba es un procedimiento adecuado para estudiar supuestos de normalidad siempre que el tamaño de la muestra sea grande.

Se ha comprobado que con un tamaño de muestra $n = 5$ veces el número de clases, los resultados que se obtienen son aceptables. En el caso de que alguna clase tuviese una frecuencia menor a 5, se agruparían las clases vecinas.

A no ser que se especifique una hipótesis alternativa que suponga un modelo alternativo particular $F_1(x)$, la probabilidad de encontrar un valor en la región crítica siendo falsa la hipótesis nula es muy complicada de determinar. Se ha comprobado que la potencia tiende a 1 cuando n tiende a infinito. Esto hace pensar que cuando n sea muy grande se rechazará la hipótesis nula, ya que será complicado especificar una $F_0(x)$ lo bastante cercana a la distribución.

Denominando T al estadístico del parámetro desconocido de Θ de $F_0(x)$, tanto la frecuencia observada, N_i , como la frecuencia esperada, $np_i(T)$, serán variables aleatorias en las que $p_i(T)$ determinará que la probabilidad bajo la H_0 sea función del estadístico T de Θ .

Se puede demostrar que si T es el estimador de máxima verosimilitud de Θ , entonces se cumplirá que:

$$\sum_{i=1}^k \frac{[N_i - np_i(T)]^2}{np_i(T)} \rightarrow \chi^2_{k-1-r}$$

Donde r es el número de parámetros que se intenta estimar.

3.4. Introducción a la estadística bayesiana

En general, las probabilidades se emplean de manera informal cuando se expresa incertidumbre o información que se tiene sobre observaciones de cantidades que no se conocen, pero este uso de la probabilidad también puede hacerse de modo formal.

Puede demostrarse matemáticamente que con el cálculo de probabilidades se es capaz de representar el conjunto racional de creencias de modo numérico, de manera que exista una relación directa entre información y probabilidad. A este proceso de aprendizaje inductivo a través de la regla de Bayes se le denomina inferencia bayesiana.

Se puede decir que los métodos bayesianos analizan los datos que derivan de los principios de la inferencia bayesiana; dichos métodos proporcionan:

- Estimadores de los parámetros con buenas propiedades estadísticas.
- Una descripción simple de los datos que se observan.
- Un método computacional potente para poder estimar, seleccionar y validar los modelos.

El método bayesiano consiste en tres pasos:

1. Especificación de un modelo de probabilidad que debe incluir algún tipo de conocimiento a priori acerca de los parámetros de dicho modelo.

2. Actualización del conocimiento sobre los parámetros que se desconocen condicionando dicho modelo probabilístico a los datos que se observan.
3. Evaluación del ajuste del modelo a la sensibilidad y los datos de las conclusiones a cambios en los supuestos del modelo.

La principal diferencia entre la estadística clásica y la estadística bayesiana radica en el concepto de probabilidad. En la estadística clásica es un concepto que se encuentra en la naturaleza, mientras que en la estadística bayesiana es un concepto objetivo, es decir, que se encuentra en el observador. Así, en la estadística clásica, únicamente se recaba la información de la muestra que se obtiene, suponiendo que se puede tener un tamaño infinito; en la estadística bayesiana, además de la muestra, también se tiene en cuenta la información externa que se tiene respecto a los fenómenos que se intentan modelizar.

La estadística bayesiana se basa en la probabilidad condicional, cuya fórmula es:

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

En la práctica, todas las probabilidades son condicionales desde el punto de vista de la estadística bayesiana, ya que la mayoría de las veces se tendrá cierta experiencia o un conocimiento previo sobre los sucesos.

La ley de la probabilidad total indica que si se tiene un suceso A y una **partición** B_1, B_2, \dots, B_k , se cumple que:

$$P(A) = \sum_{i=1}^k P(A | B_i) P(B_i)$$

Aplicando el teorema a variables discretas:

$$f(x) = \sum_y f(x | Y = y) P(Y = y)$$

O a variables continuas:

$$f(x) = \int f(x | y) f(y) dy.$$

Teorema del cambio de variable

Si se tiene una variable aleatoria X cuya función de densidad viene definida por f_x , y siendo g una función diferenciable y monótona cuya inversa g^{-1} también es diferenciable, entonces la función de densidad de $U = g(X)$ será:

$$f_u(u) = f_x[g^{-1}(u)] \left| \frac{d}{du} g^{-1}(u) \right|$$

En el caso de que g no sea biyectiva, se podrían encontrar subconjuntos de R en los cuales g fuese biyectiva para poder así aplicar el teorema de manera separada a cada uno de los subconjuntos.

La media y la varianza condicionadas

Si se tienen dos variables X e Y, se definirá la media y la varianza de X cuando $Y = y$ así:

$$E[X | Y = y] = \int x f(x | y) dx$$

$$\text{Var}[X | Y = y] = \int (x - E[X | Y = y])^2 f(x | y) dx$$

El siguiente teorema relaciona las esperanzas y las varianzas marginales y las condicionadas:

Se tienen dos variables X e Y , tales que:

- $E_x[X] = E_y[E_x(X | Y)]$

Demostración:

De manera general, se tiene que:

$$E(g(x)) = \int g(x) f(x) dx$$

Como $E[X | Y]$ es una función de Y :

$$\begin{aligned} E_y[E_x(X | Y)] &= \int E_x(X | y) f(y) dy = \\ \int (\int x f(x | y) d(x) f(y) dy) &= \int x (\int f(x | y) d(x) f(y) dy) dx = \\ \int x (\int f(x, y) dy) dx &= \int x f(x) dx = E_x[X] \end{aligned}$$

- $Var_x[X] = E_x[Var_x[X | Y]] + Var_y[E_x[X | Y]]$

Demostración:

De forma general, se tiene que:

$$E[E[X^2 | Y]] = E[X^2]$$

Así que:

$$E_y[Var_x[X | Y]] = E[E[X^2 | Y] - (E[X | Y])^2] = E[E[X^2 | Y]] - E[E[X | Y]^2] = E[X^2] - E[E[X | Y]^2]$$

Además:

$$Var_y[E_x[X | Y]] = E[E[X | Y]^2] - (E[E[X | Y]])^2 = E[E[X | Y]^2] - (E[X])^2$$

Sumando ambas expresiones, se obtiene:

$$E_y[Var_x[X | Y]] + Var_y[E_x[X | Y]] = E[X^2] - (E[X])^2 = Var[X]$$

3.4.1. Inferencia bayesiana

Si se tiene una hipótesis, H , sobre cualquier población, la inferencia bayesiana la actualizará una vez se hayan observado datos mediante la siguiente fórmula:

$$Pr(H|D) = \frac{Pr(Datos|H) Pr(H)}{Pr(Datos)}$$

Donde:

- $Pr(H)$ será la probabilidad a priori de que la hipótesis sea cierta.
- $Pr(H | D)$ será la probabilidad a *posteriori* de que la hipótesis sea cierta, es decir, la probabilidad de que la hipótesis sea cierta después de haber observado los datos.

- $\Pr(\text{Datos} | H)$ será la probabilidad de que se hayan observado dichos datos cuando la hipótesis es cierta, es decir, la verosimilitud.
- $\Pr(\text{Datos})$ será la verosimilitud marginal, que se define como la probabilidad de que se hayan observado los datos siendo la hipótesis cierta o no.

Pasos en la inferencia bayesiana:

Se quiere estimar un **parámetro** θ a partir de unos datos llamados $x = \{x_1, x_2, \dots, x_n\}$; teniendo en cuenta la estadística bayesiana, θ será una variable aleatoria. Los pasos a seguir son:

1. Se fija una distribución *a priori* para θ , conocida como $\pi(\theta)$.
2. Cuando se tienen los datos x , se escoge un modelo estadístico que describa su distribución; será la verosimilitud $f(x | \theta)$.
3. A partir del teorema de Bayes, se calcula la distribución a posteriori:

$$\pi(\theta | x) = \frac{f(x | \theta) \pi(\theta)}{f(x)}$$

La distribución marginal o verosimilitud marginal de los datos vendrá dada por lo siguiente:

$$f(x) = \int f(x | \theta) \pi(\theta) d\theta$$

Será una constante de integración, que asegurará que la distribución a posteriori de θ no dependerá de θ .

Dicha constante no proporcionará información adicional alguna acerca de la distribución a posteriori y estará definida por lo siguiente:

$$\pi(\theta | x) \propto f(x | \theta) \pi(\theta)$$

Dicha expresión será la distribución *a posteriori* sin normalizar, y será proporcional a la multiplicación de la verosimilitud por la distribución *a priori*.

Inferencia bayesiana en variables binarias

Se va a suponer que se lanza una moneda y se quiere calcular la probabilidad de que salga cara.

$$\Pr(X = \text{cara} | \theta) = \theta$$

$$\Pr(X = \text{cruz} | \theta) = 1 - \theta$$

Se va a imaginar que la creencia a priori sobre θ puede describirse como una distribución uniforme $U(0, 1)$, por lo que la distribución a priori de θ será:

$$\pi(\theta) = 1; \quad 0 < \theta < 1$$

Para la actualización de la distribución de θ , se lanzará 12 veces la moneda, de lo que se obtendrá 3 cruces y 9 caras. Dados estos datos, la verosimilitud será:

$$f(x | \theta) = \binom{12}{9} \theta^9 (1 - \theta)^3$$

La distribución *a posteriori* de θ será proporcional a lo siguiente:

$$\pi(\theta | x) \propto \theta^9 (1 - \theta)^3$$

Que será el núcleo de una distribución β :

$$\pi(\theta | x) = \frac{\theta^9 (1 - \theta)^3}{\int_0^1 \theta^9 (1 - \theta)^3 d\theta} = \frac{1}{B(10, 4)} \theta^{10-1} (1 - \theta)^{4-1}$$

Se puede deducir que:

$$\theta | x \sim B(10, 4)$$

Hay que recordar que la densidad de una beta $B(\alpha, \beta)$ está definida por lo siguiente:

$$\pi(\theta | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

De donde:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

Si se tiene en cuenta que la media de $\theta \sim B(\alpha, \beta)$ es:

$$E(\theta | \alpha, \beta) = \frac{\alpha}{\alpha + \beta}$$

La media *a posteriori* de θ será:

$$E[\theta | x] = \frac{10}{14} = \frac{5}{7} = \frac{1}{7} \cdot \frac{1}{2} + \frac{6}{7} \cdot \frac{9}{12}$$

Entonces:

$$E[\theta | x] = \frac{1}{7} E[\theta] + \frac{6}{7} \hat{\theta}_{MV}$$

Donde la media *a priori* de $U(0, 1)$ será $E[\theta] = \frac{1}{2}$, y $\hat{\theta}_{MV} = \frac{9}{12}$ será la estimación de máxima verisimilitud.

Se puede deducir entonces que la media *a posteriori* será una media ponderada de la media de las creencias *a priori* y de la estimación MV.

Hay que destacar que la distribución uniforme $U(0, 1)$ es un caso particular de la distribución β en los casos en que $\alpha = \beta = 1$, así que las distribuciones *a priori* y *a posteriori* serán distribuciones beta y se denominarán conjugadas.

De una manera más genérica, si se asume *a priori* una distribución $B(\alpha, \beta)$, se tendrá que:

$$\pi(\theta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

Y la verosimilitud será:

$$f(x | \theta) \propto \theta^k (1 - \theta)^{n-k}$$

La distribución *a posteriori* es:

$$B(\alpha + k, \beta + n - k)$$

$$\pi(\theta | x) \propto \theta^{\alpha+k-1} (1 - \theta)^{\beta+n-k-1}$$

Así pues, se tendrá:

$$E(\theta | x) = \frac{\alpha + k}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} E[\theta] + \frac{n}{\alpha + \beta + n} \hat{\theta}_{MV}$$

Si se toman α y β cada vez menores: $\alpha, \beta \rightarrow 0$, la media *a posteriori* de θ convergerá a $\hat{\theta}_{MV}$. Tomando dicha elección, la distribución *a priori* será:

$$\pi(\theta) \propto \frac{1}{\theta(1-\theta)}$$

Que no será una función de densidad, ya que la integral de $\pi(\theta)$ valdrá ∞ ; se dirá que se trata de una distribución *a priori* impropia.

Dicha elección *a priori* será válida, ya que dará lugar a una distribución *a posteriori* propia:

$$\theta | x \sim B(k, n - k)$$

Las distribuciones impropias permitirán *a priori* no imponer información subjetiva. No obstante, se hace importante verificar que la distribución *a posteriori* sea propia.

Inferencia bayesiana en variables normales

Se supone que se está interesado en hacer la inferencia para una **población** normal $X | \mu, \sigma^2 \sim N(\mu, \sigma^2)$ con una densidad definida por lo siguiente:

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Se recomienda trabajar con la precisión ($\Phi = 1/\sigma^2$) en lugar de hacerlo con la varianza:

$$f(x | \mu, \Phi) = \frac{\Phi^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{\Phi}{2} (x-\mu)^2\right)$$

Se tiene una muestra de datos $x = \{x_1, x_2, \dots, x_n\}$ de una variable $N(\mu, 1/\Phi)$ y una distribución *a priori* sobre (μ, Φ) , y lo que se pretende es la obtención de su distribución *a posteriori*.

Si se dan los datos de x , la verosimilitud de (μ, Φ) será:

$$f(x | \mu, \Phi) = (2\pi)^{-n/2} \Phi^{n/2} \exp\left(-\frac{\Phi}{2} [(n-1)s^2 + n(\bar{x} - \mu)^2]\right)$$

Donde s^2 vendrá definido por lo siguiente:

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

En principio, se puede emplear la distribución normal-gamma, que será conjugada y vendrá dada por lo siguiente:

$$\mu | \Phi \sim N(m, 1/\alpha\Phi)$$

$$\Phi \sim \mathcal{G}\left(\frac{a}{2}, \frac{b}{2}\right)$$

O también:

$$\pi(\mu, \Phi) = \pi(\mu | \Phi) \times \pi(\Phi) \propto \Phi^{1/2} \exp\left(-\frac{\alpha\Phi}{2} (\mu - m)^2\right) \times \Phi^{a/2-1} \exp\left(-\frac{b}{2}\Phi\right)$$

La distribución a posteriori también será una normal-gamma donde:

$$\mu | \Phi, x \sim N(m^*, 1/\alpha^* \Phi)$$

$$\Phi | x \sim \mathcal{G}\left(\frac{a^*}{2}, \frac{b^*}{2}\right)$$

Siendo:

$$m^* = \frac{\alpha m + n\bar{x}}{\alpha + n}$$

$$\alpha^* = \alpha + n$$

$$a^* = a + n$$

$$b^* = b + (n-1)s^2 + \frac{\alpha n}{\alpha + n} (m - \bar{x})^2$$

Hay que destacar que m^* será la media ponderada de la media *a priori* de μ y su estimador MV.

En la práctica, el interés radica en la distribución marginal *a posteriori* de μ , que vendrá dada por lo siguiente:

$$\mu | x \sim \mathcal{T}_a^*\left(m^*, \frac{b^*}{a^* \alpha^*}\right)$$

Donde $\mathcal{T}_a^*(\theta, \lambda)$ representa una distribución t de Student no estandarizada, de modo que:

$$\frac{\mathcal{T} - \theta}{\sqrt{\lambda}} \sim \mathcal{T}_a$$

Donde \mathcal{T}_a es una distribución t de Student estándar con α grados de libertad.

Dados los datos x , un intervalo creíble al $(1 - \alpha) \%$ para μ será:

$$m^* \pm t_{\alpha/2, a^*} \sqrt{\frac{b^*}{a^* \alpha^*}}$$

Si se considera una distribución *a priori* impropia:

$$\pi(\alpha, \beta) \propto \frac{1}{\Phi}$$

Se obtendrá que *a posteriori*:

$$\mu | \Phi, x \sim N(\bar{x}, 1/n\Phi)$$

$$\Phi | x \sim \mathcal{G}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

Cuya distribución marginal de μ a posteriori será:

$$\mu | x \sim \mathcal{T}_{n-1}\left(\bar{x}, \frac{(n-1)s^2}{(n-1)n}\right) \equiv \mathcal{T}_{n-1}\left(\bar{x}, \frac{s^2}{n}\right)$$

De este modo, un intervalo creíble al $(1 - \alpha) \%$ para μ será:

$$x \pm \mathcal{T}_{\alpha/2, n-1} \sqrt{\frac{s^2}{n}}$$

Que será equivalente al intervalo de confianza clásico para una media de una distribución normal.

Resumen

La inferencia estadística es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra, cuál es el comportamiento de una determinada población con un riesgo de error que se puede medir en términos de probabilidad.

Los métodos paramétricos de la inferencia estadística se pueden clasificar en dos tipos: métodos de estimación de parámetros y métodos de contraste de hipótesis. Ambos métodos se basan en el conocimiento teórico de la distribución de probabilidad del estadístico muestral que se utiliza como estimador de un parámetro.

La estimación de parámetros consiste en asignar un valor concreto al parámetro o parámetros que caracterizan a la distribución de probabilidad de la población. Cuando se estima un parámetro poblacional, aunque el estimador que se utiliza posea todas las propiedades deseables, se comete un error de estimación que será la diferencia entre la estimación y el verdadero valor del parámetro. El error de estimación es desconocido, por lo que resulta imposible saber en cada caso cuál ha sido la magnitud o el signo del error; para valorar el grado de precisión asociado con una estimación puntual, se parte de dicha estimación para construir un intervalo de confianza. En resumen, un intervalo de confianza está formado por un conjunto de valores numéricos tal que la probabilidad de que este contenga al verdadero valor del parámetro puede fijarse tan grande como se desee. Esta probabilidad se denomina grado de confianza del intervalo, y la amplitud de este constituye una medida del grado de precisión con el que se estima el parámetro.

Los métodos de contraste de hipótesis tienen como objetivo comprobar si determinado supuesto referido a un parámetro poblacional, o a parámetros análogos de dos o más poblaciones, es compatible con la evidencia empírica contenida en la muestra. Los supuestos que se establecen respecto a los parámetros se llaman hipótesis paramétricas. Para cualquier hipótesis paramétrica, el contraste se basa en establecer un criterio de decisión, que depende en cada caso de la naturaleza de la población, de la distribución de probabilidad del estimador de dicho parámetro y del control que se desea fijar *a priori* sobre la probabilidad de rechazar la hipótesis contrastada en el caso de ser esta cierta.

En todo contraste intervienen dos hipótesis. La hipótesis nula (H_0) es aquella que recoge el supuesto de que el parámetro tome un valor determinado y es la que soporta la carga de la prueba. La decisión de rechazar la hipótesis nula, que en principio se considera cierta, va en función de que sea o no compatible con la evidencia empírica contenida en la muestra. El contraste clásico permite controlar *a priori* la probabilidad de cometer el error de rechazar la hipótesis nula siendo esta cierta; dicha probabilidad se llama nivel de significación del contraste (α) y suele fijarse en el 1 %, el 5 % o el 10 %.

La proposición contraria a la hipótesis nula recibe el nombre de hipótesis alternativa (H_1) y suele presentar un cierto grado de indefinición: si la hipótesis alternativa se formula simplemente como "la hipótesis nula no es cierta", el contraste es bilateral o a dos colas; por el contrario, cuando se indica el sentido de la diferencia, el contraste es unilateral o a una sola cola.



Capítulo 4

Ajuste y regresión

Objetivos

- Explicar la relación existente entre una variable dependiente y un conjunto de variables independientes.
- Saber construir un modelo de regresión lineal.
- Calcular estimaciones puntuales de los parámetros.
- Adquirir los conocimientos necesarios para predecir futuros de la variable respuesta.

Introducción

- La correlación lineal y la regresión lineal simple son métodos estadísticos que estudian la relación lineal que existe entre dos variables. Es de suma importancia señalar algunas diferencias:
- La correlación cuantifica cómo de relacionadas están dos variables, mientras que la regresión lineal genera una ecuación que se basa en la relación que existe entre ambas variables y que permitirá predecir el valor de una a partir de la otra.

- El cálculo de la correlación entre dos variables será independiente del orden de cada variable, solo medirá la relación entre ambas. Para el caso de la regresión lineal, el modelo variará dependiendo de la variable que se considere dependiente de la otra.
- La correlación suele emplearse cuando ninguna de las variables se ha controlado, únicamente se han medido ambas y se quiere saber si existe o no relación entre ellas. Es común que en el caso de la regresión lineal sea una de las variables la que se controle y se mida la otra.
- De manera genérica, los estudios de correlación lineal preceden a la generación de modelos de regresión lineal. El primer paso es analizar si ambas variables están correlacionadas y, en tal caso, se procede a generar el modelo de regresión.

Uno de los objetivos de toda ciencia es encontrar relaciones entre los hechos que se estudian. Estas relaciones se traducen en expresiones matemáticas. De este modo, si se observan varias veces el tiempo que tarda un móvil en recorrer una distancia y su velocidad (suponiendo que tenga velocidad uniforme), los valores observados estarán relacionados, y esa relación podrá expresarse matemáticamente como $v = s/t$.

Sin embargo, existen otras variables como la inflación, los tipos de interés, la oferta y la demanda, el ahorro y la renta, etc. entre las que no cabe duda de que existe una relación, pero no hay ninguna función matemática que la verifique.

En el primer caso, tiempo y velocidad, se dirá que existe una dependencia funcional, mientras que en el segundo caso, inflación y tipos de interés, se da una dependencia estadística.

La diferencia radica en que, en el primer caso, la relación entre las variables es estricta y perfecta, y en el segundo caso, el modelo matemático al que se llegue deberá aproximar la relación entre variables de forma razonable, por lo que se debe determinar su forma y contrastar su “bondad”.

A las técnicas estadísticas empleadas para la determinación de modelos o expresiones que relacionen el comportamiento de varias variables se las denomina técnicas de regresión.

Antes de aplicar técnicas de regresión, es preciso hacer un análisis teórico que relacione las variables objeto de estudio y que dé consistencia al análisis estadístico. Este análisis es necesario, ya que es posible distinguir diferentes tipos de dependencia entre variables:

- Al azar.
- Una tercera variable influye sobre las dos variables consideradas.
- Una variable influye sobre la otra.

A la variable que se quiere predecir se la denomina dependiente o endógena, y a la variable a partir de la cual se quiere hacer la predicción se la denomina independiente, exógena o explicativa.

Cuando únicamente se utilice una variable independiente, se dirá que se está ante una regresión y correlación simple; sin embargo, si interviene más de una, la regresión o correlación será múltiple.

Hay que destacar que una relación estadística fuerte entre varias variables no implica la existencia de una relación causa-efecto entre ellas.

4.1. Ajuste lineal

En el estudio de la relación lineal que existe entre dos variables continuas, es necesario analizar ciertos parámetros que permitirán la cuantificación de esta relación. Uno de ellos es la covarianza, que cuantifica el grado de variación de dos variables aleatorias conjuntas.

$$\text{Covarianza muestral: } \text{Cov}(X,Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Siendo \bar{x} y \bar{y} la media de cada una de las variables y x_i y y_i el valor que toman las variables para la observación i .

La covarianza depende de las escalas en las que se midan las variables que se estudian, por lo que no se puede comparar entre diferentes pares de variables. Para poder compararlas se hace una estandarización de la covarianza, de modo que se generan los coeficientes de correlación.

Hay varios tipos de coeficientes de correlación; entre los más importantes están el **coeficiente de correlación** de Pearson, la Tau de Kendall y el Rho de Spearman.

Todos estos coeficientes varían entre -1 y $+1$, siendo $+1$ una correlación perfecta positiva y -1 una correlación perfecta negativa. Se emplean como tamaño del efecto o medida de fuerza de asociación, siendo:

- 0: asociación nula.
- 0.1: asociación baja.
- 0.3: asociación media.
- 0.5: asociación moderada.
- 0.7: asociación alta.
- 0.9: asociación muy alta.

Las diferencias principales entre los coeficientes de correlación son:

- La correlación de Pearson es la idónea para variables cuantitativas con distribución normal. Es más sensible a los valores extremos que las otras correlaciones.
- La correlación de Spearman se usa cuando los datos son de intervalos, ordinales, o cuando no satisfacen la condición de normalidad para variables continuas y estos datos se pueden cambiar a rangos. Se trata de un método no paramétrico.
- La correlación de Kendall es otra opción no paramétrica en la que se trabaja con rangos. Se utiliza cuando no se dispone de muchos datos y hay bastantes ligaduras, es decir, muchos de los datos ocupan la misma posición en el rango.

No solo es importante el valor del coeficiente de correlación, también hay que calcular su significancia. Únicamente en el caso en que el p-valor sea significativo se podrá aceptar la existencia de correlación, la cual tendrá la magnitud que indique el coeficiente.

El test paramétrico de significancia estadística que se emplea para hallar el coeficiente de correlación se denomina t-test. Como ocurre al trabajar con muestras, por un lado se tendrá el valor estimado (que será el coeficiente de correlación) y por otro la significancia en la consideración de la población entera. Al calcular el coeficiente de correlación entre X e Y en distintas muestras de igual población, el valor variará en función de las muestras que se empleen. Por este motivo, se hace necesario el cálculo de la significación de la correlación que se obtiene y su intervalo de confianza.

$$t = \frac{r \sqrt{N-2}}{\sqrt{1-r^2}}, \quad df = N - 2$$

En este test de hipótesis, la hipótesis nula será considerar que las variables son independientes, con lo que el coeficiente de correlación poblacional valdrá 0; mientras que la hipótesis alternativa será considerar que existe una relación, es decir, que el coeficiente de correlación poblacional es distinto de 0.

La correlación lineal que existe entre dos variables, además del valor que tenga el coeficiente de correlación y de su significancia, también tiene un tamaño de efecto asociado que se define como coeficiente de determinación R².

El modelo de regresión lineal simple consiste en explicar la relación de la variable respuesta Y con la variable explicativa X.

A partir de las técnicas de regresión de una variable Y sobre una variable X, se buscará una función que se aproxime a una nube de puntos (x_i, y_i) mediante una curva.

El modelo de regresión lineal simple tendrá la siguiente expresión:

$$Y = a + \beta X + \varepsilon$$

Donde a será la ordenada en el origen, β será la pendiente de la recta e indicará cómo varía Y al aumentar X en una unidad, y ε será una variable que incluirá un conjunto de factores, los cuales influirán en la respuesta en una pequeña cantidad, denominada error. X e Y serán las variables aleatorias.

4.1.1. Método de mínimos cuadrados

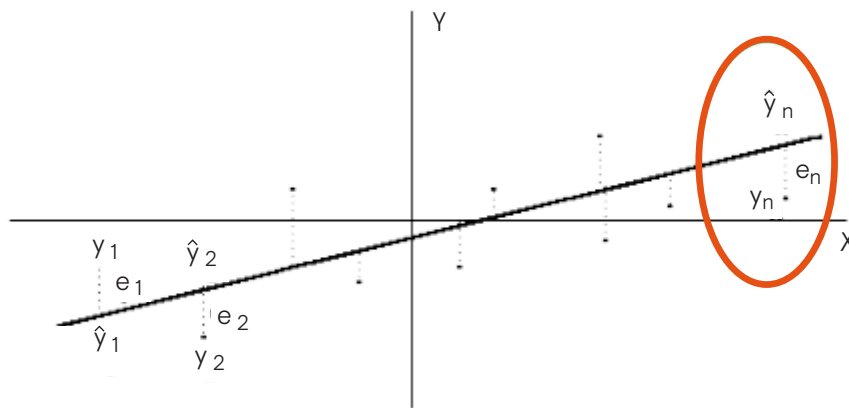
Para poder hacer una estimación del modelo de regresión lineal simple, será necesario calcular una recta de la forma:

$$\hat{Y} = \hat{a} + \hat{\beta}X = a + bX$$

De tal modo que se pueda ajustar a una nube de puntos. Para ello se empleará el método de mínimos cuadrados. Este método consiste en minimizar la suma de los cuadrados de los errores:

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

Es decir, la sumatoria de los cuadrados de las diferencias entre los valores reales que se observan (y_i) y los que se estiman (ŷ_i).



A partir de este método, las expresiones que se obtendrán para a y b serán:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{S_{XY}}{S_x^2}$$

Donde \bar{x} y \bar{y} serán las medias muestrales de X y de Y respectivamente, S_x^2 S_y^2 será la varianza muestral de X y S_{XY} será la covarianza muestral que existe entre X e Y .

Para el cálculo de estos términos se procederá de la siguiente manera:

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

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

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

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

$$S_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Al coeficiente b se le denomina coeficiente de la regresión de Y sobre X , y se denota como $b_{Y/X}$.

El cálculo de la recta de regresión de X sobre Y se efectuará mediante la aproximación de x por \hat{X} , de la siguiente forma:

$$\hat{X} = a + bY$$

De donde:

$$a = \bar{x} - b\bar{y}$$

$$b = \frac{S_{XY}}{S_y^2}$$

Para el cálculo de la recta de regresión de X sobre Y es erróneo despejar x de la ecuación:

$$Y = a + bX$$

Es preciso tener en cuenta que la recta de regresión siempre pasará por el punto (\bar{x}, \bar{y}) , es decir, por el centro de gravedad de la nube de puntos.

Como se ha dicho anteriormente, el método de los mínimos cuadrados en la obtención de los parámetros a y b consiste en minimizar la expresión:

$$M = \sum_{i=1}^n e_i^2 = \sum_{i,j} (\hat{y}_i - y_i)^2$$

Para poder simplificar el mecanismo con el que se obtiene la recta de regresión de Y sobre X, se supondrá que cada par de valores se repetirá una sola vez, y se descartarán multiplicidades. Además, en la minimización de M se tendrá en cuenta que los valores teóricos sobre la recta serán:

$$\hat{y} = a + bx_i \rightarrow M = \sum_{i=1}^n e_i^2 = \sum_{i,j} (\hat{y}_i - y_i)^2 = \sum_{i,j} (a + bx_i - y_j)^2$$

Cuando se calculan las derivadas parciales de a y b y se igualan las expresiones a 0 se obtiene:

$$\frac{\partial M}{\partial a} = 2 \sum_{i,j} (a + bx_i - y_j) = 0$$

$$\frac{\partial M}{\partial b} = 2 \sum_{i,j} (a + bx_i - y_j) x_i = 0$$

El sistema de ecuaciones que resulta es:

$$\begin{cases} \sum_{i,j} (a + bx_i - y_j) = 0 \\ \sum_{i,j} (a + bx_i - y_j) x_i = 0 \end{cases}$$

Con las propiedades del sumatorio, se obtendría:

$$\begin{cases} \sum_{i,j} (a + bx_i - y_j) = \sum_i a + b \sum_i x_i - \sum_i y_i = 0 \\ \sum_{i,j} (a + bx_i - y_j) x_i = a \sum_i x_i + b \sum_i x_i^2 - \sum_{i,j} x_i y_j = 0 \end{cases}$$

$$\begin{cases} \sum_j y_j = Na + b \sum_i x_i \\ \sum_{i,j} x_i - y_j = a \sum_i x_i + b \sum_i x_i^2 - \sum_j y_j \end{cases}$$

Al dividir las dos expresiones entre N, y suponiendo que la frecuencia absoluta de cada par sea 1, se obtiene que:

$$\begin{cases} \frac{a \sum_i 1}{N} + \frac{b \sum_i x_i}{N} - \frac{\sum_i y_i}{N} = 0 \\ a \frac{\sum_i x_i}{N} + b \frac{\sum_i x_i^2}{N} - \frac{\sum_{ij} x_i y_j}{N} = 0 \end{cases}$$

Al considerar el valor de los momentos, se obtiene:

$$a_{10} = \bar{x} = \frac{\sum_i x_i}{N}$$

$$a_{20} = \bar{x^2} = \frac{\sum_i x_i^2}{N}$$

$$a_{10} = \frac{\sum_{ij} x_i y_j}{N}$$

De donde:

$$b = \frac{a_{11} - \bar{x} \bar{y}}{a_{20} - \bar{x}^2} = \frac{m_{11}}{\sigma_x^2}$$

Como $y = a + bx$, la ecuación de la recta de regresión de Y sobre X será:

$$y - \bar{y} = \frac{m_{11}}{\sigma_x^2} (x - \bar{x})$$

El coeficiente de regresión y de correlación lineal

El coeficiente de regresión proporciona información acerca del comportamiento de la variable Y frente a la variable X, de tal manera que:

- Si $b_{Y/X} = 0$, para cualquier valor de X la variable Y será constante.
- Si $b_{Y/X} > 0$, cuando aumente el valor de X también aumentará el valor de Y.
- Si $b_{Y/X} < 0$, cuando aumente X, el valor de Y disminuirá.

El coeficiente de correlación lineal entre X e Y viene dado por la siguiente expresión:

$$r = \frac{S_{XY}}{S_X S_Y}$$

Esta expresión mide la dependencia lineal que existe entre las dos variables. El cálculo de su cuadrado se denomina coeficiente de determinación.

Las propiedades del coeficiente de correlación son:

- Es adimensional, y siempre tomará valores comprendidos entre -1 y +1.
- En el caso de variables independientes, el valor de r será 0.

- En el caso de que las variables X e Y tengan una relación lineal exacta, el valor de r será +1 si se da una relación directa, y si la relación es inversa valdrá -1.
- Si el valor de r es mayor que 0, la relación entre las variables será directa (si aumenta X, aumentará Y).
- Si el valor de r es menor que 0, la relación entre las variables será inversa (si aumenta una, disminuirá la otra).

4.1.2. Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es una prueba que mide la relación estadística que existe entre dos variables continuas. Se trata de una medida que aporta información sobre la dirección de la relación y la intensidad de esta, es decir, mide el grado de covariación entre diferentes variables que se encuentran relacionadas linealmente. En el caso de que la asociación de los elementos no sea lineal, el coeficiente no se encontrará representado de forma adecuada.

El valor del coeficiente puede variar entre -1 y +1, siendo un valor de 0 indicativo de que no existe asociación entre las variables. Si el valor es positivo, indica que a medida que aumenta el valor de una de las variables, también aumenta el de la otra. Si, por el contrario, el valor es menor que 0, indica que la asociación es negativa y si aumenta el valor de una variable, disminuye el valor de la otra.

En la representación de las relaciones de las distintas variables que se combinan de forma lineal, se puede emplear la denominada matriz de varianzas-covarianzas o la matriz de correlaciones; en la diagonal de la primera matriz se encontrarán los valores de la varianza, y en la segunda matriz, la diagonal tendrá valores de 1.

La correlación de Pearson debe cumplir una serie de características:

- La escala de medida deberá ser una escala de relación o intervalos.
- Las variables deberán ser distribuidas de manera aproximada.
- Debe haber una asociación lineal.
- No deben existir valores atípicos en los datos.

Si se tiene un conjunto de pares de observaciones, se puede representar el primer elemento del par por x y el segundo por y; así, el conjunto de las x tendrá una desviación estándar definida por la siguiente expresión:

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

Del mismo modo, la desviación estándar para y será:

$$S_y = \sqrt{\frac{\sum (y - \bar{y})^2}{N - 1}}$$

Normalizando ambas variables, se obtiene:

$$z_x = \frac{\bar{x} - \bar{y}}{s_x} ; z_y = \frac{y - \bar{y}}{s_y}$$

Así pues, el coeficiente de correlación de Pearson se define del siguiente modo:

$$r = \frac{\sum z_x z_y}{N - 1}$$

Donde $\sum z_x z_y$ es la sumatoria resultante de multiplicar las dos variables normalizadas.

Es habitual encontrar otra definición equivalente del coeficiente de correlación de Pearson:

$$r = \frac{N \sum xy - \sum x \sum y}{\sqrt{[N \sum x^2 - (\sum x)^2] \cdot [N \sum y^2 - (\sum y)^2]}}$$

Entre las ventajas del coeficiente de correlación de Pearson destacan:

- El valor será independiente de cualquier unidad que se emplee para la medida de las variables.
- En el caso de muestras grandes, la exactitud de la estimación es más probable.

Entre las desventajas destacan:

- Se hace necesario que ambas variables se midan a un nivel cuantitativo continuo.
- La distribución de las variables deberá ser similar a la curva normal.

Cuando se eleva al cuadrado el coeficiente de correlación de Pearson, el significado varía; se interpretará su valor en relación con los pronósticos. Pueden darse cuatro significados o interpretaciones:

- **Varianza asociada.** Indica la proporción de la varianza de una variable asociada a la variación de la otra.
- **Diferencias individuales.** Al multiplicar por 100 el coeficiente, indica el porcentaje de las diferencias individuales.
- **Índice de reducción del error.** El coeficiente elevado al cuadrado podría interpretarse como un índice de la reducción de error en los pronósticos, es decir, se trataría de la proporción del error cuadrático medio que se elimina usando Y en lugar de la media de Y como pronóstico. En estos casos se debe multiplicar por 100 para indicarlo como porcentaje.
- **Índice de aproximación de los puntos.** Al elevar el coeficiente al cuadrado, se indica la proximidad de los puntos a la recta de regresión; esto es, cuanto más cercano a 1 sea el valor del coeficiente, los puntos estarán más próximos a la recta de regresión.

4.1.3. Intervalos de confianza

- **Intervalo de confianza para la pendiente.**

Si se quiere obtener el intervalo de confianza para β_1 con un nivel $1 - \alpha$, como el valor de la varianza, σ^2 , será desconocido, se estimará con S_R^2 . El resultado cuando la varianza es desconocida se determina por la siguiente expresión:

$$\frac{\hat{\beta} - \beta_1}{\sqrt{\frac{S_R^2}{(n-1) S_x^2}}} \sim t_{n-2}$$

Que permite la obtención del intervalo de confianza para β_1 :

$$\hat{\beta} \pm t_{n-2, \alpha/2} \sqrt{\frac{S_R^2}{(n-1) S_x^2}}$$

Donde $t_{n-2, \alpha/2}$ es el percentil de la distribución t de Student con $n - 2$ grados de libertad, que dejará a su derecha un área $\alpha/2$.

Contraste de hipótesis para la pendiente

El test para $H_0: \beta_1 = 0$ se conoce como test de no asociación o de independencia, ya que indica si las variables son asociadas o no:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

La región de rechazo de H_0 será:

$$\left| \frac{\hat{\beta}_1}{\sqrt{\frac{S_R^2}{(n-1) S_x^2}}} \right| > t_{n-2, \alpha/2}$$

De manera equivalente, si el 0 se encuentra fuera del intervalo de confianza para β_1 de nivel $1 - \alpha$, se rechazará la H_0 a ese nivel. El p-valor de ese contraste será:

$$p\text{-valor} = 2 \Pr \left(t_{n-2, \alpha/2} > \left| \frac{\hat{\beta}_1}{\sqrt{\frac{S_R^2}{(n-1) S_x^2}}} \right| \right)$$

• **Intervalo de confianza para el intercepto.**

Si se quiere obtener el intervalo de confianza para β_0 con un nivel $1 - \alpha$, como el valor de la varianza, σ^2 , será desconocido, se estimará con S_R . El resultado cuando la varianza es desconocida se determina por la siguiente expresión:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{S_R^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2} \right)}} \sim t_{n-2}$$

Siendo el intervalo de confianza para β_0 :

$$\hat{\beta}_0 \pm t_{n-2, \alpha/2} \sqrt{S_R^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2} \right)}$$

La longitud del intervalo se hará más pequeña si:

- El tamaño de la muestra aumenta.
- La varianza de las x_i aumenta.
- La varianza residual disminuye.
- La media de las x_i disminuye.

Contraste de hipótesis para el intercepto

Si el valor verdadero de $\beta_0 = 0$, entonces la recta de regresión pasará por el origen de coordenadas. El contraste que se estudiará será:

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$

La región de rechazo de H_0 será:

$$\left| \frac{\hat{\beta}_0}{\sqrt{S_R^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2} \right)}} \right| > t_{n-2, \alpha/2}$$

En el caso de que el 0 se encuentre fuera del intervalo de confianza para β_0 con un nivel $1 - \alpha$, se rechazará la H_0 a ese nivel. El p-valor de ese contraste será:

$$p\text{-valor} = 2 \Pr \left(|t_{n-2}| > \left| \frac{\hat{\beta}_0}{\sqrt{S_R^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2} \right)}} \right| \right)$$

- **Intervalo de confianza para la varianza.**

El resultado básico será:

$$\frac{(n-2)S_R^2}{\sigma^2} \sim \chi_{n-2}^2$$

Utilizando este resultado se puede construir un intervalo de confianza para la varianza, que será:

$$\frac{(n-2)S_R^2}{\chi_{n-2, \alpha/2}^2} \leq \sigma^2 \leq \frac{(n-2)S_R^2}{\chi_{n-2, 1-\alpha/2}^2}$$

- **Intervalo de confianza para la respuesta promedio.**

Teniendo en cuenta que:

$$\text{Var}(\hat{y}_0) = \text{Var}(\bar{y}) + (x_0 - \bar{x})^2 \text{Var}(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2} \right)$$

El intervalo de confianza para la respuesta promedio será:

$$\hat{y}_0 \pm t_{n-2, \alpha/2} \sqrt{S_R^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2} \right)}$$

- **Intervalo de confianza para la predicción de una nueva respuesta.**

La varianza de la predicción de una nueva respuesta será el error cuadrático medio de la predicción, es decir:

$$E[(y_0 - \hat{y}_0)^2] = \text{Var}(y_0) + \text{Var}(\hat{y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2} \right)$$

El intervalo de confianza para la predicción de una nueva respuesta será:

$$\hat{y}_0 \pm t_{n-2, \alpha/2} \sqrt{S_R^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)S_x^2} \right)}$$

4.1.4. Predicciones

Las predicciones son una de las aplicaciones más importantes de la regresión. Una predicción consiste en la determinación, a partir del modelo ajustado, del valor de la variable dependiente para un valor dado de la variable independiente.

Pueden darse dos tipos de predicciones que estarán sujetas a niveles muy distintos de incertidumbre:

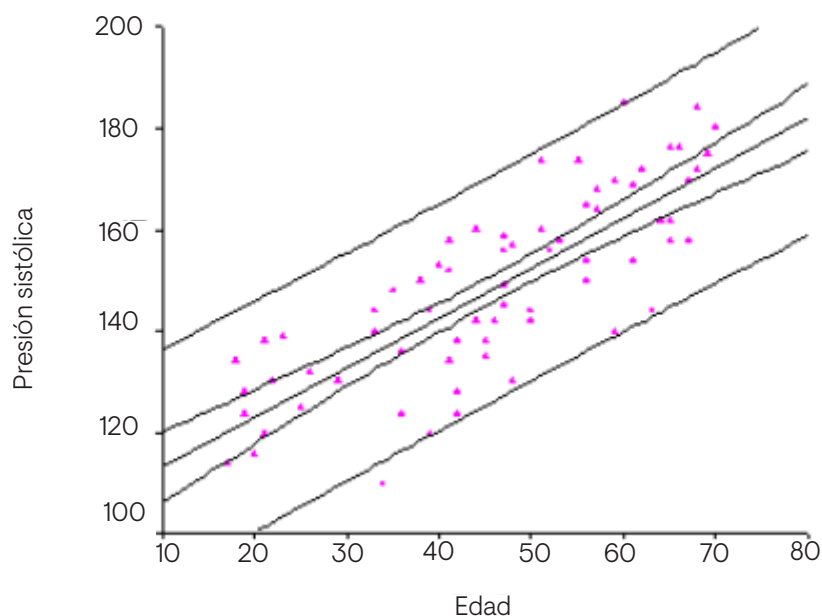
- Las interpolaciones, que son predicciones que se hacen dentro del rango de los datos que se usan para construir el modelo.
- Las extrapolaciones, que son predicciones que se hacen fuera del rango y son mucho más problemáticas, de modo que la elección del modelo es todavía más crítica, ya que distintos modelos generan con frecuencia predicciones diferentes.

En los casos de **interpolación**, los valores que se pronostiquen tendrán una fiabilidad mayor cuando se haga un buen ajuste de la recta, es decir, cuanto mayor sea el valor de R^2 , siempre y cuando exista una relación-asociación entre ambas variables.

Al verificar las hipótesis sobre las que se basa el modelo, la recta de regresión se puede emplear en la predicción del valor medio de la variable Y para cada valor concreto de la variable X . Si se calcula la esperanza matemática, se tendrá:

$$E(Y | X) = a + bX + E(\epsilon) = a + bX$$

De tal modo que la línea de regresión proporcionará un estimador del valor medio de Y para cada valor de X . Como estimador, se debe considerar la incertidumbre que se asocia a dicha recta, que se puede reflejar a través de regiones de confianza que contengan la recta. En la siguiente imagen se puede ver, superpuesta al diagrama de dispersión, la recta de regresión, además de una región de confianza para esta, la cual contiene la verdadera relación entre la presión sistólica y la edad con una fiabilidad del 95 %:



Intervalos de confianza al 95 % para la recta de regresión y para la predicción de la presión sistólica en un individuo.

En el caso de que se emplee la recta de regresión como un estimador del valor de Y en un individuo concreto, cabe esperar una incertidumbre mayor en la estimación que en la predicción de una tendencia media.

Ejemplo:

Se registran los pesos de 5 niños de diferentes edades, de modo que se obtiene la siguiente tabla:

Edad (años)	Peso (kg)
2	14
3	20
5	32
7	42
8	44

Se quiere calcular la ecuación de la recta de regresión de la edad sobre el peso y el peso aproximado que tendría un niño de 6 años.

Solución:

Se construye una tabla con los datos necesarios:

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
2	14	4	196	28
3	20	9	400	60
5	32	25	1024	160
7	42	49	1764	294
8	44	64	1936	352
$\Sigma x_i = 25$	$\Sigma y_i = 152$	$\Sigma x_i^2 = 151$	$\Sigma y_i^2 = 5320$	$\Sigma x_i \cdot y_i = 894$

El resultado de los promedios será:

$$\bar{x} = 25/5 = 5$$

$$\bar{y} = 152/5 = 30.4$$

El cálculo de la covarianza y la varianza de y será:

$$\sigma_{xy} = \frac{894}{5} - 5 \cdot 30.4 = 26.8$$

$$\sigma_y^2 = \frac{5320}{5} - (30.4)^2 = 139.84$$

La recta de regresión de la edad sobre el peso será aquella que pase por el punto (,) y cuya pendiente valga σ_{xy} / σ_y^2 :

$$x - 5 = 0.192 (y - 30.4)$$

Se despeja para obtener la recta de regresión:

$$y = \frac{x + 0.76}{0.192}$$

Para estimar el peso aproximado de un niño de 6 años, se sustituye en $X = 6$, y se obtiene:

$$y = \frac{x + 0.76}{0.192} \longrightarrow y = \frac{6 + 0.76}{0.192} \quad y = 35.21 \text{ kg}$$

4.2. Ajustes no lineales

La teoría estadística que se dedica a los ajustes lineales es bastante amplia, ya que tiene muchas aplicaciones y los resultados son fácilmente interpretables. Sin embargo, existen fenómenos que no se pueden explicar mediante modelos lineales, como por ejemplo el desarrollo de una teoría en la física o en la química; a esas situaciones se ajusta mejor un modelo no lineal.

Actualmente, gracias al avance de la informática, la aplicación de modelos no lineales en situaciones en las que los parámetros no pueden expresarse de forma lineal es mucho más viable que hace años.

En los modelos lineales, los valores o parámetros poblacionales desconocidos entran de forma lineal en la ecuación. No tiene importancia que las variables independientes se encuentren en la ecuación de forma lineal, ya que las ecuaciones serán fijas, tanto por el diseño como por observación.

Los modelos no lineales son modelos de regresión en los que los parámetros aparecerán de manera no lineal en la ecuación, por ejemplo:

$$\mu_y = \beta_0 + \beta_1 e^{-\beta_2 x}$$

4.2.1. Ajuste parabólico

Se considerará la función de ajuste $f(x, a, b, c) = a + bx + cx^2$, una parábola de segundo grado. Para la obtención de los términos a , b y c se empleará el método de los mínimos cuadrados y se minimizará la función H definida por la siguiente expresión:

$$H(a, b, c) = \sum_{i=1}^n [y_i (a + bx_i + cx_i^2)]^2$$

Hay que calcular las derivadas parciales de H respecto de a, b y c e igualarlas a 0:

$$\begin{aligned}\frac{\partial H}{\partial a} &= -2 \sum_{i=1}^n [y_i(a + bx_i + cx_i^2)] = 0 \\ \frac{\partial H}{\partial b} &= -2 \sum_{i=1}^n [y_i(a + bx_i + cx_i^2)] x_i = 0 \\ \frac{\partial H}{\partial c} &= -2 \sum_{i=1}^n [y_i(a + bx_i + cx_i^2)] x_i^2 = 0\end{aligned}$$

Operando y despejando los términos independientes, se obtendrá el sistema de ecuaciones:

$$\begin{aligned}\sum_{i=1}^n y_i &= an + b \sum_{i=1}^n x_i + c \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n y_i x_i &= a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 + c \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n y_i x_i^2 &= a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i^3 + c \sum_{i=1}^n x_i^4\end{aligned}$$

Del mismo modo, si se define:

$$\begin{aligned}y^* &= a + bx_i + cx_i^2 \\ e_i &= y_i - y^*_i\end{aligned}$$

El sistema de ecuaciones normales se podrá expresar de la siguiente forma:

$$\begin{aligned}\sum_{i=1}^n e_i &= 0 \\ \sum_{i=1}^n e_i x_i &= 0 \\ \sum_{i=1}^n e_i x_i^2 &= 0\end{aligned}$$

El valor de la varianza residual será:

$$S_c^2 = \frac{\sum_{i=1}^n y_i^2 - a \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i y_i - c \sum_{i=1}^n x_i^2 y_i}{n}$$

Ejemplo:

En la siguiente tabla se muestra el número de hectáreas por granja en un determinado país a lo largo de varios años.

Se quiere determinar el modelo cuadrático de ajuste $y = a + bt + ct^2$ que se ajusta a los datos, donde y sea el número medio de hectáreas y t el año, tomando como origen ($t = 0$) en 1950. Además, se quiere averiguar el número medio de hectáreas para el año 2000.

Año	Hectáreas
1950	213
1960	297
1970	374
1980	426
1990	461
1994	478

Solución:

Variable independiente: año en que se registran los datos (T).

Variable dependiente: número medio de hectáreas que se otorgan por granja (Y).

Se construye la tabla con los datos:

	y	t	t ²	t ³	t ⁴	ty	t ² y
	213	0	0	0	0	0	0
	298	1	1	1	1	298	298
	374	2	4	8	16	748	1496
	426	3	9	27	81	1278	3834
	461	4	16	64	256	1844	7376
	478	4,4	19,36	85,18	374,8	2103	9254
Suma	2250	14,4	49,36	185,2	728,8	6271	22258

Se reemplazan los datos en las ecuaciones:

$$\sum_{i=1}^n y_i = an + b \sum_{i=1}^n x_i + c \sum_{i=1}^n x_i^2 \longrightarrow 2250 = 6a + 14.4b + 49.36c$$

$$\sum_{i=1}^n y_i x_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 + c \sum_{i=1}^n x_i^3 \longrightarrow 6271 = 14.4a + 49.36b + 185.2c$$

$$\sum_{i=1}^n y_i x_i^2 = a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i^3 + c \sum_{i=1}^n x_i^4 \rightarrow 22258 = 49.36 a + 185.2 b + 728.8 c$$

Para resolver el sistema, se empleará el método de determinantes:

$$\Delta = \begin{vmatrix} 6 & 14.4 & 49.36 \\ 14.4 & 49.36 & 185.2 \\ 49.36 & 185.2 & 728.8 \end{vmatrix} = 1936.42$$

$$\Delta_a = \begin{vmatrix} 2250 & 14.4 & 49.36 \\ 6271 & 49.36 & 185.2 \\ 22258 & 185.2 & 728.8 \end{vmatrix} = 411279.955 \rightarrow \frac{\Delta_a}{\Delta} = \hat{a} = 212.39$$

$$\Delta_b = \begin{vmatrix} 6 & 2250 & 49.36 \\ 14.4 & 6271 & 185.2 \\ 49.36 & 22258 & 728.8 \end{vmatrix} = 185836.8704 \rightarrow \frac{\Delta_b}{\Delta} = \hat{b} = 95.97$$

$$\Delta_c = \begin{vmatrix} 6 & 14.4 & 2250 \\ 14.4 & 49.36 & 6271 \\ 49.36 & 185.2 & 22258 \end{vmatrix} = 15939.936 \rightarrow \frac{\Delta_c}{\Delta} = \hat{c} = -8.23$$

El modelo de predicción vendrá dado por lo siguiente:

$$\hat{y} = 212.39 + 95.97t - 8.23 t^2$$

Para el año 2000 ($t = 5$), el número medio de hectáreas estimadas será:

$$\hat{y} = 212.39 + 95.97 \cdot 5 - 8.23 \cdot (5)^2 = 486.49$$

4.2.2. Ajuste hiperbólico

Si se predice la variable y a través de la hipérbola que tiene una función:

$$y = a + b \left(\frac{1}{x} \right)$$

El valor esperado será:

$$\hat{y}_i = a + b \left(\frac{1}{x_i} \right)$$

El error que se cometerá estará definido por lo siguiente:

$$e_i = y_i - \hat{y}_i$$

$$e_i = y_i - a - b \left(\frac{1}{x_i} \right)$$

Empleando el método de los mínimos cuadrados, se llegará a la hipérbola que hará mínima la función:

$$S(a,b) = \sum e_i^2 = \sum \left(y_i - a - b \left(\frac{1}{x_i} \right) \right)^2$$

Se podría hacer un razonamiento definiendo la variable:

$$z = \frac{1}{x}$$

Transformando así la hipérbola en la recta de ecuación:

$$y = a + bz$$

De forma que se obtiene un sistema de ecuaciones del tipo:

$$\begin{cases} \sum y_i = n a + b \sum \left(\frac{1}{x_i} \right) \\ \sum y_i \left(\frac{1}{x_i} \right) = a \sum \left(\frac{1}{x_i} \right) + b \sum \left(\frac{1}{x_i^2} \right) \end{cases}$$

También se puede desarrollar del modo en que se ha desarrollado en el modelo parabólico, de manera que se ajustará a la nube de puntos una función de la forma:

$$y = a + b \left(\frac{1}{x} \right)$$

La función que se desea minimizar será:

$$M = \sum_{i,j} d_{i,j}^2 = \sum_{i,j} (\hat{y}_i - y_j)^2 \text{ donde } \hat{y}_i = a + \frac{b}{x_i}$$

Entonces:

$$M = \sum_{i,j} \left(a + \frac{b}{x_i} - y_j \right)^2$$

Para minimizar la expresión, se calcularán las derivadas parciales respecto a los parámetros a y b y se igualarán a 0:

$$\begin{cases} \frac{\partial M}{\partial a} = 2 \sum_{i,j} \left(a + \frac{b}{x_i} - y_j \right) = 0 \\ \frac{\partial M}{\partial b} = 2 \sum_{i,j} \left(a + \frac{b}{x_i} - y_j \right) \left(\frac{1}{x_i} \right) = 0 \end{cases}$$

Así pues, las ecuaciones normales serán:

$$\begin{cases} \sum_{i,j} \left(a + \frac{b}{x_i} - y_j \right) = 0 \\ \sum_{i,j} \left(a + \frac{b}{x_i} - y_j \right) \left(\frac{1}{x_i} \right) = 0 \end{cases} \quad \leftrightarrow \quad \begin{cases} a N + b \sum \frac{1}{x_i} = \sum y_j \\ a \sum \frac{1}{x_i} + b \sum \frac{1}{x_i^2} = \sum \frac{y_j}{x_i} \end{cases}$$

4.2.3. Ajuste potencial

El ajuste potencial tendrá como ecuación predictora:

$$Y = \alpha \cdot X^{\beta}$$

Y la regresión recíproca será:

$$Y = \frac{1}{\alpha + \beta \cdot X}$$

En el primer caso, los valores seguirán una ley potencial. Así, si se toman logaritmos en los dos miembros de la ecuación predictora, la ecuación quedará:

$$\log Y = \log \alpha + \beta \cdot \log X$$

Donde las constantes α y β quedarán fijadas al resolver simultáneamente las ecuaciones:

$$\begin{cases} \sum \log Y = \log \alpha \cdot N + \beta \sum \log X \\ \sum \log X \cdot \log Y = \log \alpha \cdot \sum \log X + \beta \sum (\log X)^2 \end{cases}$$

Para el segundo caso, si la ecuación predictora viene dada por la siguiente expresión:

$$Y = \frac{1}{\alpha + \beta \cdot X}$$

Entonces, al invertir la expresión, se obtendrá:

$$\frac{1}{Y} = \frac{\alpha + \beta \cdot X}{1}$$

Es decir:

$$Y = \frac{1}{\alpha + \beta \cdot X} \rightarrow \frac{1}{Y} = \alpha + \beta \cdot X$$

Donde las constantes α y β quedarán fijadas al resolver el sistema:

$$\begin{cases} \sum \frac{1}{Y} = \alpha \cdot N + \beta \sum X \\ \sum X \cdot \frac{1}{Y} = \alpha \cdot \sum X + \beta \sum X^2 \end{cases}$$

Ejemplo:

La siguiente tabla muestra las lecturas de un experimento en el que la variable X es el volumen en litros de un determinado gas y la componente Y es la presión en atm de dicho gas.

Se quiere ajustar los datos a una curva exponencial aplicando el método de los mínimos cuadrados y estimar la presión de un gas con un volumen de 9 litros.

X (en litros)	Y (en atm)
1	7
2	30
3	90
4	170
5	290
6	450
7	650

Solución:

Se construye una tabla de datos:

X	Y	log X	log Y	log X · log Y	(log X) ²
1	7	0,0000	0,8451	0,0000	0,0000
2	30	0,3010	1,4771	0,4447	0,0906
3	90	0,4771	1,9542	0,9324	0,2276
4	170	0,6021	2,2304	1,3429	0,3625
5	290	0,6990	2,4624	1,7211	0,4886
6	450	0,7782	2,6532	2,0646	0,6055
7	650	0,8451	2,8129	2,3772	0,7142
S X = 28		S log X = 3,7024	S log Y = 14,4354	S log X · log Y = 8,8829	S(log X) ² = 2,4890

Se sustituyen los valores en el sistema de ecuaciones:

$$\begin{cases} \sum \log Y = \log \alpha \cdot N + \beta \sum \log X \\ \sum \log X \cdot \log Y = \log \alpha \cdot \sum \log X + \beta \sum (\log X)^2 \end{cases}$$

$$\begin{cases} 14.4354 = \log \alpha \cdot N + \beta \cdot 3.7024 \\ 8.8829 = \log \alpha \cdot 3.7024 + \beta \cdot 2.4890 \end{cases} \rightarrow \begin{cases} 7 \log \alpha + 3.7024 \beta = 14.4354 \\ 3.7024 \log \alpha + 2.4890 \beta = 8.8829 \end{cases}$$

Resolviendo el sistema, se obtendrá:

$$\log \alpha = 0.819$$

$$\beta = 2.351$$

Sustituyendo estos valores en la ecuación predictora se obtendrá:

$$\log Y = \log \alpha + \beta \cdot \log X$$

$$\log Y = 0.819 + 2.351 \cdot \log X$$

$$\log \alpha = 0.819 \rightarrow \alpha = 6.592$$

La ecuación general quedará de la siguiente manera:

$$Y = \alpha \cdot X^{\beta}$$

$$Y = 6.592 \cdot X^{2.351}$$

Para conocer la presión que ejercerá un gas con un volumen de 9 litros, se sustituye el valor de $X = 9$ en la ecuación:

$$Y = 6.592 \cdot X^{2.351}$$

$$Y = 6.592 \cdot 9^{2.351} = 875.35 \text{ atm}$$

4.2.4. Ajuste exponencial

En los casos en que la curva de regresión de Y sobre X sea de tipo exponencial, es decir, para cualquier X que se considere, la media de la distribución vendrá dada por la siguiente ecuación predictora:

$$Y = \alpha \beta^X$$

Al aplicar logaritmos en ambos miembros de la ecuación, se obtiene:

$$\log Y = \log \alpha + X \cdot \log \beta$$

Ahora se pueden calcular los valores de $\log Y$ y de $\log \beta$, y de esos valores, obtener los términos α y β tras aplicar los métodos de los mínimos cuadrados.

Los valores de las constantes α y β quedarán fijados al resolverse el siguiente sistema:

$$\begin{cases} \sum \log Y = \log \alpha \cdot N + \log \beta \cdot \sum x \\ \sum x \cdot \log Y = \log \alpha \cdot \sum x + \log \beta \cdot \sum x^2 \end{cases}$$

Ejemplo:

En la siguiente tabla se muestran los datos del porcentaje de ruedas que se pueden usar de manera óptima tras haber recorrido cierto número de kilómetros:

Km	Porcentaje
1	99
2	95
5	85
15	55
25	30
30	24
35	20
40	15

Se pide ajustar la curva aplicando el método de los mínimos cuadrados y predecir el porcentaje de ruedas que podrán usarse tras 50 km.

Solución:

Se construye una tabla de datos:

X	Y	log Y	X ²	X · logY
1	99	1,996	1	1,996
2	95	1,978	4	3,955
5	85	1,929	25	9,647
15	55	1,740	225	26,105
25	30	1,477	625	36,928
30	24	1,380	900	41,406
35	20	1,301	1225	45,536
40	15	1,176	1600	47,044
Σ X = 153		Σ log Y = 12,97759	Σ X ² = 4605	Σ X · logY = 212,61769

Se reemplazan los valores en las ecuaciones del sistema:

$$\begin{cases} \Sigma \log Y = \log \alpha \cdot N + \log \beta \cdot \Sigma x \\ \Sigma x \cdot \log Y = \log \alpha \cdot \Sigma x + \log \beta \cdot \Sigma x^2 \end{cases}$$

$$\begin{cases} 12.97759 = \log \alpha \cdot 8 + \log \beta \cdot 153 \\ 212.61769 = \log \alpha \cdot 153 + \log \beta \cdot 4605 \end{cases}$$

Reorganizando los datos se obtiene:

$$\begin{cases} 8 \log \alpha + 153 \log \beta = 12.97759 \\ 153 \log \alpha + 4605 \log \beta = 212.61769 \end{cases}$$

Al resolver el sistema, resulta:

$$\log \alpha = \frac{\Delta_{\alpha}}{\Delta} = \frac{\begin{vmatrix} 12.97759 & 153 \\ 212.61769 & 4605 \end{vmatrix}}{\begin{vmatrix} 8 & 153 \\ 153 & 4605 \end{vmatrix}} = \frac{59761.80195 - 32530.50657}{36840 - 23409} = \frac{27231.295389}{13431}$$

$$\log \alpha = 2.027495747$$

$$\log \beta = \frac{\Delta_{\beta}}{\Delta} = \frac{\begin{vmatrix} 8 & 12.97759 \\ 153 & 212.61769 \end{vmatrix}}{13431} = \frac{1700.944152 - 1985.57127}{13431} = \frac{-284.627118}{13431}$$

$$\log \beta = -0.02119180389$$

Sustituyendo los valores se obtiene:

$$\log Y = \log \alpha + X \cdot \log \beta$$

$$\log Y = 2.027496 - 0.02119X$$

Al reemplazar los datos en la ecuación predictora, se tiene:

$$Y = \alpha\beta^X$$

$$Y = 106.536 \cdot 0.952^X$$

Para el cálculo de la estimación del porcentaje de ruedas que durarán 50 km, se debe reemplazar el valor de $X = 50$ en la ecuación predictora:

$$Y = 106.536 \cdot 0.952^X$$

$$Y = 106.536 \cdot 0.952^{50} = 9.106$$

Así pues, el porcentaje sería del 9.106 %.

Resumen

Antes de proceder a llevar a cabo una regresión entre dos variables, es necesario reflexionar sobre la posible relación que existe entre ellas, ya que hacerla directamente sobre una de las variables puede llevar a situaciones absurdas; por ejemplo, se puede estimar la función de regresión de la antigüedad de los trabajadores de una empresa sobre el salario que perciben, pero no tiene sentido económico que la antigüedad dependa del salario (es justo al contrario).

También puede darse la situación de que dos variables estén relacionadas estadísticamente pero no tengan ningún tipo de relación teórica. Por ejemplo, podría existir una relación entre el número de accidentes laborales al año en el sector de la construcción en España y las toneladas de carbón que se producen anualmente en Reino Unido. Sin embargo, esa relación estadística se debería al azar.

También podría ocurrir que existiese una relación lineal entre los sueldos anuales y el número de accidentes laborales al año, ambas variables referidas al sector de la construcción. De nuevo, no hay relación teórica entre ellas, pero si se llevase a cabo una regresión lineal, se observaría que existe una correlación positiva entre ambas. Esto significaría que una tercera variable, por ejemplo la actividad en el sector, está influenciando tanto los salarios (que se mueven en función del nivel de actividad) como el número de accidentes laborales que se producen en este (que también están relacionados directamente con la actividad).

Se puede deducir que un paso previo a la regresión entre dos variables es la reflexión y la constatación de que existe fundamento teórico para llevarla a cabo.

Por otro lado, una de las labores más solicitadas a los analistas es la predicción o pronóstico a futuro, tarea que se ha encomendado durante mucho tiempo (y se sigue encomendando) a los modelos de regresión. Por ejemplo, se puede llevar a cabo una regresión del salario medio en España en función del tiempo y usar la función de regresión estimada para predecir el salario medio en España en años venideros.

Es preciso tener en cuenta que la función de regresión constituye únicamente un muro del complicado edificio de la predicción; como dicen muchos autores: tratar de prever lo que va a suceder en el futuro requiere una comprensión perfecta del fenómeno a predecir, un conocimiento minucioso de los últimos acontecimientos en las actividades afines y el reconocimiento de las limitaciones de cualquier artificio mecánico a predecir”.

Por último, cabe destacar que las estimaciones de las funciones de regresión necesitan de datos, y que de la exactitud y/o veracidad de esos datos dependerá el éxito de la labor de regresión. Si los datos a partir de los cuales se estima la función de regresión son malos, la regresión y sus estimaciones no serán válidos. En consecuencia, otro paso previo a la estimación de la función de regresión es el análisis de la calidad de la información disponible.

Glosario

Adimensional

Cantidad que carece de una dimensión física asociada, siendo un número puro que permitirá describir una característica física sin dimensión ni unidad de expresión explícita.

Coeficiente de correlación

Medida específica que cuantifica la intensidad de la relación lineal entre dos variables en un análisis de correlación.

Coeficiente de variación

Relación entre la media y la desviación típica de una muestra.

Covarianza

Medida estadística de dependencia lineal. Valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias.

Cuantiles

Puntos tomados a intervalos regulares de la función de distribución de una variable aleatoria.

Desviación típica

Medida que se utiliza para cuantificar la variación o la dispersión de un conjunto de datos numéricos.

Discretización

Proceso de transferir funciones continuas, variables, modelos y ecuaciones a contrapartes discretas.

Dispersión o variabilidad

Es la mayor o menor separación de los valores de una variable.

Disperso

Que está constituido por elementos separados entre sí.

Distribución de probabilidad

Es una tabla, fórmula o gráfica que presenta conjuntamente los valores de la variable y sus probabilidades respectivas.

Esperanza matemática

Es el valor medio de un fenómeno aleatorio que se obtiene al multiplicar cada valor de la variable por su probabilidad y sumar los resultados.

Estadística

Es la ciencia que estudia las técnicas de recogida, descripción y análisis de datos.

Estandarización

Implantación de normas claras y precisas acerca de los métodos y formas de ejecutar un proceso concreto: un procedimiento de trabajo, la forma de actuar de un equipo, etc.

Estimación

Es la obtención del valor aproximado de un parámetro.

Estimador

Es una variable aleatoria cuyo valor cambia con la muestra.

Éxito

Suceso de interés en un experimento binominal.

Experimento

Es el proceso por medio del cual una observación (o medición) es registrada.

Experimento aleatorio

Es un experimento cuyo resultado no se puede predecir con seguridad.

Experimento muestral

Es la colección de todos los posibles resultados de un experimento.

Frecuencia

Número total de veces que un dato se repite en un evento.

Frecuencia absoluta

Número de veces que se repite un valor.

Frecuencia relativa

Frecuencia absoluta dividida entre el número total de datos.

Función de densidad

Es la densidad media de la probabilidad correspondiente a un intervalo infinitesimal de valores de una variable aleatoria continua. Por tanto, no proporciona la probabilidad a no ser que se integre.

Función de distribución

Es una función que, para cada valor, nos da la probabilidad de que la variable aleatoria tome un valor igual o inferior a ese valor.

Histograma

Representación gráfica de una variable en forma de barras.

Inferencia estadística

Es la parte de la estadística que consiste en el análisis de una muestra extraída de una población al objeto de obtener conclusiones sobre la población en su totalidad.

Interpolación

Obtención de nuevos puntos a partir del conocimiento de un conjunto de puntos.

Media

Medida de tendencia central. Su valor puede representar por sí solo a todo el conjunto.

Media aritmética

Es la medida de posición central más utilizada, y se define como la suma de todos los datos de una variable dividida por el número de ellos.

Mediana

Valor que ocupa el lugar central, supuesto un número impar de datos, y ordenados estos de menor a mayor.

Moda

Valor con mayor frecuencia en una de las distribuciones de datos.

Modelo de probabilidad

Conjunto dado de distribuciones que tienen sus funciones (de distribución, de cuantía o de densidad) definidas con la misma estructura funcional matemática.

Momentos de una distribución

Son unos valores que caracterizan a la distribución, de tal modo que dos distribuciones son iguales si tienen todos sus momentos iguales, y son tanto más parecidas cuanto mayor sea el número de momentos iguales. Se trata de expresiones intermedias que se utilizan para calcular medidas de posición, dispersión y forma.

Muestra

Subconjunto representativo de la población.

Partición

Es una colección de sucesos que son disjuntos y cuya unión forma el espacio muestral.

Parámetro

Valor teórico que determina una distribución concreta perteneciente a un modelo de probabilidad.

Población

Conjunto de eventos o elementos similares que resultan de interés para algún experimento o pregunta.

Probabilidad

Es la expectación que se tiene de la ocurrencia de un suceso.

Probabilidad condicionada

Es la probabilidad de un suceso modificada por la ocurrencia de otro que se llama condicionante.

Sector circular

Porción del círculo determinada por un ángulo central, quedando así delimitada por un arco y dos radios.

Suceso imposible

Es aquel que nunca ocurre.

Sucesos

Diferentes resultados de un experimento.

Sucesos disjuntos

Son sucesos que no tienen resultado común y, por tanto, no pueden ocurrir a la vez.

**Sucesos independientes**

Si uno no influye en el otro, entonces la probabilidad de que ocurran a la vez es producto de sus probabilidades.

Sucesos seguros

Son aquellos que siempre acontecen.

Variabilidad o dispersión

Es la mayor o menor separación de los valores de una variable.

Variable

Característica que puede fluctuar y cuya variación puede adoptar diferentes valores.

Variable aleatoria

Es una función para asignar a cualquier resultado de un experimento una media cuantitativa.

Variable binomial

Es el resultado de sumar n variables de Bernouilli independientes del parámetro p y, por tanto, es el número de éxitos en n repeticiones de un experimento Bernouilli.

Variable cualitativa

Son características de la población cuya observación no suministra un número, sino la pertenencia a una clase determinada.

Variable de Bernouilli

Es aquella asociada a un experimento consistente en realizar una sola observación que puede dar lugar a dos resultados posibles: éxito y fracaso.

Variable normal

Se define para una variable aleatoria continua que puede tomar cualquier valor a lo largo del eje real.

Variable Poisson

Es el número de hechos de cierto tipo que se pueden producir en un intervalo de tiempo o de espacio.

Varianza

Medida de dispersión definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media.

Bibliografía



- Canavos, G. C. (1988). *Probabilidad y estadística: aplicaciones y métodos*. Madrid: McGraw-Hill.
- Everitt, B. S. y Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. Nueva York: Cambridge University Press.
- Fernández Abascal, H., Guijarro, M. M., Rojo, J. L. y Sanz, J. A. (1994). *Cálculo de probabilidades y estadística*. Barcelona: Editorial Ariel.
- Galindo Villardón, M. P. (1984). *Exposición intuitiva de métodos estadísticos*. Salamanca: Ediciones Universidad Salamanca.
- Hermoso Gutiérrez, J. A. y Hernández Bastida, A. (1997). *Curso básico de estadística descriptiva y probabilidad*. Madrid: Némesis.
- Martín Pliego, F. J. (1989). *Curso práctico de estadística económica*. Madrid: AC.
- Martín Pliego, J. y Ruiz Maya, L. (2004). *Estadística I: Probabilidad*. 2.^a ed. Madrid: Thomson.
- Meyer, P. L. (1973). *Probabilidad y aplicaciones estadísticas*. 2.^a ed. Bogotá: Fondo Educativo Iberoamericano.
- Montero Lorenzo, J. M. (2007). *Problemas resueltos de estadística descriptiva para ciencias sociales*. Madrid: Paraninfo.
- Pardo, A. y Ruiz, M. A. (2002). *SPSS: Guía para el análisis de datos*. Madrid: McGraw-Hill.
- Pardo, A. y San Martín, R. (1998). *Análisis de datos en psicología II*. 2.^a ed. Madrid: Pirámide.
- Ramírez, A. y Cox, C. (2012). Improving on the range rule of thumb. *Rose-Hulman Undergraduate Mathematics Journal*, 13(2), art. 1.
- Suárez, M. (2004). *Interaprendizaje holístico de matemáticas*. Ecuador: Gráficas Planeta.
- Uña, I., San Martín, J. y Tomeo, V. (2009). *Cálculo de probabilidades*. Madrid: Garceta.