

Módulo 1: Introducción al análisis de datos.

UD1: Iniciación al análisis estadístico de datos.

UD1: Iniciación al análisis estadístico de datos.....	4
Conceptos estadísticos.	5
Tipos de datos.....	5
Escala de medición.	6
Otros tipos de datos.....	7
Procesamiento de datos.....	7
Extracción de datos.	7
Transformación de datos.....	8
Datos vacíos o faltantes.....	8
Creación de variables derivadas.....	8
Creación de variables ficticias (dummy).....	9
Reemplazar valores atípicos (<i>outliers</i>).	9
Valores erróneos.	10
Transformación de variables.....	10
Carga de datos.	14
Estadísticas descriptivas y tablas de contingencia.	15
Estadísticas descriptivas.....	15
Medidas de tendencia central.	15
Medidas de variabilidad o dispersión.	17
Medidas de posición o localización.....	18
Medidas de simetría.	20
Medidas de concentración.	20
Medidas de precisión.	22
Tablas.....	22
Tablas de frecuencia absoluta.....	23
Tablas de frecuencias relativas.	24
Tablas de contingencia.	24
Lista de datos ordenados.	25
Diagrama de tallo y hojas.....	25
Test estadísticos.	26
Test de proporciones.	28
Test exacto de Fisher.....	28

Test de McNemar.	28
Ji-cuadrado de Pearson (χ^2).	30
Test de Kolmogorov-Smirnov (KS).	33
Análisis de clúster.	33
Resumen.....	35
Mapa de contenidos.....	36
Recursos bibliográficos.....	37

UD1: Iniciación al análisis estadístico de datos

El análisis estadístico de datos ha ganado relevancia en los últimos años en diversos sectores, y el transporte, que es el sector que vertebra a un país, no podría ser menos.

Ya en el antiguo Egipto se recopilaban datos sobre la población y riqueza del país. A medida que se han ido desarrollando las civilizaciones, la cantidad de datos que se producen no han hecho más que aumentar. Con la llegada de las democracias, era necesario tener actualizado el censo de votantes. La imposición de impuestos requiere conocer la riqueza de cada individuo y la naturaleza de las actividades con las que se lucra.

Así, el volumen de datos que se manejan no hace más que incrementarse. Desde mediados del siglo XX se ha desarrollado la informática, popularizándose a partir de la década de los 80. Con ella, se ha podido procesar mucha más información. Desde principios del siglo XXI la red Internet se ha globalizado. El crecimiento de los datos disponibles ha sido exponencial.

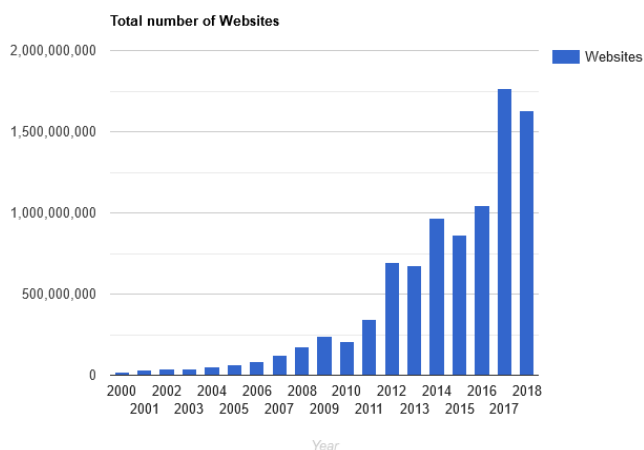


Figura 1: Número total de webs en el mundo. Fuente: <https://www.internetlivestats.com/total-number-of-websites/>.

En la última década, la popularización de los teléfonos móviles ha permitido que cada persona lleve en su bolsillo un teléfono inteligente que, en realidad, es un ordenador.

Los dispositivos inteligentes se caracterizan por poder poseer 3 funciones que resultan relevantes:

- Sensores, que permiten medir datos constantemente.
- Capacidad de cómputo para analizar y procesar los datos.
- Sistemas de comunicación.

Los diferentes sensores de un dispositivo permiten extraer muchos datos: posición, altitud, temperatura, orientación, iluminación, sonidos. Mediante la capacidad de cálculo de los dispositivos se pueden procesar algunos de estos datos para obtener otros más elaborados. Por ejemplo, con la medición consecutiva de 2 posiciones GPS, y el tiempo que pasa entre ambas, resulta trivial calcular la velocidad a la que circula un vehículo. Además, si estos datos se pueden compartir con un centro de proceso de datos, se obtiene la potencia de análisis de un (super)ordenador.

En los 80 se popularizó el ordenador personal. En los 2000 se generalizó el teléfono móvil. En los últimos años vemos como nuestras casas se llenan de dispositivos inteligentes. Resulta fácil entender que el volumen de datos que se está produciendo no hace más que crecer, y lo hace, además, de forma exponencial.

La sobreabundancia de datos refuerza la necesidad de técnicas cada vez más sofisticadas de análisis.

No existe una definición oficial de análisis estadístico de datos, pero generalmente se refiere a cómo se pueden analizar y utilizar los datos disponibles para mejorar los negocios y las organizaciones.

Las técnicas de análisis incluyen tablas de contingencia, estadísticas descriptivas, pruebas estadísticas y diversas formas gráficas de visualizar los datos.

El análisis de datos está influido por la calidad de los datos de los que se parte. Interesa que sean confiables y de buena calidad. En ocasiones, el proceso de extracción y limpieza de los datos es uno de los que más tiempo consume en todo el análisis.

Conceptos estadísticos.

Estadística: Rama de las matemáticas que se ocupa de recopilar datos, organizarlos y analizarlos para lograr un objetivo.

Población (o universo): Conjunto de todos los individuos estudiados.

Muestra: Parte seleccionada de los individuos de la población.

Individuo: Cada uno de los elementos que forman parte de la población, o de la muestra.

Muestreo: Técnica de obtención de una, o más, muestras de una población. Por ejemplo, seleccionar al azar un grupo de individuos de la población para formar una muestra. Si cada individuo tiene la misma probabilidad de ser seleccionado, se llama "muestreo aleatorio simple".

Probabilidad: Posibilidad de ocurrencia de un fenómeno. Se mide en tanto por ciento o tanto por 1.

Muestra representativa: Muestra cuya composición es similar a la de la población de partida. Los resultados obtenidos para esta muestra deben ser similares a los resultados obtenidos en la población.

Variable: Característica que se quiere medir.

Dato: Valor que toma la variable medida.

Estadística descriptiva: Trata de analizar las características de un conjunto de datos.

Estadística inductiva (o inferencia estadística): Analiza muestras para tratar de inferir características de las poblaciones.

Error estadístico: El error que tienen los resultados del estudio debido a factores no controlados.

Tipos de datos.

La utilización de los datos en los modelos o en los programas estadísticos se hace en forma de variables. Estas son posiciones en las que se pueden incluir datos de "varios" tipos. Conocer el tipo de una variable puede ayudar a obtener mejores resultados. Por ejemplo: una variable de tipo desconocido puede albergar cualquier tipo de datos: letras (a, b, c, ...), números (1, 2, 3, ...), etc. Sin embargo, si se especifica el tipo (y, por tanto, se restringe el tipo de datos a los que son

del tipo definido) se pueden hacer más cosas. Por ejemplo, los datos de tipo "letra", no se pueden sumar, pero los datos de tipo "número" si se pueden sumar. Si el tipo de dato es "número" se suma de forma diferente que si es de tipo "fecha".

En estadística hay 2 tipos de variables:

- **Variables cualitativas** (o categóricas): No se pueden medir. Se expresan con palabras. Ejemplo: color. El color puede ser "azul", "verde" o "rojo". Miden una "cualidad", un atributo. Hay 2 subgrupos de variables cualitativas:
 - o **Nominales**: Cuando no existe un orden natural entre las categorías. El nombre de la variable es sólo una etiqueta, un nombre. Ejemplo: color de ojos (castaño, azul, verde).
 - o **Ordinales**: Cuando hay un orden o jerarquía entre las categorías de las variables. Ejemplo: Conformidad (en completo desacuerdo, poco de acuerdo, indiferente, algo de acuerdo, muy de acuerdo).
- **Variables cuantitativas** (o numéricas): Se pueden medir; se pueden expresar con números. Por ejemplo, el número de hijos de una familia o la longitud de onda de la luz. Miden una "cantidad". Dentro de estas variables, hay 2 subtipos:
 - o **Cuantitativas discretas**: Sólo pueden tomar algunos valores. No hay continuidad entre estos valores. Ej. el número de hijos de una familia pueden ser 2 o 3, pero no 2,75.
 - o **Cuantitativas continuas**: Pueden tomar valores infinitos. Por ejemplo, la longitud de onda de un haz de luz puede tener infinitos decimales. Dependerá de la precisión del instrumento de medida.

Escalas de medición.

Cuando se miden los datos, se pueden utilizar varias escalas. Son estas:

Nominal: Clasifica los datos en categorías no numéricas, excluyentes y no ordenadas. Por ejemplo, el color de las estanterías de un almacén puede ser "azul", "amarillo" y "rojo". Si se recogen datos de los productos del almacén y, en una variable, se almacena el color de la estantería donde se encuentra, esa variable no indica el orden (en el sentido de que "rojo" no es mayor ni menor que cualquiera de los otros 2).

Ordinal: Clasifica los datos en categorías no numéricas, excluyentes y ordenadas. Aunque existe orden en los datos, se desconoce la magnitud exacta de la diferencia entre los datos. Por ejemplo: en el almacén hay 3 tipos de embalajes: cajas "pequeñas", "medianas" y "grandes". Si se almacena una variable en los datos de cada producto con el tipo de caja, sí se puede establecer un orden ya que: "grande" > "mediana" > "pequeña".

Intervalo: Clasifica los datos de forma ordenada que, además, cuantifica la distancia entre ellos. Las escalas de intervalo carecen de valor 0 absoluto. Por ejemplo, la temperatura de un almacén se puede registrar como 15 °C o 30 °C. Hay 15 °C de diferencia entre ambos registros. Hay que tener en cuenta que el valor 0 °C no es un 0 absoluto, ya que puede haber números negativos. Además, en un almacén que haya 30° no hay el doble de calor que en uno que tenga 15 °C. Para comprobarlo, basta con cambiar la escala de los grados a la escala Kelvin, que tiene la misma proporción que la Celsius pero el 0 absoluto lo tiene en -273 °C (aprox). Así, 0 °C equivalen a 273 °K, 15 °C equivalen a 288 °K y 30 °C equivalen a 303 °K. Efectivamente, 303 no es el doble que 288.

Razón (o proporción): Tiene las mismas características que la escala de intervalo: ordenada y distancia cuantificada, pero si tiene valor 0 absoluto. La escala de °K vista en el ejemplo anterior es una escala de razón. Otro ejemplo puede ser la altura de los productos del almacén. Un producto de 20 cm es el doble de alto que uno de 10. Además, no puede haber una altura inferior a 0.

Desde un punto de vista práctico, los programas de estadística suelen contemplar 3 valores: Nominal, Ordinal y Escala. Los de tipo “Ordinal” son como “Nominales” a los que se les puede establecer un orden. Los de tipo “Escala” son variables cuantitativas.



Figura 2: Selector de medidas del programa SPSS.

Es interesante elegir el tipo de medida adecuada. Con medidas de Tipo “Escala” se suelen obtener resultados más precisos que con medidas de tipo “Ordinal” y estas, a su vez, son más precisas que las de tipo “Nominal”.

Otros tipos de datos.

Algunos tipos de datos, por su naturaleza, tienen un tratamiento especial. Cabe destacar estos:

Cronológicos: Los valores de las variables expresan instantes o periodos en el tiempo. Aunque internamente se pueden tratar como valores numéricos, conviene marcarlos como cronológicos dado que la forma de procesarlos puede tener peculiaridades. Ej.: una fecha, como el 28 de febrero.

Geográficos: Los valores estar referidos a una localización geográfica. Por ejemplo, una coordenada GPS o una dirección postal.

Procesamiento de datos.

Los datos sin procesar no suelen ser confiables porque pueden contener errores y ruido. Además, puede que no sean adecuados, o no estén optimizados, para su procesamiento posterior. Por ejemplo, puede que la distribución de algunas variables no sea normal. Además, el procesamiento de los datos pretende dejarlos en condiciones óptimas para su explotación. El procesamiento de datos se divide en 3 fases: extracción, transformación y carga.

Extracción de datos.

La extracción de datos pretende obtener los datos de sus fuentes originales y dejarlos disponibles para continuar con el resto de las fases del procesamiento.

La fase de extracción se encarga de la recopilación, filtración e integración de datos. Se trata de detectar los problemas que puedan tener los datos y, si se puede, atajarlos en esta fase. Por ejemplo, se pueden detectar valores anómalos (outliers) o valores faltantes.

En la extracción de datos se detectan errores que deben ser corregidos a continuación, en la fase de transformación de datos.

La heterogeneidad de las fuentes de datos puede dificultar el proceso de extracción. Por ejemplo, se pueden obtener datos de fuentes públicas, privadas, APIs, streaming, web scraping (Hista, 2022), ...

Además, los datos pueden llegar en formatos muy diversos (texto, de longitud fija, csv, xlsx, json, html, odf, svg, pdf, formatos cerrados, ...). El formato debería ser unificado para continuar el procesamiento de forma coherente.

Transformación de datos

La transformación de datos genera datos limpios y confiables para realizar las operaciones que permitan manejarlos y extraer el valor que tienen. Por ejemplo, se pueden realizar transformaciones de variables para que sus distribuciones cumplan los requisitos de normalidad.

Hay algunas técnicas que se pueden emplear para corregir errores o acondicionar los datos.

Datos vacíos o faltantes.

Es común que las variables contengan datos vacíos en algunas de las mediciones. Los motivos para que falte un dato son diversos. Puede que haya fallado puntualmente el instrumento de medida, que el dato se haya eliminado en un procesamiento previo, que ese punto no haya sido recogido debido al diseño del experimento, etc. En ocasiones, el dato no está vacío sino que se ha introducido un código que indica que el dato no se pudo recoger.

La forma adecuada de lidiar con estos casos es crear una variable derivada en la que se aplique una técnica de corrección (así no se pierde la información original del dato). En esta variable se pueden reemplazar los valores faltantes por la media, la mediana o la moda de los datos cuando la variable es numérica y los registros puedan ser similares al valor faltante. También puede crearse una variable categórica que sólo indique si el valor es faltante o válido. Dicha variable podrá utilizarse posteriormente para filtrar sólo valores válidos en los datos.

Si se dispone de un modelo predictivo, se puede utilizar este modelo para predecir los valores de los datos faltantes. Por ejemplo, si se sabe que la variable sigue una progresión exponencial, se puede utilizar la función exponencial para predecir un punto que falte.

Creación de variables derivadas.

Esta técnica consiste en realizar cálculos o transformaciones sobre una variable y almacenarlos en el conjunto de datos en otra variable (esta segunda variable es derivada de la primera).

La variable derivada se crea para que sea más útil que la variable original.

Para crear las variables derivadas se emplean funciones o métodos como los siguientes:

- **Log** de una variable: Consiste en aplicar el logaritmo a sus valores. Permite reducir el efecto de los valores extremos o linealizar crecimientos exponenciales.
- **Data binning**: (Del inglés; podría traducirse como "agrupación de datos") Los valores de una variable se agrupan en intervalos o categorías. Este proceso puede simplificar la variable y hacerla más homogénea. Aunque se puede perder algo de información, suele ganarse en claridad. Por ejemplo, la renta de las familias de un barrio podrías ser:

Renta = {15.000, 25.000, 23.500, 80.000, 13.000, 7.000, 400.000, 30.000, 55.000}

Se podría crear una variable derivada "Grupo" como esta:

Grupo= {baja, media, media, alta, baja, baja, alta, media, alta}

- Otras **transformaciones**: Se pueden aplicar operaciones que transforman linealmente una variable (sumar, restar, multiplicar, dividir... por una constante) o no linealmente (potencia, raíz, seno, tangente, ...). Con estas transformaciones se puede cambiar la escala, la forma o la distribución de la variable original.

Creación de variables ficticias (dummy).

Consiste en convertir variables categóricas en numéricas. Este método es útil para utilizar variables categóricas en modelos de datos que requieren valores numéricos. Por ejemplo, en el caso de la variable Grupo anterior se podrían crear 3 variables *dummy* así:

Grupo	alta	media	baja
baja	0	0	1
media	0	1	0
media	0	1	0
alta	1	0	0
baja	0	0	1
baja	0	0	1
alta	1	0	0
media	0	1	0
alta	1	0	0

Tabla 1: Variables "dummy" creadas a partir de la variable categórica "Grupo".

Reemplazar valores atípicos (outliers).

Los valores *outliers* son valores que difieren demasiado del resto de datos de un conjunto. Se pueden producir, por ejemplo, cuando hay otra población de muestras distinta a la principal; o simplemente por errores de medición. Pueden causar desviaciones en los análisis. Se pueden visualizar, fácilmente, con gráfico de tipo box-plot o un histograma. Hay métodos de detección numéricos que suelen establecer un límite para considerar un dato como atípico. Por ejemplo: 3 desviaciones estándar de la media.

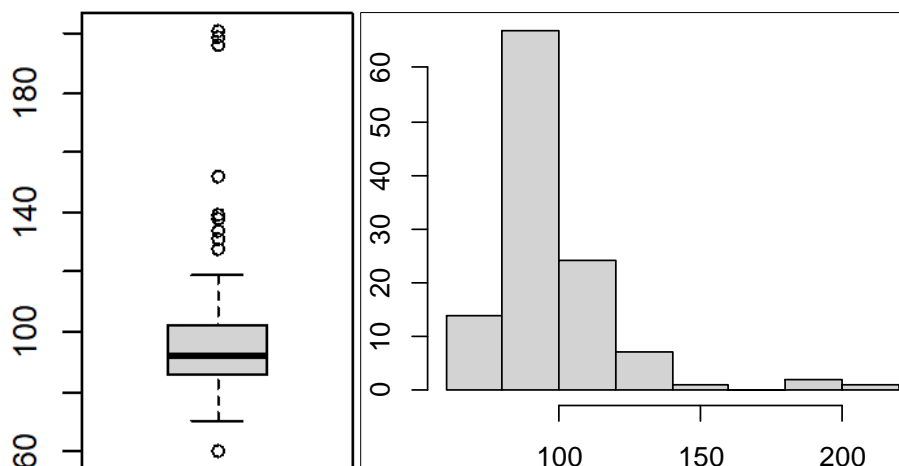


Figura 3: boxplot (izquierda) e histograma (derecha) de una variable con valores atípicos (representados como círculos en el boxplot).

Se pueden reemplazar utilizando los mismos procedimientos que se emplearían con valores faltantes.

Valores erróneos.

Los valores de los datos pueden ser erróneos en formato o en contenido. Los errores de formato son fáciles de detectar, basta con aplicar las reglas de formato establecidas y ver qué datos las incumplen. Si el sistema de recogida de datos está bien diseñado, no deberían poder introducirse valores erróneos, al menos, con errores de formato de datos. Los errores de contenido son mucho más difíciles de detectar, aunque un estudio pormenorizado de los datos puede detectar valores “imposibles”. Por ejemplo, un cordero no puede nacer antes de la fecha de nacimiento de su madre, o un coche no puede estar matriculado antes de la fecha de comercialización del modelo.

Es conveniente detectar los valores erróneos lo más rápidamente posible porque podrían ser camuflados por transformaciones posteriores.

Transformación de variables.

Algunas técnicas estadísticas, usualmente las más potentes, requieren que los datos cumplan los requisitos de normalidad, linealidad y homocedasticidad. Por ejemplo: regresión lineal, correlación, ANOVA. Estas son las características de los requisitos:

- **Normalidad:** Implica que los residuos del modelo (errores) siguen una distribución normal.
- **Linealidad:** Existe una relación lineal entre las variables dependiente e independientes. Un cambio en una de ellas se traduce en un cambio proporcional en las otras.
- **Homocedasticidad:** La varianza de los errores es constante en todas las variables independientes.

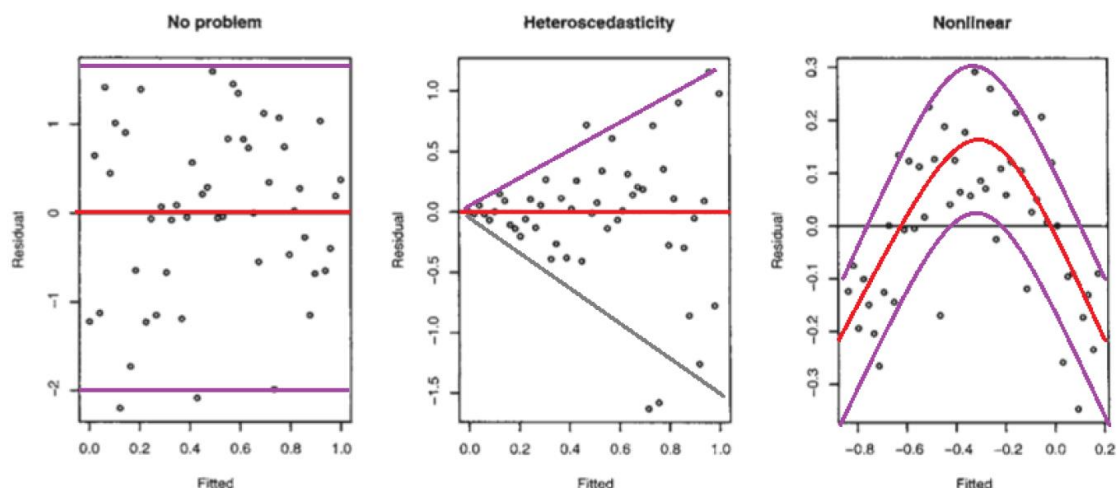


Figura 4: Gráficas de residuos de una variable que no presenta problemas (izquierda), que tiene problemas en la homocedasticidad (centro) o que tiene problemas de linealidad (derecha).

Sin embargo, en la vida real es común que muchas variables no sigan la distribución normal o que incumplan algún otro requisito. Por tanto, no se pueden utilizar muchas de las técnicas estadísticas con estas variables sin perder la fiabilidad de los resultados. Lo que se puede hacer es transformar las variables inadecuadas en otras que si cumplan los requisitos. Una vez realizados los cálculos se revertirá la transformación y se obtendrá el resultado final.

La transformación puede consistir en el reemplazo de una variable por una función. Por ejemplo, se puede sustituir X por raíz de X o por el logaritmo de X .

La existencia, o falta, de linealidad se puede observar mediante una gráfica xy.

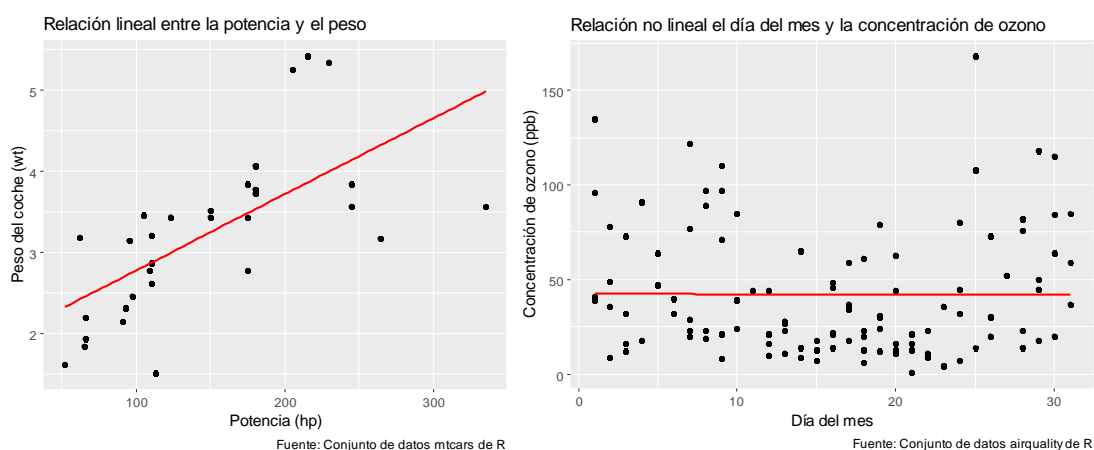


Figura 5: Gráficas donde se puede observar relación lineal entre las variables x e y (izquierda) y donde no la hay (derecha)

La normalidad se puede apreciar en un histograma. En este se espera que la distribución de frecuencia se aproxime a una campana de Gauss.

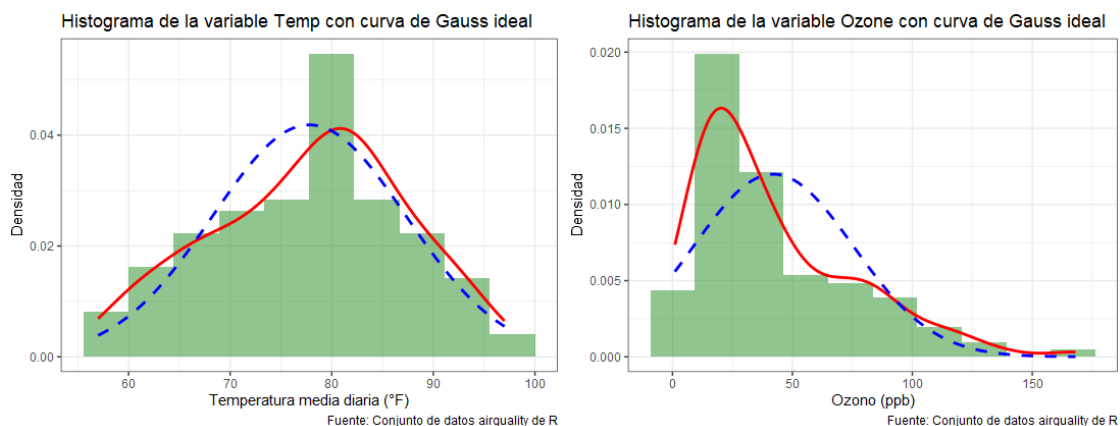


Figura 6: Histograma de una variable que sigue bastante bien una distribución normal (izquierda) y no la sigue bien (derecha). La distribución de la variable se representa en rojo. En azul, intermitente, se superpone la curva de la distribución normal ideal.

La distribución normal.

La distribución normal es aquella cuyos datos se distribuyen como una campana de Gauss. La mayoría de sus valores se concentran en el centro de la distribución y disminuyen de forma progresiva y simétrica hacia las colas de valor alto y bajo.

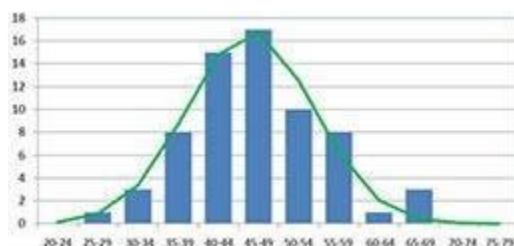


Figura 7: Distribución normal de los datos. En verde se muestra la campana de Gauss.

Tipos de transformaciones.

Estas son algunas de las transformaciones que se pueden hacer con los datos para cumplir con los requisitos:

Cambio de escala o normalización: Consiste en ajustar los valores de una variable a una escala común con otras con las que se quiere comparar. Algunos test requieren que las variables estén normalizadas. También resulta útil para comparar variables con diferentes rangos o unidades de medida.

Para realizarla se pueden reescalar los valores para que caigan en el rango entre el 0 y el 1.

$$\frac{(\text{valor} - \text{mínimo})}{(\text{máximo} - \text{mínimo})}$$

Otra normalización es la estandarización: consiste en reescalar la variable para que tenga de media 0 y de desviación estándar 1.

$$\frac{\text{valor} - \text{media}}{\text{desviación_estándar}}$$

Búsqueda de relaciones lineales: Trata de convertir una relación no lineal en una relación lineal. Suele aplicar funciones logarítmicas o raíces a la variable. Por ejemplo, cualquier progresión exponencial se puede convertir en lineal aplicando un logaritmo.

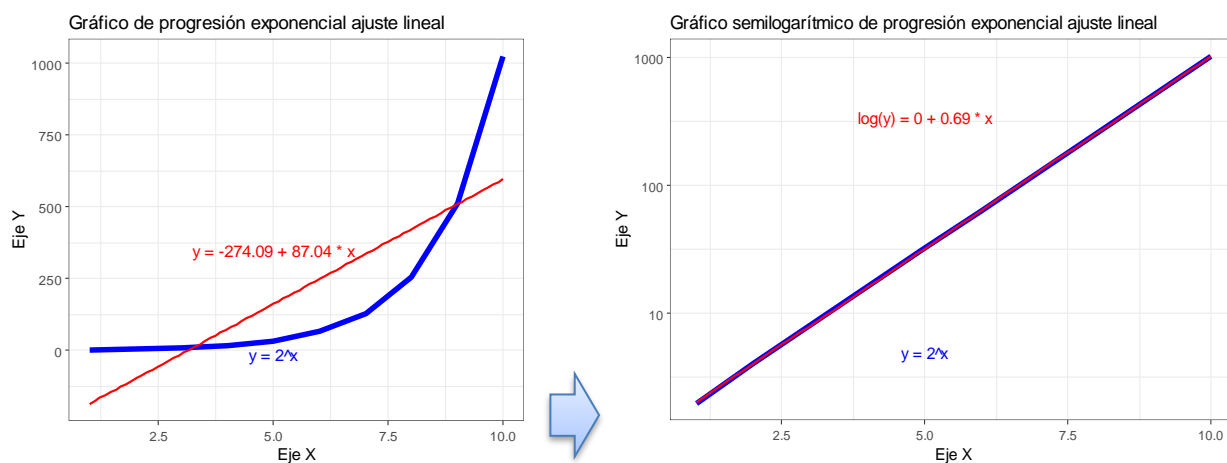


Figura 8: Efecto de aplicar una escala logarítmica a una variable exponencial (la del eje Y). Se muestra, en rojo, la recta de regresión y su ecuación. En la gráfica de la izquierda tanto el eje X como el Y están representados en escala lineal. En el gráfico de la derecha el eje Y está representado en escala logarítmica mientras que el eje X sigue estando representado en escala lineal.

Este tipo de gráfica, donde uno de los ejes se representa en escala logarítmica y el otro en escala lineal se llama gráfica semilogarítmica.

Transformar a distribución normal: Hay algunas transformaciones que pueden convertir variables no normales en normales. En función del tipo de histograma puede probarse unas u otras. En la siguiente tabla se recogen algunos ejemplos.

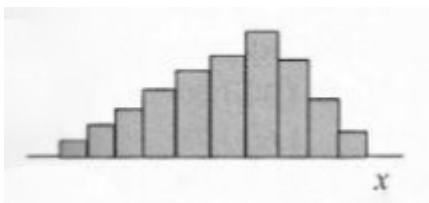
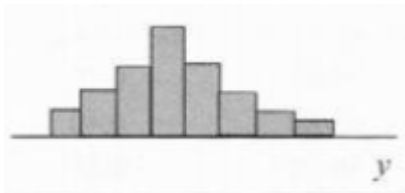
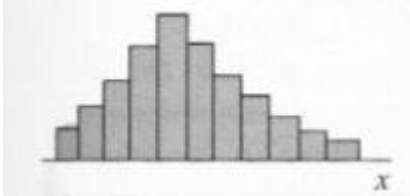
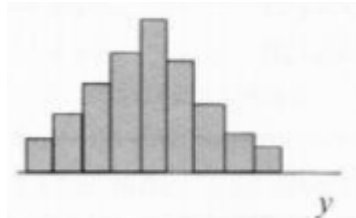
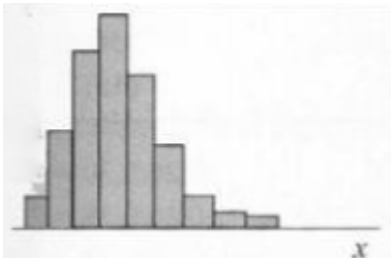
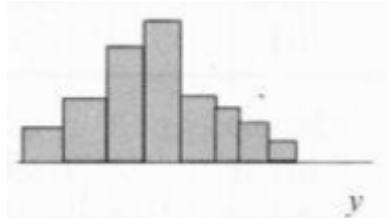
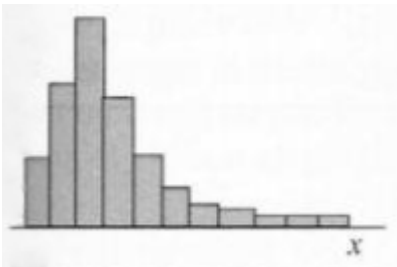
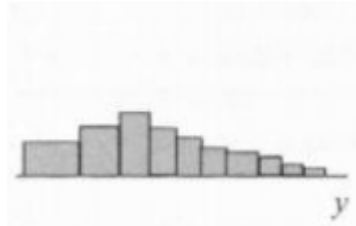
Variable inicial	Transformación	Variable final
 Distribución sesgada a la derecha	$y = x^2$	
 Distribución sesgada a la izquierda	$y = \sqrt{x}$	
	$y = \ln(x)$	
	$y = \frac{1}{x}$	

Tabla 2: Transformaciones que se pueden emplear para normalizar variables. Adaptado de: (Peña, 1997).

Carga de datos.

La carga de datos consiste en enviar los datos extraídos y transformados a sus almacenamientos definitivos. Se cargan, por tanto, en las bases de datos donde serán almacenados.

El sistema definitivo es el que suele estar diseñado para realizar las consultas de datos y los análisis posteriores.

Durante la carga se suelen realizar operaciones como:

- **Transferencia de datos.** Debe ser una transferencia segura para evitar pérdidas o corrupción de los datos. También interesa asegurar la privacidad y el secreto industrial que puedan contener. Debe ser eficiente, sobre todo en entornos donde se genera un gran volumen de datos (Big Data).
- **Optimización del almacenamiento.** Se puede optimizar para ahorrar espacio (compresión) o para aumentar la velocidad de acceso a los datos (indexado).

- **Optimización de la carga.** Se debe elegir si se cargan todos los datos (carga completa) o sólo los que han sido modificados desde la última vez (carga incremental).
- **Programación.** Puede ser interesante programar los horarios de carga de datos para afectar lo más mínimo posible al rendimiento del servidor.
- **Integración.** Los datos recibidos se integran con las herramientas de consulta, de análisis y de reporte. Así quedan disponibles para su consulta o utilización por parte de las herramientas de análisis de datos, de inteligencia de negocio (BI) o de aprendizaje automático (*machine learning*).

Estadísticas descriptivas y tablas de contingencia.

Cuando ya se dispone de un conjunto de datos fiable estamos en condiciones de extraer la información valiosa de los ellos. En primer lugar, es necesario comprender la estructura y naturaleza de los datos. Para ello se realizan resúmenes de los numéricos mediante estadística descriptiva (describe los datos) y tablas de contingencia.

Estadísticas descriptivas.

Sirven para resumir el conjunto de datos. Habitualmente son sencillas y muy fáciles de realizar en el ordenador. Hay varios tipos.

Medidas de tendencia central.

Indican hacia donde se agrupan los datos. Para ello, proporcionan un valor central o típico.

Media.

Suma todos los valores y divide por el número de valores. Para una variable compuesta por los datos x_1, x_2, \dots, x_n :

$$\text{Media} = \frac{1}{n} \sum_{i=1}^n x_i$$

💡 Se pueden distinguir 2 clases de media, la media aritmética de la población (μ) y la media aritmética de la muestra (\bar{x}), que podrían no ser iguales. μ es el valor medio real, \bar{x} es un estimador de μ .

Mediana.

Valor central de la variable cuando los datos están ordenados. El 50% de los valores son menores o iguales a ella. En caso de que el número de datos de la variable sea par no hay un valor central. En este caso se establece la mediana como la media de los 2 valores centrales: Ejemplo: para el conjunto de datos $d = \{1, 5, 4, 7, 3, 6, 9\}$, la mediana será: 5. Explicación: $d = \{1, 3, 4, 5, 6, 7, 9\}$. Porque 5 es el valor central cuando se ordena la variable. Si al conjunto se le añade 15, la mediana será: 5,5. Explicación: $\{1, 3, 4, 5, 6, 7, 9, 15\}$. Porque 5,5 es el promedio de los valores centrales 5 y 6.

Moda.

Es el valor más frecuente en el conjunto de datos (mayor frecuencia absoluta). Este puede ser unimodal (sólo tiene una moda), multimodal (tiene varias modas) o amodal (no tiene moda). Ejemplo: la moda del conjunto $d = \{1, 8, 5, 4, 7, 5, 4, 3, 5, 6, 8, 9\}$ será 5 porque aparece 3 veces.

Media geométrica.

Es la raíz n -ésima del producto de los n valores de una variable.

$$\text{Media geométrica} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Por ejemplo, para la variable $d = \{1, 3, 4, 5, 6, 7, 9, 15\}$, la media geométrica es: 4,914358.

La media geométrica es mucho más insensible que la media aritmética al efecto de los valores extremos.

Media aritmética ponderada.

Es el promedio de datos que no tienen el mismo valor sino que cada uno tiene una importancia específica, llamada peso. Para un conjunto de datos x_1, x_2, \dots, x_n , que tienen unos pesos w_1, w_2, \dots, w_n :

$$\text{Media ponderada} = \frac{\sum_{i=1}^n w_i \cdot x_i}{\sum_{i=1}^n w_i}$$

(w_i es el peso de la observación x_i).

Media móvil simple.

Es el promedio de un conjunto de datos que se va moviendo con frecuencia fija. Permite obtener una serie de tiempo de medias. Se calcula sumando un número determinado (k) de valores de la variable x y dividiendo dicha suma por el número de valores (k). La "ventana" de la media móvil es el número de valores (k), anteriores al valor que se está calculando, que se incluyen en el cálculo. Esta ventana se va desplazando a medida que se dispone de nuevos datos.

$$\bar{x}_n = \frac{x_{n-1} + x_{n-2} + \dots + x_{n-k}}{k}$$

Ejemplo, para el conjunto de datos $d = \{-1, 3, 4, 5, 6, 7, 8, 12\}$, la media móvil, con una ventana de tamaño 3 es: $\{2, 4, 5, 6, 7, 9\}$.

Esta medida se emplea en el análisis de series temporales porque suaviza las oscilaciones a corto plazo y realza las tendencias a largo.

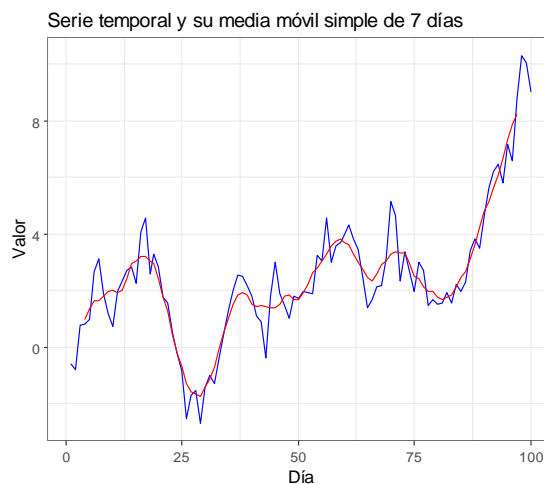


Tabla 3: Variable correspondiente a una serie temporal (azul) y su media móvil a 7 días (rojo).

Medidas de variabilidad o dispersión.

Estas medidas indican lo próximos que están los valores entre sí. Suponga que se miden las presiones de los neumáticos de un taxi. En el primero las presiones son 2, 2.2, 1.8 y 2 bares. En el segundo las presiones son 0.5, 2.5, 4 y 1 bares. En ambos casos la media es de 2 bares, pero la diferencia entre las mediciones del segundo coche es mucho mayor que la del primero. Las medidas de variabilidad sirven para caracterizar este tipo de fenómenos.

Rango.

El rango R es la diferencia entre el valor más alto menos el más pequeño.


$$R = \max - \min$$

Varianza poblacional (σ^2).

Desviación de los valores individuales de una población respecto a la media. Para una población x_1, x_2, \dots, x_n la varianza poblacional es:

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

Siendo μ la media poblacional.

 Elevar al cuadrado la diferencia evita que las diferencias positivas y negativas se cancelen.

Varianza muestral (s^2).

Si la variable no representa a la población completa sino que es sólo una muestra de ella, no se conoce la media poblacional real μ , sino que se conoce la media muestral \bar{x} . La fórmula más precisa para calcularla es:

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

s^2 es un estimador de σ^2 . En su cálculo se utiliza la media muestral (\bar{x}) ya que μ es la media poblacional. La mayor diferencia es el denominador ($n-1$) que hace el cálculo más preciso para muestras pequeñas. Por esto, para el cálculo de la varianza muestral se usa $n-1$. En muestras grandes los resultados de dividir la suma de diferencias entre $n-1$ o entre n se parecen más.

Desviación estándar.

La varianza se mide en la misma unidad de medida que los datos de la variable, pero elevada al cuadrado. Por ejemplo, si una variable se mide en centímetros, su varianza se expresa en cm^2 . Casi siempre resulta más conveniente expresar la dispersión de una variable en la misma unidad que la variable. Para conseguirlo basta con extraer la raíz cuadrada de la varianza. A esta nueva medida se le conoce como desviación estándar poblacional (σ) o desviación estándar muestral (s).

$$\sigma = \sqrt{\sigma^2}; s = \sqrt{s^2}$$

Coefficiente de variación de Pearson.

Es el cociente entre la desviación estándar y la media aritmética:

$$V_x = \frac{\sigma}{|\mu|}$$

Las unidades de σ y de μ son las mismas que las unidades en las que esté mediada la variable. Por tanto, al dividir 2 unidades iguales estas se cancelan. Así que el coeficiente de variación no tiene unidades. Además, se emplea el módulo de la media para evitar valores negativos ya que lo que se está midiendo es una proporción entre la desviación y la media. De hecho, en ocasiones se puede expresar en porcentaje.

Por todo eso, se puede utilizar el coeficiente de variación para comparar 2 variables, aunque no estén medidas en las mismas unidades o tengan la misma escala.

Dadas 2 variables x e y , si $V_x < V_y$ significa que x varía menos que y . O sea, que x es más representativa que y . O que la media de X representa mejor a su distribución que la media de Y a la suya. En general, se considera que de V_x debería ser menor o igual a 0,3.

El coeficiente de variación debe utilizarse con variables medidas en escala de razón cuya distribución sea aproximadamente normal.

Medidas de posición o localización.

Son medidas que dividen la distribución en partes iguales. Se conocen con el nombre de **cuantiles**. Por ejemplo, el cuantil 33, denotado como q_{33} , es el punto de la distribución bajo el que se encuentra el 33% de los datos.

Los cuantiles más habituales se describen a continuación.

Cuartiles

Son un tipo de cuantiles que dividen la distribución en 4 partes iguales. Hay 3 cuartiles Q_1 , Q_2 , y Q_3 , que limitan los puntos donde la distribución acumula, respectivamente, el 25 %, el 50 % y el 75% de los valores. Así, Q_2 se corresponde con la mediana de la distribución.

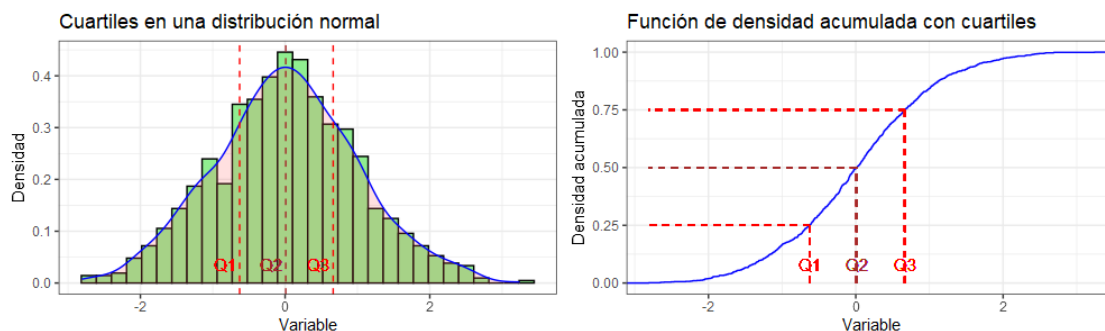


Figura 9: En la gráfica de la izquierda se muestran las posiciones de los cuartiles sobre una distribución normal. A la derecha se muestra la gráfica de densidad acumulada de la misma distribución. Se marca la posición de los 3 cuartiles y su reflejo sobre el eje y.

Deciles

Son los cuantiles que dividen la distribución en 10 partes iguales.

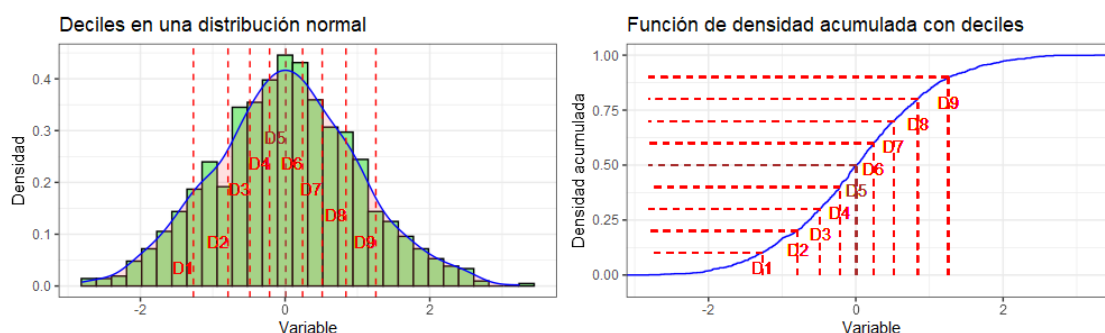


Figura 10: Deciles de la distribución anterior.

Análogamente, el decil 5 se corresponde con la mediana.

Percentiles

Son los percentiles que dividen la población en 100 partes iguales. El percentil 50 se corresponde con la mediana. Los percentiles 25 y 75 se corresponden con los cuartiles 1 y 3 respectivamente.

La diferencia entre los percentiles 75 y 25 se denomina **recorrido intercuartil**. Este recorrido se utiliza en la construcción de diagramas de caja (*box-plots*).

La diferencia entre los percentiles 90 y 10 se denomina **recorrido interdecil**.

La ventaja de estos rangos es que excluyen los valores extremos de la distribución por lo que puede que los cálculos que se hagan con ello sean más precisos.

Medidas de simetría.

Estas medidas se emplean para describir la distribución de los datos respecto al valor central. Los datos pueden distribuirse de forma simétrica respecto al valor central o bien concentrarse al lado derecho o al izquierdo de dicho valor.

Medida de asimetría (skewness).

Sirve para conocer hacia qué lado se desvía la curva con respecto al valor central. La asimetría es positiva cuando la moda se mueve hacia la izquierda de la media produciendo una cola más grande a la derecha. En el caso contrario: cuando la cola mayor es la izquierda porque la moda se ha desplazado a la derecha de la media se dice que hay asimetría negativa. En caso de que ambas colas sean iguales y la moda coincida con la media (y con la mediana) se dice que la distribución es simétrica.

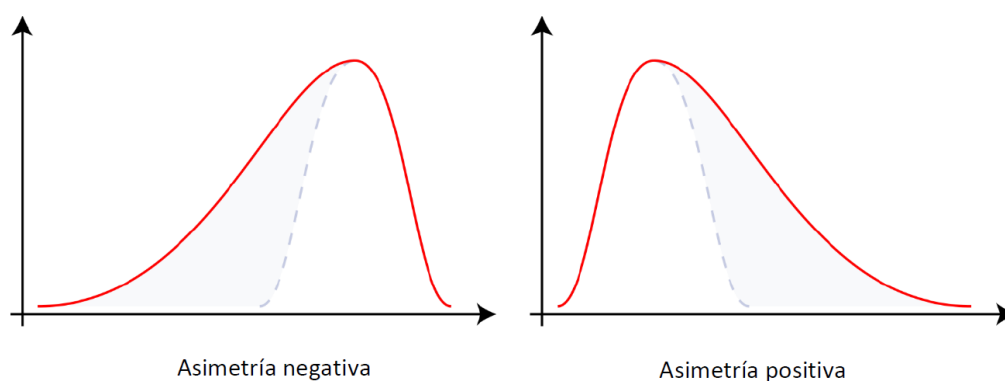


Figura 11: Gráficas con distribuciones asimétricas; negativa y positivamente. Fuente: (Goss-Sampson & Meneses, 2019)

Medidas de apuntamiento (curtosis)

Indica lo "apuntada" que es una distribución. La distribución es muy apuntada si los valores se acumulan mucho alrededor de la media. Un valor de curtosis positivo indica que la distribución es más puntiaguda que la distribución normal. Un valor negativo indica que es más achatado que la normal. El valor 0 indica que es similar a la normal.

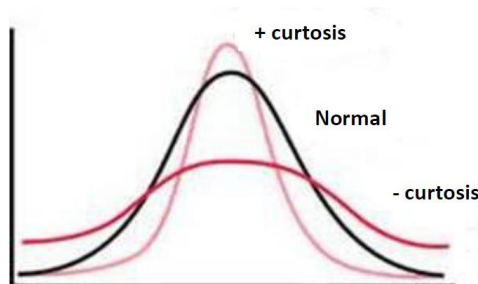


Figura 12: Comparación de gráficas con curtosis positiva y negativa con la distribución normal. Fuente: (Goss-Sampson & Meneses, 2019)

Medidas de concentración.

Se emplean para cuantificar la igualdad en el reparto.

Curva de Lorenz.

Es una representación gráfica que indica cómo se distribuye una variable en una población. Por ejemplo, cómo se distribuye el dinero entre los habitantes de una ciudad. En el eje x se representa el porcentaje acumulado de una la variable; en el eje y se representa el porcentaje acumulado de la otra variable. Si en la ciudad el 80 % de la población dispusiese del 20% del dinero. El punto de corte entre ambos valores sería un punto de la curva de Lorenz.

En un mundo ideal, con la riqueza repartida por igual entre toda la población, la curva de Lorenz sería una línea recta, diagonal, de 45°. En un mundo tiránico en que únicamente una persona tuviese toda la riqueza, la curva de Lorenz serían los 2 catetos del triángulo rectángulo. En la vida real, las curvas de Lorenz no son tan extremas; se parecen más a este ejemplo:

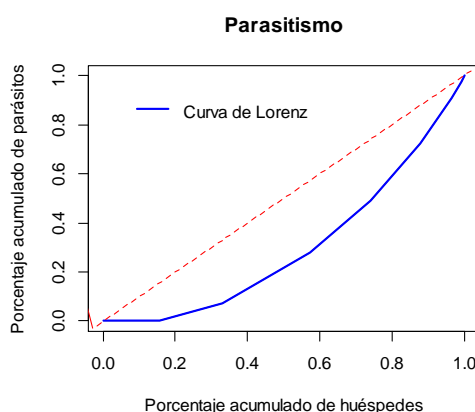


Figura 13: Ejemplo de curva de Lorenz, con datos inventados, de la concentración de los parásitos en la población de hospedadores.

Índice de Gini

Es un índice numérico muy relacionado con la curva de Lorenz porque calcula el porcentaje del área del triángulo entre la diagonal y la esquina que ocupa el área entre la diagonal y la curva. Esta es su fórmula.

$$I_g = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n \cdot \sum_{i=1}^n x_i}$$

Donde:

x_i es el valor del dato i-ésimo.

x_j es el valor del dato j-ésimo.

Un índice próximo a 0 indica que la variable está menos concentrada (todas las muestras tienen el mismo valor; el área de la diferencia entre la curva y la diagonal es próxima a 0).

Un índice próximo a 1 indica que la variable está más concentrada porque hay grandes diferencias entre unos valores y otros.

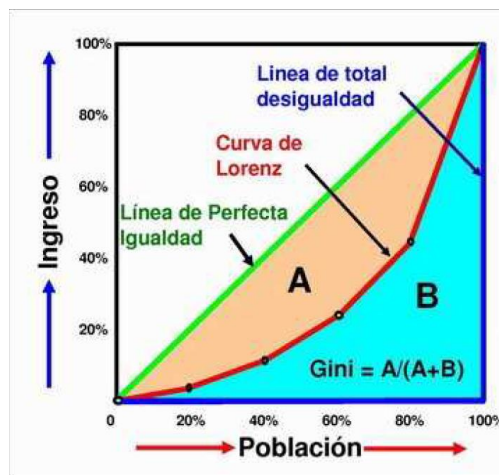


Figura 14: Relación entre el índice de Gini y la curva de Lorenz. Tomado de (Ruiz et al., 2016).

Medidas de precisión.

La precisión es la dispersión que tienen los valores de una medida repetida. A mayor dispersión, menor precisión. No debe confundirse con la exactitud que mide lo cerca, o lejos, que están esas medidas del valor real de la magnitud.

Intervalo de confianza

El intervalo de confianza para la media es una inferencia estadística para estimar el valor real de la media con una cierta probabilidad. Es un rango de los valores mínimo y máximo entre los que esperamos calcular el valor real de la magnitud con la probabilidad deseada. Se calcula así:

$$\text{Intervalo de confianza de } \mu = \left(\bar{x} - \frac{\sigma}{\sqrt{n}} \cdot Z_{\frac{\alpha}{2}}; \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot Z_{\frac{\alpha}{2}} \right)$$

Donde:

\bar{x} es la media aritmética.

σ es la desviación estándar.

α es la probabilidad de error que se permita cometer. Usualmente 0,05

$Z_{\frac{\alpha}{2}}$ es el valor de la constante que deja un valor de $\alpha/2$.

Tablas.

Las tablas son una buena manera de presentar los datos, siempre que se haga de manera clara. Deben estar adaptadas a los resultados que se pretendan presentar. Deben ser fáciles de interpretar por la audiencia que las va a leer. Cuanto más sencillas, son más fáciles de interpretar.

Son un recurso muy utilizado para el análisis de datos que no siempre es adecuado para la representación de los resultados. Pero conviene evaluar hay representaciones gráficas más eficaces que las tablas para transmitir los hallazgos.

Los elementos de una tabla son los siguientes:

Título de la tabla: Descripción clara y concisa de los datos.

Encabezado de las columnas: Se alinean en una fila de la parte superior de cada columna. Deben indicar qué datos hay en cada columna. A menudo son el nombre de una variable del conjunto de datos.

Encabezado de las filas: Se alinean en la primera fila de la tabla. Identifican cada dato. Suelen ser el código de identificación de cada dato (también llamado registro).

Datos: Zona de datos donde se recoge el valor de cada variable para cada registro en la celda intersección entre el código del registro y el nombre de la variable.

Notas al pie: Se colocan en la parte inferior de la zona de datos. Proporcionan información adicional o aclaraciones sobre los datos.

Fuente: En la parte más inferior de la tabla se puede introducir la fuente de procedencia de los datos.

Título de la tabla **Tabla de ejemplo.**

<i>Encabezado de las columnas</i>		
<i>Encabezado de las filas</i>	Variable 1	Variable 2
	Dato 1,1	Dato 1,2
	Dato 2,1	Dato 2,2
<i>Notas al pie</i>		
La tabla de datos está rodeada por bordes azules.		
En cursiva y gris se han añadido etiquetas que indican cada función.		
Los nombres de las filas y columnas están en negrita.		
<i>Fuente</i>	Fuente: Elaboración propia.	

Figura 15: Tabla de ejemplo que denota las partes de una tabla.

Es habitual reservar la última fila y la última columna para mostrar los totales de filas y columnas. Las tablas se clasifican en varios tipos en función de cómo se presentan y analizan sus datos.

Tablas de frecuencia absoluta.

La frecuencia absoluta es el número de veces que una categoría se registra en un conjunto de datos. Las tablas que listan estas frecuencias se llaman tablas de frecuencia absoluta. Por ejemplo, una empresa recoge las características de su flota de vehículos. La variable "combustible" tiene estos valores:

combustible = {"gasoil", "gasolina", "gasoil", "eléctrico", "gasoil", "gasoil", "gasolina"}

La tabla de frecuencias absolutas es esta:

combustible	Freq
eléctrico	1
gasoil	4
gasolina	2

Tabla 4: Tabla de frecuencias absolutas de la variable "combustible".

Las tablas de frecuencia absoluta sirven para resumir el contenido de variables categóricas. También se pueden emplear con variables numéricas, pero conviene que sean pequeñas o estén agrupadas y transformadas en una variable categórica nueva. Por ejemplo, si la edad de los coches anteriores (en años) fuese:

edad = {23, 7, 10, 2, 1, 15, 5}

y se estableciese que los coches de más de 10 años son "viejos", los de menos de 5 son "nuevos" y los demás "intermedios", se podría crear una nueva variable categórica "grupo" que recogiese estos valores. La tabla de frecuencias absoluta de la edad de los coches sería:

Grupo	Freq
intermedios	3
nuevos	2
viejos	2

Tabla 5: Tabla de frecuencias absolutas de la variable "grupo".

Esta técnica es especialmente útil con variables numéricas de gran tamaño.

Tablas de frecuencias relativas.

La frecuencia relativa es el cociente entre la frecuencia absoluta y el total de datos. Se puede expresar en tanto por 1 o multiplicar por 100 y expresarlo en tanto por 100. Por ejemplo, para el caso de la tabla de combustibles, la tabla de frecuencias relativas es esta:

combustible	Freq
eléctrico	14 %
gasoil	57 %
gasolina	29 %

Tabla 6: Tabla de frecuencias relativas de la variable "combustible".

Tablas de contingencia.

Son herramientas que permite organizar, y analizar, la relación entre dos, o más, variables categóricas mediante el recuento de las frecuencias de las diferentes combinaciones de categorías.

Las más simples son tablas de contingencia 2x2, que relacionan únicamente 2 variables, con 2 niveles cada una. Añadiendo más variables se obtienen tablas de contingencia multidimensionales. En los márgenes de la tabla se pueden recoger, también, los totales de las filas y columnas.

		Control		
		Enferma	Sana	TOTAL
Medicina	Enferma	50	35	85
	Sana	40	15	55
	TOTAL	90	50	140

Tabla 7: Ejemplo de tabla de contingencia.

Lista de datos ordenados.

Con variables numéricas es común ordenar los valores de mayor a menor, o viceversa.

Estas listas son interesantes para detectar los valores más importantes o para hallar parámetros como la mediana.

combustible:
gasoil
gasolina
eléctrico

Tabla 8: Lista de coches combustibles de coche ordenada de mayor a menor frecuencia.

Diagrama de tallo y hojas.

Este es un tipo de gráficos que representa de forma sencilla la distribución de una variable. Contiene 3 partes:

Tallo: la cifra o cifras más significativas de cada dato (Ej. decenas, centenas, ...).

Barra de separación: Una barra vertical.

Hojas: La cifra menos significativa (Ej. unidades).

Las cifras del tallo se listan ordenadas verticalmente, después se coloca la barra vertical y después las hojas. La cifra de las hojas se ordena horizontalmente y se introduce una hoja por cada dato, de forma que la anchura de las hojas es proporcional a la frecuencia de cada categoría (tallo). Por ejemplo: Un invernadero produce racimos de flores. Se muestrean 70 racimos para obtener el recuento de flores de cada uno. Se obtienen estos resultados:

$\text{flores} = \{55, 40, 48, 51, 46, 52, 52, 50, 38, 63, 49, 53, 51, 55, 54, 49, 52, 45, 54, 48, 62, 42, 48, 54, 53, 47, 45, 51, 50, 41, 50, 51, 51, 45, 52, 57, 52, 44, 48, 52, 47, 39, 54, 46, 47, 58, 46, 54, 44, 48\}$

El diagrama de tallos y hojas de esta variable es:

3 | 89

4 | 01244

4 | 5556667778888899

5 | 000111112222223344444

5 | 5578

6 | 23

En este diagrama se puede apreciar que la mayoría de racimos rondan las 50 flores, el mínimo tiene 38 flores y el máximo 63. La distribución de los datos tiene forma de variable normal.

Test estadísticos.

Los test estadísticos son herramientas que permiten contrastar hipótesis para aceptarlas o rechazarlas. Se basan en calcular la probabilidad de que la hipótesis sea cierta. Si la probabilidad de que la hipótesis sea falsa es inferior a un nivel preestablecido se admite la hipótesis que se esté contrastando. En caso contrario, se acepta la hipótesis nula (que es la contraria).

Así, para realizar un test estadístico es necesario establecer el nivel de significancia. Por convención, casi siempre se elige el valor 0.5. En casos puntuales, y bien justificados, se puede elegir un nivel inferior, como 0,1 ó 0,01. En cualquier caso, si no hay otra instrucción, por defecto se elige 0,5. Esto significa que se permiten 5 errores de cada 100 test que se hagan. O lo que es lo mismo, se permite un fallo de cada 20.

El segundo paso es definir las hipótesis que se van a contrastar:

Hipótesis nula (H_0): Esta hipótesis indica que no hay diferencia estadísticamente significativa entre el fenómeno que se esté estudiando y el resultado que ocurriría al azar.

Hipótesis alternativa (H_A): Esta hipótesis indica que el efecto que se está evaluando no puede haber sido producido al azar.

Como ejemplo de estos conceptos se puede usar el caso de un estudio sobre la efectividad de un medicamento para prevenir el contagio de una enfermedad. Se harían, al menos, 2 grupos: a uno se les suministraría el medicamento y a otro un placebo. Tras un tiempo, se recoge el número de enfermos en ambos grupos. La hipótesis nula sería "que el medicamento no tiene efecto sobre la enfermedad". La hipótesis alternativa se formularía como que "el medicamento afecta a la cantidad de enfermos en la población" (preferiblemente, la reduce) y, además, esta reducción no se debe al azar.

Según las leyes del azar, es posible que al lanzar una moneda al aire 100 veces seguidas se obtengan 100 caras y ninguna cruz, pero este es un resultado extremadamente improbable. Lo esperable sería que el número de caras se aproximase a 50. Si el resultado fuese 55, sería fácil pensar que se debe al azar. Pero, ¿qué pensarías si en un experimento como este se obtienen 60 caras?, ¿y 70?, ¿y 80? ¿Dónde está el límite para considerar que el resultado se ha producido por azar? Precisamente a esta última pregunta responde el establecimiento del nivel de significancia (α). Como se ha establecido un nivel de significancia de 0,05, se considerarán como debidos al azar aquellos resultados que ocurran por azar en el 95 % de los casos y se aceptará la hipótesis alternativa en aquellos resultados que tengan menos de un 5 % de probabilidad de ocurrir.

El valor numérico del nivel de significancia depende de la forma de la distribución. Determinar el tipo de distribución de una serie de datos puede ser complicado. Puede ser necesario conocer algunas características de la distribución como si es continua o discreta, simétrica o asimétrica, etc. En la Figura 16 se muestra un árbol de decisión para ayudar a la determinación del tipo de distribución de una variable. En ella se aprecia, además, que la forma de algunas distribuciones puede variar notablemente en función de los parámetros que se seleccionen. En general, cuando el número de datos de una distribución es alto, suele asimilarse a una distribución normal.

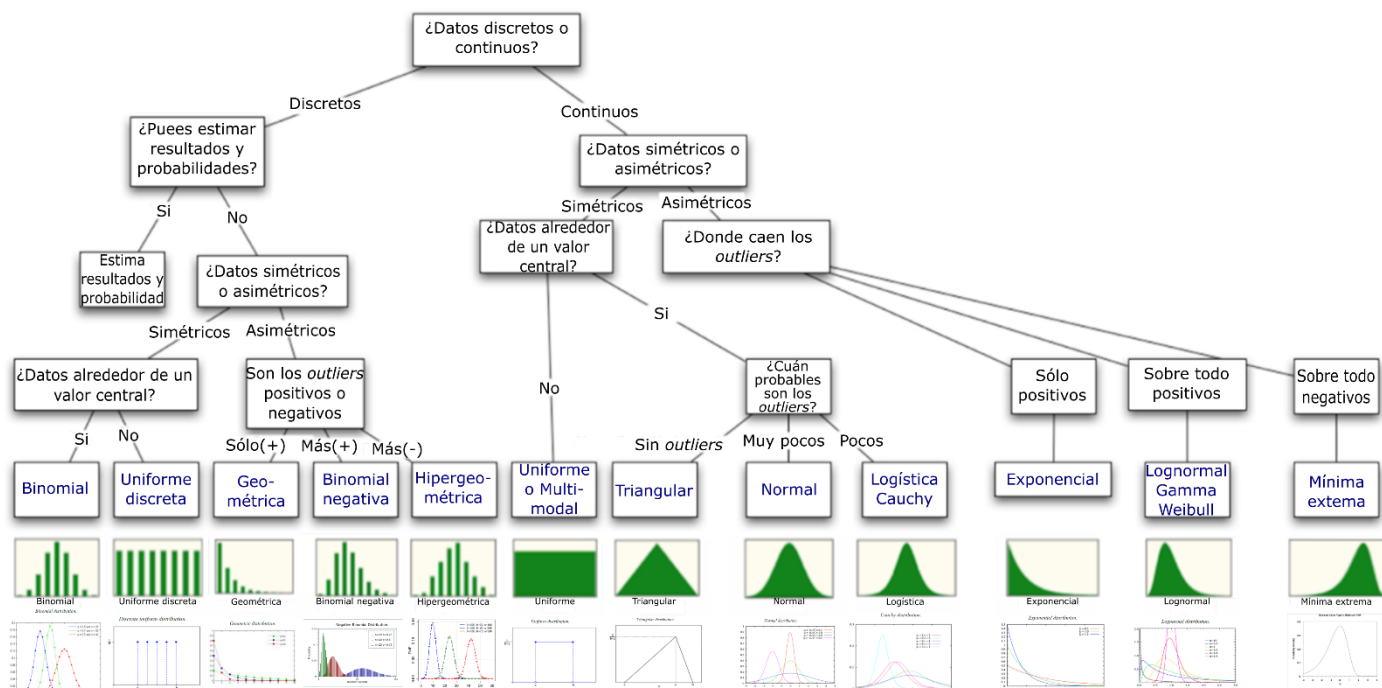


Figura 16: Árbol de decisión para caracterizar las principales distribuciones de probabilidad. La fila de imágenes verdes muestra la forma habitual de la distribución. Debajo se representan algunas otras formas, que dependen de los valores de los parámetros de cada distribución. Adaptado de (Damodaran, 2007), (Statistics How To, 2024) y (Wikipedia, 2024).

El test estadístico a utilizar para cada contraste de variables se elige en función de las características de las 2 variables. En la Figura 17 se muestra un árbol de decisión para seleccionar el test que se debe utilizar para el contraste de hipótesis de 2 poblaciones.

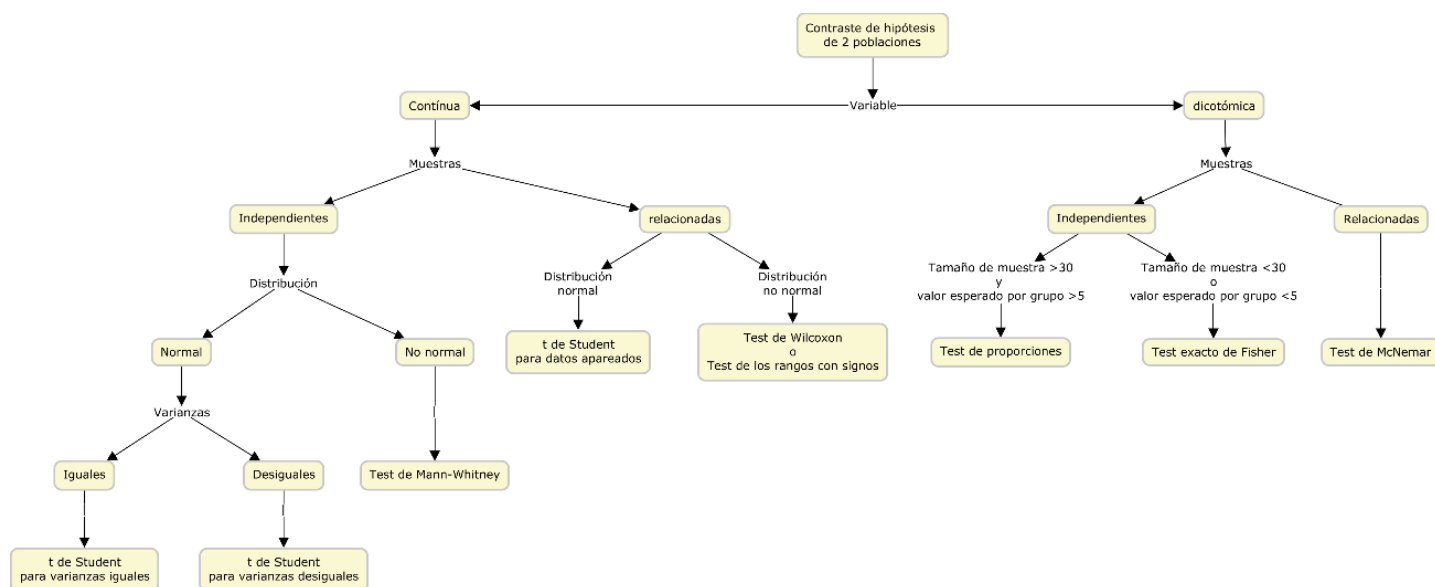


Figura 17: Árbol de decisión para elegir el test adecuado para realizar el contraste de hipótesis de 2 poblaciones.

A continuación, se estudiarán las características de algunos test.

Test de proporciones.

Se emplea para comparar las proporciones de una variable entre 2 poblaciones independientes. La variable es binaria.

La hipótesis nula (H_0) supone la igualdad de proporciones entre las 2 poblaciones ($p_1 = p_2$). Las poblaciones pueden ser independientes. La hipótesis alternativa (H_A) supone la desigualdad de proporciones ($p_1 \neq p_2$). Para establecer el límite en que la desigualdad es significativa se tiene en cuenta el tamaño de la muestra.

Test exacto de Fisher.

Permite evaluar la asociación entre 2 variables dicotómicas, especialmente cuando las muestras son muy pequeñas o frecuencias esperadas menores de 5. En condiciones en las que el test de χ^2 no se pueden utilizar. Suele aplicarse sobre tablas de contingencia de las distribuciones conjuntas de ambas variables categóricas.

La hipótesis nula (H_0) establece que no hay asociación entre las variables. H_A sostiene que si existe asociación.

Este test calcula la probabilidad de obtener los datos observados bajo los supuestos de la hipótesis nula. El resultado se compara con el umbral marcado por el nivel de significación (Ej. $\alpha = 0,05$) para la distribución hipergeométrica.

Test de McNemar.

Es una prueba no paramétrica que se utiliza para comparar las proporciones en variables relacionadas, apareadas, que han sido sometidas a los 2 tratamientos que se desea comparar. Utiliza una tabla de contingencia de 2x2 que resume los resultados de observaciones emparejadas. La hipótesis nula, H_0 establece que no hay diferencia entre las proporciones de la

variable, binaria, en las 2 condiciones. La hipótesis alternativa, H_A , sostiene que si existe una diferencia significativa.

Para 2 variables categóricas relacionadas, que llamaremos tratamiento y control, que pueden tomar 2 valores cada una, que llamaremos Si y No, se puede crear una tabla de contingencia como la siguiente, donde se han nombrado las casillas de datos con letras minúsculas:

		Control	
		Si	No
Tratamiento	Si	a	b
	No	c	d

Tabla 9: Tabla de contingencia genérica para el test de McNemar.

Hay 2 tipos de casilla: aquellas donde el resultado coincide en ambas variables: a y d (Si, Si ó No, No, respectivamente) y aquellas donde ha habido diferencias de pasar del control al tratamiento: b y c. La hipótesis H_0 del test de McNemar establece que: $b = c$ y se calcula usando la esta fórmula:

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

Ejemplo: se puede emplear para comparar la efectividad de 1 fitosanitario en 1 grupo de plantas. Se recogen los recuentos de plantas enfermas antes y después de usar el fitosanitario.

		Control		
		Enferma	Sana	TOTAL
Fitosanitario	Enferma	50	35	85
	Sana	40	15	55
TOTAL		90	50	140

Figura 18: Tabla de contingencia que muestra el número de plantas enfermas y sanas antes de usar el fitosanitario (Control) y después (Fitosanitario).

En este caso, si denominamos a las casillas de esta manera:

		Control	
		Enferma	Sana
Fitosanitario	Enferma	a	b
	Sana	c	d

Figura 19: Esquema con los nombres de las posiciones de cada casilla (letras minúsculas).

la prueba de McNemar sigue esta fórmula:

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

Siendo "b" el número de pares discordantes: plantas que al aplicar el fitosanitario no estaban enfermas.

Y "c" el otro par discordante: plantas que no aplicando el fitosanitario si estaban enfermas. Por tanto:

$$\chi^2 = \frac{(35 - 40)^2}{75} = \frac{25}{75} = 0,333$$

Este valor de χ^2 es inferior al valor para $\alpha=0,05$, con 1 grado de libertad, que es 3.84. Por tanto, determinamos que no podemos descartar la hipótesis nula (H_0) y el efecto del fitosanitario no se diferencia del que se podría haber obtenido por azar.

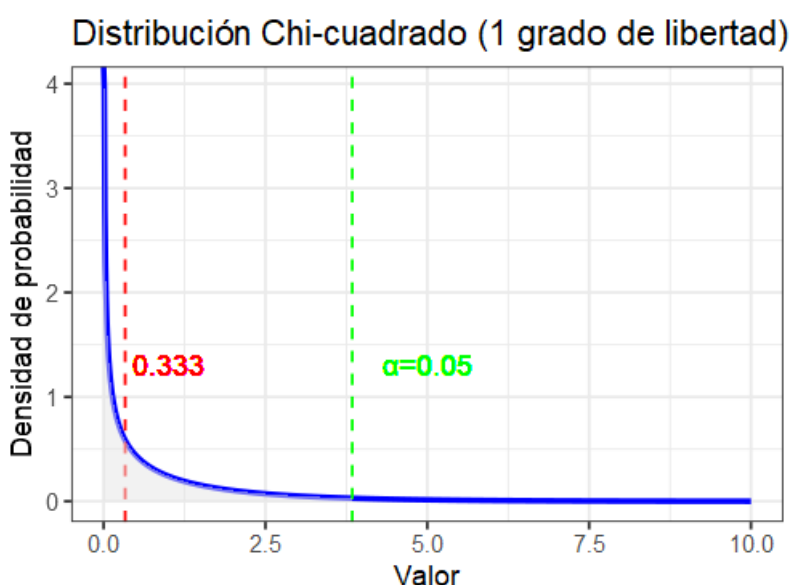


Figura 20: Distribución de probabilidad χ^2 del ejemplo (en azul). En gris se marca la zona de probabilidad correspondiente al azar. En verde el valor para un $\alpha = 0,05$ y en rojo se marca el valor obtenido para los datos del ejemplo.

Ji-cuadrado de Pearson (χ^2).

Se utiliza para determinar si existen diferencias significativas entre los datos observados y su distribución teórica (esperada) para una o más categorías. Es un método no paramétrico. Requiere que todos los valores esperados sean ≥ 5 .

La fórmula es:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Donde:

O_{ij} es la frecuencia observada en la celda ij.

E_{ij} es la frecuencia esperada (bajo la hipótesis nula) en la celda ij .

El valor obtenido se contrasta con el valor crítico para una distribución chi-cuadrado de los grados de libertad pertinentes.

Por ejemplo: Se ha recogido en una tabla de contingencia el sexo y color preferido de 115 personas:

	Hombre	Mujer	TOTAL
Azul	30	20	50
Rojo	10	15	25
Verde	15	10	25
Amarillo	5	10	15
TOTAL	60	55	115

Tabla 10: Tabla de contingencia de preferencias de color por sexo.

Se pretende saber si el sexo influye en las preferencias de color. O sea:

H_0 es que "el sexo no influye" en la preferencia de color.

H_A es que "el sexo influye" en la preferencia de color.

Para elegir una de las 2 hipótesis debemos seguir estos pasos:

Paso 1: Calcular la tabla de frecuencias esperadas (dado que las frecuencias observadas son las mostradas en la Tabla 10). Para ello, se calcula para cada celda la frecuencia que debería tener en función de los totales de su fila y columna. Así:

$$E_{ij} = \frac{Total_i \times Total_j}{TOTAL}$$

Siendo $Total_i$ el total para la fila de la celda; $Total_j$ el total para la columna de la celda y $TOTAL$ el total general de toda la tabla (115).

La tabla de resultados esperados queda así:

	Hombre	Mujer	TOTAL
Azul	26,09	23,91	50
Rojo	13,04	11,96	25
Verde	13,04	11,96	25
Amarillo	7,83	7,17	15
TOTAL	60	55	115

Tabla 11: Tabla de resultados esperados.

Paso 2: Se calcula el estadístico χ^2 aplicando la fórmula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Como resultado se genera este valor del estadístico para cada celda y se suman para obtener el valor total del estadístico:

	Hombre	Mujer	TOTAL
Azul	0,587	0,640	
Rojo	0,710	0,775	
Verde	0,293	0,320	
Amarillo	1,021	1,113	
TOTAL			5,460

Tabla 12: Tabla de valores del estadístico χ^2 para cada celda y valor total para el experimento.

Paso 3: Determinar el valor de probabilidad asociado a un $\alpha = 0,05$ con 3 grados de libertad (grados_de_libertad = n°_lineas - 1 * n°_columnas - 1) en una distribución χ^2 y compararlo con el valor obtenido.

El valor de χ^2 para un $\alpha = 0,05$ y 3 grados de libertad $\approx 7,82$. Como 5,46 es inferior que 7,82, no podemos rechazar la hipótesis nula y, por tanto, podemos concluir que el sexo no influye en las preferencias de color de esta población.

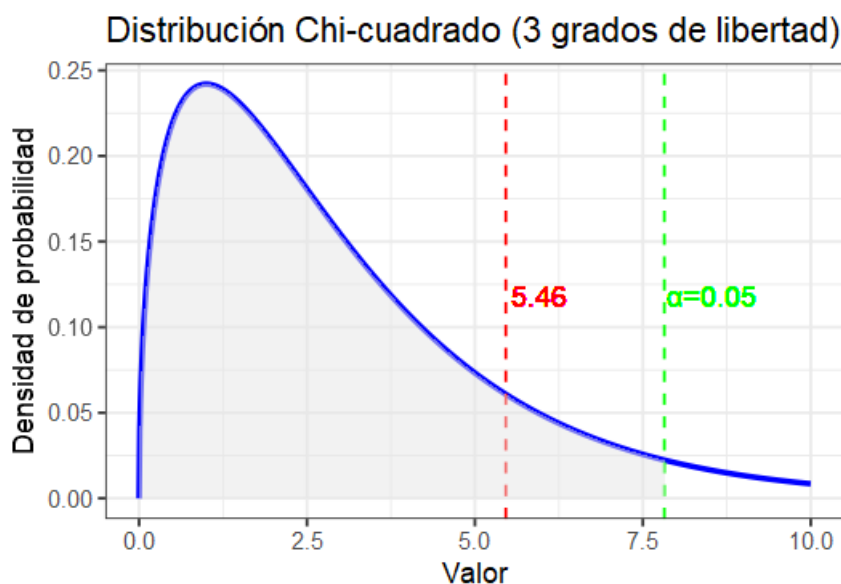


Figura 21: Imagen de la distribución χ^2 (en azul) con los datos del ejemplo. En gris se marca la zona de probabilidad correspondiente al azar. En verde el valor para un $\alpha = 0,05$ y en rojo se marca el valor obtenido para los datos del ejemplo.

Nota: este tipo de test puede realizarse rápidamente con programas estadísticos como R. Para ello basta con utilizar este código:

```
# Crear la tabla de contingencia
```



```
datos_chi <- matrix(c(30, 20, 10, 15, 15, 10, 5, 10), nrow = 4, byrow = TRUE)

colnames(datos_chi) <- c("Hombre", "Mujer")

rownames(datos_chi) <- c("Azul", "Rojo", "Verde", "Amarillo")

# Realizar el test de chi-cuadrado

resultado_chi_cuadrado <- chisq.test(datos_chi)

# Mostrar los resultados

print(resultado_chi_cuadrado)
```

Que produce este resultado:

Pearson's Chi-squared test

data: datos_chi x-squared = 5.4596, df = 3, p-value = 0.1411

Que indica que el valor del estadístico obtenido es de 5,4595, que hay 3 grados de libertad y que la probabilidad de que de obtenga, por azar, un resultado como el observado es del 14,11 % (luego es superior al límite del 5% que tenemos marcado).

Test de Kolmogorov-Smirnov (KS).

Es una prueba no paramétrica que indica la bondad de ajuste de 2 distribuciones continuas entre sí. Se puede aplicar a muestras pequeñas. Sirve para testear si la distribución de una muestra se corresponde con alguna de las distribuciones conocidas. Puede utilizarse con valores esperados < 5 .

La hipótesis nula sería que “no hay diferencias significativas” en el ajuste de las distribuciones. O sea, que la distribución empírica se ajusta a la esperada o que ambas distribuciones son iguales.

La hipótesis alternativa indica que “hay diferencias significativas” entre las 2 distribuciones.

Análisis de clúster.

El análisis de clúster es una técnica de aprendizaje no supervisado que permite agrupar los datos. Los grupos generados se llaman clústeres. Se trata de crear grupos de elementos homogéneos, lo más similares entre sí.

Permite detectar agrupamientos y patrones ocultos en los datos. Cuando se analizan grandes volúmenes de datos es común que haya agrupaciones que no se visualicen fácilmente.

Existen varios algoritmos de análisis de clúster. Quizás el más famoso es el algoritmo de K-means que únicamente requiere indicar el número de clústeres (K) que se quiere obtener; el sólo agrupa los datos de forma que se minimiza la media aritmética de distancias al centro de cada uno de los grupos, mediante un proceso reiterativo.

Los pasos de este algoritmo son:

1. Seleccionar el número K de clústeres deseado. Este paso se realiza manualmente por el analista una vez, antes de comenzar, realmente, el proceso.
2. Establecer los K centroides. La primera vez se realiza al azar. Después, en función de los datos agrupados.

3. Clasificar cada dato como perteneciente a un clúster en función de la distancia a cada uno de los centroides. Se asignan al grupo del centroide más cercano.
4. Se recalculan los centroides en función de los datos agrupados en cada clúster. El centroide será el punto medio de los datos agrupados.
5. Los centroides obtenidos equivalen a los del punto 2. Si ha habido cambio de posición de los centroides, se continúa con el punto 3. En caso de convergencia, el algoritmo ha finalizado.

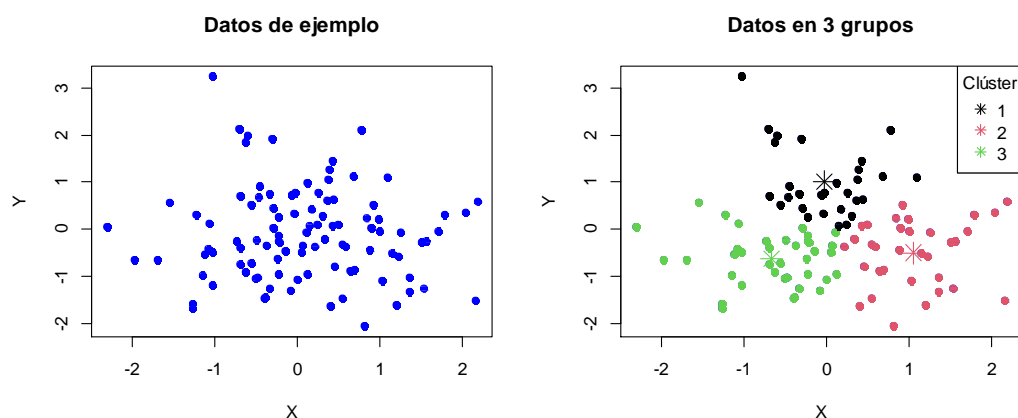


Figura 22: Gráfica de dispersión de 2 variables X e y (izquierda). Tras emplear el algoritmo de K -means, indicándole que busque 3 grupos, se ha generado el gráfico de la derecha donde se marca cada grupo de un color y, además, se añade una estrella en la posición del centroide.

Una modificación de este algoritmo, K -medians, se comporta de forma similar a K -means, pero emplea el valor de la mediana en vez de la media para formar los clústeres. Tiene ventajas ya que hay métodos matemáticos de calcular la mediana, mucho más eficientes que los empleados en el cálculo de la media, que permiten utilizarlo en bases de datos gigantescas.

Resumen

En esta unidad se ha introducido el análisis estadístico de los datos. El análisis de datos tiene raíces muy profundas y llega hasta nuestros días con más fuerza que nunca ya que la cantidad de datos disponibles está creciendo de forma exponencial.

Se han introducido los conceptos básicos de estadística, necesarios para entender el resto de la asignatura. Son especialmente remarcables conocer los tipos de variables y las escalas de medición.

Se ha presentado el procedimiento habitual de manejo de datos que se divide en las fases de extracción, transformación y carga. Se hace hincapié en la importancia de disponer de datos de calidad y se presentan algunas técnicas de transformación de datos que pueden ser útiles.

Se han presentados las estadísticas descriptivas más importante: las medidas de tendencia central, de dispersión, de posición, de simetría, de concentración y de precisión. También se ha hablado de tablas y test estadísticos sencillos y se ha presentado uno más avanzado.

Mapa de contenidos.

1	UD1: INICIACIÓN AL ANÁLISIS ESTADÍSTICO DE DATOS
1.1	CONCEPTOS ESTADÍSTICOS.
1.2	TIPOS DE DATOS:
1.3	ESCALAS DE MEDICIÓN.
1.4	OTROS TIPOS DE DATOS.
1.5	PROCESAMIENTO DE DATOS.
1.5.1	Extracción de datos.
1.5.2	Transformación de datos
1.5.2.1	Datos vacíos o faltantes.
1.5.2.2	Creación de variables derivadas.
1.5.2.3	Creación de variables ficticias (<i>dummy</i>).
1.5.2.4	Reemplazar valores atípicos (<i>outliers</i>).
1.5.2.5	Valores erróneos.
1.5.2.6	Transformación de variables.
1.5.2.6.1	La distribución normal.
1.5.2.6.2	Tipos de transformaciones.
1.5.3	Carga de datos.
1.6	ESTADÍSTICAS DESCRIPTIVAS Y TABLAS DE CONTINGENCIA
1.6.1	Estadísticas descriptivas.
1.6.1.1	Medidas de tendencia central.
1.6.1.2	Medidas de variabilidad o dispersión.
1.6.1.2.1	Rango.
1.6.1.2.2	Varianza poblacional (σ^2).
1.6.1.2.3	Varianza muestral (s^2).
1.6.1.2.4	Desviación estándar.
1.6.1.2.5	Coeficiente de variación de Pearson.
1.6.1.3	Medidas de posición o localización.
1.6.1.3.1	Cuartiles
1.6.1.3.2	Deciles
1.6.1.3.3	Percentiles
1.6.1.4	Medidas de simetría.
1.6.1.4.1	Medida de asimetría (<i>skewness</i>).
1.6.1.4.2	Medidas de apuntamiento (<i>curtosis</i>)
1.6.1.5	Medidas de concentración.
1.6.1.5.1	Curva de Lorenz.
1.6.1.5.2	Índice de Gini
1.6.1.6	Medidas de precisión.
1.6.1.6.1	Intervalo de confianza
1.6.2	Tablas.
1.6.2.1	Tablas de frecuencia absoluta.
1.6.2.2	Tablas de frecuencias relativas.
1.6.2.3	Tablas de contingencia.
1.6.3	Lista de datos ordenados.
1.6.4	Diagrama de tallo y hojas.
1.7	TEST ESTADÍSTICOS.
1.7.1	Test de proporciones.
1.7.2	Test exacto de Fisher.
1.7.3	Test de McNemar.
1.7.4	Ji-cuadrado de Pearson (χ^2).
1.7.5	Test de Kolmogorov-Smirnov (KS).
1.8	ANÁLISIS DE CLÚSTER.

Recursos bibliográficos

Bibliografía básica
<p>M. A. Goss-Sampson y J. Meneses, «Análisis estadístico con JASP: una guía para estudiantes», 2019, Accedido: 8 de febrero de 2024. [En línea]. Disponible en: http://hdl.handle.net/10609/102926</p> <p>G. C. Canavos, «Probabilidad y Estadística. Aplicaciones y Métodos». McGRAW-HILL, 1984.</p>
Bibliografía complementaria
<p>D. Peña, <i>Introducción a la estadística para las ciencias sociales</i>. McGraw-Hill, 1997.</p>
Otros recursos
<p>T. Vigen, «Correlaciones espurias». 12 de febrero de 2024. [En línea]. Disponible en: https://www.tylervigen.com/spurious-correlations</p> <p>DATAtab, «Prueba de Chi-cuadrado - Explicación sencilla - DATAtab». Accedido: 15 de febrero de 2024. [En línea]. Disponible en: https://datatab.es/tutorial/chi-square-test</p> <p>Hista, (2022). ¿Qué Es el Web Scraping? Cómo Extraer Legalmente el Contenido de la Web [web]. URL: https://kinsta.com/es/base-de-conocimiento/que-es-web-scraping/</p> <p>Damodaran, A., (2007). PROBABILISTIC APPROACHES : SCENARIO ANALYSIS, Decision Trees and Simulation. Financial Times 1-61.</p> <p>Ruiz, J., Hernández, L., Hernández-Rodríguez, G., (2016). DESIGUALDAD SALARIAL DE LOS ECONOMISTAS DE LA UNIVERSIDAD VERACRUZANA.</p> <p>Statistics How To, (2024). Negative Binomial Experiment / Distribution: Definition, Examples [web]. URL: https://www.statisticshowto.com/negative-binomial-experiment/</p> <p>Wikipedia, (2024). Hypergeometric distribution [web]. URL: https://en.wikipedia.org/wiki/Hypergeometric_distribution</p>