
Métodos de investigación cuantitativa

PID_00258288

Neus Calaf Gozalo

Tiempo mínimo de dedicación recomendado: 4 horas



Neus Calaf Gozalo

La revisión de este recurso de aprendizaje UOC ha sido coordinada por el profesor: Sergi Fàbregues Feijóo (2019)

Primera edición: febrero 2019
© Neus Calaf Gozalo
Todos los derechos reservados
© de esta edición, FUOC, 2019
Av. Tibidabo, 39-43, 08035 Barcelona
Diseño: Manel Andreu
Realización editorial: Oberta UOC Publishing, SL

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.

Índice

Introducción.....	5
Objetivos.....	6
1. Conceptos estadísticos básicos.....	7
1.1. ¿Qué es la estadística?	7
1.2. Variables y matrices de datos	9
2. Estadística descriptiva.....	11
2.1. Variables cuantitativas	11
2.1.1. Histograma	11
2.1.2. Medidas de tendencia central	12
2.1.3. Medidas de dispersión	14
2.1.4. Medidas de posición	15
2.1.5. Resumen de los cinco números	17
2.1.6. Diagrama de caja o <i>box plot</i>	17
2.2. Variables cualitativas	18
2.2.1. Tabla de frecuencias	18
2.2.2. Diagrama de barras y de columnas	19
2.2.3. Gráfico de sectores	20
2.2.4. Tabla de contingencia	21
3. Probabilidad.....	22
3.1. Experimento aleatorio, espacio muestral y evento	22
3.2. Concepto de probabilidad	22
3.3. Probabilidad condicionada e independencia	24
3.3.1. Probabilidad condicionada	24
3.3.2. Independencia	25
3.4. Variables aleatorias	27
3.4.1. Variables aleatorias discretas	27
3.4.2. Variables aleatorias continuas	28
3.4.3. Esperanza matemática	29
3.4.4. Varianza de una variable aleatoria	29
3.5. Modelos de probabilidad	29
3.5.1. Modelos de probabilidad para variables aleatorias discretas	30
3.5.2. Modelos de probabilidad para variables aleatorias continuas	31
4. Estadística inferencial.....	34
4.1. Estimación de parámetros	34

4.1.1.	Distribución muestral de un estadístico	34
4.1.2.	Intervalos de confianza para la estimación de parámetros	36
4.2.	Introducción al contraste de hipótesis	40
4.2.1.	Contraste de hipótesis: tomar decisiones	40
4.2.2.	Hipótesis nula y alternativa	41
4.2.3.	Uso de los intervalos de confianza para llevar a cabo un contraste de hipótesis	42
4.2.4.	Contraste de hipótesis y pruebas de significación	43
4.2.5.	Errores de tipo I y de tipo II	45
4.2.6.	Potencia de un contraste de hipótesis o prueba de significación	46
Bibliografía.....		49

Introducción

Cuando una persona se plantea iniciar los estudios de Logopedia, en muchos casos no se imagina que deberá cursar materias tan diversas, y menos aún que en algún momento tendrá que estudiar estadística para obtener la titulación. Sin embargo, la estadística es una herramienta imprescindible para resolver problemas y tomar decisiones en cualquier contexto científico. Por lo tanto, será necesario que apartéis el posible «miedo a los números» que podáis tener y considerar que los datos que obtendréis en vuestras investigaciones, o a lo largo de la vida laboral, son los únicos que os podrán asegurar que conseguís los objetivos que os habéis propuesto.

Pero los datos «brutos» o «crudos», sin ordenar ni reducir, no permiten mostrar su significado, sino que conforman una cantidad más o menos ingente de información caótica que hay que estructurar y sintetizar mediante las diferentes técnicas de análisis estadístico. De hecho, esta es la finalidad última del módulo: que el alumnado aprenda a **dar sentido a los datos y a hacerlos interpretables**.

Dado el carácter instrumental de este módulo, además de proporcionar las herramientas conceptuales que permitirán al alumnado analizar los datos obtenidos en sus investigaciones, también proporcionaremos los elementos necesarios para poder analizar de manera crítica los resultados y los procedimientos estadísticos utilizando los diferentes informes de investigaciones (artículos, informes, libros, tesis, etc.) que puedan ser de interés.

Al finalizar este módulo, el alumnado tendrá conocimientos elementales de estadística descriptiva, probabilidad e inferencia estadística, de modo que podrá planificar, analizar e interpretar, rigurosamente y con las técnicas estadísticas apropiadas, los datos obtenidos con los diseños de investigación en cualquiera de las modalidades en un nivel elemental.

Probablemente, este es el primer contacto que muchos de vosotros debéis de tener con la estadística. Esperamos que sea provechoso y motivador para todo el mundo.

Lectura recomendada

A. Cosculluela; A. Fornieles; J. Turbany (2014). *Tècniques d'anàlisi de dades quantitatives*. Material docente de la UOC. Universitat Oberta de Catalunya.

Objetivos

Al finalizar el módulo, el alumnado podrá:

- 1.** Comprender y formular problemas sustantivos de investigación e identificar las variables que intervienen en ellos.
- 2.** Conocer los conceptos básicos para analizar los datos de las investigaciones necesarias para solucionar los problemas de la manera más rigurosa posible, de acuerdo con los criterios científicos.
- 3.** Saber elegir a los sujetos mediante un muestreo cuidadoso que permita obtener muestras representativas de su población.
- 4.** Conocer las características principales de las distribuciones de los datos, la descripción de variables, tanto cuantitativas como categóricas, y la presentación de los resultados mediante la utilización de tablas, índices estadísticos y gráficas.
- 5.** Hacer inferencias y estudiar asociaciones entre variables cuantitativas, teniendo en consideración, asimismo, el concepto de probabilidad que hay detrás de estas operaciones.
- 6.** Conocer la hoja de cálculo de la aplicación gratuita Google Sheets para realizar las operaciones estadísticas y obtener los índices necesarios.

1. Conceptos estadísticos básicos

1.1. ¿Qué es la estadística?

La palabra *estadística* deriva de la palabra *estado*. Durante el siglo XIX, la estadística se consideraba la ciencia del estado. Después fue más allá de este límite y adquirió una aplicación más universal. De hecho, lo cierto es que la estadística penetra en casi todos los aspectos de nuestra vida, y se puede utilizar para conseguir una mejor interpretación de cualquier fenómeno que observemos.

La estadística se basa en la recopilación y el análisis de datos. En el apartado siguiente veréis que la distinción entre los dos tipos de datos posibles (cuantitativos o numéricos, y cualitativos o categóricos) es crucial, dado que el análisis de datos que se puede llevar a cabo depende del tipo de variable. La hora en la que cae el primer rayo en una tormenta, por ejemplo, es una variable numérica en el estudio meteorológico; la presencia o la ausencia de un organismo marino es una variable categórica en el estudio ambiental, y la asignación de trabajos es una variable categórica en el estudio de las leyes discriminatorias.

En cada situación hay un objetivo específico en la recopilación de datos. Por ejemplo, en la recopilación de datos sobre el primer rayo que cae, el meteorólogo o la meteoróloga quiere averiguar a qué hora del día es más probable que caiga un rayo, y el estudio propone prepararse mejor para los peligros que ello supone. En cambio, al reunir datos sobre la altura de un niño, el personal médico quiere determinar el ritmo de crecimiento y comprobar que este es normal.

También podemos observar dos maneras diferentes de recopilar datos. Por un lado, simplemente se observan los datos tal como se dan naturalmente; por ejemplo, cae un rayo y nosotros observamos la hora o el lugar donde cae. Por otro lado, se pueden reunir datos mediante la experimentación. Por ejemplo, en un estudio sobre un fármaco no se estudian a 20.000 personas y se observa simplemente cuáles tienen un ataque de corazón y cuáles han tomado el fármaco para ver si hay una conexión. En este caso, se divide la gente en dos grupos aleatoriamente (como cara o cruz) y después se determina que un grupo tome el fármaco y el otro no. No siempre podemos (o no siempre tiene sentido) llevar a cabo experimentos de este tipo, pero son más potentes a la hora de demostrar resultados verdaderos o causales.

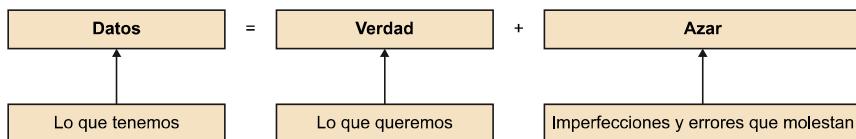
La estadística se utiliza para describir y analizar los datos. Por ejemplo, en el estudio del crecimiento de una niña se observan dos variables, la altura y la edad, y se representan los datos de la altura frente a los de la edad en lo

que denominamos **diagrama de dispersión**, que consiste en una descripción de los datos. Sin embargo, en estudios previos los médicos han establecido el ritmo de crecimiento normal para los niños. Por medio de estos gráficos, el personal médico puede deducir si hay una probabilidad alta de que un niño o una niña no crezca suficientemente deprisa. Este análisis visual de los datos lleva a una conclusión (instaurar el tratamiento).

Otro aspecto que hay que considerar es que **los datos que observamos no son perfectos**. Puede existir todo tipo de errores, tanto en la observación como en la categorización, o en el registro de la información. En las investigaciones también hay que tener en cuenta cuántos sujetos se han elegido para el estudio y cómo se han elegido. Si pudiéramos preguntar a todas las personas de Cataluña, de una en una, si trabajan o no, entonces tendríamos una medida perfecta del grado de ocupación de la población (**población** es el total de elementos sobre los cuales queremos extrapolar nuestro estudio). Sin embargo, habitualmente debemos recurrir a formular la pregunta a una muestra de la población (**muestra** es un subconjunto de la población sobre el cual hacemos nuestro análisis de datos), lo que significa que nuestros datos no son perfectos.

Todos los datos constan de una parte verdadera y de otra errónea, que nosotros denominamos «azar» (figura 1), es decir, un elemento que es imprevisible y que está fuera de nuestro control (aunque esperamos y hagamos todo lo posible para que sea muy pequeño). **El análisis estadístico tiene el propósito de separar la verdad del azar**, de modo que podamos extraer conclusiones firmes de cuanto observamos. Se trata de un tema recurrente en esta asignatura y del cual hablaremos con frecuencia.

Figura 1



Hay una secuencia de acontecimientos común en cualquier investigación que concierne a la estadística:

- En primer lugar, se encuentra la definición de un problema y sus objetivos.
- En segundo lugar, se reúnen datos de las variables relevantes.
- En tercer lugar, se describen y posiblemente se analizan los datos, lo que lleva a una conclusión con relación al objetivo del estudio.

Este módulo se centra principalmente en la tercera parte: la descripción y el análisis de los datos dirigidos a tomar decisiones.

1.2. Variables y matrices de datos

Al planificar una investigación, hay que delimitar los aspectos de la realidad que se quieren investigar. Cuando operamos con dimensiones (características, fenómenos, etc.) que pueden tomar diferentes valores que son medibles, hablamos de **variables**. Recordad que en el módulo «Introducción a la investigación en logopedia» hemos estudiado las diferentes maneras de clasificar las variables. En este módulo estudiaremos cómo analizarlas.

En el plano de notación, cuando consideramos un conjunto de n observaciones numéricas de una variable X , denotamos los valores genéricos con los símbolos x_1, x_2, x_3 , etc., hasta x_n . Denotamos este conjunto de observaciones con $x_1, x_2, x_3 \dots x_n$; o con $x_i, i = 1 \dots n$, en el cual el símbolo i utilizado en los subíndices se denomina **índice**. Así, para los datos de la tabla 1: $x_1 = 9, x_2 = 5, x_3 = 6 \dots x_{27} = 12$.

Tabla 1

9	5	6	8	8	9	12	3	7
3	11	8	4	5	2	6	4	8
17	3	13	11	7	7	4	8	12

A la hora de ordenar las observaciones, de menor a mayor, denotamos el nuevo conjunto de cantidades con los símbolos $x_{(1)}, x_{(2)}, x_{(3)}$, etc., hasta $x_{(n)}$. Por lo tanto, $x_{(1)}$ es el valor más bajo y $x_{(n)}$ es el más alto. En nuestro ejemplo: $x_{(1)} = 2$, y $x_{(27)} = 17$.

Una vez recopilados los datos, el primer paso es tabularlos, es decir, introducirlos en una matriz de sujetos (filas) para variables (columnas). Esta es precisamente la estructura que tienen las hojas de cálculo, como por ejemplo Excel. En la figura 2 se muestra un ejemplo de matriz de datos que recoge los valores de las variables sexo, edad y hándicap vocal, medido con el VHI-10 (Rosen y otros, 2004), de 10 sujetos con alteraciones de la voz que están siguiendo tratamiento logopédico.

Referencia bibliográfica

C. A. Rosen; A. S. Lee; J. Osborne; T. Zullo; T. Murry (2004). «Development and validation of the voice handicap index-10». *The Laryngoscope* (núm. 114, vol. 9, págs. 1549-1556).

Figura 2

Variable categórica Variable numérica

		Matriz de datos			
		Sujeto	Sexo	Edad	VHI-10
Caso	1	Mujer	53	20	
	2	Mujer	63	12	
	3	Hombre	25	12	
	4	Hombre	57	15	Dato
	5	Mujer	19	22	
	6	Mujer	58	5	
	7	Mujer	65	14	
	8	Hombre	58	22	
	9	Mujer	43	29	
	10	Hombre	68	12	

Como veremos más adelante, en el uso de determinadas herramientas estadísticas en el estudio de variables categóricas dicotómicas (como por ejemplo, el sexo), nos hará falta recodificar las variables con 0 y 1, asignando el 1 al valor del atributo estudiado.

2. Estadística descriptiva

Las diferentes técnicas de análisis estadístico tienen el objetivo de dar sentido a los datos y hacerlos interpretables. Por un lado, los conjuntos de datos se pueden organizar, simplificar y resumir mediante procedimientos de **estadística descriptiva**. Por otro lado, se pueden inferir o deducir posibles resultados de una población sometida a estudio a partir del análisis de muestras de la misma población, utilizando procedimientos de **estadística inferencial**.

En este apartado del módulo presentaremos varios procedimientos de estadística descriptiva y los ordenaremos en dos grandes subapartados: procedimientos para variables cuantitativas y procedimientos para variables cualitativas.

2.1. Variables cuantitativas

2.1.1. Histograma

Un **histograma** es una manera de representar gráficamente una distribución de frecuencias de datos cuantitativos. Es un gráfico muy útil cuando queremos ver el aspecto del conjunto de la distribución de un gran número de observaciones.

Para construirlo, se dibuja una barra vertical que muestra el número de valores de nuestros datos que están dentro de cada clase, intervalo o segmento del histograma. Las clases, intervalos o segmentos de un histograma cubren toda la escala de valores de la variable. Hay mucha libertad a la hora de decidir las clases de un histograma. Sin embargo, hay dos consideraciones importantes que se deben tener en cuenta:

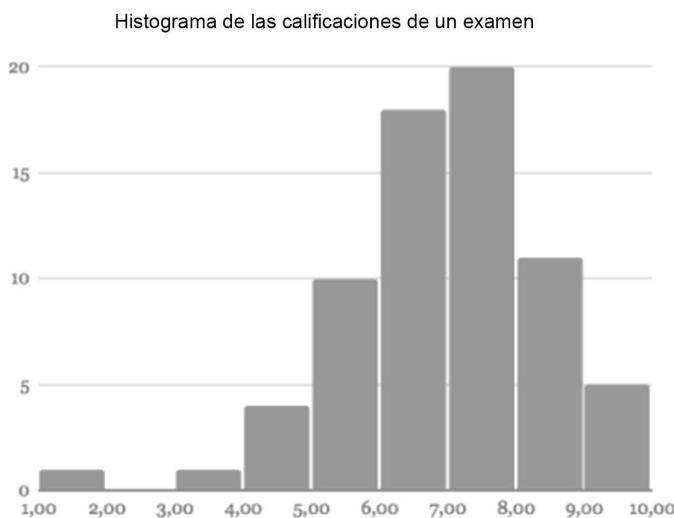
- Todas las clases han de tener la misma amplitud.
- El número de clases depende de la cantidad de datos y del detalle con el que interese ver la distribución. No existe ninguna regla estricta para hacerlo, únicamente es una cuestión de sentido común.

Se puede dibujar y personalizar un histograma con las opciones «Inserta > Gráfico» y «Tipo de gráfico: Gráfico» de histogramas de la aplicación Google Sheets.

Ejemplo

En la figura 3 podéis encontrar un ejemplo de histograma de las calificaciones que un grupo de setenta estudiantes ha obtenido en un examen del grado de Logopedia:

Figura 3



Para interpretar un histograma hay que examinar los patrones generales y después buscar las desviaciones. En nuestro histograma, por ejemplo, vemos que el centro de la distribución está entre los 6 y los 8 puntos. Las cantidades pequeñas de valores que se separan de la distribución se denominan **valores alejados** o **insólitos**. En nuestro ejemplo vemos uno de estos valores insólitos, correspondiente a una calificación de entre 1 y 2.

Si el histograma no es simétrico, la parte larga y arrastrada de la distribución asimétrica se denomina **cola**. Una distribución puede ser asimétrica por la izquierda o asimétrica por la derecha. En la práctica, es más frecuente encontrar la asimetría por la derecha.

2.1.2. Medidas de tendencia central

En este apartado veremos tres maneras de medir en un solo valor el centro de una distribución: la mediana, la media aritmética y la moda.

Mediana u observación central

La **mediana** es el valor que **divide la distribución de los datos en dos partes iguales** (deja un 50% de valores por encima y otro 50% por debajo). Se trata, pues, de un **índice de posición**.

Para encontrar la mediana, debemos ordenar los datos de más pequeños a más grandes y encontrar la observación que queda exactamente en medio, lo que implica que la mitad de las observaciones se sitúa por debajo de este valor y la otra mitad por encima. Dado que es un índice de posición, la mediana no queda afectada por la presencia de valores extremos. Por eso decimos que es un índice resistente o robusto.

El valor de la mediana de un conjunto de datos numéricos se puede obtener con la función «MEDIAN» de la aplicación Google Sheets.

Media aritmética o valor medio

La **media aritmética**, por el contrario, es un **índice de peso basado en el momento de la distribución** (en realidad, la podemos definir como el centro de gravedad de la distribución) y se calcula sumando todos los valores de los datos y dividiendo este sumatorio por el número de observaciones (n).

El valor medio numérico de un conjunto de datos se puede obtener con la función «AVERAGE» de la aplicación Google Sheets.

Tanto la mediana como la media aritmética miden el centro de la distribución, pero lo hacen de manera diferente. Solo cuando la distribución es simétrica, las dos medidas coinciden. La principal diferencia entre ambas es cómo están afectadas por las asimetrías o por los datos alejados.

Ejemplo

Imaginemos que a lo largo de un periodo de veintisiete días anotáis el rato que debéis esperar hasta que el autobús llega por la mañana. Los datos, en minutos, se muestran en la tabla 2.

Tabla 2: Tiempo de espera hasta que llega el autobús, en minutos

9	5	6	8	8	9	12	3	7
3	11	8	4	5	2	6	4	8
17	3	13	11	7	7	4	8	12

La mediana de estos valores es 7 (función «MEDIAN»), mientras que la media aritmética, en cambio, es 7,41 (función «AVERAGE»). En el caso de la media aritmética, hay que tomar precauciones con los datos alejados o insólitos. Cuando la distribución es asimétrica, como en este caso, la media aritmética siempre se desplaza hacia la cola de la distribución.

La presencia de un valor muy elevado no afecta a la mediana, pero influye mucho en la media aritmética. Decimos que la mediana «resiste» a los datos alejados. Por ejemplo, imaginemos que, en lugar de 17 minutos, el valor más alto en los datos del ejemplo fuera 45 minutos, que es una espera muy larga para un solo día. Este cambio no afecta a la mediana, de hecho se mantiene igual, incluso si lo cambiáramos por un valor mucho más elevado. La media aritmética, en cambio, quedaría afectada, puesto que la suma de todas las observaciones sería 228, que, dividido por 27, da un valor de 8,44 minutos. Este incremento de una observación hace subir la media aritmética del tiempo de espera en un minuto, a pesar de que los otros veintiséis valores se mantengan intactos. En una situación como esta, la media aritmética pierde la condición de ser un valor representativo.

Moda

La **moda** es el valor que se repite más veces en un conjunto de datos. Se puede aplicar tanto a variables cuantitativas como cualitativas.

El valor que aparece con más frecuencia en un conjunto de datos se puede obtener con la función «MODE» de la aplicación Google Sheets. En nuestro ejemplo de los minutos de espera del autobús, el valor que se repite más es 8.

2.1.3. Medidas de dispersión

En el apartado anterior hemos definido tres maneras de calcular los índices del centro de una distribución. No obstante, para completar la descripción de las variables cuantitativas hay que añadir los índices de dispersión, que indican hasta qué punto las observaciones se dispersan alrededor del centro.

Varianza

La **varianza** se puede definir como la media aritmética de los cuadrados de las diferencias que hay entre cada valor y la media aritmética. Esto hace que, cuanto mayores sean estas diferencias o distancias (más dispersa o heterogénea sea la variable), mayor será el valor de la varianza.

El hecho de que las diferencias se eleven al cuadrado:

- Evita la presencia de valores negativos (si no se elevaran las diferencias al cuadrado, al haber algunos valores por encima y otros por debajo de la media, el sumatorio sería 0).
- Hace que las diferencias mayores pesen más en el valor del índice.
- Implica que la varianza sea siempre de signo positivo y esté en la unidad de medida de la variable elevada al cuadrado (por ejemplo, el cociente intelectual, CI, tiene en la población una media $\mu = 100$ puntos de CI, y una varianza $\sigma^2 = 225$ puntos² de CI).

La varianza, basándose en una muestra, se puede obtener utilizando la función «VAR» de la aplicación Google Sheets. En nuestro ejemplo de los minutos de espera del autobús, la varianza es de 12,94 minutos.

Para facilitar la interpretación, en lugar de la varianza se suele presentar la raíz cuadrada, que, por lo tanto, ya se encuentra en las mismas unidades de medida que la variable. Este índice se denomina **desviación estándar, tipo o típica**.

Desviación estándar, tipo o típica

La desviación **estándar, tipo o típica** es la raíz cuadrada de la varianza. Se trata de un valor único que resume la dispersión de los datos, en concreto, la dispersión alrededor de la media aritmética. Es uno de los índices de dispersión más utilizados. Por ejemplo, la desviación tipo del CI en la población es de $\sigma = 15$ puntos de CI.

La desviación estándar, basándose en una muestra, se puede obtener utilizando la función «STDEV» de la aplicación Google Sheets. En nuestro ejemplo de los minutos de espera del autobús, la desviación estándar es de 3,6 minutos.

Las extensiones de diferentes distribuciones se pueden comparar simplemente comparando las respectivas desviaciones estándar. Ahora bien, cuando se trata de comparar variables medidas en unidades diferentes, debemos utilizar el **coeficiente de variación (CV)**, que elimina las unidades de las variables, lo cual facilita la comparación. El cálculo de este coeficiente es muy sencillo: se obtiene dividiendo la desviación estándar entre la media aritmética. El resultado es típicamente menor que 1, pero para su mejor interpretación se expresa como porcentaje multiplicando el resultado por 100.

2.1.4. Medidas de posición

En el apartado anterior ya hemos visto una medida de posición: la mediana. Recordemos que la mediana es el valor que divide la distribución de los datos en dos partes iguales, dejando un 50% de valores por encima y otro 50% por debajo. Otras medidas de posición son los **percentiles** y los **cuartiles**.

Percentiles

Los **percentiles** ($P_1, P_2, P_3, \dots, P_{99}$) son los 99 valores que dividen la distribución en 100 partes iguales. En cada parte estará el 1% de la distribución. El percentil k (P_k) es el valor que deja por debajo el k por ciento de las puntuaciones de una distribución. Por ejemplo, el percentil 90 (P_{90}) es el valor que deja por debajo el 90% de los datos de una distribución y es superado por el 10% restante de los datos.

El valor correspondiente a un percentil determinado de un conjunto de datos se puede obtener con la función «PERCENTILE» de la aplicación Google Sheets. En nuestro ejemplo del tiempo de espera del autobús, el percentil P_{95} es 12,7 minutos.

El **rango percentil**, en cambio, es la medida inversa del percentil. El rango percentil de un valor determinado se define como el porcentaje de datos con valores inferiores a este valor, y permite evaluar la posición relativa de un valor en un conjunto de datos.

La clasificación porcentual (percentil) de un valor especificado de un conjunto de datos se puede obtener con la función «PERCENTRANK» de la aplicación Google Sheets. En nuestro ejemplo del tiempo de espera del autobús, el rango percentil de 8 minutos es 0,54 (esto significa que el valor 8 minutos tiene el 54% de los valores por debajo).

Ejemplo

En la evaluación logopédica de los trastornos del lenguaje, los percentiles se utilizan para situar a un usuario respecto a la media poblacional y así poder valorar la severidad del trastorno que presenta. Esto es posible gracias a la estandarización y a la normalización de los instrumentos de evaluación, que permiten comparar las puntuaciones obtenidas por el usuario con las puntuaciones típicas del mismo grupo de edad.

Cuartiles

Los **cuartiles** (Q_1 , Q_2 y Q_3) son los tres valores que dividen la distribución en cuatro partes iguales. En cada parte estará el 25% de la distribución. Equivalen a los percentiles P_{25} , P_{50} y P_{75} , respectivamente (o, dicho de otro modo, al 25.^º, 50.^º y 75.^º percentiles). El primer cuartil es el valor que deja el 25% de las observaciones por debajo, el segundo coincide con la mediana y, por lo tanto, es el valor que divide la distribución en dos partes iguales, y el tercer cuartil corresponde al valor que deja el 75% de los valores por debajo (y, lógicamente, queda el 25% por encima).

El cálculo de los cuartiles es muy sencillo, dado que podemos decir que los cuartiles 1.^º y 3.^º son la mediana de las dos mitades de la distribución que quedan definidas por la mediana. Una vez calculados los cuartiles, restando el 3.^º del 1.^º ($Q_3 - Q_1$), podemos obtener el **rango intercuartílico**, que nos indica cuál es la dispersión del 50% central de las observaciones.

El valor más próximo a un cuartil especificado de un conjunto de datos se puede obtener con la función «QUARTILE» de la aplicación Google Sheets. En nuestro ejemplo del tiempo de espera del autobús, el valor del tercer cuartil Q_3 es de 9 minutos, y el valor del primer cuartil Q_1 es de 4,5 minutos. El rango intercuartílico ($Q_3 - Q_1$) es de 4,5 minutos.

Ejemplo

En las publicaciones científicas se utilizan los cuartiles para evaluar el posicionamiento relativo de una revista dentro del total de revistas de la misma materia. Así pues, las revistas del Q_1 son las que están entre el 25% de revistas con un factor de impacto más elevado (entendiendo el factor de impacto de una revista como la media de veces que, en un año en concreto, se citaron los artículos publicados por la revista durante los dos años anteriores).

2.1.5. Resumen de los cinco números

El **resumen de los cinco números** de los datos de una distribución (o sumario de Tukey) da una idea rápida de la tendencia central y de la dispersión de un conjunto de datos. Estos cinco números son los siguientes:

- **Mínimo:** valor más pequeño de la muestra.
- Q_1 : primer cuartil o percentil P_{25} .
- **Mediana:** Q_2 , segundo cuartil o percentil P_{50} .
- Q_3 : tercer cuartil o percentil P_{75} .
- **Máximo:** máximo valor de la muestra.

El valor mínimo de un conjunto de datos numéricos se puede obtener con la función «MIN» de la aplicación Google Sheets, y el valor máximo, con la función «MAX». Como hemos visto, el valor de los cuartiles se obtiene con la función «QUARTILE», y el de la mediana, con la función «MEDIAN».

A partir de estos números también se puede obtener el **rango intercuartílico** ($Q_3 - Q_1$) y el **rango** (o **recorrido** o **amplitud**) de una variable, que se obtiene calculando la diferencia entre el valor máximo y el mínimo. El rango intercuartílico nos indica cuál es la dispersión del 50% central de las observaciones. El rango (o recorrido) indica la dispersión del 100% de las muestras y, por desgracia, se trata de un índice de escasa utilidad, dado que un único valor extremo o insólito puede hacer que pierda gran parte de su sentido informativo.

2.1.6. Diagrama de caja o *box plot*

El resumen de los cinco números se puede representar gráficamente mediante un **diagrama de caja** o *box plot*. Este gráfico es de gran utilidad porque, además de ser una representación gráfica de la variable, permite comparar distribuciones de la misma variable provenientes de diferentes muestras o subgrupos.

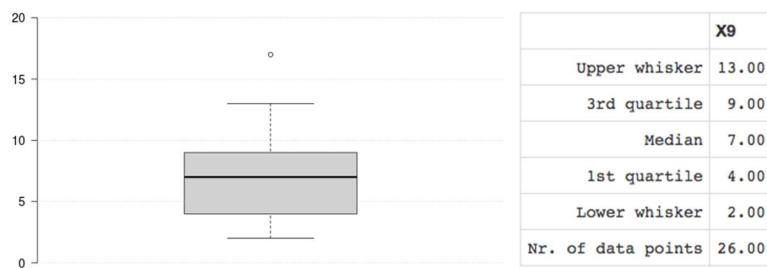
Una manera muy sencilla de generar diagramas de caja es por medio de la herramienta web BoxPlotR¹.

⁽¹⁾Herramienta web disponible en:
<http://shiny.chemgrid.org/boxplotr/>.

Ejemplo

En la figura 4 se muestra el *box plot* de los datos de espera del autobús en minutos generado con BoxPlotR. A la izquierda vemos el gráfico y, a la derecha, los datos estadísticos a partir de los cuales se genera el diagrama. Fijaos en que la herramienta genera el diagrama con 26 puntos en vez de 27. Esto es así porque considera que el dato máximo (17 minutos) es un valor alejado o insólito, y lo marca con un punto fuera del diagrama de caja.

Figura 4



2.2. Variables cualitativas

2.2.1. Tabla de frecuencias

La **tabla de frecuencias** o distribución de frecuencias es una tabla de los datos estadísticos donde se asigna a cada dato su correspondiente frecuencia.

Ejemplo

Considerad, por ejemplo, la tabla de frecuencias (tabla 3) en la cual aparece la nacionalidad de los asistentes a un congreso de logopedia que se ha hecho en París. Los datos ya están recogidos en forma de frecuencia, con las proporciones y los porcentajes calculados. La variable categórica es **país**, con diecisiete países europeos como categorías.

Tabla 3

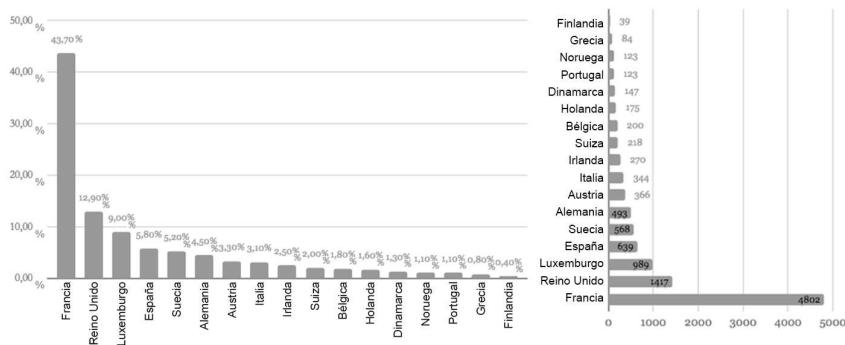
País	Número de asistentes	Proporción	Porcentaje
Francia	4.802	0,437	43,7%
Reino Unido	1.417	0,129	12,9%
Luxemburgo	989	0,090	9,0%
España	639	0,058	5,8%
Suecia	568	0,052	5,2%
Alemania	493	0,045	4,5%
Austria	366	0,033	3,3%
Italia	344	0,031	3,1%
Irlanda	270	0,025	2,5%
Suiza	218	0,020	2,0%
Bélgica	200	0,018	1,8%
Holanda	175	0,016	1,6%
Dinamarca	147	0,013	1,3%
Noruega	123	0,011	1,1%
Portugal	123	0,011	1,1%

País	Número de asistentes	Proporción	Porcentaje
Grecia	84	0,008	0,8%
Finlandia	39	0,004	0,4%
Sumatorios	10.997	1	100%

2.2.2. Diagrama de barras y de columnas

Podemos representar los datos del apartado anterior tanto con un gráfico de columnas como con un gráfico de barras (ved figura 5). En el primer caso, hemos optado por hacer un gráfico de columnas con porcentajes y, en el segundo, hemos hecho un gráfico de barras con las frecuencias.

Figura 5



Se puede dibujar y personalizar un diagrama de barras o de columnas con las opciones «Inserta > Gráfico» y «Tipo de gráfico: Gráfico de barras/de columnas» de la aplicación Google Sheets.

El histograma y el diagrama de barras son dos sistemas de representación muy similares que permiten visualizar la distribución de una variable. Una diferencia entre ambos es que el histograma se construye para una variable cuantitativa después de decidir un conjunto de clases adecuadas, mientras que el diagrama de barras se construye para una variable categórica en la cual las clases ya están hechas. Otra diferencia es que en el histograma las barras se tocan, mientras que en un diagrama de barras están separadas por espacios. Otra diferencia es que las categorías no están en ningún orden específico y, por lo tanto, podemos ordenarlas para que el diagrama de barras sea más fácil de interpretar.

2.2.3. Gráfico de sectores

Otra posible representación es con **gráficos de sectores** (los popularmente denominados «quesitos» o «pasteles»). Estos gráficos pueden expresarse tanto con las frecuencias (número de elementos en cada categoría) como con los porcentajes.

Se puede dibujar y personalizar un gráfico de sectores con las opciones «Inserta > Gráfico» y «Tipo de gráfico: Gráfico circular» de la aplicación Google Sheets.

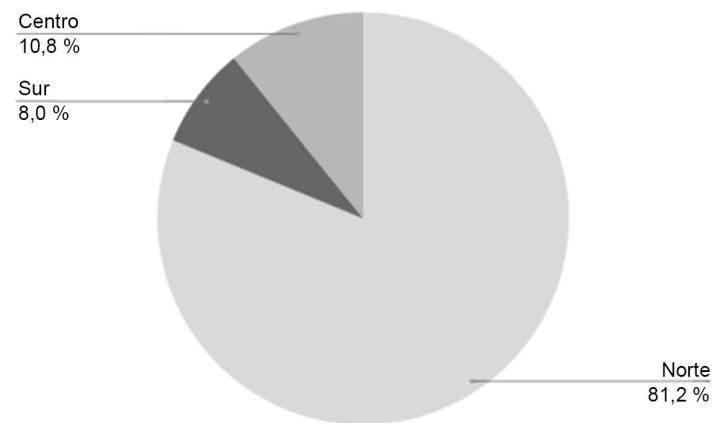
Ejemplo

Para ejemplificar este tipo de gráficos (ved la tabla 4 y la figura 6), agruparemos los países en «Norte» (Suecia, Dinamarca, Noruega y Finlandia), «Centro» (Francia, Reino Unido, Luxemburgo, Alemania, Austria, Irlanda, Suiza, Bélgica y Holanda) y «Sur» (España, Italia, Portugal y Grecia), y haremos la tabla de frecuencias correspondiente y el gráfico de sectores.

Tabla 4

Zona	Número de asistentes	Proporción	Porcentaje
Norte	8.930	0,81	81,2%
Sur	877	0,08	8,0%
Centro	1.190	0,11	10,8%
Sumas	10.997	1	100%

Figura 6



2.2.4. Tabla de contingencia

La **tabla de contingencia** es una tabla de doble entrada en la que se representan de manera conjunta dos variables categóricas, una en las filas y otra en las columnas. Se identifica por su orden, que es igual al número de categorías de la variable dispuesta en filas (k) y por el número de categorías de la variable dispuesta en columnas (l).

Ejemplo

La siguiente tabla de contingencia 3×2 representa las variables «estado civil» y «estudios universitarios», y muestra las proporciones de cada una de las categorías.

Tabla 5

	Est. univ. No	Est. univ. Sí	Total
Soltero/a	0,18	0,12	0,30
Casado/a	0,20	0,23	0,43
Otros	0,11	0,16	0,27
Total	0,49	0,51	1

Dentro de la tabla, es decir, en los cruces de las categorías de una variable con las de la otra variable, hallamos las proporciones conjuntas. En las filas y en las columnas «Total» encontramos las proporciones que corresponden a cada una de las categorías de las dos variables, denominadas **marginales de la tabla**. Así pues, vemos que, por ejemplo, la proporción de solteros con estudios universitarios es $P(S \text{ y } S) = 0,12$. También vemos que la proporción de sujetos casados es de 0,43, o que la proporción de personas sin estudios universitarios es de 0,49.

Del mismo modo que hemos utilizado las proporciones, podemos utilizar los porcentajes (multiplicando las proporciones por 100), o las frecuencias.

3. Probabilidad

La importancia del estudio de la probabilidad en el ámbito de la estadística se deriva del hecho de que es uno de los pilares teóricos fundamentales sobre los cuales se asientan el desarrollo y la aplicación de la **estadística inferencial** (ved el apartado siguiente).

3.1. Experimento aleatorio, espacio muestral y evento

Un **experimento aleatorio** es aquel en el que no se puede predecir el resultado que saldrá, pero del cual sí que se conocen todos los resultados posibles (por ejemplo, lanzar un dado o una moneda al aire).

Se denomina **espacio muestral** a todo el conjunto de resultados posibles en una determinada situación. El espacio muestral, que también recibe el nombre de **conjunto muestral**, se simboliza con la letra griega Ω .

Un ejemplo de espacio muestral son los seis resultados posibles cuando se lanza un dado. En este caso:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Otro ejemplo es lanzar dos monedas al aire. En este caso:

$$\Omega = \{\text{CC}, \text{CX}, \text{XC}, \text{XX}\}$$

Un **suceso o evento** de un experimento aleatorio es un subconjunto del conjunto de posibles resultados de Ω ; por ejemplo, sacar un número par cuando se lanza un dado. Cuando el evento contiene un solo punto muestral se denomina **evento elemental**; por ejemplo, sacar un cuatro cuando se lanza un dado.

3.2. Concepto de probabilidad

La **probabilidad** mide la posibilidad (probabilidad) de que pueda ocurrir un determinado evento cuando se lleva a cabo un experimento aleatorio. Según la definición clásica, la probabilidad es el cociente entre el número de casos favorables y el número total de casos posibles:

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

El inconveniente de esta definición es que solo es válida o aplicable en situaciones en las que todos los casos posibles son equiprobables, es decir, tienen la misma probabilidad de aparecer.

La definición axiomática de probabilidad intenta resolver el problema de la equiprobabilidad y, a partir de un espacio muestral determinado Ω , asigna a cada evento A un número real, simbolizado por $p(A)$, para que cumpla los axiomas siguientes:

La probabilidad de un evento A cualquiera oscila entre 0 y 1, es decir:

$$0 \leq p(A) \leq 1$$

La probabilidad del evento seguro es igual a 1 y la del evento imposible es igual a 0, es decir:

$$p(\Omega) = 1 \text{ e } p(\emptyset) = 0$$

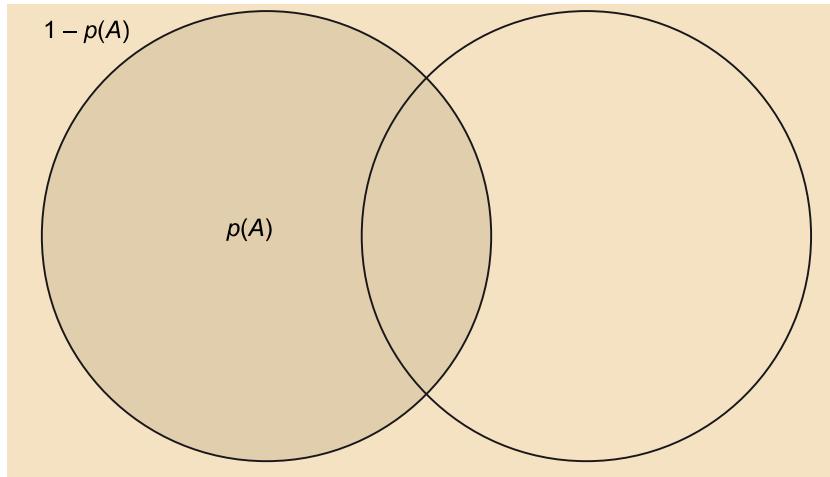
Si $A_1, A_2 \dots A_n$ son eventos mutuamente excluyentes, por lo tanto:

$$p(A_1 \cup A_2 \cup \dots \cup A_n) = p(A_1) + p(A_2) + \dots + p(A_n)$$

La probabilidad del complementario del evento A es igual a 1 menos la probabilidad del evento A , es decir:

$$p(A^C) = 1 - p(A).$$

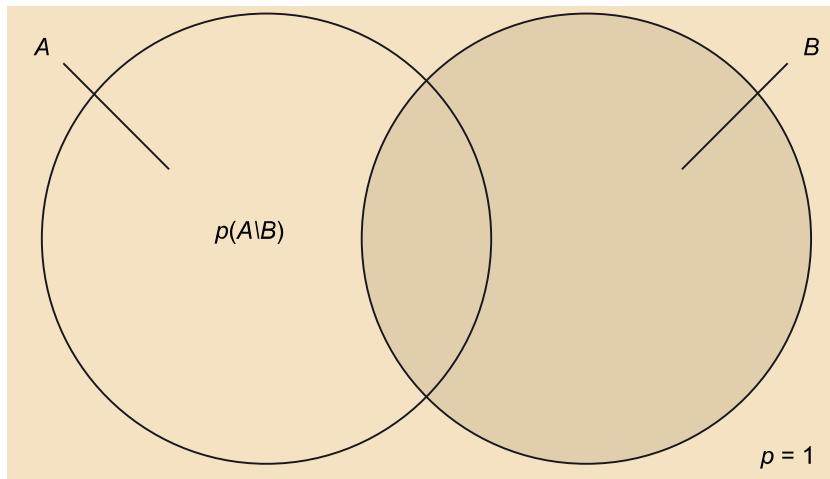
Figura 7



La probabilidad del complementario relativo del evento B respecto al evento A es igual a la probabilidad del evento A menos la probabilidad de la intersección de los dos eventos, es decir:

$$p(A \setminus B) = p(A) - p(A \cap B)$$

Figura 8



Si el evento $A <$ evento B , entonces $p(A) < p(B)$.

Si $A \subset B$, entonces $p(B) \geq p(A)$ y $p(B-A) = p(B) - p(A)$

$$p(A \cap B) \leq p(A \cup B)$$

3.3. Probabilidad condicionada e independencia

3.3.1. Probabilidad condicionada

Dados los eventos A y B , y siendo la probabilidad asociada al evento B superior a 0 [$p(B) > 0$], la probabilidad de que aparezca el evento A si se ha producido el evento B se denomina **probabilidad condicionada** de A dado B , [$p(A/B)$], y se define como:

$$p(A / B) = \frac{p(A \cap B)}{p(B)}$$

Gráficamente queda representado en las figuras 9 y 10.

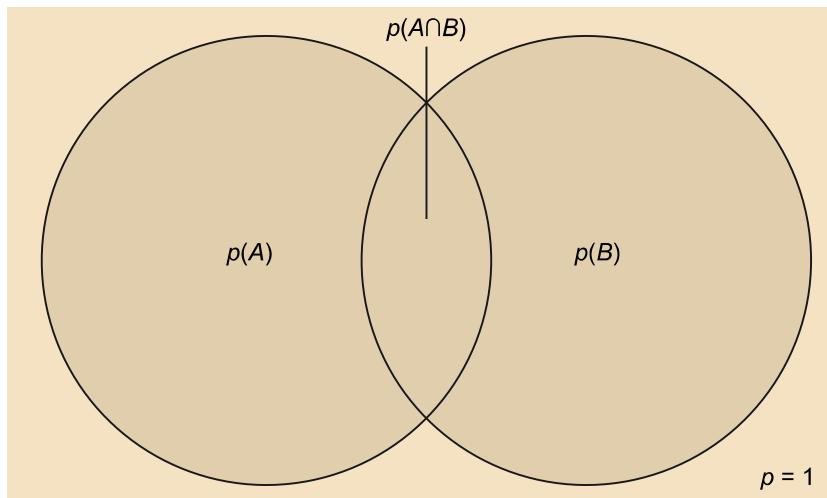
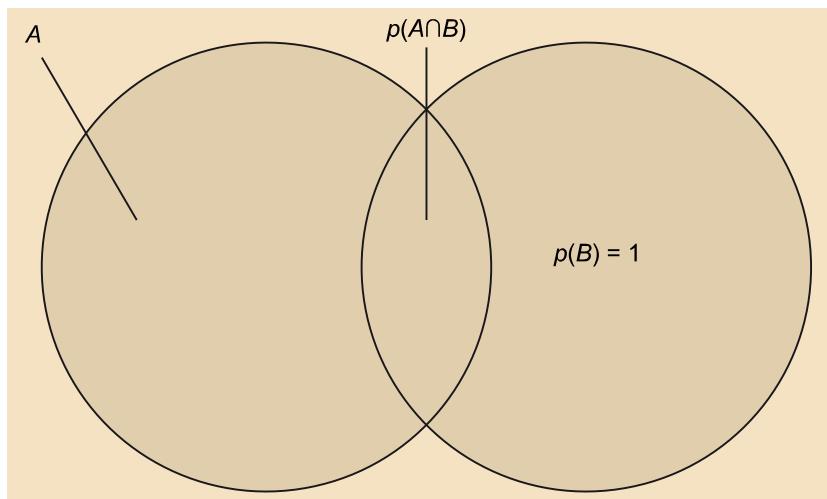
Figura 9. Espacio muestral Ω 

Figura 10. Espacio muestral reducido



En esta representación gráfica se observa claramente que el espacio muestral inicial (Ω) queda reducido solo en el espacio asociado al evento B , porque se sabe que este ha ocurrido.

3.3.2. Independencia

Dados los eventos A y B , se dice que son independientes cuando, sabiendo que se ha producido uno, no se ve alterada la probabilidad inicial del otro, es decir:

$$p(A/B) = p(A) \text{ o } p(B/A) = p(B)$$

Teorema de la probabilidad total

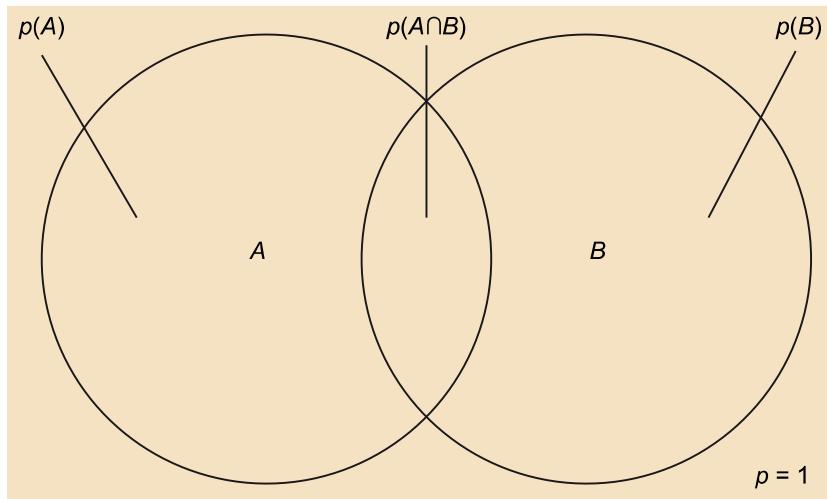
La probabilidad de la unión de dos eventos A y B es igual a la probabilidad del evento A más la probabilidad del evento B menos la probabilidad de la intersección de los dos eventos, es decir:

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

Si los eventos A y B son independientes o mutuamente excluyentes, la probabilidad de la unión de los dos eventos es igual a la suma de la probabilidad asociada al evento A y la probabilidad asociada al evento B , es decir:

$$p(A \cup B) = p(A) + p(B)$$

Figura 11



Teorema del producto

La probabilidad de la intersección de dos eventos independientes es igual al producto de las probabilidades asociadas a los dos eventos, es decir:

$$p(A \cap B) = p(A) \cdot p(B)$$

La probabilidad de la intersección de dos eventos no independientes es igual al producto de la probabilidad asociada al evento A por la probabilidad condicionada del evento B , si se ha dado el evento A o, lo que es lo mismo, la probabilidad asociada al evento B por la probabilidad condicionada del evento A , si se ha dado el evento B :

$$p(A \cap B) = p(A) \cdot p(B/A)$$

$$p(A \cap B) = p(B) \cdot p(A/B)$$

3.4. Variables aleatorias

Una variable aleatoria X es una función definida sobre un espacio muestral Ω , de modo que a cada suceso elemental de Ω le hace corresponder un número real (R). Este espacio muestral queda definido por un modelo o ley de probabilidad, que se establece a partir de la asociación de cada uno de los valores de la variable aleatoria con su probabilidad correspondiente.

Se han de diferenciar dos tipos básicos de variables aleatorias: las variables **discretas** y las variables **continuas**.

3.4.1. Variables aleatorias discretas

Una variable aleatoria se considera discreta cuando el rango de la variable es finito. Es decir, cuando se asignan probabilidades a cada valor concreto de la variable X . Una variable aleatoria discreta queda definida por **dos funciones**: la de **probabilidad** y la de **distribución**.

La **función de probabilidad** asigna a cada valor de la variable discreta su probabilidad. La probabilidad asociada a cada valor siempre estará entre 0 y 1. Además, la suma de todas las probabilidades siempre es 1.

Ejemplo

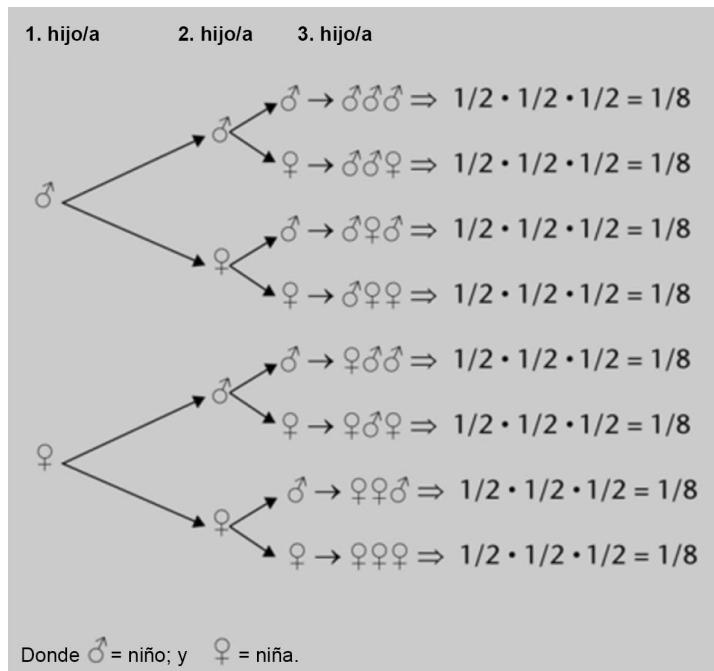
Si una familia tiene tres hijos y la probabilidad de ser niño o niña es de 0,5, la función de probabilidad de la variable «número de niños (sexo masculino)» queda reflejada en la tabla 6.

Tabla 6

Número de niños (sexo masculino)			
0	1	2	3
$1/8 = 0,125$	$3/8 = 0,375$	$3/8 = 0,375$	$1/8 = 0,125$

A partir del diagrama de árbol se ve claramente cómo se obtienen estas probabilidades (figura 12).

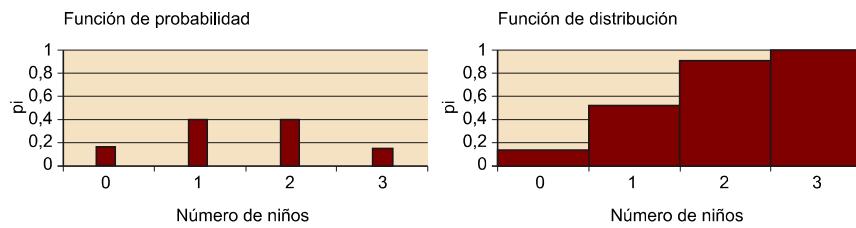
Figura 12



En este gráfico se observa que la probabilidad de que de los tres hijos que tiene la familia los tres sean niño es de $1/8$; que dos sean niño y uno niña es de $3/8$ ($1/8 + 1/8 + 1/8$); que uno sea niño y dos sean niña es de $3/8$ ($1/8 + 1/8 + 1/8$); y que los tres sean niña es de $1/8$.

La **función de distribución** asigna a cada valor de la variable discreta la probabilidad de obtener un valor inferior o igual a aquel valor concreto. Estas dos funciones se representan en la figura 13.

Figura 13

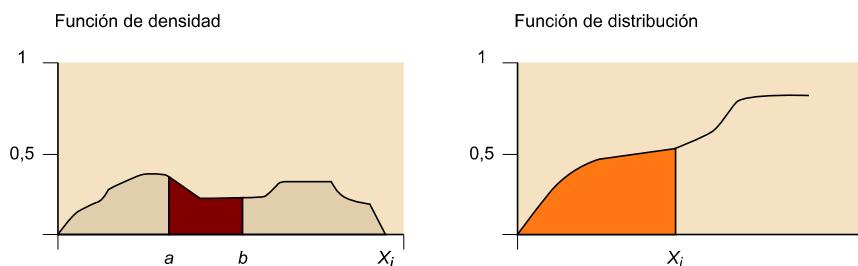


3.4.2. Variables aleatorias continuas

Una variable aleatoria es continua cuando entre dos valores de la variable hay un número infinito de valores posibles, es decir, el conjunto imagen es un conjunto continuo de números, como en un intervalo. Por ejemplo, el peso es una variable aleatoria continua porque entre dos valores cualesquiera (50-51 kg) hay un número infinito de valores. Las dos funciones que definen una variable aleatoria continua son la **función de densidad** y la **función de distribución**.

La **función de densidad** asigna una probabilidad determinada a un rango o intervalo de valores de la variable aleatoria continua $[a, b]$. La **función de distribución** asigna la probabilidad de encontrar un valor igual o inferior a x_i ; por lo tanto, conceptualmente es lo mismo que la función de distribución para variables aleatorias discretas. Estas dos funciones están representadas en la figura 14:

Figura 14



3.4.3. Esperanza matemática

El concepto de **esperanza matemática** (o **esperanza o media poblacional**) es equivalente al concepto de media en estadística descriptiva, pero aplicado a las variables aleatorias. La esperanza matemática es el valor medio teórico de todos los valores que puede tomar la variable aleatoria. La media de los datos obtenidos con un experimento aleatorio tenderá más al valor de la esperanza matemática cuanto más veces repitamos el experimento.

3.4.4. Varianza de una variable aleatoria

El concepto de **varianza de una variable aleatoria** es equivalente al concepto de varianza en estadística descriptiva, pero aplicado a las variables aleatorias. La varianza de una variable aleatoria mide la dispersión media de los valores de una variable aleatoria respecto a su esperanza matemática. La varianza de los datos obtenidos con un experimento aleatorio tenderá más al valor de la varianza de la variable aleatoria cuantas más veces repitamos el experimento. El valor positivo de la raíz cuadrada de la varianza es la desviación típica de la distribución de la variable.

3.5. Modelos de probabilidad

Un modelo o ley de probabilidad es la correspondencia que se establece entre cada valor de la variable aleatoria y las probabilidades correspondientes. En consecuencia, cada modelo de probabilidad queda definido por una función determinada (de probabilidad, o de densidad y de distribución). En algunos casos, los datos se ajustan a un modelo de probabilidad perfectamente conocido. Por ello en este apartado se pretenden explicar algunos de los principales modelos teóricos de probabilidad: la ley binomial, la ley de Poisson y la ley

normal o ley de Gauss-Laplace. Las dos primeras se utilizan cuando se trabaja con variables aleatorias discretas, y la última, cuando se trabaja con variables aleatorias continuas.

3.5.1. Modelos de probabilidad para variables aleatorias discretas

Distribución binomial

El modelo de la ley binomial se puede aplicar en el caso de trabajar con variables cuantitativas discretas generadas a partir de una variable cualitativa dicotómica. La variable dicotómica recibe el nombre de **experimento de Bernoulli** (Jacob Bernoulli, 1654-1705). Uno de los valores posibles recibe el nombre de **éxito** (E) y el otro recibe el nombre de **fracaso** (F).

Cada uno de los dos valores tiene una determinada probabilidad:

$$p(E) = \pi$$

$$p(F) = (1 - \pi)$$

Si se repite el experimento de Bernoulli un número de veces n y los experimentos son independientes, es decir, el valor de π no se altera en cada repetición, se genera una variable aleatoria discreta con $n + 1$ valores posibles.

Por ejemplo, para el caso del lanzamiento de una moneda, imaginemos que el caso de cara sea el «éxito», con una probabilidad de 0,5. Se puede lanzar la moneda tantas veces como se quiera (n). De lanzamiento a lanzamiento, el resultado es independiente, es decir, el hecho de que haya salido una cara en el primer lanzamiento no influye nada en cuanto al resultado cara o cruz de los lanzamientos sucesivos. Al final del proceso se ha generado la variable aleatoria discreta: número de caras obtenidas en n lanzamientos. Esta variable se distribuye según la ley binomial con $n + 1$ valores (0, 1, 2... n).

Se puede obtener la probabilidad de la distribución binomial con la función «BINOMDIST» de la aplicación Google Sheets.

Ejemplo

La probabilidad de sufrir ansiedad ante un examen entre la población universitaria es de 0,7, y se dispone de una muestra de cuatro personas.

- ¿Cuál es la probabilidad de que solo una de las cuatro personas de la muestra sufra ansiedad? Utilizando la función «BINOMDIST» con los parámetros conocidos (número_resultados_correctos: 1; número_pruebas: 4; probabilidad_resultados_correctos: 0,7; y acumulativa: 0) obtenemos la probabilidad pedida: 0,0756 (un 7,56%).
- ¿Cuál es la probabilidad de que haya una o menos personas de la muestra que sufra ansiedad? Utilizando la función «DISTR.BINOM» con

los parámetros conocidos (número_resultados_correctos: 1; número_pruebas: 4; probabilidad_resultados_correctos: 0,7; y acumulativa: 1) obtenemos la probabilidad pedida: 0,0837 (un 8,37%).

Distribución de Poisson

La distribución de Poisson se formuló como una particularidad de la distribución binomial por el hecho de que n tiende a ser muy grande ($n \rightarrow \infty$) y π a ser muy pequeño ($\pi \rightarrow 0$).

Se puede obtener la probabilidad de la distribución de Poisson con la función «POISSON» de la aplicación Google Sheets.

Ejemplo

La probabilidad de que un bebé nazca con labio leporino es de 0,0002 neonatos.

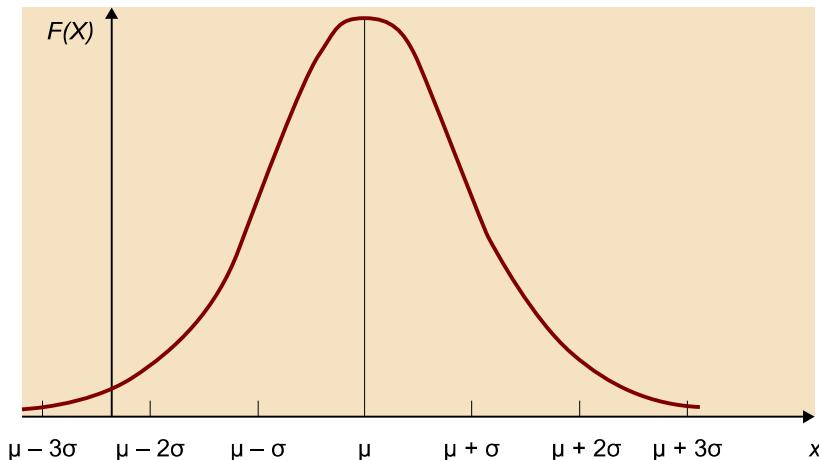
- ¿Cuál es la probabilidad de que en una población de seis mil neonatos nazcan seis bebés con esta malformación? Utilizando la función «POISSON» con los parámetros conocidos [x: 6; media (o esperanza matemática): $E(X) = 0,0002 \times 6.000 = 1,2$; y acumulativa: 0] obtenemos la probabilidad pedida: 0,0012 (un 0,12%).
- ¿Cuál es la probabilidad de que en una población de seis mil neonatos nazcan seis bebés, o menos, con esta malformación? Utilizando la función «POISSON» con los parámetros conocidos [x: 6; media (o esperanza matemática): $E(X) = 0,0002 \times 6.000 = 1,2$; y acumulativa: 1] obtenemos la probabilidad pedida: 0,999 (un 99,9%).

3.5.2. Modelos de probabilidad para variables aleatorias continuas

Distribución normal o de Gauss-Laplace

La distribución normal es un modelo teórico de probabilidad al que se ajustan determinadas variables cuantitativas continuas. Este modelo también recibe el nombre de **campana de Gauss**, en honor a Carl Friedrich Gauss, por la forma que presenta y que se muestra en la figura 15.

Figura 15. Campana de Gauss



Las características que definen la distribución normal son las siguientes:

- Sus parámetros son la media (μ) y la varianza (σ^2).
- La variable X sigue una distribución normal con media μ y varianza σ^2 .
- $X \sim N(\mu, \sigma^2)$.
- Tiene un punto de máxima altura que coincide con la media (μ), la mediana (Md) y la moda (Mo).
- Es simétrica respecto al eje de ordenadas. El eje de simetría está situado en μ .
- Es asintótica respecto al eje de abscisas, por lo tanto fluctúa entre $-\infty$ y $+\infty$.
- Tiene dos puntos de inflexión situados a $\pm 1\sigma$ de μ . Entre estos dos puntos de inflexión se encuentra el 68,26% del total de área bajo la curva normal.

Se puede obtener el valor de la función de distribución normal (o de la función de distribución acumulativa normal) para un valor, una media y una desviación estándar especificados con la función «NORMDIST» de la aplicación Google Sheets.

Se puede obtener el valor de la función de distribución normal inversa para un valor, una media y una desviación estándar especificados con la función «NORMINV» de la aplicación Google Sheets.

Ejemplo

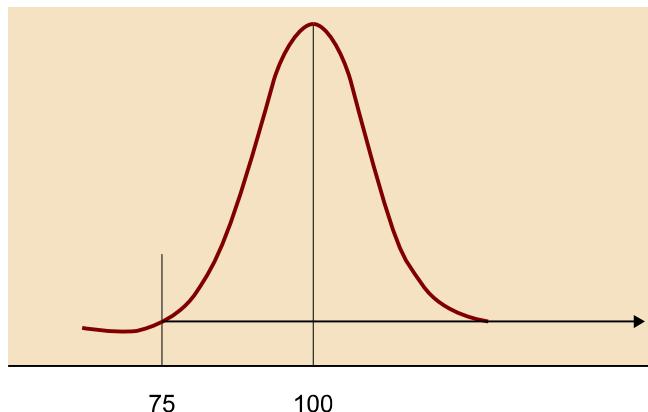
El cociente intelectual (CI) se distribuye según la ley normal con media 100 y desviación tipo 15.

¿Cuál es la probabilidad de que una persona presente un CI superior a 75?

Utilizando la función «NORMDIST» con los parámetros conocidos ($x: 75$; media: 100; desviación_estándar: 15; y acumulativa: 1) obtenemos que la probabilidad de que una persona presente un CI menor o igual a 75 es de 0,0477. La probabilidad de que una

persona presente un CI superior a 75 es 1 menos la probabilidad de que presente un CI menor o igual a 75, es decir, 0,9522 (95,22%).

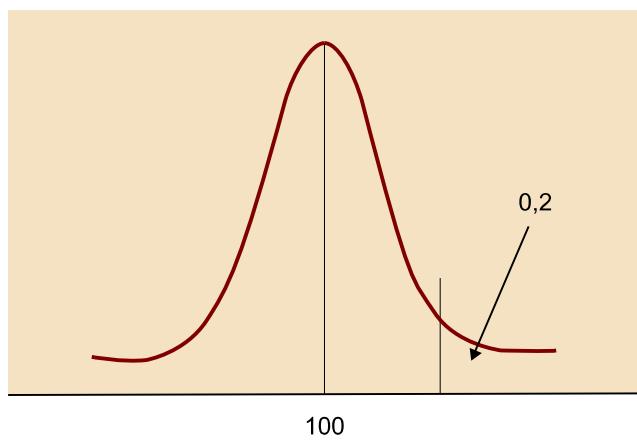
Figura 16



¿Cuál es el CI que delimita el 20% de los sujetos más inteligentes?

Utilizando la función «NORMINV» con los parámetros conocidos (probabilidad: 0,8; media: 100; y desviación_estándar: 15) obtenemos que el CI que deja por encima de este al 20% de los sujetos más inteligentes es de 112,6.

Figura 17



4. Estadística inferencial

Mientras que la estadística descriptiva resume y dibuja la información oculta en una matriz de datos para ayudarnos a entenderla, la **estadística inferencial** (o estadística inductiva) pretende aportar conclusiones generales aplicables a la población de estudio a partir del análisis de diferentes muestras de esta población.

Así pues, el principal objetivo de la estadística inferencial es estudiar las características numéricas de una población o verificar afirmaciones sobre estas características calculándolas en una o varias muestras elegidas al azar. El proceso utilizado en este tipo de estudios nos permite inferir o pronosticar el valor de los parámetros poblacionales (μ , σ , π , etc.) a partir del valor de los estadísticos normales (\bar{x} , s_x , p , etc.).

Por ello deberemos distinguir entre los conceptos de **población** y **muestra**. La **población** es el conjunto de individuos, de elementos o de cosas con alguna característica común del que se lleva a cabo un estudio estadístico. No obstante, cuando no es posible observar todos los elementos de una población hay que seleccionar una **muestra** que ponga de manifiesto las características estudiadas y que permita inferir datos de la población. El **muestreo** es el procedimiento seguido para la extracción de la muestra. Un **muestreo aleatorio**, es decir, con la elección de los elementos de la muestra al azar, favorece una mejor representatividad de la muestra y evita un posible sesgo.

Entre las diferentes maneras de trabajar la inferencia estadística que existe, destacamos la **estimación de parámetros** y el **contraste de hipótesis**.

4.1. Estimación de parámetros

4.1.1. Distribución muestral de un estadístico

La **distribución muestral de un estadístico** (media aritmética, varianza, proporción, etc.) es la distribución del estadístico, calculada en muestras infinitas de la misma medida n , elegidas al azar, de una determinada población. Así pues, si de la población de estudiantes de la UOC eligiéramos muestras aleatorias de la misma medida (por ejemplo, 30), y de cada muestra calculáramos la media de edad de los sujetos, obtendríamos una distribución de medias de edad que denominamos **distribución muestral de la media**.

La distribución muestral puede obtenerse para cualquier otro estadístico de los estudiados anteriores. Así, también podemos hablar de distribución muestral de la varianza, distribución muestral de la proporción, distribución muestral de la mediana, etc. Todas ellas se obtendrían calculando el valor del estadístico correspondiente a cada una de las infinitas muestras.

Dado que el muestreo es aleatorio, el valor del estadístico calculado en cada muestra también variará aleatoriamente de una a otra y, en consecuencia, podemos considerar la distribución muestral de este estadístico como la distribución de una variable aleatoria que puede ajustarse a uno de los modelos de distribución de probabilidad estudiados en el apartado anterior.

Distribución muestral de la media aritmética

El estadístico más ampliamente utilizado, como representativo de un conjunto de datos, es la media aritmética. La distribución muestral de la media tiene, a su vez, su media aritmética (denominada **media de la distribución muestral de la media**, y representada por $\mu_{\bar{x}}$), y su desviación estándar (denominada **desviación estándar de la distribución muestral de la media, error típico o error estándar de la media**, y representada por $\sigma_{\bar{x}}$).

Lógicamente, en poblaciones muy amplias o infinitas, el número de muestras diferentes posibles es también prácticamente infinito (o infinito realmente). La media de la distribución muestral de medias tiende hacia la media de la población (o coincide en poblaciones finitas), y la desviación estándar de la distribución (el error estándar de la media) disminuye a medida que aumenta la medida muestral.

El **teorema central del límite** nos dice que, aunque la distribución de una variable no sea normal, la distribución muestral de la media basada en muestras de medida n será aproximadamente normal. Este teorema es más cierto cuanto mayores son las medidas muestrales, así que para n «pequeños» (por ejemplo, menos de 10), la distribución muestral de la media solo es aproximadamente normal, mientras que para n «grandes» (por ejemplo, de 30), la distribución es prácticamente normal.

Distribución muestral de una proporción

Cuando trabajamos con una variable categórica, no tenemos valores numéricos para cada observación, sino la presencia o la ausencia de un atributo determinado. Así pues, para la variable «sexo» de los sujetos, lo que tenemos para cada sujeto es si es hombre o mujer, igual que para la variable «estado civil» tendremos si está casado o no. Para estas variables dicotómicas, el estadístico más representativo es la proporción (P), que también será una característica de la población de referencia. En este contexto hablaremos de la **proporción poblacional** como un parámetro que se representa por π .

Si elegimos al azar diferentes observaciones de una variable categórica y asignamos a uno de sus atributos el valor 1 (habitualmente, el que es centro de nuestro interés), y al otro atributo el valor 0, **su distribución de probabilidad se ajustará a una distribución binomial**. Al igual que con la distribución muestral de la media aritmética, también es posible calcular la **media** y la **desviación estándar** de la **distribución muestral de la proporción**.

4.1.2. Intervalos de confianza para la estimación de parámetros

Una de las aplicaciones más inmediatas es la **estimación del valor de un parámetro poblacional** a partir de la obtención de una única muestra, elegida aleatoriamente, de la mencionada población.

Cuando elegimos una muestra aleatoria de observaciones y utilizamos la media o la proporción de la muestra para estimar el valor poblacional, sabemos que si la muestra hubiera sido más amplia, la variabilidad sería más pequeña y, por lo tanto, la estimación sería más precisa. Pero ¿cómo podemos medir la precisión de nuestras estimaciones?

La manera de hacerlo es no dando una única estimación del valor poblacional, sino un intervalo, y después reforzar este intervalo de valores por medio de una declaración de nuestro nivel de confianza de que el verdadero valor esté dentro de este intervalo. Esto es lo que se denomina **intervalo de confianza**.

El intervalo de confianza es un rango entre dos valores alrededor de un parámetro muestral entre los cuales, con una probabilidad determinada (o **nivel de confianza**), se situará aquel parámetro en la población. El **nivel de confianza** ($1 - \alpha$) se presenta habitualmente como porcentaje, al multiplicar el valor de $(1 - \alpha)$ por 100, donde α es el **nivel de error** (o **nivel de significación**).

Hay un intercambio entre la precisión que se puede expresar en un intervalo de confianza y el nivel de confianza (ved tabla 7). Cuanto más bajo sea el nivel de confianza, más pequeño será el intervalo de confianza. Por lo tanto, el resultado será más preciso, pero la probabilidad de que el intervalo no incluya el verdadero valor del parámetro será más elevada.

Tabla 7

Intervalo de confianza	Nivel de confianza ($1 - \alpha$)	Nivel de error o de significación (α)	Precisión
↓	↓	↑	↑
↑	↑	↓	↓

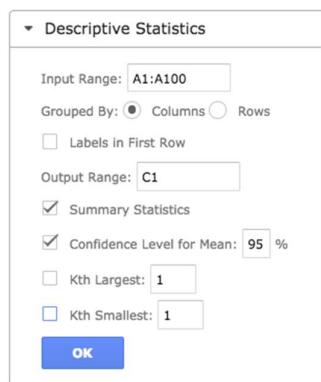
La única manera de mejorar tanto la precisión como el nivel de confianza es reduciendo el error típico. Si la desviación estándar poblacional es fija, solo podremos reducir el error típico aumentando el tamaño de la muestra. Alternativamente, si se mantiene el margen de error fijo, incrementar el tamaño de la muestra supone un incremento del nivel de confianza.

Intervalo de confianza para la media aritmética

Se puede obtener un intervalo de confianza de la media aritmética con el complemento «XLMiner Analysis ToolPak» de la aplicación Google Sheets entrando al menú del cuadro de diálogo «Descriptive Statistics».

Cuando se inicia el complemento, se abre un cuadro de diálogo en el que hay que seleccionar la opción «Descriptive Statistics» (ved figura 18).

Figura 18



- «Input range»: indicar las casillas donde se ubican los valores en la matriz de datos.
- «Labels in First Row»: activar si tenemos etiquetada la variable en la primera fila.
- «Output Range»: indicar la casilla a partir de la cual queremos que se nos muestren los resultados.
- «Summary Statistics»: activar para obtener los resultados descriptivos.
- «Confidence Level for Mean»: activar para obtener el intervalo de confianza. Cuando se activa esta opción, se establece, por defecto, un nivel de confianza del 95%, pero puede sustituirse este valor por cualquier otro.

Una vez ejecutada esta opción, haciendo clic en «OK» se obtienen los siguientes resultados (el ejemplo corresponde a los datos obtenidos pasando el MAS² a cien sujetos de un municipio):

(²)Test de ansiedad manifiesta de Taylor (Taylor, 1958).

Tabla 8. Análisis resultados (MAS)

Mean	22,060
Standard Error	1,288
Median	21,000
Mode	31,000
Standard Deviation	12,878
Sample Variance	165,855
Kurtosis	-0,855
Skewness	0,135
Range	48,000
Minimum	0
Maximum	48,000
Sum	2.206,000
Count	100
Confidence Level (95%)	2,555

La mayoría de la información que nos proporciona este análisis corresponde a la descripción de la variable. Para nuestro propósito en este apartado, la información que nos interesa es la que se halla en las casillas de **media** («Mean») y **nivel de confianza** («Confidence Level»).

Ya conocemos ampliamente la interpretación de la media, pero desconocíamos hasta ahora la interpretación del nivel de confianza. El valor que nos proporciona es el del **margen de error del intervalo de confianza para el nivel de confianza elegido**. Así pues, en nuestro caso, para el nivel de confianza del 95% este margen de error es de 2,555.

Con estos resultados podemos concluir que, con un nivel de confianza del 95%, el grado de ansiedad medio de todos los habitantes del municipio estará entre 19,505 y 24,615 puntos de la escala de MAS.

Intervalo de confianza para la proporción

Ya hemos visto anteriormente que para variables dicotómicas se puede considerar la proporción de una de sus dos modalidades como la media del conjunto de valores previamente codificados como 0 y 1, asignando el 1 a la modalidad cuya proporción queremos estudiar.

Se puede obtener un intervalo de confianza de una proporción con el complemento «XLMiner Analysis ToolPak» de la aplicación Google Sheets entrando al menú del cuadro de diálogo «Descriptive Statistics».

El ejemplo que presentamos corresponde al estudio de la proporción de hombres en un municipio. Por lo tanto, lo primero que haremos en nuestra matriz de datos será aislar la variable «sexo» y recodificar sus valores, asignando un 1 a los hombres y un 0 a las mujeres. Cuando iniciemos el complemento «XLMiner Analysis ToolPak» de la aplicación Google Sheets, se abrirá el mismo cuadro de diálogo que en el ejemplo anterior. Una vez ejecutada esta opción, haciendo clic en «OK» se obtienen los siguientes resultados:

Tabla 9. Análisis de resultados (sexo masculino)

Mean	0,400
Standard Error	0,049
Median	0,000
Mode	0,000
Standard Deviation	0,492
Sample Variance	0,242
Kurtosis	-1,866
Skewness	0,414
Range	1,000
Minimum	0,000
Maximum	1,000
Sum	40,000
Count	100,000
Confidence Level (95%)	0,098

La mayor parte de la información que nos proporciona este análisis no es pertinente porque estamos analizando una variable categórica como es el sexo, pero para nuestro objetivo sí que tiene la información necesaria. Así, la **media de la distribución** (0,40) es la **proporción** (P) de hombres de la muestra. También coincide el error típico de la proporción con el error típico de la muestra.

Deberemos **recalcular el margen de error** multiplicando el valor del error típico por el de la puntuación z correspondiente al nivel de confianza establecido. En nuestro ejemplo, el margen de error exacto será $0,049 \times 1,96 = 0,096$.

Con este resultado podemos concluir que, con un nivel de confianza del 95%, la proporción de hombres del municipio estudiado está entre 0,304 y 0,496, es decir, un porcentaje de hombres entre el 30,4% y el 49,6%.

4.2. Introducción al contraste de hipótesis

El **contraste de hipótesis**, también denominado **prueba de significación** o **prueba estadística**, es un procedimiento que nos permite decidir si una afirmación sobre cierta característica o características de la población puede ser mantenida o debe ser rechazada, de acuerdo con los datos obtenidos en una muestra de la población o en varias muestras.

Con esta breve introducción al contraste de hipótesis se establecen las bases para un amplio conjunto de pruebas estadísticas que están fuera del alcance de esta asignatura; todas, sin embargo, parten de los mismos postulados y siguen un mismo esquema para su resolución.

Una de las aplicaciones más habituales de los contrastes de hipótesis es cuando se quiere comprobar el efecto de una determinada intervención o tratamiento. Por ejemplo, podríamos plantear un estudio sobre cómo ha influido la nueva ley del tabaco en el consumo de cigarrillos en el municipio. Podríamos comparar la proporción de fumadores antes y después de la promulgación de la mencionada ley. La afirmación que pondríamos a prueba sería la siguiente: la nueva ley del tabaco ha disminuido la proporción de fumadores. El contraste de hipótesis nos permite tomar una decisión sobre si aceptamos o rechazamos la afirmación anterior (que denominaremos **hipótesis**), de acuerdo con los datos que hemos obtenido de una muestra de 100 sujetos.

4.2.1. Contraste de hipótesis: tomar decisiones

En el apartado anterior hemos visto las distribuciones muestrales de la media y cómo estas distribuciones nos permiten definir un intervalo en el cual confiamos que estará la media de la población.

Un ejemplo que permite ilustrar los conceptos implicados en la toma de decisiones estadísticas podría ser el de las pruebas de un laboratorio médico con las que intentamos detectar el virus del sida. Imaginemos que se envía una muestra de sangre a un laboratorio para que haga la prueba de anticuerpos VIH. Tenemos dos posibilidades que nos interesan: que los anticuerpos estén presentes en la sangre o que no lo estén y, en realidad, solo una de las posibilidades es cierta.

En la tabla 10 representamos estas dos posibilidades para la situación real en forma de dos filas de la tabla. Cuando se aplica la prueba de laboratorio determinada a la muestra de sangre, se llega a una cierta conclusión: la prueba es positiva (el virus está presente en la sangre) o negativa (el virus está ausente). Ambas posibilidades se representan en las dos columnas de la tabla 10.

Tabla 10

Verdad	Prueba	
	Negativa	Positiva
Ausencia del virus	Correcta	Errónea
Presencia del virus	Errónea	Correcta

Las filas indican el verdadero estado de la muestra de sangre, mientras que las columnas indican la conclusión que el laboratorio ha sacado, que podría ser errónea por muchas razones, por ejemplo porque el procedimiento del laboratorio sea incorrecto o porque haya falta de detectabilidad del virus. La tabla muestra las cuatro posibilidades diferentes, que dependen de la conclusión a la que se ha llegado y de cuál es la verdad:

- 1.^a fila, 1.^a columna: la prueba es negativa y es cierto que no hay presencia del virus del sida; es una conclusión correcta.
- 1.^a fila, 2.^a columna: la prueba es positiva, pero también es verdad que no hay presencia del virus del sida; es una conclusión falsa y los investigadores médicos a menudo hablan de un falso positivo.
- 2.^a fila, 1.^a columna: la prueba es negativa, pero es cierto que existe el virus y que la prueba ha fracasado a la hora de detectarlo; esta es una conclusión falsa y se denomina falso negativo.
- 2.^a fila, 2.^a columna: la prueba es positiva y verdaderamente hay presencia del virus; esta es una conclusión correcta.

En el contraste de hipótesis tenemos la misma situación. Nosotros consideramos dos posibles estados de la población, que denominamos **hipótesis**. A partir de los datos de una muestra, hay que decidir cuál de las hipótesis es la correcta. Nuestra decisión también puede ser correcta de dos maneras cuando decidimos a favor de la hipótesis que es verdaderamente correcta, y puede ser equivocada de dos maneras cuando decidimos a favor de la hipótesis falsa.

4.2.2. Hipótesis nula y alternativa

La **hipótesis nula**, representada por H_0 , es la expresión formal que se pone a prueba en un contraste de hipótesis. Indica la «no diferencia», o el «sin efecto», y es la que suponemos a la hora de valorar si el resultado se debe al azar.

H_0 expresa, por ejemplo, que un parámetro de la población, como puede ser la media, toma un valor específico, o que esta media es igual en dos grupos diferentes de sujetos.

La **hipótesis alternativa**, representada por H_1 , es la expresión del efecto, el cambio o la diferencia que puede existir en los datos estudiados (y que muchas veces, aunque no en todas, es la que esperamos o sospechamos). La hipótesis alternativa dice, por ejemplo, que un parámetro de la población, como por ejemplo la media, difiere de un valor especificado, o que el mismo parámetro obtenido en dos grupos diferentes difiere en su valor, en una dirección determinada (unilateral o de una cola) o en las dos direcciones (bilateral o de dos colas).

Fijaos en otra característica que diferencia la hipótesis nula de la alternativa:

- La hipótesis nula, habitualmente (pero no siempre), consiste en una igualdad simple entre parámetros o entre un parámetro y un valor fijo; en este caso, la igualdad a la media del grado de ansiedad.
- La hipótesis alternativa consiste, normalmente, en muchas posibilidades; en este caso, que la media de ansiedad de los habitantes del municipio sea diferente de la de la población general.

4.2.3. Uso de los intervalos de confianza para llevar a cabo un contraste de hipótesis

Siguiendo nuestro ejemplo anterior, podríamos analizar si el grado de ansiedad (medido con el MAS) de los habitantes del municipio es igual al de la población general (hipótesis nula), o es diferente (hipótesis alternativa).

Si definimos la media en ansiedad de la población general como $\mu = 25$, las expresiones formales de la hipótesis nula y la alternativa serían las siguientes:

- Hipótesis nula. $H_0: \mu = 25$
- Hipótesis alternativa. $H_1: \mu \neq 25$

Una vez más, podemos representar las diferentes conclusiones y las situaciones reales en una tabla:

Tabla 11

Conclusiones a partir de nuestro estudio		
Verdad	Igual grado medio de ansiedad	Diferente grado medio de ansiedad
Igual grado medio de ansiedad	Correcta	Errónea

	Conclusiones a partir de nuestro estudio	
Diferente grado medio de ansiedad	Errónea	Correcta

Hemos calculado un **intervalo de confianza** para la media en ansiedad de los sujetos del municipio a partir de la muestra de los 100 sujetos de la que disponemos. Este intervalo, con un grado de confianza del 95%, está entre 19,505 y 24,615. Por lo tanto, confiamos en un porcentaje del 95% en que la verdadera media de la población de referencia (los habitantes del municipio) está entre estos límites, y vemos que este intervalo no contiene el valor 25. Dado que queremos comprobar que la media de la población es 25, podemos decir con mucha seguridad que no es de 25. A partir de los cálculos realizados para establecer los intervalos de confianza, decimos que rechazamos la hipótesis nula, donde $\mu = 25$ y, por lo tanto, aceptamos la hipótesis alternativa, donde $\mu \neq 25$. Interpretamos estos datos, dentro del contexto de nuestro ejemplo general, de modo que el grado medio de ansiedad de los habitantes del municipio no es igual al grado medio de la población general.

4.2.4. Contraste de hipótesis y pruebas de significación

Un **estadístico** de contraste es un instrumento estadístico creado para tomar decisiones sobre la hipótesis nula con cierta probabilidad. Se caracteriza por tener una distribución muestral conocida (normal, t de Student, X^2 , etc.). Para cada tipo de contraste (de una media, de dos medias, de dos proporciones, etc.), tenemos su estadístico de contraste correspondiente.

Los pasos que hay que seguir son los siguientes:

- 1) Plantear la hipótesis nula y la alternativa.
- 2) Obtener, a partir de los datos muestrales, el estadístico de contraste correspondiente.
- 3) Obtener las regiones de aceptación y de rechazo de la hipótesis nula a partir del valor del estadístico de contraste teórico.

A partir de la distribución correspondiente del estadístico de contraste (normal, t de Student, X^2 , etc.), y especificando el nivel de confianza asumido (o más habitualmente, su complementario, que es el nivel de significación α), se obtienen los dos valores de este estadístico de contraste, que incluyen el porcentaje correspondiente en el nivel de confianza (95%). Estos dos valores se denominan **valores críticos (superior e inferior) del estadístico de contraste**, y la región comprendida entre estos dos valores se denomina **región**

de aceptación de la hipótesis nula. La **región de rechazo de la hipótesis nula** se sitúa por encima del valor crítico superior y por debajo del valor crítico inferior.

4) Tomar la decisión de aceptar o de rechazar la hipótesis nula.

Tomamos esta decisión comparando el estadístico de contraste, calculado con los datos muestrales, con los valores críticos de la distribución correspondiente. Así, si nuestro estadístico de contraste calculado queda entre estos valores críticos, aceptamos la hipótesis nula (dado que estamos en la región de aceptación de la mencionada hipótesis), mientras que si nuestro estadístico de contraste calculado es mayor que el valor crítico superior o menor que el inferior, rechazamos la hipótesis nula y aceptamos la alternativa.

5) Interpretamos el resultado en el contexto del estudio realizado.

Las dos pruebas anteriores son equivalentes y, lógicamente, nos llevan al mismo resultado.

El contraste de hipótesis, tal como lo hemos visto hasta ahora, puede tener una variante, que es la utilizada habitualmente en los paquetes estadísticos informatizados: la **prueba de significación**, que consiste en obtener directamente la probabilidad del estadístico de contraste muestral calculado. Esta probabilidad se denomina valor p . El **valor p** es la probabilidad de observar el resultado, o un resultado más extremo, cuando la hipótesis nula es cierta. Cuanto más pequeño es el valor p , más acentuada es la prueba contra la H_0 proporcionada por los datos. Los valores p por debajo del nivel de significación (habitualmente de 0,05) se denominan convencionalmente **significativos**.

El **nivel de significación** es la probabilidad máxima de cometer un error de tipo I (ved más adelante). Se establece en función del riesgo que se está dispuesto a asumir antes de reunir y analizar los datos. Si el valor p del contraste es menor que α , la hipótesis nula se rechaza y decimos que el resultado observado es estadísticamente significativo en el nivel de α .

Para tomar una decisión respecto a la hipótesis nula, simplemente comparamos este valor p con el nivel de significación (α). Si el valor p es superior a α , aceptamos la hipótesis nula, y si es más pequeño, la rechazamos y aceptamos la alternativa (entonces decimos que el resultado es estadísticamente significativo).

Podemos esquematizar los pasos para el contraste de hipótesis o la prueba de significación en la tabla 12.

Tabla 12

Contraste de hipótesis	Prueba de significación
Determinar la hipótesis nula y la alternativa	
Calcular el valor del estadístico de contraste con los datos muestrales	
Determinar las regiones de aceptación y de rechazo de la hipótesis nula (con los valores críticos del estadístico de contraste)	Obtener la probabilidad (valor p) del estadístico de contraste
Tomar una decisión comparando el valor del estadístico de contraste muestral o el observado con los valores críticos de la distribución correspondiente	Tomar una decisión comparando el valor p del estadístico de contraste muestral o el observado con el nivel de significación asumido
Interpretar la decisión anterior en el contexto del estudio realizado	

4.2.5. Errores de tipo I y de tipo II

Tal como hemos visto en apartados anteriores, siempre que tomamos una decisión en un contraste de hipótesis o en una prueba de significación podemos haber acertado o podemos habernos equivocado, dado que siempre hay una probabilidad de que la hipótesis nula sea cierta, aunque la hayamos rechazado (esta probabilidad es el valor p), o que no sea cierta aunque la hayamos aceptado, de acuerdo con nuestros datos muestrales.

Podemos representar en una tabla la situación en el contraste de hipótesis estadísticas (tabla 13).

Tabla 13

		Situación cierta	
Decisiones basadas en los datos		H_0	H_1
H_0		Decisión correcta Probabilidad $1 - \alpha$	Decisión incorrecta Error de tipo II Probabilidad β
H_1		Decisión incorrecta Error de tipo I Probabilidad α	Decisión correcta Probabilidad $1 - \beta$

En esta tabla hemos presentado dos términos que se utilizan para las decisiones incorrectas que se pueden tomar en esta situación:

- **Error de tipo I:** es el que se comete al pronunciarnos a favor de la hipótesis alternativa (es decir, al rechazar la hipótesis nula) cuando de hecho la hipótesis cierta es la nula. La probabilidad de este error de tipo I es igual a α (nivel de significación), o al valor p .

- **Error de tipo II:** es el que se comete al pronunciarnos a favor de la hipótesis nula cuando la hipótesis alternativa es la cierta. La probabilidad de cometer un error de tipo II se representa por β e inicialmente es desconocida.

4.2.6. Potencia de un contraste de hipótesis o prueba de significación

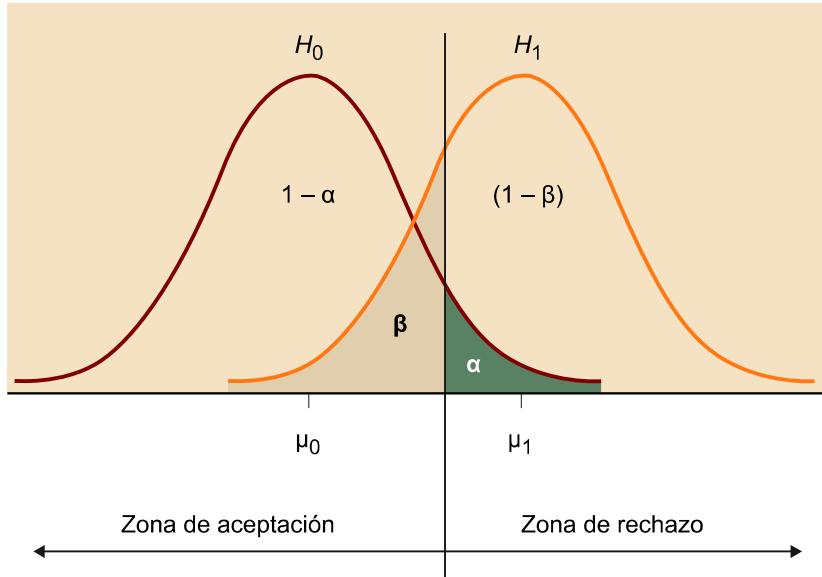
Como hemos visto en el apartado anterior, la probabilidad de cometer un error de tipo II se denomina β , y en principio el investigador la desconoce. El complementario de este valor β , es decir, $1 - \beta$, es la denominada **potencia de la prueba estadística**, y es la probabilidad de no equivocarnos cuando rechazamos una hipótesis nula y aceptamos, por lo tanto, la alternativa. Dicho de otro modo, es la seguridad que tenemos de no equivocarnos al aceptar una hipótesis alternativa (que bastantes veces representa la hipótesis de la efectividad de una intervención determinada porque expresa la diferencia entre dos o más grupos o muestras de datos).

Dado que el valor de β es desconocido inicialmente, también lo es el valor de la potencia de la prueba ($1 - \beta$), aunque sí que sabemos la relación que tiene con el grado de significación y con el tamaño de la muestra, por ejemplo.

La relación entre α y β , es decir, entre la probabilidad de cometer un error de tipo I y un error de tipo II, es otro de los intercambios característicos en estadística, porque esta relación es inversa. Así, si queremos disminuir la probabilidad de cometer un error de tipo I (disminuyendo α), estamos aumentando la probabilidad de cometer un error de tipo II (aumentando β), y disminuimos en consecuencia la potencia de la prueba estadística. Este intercambio entre α y β se puede apreciar mejor en la figura 19.

Como se puede observar en la figura 19, la hipótesis alternativa, cuando es cierta, también tiene su distribución de densidad (como la hipótesis nula), y estas dos distribuciones se encabalgan (en este caso por el lado derecho de la hipótesis nula, porque la prueba es unilateral por la cola derecha). Así, el valor crítico del estadístico de contraste es la línea vertical que divide la gráfica en dos. Por debajo (o a la izquierda) de este valor se encuentra la región (o zona) de aceptación de la hipótesis nula, y por encima (o por la derecha), la región de rechazo de la mencionada hipótesis nula.

Figura 19



Esto determina dos áreas rayadas: una verticalmente, que es la proporción de la distribución de la H_0 por encima del valor crítico del estadístico de contraste, y que corresponde al grado de significación α o al valor p ; y una horizontalmente, que es la proporción de la H_1 por debajo de este valor crítico y que denominamos β . Si disminuimos la zona rayada verticalmente (es decir, el grado de significación α), desplazamos la raya vertical hacia la derecha, y esto implica que aumenta la zona rayada horizontalmente, es decir β , con lo que disminuye, en consecuencia, la región $1 - \beta$, que denominamos potencia de la prueba.

La única manera de disminuir tanto la probabilidad de cometer un error de tipo I como de tipo II y aumentar la potencia de la prueba estadística es, una vez más, **aumentando los tamaños muestrales**. Así pues, aumentar el número de sujetos de las muestras es la única manera que tenemos de disminuir las probabilidades de cometer un error (sea del tipo I o II) en una prueba estadística de significación.

Bibliografía

Adielsson, M.; Barnes, R.; Kupfer, P.; Roberts, I.; Weber, J. H. (2005). «Google Sheets function list». <<https://support.google.com/docs/table/25273?hl=en>>

Cosculluela, A.; Fornieles, A.; Turbany, J. (2014). *Tècniques d'anàlisi de dades quantitatives*. Material docente de la UOC. Barcelona: Universitat Oberta de Catalunya.

