

Fastbook 05

Estadística Aplicada al Marketing

Introducción a la estadística



La estadística es la rama de las matemáticas que estudia cómo recopilar y resumir información para extraer conclusiones.

El origen de la estadística es anterior al año 3000 a. C. y se remonta a las primeras civilizaciones, si bien es cierto que la estadística en el sentido práctico de la palabra data del siglo XVIII. La palabra estadística tiene su origen en la palabra alemana *Statistik*, basado en el latín *statisticus* (ciencia del estado).

Con este fastbook vamos a introducirnos en el mundo de la estadística: conoceremos los conceptos de población y muestra, tipos de variables existentes, y las principales medidas de centralización, dispersión y posición. Como hemos ido viendo en los fastbooks anteriores, tan importante es tener y recopilar los datos como analizarlos mediante técnicas estadísticas y sacar conclusiones para poner en práctica y mejorar nuestro negocio.

Autora: Patricia Martín González

Población y muestra

Variables y tipos

Medidas de centralización

Medidas de dispersión

Medidas de posición

Resumen

Población y muestra

X Edix Educación

El concepto sobre el que se sustenta casi toda la estadística es el de población y muestra.

Población

La población es el conjunto (finito o infinito) de **elementos que son objeto del estudio**. Por ejemplo, las tiendas de Ikea, las emisiones de las campañas publicitarias en televisión o el conjunto de todos los números primos.

Individuo

Un individuo u observación hace referencia a **cada uno de los elementos de una población**. A pesar del nombre, hace referencia a cualquier elemento o cosa, no solamente a personas. Sobre los ejemplos anteriores, si estamos estudiando la distribución de las tiendas de Ikea en todo el mundo, cada una sería un individuo; si estamos estudiando la distribución de las emisiones publicitarias, cada una sería un individuo; o si estamos buscando la descomposición de un número, cada número primo será un individuo.

Tamaño poblacional

El **número total de individuos** que constituyen la población recibe el nombre de tamaño poblacional. Se suele representar por la **letra N**. Cuando la población es muy grande, se puede considerar población infinita, como, por ejemplo, el conjunto de todos los números primos.

Sobre nuestros ejemplos, el tamaño poblacional de las tiendas de Ikea a nivel mundial es 422, y el número de emisiones publicitarias que ha lanzado nuestra marca es 150.000.

La variabilidad del mundo real es el origen de la estadística.

Cuando la población es muy grande, la **observación o medición de todos los elementos multiplica la complejidad** en términos de trabajo, tiempo y coste. Aunque a veces es obligatorio estudiar a todos los individuos de la población (por ejemplo, las revisiones médicas), en general no es necesario. Para reducir el tamaño de estudio, se selecciona solo una parte de la población con el objetivo de tener una muestra representativa en su totalidad. Esto es lo que se conoce como **muestra estadística**.

Muestra

La muestra es una **representación significativa de una población**, es decir, es una parte de la población. Una muestra representativa contiene las características relevantes de la población en las mismas proporciones que están incluidas en tal población.

Sobre los ejemplos anteriores, para Ikea podrían estudiarse 85 tiendas tipo que representen la totalidad, y para las emisiones publicitarias podría reducirse a 1000.

Individuo

Un individuo de un espacio muestral tiene el mismo significado que de una población, pero siendo el espacio de estudio en este caso la muestra.

Sobre los ejemplos anteriores, cada una de las 85 tiendas será un individuo y cada una de las 1000 emisiones publicitarias será un individuo.

Tamaño muestral

El tamaño muestral es el número de individuos de la población del que se compone la muestra y se suele representar por la **letra n**.

Sobre los ejemplos anteriores, el tamaño muestral de las tiendas de Ikea sería 85, y el de las emisiones publicitarias 400.

Trabajar sobre una muestra en lugar de sobre la población total tiene mucha importancia cuando nos enfrentamos a un problema por primera vez, en el que sea necesario aplicar técnicas de machine learning y que el tiempo computacional para ajustar los algoritmos sea tan elevado que hace inviable el estudio. En estos casos, se coge una muestra de la población total con la que se ajusta y valida el modelo, y una vez tenemos un modelo robusto se aplica a toda la población para comprobar su validez.

Un ejemplo podría ser el **ajuste de un modelo de customer lifetime value**: si queremos conocer la predicción en diferentes momentos temporales futuros del valor de un cliente con diferentes históricos de partida y para diferentes situaciones, es muy probable que el coste computacional sea muy elevado y haya que empezar por una parte y luego extrapolar.

La elección de la muestra es muy importante para que los resultados que se extraigan de ella se puedan generalizar a toda la población. Debe haber los suficientes individuos para que no sea muy costosa, pero con la condición de que sea una muestra representativa.

En el ejemplo de las tiendas de Ikea, si solo seleccionamos tiendas de gran superficie, la muestra no sería representativa de toda la población ya que no se habría incluido el comportamiento de tiendas pequeñas.

¿Cuántos individuos son necesarios en la muestra para tener un conocimiento suficiente de la población?

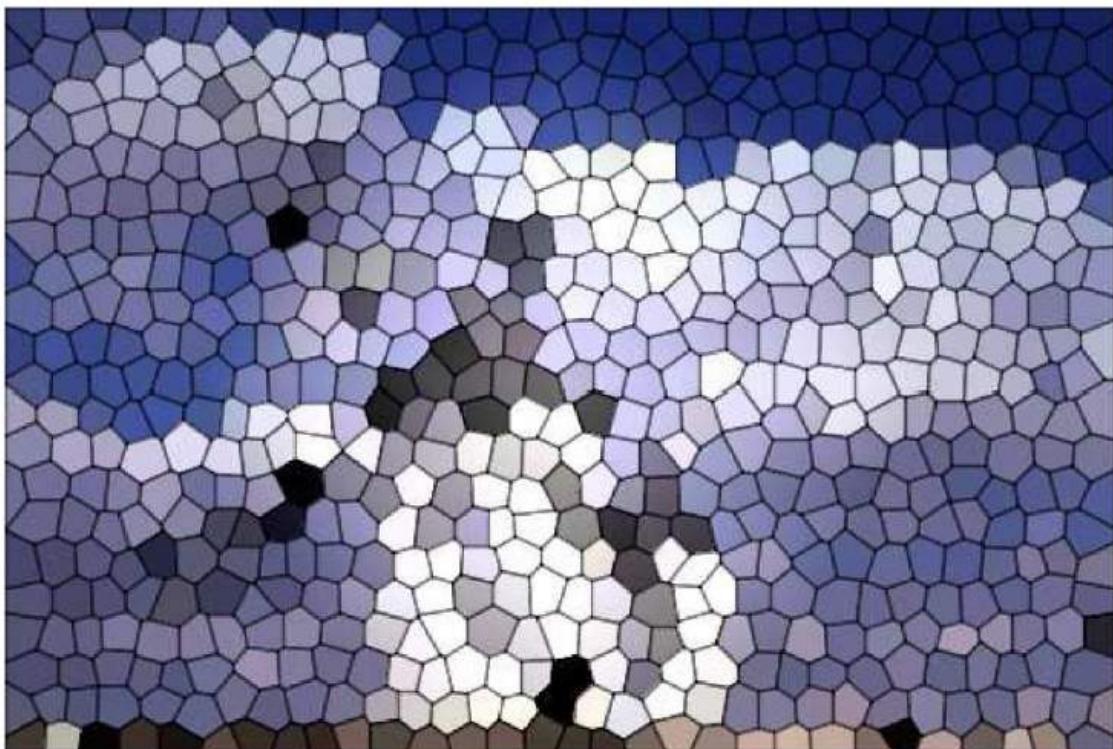
Depende de muchos factores.

Entre ellos, destacan la **variabilidad de la población, la fiabilidad deseada y el tamaño original de la población**. Existen fórmulas que permiten estimar el tamaño muestral esperado gracias al concepto de **error muestral**.

El error muestral es el error en el que se incurre con la disminución del número de individuos de la población a la muestra, generalmente $\leq 5\%$.

Veamos un ejemplo de la importancia de tomar una muestra suficiente:

Si se toma una **muestra pequeña de los píxeles** de una imagen será difícil averiguar el contenido.



Si se toma una **muestra mayor de los píxeles** de una imagen, es más fácil averiguar el contenido.



Sin necesidad de tener todos los píxeles para averiguar la imagen.



La técnica para escoger los individuos de la población para formar una o más muestras se llama muestreo. Podemos diferenciar dos tipos:

- **Muestreo probabilístico**

Son aquellos basados en el principio de **equiprobabilidad**, es decir, aquellos en los que los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra. Es el tipo de muestreo más usual.

Existen varias técnicas para ello: muestreo aleatorio simple, muestreo aleatorio sistemático, muestreo aleatorio estratificado, muestreo aleatorio por conglomerados... Pero la más extendida es **muestreo aleatorio simple**. Consiste en **elegir al azar los individuos** de la población que queremos estudiar en la que todos tienen la misma probabilidad de ser elegidos. Por ejemplo, para calcular el promedio de páginas de los libros de una biblioteca, elegiremos libros puramente al azar.

- **Muestreo no probabilístico**

Es una técnica de muestreo que a diferencia de la probabilística **no permite** que todos los individuos de la población tengan las **mismas oportunidades** de selección. En este tipo de muestreo suelen predominar aquellos individuos que al cumplir con cierta cualidad o característica benefician el estudio (en las próximas asignaturas conocerás los conceptos de *undersampling* y *oversampling*).

Entre sus técnicas se encuentra el **muestreo por conveniencia**, el **muestreo por cuotas** o el **muestreo accidental**. Por ejemplo: un sociólogo está estudiando la relación de los adultos con los adolescentes y crea su muestra mayoritariamente con gente que tiene hijos.

Variables y tipos

X Edix Educación

Una **variable** es una característica, calidad o propiedad objeto de estudio. Por ejemplo, el código postal de un cliente, el canal en el que se emitieron los anuncios publicitarios de TV, o la tasa de paro de cada uno de los países de la UE. Las variables las podemos dividir de múltiples formas, pero la forma más extendida es la división entre **cualitativa** y **cuantitativa**.

1

Variables cualitativas

Las **variables cualitativas** son aquellas que se refieren a características o cualidades que no pueden ser expresadas con valores numéricos. Podemos distinguir **tres tipos**:

Nominal

—

Son aquellas variables no numéricas que no siguen ningún orden específico y tienen un número finito de posibilidades.

Por ejemplo: la variable *medios publicitarios* tomaría los valores televisión, radio, redes sociales, emailing, OOH, etc.; o *estado civil*: soltero, casado, divorciado, separado o viudo. En base de datos suelen aparecer como **varchar o text**. En múltiples ocasiones se las suele atribuir un valor numérico identificativo, usualmente *numeric*.

Ordinal

Variables no numéricas que siguen un orden o jerarquía, como por ejemplo, el nivel socioeconómico (alto, medio, bajo) o el nivel de estudios (secundaria, bachillerato, estudios de formación profesional, universitario, máster o doctorado). Estas variables suelen ser almacenadas como **factor**.

Binaria

También conocidas como **dummies**, son variables que solamente pueden tomar dos valores, por ejemplo, verdadero o falso, 1 o 0 y sí o no. Son variables que, aunque no parezcan muy importantes, están muy presentes en bases de datos por la facilidad de almacenaje y la información que aportan.

Algunos ejemplos son: hijos sí-no, impactado por un anuncio publicitario verdadero-falso, desempleado 1-0 (formato de almacenamiento usual en base de datos donde 1 sería sí o verdadero, y 0 lo contrario). Estas variables suelen guardarse como tipo factor o *numeric* (en caso de ser 1-0 para facilitar operaciones aritméticas como la suma).

2

Variables cuantitativas

Las variables **cuantitativas** toman valores numéricos, por lo que se puede aplicar operaciones aritméticas. Entre ellas distinguimos **dos tipos: discretas y continuas**.

Las variables discretas

No pueden tomar valores intermedios entre dos valores consecutivos, es decir, son números enteros (sin decimales). Por ejemplo: número de hijos (0, 1, 2, 3...), número de visitas a la web, impresiones, etc. Suelen almacenarse como **integer**.

Las variables continuas

Pueden tomar valores intermedios entre dos valores tan próximos como deseemos, es decir, números reales.

En esta categoría podemos encontrar la tasa de paro, la puntuación promedio a una encuesta, ingresos, etc. Estas variables pueden almacenarse con varios nombres: ***double, real, numeric, etc.***

Ahora, responde a las siguientes preguntas.

1. Las variables cualitativas pueden ser...

- No numéricas.
- Numéricas.
- Enteras.

2. Las variables cuantitativas pueden ser...

- No numéricas.
- Numéricas.
- Reales.

3. La variable cualitativa nominal es aquella que:

- Sigue un orden específico.
- No sigue ningún orden.
- Siempre se responde con verdadero o falso.

4. La variable cualitativa denominada ordinal o cualitativa es aquella que:

- Es jerárquica.
- Es binaria.
- Está compuesta por números enteros.

5. La variable cualitativa binaria es aquella que:

- No sigue un orden.
- Solo se pueden tomar dos valores.
- Es real.

6. La variable cuantitativa discreta es aquella que:

- Puede tomar valores intermedios.
- Consta de valores decimales.
- Solo pueden tomarse números enteros.

7. La variable cuantitativa continua es aquella que:

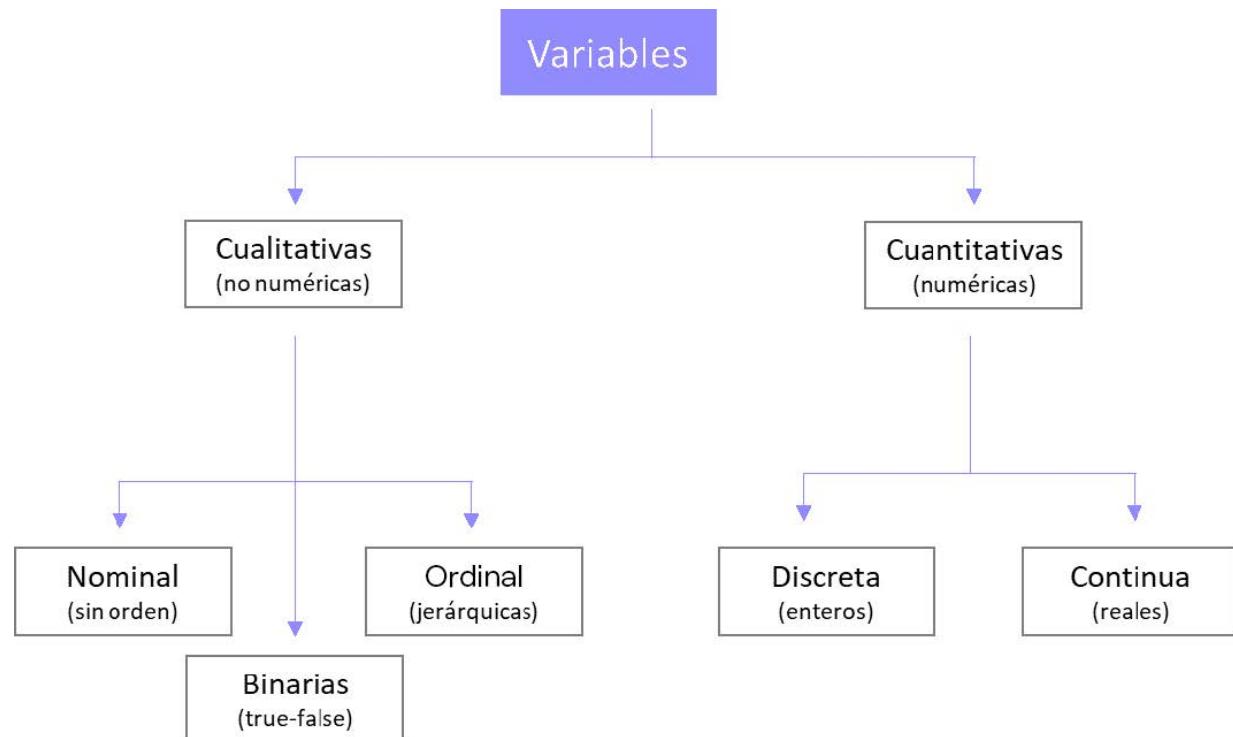
- Puede tomar valores intermedios.
- Consta de valores decimales.
- Solo pueden tomarse números enteros.



Respuestas: 1-A, 2-B, 3-B, 4-A, 5-B, 6-C, 7-A

Existe otro tipo de variables conocidas como variables de **texto libre**: recogen comentarios o textos escritos por los usuarios de forma natural (variables utilizadas mayoritariamente en técnicas de procesamiento del lenguaje natural). Estas variables suelen encontrarse en preguntas abiertas de encuestas, cuestionarios, etc. Por ejemplo: ¿Qué opina sobre nuestro producto? ¿Qué mejoraría de su experiencia de compra? En base de datos suelen almacenarse como *varchar* o *text*.

A continuación os dejo este esquema con las respuestas al test propuesto en este apartado.



Medidas de centralización

X Edix Educación

Una vez que ya conocemos los conceptos de población y muestra y los tipos de variables, estamos preparados para empezar a analizar los datos. Empezaremos con las **medidas de centralización** o de tendencia central, las cuales expresan el valor en torno al cual se sitúan los datos de una muestra. Nos centraremos en tres medidas: **media, mediana y moda**. Todas ellas se pueden **calcular para variables cuantitativas (numéricas)**.

No existe una medida mejor que otra, dependerá de la distribución de los datos y del objetivo. Muchas veces se calculan las tres para entender qué está pasando, ya que cada una aporta un conocimiento diferente. En función del comportamiento de la variable, pueden tomar el mismo valor o presentar grandes diferencias.

En este fastbook nos centraremos en la forma de **calcular dichas medidas**, pero es muy importante graficar siempre los datos para confirmar la completa compresión de la distribución de los datos. En el siguiente fastbook, veremos algunos gráficos de apoyo, aunque tendrás una asignatura en la que aprenderás a crear cada gráfico, así como su utilidad.

1

Media

La media aritmética (o promedio) es la medida de posición central más utilizada, más conocida y sencilla de calcular. Representa el punto promedio del conjunto de puntos, y se define como la suma de todos los valores observados dividido por el número total de observaciones. Se suele expresar con una barra encima, por ejemplo: \bar{X} o \bar{Y} , y su fórmula es:

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

Leyenda:

- n es el número de observaciones.
- x_i es el valor de cada observación.

Ventajas



Es fácilmente interpretable y tiene en cuenta todas las observaciones de la variable.

Limitaciones



Es independiente de la amplitud de los datos, por lo que no tiene en cuenta la dispersión (es decir, si están concentrados o separados). Esto provoca que para variables con distribuciones muy dispersas, la media se vuelva poco representativa.

Desventajas



Presenta mucha sensibilidad a los valores extremos demasiado grandes o pequeños (no tienen por qué llegar a ser outliers).

Veamos un ejemplo: supongamos que acabamos de terminar el primer año de una carrera universitaria y queremos calcular la nota promedio. Las calificaciones de cada asignatura han sido:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	id9	i
Calificación	6	9	8	0	5	4	7	10	6	7

La media será:

$$\text{Nota promedio} = \frac{\sum_{i=1}^n x_i}{n} = \frac{6+9+8+0+5+4+7+10+6+7}{10} = 6,2$$

Sabemos que el 0 de la asignatura id4 es porque no nos hemos presentado, por lo que queremos calcular la media solamente de las asignaturas a las que nos hemos presentado:

$$\text{Nota promedio} = \frac{6+9+8+5+4+7+10+6+7}{9} = 6,89$$

Como vemos, la **sensibilidad de la media a valores extremos es muy alta**, pues la media desechando el 0 ha subido de 6,2 a 6,89.

Media ponderada

Hasta ahora hemos hablado de **media aritmética** en la que todas las observaciones valen lo mismo, pero no siempre es lo que necesitamos.

La media ponderada es una medida de tendencia central que es apropiada cuando en un conjunto de datos cada uno de ellos tiene una importancia relativa (o peso) respecto de los demás: cada observación se multiplica por un valor, denominado peso, que refleja cuánto contribuye a la media.

$$\bar{X} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n}$$

Leyenda:

- n es el número de observaciones.
- w_i es el peso de cada observación.
- x_i el valor de cada observación.

Cuando todos los w_i tienen el mismo valor ($w_i=w$ para $i=1,2,\dots,n$), la media ponderada coincide con la media aritmética.

Diremos que los **pesos están normalizados cuando la suma es 1** (o 100 si hablamos de porcentaje). En general, se suelen normalizar y llevar a porcentaje para facilitar la comprensión.

Para cada peso original w_i , el nuevo peso normalizado w'_i será su peso original dividido entre la suma de todos ellos. **Matemáticamente:**

$$w'_i = \frac{w_i}{\sum_{k=1}^n w_k} = \frac{w_i}{w_1+w_2+\dots+w_k}$$

Con los pesos normalizados, calcular la media ponderada es más fácil, ya que la suma de los pesos es:

$$1 \left(\sum_{i=1}^n w_i = 1 \right)$$

Y, por tanto, quedaría:

$$\bar{X} = \sum_{i=1}^n w'_i x_i = w'_1 x_1 + w'_2 x_2 + \dots + w'_n x_n$$

Continuemos con el ejemplo anterior. Como todas las asignaturas no tienen los mismos créditos (esfuerzo requerido por los alumnos especificado para cada asignatura), necesitamos **calcular la media real, es decir, la media ponderada**. Los créditos de cada asignatura son:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	id9	i
Calificación	6	9	8	0	5	4	7	10	6	7
Créditos	3	6	9	3	9	6	9	6	3	6

Normalizamos los pesos para comprender la importancia de cada asignatura. Para la asignatura id1 el peso normalizado sería:

$$w'_1 = \frac{w_1}{\sum_{k=1}^n w_k} = \frac{3}{3+6+9+3+9+6+9+6+3+6} = \frac{3}{60} = 0,05 = 5\%$$

Si replicamos para el resto de pesos, obtenemos la siguiente tabla:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	
Calificación	6	9	8	0	5	4	7	10	
Créditos (peso normalizado)	5%	10%	15%	5%	15%	10%	15%	10%	

Calculando ahora la media ponderada tendríamos:

$$\text{Nota media ponderada} = \sum_{i=1}^n w_i' x_i = 0,05 * 6 + 9 * 0,1 + \dots + 7 * 0,1 = 6,6$$

A la vista de los resultados podemos comprobar que, aunque la media fuera de 6,2, la media real es de 6,6.

2

Mediana.

La mediana representa el valor que se encuentra exactamente el centro de los datos una vez la serie ha sido ordenada: el 50% queda por encima y el 50% por debajo.

Usualmente la mediana se representa como Me .

Los pasos a seguir para su cálculo son:

Ordenar la serie de menor a mayor.

Calcular 'n' (tamaño muestral).

1. Si 'n' es impar (hay un número impar de observaciones), la mediana será el valor central de la distribución, aquel que ocupa la posición $(n+1)/2$.
2. Si 'n' es par (hay un número par de observaciones), la mediana será el promedio de los dos valores centrales, aquellos que ocupan las posiciones $n/2$ y $n/2+1$.

Ventajas

Resuelve el problema de los valores extremos de la media y es fácilmente interpretable.

Limitaciones

Es muy sensible a los valores de la distribución: si se añaden observaciones o se prescinde de alguna, la mediana podría cambiar de forma significativa. Es independiente de la amplitud de los datos.

Desventajas

Dos distribuciones diferentes pueden tener la misma mediana.

Continuando con el ejemplo de ‘nuestras’ calificaciones, vamos a **calcular la mediana**. Primero recordamos que las notas:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	id9	i
Calificación	6	9	8	0	5	4	7	10	6	7

Si ordenamos de menor a mayor, obtendríamos:

$$(0, 4, 5, 6, 6, 7, 7, 8, 9, 10)$$

Como tenemos 10 calificaciones ($n=10$), la mediana será el promedio de los valores que se encuentran entre las posiciones $n/2=5$ y $n/2+1=6$, es decir, calificaciones de 6 y 7.

$$\text{Mediana} = \text{media } 6,7=6,5$$

3

Moda

La moda nos indica el valor que más veces se repite dentro de los datos. Si hay dos o más valores que tengan la máxima frecuencia, se dice que es una distribución multimodal.

A diferencia de la media aritmética, la moda no se ve afectada por la ocurrencia de los valores extremos.

Ventajas

Fácil de calcular e interpretar. No se ve afectada por valores extremos.

Limitaciones

En distribuciones muy uniformes no nos da información relevante. Es independiente de la amplitud de los datos.

Desventajas

Puede darse el caso de no tener ninguna moda o muchas.

Si continuamos con el ejemplo anterior, recordando que los datos son:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	id9	i
Calificación	6	9	8	0	5	4	7	10	6	7

Calculamos el número de veces que hemos tenido cada nota, ordenando los datos de menor a mayor previamente para facilitar su cálculo:

Calificación	0	4	5	6	7	8	9	10
Número de veces	1	1	1	2	2	1	1	1

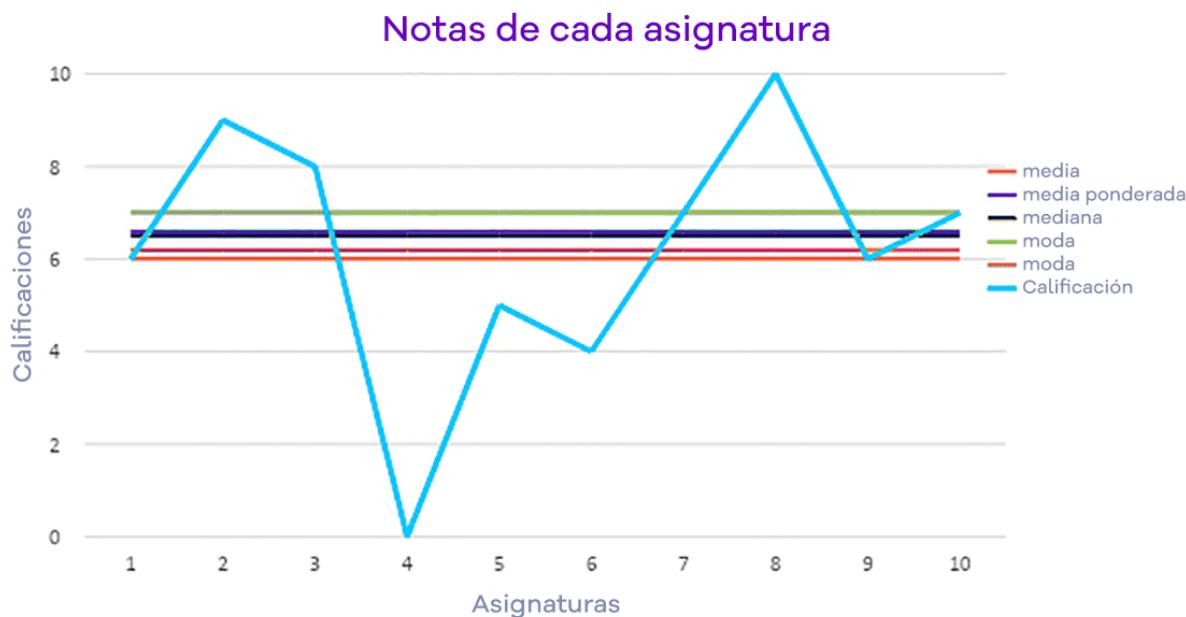
Las calificaciones más repetidas con 6 y 7, por lo que tenemos una distribución bimodal.

En resumen

Las distintas métricas nos han aportado distinto conocimiento. Si ponemos todas las medidas juntas en una tabla podremos comparar las fluctuaciones:

Medida	Valor
Media	6,2
Media ponderada	6,6
Mediana	6,5
Moda	{6, 7}

Puede parecer que las diferencias no son muy grandes, pero si graficamos junto con las calificaciones:



En este gráfico podemos ver la diferencia entre todas las medidas. ¿Te sorprende que las medidas parezcan tan altas en comparación con la distribución de las calificaciones?

Como hemos dicho nada más empezar, **graficando los datos vemos de un simple vistazo comportamientos que pueden resultar atípicos** y necesitan explicación (como puede ser en nuestro caso la asignatura id4).

Medidas de dispersión

X Edix Educación

Las medidas de dispersión son tan útiles y utilizadas como las medidas de tendencia central. El uso combinado de ambas ofrece una visión más completa del comportamiento de los datos.

Las medidas de dispersión nos ayudan a medir la variación de los datos respecto a su media.

Aunque hay varias métricas, nosotros nos centraremos en la varianza y la desviación estándar, ambas aplicables solamente a variables cuantitativas.

Aunque en este fastbook nos centraremos en la forma de calcular la media y la varianza, es vital apoyar el conocimiento extraído mediante gráficos.

1

Varianza

La varianza nos permite identificar la diferencia promedio que hay entre cada uno de los valores respecto a su media. La unidad de medida de la varianza es el cuadrado de la unidad de medida de la variable (por ejemplo: si estamos midiendo la altura en metros, la varianza se expresa en metros al cuadrado).

Su fórmula matemática es:

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

Leyenda:

- x representa el promedio de la variable.
- x_i es cada una de las observaciones.
- n es el tamaño muestral.

La principal desventaja de la varianza es que es muy sensible a valores atípicos, ya que, al calcularse mediante la diferencia respecto al promedio al cuadrado, hace que esos valores se disparen.

Continuamos con el ejemplo de las medidas de centralización. Recordamos que los datos son:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	id9	i
Calificación	6	9	8	0	5	4	7	10	6	7

El promedio lo calculamos anteriormente y es 6,2.

Calculamos ahora la varianza:

$$\sigma^2 = \frac{\sum_{i=1}^n (\bar{x}-x_i)^2}{n-1} = \frac{(6-6,2)^2 + (9-6,2)^2 + \dots + (7-6,2)^2}{10-1} = \frac{71,6}{9} = 7,96$$

Si quitamos la calificación de la asignatura id4 ya que sabemos que es no presentado, la varianza quedaría:

$$\sigma^2 = \frac{\sum_{i=1}^n (\bar{x}-x_i)^2}{n-1} = \frac{(6-6,2)^2 + (9-6,2)^2 + \dots + (7-6,2)^2}{9-1} = \frac{33,16}{8} = 3,61$$

Como vemos, la varianza presenta alta sensibilidad a la calificación de 0 (valor extremo de nuestra distribución) al reducirse la varianza en más de la mitad.

2

Desviación estándar

La desviación estándar o desviación típica (abreviada normalmente como 'sd' por su nombre en inglés *standard deviation*) es la raíz cuadrada de la varianza. Representa el promedio de la diferencia de los datos respecto a su media, y a diferencia de la varianza, se expresa en las mismas unidades que los datos a partir de los que se calcula.

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

Una desviación estándar baja (teniendo en cuenta el orden de magnitud de cada variable) indica que la mayor parte de los datos de una muestra tienden a estar agrupados cerca de su media, mientras que una desviación estándar alta indica que los datos se extienden sobre un rango de valores más amplio.

Si seguimos con el ejemplo anterior, recordemos que $\sigma^2=7,96$. La desviación típica será:

$$\sigma = \sqrt{\sigma^2} = \sqrt{7,96} = 2,82$$

Aunque la desviación estándar sea baja, los valores de nuestra variable fluctúan entre 0 y 10, por lo que podemos considerar que la dispersión es alta.

Medidas de posición

X Edix Educación

Las **medidas de posición** permiten estudiar la **distribución ordenada** de los datos en grupos similares. Son igual de importantes que las medidas de centralización, ya que dan una visión complementaria: **permiten entender cómo se divide la muestra** en grupos ordenados del mismo tamaño.

Este tipo de medidas solamente se puede aplicar a **variables cuantitativas**. Para el cálculo de estas medidas es necesario **ordenar los datos de menor a mayor**. Una vez hecho esto, podremos calcular los **percentiles, deciles y cuartiles**.

Al igual que para las medidas anteriores, en este fastbook nos centraremos en su forma de calcularlas, aunque deben llevar asociado gráficos que ayuden a entender la distribución de los datos y completen la visión global.

1

Cuartiles

Los cuartiles dividen la muestra en cuatro trozos iguales y se usan para clasificar una observación dentro de una población o muestra. Suelen representarse mediante la letra Q: Q_1 , Q_2 , Q_3

Q_1 (primer cuartil) es el valor que tiene por debajo el 25% de las observaciones (el primer cuarto), y de manera análoga Q_3 (tercer cuartil) deja por encima el 25% de las observaciones. Q_2 (segundo cuartil) separa la muestra en dos conjuntos iguales, 50% por debajo y 50% por encima, por lo que $Q_2 = \text{mediana} (Me)$.

La diferencia entre el tercer cuartil y el primero se conoce como rango intercuartílico.

Es una medida muy usada para crear gráficos (box-plots), detectar outliers, dispersión de gráficos, etc. Se define como $IQR=Q_3-Q_1$



Hay múltiples metodologías para calcular los cuartiles, nosotros nos quedaremos con el **método de Tukey**. Primero se ordenan los datos y se calcula $Q_2 = Me$, como aprendimos en el apartado anterior. Una vez conocida la mediana Q_2 , se divide el conjunto en dos grupos iguales y se calcula nuevamente la mediana de cada uno de estos subgrupos. La mediana del primer grupo se corresponde con el primer cuartil (Q_1) y la mediana del segundo grupo con el tercer cuartil (Q_3).

Sigamos con el ejemplo de las notas. Recordemos que los datos eran:

Asignatura	id1	id2	id3	id4	id5	id6	id7	id8	id9	i
Calificación	6	9	8	0	5	4	7	10	6	7

Ordenamos los datos: (0,4,5,6,6,7,7,8,9,10).

Anteriormente calculamos la mediana, $Q_2 = Me = 6,5$.

Dividimos mediante la mediana al conjunto en dos: (0,4,5,6,6) y (7,7,8,9,10).

Calculamos las medianas para cada conjunto, de tal forma que tenemos que la mediana del primer grupo es 5, y la del segundo 8.

Finalmente, tendríamos $Q_1=5$, $Q_2 = 6,5$, y $Q_3=8$. El rango intercuartílico es $IQR=Q_3-Q_1=8-5=3$.

2

Deciles

Los deciles muestran la división ordenada de la muestra en diez trozos, y se denotan como D_k , siendo $k=1,2,\dots,10$ cada uno de los deciles. D_1 tendrá un 10% de los datos ordenados por debajo, D_2 tendrá el conjunto de datos ordenados entre el 10% y el 20% inferior, etc. En este caso, $D_5 = Q_2 = Me$.

La forma de calcular los deciles es un poco más compleja, pero con el ejemplo se entenderá mejor:

Ordenar los datos de menor a mayor.

Calcular el tamaño muestral 'n'.

Calculamos la posición del valor del decil: $PD_i = (i*(n+1))/10$.

1. Si PD_i es entero, D_i es el valor de la posición PD_i .
2. Si PD_i es decimal, D_i se calcula con los valores de las posiciones más cercanas a PD_i , es decir, cogemos los valores por debajo de PD_i (digamos X_{inf_i} al valor que está en la posición inmediatamente inferior a PD_i) y por arriba (digamos X_{sup_i} a la posición inmediatamente superior). Aplicamos la fórmula
$$D_i = X_{in_i} + d * (X_{sup_i} - X_{in_i})$$
; donde 'd' es la parte decimal de PD_i .

Veamos un ejemplo: supongamos que queremos ver la distribución de la edad de los 23 empleados de una empresa ($n=23$). Recogemos sus edades, ordenamos y obtenemos los siguientes datos:

ID	id1	id2	id3	id4	id5	id6	id7	id8	id9	id10	
Edad	23	24	27	30	32	32	32	33	36	36	
ID	id13	id14	id15	id16	id17	id18	id19	id20	id21		
Edad	51	54	55	55	56	57	60	62	63		

Calcularemos los deciles D_1 , D_2 y D_5 como ejemplo.

Empecemos con D_1 :

$$PD_1 = \frac{1*(n+1)}{10} = \frac{1*(23+1)}{10} = 2,4$$

Como PD_1 no es entero, calculamos el valor de D_1 con los valores que se encuentren en la posición 2 y 3, es decir, 24 y 27, respectivamente. La parte decimal de PD_1 es $d=0,4$. Ahora ya podemos calcular D_1 :

$$D_1 = X_{inf_1} + d * (X_{sup_1} - X_{inf_1}) = 24 + 0,4 * (27 - 24) = 25,2$$

Calculamos ahora D_2 de igual forma:

$$PD_2 = \frac{2*(n+1)}{10} = \frac{2*(23+1)}{10} = 4,8$$

Como PD_2 no es entero, calculamos PD_2 con los valores 30 y 32 (en las posiciones 4 y 5, respectivamente).

$$D_2 = X_{inf_2} + d * (X_{sup_2} - X_{inf_2}) = 30 + 0,8 * (32 - 30) = 31,6$$

Calculamos ahora D_5 :

$$PD_5 = \frac{5*(n+1)}{10} = \frac{5*(23+1)}{10} = 12$$

Como $PD_5=12$ es entero, D_5 es el valor que se encuentre en la posición 12, es decir, $D_5=45$.

3

Percentiles

Por último, **con los percentiles tenemos la división en cien partes iguales de la muestra.** En este caso se representa como P_k , donde $k=1, 2, \dots, 100$. En este caso las equivalencias son: $P_{50} = D_5 = Q_2 = Me$.

La forma de calcular los percentiles es la misma que la de los deciles, excepto por el cálculo de la posición en el paso 3, donde $PD_i = (i * (n+1)) / 100$.

Resumen

X Edix Educación

Con este fastbook nos hemos introducido en el mundo de la estadística conociendo los **principales conceptos y técnicas estadística**:

- Hemos conocido el **origen de la palabra estadística**.
- Nos hemos familiarizado con los conceptos de **población, muestra, individuo y tamaño**. Hemos conocido diferentes técnicas de muestreo, y la importancia que tiene el tamaño en un estudio estadístico.
- Hemos conocido los principales tipos de **variables existentes** y cómo se podrían almacenar en bases de datos.
- A través de las **medidas de centralización**, sabemos calcular la media, media ponderada, mediana y moda de cualquier distribución, así como su significado, utilidad y puntos a tener en consideración (ventajas, limitaciones, desventajas).
- Mediante las **medidas de dispersión**, hemos conocido el significado de varianza y desviación estándar, su forma de calcularlo y el significado que tienen.
- Por último, nos hemos familiarizado con las **medidas de posición**, conocemos el significado de **cuartiles (y rango intercuartílico), deciles y percentiles**. Sabemos cómo calcular cada una de las medidas, así como las equivalencias entre ellas y con la mediana.

¡Enhorabuena! Fastbook superado

edix

Creamos Digital Workers