Apuntes - Estadística descriptiva

Curso: Universidad 1

La estadística descriptiva organiza y resume datos para describir una muestra: tablas de frecuencias, medidas de posición y dispersión y relación entre dos variables. Es el paso previo a la probabilidad y la inferencia, y se usa en toda ciencia experimental.

Actividad de introducción

Una cadena de supermercados de Castilla y León recibe cada día miles de tiques de compra. Con tantos datos, nadie puede leerlos uno a uno: hace falta resumirlos en unas pocas cifras y gráficos que digan cuánto gasta un cliente típico, cuánto varían las compras o si las ventas de un producto suben cuando baja su precio. Elegir mal el resumen puede llevar a conclusiones falsas, como cuando un único dato extremo dispara la media. La estadística descriptiva proporciona las herramientas para organizar, resumir y representar datos de forma honesta, y es el primer paso de cualquier estudio científico, médico, económico o social.

Variables y tablas de frecuencias

Definición: variables y tablas de frecuencias

Una **variable estadística** es una característica que se mide en cada individuo. Puede ser *cualitativa* (color, carrera) o *cuantitativa*, discreta (número de hermanos) o continua (estatura). Para una variable discreta con valores x_i, la **frecuencia absoluta** f_i es el número de veces que aparece, la **relativa** $h_i = f_i/n$ y las **acumuladas** F_i, H_i van sumando. Para datos continuos se agrupan en intervalos y se trabaja con la marca de clase (punto medio).

Medidas de posición y dispersión

Definición: medidas de posición y dispersión

La **media** es $\bar{x} = \frac{1}{n}\sum f_i x_i$; la **mediana** es el valor central (deja la mitad de los datos a cada lado); la **moda** es el valor más frecuente. Los **cuartiles** Q_1, Q_2, Q_3 dividen los datos ordenados en cuatro partes; el **rango intercuartílico** es $Q_3 - Q_1$. Medidas de dispersión: **varianza** $\sigma^2 = \frac{1}{n}\sum f_i (x_i - \bar{x})^2 = \frac{1}{n}\sum f_i x_i^2 - \bar{x}^2$, **desviación típica** $\sigma = \sqrt{\sigma^2}$ y **coeficiente de variación** $CV = \sigma / |\bar{x}|$, que permite comparar la dispersión de variables con distintas unidades.

Propiedad

Propiedades. Si $y = ax + b$, entonces $\bar{y} = a\bar{x} + b$ y $\sigma_y^2 = a^2\sigma_x^2$ (sumar una constante no cambia la dispersión). La media de dos grupos es la media ponderada por los

tamaños. La media es sensible a valores atípicos; la mediana no. Un dato es *atípico* (criterio de Tukey) si está fuera de $[Q_1 - 1{,}5\,IQR,\ Q_3 + 1{,}5\,IQR]$. La tipificación $z = \frac{x-\bar{x}}{\sigma}$ mide cuántas desviaciones típicas se aleja un dato de la media. Desigualdad de Chebyshev: al menos el $1 - \frac{1}{k^2}$ de los datos está a menos de k desviaciones de la media.

Dos variables: correlación y regresión

Propiedad

Para datos (x_i, y_i) la **covarianza** es $\sigma_{xy} = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$ y el **coeficiente de correlación** de Pearson es $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \in [-1, 1]$. Valores cercanos a ± 1 indican relación lineal fuerte; cerca de 0, ausencia de relación lineal (puede haber otra). La **recta de regresión de y sobre x** es $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$; minimiza la suma de cuadrados de los residuos. El **coeficiente de determinación** r^2 es la proporción de varianza explicada. Predecir fuera del rango de datos (extrapolar) es poco fiable, y correlación no implica causalidad.

Método: recta de regresión

1. Calcula las medias de x e y.
2. Calcula la covarianza y la varianza de x.
3. Obtén la pendiente b = covarianza / varianza.
4. Escribe y − media(y) = b (x − media(x)) y calcula r.

Ejemplo 1

Datos 3, 5, 5, 6, 7, 8, 10, 12.

Media $56/8 = 7$; mediana $(6+7)/2 = 6{,}5$; moda 5; $\sigma^2 = \frac{60}{8} = 7{,}5$ y $\sigma \approx 2{,}74$.

Ejemplo 2

Tabla $x = 1, 2, 3, 4, 5$ con $f = 2, 5, 8, 3, 2$ ($n = 20$): $\bar{x} = 58/20 = 2{,}9$; $\sigma^2 = 192/20 - 2{,}9^2 = 1{,}19$.

Ejemplo 3

Puntos $(1,2), (2,3), (3,5), (4,4), (5,6)$: $\bar{x} = 3$, $\bar{y} = 4$, $\sum(x-\bar{x})(y-\bar{y}) = 9$, $\sum(x-\bar{x})^2 = 10$.

Pendiente $0{,}9$ y recta $y = 1{,}3 + 0{,}9x$.

¡Atención!

Confundir varianza poblacional (dividida por n) con la muestral (dividida por $n-1$); usar los extremos de un intervalo en lugar de la marca de clase; interpretar una correlación alta como causalidad; olvidar elevar al cuadrado al cambiar de escala la varianza; calcular la media de medias sin ponderar.

Más herramientas descriptivas

Método: estudio descriptivo de una variable

1. Clasifica la variable: cualitativa, cuantitativa discreta o continua.
2. Construye la tabla de frecuencias; si hay muchos valores distintos, agrúpalos en intervalos de igual amplitud.
3. Calcula medidas de posición (media, mediana, moda, cuartiles) y de dispersión (rango, varianza, desviación típica, rango intercuartílico).
4. Detecta posibles datos atípicos con el criterio de Tukey y estudia la forma: simetría o asimetría.
5. Elige un gráfico adecuado (barras, histograma, diagrama de caja) y redacta una conclusión en el lenguaje del problema.

Definición: cuasivarianza

En muchos programas estadísticos y calculadoras aparece, además de la varianza σ^2, la cuasivarianza $s^2 = \frac{1}{n-1}\sum(x_i-\bar{x})^2$. Cuando los datos son una muestra y se quiere estimar la varianza de toda la población, dividir entre $n-1$ en lugar de n corrige un pequeño sesgo, como se estudia en inferencia. Para muestras grandes ambas apenas difieren, pero conviene saber cuál usa la calculadora.

Propiedad: forma de la distribución

Si la distribución es simétrica, la media y la mediana coinciden aproximadamente. Si tiene una cola larga hacia la derecha (asimetría positiva), como los salarios o los precios de la vivienda, la media es mayor que la mediana; si la cola va hacia la izquierda, ocurre lo contrario. Comparar media y mediana es la forma más sencilla de detectar la asimetría sin necesidad de dibujar.

Ejemplo 4

Calcula la media y la desviación típica de 3, 5, 7, 9 y 11.

$\bar{x} = \frac{35}{5} = 7$. Las desviaciones son $-4, -2, 0, 2, 4$, cuyos cuadrados suman 40; $\sigma^2 = 8$ y $\sigma = 2\sqrt{2} \approx 2{,}83$. Los datos están equiespaciados y la media coincide con la mediana.

Ejemplo 5

Las notas de una clase tienen media 5,5 y desviación típica 1,5. El profesor sube todas un 10 % y después suma medio punto. ¿Cuáles son la nueva media y la nueva desviación?

La transformación es $y = 1{,}1x + 0{,}5$. La media pasa a $1{,}1 \cdot 5{,}5 + 0{,}5 = 6{,}55$ y la desviación típica a $1{,}1 \cdot 1{,}5 = 1{,}65$; el medio punto sumado no cambia la dispersión.

Ejemplo 6

En una frutería de Lleida se anotan el precio x (€/kg) y los kilos vendidos y de manzanas en 4 días: $(1, 40)$, $(1{,}5, 32)$, $(2, 26)$, $(2{,}5, 18)$. ¿Qué indica la correlación?

Los datos bajan de forma casi regular: cada 0,5 € de subida se venden entre 6 y 8 kg menos. La covarianza es negativa y $r \approx -0{,}998$: relación lineal negativa casi perfecta, como cabe esperar de una curva de demanda.

Representaciones gráficas y buenas prácticas

Un buen gráfico permite ver de un vistazo lo que las tablas esconden. Para variables cualitativas se usan diagramas de barras o de sectores; para variables discretas, diagramas de barras; y para continuas agrupadas, histogramas, en los que el área de cada rectángulo es proporcional a la frecuencia. El diagrama de caja resume cinco números (mínimo, primer cuartil, mediana, tercer cuartil y máximo) y señala los atípicos, por lo que es ideal para comparar varios grupos. Para dos variables cuantitativas, el diagrama de dispersión es imprescindible antes de calcular cualquier correlación: muestra si la relación es lineal, curva o inexistente, y si hay datos que distorsionan los cálculos. Un buen análisis también evita trampas frecuentes: ejes que no empiezan en cero para exagerar diferencias, porcentajes calculados sobre bases distintas o conclusiones causales a partir de correlaciones. La correlación entre el consumo de helados y los ahogamientos en las playas es alta, pero ambos se deben a una tercera variable, el calor del verano. Por eso la estadística descriptiva describe, pero no explica: para hablar de causas hacen falta experimentos bien diseñados, en los que se controlan las demás variables y los individuos se asignan al azar a cada grupo. Aun así, describir bien los datos es imprescindible: permite detectar errores de registro, plantear hipótesis razonables y elegir el modelo de probabilidad adecuado para la inferencia posterior.

¡Atención!

Con datos agrupados, la media y la varianza calculadas con las marcas de clase son aproximaciones. La varianza nunca puede ser negativa: si sale negativa, hay un error de cálculo, casi siempre al restar $\bar{x}^2$. Un coeficiente de correlación cercano a 0 solo indica que no hay relación lineal; puede existir una relación curva muy fuerte.

¿Sabías que?

Francis Galton estudió a finales del siglo XIX la estatura de padres e hijos y observó que los hijos de padres muy altos tienden a ser algo más bajos que ellos. Llamó a este fenómeno regresión a la media, y de ahí procede el nombre de la recta de regresión. Su discípulo Karl Pearson formalizó el coeficiente de correlación que usamos hoy. En España, el Instituto Nacional de Estadística, creado en 1945, publica cada mes indicadores como el índice de precios de consumo o la tasa de paro, que son medias y proporciones calculadas con estas mismas técnicas a partir de encuestas a miles de hogares y empresas. Saber leer esos indicadores, entender qué miden y qué dejan fuera es una competencia básica para cualquier ciudadano, y no solo para los especialistas.

El consejo de Otto

antes de calcular, dibuja los datos (diagrama de puntos o de cajas). Un gráfico te dice si la media tiene sentido o si hay atípicos que la distorsionan.