



Seminario: Aspectos estadísticos fundamentales para un análisis crítico de la literatura científica: análisis de los datos

Autor: D. Pablo González Navarro

1. Introducción

La estadística es una disciplina matemática y científica que se encarga de recopilar, organizar, analizar e interpretar datos numéricos para describir y entender fenómenos en diferentes campos del conocimiento. Esta herramienta es esencial para la toma de decisiones informadas, la identificación de patrones, la detección de tendencias y la realización de inferencias basadas en evidencia empírica. La estadística juega un papel fundamental en las ciencias de la vida al permitirnos analizar y comprender la variabilidad inherente en los datos biológicos y de salud.

En las ciencias de la vida, la estadística desempeña un papel crucial al ayudar a los investigadores a analizar datos provenientes de experimentos, estudios clínicos y observaciones. La variabilidad inherente en los sistemas biológicos y de salud hace que la estadística sea esencial para distinguir entre resultados genuinos y fluctuaciones aleatorias. Además, la estadística permite tomar decisiones basadas en la evidencia, evaluar la eficacia de tratamientos médicos y comprender las relaciones entre variables biológicas.

La estadística descriptiva se enfoca en resumir y presentar los datos de manera concisa y comprensible. Utiliza herramientas como medidas de tendencia central (como la media, la mediana y la moda) para identificar valores promedio y representativos en un conjunto de datos. Además, utiliza medidas de dispersión (como la desviación estándar y el rango intercuartílico) para comprender la variabilidad en los datos. Los gráficos y diagramas también son parte integral de la estadística descriptiva, ya que visualmente representan la distribución y patrones de los datos.



La estadística inferencial, por otro lado, se ocupa de hacer inferencias y conclusiones sobre una población más amplia basándose en una muestra de datos. A través de métodos como la estimación de parámetros y las pruebas de hipótesis, la estadística inferencial permite tomar decisiones sobre poblaciones completas a partir de información limitada. Esto es esencial en la investigación científica, ya que no siempre es posible examinar cada individuo o elemento de una población.

2. Tipos de variables en estadística

En estadística, las variables son elementos fundamentales que representan características o cualidades que pueden variar en una población o muestra. Estas variables son la base para realizar análisis, inferencias y conclusiones sobre los datos recopilados. Los tipos de variables se dividen en dos categorías principales: variables cualitativas y variables cuantitativas.

2.1 Variables cualitativas

Las variables cualitativas, también conocidas como variables categóricas, representan atributos o categorías que no pueden medirse numéricamente, sino que se clasifican en distintas clases. Estas se dividen en dos subtipos: nominales y ordinales. Las variables cualitativas suelen representarse gráficamente mediante gráficos de barras o gráficos circulares (gráficos de pastel).

2.1.1 Cualitativas nominales

Las variables nominales son aquellas en las que las categorías no tienen un orden intrínseco. Ejemplos incluyen el género, estado civil, color de pelo, presencia o ausencia de mutación, etc. Se describen mediante frecuencias absolutas y relativas, y la medida de tendencia central más común es la moda, que es la categoría más frecuente en el conjunto de datos.



2.1.2 Cualitativas ordinales

Las variables ordinales representan categorías con un orden inherente, pero la distancia entre las categorías no es cuantificable. Ejemplos incluyen la clasificación socioeconómica, niveles de satisfacción, niveles de estudios, etc. Se describen mediante frecuencias absolutas y relativas. Para estas variables, además de la moda, se puede calcular la mediana como medida de tendencia central. En términos de dispersión, se pueden utilizar rangos intercuartílicos para evaluar la extensión de los valores.

2.2 Variables cuantitativas

Las variables cuantitativas, también llamadas variables numéricas, representan cantidades o valores que pueden ser medidos con precisión. Estas se dividen en dos subtipos: discretas y continuas.

2.2.1 Cuantitativas discretas

Las variables discretas toman valores enteros y no pueden tener valores intermedios entre ellos. Ejemplos de variables discretas incluyen el número de hijos en una familia, número de intervenciones, etc. Para estas variables, se puede calcular la moda y la mediana como medidas de tendencia central. La dispersión se mide utilizando la desviación estándar o el rango intercuartílico. Las variables cuantitativas discretas se pueden representar con histogramas, donde las barras representan los valores y sus frecuencias.

2.2.2 Cuantitativas continuas

Las variables continuas pueden tomar cualquier valor dentro de un rango específico. Ejemplos de variables continuas son la altura, el peso, la temperatura, etc. Las medidas de tendencia central típicas para variables continuas son la media y la mediana. Para evaluar la dispersión, se utilizan la desviación estándar o el rango y el rango



intercuartílico. Las variables cuantitativas continuas se pueden representar con histogramas o gráficos de densidad, que muestran la distribución de probabilidad.

2.3 Parámetros de centralidad

La centralidad se refiere a los valores que representan el centro o punto medio de un conjunto de datos. Estos valores son útiles para entender dónde se concentran los datos en torno a un valor central.

Media: La media, también conocida como promedio, se calcula sumando todos los valores en un conjunto de datos y luego dividiendo esa suma entre el número total de valores. Es un indicador sensible a valores extremos, ya que, si hay valores atípicos, puede verse afectada significativamente.

Mediana: La mediana es el valor medio de un conjunto de datos ordenados. Para encontrarla, primero se ordenan los datos y luego se elige el valor que está justo en el medio. La mediana no se ve tan influenciada por valores extremos como la media, lo que la convierte en una medida de centralidad más robusta en presencia de datos atípicos.

Moda: La moda es el valor que aparece con mayor frecuencia en un conjunto de datos. Puede haber una, ninguna o varias modas en un conjunto de datos. La moda es especialmente útil en datos categóricos o discretos.

2.4 Parámetros de dispersión

La dispersión describe cuán dispersos o extendidos están los valores en un conjunto de datos alrededor de su centro. Los parámetros de dispersión ayudan a comprender la variabilidad y la amplitud de los datos.

Desviación estándar: La desviación estándar mide la dispersión promedio de los valores individuales con respecto a la media. Valores más altos de desviación estándar indican mayor dispersión de los datos. Es una medida útil para entender la variabilidad en relación con la media.



Rango intercuartílico: El rango intercuartílico es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) de un conjunto de datos ordenados. Los cuartiles dividen los datos en cuatro partes iguales. El rango intercuartílico es útil para identificar la dispersión de los datos mientras se minimiza la influencia de valores atípicos.

3. Estadística inferencial

3.1 Población y Muestra

Población: Se refiere al conjunto completo de elementos o individuos que poseen una característica específica que estamos estudiando. Por ejemplo, si estamos interesados en la altura de todos los estudiantes de una escuela, la población sería el conjunto de todas las alturas de esos estudiantes.

Muestra: Es un subconjunto representativo de la población que se selecciona para ser estudiado. Debido a que suele ser impracticable analizar a toda la población, se toma una muestra para hacer inferencias sobre la población completa. La calidad de la muestra y su representatividad son esenciales para obtener conclusiones fiables.

3.2 Formulación de Hipótesis

Hipótesis nula (H₀): Es una afirmación inicial que asume que no hay efecto, relación o diferencia significativa entre variables. Por ejemplo, en un experimento de medicamentos, la H₀ podría afirmar que el medicamento no tiene ningún efecto.

Hipótesis alternativa (H₁ o H_a): Es la afirmación contraria a la hipótesis nula. Afirma la existencia de un efecto, relación o diferencia significativa entre variables. Siguiendo el ejemplo del medicamento, la H₁ podría afirmar que el medicamento sí tiene un efecto.

3.3 Tipos de Error en los Contrastes de Hipótesis

Error tipo I (α): También conocido como "falso positivo", ocurre cuando se rechaza incorrectamente la hipótesis nula cuando en realidad es verdadera. La probabilidad de cometer este error se denota como α .



Error tipo II (β): También llamado "falso negativo", ocurre cuando se falla en rechazar la hipótesis nula cuando en realidad es falsa. La probabilidad de cometer este error se denota como β .

3.4 Valor p

El valor p es una medida clave en la estadística inferencial que ayuda a determinar si los resultados observados en una muestra son estadísticamente significativos. Representa la probabilidad de obtener los resultados observados (o más extremos) si la hipótesis nula es verdadera. Si el valor p es pequeño (generalmente menor que 0.05), se considera evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alternativa.

3.5 Intervalos de Confianza y Estimaciones Puntuales

Intervalos de Confianza: Son rangos de valores que se calculan a partir de los datos de la muestra y que se utilizan para estimar el valor real de una característica en la población. Por ejemplo, un intervalo de confianza del 95% para la media de una variable indica que hay un 95% de probabilidad de que la verdadera media esté dentro de ese intervalo.

Estimaciones Puntuales: Son valores específicos que se calculan a partir de los datos de la muestra y que se utilizan para hacer una estimación de una característica en la población. Por ejemplo, la media muestral es una estimación puntual de la media poblacional.