estadística sesgada – biased statistic

Authored by
memjavad

November 7, 2025

Estadístico Sesgado (Biased Statistic)

Primary Disciplinary Field(s): Estadística, Inferencia Estadística.

1. Definición Central

En el ámbito fundamental de la inferencia estadística, un **estadístico sesgado** es aquel estimador cuyo valor esperado difiere del verdadero valor del parámetro poblacional que intenta medir. Formalmente, si T es un estimador para el parámetro poblacional $theta$, el sesgo (B) se define como la diferencia entre el valor esperado del estimador y el parámetro real: $B(T) = E - theta$. Si esta diferencia es distinta de cero, el estadístico T es considerado sesgado. Este fenómeno indica una desviación sistemática, no aleatoria, en la estimación, lo que implica que, incluso si se repitiera el proceso de muestreo un número infinito de veces, el promedio de las estimaciones no convergería al valor poblacional verdadero. La existencia de sesgo es un problema crítico, ya que socava la validez y la fiabilidad de las conclusiones extraídas de los datos muestrales, resultando en estimaciones que consistentemente se encuentran por encima o por debajo del valor real.

Es crucial distinguir el concepto de sesgo de la varianza. Mientras que la varianza mide la dispersión o la aleatoriedad de las estimaciones alrededor de su media (que puede o no ser el parámetro verdadero), el sesgo mide la precisión o exactitud sistemática. Un estimador puede ser altamente sesgado pero tener baja varianza (siempre da un valor erróneo, pero consistente), o puede ser insesgado pero tener alta varianza (su promedio es correcto, pero las estimaciones individuales varían mucho). El objetivo ideal de la inferencia estadística es encontrar estimadores que posean tanto un sesgo bajo como una varianza baja, propiedades que a menudo entran en conflicto, llevando al concepto de compromiso entre sesgo y varianza (**Bias-Variance Tradeoff**). En muchos contextos aplicados, una ligera cantidad de sesgo es aceptable si conduce a una reducción sustancial de la varianza, mejorando la estabilidad de la estimación.

2. El Sesgo y el Error Cuadrático Medio

El sesgo, entendido como la desviación esperada, se relaciona intrínsecamente con el error total de un estimador, cuantificado generalmente a través del Error Cuadrático Medio (ECM). El ECM de un estimador T se descompone matemáticamente en la suma de su varianza y el cuadrado de su sesgo: $text{ECM}(T) = text{Var}(T) + ^2$. Esta identidad fundamental demuestra que un estadístico sesgado contribuye directamente, a través del término del sesgo al cuadrado, al error total de la estimación. La minimización del ECM es, por lo tanto, un criterio fundamental en la selección de un estimador, y es el fundamento matemático que justifica la preferencia por estimadores ligeramente sesgados si su menor varianza global resulta en un ECM inferior al de cualquier estimador insesgado disponible.

La presencia de sesgo en un estadístico no implica automáticamente que sea inútil. De hecho, muchos estimadores de uso común son inherentemente sesgados en muestras pequeñas. Un ejemplo clave es el estimador de máxima verosimilitud (MLE), que posee propiedades de optimalidad bajo ciertas condiciones (como la eficiencia asintótica), pero que es sesgado para tamaños muestrales finitos. La teoría estadística moderna se centra en estimadores que son al menos **asintóticamente insesgados**, lo que significa que el sesgo tiende a cero a medida que el tamaño de la muestra (n) se acerca al infinito. Para aplicaciones prácticas con tamaños muestrales limitados, la evaluación del ECM se convierte en el estándar de oro para determinar la calidad de un estimador, incluso si este tiene un sesgo finito.

3. Tipos Comunes de Sesgo en la Estimación

El sesgo puede surgir de diversas fuentes, dependiendo tanto de la naturaleza intrínseca de la fórmula del estimador como del proceso mediante el cual se recopilan los datos. La identificación correcta de la fuente es esencial para la mitigación.

Sesgo del Estimador (Estimator Bias): Este es el sesgo formal definido matemáticamente ($E neq theta$). Ocurre cuando la fórmula utilizada para calcular el estadístico es inherentemente defectuosa o insuficiente para el tamaño muestral dado. El ejemplo paradigmático es el estimador de la varianza poblacional que utiliza el divisor n, el cual es sistemáticamente sesgado a la baja, subestimando el parámetro verdadero.
Sesgo de Muestreo (Sampling Bias): Este tipo de sesgo surge del diseño experimental o del método de recolección de datos, no de la fórmula matemática. Ocurre cuando la muestra seleccionada no es representativa de la población de interés. Esto puede incluir el sesgo de conveniencia, donde solo se muestrean elementos de fácil acceso, o el sesgo de supervivencia, donde solo se observan los elementos que han "sobrevivido" a un proceso selectivo, ignorando a los que fallaron o desaparecieron.
Sesgo de Variables Omitidas (Omitted Variable Bias - OVB): Prevalente en los modelos de regresión, el OVB surge cuando una variable que influye tanto en la variable dependiente como en al menos una de las variables independientes incluidas, es excluida del modelo. La exclusión hace que el modelo atribuya incorrectamente el efecto de la variable omitida a las variables incluidas, resultando en coeficientes de regresión sesgados e inconsistentes. Este es un problema grave de especificación del modelo.
Sesgo de Medición (Measurement Error Bias): Resulta de errores sistemáticos o no aleatorios en la forma en que se miden las variables. Si el error de medición no es aleatorio (por ejemplo, un instrumento de medición que consistentemente subestima el peso), introduce un sesgo directo en la estimación de los parámetros poblacionales relacionados con esa variable.

4. Propiedades Deseables y el Insesgamiento

El insesgamiento es una de las principales propiedades que los estadísticos buscan en un estimador, aunque debe equilibrarse con otras cualidades para lograr la optimalidad.

Para que un estimador sea considerado de alta calidad, generalmente se evalúan cuatro propiedades clave. El **insesgamiento** ($E = \theta$) asegura que, a largo plazo, el método de estimación no produce errores sistemáticos. La **consistencia** es una propiedad más débil, asegurando que el estimador se acerca al valor verdadero a medida que el tamaño de la muestra crece. La **eficiencia** se refiere a la baja varianza del estimador; dentro de la clase de estimadores insesgados, el más eficiente es aquel con la menor varianza, a menudo limitado por la Cota de Cramér-Rao. Finalmente, la **suficiencia** garantiza que el estimador utiliza toda la información relevante que la muestra puede proporcionar sobre el parámetro.

Históricamente, la escuela de pensamiento frecuentista ha dado una gran importancia al insesgamiento. Sin embargo, en la práctica moderna, especialmente en el aprendizaje automático y la estadística bayesiana, la minimización del ECM (que permite un compromiso entre sesgo y varianza) es a menudo el criterio dominante. Un estimador insesgado no siempre es el "mejor" en términos de rendimiento predictivo o estabilidad, especialmente en situaciones de alta dimensionalidad donde la varianza puede ser incontrolable.

5. Ejemplos Canónicos y la Corrección de Bessel

El ejemplo más didáctico y fundamental de un estadístico sesgado corregible es el estimador de la varianza poblacional. Si se utiliza la media muestral ($\bar{X}$) en lugar de la media poblacional verdadera (μ) para calcular las desviaciones cuadráticas, la fórmula tradicional para la varianza muestral, $\sigma^2_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$, resulta ser un estimador sesgado a la baja (subestima σ^2). Este sesgo se debe a que la media muestral se ajusta a los datos de la muestra, minimizando artificialmente la suma de las desviaciones cuadráticas.

La solución estándar a este problema es la aplicación de la **Corrección de Bessel**, que reemplaza el divisor n por $n-1$. El estimador resultante, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, es un **estadístico insesgado** de la varianza poblacional, ya que $E = \sigma^2$. La cantidad $n-1$ representa los grados de libertad, indicando el número de observaciones independientes disponibles para estimar la varianza una vez que se ha utilizado la media muestral para estimar la media poblacional. Este ajuste es un procedimiento universal en la estadística inferencial.

6. Métodos de Reducción y Eliminación del Sesgo

La estrategia para manejar el sesgo depende de su origen. Si el sesgo es conocido y analíticamente manejable (como en el caso de la varianza), se aplica una simple corrección algebraica. Sin embargo, para sesgos más complejos, se requieren métodos más sofisticados.

Uno de los enfoques más poderosos es el **método *Bootstrap***. Esta técnica de remuestreo permite estimar empíricamente la distribución muestral de un estadístico y, por lo tanto, calcular su sesgo. Una vez que se estima el sesgo, se puede restar del estimador original para obtener un estimador corregido por sesgo. Además, las técnicas de regularización, como la regresión *ridge* o el *lasso*, introducen intencionalmente un sesgo en los estimadores de los coeficientes. Aunque estos estimadores son formalmente sesgados, su capacidad para reducir la varianza en modelos complejos con alta correlación entre predictores a menudo resulta en un ECM sustancialmente menor y una mejor capacidad de generalización.

En el campo de la econometría, cuando el sesgo surge de la endogeneidad o la correlación entre los regresores y el error (común en el OVB o en la causalidad inversa), se emplean técnicas avanzadas como la regresión de variables instrumentales (IV) o el método de Mínimos Cuadrados en Dos Etapas (2SLS). Estos métodos están diseñados para producir estimadores **consistentes** (que convergen al valor verdadero a medida que $n to infty$), aunque a menudo son sesgados en muestras pequeñas, priorizando la consistencia sobre el insesgamiento finito para garantizar la validez asintótica de las conclusiones.

7. Consecuencias en la Inferencia y la Toma de Decisiones

La presencia de un estadístico sesgado tiene profundas implicaciones para la validez de la inferencia estadística, llevando a conclusiones sistemáticamente distorsionadas que pueden tener consecuencias prácticas severas. Si un estimador de un efecto de tratamiento está sesgado al alza, por ejemplo, los investigadores podrían sobreestimar consistentemente la eficacia de una política o medicamento, resultando en una asignación ineficiente de recursos o, peor aún, en la adopción de intervenciones ineficaces o dañinas.

Además, el sesgo afecta la cobertura de los intervalos de confianza. Los intervalos de confianza se construyen asumiendo que el estimador es insesgado o, al menos, asintóticamente insesgado. Si el estimador es significativamente sesgado, el centro del intervalo de confianza se desplaza sistemáticamente lejos del verdadero parámetro poblacional. En consecuencia, la probabilidad real de que el intervalo contenga el parámetro verdadero (la cobertura) será menor que el nivel nominal declarado (por ejemplo, 95%), invalidando la seguridad estadística que se intenta transmitir. Por lo tanto, la detección y corrección del sesgo no es solo una preocupación teórica, sino un requisito fundamental para garantizar la robustez y la credibilidad de la ciencia basada en datos.

8. Lecturas Adicionales

Estadístico insesgado (Wikipedia)
Error Cuadrático Medio (Wikipedia)
Varianza Muestral y Corrección de Bessel (Wikipedia)
Método Bootstrap (Wikipedia)