Guía de estudio - Estadistica Inferencial

¡Hola! Te damos la bienvenida a esta nueva guía de estudio.

¿En qué consiste esta guía?

La siguiente guía de estudio tiene como objetivo profundizar contenidos adicionales, además de recordar y repasar los temas que hemos tratado en clase.

¡Vamos con todo!

Tabla de contenidos

¡Comencemos!

Estadística Inferencial

La estadística inferencial es una rama de la estadística que se centra en hacer conclusiones o inferencias acerca de una población más amplia basándose en la información recopilada de una muestra representativa de esa población. En otras palabras, la estadística inferencial busca sacar conclusiones generales sobre una población utilizando datos recopilados de una muestra más pequeña, en lugar de examinar cada individuo de la población en sí.

Los principales objetivos de la estadística inferencial son estimar parámetros desconocidos de la población y probar hipótesis sobre esas poblaciones utilizando información muestral. Para lograr estos objetivos, se emplean métodos probabilísticos y técnicas de análisis que permiten medir la incertidumbre asociada a las conclusiones extraídas de la muestra.

Supongamos que eres un científico que trabaja en una empresa agrícola y deseas determinar si un nuevo fertilizante tiene un efecto significativo en el rendimiento de cultivos de tomate en comparación con el fertilizante existente. Para llevar a cabo este estudio, decides realizar un experimento en el que cultivas dos grupos de plantas de tomate idénticas. Un grupo recibe el nuevo fertilizante y el otro grupo recibe el fertilizante existente.

¿En qué se diferencia de la estadística descriptiva?

La estadística inferencial y la estadística descriptiva son dos ramas interrelacionadas pero distintas de la estadística que cumplen roles diferentes en el análisis de datos.

La estadística descriptiva se centra en resumir, organizar y presentar datos de manera informativa, y así proporcionar una comprensión clara y concisa de los datos utilizando medidas como la media, la mediana, la moda, la desviación estándar y gráficos como histogramas, gráficos de barras y diagramas de dispersión.

Mientras tanto, la estadística inferencial busca hacer suposiciones sobre una población a partir de una muestra. Su objetivo es tomar decisiones o hacer afirmaciones sobre propiedades desconocidas de la población a partir de la información extraída de la muestra.

Definiciones básicas de la estadística inferencial

Revisemos algunos conceptos básicos de la estadística inferencial:

Población

La **población** se refiere al conjunto completo de elementos o individuos que poseen una característica particular y que están sujetos a estudio o análisis. Es el grupo más amplio sobre el cual se desea hacer una inferencia estadística.

Por ejemplo, si estás interesado en estudiar las alturas de todos los estudiantes en una escuela, la población sería el conjunto total de estudiantes en esa escuela.

Muestra

Una muestra es un subconjunto seleccionado de la población con el propósito de estudiarla y obtener información sobre la población en su conjunto. Las muestras son más manejables en términos de recolección y análisis de datos, en comparación con la población completa. Una muestra debe ser representativa de la población para que las conclusiones basadas en ella sean válidas y generalizables.

Por ejemplo, si tomas una muestra de 100 estudiantes de una escuela para estudiar sus alturas, esos 100 estudiantes conformarán tu muestra.

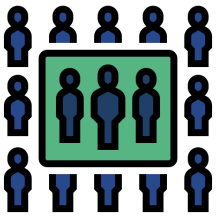

Figura 01: Población y muestra
Fuente: flatIcon

Parámetros y estimadores estadísticos

Los **parámetros estadísticos** son valores numéricos que describen características específicas de una **población.** Estos valores son generalmente desconocidos y son el enfoque de la inferencia estadística. Los parámetros son valores fijos y constantes para una población dada.

Cuando seleccionamos una muestra podemos hacer algunos cálculos que llamamos **estimadores**, ya que confiamos en que nos permiten "aproximarnos" a los valores de los parámetros. Utilizaremos la siguiente notación:

Nombre	Parámetro estadístico	Estimador muestral
Media	μ	$\bar{x}$
Varianza	σ^2	S^2
Desviación	σ	S
Proporción	**p**	$\hat{p}$

Nota: proporción se relaciona con la proporción de la población o la muestra que presenta una característica determinada.

Estadísticos de prueba

Cuando vimos la Ley de los grandes números y el teorema del límite central pudimos constatar que una variable aleatoria puede seguir una distribución cualquiera, pero sus medias muestrales siguen una distribución normal.

Así, podemos definir un **estadístico de prueba** como una variable aleatoria de distribución conocida, que relaciona un parámetro de interés con un estimador de ese parámetro.
Algunos ejemplos fundamentales de esto son:

- Si una variable aleatoria X es tal que $X \sim N(\mu, \sigma)$, conocemos el valor de σ y $\{x_1, x_2, ..., x_n\}$ es una muestra aleatoria de X, tenemos que:

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

- Si una variable aleatoria X es tal que $X \sim N(\mu, \sigma)$, $\{x_1, x_2, ..., x_n\}$ es una muestra aleatoria de X, **no** conocemos el valor de σ (pero lo estimamos como S) tenemos que:

$$\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

  t_{n-1} corresponde a una distribución llamada t de student con n - 1 grados de libertad. No nos adentraremos mucho en su explicación teórica, pero la veremos en uso cuando apliquemos algo de código en Python.
- Si una variable aleatoria X tiene distribución desconocida, $\{x_1, x_2, ..., x_n\}$ es una muestra aleatoria de X, **no** conocemos el valor de σ (pero lo estimamos como S) tenemos que:

$$\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} \approx N(0,1)$$

  Notemos que en este caso se trata de una aproximación.
- Si una variable aleatoria X tiene distribución Bernoulli, con probabilidad p, $\{x_1, x_2, ..., x_n\}$ es una muestra aleatoria de X con n>30, tenemos que:

$$\frac{\hat{p}-p}{\sqrt{p(1-p)}} \approx N(0,1)$$

Error Muestral

El error muestral se refiere a la discrepancia o diferencia entre los resultados obtenidos de una muestra de datos y los resultados que se habrían obtenido si se hubiera recopilado y analizado toda la población en lugar de solo una muestra. En otras palabras, el error muestral es una medida de cuánto los resultados de la muestra pueden variar con respecto a los resultados que se obtendrían si se tuviera acceso a todos los datos de la población completa.

$$\text{Error Muestral} = \bar{x} - \mu$$

El error muestral es una consecuencia natural de trabajar con muestras en lugar de poblaciones completas. Siempre existe cierta incertidumbre debido a que los datos de la muestra son solo una parte de la imágen completa. Sin embargo, mediante el uso adecuado de técnicas estadísticas, es posible estimar la magnitud del error muestral y cuantificar la confianza en las conclusiones que se extraen de la muestra.

Por ejemplo, la talla promedio que surge de la muestra no necesariamente coincidirá con la de la población total, aunque si la muestra no es sesgada el valor ha de ser cercano. Esa diferencia entre el valor medio de la muestra y el de la población total es el error muestral.

Pruebas de Hipótesis

Las pruebas de hipótesis son procedimientos estadísticos que se utilizan para tomar decisiones basadas en la evidencia de una muestra de datos. Estas pruebas permiten evaluar si una afirmación o suposición sobre una población es plausible o no, utilizando información recopilada de una muestra representativa de esa población. El proceso implica comparar los resultados observados en la muestra con lo que se esperaría bajo una hipótesis específica.

Hipótesis Nula y alternativa

La hipótesis nula Ho es una suposición inicial o afirmación que se formula en una prueba de hipótesis con el propósito de ser evaluada y probada. Representa la idea de que no hay efecto, no hay diferencia, o no hay cambio en un parámetro poblacional específico o en una relación que se está investigando. La hipótesis nula es esencial para establecer un marco para la comparación con los resultados observados en la muestra.

En términos más simples, la hipótesis nula es una afirmación de "ningún efecto" o "ninguna diferencia" que se somete a prueba estadística para determinar si los datos muestrales

proporcionan evidencia suficiente para rechazarla en favor de la hipótesis alternativa (H1 o Ha).

Ejemplos de hipótesis nulas en diferentes contextos:

- **Ciencias Médicas:** Ho : El nuevo medicamento no tiene ningún efecto sobre la tasa de recuperación.
- **Economía:** Ho : La política fiscal actual no afecta el crecimiento económico.
- **Educación:** Ho : No hay diferencia en el rendimiento académico entre los dos métodos de enseñanza.
- **Tecnología:** Ho : El nuevo software no mejora la eficiencia del proceso.
- **Marketing:** Ho : No hay cambio en la preferencia del consumidor después de la campaña publicitaria.

La hipótesis alternativa (denotada como H1 o Ha) es una afirmación contraria a la hipótesis nula (Ho) en una prueba de hipótesis.

Representa la idea de que hay algún tipo de efecto, diferencia o cambio en un parámetro poblacional específico o en una relación que se está investigando. En otras palabras, la hipótesis alternativa sugiere que algo está sucediendo y que la situación no es igual a la suposición de la hipótesis nula. La hipótesis alternativa es crucial en una prueba de hipótesis, ya que contrasta con la hipótesis nula y establece la dirección en la que se está buscando evidencia.

Dependiendo del tipo de hipótesis alternativa, una prueba de hipótesis puede ser de cola (unilateral) o de dos colas (bilateral).

Existen tres tipos comunes de hipótesis alternativas:

- **Hipótesis Alternativa Unilateral (Cola):** En este caso, la hipótesis alternativa se formula de manera que sugiere un efecto o diferencia en una dirección específica. Puede ser "mayor que" o "menor que" la suposición de la hipótesis nula. Por ejemplo: el nuevo tratamiento reduce la tasa de error en comparación con el tratamiento existente.
- **Hipótesis Alternativa Bilateral (Dos Colas):** En esta situación, la hipótesis alternativa sugiere que hay un efecto o diferencia, pero no especifica una dirección en particular. Está abierta a cambios en ambas direcciones. Por ejemplo: la media del nuevo método es diferente de la media del método existente.
- **Hipótesis Alternativa de No Igualdad (Dos Colas):** Similar a la hipótesis bilateral, esta versión específica de la hipótesis alternativa sugiere que hay una diferencia, pero no especifica en qué dirección. Sin embargo, se utiliza con pruebas que se enfocan en la igualdad de parámetros. Por ejemplo: la tasa de efectividad de un método no es igual a 0,10.

Tipos de errores

El error tipo I y el error tipo II son dos tipos de errores que pueden ocurrir en las pruebas de hipótesis. Estos errores están asociados con las decisiones que se toman al rechazar o no rechazar la hipótesis nula.

Error Tipo I (Error alfa, Falso Positivo)

El error tipo I ocurre cuando se rechaza incorrectamente la hipótesis nula (Ho) cuando en realidad es verdadera. En otras palabras, se concluye que hay un efecto, una diferencia o un cambio en la población cuando en realidad no lo hay. El error tipo I se relaciona con el nivel de significancia (α) establecido previamente.

El nivel de significancia, o probabilidad de error de tipo 1 (α) corresponde a la probabilidad de error de tipo I. Se trata siempre de minimizarlo ya que puede conducir a conclusiones incorrectas y decisiones basadas en evidencia falsa. La elección de un nivel de significancia más bajo disminuye la probabilidad de cometer un error tipo I, pero aumenta la posibilidad de cometer un error tipo II.

Error Tipo II (Error Beta)

El error tipo II ocurre cuando se acepta incorrectamente la hipótesis nula (Ho) cuando en realidad es falsa. En otras palabras, se concluye que no hay un efecto, una diferencia o un cambio en la población cuando en realidad sí lo hay. El error tipo II se relaciona con la **potencia estadística**, que es la capacidad de detectar un efecto real.

La potencia estadística es la probabilidad de rechazar la hipótesis nula cuando la hipótesis alternativa es verdadera. Una mayor potencia significa una menor probabilidad de cometer un error tipo II. Para aumentar la potencia, se puede aumentar el tamaño de muestra, utilizar pruebas más sensibles o modificar el nivel de significancia, pero hacerlo puede aumentar el riesgo de cometer un error tipo I.

		Situación real	
		Hay relación	No hay relación
Prueba	Determina relación (H1)	Verdadero positivo	Falso positivo Error I
	Determina relación (H1)	Falso Negativo Error II	Verdadero negativo

Nivel de significancia y región crítica

El nivel de significancia y la región crítica son conceptos importantes en las pruebas de hipótesis, ya que influyen en cómo se toman decisiones basadas en los resultados de la prueba. Juntos, ayudan a determinar si hay evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alternativa.

El nivel de significancia (α) es la probabilidad máxima que se está dispuesto a aceptar de cometer un error de tipo I en una prueba de hipótesis. Un error de tipo I ocurre cuando se rechaza incorrectamente la hipótesis nula cuando es verdadera. En otras palabras, es la probabilidad de concluir que hay un efecto o diferencia cuando en realidad no lo hay.

El nivel de significancia es un valor establecido previamente por el investigador antes de realizar la prueba. Un valor comúnmente utilizado es 0.05 (5%), pero también se pueden usar otros valores según el contexto. Un nivel de significancia más bajo implica que el investigador es más conservador y requiere evidencia más sólida para rechazar la hipótesis nula.

La región crítica es el rango de valores de la estadística de prueba en el cual se rechaza la hipótesis nula en una prueba de hipótesis. La ubicación de la región crítica depende del tipo de prueba (de cola o de dos colas) y del nivel de significancia seleccionado.

- **En una prueba de cola (unilateral),** la región crítica se encuentra en uno de los extremos de la distribución de la estadística de prueba.
- **En una prueba de dos colas (bilateral),** la región crítica se divide en dos partes, una en cada extremo de la distribución de la estadística de prueba.

Cuando el estadístico de prueba cae dentro de la región crítica se rechaza la hipótesis nula, lo que sugiere que hay suficiente evidencia para respaldar la hipótesis alternativa.

Inferencia estadística

Las distribuciones de probabilidad son fundamentales para la inferencia estadística. Cuando se trabaja con muestras, es común asumir que las observaciones siguen una cierta distribución de probabilidad (como la distribución normal) para poder aplicar métodos estadísticos.

Por ejemplo, al realizar una prueba de hipótesis, se compara una estadística de prueba calculada a partir de la muestra con una distribución de probabilidad conocida bajo la hipótesis nula. También se utilizan distribuciones de probabilidad para calcular intervalos de confianza y estimar la probabilidad de ciertos resultados.

La estimación puntual y la estimación por intervalos son dos enfoques utilizados en estadística para estimar valores desconocidos de una población utilizando datos muestrales. Ambos métodos tienen sus propias ventajas y desventajas, y su elección depende del grado de precisión y confianza deseado en la estimación.

Estimación Puntual

La estimación puntual implica proporcionar un único valor como la mejor suposición o "mejor conjetura" para el parámetro desconocido de la población. En otras palabras, se proporciona una estimación única para el valor exacto del parámetro que se está tratando de inferir.

Ejemplo: Si estás tratando de estimar la media poblacional de edades de una población usando una muestra, la estimación puntual sería simplemente la media muestral de las edades.

Estimación por Intervalos (Intervalos de Confianza)

La estimación por intervalos, también conocida como intervalos de confianza, proporciona un rango de valores en el cual es probable que se encuentre el verdadero valor del parámetro poblacional. Este rango se construye de manera que haya una alta probabilidad (generalmente especificada por un nivel de confianza) de que el valor del parámetro esté contenido dentro del intervalo.

Ejemplo: Si calculas un intervalo de confianza del 95% para la media de las edades de una población, este intervalo indicará que con un 95% de confianza, el valor verdadero de la media estará dentro del intervalo.

Inferencia sobre la media de la población

Nuestra idea en este caso es poder afirmar con un cierto grado de certeza que, si contamos con la media muestral $\bar{x}$, la media poblacional μ debe estar en un intervalo determinado $[\bar{x} - E, \bar{x} + E]$.

Considerando entonces una muestra de tamaño n, podemos calcular de ella la media muestral $\bar{x}$ y la desviación estándar S. Si definimos un nivel de significancia α nos interesarán los valores críticos para los cuales nuestra media muestral quedaría fuera de la zona crítica, pensando en ambos extremos (dos colas). A este valor crítico (determinable por tabla o software) le llamamos $z_{\frac{\alpha}{2}}$ y calculamos el margen de error $E = z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$. Con esto obtenemos nuestro intervalo de confianza $[\bar{x} - E, \bar{x} + E]$

Podemos abordar, también, el problema inverso: si queremos determinar el tamaño de nuestra muestra, necesario para nuestro intervalo de confianza con nivel de significancia y error dados, podemos aplicar la siguiente fórmula:

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2$$

Inferencia sobre la proporción

Para estimar la proporción p de la población que presenta una característica determinada, podemos utilizar la proporción muestral $\widehat{p}$ con el intervalo de confianza

$$\left(\widehat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \right)$$

Y también podemos determinar el tamaño de la muestra necesaria, con la fórmula

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \widehat{p} \cdot (1-\widehat{p})}{E} \right)^2$$

Prueba de hipótesis para muestras independientes

Las pruebas de hipótesis para muestras independientes son métodos estadísticos utilizados para comparar dos o más grupos o muestras que son independientes entre sí, lo que significa que los datos en un grupo no están relacionados ni emparejados con los datos en el otro grupo. Estas pruebas se utilizan para determinar si hay evidencia estadística suficiente para afirmar que hay diferencias significativas entre las poblaciones o grupos representados por las muestras.

En un sentido práctico, nos interesa averiguar si las diferencias muestrales pueden deberse al azar, o es más probable que se deban a una diferencia efectiva entre los grupos.

Para verificar estas hipótesis, algunas de las pruebas más comunes para muestras independientes son:

- **Prueba t independiente:** La prueba t independiente se utiliza para comparar las medias de dos grupos independientes y determinar si existe una diferencia significativa entre ellas. Por ejemplo, puede utilizarse para comparar las puntuaciones promedio de dos grupos de estudiantes que tomaron diferentes métodos de enseñanza.

- **ANOVA (Análisis de Varianza):** El ANOVA se utiliza cuando se comparan las medias de tres o más grupos independientes. Se puede utilizar para determinar si hay diferencias significativas entre los grupos y, si es así, realizar pruebas post hoc para identificar cuáles de los grupos difieren entre sí.

El ANOVA se divide en varios tipos, dependiendo del diseño experimental y la cantidad de factores involucrados:

- **ANOVA de un factor (One-Way ANOVA):** Esta es la forma más básica de ANOVA y se utiliza cuando hay un solo factor independiente o variable categórica que divide a los datos en diferentes grupos. Se compara la variabilidad dentro de los grupos con la variabilidad entre los grupos.
- **ANOVA de dos factores (Two-Way ANOVA):** En este caso, hay dos factores independientes o variables categóricas que se utilizan para dividir los datos en diferentes grupos. El análisis evalúa la influencia de ambos factores en la variable dependiente y sus interacciones.
- **ANOVA de medidas repetidas (Repeated Measures ANOVA):** Se utiliza cuando las mismas unidades experimentales se someten a diferentes condiciones en varios momentos. Evalúa los efectos de una variable independiente en múltiples mediciones repetidas.
- **ANOVA de efectos mixtos (Mixed-Design ANOVA):** Combina elementos del ANOVA de un factor y del ANOVA de medidas repetidas. Se aplica cuando hay un factor entre sujetos y un factor dentro de sujetos.

El ANOVA produce una estadística F y su correspondiente valor p. Si el valor p es menor que el nivel de significancia elegido (α), se rechaza la hipótesis nula y se concluye que al menos un grupo tiene una media significativamente diferente. En caso contrario, no hay suficiente evidencia para concluir que hay diferencias significativas.

El análisis de varianza es especialmente útil cuando se trabaja con múltiples grupos y se desea comparar sus medias de manera eficiente en lugar de realizar múltiples pruebas de hipótesis individuales. También es útil para detectar patrones de variación y relaciones entre variables en un contexto experimental.

- **Prueba de Chi-cuadrado (Chi-squared):** La prueba de Chi-cuadrado se utiliza para comparar la distribución de frecuencias de dos o más grupos independientes en función de una variable categórica. Se usa comúnmente en estudios de asociación entre variables categóricas.

- **Prueba de Mann-Whitney U:** Esta prueba se utiliza cuando los datos no cumplen con los supuestos de normalidad requeridos por la prueba t independiente. Se utiliza para comparar dos grupos independientes en términos de sus distribuciones de rangos.

- **Prueba de Kruskal-Wallis:** Similar a la prueba de Mann-Whitney U, pero aplicable a tres o más grupos independientes. Se utiliza cuando se desea comparar grupos en términos de sus distribuciones de rangos y no se cumplen los supuestos de normalidad.

Estas son solo algunas de las pruebas de hipótesis para muestras independientes disponibles en estadísticas. La elección de la prueba dependerá de la naturaleza de los datos y de la pregunta de investigación que estés tratando de responder. Cada prueba tiene sus propios supuestos y condiciones de uso, por lo que es importante seleccionar la prueba adecuada en función de tus datos y objetivos de investigación.

Preguntas de proceso

Reflexiona:

- ¿En qué áreas habías visto o escuchado aplicarse los conceptos vistos en esta guía?
- ¿Qué dificultades de interpretación pueden tener los conceptos vistos? ¿Los has visto mal aplicados?