Pruebas de hipótesis

Cristian Guarnizo-Lemus
cristianguarnizo@itm.edu.co

Maestria en Automatización y Control Industrial

Definiciónes

- Por medio de la pruebas de hipótesis podemos interpretar o sacar concluciones sobre una población usando datos muestreados.
- Una prueba de hipótesis evalúa dos proposiciones mutuamente excluyentes sobre una problación para determinar cual proposición esta mejor soportada por los datos muestreados.

Parámetro y estadístico

Un parámetro es una descripción resumida de una característica fija o medida de una población objetivo. Un parámetro denota el verdadero valor que podría ser obtenido sí un censo en ves de una muestra fueran tomados.

Ejemplos: Media (μ), Varianza (σ^2), Desviación estándar (σ).

Población

- La población es una colección de objetos que queremos estudia/probar. La colección puede ser Ciudades, Empresas, personas, estudiantes, etc. Depende del estudio a mano.
- En el mundo real es una tarea difícil obtener la información completa de una población. Entonces, extraemos una muestra de esa población y derivamos las mismas medidas estadísticas mencionadas antes. Y estas medidas son llamadas Estadísticas muestrales.

Estadístico muestral

Un estadístico es una descripción resumida de una característica o medida de una muestra. El estadístico muestral es utilizado como una estimación del parámetro de la población.

Ejemplo: Media muestral ($\bar{x}$), Varianza muestral (S^2 o σ_n^2).

Distribución muestral

Una distribución muestral es una distribución de probabilidad de un estadístico obtenido a través de un numero grande de muestras tomadas de una población específica.

En general, la media de una distribución muestral sera aproximadamente equivalente a la media de la población, esto es, $\mathbb{E}(\bar{x}) = \mu$.

Standard Error (SE)

El error estándar es muy similar a la desviación estándar. Ambos son medidas de dispersión. Mientras el error estándar emplea estadísticos (datos muestrados), la desviación estándar usa parámetros (datos poblacionales).

El error estándar indica que tan lejos su estádistico muestral se desvía del valor actual de la población. Entre mas grande su tamaño muestral, mas cercano estará su estadístico muestral del poblacional.

Hipótesis nula (H_0)

Una declaración en la cual no se espera diferencia o efecto. Sí la hipótesis nula no es rechazada, no habrán cambios.

Hipótesis alterna (H_1)

Una declaración que se espera alguna diferencia o efecto. Aceptar la hipótesis alternativa nos lleva a cambios en opiniones o acciones.

Pruebas de una sola cola

- Es una prueba de hipótesis estadística en la cual el área critica de una distribución es a un lado, entonces es también mas grande que o mas pequeño que un cierto valor, pero no ambos.
- Sí la muestra que esta siendo probada cae en el lado crítico, la hipótesis alternativa será aceptada en vez de la hipótesis nula.
- La región critica son lo valores que corresponden al rechazo de la hipótesis nula a un nivel de probabilidad dado.

Pruebas de dos colas

- Es un método en el cual el área critica de una distribución es dos lados y prueba cuando una muestra es mayor que o menor que cierto rango de valores.
- Sí la muestra que esta siendo probada cae en alguno de los lados críticos, la hipótesis alternativa será aceptada en vez de la hipótesis nula
- Típicamente, las prueba de dos colas son utilizadas para determinar un nivel de significancia del 5%, lo que significa que cada lado de la distribución es cortado al 2.5%.

Ejemplo: Pruebas con colas

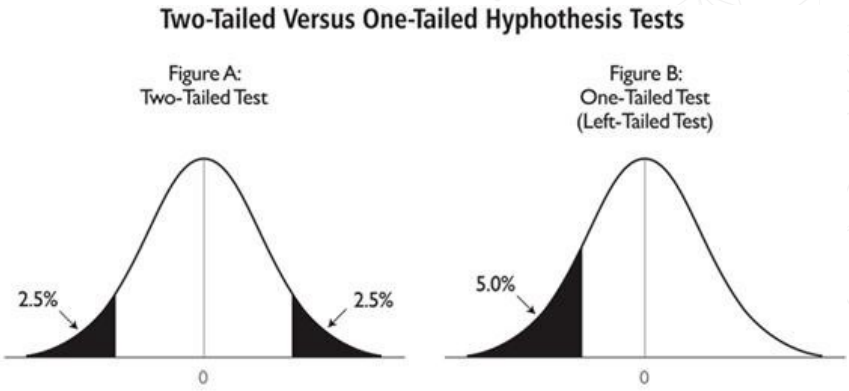
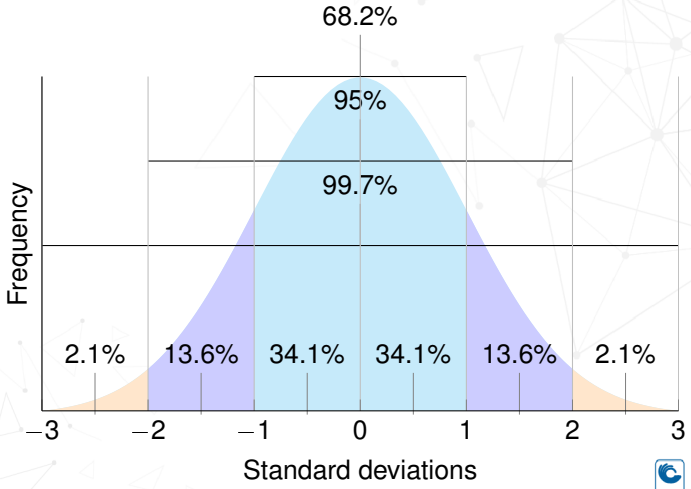


Figure: Pruebas de dos colas, y pruebas de una cola.

Ejemplo: Pruebas con colas



Pruebas estadísticas

- La prueba estadística mide que tan cerca está la muestra de la hipótesis nula.
- Una prueba estadística contiene información acerca de los datos que es relevante para decidir si se rechaza la hipótesis nula o no.
- Las pruebas mas usadas son Z-test, t-test, ANOVA, Chi-square.

Error tipo I

- Ocurren cuando los resultados muestrales, llevan al rechazo de la hipótesis nula cuando de hecho es verdadera.
- Los errores tipo I son equivalentes a falsos positivos.
- Podemos controlar estos errores a partir del valor α que representa el nivel de significancia que seleccionamos.

Error tipo II

- Ocurren cuando basados en los resultados muestrales, la hipótesis nula no es rechazada cuando de hecho es falsa.
- Los errores tipo II son equivalentes a falsos negativos.
- Podemos controlar estos errores a partir del valor α que representa el nivel de significancia que seleccionamos.

Tipos de errores

		Decision based on the data	
		Reject H_0	Do not reject H_0
Truth (unknown)	H_0 is true	Type I error (false rejection) Probability $=\alpha$	Correct decision Probability $=1-\alpha$
	H_0 is false	Correct decision Probability $=1-\beta$	Type II error (false acceptance) Probability $=\beta$

Nivel de significancia

- La probabilidad de realizar un error tipo I y esta denotada por α.
- α es la probabilidad máxima que tenemos un error tipo I.
- Para 95% de nivel de confidencia (confianza), el valor α es 0.05. Esto significa que existe un 5% de probabilidad de rechazar la hipótesis nula.

P-values

- Evaluan que tan bien los datos muestreados soportan el argumento que la hipótesis nula es verdadera.
- Mide que tan compatible son sus datos con la hipótesis nula.
- Dada que la hipótesis nula es verdadera, el valor *p* es la probabilidad de obtener un resultado como o mas extremo que el resultado muestral de manera aleatoria.
- Valores altos de *p*: Sus datos son verosímiles con un verdadero nulo.
- Valores bajos de *p*: Sus datos son no verosímiles con un verdadero nulo.

Pruebas de hipótesis para la media

- Se toma una decisión entre 2 hipótesis sobre la base del valor de una prueba estadística (test), la cual es función de las observaciones de una variable aleatoria.

$$\text{Prueba Estadistica} = \frac{\text{Estadistico relevante} - \text{Parametro hipotetico}}{\text{Error estandar de los estadisticos relevantes}}$$

- Al considerar la media en una población a partir de una muestra de tamaño n:

$$\text{Prueba Estadistica} = \begin{cases} \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} & \text{Sí } \sigma_X \text{ es conocida} \\ \frac{\bar{X}_n - \mu_X}{\sigma_n/\sqrt{n}} & \text{Sí } \sigma_X \text{ es desconocida} \end{cases}$$

Pruebas de hipótesis para la media

- Caso I: Población con distribución Gaussiana y varianza conocida (σ_X^2).
- Caso II: Si el tamaño muestral n es grande, se desconoce σ_X^2, pero se conoce la varianza muestral σ_n^2.
- Caso III: Población con distribución Gaussiana y se desconoce σ_X^2, pero se conoce σ_n^2.

Hipótesis Alternativa H_1	Tipo de Prueba	Condiciones para rechazar H_0 Caso-I	Caso-II	Caso-III
$\mu_X < \mu_0$	Left-tailed	$Z \leq -z_\alpha$	$Z \leq -z_\alpha$	$T \leq -t_\alpha$
$\mu_X > \mu_0$	Right-tailed	$Z \geq z_\alpha$	$Z \geq z_\alpha$	$T \geq t_\alpha$
$\mu_X \neq \mu_0$	Two-tailed	$\lvert Z \rvert \geq z_{\alpha/2}$	$\lvert Z \rvert \geq z_{\alpha/2}$	$\lvert T \rvert \geq t_{\alpha/2}$

Contenido

Z-test

- Un test Z es cualquier prueba estadística para el cual la distribución de la prueba estadística bajo la hipótesis nula puede ser aproximada por una distribución normal.
- Se prueba la media de una distribución de la cual ya se conoce su varianza (σ^2).
- Sí se desconoce la varianza de la población (y entonces se debe estimar de la misma muestra) y el tamaño muestral no es grande ($n < 30$), la prueba Student-t puede ser mas apropiada.

Z-test

- Un test Z es cualquier prueba estadística para el cual la distribución de la prueba estadística bajo la hipótesis nula puede ser aproximada por una distribución normal.
- Se prueba la media de una distribución de la cual ya se conoce su varianza (σ^2).
- Sí se desconoce la varianza de la población (y entonces se debe estimar de la misma muestra) y el tamaño muestral no es grande ($n < 30$), la prueba Student-t puede ser mas apropiada.

Contenido

T-test

- Se emplea para determinar sí existe una diferencia significativa entre la media de dos grupos, los cuales pueden estar relacionados en ciertas características.
- Se usa en datasets donde las variables de interés sigan una distribución normal y se desconozca las varianzas.
- Existen 2 tipos, de una muestra y 2 muestras.

T-test de una muestra

- Determina cuando la media muestral es estadísticamente diferente de la media hipotética o conocida de una población.

T-test de dos muestras

- Compara la media de dos grupos independientes con el fin de determinar sí existe evidencia estadística que las medias de las poblaciones son significativamente diferentes.

Paired sampled t-test

- Conocido tambien como la prueba t de muestra dependiente.
- Prueba si existe una diferencia significativa entre 2 variables relacionadas.

F-test

- El análisis de la varianza o ANOVA es una prueba de inferencia estadística que permite comparar multiples grupos al mismo tiempo.
- A diferencia de las distribuciones T y Z, la distribución F no tiene valores negativos debido que la relación de abajo siempre es positiva.

$$F = \frac{\text{Variabilidad entre grupos}}{\textit{Variabilidadintragrupo}}$$

F-test

- Variabilidad entre grupos:

$$\sum_{i=1}^{K} n_i \left(\bar{Y}_i - \bar{Y}\right)^2 / (K-1)$$

- Variabilidad intra grupos:

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2 / (N-K)$$

One way F-test

- Indica si dos o mas grupos son similares o no, dependiendo de la similaridad en su media y F-score.

Two way F-test

- Es una extensión del método anterior.
- Se emplea cuando tenemos 2 variables independientes y mas de 2 grupos.
- No dice cual variable es dominante.

Chi-square test

- Se emplea cuando se tienen 2 variables categóricas de una sola población.
- Determina si existe un asociación significativa entre ambas variables.

Referencias

- Ali Grami. “ Probability, Random Variables, Statistics, and Random Processes: Fundamentals & Applications”, 2019.
- Y. Agrawal. “ Hypothesis testing in Machine learning using Python ”, 2019. `https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce`
- Ed Bullen’s repository: `https://github.com/edbullen/Hypothesis`