# UNIDAD 4. INFERENCIA ESTADÍSTICA

## 1. Estimación por Intervalos

Se puede establecer un intervalo de estimación para la media $\mu$, si la muestra se selecciona de una **población normal** o si **n es grande** ($n \geq 30$), considerando la distribución muestral de $\overline{X}$.

• Intervalo de Confianza para $\mu$, si se conoce $\sigma$.

Se selecciona una muestra aleatoria de tamaño n de población cuya varianza $\sigma^2$ se conoce y la media $\overline{x}$ se calcular para obtener el siguiente intervalo de confianza del (1-$\alpha$)% para $\mu$:

$$\overline{x} - Z_{\alpha/2} \, \sigma/\sqrt{n} < \mu < \overline{x} + Z_{\alpha/2} \, \sigma/\sqrt{n}$$

donde $Z_{\alpha/2}$ es el valor de Z a la derecha del cual se tiene un área de $\alpha/2$.

**Error**: si $\mu$ es realmente el valor del centro del intervalo, entonces $\overline{x}$ estima a $\mu$ sin error. La mayor parte de las veces, sin embargo, $\overline{x}$ no será exactamente igual a $\mu$ y la estimación puntual es errónea. El tamaño de este error (e) será el valor absoluto de la diferencia entre $\overline{x}$ y $\mu$ y se puede tener una confianza del (1-$\alpha$)% que esta diferencia no excederá de $Z_{\alpha/2} \, \sigma/\sqrt{n}$.

**Tamaño de la muestra**: si se utiliza a $\overline{x}$ como una estimación de $\mu$, se puede tener una confianza del (1-$\alpha$)% de que el error no excederá una cantidad específica e cuando el tamaño de la muestra es:

$$n = \left( \frac{Z_{\alpha/2}\sigma}{e} \right)^2$$ redondeada al entero superior más cercano

• Intervalo de Confianza para $\mu$, si no se conoce $\sigma$.

Si $\overline{x}$ y s son la media y la desviación estándar de una muestra aleatoria tomada de una población normal con varianza desconocida $\sigma^2$, un intervalo de confianza del (1-$\alpha$)% para $\mu$ es:

$$\overline{x} - t_{\alpha/2} \, s/\sqrt{n} < \mu < \overline{x} + t_{\alpha/2} \, s/\sqrt{n}$$

donde $t_{\alpha/2}$ es el valor de la Distribución t de Student con n-1 grados de libertad, que deja un área de $\alpha/2$ a la derecha.

• Intervalo de Confianza para $\mu$, de muestra grande.

Incluso cuando no se puede suponer la normalidad de la población, con $\sigma$ desconocida y $n \geq 30$, s podría reemplazar a $\sigma$ y se podría utilizar el intervalo de confianza:

$$\overline{x} - Z_{\alpha/2} \, s/\sqrt{n} < \mu < \overline{x} + Z_{\alpha/2} \, s/\sqrt{n}$$

donde $Z_{\alpha/2}$ es el valor de Z a la derecha del cual se tiene un área de $\alpha/2$.

Estimación de la diferencia entre dos medias

Si se tienen dos poblaciones con medias $\mu_1$ y $\mu_2$ y varianzas $\sigma_1^2$ y $\sigma_2^2$ respectivamente, el estimador puntual de la diferencia entre $\mu_1$ y $\mu_2$ lo da el estadístico $\overline{X}_1 - \overline{X}_2$.

• Intervalo de Confianza para $\mu_1 - \mu_2$ conociendo $\sigma_1^2$ y $\sigma_2^2$:

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes de tamaños $n_1$ y $n_2$ de poblaciones con varianzas conocidas $\sigma_1^2$ y $\sigma_2^2$, respectivamente, un intervalo de confianza del (1-$\alpha$)% para $\mu_1 - \mu_2$ es:

$$(\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

donde $Z_{\alpha/2}$ es el valor de Z que tiene a su derecha un área bajo la curva igual a $\alpha/2$.

• Intervalo de Confianza para $\mu_1 - \mu_2$, con $\sigma_1^2=\sigma_2^2$ pero desconocidas

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes de tamaños $n_1$ y $n_2$, respectivamente, de poblaciones aproximadamente normales con varianzas iguales pero desconocidas, un intervalo de confianza del (1-$\alpha$)% para $\mu_1 - \mu_2$ es:

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

donde $S_p$ es la estimación común de la desviación estándar poblacional y $t_{\alpha/2}$ es el valor de la distribución t con $v = n_1 + n_2 - 2$ grados de libertad, que tiene a su derecha un área bajo la curva igual a $\alpha/2$.

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

• Intervalo de Confianza para $\mu_1 - \mu_2$ con $\sigma_1^2 \neq \sigma_2^2$ pero desconocidas

Si $\bar{x}_1$ y $s_1^2$ y $\bar{x}_2$ y $s_2^2$ son las medias y varianzas de muestras aleatorias pequeñas e independientes de tamaños $n_1$ y $n_2$, respectivamente, de poblaciones aproximadamente normales con varianzas diferentes y desconocidas, un intervalo de confianza del (1-$\alpha$)% para $\mu_1 - \mu_2$ es:

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

donde $t_{\alpha/2}$ es el valor de la distribución t con v grados de libertad, que tiene a su derecha un área bajo la curva igual a $\alpha/2$.

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left[(s_1^2/n_1)^2/(n_1 - 1)\right] + \left[(s_2^2/n_2)^2/(n_2 - 1)\right]}$$

Estimación de la proporción

Un estimador puntual de la proporción p en un experimento binomial está dado por el estadístico $\hat{P} = X / n$, donde X representa el número de éxitos en n intentos. Por lo tanto, la proporción muestral $\hat{p} = x / n$ se utiliza como estimación puntual del parámetro p.
Si no se espera que la proporción desconocida p se acerque demasiado a cero o a uno, se puede establecer un intervalo de confianza para p considerando la distribución muestral de $\hat{p}$.

- Intervalo de Confianza para p de una muestra grande

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, y $\hat{q} = 1 - \hat{p}$, un intervalo de confianza aproximado de (1-α)% para el parámetro binomial p es:

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

donde $Z_{\alpha/2}$ es el valor de Z que tiene a su derecha un área bajo la curva igual a α/2.

Cuando n es pequeño y se cree que la proporción desconocida p se acerca a 0 o a 1, este intervalo no es confiable y, por lo tanto, no debe utilizarse. Para estar seguro se requiere que ambos n $\hat{p}$ y n $\hat{q}$ sean mayores o iguales a 5.

Error: Si $\hat{p}$ se utiliza como una estimación puntual de p, se puede tener una confianza del (1-α)% de que el error no excederá de $Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

Tamaño de n: Si $\hat{p}$ se utiliza como una estimación puntual de p, se puede tener una confianza del (1-α)% de que el error será menor que una cantidad específica e cuando el tamaño de la muestra sea aproximadamente:

$$n = \frac{Z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}$$ Cuando no se conoce una estimación de p: $$n = \frac{Z_{\alpha/2}^2}{4e^2}$$

Estimación de la diferencia entre dos proporciones

- Intervalo de Confianza para $p_1 - p_2$ de una muestra grande

Si $\hat{p}_1 \ y \ \hat{p}_2$ son las proporciones de éxito en muestras aleatorias de tamaños $n_1$ y $n_2$ de dos poblaciones binomiales, un intervalo de confianza del (1-α)% para la diferencia entre los dos parámetros binomiales, $p_1$-$p_2$, es:

$$(\hat{p}_1 - \hat{p}_2) - Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

donde $Z_{\alpha/2}$ es el valor de Z que tiene a su derecha un área bajo la curva igual a α/2.
Se requiere que: $n_1 \hat{p}_1$, $n_1 \hat{q}_1$, $n_2 \hat{p}_2$ y $n_2 \hat{q}_2$ sean mayores o iguales a 5.

# 2. Prueba de Hipótesis

## 1. Fundamentos de las pruebas de hipótesis

Las pruebas de hipótesis consisten en la formulación de un procedimiento de decisión basado en datos experimentales, que pueda producir una conclusión acerca del sistema en estudio.

Una **hipótesis estadística** es una afirmación o conjetura acerca de una o más poblaciones.

Se plantean las conjeturas o hipótesis, se utilizan datos experimentales (muestra aleatoria) y se toman decisiones con base en ellos.

Nunca se sabe con absoluta certeza la verdad o falsedad de una hipótesis estadística, a no ser que se examine la población entera, pero como se sabe esto no es posible en la mayoría de los casos prácticos, por ello se toma una muestra aleatoria de la población de interés y se utilizan dichos datos para <u>proporcionar evidencias</u> que confirmen o no la hipótesis. La evidencia de la muestra que es inconsistente con la hipótesis planteada conduce a un rechazo de la misma, mientras que la evidencia que apoya la hipótesis conduce a su aceptación.

La aceptación de una hipótesis implica tan solo que los datos no proporcionan evidencias suficientes para refutarla. Por otro lado, el rechazo implica que la evidencia de la muestra la refuta.

Si el científico se interesa en respaldar con fuerza un argumento, espera llegar al argumento en la forma de rechazo de una hipótesis. Por ejemplo, si se desea mostrar que el tomar café aumenta el riesgo de cáncer, la hipótesis a probar deberá ser "no hay aumento en el riesgo de padecer de cáncer debido a la ingestión de café"; como resultado el argumento se alcanza vía rechazo.

## 2. Formulación de hipótesis. Hipótesis simple y compuesta.

La estructura de la prueba de hipótesis se formulará utilizando el término hipótesis nula. Esto se refiere a cualquier hipótesis que se desee probar y se representa por $H_0$. el rechazo de $H_0$ da como resultado la aceptación de una hipótesis alternativa, que se representa por $H_1$.

Una hipótesis nula referida a un parámetro poblacional siempre será establecida en forma tal que especifique un valor exacto del parámetro, mientras que la hipótesis alternativa admite la posibilidad de varios valores. Por ejemplo: Si $H_0$: $p=0.5$, la hipótesis alternativa puede ser $H_1$: $p< 0.5$, ó $H_1$: $p > 0.5$, ó $H_1$: $p \neq 0.5$.

## 3. Prueba de una hipótesis estadística

La decisión se basa en un **estadístico de prueba**, por ejemplo X: cantidad de éxitos si se está trabajando con una Distribución Binomial. Y en el establecimiento de la **región crítica**, que es la zona donde se rechaza $H_0$.


**4. Tipos de errores que se pueden cometer en las Pruebas de Hipótesis.**

Este procedimiento que se acaba de describir podría conducir a cualquiera de dos conclusiones erróneas, que se resumen en la siguiente tabla:

| | $H_0$ es verdadera | $H_0$ es falsa |
|---|---|---|
| Se acepta $H_0$ | Decisión correcta | Error tipo II ($\beta$) |
| Se rechaza $H_0$ | Error tipo I ($\alpha$) | Decisión correcta |

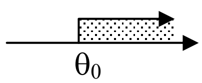
Nivel de significancia ($\alpha$): es la probabilidad de cometer un error tipo I.

Para una muestra de tamaño fijo, la disminución en la probabilidad de cometer un error casi siempre resulta en un incremento en la probabilidad de cometer el otro error. Afortunadamente, la probabilidad de cometer ambos tipos de errores puede reducirse incrementando el tamaño de la muestra.

**5. Tipos de pruebas**

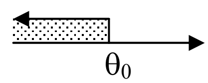
Pruebas de una cola: donde $H_1$ es unilateral, por ejemplo:

$H_0$: $\theta = \theta_0$
$H_1$: $\theta > \theta_0$



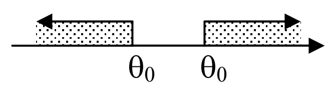

ó

$H_0$: $\theta = \theta_0$
$H_1$: $\theta < \theta_0$




Pruebas de dos colas: donde $H_1$ es bilateral

$H_0$: $\theta = \theta_0$ La región crítica se divide en 2 partes,
$H_1$: $\theta \neq \theta_0$ ya que $H_1$ establece que $\theta < \theta_0$ y $\theta > \theta_0$




6. **Pruebas a estudiar**: en cada prueba el procedimiento es el mismo, lo único que cambia es el estadístico de prueba a utilizar y el establecimiento de la región crítica.

<u>Guías para establecer las hipótesis</u>:

- Leer cuidadosamente el problema para determinar la afirmación que desea probarse. Esta afirmación debe seguir una sola dirección como, por ejemplo, más que, menos que, inferior a, superior a, etc. Entonces en $H_1$ se establecerá el signo de desigualdad (< ó >) correspondiente a la dirección sugerida.
- Si la afirmación sugiere una dirección compuesta (=, $\leq$, $\geq$) como, por ejemplo, al menos, igual o mayor que, no mayor que, etc., entonces esta dirección compuesta completa ($\leq$, $\geq$) se expresa como $H_0$ pero usando solo el signo de igualdad y $H_1$ se expresa en la dirección opuesta.
- Si no se sugiere ninguna dirección en la afirmación, entonces $H_1$ se establece utilizando el signo diferente que $\neq$.

Etapas en la contrastación de hipótesis

1. Establecer la hipótesis nula $H_0$: $\theta = \theta_0$
2. Seleccionar una hipótesis alternativa apropiada $H_1$ de una de las alternativas: $\theta < \theta_0$, $\theta > \theta_0$, o $\theta \neq \theta_0$.
3. Seleccionar un nivel de significancia de tamaño $\alpha$.
4. Seleccionar el estadístico de prueba apropiado y establecer la región crítica.
5. Calcular el valor del estadístico de prueba a partir de los datos muestrales.
6. Decidir: rechazar $H_0$ si el estadístico de prueba se ubica dentro de la región crítica, de otra forma, no rechazar $H_0$.

Pruebas de hipótesis a estudiar:

- Prueba de hipótesis para la media con $\sigma^2$ conocido.
  - Estadístico de prueba: $Z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
- Prueba de hipótesis para la media con $\sigma^2$ desconocido.
  - Estadístico de prueba: $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$ con $\nu = n - 1$ grados de libertad
- Prueba de hipótesis para las proporciones: consiste en probar la hipótesis de que la proporción de éxitos en un experimento Binomial es igual que un valor especificado. Se usa como estadístico de prueba la variable Binomial X con $p=p_0$.

- Si n no es grande: Debido a que X es una v.a. Binomial discreta, es poco probable que pueda determinarse una región crítica cuyo tamaño sea exactamente igual a $\alpha$. Por esta razón es preferible, al tratar con muestras pequeñas, basar las decisiones en el valor denominado **P**.

* Si $H_0$: $p = p_0$, $H_1$: $p < p_0$ $\left\{ \begin{array}{l} \mathbf{P} = P(X \leq x) \text{ cuando } p=p_0 \\ \text{donde x es el número de éxitos en la muestra de tamaño n} \end{array} \right.$

Si $P \leq \alpha$, la prueba es significativa en el nivel $\alpha$ y se rechaza $H_0$ a favor de $H_1$.

* Si $H_0$: $p = p_0$, $H_1$: $p > p_0$ $\left\{ \begin{array}{l} \mathbf{P} = P(X \geq x) \text{ cuando } p=p_0 \\ \text{donde x es el número de éxitos en la muestra de tamaño n} \end{array} \right.$

Si $P \leq \alpha$, la prueba es significativa en el nivel $\alpha$ y se rechaza $H_0$ a favor de $H_1$.

* Si $H_0$: $p = p_0$, $H_1$: $p \neq p_0$ $\left\{ \begin{array}{l} \mathbf{P} = 2P(X \leq x) \text{ cuando } p=p_0 \text{ si } x < np_0 \\ \mathbf{P} = 2P(X \geq x) \text{ cuando } p=p_0 \text{ si } x > np_0 \\ \text{donde x es el número de éxitos en la muestra de tamaño n} \end{array} \right.$

Si $P \leq \alpha$, la prueba es significativa en el nivel $\alpha$ y se rechaza $H_0$ a favor de $H_1$.

Si n es grande y $p_0$ no se acerca a cero o uno, se utiliza la aproximación normal y el valor z para probar que $H_0$: $p = p_0$ está dado por:

$$z = \frac{x - np_0}{\sqrt{np_0 q_0}}$$

el cual es un valor de la variable aleatoria normal estándar Z.

Si se tiene una prueba de dos colas ($H_1$: $p \neq p_0$), la región crítica es $z < -z_{\alpha/2}$ y $z > z_{\alpha/2}$.

Si se tiene una prueba de una sola cola la región crítica es: Si $H_1$: $p < p_0 \rightarrow z < -z_\alpha$

Si $H_1$: $p > p_0 \rightarrow z > z_\alpha$