

Pruebas de Hipótesis

Estadística Aplicada

Lic. Yolanda Segura García

Introduccion

- Muchos Problemas de ingenieria requieren que se tome una decisión entre aceptar o rechazar una proposición.
- Esta proposición recibe el nombre de hipótesis.
- El procedimiento de toma de decisión recibe el nombre de prueba de hipótesis.

- La prueba de hipótesis se considera, como la etapa de análisis de datos de un **experimento comparativo**, por ejemplo comparar la media de una población con un valor específico.
- Nosotros en el presente trabajo trataremos los experimentos comparativos donde intervienen una o dos poblaciones y la finalidad es probar hipótesis con respecto a los parámetros de las poblaciones

- Una hipótesis estadística es una proposición sobre los parámetros de una o más poblaciones
- Por ejemplo: supóngase que se tiene interés en conocer la rapidez de la combustión de cierto combustible. La rapidez de combustión es una variable aleatoria que puede describirse con una distribución de probabilidad. Supóngase que nuestro interés se centra sobre la rapidez de combustión promedio

- el interés recae en decidir si la rapidez de combustión promedio es o no 50 cm/s
- $H_0 : \mu = 50 \text{ cm/s}$
- $H_1 : \mu \neq 50 \text{ cm/s}$
- H_0 se conoce como Hipótesis Nula y H_1 recibe el nombre de hipótesis alternativa.
- muestra de $n=10$ especímenes y que se observa que la rapidez promedio de las 10 muestras es un estimador de la media verdadera de la población μ .

- Un valor de media muestral \bar{x} este próximo a un valor hipotético de $\mu=50$ cm/s es una evidencia de que el verdadero valor de la media μ es realmente 50 cm/s , tal evidencia apoya la hipótesis nula H_0
- una media muestral muy diferente de 50 cm/s constituye una evidencia que apoya la hipótesis alternativa H_1
- Error tipo I se define como el rechazo de la hipótesis nula H_0 cuando esta es verdadera.

Errores asociados a una hipótesis

		Decisión	
		Verdadera	Falsa
Hipótesis	No Rechaza	Correcto $1 - \alpha$	Error II β
	Rechazar	Error I α	Correcto $1 - \beta$

Prueba de Hipótesis

Una prueba estadística consta de cinco elementos:

- 1. Una hipótesis nula, denotada por H_0
- 2. La hipótesis a investigar (también llamada hipótesis alternativa), denotada por H_a .
- 3. Un test estadístico.
- 4. Una región de rechazo, denotada por R.R.
- 5. Una conclusión.

Proceso de prueba de hipótesis

1. Planteamiento de la hipótesis

❖ Hipótesis de nulidad

$$H_0 : \theta = \theta_0$$

❖ Hipótesis alternativa

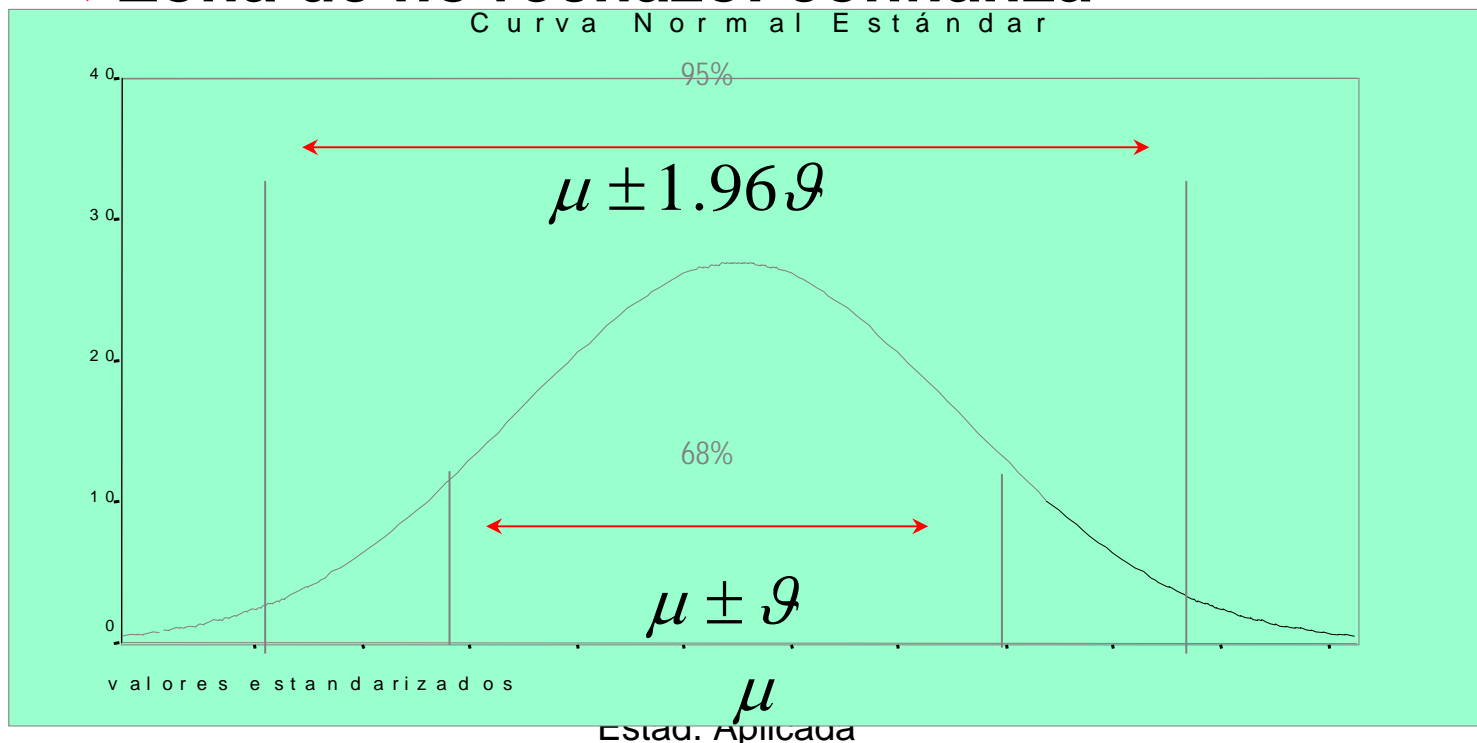
$$H_a : \theta \succ \theta_0$$

$$H_a : \theta \prec \theta_0$$

$$H_a : \theta \neq \theta_0$$

Proceso de prueba de hipótesis

- ❖ Nivel de significancia y nivel de confianza
- ❖ Zona de rechazo: significancia
- ❖ Zona de no rechazo: confianza

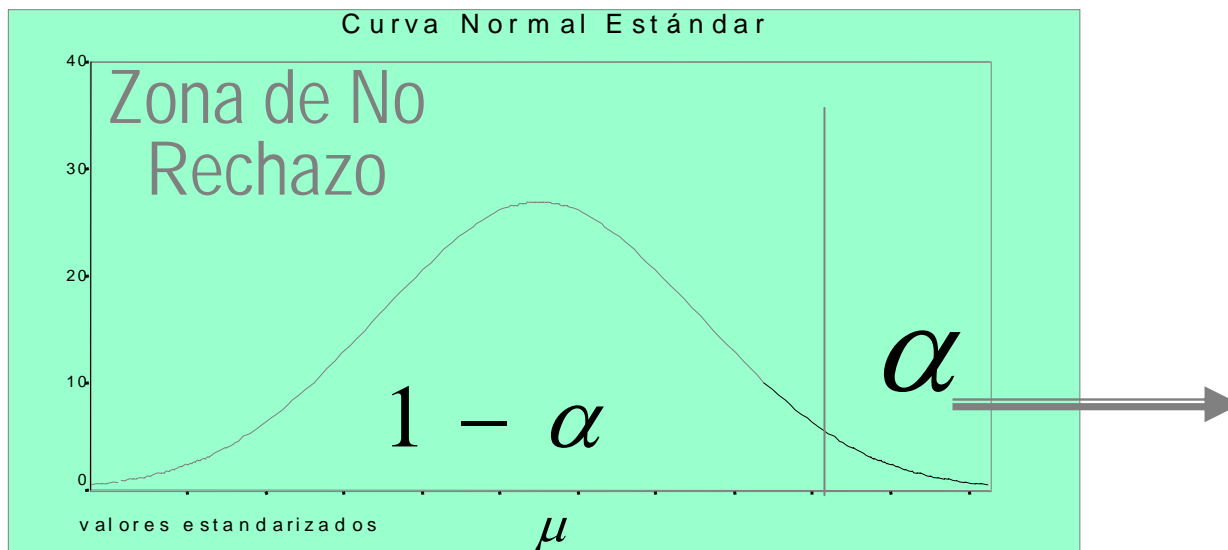


Proceso de prueba de hipótesis

3. Establecimiento del punto de decisión

❖ Zona de rechazo: significancia

❖ Zona de no rechazo: confianza



Estad. Aplicada

Proceso de prueba de hipótesis

4. Selección del método de prueba

- ❖ **Intervalo de confianza**
- ❖ **Estadístico de prueba**

Estadístico y el intervalo de prueba dependen de la hipótesis planteada

- ❖ **Z calculado**
- ❖ **T calculado**
- ❖ **F calculado (Análisis de variancia)**

Proceso de prueba de hipótesis

5. Interpretación de los resultados

- ❖ **Rechazo o no rechazo de la hipótesis nula conforme a la evidencia**
- ❖ **Interpretación en términos del problema**

Proceso de prueba de hipótesis

¿Cómo construir el estadístico o el intervalo de confianza para la prueba?

Prueba de hipótesis

❖ Tipo de hipótesis

- ❖ Univariada

- ❖ Asociación

- ❖ Una población

- ❖ Dos poblaciones

❖ Planteamiento de la hipótesis

- ❖ Hipótesis de nulidad

- ❖ Hipótesis alternativa

Prueba de hipótesis

- ❖ Nivel de significancia
 - ❖ Zona de rechazo
 - ❖ Zona de no rechazo
- ❖ Selección del método
 - ❖ Intervalo de confianza
 - ❖ Estadístico de prueba
- ❖ Intervalo de confianza dependiendo de la hipótesis planteada
 - ❖ Intervalo para p
 - ❖ Intervalo para μ

Prueba de hipótesis

- ❖ Estadístico de prueba dependiendo de la hipótesis planteada
 - ❖ **Z calculado**
 - ❖ **T calculado**
 - ❖ **Chi calculado**
- ❖ Interpretación de los resultados
 - ❖ **Rechazo o no rechazo de la hipótesis nula conforme a la evidencia**
 - ❖ **Interpretación en términos del problema**

Prueba de hipótesis

- ❖ Prueba para la media “ μ ” (varianza conocida y “ n ” mayor de 30)

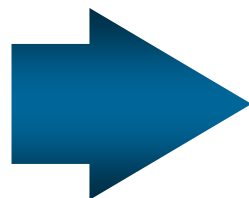
$$H_0 : \mu = \mu_0$$

Nula

$$H_1 : \mu \neq \mu_0$$

$$H_1 : \mu < \mu_0$$

$$H_1 : \mu > \mu_0$$



Alternativa

$$Z_{cal} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

Prueba de hipótesis

❖ Prueba para la media “ μ ” (varianza desconocida y “ n ” menor de 30)

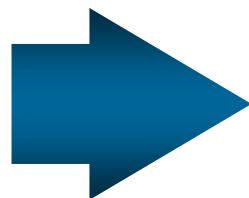
$$H_0 : \mu = \mu_0$$

Nula

$$H_1 : \mu \neq \mu_0$$

$$H_1 : \mu < \mu_0$$

$$H_1 : \mu > \mu_0$$



Alternativa

$$t_{cal} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

Si $n > 30$ T_{cal} tiende a la normal

Prueba de hipótesis

- ❖ Prueba para la media “ μ ” (varianza conocida y “ n ” mayor de 30)

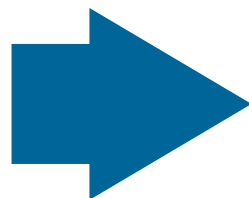
$$H_0 : \mu = \mu_0$$

Nula

$$H_1 : \mu \neq \mu_0$$

$$H_1 : \mu < \mu_0$$

$$H_1 : \mu > \mu_0$$



Alternativa

$$Z_{cal} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

Prueba de hipótesis

- ❖ Prueba para la media “ μ ” (varianza desconocida y “ n ” menor de 30)

$$H_0 : \mu = \mu_0$$

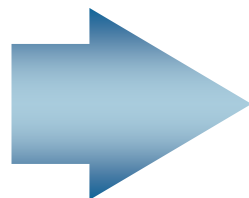
Nula

$$H_1 : \mu \neq \mu_0$$

$$H_1 : \mu < \mu_0$$

$$H_1 : \mu > \mu_0$$

Alternativa



$$t_{cal} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

Estad. Aplicada

Si $n > 30$ T_{cal} tiende a
la normal

Prueba de hipótesis

❖ Prueba para la proporción “P”

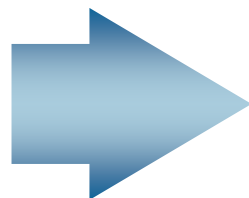
$$H_0 : P = P_0$$

Nula

$$H_1 : P \neq P_0$$

$$H_1 : P < P_0$$

$$H_1 : P > P_0$$



Alternativa

$$z_{cal} = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}$$

Ejemplo : Un problema de decisión puede estar relacionado con la valoración de un predio forestal, y podemos definir como la variable de decisión X , los metros cúbicos aserrales de pino radiata por hectárea. De este modo, el problema de decisión puede ser resuelto al realizar una prueba estadística relacionada con la variable X , al investigar la hipótesis de que el promedio de metros cúbicos aserrales por hectárea sea mayor a 520 m^3 (**hipótesis alternativa**). Para verificar la validez de esta hipótesis, lo hacemos una vez diseñadas las hipótesis nula y alternativa, debemos al contrastarla con otra hipótesis llamada **hipótesis nula**, de que obtener una muestra aleatoria de sectores de una hectárea, este promedio es menor o igual a 520 m^3 , seleccionados en forma aleatoria dentro del predio y calcular la media y la desviación estándar muestral. La decisión para aceptar la hipótesis nula o rechazarla a favor de la alternativa está basada en un **test estadístico** o una regla de decisión que se calcula a partir de la muestra. Una elección lógica como una regla de decisión para μ debería ser X o alguna función de esta media muestral. Si elegimos X como test estadístico, sabemos que su distribución de probabilidades es aproximadamente normal con media $\mu = 520$, bajo la hipótesis nula.

Dócima de hipótesis para μ .

Considerando el caso expuesto en el ejemplo anterior, probaremos la hipótesis de que los metros cúbicos promedio aserrables por hectárea es mayor a 520 metros cúbicos, para $\alpha=0.025$. Utilizando los siguientes datos, se tomó una muestra aleatoria de tamaño $n=36$ y dio como resultado una media muestral, $X=573$ y una desviación estándar, $s=124$.

- .Ho: $\mu=520$
- .Ha: $\mu>520$
- .Test estadístico: X
- .R.R: Para $\alpha=0.025$, rechazamos la hipótesis nula si X está más allá de 1.96 desviaciones estándar de la $\mu=520$

El valor crítico X_c para X se determina usando la distribución normal estándar, en forma análoga a lo realizado en la determinación de intervalos de confianza para μ .

Primero, sabemos que si $Z \sim N(0, 1)$ entonces $P(Z \geq Z_{\alpha}) = \alpha$

Remplazando Z como una función de X obtenemos:

$$P[(X - \mu_0) \sqrt{n} / \sigma \geq Z_{\alpha} \text{ / dado } H_0 \text{ es verdadero}] = \alpha$$

Despejando X , se tiene $P[X \geq Z_{\alpha} \sigma / \sqrt{n} + \mu_0 \text{ / dado } \mu_0 = 520]$

Si σ es desconocido y $n \geq 30$, entonces podemos reemplazarlo por s y la expresión anterior sigue siendo válida.

De la distribución Z obtenemos que para $\alpha = 0.025 \rightarrow Z_{\alpha} = 1.96$

Luego, $X_c = 1.96 \cdot 124/6 + 520 = 560.51$.

Por lo tanto, $X = 573 > 560.51$, se encuentra dentro de la región de rechazo para H_0 . Esto significa que debemos rechazar H_0 .

Resumen para una d cima estad stica sobre μ

$$H_0: \mu = \mu_0$$

$$H_a: 1. \mu > \mu_0$$

$$2. \mu < \mu_0$$

$$3. \mu \neq \mu_0$$

Varianza

Conocida

Desconocida (muestras peque as)

Test:

$$Z = \frac{(X - \mu)\sqrt{n}/\sigma}{\mu)\sqrt{n/s}}$$

$$T = (X -$$

R.R.: Para α dado

1. Rechazar H_0 si

$$Z > Z_{\alpha}$$

$$T > t_{\alpha, (n-1)}$$

2. Rechazar H_0 si

$$Z < -Z_{\alpha}$$

$$T < -t_{\alpha, (n-1)}$$

3. Rechazar H_0 si

$$|Z| > Z_{\alpha/2}$$

$$|T| > t_{\alpha/2, (n-1)}$$

Nota: Para $n \geq 30$ y σ desconocido usted puede reemplazarlo por s y procede con el test Z .

Prueba de hipotesis usando “p-values”

El “P-value” llamado el nivel de significación observado, es el valor de α al cual se rechazaría la hipótesis nula si se usa el valor calculado de la prueba estadística. En la práctica un “P-value” cercano a 0 indica un rechazo de la hipótesis nula. Así un “P-value” menor que .05 indicará que se rechaza la hipótesis nula.

Fórmulas para calcular “P-value”: *Depende de la forma de la hipótesis alterna*

Si $H_a: \mu > \mu_0$, entonces $P\text{-value} = \text{Prob}(Z > Z_{\text{calc}})$.

Si $H_a: \mu < \mu_0$, entonces $P\text{-value} = \text{Prob}(Z < Z_{\text{calc}})$.

Si $H_a: \mu \neq \mu_0$, entonces $P\text{-value} = 2\text{Prob}(Z > |Z_{\text{calc}}|)$.

Los principales paquetes estadísticos, entre ellos MINITAB, dan los “P-values” para la mayoría de las pruebas estadísticas.

Prueba de Hipótesis para la Varianza Poblacional

Asumiendo que la población de donde se extrae la muestra se distribuye normalmente se pueden hacer las siguientes hipótesis acerca de la varianza poblacional:

Caso I

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \sigma^2 < \sigma_0^2$$

Caso II

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \sigma^2 \neq \sigma_0^2$$

Caso III

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \sigma^2 > \sigma_0^2$$

Prueba Estadística:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

con n-1 g.l

Decisión:

Si $\chi_{cal}^2 < \chi_{\alpha}^2$ entonces
se rechaza H_0

Si $\chi_{cal}^2 < \chi_{\alpha/2}^2$ ó $\chi_{cal}^2 > \chi_{1-\alpha/2}^2$
se rechaza H_0

Si $\chi_{cal}^2 > \chi_{1-\alpha}^2$
se rechaza H_0

PRUEBA DE HIPOTESIS SOBRE LA IGUALDAD DE DOS MEDIAS (σ^2 CONOCIDAS)

Se tienen 2 poblaciones

La primera con μ_1 desconocida y σ_1^2 conocida

La segunda con μ_2 desconocida y σ_2^2 conocida

Nuestro interés recae en probar la hipótesis de que las dos medias poblacionales μ_1 y μ_2 son iguales (suponiendo que las dos poblaciones son normales)

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Se toman una muestra aleatoria de cada población

- n_1 para la primera población $X_{11}, X_{12}, \dots, X_{1n}$
- n_2 para la segunda población $X_{21}, X_{22}, \dots, X_{n2}$

X_1 se distribuye según $N(\mu_1, \sigma_1^2)$

$$X_{11} = x_{11}, X_{12} = x_{12}, \dots, X_{1n_1} = x_{1n_1}$$

$$\overline{X_1} = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1}$$

$$S_1^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \overline{X_1})^2}{n_1 - 1}$$

X_2 se distribuye según $N(\mu_2, \sigma_2^2)$

$$X_{21} = x_{21}, X_{22} = x_{22}, \dots, X_{2n_2} = x_{2n_2}$$

$$\overline{X_2} = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2}$$

$$S_2^2 = \frac{\sum_{i=1}^{n_2} (X_{2i} - \overline{X_2})^2}{n_2 - 1}$$

El procedimiento de prueba se basa en la distribución de la diferencia entre las $\bar{X}_1 - \bar{X}_2$ medias muestrales

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 / n_1 + \sigma_2^2 / n_2)$$

Donde si la $H_0: \mu_1 = \mu_2$, el estadístico de prueba es:

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

Para probar H_0 se calcula el valor numérico del estadístico Z_0 y se rechaza H_0 si:

$$Z_0 > Z_{1-\alpha/2} \quad \text{ó} \quad Z_0 < -Z_{1-\alpha/2}$$

Para probar hipótesis alternativas unilaterales se analizan de forma similar:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Se calcula el estadístico de prueba Z_0 y H_0 se rechaza si:

$$Z_0 > Z_{1-\alpha}$$

Para probar la otra hipótesis alternativa unilateral:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 < \mu_2$$

Se calcula estadístico de prueba Z_0 y H_0 se rechaza si:

$$Z_0 < -Z_{1-\alpha}$$

EJEMPLO:

Un diseñador de productos está interesado en reducir el tiempo de secado de una pintura. Se prueban dos fórmulas de pintura; la fórmula 1 tiene el contenido químico estándar, y la fórmula 2 tiene un nuevo ingrediente secante que debe reducir el tiempo de secado. De la experiencia se sabe que la desviación estándar del tiempo de secado es ocho minutos, y esta variabilidad inherente no debe verse afectada por la adición del nuevo ingrediente. Se pintan diez especímenes con la fórmula 1 y otros diez con la 2. Los dos tiempos de secado muestrales son 121 y 112 minutos respectivamente. ¿A qué conclusiones puede llegar el diseñador del producto sobre la eficacia del nuevo ingrediente, utilizando $\alpha=0.05$?

SOLUCION

- La cantidad de interés es la diferencia entre los tiempos de secado

$$\mu_1 - \mu_2$$

- Sea la prueba de hipótesis

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Se desea rechazar H_0 si el nuevo ingrediente disminuye el tiempo promedio de secado ($\alpha=0.05$) donde

$$\sigma_1^2 = \sigma_2^2 = (8)^2 = 64 \text{ y } n_1 = n_2 = 10$$

- Rechazamos $H_0: \mu_1 = \mu_2$ si $Z_0 > 1.645 = Z_{0.95}$

Calculamos el estadístico de prueba

$$Z_0 = \frac{121 - 112}{\sqrt{\left(\frac{(8)^2}{10} + \frac{(8)^2}{10} \right)}} = 2.52$$

CONCLUSION

Puesto que $Z_0 = 2.52 > 1.645$ se rechaza $H_0: \mu_1 = \mu_2$ con un nivel $\alpha = 0.05$, y se concluye que la adición del nuevo ingrediente a la pintura sí disminuye de manera significativa el tiempo de secado.

PRUEBAS DE HIPÓTESIS SOBRE LAS MEDIAS DE DOS

DISTRIBUCIONES NORMALES, VARIANZAS DESCONOCIDAS

- Se pueden presentar 2 casos:
- Caso1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Caso2: $\sigma_1^2 \neq \sigma_2^2$

Caso1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Suponer:

- μ_1 y μ_2 son dos poblaciones normales, independientes, desconocidas
- σ_1^2 y σ_2^2 varianzas desconocidas, pero iguales.
- Se desea probar:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

- Sean $X_{11}, X_{12}, \dots, X_{1n_1}$ una muestra aleatoria de n_1 observaciones tomadas de la primera población,
- y $X_{21}, X_{22}, \dots, X_{2n_2}$ una muestra aleatoria de n_2 observaciones tomadas de la segunda población.
- *Obtenemos $\bar{X}_1, \bar{X}_2, S_1^2$ y S_2^2* , que son las medias muestrales y las varianzas muestrales respectivas.
- S_1^2 y S_2^2 son los estimadores de la varianza σ^2 y pueden combinarse para formar un solo estimador (S_p^2)

- $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}$
- La estadística de prueba es:

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \approx T_{(n_1 + n_2 - 2)}$$

Si $H_0: \mu_1 = \mu_2$ es verdadero t_0 tiene distribución $t_{n_1 + n_2 - 2}$

- Ahora entonces si:

$$t_o > t_{(1-\alpha/2, n^1+n^2-2)}$$

ó

$$t_o < - t_{(1-\alpha/2, n^1+n^2-2)}$$

- Se debe rechazar

$$H_o: \mu_1 = \mu_2$$

- Ejemplo: se analizan dos catalizadores para determinar la forma en que afectan el rendimiento promedio de un proceso químico. Para la cual se realizará una prueba de hipótesis que dé una respuesta a que ¿si existe alguna diferencia entre los rendimientos promedio? con un $\alpha = 0.05$ los datos del estudio son los siguientes: $X_1 = 92.255$, $X_2 = 92.733$, $S_1 = 2.39$, $S_2 = 2.98$, $n_1 = n_2 = 8$

- Desarrollo:
- μ_1 y μ_2 parámetros que representan el rendimiento promedio del proceso con los catalizadores 1 y 2.
- $H_0: \mu_1 = \mu_2$
 $H_1: \mu_1 \neq \mu_2$
- $\alpha=0.05$
- El estadístico de prueba es:

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \approx T_{(n_1+n_2-2)}$$

- Rechazar H_0 si $t_0 > t_{0.975,14} = 2.145$ ó si $t_0 < -t_{0.975,14} = -2.145$.

$$R.R. \{ t_0 > 2.145 \text{ ó } t_0 < -2.145 \}$$

- Calculo de t_0

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)} = \frac{7(2.39)^2 + 7(2.98)^2}{8+8-2} = 7.3$$

$$S_p = \sqrt{7.3} = 2.3$$

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{92.255 - 92.733}{2.7\sqrt{(1/8 + 1/8)}} = -0.35$$

- Dado que $-2.45 < t_0 = -0.35 < 2.145$ no es posible rechazar la hipótesis nula.

Caso2: $\sigma_1^2 \neq \sigma_2^2$

- Se procede de igual forma que en el Caso1 ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) pero con las siguientes salvedades.
- Se define un nuevo estadístico dado por

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}} \approx T_{(v)}$$

V grados de libertad dados por:

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{\frac{n_1}{n_1 + 1}} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{\frac{n_2}{n_2 + 1}}} - 2$$

Estad. Aplicada

Prueba de hipótesis para igualdad de varianza

- Sean $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ parámetros desconocidos.
- Se desea probar:
- $H_0: \sigma_1^2 = \sigma_2^2$
 $H_1: \sigma_1^2 \neq \sigma_2^2$
- El estadístico es: $F_0 = \frac{S_1^2}{S_2^2} \approx F_{(n_1-1, n_2-1)}$

Tiene una distribución F con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador.

- Ahora entonces si:

$$f_0 > f_{(1-\alpha/2, n_1-1, n_2-1)}$$

o si

$$f_0 < f_{(\alpha/2, n_1-1, n_2-1)} = 1 / f_{(1-\alpha/2, n_1-1, n_2-1)}$$

Se debe rechazar $H_0: \sigma_1^2 = \sigma_2^2$

- Ejemplo: se estudian dos mezclas de diferentes gases con la finalidad de determinar con cual se obtienen mejores resultados en cuanto a la variabilidad del espesor del oxido, 20 elementos son depositados en cada gas las desviaciones estándar de cada muestra del espesor del oxido son: $S_1 = 1.96$ y $S_2 = 2.13$ ¿existe alguna preferencia por alguno de los gases? Utilice $\alpha = 0.05$

- Solución:
- Los parámetros de interés son las varianzas del espesor del oxido σ_1^2 y σ_2^2 .
- $H_0: \sigma_1^2 = \sigma_2^2$
 $H_1: \sigma_1^2 \neq \sigma_2^2$
- $\alpha = 0.05$
- El estadístico de prueba es:

$$F_0 = \frac{S_1^2}{S_2^2} = \frac{3.84}{4.54} = 0.85$$
- R.R $\{ f_0 > f_{0.975, 19, 19} = 2.53 \text{ ó } f_0 < f_{0.025, 19, 19} \}$

- $f_0 < f_{0.025,19,19} = 1/f_{0.975,19,19} = 1/2.53 = 0.40$

- Conclusión:

Como

$$f_{0.025,19,19} = 0.40 < 0.85 < f_{0.975,19,19} = 2.53,$$

No es posible rechazar la hipótesis nula

$H_0: \sigma_1^2 = \sigma_2^2$ con el nivel de significancia $\alpha=0.05$.

Por consiguiente, no hay evidencia fuerte que indique cuál de los dos gases dará como resultado una varianza más pequeña en el espesor de la capa de óxido.

Caso I

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 < \sigma_2^2$$

Caso II

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

Caso III

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 > \sigma_2^2$$

Prueba Estadística:

$$F = \frac{s_1^2}{s_2^2}$$

con n_1-1 g.l. en el numerador y n_2-1 g.l en el denominador

Decisión:

Si $F_{cal} < F_{\alpha}$ entonces
se rechaza H_0

Si $F_{cal} < F_{\alpha/2}$ o $F_{cal} > F_{1-\alpha/2}$
se rechaza H_0

Si $F_{cal} > F_{1-\alpha}$ entonces
se rechaza H_0

Comparando media de dos poblaciones usando muestras pareadas

En este caso se trata de comparar dos métodos o tratamientos, pero se quiere que las unidades experimentales donde se aplican los tratamientos sean las mismas, ó lo más parecidas posibles, para evitar influencia de otros factores en la comparación

Sea X_i el valor del tratamiento I y Y_i el valor del tratamiento II en el i -ésimo sujeto. Consideremos $d_i = X_i - Y_i$ la diferencia de los tratamientos en el i -ésimo sujeto.

Las inferencias que se hacen son acerca del promedio poblacional μ_d de las d_i . Si $\mu_d = 0$, entonces significa que no hay diferencia entre los dos tratamientos.

Intervalo de Confianza

Un intervalo de confianza del $100(1-\alpha)\%$ para la diferencia poblacional μ_d dada una muestra de tamaño n es de la forma

$$\left(\bar{d} - t(n-1, \alpha/2) \text{ sd} / \sqrt{n}, \bar{d} + t(n-1, \alpha/2) \text{ sd} / \sqrt{n} \right)$$

donde \bar{d} , es media de las diferencias muestrales d_i y es la desviación estándar.

$$s_d = \sqrt{\frac{\sum_i (d_i - \bar{d})^2}{n-1}}$$

Pruebas de Hipótesis

Caso I

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d < 0$$

Caso II

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0$$

Caso III

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d > 0$$

Prueba Estadística:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \text{ se distribuye con una } t \text{ de Student con } n-1 \text{ gl.}$$

Decisión:

Si $t < -t_{1-\alpha}$ entonces
entonces

se rechaza H_0
 H_0

Si $|t| > t_{1-\alpha/2}$ entonces

se rechaza H_0

Si $T_{cal} > t_{1-\alpha}$

se rechaza

Ejemplo

Un médico desea investigar si una droga tiene el efecto de bajar la presión sanguínea en los usuarios. El médico eligió al azar 15 pacientes mujeres y les tomó la presión, luego les recetó la medicina por un período de 6 meses, y al final del mismo nuevamente les tomó la presión. Los resultados son como siguen:

Sujetos															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Antes	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
Después	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74

Solución:

Sea μ_d que representa la media poblacional de las diferencias. Luego:

$H_o: \mu_d = 0$ (La droga no tiene ningún efecto)

$H_a: \mu_d > 0$ (La droga tiene efecto, la presión antes de usar la droga era mayor que después de usarla).

Paired T-Test and Confidence Interval

Paired T for Antes – Después

	N	Mean	StDev	SE Mean
Antes	15	75.87	6.86	1.77
Después	15	67.07	6.67	1.72
Difference	15	8.80	10.98	2.83

95% CI for mean difference:(2.72, 14.88)

T-Test of mean difference = 0 (vs > 0): T-Value = 3.11 P-Value = 0.004

Interpretación: Notando que el “P-value” es .004 menor que .05, se rechaza la hipótesis nula y se llega a la conclusión de que, efectivamente la droga reduce la presión sanguínea. Por otro lado, se puede observar que el intervalo de confianza del 95% para la diferencia de medias es (2.72, 14.88), el cual no contiene a cero, ésta es otra razón para rechazar la hipótesis nula.

Inferencia para Proporciones

Pruebas de hipótesis:

Caso I

$$H_0 : p = p_0$$

$$H_a : p < p_0$$

Caso II

$$H_0 : p = p_0$$

$$H_a : p \neq p_0$$

Caso III

$$H_0 : p = p_0$$

$$H_a : p > p_0$$

Prueba Estadística (Aproximada):

$$Z = \frac{(p - p_0)}{\sqrt{\frac{p_0 q_0}{n}}}$$


Decisión

Si $Z_{cal} < -Z_{1-\alpha}$ entonces
se rechaza H_0

Si $|Z_{cal}| > Z_{\alpha/2}$ entonces
se rechaza H_0

Si $Z_{cal} > Z_{1-\alpha}$ entonces
se rechaza H_0

Pruebas No Paramétricas

Análisis de Datos  Prueba de Bondad de Ajuste
Categoricos Prueba de Independencia
Aplicaciones de la Chi Prueba de Igualdad de proporciones
Cuadrado Prueba de Homogenidad

Objetivo: hacer inferencia acerca de k parámetros p_1, p_2, \dots, p_k que se definen como las proporciones o probabilidades de que ocurran respectivamente k resultados posibles C_1, C_2, \dots, C_k mutuamente excluyentes de un experimento multinomial

Distribución Multinomial

El vector aleatorio (X_1, X_2, \dots, X_k) tiene distribución multinomial con parámetros n, p_1, p_2, \dots, p_k

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{\prod_{i=1}^k x_i!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

$$E[X] = (np_1, np_2, \dots, np_k)$$

Ejercicio 2. La siguiente tabla reporta la distribución de la población de un país de acuerdo a su nivel educacional y el número de alcaldes elegidos en cada una de las categorías en las últimas elecciones:

Nivel Educacional	País	Alcaldes electos
Primaria	30%	6
Secundaria	45%	15
Universitaria Incompleta	12%	27
Universitaria Completa	13%	30

¿Habrá suficiente evidencia para concluir que la distribución del nivel educacional de los alcaldes electos sigue la misma distribución del país?. Usar un nivel de significación del 5%.

Ejemplo: Si se toma una muestra de 10 alcaldes

Nos interesa estudiar la probabilidad de:

- a) Encontrar 2 con estudios primarios, 4 con secundaria, 1 con estudios universitarios incompletos y 3 con estudios universitarios completos.
- b) Encontrar a lo más dos alcaldes con estudios primarios y 8 tengan estudios secundarios.
- c) Calcule el número esperado de alcaldes con estudios de primaria.

Intervalo de Confianza Para Proporciones

Para una proporción

$$p_i \in \left(\bar{p}_i \pm Z_{1-\alpha/2} \sqrt{\frac{\bar{p}_i \bar{q}_i}{n}} \right)$$

Para Diferencia de dos proporciones

$$(\bar{p}_i - \bar{p}_j) \pm Z_{(1-\alpha/2)} \sqrt{\frac{\bar{p}_i(1-\bar{p}_i) + \bar{p}_j(1-\bar{p}_j) + 2\bar{p}_i\bar{p}_j}{n}}$$

Para dos poblaciones independientes

$$(\bar{p}_i - \bar{p}_j) \pm Z_{(1-\alpha/2)} \sqrt{\left(\frac{\bar{p}_i(1-\bar{p}_i)}{n_1} + \frac{\bar{p}_j(1-\bar{p}_j)}{n_2} \right)}$$

Intervalo de Confianza Simultáneos

Nivel de confianza global $(1 - \alpha)$

Por el método de Bonferroni se deben calcular $m = C_2^k$ intervalos de estimación de diferencias de dos parámetros con nivel de confianza igual a $(1 - \alpha_o)$ donde $\alpha_o = \frac{\alpha}{m}$

$$(\bar{p}_i - \bar{p}_j) \pm Z_{(1-\alpha_o/2)} \sqrt{\frac{\bar{p}_i(1-\bar{p}_i) + \bar{p}_j(1-\bar{p}_j) + 2\bar{p}_i\bar{p}_j}{n}}$$

Intervalos de confianza

K	4
Alfa	0.05
Nro de Comparaciones	6
$\alpha_o = \frac{\alpha}{m}$	0.0083333333
$Z_{(1-\alpha_o / 2)}$	2.3939798

Ejemplo

p1	p2	p3	p4
0.076923077	0.192307692	0.34615385	0.38461538
Comparaciones	Límite inferior	Límite Superior	Conclusión
4y 1	0.289102351	0.47187436	4>1
4 y 2	-0.006872554	0.39148794	4=2
4y3	-0.193023695	0.26994677	4=3
3 y1	0.10873115	0.42973039	3>1
3 y 2	-0.040640343	0.34833265	3=2
2 y 1	-0.021742356	0.25251159	2=1

Comparaciones múltiples de proporciones : Método de Bonferroni

Para analizar las diferencias significativas de tres o más parámetros de manera que tengan un mismo nivel de confianza global.

Cuando se tiene más de dos parámetros de poblaciones que globalmente son diferentes.

Desigualdad de Bonferroni

$$P(A_1 \cap A_1 \cap \dots \cap A_m) \geq \sum_{i=1}^m P(A_i) - m + 1$$

$$P(A_1 \cap A_1 \cap \dots \cap A_m) \geq 1 - \alpha$$

$$P(A_i) = 1 - \alpha_o, \text{ donde } \alpha_o = \frac{\alpha}{m}$$

Si son independientes

$$P(A_1 \cap A_1 \cap \dots \cap A_m) = P(A_1)P(A_2)P(A_3)\dots P(A_m)$$

$$1 - \alpha = (1 - \alpha_o)^m \Rightarrow \alpha_o = 1 - (1 - \alpha)^{1/m}$$

I.- Prueba de Bondad de Ajuste:

Aquí se trata de probar si los datos de una muestra tomada siguen una cierta distribución predeterminada. Los n datos tomados deben estar divididos en K categorías.

Categoría	1	2	3	...	K	
Frecuencia observada	Obs_1	Obs_2	Obs_3		Obs_k	N

Se asume que las probabilidades p_i , de caer en la categoría i deben ser conocidos.

La hipótesis nula es $H_o: p_1 = p_{10}, p_2 = p_{20} = \dots = p_k = p_{k0}$, es decir los datos siguen la distribución deseada, y la hipótesis alternativa es H_a : al menos una de las p_i es distinta de la probabilidad dada p_{i0} .

La prueba estadística es:

$$\sum_{i=1}^k \frac{(Obs_i - np_{i0})^2}{np_{i0}}$$

donde p_{i0} representa la proporción deseada en la i -ésima categoría, Obs_i la frecuencia observada en la categoría i y n es el tamaño de la muestra. La prueba estadística se distribuye como una Chi-Cuadrado con $K-1$ grados de libertad donde, K es el número de categorías.

Si el valor de la prueba estadística es mayor que $\chi^2_{1-\alpha}$

se rechaza la hipótesis nula.

Ejemplo 1. Los siguientes datos representan los nacimientos por mes en una ciudad durante 1993. Probar si hay igual probabilidad de nacimiento en cualquier mes del año. Usar un nivel de significación del 5%.

5435	4830	5229	4932	5052	5072	5198	5712
6126	5972	5748	5936				

Solución:

La hipótesis nula es H_0 : Hay igual probabilidad de nacer en cualquier mes del año (es decir, $p_1 = p_2 = \dots = p_{12} = 1/12 = .083$).

La hipótesis alterna es que no hay igual probabilidad de nacer en cualquier mes del año.

Obs= Oi	pi	Ei =npi	$\sum_{i=1}^k \frac{(Obs_i - np_{io})^2}{np_{io}}$
5435	0.083	5436.833	0.001
4830	0.083	5436.833	67.732
5229	0.083	5436.833	7.945
4932	0.083	5436.833	46.876
5052	0.083	5436.833	27.240
5072	0.083	5436.833	24.482
5198	0.083	5436.833	10.492
5712	0.083	5436.833	13.927
6126	0.083	5436.833	87.358
5972	0.083	5436.833	52.678
5748	0.083	5436.833	17.809
5936	0.083	5436.833	45.830
65242		65242.000	402.368

Esta es la prueba de Chi-Cuadrado para Bondad de ajuste

$$\sum_{i=1}^k \frac{(Obs_i - np_{io})^2}{np_{io}} = 402.369$$

El valor de $\chi^2_{(0.95,11)} = 19.675$

Interpretación: Comparando el valor de la prueba estadística con una Chi-Cuadrado con 11 grados de libertad y nivel de significación del 5 por ciento que es 19.6751 se concluye que se rechaza la hipótesis nula, es decir no hay igual probabilidad de nacimiento para los meses.

Ejercicio 2. La siguiente tabla reporta la distribución de la población de un país de acuerdo a su nivel educacional y el número de alcaldes elegidos en cada una de las categorías en las últimas elecciones:

Nivel Educacional	País	Alcaldes electos
Primaria	30%	6
Secundaria	45%	15
Universitaria Incompleta	12%	27
Universitaria Completa	13%	30

¿Habrá suficiente evidencia para concluir que la distribución del nivel educacional de los alcaldes electos sigue la misma distribución del país?. Usar un nivel de significación del 5%.

La hipótesis Nula es:

Ho: la distribución del nivel educacional de los alcaldes es la misma que la distribución del nivel educacional del país. Es decir, $p_{el}=.30$, $p_{es}=.45$, $p_{ui}=.12$, $p_{uc}=.13$.

La hipótesis alternativa es:

Ha: la distribución del nivel educacional de los alcaldes no es la misma que la distribución del nivel educacional del país. Es decir, que al menos una de las siguientes igualdades $p_{el}=.30$, $p_{es}=.45$, $p_{ui}=.12$, $p_{uc}=.13$ no se cumple.

Nivel Educativo	Obs= Oi	pi	Ei =npi	$\sum_{i=1}^k \frac{(Obs_i - np_{io})^2}{np_{io}}$
Primaria	6	0.3	23.4	12.938
Secundaria	15	0.45	35.1	11.510
Universitaria Incompleta	27	0.12	9.36	33.245
Universitaria completa	30	0.13	10.14	38.897
	78		78	96.591

Chi-Square with 3 grados de libertad es 7.8147 y 0.95

Interpretacion: la hipotesis nula se rechaza pues el Chi-square calculado es 96.59 mayor que el Chi-square al 95%. Luego la distribucion del nivel educacional de los alcaldes no es la misma que la del pais.

Estad. Aplicada

Prueba de Bondad de Ajuste: Método Gráfico

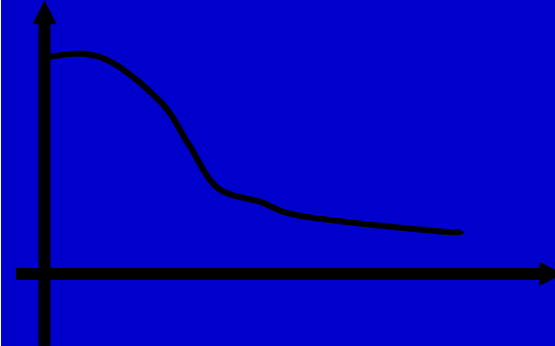
Dibujar la curvas acumuladas Teórica y empírica.

**Si ambas curvas no se desvían mucho
Entre sí, entonces la muestra tiene la
Distribución teórica.**

Aplicación del Método Gráfico: Obtención de la exponencial

$$\lambda = 1/t_m =$$
$$= 1/3.93 =$$
$$0.254 \text{ servicios/min}$$

$$f(t) = \lambda e^{-\lambda t}, t > 0$$



$$\mathbf{F(t)} = \int_0^T \lambda e^{-\lambda t} dt =$$

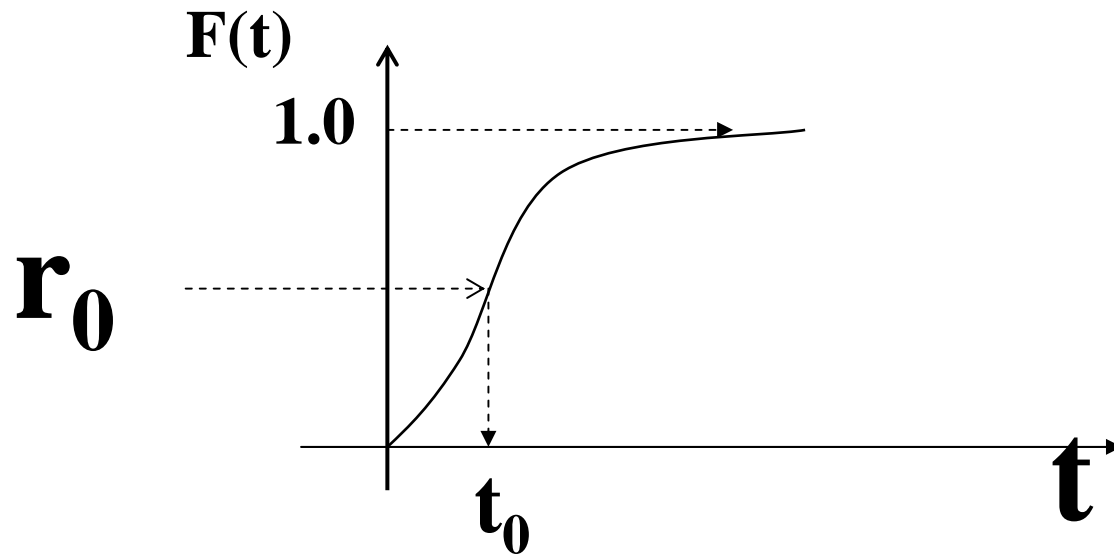
$$= -\int e^{-\lambda t} (-\lambda dt) = [e^{-\lambda t}]_0^T$$

$$= -[e^{-\lambda T} - 1] =$$

$$= 1 - e^{-\lambda T}$$

$$= 1 - e^{-0.254T}$$

Distribución Teórica (acumulada) de los componentes: La exponencial



Ejemplo: Muestra del Tiempo (en minutos) de atención de 60 componentes

.7	.4	3.4	4.8	2.0	1.0	5.5	6.2	1.2	4.4	Intervalo	Observaciones	Frecuencia	Frecuencia relativa	Frec relativa acumulada
1.5	2.4	3.4	6.4	3.7	4.8	2.5	5.5	.3	8.7	(0,1)	,	11	.1883	.1883
2.7	.4	2.2	2.4	.5	1.7	9.3	8.0	4.7	5.9	(1,2)	.	8	.1333	.3166
.7	1.6	5.2	.6	.9	3.9	3.3	.2	.2	4.9	(2,3)	.	9	.1500	.4666
9.6	1.9	9.1	1.3	10.6	3.0	.3	2.9	2.9	4.8	(3,4)	.	7	.1167	.5833
8.7	2.4	7.2	1.5	7.9	11.7	6.3	3.8	6.9	5.3	(4,5)	.	6	.	.6833



Media: $t_m = \sum f_i t_i$
 $= .1883*0.5 + 0.1333*1.5 + .. = 3.93$
Varianza: $s_i^2 = \sum f_i (t_m - t)^2$
 $= .1883(3.93-.5)^2 + .1333(3.93-1.5)^2 + ... = 8.64$

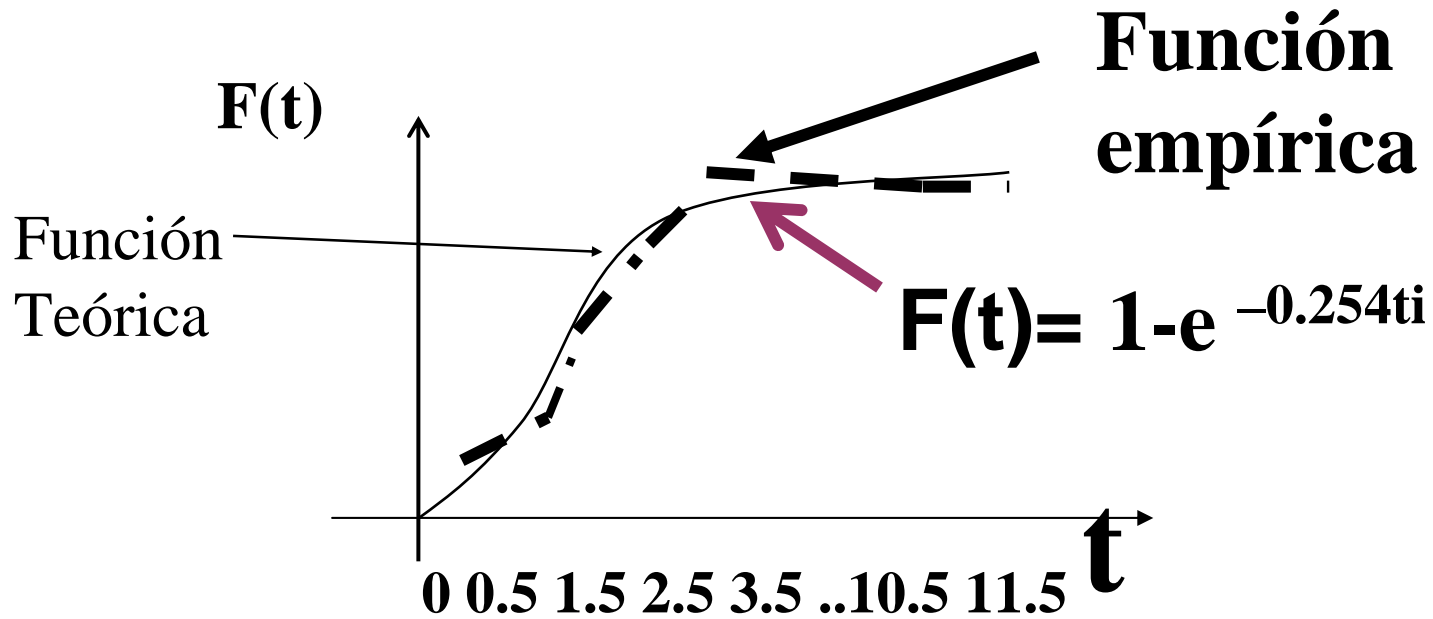
Totales	.	60	1.000	-----
---------	---	----	-------	-------

Intervalo	Marca de clase ti	Frecuencia	Frecuencia relativa acumulada	Distribución de Probabilidad acumulada: $F(t)=1-e^{-0.254t_i}$
(0,1)	0.5	11	.1883	.12
(1,2)	1.5	8	.3166	..
(2,3)	2.5	9	.4666	..
(3,4)	3.5	7	.5833	..
(4,5)	4.5	6	.6833	...
(5,6)	.	5	.	.
(6,7)	.	4	.	.
(7,8)	.	2	.	.
(8,9)	.	3	.	.
((9,10)	.	3	.	.
(10,11)	.	1	.	.
(11,12)	.	1	1.0000	1.0000
Totales	.	60	-----	-----

Estad. Aplicada

Comparar
r
Teórica
Vs
Empírica

Comparación Gráfica



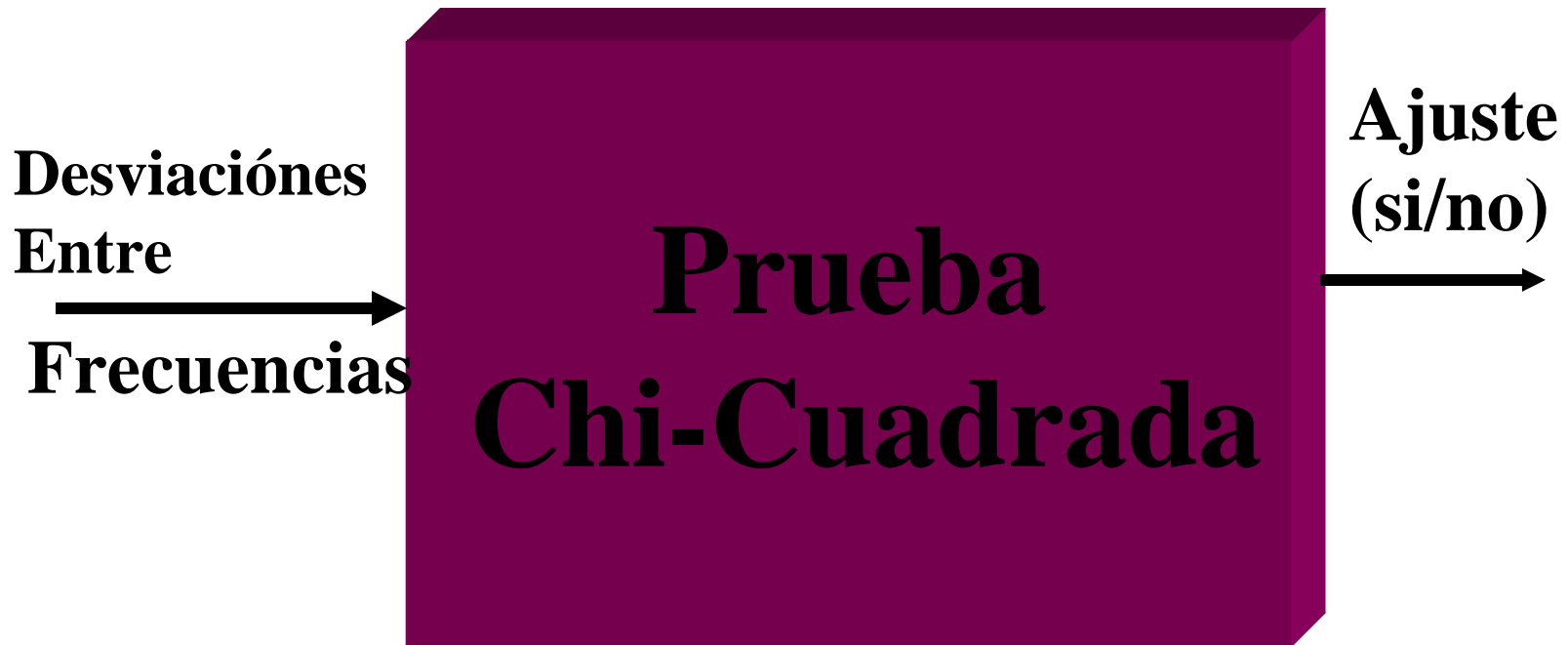
Método Estadístico



Tipos de Hipótesis

- H_0 : Hipótesis que se va a probar. Es la que se desea aceptar o rechazar.
- H_1 : Hipótesis alternativa. Es la que se toma cuando H_0 se rechaza. Ejem:
 - H_0 : la distribución es exponencial.
 - H_1 : La distribución no es exponencial

Prueba de Bondad de Ajuste



Proponer una distribución
Ej: Exponencial

Ch

Obtener Distribución
Empírica

Calcular
Dispersiones χ^2

Estimar el parámetro de
La distribución.
Ej: λ

Calcular α , p , $\chi^2_{\alpha,p}$
 $p = N - k - 1$

$\chi^2 < \chi^2_{\alpha,p}$

Ho ok

Muestra del Tiempo (en minutos) de atención de 60 componentes

.7	.4	3.4	4.8	2.0	1.0	5.5	6.2	1.2	4.4
1.5	2.4	3.4	6.4	3.7	4.8	2.5	5.5	.3	8.7
2.7	.4	2.2	2.4	.5	1.7	9.3	8.0	4.7	5.9
.7	1.6	5.2	.6	.9	3.9	3.3	.2	.2	4.9
9.6	1.9	9.1	1.3	10.6	3.0	.3	2.9	2.9	4.8
8.7	2.4	7.2	1.5	7.9	11.7	6.3	3.8	6.9	5.3



Intervalo	Observaciones	Frecuencia	Frecuencia relativa	Frecuencia relativa acumulada
(0,1)	,	11	.1883	.1883
(1,2)	.	8	.1333	.3166
(2,3)	.	9	.1500	.4666
(3,4)	.	7	.1167	.5833
(4,5)	.	6	.	.6833
(5,6)				
(6,7)				
(7,8)				
(8,9)				
((9,10)				
(10,11)				
(11,12)	.	1	.	1.0000
Totales	.	60	1.000	-----

Media: $t_m = \sum f_i t_i = 3.93$

Frecuencias Teóricas n_i

Inter- valo	Marca clase t_i	Frec.Ob- servada O_i	Frecuencia relativa Acu- mulada F_i	Dist. Prob Acumulada $F(t)=1-e^{-0.254t_i}$
(0,1)	0.5	11	.1883	.12
(1,2)	1.5.	8	.3166	..
(2,3)	2.5	9	.4666	..
(3,4)	3.5	7	.5833	..
(4,5)	4.5	6	.6833	...
(5,6)	.	5		
(6,7)	.	4		
(7,8)	.	2		
(8,9)	.	3		
((9,10)	.	3		
(10,11)	.	1	.	.
(11,12)	.	1	1.0000	1.0000
Totales	.	60	-----	-----

Frecuencias, celda i:

Teórica: n_i

Observada: O_i

Frec relativa celda $i=n_i/n$

$$n_i/n \approx \int_{I_{i-1}}^{I_i} f(t)dt$$

$$n_i = 60 \{F(I_i)-F(I_{i-1})\}$$

$$n_i = 60\{e^{-.254 I_{i-1}} - e^{-.254 I_i}\}$$

Intervalo	Marca de clase ti	Frec. Observada Oi	FrecTeórica ni	$\frac{(oi-ni)^2}{ni}$
(0,1)	0.5	11	13.47	.435
(1,2)	1.5.	8	10.44	.570
(2,3)	2.5	9	8.10	.100
(3,4)	3.5	7	6.28	.083
(4,5)	4.5	6	4.87	.
(5,6)	.	5	3.88	.
(6,7)	.	4
(7,8)	.	2
(8,9)	.	3
((9,10)	.	3	1.37	.
(10,11)	.	1	1.06	.
(12,∞)	.	1	2.75	.
Totales	.	N=5	n=60	χ^2

25

21.71

Estadística Aplicada

Desviaciones de Frecuencias

$$\chi^2 = \sum \frac{(oi - ni)^2}{ni}$$

Oi's Pequeñas (≈5)
Se agrupan

Parámetros

Tablas:

$$\chi^2_{3,\alpha}=7.815$$

$$N=5$$

$$\alpha=0.05$$

$$K=1$$

$$P=3$$

$$\chi^2=1.705$$

V_s

$$\chi^2_{3,\alpha}=7.815$$

→ Ho se

→ acepta

II.- Pruebas de Independencia y Homegeneidad

Consideremos datos de dos variables cualitativas A y B como por ejemplo, sexo:M y F y opinion: A Favor, en contra de una persona. También podrían ser dos variables cuantitativas que han sido categorizadas, como por ejemplo, Nivel de Educación y Nivel de salario

Los datos pueden organizarse en una tabla de doble entrada llamada tabla de contingencia. La siguiente es una tabla de contingencia 2 x 2

	A1	A2	Total
B1	8	6	14
B2	12	9	21
Total	20	15	35

La primera pregunta que uno se hace es si existirá o no relación entre las variables A y B, es decir si A y B son o no independientes. A y B serán independientes si cada entrada de la tabla es igual al producto de los totales marginales dividido entre el número de datos.

Claramente, esto se cumple para la tabla anterior. Por ejemplo, $8 = (14)(20)/35$. En consecuencia, no hay relación entre las variables A y B.

Otra pregunta que se puede tratar de responder es sí las proporciones de los valores de la variable B en cada columna son iguales.

Por ejemplo si

A: El estudiante graduando consigue trabajo,

B: Sexo del graduando.

Uno puede estar interesado en comparar la proporción de mujeres graduandas que consiguen trabajo con la proporción de mujeres graduandas que no consiguen trabajo.

La forma general de una tabla de contingencia es la siguiente

		VAR A					Total
		A_1	A_2	A_3	...	A_c	
VAR B	B_1	O_{11}	O_{12}	O_{13}		O_{1c}	R_1
	B_2	O_{21}	O_{22}	O_{23}		O_{2c}	R_2
	B_3	O_{31}	O_{32}	O_{33}		O_{3c}	R_3
	
	B_r	O_{r1}	O_{r2}	O_{r3}	...	O_{rc}	R_r
	Total	C_1	C_2	C_3	...	C_c	N

Aquí O_{ij} es el número de sujetos que tienen las características A_i y B_j a la vez.

R_i ($i = 1, \dots, r$) es la suma de la i -ésima fila de la tabla. Es decir, es el total de sujetos que poseen la característica B_i .

C_j ($j = 1, \dots, c$) es la suma de la j -ésima columna de la tabla. Es decir, es el total de sujetos que poseen la característica A_j .

n representa el total de observaciones tomadas.

Cuando consideramos que los valores de nuestra tabla han sido extraídos de una población, entonces nos interesaría probar las siguientes dos hipótesis:

La **prueba de Independencia**, que se efectúa para probar si hay asociación entre las variables categóricas A y B, y

La **prueba de Homogeneidad**, que es una generalización de la prueba de igualdad de dos proporciones. En este caso se trata de probar si para cada nivel de la variable B, la proporción con respecto a cada nivel de la variable A es la misma. Si A tiene 3 niveles y B tiene 2 niveles *entonces* $H_0 : p_1=p_2=p_3$, donde p_i es la proporción de uno de los valores de la variable B en cada columna de A

Por ejemplo si estamos interesados en establecer si hay relación entre el nivel de educación y el nivel de salario se usará una prueba de independencia. Si deseamos probar que para cada nivel económico hay igual proporción de personas en cada partido político se usará una prueba de homogeneidad.

Las hipótesis de independencia son:

Ho: No hay asociación entre las variables A y B (es decir hay independencia)

Ha: Si hay relación entre las variables A y B

Las hipótesis de Homogeneidad son:

Ho: Las proporciones de cada valor de la variable B son iguales en cada columna

Ha: Al menos una de las proporciones para cada valor de la variable B no son iguales en cada columna.

Ambas hipótesis se prueban usando una prueba de Ji-Cuadrado:

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde, O_{ij} es la frecuencia observada de la celda que está en la fila i , columna j y

$$E_{ij} = \frac{R_i C_j}{n}$$

es la frecuencia esperada de la celda (i, j) . La frecuencia esperada es aquella que debe ocurrir para que la hipótesis nula sea aceptada.

La prueba estadística se distribuye como una Ji-Cuadrado con $(r-1)(c-1)$ grados de libertad.

Decision: La hipótesis Nula se rechaza si

$$\chi^2_{cal} > \chi^2_{1-\alpha}$$

donde α es el nivel de significancia o equivalentemente se rechaza si el "P-value" es menor que 0.5.

Si la tabla de contingencia presenta pocas observaciones en algunas celdas (digamos menos de 5), entonces la prueba no es confiable.

Ejemplo: Usando los siguientes datos establecer si hay relación entre las variables tipo de escuela superior y el resultado (aprueba o no aprueba), de la primera clase de matemáticas que toma el estudiante en la universidad, basados en los resultados de 20 estudiantes.

Est escuela aprueba			Est escuela aprueba		
1	priv	si	11	públ	si
2	priv	no	12	priv	no
3	públ	no	13	públ	no
4	priv	si	14	priv	si
5	públ	si	15	priv	si
6	públ	no	16	públ	no
7	públ	si	17	priv	no
8	priv	si	18	públ	si
9	públ	si	19	públ	no
10	priv	si	20	priv	si

Para la prueba de Independencia las hipótesis son:
Ho: No hay relación entre el tipo de escuela y el resultado obtenido en la primera clase de Matemáticas.
Ha: Si hay relación entre ambas variables.

Para la prueba de homogeneidad las hipótesis son:
Ho: La proporción de aprobados en la primera clase de matemáticas es igual tanto para estudiantes que provienen de escuela pública como de escuela privada.
Ha: La proporción de aprobados en la primera clase de matemáticas no es la misma para ambos tipos de escuela.

Ejemplo. Usar los siguientes datos para tratar de establecer si hay relación entre el Sexo del entrevistado y su opinión con respecto a una ley del gobierno.

Sexo	Opinion	conteo
male	si	10
male	no	20
male	abst	30
female	si	15
female	no	31
female	abst	44

Las hipótesis correspondientes son:

Ho: No hay asociación entre el sexo del entrevistado y su opinión, y

Ha: Si hay relación entre las variables.

Medidas de Asociación

Asumiendo que se rechaza la hipótesis Nula H_0 : No hay relación entre las variables de la tabla, entonces el próximo paso es determinar el grado de asociación de las dos variables categóricas, para ello se usan las llamadas medidas de asociación.

Existen un gran número de estas medidas, tales como:

El Coeficiente de Contingencia,

El coeficiente de Cramer,

la medida Kappa,

los coeficientes Lambda y Tau de Goodman y Kruskal y

los coeficientes de correlación de Pearson y de Spearman

a) El Coeficiente de Contingencia:

Se define por

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

donde χ^2 es el valor calculado de la prueba de Chi-Cuadrado y n es el número de datos.

El valor de C varía entre 0 y 1. Si $C = 0$, significa que no hay asociación entre las variables. El coeficiente de contingencia tiene la desventaja de que no alcanza el valor de uno aún cuando las dos variables sean totalmente dependientes. Otra desventaja es que su valor tiende a aumentar a medida que el tamaño de la tabla aumenta.

En general, un valor de C mayor que .30, indica una buena asociación entre las variables. Sin embargo hay que tomar en consideración también el tamaño de la tabla.

Ejemplo: La siguiente tabla muestra los resultados de un estudio para mostrar la relación entre asistir a la iglesia los domingos y la ausencia a clases para jóvenes entre 13 y 18 años:

		Falta a Clases		
Va a la Iglesia		Nunca	De vez en Cuando	Frecuentemente
	Nunca	91	68	136
	De vez en Cuando	140	78	119
	Frecuentemente	296	106	90

Calcular el coeficiente de contingencia para medir la asociación entre las dos variables