Conferencia 4. Prueba de Hipótesis para dos Poblaciones

Tema 2. Inferencia Estadística.

Sumario:

- Pruebas de hipótesis para la comparación de las medias de dos poblaciones Normales.
- Pruebas de hipótesis para la comparación de las varianzas de dos poblaciones Normales
- Pruebas de hipótesis para la comparación de las proporciones de dos poblaciones Binomiales.

1. Introducción

En la práctica lo que más comúnmente se presenta son situaciones donde se poseen datos que provienen de dos o más muestras y se quieren comparar los valores de los parámetros de estas muestras.

Tal puede ser el caso que se le presenta a un investigador, cuando quiere probar que los valores de temperatura en una provincia del país son iguales a los de otra provincia, o la relación de la proporción de lluvia caída en unos meses con respecto a otros.

Para esto se realizan las pruebas de hipótesis con comparación de poblaciones.

2. Dócimas paramétricas

En este tipo de pruebas de hipótesis las dócimas paramétricas están dadas en función de los valores de los parámetros de dos poblaciones.

Llamémosle θ_1, θ_2 a los valores del parámetro θ en la población 1 y 2 respectivamente, podemos entonces definir los planteamientos siguientes:

a.	b.	c.
$H_0: \theta_1 \leq \theta_2$	$H_0: \theta_1 \geq \theta_2$	$H_0: \theta_1 = \theta_2$
$H_0: \theta_1 > \theta_2$	$H_0: \theta_1 < \theta_2$	$H_0: \theta_1 \neq \theta_2$

Basados en el criterio anterior se pueden encontrar las reglas de decisión adecuadas para los problemas de pruebas de hipótesis asociados a los parámetros $\mu\ y\ \sigma^2$ de la distribución Normal, y el parámetro p de la distribución Binomial.

La teoría relacionada con las regiones críticas, errores y estadígrafos es análoga a las pruebas de hipótesis de una población.

3. Pruebas de Hipótesis para la comparación de las medias de dos poblaciones Normales

a. Varianzas conocidas

Supongamos que se tienen dos muestras independientes $X_1, \ldots, X_{n_1}$ de X y $Y_1, \ldots, Y_{n_2}$ de Y de dos distribuciones normales de parámetros (μ_1, σ_1^2) y (μ_2, σ_2^2) respectivamente con μ_1 y μ_2 desconocidas y σ_1^2 y σ_2^2 conocidas, entonces se plantean las siguientes pruebas de hipótesis y sus respectivas reglas de decisión:

1. Hipótesis

$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

2. Estadígrafo

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

3. Región Critica (RH_0)

$Z > Z_{1-\alpha}$	$Z < -Z_{1-\alpha}$	$\lvert Z\rvert > Z_{1-\frac{\alpha}{2}}$

Ejemplo 1

Dos métodos de aprendizaje A y B deben ser comparados. De acuerdo con el criterio del especialista, el método A es superior al B. Dos muestras de 100 alumnos cada una, fueron seleccionadas aleatoriamente. Un grupo fue enseñado por el método A y el otro por el B y al final del período de aprendizaje los dos grupos fueron sometidos a un mismo examen. La puntuación promedio obtenida fue de 95 puntos para los del método A y 89 puntos para los del método B. De experiencias anteriores se sabe que estos métodos tienen varianzas de 5 y 4 respectivamente. Los datos justifican que se puede suponer una distribución Normal para las variables aleatorias X y Y que describen las calificaciones en el examen.

R/:

Datos:

μ_1: Calificación media poblacional usando A

μ_2: Calificación media poblacional usando B

H_1: Método A superior al B

$\bar{X} = 95$

$\bar{Y} = 89$

$n_1 = n_2 = 100$

$\sigma_1^2 = 5$

$\sigma_2^2 = 4$

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

1. Fijo α $\alpha = 0.05$
2. Estadígrafo

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{95 - 89}{\sqrt{\frac{5}{100} + \frac{4}{100}}} = \frac{6}{\sqrt{\frac{9}{100}}} = \frac{6}{\sqrt{0.09}} = \frac{6}{0.3} = 20$$

3. Región Critica (RH_0)

$$Z_{1-\alpha} = Z_{0.95} = 1.65$$
$$20 = Z > Z_{1-\alpha} = 1.65$$

4. Conclusión

Como $20 > 1.65$ se rechaza H_0, por lo que se acepta que el método A es superior al B con un nivel de significación del 5%.

b. Las varianzas son desconocidas pero iguales

Supongamos que se tienen dos muestras independientes $X_1, \ldots, X_{n_1}$ de X y $Y_1, \ldots, Y_{n_2}$ de Y de dos distribuciones normales de parámetros (μ_1, σ_1^2) y (μ_2, σ_2^2) respectivamente con μ_1 y μ_2 desconocidas y σ_1^2 y σ_2^2 desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2$) entonces se plantean las siguientes pruebas de hipótesis y sus respectivas reglas de decisión:

1. <u>Hipótesis</u>

$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

1. <u>Estadígrafo</u>

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1-1)S_1^2 + (n_2-1)S_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

2. <u>Región Critica (RH_0)</u>

$T_{\bar{X}-\bar{Y}} > t_{1-\alpha}(n-2)$	$T_{\bar{X}-\bar{Y}} < -t_{1-\alpha}(n-2)$	$\lvert T_{\bar{X}-\bar{Y}} \rvert > t_{1-\frac{\alpha}{2}}(n-2)$

Ejemplo 2.

Supongamos que en el ejemplo anterior no se conocen las varianzas, pero se sabe que son iguales.
Si $S_1^2 = 5.1$ $S_1^2 = 4.9$.

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1-1)S_1^2 + (n_2-1)S_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 = 2)}{n_1 + n_2}}$$

$$= \frac{95 - 89}{\sqrt{99 * 5.1 + 99 * 4.9}} \sqrt{\frac{100 * 100(100 + 100 - 2)}{100 + 100}}$$

$$= \frac{6}{\sqrt{504.9 + 485.1}} \sqrt{\frac{10000(198)}{200}} = \frac{6}{\sqrt{990}} \sqrt{\frac{19800}{2}} = \frac{6}{31.46}\sqrt{9900}$$

$$= 0.19 * 99.50 = 18.91$$

$$t_{1-\alpha}(n_1 + n_2 - 2) = t_{0.95}(198) \approx 1.65$$

$$18.91 = T_{\bar{X}-\bar{Y}} > t_{1-\alpha}(n_1 + n_2 - 2) = t_{0.95}(198) \approx 1.65$$

Como $18.91 > 1.65$ se rechaza H_0 por lo que se acepta que A es mejor que B con un nivel de significación del 5%, con una probabilidad de error de menor o igual a 0.05.

c. Las varianzas son desconocidas y diferentes

Supongamos que se tienen dos muestras independientes $X_1,\ldots,X_{n_1}$ de X y $Y_1,\ldots,Y_{n_2}$ de Y de dos distribuciones normales de parámetros (μ_1,σ_1^2) y (μ_2,σ_2^2) respectivamente con μ_1 y μ_2 desconocidas y σ_1^2 y σ_2^2 desconocidas y diferentes se plantean las siguientes pruebas de hipótesis y sus respectivas reglas de decisión:

1. Hipótesis

$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

2. Estadígrafo

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

3. Región Critica (RH_0)

$T_{\bar{X}-\bar{Y}} > t_{1-\alpha}(v)$	$T_{\bar{X}-\bar{Y}} < -t_{1-\alpha}(v)$	$\lvert T_{\bar{X}-\bar{Y}}\rvert > t_{1-\frac{\alpha}{2}}(v)$

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

Ejemplo 3.

Supongamos que en el ejemplo anterior las varianzas son desconocidas y distintas. Entonces

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{95 - 89}{\sqrt{\frac{5.1}{100} + \frac{4.9}{100}}} = \frac{6}{\sqrt{\frac{10}{100}}} = \frac{6}{\sqrt{0.1}} = \frac{6}{0.32} = 18.75$$

Fijando $\alpha = 0.05$ se tiene que $t_{1-\alpha}(v) = t_{0.95}(v)$

$$v = \left(\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)\frac{1}{n_1+1} + \left(\frac{S_2^2}{n_2}\right)\frac{1}{n_2+1}}\right) - 2 = \frac{\left(\frac{5.1}{100} + \frac{4.9}{100}\right)^2}{\frac{5.1}{100}\frac{1}{100+1} + \frac{4.9}{100}\frac{1}{100+1}} - 2$$

$$= \frac{\left(\frac{10}{100}\right)^2}{\frac{5.1+4.9}{100(100+1)}} - 2 = \frac{(0.1)^2}{\frac{10}{10100}} - 2 = \frac{0.01}{0.00099} - 2 = 10.10 - 2 \approx 8$$

Por tanto $t_{1-\alpha}(v) = t_{0.95}(v) = t_{0.95}(8) = 1.8595$

Como $18.75 > 1.86$ se rechaza H_0, por lo que se acepta que A es mejor que B con un nivel de significación del 5%.

Qué pasaría si en los ejemplos 2 y 3 no conocemos las varianzas y tampoco conocemos si estas son iguales o diferentes. Pues sería necesario realizar primero una prueba de igualdad contra diferencia en las varianzas y en dependencia del resultado de la misma realizamos la prueba de medias con varianza desconocida correspondiente.

4. Pruebas de Hipótesis para la comparación de las varianzas de dos poblaciones Normales.

Supongamos que se tienen dos muestras independientes $X_1,...,X_{n_1}$ de X y $Y_1,...,Y_{n_2}$ de Y de dos distribuciones normales de parámetros (μ_1,σ_1^2) y (μ_2,σ_2^2) respectivamente con μ_1,μ_2,σ_1^2 y σ_2^2 desconocidas entonces se plantean las siguientes pruebas de hipótesis y sus respectivas reglas de decisión:

$H_0: \sigma_1^2 \le \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 \ge \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \ne \sigma_2^2$

1. <u>Estadígrafo</u>

$$F = \frac{S_1^2}{S_2^2}$$

2. <u>Región Critica (RH_0)</u>

$F > F_{1-\alpha}(n_1-1,n_2-1)$	$F < F_{\alpha}(n_1-1,n_2-1)$	$F < F_{\frac{\alpha}{2}}(n_1-1,n_2-1)$ o $F > F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)$

Donde F_α es el percentil de la distribución F de Fisher a nivel α

Ejemplo 4

Si en el ejemplo anterior se quiere analizar si puede aceptarse o no la igualdad de las varianzas de los métodos A y B entonces realizamos la prueba

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \ne \sigma_2^2$$

Tenemos que $S_1^2 = 5.1\ S_1^2 = 4.9$.

1. <u>Fijo α</u> α = 0.05
2. <u>Estadígrafo</u>

$$F = \frac{S_1^2}{S_2^2} = \frac{5.1}{4.9} = 1.04$$

3. <u>Región Critica (RH_0)</u>

$$F_{\frac{\alpha}{2}}(n_1-1,n_2-1) = F_{\frac{0.05}{2}}(100-1,100-1) = F_{0.25}(99,99) = No\ tenemos\ tabla$$

Aplicamos Propiedad $F_\alpha(n,m) = \frac{1}{F_{1-\alpha}(m,n)}$

En este caso $F_{0.25}(99,99) = \frac{1}{F_{1-0.25}(99,99)} = \frac{1}{F_{0.75}(99,99)}$ por desgracia tampoco contamos con la tabla de $F_{0.75}$ que es Fisher al 25%.

$$F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1) = F_{1-\frac{0.05}{2}}(100-1,100-1) = F_{1-0.25}(99,99) = F_{0.75}(99,99)$$

En este caso tampoco contamos con la tabla de $F_{0.75}$ que es Fisher al 25%. Por tanto no podemos continuar con el ejercicio puesto que no podemos aceptar ni Rechazar H_0

5. Pruebas de Hipótesis para la comparación de proporciones de dos poblaciones Binomiales

Supongamos que se tienen dos muestras independientes $X_1, ..., X_{n_1}$ de X y $Y_1, ..., Y_{n_2}$ de Y donde $X \sim B(n_1, p_1), Y \sim B(n_2, p_2)$, entonces se plantean las siguientes pruebas de hipótesis y sus respectivas reglas de decisión:

$H_0: \mathrm{p}_1 \leq \mathrm{p}_2$ $H_0: \mathrm{p}_1 > \mathrm{p}_2$	$H_0: \mathrm{p}_1 \geq \mathrm{p}_2$ $H_0: \mathrm{p}_1 < \mathrm{p}_2$	$H_0: \mathrm{p}_1 = \mathrm{p}_2$ $H_0: \mathrm{p}_1 \neq \mathrm{p}_2$

1. <u>Estadígrafo</u>

$$Z_{p_1 - p_2} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{(n_1\bar{p}_1 + n_2\bar{p}_2)(n_1 + n_2 - n_1\bar{p}_1 - n_2\bar{p}_2)}{(n_1 + n_2)(n_1 * n_2)}}}$$

2. <u>Región Critica (RH_0)</u>

$Z_{p_1 - p_2} > Z_{1-\alpha}$	$Z_{p_1 - p_2} < -Z_{1-\alpha}$	$\lvert Z_{p_1 - p_2} \rvert > Z_{1-\frac{\alpha}{2}}$

Ejemplo 5

Dos grupos de 50 estudiantes fueron seleccionados aleatoriamente e interrogados acerca de la importancia de la Matemática. Lo grupos se diferenciaban en el año escolar que estaban cursando (10mo y 11no grado). En el primer grupo, 30 estudiantes atribuyeron una importancia decisiva a las Matemáticas, mientras que en el segundo 25 estudiantes atribuyeron importancia similar. ¿Se puede afirmar que los estudiantes de 10mo grado atribuyen mayor importancia a la Matemática que los de 11no grado con un nivel de significación del 1%?

R/

Datos:

Sea X: número de estudiantes de 10mo grado que atribuyen importancia a la Matemática.

$X \sim B(n_1, p_1)$

$n_1 = 50$

p_1 desconocido.

Sea Y : número de estudiantes de 11no grado que atribuyen importancia a la Matemática.

$Y \sim B(n_2, p_2)$

$n_2 = 50$

p_2 desconocido.

Se tiene que:

$\bar{p}_1 = \frac{30}{50} = 0.6$

$\bar{p}_2 = \frac{25}{50} = 0.5$

$\alpha = 0.01$

Plateamos la siguiente hipótesis

$$H_0 : \mathrm{p}_1 \leq \mathrm{p}_2$$
$$H_0 : \mathrm{p}_1 > \mathrm{p}_2$$

$$
\begin{aligned}
Z_{p_1 - p_2} &= \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{(n_1\bar{p}_1 + n_2\bar{p}_2)(n_1 + n_2 - n_1\bar{p}_1 - n_2\bar{p}_2)}{(n_1+n_2)(n_1 * n_2)}}} \\
&= \frac{0.6 - 0.5}{\sqrt{\frac{(50*0.6+50*0.5)(50+50-50*0.6-50*0.5)}{(50+50)(50*50)}}} \\
&= \frac{0.1}{\sqrt{\frac{(30+25)(50+50-30-25)}{100*2500}}} \\
&= \frac{0.1}{\sqrt{\frac{55(100-55)}{250000}}} \\
&= \frac{0.1}{\sqrt{\frac{55*45}{250000}}} \\
&= \frac{0.1}{\sqrt{\frac{2475}{250000}}} \\
&= \frac{0.1\sqrt{250000}}{\sqrt{2475}} = \frac{0.1\sqrt{250000}}{49.75} = \frac{0.1*500}{49.75} \\
&= \frac{50}{49.75} \\
&= 1.005
\end{aligned}
$$

Como $\alpha \ = \ 0.01$ entonces $Z_{1-\alpha} = Z_{0.99} \ = \ 2.33$

Como $1.005 < 2.33$ se rechaza H_0, por lo que los alumnos de $10mo$ grado conceden mayor importancia a las Matemáticas que los de $11no$ grado.