Comparación de las medias para varias muestras. Varias variables y varias muestras

Sean $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \lambda_p \geq 0$ los valores propios de $\mathbf{W}^{-1}\mathbf{B}$, donde

B=T-W se denomina matriz-entre-muestras de sumas de cuadrados y productos cruzados, porque la entrada típica es la diferencia entre una suma total de cuadrados o productos cruzados menos el término correspondiente dentro-de-las-muestras.

Entonces lambda de Wilks también puede expresarse como

$$\Lambda = \prod_{i=1}^{p} 1/(1+\lambda_i)$$

Comparación de las medias para varias muestras. Varias variables y varias muestras

- **Prueba de Roy**

El segundo estadígrafo es el mayor valor propio de $\mathbf{W}^{-1}\mathbf{B}$

Fundamento: si se encuentra la combinación lineal de las variables X_1 a X_p que maximiza la razón entre la suma de cuadrados entre-muestras y la suma de cuadrados dentro-de-la-muestra, entonces esta razón máxima es igual a λ_1. Esto implica entonces que este máximo valor propio debe ser un buen estadígrafo para probar si la variación entre-muestras es significativamente grande y que, por lo tanto, existe evidencia de que las muestras que se consideran no provienen de poblaciones con el mismo vector medio.

Este enfoque está relacionado con el análisis de la función discriminante.

Comparación de las medias para varias muestras. Varias variables y varias muestras

Software

Algunos programas informáticos llaman estadígrafo de Roy a $\lambda_1/(1-\lambda_1)$ en lugar de λ_1. En caso de duda, consulte la documentación del programa. Para evaluar si λ_1 es significativamente grande, se puede calcular numéricamente la probabilidad exacta de un valor tan grande como el observado, o se puede utilizar una distribución F para encontrar un límite inferior al nivel de significación. Los usuarios de los paquetes informáticos deben ser conscientes de cuál de estas alternativas se utiliza si se obtiene un resultado significativo. Esto se debe a que si se utiliza la distribución F, entonces el valor de λ_1 puede no ser significativamente grande para el nivel de significancia elegido. El valor F utilizado se describe en la tabla 4.4.

Comparación de las medias para varias muestras. Varias variables y varias muestras

- **Estadígrafo de trazas de Pillai**

Se puede formular en términos de valores propios como:

$$V = \sum_{i=1}^{p} \lambda_i/(1+\lambda_i)$$

Nuevamente, los grandes valores de este estadígrafo proporcionan evidencia de que las muestras que se consideran provienen de poblaciones con diferentes vectores medios. Una aproximación del nivel de significación (la probabilidad de obtener un valor tan grande o más grande que V, si las muestras provienen de poblaciones con el mismo vector medio) se proporciona en la Tabla 4.4

Comparación de las medias para varias muestras. Varias variables y varias muestras

❑ Estadígrafo de trazas de Lawes-Hotelling

Se puede formular en términos de valores propios como:

$$U = \sum_{i=1}^{p} \lambda_i$$

es decir, la suma de los valores propios de la matriz **$W^{-1}B$.**

Una vez más, los grandes valores proporcionan evidencia contra la hipótesis nula, con una prueba F aproximada proporcionada en la Tabla 4.4.

Comparación de las medias para varias muestras. Varias variables y varias muestras

Las 4 pruebas:

- Proporcionen niveles de significación similares, por lo que no hay necesidad real de elegir entre ellas.
- Suponen que la distribución de las p variables es multinormal, con la misma matriz de covarianza dentro-de-las-muestras para las m poblaciones de las que se extraen las muestras.
- Son bastante robustas si los tamaños de las muestras, son iguales o casi, para las m muestras.
- Si se cuestiona la multinormalidad o la igualdad de matrices de covarianza, los estudios realizados sugieren que el estadígrafo de Pillai puede ser más robusto.

Comparación de las varianzas para varias muestras

- **Prueba M de Box**

Es la prueba más conocida para comparar la varianza en varias muestras. Ya fue mencionada para dos muestras con varias variables, con dos o más muestras. Para **m muestras**, el **estadígrafo M** es:

$$M = \left\{\prod_{i=1}^{m} |C_i|^{(n_i-1)/2}\right\} / |C|^{(n-m)/2}$$

$$C = \sum_{i=1}^{m} (n_i - 1) C_i / (n - m)$$

donde $\mathbf{n_i}$ es el tamaño de la i-ésima muestra

$\mathbf{C_i}$ es la covarianza de la muestra para la i-ésima muestra

C es la matriz ponderada de covarianza

n$=\sum n_i$ es el número total de observaciones

Comparación de las varianzas para varias muestras

Valores grandes de M proporcionan evidencia de que las muestras no provienen de poblaciones con la misma matriz de covarianza. Una prueba F aproximada para saber si un M-valor observado es significativamente grande consiste en calcular:

$$\boldsymbol{F = -2b\ ln(M)} \quad \boldsymbol{\text{válido para} \quad c_2 > c_1^2}$$

$v_1 = \frac{p(p+1)(m-1)}{2}$; $v_2 = \frac{v_1+2}{c_2-c_1^2}$ $\quad son\ los\ g.d.l.$

$b = (1 - c_1 - \frac{v_1}{v_2})/v_1$

$c_1 = (2p^2 + 3p - 1)\left\{\sum_{i=1}^{m} \frac{1}{n_i - 1} - \frac{1}{n-m}\right\}/\{6(p+1)(m-1)\}$

$c_2 = (p-1)(p+2)\left\{\sum_{i=1}^{m} \frac{1}{(n_1-1)^2} - \frac{1}{(n-m)^2}\right\}/\{6(m-1)\}$

Comparación de las varianzas para varias muestras

$$F = \frac{\{2b_1 v_2 \ln(M)\}}{\{v_1 + 2b_1 \ln(M)\}}$$ **válido para** $c_2 < c_1^2$

donde $b_1 = (1 - c_1 - \frac{2}{v_2})/v_2$

Esto se prueba contra la distribución F con v_1 y v_2 g.d.l. para ver si es significativamente grande.

Comparación de las varianzas para varias muestras

Se sabe que **la prueba de Box es sensible a las desviaciones de la normalidad en la distribución de las variables consideradas**. Por esta razón, aquí se recomiendan alternativas robustas a la prueba de Box, que son generalizaciones de lo que se sugirió para la situación de dos muestras. Por lo tanto, se pueden calcular desviaciones absolutas de medianas de muestra para los datos en m muestras. Para **una sola variable**, éstas pueden ser tratadas como las observaciones para un análisis de varianza de un factor. Una razón F significativa es entonces evidencia de que las muestras provienen de poblaciones con diferentes desviaciones medias, es decir, poblaciones con diferentes matrices de covarianza. Con **más de una variable**, cualquiera de las cuatro pruebas descritas anteriormente puede aplicarse a los datos transformados, y un resultado significativo indica que la matriz de covarianza no es constante para las m poblaciones muestreadas.

Comparación de las varianzas para varias muestras

Alternativamente, las variables pueden estandarizarse para tener varianza unitaria para todos los datos agrupados juntos, y se pueden calcular los valores de d. Estos valores d pueden entonces ser analizados por un análisis de varianza de un factor. Esto **generaliza la prueba de Van Valen**, que se sugirió para comparar la variación en dos muestras multivariantes. Una proporción de F significativa a partir del análisis de varianza indica que algunas de las m poblaciones muestreadas son más variables que otras. Como en la situación de dos muestras, esta prueba sólo es realmente apropiada cuando algunas muestras pueden ser más variables que otras para todas las mediciones que se están considerando.

Comparación de las varianzas para varias muestras. Ejemplo: comparación de las muestras de los cráneos egipcios

Datos: 4 mediciones en cráneos egipcios masculinos para 5 muestras de varias edades pasadas

Análisis de la varianza de 1-factor

4 y 145 g.d.l.

Variable	F	Nivel de significancia
X_1	5,95	0,1 %
X_2	2,45	5 %
X_3	8,31	0,1 %
X_4	1,51	No significativo

Clara evidencia de que las **medias** de X_1, X_2 y X_3 cambiaron con el tiempo.

Comparación de las varianzas para varias muestras. Ejemplo: comparación de las muestras de los cráneos egipcios

Consideremos ahora las 4 variables en forma conjunta. Si se combinan las 5 muestras, entonces la matriz T de sumas de cuadrados y productos para las 150 observaciones y la matriz W de sumas de cuadrados y productos cruzados dentro-de-las-muestras son:

$$T = \begin{bmatrix} 3563{,}89 & -222{,}81 & -615{,}16 & 426{,}73 \\ -222{,}81 & 3635{,}17 & 1046{,}28 & 346{,}47 \\ -615{,}16 & 1046{,}28 & 4309{,}27 & -16{,}40 \\ 426{,}73 & 346{,}47 & -16{,}40 & 1533{,}33 \end{bmatrix} \qquad |T| = 7{,}306 \times 10^{13}$$

$$W = \begin{bmatrix} 3061{,}07 & 5{,}33 & 11{,}47 & 291{,}30 \\ 5{,}33 & 3405{,}27 & 754{,}00 & 412{,}53 \\ 11{,}47 & 754{,}00 & 3505{,}97 & 164{,}33 \\ 291{,}30 & 412{,}53 & 164{,}33 & 1472{,}13 \end{bmatrix} \qquad |W| = 4{,}848 \times 10^{13}$$

El **estadígrafo lambda de Wilks** es:

$\boldsymbol{\Lambda}=|\boldsymbol{W}|/|\boldsymbol{T}|=0{,}6636$

Comparación de las varianzas para varias muestras. Ejemplo: comparación de las muestras de los cráneos egipcios

- La prueba F para evaluar si este valor es significativamente pequeño se proporciona en la Tabla 4.4.

p = 4 variables, m = 5 muestras y n = 150 observaciones

$df_1 = p(m-1) = 16$

$w = n - 1 - \frac{p+m}{2} = 150 - 1 - \frac{4+5}{2} = 144{,}5$

$t=[(df_1^2-4)/\{p^2+(m-1)^2-5\}]^{1/2} = [(16^2-4)/\{4^2+(5-1)^2-5\}]^{1/2}=3{,}055$

$df_2 = wt - \frac{df_1}{2} + 1 = 144{,}5x3{,}055 - \frac{16}{2} + 1 = 434{,}5$

$F=\{(1-\Lambda^{1/t})/\Lambda^{1/t}\}({}^{df_2}/_{df_1})=\{(1-0{,}6636^{1/3{,}055})/0{,}6636^{1/3{,}055}\}(434{,}5/16) = 3{,}90$

Con 16 y 434,5 df (g.d.l.) Esto es significativamente grande para el nivel de 0,1% (p <0,001). Por lo tanto, evidencia que el vector de valores medios de las 4 variables cambió con el tiempo

Comparación de las varianzas para varias muestras. Ejemplo: comparación de las muestras de los cráneos egipcios

Prueba de Roy

El máximo valor propio de la matriz $\mathbf{W^{-1}B}$ es $\lambda_1=0{,}4251$.

$F=(df_2/df_1)\ \lambda_1=(140/4)0{,}4251=14{,}88$

Esto es nuevamente significativamente grande ($p<0{,}001$).

Estadígrafo de trazas de Pillai

$V=0{,}3533$

$F=(n-m-p+s)V/\{d(s-V)\}=3{,}51$

Con $sd=16$ y $s(n-m-p+s)=580$ g.d.l.

Este es otro resultado muy significativo ($p<0{,}0001$)

Comparación de las varianzas para varias muestras. Ejemplo: comparación de las muestras de los cráneos egipcios

Estadígrafo de trazas de Lawley-Hotelling

$s=4$, $A=-0{,}5$, $B=70$ y $U=0{,}4818$

$df_1=s(2^a+s+1)=16$ y $df_2=2(sB+1)=562$

con lo que:

$F=df_2U/(sdf_1)=(560x0{,}4818)/(4x16)=4{,}23$

Nuevamente, otro resultado muy significativo ($p<0{,}0001$)

Comparación de las varianzas para varias muestras. Ejemplo: comparación de las muestras de los cráneos egipcios

Prueba M de Box

$M = 2,869 \times 10^{-11}$ $b=0,035$

$F=-2\ b\ \ln(M) = 1,14$ con $v_1=40$ y $v_2=46,379$ g.d.l.

Esto no es en absoluto significativamente grande (p=0,250), por lo que esta prueba no evidencia que la matriz de covarianza haya cambiado con el tiempo.

En suma, existe una evidencia muy fuerte de que los valores de las medias cambian con el tiempo para las cuatro variables que se consideran, sin embargo no hay evidencia de que la varianza haya cambiado.

Pruebas de significancia con datos multivariados - RESUMEN

- ✓ La **prueba t** compara los **valores medios** de dos muestras de poblaciones con **varianzas iguales**. La **prueba de Welch** aplica a situaciones en las que las **varianzas** de población **no son iguales**. Ambas pruebas suponen que **los valores de los datos se distribuyen normalmente.**
- ✓ La **prueba T^2 de Hotelling** generaliza la prueba t cuando se mide más de una variable para dos muestras.
- ✓ Se discutió el problema de la **prueba múltiple** (es decir, la tasa de error tipo-1 que se infla si se llevan a cabo varias pruebas univariadas al mismo tiempo), y la ventaja de usar una **prueba multivariante** en su lugar.

Pruebas de significancia con datos multivariados - RESUMEN

- ✓ La **prueba F** y la **prueba** robusta **de Levene** comparan la varianza de dos muestras.
- ✓ La **prueba M de Box**, una alternativa robusta basada en el enfoque utilizado en el test de Levene, y la **prueba de Van Valen**, comparan la **varianza de dos muestras**.
- ✓ 4 pruebas para comparar los **vectores medios de varias muestras multivariantes.** Estas son: **estadígrafo lambda de Wilks, la prueba** del mayor valor propio **de Roy y los estadígrafos de trazas de Pillai y de Lawes-Hotelling.** Se presentaron además **pruebas F** aproximadas para evaluar la importancia de las estadísticas involucradas en cada caso.

Pruebas de significancia con datos multivariados - RESUMEN

- ✓ La **prueba M de Box** también se utiliza para comparar la **varianza de varias muestras multivariantes.** Se observa que es muy sensible a la suposición de que los datos multivariados están normalmente distribuidos. Por lo tanto, se describieron alternativas robustas utilizando el principio de la prueba de Levene.