Tema 4. Contrastes de hipótesis paramétricos para varias muestras independientes

Andrés M. Alonso (Profesor - Grupos 27 y 28)
andres.alonso@uc3m.es

Grado en Estadística y Empresa
Curso 2020/21

Contenidos

Introducción: Contrastes para múltiples muestras

En este tema vamos a abordar el problema de homogeneidad a partir de k muestras independientes:

$$\begin{aligned} \text{Muestra de } Y_1 &: \{y_{11}, \ldots, y_{1n_1}\} \\ \text{Muestra de } Y_2 &: \{y_{21}, \ldots, y_{2n_2}\} \\ &\vdots \\ \text{Muestra de } Y_k &: \{y_{k1}, \ldots, y_{kn_k}\} \end{aligned}$$

Como se trata de k muestras independientes, los tamaños de cada muestra , n_1, n_2,...,n_k, pueden ser diferentes.

Introducción: Contrastes para múltiples muestras

Ejemplo

Se toman medidas del peso de un tipo de estorninos en 4 regiones para examinar si existen diferencias entre las variedades de cada región:

	Peso									
Loc. 1	78,	88,	87,	88,	83,	82,	81,	80,	80,	89
Loc. 2	78,	78,	83,	81,	78,	81,	82,	76,	76	
Loc. 3	79,	73,	79,	75,	77,	78,	80,	78,	83,	84
Loc. 4	77,	69,	75,	70,	74,	83,	80			

Contrastes paramétricos ANOVA

Suponemos que las k variables son normales con la misma varianza:

$$Y_1 \sim N\left(\mu_1, \sigma^2\right)$$
$$Y_2 \sim N\left(\mu_2, \sigma^2\right)$$
$$\vdots$$
$$Y_k \sim N\left(\mu_k, \sigma^2\right)$$

Queremos resolver el siguiente contraste:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \text{Alguna media es diferente}$$

Contrastes paramétricos ANOVA

De este modo, se puede expresar que cada observación es:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

donde:

- y_{ij} representa la observación j-ésima del grupo i.
- μ_i es la media del grupo i.
- ϵ_{ij} es el error de la la observación j-ésima del grupo i.

Se asume que los errores son normales, independientes con la misma varianza:

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

Contrastes paramétricos ANOVA

- El Análisis de la Varianza (ANOVA) decide si los grupos son iguales comparando la distancia entre las medias en función de varianza de los grupos.
- Grupos con la misma diferencia de medias serán probablemente distintos si sus datos tienen menos variabilidad.

Ejemplo

Pintar un gráfico que presente los boxplots de los pesos de cada región.

Contrastes paramétricos ANOVA

Calculamos la media de cada grupo y la media total:

$$\text{Muestra de } Y_1 : \{y_{11}, \dots y_{1n_1}\} \to \bar{y}_{1\cdot} = \frac{\sum_{j=1}^{n_1} y_{1j}}{n_1}$$

$$\text{Muestra de } Y_2 : \{y_{21}, \dots y_{2n_2}\} \to \bar{y}_{2\cdot} = \frac{\sum_{j=1}^{n_2} y_{2j}}{n_2}$$

$$\dots$$

$$\text{Muestra de } Y_k : \{y_{k1}, \dots y_{kn_k}\} \to \bar{y}_{k\cdot} = \frac{\sum_{j=1}^{n_k} y_{kj}}{n_k}$$

$$\text{Toda la muestra} : \{y_{11}, \dots, y_{kn_k}\} \to \bar{\bar{y}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}}{n_1 + \dots + n_k}$$

Contrastes paramétricos ANOVA

Vemos que cada observación es:

$$y_{ij} - \overline{\overline{y}} = (y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \overline{\overline{y}})$$

Luego, elevando al cuadrado y sumando para todas las observaciones:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \overline{\overline{y}})^2 + \underbrace{2\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \overline{\overline{y}})}_{=0}$$

Contrastes paramétricos ANOVA

El primer término se llamavariación total o suma de cuadrados total (TSS):

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{\overline{y}}\right)^2$$

El segundo término se llama variación explicada o suma de cuadrados explicado por el factor (FSS):

$$FSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(\bar{y}_{i\cdot} - \overline{\overline{y}}\right)^2$$

y el último término se llama variación no explicada o suma de cuadrados residual (RSS):

$$RSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

De modo que:

$$TSS = FSS + RSS$$

Contrastes paramétricos ANOVA

El estadístico de contraste es:

$$\frac{\frac{FSS}{k-1}}{\frac{RSS}{n-k}} \sim_{H_0} F_{k-1,n-k}$$

Toda la información se resume en la tabla ANOVA:

Fuentes	S. Cuadrados	g^{os} Lib.	Varianzas	F
Factor	$FSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(\bar{y}_{i\cdot} - \bar{\bar{y}}\right)^2$	$k-1$	$\hat{s}_F^2 = \frac{FSS}{k-1}$	$\frac{\hat{s}_F^2}{\hat{s}_R^2}$
Residual	$RSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2$	$n-k$	$\hat{s}_R^2 = \frac{RSS}{n-k}$	
Total	$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2$	$n-1$	$\hat{s}_y^2$	

$$\text{p-valor} = \Pr\left(F_{k-1,n-k} > \frac{\hat{s}_F^2}{\hat{s}_R^2}\right)$$

Contrastes paramétricos ANOVA

Ejemplo

Contrastar la hipótesis de que haya diferencias entre las medias del peso de los estorninos en las distintas localidades.

Test HSD de Tukey

Cuando se ha rechazado la hipótesis de igualdad de medias con el test ANOVA, el interés está en averiguar cuál o cuáles pares de medias son diferentes entre sí.

Se podría pensar en realizar un t-test de igualdad de medias para cada par de variables. Pero el problema es que si realizamos múltiples contrastes de hipótesis se incrementa la probabilidad de cometer un error de tipo I.

Esto es porque podemos haber realizado un error de tipo I en cualquiera de los contrastes y entonces, la probabilidad total de error de tipo I está acotada por la suma de todas las α, es decir, $\frac{k(k-1)}{2}\alpha$.

Test HSD de Tukey

La solución es usar el test HSD (Honestly-significant-difference) de Tukey que es un test de comparaciones múltiples que contrasta simultáneamente para todos los pares (i, j):

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j$$

El estadístico de contraste es:

$$\frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{\frac{\hat{s}_R}{\sqrt{2}}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \sim_{H_0} q_{k,n-k}$$

donde q es la distribución del rango estudentizado.

Ejemplo

Realizar el test de TukeyHSD de comparaciones múltiples para ver qué pares de medias del peso de los estorninos en las distintas localidades son diferentes dos a dos.