Estadística No Paramétrica

Clase 11: Test Kruskal-Wallis

Joaquin Cavieres G.

Ingeniería en Estadística

Facultad de Ciencias, Universidad de Valparaíso

Test para varias muestras independientes

El siguiente contenido del curso esta relacionado a analizar test de hipótesis para varias muestras independientes

Test para varias muestras independientes

De aquí en adelante nosotros queremos testear un H_0 contra la alternativa general de que al menos dos de los efectos del tratamiento no son iguales, es decir, en forma general:

$$H_1 : [\tau_1, ..., \tau_k \quad \text{no son todos iguales}]$$

Test de Kruskal-Wallis

El test no paramétrico de Kruskal-Wallis (de William Kruskal y W. Allen Wallis) sirve para para probar si un grupo de datos proviene de la misma población. Es muy similar al test de ANOVA con los datos reemplazados por categorías. Tambien se puede considerar como una extensión de la prueba de la U de Mann-Whitney para 3 o más grupos.

Test de Kruskal-Wallis

- Como el test es uno no paramétrico este no asume normalidad en los datos observados, en contrario al supuesto del test de ANOVA.
- Se asume que bajo H_0 los datos provienen de la misma población (misma distribución)
- Generalmente no se cumple el segundo supuesto cuando observados con una varianza no constante (heterocedásticos)

El test de Kruskal-Wallis emplea rangos para contrastar la hipótesis de que k muestras han sido obtenidas de una misma población.

Test de Kruskal-Wallis

De forma resumida y bajo ciertas condiciones, el test de Kruskal-Wallis compara las mediandas muestrales, por tanto el contraste de hipótesis puede plantearse como:

Contraste de Hipótesis

H_0 :Las muestras provienen de una misma población (distribución).

H_1 :Las muestras no provienen de una misma población (distribución)

Test de Kruskal-Wallis

Para diferenciar cuando se debe utilizar un Test de ANOVA y un test de Kruskal-Wallis vea el siguiente ejemplo:

Si queremos encontrar diferencias entre colectivos y microbuses en hacer el 'recorrido' entre Valparaíso-Viña. Una opción podría ser comparar los tiempos entre cada uno de ellos (ANOVA) o los tiempos de cada uno de ellos (Kruskal-Wallis).

El test de Kruskal-Wallis es el test adecuado cuando los datos tienen un orden natural, es decir, para darles un sentido tienen que estar ordenados o bien cuando no se satisfacen las condiciones para poder aplicar un test de ANOVA.

Test de Kruskal-Wallis

Supuestos

- No es necesario que las muestras las cuales estamos comparando provengan de una distribución Normal.
- Bajo H_0 se asume que todos los grupos pertenecen a una misma población, por lo tanto, tienen las mismas medianas. Dado lo anterior todos los grupos deben tener la misma varianza.
- Los grupos deben tener la misma distribución.

Test de Kruskal-Wallis

Si los supuestos se cumplen entonces el estadístico H se compara con:

- Si $k = 3$ y el número de observaciones en $k < 5$, se recurre a tablas tabuladas con valores teóricos de H.
- En el resto de casos se asume que el estadístico H sigue una distribución χ^2 con $k - 1$ grados de libertad (siendo k el número de grupos a comparar).

Procedimiento

Para calcular el estadístico H primero combinamos todas las N observaciones de las k muestras y las ordenamos de menor a mayor, esto es: sea r_{ij} el rango de X_{ij} en esta clasificación conjunta y se establece que:

$$R_j = \sum_{i=1}^{n_j} r_{ij} \text{ y } R_{.j} = \frac{R_j}{n_j}, \quad j = 1, ..., k$$

Así, por ejemplo, R_1 es la suma de los rangos conjuntos recibidos por las observaciones del tratamiento 1 y $R_{.1}$ es el rango promedio para estas mismas observaciones.

Test de Kruskal-Wallis

Procedimiento

- Se dispone de k grupos cada uno con N observaciones.
- Se ordenan todas las observaciones de menor a mayor y se le asigna a cada una de ellas su rango,
- Cuando se obtenga la suma de rangos para cada uno de los grupos (R_j) es de esperar que bajo H_0 todos los grupos tengan un valor similar. Cumpliendose lo anterior el test H se calcula como:

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1),$$

Test de Kruskal-Wallis

Procedimiento

Entonces:

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1),$$

para testear que:

$$H_0 : [\tau_1 = ... = \tau_k]$$
$$H_1 : [\tau_1, ..., \tau_k \text{no son iguales}]$$

Por lo tanto, con nivel de significancia α, se rechaza H_0 si $H_0 \geq h_\alpha$, de otra manera no rechazar H_0.

Approximación mediante un n grande

Cuando H_0 se cumple, entonces el estadístico H tiene un $min(n_1, ..., n_k)$ tendiendo al infinito y asintóticamente distribido χ^2 con $k-1$ grados de libertad. La aproximación mediante χ^2 es:

$$\text{Rechazar } H_0 \text{ si } H \geq \chi^2_{k-1,\alpha}; \text{ de otra manera no rechazar } H_0$$

Comparación *post-hoc*

Si el test de Kruskal-Wallis es significativo, entonces al menos dos grupos de entre los comparados son significativamente diferentes, pero no sabemos cual de ellos. Por lo tanto, es necesario comparar todos los grupos, lo que nos lleva a realizar una correción del nivel de significancia para evitar cometer el error de Tipo I. Existen dos tipos de métodos post-hoc que son los más comunmente utilizados:

- Test de Mann-Whitney: Para cada par de grupos ls corrección de significancia mediante la función `pairwise.wilcox.test()`.
- Tukey's range test: eMediante la función `kruskalmc()` de la librería `pgirmess`.

Observación

Bajo las principales condiciones de utilizar el test de Kruskal-Wallis, el vector de rangos $\boldsymbol{R}^* = (r_{11}, ..., r_{n_1 1}, r_{22}, ..., r_{n_2 2}, ..., r_{1k}, ..., r_{n_k k}$ tiene una distribución Uniforme a través de todas las $N!$ permutaciones del vector de enteros $(1, 2, ..., N)$. Esto da como resultado que:

$$\mathbb{E}_0(r_{ij}) = \frac{1}{N!}(N-1)! \sum_{i=1}^{N} \frac{N+1}{2},$$

el rango promedio que se asigna en el rango conjunto de las observaciones para todos los grupos.

Observación

Así,

$$\mathbb{E}_0(R_{.j}) = \mathbb{E}_0\left(\frac{1}{n_j}\sum_{i=1}^{n_j} r_{ij}\right) = \frac{n_j(N+1)}{2n_j} = \frac{N+1}{2}, \quad j = 1, 2, ...k,$$

por lo que nosotros deberíamos esperar que $R_{.j}$ sea bastante cercano a $(N+1)/2$ bajo H_0.

Test de Kruskal-Wallis

Ejemplo

En un estudio se comparan las cosechas de paltas 3 regiones distintas. ¿Existen diferencias significativas dependiendo de las condiciones climáticas de cada zona?

Ver ejemplo en `R`

Test de Kruskal-Wallis

Ejemplo 2

Thomson y Short (1969) han evaluado la eficiencia mucociliar a partir de la velocidad de eliminación del polvo en sujetos normales, sujetos con enfermedad obstructiva de las vías respiratorias y sujetos con asbestosis. La siguiente tabla se basa en un subconjunto de los datos de Thomson-Short. Los rangos conjuntos (r_{ij}) de las observaciones se dan entre paréntesis después de que los valores de los datos y las sumas de los rangos de tratamiento (R_1, R_2 y R_3) se proporcionan en la parte inferior de las columnas.

	Sujetos con:	
Sujetos normales	Enfermedad respiratoria	Asbestosis
2.9 (8)	3.8 (13)	2.8 (7)
3.0 (9)	2.7 (6)	3.4 (11)
2.5 (4)	4.0 (14)	3.7 (12)
2.6 (5)	2.4 (3)	2.2 (2)
3.2 (10)		2.0 (1)
$R_1 = 36$	$R_2 = 36$	$R_3 = 33$