Tests para varias muestras independientes
Análisis de la varianza de un factor

Estudiaremos el problema de posición en varias muestras. Kruskal y Wallis (1952) extendieron el test de Mann-Whitney a k muestras independientes (k > 2). La situación experimental es aquella en que k muestras aleatorias son obtenidas de k poblaciones posiblemente diferentes y se desea testear la hipótesis nula de que todas las poblaciones tienen idéntica mediana.

Test de Kruskal-Wallis: El modelo consiste en k muestras aleatorias independientes de tamaños $n_1, n_2, ..., n_k$:

$$X_{11}, X_{12}, ..., X_{1n_1}$$
$$X_{21}, X_{22}, ..., X_{2n_2}$$
$$.....$$
$$X_{k1}, X_{k2}, ..., X_{kn_k}$$

donde $X_{ij} \sim F(x-\theta_i)$ y $F \in \Omega_o$.

En el caso Normal correspondería al modelo $X_{ij} \sim \theta_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0,\sigma^2)$ independientes.

Hipótesis a testear:

H_o: $\theta_1 = \theta_2 = ... = \theta_k$ vs H_1: existe al menos un par (i,j) tal que $\theta_i \neq \theta_j$.

Estadístico del test: Sea $N = \sum_{i=1}^{k} n_i$. Se ordenan los N datos combinados y se asigna a cada una de las observaciones su correspondiente rango: $R_{ij} = R(X_{ij})$, $1 \le i \le k, 1 \le j \le n_i$.

Se calcula la suma y el promedio de los rangos en cada población:

$$R_{i.} = \sum_{j=1}^{n_i} R_{ij} \qquad \overline{R}_{i.} = R_{i.} \Big/ n_i$$

En el problema de dos muestras la suma de los rangos es $R_{1.} + R_{2.} = \frac{N(N+1)}{2}$ y el estadístico del test se basa en $R_{1.}$ pues éste contiene toda la información necesaria. En el caso de k muestras también la suma de los rangos satisface

$$\sum_{i=1}^{k} R_{i.} = \frac{N(N+1)}{2}$$

por lo tanto sería suficiente basarse en k-1 poblaciones. Sin embargo se utilizará el conjunto completo: $R_{1.}, R_{2.}, ..., R_{k.}$, pero deberá tenerse en cuenta su dependencia al estudiar la distribución del estadístico.

El estadístico del test se define como

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right) \qquad (1)$$

donde

$$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right)$$

Si no hay empates, $S^2 = \frac{N(N+1)}{12}$ y el estadístico del test se reduce a

$$T = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_{i.}^2}{n_i} - 3(N+1)$$

Distribución nula del estadístico: Para obtener la distribución exacta y asintótica del estadístico *T* bajo H_o necesitamos algunos resultados.

Teorema 1: Bajo H_o , podemos suponer que las k muestras provienen de una población con distribución $F \in \Omega_o$, y

$$P(R_{ij} = s) = \frac{1}{N} \qquad 1 \le s \le N$$

$$P(R_{ij} = s, R_{rl} = t) = \begin{cases} \dfrac{1}{N(N-1)} & \text{si } s \ne t, (i,j) \ne (r,l) \\ 0 & \text{en caso contrario} \end{cases}$$

$$E(R_{i.}) = n_i\,\frac{N+1}{2} \qquad V(R_{i.}) = n_i(N-n_i)(N+1)/12$$

$$\operatorname{cov}(R_{i.}, R_{r.}) = -n_i n_r (N+1)/12 \qquad \text{si } i \ne r$$

Demostración: Se demuestra considerando a $X_{i1}, X_{i2}, ..., X_{in_i}$ como una muestra y al resto de las observaciones como otra muestra y aplicando los teoremas demostrados para el test de Mann-Whitney.

Observemos ahora que $\left(\overline{R}_{i.} - \frac{N+1}{2}\right)$ representa la diferencia del rango promedio de la muestra i respecto de su media bajo H_o. Rechazaríamos la hipótesis nula si la suma de los alejamientos es grande y ésto sugiere definir el estadístico

$$T = \sum_{i=1}^{k} c_{iN}^2 \left(\frac{\overline{R}_{i.} - \frac{N+1}{2}}{\sqrt{V(\overline{R}_{i.})}} \right)^2$$

donde las constantes c_{iN}^2 se eligen de manera que, bajo H_o, $T \xrightarrow{d} \chi^2_{k-1}$. Estas constantes no son iguales a 1 debido a la dependencia de los $\overline{R}_{i.}$.

La distribución exacta debe obtenerse para cada valor de k, $n_1,...n_k$ y suponiendo que, bajo H_o, las $\frac{N!}{\prod_{i=1}^{k} n_i!}$ configuraciones son igualmente probables. En el ejercicio 1 de la Práctica 5 lo harán para el caso k = 3, $n_1 = 3$, $n_2 = 2$ y $n_3 = 1$ y suponiendo que no hay empates. Lo que haremos a continuación es justificar la distribución nula asintótica del estadístico.

Teorema 2: Supongamos que $F \in \Omega_o$ y $n_i \to \infty$ de modo tal que $n_i / N \to \lambda_i, 0 < \lambda_i < 1$ y además que $c_{iN} \to c_i$. Entonces bajo H_o,

$$V = \begin{pmatrix} V_1 \\ V_2 \\ \dots \\ V_k \end{pmatrix} \xrightarrow{d} N(0, B)$$

donde

$$V_i = c_{iN} \frac{1}{\sqrt{N}} \left(\overline{R}_{i.} - \frac{N+1}{2} \right) \qquad B_{ij} = \begin{cases} \frac{c_i^2 (1 - \lambda_i)}{12\lambda_i} & \text{si } i = j \\ -\frac{c_i c_j}{12} & \text{si } i \neq j \end{cases}$$

Demostración: Hettmansperger, pag.182.

Lema: Sea $Z \sim N(0, A)$, $Z \in \Re^k$ y A una matriz simétrica idempotente de rango r, entonces

$$\sum_{i=1}^{k} Z_i^2 \sim \chi_r^2$$

Nota: Recordemos que A es idempotente si y sólo si $A^2 = A$ y que si A es idempotente y simétrica de rango r todos sus autovalores son 1 ó 0 y tiene r autovalores iguales a 1.

Considerando $c_{iN} = \sqrt{1 - \frac{n_i}{N}}$ que tiende a $\sqrt{1 - \lambda_i}$, se demuestra el siguiente Teorema.

Teorema 3: Bajo H_o,

$$T = \sum_{i=1}^{k} \left(1 - \frac{n_i}{N}\right) \left[\frac{\overline{R}_{i.} - \frac{N+1}{2}}{\sqrt{\frac{(N-n_i)(N+1)}{12 n_i}}} \right]^2 = \left\{ \frac{12}{N(N+1)} \sum_{i=1}^{k} \left(R_{i.}^2 \Big/ n_i \right) \right\} - 3(N+1) \xrightarrow{d} \chi_{k-1}^2$$

Demostración: Hettmansperger, pag.184.

Si hay empates, se modifica la $V(\overline{R}_{i.})$ y se obtiene el estadístico (1).

Zona de rechazo: Se rechaza H_o si $T > \chi_{k-1,\alpha}^2$ siendo $\chi_{k-1,\alpha}^2$ el valor crítico α de la distribución chi-cuadrado con k-1 grados de libertad.

Ejemplo: Se desea comparar 4 métodos de cultivo de maíz, para lo cual se aplican en diferentes parcelas de terreno y se mide el rendimiento. Los resultados obtenidos se presentan en la siguiente tabla, juntamente con los rangos de cada observación en la muestra combinada.

Método 1		**Método 2**		**Método 3**		**Método 4**	
Observ.	Rango	Observ.	Rango	Observ.	Rango	Observ.	Rango
83	11	91	23	101	34	78	2
91	23	90	19.5	100	33	82	9
94	28.5	81	6.5	91	23	81	6.5
89	17	83	11	93	27	77	1
89	17	84	13.5	96	31.5	79	3
96	31.5	83	11	95	30	81	6.5
91	23	88	15	94	28.5	80	4
92	26	91	23			81	6.5
90	19.5	89	17				
		84	13.5				
	n_1=9		n_2=10		n_3=7		n_4=8
	$R_{1.}$= 196.5		$R_{2.}$= 153.0		$R_{3.}$= 207.0		$R_{4.}$= 38.5

La región crítica de nivel 0.05 corresponde a valores del estadístico T mayores que 7.815 (Tabla A2 de Conover). Respecto al valor de T observado es

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k} \frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4} \right) = \frac{1}{99.167}\left(\frac{196.5^2}{9} + \frac{153.0^2}{10} + \frac{207.0^2}{7} + \frac{38.5^2}{8} - \frac{34*35^2}{4} \right) = 25.63$$

pues

$$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4} \right) = \frac{1}{33}\left(13664 - \frac{34*35^2}{4} \right) = 98.53$$

Dado que el valor de T observado es mayor que el correspondiente valor crítico se rechaza la hipótesis nula a nivel 0.05. Respecto al p-valor, es

$$P(\chi_3^2 > 25.63) < 0.0001$$

Habiendo rechazado la hipótesis nula, es natural preguntarse qué métodos son significativamente diferentes.

Comparaciones múltiples: Al aplicar el test de Kruskal-Wallis para comparar k poblaciones y rechazar H_o, es posible hacer k(k-1)/2 comparaciones a fin de detectar pares de poblaciones significativamente diferentes. La comparación correspondiente al par i,j $(\theta_j - \theta_i)$ se basará en la diferencia de los rangos promedio de las respectivas muestras.

Sea $D_{ij} = \frac{1}{\sqrt{N}}\left(\overline{R}_{j.} - \overline{R}_{i.}\right)$. Bajo H_o , por el Teorema 1, $E(D_{ij}) = 0$ y, si no hay empates

$$V(D_{ij}) = \frac{N+1}{12}\left(\frac{1}{n_j} + \frac{1}{n_i}\right) \to \frac{1}{12}\left(\frac{1}{\lambda_j} + \frac{1}{\lambda_i}\right)$$

Si ahora, en el Teorema 2, consideramos $c_{iN} = 1 \ \forall i$, entonces

$$D_{ij} = \frac{1}{\sqrt{N}}(\overline{R}_{j.} - \overline{R}_{i.}) = V_j - V_i \xrightarrow{d} N(0, \sigma_{ij}^2)$$

donde $\sigma_{ij}^2 = \frac{1}{12}\left(\frac{1}{\lambda_j} + \frac{1}{\lambda_i}\right)$.

Si α es el nivel de significación global elegido, sea $\alpha' = \frac{\alpha}{\binom{k}{2}}$, diremos que θ_i es significativamente distinto de θ_j si

$$\left|D_{ij}\right| \geq z_{\alpha'/2}\sqrt{V(D_{ij})}$$

o sea si

$$\left|\overline{R}_{j.} - \overline{R}_{i.}\right| \geq z_{\alpha'/2}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_j} + \frac{1}{n_i}\right)}$$

La probabilidad de cometer por lo menos un error de tipo I con esta regla es menor o igual que α.

Conover (Conover and Iman, 1979) sugiere esta alternativa para hacer comparaciones múltiples: θ_i es significativamente distinto de θ_j si

$$\left|\overline{R}_{j.} - \overline{R}_{i.}\right| \geq t_{N-k,\alpha/2}\sqrt{\frac{S^2(N-1-T)}{N-k}\left(\frac{1}{n_j}+\frac{1}{n_i}\right)}$$

Ejemplo: En el ejemplo presentado antes, rechazamos H_o, es decir que los métodos de cultivo son significativamente diferentes. Deseamos identificar qué pares de métodos son significativamente distintos a nivel global 0.05

Métodos (i,j)	$\left\|\overline{R}_{i.} - \overline{R}_{j.}\right\|$	$z_{\alpha'/2}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_j}+\frac{1}{n_i}\right)}$
(1,2)	6.533	10.954
(1,3)	7.738	12.014
(1,4)	17.021	11.584 (*)
(2,3)	14.271	11.748 (*)
(2,4)	10.488	11.308
(3,4)	24.759	12.338 (*)

y por lo tanto los métodos significativamente diferentes, indicados con (*) en la tabla son: el 1 y el 4, el 2 y el 3 y el 3 y el 4. Es interesante notar que si se aplicase el método sugerido por Conover todos los métodos hubiesen resultado diferentes.

Ejemplo del uso del test de Kruskal-Wallis en una tabla de contingencia: Un ejemplo que muestra que el test de Kruskal-Wallis puede utilizarse aun en presencia de un gran número de empates es el caso de una tabla de contingencia en la cual las filas representan *c* categorías *ordenadas* y las columnas representan *r* poblaciones.

Cat.	**Poblaciones**					**Total**
	1	**2**	**3**		**c**	
1	O_{11}	O_{12}	O_{13}	...	O_{1c}	t_1
2	O_{21}	O_{22}	O_{23}	...	O_{2c}	t_2
3	O_{31}	O_{32}	O_{33}	...	O_{3c}	t_3
....	...	...	...	...	...	
r	O_{r1}	O_{r2}	O_{r3}	...	O_{rc}	t_r
Total	n_1	n_2	n_3	...	n_c	N

siendo O_{ij} el número de observaciones en la población j que pertenece a la categoría i. Lo importante es que se supone que las categorías son ordenadas, es decir que las observaciones de la categoría k son todas iguales entre sí y menores que las de la categoría k+1. Se puede pensar, sin pérdida de generalidad, que todas las observaciones de la categoría j toman el valor j. El rango promedio de las filas se calcularía en la forma:

$$\overline{R}_{i.} = \begin{cases} (t_1 + 1)/2 & \text{si } i = 1 \\ t_1 + (t_2 + 1)/2 & \text{si } i = 2 \\ t_1 + t_2 + (t_3 + 1)/2 & \text{si } i = 3 \\ \dots & \\ \sum_{k=1}^{r-1} t_k + (t_r + 1)/2 & \text{si } i = r \end{cases}$$

Por lo tanto la suma de los rangos de la población (columna) j será

$$R_{.j} = \sum_{i=1}^{r} O_{ij}\ \overline{R}_{i.}$$

y

$$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{r} t_i\ \overline{R}_{i.}^2 - \frac{N(N+1)^2}{4} \right)$$

Luego, el estadístico T se calcula sustituyendo estas dos expresiones en

$$T = \frac{1}{S^2}\left(\sum_{j=1}^{c} \frac{R_{.j}^2}{n_j} - \frac{N(N+1)^2}{4} \right)$$

Ejemplo: Se comparan las calificaciones otorgadas por 3 instructores durante un semestre para ver si alguno de ellos tiende a asignar menores calificaciones que los otros. Las hipótesis a testear son:

H_o: los 3 instructores califican en forma similar

H_1: algunos instructores tienden a asignar calificaciones menores que otros

Las calificaciones asignadas por los instructores son las siguientes:

Calificación	Instructor			Total fila	Rango promedio
	1	2	3		
A	4	10	6	20	10.5
B	14	6	7	27	34
C	17	9	8	34	64.5
D	6	7	6	19	91
F	2	6	1	9	105
Total columna	43	38	28	109	

La suma de los rangos por columna (instructor) son:

$$R_{.1} = 2370.5 \qquad R_{.2} = 2156.5 \qquad R_{.3} = 1468$$

Finalmente, el valor de S^2 = 941.71 y el valor del estadístico T = 0.3209. La región crítica de nivel 0.05 corresponde a valores de T mayores que el valor crítico $\chi^2_{2,0.05}$ = 5.991 y por lo tanto no se rechaza H_o a este nivel. Es decir que no hay evidencia de que alguno de los instructores tienda a asignar menores calificaciones que los otros.

$$\text{p-valor} = P(\chi^2_2 > 0.3209) = 0.85$$

Eficiencia: Bajo la alternativa, el estadístico T tiende a una distribución chi-cuadrado no central. A partir del parámetro de no centralidad se define una noción de eficiencia. Comparando el test de Kruskal-Wallis con el test F clásico, se obtiene

$$e(T,F) = 12\sigma^2 \left(\int f^2(x)dx\right)^2$$

que es la eficiencia de Pitman de Wilcoxon respecto al test de t para muestras apareadas y la del test de Mann-Whitney respecto al test de t para muestras independientes. Por lo tanto la eficiencia del test de Kruskal-Wallis relativa al test F nunca es menor que 0.864. Si las poblaciones son Normales, la eficiencia relativa es 0.955, si son uniformes es 1.0 y si son doble exponenciales es 1.5.

Comparado con el test de la mediana (Mood) la eficiencia del test de Kruskal-Wallis es 1.5, 3 y 0.75 para las distribuciones Normal, uniforme y doble exponencial respectivamente.

Scores generales: Para mejorar la eficiencia es posible definir scores, reemplazando $R(X_{ij})$ por

$$A_{ij} = \psi\left(\frac{R(X_{ij})}{N+1}\right)$$

siendo Ψ una función generadora de scores. Por ejemplo, si $\Psi = \Phi^{-1}$, se obtiene el estadístico de scores normales, para el cual $\overline{\psi} = \int_0^1 \psi(u)du = 0$ y por lo tanto no es necesario restar nada.

Entonces, se rechazará H_o, si:

$$T^* = \frac{1}{S^2}\sum_{i=1}^{k}\frac{A_{i.}^2}{n_i} > \chi^2_{k-1,\alpha}$$

donde

$$S^2 = \frac{1}{N-1}\sum_{i,j} A_{ij}^2$$

En el caso de rechazar H_o, θ_i será significativamente distinto de θ_j si

$$\left|\overline{A}_{i.} - \overline{A}_{j.}\right| > t_{N-k,\alpha/2}\sqrt{S^2\left(\frac{N-1-T_1}{N-k}\right)\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Tests para alternativas ordenadas: Muchas veces la alternativa de locaciones distintas no es la adecuada. Nos podría interesar detectar un efecto creciente o decreciente, lo que es análogo a las alternativas unilaterales en los problemas de una o dos muestras. En este caso, el test de Kruskal-Wallis no es adecuado ya que está diseñado para detectar alejamientos respecto de la igualdad. Es posible construir tests con mayor potencia para detectar alternativas crecientes.

Supongamos que se cuenta con datos provenientes de k poblaciones y que se desea testear:

H_o: $\theta_1 = \theta_2 = ... = \theta_k$ vs H_1: $\theta_1 \le \theta_2 \le ... \le \theta_k$ con al menos una desigualdad estricta

Consideremos el siguiente estadístico

$$L = \frac{1}{\sqrt{N}}\sum_{i=1}^{k}\left(i - \frac{k+1}{2}\right)\left(\overline{R}_{i.} - \frac{N+1}{2}\right)$$

Valores grandes de *L* proveen evidencia a favor de la alternativa.

Bajo H_o, E(*L*) = 0 y, si no hay empates,

$$Var(L) = \frac{N+1}{12}\sum_{i=1}^{k}\frac{1}{n_i}\left(i - \frac{k+1}{2}\right)^2$$

En el caso particular en que $n_i = n$ para todo i (diseño balanceado),

$$Var(L) = \frac{\left(k^2-1\right)(nk+1)k}{144n}$$

Se puede probar que la distribución asintótica de *L* es Normal y por lo tanto se rechaza H_o a nivel α si

$$L \ge z_\alpha\sqrt{Var(L)}$$

Ejemplo: En el estudio Stanford sobre trasplantes de corazón se registraron varias variables cualitativas y cuantitativas sobre cada paciente. Una de las medidas, el puntaje de incompatibilidad, indica el grado de incompatibilidad de los tejidos entre donante y paciente trasplantado. Los puntajes de incompatibilidad se clasifican en bajos (0-1), medianos (1-2) y altos (mayor que 2). Se podría plantear la hipótesis de que el tiempo de sobrevida aumentará cuando es menor el puntaje de incompatibilidad. Los datos (Mosteller y Tukey, 1977) son los siguientes:

Incompatibilidad		
Baja	**Mediana**	**Alta**
44	15	3
551	280	136
127	1024	65
1	253	25
297	66	64
46	29	322
60	161	23
65	624	54
12	39	63
1350	51	50
730	68	10
47	836	48
994	51	
26		

Si denominamos θ_B, θ_M y θ_A a las medianas poblacionales de los tiempos de sobrevida correspondientes a los tres grupos, las hipótesis a testear son:

H_o: $\theta_A = \theta_M = \theta_B$ vs H_1: $\theta_A \leq \theta_M \leq \theta_B$ con al menos una desigualdad estricta

En este caso $n_1 = 12$, $n_2 = 13$ y $n_3 = 14$, N=39 y Var(L)=0.52 . Por lo tanto se rechazaría H_o si

$L >$ 1.18. El valor observado de L es 0.80 y por lo tanto no se rechaza H_o a este nivel.

Terpstra (1952) y Jonckheere (1954) propusieron otro test para alternativas ordenadas basado en estadísticos de tipo Mann-Whitney-Wilcoxon. La ventaja de este método es que la comparación entre las muestras i y j no depende de las demás muestras. Para testear

H_o: $\theta_1 = \theta_2 = ... = \theta_k$ vs H_1: $\theta_1 \leq \theta_2 \leq ... \leq \theta_k$ con al menos una desigualdad estricta

se define $J = \sum_{1 \leq i < j \leq k} W_{ij}$

donde $W_{ij} = \sum_{u=1}^{n_j} \sum_{t=1}^{n_i} s(X_{ju} - X_{it})$. Se rechazará H_o si J es grande.

Bajo H_o y si no hay empates,

$$E(J) = \sum_{i<j} \frac{n_i\, n_j}{2} \qquad \text{y} \qquad Var(J) = \frac{1}{72}\left(N^2(2N+3) - \sum_{i=1}^{k} n_i^2 (2\, n_i + 3) \right)$$

y se rechazará H_o a nivel α si $J \geq E(J) + z_\alpha \sqrt{Var(J)}$.