CONSTRUCCION Y USO DE ESTADISTICOS NO PARAMETRICOS MAS COMUNES (PARTE II)

PABLO ANDRES GONZALEZ GONZALEZ

T E S I S

PRESENTADA COMO REQUISITO PARCIAL
PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS
EN ESTADISTICA EXPERIMENTAL

Universidad Autonoma Agraria
Antonio Narro

PROGRAMA DE GRADUADOS
Buenavista, Saltillo, Coah.

DICIEMBRE DE 1991

Tesis elaborada bajo el supervisión del comité particular de asesoria y aprobada como requisito parcial, para optar el grado de

MAESTRO EN CIENCIAS EN
ESTADISTICA EXPERIMENTAL

COMITE PARTICULAR

Asesor Principal : _____
M.C. Emilio Padrón Corral

Asesor : _____
M.C. Regino Morones Reza

Asesor : _____
M.C. Félix de Jesús Sánchez P.

_____
Dr. José Manuel Fernández Brondo
Subdirector de Asuntos de Postgrado

Buenavista, Saltillo Coah.

Diciembre de 1991

AGRADECIMIENTOS

A DIOS

Y a cada uno de los que con su apoyo generoso, representan su gratificante manifestación, e hicieron posible que el paso de todo este tiempo, culminara con este trabajo.

COMPENDIO

Construcción y uso de los estadistísticos no paramétricos mas comunes (parte II)

POR

PABLO ANDRES GONZALEZ GONZALEZ

MAESTRIA
ESTADISTICA EXPERIMENTAL
UNIVERSIDAD AUTONOMA AGRARIA ANTONIO NARRO
BUENAVISTA, SALTILLO, COAHUILA Dic. 1991

M. C. Emilio Padrón Corral - Asesor -

Palabras clave : K muestras relacionadas, k muestras independientes, prueba de Cochran, prueba de Friedman, prueba de la X^2, Prueba de la mediana, prueba de Kruskal y Wallis, medidas de asociación, coeficiente de correlación de rangos de Spearman, coeficiente de correlación de rangos de Kendall, escala nominal, escala ordinal, métodos no paramétricos, estadístico de prueba, distribución libre.

El presente trabajo se inicia con un breve planteamiento de las pruebas de distribución libre o métodos no paramétricos. Posteriormente se trata de hacer un desarrollo sencillo de los estadísticos de prueba no paramétricos más comunes, tanto para los casos en que se presentan k muestras relacionadas, como son la prueba de Cochran y la prueba de Friedman; así como para los casos de k muestras independientes: La prueba de la X^2, prueba de la mediana y la prueba de Kruskal Wallis.

Además del desarrollo del estadístico se mencionan brevemente las características de cada prueba, donde se resalta la escala de medición necesaria para poder utilizar dicha prueba, como son la escala nominal y la ordinal.

Por último se presenta un ejemplo en cada prueba, para tratar de ilustrar la aplicación del estadístico, así como los posibles arreglos que se pudieran hacer cuando se usa, como es el caso cuando se presentan observaciones ligadas.

ABSTRACT

Construction and use of the most common nonparametric statisticals. (Part. II)

BY

PABLO ANDRES GONZALEZ GONZALEZ

MASTER OF SCIENCE

EXPERIMENTAL STATISTICS

UNIVERSIDAD AUTONOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA. MEXICO. DECEMBER 1991

M. C. Emilio Padrón-Corral - Advisor -

Key words: K relation samples, K independent samples, Cochran test, Friedman test, X^2 test, Median test, Kruskal and Wallis test, association measures, Spearman's coefficient correlation ranks, Kendall's coefficient correlation ranks, nominal scale, ordinal scale, nonparametric methods, statistical test, free distribution.

The present work begins with a brief establishement of free distribution tests or nonparametric methods.

Afterwards, a simple development of the most common nonparametric test statisticals is made for the cases where k relation samples are present, as in the Cochran test and Friedman test, as well as in the cases of k independent samples, X^2 test, median test and Kruskal-Wallis test.

In addition to this statistical development, the characteristics of each test are briefly mentioned and the necessary measurement scale to use the test, as it occurs in the nominal and ordinal scales is pointed out.

Finally, an example of each test is presented, in order to illustrate the application of the statistical, as well as the possible arrangements that could be made when it is used, like in the case where binding observations are presents.

INDICE DE CONTENIDO

INTRODUCCION

Es común que en la gran diversidad de actividades de investigación científica, y no científica; se presenten casos en los que el patrón de comportamiento de las poblaciones en estudio resulte desconocido. Es decir la distribución de la población no es normal, por lo tanto no se puede caracterizar por medio de los parámetros más comunes, tales como la media y la varianza. Y el utilizar algún método estadístico paramétrico en lugar de llevarnos a una conclusión nos arrojaría más dudas.

Afortunadamente para salvar las situaciones como la anterior existen técnicas llamadas no paramétricas y/o de distribución libre. Estos dos últimos términos mencionados frecuentemente de manera indistinta no son sinónimos, ya que en el primero, no se plantea hipótesis alguna sobre el valor del parámetro de la distribución. Y en el segundo explícitamente el método no depende de la forma de distribución de la población.

Por lo anterior son importantes las presunciones antes de seleccionar métodos estadísticos, ya que la calidad de la inferencia depende de la relación entre el método de análisis y las características de la población, método de muestreo y otras influencias en los datos.

Los usos de métodos no paramétricos tienen varias ventajas sobre los métodos clásicos, pero también tienen algunos aspectos desfavorables. En general los métodos clásicos tienen la potencia máxima cuando las condiciones sobre la distribución y los parámetros son validos. Cuando no sucede esto su potencia disminuye y en algunas situaciones los métodos no paramétricos tienen mayor potencia. En su mayoría los métodos no paramétricos son más fáciles de aplicar, pero hay una gran variedad de cada tipo que confunde la selección del mejor; sus cálculos son sencillos y directos, pero cuando se aumenta el tamaño de la muestra, el trabajo se hace pesado. Sin embargo para muestras grandes la mayoría de las pruebas de libre distribución tienen fórmulas que producen una relación de convergencia con respecto a la distribución normal.

Sobre los métodos estadísticos no paramétricos existe poca literatura en español que conjunte el aspecto teórico de las pruebas, asi como lo práctico. Lo primero se pensó como una forma de reforzar la materia de estadística no paramétrica de la Maestría de Estadística Experimental de la UAAAN, y lo segundo para poner a disposición de los investigadores de la institución en una forma clara esta herramienta de la investigación, si no nueva pero útil.

2. PROCEDIMIENTO DE UNA PRUEBA DE DISTRIBUCION LIBRE

Sea X_1, X_2, ..., X_n que representan una muestra aleatoria de una población con función de distribución acumulada contínua F_x. Dado que F_x es supuesta contínua, la probabilidad de que dos o más variables aleatorias asuman igual magnitud es cero. Además existe un único arreglo ordenado dentro de la muestra. Suponga que $X_{(1)}$ denota la mas pequeña del conjunto de observaciones X_1, X_2,..., X_n; $X_{(2)}$ la segunda mas pequeña, etc; y $X_{(n)}$ denota la mayor. Entonces $X_{(1)} < X_{(2)} <...< X_{(n)}$ denota la muestra aleatoria original después de arreglar en orden de incremento de magnitud, y estos son llamados colectivamente estadísticos de orden de la muestra aleatoria X_1, X_2, ..., X_n. Para $1 \leq r \leq n$, $X_{(r)}$ es llamado el r-ésimo estadístico de orden. El tema de estadísticos de orden generalmente trata con las propiedades de $X_{(r)}$ por si mismo o funciones de algún subconjunto de n estadísticos de orden.

Los estadísticos de orden son particularmente útiles en estadística no paramétrica por que la transformación $U(r) = F_x(X(r))$ produce una variable aleatoria la cual es el r-ésimo estadístico de orden de la población uniforme sobre el intervalo (0,1), y además U(r) es libremente distribuído. Esta propiedad es adecuada para la llamada "probabilidad de transformación integral" la cual se demostrará a continuación.

Sea la variable aleatoria X que tiene la función de distribución acumulada F_x. Si F_x es contínua, la variable aleatoria Y producida por la transformación $y = F_x(X)$ tiene la distribución de probabilidad uniforme sobre el intervalo (0,1).

Dado $0 \leq F_x(x) \leq 1$ para todo x, tenemos $F_Y(y) = 0$ para $y \leq 0$ y $F_Y(y) = 1$ para $y \geq 1$. Para $0 < y < 1$, definimos **u** como el mayor número que satisface $F_x(u) = y$. Entonces $F_x(x) \leq y$ si y solamente si $X \leq u$, y se sigue que

$$F_Y(y) = P[F_x(x) \leq y] = P(x \leq u) = F_x(u) = y$$

la cual es la distribución uniforme.

Como resultado podemos concluir que si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de alguna población con distribución F_x contínua, entonces $F_x(X_1), F_x(X_2), \ldots, F_x(X_n)$ constituyen una muestra aleatoria de la población uniforme. Similarmente, si $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ son los estadísticos de orden para la muestra original, entonces

$$F_x(X_{(1)}) < F_x(X_{(2)}) < \ldots < F_x(X_{(n)})$$

son estadísticos de orden de la distribución uniforme sobre (0,1). Estos estadísticos de orden pueden ser llamados de distribución libre, en el sentido de que su distribución de probabilidad es conocida como uniforme, independientemente de la distribución original F_x.

Cuando la observación $X_{(r)}$ es reemplazada por su rango r toma el nombre de estadísticos de orden por rango, los cuales tienen un uso importante en estadística no paramétrica para hacer pruebas. Los estadísticos de orden por rango para una muestra aleatoria, son algún conjunto de constantes las cuales indican el orden de las observaciones. La magnitud de alguna observación es usada solamente en la determinación de su posición relativa en el arreglo muestral, y después se ingnora en el análisis basado en los estadísticos de orden de rango. Si el estadístico de orden por rango de una muestra aleatoria X_1, $X_2,\ldots,X_n$ son denotados por $r(X_1)$, $r(X_2),\ldots,r(X_n)$; r es una función tal que $r(X_i) \leq r(X_j)$ siempre y cuando $X_i \leq X_j$. Como con los estadísticos de orden, los estadísticos de orden por rango son invariantes bajo transformaciones monótonas, y entonces, si $r(X_i) \leq r(X_j)$, entonces $r[F(X_i)], \leq r[F(X_j)]$, en adición a $F[r(X_i)] \leq F[r(X_j)]$, donde F es alguna función no decreciente.

Para algún conjunto de **N** diferentes observaciones muestrales, el conjunto más simple de números utilizados para indicar la posición relativa son los **N** primeros enteros positivos. Además podemos asumir que los estadísticos de orden por rango son una permutación de los **N** primeros enteros.El i-ésimo estadístico de orden por rango $r(X_i)$ es conocido como el rango de la i-ésima observación en la muestra desordenada.

El valor asumido, $r(X_i)$, es el número de observaciones X_j, $j = 1, 2,...N$, tales que $X_j \leq X_i$. Por ejemplo, el rango del i-ésimo estadístico de orden es igual a i, o $r(X_{(i)})=i$.

La variable aleatoria $r(X_i)$ es discreta y para una muestra aleatoria de una población continua esta sigue la distribución uniforme discreta, ó

$$P[r(X_i) = j] = \frac{1}{N} \text{ para } j = 1,2,...,N$$

Por otro lado una función de los estadísticos de orden por rango es conocida por estadísticos de rango, los cuales son particularmente útiles en la inferencia no paramétrica dado que están usualmente distribuídos libremente. Los métodos son aplicables a una amplia variedad de situaciones de prueba de hipótesis dependiendo de la función particular utilizada.

Para hacer la prueba de hipótesis se escoge una muestra aleatoria de la población de interés bajo las condiciones experimentales, exactamente como en las pruebas clásicas. Se tienen algunos supuestos sobre los datos, pero no son sobre la distribución de la población. Con estos datos (convertidos a estadísticos de rango) formamos una estadística, pueden ser sumas, diferencias o productos, de los rangos de los datos, y se estudia la probabilidad de encontrar valores de esa estadística igual o más extremos con respecto a la hipótesis alternativa bajo la hipótesis nula. Si la probabilidad es menor o igual que el valor del

nivel de confianza (α) se rechaza la hipótesis nula.

Expondremos un ejemplo para aclarar esto último :

Se tienen tres muestras y dos de las cuales son conocidas A y B, la tercera es desconocida, y se quiere saber si es como A o como B. Una persona hizo la identificacion correcta en nueve de las diez pruebas presentadas. ¿ Se puede afirmar que esa persona tiene la capacidad de distinguir entre las dos clases A y B con suficiente regularidad? Observamos en cada una de las diez comparaciones que hay dos posibilidades A o B, y bajo la hipótesis nula no tenemos ninguna preferencia. En las diez oportunidades, hay 2^{10} posibilidades o 1024. Si se estableció la hipótesis nula de que la persona no sabe nada con respecto al método de identificación de la muestra desconocida con alternativa unilateral observamos que las dos posibilidades igual a la muestra o más extremas son nueve correctas en diez o diez en diez. Podemos calcular que hay diez oportunidades de identificar nueve de diez y una de identificar diez de diez, o once posibilidades satisfactorias. La probabilidad de encontrar resultados de este tipo o más extremos bajo la hipótesis nula es 11/1024 ó .017. Si podemos aceptar un riesgo hasta de 0.05; podemos rechazar la hipótesis nula y aceptar la hipótesis alternativa, con la conclusión que de la persona tiene capacidad de detectar diferencias entre A y B. Si se hubiera encontrado solamente ocho correctas de diez, la posibilidad hubiera sido 56/1024 > 0.05 y no

podriamos rechazar la hipótesis nula. Con pruebas de esta clase se establece el valor crítico de nueve en diez al nivel de confianza de 0.05. Esta prueba no depende, de ninguna presunción sobre los datos, del método de detección usado, ni de aspectos de la distribución de los datos.

En el estudio de los métodos no paramétricos, necesitamos considerar tres factores importantes en la selección de la prueba apropiada.

a) La clase de los datos con respecto a la escala de medición (nominal, ordinal, de intervalo ó de proporción).
b) El número y tipo de variables
c) La hipótesis de interés (Comparación de la forma de distribución, escala del parámetro, correlación, bondad de ajuste, etc.).

3. PRUEBAS PARA EL CASO DE K MUESTRA RELACIONADAS

Existen situaciones en las que es necesario probar **k** tratamientos, donde $K \geq 3$, que produzcan o no los mismos efectos en ciertas condiciones. Si del grupo de **K** tratamientos (el cual prodría considerarse como un bloque) repetiremos **r** veces los tratamientos obtendriamos r-bloques, haciendo un total de $N = Kr$ unidades experimentales. Donde las unidades experimentales son muy semejantes entre sí. Dentro de la estadística paramétrica estos diseños se conocen como bloques al azar.

Para probar la significancia de la diferencia entre el total de las **K** muestras se presentarán dos técnicas no paramétricas (Prueba de Cochran y la prueba de Friedman).

En la primera de ellas se supone una escala nominal (es decir a las observaciones se les categoriza de acuerdo a nombres o puede ser por números pero entre uno y otro no se establece una magnitud). Y en la segunda se supone una escala ordinal (aquí las observaciones se pueden ordenar de menor a mayor pero no existe una magnitud entre ellos).

La Prueba Q de Cochran

La prueba **Q** de Cochran para **K** muestras relacionadas proporciona un método para examinar si tres o más conjuntos

igualados de frecuencias o proporciones difieren significativamente entre sí. La igualación puede basarse en características relevantes de los diferentes sujetos o en el hecho de que los mismos sujetos se usan en condiciones diferentes.

Esta prueba es particularmente adecuada cuando los datos están en una escala nominal o se ha dicotomizado la información ordinal (el resultado de las observaciones puede ser clasificado como cero o uno).

Si los datos de investigaciones se colocan en una tabla de dos clasificaciones formada de R hileras y C columnas (como se muestra en la tabla siguiente), es posible probar la hipótesis de nulidad Ho: de que la proporción o frecuencia de respuesta de una clase particular es la misma en cada columna, excepto por diferecias aleatorias

Tabla de doble clasificación

	Tratamientos						
Bloques	1	2	3	4	... C		Total de filas
1	X_{11}	X_{12}	X_{13}	X_{14}	...	X_{1c}	R_1
2	X_{21}	X_{22}	X24	X24	...	X_{2c}	R_2
3	...	...	...	...	...	...	...
r	X_{r1}	X_{r2}	X_{r3}	X_{24}	...	X_{rc}	R_r
Total Columnas	C_1	C_2	C_3	C_4	...	C_c	N = Gran Total

En donde X_{ij} es uno o cero, y es el resultado obtenido para la i-ésima fila en la j-ésima columna. Además cada fila respresentaría un bloque con C tratamientos.

Como X_{ij} es cero o uno, es una variable aleatoria con distribución Bernoulli con parámetros P_{ij}. Entonces el total de columna C_j, definido por $C_{ij} = \sum_{i=1}^{r} X_{ij}$ Es también una variable aleatoria; si los P_{ij} en columna fueran el mismo, C_j seguiría una distribución Binomial. Pero aun cuando H_0 es asumida como cierta, las P_{ij} bajo la columna pueden diferir cada uno de otro. La hipótesis solamente enuncia que los P_{ij} a través de cada fila son iguales, pero pueden variar de una fila (bloque) a otra. Sin embargo dado que la variable aleatoria C_j es la suma de las r variables aleatorias independientes, el teorema del límite central se aplica, y para r grande la distribución de C_j es aproximadamente normal

Asi tenemos que

$$\frac{C_j - E(C_j)}{\sqrt{var(C_j)}} \sim N(0,1) \quad \text{Aproximadamente} \qquad (1)$$

De modo que la suma

$$\sum_{j=1}^{c} \left[\frac{C_j - E(C_j)}{\sqrt{var\ C_j}} \right]^2 = \sum_{j=1}^{c} \frac{[C_j - E(C_j)]^2}{Var\ (C_j)} \qquad (2)$$

Puede ser aproximada por la distribución J_i cuadrada con c grados de libertad. Sin embargo, los parámetros $E(C_j)$ y Var (C_j)son desconocidos, por lo que su cálculo resulta en la pérdida de un grado de libertad.

La media de C_j puede ser estimada por la media muestral

$$\frac{1}{c} \sum_{j=1}^{c} C_j = \frac{N}{c} = E\ (C_j) \qquad (3)$$

El mismo estimador es usado para la media de cada C_j. La varianza de C_j es igual a la suma de las varianzas de X_{ij} en la j-ésima columna.

$$\text{Var}\ (C_j) = \sum_{i=1}^{r} \text{Var}\ (X_{ij}) \qquad (4)$$

Dado que los bloques son independientes la varianza de X_{ij} es

$$\text{Var}\ (X_{ij}) = P_{ij}\ (1 - P_{ij}) \qquad (5)$$

Bajo Ho, la probabilidad de un "éxito" P_{ij} es la misma para todas las columnas dentro de un bloque, y además es natural para estimar P_{ij} por el número promedio de éxitos en la fila i,

$$P_{ij} = \frac{R_i}{c} \qquad (6)$$

Sustituímos (6) en (5)

$$\text{Var } (X_{ij}) = \frac{R_i}{c}\left(1 - \frac{R_i}{c}\right) \qquad (7)$$

Sin embargo tal estimador tiende a ser pequeño y es incrementado multiplicandolo por $\frac{c}{c-1}$, entonces la varianza (X_{ij}) es estimada por .

$$\text{Var } (X_{ij}) = \frac{R_i}{c}\left(1 - \frac{R_i}{c}\right)\frac{c}{c-1} = \left(\frac{R_i}{c} - \frac{R_i^2}{c^2}\right)\frac{c}{c-1} =$$

$$\frac{c\,R_i}{c\,(c-1)} - \frac{c\,R_i^2}{c^2-(c-1)} = \frac{R_i}{c-1} - \frac{R_i^2}{c(c-1)} =$$

$$\frac{c\,R_i - R_i^2}{c\,(c-1)} = \frac{R_i\,(c-R_i)}{c\,(c-1)} \qquad (8)$$

Sustituímos (8) en (4)

$$\text{Var } (C_j) = \frac{1}{c\,(c-1)} \sum_{i=1}^{r} R_i\,(c - R_i) \qquad (9)$$

El cual no depende sobre **j** y asi es usado para todos los $C_{j's}$.

Por último sustituímos (3) y (9) en (2)

$$\frac{\sum_{j=1}^{c}\left(C_j - \frac{N}{c}\right)^2}{\sum_{i=1}^{r}\frac{R_i\,(c-R_i)}{c\,(c-1)}} = \frac{c(c-1)\sum_{j=1}^{c}\left(C_j - \frac{N}{c}\right)^2}{\sum_{i=1}^{r} R_i\,(c - R_i)} \qquad (10)$$

De esta manera obtenemos el estadístico de Cochran que se distribuye como una ji - cuadrada con c-1 grados de libertad.

La distribución del estadístico es difícil de tabular, así que la aproximación de muestras grandes es usada en su lugar. El número de bloques (r) es supuesto grande. Luego la región crítica de tamaño aproximado α corresponde a todos los valores del estadístico mayores que X^2 $(1-\alpha)$, α cuantil de una variable aleatoria j_i - cuadrada con (c-1) grados libertad, obtenido de la tabla de valores críticos de la ji cuadrada. Si el valor del estadístico excede X^2 $(1-\alpha)$, rechazamos Ho. De otra manera no rechazamos la hipótesis nula de no diferencia en la efectividad de los diferentes tratamientos.

Enseguida presentamos un ejemplo de la prueba para su ilustración.

Se quieren comparar 4 vacunas contra una enfermedad en cerdos. Para compararlos en las condiciones más homogéneas posibles se eligen 8 camadas de 4 cerdos cada una. Cada camada representará un bloque. Al final se anota si se presentó la enfermedad ("o") o no se presentó ("1").

Los resultados se dan enseguida.

Camada (Boques)	Vacuna (Tratamiento)					
	A	B	C	D	R_i	R_i^2
1	0	1	0	1	2	4
2	0	1	0	1	2	4
3	1	1	0	1	3	9
4	1	0	0	1	2	4
5	0	1	1	1	3	9
6	0	1	0	1	2	4
7	0	1	0	0	1	1
8	0	0	0	1	1	1
C_J	2	6	1	7	16	36
C_j^2	4	36	1	49	90	

Para obtener el valor del estadístico usamos una forma más apropiada para facilitar los cálculos, así que empleamos la siguiente fórmula:

$$Q = \frac{(c-1)c \sum_{j=1}^{c} c_j^2 - (c-1)\left(\sum_{j=1}^{c} c_j\right)^2}{c \sum_{i=1}^{r} R_i - \sum_{i=1}^{r} R_i^2}$$

Asi

$$Q = \frac{(3)\ (4)\ (90) - (3)\ (16)^2}{4\ (16) - 36} = \frac{312}{28} = 11.14$$

Con $\alpha = .05$ obtenemos de la tabla de X^2 :

$X^2.05\ (03) = 7.815$

Puesto que $Q > 7.815$ rechazamos H_0 y concluímos que si hay diferencia entre los tratamientos con $\alpha = .05$.

A continuación presentamos un resumen de la prueba.

a) Suposiciones:

1.- Los bloques fueron seleccionados aleatoriamentede de las poblaciones de todos los bloque posibles.

2.- Los resultados de los tratamientos pueden ser dicotomizados, clasificados con cero o uno. La escala es nominal.

b) Hipótesis:

H_0: Los tratamientos son igualmente efectivos

Ha: Hay una diferencia en efectividad entre los tratamientos.

$$P_{ij} = P\ (X_{ij} = 1),\ i = 1, \ldots, r;\ j=1, \ldots, c$$

Entonces la efectividad semejante entre los tratamientos implica.

$$P_{i1} = P_{i2} = \ldots = P_{ic},\ \text{para cada } i \text{ de } 1 \text{ a } r$$

Esto es, para cada bloque la probabilidad de que un tratamiento sea un éxito no depende sobre cual tratamiento es usado. Luego las hipótesis pueden ser como sigue:

Ho: $P_{i1} = P_{i2} = \ldots = P_{ic}$, para cada i de 1 a r

Ha: $P_{ij} \neq P_{ik}$ para algún **j** y **k**, y alguna **i**

c) Estadística de prueba:

$$Q = \frac{\left[\sum_{j=1}^{c} c(c-1)\ \left(C_j - \frac{N}{c}\right)\right]^2}{\sum_{i=1}^{r} R_i (c-R_i)} \qquad \text{o equivalentemente}$$

$$Q = \frac{(c-1)c\sum_{j=1}^{c} C_j^2 - (c-1)\left(\sum_{j-1}^{c} C_j\right)^2}{\sum_{i=1}^{r} R_i - \sum_{i=1}^{r} R_i^2}$$

d) Distribución de la estadística: X^2 (C-1) g.l.

e) Regla de decisión: Si $Q > X^2\ (C-1, \alpha\ )$ se rechaza H_0

3.2 Prueba de Friedman

Cuando los datos de K muestras igualadas obtenidas de un experimento o situación estan dados en forma ordinal

al menos, y el tratar de dicotomizarlo puede hacer que se pierda parte de la información. Es por ello que se presenta esta prueba desarrollada por Friedman en 1937.

Puesto que las muestras han sido igualadas, el número de casos es el mismo en cada una de las muestras. La igualación se hace como en la prueba anterior estudiando el mismo grupo de sujetos en cada una de las **k** condiciones.

Los datos de la prueba se colocan en una tabla de dos clasificaciones con **r** hileras y **k** columnas, las hileras representan a los diferentes sujetos o conjuntos de sujetos igualados y las columnas representan las diferentes condiciones. Si se estudian los puntajes de los sujetos utilizados en todas las condiciones, en cada hilera están los puntajes de un sujeto en las **k** condiciones.

Los puntajes se han ordenado por rangos para cada hilera por separado. Esto es, con **k** condiciones en estudio, los rangos de cualquier hilera van de 1 a **k**. Determinando la prueba de Friedman la probabilidad de que las diferentes columnas de rangos (muestras) procedan de la misma población.

En la siguiente tabla se presenta el arreglo de la información original.

		Tratamientos				
		1	2	3	...	K
Bloques	1	X_{11}	X_{12}	X_{13}	...	X_{1k}
	2	X_{21}	X_{22}	X_{23}	...	X_{2k}
	...	...	...	...	...	...
	r	X_{r1}	X_{r2}	X_{r2}	...	X_{rk}

Donde X_{ij} denota el i-ésimo sujeto de la j-ésima condición. De donde, para hacer posible la comparación de la igualdad de tratamientos, a cada X_{ij} que pertenece a la i-ésima hilera, se le asigna un rango de acuerdo a su magnitud dentro de las observaciones de esa misma hilera. Generando así el siguiente arreglo.

	Tratamientos				
Bloques	R_{11}	R_{12}	R_{13}	...	R_{1n}
	R_{21}	R_{22}	R_{23}	...	R_{2n}
	...	...	...	...	...
	R_{m1}	R_{m2}	R_{m3}	...	R_{mn}
Total	R_1	R_2	R_3	...	R_n

Donde R_{ij} es el rango asociado por renglón al lugar ij-ésimo, de tal forma que los renglones sumen $\frac{n(n+1)}{2}$ ó sea $\sum_{j=1}^{n} R_{ij} = \frac{n(n+1)}{2}$. Además sea $R_j = \sum_{i=1}^{m} R_{ij}$, el total de la columna j-ésima. Entonces si la hipótesis de igualdad de tratamientos es cierta, se tiene que el total de cada columna sea aproximadamente igual al promedio de todas ellas, es decir que R_j y $\frac{m(n+1)}{2}$ son aproximadamente

igual. De esta manera suponemos que los rangos asociados por renglón, lo fueron aleatoriamente, y que en promedio todos los totales por columnas son iguales.

Por otra parte, la distribución muestral de las medias de los rangos por columna será aproximadamente normal (0,1) tanto como el número de hileras sea grande; esto por el teorema del límite central.

Asi tenemos que

$$\frac{R_j - E(R_j)}{\sqrt{Var(R_j)}} \sim N(0,1) \qquad \text{Aproximadamente} \qquad (1)$$

Luego la suma

$$\sum_{j=1}^{n}\left[\frac{R_j - E(R_j)}{\sqrt{Var(R_j)}}\right]^2 = \sum_{j=1}^{n}\frac{[R_j - E(R_j)]^2}{Var(R_j)} \qquad (2)$$

Se aproxima por la disitribución Ji - cuadrada con **n** grados de libertad. Donde hay que estimar $E(R_j)$ y $Var(R_j)$, perdiendo un grado de libertad.

En el capítulo 2 mencionamos acerca de la distribución de los estadísticos de orden y su fuerte relación con los rangos, apelando a esto obtendremos los parámetros que necesitamos.

Asi $$E(R_j) = E\left[\sum_{i=j}^{m} R_{ij}\right] = \sum_{i=1}^{m} E[R_{ij}] = \sum_{i=1}^{m}\sum_{j=1}^{n} j\left(\frac{1}{n}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n} j$$

$$= \frac{1}{n}\sum_{i=1}^{m}\frac{n(n+1)}{2} = \frac{m}{n}\frac{n(n+1)}{2}$$

$$= \frac{m(n+1)}{2} \qquad (3)$$

Y

$$E(R_j^2) = E\left[\sum_{i=1}^{m} R_{ij}\right]^2 = E\left[\sum_{i=1}^{m} R_{ij}^2 + \sum_{i=1}^{m}\sum_{\substack{i'=1\\ i\neq i'}}^{m-1} R_{ij}\, R_{i',j}\right]$$

$$= \sum_{i=1}^{m} E[R_{ij}]^2 + \sum_{i=1}^{m}\sum_{\substack{i'=1\\ i\neq i'}}^{m=1} E[R_{ij}\, R_{ij}]$$

Desarrollamos el primer término

$$\sum_{i=1}^{m} E[R_{ij}]^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} j^2\left(\frac{1}{n}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n} j^2 = \frac{1}{n}\sum_{i=1}^{m}\frac{n(n+1)(2n+1)}{6}$$

$$= \frac{mn(n+1)(2n+1)}{6n} = \frac{m(n+1(2n+1)}{6}$$

Ahora desarrollamos el segundo término

$$\sum_{i=1}^{m}\sum_{\substack{i'=1\\ i\neq i'}}^{m=1} E[R_{ij}\, R_{i',j}] = \sum_{i=1}^{m} E[R_{ij}]\sum_{i=1}^{m=1} E[R_{i',j}]$$

$$= \sum_{i=1}^{m}\left(\frac{n+1}{2}\right)\sum_{i'=1}^{m-1}\left(\frac{n+1}{2}\right) = \frac{(n+1)^2}{4}\, m\,(m-1)$$

$$= \frac{12(n-1)}{mn\ (n^2-1)} \sum_{j=1}^{n} \left[R_j^2 - 2\ \frac{m(n+1)}{2}\ \frac{m(n+1)}{2} + \frac{m^2(n+1)^2}{4} \right]$$

$$= \frac{12(n-1)}{mn\ (n+1)(n-1)} \sum_{j=1}^{n} \left[R_j^2 - \frac{m^2\ (n+1)^2}{2} + \frac{m^2(n+1\ )^2}{4} \right]$$

$$= \frac{12}{mn\ (n+1)} \sum_{j=1}^{n} \left[R_j^2 - \frac{m^2\ (n+1)^2}{4} \right]$$

$$= \frac{12}{mn\ (n+1)} \sum_{j=1}^{n} R_j^2 - \frac{^2 mn\ (n+1)^2}{4}\ \frac{12}{mn(n+1)}$$

$$= \frac{12}{mn\ (n+1)} \sum_{j=1}^{n} R_j^2 - 3\ m\ (n+1) \qquad (5)$$

Obteniendo así en (5) el esadístico de Friedman (X_r^2).

La regla de decisión es, si X_r^2 es mayor que el cuantil $1-\alpha$ de una X^2 con n-1 grados de libertad, querrá decir que las diferencias son grandes y por lo tanto la hipótesis de igualdad de tratamientos deberá de rechazarse. En caso contrario, las diferencias no son significativas y la hipótesis deberá aceptarse.

Cuando el número de hileras y columnas es menor que el mínimo requerido para utilizar una tabla de la X^2, deben usarse las tablas de Probabilidades exactas.

Presentaremos un ejemplo para ilustrar el uso del estadístico de Friedman.

Ejemplo: Cuatro variedades de fríjol de soya fueron plantados, en tres parcelas, obteniendo los siguientes resultados.

Variedades de fríjol de soya

Parcelas	A	B	C	D
1	45	48	43	41
2	49	45	42	39
3	38	39	35	36

Por lo que la hipótesis a probar es

H_0: El promedio de cosechas de las cuatro variedades es el mismo

Teniéndose como alternativa

H_a Existen diferencias significativas en los promedios de cosechas.

Hacemos un ordenamieto por rangos de cada uno de los tres grupos formados.

	A	B	C	D
1	3	4	2	1
2	4	3	2	1
3	3	4	1	2
R_j =	10	11	5	4

Así

$$X_r^2 = \frac{12}{3\ (4)(5)}\ (262) - 3(3)(4+1) = 7.4$$

Si $\alpha = 0.05$, el cuantil 0.95 de una X^2 con 3 grados de

libertad es 7.815 la cual lleva a no rechazar la hipótesis H_0.

Finalmente presentamos un resumen de la prueba.

a) Suposiciones: La escala es ordinal ya que ordenamos por rangos cada una de las hileras de la tabla de doble clasificación.

b) Hipótesis:

H_0 los tratamientos son iguales

c) Distribución de la estadística: X^2 (n-1)

d) Decisión: si $X_r^2 > X^2$ (n-1,1-α) se rechaza H_0

e) Estadística de prueba:

$$X_r^2 = \frac{12}{mn(n+1)} \sum_{j=1}^{n} R_j^2 - 3\ (m)\ (n+1)$$

4. PRUEBAS PARA EL CASO DE K MUESTRAS INDEPENDIENTES

Cuando se van a analizar datos de alguna investigación, regularmente se hace el cuestionamiento si varias muestras independientes deben considerarse como procedentes de la misma población. Es decir, se trata de verificar si tras una diferencia entre muestras observadas existen diferencias verdaderas entre las poblaciones o si solo son atribuibles al azar, las cuales serían de esperarse entre muestras aleatorias de una misma población.

Si las suposiciones asociadas con el modelo estadístico clásico "La prueba de F", (que es la técnica paramétrica usual para probar si varias muestras independientes proceden de la misma población) digamos que estén distribuídas normalmente y con varianzas iguales y que la medida de la variable estudiada sea de intervalo, no se cumplieran; el investigador pudiera usar una de las pruebas estadísticas no paramétricas, las cuales permiten el análisis de datos que se refieran solamente por clasificación o sea en una escala nominal, o por medio de rangos (escala ordinal).

Dichas técnicas no paramétricas que se analizarán a continuación son: La prueba J_i- cuadrada, KrusKal-Wallis y la prueba de la mediana.

Prueba de la X^2

Antes que nada, para evitar confusión se desea aclarar que el símbolo X^2 es usado para la cantidad que será obtenida del estadístico que se pretende presentar, conocido como prueba de la X^2.

Las palabras j_i-cuadrada se referirán a la variable aleatoria que sigue la distribución j_i- cuadrada.

Ahora bien, cuando los datos de investigación consisten en frecuencias de categorías discretas, puede usarse la prueba X^2 para determinar la significación de las diferencias entre **k** grupos independientes, la técnica es del tipo de la bondad de ajuste, que puede usarse para probar la existencia de una diferencia significativa entre un número observado de objetos o respuestas de cada categoría y un número esperado; la medición de las variables puede ser en una escala nominal.

La hipótesis de nulidad (H_0) que se pondrá a prueba, será que los k grupos no difieren con respecto a alguna característica, y por lo tanto, con respecto a la frecuencia relativa con que los miembros de cada grupo son encontrados en diferentes categorías. Para probar la hipótesis contamos el número de casos de cada grupo en cada categoría y comparamos la proporción de casos en las diferentes categorías de un grupo con la de otros grupos.

A continuación se presenta el desarrollo del estadístico basado en una sola muestra, al cual se le hacen pequeñas modificaciones para comparar **k** muestras.

Supongamos que una muestra aleatoria de tamaño n es extraída de una población con función de distribución acumulada continua Fx. Nosotros deseamos probar la hipótesis nula

$$H_0: F_x(x) = F_0(x) \text{ para todo } x$$

Donde $F_0(x)$ es especificado; contra la hipótesis alterna

$$Ha: F_x(x) \neq F_0(x) \text{ para alguna } x$$

En orden a lo anterior los datos muestrales deben primero ser agrupados de acuerdo a algún criterio en k categorías a fín de formar una distribución de frecuencias.

Asumiendo que la distribución de la población es especificada como se mencionó anteriormente, por la hipótesis nula, se puede calcular la probabilidad de que una observación aleatoria sea clasificada dentro de una de las k categorías. Estas probabilidades multiplicadas por n dan las frecuencias para cada categoría, las cuales serían esperadas si la hipótesis nula fuera verdadera. Excepto para variaciones muestrales, las frecuencias esperadas y las observadas serían concordantes si los datos muestrales son compatibles con $F_0(X)$.

Una vez que la muestra es clasificada en distribución de frecuencias, las variables aleatorias de importancia son las frecuencias de clase F_1, $F_2,\ldots,F_k$. Estas constituyen un conjuno de variables aleatorias de la distribución multinomial **k** - variada con **k** posibles resultados, siendo el i-ésimo resulado la i-ésima cateogría en el sistema de clasificación. Con Θ_1, $\Theta_2,\ldots,\Theta_k$ denotando las probabilidades de los resultados respectivos, entonces la función de verosimilitud de la muestra es

$$L\ (\Theta_1,\ \Theta_2,\ldots,\Theta_k) = \prod_{i=1}^{k} \Theta_i^{f_i} \qquad f_i = 0,1,\ldots,n$$

$$\sum_{i=1}^{k} f_i = n\ ;\ \sum_{i=1}^{k} \Theta_i = 1 \qquad (1)$$

La hipótesis nula fue asumida para especificar la distribución de la población completa, de la cual las Θ_i pueden ser calculadas. Estas hipótesis conciernen ahora solamente a los valores de esos parámetros, y pueden ser enunciados equivalentemente como:

$$H_0:\ \Theta_i = \Theta_i^{\,0} = \frac{e_i}{n} \quad \text{para } i = 1,2,\ldots,k$$

Un estimador de verosimilitud máxima de los parámetros en (1), es $\hat{\Theta}_i = \frac{f_i}{n}$ Luego el estadístico de razón de verosimilitud para esta hipótesis es

$$T = \frac{L\ (\mathbb{A})}{L\ (\hat{\cap})} = \frac{L(\Theta_1^{\circ},\ \Theta_2^{\circ},\ldots,\ \Theta_k)}{L(\hat{\Theta}1,\ \hat{\Theta}2,\ldots,\hat{\Theta}_k} = \prod_{i=1}^{k} \left(\frac{\Theta_i^{\circ}}{\hat{\Theta}_i} \right)^{f_i}$$

De donde es ya conocido que la distribución de la cantidad -2 log T se aproxima a la distribución j_i-cuadrada. Los grados de libertad son K-1, puesto que la restricción $\sum_{i=1}^{k} \Theta_i = 1$ permite solamente k-1 parámetros en $\cap$ para ser estimados independientemente. Se tiene que.

$$-2 \log t = -2 \sum_{i=1}^{k} f_i \left(\log \Theta_i^{o} - \log \frac{f_i}{n}\right) \qquad (2)$$

Ahora por expansión de series de Taylor del log Θ_i sobre

$$\frac{f_i}{n} = \hat{\Theta}_i$$

Tenemos

$$\log \Theta_i = \log \hat{\Theta}_i + (\Theta_i - \hat{\Theta}_i) \frac{1}{\hat{\Theta}_i} + \frac{(\Theta_i - \hat{\Theta}_i)^2}{2\,!} \left(-\frac{1}{\hat{\Theta}_i^2}\right) + E$$

ó

$$\log \Theta_i - \log \frac{f_i}{n} = \left(\Theta_i - \frac{f_i}{n}\right) \frac{n}{f_i} - \left(\Theta_i - \frac{f_i}{n}\right)^2 \frac{n^2}{2 f_i^2} + E$$

$$= \frac{(n \Theta_i - f_i)}{f_i} - \frac{(n \Theta_i - f_i^2)}{2 f_i^2} + E \qquad (3)$$

donde **E** representa la suma de términos alternados en el signo, como se presenta enseguida:

$$\sum_{j=3}^{\infty} (-1)^{j+1} \left(\Theta_i - \frac{f_i}{n}\right)^j \frac{n^j}{j! f_i^j}$$

Sustituímos (3) en (2)

$$-2 \log t = -2 \sum_{i=1}^{k} (n \theta_i^{o} - f_i) + \sum_{i=1}^{k} \frac{(n\theta_i - f_i)^2}{f_i} + \sum_{i=1}^{k} E'$$

$$= 0 + \sum_{i=1}^{k} \frac{(n \theta_i^{o} - f_i)^2}{f_i} + E''$$

Por la ley de los grandes números la variable aleatoria F_i/n es un estimador consistente de θ_i, ó

$$\lim_{n \to \infty} [\frac{1}{n} P(|F_i - n \theta_i| > E] = 0 \text{ para todo } E > 0$$

Así vemos que - 2 log T se aproxima asintóticamente al estadístico

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

Sugerido por Pearson en 1900, donde **n** observaciones han sido agrupadas en **k** categorías mutuamente excluyentes, y denota por f_i y e_i las frecuencias observadas y esperadas respectivamente, para el i-ésimo grupo, i=1,2,...,k.

Otra forma simple de llegar al estadístico, es la siguiente, que bajo cualquier distribución se pueden calcular las frecuencias esperadas e_i (i=1,2,...k, el número de clases) y se comparan con las frecuencias observadas f_i en las clases correspondientes: Se sabe que la suma de cuadrados tiene una forma parecida a la distribución de j_i-cuadrada con **n** infinita, en donde e_i

debe ser mayor que cinco para toda clase; para normalizar la suma de las desviaciones cuadradas, se dividen los cuadrados de las diferencias entre las esperanzas:

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

el que utilizamos como una estadística no paramétrica, y comparamos su valor con el valor crítico de una J_i cuadrada al nivel α con grados de libertad apropiados. Los g.l. son, en general k-r, donde **r** es el número de restricciones en los cálculos de los e_i.

Como se había mencionado anteriormente el estadístico desarrollado fue para una muestra , y que para comparar **k** muestras las cuales clasificadas cada una en C categorías tomaría pequeños cambios, lo cual se menciona a continuación.

El uso de la prueba X^2 es normalmente en tablas de contingencia, o en general, en tablas de muestras multiples con dos o más categorías o clasificaciones. Hay **R** renglones y **C** columnas. Se tienen f_{ij} observaciones en las celdas y las sumas $n_i.$ por cada renglón R_i, y $n._j$ por cada columna C_j. El total es **n..** , las frecuencias esperadas e_{ij} se obtienen $e_{ij} = \frac{(n_i.)(n.j)}{n..}$

A continuación se presenta una tabla para ilustrar lo anterior.

Población	Categoría 1	2	. . .	C	Σ
1	f_{11}	f_{12}	. . .	f_{1c}	$n_{1.}$
2	f_{21}	f_{22}	. . .	f_{2c}	$n_{2.}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
R	f_{r1}	f_{r2}	. . .	f_{rc}	$n_{r.}$
Σ	$n._1$	$n._2$	. . .	n.c	n..

Donde

f_{ij} = Número de individuos en la muestra de la población i-ésima que pertenecen a la categoría j-ésima.

$$(i = 1,2, \ldots, r \ ; \ j = 1,2, \ldots c)$$

$$n_{i.} = \sum_{j=1}^{c} f_{ij}$$

$$n._j = \sum_{i=1}^{r} f_{ij}$$

$$\sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n._j = n..$$

Quedando el estadístico como

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

El cual también se distribuye como una j_i - Cuadrada pero con (r-1) (c-1) grados de libertad.

A continuación presentamos un ejemplo para ilustrar la técnica.

En una comunidad campesina se tienen tres tipos de productores: Ejidatarios, arrendatarios y pequeños propietarios. Se toma una muestra de cada grupo y se clasifica a cada individuo de acuerdo al nivel de tecnología que utiliza. Se tienen 4 niveles de tecnología: N(nulo); P(pobre); M(mediano); A(alto). En la siguiente tabla se muestran los datos

	Clase				
Población	N	P	M	A	
Ejidatario	32.3 | 40	12 | 13	6.5 | 5	9.2 | 2	60
Arrendatario	16.2 | 20	6 | 5	3.2 | 3	4.6 | 2	30
Pequeño Prop.	21.5 | 10	8 | 8	4.3 | 6	6.2 | 16	40
	70	26	1	20	130

Las hipótesis a probar son

H_0: No hay diferencia en el nivel de tecnología en las tres poblaciones

Ha: Si hay diferencia en el nivel de tecnología en las tres poblaciones

Los valores esperados están anotados en el recuadro de las celdillas y fueron calculados como se indica.

$$e_{11} = \frac{60x70}{130} = 32.3$$

$$e_{12} = \frac{60x26}{130} = 12$$

$$e_{13} = \frac{60x14}{130} = 6.5$$

$$e_{14} = \frac{60x20}{130} = 9.2$$

Con estos cálculos y los sucesivos llenamos la tabla siguiente:

f_{ij}	e_{ij}	$f_{ij}-e_{ij}$	$(f_{ij}-e_{ij})^2$	$(f_{ij}-e_{ij})^2/e_{ij}$
40	32.3	7.7	59.29	1.84
13	12	1.0	1.00	.08
5	6.5	-1.5	2.25	.35
2	9.2	-7.2	51.84	5.63
20	16.2	3.8	14.44	.89
5	6.0	-1.0	1.00	.17
3	3.2	-0.2	.04	.01
2	4.6	-2.6	6.76	1.47
10	21.5	-11.5	132.25	6.15
8	8	0	0	0
6	4.3	1.7	2.89	.67
16	6.2	9.8	96.04	15.49
Σ 130.0	130.0	0		32.75

Asi $$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = 32.75$$

Con $\alpha = .01$, el valor crítico en la tabla de la j_i - cuadrada con (3-1) (4-1) = 6 g.l.

es X^2 .01, (6) = 16.81

$X^2 > X^2.01$ (6) por lo que rechazamos H_0 y se

concluye que el nivel de tecnología no es el mismo en las tres poblaciones.

Para usar esta prueba se deben tener las siguientes precauciones.

i) La aproximación es buena si $e_{ij} > 5$ en todas las celdas.

ii) Si algunas $e_{ij} < 1$ o más del 20 porciento son menores de 5, la aproximación es inadecuada.

iii) Si tanto r como c son grandes y las e_{ij} son aproximadamente iguales no importa que los valores de e_{ij} sean pequeños.

A continuación hacemos un resumen de la prueba.

a) Suposiciones: Cada muestra es aleatoria, los resultados de las diferentes muestras son independientes, cada observación puede catalogarse como miembro de c categorías mutuamente excluyentes y la escala es al menos nominal.

b) Hipótesis

$$H_0: P_{1j} = P_{2j} = \ldots = P_{kj}$$

c) Estadística de Prueba

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

d) Distribución de la estadística

$$X^2 \ (r-1)(c-1) \text{ g.l.}$$

e) Decisión

Si $X^2 > X^2 [\alpha, (r-1)(c-1)]$ se rechaza H_0

La Prueba de la Mediana

Esta prueba sirve para determinar si **k** grupos independientes, los cuales pueden no ser de tamaño igual, fueron tomados de la misma población o de poblaciones con medianas iguales. Además es utilizable cuando la variable en estudio se puede medir por lo menos en una escala ordinal.

Bajo la hipótesis nula H_0 de poblaciones idénticas (todas tienen la misma mediana),tenemos una sola muestra aleatoria de tamaño $\sum_{i=1}^{k} n_i = N$ de la población común. La mediana general δ de las muestras combinadas es un estimador de la mediana de esta población común. El conjunto de **N** observaciones mantendrá la hipótesis nula, si para cada una de las **k** muestras, alrededor de la mitad de las observaciones en esa muestra son menores que la mediana general. Esta mediana general será definida como la observación de las muestras combinadas ordenadas, la cual tiene rango $(n+1)/2$ si **N** es impar, y algún número entre las dos observaciones con rangos $\frac{N}{2}$ y $(N+2)/2$ si **N** es par. Luego para cada muestra separadamente,las observaciones son dicotomizadas de acuerdo como ellas sean menores o mayores que δ.

Ahora definimos la variable u_i como el número de obsevaciones en la i-ésima muestra las cuales son menores que δ, y t denota el número total de observaciones las cuales son menores que δ. Entonces por la definición de δ,

tenemos que

$$t = \sum_{i=1}^{k} u_i = \begin{cases} \frac{N}{2} & \text{si } N \text{ es par} \\ \frac{N-1}{2} & \text{si } N \text{ es impar} \end{cases}$$

Bajo la hipótesis nula, cada una de las $\binom{N}{t}$ posibles conjuntos de t observaciones son igualmente posibles para estar en la categoría menor que δ, y el número de dicotomizaciones con esta muestra particular es $\prod_{i=1}^{k} \binom{n_i}{u_i}$ por lo tanto la distribución de probabilidad nula de las variabales aleatorias es la extensión multivariada de la distribución hipergeométrica.

$$f(u_i, u_2, \ldots, u_k/t) = \frac{\binom{n_1}{u_1}\binom{n_2}{u_2}\cdots\binom{n_k}{u_k}}{\binom{N}{t}} \qquad (1)$$

Si alguna o todas las u_i difieren mucho de su valor esperado de $n_i\ominus$, donde $\ominus$ denota la probabilidad de que una observación de la población común sea menor que δ, la hipótesis nula sería rechazada. Generalmente sería impractico establecer regiones de rechazo conjuntas para los estadísticos de prueba $u_1, u_2, \ldots, u_k$; por que de la gran variedad de combinaciones de los tamaños de muestra n_1 $n_2, \ldots, n_k$ si una prueba exacta es deseada, esto es para calcular el valor de (1) para el observado real y más extremo $u_1, u_2, \ldots u_k$ y acumular estos puntos probables. Si la suma es menor que el nivel de significancia deseado, la hipótesis nula es rechazada.

Esta prueba es raramente llevada por este procedimiento por que es tedioso y lento, igualmente con las tablas de coeficientes de la binomial, a menos que K y n_i sean pequeños. Afortunadamente se pude usar otra prueba de hipótesis, la cual se hace mediante una aproximación, que es razonablemente exacta para N tan pequeña como 25, si cada muestra consiste de al menos 5 observaciones. Cada uno de los N elementos en la muestra combinada es clasificado de acuerdo a dos criterios, número de muestra y magnitud relativa a δ. Estas dos categorías son denotadas por (i,j), donde i=1,2,...,k de acuerdo al número de muestra; y j=1 si la observación es menor que δ y j=2 de otra manera. Denotaremos las frecuencias observadas O_{ij} y esperadas E_{ij} para (i,j) categoría.

Luego $o_{i1} = u_i$

Para i = 1,2,...,K

$o_{i2} = n_i - u_i$

y las frecuencias esperadas bajo H_0 son estimadas de los datos

$$e_{i1} = \frac{n_i t}{N}$$

Para i = 1,2,...,k

$$e_{i2} = \frac{n_i (N-t)}{N}$$

Así

$$M = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{(o_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{k} \frac{(o_{i1} - e_{i1})^2}{e_{i1}} + \sum_{i=1}^{k} \frac{(o_{i2} - e_{i2})^2}{e_{i2}}$$

Donde M representa la mediana

Sustituímos los e_{ij} y los o_{ij}

$$M = \sum_{i=j}^{k} \frac{(u_i - n_i t/N)^2}{n_i t/N} + \sum_{i=1}^{k} \frac{(n_i - u_i - n_i (N-t)/N)^2}{n_i (N-t)/N}$$

$$= N \sum_{i=1}^{k} \frac{(u_i - n_i t/N)^2}{n_i t} + N \sum_{i=1}^{k} \frac{(n_i t/N - u_i)^2}{n_i (N-t)}$$

$$= N \sum_{i=1}^{k} \frac{(u_i - n_i t/N)^2}{n_i} \left(\frac{1}{t} + \frac{1}{N-t} \right)$$

$$= \frac{N^2}{t(N-t)} \sum_{i=1}^{k} \frac{(u_i - n_i t/N)^2}{n_i} \qquad (2)$$

Donde (2) tiene aproximmadamente la distribución j_i cuadrada. Los parámetros estimados de los datos son las 2 k probabilidades de que una observación es menor que δ para cada una de las k muestras, y que no es menor que δ. Pero para cada muestra estas probabilidades suman 1.0, y así hay solamente **k** parámetros independientes estimados.

El número de grados de libertad para (2) es entonces 2k-1-k, ó k-1. Teniendo como región de rechazo para prueba $M \geq X^2 (k-1, 1-\alpha)$

A continuación se presenta un ejemplo para aplicar la prueba:

Cuatro diferentes fertilizantes son usados cada uno en seis diferentes campos, y el experimento entero es repetido usando tres diferentes tipos de semilla. El rendimiento por acre es calculado bajo cada una de las (4)(6)(3) = 72 diferentes condiciones con los siguientes resultados.

	Semilla 1 Fertilizantes				Semilla 2 Fertilizantes			
	1	2	3	4	1	2	3	4
1	80.5	90.1	87.0	88.0	79.1	87.0	82.6	81.5
2	87.0	83.4	89.1	90.3	77.6	82.0	81.4	87.9
Campo 3	86.1	88.4	91.0	86.1	84.1	80.6	89.0	80.4
4	82.1	84.9	84.4	83.1	83.3	79.5	86.3	83.1
5	79.3	87.1	92.2	90.8	76.6	86.2	84.0	87.4
6	84.2	89.3	85.3	84.7	81.0	84.1	88.1	85.0

Semilla 3 Fertilizantes			
1	2	3	4
85.4	92.3	92.0	89.3
89.2	90.1	90.2	93.6
90.0	88.1	87.2	90.8
83.1	85.3	94.3	87.6
87.4	86.3	88.4	93.7
82.3	92.9	95.1	32.9

Probaremos la hipótesis nula:

H_0 No hay diferencia en la mediana de los rendimientos esperados para los diferentes fertilizantes.

Para probar la hipótesis se deben ordenar las 72 observaciones y calcular la mediana de la muestra combinada (conjunta). Como se muestra enseguida.

76.6(1)	77.6(1)	79.1(1)	79.3(1)	79.5(2)	80.4(4)
80.5(1)	80.6(2)	81.0(1)	81.4(3)	81.5(4)	82.0(2)
82.1(1)	82.3(1)	82.4(2)	82.6(3)	82.9(4)	83.1(4)
83.1(4)	83.3(1)	83.4(2)	83.4(1)	84.0(3)	84.1(1)
84.1(2)	84.2(1)	84.4(3)	84.7(4)	84.9(2)	85.0(4)
85.3(2)	85.3(3)	85.4(1)	86.1(1)	86.1(4)	86.2(2)
86.3(3)	86.3(2)	87 (1)	87.0(3)	87.0(2)	87.1(2)
87.1(1)	87.2(3)	87.4(4)	87.6(4)	87.4(4)	88 (4)
88.1(2)	88.1(3)	88.4(3)	89.0(3)	89.1(3)	89.2(1)
89.3(2)	89.3(4)	90.0(1)	90.1(2)	90.1(2)	90.2(3)
90.3(4)	90.8(4)	90.8(4)	91.0(3)	92.0(3)	92.2(3)
92.3(2)	92.9(2)	93.6(4)	93.7(4)	94.3(3)	95.1(3)

El número entre paréntesis significa la muestra relativa a cada uno de los fertilizantes.

Ahora clasificamos las observaciones de acuerdo a la muestra de que provienen y si son mayores ó menores (ó igual) a la mediana general.

	Fertilizantes			
	1	2	3	4
No de valores mayores a la mediana general	4	9	13	10
No de valores menores ó iguales a la mediana general	14	9	5	8

donde la mediana general = 86.25

Ahora

$$M = \frac{(72)}{36(72.36)} \left\{ \frac{\left[4 - \frac{(18)(36)}{72}\right]^2 + \left[9 - \frac{(18)(36)}{72}\right]^2 + \left[13 - \frac{(18)(36)}{72}\right]^2}{18} + \frac{\left[10 - \frac{(18)(36)}{72}\right]^2}{} \right\}$$

$$= \frac{5184}{1296} \left\{ \frac{(4-9)^2 + (9-9)^2 + (13-9)^2 + (10-9)^2}{18} \right\}$$

$$= 4 \left\{ \frac{25+16+1}{18} \right\} = 9.33$$

Si trabajamos con un nivel de significancia de $\alpha = 0.05$ tenemos que $X^2\ 0.95,3 = 7.81$

entonces $M > 7.81$

Por lo que rechazamos H_0, y concluímos que hay diferencias en los resultados por el fertilizante utilizado.

A continuación un resumen de la prueba:

a) Suposiciones: se toman **K** muestras independientes de tamaño u_i

b) Hipótesis: H_0: $m_1 = m_2 = \ldots = m_k$

c) Estadística $\frac{N^2}{t(N-t)} \sum_{i=1}^{k} \frac{(u_i - n_i\ t/N)^2}{n_i}$

d) Distribución X^2 (k-1) g.l.

e) Regla de decisión: Si $M > X^2\ (K-1,\ 1-\alpha)$ se rechaza H_0

Prueba de Kruskal y Wallis

El análisis de varianza de una clasificación por rangos de Kruskal-Wallis es una prueba para decidir si **k** muestras independientes son de poblaciones diferentes. Esta técnica examina la hipótesis de nulidad que supone que las k muestras proceden de la misma población o poblaciones idénticas con respecto a los promedios. Definiendo que las diferencias entre las muestras signifiquen verdaderas diferencias de población o simples variaciones aleatorias.

La prueba supone que la variable en estudio tiene como base una distribución continua y, requiere por lo menos, una medida ordinal de la variable.

La distribución exacta del estadístico es encontrada bajo la suposición de que todas las observaciones fueron obtenidas de la misma o idénticas poblaciones. Esto es, por la suposición precedente, cada ordenamiento de los rangos 1 a N (donde N es el rango máximo de las observaciones ordenadas) dentro de los grupos de tamaños n_1 n_2,..., n_k; respectivamente, es igualmente posible, y ocurre con probabilidad: $n_1!$ $n_2!...n_k!$ /N!, lo cual es el recíproco del número de formas que los N rangos pueden ser divididos dentro de los grupos de tamaño $n_1,n_2,...,n_k$. Las probabilidades asociadas con valores iguales del estadístico son sumadas para obtener la distribución de probabilidad del estadístico.

Por ejemplo, si $n_1=2$, $n_2=1$ y $n_3=1$ en el caso de tres muestras, estos son doce ordenamientos igualmente posibles de los cuatro rangos; por lo tanto cada ordenamiento tiene probabilidad $\frac{1}{12}$. Los doce ordenamientos con el valor asociado del estadístico, son los siguientes:

ordenamiento	muestras			valor
	1	2	3	
1	1,2	3	4	2.7
2	1,2	4	3	2.7
3	1,3	2	4	1.8
4	1,3	4	2	1.8
5	1,4	2	3	0.3
6	1,4	3	2	0.3
7	2,3	1	4	2.7
8	2,3	4	1	2.7
9	2,4	1	3	1.8
10	2,4	3	1	1.8
11	3,4	1	2	2.7
12	3,4	2	1	2.7

Por consiguiente la función de probabilidad f(x) y la función de distribución F(x) son dados como sigue para $n_1=2$, $n_2=1$ y $n_3=1$

X	f(x) = P(T = X)	F(x) = P(T ≤ x)
0.3	2/12 = 1/6	1/6
1.8	4/12 = 1/3	1/2
2.7	6/12 = 1/2	1.0

Ahora, tenemos datos que consisten de k muestras aleatorias posiblemente de tamaños diferentes. Denotamos la i-ésima muestra aleatoria de tamaño ni por X_{i1} $X_{i2},\ldots,X_{in}$. Luego los datos pueden ser ordenados en columnas.

Muestra 1	Muestra 2	...Muestra k
$X_{1,1}$	$X_{2,1}$	$X_{k,1}$
$X_{1,2}$	$X_{2,2}$	$X_{k,2}$
...	...	...
$X_{1,n}$	$X_{2,n}$	X_{kn}

Sea N el total del número de observaciones

$$N = \sum_{i=1}^{k} n_i$$

Asignamos el rango 1 a la más pequeña del total de N observaciones, rango 2 a la segunda más pequeña, asi hasta la más grande de todas las observaciones, la cual recibe el rango **N**. Sea $R(X_{ij})$ el rango asignado a X_{ij}. Y R_i la suma de los rangos asignados a la i-ésima muestra.

$$R_i = \sum_{i=1}^{ni} R(x_{ij}) \qquad i = 1,2,\ldots,k$$

En el caso de que existan rangos empatados, se les asignará el promedio de los rangos que les hubiera correspondido a cada una de las observaciones si no existiera empate.

La aproximación para la distribución del estadístico en una muestra grande se basa en que R_i es la suma de n_i variables aleatorias, y para n_i grande el teorema del límite central puede ser aplicado.

Esto es

$$\frac{R_i - E(R_i)}{\sqrt{\text{var}\ (R_i)}} \qquad (1)$$

es aproximadamente distribuído como una variable aleatoria normal estándar, cuando H_o es verdadera.

Donde

$$E\ [R_i] = E\left[\sum_{i=1}^{ni} R\ (X_{ij})\right] = E\left[\sum_{i=1}^{ni} R_{ij}\right]$$

$$= \sum_{i=1}^{ni} E\left[R_{ij}\right]$$

$$= \sum_{i=1}^{ni}\sum_{j=1}^{N} j\ \left(\frac{1}{N}\right) = \frac{1}{N}\sum_{i=1}^{ni}\sum_{j=1}^{N} j$$

$$= \frac{1}{N}\sum_{i=1}^{ni} \frac{N(N+1)}{2}$$

$$= \frac{ni\ (N+1)}{2}$$

y

$$\text{Var}\ (R_i) = \sum_{i=1}^{ni} \text{Var}\ [R_i(X_{ij})] + \sum_{i=1}^{ni} \sum_{\substack{j=1 \\ i \neq j}}^{ni} \text{Cov}\ [R_i(X_{ij})][R_j(X_{ij})]$$

Luego

$$\text{Var}\ (R_i(X_{ij})) = \sum_{i=1}^{N} \left(i - \frac{N+1}{2} \right)^2 \frac{1}{N}$$

$$= \frac{1}{N} \left[\sum_{i=1}^{N} \left(i^2 - 2i \left(\frac{N+1}{2} \right) + \left(\frac{N+1}{2} \right)^2 \right) \right]$$

$$= \frac{1}{N} \left[\sum_{i=1}^{N} i^2 - (N+1) \sum_{i=i}^{N} i + N \left(\frac{N+1}{4} \right)^2 \right]$$

$$= \frac{1}{N} \left[\frac{N(N+1)(2N+1)}{6} - \frac{(N+1)N(N+1)}{2} + \frac{N(N+1)^2}{4} \right]$$

$$= \frac{1}{N} \left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{2} + \frac{N(N+1)^2}{4} \right]$$

$$= \frac{1}{N} \left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4} \right]$$

$$= \frac{1}{N} \left[\frac{N(N+1) \quad [2\ (2N+1) - 3(N+1)]}{12} \right]$$

$$= \frac{1}{N} \left[N(N+1) \ (\frac{4N+2 \ - \ 3N-3)}{12} \right]$$

$$= \frac{1}{N} \ [\frac{N(N+1)\,(N-1)}{12}] \ = \ \frac{(N+1)\,(N-1)}{12}$$

Por lo que

$$\sum_{i=1}^{n_i} \text{Var} \left[R_i(X_{ij}) \right] = \frac{n_i\,(N+1)\,(N-1)}{12} \qquad (4)$$

Por otro lado por definición de la covarianza.

$$\text{Cov} \left[R_i(X_{ij}), R_j(X_{ij})\right] = E \left\{ [(R_i(X_{ij}) - E(R_i(X_{ij}))][R_j(X_{ij}) \right.$$

$$\left. - E(R_j(X_{ij}))] \right\}$$

$$= E\left[R_i(X_{ij}) \ R_j(X_{ij})\right] \ - \ E[R_i(X_{ij})] \ E \ [R_j \ (X_{ij})]$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \left(i - \frac{N+1}{2}\right)\left(j - \frac{N+1}{2}\right)\frac{1}{(N-1)\ (N)} - \sum_{i=1}^{N} \left(i \ - \ \frac{N+1}{2}\right)^2 \ \frac{1}{(N-1)N}$$

$$= \frac{1}{(N-1)N} \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right) \sum_{j=1}^{N}\left(j - \frac{N+1}{2}\right) - \frac{1}{N-1} \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2 \frac{1}{N} \qquad (5)$$

En la ecuación (5) observamos que

$$\sum_{i=1}^{N} \left(i - \frac{N+1}{2} \right) = \sum_{i=1}^{N} i - \sum_{i=1}^{N} \frac{N+1}{2} = \frac{N(N+1)}{2} - \frac{N(N+1)}{2} = 0$$

Por lo tanto el primer término de la ecuación (5) es cero. Además de la definición de varianza se tiene que

$$\text{Var } (R_i(X_{ij})) = \sum_{i=1}^{N} \left(i - \frac{N+1}{2} \right)^2 \frac{1}{N} = \frac{(N+1)(N-1)}{12}$$

y sustituyendo esta definición en la ecuación (5), tenemos:

$$= - \frac{1}{(N-1)} \frac{(N+1)(N-1)}{12}$$

Así

$$\text{Cov } (R_i(X_{ij}), R_j(X_{ij}) = - \frac{N+1}{12} \qquad (6)$$

Por lo que

$$\sum_{i=1}^{ni} \sum_{\substack{j=1 \\ i \neq j}}^{ni-1} \text{Cov } (R_i(x_{ij}), R_j(x_{ij})) = n_i(n_i-1) \left(- \frac{N+1}{12} \right) \qquad (7)$$

Sustituímos (4) y (7) en (3)

$$\text{Var } (R_i) = \frac{n_i (N+1)(N-1)}{12} + n_i (n_i-1) \left(- \frac{N+1}{12} \right)$$

$$= \frac{n_i (N^2-1)}{12} + \left(- \frac{Nn_i^2 - n_i^2 + Nn_i + n_i}{12} \right)$$

$$= \frac{N^2 n_i - n_i - N n_i^2 - n_i^2 + N n_i + n_i}{12}$$

$$= \frac{N^2 n_i - N n_i^2 - n_i^2 + N n_i}{12}$$

$$= \frac{n_i (N_i^2 - N n_i - n_i + N)}{12}$$

$$= \frac{n_i (N+1)(N-n_i)}{12} \qquad (8)$$

Por último sustituímos (4) y (8) en (1)

$$\frac{R_i - n_i (N+1)/2}{\sqrt{n_i (N+1)(N-n_i)/12}}$$

Así la suma

$$\sum_{i=1}^{k} \frac{[R_i - n_i (N+1)/2]^2}{n_i (N+1)(N-n_i)/12}$$

Es aproximadamente distribuída como una v.a. j_i-cuadrada con k-1 g.l., en caso de que los Ri's fueran independientes. Sin embargo, la suma de los Ri's es $\frac{N(N+1)}{2}$, entonces hay una dependencia entre los Ris!

Donde (9) se conoce como el estadístico de prueba H, de Kruskal y Wallis.

La prueba de rangos presentada aquí requiere que todas las observaciones se categoricen juntas, además de la suma de los rangos obtenidos para cada muestra.

El estadístico de prueba es calculado si no hay empates (esto es, si dos observaciones no son iguales).

Por $$H = \frac{12}{N(N+1)} \sum_{i\ 1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

donde K = número de muestras

n_i = número de observaciones en la i-ésima muestra

$N = \Sigma n_i$ número de observaciones en toda la muestra combinada

Ri = Suma de los rangos en la i-ésima muestra.

Valores grandes de H llevan el rechazo de la hipótesis nula.

Si las muestras vienen de poblaciones idénticas y las n_i no son pequeñas, H es distribuída como j_i- cuadrada con k-1 g.l., permitiendo usar las tablas de la X^2. Para

c=3 y $n_i \leq 5$, existen tablas de exactitud para estos valores.

Cuando existen empates entre las observaciones, se calcula H y se divide por $1 - \frac{\Sigma T}{N^3 - N}$

Donde la sumatoria es sobre todos las grupos de empates y $T = (t-1)\ t\ (t+1) = t^3 - t$, para cada grupo de empates, t es el número de observaciones empatadas en el grupo.

La corrección del efecto de las ligas resulta en un incremento del valor de H, y de este modo el resultado es aún más significativo de lo que habría sido sin la corrección. Por consiguiente si se puede rechazar H_0 sin la corrección, se podrá rechazar H_0 en un nivel de significación aún más severo por medio de la corrección.

En la mayoría de los casos, el efecto de la corrección es insignificante. Si no hay de un veinticinco porciento de observaciones ligadas, la probabilidad asociada con H calculada sin la corrección, rara vez cambia en más del diez porciento cuando se corrige el efecto de las ligas.

A continuación presentamos un ejemplo para utilizar las tablas de probabilidad exactas. Con $k=3$ y $n_i \leq 5$

Se tienen tres variedades de trigo y se plantan tres unidades experimentales de la variedad A, tres de la

variedad B y dos de la variedad C. Se quieren comparar los días a la floración de las tres variedades, los datos observados son:

	Variedad		
	A	B	C
Días a floración	70	60	62
	72	66	63
	78	50	

Hacemos un cambio de las observaciones por sus respectivos rangos.

	A	B	C
	6	2	3
	7	5	4
	8	1	
R_j =	21	8	7
n_j =	3	3	2

$$N = \sum_{j=1}^{3} n_j = 8$$

$$H = \frac{12}{8(9)} \left[\frac{(21)^2}{3} + \frac{(8)^2}{3} + \frac{(7)^2}{2} \right] - 3(9) = 5.1389$$

De la tabla para $k = 3$ y $n_i \leq 5$

Buscamos con $n_1 = 3$, $n_2 = 3$, $n_3 = 2$

$$\hat{\alpha} = P\ (H \leq 5.1389) = .061$$

De manera que rechazamos H_0 con cualquier $\alpha > .061$

Así que con $\alpha = .05$ no rechazamos H_0

Enseguida presentamos un ejemplo para $K \neq 3$ y $n_i > 5$.

Un investigador anotó los pesos que tenían al nacer los miembros de ocho camadas diferentes de cerdos, para determinar si el peso (en lb.) al nacer es afectado por el tamaño de la camada.

Camadas

	1	2	3	4	5	6	7	8
	2.0	3.5	3.3	3.2	2.6	3.1	2.6	2.5
	2.8	2.8	3.6	3.3	2.6	2.9	2.2	2.4
	3.3	3.2	2.6	3.2	2.9	3.1	2.2	3.0
Pesos	3.2	3.5	3.1	2.9	2.0	2.5	2.5	1.5
	4.4	2.3	3.2	3.3	2.0		1.2	
	3.6	2.4	3.3	2.5	2.1		1.2	
	1.9	2.0	2.9	2.6				
	3.3	1.6	3.4	2.8				
	2.8		3.2					
	1.1		3.2					

Luego hacemos un ordenamiento general de las observaciones

Camadas

	1	2	3	4	5	6	7	8
	8.5	52.5	47.5	41.0	23.0	36.0	23.0	18.5
	27.5	27.5	54.5	47.5	23.0	37.5	12.5	15.5
	47.5	41.0	23.0	41.0	31.5	36.0	12.5	34.0
Pesos	41.0	52.5	36.0	31.5	8.5	18.5	18.5	4.0
	56.0	14.0	41.0	47.5	8.5		2.5	
	54.5	15.5	47.5	18.5	11.0		2.5	
	6.0	8.5	31.5	23.0				
	47.5	5.0	51.0	27.5				
	27.5		41.0					
	1.0		41.0					
	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8

$R_1 = 317.0$, $R_2 = 216.5$, $R_3 = 414.0$, $R_4 = 277.5$, $R_5 = 122.0$, $R_6 = 122.0$, $R_7 = 71.5$, $R_8 = 72.0$

Calculamos el valor de H.

$$H = \frac{12}{56(56+1)}\left[\frac{(317)^2}{10} + \frac{(216)^2}{8} + \frac{(414)^2}{10} + \frac{(277.5)^2}{6} + \frac{(122)^2}{4} + \frac{(122)^2}{6} + \frac{71.5)^2}{6} + \frac{(72)^2}{4}\right] - 3\,(56+1)$$

$= 18.464$

La referencia de la tabla de la X^2 indica que una H $\geq$ 18.464 con g.l. = k-1 = 7

Tiene una probabilidad de ocurrencia conforme a H_0 de $p < 0.02$

Como se presentaron ligas en el experimento veremos cual fué su efecto calculando el factor de corrección.

Hubo una primera liga entre dos cerdos de la camada siete, de dos y tres por lo tanto se les signa el promedio 2.5; como fueron solo dos observaciones el valor de t = 2, así $T = t^3 - t = 8 - 2 = 6$

En total se presentaron trece grupos ligados, los cuales presentamos a continuación.

t	2	4	2	2	4	5	4	4	3	7	6	2	2
T	6	60	6	6	60	120	60	60	24	336	210	6	6

Con estos datos calculamos: $1 - \frac{\Sigma T}{N^3 - N}$

$$= 1 - \frac{(6+60+6+6+60+120+60+60+24+336+210+6+6)}{(56)^3 - 56}$$

$$= 0.9945$$

y corregimos H

$$H = \frac{18.464}{0.9945} = 18.566$$

La tabla muestra que la probabilidad asociada con la ocurrencia conforme a H_0 de un valor tan grande como H = 18.566, gl.= 7, es p < 0.01 como este valor es menor que el nivel de significancia fijado de $\alpha = 0.05$, se rechaza H_0. Y se puede concluir que el peso al nacimiento de los cerdos varía significativamente de acuerdo con el tamaño de la camada.

A continuación resumimos la prueba

Esta técnica sería el equivalente de bloques completamente al azar en la versión paramétrica.

a) Suposiciones:

Todas las muestras son muestras aleatorias de sus respectivas poblaciones.

Todas las variables X_{ij} son continuas

La escala de medición es al menos ordinal.

b) Hipótesis: H_0: Las K poblaciones son iguales

Ha: Al menos una de las poblaciones tiene valores mayores que las otras.

c) Estadístico: $H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\ (N+1)$

c) Distribución: Si K = 3 y $n_i \leq 5$ de las tablas exactas

Si K ≠ 3 o $n_i > 5$ $X^2\ (k-1)$

Decisión:

Si $K = 3$ y $n_i \leq 5$ rechazamos H_0 si $H > \alpha$

Si $K \neq 3$ ó $n_i > 5$ rechazamos H_0 si $H > X^2 (k-1, 1-\alpha)$

5. MEDIDAS DE ASOCIACION

(Medidas de correlación de rangos)

Una medida de correlación es una variable aleatoria que es usada en situaciones donde los datos consisten de pares de números (X_1,Y_1) , $(X_2,Y_2),\ldots,(Y_n, Y_n)$, que representan n pares de observaciones. Los (X_i,Y_i) para $i= 1,2,\ldots,n$ tienen distribuciones bivariadas idénticas.

Para ser aceptada una medida de correlación entre X y Y, debe satisfacer lo siguiente.

a) Si los valores grandes de X e Y tienden a aparearse, y lo mismo sucede con valores chicos de X e Y; la medida es positiva y tiende a uno, a medida que la relación sea más fuerte. Por ejemplo, si para cualesquiera dos pares independientes $(X_i\ Y_i)$ y (X_j, Y_j) de variables aleatorias independientes; $X_i < X_j$ siempre que $Y_i < Y_j$ (o bien $X_i > X_j$ siempre que $Y_i > Y_j$), la medida tiende a ser uno, y disminuye cuando la relación no es tan marcada, pero permanece como un valor positivo.

b) Si valores grandes de X tienden a aparearse con valores chicos de Y o viceverza, la medida es negativa, y según sea más marcada la relación tenderá a menos uno.

Por ejemplo si tenemos dos pares (X_i,Y_i) y (X_j,Y_j), luego $X_i < X_j$ siempre que $Y_i > Y_j$, entonces el valor de la medida tenderá a ser menos uno, y aumentará a medida que la tendencia disminuya, pero siempre siendo negativa.

c) Como se pudo haber observado la medida solo tomará valores entre [-1,1].

d) Si las variables (X,Y) fueron obtenidas aleatoriamente, es decir que no hay tendencia (son independientes), la medida se aproxima a cero.

Cuando se presenta el caso (a), existe correlación positiva entre X e Y, y si el valor de la medida es uno, la concordancia es perfecta.

Ahora, si sucede el caso (b), la correlación entre X e Y es negativa, y si su valor es menos uno, hay discordancia perfecta. Por último si ocurre.(d), no existe correlación entre X e Y.

Lo mencionado anteriormente es común a las medidas de asociación en general, pero existen algunas medidas de asociación cuya función de distribución no depende de la función de distribución bivariada de (X,Y), es decir estas medidas de asociación están en función de los rangos asignados a las observaciones, tales medidas son las

medidas de correlación de rangos, y son usadas en pruebas de independencia de estadística no paramétrica.

La medida de correlación, dá el grado de asociación que existe entre las dos variables; pero, también es importante afirmar si el encontrar correlación en la muestra, significa que las poblaciones estén relacionadas o son fluctuaciones del azar, por lo que la construcción de hipótesis y la consiguiente prueba de hipótesis, se hace necesario.

Coeficiente de Correlación de Rango de Spearman: ρs

Una muestra aleatoria de n pares (X_1, Y_1), $(X_2, Y_2), \ldots (X_n, Y_n)$ es extraída de una población bivariada, con el coeficiente de correlación ρ producto-momento de Pearson.

En estadística clásica, el estimador comunmente usado para ρ, es el coeficiente de correlación muestral definido por

$$R = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2\right]^{\frac{1}{2}}} \qquad (1)$$

En general, la distribución muestral de la R depende de la forma de la población bivariada, de la cual la muestra de pares es extraída. Pero, ahora supongamos que las X observaciones son ordenadas de menor a mayor usando los enteros 1,2,...,n,; y las Y observaciones son ordenadas por separado, usando el mismo esquema de ordenamiento. Es decir a cada observación se le asigna un rango de acuerdo a su magnitud relativa a otras dentro de su mismo grupo. Si las distribuciones marginales de X e Y se suponen continuas, el conjunto único de rangos existe teoricamente. Entonces los datos consisten de n grupos de pares de rangos a partir de los cuales, como se definió en (1), R puede ser calculado. El estadístico resultante es el "coeficiente de correlación de rango de Spearman". Este mide el grado de correspondencia entre los ordenamientos, en lugar de medirlo entre los diferentes valores, pero puede ser considerado sin embargo una medida de asociación entre las muestras y un estimador de la asociación entre X e Y en la población contínua bivariada.

El hecho que conozcamos además de los valores numéricos de las observaciones derivadas a partir de los cuales R de Spearman es calculado, su esquema de apareamiento, significa que la expresión en (1) puede ser simplificado considerablemente.

Denotando los rangos respectivos de las variables aleatorias en las muestras por:

$$R_i = \text{rango } (X_i) \qquad y \qquad S_i = \text{rango } (Y_i)$$

Las observaciones muestrales derivadas de n pares son $(r_1,s_1), (r_2,s_2),\ldots, (r_n,s_n)$

luego

$$r_i,s_i = 1,2,\ldots,n \quad \text{para } i = 1,2,\ldots,n$$

Observamos los valores constantes para todas las muestras

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \quad y \quad \bar{r} = \bar{s} = \frac{n+1}{2} \qquad (2)$$

Ahora $$\sum_{i=1}^{n} (r_i - \bar{r})^2 = \sum_{i=1}^{n} (s_i - \bar{s})^2 = \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right)^2$$

$$= \sum_{i=1}^{n} \left[i^2 - 2i \frac{(n+1)}{2} + \frac{(n+1)^2}{4} \right]$$

$$= \sum_{i=1}^{n} i^2 - (n+1) \sum_{i=1}^{n} i + n \frac{(n+1)^2}{4}$$

Donde $$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Por lo que

$$= \frac{n(n+1)(2n+1)}{6} - (n+1)\ \frac{[n(n+1)]}{2} + \frac{n(n+1)^2}{4}$$

$$= \frac{2n^3+ 3n^2+n}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)^2}{4}$$

$$= \frac{4n^3+ 6n^2+ 2n - 3n^3 - 6n^2 - 3n}{12}$$

$$= \frac{n^3- n}{12}$$

$$= \frac{n(n^2- 1)}{12} \qquad (3)$$

Sustituímos en **(1)** las constantes obtenidas

$$\rho s = \frac{\sum_{i=1}^{n} (R_i-\bar{R})(S_i-\bar{S})}{\left\{ \left[\frac{n(n^2-1)}{12} \right] \left[\frac{n(n^2-1}{12} \right] \right\}^{\frac{1}{2}}}$$

$$= \frac{\sum_{i=1}^{n} R_i S_i - n\bar{S} \sum_{i=1}^{n} R_i - n\bar{R} \sum_{i=1}^{n} S_i + n\ \bar{S}\ \bar{R}}{\frac{n(n^2-1)}{12}}$$

$$= \frac{\sum_{i=1}^{n} R_i S_i - n\ \bar{R}\ \bar{S} - n\ \bar{R}\ \bar{S} + n\ \bar{S}\ \bar{R}}{\frac{n\ (n^2-1)}{12}}$$

$$= \frac{12 \left[\sum_{i=1}^{n} R_i S_i - \frac{n(n+1)^2}{4} \right]}{n\ (n^2-1)} \qquad (4)$$

Si hacemos

$$D_i = R_i - S_i = (R_i - \bar{R}) - (S_i - \bar{S})$$

Luego

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (R_i - \bar{R})^2 + \sum_{i=1}^{n} (S_i - \bar{S})^2 - 2 \sum_{i=1}^{n} (R_i - \bar{R})(S_i - \bar{S})$$

$$= \frac{n(n^2-1)}{12} + \frac{n(n^2-1)}{12} - 2 \Sigma R_i S_i + \frac{n(n+1)^2}{2}$$

$$= \frac{2n(n^2-1)}{12} + \frac{n(n+1)^2}{2} - 2 \Sigma R_i S_i$$

$$\Sigma R_i S_i = \frac{n(n^2-1)}{12} + \frac{n(n+1)^2}{4} - \frac{\sum_{i=1}^{n} D_i^2}{2} \qquad (5)$$

Sustituímos (5) en (4)

$$\rho s = 12 \left[\frac{n(n^2-1)}{12} + \frac{n(n+1)^2}{4} - \frac{\sum_{i=1}^{n} D_i^2}{2} - \frac{n(n+1)^2}{4} \right] / n(n^2-1)$$

$$= 12 \left[\frac{n(n^2-1) - 6(\Sigma D_i^2)}{12} \right] / n(n^2-1) = \frac{n(n^2-1) - 6 \sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

$$\rho s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad (6)$$

Ahora, puesto que se está trabajando con rangos pueden existir dos o más observaciones iguales, y asociarles el mismo rango (esto es el promedio de los

rangos que les hubiera tocado si no fueran iguales las observaciones). En el caso de ρs, estos hacen que el valor de $\sum_{i=1}^{n}(R(X_i)-(\bar{R}(X))^2$ disminuya, si es que los empates ocurren en los valores $X_1,\ldots,X_n$ (de igual forma para Y), afectando el valor de ρs. La suma de cuadrados se corrige restándole $\Sigma T = \Sigma\frac{t^3-t}{12}$ donde t es el número de observaciones ligadas en un rango dado, y ΣT indica que se suman los diversos valores de T para todos los grupos de observaciones ligadas.

Veamos como es que se reduce la suma de cuadrados al presentarse las ligas entre las observaciones.

Tenemos: $$\sum_{i=1}^{n}(S_i - \bar{S})^2 = \sum_{i=1}^{n} S_i^2 - \frac{n(n+1)^2}{2}$$

Donde hay uno o más grupos de t observaciones ligadas dentro de la muestra Y, y a cada una es asignada el promedio de los rangos correspondientes en caso de que no hubiera habido ligas. En cada grupo de t observaciones ligadas, si no hubiera ligas, se les asignarían los rangos P_k+1, $P_k+2,\ldots,P_k+t$, el rango asignado al total es.

$$\sum_{i=1}^{t}\left(\frac{P_k + i}{t}\right) = P_k + \frac{(t+1)}{2} \qquad (7)$$

Luego la suma de cuadrados para esos rangos ligados es

$$t\left(P_k + \frac{t+1}{2}\right)^2 = t\left[P_k^2 + P_k\,(t+1) + \frac{(t+1)^2}{4}\right] \qquad (8)$$

y la correspondiente suma en la ausencia de ligas sería

$$\sum_{i=1}^{t}(P_k + i)^2 = t\,P_k^2 + P_k t(t+1) + \frac{t(t+1)(2t+1)}{6} \qquad (9)$$

Como se puede observar el grupo particular de t observaciones ligadas disminuye la suma de cuadrados por la diferencia entre (9) y (8), en sus últimos términos ,o sea

$$\frac{t(t+1)(2t+1)}{6} - \frac{t(t+1)^2}{4} = \frac{t(t^2-1)}{12} = \frac{t^3-t}{12} \qquad (10)$$

Así, (10) es el factor de corrección a utilizar.

Una vez obtenído el factor de corrección la fórmula apropiada para aplicarlo es la siguiente.

$$\rho s = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d_i^2}{2\sqrt{\Sigma x^2 \; \Sigma y^2}} \qquad (11)$$

donde $\Sigma x^2 = \frac{N^3 - N}{12} - \Sigma Tx$

$\Sigma y^2 = \frac{N^3 - N}{12} - \Sigma Ty$

La fórmula (11) es conformada de la siguiente manera: si $x = x - \bar{x}$, donde $\bar{x}$ es la media de los puntajes en la variable x, y si $y = Y - \bar{Y}$, la expresión general para un coeficiente de correlación puede ser:

$$\rho s = \frac{\Sigma\, x\, y}{\sqrt{\Sigma x^2\ \Sigma y^2}}$$

Si hacemos

$$d = x-y$$

$$d^2 = (x-y)^2 = x^2 - 2xy + y^2$$

$$\Sigma\, d^2 = \Sigma\, x^2 + \Sigma\, y^2 - 2\Sigma\, xy \qquad (12)$$

Además

$$\rho s = \frac{\Sigma\, x\, y}{\sqrt{\Sigma\, x^2\ \Sigma\, y^2}}$$

Luego $\Sigma\, xy = \rho s \sqrt{\Sigma\, x^2\ \Sigma\, y^2}$ (13)

Sustituímos (13) en (12)

$$\Sigma\, d^2 = \Sigma\, x^2 + \Sigma\, y^2 - 2\ \rho s \sqrt{\Sigma\, x^2\ \Sigma\, y^2}$$

Así

$$\rho s = \frac{\Sigma\, x^2 + \Sigma\, y^2 - \Sigma\, d^2}{\sqrt{\Sigma\, x^2\ \Sigma\, y^2}} \qquad (14)$$

Con X y Y en rangos, se puede sustituir

$$\Sigma\, x^2 = \frac{N^3-N}{12} = \Sigma\, y^2$$

Y como $d = x-y = (x-\bar{x}) - (y-\bar{y}) = (x-y)$, y además $\bar{x} = \bar{y}$ en rangos, entonces

$$\rho s = \frac{\Sigma x^2 + \Sigma y^2 - \sum_{i=1}^{n} d_i^2}{2\sqrt{\Sigma x^2 \Sigma y^2}} \qquad (15)$$

Si los sujetos cuyos puntajes se usaron al calcular ρs fueron tomados al azar de una población, podemos usar sus puntajes para determinar si las dos variables están asociadas en la población para exáminar la hipótesis de nulidad, que supone que las dos variables en estudio no están asociadas en la población, y que la diferencia de cero del valor observado se debe solamente al azar. Para probar lo anterior se utilizará la estadística.

$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left[R(x_i) - R(y_i)\right]^2$ o se puede utilizar a ρs

Estableciendose las hipótesis

A) H_0: Las X_i y Y_i son independientes

H_a: Existe una tendencia en aparear a valores grandes de X con valores grandes de Y, o bien aparear valores chicos de X con valores chicos de Y.

Si se usa a ρs como estadística de prueba existen tablas (tabla de apéndice), donde aparecen los cuantiles de ρs bajo la hipótesis H_0 de donde si ρs es mayor que el

$1-\alpha/2$ cuantil o menor que el $\alpha/2$ cuantil, la hipótesis H_0 se rechaza; donde α representa el nivel de significancia.

Si se utilizará $\sum_{i=1}^{n} D_i^2$, hay tablas (tablas del apéndice) para los cuantiles, rechazándose H_0 si el valor de $\sum_{i=1}^{n} D_i^2$ es mayor que $1- \alpha/2$ cuantil o menor que el $\alpha/2$ cuantil correspondiente.

Cuando se quiera hacer pruebas de una cola, se observa que ρs crece cuando $\sum_{i=1}^{n} D_i^2$ disminuye y viceversa, por lo que hay que tener cuidado cuando se define la región de rechazo en esos casos.

Las hipótesis a plantear para estos casos son

B) H_0: Las variables X y Y son independientes

H_a: La asociación entre X y Y es positiva

C) H_0: Las variables X y Y son independientes

H_a: La asociación entre X y Y es negativa

A continuación presentamos un ejemplo para ilustrar la aplicación del coeficiente de correlación de Spearman.

En la siguiente tabla se presentan el número de habitantes por kilómetro cuadrado y la calidad de aire en 8 ciudades. La escala para la calidad del aire es ordinal: 1 indica baja calidad,...,4 alta calidad.

Ciudad	Habitantes (X_1)	Calidad (Y_1)	$R(x_i)$	$R(y_i)$	D_i	D_i^2
N.York	2782	2	8	3	5	25
Boston	869	2	7	3	4	16
Chicago	724	1	6	1	5	25
L. angeles	668	3	5	5.5	-0.5	0.25
Sn. Fco.	573	4	4	7.5	-3.5	12.25
Filadelfia	524	3	3	5.5	-2.5	6.25
Detroit	464	2	2	3	- 1	1
Whasington	399	4	1	7.5	-6.5	42.25
						128.0

Aún cuando existen empates haremos el cálculo para ρs como si no los hubiera, y luego haremos las correcciones pertinentes y compararemos los dos coeficientes obtenidos.

$$\rho s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6(128)}{8(64-1)} = 1 - \frac{768}{504}$$

$$\rho s = - .5238$$

Ahora calcularemos ρs con su respectiva corrección por observaciones ligadas.

$$\rho s = \frac{\Sigma x^2 + \Sigma y^2 - \sum_{i=1}^{n} D_i^2}{2\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$\Sigma x^2 = \frac{N^3-N}{12} - \Sigma Tx = \frac{8^3-8}{12} - 0 = 42$$

$$T = \frac{t^3-t}{12}$$

Como no hay observaciones ligadas en X_i, $T = 0$

$$\Sigma y^2 = \frac{N^3-N}{12} - \Sigma Ty = \frac{8^3-8}{12} - 3 = 39$$

dado que

$$\Sigma Ty = \frac{3^3-3}{12} + \frac{2^3-2}{12} + \frac{2^3-2}{12} = (2 + 0.5 + 0.5) = 3$$

Sustitímos los valores obtenidos

$$\rho s = \frac{42 + 39 - 128}{2\sqrt{(42)(39)}} = \frac{-47}{80.9444} = - .5806$$

Debido a la corrección por empates se obtuvo un valor de ρs más alto, por lo tanto tomamos ese último. En este caso el valor de ρs indica el deterioro de la calidad del aire cuando aumenta el número de habitantes por kilómetro cuadrado.

Ahora bien, con los datos con que se calculó $\rho s = -0 .5806$ la hipótesis adecuada es del tipo C. Con $\alpha = 0.05$ y $n = 8$, se tiene que el valor crítico es de - 0.6190 por lo que el valor observado no es significativo al cinco porciento. Para $\alpha = 0.10$, el valor crítico es de - . 5000 por lo que si es significativo al 10 porciento.

Es importante mencionar que no solo por que se obtenga un valor grande de ρs, es suficiente para concluir que existe una relación de causa-efecto entre las variables. La relación causal se establece solo cuando así lo demuestra la lógica de los fenomenos en estudio.

Resumimos la prueba a continuación:

a) Suposiciones: La escala es al menos ordinal, las muestras son dos del mismo tamaño

b) Hipótesis: Independencia

c) Estadística: $$\rho s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

d) Distribución: ρs = especial (tabla del apendice)

e) Decisión: S_i $\rho s < \alpha/2$ ó $\rho s > 1 - \alpha/2$ se rechaza H_0

El Coeficiente de Correlación de Rango de Kendall: τ

El coeficiente de correlación de Kendall es otra medida de asociación muy usada en los casos no paramétricos, debido a que también se basa en el orden de las observaciones, en lugar de su valor; requiriéndose por lo mismo, una escala ordinal para su utilización. Una diferencia importante entre este coeficiente y el coeficiente de correlación de Spearman, es que cuando el número de parejas crece, la función de distribución del

coeficiente de correlación de Kendall, se aproxima más rápido a una distribución Normal que la función de distribución del coeficiente de Spearman.

Para adentrarnos en la naturaleza del coeficiente de correlación de Kendall, lo haremos en forma práctica suponiendo lo siguiente:

Ordenamos cuatro objetos (a,b,c y d) de acuerdo a dos criterios (X e Y) pensando en una característica que deseamos de dichos objetos, obteniendo el siguiente arreglo:

	O b j e t o s			
Criterio	a	b	c	d
x	3	4	2	1
y	3	1	4	2

Ahora reordenamos los rangos del criterio X para que ocupen su orden natural:

	O b j e t o s			
Criterios	d	c	a	b
x	1	2	3	4
y	2	4	3	1

Con este nuevo orden podremos determinar la correspondencia que pueda existir entre los dos criterios. Es decir con los rangos del criterio en orden natural veremos cuales pares de rangos en el criterio Y se encuentran en el orden natural. De esta manera empezaremos con el rango dos en el criterio Y, formaremos parejas con este rango y cada uno de los rangos subsiguientes a su derecha. Así los pares formados son (2,4), (2,3), (2,1); observamos que solo los dos primeros pares están en el orden natural, asignándoles a cada uno un valor de + 1; El tercer par no está en orden natural por lo que le asignaremos un valor de -1. Luego para todos los pares que incluyen el rango 2, sumamos sus puntajes : (+1)+(1)+(-1) = +1. Este mismo proceso es repetido para los rangos cuatro y tres; sumamos los subtotales obtenidos en cada rango: (+1)+(-2)+(-1) = - 2. De esta forma nosotros hemos obtenido, la diferencia entre el número de pares de rangos que se encontraban en orden natural y el número de pares que no estaban en el orden natural. El número total de pares de rangos que se pueden formar independientemente del valor que se les pueda asignar es $\binom{4}{2} = 6$, ó en su forma general.

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

Como se puede observar, si los criterios X e Y fueran perfectamente concordantes la suma de valores

obtenidos sería 6 ó -6, si fueran perfectamente discordantes.

El coeficiente de Kendall sería obtenido como el cociente de la suma de los subtotales (s) y el total de pares posibles $\binom{n}{2}$:

$$\tau = \frac{S}{\binom{n}{2}} = \frac{S}{\frac{n(n-1)}{2}} = \frac{-2}{6} = -.33$$

De acuerdo con lo anterior se puede considerar a τ como una función del número mínimo de inversiones o intercambios que se requieren entre rangos vecinos para transformar una ordenación en otra. O dicho de otra manera τ es un tipo de coeficiente de desorden.

El cálculo de S puede acortarse considerablemente si tomamos el primer número de la izquierda y contamos el número de rangos a su derecha que son mayores, luego sustraemos de estos el número de rangos a su derecha que son menores. Si hacemos esto con todos los rangos y luego sumamos los resultados, obtenemos S.

Como en cualquier prueba que se basa en rangos, también en el coeficiente de Kendall, pueden ocurrir ligas entre estos, Por lo que puede existir la necesidad de hacer correcciones en el denominador $\frac{n(n-1)}{2}$, debido al efecto de

empates entre rangos. Para hacer esto, podemos reemplazar $\frac{n(n-1)}{2}$ por $\frac{1}{2}\sqrt{\Sigma a_{ij}^2\ \Sigma b_{ij}^2}$, donde a_{ij} es el puntaje del i-ésimo y j-ésimo miembros apareados en un ordenamiento y b_{ij} es el correspondiente puntaje en el otro.

Donde no existen ligas, algún término a_{ij}^2 es único, de modo que Σa_{ij}^2 se reduce al número de posibles términos, o sea $\frac{n(n-1)}{2}$; de igual manera sucede para Σb_{ij}^2.

Lo anterior es debido a que $a_{ij} = +1$ si $p_i > p_j$

$a_{ij} = -1$ si $p_i < p_j$

Donde p_i es el rango del i-ésimo miembro; igualmente para b.

Si hay una liga de t miembros consecutivos, todos los puntajes que provengan de algún par elegido de ellos, es cero. Hay $\frac{t(t-1)}{2}$ de tales pares. Consecuentemente la suma Σa_{ij}^2 es $\frac{n(n-1)}{2} - \frac{\Sigma t\ (t-1).}{2}$

Podemos escribir así para cada ordenamiento

$\Sigma Tx = \frac{1}{2}\ \Sigma t(t-1)$ t = número de observaciones

$\Sigma Ty = \frac{1}{2}\ \Sigma t(t-1)$ ligadas en un rango

Así cuando se presentan ligas la alternativa para calcular τ puede ser

$$\tau = \frac{S}{\sqrt{\left[\frac{1}{2}\ n(n-1) - \Sigma Tx\right]\left[\frac{1}{2}\ n(n-1) - \Sigma Ty\right]}}$$

Para probar la significancia de τ, bajo la hipótesis nula (Ho) de que las dos muestra son independientes si tomamos al azar una muestra de una población en donde **X** e **Y** no se relacionen, y ordenamos los miembros de esa muestra en **X** e **Y** , tenemos que para cualquer orden obtenido en **X**, todos los posibles ordenes de los rangos de **Y** son igualmente probables.

Si ordenamos los rangos de **X** en su orden natural, todos los ordenes posibles (**N!**) de **Y** para ese ordenamiento son igualmente probables conforme a H_0. Por lo que cada uno de esas posibles ordenes tienen probabilidad de ocurrencia conforme a $\mathbf{H_0}$ de $\frac{1}{N!}$. Y a cada uno de las N! ordenaciones posibles de **Y** está asociado un valor de τ; los cuales van de +1 a -1, como se muestra en la tabla siguiente para un N = 4.

Frecuencia de la ocurrencia conforme a H_0	Probabilidad de la ocurrencia	Valor de τ
1	$\frac{1}{24}$	- 1.0
3	$\frac{3}{24}$	- 0.67
5	$\frac{5}{24}$	- 0.33
6	$\frac{6}{24}$	0
5	$\frac{5}{24}$	0.33
3	$\frac{3}{24}$	0.67
1	$\frac{1}{24}$	1.0

Este método sería muy tedioso para valores grandes de **N**, pero para **N** ≥ **8**, la distribución se aproxima a la normal.

Para **N** ≤ **10** puede usarse la tabla especial del apéndice para determinar la probabilidad exacta asociada con la ocurrencia (de una cola) conforme a H_0 de un valor tan extremo como una **S** observada.

La distribución muestral de **S** y de τ son idénticas en cuanto a su probabilidad. Puesto que τ es una función de **S**, se puede usar cualquiera de las dos.

Cuando **N** es mayor que 10, T se aproxima a una distribución normal con $\mu_\tau = 0$ y $\sigma^2_\tau = \frac{2(2N+5)}{9N(N-1)}$

$$\text{Así} \qquad Z = \frac{\tau - \mu_\tau}{\sigma_\tau} = \frac{\tau}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}} = \frac{3\sqrt{N(N-1)}\ \tau}{\sqrt{2(2N+5)}}$$

donde $\sigma^2\tau$ es obtenida de la siguiente manera:
Tenemos el siguiente estimador insesgado de τ

$$T = \sum_{1\le i\le j\le n} \frac{A_{ij}}{\binom{n}{2}} = \sum_{1\le i\le j\le n} \frac{A_{ij}}{\frac{n(n-1)}{2}} = 2 \sum_{1\le i\le j\le n} \frac{A_{ij}}{n(n-1)} \qquad (1)$$

Donde A_{ij} es el indicador de variables para cada uno de los grupos de pares (x_i, y_i), (x_j, y_j) de la muestra de observaciones y es igual a

$$\text{sgn}\ (x_j - x_i)\ \text{sgn}\ (y_j - y_i); \tag{2}$$

además

$$\text{sgn}\ (u) = \begin{cases} -1 & \text{si } u < 0 \\ & \text{ó } u = 0 \\ 1 & \text{si } u > 0 \end{cases}$$

entonces los valores asumidos por A_{ij} son

$$a_{ij} = \begin{cases} 1 & \text{si los pares son concordantes (pc)} \\ -1 & \text{si los pares son discordantes (pd)} \\ 0 & \text{si los pares no son concordantes ni discordantes, existe una liga en algún componente} \end{cases}$$

La distribución de probabilidad marginal de ese indicador de variables es

$$f_{A_{ij}}\ (a_{ij}) = \begin{cases} Pc & \text{si } a_{ij} = 1 \\ Pd & \text{si } a_{ij} = -1 \\ 1-Pc-Pd & \text{si } a_{ij} = 0 \end{cases} \tag{3}$$

$$E\ (A_{ij}) = 1Pc + (-1)\ Pd = Pc - Pd$$

Así para determinar la varianza de T, las varianzas y covarianzas de los A_{ij} pueden ser evaluados puesto que T

es una combinación lineal de este indicador de variables.

Por lo que de (1)

$$\text{Var}\ (T) = \text{Var}\left[\ 2 \sum_{1\le i\le j\le n} \sum \frac{A_{ij}}{n(n-1)}\ \right]$$

$$n^2(n-1)^2\ \text{Var}\ (T) = 4\left[\ \sum_{1\le i\le j\le n} \sum \text{Var}(A_{ij}) + \sum_{\substack{1\le i\le j\le n \\ 1\le h\le k\le n \\ i\ne h\ \acute{o}\ j\ne k}} \sum \sum \sum \text{Cov}\ (A_{ij}, A_{hk})\right] \qquad (4)$$

Puesto que los A_{ij} son idénticamente distribuídos para todo $i < j$, y A_{ij} y A_{hk} son independientes para todo $i\ne h$ y $j\ne k$ (pares no comunes), (4) se puede escribir como.

$$n^2(n-1)^2\ \text{Var}\ (T) = 4\left[\left(\begin{array}{c} n \\ 2 \end{array}\right)\underset{i<j}{\text{Var}}(A_{ij}) + \right.$$

$$\sum_{\substack{i=1 \\ j\ne k}}^{n-1} \sum_{j=i+1}^{n} \sum_{k=i+1}^{n} \text{Cov}\ (A_{ij}, A_{ik}) + \sum_{j=2}^{n} \sum_{\substack{i=1 \\ i\ne k}}^{j-1} \sum_{k=1}^{j-1} \text{Cov}\ (A_{ij}, A_{kj})$$

$$\left. + \sum_{j=2}^{n} \sum_{\substack{i=1 \\ i\ne k}}^{j-1} \sum_{k=j+1}^{n} \text{Cov}\ (A_{ij}, A_{jk}) + \sum_{\substack{i=2 \\ j\ne k}}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{i-1} \text{Cov}\ (A_{ij}, A_{ki})\right] \qquad (5)$$

Por simetría, todos los términos de covarianzas en (5) son iguales. Ellos están agrupados de acuerdo a cual de los (X,Y) pares son comunes a los (A_{ij}, A_{hk}) para facilitar la cuenta del número de términos en cada grupo de sumatorias. Dentro de los primeros grupos tenemos dos

distintas permutaciones, (A_{ij}, A_{ik}) y (A_{ik}, A_{ij}), para cada una de las $\binom{n}{2}$ elecciones de $i \neq j \neq k$, y similarmente para el segundo grupo. Pero el tercer y cuarto grupo no permiten invertir los términos de los A_{ij} y A_{hk}, puesto que estos hacen diferentes (X,Y) pares en común, y así hay solamente $\binom{n}{3}$ términos de covarianza en cada una de estas sumatorias.

El número total de términos distinguibles de covarianzas es $(2 + 2 + 1 + 1)\ \binom{n}{3} = 6\ \binom{n}{3}$, y (5) puede ser escrito simplemente como

$$n^2(n-1)^2 \text{ Var } (T) = 4 \left[\binom{n}{2} \text{ Var } (A_{ij}) + 6 \binom{n}{3} \text{ Cov } (A_{ij}, A_{ik}) \right]$$

ó

$$n^2(n-1)^2 \text{ Var } (T) = 4 \left[\frac{n(n-1)}{2} \text{ Var } (A_{ij}) + 6 \frac{n(n-1)(n-2)}{6} \text{ Cov } (A_{ij}, A_{ik}) \right]$$

$$n(n-1) \text{ Var } (T) = 2 \text{ Var } (A_{ij}) + 4 (n-2) \text{ Cov } (A_{ij}, A_{ik}) \qquad (6)$$

para algún $i < j \quad i < k \quad j \neq k \qquad i = 1,2,\ldots,n-1$

$j = 2,3,\ldots,n$

$k = 2,3,\ldots,n$

Usando la distribución de probabilidad marginal de A_{ij} dada en (3), la varianza de A_{ij} es evaluada fácilmente como sigue:

$$E\ (A_{ij})^2 = 1Pc + (-1)^2 Pd = Pc + Pd$$

$$\text{Var } (A_{ij}) = (Pc + Pd) - (Pc - Pd)^2 \qquad (7)$$

Ahora, la expresión de covarianza, requiere conocer la distribución conjunta de A_{ij} y A_{ik}, la cual puede ser expresada como sigue:

$$fA_{ij},\ A_{ik}\ (a_{ij},\ a_{ik}) = \begin{cases} Pcc & \text{si } a_{ij} = a_{ik} = 1 \\ Pdd & \text{si } a_{ij} = a_{ik} = -1 \\ Pcd & \text{si } a_{ij} = 1,\ a_{ik} = -1 \text{ ó} \\ & \quad a_{ij} = -1,\ a_{ik} = 1 \\ 1-Pcc-Pdd-2Pcd & \text{si } a_{ij} = 0, a_{ik} = -1,0,1 \\ & \quad \text{ó } a_{ij} = -1,0,1,\ a_{ik} = 0 \\ o & \text{de otra manera} \end{cases}$$

Para todo $i < j$, $i < k$, $j \neq k$, $i = 1,2,,...,n$, y algún $0 \leq Pcc, Pdd, Pcd \leq 1$.

Así podemos evaluar

$$E\ (A_{ij},\ A_{ik}) = 1^2Pcc + (-1)^2Pdd + 2\ (-1)Pcd$$

y

$$Cov\ (A_{ij},\ A_{ik}) = Pcc + Pdd - 2Pcd - (Pc - Pd)^2 \qquad (8)$$

Sustituímos (7) y (8) en (6)

$$n(n-1)Var(T) = 2(Pc+Pd) + 4(n-2)(Pcc+Pdd-2Pcd) - 2(2n-3)(Pc-Pd)^2 \qquad (9)$$

Los resultados obtenidos hasta aquí son completamente generales, aplicado a todas las variables aleatorias. Si la distribuición marginal de X y Y son continuas, la $P(A_{ij}=0) = 0$ y las resultantes identidades

$Pc + Pd = 1$ y $Pcc + Pdd + 2Pcd = 1$

Permitámos simplificar (9) a una función de Pc y Pcd solamente:

$$n(n-1)\ Var(T) = 2-2(2n-3)(2Pc-1)^2 + 4(n-2)(1-4\ Pcd)$$
$$= 8(2n-3)Pc\ (1-Pc) - 16(n-2)Pcd \qquad (10)$$

Puesto que para X e y continuas tenemos

$$Pcd = P\ (A_{ij} = 1 \cap A_{ik} = -1)$$

$$= P\ (A_{ij} = 1) - P(A_{ij} = 1 \cap A_{ik} = 1)$$

$$= Pc - Pcc$$

Y otra expresión equivalente a (10) es

$$n(n-1)\ Var\ (T) = 8\ (2n-3)\ Pc(1-Pc) - 16(n-2)(Pc-Pcc)$$
$$= 8\ Pc(1-Pc) + 16(n-2)(Pcc-Pc^2) \qquad (11)$$

Ya se ha interpretado Pc como la probabilidad de que el par (X_i, Y_i) es concordante con (X_j, Y_j). Puesto que el parámetro Pcc es

$$Pcc = P(A_{ij} = 1 \cap A_{ik} = 1)$$

$$= P\left[(X_j - X_i)(Y_j - Y_i) > 0 \cap (X_k - X_i)(Y_k - Y_i) > 0\right] \tag{12}$$

Para todo $i<j, i<k, j\neq k, i=1,2\ldots,n$, interpretaramos Pcc como la probabilidad de que el par (X_i, Y_i), es concordante con (X_j, Y_j) y (X_k, Y_k). Las expresiones integrales pueden ser obtenidas como sigue, para las probabilidades Pc y Pcc para las variables aleatorias X e Y de alguna población continua bivariada Fx, y (x,y).

$$Pc = P\left[(x_i < x_j) \cap (y_i < y_j)\right] + P\left[(x_i > x_j) \cap (y_i > Y_j)\right]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P\left[(x_i < x_j) \cap (y_i < y_i)\right] fx_j\, y_j(x_j\, y_j)\, dx_j dy_j$$

$$+ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P\left[(x_j < x_i) \cap (y_j < y_i)\right] fx_i\, y_i(x_i\, y_i)\, dx_i dy_i$$

$$= 2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Fx,y\ (x,y) f(x,y)\, dxdy \tag{13}$$

$$Pcc = P\left(\left\{ \left[(x_i < x_j) \cap (y_i < y_j)\right] U \left[(x_i > x_j) \cap (y_i > y_j)\right] \right\} \right.$$

$$\left. \cap \left\{ \left[(x_i < x_k) \cap (y_i < y_k)\right] U \left[(x_i > x_k) \cap (y_i > y_k)\right] \right\} \right)$$

$$= P\left[(A \cup B) \cap (C \cup D)\right] \qquad \text{digamos}$$

$$= P\left[(A \cap C) \cup (B \cap D) \cup (A \cap B) \cup (B \cap C)\right]$$

$$= P\left[A \cap C)\right] + P\left[B \cap D\right] + 2\ p\left[A \cap D\right]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big\{ P\left[(x_j > x_i) \cap (y_j > y_i) \cap (x_k > x_i) \cap (y_k > y_i)\right]$$

$$+ P\left[(x_j > x_i) \cap (y_j < y_i) \cap (x_k < x_i) \cap (y_k < y_i)\right]$$

$$+ 2\ P\left[(x_j > x_i) \cap (y_i > y_j) \cap (x_k < x_i) \cap (y_k < y_i)\right] \Big\}$$

$$f x_i y_i (x_i y_i)\, dx_i dy_i$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big(\Big\{ P\left[(X > x) \cap (Y > y)\right] \Big\}^2 \Big\{ P\left[(X < x) \cap (Y < y)\right] \Big\}^2$$

$$+ 2\ P\left[(X > x) \cap (Y > y)\right] P\left[(X < x) \cap (Y < y)\right] \Big)$$

$$f x,y(x,y)\, dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big(P\left[(X > x) \cap (Y > y)\right] + P\left[(X < x) \cap (Y < y)\right] \Big\}^2$$

$$f x,y(x,y)\, dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[1 - Fx(x) - FY(y) + 2\ Fx,y(x,y)\right]^2 Fx,y(x,y)\, dxdy \qquad (14)$$

Ahora, por el teorema de la transformación integral de probabilidad visto en el capítulo (2), podemos asumir

que X y Y se distribuyen idénticamente de acuerdo a la distribución uniforme en el intérvalo (0,1).

Luego en (13) y (14) se tiene

$$Pc = 2 \int_0^1 \int_0^1 xydxdy = \frac{1}{2}$$

$$Pcc = \int_0^1 \int_0^1 (1-x-y + 2xy)^2 \, dx \, dy = 5/18 \qquad (15)$$

Sustituímos estos resultados en (11), y se obtiene

$$n(n-1) \text{ Var } (T) = 2 + \frac{16(n-2)}{36}$$

$$n(n-1) \text{ Var } (T) = 2 + \frac{16(n-2)}{36}$$

$$n(n-1) \text{ Var } (T) = \frac{72+16n-32}{36}$$

$$\text{Var } (T) = \frac{2(2n-5)}{9n(n-1)}$$

Así

$$\text{Var } (\hat{\tau}) = \frac{2(2n - 5)}{9n \ (n-1)}$$

Bajo H_0 $E(T) = 0$, veamos como es esto:

Si X y Y son variables aleatorias continuas e independientes, $P(X_i < X_j) = P \ (X_i > X_j)$, y además las probabilidades conjuntas en Pc o Pd son el producto de las

probabilidades individuales; con estas relaciones podemos escribir

$$Pc = P(X_i < X_j)p(Y_i < Y_j) + P(X_i > X_j)P(Y_i > Y_j)$$

$$= P(X_i > X_j)p(Y_i < Y_i) + P(X_i < X_j)P(Y_i > Y_j)$$

$$= Pd \qquad (17)$$

$$y \quad E(T) = 1\ Pc + (-1)Pd = Pc - Pd = 0$$

Ahora, bien la regla de decisión para la hipótesis

$$H_0: \tau = \tau_0 \text{ contra } H_a\ \tau \neq \tau_0$$

Con un nivel de significancia α

rechazamos H_0 cuando $\dfrac{\tau-\tau_o}{\sigma_\tau} \geq Z\ \alpha/2$ (18)

A continuación presentamos un ejemplo (el mismo que se utilizó para ρs) para ilustrar la aplicación del coeficiente de rangos de Kendall.

En la siguiente tabla aparecen ordenados en forma natural los rangos de la variable X con sus respectivas parejas de Y.

No. de habitantes (X)	1	2	3	4	5	6	7	8
calidad del aire (Y)	4	2	3	4	3	1	2	2

Calculamos el valor de S

$$S = (0-6) + (3-1) + (1-3) + (0-4) + (0-3) + (2-0) = -11$$

Como hay ligas de observaciones en algunos rangos de Y calculamos el factor de corrección correspondiente.

$$\Sigma Ty = 2(2-1) + 3(3-1) + 2(2-1) = 10$$

Ahora aplicamos la formula para calcular τ cuando ocurren observaciones ligadas.

$$\tau = \frac{-11}{\sqrt{\frac{1}{2}(8)(8-1)-0}\quad\sqrt{\frac{1}{2}(8)(8-1)-10}}$$

$$\tau = \frac{-11}{(5.29)(4.24)} = \frac{-11}{22.43} = -.4904$$

En la tabla correspondientes del ápendice podemos ver que no rechazamos H_0 a un nivel α de .05 ya que $S < \alpha/2$ ó $S > 1-\alpha/2$ no se cumple; pues $-16 < S < 16$.

Se ha usado la tabla especial para distribuciones exactas ya que $N < 10$. En caso de que $N > 10$ se hubiera utilizado el estadístico (18) con aproximación normal y

utilizar la tabla de la distribución normal.

Por último presentamos un resumen de la prueba .

a) Suposiciones: Dos muestras de tamaño n

b) Escala: Ordinal

c) Hipótesis Independencia

d) Distribución: Especial para $N < 10$

Normal para $N > 10$

e) Decisión: Si $S < \alpha/2$ ó $S > 1- \alpha/2$ se rechaza Ho.

$$\begin{cases} \text{para } N > 10 \\ \text{rechazamos Ho si,} \\ \dfrac{\tau - \tau_o}{\sigma\, \tau} \geq Z\ \alpha/2 \end{cases}$$

Por último, esperamos que las técnicas no paramétricas presentadas aquí, sean de utilidad a los estudiantes de estadística así como a los investigadores, lo cual fue el objetivo trazado al inicio de este trabajo; además de tratar de difundir este tipo de técnicas no tan populares, mas sin embargo muy importantes dentro de la investigación.

LITERATURA CITADA

Blomqvist, N. 1950 On a measure of dependence betwen two random variables. Ann. Math Statist. Vol. 21, pp 593-600.

_____1951. Some tests based on dichotomization. Ann. Math. Statist. Vol. 22, pp. 362-371.

Bradley, J.V. 1968. Distribution-Free statistical test. ed. Pretice-Hall. New Jersey. 388 p.

Brunk, H.D. 1965. An introduction to mathematical statistics 2 ed. Blaisdell Publishing. 429 p.

Cochran, W.G. 1950. The comparision of percentages in matched samples. Biometrica. Vol. 37, pp. 256-266.

_____1952. The X^2 test of goodnes of fit. Ann. Math. Statist. Vol. 23, pp. 315-345.

Conover, W.J. 1980. Practical nonparametric statistics. 2 ed Wiley. New York. 493 p.

Fieller, E.C. & E.S. Pearson. 1961. Tests for rank correlation coefficientes. Biometrica Vol. 48, pp. 29-40.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Jour. Amer. Stat. Assn. Vol. 32, pp. 675-701.

Gibbons, J.D. 1985. Nonparametric statistical

inference. 2 ed. Marcel Dekker. New York. 408 p.

Hájek, J. 1969. Nonparametric statistic. ed. Holden-day. San Francisco. 184 p.

Hogg, R.V. & A.T. Craig. 1978. Introduction to mathematical statistics. 5a. ed. Mc. Millan. New York. 438 p.

Hollander, M. & D.A. Wolfe. 1973. Nonparametric statistical methods. ed. Wiley. New York. 503 p.

Hotteling, H. & M.R. Pabst. 1939. Rank correlation and test of significance involving no assumption of normality. Ann. Math. Statist. Vol. 7, pp. 29-43.

Iman, R.L. & W.J. Conover. 1978. Aproximations of the critical region for Spearman's rho with and without ties present. Commun. Statist. - simula computa. Vol. B_7 (3) pp. 269-282.

Kendall, M.G. 1938. A new measure of rank correlation. Biometrica. Vol. 30, pp. 81-93.

Kendall. M.G. 1962. Rank correlation methods. 3 ed. Charles Griffin. London. 199 p.

Kendall. M.G. & B.B.Smith 1939. the problem of m rankings. Ann. Math. Statist. Vol. 10, pp. 275-287.

KrusKal, W.H. 1938. Ordinal measures of association. Jour. Amer. Stat. Assn. Vol. pp. 814-861.

_____1952. A nonparametric test for the several sample problem. Ann. Math. Statist. Vol. 23, pp.

525-540.

Kruskal, W.H. & W.A. Wallis 1952. Use of ranks on one criterion variance analysis. Jour. Amer. Statist. Assn. Vol. 47, pp. 583-621.

Milton, E.T. 1952. Some rank tests which are most power ful against specific parametric alterntives. Ann. Math. statist. Vol. 23, pp. 346-366

Mood, A.M. 1950. Introduction to the theory of statistics. ed. Mc. Graw Hill. New York. pp. 394-406.

Moran, P.A.P. 1950. Recent developments in ranking theory. Jour. Royal Statist. Soc. Ser. B, Vol. 12, pp. 153-162.

Neave, H.R. & P.L. Worthington. 1988. Ditribution-free test. ed. Unwin Hyman. London. 430 p.

Pearson, K. 1900. On the criterion that a given system of a correlated system of variables is such that it can be reasonably supposed to haven arisen from random sampling. Philos. Mag. Series 5, Vol. 50, pp. 157-172.

Rohatgi, V.K. 1976. An introduction to probability theory and mathematical statistics. ed. Wiley. 684.

Rueda, R. 1980. Estadística no paramétrica, un enfoque intuitivo. UNAM. Comunicación Interna No. 3, 231 p.

Said, I. 1980. Metodos estadísticos no paramétricos. Centro de Estadística y Cálculo, Colegio de Postgraduados. Chapingo. 213 p.

Siegel, S. 1972. Estadística no paramétrica. 2 ed. Trillas. México. 344 p.

Spearman, C. 1904. The proof and measurement of association between two things. Jour. Amer. Psych. Vol. 15, pp. 72-101.

Spearman, C. 1906. Foot rule for measuring correlation. British Journal of Psychology. Vol. 2, pp. 89-108.

Tate, M.W. & S. M. Brown. 1970. Note on the Cochran Q test. Jour. Amer. Stat. Assn. Vol. 65, pp. 155-160.

Tucker, H. 1970. Apuntes sobre estadística no paramétrica. Instituto Mexicano de Estudios Sociales A.C. 83 p.

Wayne, W.D. 1978. Applied nonparametric statistics. ed Houghton Mifflin. Georgia. 503 p.

Wolfowitz, W. & J. Wolfowitz 1944. Statistical tests based on permutations of the observations. Ann. Math. Statist. Vol. 15, pp. 358-372.

APENDICE

TABLA A.1 Tabla de probabilidades asociadas con valores tan extremos como los valores observados de z en la distribución normal

El cuerpo de la tabla contiene las probabilidades de una cola conforme a H_a de z. La columna marginal izquierda tiene diferentes valores de z hasta de una cifra decimal. La hilera superior tiene diferentes valores hasta la segunda cifra decimal. Así por ejemplo, la p de una cola de $z \geq 0.11$ o $z \leq -0.11$ es $p = 0.4562$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007									
3.3	.0005									
3.4	.0003									
3.5	.00023									
3.6	.00016									
3.7	.00011									
3.8	.00007									
3.9	.00005									
4.0	.00003									

TABLA A-2 Tabla de valores críticos de chi cuadrada*

gl	Probabilidad conforme a H_0 de que $x^2 \geq$ chi cuadrada													
	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.00016	.00063	.0039	.016	.064	.15	.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	.02	.04	.10	.21	.45	.71	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	.12	.18	.35	.58	1.00	1.42	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	.30	.43	.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	.55	.75	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	.87	1.13	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	4.76	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	6.61	7.96	9.31	11.15	12.62	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.29
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.75
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	9.54	10.60	12.34	14.04	16.31	18.10	21.24	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	11.29	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

* La tabla C es la tabla IV de Fisher y Yates abreviada: *Tablas estadísticas para biología, agricultura e investigación médica,* publicadas por Oliver y Boyd Ltd., Edinburgo, con permiso de los autores y editores.

TABLA A-3 **Tabla de probabilidades asociadas con valores tan grandes como los valores observados de χ_r^2 en el análisis de varianza de dos clasificaciones por rangos de Friedman***

Tabla N_r. $k = 3$

$N = 2$		$N = 3$		$N = 4$		$N = 5$	
χ_r^2	p	χ_r^2	p	χ_r^2	p	χ_r^2	p
0	1.000	.000	1.000	.0	1.000	.0	1.000
1	.833	.667	.944	.5	.931	.4	.954
3	.500	2.000	.528	1.5	.653	1.2	.691
4	.167	2.667	.361	2.0	.431	1.6	.522
		4.667	.194	3.5	.273	2.8	.367
		6.000	.028	4.5	.125	3.6	.182
				6.0	.069	4.8	.124
				6.5	.042	5.2	.093
				8.0	.0046	6.4	.039
						7.6	.024
						8.4	.0085
						10.0	.00077

$N = 6$		$N = 7$		$N = 8$		$N = 9$	
χ_r^2	p	χ_r^2	p	χ_r^2	p	χ_r^2	p
.00	000	.000	1.000	.00	1.000	.000	1.000
.33	956	.286	.964	.25	.967	.222	.971
1.00	740	.857	.768	.75	.794	.667	.814
1.33	570	1.143	.620	1.00	.654	.889	.865
2.33	430	2.000	.486	1.75	.531	1.556	.569
3.00	252	2.571	.305	2.25	.355	2.000	.398
4.00	184	3.429	.237	3.00	.285	2.667	.328
4.33	142	3.714	.192	3.25	.236	2.889	.278
5.33	072	4.571	.112	4.00	.149	3.556	.187
6.33	052	5.429	.085	4.75	.120	4.222	.154
7.00	029	6.000	.052	5.25	.079	4.667	.107
8.33	012	7.143	.027	6.25	.047	5.556	.069
9.00	0081	7.714	.021	6.75	.038	6.000	.057
9.33	0055	8.000	.016	7.00	.030	6.222	.048
10.33	0017	8.857	.0084	7.75	.018	6.889	.031
12.00	00013	10.286	.0036	9.00	.0099	8.000	.019
		10.571	.0027	9.25	.0080	8.222	.016
		11.143	.0012	9.75	.0048	8.667	.010
		12.286	.00032	10.75	.0024	9.556	.0060
		14.000	.000021	12.00	.0011	10.667	.0035
				12.25	.00086	10.889	.0029
				13.00	.00026	11.556	.0013
				14.25	.000061	12.667	.00066
				16.00	.0000036	13.556	.00035
						14.000	.00020
						14.222	.000097
						14.889	.000054
						16.222	.000011
						18.000	.0000006

* Tomada de Friedman, M. 1937. El uso de rangos para evitar la suposición implícita de normalidad en el análisis de varianza. *J. Amer. Statist. Ass.*, **32**, 688-689, con el amable permiso del autor y el editor.

TABLA A-3 Tabla de probabilidades asociadas con valores tan grandes como los valores observados de χ_r^2 en el análisis de varianza de dos clasificaciones por rangos de Friedman*

(Continuación)

Tabla N_{II}. $k = 4$

$N = 2$		$N = 3$		$N = 4$			
χ_r^2	p	χ_r^2	p	χ_r^2	p	χ_r^2	p
.0	1.000	.2	1.000	.0	1.000	5.7	.141
.6	.958	.6	.958	.3	.992	6.0	.105
1.2	.834	1.0	.910	.6	.928	6.3	.094
1.8	.792	1.8	.727	.9	.900	6.6	.077
2.4	.625	2.2	.608	1.2	.800	6.9	.068
3.0	.542	2.6	.524	1.5	.754	7.2	.054
3.6	.458	3.4	.446	1.8	.677	7.5	.052
4.2	.375	3.8	.342	2.1	.649	7.8	.036
4.8	.208	4.2	.300	2.4	.524	8.1	.033
5.4	.167	5.0	.207	2.7	.508	8.4	.019
6.0	.042	5.4	.175	3.0	.432	8.7	.014
		5.8	.148	3.3	.389	9.3	.012
		6.6	.075	3.6	.355	9.6	.0069
		7.0	.054	3.9	.324	9.9	.0062
		7.4	.033	4.5	.242	10.2	.0027
		8.2	.017	4.8	.200	10.8	.0016
		9.0	.0017	5.1	.190	11.1	.00094
				5.4	.158	12.0	.000072

* Tomada de Friedman, M. 1937. El uso de rangos para evitar la suposición implícita de normalidad en el análisis de varianza. *J. Amer. Statist. Ass.*, **32** 688 689, con el amable permiso del autor y el editor.

TABLA A-4 Tabla de probabilidades asociadas con valores tan grandes como valores observados de *H* en el análisis de varianza de una clasificación por rangos de Kruskal-Wallis*

Tamaño de muestras			*H*	*p*	Tamaño de muestras			*H*	*p*
n_1	n_2	n_3			n_1	n_2	n_3		
2	1	1	2.7000	.500	4	3	2	6.4444	.008
								6.3000	.011
2	2	1	3.6000	.200				5.4444	.046
2	2	2	4.5714	.067				5.4000	.051
			3.7143	.200				4.5111	.098
								4.4444	.102
3	1	1	3.2000	.300					
					4	3	3	6.7455	.010
3	2	1	4.2857	.100				6.7091	.013
			3.8571	.133				5.7909	.046
								5.7273	.050
3	2	2	5.3572	.029				4.7091	.092
			4.7143	.048				4.7000	.101
			4.5000	.067					
			4.4643	.105	4	4	1	6.6667	.010
								6.1667	.022
3	3	1	5.1429	.043				4.9667	.048
			4.5714	.100				4.8667	.054
			4.0000	.129				4.1667	.082
								4.0667	.102
3	3	2	6.2500	.011					
			5.3611	.032	4	4	2	7.0364	.006
			5.1389	.061				6.8727	.011
			4.5556	.100				5.4545	.046
			4.2500	.121				5.2364	.052
								4.5545	.098
3	3	3	7.2000	.004				4.4455	.103
			6.4889	.011					
			5.6889	.029					
			5.6000	.050	4	4	3	7.1439	.010
			5.0667	.086				7.1364	.011
			4.6222	.100				5.5985	.049
								5.5758	.051
4	1	1	3.5714	.200				4.5455	.099
								4.4773	.102
4	2	1	4.8214	.057					
			4.5000	.076	4	4	4	7.6538	.008
			4.0179	.114				7.5385	.011
								5.6923	.049
4	2	2	6.0000	.014				5.6538	.054
			5.3333	.033				4.6539	.097
			5.1250	.052				4.5001	.104
			4.4583	.100					
			4.1667	.105	5	1	1	3.8571	.143
4	3	1	5.8333	.021	5	2	1	5.2500	.036
			5.2083	.050				5.0000	.048
			5.0000	.057				4.4500	.071
			4.0556	.093				4.2000	.095
			3.8889	.129				4.0500	.119

TABLA A-4 Tabla de probabilidades asociadas con valores tan grandes como los valores observados de *H* en el análisis de varianza de una clasificación por rangos de Kruskal-Wallis*
(Continuación)

Tamaño de muestras			*H*	*p*	Tamaño de muestras			*H*	*p*
n_1	n_2	n_3			n_1	n_2	n_3		
5	2	2	6.5333	.008				5.6308	.050
			6.1333	.013				4.5487	.099
			5.1600	.034				4.5231	.103
			5.0400	.056	5	4	4	7.7604	.009
			4.3733	.090				7.7440	.011
			4.2933	.122				5.6571	.049
5	3	1	6.4000	.012				5.6176	.050
			4.9600	.048				4.6187	.100
			4.8711	.052				4.5527	.102
			4.0178	.095	5	5	1	7.3091	.009
			3.8400	.123				6.8364	.011
5	3	2	6.9091	.009				5.1273	.046
			6.8218	.010				4.9091	.053
			5.2509	.049				4.1091	.086
			5.1055	.052				4.0364	.105
			4.6509	.091	5	5	2	7.3385	.010
			4.4945	.101				7.2692	.010
5	3	3	7.0788	.009				5.3385	.047
			6.9818	.011				5.2462	.051
			5.6485	.049				4.6231	.097
			5.5152	.051				4.5077	.100
			4.5333	.097	5	5	3	7.5780	.010
			4.4121	.109				7.5429	.010
5	4	1	6.9545	.008				5.7055	.046
			6.8400	.011				5.6264	.051
			4.9855	.044				4.5451	.100
			4.8600	.056				4.5363	.102
			3.9873	.098	5	5	4	7.8229	.010
			3.9600	.102				7.7914	.010
5	4	2	7.2045	.009				5.6657	.049
			7.1182	.010				5.6429	.050
			5.2727	.049				4.5229	.099
			5.2682	.050				4.5200	.101
			4.5409	.098	5	5	5	8.0000	.009
			4.5182	.101				7.9800	.010
5	4	3	7.4449	.010				5.7800	.049
			7.3949	.011				5.6600	.051
			5.6564	.049				4.5600	.100
								4.5000	.102

* Versión abreviada de Kruskal, W. H., y Wallis, W. A. 1952. Uso de los rangos en el análisis de varianza de un criterio. *J. Amer. Statist. Ass.*, **47**, 614-617. Con el amable permiso de los autores y editores. (Las correcciones de esta tabla dadas por los autores en Errata, *J. Amer. Statist. Ass.*, **48**, 910, se han incorporado.)

Tabla A–5 Cuantiles de la Estadística de Spearman.

n	$p = .900$	.950	.975	.990	.995	.999
4	.8000	.8000				
5	.7000	.8000	.9000	.9000		
6	.6000	.7714	.8286	.8857	.9429	
7	.5357	.6786	.7450	.8571	.8929	.9643
8	.5000	.6190	.7143	.8095	.8571	.9286
9	.4667	.5833	.6833	.7667	.8167	.9000
10	.4424	.5515	.6364	.7333	.7818	.8667
11	.4182	.5273	.6091	.7000	.7455	.8364
12	.3986	.4965	.5804	.6713	.7273	.8182
13	.3791	.4780	.5549	.6429	.6978	.7912
14	.3626	.4593	.5341	.6220	.6747	.7670
15	.3500	.4429	.5179	.6000	.6536	.7464
16	.3382	.4265	.5000	.5824	.6324	.7265
17	.3260	.4118	.4853	.5637	.6152	.7083
18	.3148	.3994	.4716	.5480	.5975	.6904
19	.3070	.3895	.4579	.5333	.5825	.6737
20	.2977	.3789	.4451	.5203	.5684	.6586
21	.2909	.3688	.4351	.5078	.5545	.6455
22	.2829	.3597	.4241	.4963	.5426	.6318
23	.2767	.3518	.4150	.4852	.5306	.6186
24	.2704	.3435	.4061	.4748	.5200	.6070
25	.2646	.3362	.3977	.4654	.5100	.5962
26	.2588	.3299	.3894	.4564	.5002	.5856
27	.2540	.3236	.3822	.4481	.4915	.5757
28	.2490	.3175	.3749	.4401	.4828	.5660
29	.2443	.3113	.3685	.4320	.4744	.5567
30	.2400	.3059	.3620	.4251	.4665	.5479

Para $n > 30$ se usa la aproximación:

$$w_p \simeq \frac{x_p}{\sqrt{n-1}}$$

donde X_p es el p-ésimo cuantil de una normal.

Los cuantiles inferiores se obtienen mediante la relación:

$$W_p^* = W_{1-p}$$

Table A-6 Cuantiles de la Estadística Hotelling-Pabst.

n	$p = .001$	.005	.010	.025	.050	.100	$\frac{1}{3}n(n^2-1)$
4					2	2	20
5			2	2	4	6	40
6		2	4	6	8	14	70
7	2	6	8	14	18	26	112
8	6	12	16	24	32	42	168
9	12	22	28	38	50	64	240
10	22	36	44	60	74	92	330
11	36	56	66	86	104	128	440
12	52	78	94	120	144	172	572
13	76	110	130	162	190	226	728
14	106	148	172	212	246	290	910
15	142	194	224	270	312	364	1120
16	186	250	284	340	390	450	1360
17	238	314	356	420	480	550	1632
18	300	390	438	512	582	664	1938
19	372	476	532	618	696	790	2280
20	454	574	638	738	826	934	2660
21	546	686	758	870	972	1092	3080
22	652	810	892	1020	1134	1270	3542
23	772	950	1042	1184	1312	1464	4048
24	904	1104	1208	1366	1510	1678	4600
25	1050	1274	1390	1566	1726	1912	5200
26	1212	1462	1590	1786	1960	2168	5850
27	1390	1666	1808	2024	2216	2444	6552
28	1586	1890	2046	2284	2494	2744	7308
29	1800	2134	2306	2564	2796	3068	8120
30	2032	2398	2584	2868	3120	3416	8990

Para $n > 30$ los cuantiles pueden ser aproximados por:

$$w_p \cong \tfrac{1}{3}n(n^2-1) + x_p \cdot \frac{1}{6}\sqrt{\frac{n(n^2-1)}{n-1}}$$

donde X_p es el p-ésimo cuantil de una normal.

Cuantiles superiores pueden ser obtenidos usando:

$$W_{1-p} = n(n^2-1)/3 - W_p$$

Tabla A-7 Cuantiles de la Estadística de Kendall.

n	p = .900	.950	.975	.990	.995
4	4	4	6	6	6
5	6	6	8	8	10
6	7	9	11	11	13
7	9	11	13	15	17
8	10	14	16	18	20
9	12	16	18	22	24
10	15	19	21	25	27
11	17	21	25	29	31
12	18	24	28	34	36
13	22	26	32	38	42
14	23	31	35	41	45
15	27	33	39	47	51
16	28	36	44	50	56
17	32	40	48	56	62
18	35	43	51	61	67
19	37	47	55	65	73
20	40	50	60	70	78
21	42	54	64	76	84
22	45	59	69	81	89
23	49	63	73	87	97
24	52	66	78	92	102
25	56	70	84	98	108
26	59	75	89	105	115
27	61	79	93	111	123
28	66	84	98	116	128
29	68	88	104	124	136
30	73	93	109	129	143

Cuantiles inferiores se obtienen usando:

$$W_p = -W_{1-p}.$$

Tabla A-7 (Continuación).

n	$p = .900$	.950	.975	.990	.995
31	75	97	115	135	149
32	80	102	120	142	158
33	84	106	126	150	164
34	87	111	131	155	173
35	91	115	137	163	179
36	94	120	144	170	188
37	98	126	150	176	196
38	103	131	155	183	203
39	107	137	161	191	211
40	110	142	168	198	220

Para n > 40 los cuantiles pueden ser aproximados usando:

$$w_p \cong x_p \sqrt{\frac{n(n-1)(2n+5)}{18}}$$

donde X_p es el p-ésimo cuantil de una normal.