ESTADÍSTICA NO PARAMÉTRICA

UNIDAD N°4

PRUEBAS NO PARAMÉTRICAS

La mayor parte de los procedimientos de prueba de hipótesis se basa en la suposición de que las muestras aleatorias se seleccionan de poblaciones normales.

La mayor parte de estas pruebas aún son confiables cuando experimentamos ligeras desviaciones de la normalidad, en particular cuando el tamaño de la muestra es grande.

Estos procedimientos se denominan ***métodos paramétricos***

En los procedimientos ***no paramétricos*** o ***métodos de distribución libre***, que a menudo no suponen conocimiento de ninguna clase acerca de las distribuciones de las poblaciones fundamentales, excepto que éstas son continuas.

Estos procedimientos usan más bien una escala ordinal, por lo que se asigna rangos a los datos.

DESVENTAJAS: Es menos eficiente que la prueba paramétrica cuando es posible aplicar cualquier prueba sobre la muestra, por lo que para pruebas no paramétricas deben utilizarse muestras más grandes.

VENTAJAS: Las divergencias sobre la normalidad, hacen al método no paramétrico más eficiente.

PRUEBA DE SIGNO

LA PRUEBA DE SIGNO SE UTILIZA PARA PROBAR LA HIPÓTESIS SOBRE UNA *MEDIANA* POBLACIONAL.

LA MEDIA SE REEMPLAZA POR LA MEDIANA COMO EL PARÁMETRO DE UBICACIÓN PERTINENTE BAJO LA PRUEBA.

SE PUEDE APLICAR A DATOS DICOTÓMICOS QUE NO SE PUEDEN REGISTRAR EN UNA ESCALA NUMÉRICA PERO QUE SE PUEDEN REPRESENTAR MEDIANTE RESPUESTAS POSITIVAS Y NEGATIVAS.

PRUEBA DE SIGNO

El parámetro poblacional se denota con $\tilde{\mu}$ (mediana)

Dada una variable aleatoria X, $\tilde{\mu}$ se define de modo que

$$P(X > \tilde{\mu}) = P(X < \tilde{\mu}) = 0{,}5$$

Al probar la hipótesis nula H_0 de que $\tilde{\mu} = \tilde{\mu}_0$ contra una alternativa apropiada, sobre la base de una muestra aleatoria de tamaño *n*, reemplazamos cada valor de la muestra que exceda a $\tilde{\mu}_0$ con un signo *más* y cada valor de la muestra menor que $\tilde{\mu}_0$ con un signo *menos*

Si la hipótesis nula es verdadera y la población es simétrica, la suma de los signos *más* debe ser aproximadamente igual a la suma de los signos *menos*. Cuando un signo aparece con mayor frecuencia de lo que debería, con base sólo en el azar, rechazamos la hipótesis de que la mediana poblacional $\tilde{\mu}$ es igual a $\tilde{\mu}_0$

La prueba de signo se aplica sólo en situaciones donde $\tilde{\mu}_0$ no puede ser igual al valor de cualquiera de las observaciones.

Cuando se observan valores muestrales iguales a $\tilde{\mu}_0$, se excluyen del análisis y el tamaño de la muestra se reduce en consecuencia

PRUEBA DE SIGNO

La estadística de prueba apropiada para la prueba de signo es la variable aleatoria binomial X, que representa el número de signos ***más*** en una muestra aleatoria.

Si la hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$ es verdadera, la probabilidad de que un valor muestral tenga como resultado un signo *más* o uno *menos* es igual a $\frac{1}{2}$

Para probar la hipótesis nula de que $\tilde{\mu} = \tilde{\mu}_0$, en realidad probamos la hipótesis nula de que el número de signos *más* es un valor de una variable aleatoria que tiene la distribución binomial con el parámetro $p = \frac{1}{2}$
Los valores P para las alternativas unilateral y bilateral se pueden calcular entonces con el uso de esta distribución binomial

PRUEBA DE SIGNO

Para probar la hipótesis:

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

$$H_1: \tilde{\mu} < \tilde{\mu}_0$$

- Rechazaremos H_0 a favor de H_1 sólo si la proporción de los signos más es bastante menor que $\frac{1}{2}$; es decir, cuando el valor *x* de nuestra variable aleatoria es pequeño. De aquí, si el valor *P* se calcula

$$P = P\left(X \leq x \; cuando \; p = \frac{1}{2}\right)$$

- P es menor o igual a algún nivel de significancia α preestablecido, rechazamos H_0 a favor de H_1

Para probar la hipótesis:

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

$$H_1: \tilde{\mu} > \tilde{\mu}_0$$

- Rechazaremos H_0 a favor de H_1 sólo si la proporción de los signos más es bastante mayor que $\frac{1}{2}$; es decir, cuando *x* es grande. De aquí, si el valor *P* se calcula

$$P = P\left(X \geq x \; cuando \; p = \frac{1}{2}\right)$$

- P es mayor que α, rechazamos H_0 a favor de H_1

Para probar la hipótesis:

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

$$H_1: \tilde{\mu} \neq \tilde{\mu}_0$$

- Rechazaremos H_0 a favor de H_1 cuando la proporción de signos *más* es significativamente menor o mayor que $\frac{1}{2}$; es decir, cuando *x* sea bastante pequeña o bastante grande. Por tanto, si $x < \frac{1}{2}$ y el valor *P* calculado

$$P = 2P\left(X \leq x \; cuando \; p = \frac{1}{2}\right)$$

- P es menor que o igual a α, rechazamos H_0 a favor de H_1

PRUEBA DE SIGNO

Siempre que $n > 10$, las probabilidades binomiales con $p = \frac{1}{2}$ se pueden aproximar a partir de la curva normal, pues $np = nq > 5$

Como en la distribución binomial la variable X es discreta, en tanto que en la distribución normal es continua, se hace una *corrección por continuidad.*

Esto es equivalente a restarle 0,5 al valor de X, si $X > Np$ y sumarle 0,5 al valor de X, si $X < Np$

Supongamos que queremos probar la hipótesis:

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

$$H_1: \tilde{\mu} < \tilde{\mu}_0$$

en el nivel de significancia $\alpha = 0{,}05$ para una muestra aleatoria de tamaño $n = 20$ que produce $x = 6$ signos más. Con el uso de la aproximación de la curva normal con

$$\tilde{\mu} = np = 20 \cdot 0{,}5 = 10$$

$$\sigma = \sqrt{npq} = \sqrt{20 \cdot 0{,}5 \cdot 0{,}5} = 2{,}236$$

$$z = \frac{6{,}5 - 10}{2{,}236} = -1{,}57$$

$$P = P(X \leq 6) \cong P(Z < -1{,}57) = 0.0582$$

Que conduce a no rechazar la hipótesis nula

EJEMPLO 1

- LOS SIGUIENTES DATOS REPRESENTAN EL NÚMERO DE HORAS QUE UN COMPENSADOR OPERA ANTES DE REQUERIR RECARGA: 1,5 – 2,2 – 0,9 – 1,3 – 2,0 – 1,6 – 1,8 – 1,5 – 2,0 – 1,2 – 1,7
- UTILICE LA PRUEBA DE SIGNO PARA PROBAR LA HIPÓTESIS EN EL NIVEL DE SIGNIFICANCIA DE 0,05 QUE ESTE COMPENSADOR PARTICULAR OPERA CON UNA MEDIANA DE 1,8 HORAS ANTES DE REQUERIR UNA RECARGA

- ***SOLUCIÓN***

1. $H_0\colon \tilde{\mu} = 1{,}8$
2. $H_1\colon \tilde{\mu} \neq 1{,}8$
3. $\alpha = 0{,}05$
4. Estadística de prueba: variable binomial X con $p = \frac{1}{2}$
5. Cálculos: reemplazar cada valor con el símbolo "+" si excede 1,8, con el símbolo "-" si es menor que 1,8 y descartar las mediciones que sean iguales a 1,8

EJEMPLO 1

X	X - $\tilde{\mu} = d_i$	Signo
1,5	-0,3	-
2,2	0,4	+
0,9	-0,9	-
1,3	-0,5	-
2,0	0,2	+
1,6	-0,2	-
1,8	0	Se descarta
1,5	-0,3	-
2,0	0,2	+
1,2	-0,6	-
1,7	-0,1	-

- A PARTIR DE ESTA INFORMACIÓN, TENEMOS:

$$n = 10, \qquad x = 3 \qquad y \qquad \frac{n}{2} = 5$$

$$P = 2P\left(X \leq 3\ cuando\ p = \frac{1}{2}\right)$$

$$= 2 \cdot DISTR.BINOM\left(3; 10; \frac{1}{2}; VERDADERO\right)$$

$$= 2 \cdot 0{,}1719 = 0{,}3438 > 0{,}05$$

6. DECISIÓN: NO RECHAZAR LA HIPÓTESIS NULA Y CONCLUIR QUE EL TIEMPO MEDIANO DE OPERACIÓN NO ES SIGNIFICATIVAMENTE DIFERENTE A 1,8 HORAS.

PRUEBA DE SIGNO PARA MUESTRAS PAREADAS*

También se puede utilizar la prueba de signo para probar la hipótesis nula $\tilde{\mu}_1 - \tilde{\mu}_2 = d_0$ para observaciones pareadas

Aquí se reemplaza cada diferencia, d_i, con un signo más o menos dependiendo si la diferencia ajustada, $d_i - d_0$, es positiva o negativa

Suponemos que las poblaciones son simétricas. Sin embargo, aun si las poblaciones son asimétricas se puede llevar a cabo el mismo procedimiento de prueba, pero las hipótesis se refieren a las medianas poblacionales en lugar de las medias.

*Se llaman muestras pareadas todas aquellas que comparan un hecho anterior con un hecho posterior, utilizando p.ej. las mismas personas

EJEMPLO 2

- UNA COMPAÑÍA DE TAXIS TRATA DE DECIDIR SI EL USO DE LLANTAS RADIALES EN LUGAR DE LLANTAS REGULARES CON CINTURÓN MEJORA LA ECONOMÍA DE COMBUSTIBLE. SE EQUIPAN 16 AUTOMÓVILES CON LLANTAS RADIALES Y SE MANEJAN POR UN RECORRIDO DE PRUEBA ESTABLECIDO. SIN CAMBIAR DE CONDUCTORES, SE EQUIPAN LOS MISMOS AUTOS CON LLANTAS REGULARES CON CINTURÓN Y SE MANEJAN UNA VEZ MÁS POR EL RECORRIDO DE PRUEBA. SE REGISTRA EL CONSUMO DE GASOLINA, EN KILÓMETROS POR LITRO, DE LA SIGUIENTE MANERA:

Automóvil	Llantas radiales	Llantas con cinturón
1	4.2	4.1
2	4.7	4.9
3	6.6	6.2
4	7.0	6.9
5	6.7	6.8
6	4.5	4.4
7	5.7	5.7
8	6.0	5.8
9	7.4	6.9
10	4.9	4.9
11	6.1	6.0
12	5.2	4.9
13	5.7	5.3
14	6.9	6.5
15	6.8	7.1
16	4.9	4.8

¿Se puede concluir en el nivel de significancia de 0.05 que los autos equipados con llantas radiales obtienen mejores economías de combustible que los equipados con llantas regulares con cinturón?

EJEMPLO 2: SOLUCIÓN

1. H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 = 0$
2. H_1: $\tilde{\mu}_1 - \tilde{\mu}_2 > 0$
3. $\alpha = 0{,}05$
4. ESTADÍSTICA DE PRUEBA: VARIABLE BINOMIAL X CON $p = \frac{1}{2}$
5. CÁLCULOS: PLANTEAMOS LAS DIFERENCIAS, Y OBTENEMOS LA SECUENCIA (CUARTA COLUMNA DE LA TABLA)PARA LA QUE $n = 14$ Y $x = 11$. CON EL USO DE LA APROXIMACIÓN DE LA CURVA NORMAL ENCONTRAMOS QUE:

$$z = \frac{10{,}5 - 7}{\sqrt{14 \cdot 0{,}5 \cdot 0{,}5}} = 1{,}87$$

$$P = P(X \geq 11) \cong P(Z \geq 1{,}87) =$$
$$= 1 - DISTR.NORM.ESTAND(1{,}87) = 0{,}0307$$

6. DECISIÓN: RECHAZAR H_0 Y CONCLUIR QUELAS LLANTAS RADIALES MEJORAN LA ECONOMÍA DE COMBUSTIBLE

Automóvil	Llantas radiales	Llantas con cinturón	d
1	4.2	4.1	+
2	4.7	4.9	-
3	6.6	6.2	+
4	7.0	6.9	+
5	6.7	6.8	-
6	4.5	4.4	+
7	5.7	5.7	0
8	6.0	5.8	+
9	7.4	6.9	+
10	4.9	4.9	0
11	6.1	6.0	+
12	5.2	4.9	+
13	5.7	5.3	+
14	6.9	6.5	+
15	6.8	7.1	-
16	4.9	4.8	+

PRUEBA DE RANGO CON SIGNO DE WILCOXON

LA PRUEBA DE SIGNO UTILIZA SÓLO LOS SIGNOS MÁS Y MENOS DE LAS DIFERENCIAS ENTRE LAS OBSERVACIONES Y μ_0 EN EL CASO DE UNA MUESTRA, O LOS SIGNOS MÁS Y MENOS DE LAS DIFERENCIAS ENTRE LOS PARES DE OBSERVACIONES EN EL CASO DE LA MUESTRA PAREADA, PERO NO TOMA EN CONSIDERACIÓN LA MAGNITUD DE ESTAS DIFERENCIAS. UNA PRUEBA QUE UTILIZA DIRECCIÓN Y MAGNITUD, SE LLAMA **PRUEBA DE RANGO CON SIGNO DE WILCOXON**

PRUEBA DE RANGO CON SIGNO DE WILCOXON

Esta prueba se aplica en el caso de una distribución continua simétrica. Bajo esta condición se puede probar la hipótesis nula $\mu = \mu_0$

Primero se resta μ_0 de cada valor muestral y se descarta todas las diferencias iguales a cero. Se asigna un rango de 1 a la diferencia absoluta más pequeña, un rango de 2 a la siguiente más pequeña, y así sucesivamente

Cuando el valor absoluto de dos o más diferencias es el mismo, se asigna a cada uno el promedio de los rangos que se asignarían si las diferencias se distinguieran

Por ejemplo, si la quinta y sexta diferencia son iguales en valor absoluto, a cada una se le asignaría un rango de 5,5

Si la hipótesis $\mu = \mu_0$ es verdadera, el total de los rangos que corresponden a las diferencias positivas debe ser casi igual al total de los rangos que corresponden a las diferencias negativas

Se representan esos totales como w_+ y w_-, respectivamente. Se designa el menor de w_+ y w_- con w

PRUEBA DE RANGO CON SIGNO DE WILCOXON

Al seleccionar muestras repetidas esperaríamos que variarían w_+ y w_-, y por tanto w. De esta manera se puede considerar a w_+ y w_-, y w como valores de las correspondiente variables aleatorias W_+, W_-, y W

La hipótesis nula $\mu = \mu_0$ se puede rechazar a favor de la alternativa $\mu < \mu_0$ sólo si w_+ es pequeña y w_- es grande

La alternativa $\mu > \mu_0$ se puede aceptar sólo si w_+ es grande y w_- es pequeña

Para una alternativa bilateral se puede rechazar H_0 a favor de H_1 si w_+ o w_- y por tanto w son suficientemente pequeñas. No importa cuál hipótesis alternativa puede ser, rechazar la hipótesis nula cuando el valor de la estadística apropiada W_+, W_-, o W es suficientemente pequeño

DOS MUESTRAS CON OBSERVACIONES PAREADAS

Para probar la hipótesis nula de que se muestrean dos poblaciones simétricas continuas con $\mu_1 = \mu_2$ para el caso de una muestra pareada, se clasifican las diferencias de las observaciones paradas sin importar el signo y se procede como en el caso de una muestra. Los diversos procedimientos de prueba para los casos de una sola muestra y de una muestra pareada se resumen en la siguiente tabla:

Prueba de rango con signo		
Para probar H_0	Contra H_1	Calcular
$\mu = \mu_0$	$\begin{cases} \mu < \mu_0 \\ \mu > \mu_0 \\ \mu \neq \mu_0 \end{cases}$	w_+ w_- w
$\mu_1 = \mu_2$	$\begin{cases} \mu_1 < \mu_2 \\ \mu_1 > \mu_2 \\ \mu_1 \neq \mu_2 \end{cases}$	w_+ w_- w

DOS MUESTRAS CON OBSERVACIONES PAREADAS

Siempre que n < 5 y el nivel de significancia no exceda 0.05 para una prueba de una cola o 0.10 para una prueba de dos colas, todos los valores posibles de w_+ , w_- , o w conducirán a la aceptación de la hipótesis nula

Cuando $5 \leq n \leq 30$, la tabla A.16 muestra valores críticos aproximados de W_+ y W_- para niveles de significancia iguales a 0.01, 0.025 y 0.05 para una prueba de una cola, y valores críticos de W para niveles de significancia iguales a 0.02, 0.05 y 0.10 para una prueba de dos colas. La ***hipótesis nula se rechaza*** si el valor calculado w_+ , w_- , o w **es menor o igual que** el valor de tabla apropiado

Por ejemplo, cuando n = 12 la tabla A.16 muestra que se requiere un valor de $w_+ \leq 17$ para que la alternativa unilateral $\mu < \mu_0$ sea significativa en el nivel 0.05

La prueba de rango con signo también se puede utilizar para probar la hipótesis nula $\mu_1 - \mu_2 = d_0$. En este caso las poblaciones no necesitan ser simétricas. Como con la prueba de signo, restamos d_0 de cada diferencia, clasificamos las diferencias ajustadas sin importar el signo.

EJEMPLO 3

- Los siguientes datos representan el número de horas que un compensador opera antes de requerir una recarga: 1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2 y 1.7. Utilice la prueba de rango con signo para probar la hipótesis en el nivel de significancia de 0.05 que este compensador particular opera con una media de 1.8 horas antes de requerir una recarga

SOLUCIÓN

1. H_0: $\mu = 1{,}8$
2. H_1: $\mu \neq 1{,}8$
3. $\alpha = 0{,}05$
4. SE PROCEDERÁ A EFECTUAR LAS DIFERENCIAS Y A PONER RANGO CON SIGNO A LOS DATOS (VER TABLA)
5. REGIÓN CRÍTICA: COMO $n = 10$, DESPUÉS DE DESCARTAR LA MEDICIÓN QUE ES IGUAL A 1,8. LA TABLA A.16 MUESTRA QUE LA REGIÓN CRÍTICA ES $w \leq 8$
6. CÁLCULOS:
   - $W_+ = 7 + 3 + 3 = 13$
   - $W_- = 5.5 + 10 + 8 + 3 + 5.5 + 9 + 1 = 42$
   - POR LO QUE $W = 13$ (MENOR ENTRE W_+ Y W_-).
7. DECISIÓN Y CONCLUSIÓN: COMO 13 NO ES MENOR QUE 8, NO SE RECHAZA H_0 Y SE CONCLUYE QUE EL TIEMPO PROMEDIO DE OPERACIÓN NO ES SIGNIFICATIVAMENTE DIFERENTE DE 1.8 HORAS

Dato	d_i = dato - 1.8	Rangos
1.5	-0.3	**5.5**
2.2	0.4	**7**
0.9	-0.9	**10**
1.3	-0.5	**8**
2.0	0.2	**3**
1.6	-0.2	**3**
1.8	0	**Se anula**
1.5	-0.3	**5.5**
2.0	0.2	**3**
1.2	-0.6	**9**
1.7	-0.1	**1**

EJEMPLO 4

- SE AFIRMA QUE UN ESTUDIANTE UNIVERSITARIO DE ÚLTIMO AÑO PUEDE AUMENTAR SU CALIFICACIÓN EN EL ÁREA DEL CAMPO DE ESPECIALIDAD DEL EXAMEN DE REGISTRO DE GRADUADOS EN AL MENOS 50 PUNTOS SI DE ANTEMANO SE LE PROPORCIONAN PROBLEMAS DE MUESTRA. PARA PROBAR ESTA AFIRMACIÓN, SE DIVIDEN 20 ESTUDIANTES DEL ÚLTIMO AÑO EN 10 PARES DE MODO QUE CADA PAR TENGA CASI EL MISMO PROMEDIO DE PUNTOS DE CALIDAD GENERAL EN SUS PRIMEROS AÑOS EN LA UNIVERSIDAD. LOS PROBLEMAS Y RESPUESTAS DE MUESTRA SE PROPORCIONAN AL AZAR A UN MIEMBRO DE CADA PAR UNA SEMANA ANTES DEL EXAMEN. SE REGISTRAN LAS SIGUIENTES CALIFICACIONES DEL EXAMEN:

- PRUEBE LA HIPÓTESIS NULA EN EL NIVEL DE SIGNIFICANCIA DE 0.05 DE QUE LOS PROBLEMAS AUMENTAN LAS CALIFICACIONES EN 50 PUNTOS CONTRA LA HIPÓTESIS ALTERNATIVA DE QUE EL AUMENTO ES MENOR A 50 PUNTOS.

Par	Con problemas de muestra	Sin problemas de muestra
1	531	509
2	621	540
3	663	688
4	579	502
5	451	424
6	660	683
7	591	568
8	719	748
9	543	530
10	575	524

EJEMPLO 4: SOLUCIÓN

- REPRESENTAMOS CON μ_1 Y μ_2 LA CALIFICACIÓN MEDIA DE TODOS LOS ESTUDIANTES QUE RESUELVEN EL EXAMEN EN CUESTIÓN CON Y SIN PROBLEMAS DE MUESTRA, RESPECTIVAMENTE.

1. $H_0: \mu_1 - \mu_2 = 50$
2. $H_1: \mu_1 - \mu_2 < 50$
3. $\alpha = 0{,}05$
4. SE PROCEDERÁ A EFECTUAR LAS DIFERENCIAS CORRESPONDIENTES Y A PONER RANGO A LOS DATOS (VER TABLA)
5. <u>CÁLCULOS</u>: <u>REGIÓN CRÍTICA</u>: COMO $n = 10$. LA TABLA A.16 MUESTRA QUE LA REGIÓN CRÍTICA ES $w \leq 11$
   - $W_+ = 6 + 3{,}5 + 1 = 10{,}5$
   - $W_- = 5 + 9 + 2 + 8 + 3{,}5 + 10 + 7 = 44{,}5$
   - POR LO QUE $W = 10{,}5$ (MENOR ENTRE W_+ Y W_-).
6. <u>DECISIÓN Y CONCLUSIÓN</u>: RECHAZAR H_0 Y CONCLUIR QUE LOS PROBLEMAS DE MUESTRA, EN PROMEDIO, NO AUMENTAN LAS CALIFICACIONES DE REGISTRO DE GRADUADOS EN 50 PUNTOS.

Par	Con prob. de muestra	Sin prob. de muestra	d_i	$d_i - d_0$ ($d_i - 50$)	Rangos
1	531	509	22	-28	5
2	621	540	81	31	6
3	663	688	-25	-75	9
4	579	502	77	27	3.5
5	451	424	27	-23	2
6	660	683	-23	-73	8
7	591	568	23	-27	3.5
8	719	748	-29	-79	10
9	543	530	13	-37	7
10	575	524	51	1	1

APROXIMACIÓN NORMAL PARA MUESTRAS GRANDES

Cuando $n \geq 15$, la distribución muestral de W_+ ó W_- se aproxima a la distribución normal con los siguientes parámetros:

Cuando *n* excede el valor más grande de la tabla A.16, se puede utilizar la estadística:

Media:

$$\mu_{W_+} = \frac{n(n+1)}{4}$$

Varianza:

$$\sigma^2{}_{W_+} = \frac{n(n+1)(2n+1)}{24}$$

$$z = \frac{W_+ - \mu_W}{\sigma_{W_+}}$$

para determinar la región crítica para nuestra prueba

PRUEBA DE LA SUMA DE RANGOS DE WILCOXON

ESTA PRUEBA ES UNA ALTERNATIVA PARA EL USO DE LA DISTRIBUCIÓN *T* DE DOS MUESTRAS.

PRUEBA DE LA SUMA DE RANGOS DE WILCOXON

Se quiere probar la hipótesis nula H_0 de que $\mu_1 = \mu_2$ contra alguna alternativa adecuada.

n_1 es el número de observaciones en la muestra más pequeña y n_2 es el número de observaciones en la muestra más grande

Si n_1 y n_2 tienen el mismo tamaño, se asigna de manera aleatoria

Luego se ordenan las $n_1 + n_2$ observaciones de las muestras combinadas en orden ascendente y sustituir un rango de $1, 2, \dots, n_1 + n_2$ para cada observación

En caso de empate, se promedian los rangos que tendrían las observaciones

La suma de los rangos que corresponden a las n_1 observaciones de la muestra más pequeña se denota con w_1

El valor w_2 representa la suma de los n_2 rangos que corresponden a la muestra más grande

El total $w_1 + w_2$ depende del número de observaciones de las dos muestras

$$w_1 + w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

la suma aritmética de los enteros $1, 2, \dots, n_1 + n_2$

Al elegir muestras repetidas de tamaño n_1 y n_2, esperaríamos que variara w_1 y por lo tanto w_2

Podemos considerar w_1 y w_2 como valores de variables aleatorias W_1 y W_2 respectivamente

PRUEBA DE LA SUMA DE RANGOS DE WILCOXON

La hipótesis nula $\mu_1 = \mu_2$ se rechazará a favor de la alternativa $\mu_1 < \mu_2$ sólo si w_1 es pequeña y w_2 es grande

La alternativa $\mu_1 > \mu_2$ se puede aceptar sólo si w_1 es grande y w_2 es pequeña

Para una prueba de dos colas, podemos rechazar H_0 a favor de H_1 si w_1 es pequeña y w_2 es grande o si w_1 es grande y w_2 es pequeña

Basamos nuestra decisión en los siguientes valores:

$$u_1 = w_1 - \frac{n_1(n_1+1)}{2} \quad y \quad u_2 = w_2 - \frac{n_2(n_2+1)}{2}$$

de la estadística relacionada U_1 o U_2, o en el valor u de la estadística U, el mínimo de U_1 y U_2,

U_1 y U_2 tienen distribuciones muestrales simétricas y toman valores en el intervalo de 0 a $n_1 n_2$ tales que $u_1 + u_2 = n_1 n_2$

PRUEBA DE LA SUMA DE RANGOS DE WILCOXON

De las fórmulas para u_1 y u_2 vemos que u_1 será pequeña cuando w_1 es pequeña y u_2 será pequeña cuando w_2 es pequeña

La **hipótesis nula se rechazará** siempre que **las estadísticas** apropiadas U_1, U_2 y U **tomen un valor menor o igual que el valor crítico dado** en la tabla A.17

Prueba de la suma de rangos		
Para probar H_0	Contra H_1	Calcular
$\mu_1 = \mu_2$	$\begin{cases} \mu_1 < \mu_2 \\ \mu_1 > \mu_2 \\ \mu_1 \neq \mu_2 \end{cases}$	u_1 u_2 u

EJEMPLO 5

- SE ENCUENTRA QUE EL CONTENIDO DE NICOTINA DE DOS MARCAS DE CIGARRILLOS, MEDIDO EN MILIGRAMOS, ES EL SIGUIENTE:

Marca A	2,1	4,0	6,3	5,4	4,8	3,7	6,1	3,3		
Marca B	4,1	0,6	3,1	2,5	4,0	6,2	1,6	2,2	1,9	5,4

- PRUEBE LA HIPÓTESIS, EN EL NIVEL DE SIGNIFICANCIA DE 0,05, QUE EL CONTENIDO PROMEDIO DE NICOTINA DE LAS DOS MARCAS ES IGUAL CONTRA LA ALTERNATIVA DE QUE SON DIFERENTES.

EJEMPLO 5: SOLUCIÓN

1. $H_0: \mu_1 = \mu_2$
2. $H_1: \mu_1 \neq \mu_2$
3. $\alpha = 0,05$
4. REGIÓN CRÍTICA:

   $n_1 = 8$ Y $n_2 = 10$

   $u \leq 17$. (DE LA TABLA A.17, PRUEBA A DOS COLAS)
5. CÁLCULOS: LAS OBSERVACIONES SE ACOMODAN EN ORDEN ASCENDENTE Y SE LES ASIGNAN RANGOS DEL 1 AL 18

$$w_1 = 4 + 8 + 9 + 10,5 + 13 + 14,5 + 16 + 18 = 93$$

$$w_2 = \frac{(8+10)(8+10+1)}{2} - 93 = 78$$

$$u_1 = 93 - \frac{8(8+1)}{2} = 57 \qquad u_2 = 78 - \frac{10(10+1)}{2} = 23$$

6. DECISIÓN: NO RECHAZAR H_0 Y CONCLUIR QUE NO HAY DIFERENCIA SIGNIFICATIVA EN EL CONTENIDO PROMEDIO DE NICOTINA EN LAS DOS MARCAS DE CIGARRILLOS.

Datos Originales	Rangos
0,6	1
1,6	2
1,9	3
2,1	4
2,2	5
2,5	6
3,1	7
3,3	8
3,7	9
4,0	10,5
4,0	10,5
4,1	12
4,8	13
5,4	14,5
5,4	14,5
6,1	16
6,2	17
6,3	18

- Rangos muestra A
- Rangos muestra B

TEORÍA NORMAL DE APROXIMACIÓN PARA DOS MUESTRAS

Cuando n_1 y n_2 exceden 8, la distribución muestral de U_1 (o U_2) se aproxima a la distribución muestral

Parámetros: Media y Varianza

$$\mu_{U_1} = \frac{n_1 n_2}{2}$$

$$\sigma^2{}_{U_1} = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Cuando n_2 es mayor que 20 y n_1 es al menos 9, se puede usar la estadística:

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

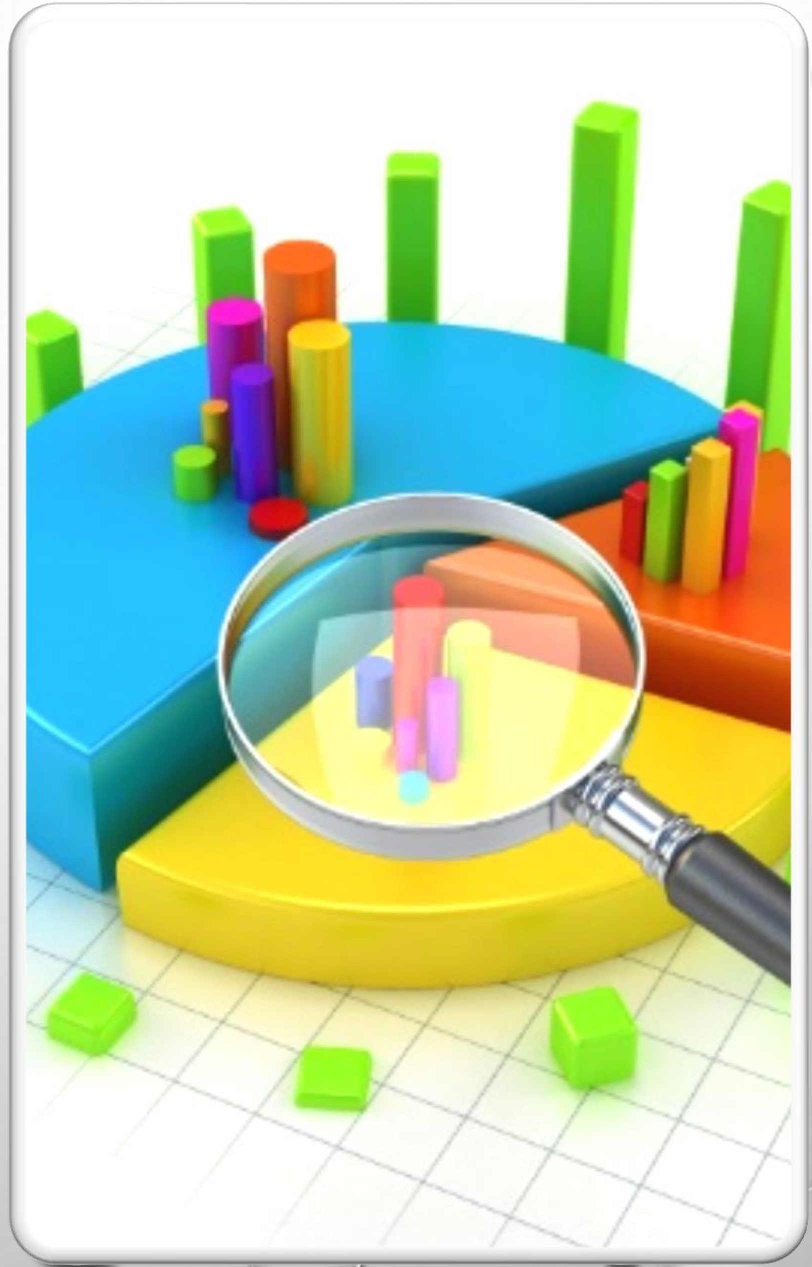
Para nuestra prueba, con la región crítica que cae ya sea en alguna o en ambas colas de la distribución normal, dependiendo de la distribución H_1

PRUEBA DE KRUSKAL WALLIS

LA PRUEBA DE KRUSHAL – WALLIS, ES UNA GENERALIZACIÓN DE LA PRUEBA DE LA SUMA DE RANGOS PARA EL CASO DE K > 2 MUESTRAS.

SE UTILIZA PARA PROBAR LA HIPÓTESIS NULA H_0 DE QUE *K* MUESTRAS INDEPENDIENTES SON DE POBLACIONES IDÉNTICAS.

LA PRUEBA ES UN PROCEDIMIENTO NO PARAMÉTRICO PARA PROBAR LA IGUALDAD DE LAS MEDIAS EN EL ANÁLISIS DE VARIANZA DE UN FACTOR CUANDO EL EXPERIMENTADOR DESEA EVITAR LA SUPOSICIÓN DE QUE LAS MUESTRAS SE SELECCIONARON DE POBLACIONES NORMALES.



PRUEBA DE KRUSKAL WALLIS

Sean n_i $(i = 1, 2, \ldots, k)$ el número de observaciones en la i-ésima muestra

Se combinan todas las k muestras y acomodamos las $n = n_1 + n_2 + \cdots + n_k$ observaciones en orden ascendente, y se sustituyen el rango apropiado de $1, 2, \ldots, n$ para cada observación

En caso de empates, se reemplazan las observaciones por las medias de los rangos

La suma de los rangos que corresponde a las n_i observaciones en la i-ésisma muestra se denota mediante la variable aleatoria R_i:

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

que se aproxima muy bien mediante una distribución ji cuadrada con $k - 1$ grados de libertad cuando H_0 es verdadera y si cada muestra consiste en al menos 5 observaciones

PRUEBA DE KRUSKAL WALLIS

La estadística H toma el valor h.
R_1 toma el valor r_1, R_2 toma el valor r_2, etc.
El hecho de que h sea grande cuando las muestras independientes provienen de poblaciones que no son idénticas nos permite establecer el siguiente criterio de decisión para probar H_0

Para probar la hipótesis H_0 de que k muestras independientes son de poblaciones idénticas, calcular:

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1)$$

Si h cae en la región crítica $H > \chi_\alpha^2$ con $v = k - 1$ grados de libertad, rechazar H_0 en el nivel de significancia α; en cualquier otro caso, aceptar H_0

EJEMPLO 6

- EN UN EXPERIMENTO PARA DETERMINAR CUÁL DE TRES DIFERENTES SISTEMAS DE MISILES ES PREFERIBLE, SE PIDE LA TASA DE UTILIZACIÓN DEL PROPULSOR. LOS DATOS, DESPUÉS DE CODIFICARLOS, SE DAN EN LA SIGUIENTE TABLA.

Tasas de utilización del propulsor			
Sistemas de misiles:	**1**	**2**	**3**
	24,0	23,2	18,4
	16,7	19,8	19,1
	22,8	18,1	17,3
	19,8	17,6	17,3
	18,9	20,2	19,7
		17,8	18,9
			18,8
			19,3

- UTILICE LA PRUEBA DE KRUSKAL – WALLIS Y UN NIVEL DE SIGNIFICANCIA $\alpha = 0{,}05$ PARA PROBAR LA HIPÓTESIS DE QUE LAS TASAS DE UTILIZACIÓN DEL PROPULSOR SON LAS MISMAS PARA LOS TRES SISTEMAS DE MISILES

EJEMPLO 6: SOLUCIÓN

1. $H_0: \mu_1 = \mu_2 = \mu_3$
2. H_1: LAS TRES MEDIAS SON DIFERENTES
3. $\alpha = 0{,}05$
4. REGIÓN CRÍTICA: $h > \chi^2_\alpha = 5{,}991$, PARA $\upsilon = 3 - 1 = 2$ GRADOS DE LIBERTAD

EN EXCEL, PERO CON LA DISTRIBUCIÓN GAMMA, CONSIDERANDO QUE $\alpha = \upsilon/2$ Y β=2 (=DISTR.GAMMA.INV(0,95; 1; 2)

5. CÁLCULOS: CONVERTIMOS LAS OBSERVACIONES A RANGOS Y SUMAMOS LOS RANGOS PARA CADA SISTEMA DE MISILES

$$n_1 = 5 \quad n_2 = 6 \quad n_3 = 8$$

$$r_1 = 61 \quad r_2 = 63{,}5 \quad r_3 = 65{,}5$$

$$h = \frac{12}{19(19+1)}\left(\frac{61^2}{5} + \frac{63{,}5^2}{6} + \frac{65{,}5^2}{8}\right) - 3(19+1) = 1{,}66$$

6. DECISIÓN: COMO $h = 1{,}66$ NO CAE EN LA REGIÓN CRÍTICA $h > 5{,}991$, TENEMOS INSUFICIENTE EVIDENCIA PARA RECHAZAR LA HIPÓTESIS DE QUE LAS TASAS DE UTILIZACIÓN DEL PROPULSOR SON LAS MISMAS PARA LOS TRES SISTEMAS DE MISILES.

Tasas de utilización del propulsor			
Sistemas de misiles:	**1**	**2**	**3**
	19	18	7
	1	14,5	11
	17	6	2,5
	14,5	4	2,5
	9,5	16	13
	$r_1 = 61$	5	9,5
		$r_2 = 63{,}5$	8
			12
			$r_3 = 65{,}5$

PRUEBA DE CORRIDAS (ALEATORIEDAD)

LAS PRUEBAS DE CORRIDAS SE BASAN EN EL ORDEN EN EL QUE SE OBTIENEN LAS OBSERVACIONES MUESTRALES, ES UNA TÉCNICA ÚTIL PARA PROBAR LA HIPÓTESIS NULA H_0 DE QUE LAS OBSERVACIONES EN REALIDAD SE EXTRAEN AL AZAR.

PRUEBA DE CORRIDAS (ALEATORIEDAD)

Para entender qué es una corrida, considérese una secuencia formada por dos símbolos, *a* y *b*, por ejemplo:

$$a\ a \mid b\ b\ b \mid a \mid b\ b \mid a\ a\ a\ a\ a \mid b\ b\ b \mid a\ a\ a\ a \mid$$

Una corrida se define como un conjunto de símbolos idénticos (o semejantes) que se encuentran entre dos símbolos diferentes o entre ningún símbolo (como el principio y final de la secuencia)

Las corridas se leen de izquierda a derecha, el final de la primer corrida está señalado por una línea vertical, y así seguimos contando las diferentes corridas

En el ejemplo, hay siete corridas

PRUEBA DE CORRIDAS (ALEATORIEDAD)

Parece claro que debe existir alguna relación entre aleatoriedad y cantidad de rachas (corridas)

En la siguiente secuencia:

$$a|b|a|b|a|b|a|b|a|b|a|b|$$

se observa un ***patrón cíclico***, por lo que es difícil pensar que es aleatorio

En este caso, se tienen *demasiadas corridas* (se tiene la cantidad máxima posible, dada la cantidad de letras *a* y *b*)

Por otro lado, en la siguiente secuencia:

$$a\ a\ a\ a\ a\ a|b\ b\ b\ b|a\ a\ a\ a\ a|b\ b\ b|$$

parece haber un ***patrón de tendencia*** en el que se agrupan (o acumulan) la letras *a* y las letras *b*

En este caso hay *muy pocas* rachas y no se puede considerar que la secuencia sea aleatoria

Se considerará que una secuencia no es aleatoria si hay demasiadas o muy pocas rachas; si ni es así se considera que la secuencia es aleatoria

PRUEBA DE CORRIDAS (ALEATORIEDAD)

Para cuantificar se consideran todas las secuencias posibles que tengan una cantidad N_1 del símbolo 1 y cantidad N_2 del símbolo 2

El total de símbolos está dado por:

$$N = N_1 + N_2$$

La colección de todas estas secuencias proporciona una distribución muestral: a cada secuencia le corresponde una cantidad de rachas, denotada V

De esta manera se llega a la distribución muestral del estadístico V

La distribución muestral V tiene las siguiente media y varianza:

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1$$

$$\sigma_V^2 = \frac{2\,N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)}$$

Se puede probar la hipótesis de aleatoriedad al nivel de significancia adecuado

Se encuentra que si tanto N_1 como N_2 son por lo menos 8, entonces la distribución muestral de V se aproxima a una distribución normal:

$$z = \frac{V - \mu_V}{\sigma_V}$$

Para valores menores, debe trabajarse con datos tabulados en tablas.

OTRAS APLICACIONES DE LAS PRUEBAS DE LAS CORRIDAS

Prueba mayor y menor que la mediana para la aleatoriedad de datos numéricos:

- Para determinar si un conjunto de datos numéricos es aleatorio,
  1) Se colocan los datos en el mismo orden que se obtuvieron
  2) Se encuentra la mediana de esos datos y cada dato se reemplaza por una *a*, si su valor es mayor que la mediana o por una *b* si su valor es menor que la mediana

  La muestra es o no aleatoria según si la secuencia de letras *a* y *b* sea o no aleatoria

Diferencias entre poblaciones de las que se ha tomado una muestra:

- Supongamos que dos muestras de tamaño m y n se denotan $a_1, a_2, \ldots, a_m$ y $b_1, b_2, \ldots, b_n$, respectivamente
- Para decidir si las muestras provienen o no de una misma población, primero se ordenan todos los $m + n$ valores muestrales de menor a mayor
- Si hay valores iguales, se deben ordenar mediante un proceso aleatorio
- Si la secuencia obtenida es aleatoria, se concluye que las muestras realmente no son diferentes y que, por lo tanto, provienen de la misma población

EJEMPLO 7

- EN 30 LANZAMIENTOS DE UNA MONEDA SE OBTIENE LA SIGUIENTE SECUENCIA DE CARAS (H) Y CRUCES (T)

H T T H T H H H T H H T T H T H T H H T H T T H T H H T H T

a) DETERMINAR LA CANTIDAD DE *V* RACHAS

b) AL NIVEL DE SIGNIFICANCIA 0,05 PROBAR SI ESTA SECUENCIA ES ALEATORIA

- SOLUCIÓN

a) USAMOS UNA LÍNEA VERTICAL PARA SEPARAR LAS RACHAS: HAY 22 RACHAS

H | T T | H | T | H H H | T | H H | T T | H | T | H | T | H H | T | H | T T | H | T | H H | T | H | T |

b) EN ESTA MUESTRA HAY:

$$N_1 = 16\, caras \qquad N_2 = 14\, cruces \qquad \mathrm{V} = 22 \qquad z_{\alpha/2} = 1{,}96\ (a\ dos\ colas)$$

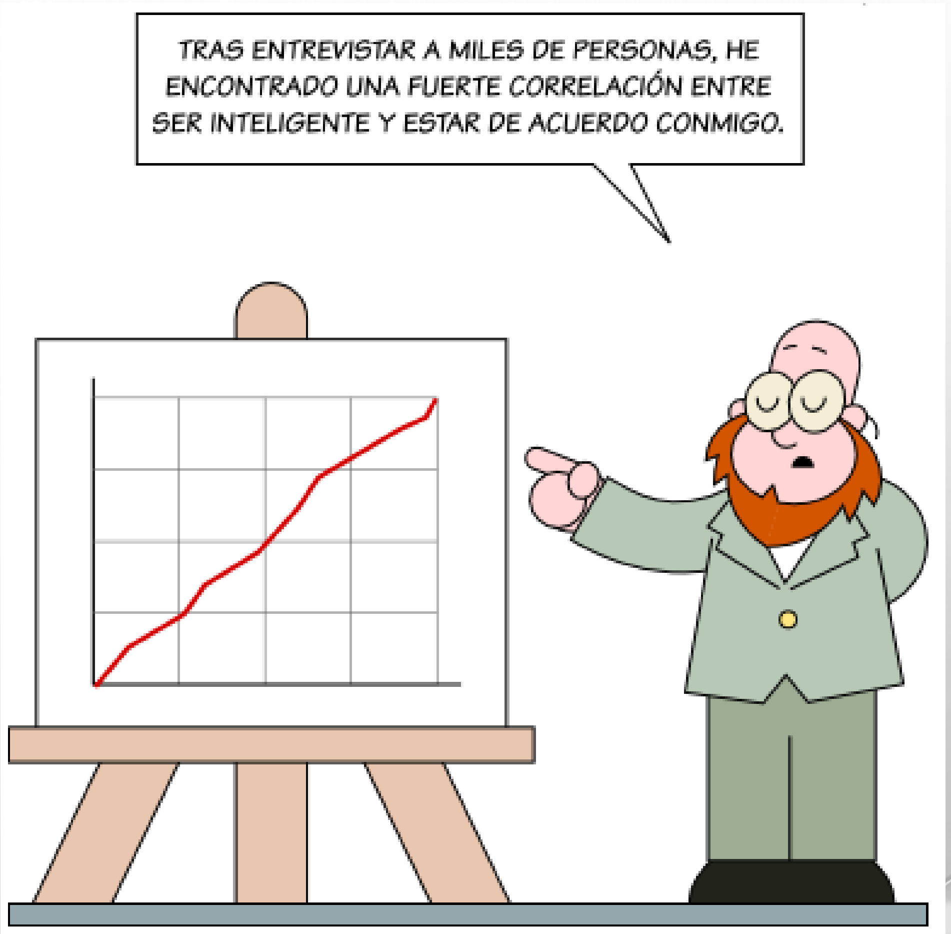
$$\mu_V = \frac{2 \cdot 16 \cdot 14}{16 + 14} + 1 = 15{,}93 \qquad \sigma_V^2 = \frac{2 \cdot 16 \cdot 14(2 \cdot 16 \cdot 14 - 16 - 14)}{(16 + 14)^2(16 + 14 - 1)} = 7{,}175 \qquad \sigma_V = 2{,}679$$

$$z = \frac{22 - 15{,}93}{2{,}679} = 2{,}27$$

EN UNA PRUEBA DE DOS COLAS, AL NIVEL DE SIGNIFICANCIA 0,05 SE ACEPTARÁ LA HIPÓTESIS H_0 DE ALEATORIEDAD SI $-1{,}96 \leq z \leq 1{,}96$ Y SE RECHAZARÁ SI NO ES ASÍ. COMO EL VALOR OBTENIDO PARA Z ES 2,27>1,96, SE CONCLUYE AL NIVEL DE SIGNIFICANCIA DE 0,05 QUE LOS LANZAMIENTOS NO SON ALEATORIOS. LA PRUEBA INDICA QUE HAY DEMASIADAS RACHAS, LO QUE INDICA UN PATRÓN CÍCLICO EN LOS LANZAMIENTOS

COEFICIENTE DE CORRELACIÓN DE RANGO DE SPEARMAN

EL COEFICIENTE DE CORRELACIÓN DE RANGOS DE SPEARMAN SE EMPLEA COMO CONTRAPARTE NO PARAMÉTRICA DEL COEFICIENTE DE CORRELACIÓN CONVENCIONAL



COEFICIENTE DE CORRELACIÓN DE RANGO DE SPEARMAN

Una medición no paramétrica de la asociación entre dos variables *X* e *Y* está dada por el ***coeficiente de correlación de rangos***

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia entre los rangos asignados x_i e y_i, y *n* es el número de pares de datos

El valor de r_s por lo general está cercano al valor que se obtienen al encontrar *r* con base en mediciones numéricas y se interpreta casi en la misma forma

* Varía entre -1 y +1
* Un valor de -1 o +1 indica una asociación perfecta entre *X* e *Y*
* El signo más para rangos idénticos y el signo menos para rangos inversos
* Cuanto más cercano es a cero, significa que las variables no están correlacionadas

EJEMPLO 8

- LAS CIFRAS QUE SE LISTAN EN LA TABLA, PUBLICADAS POR LA COMISIÓN FEDERAL DE COMERCIO, MUESTRAN LOS MILIGRAMOS DE ALQUITRÁN Y NICOTINA QUE SE ENCUENTRA EN 10 MARCAS DE CIGARRILLOS. CALCULE EL COEFICIENTE DE CORRELACIÓN DE RANGOS PARA MEDIR EL GRADO DE RELACIÓN ENTRE EL CONTENIDO DE ALQUITRÁN Y NICOTINA EN CIGARRILLOS.

Contenidos de alquitrán y nicotina		
Marca de cigarrillos	Contenido de alquitrán	Contenido de nicotina
Viceroy	14	0,9
Marlboro	17	1,1
Chesterfield	28	1,6
Kool	17	1,3
Kent	16	1,0
Raleigh	13	0,8
Old Gold	24	1,5
Philip Morris	25	1,4
Oasis	18	1,2
Players	31	2,0

EJEMPLO 8: SOLUCIÓN

- REPRESENTAMOS CON *X* E *Y* LOS CONTENIDOS DE ALQUITRÁN Y NICOTINA, RESPECTIVAMENTE. PRIMERO ASIGNAMOS RANGOS A CADA CONJUNTO DE MEDIDAS (DE MENOR A MAYOR)
- SUSTITUIMOS LOS VALORES EN LA FÓRMULA:

$$r_s = 1 - \frac{6 \cdot 5{,}5}{10(10^2 - 1)} = 0{,}967$$

ESTE RESULTADO IMPLICA QUE HAY UNA CORRELACIÓN ALTA ENTRE LA CANTIDAD DE ALQUITRÁN Y DE NICOTINA QUE SE ENCUENTRA EN LOS CIGARRILLOS

Contenidos de alquitrán y nicotina				
Marca de cigarrillos	x_i	y_i	d_i	d_i^2
Viceroy	2	2	0	0
Marlboro	4,5	4	0,5	0,25
Chesterfield	9	9	0	0
Kool	4,5	6	-1,5	2,25
Kent	3	3	0	0
Raleigh	1	1	0	0
Old Gold	7	8	-1	1
Philip Morris	8	7	1	1
Oasis	6	5	1	1
Players	10	10	0	0
				5,5