Unidad VI

PRUEBAS PARA VARIABLES CUALITATIVAS, MEDIDAS DE ASOCIACIÓN Y CORRELACIÓN

"It is easy to lie with statistics. It is hard to tell the truth without statistics"

Andrejs Dunkels

Introducción

En muchas investigaciones no se analizan solo variables cuantitativas, sino también de tipo cualitativo.
De las variables cualitativas se puede aprovechar su frecuencia, más aún cuando se cruzan dos variables de este tipo, se pueden formar tablas de contingencia.
Si esta tabla de contingencia se forma a partir de los datos de una muestra aleatoria se puede utilizar la Prueba de Independencia, la cual permite verificar si las dos variables están relacionadas; mientras que si los datos provienen de varias muestras se puede hacer uso de la Prueba de Homogeneidad de Subpoblaciones, la cual permite verificar si las subpoblaciones no provienen de una misma población.
Si con la Prueba de Independencia se demuestra que las variables están relacionadas una posterior interrogante que se desea responder es que tan fuerte es la relación existente entre las dos variables. Esto se puede determinar con una serie de indicadores que se desarrollarán en este capítulo.

Por otro lado, si se tienen dos variables que se encuentran medidas en al menos una escala ordinal y se desea analizar si estas variables se encuentran o no correlacionadas no solo se puede hacer uso de la Correlación de Pearson, pues para realizar inferencia sobre este coeficiente se debe demostrar que los datos provienen de una distribución normal bivariada. Si este requisito no se cumple se puede hacer uso de otros coeficientes de correlación como el de Spearman o de Kendall.

También, se puede considerar estudios donde adicionalmente a las dos variables que se desea analizar puede existir una tercera variable que permite segmentar grupos, a esta tercera variable usualmente se le conoce como capa. Se presentará una prueba que permita analizar este tipo de situaciones.

En este capítulo, se presentará el análisis de variables cualitativas, algunas medidas de correlación no paramétrica para variables medidas en al menos escala intervalo, así como las respectivas pruebas estadísticas que determinan la significación de la asociación observada.

1. Pruebas para variables cualitativas nominales

Cuando se utiliza variables cualitativas en una investigación se puede aprovechar la frecuencia de sus categorías.
Esto ya fue se puede haber visto en la prueba de frecuencias o prueba de proporciones para una muestra cuando se analiza una variable. Sin embargo, cuando se quiere analizar dos variables, estas se pueden cruzar obteniéndose una tabla de contingencia o una tabla de contingencia en una o varias capas o estratos.
La definición formal de una tabla de contingencia se desarrollará a continuación. Pero, vale la pena mencionar que para la evaluación de una tabla de contingencia por lo general se utiliza el estadístico Chi Cuadrado de Pearson.

Tabla de Contingencia

Es un cuadro de doble entrada en el cual se recoge la frecuencia conjunta de los datos de una o varias muestras aleatorias. Estas frecuencias son clasificadas de acuerdo a las clases ó categorías de una variable A y a las clases ó categorías de una variable B.

Sea "A" una característica con sus categorías a_1, a_2,…, a_c y "B" una característica con sus categorías b_1, b_2,..., b_f

		Característica A				**Total**
		a_1	**a_2**	**…**	**a_c**	
Carac. B	**b_1**	o_{11}	o_{12}	…	o_{1c}	$n_{1.}$
	b_2	o_{21}	o_{22}	…	o_{2c}	$n_{2.}$
	⋮					
	b_f	o_{f1}	o_{f2}	…	o_{fc}	$n_{r.}$
Total		$n_{.1}$	$n_{.2}$		$n_{.c}$	$n_{..}$

Donde:

i = 1, 2, ..., f "filas"
j = 1, 2, ..., c "columnas"

$$n_{i}. = \sum_{j=1}^{c} o_{ij} \qquad n_{i}. = \sum_{i=1}^{f} o_{ij} \qquad n.. = \sum_{i=1}^{f}\sum_{j=1}^{c} o_{ij}$$

A los totales de filas y columnas se les conoce como totales marginales.

La ij–ésima frecuencia observada $\left(o_{ij}\right)$ indica el número de veces que se repite un elemento en las categorías *i* y *j* a la vez.

1.1. Prueba de Independencia

- **Aspectos Generales**

Con frecuencia un investigador está interesado en saber si dos variables cualitativas son independientes o probablemente están relacionadas. Se dice que dos variables son independientes si la distribución de una variable no depende de la distribución del otro.

Esta prueba se aplica cuando los datos de **una muestra** aleatoria son clasificados de acuerdo a **dos características** (variables) y lo que se desea es probar si las características utilizadas como criterios de clasificación son independientes entre sí o si existe alguna relación entre ellas.

En una prueba de independencia los totales marginales de filas y columnas son aleatorios.

- **Supuestos**

- La muestra es seleccionada al azar.
- Los datos deben encontrarse en una escala nominal u ordinal y si se trabaja con variables de tipo intervalo o razón se deben categorizar.

1.2. Contraste de Homogeneidad de Sub-Poblaciones

Esta prueba se aplica cuando se desea verificar **si una característica** tiene un comportamiento semejante u homogéneo en dos o más poblaciones. Es decir, las muestras correspondientes a "C" poblaciones son clasificadas de acuerdo a las clases ó categorías de una característica "A".

En una prueba de homogeneidad de subpoblaciones uno de los totales marginales de filas y columnas es aleatorio y el otro es fijo.

La prueba Chi-cuadrado se utiliza también para contrastar la homogeneidad de varias muestras, es decir, si varias muestras pueden ser consideradas como seleccionadas de una misma población.

- **Supuestos**

- Las muestras son seleccionadas al azar.
- Los datos deben encontrarse en una escala nominal u ordinal y si se trabaja con variables de tipo intervalo o razón se deben categorizar.

- **Inferencia Estadística para ambas pruebas**

Estas pruebas se aplican cuando se desea verificar si al menos una de las frecuencias observadas (o_{ij}) perteneciente a la ij-ésima categoría (mutuamente excluyentes) difiere significativamente de su respectiva frecuencia teórica o frecuencia esperada (e_{ij}).

- Definir si la prueba se trata de un contraste de homogeneidad de sub-poblaciones o un contraste de independencia.
- Calcular las frecuencias esperadas (e_{ij}) de la siguiente manera**:**

$$e_{ij} = n_{..}p_{ij} \Rightarrow e_{ij} = n_{..}p_{i.}p_{.j} \Rightarrow e_{ij} = n_{..}\left(\frac{n_{i.}}{n_{..}}\right)\left(\frac{n_{.j}}{n_{..}}\right) \Rightarrow e_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

- Aplicar la siguiente prueba estadística
  Como medida de discrepancia, entre las frecuencias esperadas y observadas, Pearson propuso el siguiente estadístico:

$$\chi_c^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{[1-\alpha,(f-1)(c-1)]}$$

  También se puede hacer uso de la prueba de razón de verosimilitud

$$G = 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\ln\left(\frac{O_{ij}}{e_{ij}}\right) \sim \chi^2_{(1-\alpha,(f-1)(c-1))}$$

- Evaluar el valor calculado sobre la siguiente región crítica
  Valores elevados del estadístico χ^2 evidencian discrepancias relevantes entre las frecuencias observadas (o_{ij}) y las esperadas (e_{ij}), por lo que deberá rechazarse la hipótesis nula de que dicha muestra procede de una población con probabilidades teóricas π_i. Por lo tanto, si $\chi_c^2 > \chi^2_{[1-\alpha,(f-1)(c-1)]}$ se rechaza H_0.

  La hipótesis para la Prueba de Independencia es:
  H_0: Las variables X e Y son independientes (no están relacionadas)
  H_1: Las variables X e Y no son independientes (están relacionadas)

  La hipótesis para la Prueba de Homogeneidad de Subpoblaciones es:
  H_0: Las subpoblaciones provienen de una misma población
  H_1: Las subpoblaciones no provienen de una misma población

  Observaciones:
  Si se tiene un solo grado de libertad para el valor crítico, el tamaño de la muestra es pequeño (n<50) o existe un valor esperado menor a 5, se puede hacer uso de la Corrección de Yates, el cual hace un ajuste al estadístico χ^2

$$\chi_c^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(|o_{ij}-e_{ij}|-0.5)^2}{e_{ij}} \sim \chi^2_{[1-\alpha,(f-1)(c-1)]}$$

- **Aplicación**
  Ejemplo 1: Prueba de Independencia
  El jefe de una planta industrial desea determinar si existe relación entre el rendimiento en el trabajo y turno laboral del empleado. Se tomó una muestra aleatoria de 400 empleados y se obtuvo las frecuencias observadas que se presentan en la siguiente tabla de contingencia:

Rendimiento en el trabajo	**Turno Laboral**			
	Mañana	**Tarde**	**Noche**	**Total**
Deficiente	23	60	29	112
Promedio	28	79	60	167
Muy bueno	9	49	63	121
Total	60	188	152	400

Con el nivel de significación 0.01, ¿La calificación del rendimiento del trabajador está asociada con el turno en el que labora el empleado?

Solución:
H_0: El rendimiento de un empleado en el trabajo es independiente del turno en el que labora.
H_1: El rendimiento de un empleado en el trabajo no es independiente del turno en el que labora.
α= 0,01
Prueba Estadística

$$\chi_c^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{[1-\alpha,(f-1)(c-1)]}$$

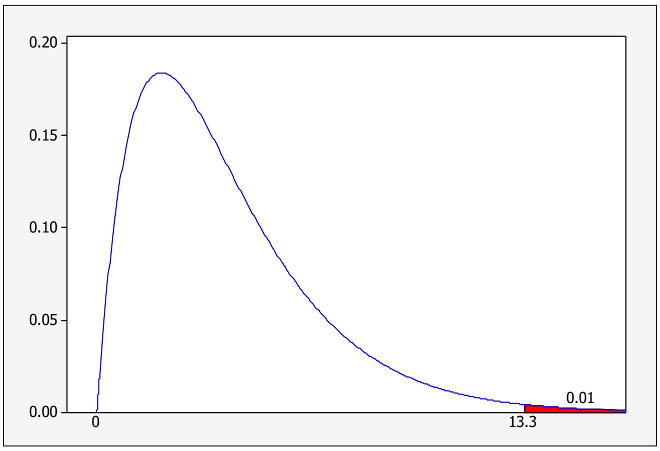
Desarrollo de la prueba

La siguiente tabla muestra tanto las frecuencias observadas como las esperadas (entre paréntesis)

Rendimiento en el trabajo	**Turno Laboral**			
	Mañana	**Tarde**	**Noche**	**Total**
Deficiente	23 (16.80)	60 (52.64)	29 (42.56)	112
Promedio	28 (25.05)	79 (78.49)	60 (63.46)	167
Muy bueno	9 (18.15)	49 (56.87)	63 (45.98)	121
Total:	**60**	**188**	**152**	**400**

$$\chi_c^2 = \frac{(23-16.80)^2}{16.80} + \frac{(28-25.05)^2}{25.05} + \ldots + \frac{(63-45.98)^2}{45.98} = 20.18$$

Criterios de decisión.



Si $\chi^2 > 13.277$ se rechaza H_0
Si $\chi^2 \leq 13.277$ no se rechaza H_0

Conclusión
Con nivel de significación 0,01 se puede afirmar que la calificación del rendimiento real de un empleado en el trabajo está relacionada con el turno en el que labora.

Ejemplo 2: Prueba de Homogeneidad

Muestras de tres tipos de materiales, sujetos a cambios extremos de temperatura, produjeron los resultados que se muestran en la siguiente tabla:

Condición	**Material A**	**Material B**	**Material C**	**Total**
Desintegrados	41	27	22	90
Permanecieron intactos	79	53	78	210
Total	120	80	100	300

Use un nivel de significancia de 0.05 para probar si, en las condiciones establecidas, la probabilidad de desintegración es diferente en al menos uno de los tres tipos de materiales.

Solución
Formulación de las hipótesis

H_0: La probabilidad de desintegración no difiere los tres tipos de materiales.
H_1: La probabilidad de desintegración es diferente en al menos uno de los tres tipos de materiales.

α=0.05
Prueba Estadística

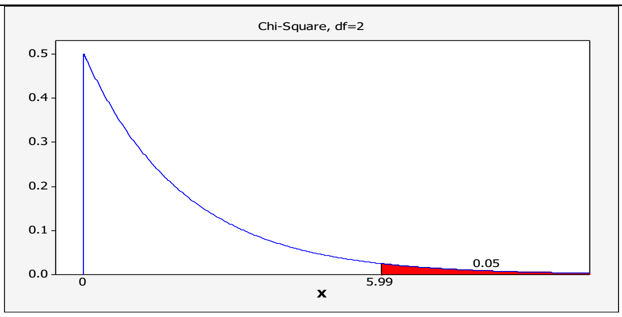
$$\chi_c^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2_{[1-\alpha,(f-1)(c-1)]}$$

Desarrollo de la Prueba

Condición	**Tipo de Material**			**Total**
	Material A	**Material B**	**Material C**	
Desintegrados	41 (36)	27 (24)	22 (30)	90
Permanecieron intactos	79 (84)	53 (56)	78 (70)	210
Total	120	80	100	300

$$\chi_c^2 = \frac{(41-36)^2}{36} + \frac{(79-84)^2}{84} + \ldots + \frac{(78-70)^2}{70} = 4.575$$

Criterios de decisión.



No se rechaza H_0 si: $\chi^2_c < 5.9915$
Se rechaza H_0 si: $\chi^2_c > 5.9915$

Conclusión
Con nivel de significación 0,05 no se rechaza la hipótesis nula.
Por lo tanto, no se puede afirmar que la probabilidad de desintegración es diferente en al menos uno de los tres tipos de materiales

- **Secuencia o funciones con programas estadísticos**

En R
En R existe una chisq.test que permite obtener el resultado para ambas pruebas.
chisq.test(x,y) o chisq.test(tabla)
La función assocstats del paquete vcd permite obtener la prueba de Razón de Verosimilutud.

- **Resultados con programas estadísticos**

Resultados con R
```
tabla<-matrix(c(23,60,29,28,79,60,9,49,63),3,3,byrow=TRUE)
chisq.test(tabla)
```

```
        Pearson's Chi-squared test
data:  tabla
X-squared = 20.1789, df = 4, p-value = 0.0004604
```

```
tabla<-matrix(c(41,27,22,79,53,78),2,3,byrow=TRUE)
library(vcd)
assocstats(tabla)
```

```
                    X^2 df P(> X^2)
Likelihood Ratio 4.7265  2 0.094113
Pearson          4.5754  2 0.101500
```

- **Algunas consideraciones de los programas estadísticos**
  En R

- Si realiza la corrección de Yates solo para tablas 2x2
- Permite hacer la prueba para datos agrupados y sin agrupar en una tabla de contingencia.

1.3. Prueba Exacta de Fisher

- **Aspectos Generales**

Es una prueba muy buena para analizar variables nominales binarias que provienen de dos muestras independientes que son pequeñas.
Las observaciones de cada una de las muestras son clasificadas en una de las dos categorías con las que cuenta la variable de interés. Es decir se formar una tabla de contingencia 2x2.
La prueba determina si los dos grupos difieren en las proporciones en la clasificación de la variable en estudio.

- **Supuestos**

- Las dos muestras son seleccionadas al azar.
- Las muestras son independientes.
- Los datos deben encontrarse en una escala nominal u ordinal y si se trabaja con variables de tipo intervalo o razón se deben categorizar.

- **Inferencia Estadística**

Para llevar a cabo la prueba se debe realizar lo siguiente:

- Clasificar las muestras en las 2 categorías de la variable de interés, de tal manera que se forme una tabla de contingencia 2x2 de la siguiente manera:

Variable	Grupo		Combinación
	I	II	
+	A	B	A+B
-	C	D	C+D
Total	A+C	B+D	n

Se desea determinar si los grupos I y II difieren significativamente en la proporción de signos más (+) y signos menos (-) pertenecientes a cada grupo.
Para ello se debe calcular la probabilidad exacta de observar un conjunto particular de frecuencias en una tabla 2x2, cuando los totales marginales se consideran fijos, la cual está dada por la distribución hipergeométrica

$$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{n}{A+B}} = \frac{(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!}{n!\,A!\,B!\,C!\,D!}$$

Hipótesis

Bilateral	Unilateral	
Caso A	Caso B	Caso C
$H_0: \pi_1 = \pi_2$	$H_0: \pi_1 = \pi_2$	$H_0: \pi_1 = \pi_2$
$H_1: \pi_1 \neq \pi_2$	$H_1: \pi_1 > \pi_2$	$H_1: \pi_1 < \pi_2$

- **Aplicación**

Se tienen dos grupos de pacientes (hombres y mujeres) a los que se les proporcionó un analgésico. Los resultados (mejoró (+) ó no mejoró (-)) luego de un periodo son los siguientes:

Variable	Grupo		Combinación
	Mujeres	Hombres	
Mejoró(+)	5	1	6
No mejoró (-)	2	7	9
Total	7	8	15

Pruebe si la proporción de mujeres que mejoró supera a la proporción de hombres que mejoró luego de proporcionado el analgésico. Use α=0.05

$H_0 : \pi_1 = \pi_2$

$H_1 : \pi_1 > \pi_2$

α=0.05

p_1=5/7 p_2=1/8 entonces p_1-p_2=0.714-0.125 = 0.589, se deben encontrar todas las combinaciones superiores a 0.589. Esto solo ocurre para las tablas I y II por lo que el pvalor es igual a 0.0014+0.0336=0.035

Tabla					P_1	P_2	$P_1 - P_2$	$P(tabla)$
I:		1	2		0.857	0	0.857	0.0014
	+	6	0	6				
	−	1	8	9				
		7	8	15				
II:		1	2		0.714	0.125	0.589	0.0336
	+	5	1	6				
	−	2	7	9				
		7	8	15				
III:		1	2		0.571	0.250	0.321	0.1958
	+	4	2	6				
	−	3	6	9				
		7	8	15				
IV:		1	2		0.429	0.375	0.054	0.3916
	+	3	3	6				
	−	4	5	9				
		7	8	15				
V:		1	2		0.286	0.500	−0.214	0.2937
	+	2	4	6				
	−	5	4	9				
		7	8	15				
VI:		1	2		0.143	0.625	−0.482	0.0783
	+	1	5	6				
	−	6	3	9				
		7	8	15				
VII:		1	2		0	.750	−0.750	0.0056
	+	0	6	6				
	−	7	2	9				
		7	8	15				

Conclusión
A un nivel de significación de 0.05, se puede afirmar que la proporción de mujeres que mejoró luego de aplicado el analgésico es superior a la proporción de hombres que mejoró luego de aplicado el analgésico.

Si la hipótesis hubiese sido
$H_0 : \pi_1 = \pi_2$
$H_1 : \pi_1 \neq \pi_2$
Para calcular el pvalor se considerarían los valores más extremos a 0.589 en valor absoluto, por lo que el valor se calcularía de la siguiente manera:
Pvalor= 0.0014+ 0.0336 + 0.0056 = 0.041

- **Secuencia o funciones con programas estadísticos**

En R
Existe la función fisher.test
fisher.test(x,y, alternativa) o fisher.test(tabla, alternativa)

- **Resultados con programas estadísticos**

En R
```
tabla<-matrix(c(5,2,1,7),2,2)
fisher.test(tabla,alternative="g")
```

```
        Fisher's Exact Test for Count Data
data:  tabla
p-value = 0.03497
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.179718      Inf
sample estimates:
odds ratio
  13.59412
```

- **Algunas consideraciones de los programas estadísticos**

En R
- Realiza los casos bilateral y unilateral.
- Se puede realizar la prueba con los datos sin agrupar o agrupados en una tabla de contingencia 2x2.

1.4. Prueba de Mantel-Haenszel-Cochran

- **Aspectos Generales**

Esta prueba utiliza tres variables; la primera es considerada como estratos (o capas) y dentro de cada una de ella se clasifican las otras dos variables.
Si cada una de las tablas que se forma en su respectivo estrato proviene de un estudio independiente, la prueba de Mantel-Haenszel-Cochran es una herramienta que estudia en forma conjunta como un metaanálisis.
Esta prueba supone que no hay interacción entre las tres variables en estudio.

- **Supuestos**

- Las muestras son seleccionadas al azar.
- Las muestras son independientes.
- Los datos deben encontrarse en una escala nominal u ordinal y si se trabaja con variables de tipo intervalo o razón se deben categorizar.

- **Inferencia Estadística**

- Clasificar dentro de cada estrato las variables de interés.
- Se denomina p_{1i} a la proporción de elementos de la primera fila que caen en la primera columna y p_{2i} a la proporción de elementos de la segunda fila que caen en la primera columna de la tabla i.
- En cada tabla i hay n_i observaciones, todas ellas pueden ser categorizadas como del tipo 1 (r_i de ellos) o del tipo 2 (n_i-r_i de ellos). Si c_i elementos son seleccionados del total de los n_i elementos, la probabilidad que exactamente x_i de los elementos seleccionados son del tipo 1 es:

$$\frac{\binom{r_i}{x_i}\binom{n_i - r_i}{c_i - x_i}}{\binom{n_i}{c_i}}$$

De igual manera, todos los elementos pueden ser categorizados como del tipo A (c_i de ellos) o del tipo B (n_i-c_i de ellos), la probabilidad de que exactamente x_i de los seleccionados son del tipo A es:

$$\frac{\binom{c_i}{x_i}\binom{n_i - c_i}{r_i - x_i}}{\binom{n_i}{r_i}}$$

De seguro que las dos probabilidades son iguales

$$\frac{\binom{r_i}{x_i}\binom{n_i - r_i}{c_i - x_i}}{\binom{n_i}{c_i}} = \frac{\binom{c_i}{x_i}\binom{n_i - c_i}{r_i - x_i}}{\binom{n_i}{r_i}}$$

Esas son probabilidades hipergeométricas con media y varianza:

$$\frac{r_i c_i}{n_i} \text{ y } \frac{r_i c_i\left(n_i - r_i\right)\left(n_i - c_i\right)}{n_i^2\left(n_i - 1\right)}$$

Los k estratos son independientes por lo que el estadístico es:

$$T = \frac{\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \frac{r_i c_i}{n_i}}{\sqrt{\sum_{i=1}^{k} \frac{r_i c_i \left(n_i - r_i\right)\left(n_i - c_i\right)}{n_i^2 \left(n_i - 1\right)}}} \sim N(0,1)$$

Hipótesis

Bilateral	Unilateral	
Caso A	Caso B	Caso C
$H_0 : \pi_{1i} = \pi_{2i}$	$H_0 : \pi_{1i} = \pi_{2i}$	$H_0 : \pi_{1i} = \pi_{2i}$
$H_1 : \pi_{1i} \neq \pi_{2i}$	$H_1 : \pi_{1i} > \pi_{2i}$	$H_1 : \pi_{1i} < \pi_{2i}$

Se desea probar si esto sucede en todos los estratos. Es decir, si la proporción de éxitos con respecto a una categoría es diferente, mayor o menor a la proporción de éxitos con respecto a la otra categoría en todos los estratos en estudio.

- **Aplicación**

Se tiene tablas 2x2 de la clasificación de personas de 3 localidades con respecto a su hábito de fumar y su diagnóstico de cáncer. Los resultados se presentan a continuación:

Localidad 1			
Tipo	**Diagnos.**		Total
	Si	**No**	
Fumador	3	1	4
No Fum	3	2	5
Total	6	3	9

Localidad 2			
Tipo	**Diagnos.**		**Total**
	Si	**No**	
Fumador	20	6	26
No Fum.	22	13	35
Total	42	19	61

Localidad 3			
Tipo	**Diagnos.**		**Total**
	Si	**No**	
Fumador	4	1	5
No Fum.	12	4	16
Total	16	5	21

Pruebe si la proporción de incidencia de cáncer para fumadores y no fumadores no coincide en las 3 localidades. Use α=0.05.

$H_0 : \pi_{1i} = \pi_{2i}$
$H_1 : \pi_{1i} \neq \pi_{2i}$ para $\forall$ i=1,2,3

α=0.05

Prueba Estadística

Desarrollo de la prueba estadística

$$\sum_{i=1}^{k} x_i = 3 + 20 + 4 = 27$$

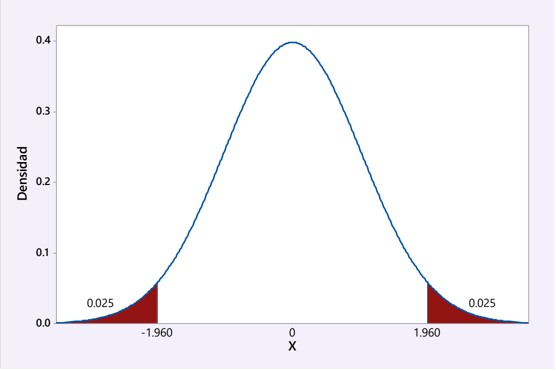
$$\sum_{i=1}^{k} \frac{r_i c_i}{n_i} = \frac{(6)(4)}{9} + \frac{(42)(26)}{61} + \frac{(16)(5)}{21} =$$

$$\sum_{i=1}^{k} \frac{r_i c_i (n_i - r_i)(n_i - c_i)}{n_i^2 (n_i - 1)} = \frac{(4)(6)(5)(3)}{(9)^2 (8)} + \frac{(26)(42)(35)(19)}{(61)^2 (60)} + \frac{(5)(16)(16)(5)}{(21)^2 (20)} = 4.533$$

$$T = \frac{\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \frac{r_i c_i}{n_i}}{\sqrt{\sum_{i=1}^{k} \frac{r_i c_i (n_i - r_i)(n_i - c_i)}{n_i^2 (n_i - 1)}}} = \frac{27 - 24.378}{\sqrt{4.533}} = 1.232$$

pvalor<- 2*(1-pnorm(1.232))
0.2179491

Criterios de decisión.



No se rechaza H_0 si: -1.96<Zcal < 1.96
Se rechaza H_0 si: Zcal > 1.96 o Zcal>-1.96

Conclusión
A un nivel de significación de 0.05, no se puede afirmar que la proporción de incidencia de cáncer para fumadores y no fumadores no coincide en las 3 localidades.

- **Secuencia o funciones con programas estadísticos**

En R
Existe la función mantelhaen.test, en donde se debe indicar el conjunto de datos como un arreglo
mantelhaen.test(tabla, alternativa)

- **Resultados con programas estadísticos**

Resultados con R

```
tabla<-array(c(3,3,1,2,20,22,6,13,4,12,1,4),dim=c(2,2,3))
mantelhaen.test(tabla)
```

```
      Mantel-Haenszel  chi-squared  test  with  continuity
correction
data:  tabla
Mantel-Haenszel X-squared = 0.9933, df = 1, p-value = 0.3189
alternative hypothesis: true common odds ratio is not equal
to 1
95 percent confidence interval:
 0.6984315 4.9240804
sample estimates:
common odds ratio
          1.85449
```

```
mantelhaen.test(tabla,correct=FALSE)
```

```
        Mantel-Haenszel chi-squared test without continuity
correction
data:  tabla
Mantel-Haenszel X-squared = 1.5166, df = 1, p-value = 0.2181
alternative hypothesis: true common odds ratio is not equal
to 1
95 percent confidence interval:
 0.6984315 4.9240804
sample estimates:
common odds ratio
          1.85449
```

- **Algunas consideraciones de los programas estadísticos**
  En R
- Se puede realizar la prueba con los datos sin agrupar.
- Analiza los casos bilateral y unilateral.

2. Medidas de Asociación

En el proceso de investigación, se puede desear conocer si dos variables están relacionadas y si es así determinar cuál es su grado de relación.

En esta sección se presentará medidas de correlación no paramétrica y sus respectivas pruebas estadísticas que permiten determinar la significación de la asociación observada. El problema de medir el grado de asociación entre dos variables es más general que el de probar la existencia de algún grado de asociación.

En el caso paramétrico, la medida usual de correlación es el coeficiente de Pearson. Este estadístico requiere que las variables estén medidas en al menos una escala de intervalo, para una adecuada interpretación del estadístico.

Si deseamos probar la significación del este coeficiente, debemos no sólo utilizar la medida requerida, sino también verificar que las observaciones provengan de una distribución normal bivariada.

El coeficiente de correlación de Pearson mide el grado en el cual existe una relación lineal entre las variables.

Si para un conjunto de datos los supuestos antes mencionados no son sostenibles, entonces se debe usar un coeficiente de correlación alternativo como es el caso de los coeficientes de Spearman o de Kendall.

2.1 Coeficiente V de Cramer

- **Aspectos Generales**

Es una medida del grado de asociación o relación entre dos variables cualitativas. Se usa únicamente cuando se tiene datos categóricos en escala nominal. El coeficiente de Cramer, al ser calculado de una tabla de contingencia, proporciona los mismos valores sin considerar cómo fueron ordenadas las categorías en las filas y columnas.

- **Supuestos**

  - La muestra es seleccionada al azar.
  - Los datos deben encontrarse en una escala nominal u ordinal y si se trabaja con variables de tipo intervalo o razón se deben categorizar.

- **Inferencia Estadística**

  - Con las variables A, con categorías A_1, A_2, …,A_k y B con categorías B_1, B_2, …, B_r, obtener la siguiente tabla de contingencia:

	A_1	A_2	…	A_k	Total
B_1	n_{11}	n_{12}	…	n_{1k}	R_1
B_2	n_{21}	n_{22}	…	n_{2k}	R_2
:	:	:		:	:
B_r	n_{r1}	n_{r2}	…	n_{rk}	R_r
Total	C_1	C_2	…	C_k	n

Los datos pueden consistir en cualquier número de categorías, es decir, se puede calcular un coeficiente V de Cramer para datos en una tabla rxk.

- Calcular el coeficiente de Cramer mediante:

$$V = \sqrt{\frac{\chi^2}{n(L-1)}}$$

Donde:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}}{e_{ij}} - n$$ y *L = min(r,k)*

Mientras mayor sea la asociación entre las dos variables será más grande el valor del coeficiente de Cramer. El coeficiente de Cramer varía entre 0 y 1.

Hipótesis
H_0: No existe asociación entre las variables X e Y. H_0: $\nu = 0$
H_1: Existe asociación entre las variables X e Y. H_1: $\nu \neq 0$

Podemos probar si una V observada difiere significativamente de cero simplemente al determinar la significación del estadístico χ^2 para la tabla de contingencia asociada, debido a que V es una función lineal de χ^2. Ya que sabemos que la distribución muestral de χ^2, conocemos la de V^2 y por tanto, la de V.
Para cualquier tabla de contingencia rxk, podemos determinar la significación del grado de asociación (la significación de V) averiguando la probabilidad asociada con la ocurrencia, cuando H_0 es cierta, de valores tan grandes a los valores observados de χ^2, con (r-1)(k-1) grados de libertad. Si la χ^2 para el estadístico de la muestra es significativo, entones podemos concluir que en la población la asociación entre las dos series de atributos no es cero, esto es, que los atributos o las variables no son independientes.

En general, es deseable que un índice de asociación muestre al menos las siguientes características:

- Cuando las variables sean independientes y exista una carencia completa de asociación entre las variables, el valor del índice debe ser cero.
- Cuando las variables muestren completa dependencia una de la otra, esto es, cuando estén perfectamente asociadas, el estadístico debe ser igual a la unidad.

El coeficiente V de Cramer tiene algunas limitaciones y es por esa razón que han aparecido otros coeficientes alternativos como: Coeficiente de contingencia corregido de Pawlik, Cuadrado medio de contingencia, Coeficiente de Tschuprow, entre otros.
Algunas limitaciones del coeficiente V de Cramer son:

- El coeficiente V de Cramer tiene la primera característica es igual a cero cuando no existe asociación entre las variables en la muestra. Sin embargo, cuando es igual a la unidad, pudiera no ser una asociación “perfecta” entre las variables.

- Una segunda limitación de V es que los datos deben ser fáciles de usar con el estadístico χ^2, con el propósito que su significación pueda ser interpretada apropiadamente, esto es la prueba Chi Cuadrado solo debe aplicarse sólo si menos del 20% de las celdas en la tabla de contingencia tienen frecuencias esperadas menores que cinco y ninguna celda tiene una frecuencia esperada menor que uno.

- Una tercera limitación de V es que no resulta directamente comparable con cualquier otra medida de correlación, por ejemplo, la r de Pearson, la r_s de Spearman o la T de Kendall). Estas medidas se aplican a variables ordenadas, mientras que el coeficiente de Cramer es apropiado para usarse con variables categóricas (escala nominal).

A pesar de estas limitaciones, el coeficiente de Cramer es una medida de asociación extremadamente útil debido a su amplia aplicabilidad. Dicho coeficiente no hace suposiciones acerca de la forma de las distribuciones poblacionales de donde provienen las variables que están siendo evaluadas.

Otra ventaja del coeficiente V de Cramer es que permite al investigador comparar tablas de contingencia de diferentes tamaños y lo más importante, tablas basadas en diferentes tamaños de muestra. Aunque el estadístico χ^2 no mide la independencia de dos variables, es sensible al tamaño de la muestra. El coeficiente V de Cramer hace que las comparaciones de las relaciones obtenidas en diferentes tablas resulten más fáciles.

- **Aplicación**

Koch & Edwards (1988) realizaron un ensayo clínico doble ciego que investiga un nuevo tratamiento para la artritis reumatoide. En un experimento doble ciego, ni los individuos participantes ni los investigadores saben quién pertenece al grupo de control (el que recibe placebos) y quién es el grupo experimental. Solamente después de haberse recolectado todos los datos, y concluido el experimento, los investigadores conocen qué individuos pertenecen a cada grupo.

Utilice las variables Treatment y Improved del conjunto de datos Arthritis del paquete vcd para obtener el coeficiente de Cramer y evaluar su significancia a un α=0.05.
H_0: $\nu = 0$
H_1: $\nu \neq 0$
α=0.05

$$V = \sqrt{\frac{\chi^2}{n(L-1)}} = \sqrt{\frac{13.055}{84(2-1)}} = 0.3942$$

χ^2 = 13.055
Pvalor=0.001 < α se rechaza H_0
Conclusión
A un α=0.05, se puede afirmar que el coeficiente de asociación V de Cramer es significativo.

- **Secuencia o funciones con programas estadísticos**
  En R
  Existe la función cramersV del paquete lsr
  cramersV(tabla)
  La función assocstats del paquete vcd también permite obtener el coeficiente V de Cramer y otras medidas de asociación
  assocstats(tabla)
  Las funciones Assocs y CramerV del paquete DescTools también permiten obtener el coeficiente V de Cramer
  CramerV(tabla)
  Assocs(tabla)

- **Resultados con programas estadísticos**
  Resultados con R

```
library(vcd)
data("Arthritis")
tabla<-table(Arthritis[,2],Arthritis[,5])
assocstats(tabla)
```

```
                    X^2 df  P(> X^2)
Likelihood Ratio 13.530  2 0.0011536
Pearson          13.055  2 0.0014626

Phi-Coefficient   : 0.394
Contingency Coeff.: 0.367
Cramer's V        : 0.394
```

```
library(lsr)
cramersV(tabla)
```

```
[1] 0.3942295
```

```
library(DescTools)
CramerV(tabla1)
```

```
[1] 0.3942295
```

```
Assocs(tabla1)
```

```
                      estimate      lwr.ci     upr.ci
Phi Coeff.          3.9420e-01           -          -
Contingency Coeff.  3.6680e-01           -          -
Cramer V            3.9420e-01  1.5650e-01 5.9580e-01
```

- **Algunas consideraciones de los programas estadísticos**
  En R

- Brinda el V de Cramer.
- Presenta en el resultado del V de Cramer la evaluación de su significancia para la función assocstats.

2.2 Coefficiente de Contingencia de Pearson

- **Aspectos Generales**

Es una medida del grado de asociación alternativo al V de Cramer. Para poder estimarlo se debe construir primero una tabla de contingencia.

- **Supuestos**

- La muestra es seleccionada al azar.
- Los datos deben encontrarse en una escala nominal u ordinal y si se trabaja con variables de tipo intervalo o razón se deben categorizar.

- **Inferencia Estadística**

- Construir la tabla de contingencia.
- Calcular el estadístico Chi Cuadrado
- Calcular el coeficiente de Contingencia de Pearson mediante:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Donde:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}}{E_{ij}} - n$$

Mientras mayor sea la asociación entre las dos variables será más grande el valor del coeficiente de contingencia de Pearson. El coeficiente de Contingencia de Pearson varía entre 0 y C_{max}.

El máximo valor del coeficiente de contingencia depende de la dimensión de la tabla de contingencia.

Si la tabla de contingencia es cuadrada (rxr), entonces $C_{\max} = \sqrt{\frac{r-1}{r}}$

Si la tabla de contingencia es de dimensión (rxk), entonces L=min(r,k) $C_{\max} = \sqrt{\frac{L-1}{L}}$

Hipótesis

H_0: No existe asociación entre las variables X e Y. H_0: $\kappa = 0$
H_1: Existe asociación entre las variables X e Y. H_1: $\kappa \neq 0$

Al igual que el coeficiente V de Cramer, para probar si κ difiere significativamente de cero simplemente al determinar la significación del estadístico χ^2 para la tabla de contingencia asociada.

- **Aplicación**

Utilice las variables Treatment y Improved del conjunto de datos Arthritis del paquete vcd provenientes del estudio de Koch & Edwards (1988) para obtener el coeficiente de Contingencia. Evalúe su significancia a un α=0.05.

H_0: $\kappa = 0$
H_1: $\kappa \neq 0$
α=0.05

$$V = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{13.055}{13.055 + 84}} = 0.367$$

χ^2 = 13.055
Pvalor=0.001 < α se rechaza H_0

Conclusión
A un α=0.05, se puede afirmar que el coeficiente Contingencia es significativo.

- **Secuencia o funciones con programas estadísticos**
  En R
  Existe la función assocstats del paquete vcd
  assocstats(tabla).
  También dentro del paquete DescTools, se pueden utilizar las funciones ContCoef o Assocs
  ContCoef(tabla)
  Assocs(tabla)

- **Resultados con programas estadísticos**
  Resultados con R

```
library(vcd)
assocstats(tabla1)
```

```
                    X^2 df  P(> X^2)
Likelihood Ratio 13.530  2 0.0011536
Pearson          13.055  2 0.0014626

Phi-Coefficient   : NA
Contingency Coeff.: 0.367
Cramer's V        : 0.394
```

```
library(DescTools)
ContCoef(tabla1)
```

```
[1] 0.3667581
```

```
Assocs(tabla1)
                         estimate      lwr.ci     upr.ci
Phi Coeff.             3.9420e-01           -          -
Contingency Coeff.     3.6680e-01
```

- **Algunas consideraciones de los programas estadísticos**
  En R

- Solo la función assocstats permite evaluar la significancia del coeficiente de Contingencia.

2.3 Coeficiente Phi

- **Aspectos Generales**

Es una evaluación de la asociación o relación entre dos variables medidas en una escala nominal, cada uno de los cuales puede tomar sólo dos valores. De hecho, es idéntico en valor al coeficiente de Cramer.

- **Supuestos**

- La muestra es seleccionada al azar.
- Los datos deben encontrarse en una escala nominal u ordinal y si se trabaja con variables de tipo intervalo o razón se deben categorizar en una variable binaria.

- **Inferencia Estadística**

- Arreglar los datos en una tabla 2x2. Ya que los datos son dicotómicos, supondremos que los datos son codificados como cero y uno para cada variable, aunque puede ser usada cualquier asignación del valor binario.

Variable Y	Variable X		Total
	0	1	
1	A	B	A+B
0	C	D	C+D
Total	A+C	B+D	N

- El coeficiente Phi para una tabla 2x2 es definido como:

$$r_{\phi}=\frac{|AD-BC|}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Cuyo rango puede ser desde cero hasta uno.

- El coeficiente Phi está relacionado con el estadístico χ^2 que se usa para probar la independencia de variables categóricas (medidas nominalmente). De aquí que la significación del coeficiente Phi puede probarse al usar el estadístico χ^2.

$$\chi^2=\frac{n\left(|AD-BC|-n/2\right)^2}{(A+B)(C+D)(A+C)(B+D)}\sim\chi^2_{(1-\alpha,1)}$$

Hipótesis

H_0: No existe relación entre las variables X e Y. H_0: $\phi = 0$
H_1: Existe relación entre las variables X e Y. H_1: $\phi \neq 0$

- **Aplicación**

En una segunda vuelta electoral para la elección presidencial se quiere analizar si existe relación entre los candidatos y el género del elector. Se seleccionó una muestra aleatoria de electores, obteniéndose los siguientes resultados:

Género	Candidato	
	A	B
Masculino	29	12
Femenino	44	26

Calcule el coeficiente phi y evalúe su significancia a un α=0.05.
H_0: $\phi = 0$
H_1: $\phi \neq 0$
α=0.05
r_ϕ=0.08
χ^2= 0.712
Pvalor=0.399
Conclusión
A un α=0.05, no se puede afirmar que existe relación entre el género y el candidato de preferencia en la segunda vuelta electoral.

- **Secuencia o funciones con programas estadísticos**
  En R
  Existe la función phi del paquete psych
  phi(tabla)
  La función assocstats del paquete vcd también permite obtener el coeficiente Phi y otras medidas de asociación
  assocstats(tabla).
  El paquete DescTools con sus funciones Assocs y Phi también permiten obtener el Coeficiente Phi.

```
Assocs(tabla)
Phi(tabla)
```

- **Resultados con programas estadísticos**
  Resultados con R

```
library(vcd)
tabla<-matrix(c(29,44,12,26),2,2)
assocstats(tabla)
```

```
                    X^2 df P(> X^2)
Likelihood Ratio 0.72046  1  0.39599
Pearson          0.71212  1  0.39874

Phi-Coefficient   : 0.08
Contingency Coeff.: 0.08
Cramer's V        : 0.08
```

```
library(psych)
phi(tabla)
```

```
[1] 0.08
```

```
library(DescTools)
```

```
Phi(tabla2)
```

```
[1] 0.01800945
```

```
Assocs(tabla2)
```

```
                         estimate      lwr.ci     upr.ci
Phi Coeff.             1.8000e-02           -          -
Contingency Coeff.     1.8000e-02           -          -
```

- **Algunas consideraciones de los programas estadísticos**
  En R

- Brinda el coeficiente Phi y su significancia para la función assocstats.

3. Medidas de Correlación

1.1. Coeficiente de Correlación r_S de Spearman de rangos ordenados

➢ **Aspectos Generales**

El coeficiente de correlación de Spearman mide el grado de asociación entre dos variables cuantitativas que siguen una tendencia siempre creciente o decreciente. Es decir, es más general que el coeficiente de correlación de Pearson, el cual asume que la relación entre las dos variables es lineal, la correlación de Spearman en cambio se puede calcular para las relaciones exponenciales o logarítmicas entre las variables.
Es una medida de asociación entre dos variables que requiere que ambas estén medidas en al menos una escala ordinal, de tal manera que los elementos en estudio puedan ser colocados en rangos en dos series ordenadas.

➢ **Supuestos**

- La muestra es seleccionada al azar.
- Los datos deben encontrarse en una escala al menos ordinal.

➢ **Inferencia Estadística**

- Se obtiene los rangos para cada una de las variables (X e Y) de manera independiente.
- Se calcula la diferencia de rangos di para cada pareja de observaciones, restando el rango de Y_i menos el rango de X_i.
- Se eleva al cuadrado cada d_i y se calcula la suma de estos valores.
- Se calcula:

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

Cuando ocurren puntuaciones empatadas, a cada una de ellas se le asigna el promedio de los rangos.
Si la proporción de las observaciones empatadas no es grande, su efecto sobre r_S es insignificante y puede usarse la expresión presentada anteriormente. Si la proporción de empates es grande, entonces debe incorporarse un factor de corrección en el cálculo de r_S.

$$r_s = \frac{\left(n^3-n\right) - 6\sum_{i=1}^{n} d_i^2 - \left(T_x + T_y\right)/2}{\sqrt{\left(n^3-n\right)^2 - \left(T_x+T_y\right)\left(n^3-n\right) + T_x T_y}}$$

Donde

$T_x = \sum_{i=1}^{g}\left(t_i^3 - t_i\right)$, donde g es el número de grupos de diferentes rangos empatados y t_i es número de elementos empatados en el i-ésimo grupo.

Prueba de significación de r_S
Se puede probar la hipótesis nula de que las dos variables en estudio no están asociadas (son independientes) contra la hipótesis H_1 que existe asociación entre X e Y (una prueba bidireccional) o existe una asociación positiva (o negativa) entre X e Y (una prueba unidireccional).

Cuando n es superior a 20, la significación de r_S puede ser probada mediante el estadístico

$$z = r_s\sqrt{n-1} \sim N(0,1)$$

También se puede hacer uso del estadístico

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} \sim t_{(n-2)}$$

Hipótesis

Bilateral	Unilateral	
Caso A	Caso B	Caso C
$H_0 : \rho_s = 0$	$H_0 : \rho_s = 0$	$H_0 : \rho_s = 0$
$H_1 : \rho_s \neq 0$	$H_1 : \rho_s > 0$	$H_1 : \rho_s < 0$

Las hipótesis especificadas en el número a) conducen a una prueba bilateral y se utilizan cuando se desea descubrir cualquier desviación de la independencia. Las pruebas unilaterales indicadas en los números b) y c) se utilizan, respectivamente, cuando el investigador desea saber si puede concluir que las variables están directa o inversamente correlacionadas.

- **Aplicación**
La tabla siguiente muestra los consumos de calorías (cal/día/Kg) y de oxígeno VO_2 (ml/min/Kg.) de 10 niños.

N° de niño	**Consumo de calorías (X)**	**VO_2 (Y)**	**Rango (X)**	**Rango (Y)**	***d_i***	d_i^2
1	50	7.0	2	1	-1	1
2	70	8.0	3	2	-1	1
3	90	10.5	5	6	1	1
4	120	11.0	8	8	0	0
5	40	9.0	1	3	2	4
6	100	10.8	6	7	1	1
7	150	12.0	9	10	1	1
8	110	10.0	7	5	-2	4
9	75	9.5	4	4	0	0
10	160	11.9	10	9	-1	1
					Total	14

Pruebe la hipótesis nula de que las dos variables son mutuamente independientes, contra la alternativa de que están directamente relacionadas. Use α=0.05.

Solución
H_0: Los consumos de calorías y de oxígeno VO2 son mutuamente excluyentes.
H_0: $\rho_S = 0$
H_1: Los consumos de calorías y de oxígeno VO_2 están directamente relacionadas. H_1: $\rho_S > 0$
α=0.05
Prueba Estadística

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

Criterio de Decisión
No se rechaza H_0 si $r_S \leq$ 0.5515
Se rechaza H_0 si $r_S >$ 0.5515

Desarrollo de la Prueba

$$r_s = 1 - \frac{6(14)}{10(10^2-1)} = 1 - \frac{84}{990} = 0.915$$

Conclusión
Existe suficiente evidencia estadística a un nivel de significación de 0.05 para rechazar la H_0.
Por lo tanto, podemos afirmar que los consumos de calorías y de oxígeno VO_2 están directamente relacionados.

- **Secuencia o funciones con programas estadísticos**
  Existe la función cor.test del paquete Stat
  `cor.test(x,y,método=spearman, alternativa)`
  También existe el paquete pspearman con la función spearman.test
  `spearman.test(x,y,alternativa,aproximación)`

- **Resultados con programas estadísticos**
  Resultados con R

```
x<-c(50,70,90,120,40,100,150,110,75,160)
y<-c(7,8,10.5,11,9,10.8,12,10,9.5,11.9)
cor.test(x,y,method="spearman",alternative="g")
```

```
        Spearman's rank correlation rho

data:  x and y
S = 14, p-value = 0.0002334
alternative hypothesis: true rho is greater than 0
sample estimates:
      rho
0.9151515
```

- **Algunas consideraciones de los programas estadísticos**
- Permite analizar los casos unilaterales y bilaterales.

1.2. Coeficiente de Correlación Txy de Kendall

➢ **Aspectos Generales**

Otro indicador para poder analizar la correlación entre dos variables que se encuentran medidas en al menos escala ordinal es el coeficiente de correlación de Kendall

Una ventaja de T sobre el coeficiente de correlación de Spearman es que T puede ser generalizada a un coeficiente de correlación parcial.

➢ **Supuestos**

- La muestra es seleccionada al azar.
- Los datos deben encontrarse en una escala al menos ordinal.

➢ **Inferencia Estadística**

- Primero se debe calcular el coeficiente de correlación de Kendall como el número de acuerdos menos el número de desacuerdos entre el número total de combinaciones tomados en dos.

  Por ejemplo:

  Supóngase que para poner el rango de calidad de cuatro objetos (a, b, c y d) preguntamos a los jueces X e Y.

Ensayo	a	b	c	d
Juez X	3	4	2	1
Juez Y	3	1	4	2

Si arreglamos el orden de los ensayos de tal modo que los rangos del juez X aparezcan en orden natural (1, 2, ... , n) tenemos:

Ensayo	d	c	a	b
Juez X	1	2	3	4
Juez Y	2	4	3	1

Ahora se puede determinar el grado de correspondencia entre los jueces X e Y, es decir, cuántos pares de rangos en el conjunto del juez Y están en su orden correcto, respecto a aquellos del juez X. Considérese primero todos los posibles pares de rangos en los cuales el rango del juez Y es 2 (el primer rango en este conjunto) y los miembros posteriores del lado derecho, se le asigna un +1 si el orden es correcto y -1 si el orden es incorrecto. Las comparaciones lo podríamos resumir en la siguiente tabla:

Juez X	**1**	**2**	**3**	**4**	
Juez Y	**2**	**4**	**3**	**1**	**Total**
	2→	+	+	-	1
		4→	-	-	-2
			3→	-	-1
				1→	0
	Gran total				-2

Así el número total de acuerdos en el ordenamiento menos el número desacuerdos en el ordenamiento entre los rangos es -2. El número total de posibles comparaciones es:

$$\binom{n}{2} = \binom{4}{2} = 6$$

El coeficiente de correlación por orden de rangos de Kendall es la razón:

$$T = \frac{\text{\#de acuerdos - \# de desacuerdos}}{\text{\# total de pares}} = \frac{-2}{6} = -\frac{1}{3} = -0.333$$

En general, el máximo posible total será

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

Si denominamos la suma observada de puntuaciones +1 (acuerdos) y puntuaciones -1 (desacuerdos) para todos los pares como S, entonces el coeficiente de correlación de Kendall es:

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

Cuando dos o más observaciones están empatadas ya sea en la variable X o Y, utilizamos nuestro procedimiento usual de colocar los rangos a las puntuaciones empatadas; se les da a la observación ligadas el promedio de los rangos que deberían haber recibido si no hubiera habido empates.
El efecto de los empates es cambiar el denominador de nuestra ecuación para T. En el caso de empates, T se convierte en:

$$T = \frac{2S}{\sqrt{n(n-1)-T_x}\sqrt{n(n-1)-T_y}}$$

Donde
$T_x = \sum t(t-1) \quad T_y = \sum t(t-1)$
Siendo t el número de observaciones empatadas en cada grupo de empates en la variable X e Y respectivamente

```
x<-c(3,4,2,1)
y<-c(3,1,4,2)
cor(x,y,method="kendall")
[1] -0.3333333
```

El coeficiente de correlación de Spearman se interpreta de igual manera que el coeficiente de Pearson, calculado entre variables cuyos valores consisten en rangos. Por otra parte, el coeficiente de correlación de rangos de Kendall tiene interpretación diferente, esta es la diferencia entre la probabilidad de que, en los datos observados X e Y estén en el mismo orden y la probabilidad de que los datos de X e Y estén en un orden diferente.

- Para evaluar la significancia del coeficiente de Kendall, se considera que si una muestra aleatoria se extrae de alguna población en la cual X e Y no están relacionadas y se les ponen rangos a los miembros de la muestra en X e Y, entonces para cualquier orden dado de los rangos de X, todos los posibles ordenes de rangos de Y son igualmente probables. Supóngase que ordenamos los rangos de X en orden natural, 1, 2,.., n; para este orden, todos los n! posibles órdenes de rangos de Y son igualmente probables según H_0. Por tanto, cualquier orden particular de los rangos de Y tiene una probabilidad de ocurrencia, cuando H_0 es cierta, de 1/n!.
  Para cada uno de los n! posibles rangos de Y, existirá un valor asociado, estos posibles valores de T variarán desde -1 hasta +1 y pueden ser obtenidos en una distribución de frecuencias, pero naturalmente al aumentar el valor de n este método se vuelve más tedioso.
  Si la muestra es grande, la distribución de T se aproxima a la distribución normal:

$$z = \frac{3T\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \sim N(0,1)$$

Hipótesis

Bilateral	Unilateral	
Caso A	Caso B	Caso C
$H_0: \tau_{xy} = 0$	$H_0: \tau_{xy} = 0$	$H_0: \tau_{xy} = 0$
$H_1: \tau_{xy} \neq 0$	$H_1: \tau_{xy} > 0$	$H_1: \tau_{xy} < 0$

- **Aplicación**
  A continuación, se presenta las calificaciones de 12 estudiantes a dos temas de interés. Pruebe a un α=0.05 si existe relación entre estos dos temas de interés

Tema1	3	4	2	1	8	11	10	6	7	12	5	9
Tema2	2	6	5	1	10	9	8	3	4	12	7	11

H_0: τ_{xy}= 0
H_1: $\tau_{xy} \neq 0$
α=0.05
Pvalor = 0.0018

Conclusión
Existe suficiente evidencia estadística a un nivel de significación de 0.05 para rechazar la H_0.
Por lo tanto, podemos afirmar que existe relación entre los dos temas de interés.

- **Secuencia o funciones con programas estadísticos**
  En R
  Existe la función cor.test

```
cor.test(x,y,método=kendall, alternativa)
```

- **Resultados con programas estadísticos**
  Resultados con R

```
Tema1<-c(3,4,2,1,8,11,10,6,7,12,5,9)
Tema2<-c(2,6,5,1,10,9,8,3,4,12,7,11)
cor.test(Tema1,Tema2,method="kendall")
```

```
        Kendall's rank correlation tau

data:  Tema1 and Tema2
T = 55, p-value = 0.001803
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.6666667
```

- **Algunas consideraciones de los programas estadísticos**
  En R
- Permite analizar los casos unilaterales y bilaterales.
- Presenta el estadístico de prueba.

1.3. Coeficiente de Correlación Parcial $T_{xy.z}$ de Kendall de rangos

- **Aspectos Generales**

Cuando se observa correlación entre dos variables, existe siempre la posibilidad de que la correlación se deba a la asociación entre cada una de las dos variables y una tercera variable.
Estadísticamente, este problema puede ser atacado por métodos de correlación parcial. En la correlación parcial, se eliminan los efectos de variación en una tercera variable sobre la relación entre las variables X e Y. En otras palabras se encuentra la correlación entre X e Y manteniéndose constante la tercera variable Z.

- **Supuestos**
- La muestra es seleccionada al azar.
- Los datos deben encontrarse en una escala ordinal.

- **Inferencia Estadística**
- Se deben calcular todas las posibles correlaciones de Kendall entre las tres variables T_{xy}, T_{xz} y T_{yz}.
- Calcular el coeficiente de correlación parcial de Kendall mediante la siguiente expresión

$$T_{xy.z} = \frac{T_{xy} - T_{xz}T_{yz}}{\sqrt{\left(1-T_{xz}^2\right)\left(1-T_{yz}^2\right)}}$$

Si la muestra es suficientemente grande (n>50), se puede hacer uso del siguiente estadístico de prueba para evaluar la significancia del coeficiente de correlación parcial de Kendall:

$$z = \frac{3T_{xy.z}\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \sim N(0,1)$$

Hipótesis

Bilateral	Unilateral	
Caso A	Caso B	Caso C
$H_0 : \tau_{xy.z} = 0$	$H_0 : \tau_{xy.z} = 0$	$H_0 : \tau_{xy.z} = 0$
$H_1 : \tau_{xy.z} \neq 0$	$H_1 : \tau_{xy.z} > 0$	$H_1 : \tau_{xy.z} < 0$

- **Aplicación**

En un estudio de psicología se ha evaluado las puntuaciones de tres temas: autoritarismo (X), estatus de lucha (Y) y la conformidad a la presión de grupo (z). Los resultados de la evaluación a doce personas se presentan a continuación:

X	3	4	2	1	8	11	10	6	7	12	5	9
Y	2	6	5	1	10	9	8	3	4	12	7	11
Z	1.5	1.5	3.5	3.5	5	6	7	8	9	10.5	10.5	12

Se desea verificar si existe relación entre el autoritarismo y estatus de lucha debido a la conformidad a la presión de grupo.

H_0: $\tau_{xy.z} = 0$
H_1: $\tau_{xy.z} \neq 0$
α=0.05

Z = 2.776
Pvalor=0.0055<α se rechaza H_0.

Conclusión
Existe suficiente evidencia estadística a un nivel de significación de 0.05 para rechazar la H_0.
Por lo tanto, podemos afirmar que existe relación si existe relación entre el autoritarismo y estatus de lucha debido a la conformidad a la presión de grupo.

- **Secuencia o funciones con programas estadísticos**

En R
Existe la función cor y a partir de ella se debe obtener la correlación parcial
```
cor(x,y,método=kendall, alternativa)
```
Tambien existe la función pcor.test del paquete ppcor
```
pcor.test(X,Y,Z,method="kendall")
```

- **Resultados con programas estadísticos**

Resultados con R
```
X<-c(3,4,2,1,8,11,10,6,7,12,5,9)
Y<-c(2,6,5,1,10,9,8,3,4,12,7,11)
Z<-c(1.5,1.5,3.5,3.5,5,6,7,8,9,10.5,10.5,12)
```

```
XY<-cor(X,Y,method="kendall")
XZ<-cor(X,Z,method="kendall")
YZ<-cor(Y,Z,method="kendall")
Txyz<-(XY-XZ*YZ)/sqrt((1-XZ^2)*(1-YZ^2))
n<-length(X)
zcal<-(3*Txyz*sqrt(n*(n-1)))/sqrt(2*(2*n+5))
[1] 2.776892
2*(1-pnorm(zcal))
[1] 0.005488142

pcor.test(X,Y,Z,method="kendall")
```

```
estimate     p.value statistic  n gp  Method
1 0.6135709 0.008610245  2.627154 12  1 kendall
```

1.4. Otros coeficientes basados en la concordancia de observaciones

- **Aspectos Generales**

El concepto de concordancia se utiliza para estimar índices como: Tau-b, Tau-c, Gamma y D Somers para variables ordinales.

Análisis de concordancias

Se traducen a rangos los valores de las variables originales X e Y.
Por ejemplo: Dadas las variables A, B y C y sus respectivos rangos RA, RB y RC.

A	B	C	RA	RB	RC
1	11	34	1	1	5
4	12	32	2	2	4
7	13	30	3	3	3
8	56	21	4	4	2
9	58	15	5	5	1

Si se calcula el coeficiente de Spearman obtendríamos el valor 1 para la pareja de variables A y B, y -1 para las parejas A y C, y B y C. Se puede utilizar una técnica de análisis más intuitiva:
Se pueden contar el número de concordancias, discordancias y empates entre parejas de casos.
Si pasamos del caso 1 al caso 2 de A, vemos que el valor del rango aumenta, y lo mismo ocurre al pasar del caso 1 de B al caso 2 de B, entonces decimos que ha ocurrido una concordancias en la pareja A&B (simbolizada con C), en cambio, al pasar del caso 1 al caso 2 de A, ocurre un aumento de sus rangos, y al pasar del caso 1 al caso 2 de C acurre una disminución de sus rangos, decimos que ha ocurrido una discordancia en la pareja A&C (simbolizada con D).
Si en todas las M parejas posibles de valores hay M concordancias, la relación entre las variable es la máxima positiva. Si de todas las M parejas posibles de valores hay M discordancias, la relación entre las dos variables es máxima negativa. Si existen M/2 discordancias y M/2 concordancias, cabe esperar una relación nula.
Un empate ocurre cuando al menos una de las dos variables presenta el mismo valor en ambos casos. Hay tres tipos de empates: el empate en la variable A y

no en B, el empate en la variable B y no en A, y el empate en ambos. Se simbolizan respectivamente, como E_A, E_B y E_D.

- **Supuestos**
  - La muestra es seleccionada al azar.
  - Los datos deben encontrarse en al menos una escala ordinal.

a) ÍndiceTau de Kendall

$$\tau = \frac{C-D}{\frac{n(n-1)}{2}} = \frac{2(C-D)}{n(n-1)}$$

$$-1 \leq \tau \leq 1$$

Su interpretación es similar a la correlación de Pearson. Un inconveniente es que no considera los empates, que sí están contados en el denominador.

b) Índice Gamma de Goodman y Kruskal (γ)

$$\gamma = \frac{C-D}{C+D}$$

Tampoco considera los empates, pero si D = 0, se obtiene el valor 1, máxima relación positiva, si C = 0, se obtiene el valor -1, máxima relación negativa.
Si C = D, se obtiene un coeficiente de cero, no existe relación lineal entre las variables.

c) Índice D de Sommers

Este índice incluye los empates en su fórmula:

$$D^* = \frac{C-D}{\frac{(C+D+E_A)+(C+D+E_B)}{2}} = \frac{C-D}{C+D+\frac{E_A+E_B}{2}}$$

Alcanza los valores máximos (1 o -1) cuando no hay empates.

d) Índices Tau-b y Tau-c de Kendall (τ_B y τ_C)

La tau-b, denominada comúnmente tau de Kendall y Stuart, utiliza el mismo criterio de la D de Sommers, sólo que en lugar de usar en el denominador una media aritmética, usa una media geométrica.

$$\tau_b = \frac{C-D}{\sqrt{(C+D+E_A)(C+D+E_B)}}$$

La tau-c de Kendall, en lugar de manipular el número de empates, utiliza el valor de V, que es el número más pequeño entre los diferentes valores que toma cada variable.

$$\tau_c = \frac{2V(C-D)}{n^2(V-1)}$$

- **Aplicación**

Ejemplo: Se tiene las siguientes 4 variables con 6 casos cada una.

X_1	X_2	X_3	X_4
1	1	1	1
2	1	1	2
3	2	1	3
4	3	2	4
5	5	4	4
6	4	3	4

Asignando rangos tenemos:

RX_1	RX_2	RX_3	RX_4
1	1.5	2	1
2	1.5	2	2
3	3	2	3
4	4	4	5
5	6	6	5
6	5	5	5

El número total de parejas entre n datos es n(n-1)/2. Luego en este caso existen (6)(5)/2 = 15 parejas.
Para cada par de variables analizaremos el número de concordancias, discordancias y empates.

Variables	C	D	E_A	E_B	E_D
1-2	13	1	0	1	0
1-3	11	1	0	3	0
1-4	12	0	0	3	0
2-3	12	0	0	2	1
2-4	11	0	1	3	0
3-4	9	0	3	3	0

- Calcular el Índice Tau de Kendall para las variables X_1 y X_2:

$$\tau = \frac{2(C-D)}{n(n-1)} = \frac{2(13-1)}{6(5)} = 0.8$$

Lo cual indica una relación lineal fuerte y directa entre las variables X_1 y X_2.

- Calcular la Gamma de Goodman y Kruskal para las variables X_1 y X_2:

$$\gamma = \frac{C-D}{C+D} = \frac{13-1}{13+1} = 0.8571$$

- Calcular la D de Sommers para las variables X_1 y X_2:

$$D^* = \frac{C-D}{C+D+\frac{E_A+E_B}{2}} = \frac{13-1}{13+1+\frac{0+1}{2}} = 0.8276$$

- Calcular la Tau-b de Kendall para las variables X_1 y X_2:

$$\tau_b = \frac{C-D}{\sqrt{(C+D+E_A)(C+D+E_B)}} = \frac{13-1}{\sqrt{(13+1+0)(13+1+1)}} = 0.8281$$

- Calcular la Tau-c de Kendall para las variables X_1 y X_2:

$$\tau_c = \frac{2V(C-D)}{n^2(V-1)} = \frac{2(5)(13-1)}{6^2(5-1)} = 0.833$$

V=min(6,5)=5

- **Secuencia o funciones con programas estadísticos**
  En R
  En el paquete vcdExtra se encuentra la función GKgamma con la cual se puede obtener el índice de Gamma y Kruskal
  GKgamma(tabla)
  También existe el paquete ryouready que presenta varias funciones que permite obtener varios índices como:
  Índice de Goodman y Kruskal
  ord.gamma(tabla)
  La D de Sommers
  ord.somers.d(tabla)
  Las Tau-b y Tau-c de Kendall
  ord.tau(tabla)

- **Resultados con programas estadísticos**
  Resultados con R

```
x1<-1:6
x2<-c(1.5,1.5,3,4,6,5)
library(vcdExtra)
tabla<-table(x1,x2)
GKgamma(tabla)

gamma        : 0.857
std. error   : 0.159
CI           : 0.545 1
```

```
library(ryouready)
ord.gamma(tabla)
Goodman-Kruskal Gamma: 0.857
ord.somers.d(tabla)
Somers' d:
        Columns dependent: 0.800
        Rows dependent: 0.857
        Symmetric: 0.828

ord.tau(tabla)
Kendall's (and Stuart's) Tau statistics
        Tau-b: 0.828
        Tau-c: 0.833>
```