resumen Estadistica descriptiva

Conceptos básicos de estadística

Muestreo

Muestreo probabilistico: misma probabilidad para todos los elementos de ser seleccionados en la poblacion muestral.

Muestreo no probabilistico: Seleccion arbitraria seleccionada por el investigador

Población (universo o colectivo)

Es el conjunto total de ELEMENTOS de la misma naturaleza cualquiera que sea, que son de interés para un problema dado

- N = Representación de el tamaño de la población

muestra

Variable aleatoria:

Son fenómenos o características de los elementos de la población.

Función de valor real que tiene como dominio el espacio muestral de un experimento aleatorio.

Variables sobre las cuales tenemos un grado de icertidumbre respecto a los valores que puede tomar

Datos

Son los resultados observados de las variables aleatorias (Cuando se hace una medición)

Parámetro

Es la medición global de cualesquer característica de los elementos de la población.

Es un valor teórico asociado a la población.

Ejemplos

Población: Los niños y niñas de 0 a 5 años de edad localozados en Bogotá

Variables: género, edad, peso, talla, estrato, localidad, fecha, lugar de nacimiento..

Parámetros: IMC

Clasificacion de variables

Cualitativas (categóricas)

Cuantitativas

Los valores de las observacione so niméricas y en consecuencia, ordenables.

Discreta

Recorridos finitos numerables sin tomar valores intermedios e.g. conteos.

Continua

Recorridos infinitos no numerables e.g. la distribución normal

Escalas de medición

Cualitativas

Nominales: Clasificación de objetos o fenómenos mediante símbolos o signos (No hay orden o dirección). e.g.

- Nombre
- Número de la cédula
- Tipo de sangre
- Color de los ojos

- Número de camiseta de los jugadores

Los números en la lista anterior no pueden ser sometidos a operaciones matemáticas

Ordinales

Categorías ordenadas (Rangos, órdenes, escalamientos)

- Sabor de un yogurt

Cuantitativas

Intervalo

Los datos medidos en una escala orrdinal para los cuales pueden clasificarse las distancias entre valores pero no existe un cero absoluto o no exista ausencia total de la característica

- Temperatura: a 0°C no deja de existir la temperatura
- Notas: se corre la escala e inicia desde 3.

Razón

Tiene todas las características de un intervalo, y además tiene un cero absoluto

Resumen y descripción de datos de una variable

Datos en bruto en forma de listas (o bases no son fáciles de usar para tomar decisiones)

- Se necesita algún tipo de organización

Para esto podemos utilizar gráficos de barras, graficos de torta, o tablas de frecuencias.

Como agrupar los datos: Sturgues

Si n no es demasiado grande, intervalos = $\sqrt{n}$

En caso contrario:

$$k = 1 + 3.322log(n) \tag{1}$$

k = intervalos de clase

Para la longitud de los intervalos:

$$L = \frac{Dato\ mayor - Dato\ menor}{n} \tag{2}$$

- A menudo es prueba y error

Tipos de frecuencias

- Absoluta: Conteo de observaciones que cae en cada intervalo.
- Relativa: $\frac{Absoluta}{n}$.
- Acumulada: Suma de las frecuencias absolutas
- Relativa acumulada: Suma de las frecuencias relativas.

Caracteristicas a revisar de las distribuciones

- Distribucion
- Localizacion (sesgo)
- Dispersion

Medidas de localización

Media aritmética:

Si $x_1, x_2, x_3, ...x_n$ es una muestra de una poblacion de tamaño N entonces la media es N

Media poblacional

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} \tag{3}$$

Estimador muestral

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

Caracteristicas:

- Es facil de obtener
- Medida no robusta: Afectada por valores extremos o datos atípicos.

Propiedades de la media aritmetica:

Si $x_1, x_2, x_3, ...x_n$ es una muestra de una poblacion de tamaño N entonces la media es N, entonces

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n \tag{5}$$

Si $x_i = c$ y a su vez c es constante, entonces

$$\sum_{i=1}^{n} x_i = \sum_{i=1} c = c + c + c + \cdots \tag{6}$$

Entonces

$$\sum_{i=1}^{n} x_i = nc \tag{7}$$

- *Ejemplo:*

$$\sum_{i=1}^{5} 2 = 2 + 2 + 2 + 2 + 2 \tag{8}$$

Si c es una constante que multiplica las observaciones:

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i \tag{9}$$

$$\sum_{i=1}^{n} cx_i = c \cdot x_1 + c \cdot x_2 + c \cdot x_3 + \cdots + c \cdot x_n \tag{10}$$

$$\sum_{i=1}^{n} cx_i = c\,(x_1 + x_2 + x_3 + \cdots + x_n) \tag{11}$$

$$\sum_{i=1}^{n} cx_i = c \sum_{i=1}^{n} x_i \tag{12}$$

Si $x_1, x_2, x_3, ...x_n$ y $y_1, y_2, y_3, ...y_n$ son sucesiones de numeros;

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{13}$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) + \cdots + (x_n + y_n) \tag{14}$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n) \tag{15}$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{16}$$

Si $x_1, x_2, x_3, ...x_n$ y $y_1, y_2, y_3, ...y_n$ son sucesiones de numeros;

$$\sum_{i=1}^{n} (x_i - y_i) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \tag{17}$$

5.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{18}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \tag{19}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \frac{n}{n} \sum_{i=1}^{n} x_i - n\bar{x} \tag{20}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = n\bar{x} - n\bar{x} \tag{21}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{22}$$

6. promedio de y en funcion de promedio de x en regresion lineal simple

Si $y_i = a + bx_i$ siendo a y b constante

$$\bar{y} = a + b\bar{x} \tag{23}$$

En efecto:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(a + bx_i) \tag{24}$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i \tag{25}$$

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \tag{26}$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = \frac{na}{n} + b\frac{\sum_{i=1}^{n} x_i}{n} \tag{27}$$

$$\bar{y} = a + b\bar{x} \tag{28}$$

La mediana

Es el valor central (es el dato de la variable que esta en el centro de la misma). Deja por encima y por debajo mitad y mitad de las observaciones.

Calculo de la mediana

Depende si el conjunto es par o impar:

Si$x_1, x_2, x_3, ...x_n$ Son los valores ordenados en una muestra de una poblacion de tamaño N:

$\hat{x} = \frac{x_{n/2} + x_{n+1/2}}{2}$ si n es par

$\hat{x} = x_{n=1/2}$ si n es impar

Es un estimador robusto, no se ve afectado por valores extremos

Ejemplo

Edad de ninos

```
x1 <- c(6,7,8,9,10)
```

n es impar, entonces $\hat{x} = x_{n+1/2} = x_{6/2} = x_3 = 8$

De la muestra analizada la mitad de los ninos tienen entre 6 y 8 años, y la otra mitad entre 8 y 10 años.

Moda

- El valor que más se repite
- Usada para valores numéricos o categóricos

e.g. Cual es el color más frecuente en los ojos.

Medidas de dispersión o variación

Varianza

Uno de los problemas es que la unidad de medida queda al cuadrado, e.g. si se miden cm, la varianza tiene unidades de cm^2:

- Varianza poblacional:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{29}$$

- Varianza muestral:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{30}$$

$$s^2 = \sum_{i=1}^{n} \frac{x_i^2 - 2x_i\bar{x} + \bar{x}^2}{n-1} \tag{31}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2}{n-1} \tag{32}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{n}{n}\sum_{i=1}^{n} x_i + n\bar{x}^2}{n-1} \tag{33}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2}{n-1} \tag{34}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2}{n-1} \tag{35}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} \tag{36}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n\bar{x}^2}{n-1} \tag{37}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n}{n-1}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 \tag{38}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{1}{n-1}\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \tag{39}$$

En algunos casos puede ser más conveniente calcular la varianza de esta forma.

Coeficiente de variación

$CV = \frac{s}{\hat{x}} \cdot 100\%$

Si CV es igual o menor a 5% hay homogeneidad

si esta entre 5% y 20% los datos son medianamente homogeneos

Si CV mayor a 20% hay heterogeneidad
Rango

medida no robusta, si hay datos atipicos se ve muy afectado

rango intercuartilico

boxplot

Cuartiles

Se divide en cuatro partes porcentiales el conjunto de observaciones.

Se calcula de la sigueinte manera.

Se ordena la muestra y se toma la posicion que corresponde.

$Q_k = k \cdot \frac{n}{4}$ $k = 1, 2, 3$

Deciles

Se divide en diez partes porcertialmente iguales

$D_k = \frac{n}{10}$ $1, 2, 3, ..., 9$

Percentiles

es mas detallado, nos da mas acceso a distintos puntos de la distribucion

$P_k = \frac{n}{100}$ $1, 2, 3, ..., 99$

Coeficientes de asimetría de Fisher

Permite interpretar la forma de la distribución, respecto a ser o no asimétrica

Coeficiente de curtosis

Mide el grado de aplastamiento o apuntamiento de la gráfica de la distribución.

Otras medidas de centralización

Desviacion media absoluta

$$DM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{40}$$

Media ponderada

Media geometrica

Media armonica

Sirve en diseño de experimentos par aproximar el numero de replicas en todo el expermineto cuando el diseño es desbalanceado.

Ejemplo en excel

Tabla de pesos de mujeres en una empresa

1. construcción de la tabla de frecuencias

- Definición de los intervalos: podemos utilizar la siguiente ecuación:

$$k = 1 + 3.322log_{10}(n) \tag{41}$$

Para esta muestra de 50 pesos de mujeres:

$$k = 1 + 3.322log_{10}(50) \tag{42}$$

$$k = 6.6 \approx 7 \tag{43}$$

Vamos a usar 7 intervalos, para saber la longitud dividimos el rango en el numero de intervalos.

Para hallar el rango podemos importar los datos a R:

```
tablaPesos <- read.table("TABLAPESOS.txt", header = T)
```

Y luego preguntar sobre el valor maximo y el minimo:

```
max(tablaPesos)
```

```
## [1] 72
```

```
min(tablaPesos)
```

```
## [1] 53
```

Podemos entonces calcular la longitud de cada intervalo:

```
(max(tablaPesos)-min(tablaPesos))/7
```

```
## [1] 2.714286
```

Obtenemos una longitud de intervalo de $2.71 \approx 3$

Los intervalos entonces serían:

```
##   intervalo valores
## 1         1   53 55
## 2         2   56 58
## 3         3   59 61
## 4         4   62 64
## 5         5   65 67
## 6         6   68 70
## 7         7   71 73
```

En excel:

- Usamos los limites superiores de los intervalos
  - Datos, analisis de datos, histograma, aceptar, seleccionar rango de entrada, rango de clases son los limites superiores. Al hacer esto, excel genera una tabla como la siguiente:

Clase	Frecuencia
55	2
58	5
61	9
64	15
67	12
70	5
73	2
y mayor...	0

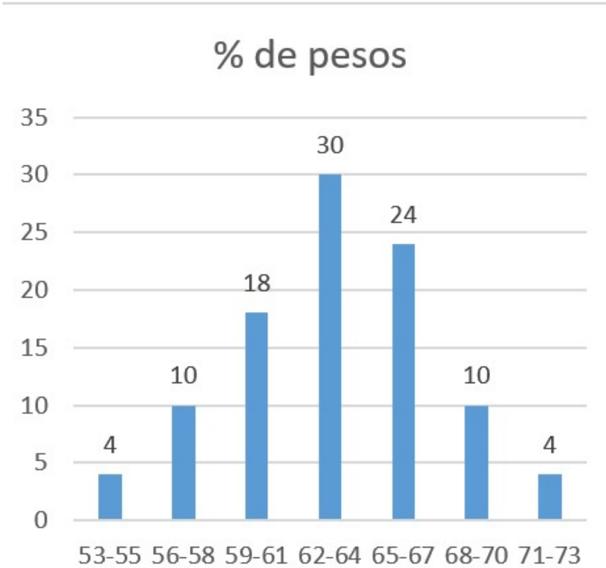
Figure 1: tabla de frecuencias en excel

Luego, a partir de esta tabla podemos calcular todas las frecuencias, la frecuencia absoluta (f_i), recuencia relativa (f_r), la frecuencia absoluta acumulada (F_i) y la frecuencia relativa acumulada (F_r):

	absoluta	relativa	abs. acumul.	rel. acumul.
intervalo	fi	fr	Fi	Fr
53-55	2	0.04	2	0.04
56-58	5	0.1	7	0.14
59-61	9	0.18	16	0.32
62-64	15	0.3	31	0.62
65-67	12	0.24	43	0.86
68-70	5	0.1	48	0.96
71-73	2	0.04	50	1

AF Andrés Felipe Beltrán Rodri...
El 62% de las mujeres de la muestra
pesan entre 53 y 64 kg
10/26/2021 6:51 AM

Luego a partir de esta tabla de frecuencias, utilizando las columnas de intervalo y % de frecuencia podemos construir un histograma como el siguiente:



Para hacer una operación análoga en R podemos crear los intervalos de la siguiente manera:

```
mins <- seq(53,71, by = 3)
maxs <- seq(55,73, by = 3)
```

Luego, podemos juntar las dos columnas de limites inferiores y superiores de intervalo en la tabla `TDF`:

```
TDF <- data.frame(min = mins,
                                             max = maxs )
TDF
```

```
##   min max
## 1  53  55
## 2  56  58
## 3  59  61
## 4  62  64
## 5  65  67
## 6  68  70
## 7  71  73
```

Luego podemos iterar a lo largo de las filas de `tablasDefrecuencias` buscando cuantos elementos de `tablaDePesos` están dentro del intervalo definido por cada fila:

```
for(i in 1:nrow(TDF)) {
  TDF$Freq[i] <-
      length(
          which(
mins[i] <=  tablaPesos & tablaPesos<= maxs[i]
}
```

Podemos luego calcular la frecuencia relativa dividiendo por el total de observaciones:

```
TDF$fr <-TDF$Freq/
    sum(TDF$Freq)

TDF
```

```
##   min max Freq   fr
## 1  53  55    2 0.04
## 2  56  58    5 0.10
## 3  59  61    9 0.18
## 4  62  64   15 0.30
## 5  65  67   12 0.24
## 6  68  70    5 0.10
## 7  71  73    2 0.04
```

Luego podemos calcular la frecuencia absoluta y relativa acumuladas para cada intervalo de manera ascendiente:

```
for(i in 1:nrow(TDF)){

TDF$Fi[i] <- sum(TDF$Freq[1:i])

}
```

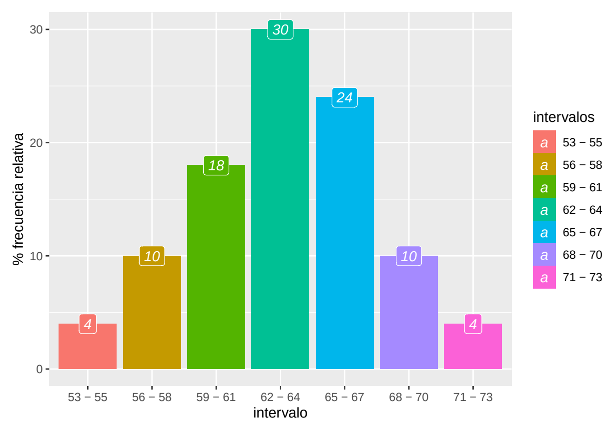
Podemos hacer lo mismo para la frecuencia relativa acumulada (F_r):

```
for(i in 1:nrow(TDF)){

TDF$Fr[i] <- sum(TDF$fr[1:i])

}
```

Una vez tenemos la tabla de frecuencias completa, podemos hacer la gráfica de frecuencias porcentuales:

```
library(ggplot2)
intervalos <- factor(paste(TDF$min,'-',TDF$max))

ggplot(TDF,
       aes(x = intervalos,
           y = fr*100,
           fill=intervalos,
           label = round(fr*100,2)
           )
       ) +
    geom_bar(stat="identity") +
  xlab("intervalo") +
    ylab('% frecuencia relativa')+
```

```
  geom_label(aes(fill = intervalos),
             colour = "white",
             fontface = "italic")
```



Calculo de medidas descriptivas

```
mean(tablaPesos$pesos)
```

```
## [1] 63.2
```

```
quantile(tablaPesos$pesos)
```

```
##   0%  25%  50%  75% 100%
## 53.0 61.0 63.5 66.0 72.0
```

El 75% de las mujeres pesan entre 53 y 66 kg. El 25% restante pesa entre 66 kg y 72 kg

- la varianza: Es importante calcular la varianza muestral y no la varianza poblacional, dado que no se puede saber la poblacional (σ^2) si no su estimador (s^2).
- Algunos paquetes hacen calculos con la varianza poblacional

En R: 'The denominator n - 1 is used which gives an unbiased estimator of the (co)variance for i.i.d. observations.'

```
var(tablaPesos$pesos)
```

```
## [1] 17.10204
```

Para calcular la varianza poblacional:

```
pesos <- tablaPesos$pesos
sum((pesos - mean(pesos))^2)/length(pesos)
```

```
## [1] 16.76
```

para el calculo de la desviacion estandar:

```
sd(pesos)
```

```
## [1] 4.135461
```

```
sqrt(var(pesos))
```

```
## [1] 4.135461
```

Interpretación:

Varianza: Tiene unidades al cuadrado. Tendriamos que tener otro grupo de comparación, con una medición similar.

Desviacion estandar: la desviacion de las observaciones respecto al promedio es de 4.14 unidades de masa (kg).

Para calcular otros estadísticos descriptivos podemos utilizar en excel:

- Datos > Análisis de datos > Estadística descriptiva > Se selecciona rango de entrada y de salida.

Se obtiene la siguiente tabla:

Dentro de esta tabla está el error típico

Error típico - error estándar

$$\frac{s}{\sqrt(n)} \tag{44}$$

En este caso:

```
sd(pesos)/sqrt(50)
```

```
## [1] 0.5848426
```

Se utiliza para inferencia, para intervalos de confianza. En diseño de experimentos sirve para el cálculo de tamaño de muestra. Se espera que no aumente el número de réplicas si no disminuye lo suficiente el error típico.

Columna1	
Media	63.2
Error típico	0.5848426
Mediana	63.5
Moda	64
Desviación estándar	4.1354614
Varianza de la muestra	17.102041
Curtosis	-0.0713723
Coeficiente de asimetría	-0.2023508
Rango	19
Mínimo	53
Máximo	72
Suma	3160
Cuenta	50

Figure 2: Cuadro de estadística descriptiva en excel

```
library(e1071)
kurtosis(pesos)
```

```
## [1] -0.2936688
```

```
skewness(pesos)
```

```
## [1] -0.1903716
```

La distribucion de pesos tiene una curtosis < 3, lo cual indica que es mas aplanada que una distribucion normal, o tiene hombros más pesados.

Además, tiene un coeficiende de asimetría cercano a cero, lo cual indica un ligero sesgo con cola hacia valores menores de peso.

```
range(pesos)
```

```
## [1] 53 72
```

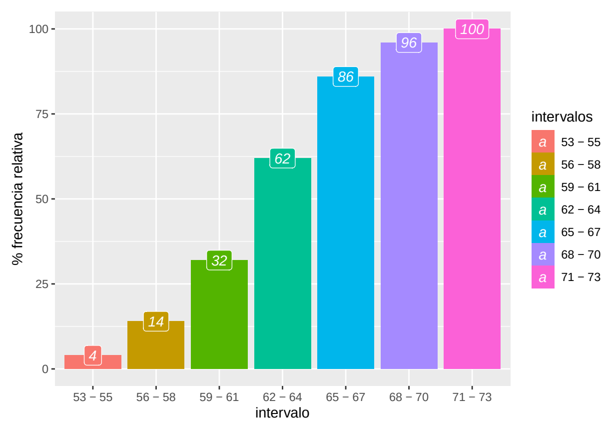
```
range(pesos)[2] - range(pesos)[1]
```

```
## [1] 19
```

La persona que mas peso tiene, tiene 19 kg mas que la persona de menos peso.

Tambien podemos graficar la frecuencia acumulada:

```
library(ggplot2)
intervalos <- factor(paste(TDF$min,'-',TDF$max))

ggplot(TDF,
       aes(x = intervalos,
           y = Fr*100,
           fill=intervalos,
           label = round(Fr*100,2)
           )
       ) +
    geom_bar(stat="identity") +
  xlab("intervalo") +
    ylab('% frecuencia relativa')+
  geom_label(aes(fill = intervalos),
             colour = "white",
             fontface = "italic")
```



el 86% de las mujeres pesan entre 53 y 67 kg.

Desviación media absoluta

Ejemplo de edades

Este ejemplo está en el código llamado `medidasdescriptivas.R`

Aun asi podemos

Proporción

Es similar al promedio, para variables de tipo cualitativo:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Con esta ecuación es posible calcular las proporciones.

$$x_i = \begin{cases} 1 & \text{Si cumple la condicion} \\ 0 & \text{Si no} \end{cases}$$

Para calcular la proporcion de varias variables cualitativas podemos utilizar la funcion `crosstable` del paquete `gmodels`.

Tambien se puede buscar asociacion entre las variables. Existen pruebas de asociación tales como la de *ji* cuadrado.

Asociacion

'La existencia de asociación entre dos variables indicaría que la distribución de los valores de una de las dos variables difiere en función de los valores de la otra'

La asociación entre 2 variables de diferente tipo se puede encontrar:

- El caso de dos variables categoricas # Prueba de independencia chi cuadrado

La prueba de independencia de chi cuadrado de termina si hay alguna asociacion entre variables categóricas (Si están asociadas o son independientes) Es una prueba no paramétrica.

Esta prueba utiliza una tabla de contingencia para analizar los datos. Ésta tabla es un arreglo en el cual los datos son clasificados de acuerdo a dos variables categóricas. Las categorías de una variable aparecen en las filas, y las categorías para la otra variable a aparecen en las columnas. Cada variable debe tener dos o más categorías o niveles. Cada celda refleja el conteo total de casos para un par especifico de categorías.

> **NOTA**
> Existen varias pruebas con el nombre 'prueba chi-cuadrado' además de la prueba de independencia de chi cuadrado. Es útil revisar el contexto de los datos y la pregunta de investigación para asegurar cual forma de la prueba chi cuadrado se está utilizando.

Usos de la prueba

la prueba de independencia de chi cuadrado se utiliza comúnmente para probar:

- Independencia estadística o asociación entre dos o más variables categóricas.

La prueba de independencia chi-cuadrado solo muede comparar variables categóricas. No puede hacer comparaciones entre varaibles continuas o entre variables continuas y categoricas. Además, La prueba de independencia chi-cuadrado solo *evalua asociaciones* entre variables categoricas, y no puede inferir nada sobre causalidad.

requerimientos de los datos

Los datos deben cumplir las siguientes condiciones:

1. Tener dos variables categóricas.
2. Dos o más categorías (grupos) o niveles para cada variable.
3. Independencia de las observaciones.
    - No existen relaciones entre los sujetos de cada grupo
    - Las variables categoricas no estan 'emparejadas' de manera.
4. Tamaño de muestra relativamente grande.
    - Se espera por lo menos una frecuencia de 1 en cada celda.
    - Se esperan frecuencias de por lo menos 5 en la mayoria (80%) de celdas.

Hipotesis

La hipótesis nula (H_0) y la hipótesis alternativa (H_1) de la prueba de independencia de la prueba de chi cuadrado pueden ser expresadas de dos maneras diferentes pero equivalentes:

$$H_0 : \text{La [variable 1] es independiente de la [variable 2]}$$

$$H_1 : \text{La [variable 1] no es independiente de la [variable 2]}$$

O

$$H_0 : \text{La [variable 1] No está asociada con [variable 2]}$$

$$H_1 : \text{La [variable 1] está asociada con [variable 2]}$$

Tabla de contingencia

para i = 1, ..; , k y j = 1, . . . , p se tiene que n_{ij} es el número de individuos o **frecuencia absoluta** que presentan a la vez las modalidades X_i e Y_i

Y / X	y_1	y_2	$\ldots$	y_j	$\ldots$	y_p	
x_1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2p}	$n_{2\bullet}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_i	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{ip}	$n_{i\bullet}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_k	n_{k1}	n_{k2}	$\ldots$	n_{kj}	$\ldots$	n_{kp}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet j}$	$\ldots$	$n_{\bullet p}$	$n_{\bullet\bullet}$

El numero de individuos que presentan la modalida x_i, es lo que llamamos *frecuencia absoluta marginal* de x_i y se representa como:

$$n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{ip} = \sum_{j=1}^{p} n_{ij}$$

De forma análoga se define la frecuencia absoluta marginal de la modalidad y_j como:

$$n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{kj} = \sum_{i=1}^{k} n_{ij}$$

El número total de elementos de la población)o de la muestra n lo obtenemos de cualquiera de las siguientes formas, que son equivalentes:

$$n_{\bullet\bullet} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{p} n_{\bullet j} = \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}$$

Estadistico de la prueba

El estadistico de la prueba de independencia de chi-cuadrado se denota como X^2 y se calcula de la siguiente manera:

$$X_c^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{\left(n_{ij} - \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} \cdot n_{\bullet j}}{n}}$$

El índice X^2 toma el valor de cero cuando dos variables son independientes.

Siendo mayor que cero cuando exista asociación entre ellas, tanto mayor cuanto más intensa sea esa correlación.

No tiene un límite máximo, lo cual supone una dificultad al nivel de interpretarlo.

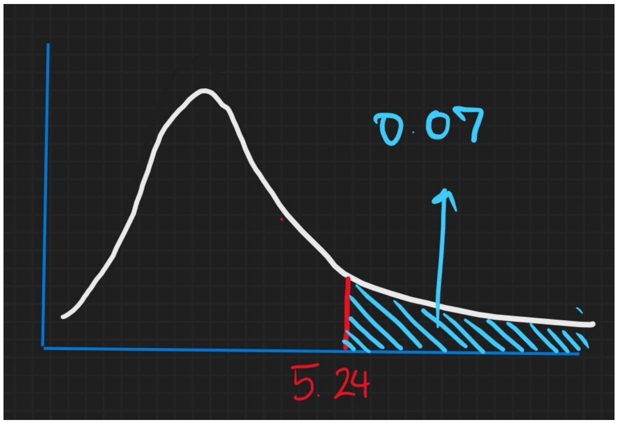
Otra forma de escribirlo es:

$$X_c^2 = \frac{\sum_{i=1}^{pk} (o_i - e_i)^2}{e_i^2} \quad \text{En donde} \quad e_i = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n_{\bullet\bullet}}$$

Ejemplo en excel

Para calcular en excel el chi cuadrado, podemos usar una tabla de contingencia ejemplo como la siguiente:

	A	B	C	D	E	F	G	H
1	USO DE CINTURON	NIVEL BAJO	NIVEL MEDIO	NIVEL ALTO	TOTAL			
2	SI	8	15	28	51			
3	NO	13	16	14	43			
4	TOTAL	21	31	42	94			
5								
6								
7	USO DE CINTURON	NIVEL BAJO	NIVEL MEDIO	NIVEL ALTO	TOTAL			
8	SI	11.39361702	16.81914894	22.787234	51			
9	NO	9.606382979	14.18085106	19.212766	43			
10	TOTAL	21	31	42	94			
11								
12								
13	chicuadrado calculado	=(B2-B8)^2/B8 + (B3-B9)^2/B9 + (C2-C8)^2/C8+ (C3-C9)^2/C9 + (D2-D8)^2/D8 + (D3-D9)^2/D9						
14		5.246551134						
15	chi critico	=INV.CHICUAD.CD(0.05,2)						
16		5.991464547						
17	p valor	=DISTR.CHICUAD.CD(B13,2)						
18	0.072564782	DISTR.CHICUAD.CD(x, grados_de_probabilidad)						



El 7 % de las veces que se obtiene un valor de chi cuadrado cumpliendose la hipótesis nula, se obtiene una suma de diferencias entre las frecuencias calculadas y las esperadas igual o mayor al experimental. No es posible rechazar la hipótesis nula. Es decir, la suma de las diferencias entre las frecuencias calculadas y las esperadas no es significativamente grande comparada con la suma de diferencias obtenida cuando la hipotesis nula se cumple.

Entonces no es posible rechazar la hipótesis nula (En la cual no existe asociación entre las variables). Podemos así concluir que no hay evidencia suficiente para sugerir una asociación entre el uso de cinturon y el nivel de escolaridad.

- Cuando la diferencia es pequeña, son independientes.
- Cuando la diferencia es grande, están asociadas.

Podemos realizar una análigo del análisis en R.

Primero, importamos la tabla:

```r
library(readxl)
datos <- read_excel("./datostabla.xlsx")
tabla <- table(datos)
chisq.test(tabla)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla
## X-squared = 34.41, df = 6, p-value = 5.607e-06
```

`Warning in chisq.test(tabla) : Chi-squared approximation may be incorrect` Aparece proque hay celdas con frecuencias de cero. Es aconsejable hacer la aproximación de fisher para muestras como esta.

Los grados de libertad de la prueba seria (filas-1)(columnas-1) $\neq 6$. Aqui quedan preguntas

Analogo manual en R

Queda pendiente para avanzar en tema, hacerlo todo manual. con ciclos debe salir rapido

ϕ # Coeficiente *phi* de pearson ()

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Puede oscilar entre 0 y $\sqrt{q-1}$ siendo q el mímimo número de modalidades entre las variables (niveles).

Si $\phi \leq 0.3$ nivel bajo de asociación Si $0.3 \leq \phi \leq 0.5$ nivel medio de asociación Si $\phi \geq 0.5$ nivel alto de asociación

Para las tablas de contingengia 2x2 oscila entre 0 y 1.

Coeficiente de contingencia de cramer

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

Donde

$$q = min[i, j]$$

Varia entre 0 y 1.

- 0 = independencia
- Cercania a 1, intensidad de la asociacion entre las variables.

uso en R

Para calcular el coeficiente de cramer podemos usar la funcion `cramerV` # Coeficiente de cohen

Variable categórica dicotómica (dos niveles [a,b]) y una variable cuantitativa Y, el índice de asociación *d de cohen* se obtiene a través de la siguiente expresión:

$$d = \frac{\bar{Y}_a - \bar{Y}_b}{s_Y}$$

En donde

- $\bar{Y}_a$: es la media de la variable cuantitativa Y en la categoría a.
- $\bar{Y}_b$: es la media de la variable cuantitativa Y en la categoría b.
- Desviación estándar de la variable Y.

Los valores que puede tomar d no est[an acotados a un rango.

Pueden ser tanto positivos como negativos.

- d = 0, las variables son independientes.
- mayor asociacion, mayor $|d|$

El caso de dos variables cuantitativas

Coeficiente de correlación de pearson

El coeficiente de correlación es la medida que describe que tan bine una variable es explicada por otra, y se calcula:

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{E(xy) - E(x) \cdot E(y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} x \cdot y \ - \ \frac{1}{n} \cdot \sum_{i=1}^{n} x \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} y}{s_x \cdot s_y}$$

$$-1 \leq r \leq 1$$

- La correlacion de pearson se recomienda para variables con distribucion normal, para variables no normales se recomienda el uso de la correlacion de spearman.

Ejemplo de correlacion

	A	B	C	D	E
1	variable	x	y		xy
2		2	0.05		0.1
3		3	0.09		0.27
4		4	0.11		0.44
5		5	0.13		0.65
6		6	0.17		1.02
7		7	0.2		1.4
8	**media**	**4.5**	**0.125**		0.64666667
9					
10		COV	=E8-(B8*C8)		
11			0.08417		
12					
13		CORR	=C10/(DESVEST.M(B2:B7)*DESVEST.M(C2:C7))		
14			0.82831		
17		CORR	=C10/(DESVEST.P(B2:B7)*DESVEST.P(C2:C7))		
18			0.99398	con varianza poblacional	

19	covarianza					
20					*x*	*y*
21				x	2.91666667	
22				y	0.08416667	0.00245833
23	Correlacion					
24			*x*	*y*		
25		x	1			
26		y	0.993977019	1		

La imagen anterior tiene un ejemplo de como calcular la correlacion utilizando tanto varianza muestral como poblacional. Los paquetes estadisticos usan la varianza poblacional.

##Coeficiende de correlacion de spearman

$$r_s = 1 - [6 \sum d_i^2 / (n^3 - n)]$$

Cuando se usa un estimador no parametrico, se le asigna a las observaciones rangos.

Ejemplo de coeficiente intelectual y horas de TV

Primero, se tiene el set de datos:

	A	B
1	CI	HORAS DE TV
2	86	0
3	97	20
4	99	28
5	100	50
6	100	28
7	103	28
8	106	7
9	110	17
10	113	7
11	113	12

Luego, se le asigna un orden o un rango a cada uno de los datos. En este caso en especial el coeficiente intelectual (CI) ya está organizado de manera ascendente, entonces podemos asignar un orden de la siguiente manera:

	A	B	C
1	CI	HORAS DE TV	ORDEN(ci)
2	86	0	1
3	97	20	2
4	99	28	3
5	100	50	4.5
6	100	28	4.5
7	103	28	6
8	106	7	7
9	110	17	8
10	113	7	9.5
11	113	12	9.5

En donde el numero 100 se repote 2 veces, entonces se reparte el rango 4 y 6 entre los dos. Por esto, los dos comparten el rango promedio que es 4,5. Lo musmo sucede con el numero 113 al final, que en las dos últimas posiciones comparte el rango 9.5, promedio de 9 y 10.

Lo mismo se puede hacer para las horas de TV:

	A	B	C	D
1	CI	HORAS DE TV	ORDEN(ci)	ORDEN(tv)
2	86	0	1	1
3	97	20	2	6
4	99	28	3	8
5	100	50	4.5	10
6	100	28	4.5	8
7	103	28	6	8
8	106	7	7	2.5
9	110	17	8	5
10	113	7	9.5	2.5
11	113	12	9.5	4

Una vez se tienen los rangos, se pueden calcular las diferencias entre los rangos:

	A	B	C	D	E	F
1	CI	HORAS DE TV	ORDEN(ci)	ORDEN(tv)		D
2	86	0	1	1		=C2-D2
3	97	20	2	6		-4
4	99	28	3	8		-5
5	100	50	4.5	10		-5.5
6	100	28	4.5	8		-3.5
7	103	28	6	8		-2
8	106	7	7	2.5		4.5
9	110	17	8	5		3
10	113	7	9.5	2.5		7
11	113	12	9.5	4		5.5

y las diferencias entre los rangos al cuadrado, asi como su suma:

	A	B	C	D	E	F	G
1	CI	HORAS DE TV	ORDEN(ci)	ORDEN(tv)		D	D^2
2	86	0	1	1		0	=F2^2
3	97	20	2	6		-4	16
4	99	28	3	8		-5	25
5	100	50	4.5	10		-5.5	30.25
6	100	28	4.5	8		-3.5	12.25
7	103	28	6	8		-2	4
8	106	7	7	2.5		4.5	20.25
9	110	17	8	5		3	9
10	113	7	9.5	2.5		7	49
11	113	12	9.5	4		5.5	30.25
12							196

Una vez tenemos este valor, podemos calcular el coeficiente de spearman como sigue:

$$r_s = 1 - [6[d_i^2/(n^3 - n)]]$$

$$r_s = 1 - [6[196/(10^3 - 10)]]$$

$$r_s = 0.1879$$

Para saber si la correlaciones significativa o no, se busca un valor tabulado para n=10 y $\alpha = 0.05$

Para este estimador se utiliza una prueba de hipótesis:

$$H_0 : r_s = 0 \text{ , No hay correlacion}$$

$$H_1 : r_s \neq 0 \text{ , hay correlacion}$$

Buscando en los valores tabulados encontramos que el valor critico para este tamaño de muestra y este nivel de significancia es:

$$0.648$$

α(2):	0.50	0.20	0.10	0.05
α(1):	0.25	0.10	0.05	0.025
n				
4	0.600	1.000	1.000	
5	0.500	0.800	0.900	1.000
6	0.371	0.657	0.829	0.886
7	0.321	0.571	0.714	0.786
8	0.310	0.524	0.643	0.738
9	0.267	0.483	0.600	0.700
10	0.248	0.455	0.564	0.648

dado que el valor experimental es menor al tabulado, no se obtiene un valor suficientemente grande respecto al obtenido cuando la hipotesis nula es verdadera. Entonces, no es posible rechazarla. Entonces, no hay evidencia suficiente para decir que hay correlacion entre las variables, o que $r_s \neq 0$.

Ejemplo en R

primero creamos las dos variables:

```
x<-c(2,3,4,5,6,7)
y<-c(0.05,0.09,0.11,0.13,0.17,0.2)
```

Podemos calcular la covarianza y correlacion con las funciones `cov` y `cor`, y seleccionando el método con el argumento `method`

```
covx<-cov(x, y, method = c("pearson"))
corx<-cor(x, y, method = c("pearson"))
corx<-cor(x, y, method = c("spearman"))
corx<-cor(x, y, method = c("kendall"))
```

A su vez, es posible realizar pruebas de hipótesis para la significancia de cada una de las corelaciones:

```
cor.test(x, y,
         alternative = c("two.sided",
                         "less",  # ver si es menor a cer
                         "greater"), # si es mayor a cero
         method = c("pearson", "kendall", "spearman"),
         exact = NULL, conf.level = 0.95, continuity = FAI
```

```
cor.test(x, y, method="kendall","two.sided")
```

```
##
##  Kendall's rank correlation tau
##
## data:  x and y
## T = 15, p-value = 0.002778
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
## tau
##   1
```

```
cor.test(x, y, method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  x and y
## S = 0, p-value = 0.002778
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
##   1
```

```r
cor.test(x, y, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 18.14, df = 4, p-value = 5.431e-05
## alternative hypothesis: true correlation i
## 95 percent confidence interval:
##  0.9435619 0.9993718
## sample estimates:
##      cor
## 0.993977
```

Ejemplo de correlacion en R - pearson

Primero creamos las variables

```r
# ejemplo Pearson
x<-c(2,3,4,5,6,7)
y<-c(0.05,0.09,0.11,0.13,0.17,0.2)
```

Luego calculamos la correlacion de pearson:

```r
corp<-cor(x, y, method = c("pearson"))
corp
```

```
## [1] 0.993977
```

y la covarianza:

```r
covp<-cov(x, y, method = c("pearson")) # covarianza Pearson
covp
```

```
## [1] 0.101
```

y tambien podemos realizar la prueba de hipotesis

```r
cor.test(x, y, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 18.14, df = 4, p-value = 5.431e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9435619 0.9993718
## sample estimates:
##      cor
## 0.993977
```

- Calculo manual

```r
sumXi <- sum(x)
sumXi2 <- sum(x^2)

sumYi <- sum(y)
sumYi2 <- sum(y^2)

sumXiYi <- sum(x*y)
r <- sumXiYi/sqrt(sumXi2*sumYi2)
r
```

```
## [1] 0.9991011
```

Ejemplo de correlacion en R - spearman

```r
CI<-c(106,86,100,100,99,103,97,113,113,110)
horastv<-c(7,0,28,50,28,28,20,12,7,17)
cor.test(CI, horastv, method="spearman","two.sided")
```

```
## Warning in cor.test.default(CI, horastv, method = "spea
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  CI and horastv
## S = 200.25, p-value = 0.5534
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2136315
```

```r
cors<-cor(CI, horastv, method = "spearman")
```

Coeficiente tau de kendall

$$\tau = \frac{S_a - S_b}{n(n-1)/2}$$

En donde: τ = estadistico de kendall n = numero de casos o tamaño de muestra S_a = sumatoria de los rangos mas altos S_b = sumatoria de los rangos mas bajos

Ejemplo correlacion de kendall

En una evaluacion de los jugadores delanteros de futbol en un pais, hay 9 de ellos catalogados como mas intensos para marcar goles. para analizar esta intensidad durante un periodo de una temporatda se registrio sistematicamente el grado de intensidad de cada uno de estos delanteros tanto en juegos a nivel nacional (NP = puntaje nacional) como a nivel internacional (IP = puntaje en juegos internacionales) Ademas, se registrarion los rangos a nivel nacional (NR = rango nacional) y a nivel internacional (IR = rango a nivel internacional)

tenemos entonces la tabla con los rangos:

	A	B	C	D	E
1	jugador	NP	IP	RN	RI
2	1	84	60	1	4
3	2	80	64	2	2
4	3	78	71	3	1
5	4	76	61	4	3
6	5	70	58	5	5
7	6	64	57	6	6
8	7	62	54	7	8
9	8	50	55	8	7
10	9	47	52	9	9

Tambien podemos crear la tabla en R:

```
jugador <- 1:9
NP <- c(84,80,78,76,70,64,62,50,47)
IP <- c(60,64,71,61,58,57,54,55,52)
RN <- 1:9
RI <- c(4,2,1,3,5,6,8,7,9)

jugadores <- data.frame(
  jugador = jugador,
  NP = NP,
  IP = IP,
  RN = RN,
  RI = RI
)
head(jugadores)
```

```
##   jugador NP IP RN RI
## 1       1 84 60  1  4
## 2       2 80 64  2  2
## 3       3 78 71  3  1
## 4       4 76 61  4  3
## 5       5 70 58  5  5
## 6       6 64 57  6  6
```

para hallar los valores S_a y S_b podemos hacer lo siguiente

calculo de S_a y S_b

- *manual:* para el rango de interes, se cuentan cuantos rangos son mayores al rango en cuestion:

	A	B	C	D	E
1	jugador	NP	IP	RN	RI
2	1	84	60	1	4
3	2	80	64	2	2
4	3	78	71	3	1
5	4	76	61	4	3
6	5	70	58	5	5
7	6	64	57	6	6
8	7	62	54	7	8
9	8	50	55	8	7
10	9	47	52	9	9

En este caso serian 5, $S_{a,1} = 5$.

- en R:

```
which(jugadores$RI[1:9] > jugadores$RI[1])
```

```
## [1] 5 6 7 8 9
```

```
length(which(jugadores$RI[1:9] > jugadores$RI[1]))
```

```
## [1] 5
```

Podemos iterar esta operacion y crear la columna S_a:

```
for(i in 1:9){
 jugadores$Sa[i] <-
   length(
   which(
     jugadores$RI[i:9] > jugadores$RI[i]))
}
```

de manera analoga S_b:

```
for(i in 1:9){
 jugadores$Sb[i] <-
   length(
     which(
       jugadores$RI[i:9] < jugadores$RI[i]))
}
jugadores
```

```
##   jugador NP IP RN RI Sa Sb
## 1       1 84 60  1  4  5  3
## 2       2 80 64  2  2  6  1
## 3       3 78 71  3  1  6  0
## 4       4 76 61  4  3  5  0
## 5       5 70 58  5  5  4  0
## 6       6 64 57  6  6  3  0
## 7       7 62 54  7  8  1  1
## 8       8 50 55  8  7  1  0
## 9       9 47 52  9  9  0  0
```

y luego hayamos las sumatorias:

- para S_a:

```
sumSa <- sum(jugadores$Sa)
sumSa
```

```
## [1] 31
```

- para S_b:

```
sumSb <- sum(jugadores$Sb)
sumSb
```

```
## [1] 5
```

Una vez tenemos las sumatorias podemos calcular el estadistico:

$$\tau = \frac{S_a - S_b}{n(n-1)/2}$$

```
n <- 9
tauJugadores <- (sumSa-sumSb)/((n*(n-1))/2)
tauJugadores
```

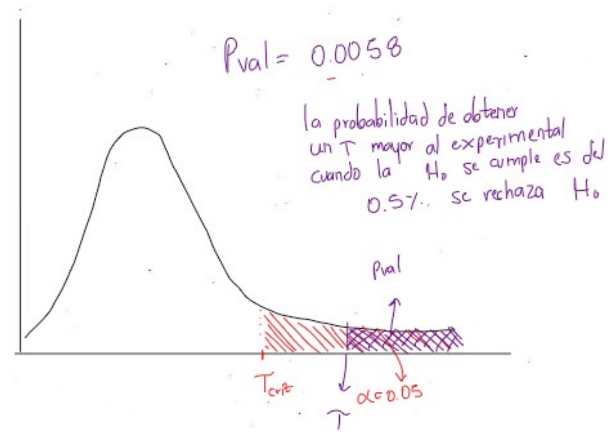
```
## [1] 0.7222222
```

el coeficiente tao de kendall es de 0.72

Calculo de tau utilizando funciones de R

```
cor.test(NP, IP, method="kendall","two.sided")
```

```
##
##  Kendall's rank correlation tau
##
## data:  NP and IP
## T = 31, p-value = 0.005886
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.7222222
```



dado que el p-value es menor al nivel de 0.05, se rechaza la hipótesis nula. Ya que solo un 0.5% de las veces de casos en los que la nula se cumple, es posible obtener un valor de tau igual o mayor al obtenido en el experimento. Es decir, es significativa la correlacion entre el puntaje nacional y el puntaje internacional, presentando un 72.2% de correlacion.

Coeficiente de kendal para concordancia

Es un promedio de un grupo de coeficientes de spearman. Es una medida del grado de acuerdo o concordancia entre m conjuntos de n rangos. varía entre 0 y 1.

Por ejemplo, para n objetos evaluados por un grupo de me jueces, es posible revisar el grado de concordancia en la evaluacion de los objetos entre los jueces.

$$W = \frac{12 * \Sigma D^2}{m^2 * n * (n^2 - 1)}$$

En donde:

$$D = \sum R - (\sum R/n)$$

$$m = \text{rango de evaluadores}$$

$$n = \text{numero de objetos}$$

$$\sum R = \text{suma de rangos}$$

las hipotesis serian:

$$
\begin{aligned}
H_0 &: \text{no hay concordancia}, \omega = 0 \\
H_1 &: \text{ hay concordancia}, \omega \neq 0
\end{aligned}
$$