METODOS NO PARAMETRICOS

Autora: María del Carmen Pría Barros. LA HABANA, 2001.

CAPITULO 1: GENERALIDADES.

1.1 INTRODUCCION A LOS METODOS NO PARAMÉTRICOS.

Cuando se trabaja en las ciencias de la salud surge con frecuencia la necesidad de tomar decisiones con relación a diferentes problemas de investigación, generalmente no contamos con el tiempo ni con los recursos para estudiar a toda la población, por esto necesitamos seleccionar una muestra aleatoria a partir de la cual tomar decisiones sobre la población.

Las Pruebas de Hipótesis tienen la finalidad de ayudar al investigador a tomar decisiones sobre la población basándose en el análisis de una muestra aleatoria de la misma, las decisiones que debemos tomar son en relación con determinadas características de la población, denominados parámetros, que necesitamos conocer, para resolver un problema.

Es por esto que hay que establecer un procedimiento objetivo que permita, sobre la base de la información muestral obtenida, tomar una decisión sobre los parámetros de la población, lo que determinará cierto grado de incertidumbre asociada a la decisión. Este procedimiento se conoce como **Prueba de Hipótesis**.

Para realizar una Prueba de Hipótesis deben considerarse una serie de pasos o etapas que se relacionan a continuación:

1. Formulación de Hipótesis:

Una **Hipótesis Estadística** es una afirmación que se hace con relación a un parámetro o a la forma de distribución de una variable. En una Prueba de Hipótesis se consideran 2 tipos:

- **Hipótesis Nula, H_o**, es una hipótesis de diferencias nulas que se formula con la intención de rechazarla, en la práctica se expresa en la misma lo contrario a lo que se desea probar.
- **Hipótesis Alternativa, H_1**, es la hipótesis en la que se formula lo que se desea probar.

Hay que aclarar que mediante éste procedimiento lo único que se hace es evidenciar si los datos apoyan o no la hipótesis estadística, no se demuestra la misma.

2. **Elección de la Prueba Estadística:**

Existen muchas Pruebas Estadísticas para resolver los diferentes problemas que surgen en la practica diaria, pero siempre tendremos que determinar ante un problema dado la prueba especifica a utilizar tomando en consideración:

- Las hipótesis a probar, por ejemplo si se desea comparar las medias de 2 poblaciones, existen pruebas estadísticas especificas para resolver éste problema.
- El tipo de variable, ya que existen pruebas para utilizar con variables nominales, ordinales y cuantitativas.
- El Numero y tipo de muestras con que se trabajara, ya que dependiendo del tipo de muestras, independientes o equiparadas estas ultimas serán explicadas en el capitulo donde se utilizan, y del numero de muestras con que se trabaja se determinara la prueba a utilizar.

3. Estadística de Prueba:

La Estadística de Prueba es una expresión matemática que tiene una distribución conocida y se calcula a partir de los datos muestrales, permitiendo así determinar de acuerdo a su valor si se rechaza o no la Hipótesis Nula.

La distribución muestral de una Estadística de Prueba es la distribución conforme a H_o de todos los posibles valores que la misma puede tomar cuando es calculada con muestras de igual tamaño tomadas al azar.

4. <u>Nivel de Significación</u>: Nosotros habíamos formulado una Hipótesis Nula con la finalidad de rechazarla, sin embargo al tomar esta decisión podríamos equivocarnos, igual que al aceptar una Hipótesis Alternativa falsa, por tanto deben definirse los 2 tipo de errores anteriores:

- **Error de Tipo I**: Es el error que se comete al rechazar una Hipótesis Nula que es cierta. La probabilidad de cometer éste tipo de error se conoce como α y es el **Nivel de Significación de la Prueba**.
- **Error de Tipo II**: Es el error que se comete al aceptar una Hipótesis Nula que es falsa. *La probabilidad de cometer éste tipo de error se conoce como B.*
- **Potencia de una prueba** es entonces la probabilidad de rechazar H_o siendo falsa, o sea, 1 – B

Ambos tipo de errores están relacionados de forma inversa, o sea a medida que aumenta la probabilidad de cometer el Error de un tipo disminuye la probabilidad de cometer el otro tipo de Error, sin embargo la única forma de disminuirlos simultáneamente es aumentando el tamaño de la muestra.

Un concepto importante es el de **Eficiencia Relativa de una Prueba de Hipótesis A con relación a otra Prueba B**, que es el numero de observaciones requeridas por la Prueba A para tener la misma potencia que la Prueba B cuando se dócima la misma hipótesis con el mismo nivel de significación.

5. <u>Región de Rechazo</u>: De acuerdo al nivel de significación que estamos dispuestos a asumir en una prueba, se va a dividir la distribución muestral de la Estadística de Prueba en 2 regiones, una llamada **Región de Aceptación** que es aquella en la que si cae el valor de la Estadística de Prueba no se rechaza H_o, ya que de hacerlo estaría asumiendo un riesgo mayor de Error de tipo I que el prefijado.

La otra región, llamada **Región de Rechazo** es aquella en la que si cae el valor de la Estadística de Prueba, rechazaría H_o ya que de hacerlo estaría asumiendo un riesgo de cometer el Error de tipo I menor que el prefijado.

Una vez definidas ambas regiones, solo queda ubicar el valor obtenido por la Estadística de Prueba en una de ellas y determinar la decisión de rechazar o no.

Debe señalarse que las Pruebas de Hipótesis pueden ser de 1 o de 2 colas, de acuerdo a la región de rechazo. Si la región de rechazo se encuentra a ambos lados de la distribución, entonces estamos ante una **Prueba de 2 colas o Bilateral**, mientras que si la región de rechazo se encuentra a un solo lado estamos ante una **Prueba de 1 cola o Unilateral**.

6. <u>Interpretación</u>:

Una vez que tomamos la decisión estadística de rechazar o no, debemos darle respuesta al problema planteado en términos clínicos, administrativos, epidemiológicos, etc. ,o sea debemos tomar la decisión adecuada en términos de solucionar el problema planteado considerando no solo los resultados de la prueba de hipótesis sino todos los elementos que nos permitan llegar a una acertada solución del problema.

Debe recordarse que una **Variable** es cualquier característica de la población que nos interesa estudiar y que puede asumir diferentes comportamientos, valores o grados de intensidad entre los diferentes elementos, individuos o unidades de análisis que conforman a esa población.

Las variables **por su naturaleza** pueden ser **cualitativas o cuantitativas**, las primeras son aquellas en que las diferencias entre los elementos estudiados esta dada por la presencia o no de una cualidad o atributo no medible en términos numéricos. Las cuantitativas son variables en las que las diferencias existentes entre los elementos de la población son medibles numéricamente.

Una vez que tenemos definidas las variables que se van a estudiar hay que determinar las escalas que se utilizaran para medir esas variables, existen diferentes tipos de **escalas de medición** que se relacionan a continuación:

1. **Escala Nominal**: Es el nivel de medición más elemental que se usa para identificar a que categoría o clase de una variable pertenecen los diferentes individuos estudiados. Se utiliza para trabajar con variables cualitativas. Ejemplo del uso de esta escala es el caso de la variable sexo, que incluye 2 posibilidades para clasificar los individuos: femenino o masculino, estas serian las 2 categorías que conforman a esta variable.

En éste tipo de escala **las categorías o clases son equivalentes** o sea la **relación** que se establece entre las mismas es **de igualdad**. Este tipo de escala puede ser **dicotómica** , cuando solo hay 2 posibilidades de clasificación como en el ejemplo del sexo y **politómica** cuando hay 3 o mas posibilidades para clasificar a los individuos , como seria en el caso de la variable estado civil, en la que los individuos pueden ser clasificados como solteros, casados ,divorciados o viudos. **Este tipo de escala permite resumir la información con frecuencias absolutas y porcentajes, además utiliza como medida de tendencia central la moda.**

2. **Escala Ordinal:** Es el nivel de medición que sigue al Nominal y que también se usa para identificar a que categoría o clase de una variable pertenecen los diferentes individuos estudiados, sin embargo en éste caso a diferencia del anterior se establece determinado orden o intensidad de la característica que se mide , en éste caso existe una relación de subordinación entre las categorías que conforman la variable. **Se utiliza básicamente para trabajar con variables cualitativas en la cual puede medirse la intensidad de la característica que se estudia, estas variables se conocen también como cuasi cuantitativas.**

Ejemplo del uso de esta escala es el caso de la variable nivel de escolaridad, que puede ser primario, medio o superior, en éste caso se sabe que la persona que tiene nivel primario tiene menor escolaridad que la que tiene nivel medio y esta tiene menor escolaridad que la que tiene nivel superior. Otro ejemplo pudiera ser la variable grado de escolaridad, donde se consideran los grados primero (1^{o}) ,segundo(2^{o}),etc., aunque se piense que esta variable es cuantitativa por estar representada por un numero , en realidad es ordinal pues en éste caso la relación que se establece entre las categorías que conforman la variable solo es de subordinación. No podemos medir la distancia que existe en términos de conocimientos entre un niño que esta en sexto grado y otro que esta en tercero, y por esto no podemos afirmar que el que esta en sexto grado sabe el doble que el niño que esta en tercero, pues esta medición se establece acorde al cumplimiento de los objetivos docentes que tiene que alcanzar el niño para aprobar el grado, por lo que esta medición es cualitativa.

Este tipo de escala permite resumir la información con frecuencias absolutas y porcentajes, además utiliza como medidas de tendencia central la mediana y la moda.

3. **Escala de Intervalo**: Cuando una escala tiene las **características de una escala ordinal pero además permite conocer la distancia entre todas las unidades de análisis objeto de medición**, o sea si podemos conocer las distancias o el intervalo que existe entre 2 mediciones cualesquiera estamos en presencia de una escala de intervalo. **Es una escala cuantitativa,** por lo que se utiliza con variables de éste tipo.

Esta escala se caracteriza por asignar un numero real a todos los pares de objetos de un conjunto ordenado, de forma que mantienen las relaciones de subordinación entre las unidades de análisis y además permite determinar la magnitud de esta. Por ejemplo podemos decir que la diferencia que hay entre 10 y 20 es la misma que la que hay entre 50 y 60 pues en éste caso tenemos una distancia unitaria y un punto cero, ambos arbitrarios. En esta escala **el punto cero y las unidades de medida son arbitrarios**, ejemplo de éste tipo de

escala son las que se usan para medir la temperatura en grados Centígrados o en grados Fahrenheit, en las que el punto cero es diferente para ambas y no implica ausencia de calor. **Este tipo de escala utiliza como medidas de tendencia central la mediana, la moda y la media aritmética**.

4. **<u>Escala de Razón</u>:** Cuando una escala **cumple todas las características de una escala de intervalo y además tiene cero real en su origen e independencia de la unidad de medida** es entonces una escala de razón.

Constituye el **nivel mas alto de medición,** se utiliza con variables cuantitativas. Ejemplo de mediciones en esta escala es cuando se mide el peso, la altura y la longitud.

Este tipo de escala utiliza como medidas de tendencia central la mediana, la moda, la media aritmética y la media geométrica.

Debe destacarse que las variables que brindan mas información son las cuantitativas y estas pueden utilizar diferentes escalas de medición según el interés del investigador, sin embargo no ocurre lo mismo con las cualitativas que no pueden abrirse a niveles de medición mayores.

Por ejemplo cuando tenemos una variable cuantitativa como el peso al nacer debe utilizarse una escala de razón, que es la que brinda mas información, sin embargo de acuerdo a los intereses del investigador podemos usar una escala de intervalo (1000 – 2499 gr, 2500 – 3999 gr , 4000 gr y más) o una ordinal

(bajo peso, peso normal , sobrepeso). También solo pudiera interesarle al investigador usar una escala nominal dicotómica (bajo Peso: sí o no).Observe que si tenemos una variable como sexo es imposible abrir mas esta información, por lo que en éste caso solo se podrá usar siempre una escala nominal dicotómica.

Cuando estamos trabajando con variables que tienen escalas de medición **Nominal y Ordinal** solo podremos utilizar para realizar el análisis estadístico **Métodos no Paramétricos o de Libre distribución**, ya que las Estadísticas Clásicas no permiten por sus restricciones usarlas en éste caso. Sin embargo cuando trabajamos con variables medidas con escalas de **Intervalo y de Razón** podemos utilizar **cualquier tipo de estadísticas**, siempre verificando previamente que se cumplan los supuestos requeridos por las Estadísticas Clásicas.

La mayoría de los métodos estadísticos tienen 3 aspectos comunes:

1. Se supone que la variable con que trabajamos tiene una distribución teórica conocida especifica (Normal, t de Student, F de Fisher, etc.)
2. Se asume la igualdad de las varianzas de las variables con que estamos trabajando.
3. Se refieren a la verificación de hipótesis o a la estimación de parámetros de una distribución teórica conocida.

Las distribuciones teóricas conocidas están definidas por uno o dos parámetros poblacionales, por esto las pruebas de hipótesis clásicas se conocen como Pruebas Paramétricas ya que para hacer un uso eficiente de estas pruebas se requieren generalmente de 2 condiciones:

1. La normalidad de la población.
2. La igualdad de las varianzas.

<u>Ejemplos de lo anterior son las situaciones siguientes</u>:

1. Al estimar un intervalo de confianza para la media aritmética, se supone que la variable estudiada tiene una distribución normal , lo que permite determinar el percentil de la distribución Normal (0,1) de acuerdo al nivel de confiabilidad con que se éste trabajando.
2. Si se desea docimar $H_{o:}\ M_1 = M_{2,}$ o sea la igualdad de las medias de 2 poblaciones, tenemos que suponer la normalidad de la variable para poder tomar la decisión al comparar el valor del estadístico de prueba con el percentil correspondiente de la distribución Normal(0,1).
3. Si se necesita realizar un Análisis de la Varianza hay que suponer normalidad de la variable e igualdad de varianzas, para que se cumpla el modelo bajo la hipótesis nula.

A diario se le presentan a los investigadores situaciones donde no pueden asumirse los supuestos requeridos por desconocerse la distribución de la variable estudiada, o bien por ser la muestra muy pequeña y aunque conozcamos la distribución de la variable y sea valido teóricamente hacer supuestos sobre la distribución y sus parámetros, no es razonable utilizar una prueba paramétrica ya que hay poca certeza de que se cumplan en éste conjunto de datos. También puede ocurrir que la variable no sea continua por lo que no se cumplen las restricciones anteriores.

En cualquiera de los casos anteriores hay que buscar técnicas alternativas que permitan darle solución a estas situaciones de forma eficiente, surge así la necesidad de desarrollar un arsenal de técnicas estadísticas que tengan un mínimo de restricciones.

Las pruebas de hipótesis que no hacen afirmaciones acerca de los parámetros de una población se conocen como **Métodos no Paramétricos**, mientras que aquellas pruebas que no hacen ningún supuesto acerca de la distribución de la población de la cual se extraen las muestras se llaman **Métodos de Libre Distribución.**

Aunque tienen significados diferentes generalmente todas estas pruebas se conocen como

Métodos no Paramétricos.

Estos métodos surgen a finales del siglo 19, en 1900 Pearson desarrolla la prueba X^2 de Independencia y la prueba para la Bondad de Ajuste. Sin embargo su verdadero inicio se establece en 1936 con la dócima del Coeficiente de Correlación de Rangos de Spearman que fue popularizado por Hotelling y Pabst, finalmente Wilcoxon propuso una Prueba para la comparación de 2 muestras independientes cuando se trabaja con una variable cuantitativa que se destaco por su simplicidad y excelentes resultados aun en aquellos casos en que se conoce que la distribución de la variable es normal. A partir de éste momento estos métodos tuvieron un gran desarrollo debido a que siempre se desea tener pruebas de hipótesis con pocas restricciones impuestas por los supuestos subyacentes.

Según Wonnacott existen 2 indicaciones para preferir las Pruebas no Paramétricas:

1. Cuando la prueba clásica correspondiente no es valida.
2. En aplicaciones donde la prueba clásica es razonablemente valida, pero un estimador no paramétrico puede ser más eficiente.

Entre sus **ventajas** se encuentran las siguientes:

1. Tienen mayor eficiencia que los Métodos Paramétricos en distribuciones asimétricas, o sea cuando hay valores atípicos o datos aberrantes.
2. Tienen validez en el sentido que su nivel de confiabilidad es realmente el especificado en la mayoría de las pruebas.
3. Generalmente son de computo mucho más fácil que las técnicas de las estadísticas clásicas.
4. Son aplicables en situaciones donde los procedimientos clásicos no son aplicables, como se explico anteriormente.
5. Aun cuando se cumplan los requisitos para realizar una prueba paramétrica, si la muestra es pequeña la eficiencia relativa de la prueba no paramétrica es alta.

Entre sus **desventajas** se citan las siguientes:

1. Fundamentalmente cuando las muestras son muy grandes las Pruebas no Paramétricas tienen una eficiencia relativa baja con relación a las Paramétricas cuando se cumplen los supuestos.
2. Las hipótesis que se plantean en las pruebas no Paramétricas son menos precisas, lo que hace que la interpretación de los resultados sea mas ambigua.
3. Su aplicación en muestras grandes se hace muy laboriosa.

4. Para un problema particular pueden existir varias pruebas, por lo que en ocasiones se hace difícil seleccionar la mejor.

En sentido general éste tipo de pruebas deben utilizarse cuando no se cumplan los supuestos para hacer una Prueba Paramétrica y cuando se duda sobre el cumplimiento de los supuestos en una muestra pequeña.

El concepto de **Potencia-Eficiencia** de una prueba se refiere al aumento porcentual del tamaño muestral que se necesita para obtener los mismos resultados con una prueba X, no tan poderosa como otra que es la prueba Y, que es la más potente que se conoce para resolver un problema especifico cuando se cumplen los supuestos para realizar la misma.

La **Potencia –Eficiencia de la prueba X** puede calcularse de la forma siguiente:

Potencia –Eficiencia de la prueba X= $\frac{N_Y}{N_X} x100$

Este concepto usted lo vera reflejado en éste material al final de cada prueba relacionando la Prueba no Paramétrica estudiada con la Paramétrica homologa, o sea aquella que existe para resolver el mismo problema en las estadísticas clásicas

1.2 EJERCICIOS.

1. Explique las diferencias que hay entre los conceptos siguientes:

a) Estadísticas Clásicas y Estadísticas no Paramétricas.
b) Estadísticas no Paramétricas y Estadísticas de Libre Distribución.

2. Señale y ejemplifique las ventajas y desventajas de los Métodos no Paramétricos o de Libre Distribución.

3. Mencione y ejemplifique los tipos de variables y escalas de medición que conoce y con relación a esto diga:

a) Relación entre tipo de variables y de escala de medición.
b) Relación entre tipo de análisis estadístico y de variables.
c) Por qué debe utilizarse el mayor nivel de información cuando realizamos un análisis estadístico.

1.3 BIBLIOGRAFIA.

1. Daniel, W.: Bioestadistica: base para el análisis de las ciencias de la salud. Editorial UTEHA, México, 1997.
2. Cho Ya Lun: Análisis Estadístico. Editorial Interamericana, México, 1972.
3. Hollander, M y D. A. Wolfe: Nonparametrics Statistical Methods. John Wiley and Sons, New York, 1973.
4. Wonnacott, T. H. y R. J. Wonnacott: Introductory Statistics. John Wiley and Sons, New York, 1973.
5. Siegel, S.: Análisis Experimental no paramétrico. Instituto Cubano del Libro, Edición Revolucionaria, La Habana, 1972.
6. Bradley, J. V.: Distribution Free Statistical Test. Prentice Hall, New Jersey, 1968.
7. López Pardo, Cándido: Notas de Clase. Curso La Estadística no paramétrica en la investigación contemporánea. Facultad de Economía. Universidad de la Habana, 1990.
8. Astrain Rodríguez, M. Elena: Análisis Estadístico. Escuela Nacional de Salud, La Habana, 2001.

CAPITULO 2: MÉTODOS DE ANÁLISIS PARA VARIABLES CUALITATIVAS EN MUESTRAS INDEPENDIENTES.

2.1 INTRODUCCION.

Es frecuente en la práctica de la investigación biomédica trabajar con variables cualitativas tales como los factores de riesgo de una enfermedad, la presencia de diferentes diagnósticos médicos, la presencia o ausencia de una enfermedad, de un signo o síntoma, la necesidad de evaluar la efectividad de diferentes tratamientos cuando la variable respuesta se expresa en relación con la desaparición de determinados signos y síntomas de la enfermedad. Muchas veces se necesita relacionar estas variables con otras del mismo tipo para determinar si existe una asociación estadística entre las mismas.

También con frecuencia es necesario evaluar determinados factores de riesgo de una enfermedad, para lo cual hay que determinar si la proporción de personas con un posible factor de riesgo de una enfermedad es mayor entre los que tienen la enfermedad que entre aquellas personas que no la tienen, para lo cual nos interesaría determinar si estas diferencias son estadísticamente significativas.

Cuando estamos ante un problema en el cual necesitamos estudiar la relación existente entre 2 variables cualitativas, la información se resume en una **Tabla de Contingencia** que no es mas que una tabla estadística en la que se entrecruza la información de 2 ó más variables cualitativas o de mas nivel, pero con escala Nominal, Ordinal o de Intervalo, que se utiliza para estudiar la relación existente entre las mismas.

La Tabla de Contingencia que se utiliza con más frecuencia es la que se utiliza para estudiar la relación entre 2 variables cualitativas, éste tipo de tabla de contingencia tiene la configuración siguiente:

VARIABLE 1 SEGÚN NIVELES DE CLASIFICACION	VARIABLE 2 SEGUN NIVELES DE CLASIFICACION			
	2	.	C	TOTAL
1	n_{12}		n_{1c}	$n_{1.}$
2	n_{22}	.	n_{2c}	$n_{2.}$
.				
R	n_{r2}	.	n_{rc}	$n_{r.}$
TOTAL	$n_{.2}$	.	$n_{.c}$	$n_{..}$

Esta es una Tabla de Contingencia de R x C, una tabla de éste tipo tiene **(R – 1) (C- 1) grados de libertad, que son las celdas que pueden llenarse libremente sin depender del resto de los valores de la tabla.**

Por ejemplo la primera fila tiene C - 1 celdas que pueden llenarse libremente, sin embargo la última tiene un valor determinado por el total de la fila menos el valor de la suma de las observaciones de las celdas de esa fila. Lo mismo sucede con las columnas donde la primera columna tiene R - 1 celdas que pueden llenarse libremente, sin embargo la ultima tiene un valor determinado por el total de esa columna menos el valor de la suma de las observaciones de las celdas de esa la columna. La tabla anterior se puede generar a partir de 2 procedimientos:

1. Se selecciona **una muestra de tamaño n..** y se clasifican los individuos según las 2 variables anteriores, la variable 1 con C categorías y la 2 con R categorías. En éste caso se fija el tamaño total de la muestra sin embargo la distribución de los individuos que constituyen la muestra en relación a las 2 variables es aleatoria. La tabla que se genera esta constituida por R x C celdas y cada una de ellas tiene en su interior la frecuencia absoluta de individuos según la combinación de niveles de medición o de categorías, por ejemplo en la celda 11 hay n_{11} individuos provenientes de la muestra n. que tienen la variable 1 con la categoría 1 y la variable 2 con la categoría 1.

Si nos interesa saber si ambas variables están asociadas estadísticamente tendríamos que realizar una Prueba X^2 de Independencia.

2. Se seleccionan **C muestras de tamaños $n_{.1}$, $n_{.2}$ . . . , $n_{.c}$** (en éste caso la variable 2 se esta utilizando de forma que se selecciona una muestra de cada categoría de esta variable) y se clasifican los individuos de cada una de estas muestras según las R categorías de la variable 1 .

La tabla que se genera esta constituida por R x C celdas y cada una de ellas tiene en su interior la frecuencia absoluta de individuos según la combinación de niveles de medición o de categorías, por ejemplo en la celda 11 hay n_{11} individuos provenientes de la muestra **$n_{.1}$** que tienen la variable 1 con la categoría 1. En éste caso se fija uno de los 2 márgenes de la tabla., lo que implica la selección de varias muestras independientes, por lo que solo queda al azar la distribución de los individuos que constituyen la muestra en relación al otro criterio de clasificación ,que en éste caso es la variable cualitativa estudiada.

Si nos interesa saber si las C muestras difieren significativamente en relación con la forma de distribución de la variable 2 tendríamos que realizar una Prueba X^2 de Homogeneidad.

A continuación se exponen en detalle cada una de estas pruebas , debe señalarse que aunque matemáticamente son iguales ,ambas pruebas son diferentes.

2.2 PRUEBA X^2 DE INDEPENDENCIA.

Esta prueba permite medir la significación de la asociación entre 2 variables de clasificación o sea entre 2 variables cualitativas. Sea una tabla de contingencia de 2 entradas:

VARIABLE 1 SEGÚN NIVELES DE CLASIFICACION	VARIABLE 2 SEGUN NIVELES DE CLASIFICACION				
	1	2	.	C	TOTAL
1	n_{11}	n_{12}		n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	.	n_{2c}	$n_{2.}$
.					
R	n_{r1}	n_{r2}	.	n_{rc}	$n_{r.}$
TOTAL	$n_{.1}$	$n_{.2}$	.	$n_{.c}$	n..

Donde ***:***

n_{11} Es el numero de individuos que tienen la categoría 1 de ambas variables.

n_{21} Es el numero de individuos que tienen la categoría 2 de la variable 1 y la 1 de la variable 2.

$n_{1.}$ Es el total de individuos que tiene la categoría 1 de la variable 1.

$n_{.1}$. Es el total de individuos que tiene la categoría 1 de la variable 2.

n.. Es el total de individuos de la muestra.

Una tabla de contingencia como la anterior se denota T.C. de C x R, o sea tabla de contingencia de C filas por R columnas.
Sea entonces:
P_{ij} Es la probabilidad de que un individuo seleccionado al azar de la población pertenezca a la celda situada en la i-esima fila y la j-esima columna.
$P_{i.}$ Es la probabilidad de que un individuo seleccionado al azar de la población pertenezca a la i-esima fila.
$p_{.j}$ Es la probabilidad de que un individuo seleccionado al azar de la población pertenezca a la j-esima columna.
Entonces expresamos la hipótesis de No Asociación entre las 2 variables mediante la siguiente formulación:
$H_o : P_{ij} = P_{i.} P_{.j}$ para toda i= 1,2,3,...,C y j= 1,2,3,...,R.
Sean n.. el numero total de individuos de la muestra y n_{ij} el numero de individuos de la celda ij constituida por la fila i-esima y la columna j-esima.
Se docimará la hipótesis establecida mediante la siguiente expresión, que bajo el supuesto que H_o sea cierta, o sea bajo la hipótesis de No Asociación, puede sustituirse P_{ij} por el producto $P_{i.} P_{.j}$, transformándose la expresión en:

$$X^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\left(\frac{(np_{ij} - np_{i.}p_{.j})^2}{np_{i.}p_{.j}}\right) \dashrightarrow X^2$$, se distribuye Chi Cuadrado

Pero como $P_{i.}$ y $P_{.j}$ son parámetros desconocidos ,es necesario estimarlos mediante sus estimadores máximo verosímiles que son $n_{i.}/n..$ y $n_{.j}/n..$, respectivamente, por lo que la expresión anterior se transforma en :

$$X^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\left(\frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n..})^2}{\frac{n_{i.}n_{.j}}{n..}}\right)$$

Debe destacarse que n_{ij} es el valor observado, o sea el numero de individuos que hay en la celda ij de la tabla de contingencia, mientras que $\left(\frac{n_{i.}n_{.j}}{n..}\right)$ es el valor esperado de esta celda, estimado bajo el modelo de Independencia, por lo que el numerador de esta expresión nos permite medir las discrepancias existentes entre los valores observados y los esperados bajo éste modelo.
La sustitución de $P_{i.}$ y $P_{.j}$ por sus estimadores conlleva a disminuir 1 grado de libertad por cada parámetro estimado, de forma que en éste caso el estadígrafo utilizado tiene una distribución X^2 con (C – 1) (R – 1) grados de libertad.
Algunos autores sugieren incluir en el estadígrafo la corrección de continuidad de Yates, que consiste en restarle ½ al numerador de la expresión para que el valor obtenido de la X^2 disminuya y sea mas difícil rechazar la hipótesis nula , sin embargo actualmente se sugiere considerar el uso de esta corrección solo en los casos de las tablas de contingencias de 2x2, para atenuar así el efecto de usar una distribución continua (X^2) para representar una distribución discreta de frecuencias muestrales, en éste caso se generaría una tabla como la siguiente:

VARIABLE 1 SEGÚN NIVELES DE CLASIFICACION	VARIABLE 2 SEGUN NIVELES DE CLASIFICACION		
	1	2	TOTAL
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
TOTAL	$n._1$	$n._2$	$n..$

En éste caso particular podría utilizarse una expresión simplificada del estadígrafo, que incluye la corrección de Yates:

$$X^2 = \left[\frac{n..\left(|n_{11}n_{22} - n_{12}n_{21}| - n../2\right)^2}{n_{1.}n_{2.}n_{.1}n_{.2}} \right] \rightarrow X^2,$$ con 1 grado de libertad.

Solo resta ahora fijar la regla de decisión:

Sí la X^2 calculada es mayor que la X^2 tabulada con (C-1)(R-1) grados de libertad para determinado nivel de significación, entonces rechazo la hipótesis nula, que en éste caso es de No Asociación = Independencia.
La lógica de esta regla de decisión se basa en que si se cumple el modelo de Independencia entonces las estimaciones de las P_{ij} a partir de éste modelo deben producir escasas discrepancias entre los valores observados presentes en la tabla de frecuencias observadas y los estimados por el modelo, produciendo un valor del estadígrafo pequeño, que no me permitiría rechazar la hipótesis nula de No Asociación. Sin embargo si el modelo no se cumple las discrepancias serán grandes y entonces el valor del estadígrafo será grande también, por lo que en éste caso si rechazo la hipótesis nula de no-asociación , se podría interpretar que hay una asociación estadística significativa entre las 2 variables estudiadas.

Limitaciones de la Prueba:
Siempre que en mas del 20% de las celdas de la tabla de contingencia existan valores esperados menores que 5 o que en una sola celda haya un valor esperado menor que 1, no debe usarse la prueba. En éste caso deben agruparse las categorías siguiendo un sentido lógico para así poder aumentar los valores observados y por ende también los valores esperados.
En el caso de la tabla de 2x2 si existe una sola celda con valor esperado menor que 5 , esto representaría un 25% de las celdas por lo que se utilizaría la Prueba de las Probabilidades exactas de Fisher en lugar de la Prueba X^2 ,ya que en éste caso no es posible agrupar categorías.

Planteamiento y solución de una situación practica.
Se desea conocer si existe asociación entre el habito de fumar y el bajo peso al nacer en una población, para lo cual se selecciona una muestra aleatoria de 100 recién nacidos, obteniéndose los resultados siguientes:

HABITO DE FUMAR	BAJO PESO AL NACER		TOTAL
	SI	NO	
PRESENTE	30	10	40
AUSENTE	20	40	60
TOTAL	50	50	100

Considere α=0.05.

En éste caso hay 1 muestra aleatoria y se quiere determinar si existe asociación estadística significativa entre 2 variables cualitativas (Bajo peso al nacer y Habito de Fumar) por lo que la prueba que debe utilizarse es la de Independencia.

Las Hipotesis a plantear serian las siguientes:

H_o: $P_{ij} = P_{i.} \times P_{.j}$ $\Rightarrow$ Independencia $\Rightarrow$ No existe Asociación

$H_{1:}$ $P_{ij} \neq P_{i.} \times P_{.j}$ $\Rightarrow$ No existe Independencia $\Rightarrow$ Existe Asociación.

Debemos verificar si se cumplen las condiciones para realizar una Prueba Chi Cuadrado, calculando los valores esperados mediante la expresión $n_{i.} \times n_{.j}/n..$, obteniendo los resultados siguientes:

HABITO DE FUMAR	BAJO PESO AL NACER		TOTAL
	SI	NO	
PRESENTE	20	20	40
AUSENTE	30	30	60
TOTAL	50	50	100

Puede observarse que en todas las celdas las frecuencias esperadas son mayores que 5, por lo que puede realizarse la prueba.

Como tenemos una tabla de contingencia de 2x2, podemos utilizar la fórmula del estadígrafo simplificado:

$$X^2 = \left[\frac{n..\left(\left[n_{11}n_{22} - n_{12}n_{21}\right] - n../2\right)^2}{n_{1.}n_{2.}n_{.1}n_{.2}} \right] = \left[\frac{100\left(\left|30x40 - 10x20\right| - 100/2\right)^2}{50x50x40x60} \right] = 15.04$$

Regla de Decisión:

Si la X^2 calculada = 15.04 es mayor que la X^2 tabulada = 3.84 con 1 grado de libertad y un α= 0.05 entonces rechazo la hipótesis nula , por lo que si rechazo que hay independencia , entonces hay asociación estadística significativa entre el habito de fumar y el bajo peso al nacer.

Si realizamos la prueba con un paquete estadístico en una computadora obtendríamos los resultados siguientes:

X^2 = 15.04 Grados de Libertad = 1 p= 0.0001052

Entonces verificaría que todos los valores esperados sean mayores que 5, y como en éste caso se cumple esta restricción podemos usar esta prueba.

Al considerar un α=0.05, como la p asociada al valor del estadígrafo

p= 0.0001052 es menor que α=0.05 , se rechaza la hipótesis nula de no-asociación, por lo que llegamos a la conclusión que hay asociación estadística significativa entre ambas variables.

Esta prueba solo permite determinar la significación de la asociación entre 2 variables cualitativas, la intensidad de esta asociación se medirá con otras medidas que se explicaran posteriormente.

2.3 PRUEBA X^2 DE HOMOGENEIDAD.

Cuando tenemos varias muestras y se desea determinar si son homogéneas con relación a la distribución en las mismas de una variable cualitativa debemos emplear esta prueba. A continuación se expondrá mediante un ejemplo la misma.

Con la finalidad de evaluar el hábito de fumar como factor de riesgo del cáncer del pulmón, se seleccionan 2 muestras aleatorias, una de pacientes con esta enfermedad y la otra de personas sin esta condición. A continuación se brinda la información obtenida:

HABITO DE FUMAR	CANCER DEL PULMON		TOTAL
	SI	NO	
PRESENTE	11	13	24
AUSENTE	10	46	56
TOTAL	21	59	80

La información que se resume en una tabla de contingencia como la anterior puede expresarse de forma general:

HABITO DE FUMAR	CANCER DEL PULMON		TOTAL
	SI	NO	
PRESENTE	n_{11}	n_{12}	$n_{1.}$
AUSENTE	n_{21}	n_{22}	$n_{2.}$
TOTAL	$n_{.1}$	$n_{.2}$	$n_{..}$

Donde:

$P_{1.} = n_{1.}/n_{..}$ es la probabilidad de fumar

$P_{2.} = n_{2.}/n_{..}$ es la probabilidad de no fumar

$P_{11} = n_{11}/n_{.1}$ es la probabilidad de fumar de los pacientes con cáncer del pulmón.

$P_{12} = n_{12}/n_{.2}$ es la probabilidad de fumar de las personas sin cáncer del pulmón.

Si el habito de fumar esta relacionado con el cáncer del pulmón se supone que la proporción de fumadores sea mayor entre las personas que tienen esta enfermedad que entre las que no la tienen, por esto nos interesa determinar como hipótesis alternativa $H_1: P_{11} \neq P_{12}$

Nos interesa entonces docimar las hipótesis siguientes:

$H_o: P_{11} = P_{12}$ $H_1: P_{11} \neq P_{12}$

La tabla de contingencia anterior tiene los márgenes fijos, pues estamos trabajando con 2 muestras y por eso fijamos los valores de $n_{.1}$ (pacientes con cáncer del pulmón) y $n_{.2}$ (personas sin esta enfermedad), eso hace que cada individuo ubicado en una celda sólo tiene 2 posibilidades de respuesta: fuman o no , por lo que cada individuo tiene una distribución Bernoulli.

En cada celda hay un conjunto de individuos, por lo que n variables Bernoulli $\Rightarrow$distribución Binomial, de esta forma tenemos que:

n_{ij} - es una variable con distribución Binomial con un Valor Esperado $E(n_{ij})$ y una Varianza $V(n_{ij})$ definidas de la forma siguiente:

$n_{ij} \rightarrow$ **b ($E(n_{ij})$ =** $n_{i.}\frac{n_{.j}}{n_{..}}$ **, $V(n_{ij})$=** $n_{i.}\frac{n_{.j}}{n_{..}}\left(1-\frac{n_{.j}}{n_{..}}\right)$**).**

Entonces se necesita conocer los valores esperados de cada una de las celdas de la tabla de contingencia y esto lo podemos hacer aplicando las expresiones anteriores:

$E(n_{11}) = \frac{n_{1.} \frac{n_{.1}}{n_{..}}}{} = 21\frac{24}{80} = 6.3$

$E(n_{12}) = \frac{n_{1.} \frac{n_{.2}}{n_{..}}}{} = 21\frac{56}{80} = 14.7$

$E(n_{21}) = \frac{n_{2.} \frac{n_{.1}}{n_{..}}}{} = 59\frac{24}{80} = 17.7$

$E(n_{22}) = \frac{n_{2.} \frac{n_{.2}}{n_{..}}}{} = 59\frac{56}{80} = 41.3$

Si ya conocemos los valores observados y los valores esperados sólo necesitamos un estadígrafo con una distribución conocida que nos permita determinar si hay diferencias significativas entre los valores observados y los valores que se esperarían bajo el supuesto que la hipótesis nula fuera cierta, o sea, si se cumpliera que la distribución de las proporciones en ambas muestras es similar.

El estadígrafo que nos permite determinar lo anterior es:

$$X^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\left(\frac{(n_{ij} - n_{..}\frac{n_{i.}n_{.j}}{n_{..}^2})^2}{n_{..}\frac{n_{i.}n_{.j}}{n_{..}^2}}\right) \rightarrow X^2$$ con (C- 1) (R – 1) grados de libertad.

Como en éste caso estamos ante una tabla de contingencia de 2x2 utilizaremos la expresión reducida que lleva implícita la corrección de continuidad:

$$X^2 = \left[\frac{n..\left(\left|n_{11}n_{22} - n_{12}n_{21}\right| - n../2\right)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}\right] \rightarrow X^2$$, con 1 grado de libertad.

Sustituyendo en esta expresión los valores de la tabla:

$$X^2 = \left[\frac{80\left(\left|11x46 - 10x13\right| - 80/2\right)^2}{24x56x21x59}\right] = 5.42$$

Solo resta ahora fijar la **regla de decisión**: Sí la X^2 calculada es mayor que la X^2 tabulada con (C-1)(R-1) grados de libertad para determinado nivel de significación, entonces rechazo la hipótesis nula, que en éste caso es que la proporción de fumadores es similar en ambas muestras, lo que sugiere que en éste caso H_1: $P_{11} \neq P_{12}$, o sea que la proporción de fumadores es diferente en ambos grupos.

La lógica de esta regla de decisión se basa en que si se cumple la hipótesis nula H_o: $P_{11} = P_{12}$ se deben producir escasas discrepancias entre los valores observados presentes en la tabla de frecuencias observadas y los valores esperados estimados, produciendo un valor del estadígrafo pequeño, que no permitiría rechazar la hipótesis nula de igualdad de proporciones.

Sin embargo si la hipótesis nula no se cumple las discrepancias serán grandes y entonces el valor del estadígrafo será grande también, por lo que en éste caso si rechazo la hipótesis nula de igualdad de proporciones, entonces interpretar estos resultados como que la proporción de fumadores es diferente en ambos grupos, o sea. que se cumple la H_1: $P_{11} \neq P_{12}$.

En nuestro ejemplo obtuvimos una X^2 calculada = 5.42 que es mayor que la X^2 tabulada = 3.84 con 1 grado de libertad y un α= 0.05 entonces se rechaza la hipótesis nula, por lo que si rechazo que la proporción de fumadores es similar en ambas muestras, entonces hay diferencias estadísticas significativas entre el habito de fumar y el cáncer del pulmón.

Si realizamos la prueba con un paquete estadístico en una computadora obtendríamos los resultados siguientes:

X^2 = 5.42 Grados de Libertad = 1 p= 0.0198649

Entonces como verificaría si todos los valores esperados son mayores que 5, como en éste caso se cumple esta restricción podemos usar esta prueba.

Al considerar un α=0.05 , como la p asociada al valor del estadígrafo

p= 0.0198649 es menor que α=0.05, se rechaza la hipótesis nula de no diferencias entre las proporciones , por lo que llegamos a la conclusión que existen diferencias estadísticas significativas entre ambas muestras en relación a la distribución de esta variable.

Limitaciones de la Prueba:

Al igual que en el caso de la Prueba de Independencia en esta Prueba existen las misma restricciones, que se señalan a continuación:

Siempre que en más del 20% de las celdas de la tabla de contingencia existan valores esperados menores que 5 o que en una sola celda haya un valor esperado menor que 1, no debe usarse la prueba. En éste caso deben agruparse las categorías siguiendo un sentido lógico para así poder aumentar los valores observados y por ende también los valores esperados.

En el caso de la tabla de 2x2 sí existe una sola celda con valor esperado menor que 5 , esto representaría un 25% de las celdas, por lo que se utilizaría la Prueba de las Probabilidades exactas de Fisher en lugar de la Prueba X^2 ,ya que en éste caso no es posible agrupar las categorías.

Planteamiento y solución de una situación practica.

Se necesita probar la efectividad de 3 tratamientos para controlar la hipertensión arterial, para lo cual se seleccionan 3 muestras aleatorias de pacientes hipertensos y se asignan aleatoriamente los tratamientos a los pacientes estudiados.

La variable respuesta es el control de la hipertensión arterial a los 6 meses de tratamiento, los resultados obtenidos se relacionan a continuación:

HIPERTENSION ARTERIAL	TRATAMIENTOS		
	1	2	3
CONTROLADA	50	160	185
NO CONTROLADA	50	40	15
TOTAL	100	200	200

Considere α=0.05.

En éste caso hay 3 muestras aleatorias y se quiere determinar si existen diferencias estadísticas significativas entre los porcentajes de pacientes controlados , por lo que la prueba que debe utilizarse es la de Homogeneidad.

Las hipótesis a plantear serian las siguientes:

H_o: $P_{tto\ 1} = P_{tto\ 2} = P_{tto\ 2} \Rightarrow$ No existen diferencias entre tratamientos

H_1: $P_{tto\ 1} \neq P_{tto\ 2} \neq P_{tto\ 2} \Rightarrow$ Existen diferencias entre tratamientos.

Debe verificarse si se cumplen las condiciones para realizar una Prueba Chi-Cuadrado, calculando los valores esperados mediante la expresión $n_{i.}$ x $n_{.j}$/n.., obteniendo los resultados siguientes:

HIPERTENSION ARTERIAL	TRATAMIENTOS		
	1	2	3
CONTROLADA	79	158	158
NO ONTROLADA	21	42	42
TOTAL	100	200	200

Se observa que las celdas las frecuencias esperadas de todas las celdas son mayores que 5, por lo que puede realizarse la prueba X^2.

Si se conocen los valores observados y los valores esperados sólo se necesita un estadígrafo con una distribución conocida que nos permita determinar si hay diferencias significativas entre los valores observados y los valores que se esperarían bajo el supuesto que la hipótesis nula fuera cierta ,o sea si se cumpliera que la distribución de las proporciones en ambas muestras es similar.

El estadígrafo que nos permite determinar lo anterior es:

$$X^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\left(\frac{(n_{ij} - n_{..}\frac{n_{i.}n_{,j}}{n_{..}^2})^2}{n_{..}\frac{n_{i.}n_{.j}}{n_{..}^2}}\right)$$

$\rightarrow X^2$ con (C- 1) (R – 1) grados de libertad

Entonces como estamos trabajando con una tabla de contingencia de 2 filas (HTA controlada o no) y 3 columnas (3 tratamientos) , tenemos una TC de 2x3 en la que podemos trabajar con el estadígrafo X^2 sin la corrección de Yates, como aparece en la expresión anterior.

Sustituyendo los datos en la expresión anterior tenemos:

$$X^2 = \frac{(50-79)^2}{79} + \frac{(160-158)^2}{158} + \frac{(185-158)^2}{158} + \frac{(50-21)^2}{21} + \frac{(40-42)^2}{42} + \frac{(15-42)^2}{42} =$$

X^2 = 10.65 + 0.02 + 4.61 + 40.05 +0.09 + 17.36 = 72.78

Se busca entonces el valor de la X^2 tabulada para una confiabilidad de un 95% con (C=2-1) x (R=3-1) grados de libertad, o sea con 1 x 2 = 2 grados de libertad que es 5.99, por lo que como 72.78 es mayor que 5.99 se rechaza la hipótesis nula de proporciones iguales. Observemos como se comportan las proporciones de pacientes controlados:

TRATAMIENTOS	PACIENTES CONTROLADOS	TOTAL DE PACIENTES	PORCENTAJE DE CONTROLADOS
1	50	100	50.0
2	160	200	80.0
3	185	200	92.5
TOTAL	395	500	79.0

Llegamos entonces a la conclusión de que la efectividad de los tratamientos es diferente, se observa que el tratamiento 1 es el menos eficiente mientras que el tratamiento 3 es el que produce mayor porcentaje de hipertensos controlados.

Si realizamos la prueba con un paquete estadístico en la computadora obtendríamos los resultados siguientes:

$X^2 = 72.78$ Grados de Libertad = 2 p= 0.000000

Entonces verificaría antes de aplicar la prueba si se cumplen restricciones, como en éste caso se cumplen la utilizaría. Al considerar un α=0.05 , como la p asociada al valor del estadígrafo p= 0.000000 es menor que α=0.05 , se rechaza la hipótesis nula de no diferencias entre las proporciones , por lo que llegamos a la conclusión que hay diferencias estadísticas significativas entre ambas poblaciones en relación a la distribución de esta variable.

2.4 EJERCICIOS.

En un servicio de geriatría se desea estudiar si existe asociación entre la presencia de hipercolesterolemia en los últimos 10 años y la aparición de aterosclerosis para lo que se selecciona una muestra aleatoria de 244 ancianos de 70 y más años.

Los resultados se brindan a continuación:

ATEROSCLEROSIS	HIPERCOLESTEROLEMIA		TOTAL
	SI	NO	
SI	56	138	194
NO	8	42	50
TOTAL	64	180	244

Considere $\propto$ = 0.05

a) ¿Qué prueba utilizaría para resolver éste problema?. Fundamente su respuesta.

b) A partir de estos resultados ¿puede usted inferir que existe asociación entre la hipercolesterolemia y la aterosclerosis? Fundamente su respuesta.

2. Se desea conocer si la toxemia durante el embarazo puede ser considerada un factor de riesgo para la mortalidad perinatal.

Se seleccionan 2 muestras aleatorias, una de embarazadas toxémicas y otra de embarazadas sin esta condición, se siguieron durante el embarazo. Se consideró como variable de respuesta si el producto de la gestación falleció o no durante el periodo perinatal. Los resultados se brindan a continuación:

TOXEMIA	MUERTE PERINATAL		TOTAL
	SI	NO	
SI	48	192	240
NO	50	430	480

Considere $\propto$=0.05

a) Que Prueba utilizaría para resolver éste problema? Fundamente su respuesta.

b) A partir de estos resultados puede usted inferir que la toxemia gravídica es un factor de riesgo para la mortalidad perinatal? Fundamente su respuesta.

2.5 BIBLIOGRAFIA.

1. Daniel, W.: Bioestadistica: base para el análisis de las ciencias de la salud. Editorial UTEHA, México, 1997.
2. Siegel, S.: Análisis Experimental no paramétrico. Instituto Cubano del Libro, Edición Revolucionaria, La Habana, 1972.
3. Astrain Rodríguez, M. Elena: Análisis Estadístico. Escuela Nacional de Salud, La Habana, 2001.
4. Martínez Canalejo, H. y S. Santana Porben: Manual de procedimientos bioestadísticos. Tomo 1. Editorial de Ciencias Médicas, La Habana, 1989.

5. López Pardo, Cándido: Notas de Clase. Curso La Estadística no paramétrica en la investigación contemporánea. Facultad de Economía. Universidad de la Habana, 1990.
6. López Pardo, Cándido: El análisis de tablas de contingencia.. Departamento de Estadística Aplicada. Facultad de Economía. Universidad de la Habana, 1990.
7. Verdecía, Dominador: Notas de Clase. Especialidad de Bioestadistica. Instituto de Desarrollo de la Salud. Ministerio de Salud Publica, 1980.
8. Mosteller, F y R. E. Rourke: Sturdy Statistics: Nonparametrics and Order Statistics. Addison Wesley, Massachusetts, 1973.
9. Klugh,H.E.: Statistics: The essentials for research. John Wiley and sons, New York, 1970.

CAPITULO 3: MÉTODOS DE ANÁLISIS PARA VARIABLES CUANTITATIVAS EN MUESTRAS INDEPENDIENTES.

3.1 INTRODUCCION.

En el trabajo de Salud Publica es frecuente la necesidad de establecer comparaciones entre los valores medios de parámetros cuantitativos de diferentes poblaciones, pueden citarse los ejemplos siguientes:

1. Determinar si existen diferencias entre los valores medios de Hemoglobina, Eritrosedimentación u otras determinaciones séricas entre las poblaciones de hombres y de mujeres.
2. Identificar si el peso medio de los recién nacidos, o si el peso y la talla de adolescentes en distintos municipios del país difieren,
3. Determinar si el ingreso per cápita de la población difiere entre las distintas provincias del país.

Cuando se conoce que la variable en estudio se distribuye Normal, si se están comparando 2 muestras independientes y se conocen las varianzas podremos utilizar la prueba Z para la comparación de 2 medias, si se cumplen todas las condiciones anteriores pero no conocemos las varianzas y es necesario estimarlas a partir de los datos muestrales, siempre que se asuma que son iguales, hay 2 soluciones posibles:

1. Si las muestras son grandes, de 30 o más observaciones, puede asumirse por el Teorema Central del Limite que la distribución de la variable se aproxima a la distribución Normal y puede entonces trabajarse con la prueba Z como en el caso anterior,
2. Si las muestras son pequeñas, menos de 30 observaciones, se podrá utilizar la prueba T para muestras pequeñas.

Hay situaciones en que no conocemos la distribución de la variable ni tampoco podemos asumir la igualdad de las varianzas y existe la necesidad de resolver problemas como los antes expuestos, en éste caso tendría que recurrir a las Pruebas no Paramétricas que brindan diferentes opciones para estas situaciones.

Las pruebas que se explican a continuación son pruebas basadas en las estadísticas de rangos, que permiten realizar análisis muy simples y efectivos.

En éste tipo de estadística se sustituye el valor de cada observación por un numero denominado Rango que representa el orden del valor de cada observación en el conjunto de datos estudiados, por lo que tienen la ventaja de que cuando existen datos aberrantes, al sustituir el valor real de la observación por el valor del rango, no se afectan los resultados obtenidos, como ocurriría en el caso de trabajar con el valor aberrante.

3.2 PRUEBA DE LA MEDIANA.

Esta Prueba permite determinar si 2 muestras independientes difieren con relación a sus medianas, o sea permite determinar si 2 muestras independientes provienen de poblaciones con la misma mediana siempre que la variable éste al menos en escala ordinal. Existe una generalización de esta prueba que permite determinar sí 3 o más muestras independientes provienen de poblaciones con la misma mediana siempre que la variable éste al menos en escala ordinal, que no será objeto de estudio en éste curso.

Nos interesa probar la hipótesis: $H_o : Me_1 = Me_2$ y $H_1 : Me_1 \neq Me_2$

Debe señalarse que en el caso que las distribuciones de ambas poblaciones sean simétricas, la mediana debe coincidir con la media aritmética, convirtiéndose esta prueba en una Prueba de Comparación de 2 medias también.

Como estamos asumiendo que ambas muestras provienen de poblaciones con la misma mediana , se determina el valor de la mediana general, o sea se combinan los valores de ambos grupos para determinar el valor de la mediana.

Una vez determinado éste valor se determina en cada una de las muestras cuantas observaciones tienen valores mayores que la mediana general y cuantas tienen valores inferiores a la misma.

Según Siegel si hay valores que coincidan con los de la mediana general en las muestras estudiadas, se procederá de la siguiente forma:

1. Si la muestra es grande y un numero pequeño de observaciones coincide con éste valor pueden excluirse del análisis estas observaciones.
2. Si la muestra no es grande o si el numero de observaciones que coincide con éste valor es grande entonces pueden contarse estos valores junto con los valores menores que la mediana. La información se resumirá en una tabla de contingencia de la forma siguiente:

Valores/Mediana	Muestra 1	Muestra 2	Total
Observaciones > Mediana General	a	b	a+b
Observaciones ≤ Mediana General	c	d	c+d
Total	a+c	b+d	n

Entonces necesitamos un estadígrafo con una distribución conocida para determinar si existen diferencias significativas entre las medianas de ambas poblaciones, para lo cual utilizaremos el estadígrafo X^2:

$$X_o^2 = \frac{n\,(/ ad - bc / - n/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Este estadígrafo bajo el supuesto de que no existen diferencias entre las medianas de ambas poblaciones se distribuye X^2 con 1 grado de libertad.

Es lógico suponer que si las medianas de ambas poblaciones son iguales entre si y por ende iguales a la mediana general, entonces como la mediana es el valor central de una distribución el numero de observaciones que en ambas muestras quedaría por encima y por debajo de la mediana general seria la mitad de cada una de las muestras , por lo que en ese caso el producto **ad** seria aproximadamente igual al producto **bc** de forma que **/ad – bc/≅ 0** y por ende el valor del estadígrafo **X_o^2 sería aproximadamente igual a cero.**

En el caso que las medianas de ambas poblaciones no coincidieran, la mediana general no coincidiría con ellas y al determinar en cada muestra el numero de observaciones por encima y por debajo de la mediana general seria muy diferente en ambos grupos, por lo que el producto **ad** seria muy diferente al **bc** de forma que **/ ad – bc/ >>> 0** , de forma que el valor del estadígrafo **X_o^2 sería elevado.**

En el caso de esta prueba la Regla de Decisión sería la siguiente:

Si $X_o^2 > X^2_{tabulada}$ *con 1 grado de libertad para un nivel de significación determinado,* ***rechazaría la hipótesis nula*** $H_o : Me_1 = Me_2$.

Otra forma de rechazar sería si la probabilidad asociada al valor del estadígrafo fuera menor que la probabilidad de error de tipo I que estamos dispuestos a asumir, o sea **si estamos trabajando con un** α**= 0.05 se podría rechazar la hipótesis nula siempre que la probabilidad de error asociada al valor de** X_o^2 **fuera menor que éste valor.**

Debemos recordar que al usar el estadígrafo X^2 hay que tener en cuenta que deben cumplirse las restricciones para la Prueba $X^{2,}$. En éste caso como hay una tabla de contingencia de 2x2, debe cumplirse que todos los valores esperados sean mayores o iguales a 5, en caso de no cumplirse esta restricción se deberá utilizar la Prueba de las Probabilidades Exactas de Fisher.

Según Siegel , Mood (1954) se ha demostrado que si la Prueba de la Mediana se aplica a datos que pueden analizarse adecuadamente por una prueba paramétrica más poderosa, la prueba t en éste caso, su potencia-eficiencia sería de cerca de un 95% cuando $n_1 + n_2 = 6$,éste valor iría disminuyendo a medida que aumenta el tamaño muestral llegando a tener una eficiencia asintótica eventual de 63%.

<u>Planteamiento y solución de una situación practica</u>.

Se desea determinar sí las medianas de las distribuciones de los valores de creatinina varia en 2 grupos de pacientes sometidos a distintos tipos de tratamientos.

Los valores de creatinina obtenidos fueron los siguientes:

Muestra A: 1.63, 1.68, 1.59, 1.64, 1.70, 1.58, 1.62, 1.71, 1.57, 1.84, 1.90, 1.96.

Muestra B: 1.65, 1.69, 1.72, 1.91, 1.74, 1.75, 1.55, 1.86, 1.87, 1.88.

¿Puede Ud. afirmar con un nivel de significación de 0.05 que las medianas de creatinina de ambos grupos de pacientes difieren entre si?

En éste ejemplo hay 2 muestras independientes y una variable cuantitativa , nos interesa determinar si las medianas de esas 2 poblaciones difieren, por lo que la prueba que debe utilizarse es la Prueba de la Mediana.

Las hipotesis a docimar son las siguientes:

$$H_o : Me_1 = Me_2$$

$$H_1 : Me_1 \neq Me_2$$

Se ordenan los valores de ambas muestras, como si provinieran de una sola población y se determina el valor central de la forma siguiente:

VALOR	ORDEN	MUESTRA
1.55	1	B
1.57	2	A
1.58	3	A
1.59	4	A
1.62	5	A
1.63	6	A
1.64	7	A
1.65	8	B
1.68	9	A
1.69	10	B
1.70	11	A
1.71	12	A
1.72	13	B

1.74	14	B
1.75	15	B
1.84	16	A
1.86	17	B
1.87	18	B
1.88	19	B
1.90	20	A
1.91	21	B
1.96	22	A

La mediana general de la serie es el valor que ocupa la posición central de la misma, luego de ordenada. En éste caso luego de ordenada la serie, se determinaría la posición central de la forma siguiente:

$$\text{Posición de la Mediana general} = \frac{22+1}{2} = 11.5$$

Como se obtuvo un número fraccionario hay que buscar los valores de los números que ocupan las posiciones 11 y 12. Se calcula entonces la semisuma de ambos y el resultado es el valor de la mediana general que en nuestro ejemplo seria:

$$\text{Mediana General} = \frac{1.70+1.71}{2} = 1.705$$

En el caso que se hubiera obtenido un numero entero al buscar la posición de la mediana, solo hubiera sido necesario buscar el numero que ocupa esa posición y ese es el valor de la mediana general. Luego de obtener la Mediana General se determinan cuantas observaciones en cada una de las muestras tiene valores superiores al de la mediana general y cuantas valores iguales o menores a la misma, se resume esa información de la forma siguiente:

Valores/Mediana	Muestra A	Muestra B	Total
Observaciones > Mediana General	4	7	11
Observaciones ≤ Mediana General	8	3	11
Total	12	10	22

Se determina entonces el valor del estadígrafo X_o^2 de la forma siguiente:

$$X_o^2 = \frac{22\left(\left|4x3-7x8\right|-22/2\right)^2}{(11x11x12x10)}$$

X_o^2 = 1.65. Ahora tendríamos que determinar si con éste valor del estadígrafo puede rechazarse la hipótesis nula que se esta probando, en esté caso la regla de decisión Seria

Si $X_o^2 \geq X^2_{\text{tabulada}}$ con 1 grado de libertad, para un nivel de significación de un 5% sería 3.84, se rechaza la hipótesis nula. En nuestro ejemplo no se cumple lo anterior, pues 1.65 < 3.84 y no se rechaza la hipótesis nula, por lo que se concluye que las medianas de las distribuciones de los valores de creatinina no varia en los 2 grupos de pacientes estudiados con una confiabilidad de un 95%. En éste ejemplo al calcular el estadígrafo con un programa estadístico se obtienen los resultados siguientes:

Muestra Numero de casos Mediana General = 1.705
B 10 A 12
p= 0.1984 (mayor que 0.05) $\Rightarrow$ No Rechazo la hipótesis nula

3.3 PRUEBA DE LA SUMA DE RANGOS DE WILCOXON.

Esta prueba constituye la base para el resto de las pruebas que utilizan rangos, a partir de ella las pruebas basadas en rangos tuvieron un gran desarrollo.

Suponga que tenemos 2 muestras independientes, una muestra A con M=3 observaciones y otra B con N=2 observaciones.

Se desea conocer si ambas muestras proceden de la misma distribución, para determinar lo anterior se ordenan en orden ascendente los valores de las 2 muestras como si fuera una sola muestra y posteriormente se le da un rango a cada observación.

Se determina entonces la suma de rangos de la muestra más pequeña, en éste caso la muestra B, y se procede a buscar cual es la probabilidad de los diferentes valores que podría tomar esta suma de rangos de la muestra B.

Si asumimos que ambas muestras provienen de la misma distribución, entonces los rangos asignados a cada una de las muestras estarán distribuidos aleatoriamente, por lo que el total de combinaciones de rangos posibles se obtendría mediante la Combinatoria de $n_A + n_B$ en n_B , lo que en éste caso seria igual a :

$\frac{(n_a + n_b)!}{n_a! n_b!} = \frac{(3+2)!}{3!2!} = \frac{5!}{3!2!} = 10$, combinaciones posibles, que se expresan en la tabla:

No. de Combinaciones i	Combinaciones de Rangos asignados a muestra A y B					Suma de rangos de muestra B
	Valores de los rangos					
	1	2	3	4	5	
1	A	A	A	B	B	9
2	A	A	B	B	A	7
3	A	B	B	A	A	5
4	B	B	A	A	A	3
5	A	B	A	B	A	6
6	B	A	A	A	B	6
7	B	A	B	A	A	4
8	B	A	A	B	A	5
9	A	B	A	A	B	7
10	A	A	B	A	B	8

Bajo el supuesto de que las 2 muestras provienen de la misma distribución, seria lógico pensar que cada combinación tiene una probabilidad de ocurrencia de 1/10. Resumiendo en un cuadro los resultados anteriores tenemos que la probabilidad de obtener diferentes valores para la Suma de rangos de B

(Σ Rangos B) será:

PΣ rangos B	1/10	1/10	2/10	2/10	2/10	1/10	1/10
Σ rangos B	3	4	5	6	7	8	9

Entonces si queremos determinar la probabilidad de que la Suma de rangos de B tenga un valor mayor o igual a 8, plantearemos entonces:

P(Σ rangos B)$^{\geq 8 / M = 2 y N = 3}$ = P(8)+P(9) = 1/10+1/10=2/10=0.2

Si consideramos que α=0.05, como 0.2 es mayor que α, entonces es muy probable que ambas muestras procedan de la misma población. En la práctica existen tablas para determinar estos valores, pues como se evidenció anteriormente el cálculo para muestras mayores seria muy engorroso.

Procedimientos:

Sean A y B, dos muestras con M y N observaciones respectivamente, donde se cumple que M$\geq N$

1.1 Se ordenan todas las observaciones de ambas muestras, como si fuera una sola muestra, en orden ascendente y se asignan los rangos a los valores ordenados.

1.2 Se identifican los valores que pertenecen a cada muestra.

1.3 Se determina el estadígrafo que en esta prueba es: T_o = Suma de rangos de B =(Σ rangos B) , recuerde que B es la muestra más pequeña.

1.4 Se calcula T_o = Suma de rangos de B =(Σ rangos B).

1.5 La Regla de Decisión será:

a) Se plantea la hipótesis que se adecue a la situación que necesitamos resolver, tomando en consideración las hipótesis siguientes:

HIPOTESIS	REGLA DE DECISION RECHAZAR H_O SI:	α MAS USADOS
$H_o: Me_B = Me_A$ $H_1: Me_B \neq Me_A$	$T_o \leq T_I$ o $T_o \geq T_S$	0.025 0.005
$H_o: Me_B \geq Me_A$ $H_1: Me_B < Me_A$	$T_o \leq T_I$	0.05 0.01
$H_o: Me_B \leq Me_A$ $H_1: Me_B > Me_A$	$T_o \geq T_S$	0.05 0.01

Donde T_I y T_S son los valores críticos obtenidos en la tabla de Valores críticos para la estadística de prueba de la Suma de Rangos de Wilcoxon, considerando un tamaño de n_A y n_B y un nivel de significación dado. Esta tabla sirve para trabajar cuando el tamaño de la muestra llega hasta 25 la muestra menor y 50 la mayor.

b)Cuando el tamaño de la muestra menor excede las 25 observaciones puede trabajarse con la aproximación a la distribución normal: T_o = Σ rangos de B, se distribuye N (μ_T, σ_T) cuando n $\rightarrow \infty$

Entonces en ese caso puede usarse el estadígrafo Z

$Z = \dfrac{(To - \mu T)}{\sigma T}$ que se distribuye Normal (0, 1) ,donde :

$$\mu T = \frac{N(M+N+1)}{2} \quad \text{y} \quad \sigma T = \left[\frac{MN(M+N+1)}{12}\right]^{1/2}$$

c) Otra forma de rechazar la hipótesis nula es cuando estamos trabajando con un programa estadístico que calcule T_o y p que es la probabilidad de error de tipo I asociada a ese valor. En éste caso si el valor de p es menor que el α prefijado se rechaza la hipótesis nula., tomando en consideración los aspectos siguientes:

HIPOTESIS	REGLA DE DECISION RECHAZAR H_O SI:
$H_o: Me_B = Me_A$ $H_1: Me_B \neq Me_A$	$Me_B \neq Me_A$ y $p \leq 0.025$ ó $p \leq 0.005$
$H_o: Me_B \geq Me_A$ $H_1: Me_B < Me_A$	$Me_B < Me_A$ y $p \leq 0.05$ ó $p \leq 0.01$
$H_o: Me_B \leq Me_A$ $H_1: Me_B > Me_A$	$Me_B > Me_A$ y $p \leq 0.05$ ó $p \leq 0.01$

Deben aclararse los aspectos siguientes:

1. Si $n_A = n_B$ se seleccionara el estadígrafo Σ rangos de la muestra tomando en consideración la hipótesis alternativa, de la forma siguiente:

Si la Hipótesis Alternativa es:	Se seleccionará la muestra:
$H_1: Me_B \neq Me_A$	Cualquiera de las 2 tiene igual solución.
$H_1: Me_B < Me_A$	T_o= la que tenga mayor suma de rangos.
$H_1: Me_{B>} Me_A$	T_o= la que tenga menor suma de rangos.

2. Las hipótesis deben plantearse tomando como primer parámetro de referencia para docimar, el de la muestra mas pequeña, en el caso de que sean de igual tamaño podrán plantearse de cualquier forma.

3. En caso de existir observaciones de igual valor se asignaran rangos promedios, por ejemplo si los 3 primeros valores son:

VALORES	15	15	15
LUGARES	1	2	3
RANGOS	1+2+3/3 =**2**	1+2+3/3 =**2**	1+2+3/3 =**2**

Potencia - Eficiencia:

Según Siegel, si la prueba de la Suma de Rangos de Wilcoxon se aplica a datos que pueden analizarse adecuadamente por una prueba paramétrica mas poderosa, la prueba t de Student en éste caso, su potencia-eficiencia seria de cerca de un 95.5% y se acerca a 95% para muestras de tamaño moderado, por lo que es una excelente alternativa ante la prueba t. Debe destacarse que esta prueba es más potente que la de la Mediana, pues a diferencia de esta ultima que utiliza solo la información de cómo están ubicadas las observaciones de cada muestra con relación al valor de la Mediana General, en el caso de la prueba de la Suma de Rangos se utiliza la información relativa a la ubicación de cada observación en las muestras, que se resume en el estadígrafo de la Suma de Rangos.

Planteamiento y solución de una situación practica.

Se desea conocer si los niveles de excreción urinaria de Sodio/Potasio varían en relación a la presencia de la enfermedad X , para lo cual se seleccionaron 2 muestras aleatorias una constituida por 16 pacientes con esta enfermedad y la otra por 12 personas sin ella. Pruebe las hipótesis siguientes:

a) Los niveles de excreción urinaria de Sodio/Potasio difieren en ambos grupos.
b) Los niveles de excreción urinaria de Sodio/Potasio en los enfermos de X son mayores que en las personas sin esta enfermedad.

Considere $\alpha = 0.05$.

Excreción urinaria de Sodio/Potasio según presencia de la Enfermedad X.

Enfermedad X	
Presente	Ausente
263	60
288	119
432	153
890	588
450	124
1270	196
220	14
350	23
283	43
274	854
580	400
285	73
524	
135	
500	
120	

Cuando estamos ante un problema de éste tipo lo primero que hacemos es identificar que prueba debe utilizarse. En éste caso tenemos: Tipo de muestras: 2 muestras independientes; Tipo de variable: cuantitativa continua. Por lo que la prueba más eficiente es la de la Suma de Rangos de Wilcoxon.

a) Se quiere probar si la excreción urinaria de Sodio/Potasio difiere según presencia de la Enfermedad X por lo que esta seria nuestra hipótesis alternativa

$$H_o : Me_B = Me_A$$
$$H_1 : Me_B \neq Me_A$$

Nuestro estadígrafo seria entonces la Suma de Rangos de la muestra más pequeña. La muestra de pacientes con la Enfermedad X es de 16 personas y la de personas sin esta condición es de 12 personas, por lo que la muestra menor es la segunda.
Entonces ahora asignamos los rangos a todas las observaciones como si procedieran de la misma muestra.

Enfermedad X			
Presente		Ausente	
Excreción Na/K	Rangos	Excreción NA/K	Rangos
263	13	60	4
288	17	119	6
432	20	153	10
890	27	588	25
450	21	124	8
1270	28	196	11
220	12	14	1
350	18	23	2
283	15	43	3
274	14	854	26
580	24	400	19
285	16	73	5
524	23	ΣSuma de Rangos	120
135	9		
500	22		
120	7		

Se buscan entonces los valores críticos para M= 16, N= 12 y α= 0.025 por ser una prueba bilateral., y se encuentra que $T_I = 131$ y $T_S = 217$
La Regla de Decisión nos dice que si: $T_o \leq T_I$ o $T_o \geq T_S$ rechazamos la hipótesis nula , en nuestro caso tenemos que : $T_O = 120$ es menor que $T_I = 131$ por lo que se rechaza la hipótesis nula con un nivel de significación de 5% y se concluye que la excreción urinaria de Sodio/Potasio difiere según la presencia de la Enfermedad X .
b)Se quiere probar que la excreción urinaria de Sodio/Potasio es menor en las personas sin la Enfermedad X por lo que esta seria nuestra hipótesis alternativa:

$$H_o : Me_B \geq Me_A$$
$$H_1 : Me_B < Me_A$$

Observe como en éste ejemplo se plantean las hipótesis en función de la muestra más pequeña, la muestra de personas sin la enfermedad.
Como el valor del estadígrafo de la Suma de Rangos de B es 120, ahora tendríamos que buscar los valores críticos para M=16, N= 12 y α= 0.05 por ser una prueba unilateral., encontramos entonces que $T_I = 138$ y $T_S = 210$

La Regla de Decisión nos dice que sí: $T_o \leq T_i$ se rechaza la hipótesis nula , en nuestro caso tenemos que : T_o= 120 es menor que $T_{I=}$ 138 por lo que se rechaza la hipótesis nula con un nivel de significación de un 5% y se concluye que la excreción urinaria de Sodio/Potasio es menor en las personas que no tienen la Enfermedad X .

En el caso de éste ejemplo al calcular el estadígrafo con un programa estadístico obtenemos los resultados siguientes:

Muestra	Numero de casos	Suma de Rangos	T_O= 120
B	12	120.00	
A	16	286.00	

p=0.012 (menor que 0.05)$\Rightarrow$ Rechazo la hipótesis nula.

3.4 PRUEBA DE MANN-WHITNEY.

Otra prueba que se utiliza para resolver el mismo caso que resuelve la prueba de la Suma de Rangos de Wilcoxon es la prueba de Mann-Whitney, el procedimiento seguido en ambas es muy parecido, las hipótesis que se dociman las mismas y el estadígrafo utilizado se parece aunque por supuesto no es igual.

Procedimientos: Sean A y B , 2 muestras con M y N observaciones respectivamente, donde se cumple que M$\geq N$.

1.Se ordenan todos las observaciones de ambas muestras ,como si fuera una sola muestra, en orden ascendente y se asignan los rangos a los valores ordenados.

2.Se identifican los valores que pertenecen a cada muestra.

3.Se calculan las Sumas de Rangos de cada muestra y se define S, que es la Suma de Rangos de menor valor y N que es el tamaño muestral de esta muestra..

4.Se determina el estadígrafo que en esta prueba es:

$$T_{o=} \quad S - \frac{N(N+1)}{2}$$

1. La Regla de Decisión será:

a) Se plantea la hipótesis que se adecue a la situación que necesitamos resolver, tomando en consideración las hipótesis siguientes:

Hipótesis	Regla de Decisión Rechazar H_O si:	α mas usados
H_o : Me_B= Me_A H_1 : $Me_B \neq Me_A$	$T_o \leq W_{\alpha/2}$ o $T_o \geq W_{1-\alpha/2}$	0.025 0.005
H_o : $Me_B \geq Me_A$ H_1 : $Me_B < Me_A$	$T_o \leq W_{\alpha}$	0.05 0.01
H_o : $Me_B \leq Me_A$ H_1 : $Me_B > Me_A$	$T_o \geq W_{1-\alpha}$ donde $W_{1-\alpha}$= NM - W_{α}	0.05 0.01

Donde W_{α} y $W_{\alpha/2}$ son los valores críticos obtenidos en la tabla de Valores críticos para la estadística de prueba de Mann- Whitney , considerando un tamaño muestral n_A , n_B y un nivel de significación dado. Esta tabla sirve para trabajar cuando el tamaño de la muestra llega hasta 20 la muestra menor y 40 la mayor.

b)Cuando el tamaño de la muestra menor excede las 20 observaciones puede trabajarse con la aproximación a la distribución normal:

$T_o = \dfrac{S - \frac{N(N+1)}{2}}{}$, se distribuye N (μ_T, σ_T) cuando n $\to \infty$

Entonces en ese caso puede usarse el estadígrafo Z

$Z = \dfrac{(To - \mu T)}{\sigma T}$ que se distribuye Normal (0, 1) ,donde :

$$\mu T = \frac{(MN)}{2}$$

y

$$\sigma T = \left[\frac{MN(M+N+1)}{12} \right]^{1/2}$$

c) Otra forma de rechazar la hipótesis nula es cuando estamos trabajando con un programa estadístico que calcule T_o y p que es la probabilidad de error de tipo I asociada a ese valor.

En éste caso si el valor de p es menor que el α prefijado se rechaza la hipótesis nula., tomando en consideración los aspectos siguientes:

Hipótesis	Regla de Decisión Rechazar H_O si:
$H_o: Me_B = Me_A$ $H_1: Me_B \neq Me_A$	$Me_B \neq Me_A$ y $p \leq 0.025$ ó $p \leq 0.005$
$H_o: Me_B \geq Me_A$ $H_1: Me_B < Me_A$	$Me_B < Me_A$ y $p \leq 0.05$ ó $p \leq 0.01$
$H_o: Me_B \leq Me_A$ $H_1: Me_B > Me_A$	$Me_B > Me_A$ y $p \leq 0.05$ ó $p \leq 0.01$

Planteamiento y solución de una situación practica.

Utilicemos el ejemplo de la prueba de la Suma de Rangos de Wilcoxon para ilustrar la prueba de Mann- Whitney. Se desea conocer si los niveles de excreción urinaria de Sodio/Potasio varían en relación a la presencia de la enfermedad X , para lo cual se seleccionaron 2 muestras aleatorias una constituida por 16 pacientes con esta enfermedad y la otra por 12 personas sin ella.

Pruebe las hipótesis siguientes:

a) Los niveles de excreción urinaria de Sodio/Potasio difieren en ambos grupos.

b) Los niveles de excreción urinaria de Sodio/Potasio en los enfermos de X son mayores que en las personas sin esta enfermedad. Considere $\alpha = 0.05$.

La información obtenida es la misma que se utilizo con la prueba de la Suma de Rangos de Wilcoxon , igual que se hizo en aquella ocasión lo primero que se hace es identificar la prueba que debe utilizarse. En éste caso tenemos: Tipo de muestras: 2 muestras independientes; Tipo de variable: cuantitativa continua; por lo que puede utilizarse la prueba de Mann- Whitney.

a) Se quiere probar si la excreción urinaria de Sodio/Potasio difiere según presencia de la Enfermedad X por lo que esta seria nuestra hipótesis alternativa:

$H_o : Me_B = Me_A$
$H_1 : Me_B \neq Me_A$

Entonces ahora asignamos los rangos a todas las observaciones como si procedieran de la misma muestra:

Enfermedad X			
Presente		Ausente	
Excreción Na/K	Rangos	Excreción NA/K	Rangos
263	13	60	4
288	17	119	6
432	20	153	10
890	27	588	25
450	21	124	8
1270	28	196	11
220	12	14	1
350	18	23	2
283	15	43	3
274	14	854	26
580	24	400	19
285	16	73	5
524	23	ΣSuma de Rangos	120
135	9		
500	22		
120	7		
ΣSuma de Rangos	286		

Nuestro estadígrafo seria entonces calculado a partir de Suma de Rangos de menor valor entre las muestras consideradas mediante la expresión:

$T_o = S - \frac{N(N+1)}{2} = 120 - \frac{12(12+1)}{2} = 42$

Se buscan entonces los valores críticos para M= 16, N= 12 y α= 0.025 por ser una prueba bilateral., y se encuentra que $W_{\alpha/2} = 54$ y $W_{1-\alpha/2} = 138$.

La Regla de Decisión nos dice que si: $T_o \leq W_{\alpha/2}$ o $T_o \geq W_{1-\alpha/2}$ rechazamos la hipótesis nula , en nuestro caso tenemos que : $T_O = 42$ es menor que $W_{\alpha/2} = 54$ por lo que se rechaza la hipótesis nula con un nivel de significación de 5% y se concluye que la excreción urinaria de Sodio/Potasio difiere según la presencia de la Enfermedad X .

b)Se quiere probar que la excreción urinaria de Sodio/Potasio es menor en las personas sin la Enfermedad X por lo que esta seria nuestra hipótesis alternativa:

$H_o : Me_B \geq Me_A$
$H_1 : Me_B < Me_A$

Observe como en éste ejemplo se plantean las hipótesis en función de la muestra que tiene la Suma de Rangos de menor valor o sea en éste caso la muestra de personas sin la

enfermedad. Como el valor del estadígrafo de Mann –Whitney es 42, ahora tendríamos que buscar los valores críticos para M=16, N= 12 y α= 0.05 por ser una prueba unilateral., encontramos entonces que $W_{\infty} = 61$ y $W_{1\ -\infty} = 131$.

La Regla de Decisión nos dice que sí: $T_0 \leq W_{\infty}$ se rechaza la hipótesis nula , en nuestro caso tenemos que : $T_0 = 42$ es menor que $W_{\infty} = 61$ por lo que se rechaza la hipótesis nula con un nivel de significación de un 5% y se concluye que la excreción urinaria de Sodio/Potasio es menor en las personas que no tienen la Enfermedad X .

En el caso de éste ejemplo al calcular el estadígrafo con un programa estadístico obtenemos los resultados siguientes:

Muestra	Numero de casos	Rangos Promedios	$T_O = 42$
B	12	10.00	
A	16	17.88	

p=0.012 (menor que 0.05)$\Rightarrow$ Rechazo la hipótesis nula.

Debe destacarse que ambas pruebas producen los mismos resultados, y que tiene la misma Potencia – Eficiencia que la prueba de la Suma de Rangos de Wilcoxon.

3.5 PRUEBA DE KRUSKAL WALLIS.

En ocasiones se presenta el problema de comparar mas de 2 muestras con el propósito de conocer si proceden de la misma población o si hay diferencias entre las medidas de tendencia central de mas de 2 poblaciones.

Cuando la variable estudiada es cuantitativa continua, se distribuye normal y además sus varianzas son iguales y existe independencia entre las observaciones el problema se resolvería mediante un Análisis de Varianza de una Vía (ANOVA 1 VIA), sin embargo cuando no se cumplen estos supuestos o no estamos seguros de que se cumplan puede realizarse la prueba de Kruskal Wallis como alternativa no paramétrica para resolver el problema , esta prueba se considera como una generalización de la prueba de la Suma de Rangos de Wilcoxon como se vera a continuación en su desarrollo.

Sean :

- X una variable ordinal que estudiamos y se observa en K muestras independientes, nuestro problema es determinar si las K muestras difieren en su localización, o sea en sus medidas de tendencia central.
- K el numero de muestras independientes existentes.
- N el total de observaciones en las K muestras.
- n_j el numero de observaciones en la j-esima muestra, y j toma valores desde 1 hasta K, de forma que puede tener diferentes o iguales tamaños para cada muestra.

 X_{ij} el valor de la i-esima observación, i toma valores desde 1 hasta n, de la j-esima muestra.

Entonces las observaciones de las K muestras pueden disponerse en una tabla como la siguiente:

INDIVIDUOS	MUESTRAS				
	1	2	3	....	K
1	X_{11}	X_{12}	X_{13}	....	X_{1K}
2	X_{21}	X_{22}	X_{23}	....	X_{2K}
3	X_{31}	X_{32}	X_{33}	.....	X_{3K}

				...	
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
n	X_{n1}	X_{n2}	X_{n3}	.....	X_{nk}

Ahora se puede hacer una asignación de rangos a cada valor , considerando las K muestras como si fuera una sola, entonces se puede pasar de la tabla anterior, de las observaciones , a una tabla similar pero de los rangos asignados, considerando R_{ij} al rango asignado a la observación X_{ij} ,, como puede observarse en la tabla siguiente:

INDIVIDUOS	MUESTRAS				
	1	2	3	.....	K
1	R_{11}	R_{12}	R_{13}	.....	R_{1K}
2	R_{21}	R_{22}	R_{23}	.....	R_{2K}
3	R_{31}	R_{32}	R_{33}	.....	R_{3K}
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
n	R_{n1}	R_{n2}	R_{n3}	.....	R_{nk}
Suma de rangos	$R._1$	$R._2$	$R._3$		$R._K$

Donde $R._j = \sum R_{ij}$ es la Suma de rangos de la muestra J.

Nos interesa docimar las hipótesis siguientes:

1) H_o : Las K poblaciones están distribuidas de igual forma.

H_1 : Al menos 1 población difiere de las restantes.

Detrás de estas hipotesis subyacen las siguientes:

2)) H_o : $Me_1 = Me_2 = \ldots\ldots\ldots\ldots = Me_K$

H_1 : $Me_i \neq Me_j$ al menos un par i,j

Entonces sean:

$\check{R}_j = R.j/nj$ el rango promedio de la muestra j.

$$\check{R} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} nij}{\sum_{j=1}^{k} nj} = \frac{N+1}{2}$$ es el promedio general de los rangos.

Si las K muestras proceden de la misma población, en el sentido de que todas las muestras tengan la misma localización, entonces el rango promedio de cada una de las muestras no debe ser diferente entre las mismas ni diferir del promedio general de los rangos de forma que las variaciones que existan deben ser atribuidas al azar.

Se define entonces una medida de la dispersión entre $\check{R}_j$ y $\check{R}$ mediante la siguiente expresión:

$D = \left(\frac{R._J}{n._j} - \frac{N+1}{2} \right)^2$ al desarrollar esta expresión tendremos:

D= $\sum_{j=1}^{k} nj\left(\frac{R.j}{nj} - \frac{N+1}{2} \right)^2$ es la suma de los cuadrados de las desviaciones entre la media muestral y la general de los rangos ponderado por el tamaño de cada muestra.

De esta forma puede definirse el estadígrafo de Kruskal -Wallis por la expresión:

H = $\frac{D}{N(N+1)/12} = \frac{12}{N(N+1)}\left[\sum_{j=1}^{k} \frac{R.j^2}{nj} \right] - 3(N+1)$ entonces

H = $\frac{12}{N(N+1)}\left[\sum_{j=1}^{k} \frac{R.j^2}{nj} \right] - 3(N+1)$ si tenemos 3 o más muestras mayores que 5, éste Estadígrafo se distribuye X^2 con K-1 grados de libertad.

Cuando solo hay 3 grupos con 5 o menos observaciones cada uno , existe una tabla para determinar la significación asociada al valor de H.

Si las muestras no difieren en la forma de su distribución podemos esperar que D sea casi nulo, pues éste termino mide las diferencias entre los rangos promedios de cada muestra y el promedio general de los rangos de todas las muestras, por ende el valor de H será muy pequeño.

De igual forma si el valor de D es elevado esto significa que existen grandes diferencias entre los rangos promedios de cada muestra y el promedio general de los rangos de todas las muestras, por ende la regla de decisión aquí seria la siguiente:

a) Cuando solo hay 3 grupos con 5 o menos observaciones cada uno , existe una tabla para determinar la probabilidad exacta asociada al valor de H.

Se busca el valor critico de H_t en la tabla tomando en consideración los tamaños de n1 , n2 y n3 ,así como el α con que estamos trabajando:

Si H es mayor que H_t se rechaza la hipótesis nula.

b) Si tenemos 3 o más muestras mayores que 5, H se distribuye X^2 con K-1 grados de libertad, entonces éste valor se busca en la tabla de acuerdo al nivel de significación que estamos considerando, en éste caso la regla de decisión será:.

Si H es mayor que X^2 se rechaza la hipótesis nula.

c) Si calculamos H en mediante un programa estadístico en una computadora obtendremos el valor de H con la probabilidad exacta de error de tipo I asociada a ese valor.

Debe señalarse que la Potencia-eficiencia de esta prueba comparada con la prueba paramétrica mas poderosa, la prueba F, considerando que se cumplen los supuestos para la misma es de un 95.5 %.

Planteamiento y solución de una situación practica.

Se desea determinar si las cifras que excretan en orina de sodio y potasio 4 tipos de ratas difieren entre si, para lo cual se hicieron las determinaciones que se expresan a continuación:

RAZAS			
A	B	C	D
2.80	2.81	1.25	1.32
2.72	2.54	1.42	1.64

2.93	1.97	1.27	1.89
2.62	1.62	1.20	1.12
2.32	1.70	1.32	1.23

Como hay mas de 2 muestras independientes con una variable cuantitativa y no se hacen supuestos sobre la distribución de la variable ni sobre la igualdad de sus varianzas, se decide entonces hacer una prueba de Kruskal Wallis. Las hipotesis a docimar serían las siguientes:

H_o : $Me_1 = Me_2 = \ldots\ldots\ldots\ldots = Me_K$

H_1 : $Me_i \neq Me_j$ al menos un par i,j

Entonces asignaríamos los rangos a las observaciones:

RAZAS							
A		B		C		D	
VALOR	RANGOS	VALOR	RANGOS	VALOR	RANGOS	VALOR	RANGOS
2.80	18	2.81	19	1.25	4	1.32	6.5
2.72	17	2.54	15	1.42	8	1.64	10
2.93	20	1.97	13	1.27	5	1.89	12
2.62	16	1.62	9	1.20	2	1.12	1
2.32	14	1.70	11	1.32	6.5	1.23	3
$R._A$	85	$R._B$	67	$R._C$	25.5	$R._D$	32.5

Entonces se utiliza el estadígrafo de Kruskal Wallis para determinar si los valores medios de los 4 tipos de ratas difieren entre si:

$$H = \frac{12}{N(N+1)}\left[\sum_{j=1}^{k}\frac{R.j^2}{nj}\right] - 3(N+1)$$

Sustituyendo los valores tendremos:

$$H = \frac{12}{20x21}\left[\frac{(85)^2}{5}+\frac{(67)^2}{5}+\frac{(25.5)^2}{5}+\frac{(32.5)^2}{5}\right]-(3x21) = 13.69$$

Como hay 4 muestras puede decirse que éste estadígrafo se distribuye X^2 con K-1 grados de libertad, así que buscamos en la tabla el valor de X^2 con 3 grados de libertad para un nivel de significación de 5% y encontramos el valor 7.81

.La regla de decisión seria: Si H es mayor que el valor tabulado de X^2 con 3 grados de libertad para un nivel de significación de 5% , rechazamos la hipótesis nula.

En éste caso seria H= 13.69 es mayor que 7.81, por lo que se rechaza la hipótesis nula y se concluye que con una confiabilidad de 95% se puede afirmar que las cifras que excretan en orina de sodio y potasio los 4 tipos de ratas difieren entre si.

En el caso de éste ejemplo al calcular el estadígrafo con un programa estadístico obtuvimos los resultados siguientes:

Muestra	Numero de casos	Rangos Promedios	H= 13.699
A	5	17.00	grados de libertad = 3
B	5	13.40	p=0.003 (menor que 0.05) $\Rightarrow$
C	5	5.10	Rechazo la hipótesis nula
D	5	6.50	

3.5 <u>EJERCICIOS</u>.

1. Resuelva el inciso a del ejemplo de la prueba de la Suma de Rangos de Wilcoxon utilizando la prueba de la Mediana.

a) Interprete los resultados.

b) A que atribuye estos resultados?. Fundamente su respuesta.

2. Se desea determinar si la edad de las personas fallecidas por Enfermedad Cerebro Vascular difiere de la de personas que sobrevivieron esa enfermedad.

Se tomo una muestra de fallecidos (F*) y una muestra de egresados vivos (E*) por esa causa en un hospital, obteniéndose la siguiente información:

F*	99	98	85	73	83	88	99	80	74	91	80	94	95	97	80
E*	78	74	69	79	57	78	79	68	59	91	89	55	60	55	79

Es posible concluir con un nivel de significación de un 5% que las edades de esas 2 poblaciones son diferentes? Considere $\alpha = 0.05$

3. Se necesita probar la efectividad de 3 productos químicos para la eliminación de moscas en las viviendas. Con cada una de ellos se fumigan en 6 ocasiones las viviendas de un área de salud y luego de pasar 1 hora cada fumigación se registra el porcentaje de moscas muertas en las viviendas, obteniéndose los resultados siguientes:

PRODUCTOS QUIMICOS		
A	B	C
72	55	64
65	59	74
67	68	61
75	70	58
62	53	51
73	50	69

Cree usted que hay diferencias en la efectividad de estos 3 productos químicos para la eliminación de moscas en las viviendas?

Considere $\alpha = 0.05$

4. Se realizaron determinaciones de glicemia en sangre a 5 grupos de pacientes sometidos a diferentes tratamientos antigluciemiantes durante 3 meses y se desea conocer al terminar el ciclo de tratamiento si los niveles de glicemia en sangre de estos pacientes difieren. Los resultados se brindan a continuación:

TRATAMIENTOS ANTIGLUCEMIANTES				
A	B	C	D	E
124	111	117	104	142
116	101	142	128	139
101	130	121	130	133
118	108	123	103	12
120	127	121	121	127
110	129	148	119	149
127	122	141	106	150
106	103	122	107	149
130	122	139	107	120
118	127	125	115	116

Proporcionan estos datos evidencia suficiente que indique con una confiabilidad de un 95 % una diferencia en el nivel medio de glucosa en sangre en las personas sometidas a los diferentes tratamientos?

3.6 BIBLIOGRAFIA.

1. López Pardo, Cándido: Notas de Clase. Curso La Estadística no paramétrica en la investigación contemporánea. Facultad de Economía. Universidad de la Habana, 1990.
2. López Pardo, Cándido: Dócimas de hipótesis no Paramétricas para datos ordinales y de mayor nivel: Caso 3 o más muestras independientes o pareadas. Departamento de Estadística Aplicada. Facultad de Economía .Universidad de la Habana, 1990.
3. Verdecía, Dominador: Notas de Clase. Especialidad de Bioestadistica. Instituto de Desarrollo de la Salud. Ministerio de Salud Publica, 1980.
4. Daniel, W. W.: Bioestadistica: base para el análisis de las ciencias de la salud. Editorial UTEHA, México, 1997.
5. Hollander, M y D. A. Wolfe: Nonparametrics Statistical Methods. John Wiley and Sons, New York, 1973.
6. Mosteller, F y R. E. Rourke: Sturdy Statistics: Nonparametrics and Order Statistics. Addison Wesley, Massachusetts, 1973.
7. Siegel, S.: Análisis Experimental no paramétrico. Instituto Cubano del Libro, Edición Revolucionaria, La Habana, 1972.
8. Bradley, J. V.: Distribution Free Statistical Test. Prentice Hall, New Jersey. 1968.

4. CORRELACION.

Con frecuencia en las ciencias de la salud se necesita conocer si 2 variables están relacionadas y si lo están es importante evaluar la intensidad de la relación.

En las estadísticas clásicas la medida de correlación usada es el Coeficiente de Correlación de Pearson que se estima por el método de los mínimos cuadrados que requiere:

1. 1. Que las variables estén medidas en una escala continua.
2. 2. Si se quiere medir su significación estadística es necesario que estas variables tengan una distribución Normal bivariada.

Cuando estos supuestos no se cumplen pueden utilizarse Medidas de Correlación no Paramétricas, que permiten además de medir la correlación existente, determinar su significación estadística.

Hay varios coeficientes de correlaciona no paramétricos, a continuación expondremos el Coeficiente de Correlación de Spearman, que fue la primera estadística de rangos que se desarrollo y que requiere que ambas variables estén medidas al menos en una escala ordinal de forma que las unidades experimentales puedan colocarse en 2 series ordenadas y que las observaciones se hayan extraído aleatoriamente de una población para poder realizar la prueba de significación de éste coeficiente.

4.1 COEFICIENTE DE CORRELACION DE SPEARMAN.

Sean n observaciones de 2 variables aleatorias X e Y, medidas ambas al menos en escala ordinal de forma simultanea en una misma unidad de observación.

Los datos así obtenidos pueden registrarse de la forma siguiente:

INDIVIDUOS	VARIABLES	
	X	Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
...	...	....
n	x_n	y_n

Observe que cada individuo queda representado por un par de valores de las variables X eY, o sea queda representado por el punto (x_i , y_i), que posteriormente puede ser llevado a un gráfico denominado Diagrama de Dispersión que se utiliza para explorar la relación lineal entre las 2 variables .

Entonces a cada valor de la variable X se asigna un rango r_{i1} y a cada valor de la variable Y se asigna un rango r_{i2}, convirtiéndose la matriz de información en una matriz de rangos como la siguiente:

INDIVIDUOS	VARIABLE X	RANGO DE X	VARIABLE Y	RANGO DE Y
1	X_1	R_{11}	Y_1	R_{21}
2	X_2	R_{12}	Y_2	R_{22}
3	X_3	R_{13}	Y_3	R_{23}
4	X_4	R_{14}	Y_4	R_{24}
...	.....	.....	.....	.....
n	X_n	R_{1n}	Y_n	R_{2n}

Se necesita entonces medir las discrepancias que hay entre los rangos de X y los de Y para cada individuo, definimos entonces a $d_{ij}= r_{1j} – r_{2j}$.

- Si la relación que hay entre las 2 variables es perfecta positiva, entonces ambas variables están ordenadas de igual forma y toman los mismos rangos para cada individuo , en éste caso todas las $d_{ij}= r_{1j} – r_{2j}$ son nulas →que a valores altos de una variable corresponden valores altos de la otra..
- Si la relación que hay entre las 2 variables es perfecta negativa ,entonces ambas variables están ordenadas de forma inversa y toman los rangos asignados de forma inversa para cada individuo , en éste caso se obtendrían las $d_{ij}= r_{1j} – r_{2j}$ máximas posibles →que a valores altos de una variable corresponden valores bajos de la otra..
- Debe recordarse que cuando analizamos un coeficiente de correlación se deben considerar 2 aspectos:

1. Su magnitud: Oscila entre 0 y 1, valores cercanos a 0 significan que no hay relación lineal simple entre ambas variables, mientras que los valores cercanos a 1 significan que la relación lineal simple entre las variables es muy intensa.

2. Su sentido: Esta dado por el signo + o – del coeficiente, el signo – significa que existe una relación inversamente proporcional, o sea a valores altos de una variable corresponde valores bajos de la otra.

El signo + significa que existe una relación directamente proporcional entre las variables, o sea a valores altos de una se corresponden valores altos de la otra y viceversa.

La formula del Coeficiente de Correlación de Rangos de Spearman se deduce a partir de la del Coeficiente de Correlación de Pearson, sustituyendo los valores de las observaciones de cada variable por la de sus rangos, llegando a obtener la siguiente expresión:

$R_s= 1-\frac{6\sum_{i=1}^{n} d_i^2}{n^3-n}$, de esta forma R_s brinda una estimación del grado de asociación entre las variables X y Y.

4.2 PRUEBA DE SIGNIFICACION PARA EL COEFICIENTE DE CORRELACION DE SPEARMAN.

Una vez estimado el Coeficiente de Correlación de Spearman es necesario determinar si éste valor puede ser inferido a la población de la cual se extrajo la muestra aleatoria.

Procedimientos:

Sean X y Y , 2 variables aleatorias medidas en n individuos , observe que cada individuo queda representado por un par de valores de las variables X y Y.

1. Entonces se ordenan las observaciones de ambas variables, asignando los rangos a los valores ordenados de cada una de ellas, o sea a cada valor de la variable X se asigna un rango r_{i1} y a cada valor de la variable Y se asigna un rango r_{i2} .
2. Se define $d_{ij}= r_{1j} – r_{2j}$ .que mide las discrepancias que hay entre los rangos de X y los de Y para cada individuo.
3. Se sustituyen los valores obtenidos en la siguiente expresión:

Coeficiente de Correlación de Rangos de Spearman = R_s = $1-\frac{6\sum_{i=1}^{n} d_i^2}{n^3-n}$

4. Se plantea la hipótesis que se adecue a la situación que necesitamos resolver, tomando en consideración las hipótesis siguientes:

HIPOTESIS	REGLA DE DECISION RECHAZAR H_O SI:	α mas usados
$H_o : R_s = 0$ $H_1 : R_s \neq 0$	$-R_s \leq -R_{tabulada}$ o $R_s \geq R_{tabulada}$	0.025 0.005
$H_o : R_s \geq 0$ $H_1 : R_s < 0$	$- R_s \leq - R_{tabulada}$	0.05 0.01
$H_o : : R_s \leq 0$ $H_1 : : R_s > 0$	$R_s \geq R_{tabulada}$	0.05 0.01

a) Donde $- R_{tabulada}$ y $R_{tabulada}$ son los valores críticos obtenidos en la tabla de Valores críticos para el Coeficiente de Correlación de Rangos de Spearman R_s, para un tamaño de muestra que oscile entre 4 y 30 y un nivel de significación dado.

b)Cuando el tamaño de la muestra excede las 30 observaciones puede trabajarse entonces con la aproximación a la distribución normal:

Z= $R_S\sqrt{n-1}$ que se distribuye Normal (0, 1)

c)Otra forma de rechazar la hipotesis nula es cuando estamos trabajando con un programa estadístico entonces se calcula R_s, y p que es la probabilidad de error de tipo I asociada a ese valor.

En éste caso si el valor de p es menor que el α prefijado se rechaza la hipótesis nula., tomando en consideración los aspectos siguientes:

HIPOTESIS	REGLA DE DECISION RECHAZAR H_O SI:
$H_o : R_s = 0$ $H_1 : R_s \neq 0$	$R_s \neq 0$ y $p \leq 0.025$ ó $p \leq 0.005$

$H_o: R_s \geq 0$ $H_1: R_s < 0$	$R_s < 0$ y $p \leq 0.05$ ó $p \leq 0.01$
$H_o: : R_s \leq 0$ $H_1: : R_s > 0$	$: R_s > 0$ y $p \leq 0.05$ ó $p \leq 0.01$

5. Debe aclararse que en caso de existir observaciones de igual valor se asignaran rangos promedios, por ejemplo si los 3 primeros valores son:

VALORES	15	15	15
LUGARES	1	2	3
RANGOS	1+2+3/3 =**2**	1+2+3/3= **2**	1+2+3/3= **2**

Potencia - Eficiencia:
Según Siegel , si la prueba de Correlación de Rangos de Spearman se aplica a datos que pueden analizarse adecuadamente por una prueba paramétrica mas poderosa, como la prueba del Coeficiente de Correlación de Pearson en éste caso, su potencia-eficiencia seria de cerca de un 91%.

Planteamiento y solución de una situación práctica.
Se desea conocer si existe una relación lineal entre la creatinina y el peso de las embarazadas, para evaluar lo anterior se selecciono una muestra aleatoria de 10 embarazadas obteniéndose los resultados siguientes:

EMBARAZADA	CREATININA (X)	PESO (Y)
1	22	122
2	19	129
3	16	115
4	20	132
5	21	119
6	19	110
7	22	127
8	23	142
9	25	137
10	24	140

Considere $\alpha = 0.05$

1. Como se quiere conocer si existe correlación entre 2 variables cuantitativas y se desconoce si tienen una distribución Normal bivariada, la prueba de elección debe ser la del Coeficiente de Correlación de Spearman. En éste caso también favorece esta elección el hecho de tener solo 10 observaciones.
2. Las hipotesis a docimar son las siguientes:

$$H_o: R_s = 0$$
$$H_1: R_s \neq 0$$

3. A partir de los datos anteriores se construye una tabla de trabajo como la siguiente:

EMBARAZADA	CREATININA X	PESO Y	RANGO X	RANGO Y	Di	D_i^2
1	22	122	6.5	4	2.5	6.25
2	19	129	2.5	6	-3.5	12.25
3	16	115	1	2	-1	1.00
4	20	132	4	7	-3	9.00
5	21	119	5	3	2	4.00
6	19	110	2.5	1	1.5	2.25
7	22	127	6.5	5	1.5	2.25
8	23	142	8	10	-2	4.00
9	25	137	10	8	2	4.00
10	24	140	9	9	0	0.00
				$\sum$		45.00

4. La Regla de Decisión:

Si $-R_s \leq -R_{tabulada}$ o $R_s \geq R_{tabulada}$ rechazo H_o

5. $R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6x45}{1000-10} = 1 - \frac{270}{990} = 0.7273 \cong 0.73$

Al buscar en la tabla se encuentra que $R_{tabulado}$ para n=10 y α=0.025 es 0.6364

Como $R_s = 0.73 \geq R_{tabulada} = 0.66364$ rechazo H_o

6. Se concluye que existe una Correlación lineal significativa en la población de embarazadas entre la creatinina y el peso de la embarazada.

Si queremos determinar si existe una correlación positiva entre las variables estudiadas en esa población plantearíamos las hipótesis siguientes:

$H_o: R_s \leq 0$

$H_1: R_s > 0$

Buscaríamos entonces el valor de $R_{tabulado}$ para n=10 y α=0.05 que es 0.5515

El valor del estadígrafo seria el mismo R_s = 0.73, entonces para esa hipótesis la regla de decisión sería:

Si $R_s = 0.73 \geq R_{tabulada} = 0.5515$ Rechazo H_o

Entonces podemos afirmar que entre la creatinina y el peso de las embarazadas existe una relación lineal directamente proporcional significativa., o sea a valores grandes de creatinina se evidencian también pesos elevados y a pesos pequeños se observan valores pequeños de creatinina , éste resultado puede inferirse a la población de embarazadas de donde fue seleccionada la muestra con una confiabilidad de un 95 %.

Si utilizamos un programa estadístico obtenemos los resultados siguientes:

R_s,= 0.726 y p =0.018 que es la probabilidad de error de tipo I asociada a ese valor.

En éste caso como el valor de p es menor que el α prefijado se rechaza la hipotesis nula., tomando en consideración los aspectos siguientes:

HIPOTESIS	RESULTADOS OBTENIDOS	REGLA DE DECISION RECHAZAR H_O SI:
$H_o: R_s = 0$ $H_1: R_s \neq 0$	$R_s = 0.726$ P=0.018	$R_s \neq 0$ y $p \leq 0.025$ ó $p \leq 0.005$

H_o : : $R_s \leq 0$ H_1 : : $R_s > 0$	$R_s = 0.726$ P = 0.009	: $R_s > 0$ y $p \leq 0.05$ ó $p \leq 0.01$

Por lo que en ambos casos rechazamos la hipótesis nula y llegamos a los mismos resultados obtenidos anteriormente.

4.3 EJERCICIOS.

1**.** A continuación se relacionan los valores obtenidos en una muestra de 12 embarazadas cuyas edades oscilan entre 30 y 45 años.

EMBARAZADAS	EDAD GESTACIONAL *	CREATININA URINARIA**	CONSUMO DE OXIGENO***
1	8	104	133
2	12	96	124
3	17	87	125
4	19	98	120
5	22	81	121
6	23	93	128
7	25	90	110
8	29	112	130
9	32	100	112
10	34	94	151
11	37	120	117
12	40	116	135

*Expresado en semanas. ** Expresado en mg/24 horas. *** Expresado en cm^3 /mm/m^2 de superficie corporal.

a) Mida la correlación que existe entre: a_1) Edad gestacional y Creatinina; a_2) Edad gestacional y Consumo de Oxigeno; a_3) Creatinina y Consumo de Oxigeno.

b) Determine si los valores hallados: b_1) Difieren significativamente de cero; b_2) Son significativamente mayores que cero; c)Interprete todos los resultados obtenidos para α=0.05.

4. BIBLIOGRAFIA.

1. Daniel, W.: Bioestadistica: base para el análisis de las ciencias de la salud. Editorial UTEHA, México, 1997.
2. Siegel, S.: Análisis Experimental no paramétrico. Instituto Cubano del Libro, Edición Revolucionaria, La Habana, 1972.
3. López Pardo, Cándido: Notas de Clase. Curso La Estadística no paramétrica en la investigación contemporánea. Facultad de Economía. Universidad de la Habana, 1990.
4. Verdecía, Dominador: Notas de Clase. Especialidad de Bioestadistica. Instituto de Desarrollo de la Salud. Ministerio de Salud Publica, 1980.
5. Hollander, M y D. A. Wolfe: Nonparametrics Statistical Methods. John Wiley and Sons, New York, 1973.
6. Dawson-Saunders, B y R. G. Trapp: Bioestadistica Medica. Editorial El Manual Moderno, S . A. ,México, D. F.,1995.