



Matemáticas

Estadística II

4° Semestre

Unidad 1. Estadística no paramétrica y pruebas de Bondad de Ajuste

Clave:

05142421/06142421

Universidad Abierta y a Distancia de México





Estadística II

Unidad 1. Estadística no paramétrica

Índice

Unidad 1. Estadística no Paramétrica y Pruebas de Bondad de Ajuste	4
Presentación de la unidad	4
Competencia específica	4
Logros de la unidad	4
1.1 Utilidad de las pruebas no paramétricas	4
1.2. Pruebas para una sola población	5
1.2.1. Prueba Binomial para una sola muestra	5
1.2.2. Prueba de la tendencia Cox Stuart	11
1.3. Pruebas para dos poblaciones independientes	16
1.3.1. Prueba U de Mann-Whitney	16
1.3.2. La prueba de la mediana	21
1.3.3. Prueba de rachas Wald-Wolfowitz	24
1.3.4. Prueba de Mac Nemar	27
1.4. Pruebas para dos poblaciones independientes	30
1.4.1. Prueba de signos	30
1.4.2. Prueba de Wilcoxon	32
1.5. Prueba de independencia y homogeneidad	35
1.5.1. Tablas de contingencia	35
1.5.2. Prueba de independencia con Ji-Cuadrada	38
1.6. Prueba de tres o más poblaciones independientes	41
1.6.1. Extensión de la prueba de la mediana	41
1.6.2. Comparación de varias poblaciones Kruskal-Wallis	44
1.7. Prueba de Bondad de Ajuste	46
1.7.1. Prueba de bondad y ajuste basada en Ji-Cuadrada	47
1.7.2. Prueba de Kolmogorov-Smirnov para una muestra	49



Estadística II

Unidad 1. Estadística no paramétrica

1.7.3. Prueba de Kolmogorov-Smirnov para dos muestras	53
1.7.4. Otras pruebas de bondad y ajuste.....	56
Cierre de la unidad	58
Para saber más	58
Fuentes de consulta	58



Estadística II

Unidad 1. Estadística no paramétrica

Unidad 1. Estadística no Paramétrica y Pruebas de Bondad de Ajuste

Presentación de la unidad

Cuando se habla de estadística paramétrica lo que se pretende es estimar, probar hipótesis acerca de uno o más parámetros de la población. En esos casos se tenía el conocimiento de la distribución de la población de la cual se extrajo la muestra.

Al hablar de estadística no paramétrica por convención se entenderán dos cosas: primero será la estadística no paramétrica propiamente que son aquellos procedimientos que no son afirmaciones de los parámetros, y segundo los procedimientos de libre distribución como aquellos en que no hacen supuesto alguno acerca de la población de la cual se extrae la muestra.

Competencia específica

Utilizar las pruebas no paramétricas para resolver problemas estadísticos de diversas poblaciones determinando sus características.

Logros de la unidad

- Identificar las diversas pruebas no paramétricas y de bondad de ajuste.
- Determinar las pruebas paramétricas y no paramétricas.
- Establecer los elementos necesarios para cada prueba no paramétrica y de bondad de ajuste.
- Aplicar pruebas no paramétricas y de bondad de ajuste.

1.1 Utilidad de las pruebas no paramétricas

La ventaja de las pruebas no paramétricas consiste en que requieren pocos supuestos acerca de la población de la cual provienen los datos. En particular olvidan el supuesto tradicional de que los datos provienen de una distribución Normal.

Lo anterior quiere decir que pueden aplicarse cuando los datos que sirven para el análisis constan simplemente de categorías o clasificaciones, es decir, los datos pueden no estar basados en una escala de medición lo suficientemente sólida como para permitir las operaciones aritméticas necesarias para llevar a cabo los procedimientos necesarios.



Estadística II

Unidad 1. Estadística no paramétrica

También son procedimientos más fáciles de usar que la contraparte en la teoría Normal y usualmente son más fáciles de entender.

Aunque es recomendable utilizar los procedimientos paramétricos cuando sea posible para evitar un desperdicio de información.

La aplicación de algunas pruebas no paramétricas pueden ser muy laboriosas, lo que es una desventaja cuando se tienen muestras grandes.

1.2. Pruebas para una sola población

En tus cursos anteriores de estadística has estudiado los tipos de variables que existen. Como las pruebas que estudiaremos en esta unidad están enfocadas a diferentes tipos de variables daremos un pequeño repaso de ellos.

Llamamos **medición** al número que asignamos a los objetos de acuerdo a un conjunto de reglas. Las cuatro principales escalas de medición son:

- **Escala nominal:** Clasifica las observaciones en varias categorías mutuamente excluyentes y colectivamente exhaustivas. Por ejemplo:
 - Masculino-Femenino
 - Sano-Enfermo
 - Menores o iguales a 56 años- Mayores a 56 años
- **Escala ordinal:** Difieren de categoría a categoría y además pueden clasificarse por grados de acuerdo con algún criterio. Por ejemplo:
 - Los pacientes convalecientes pueden clasificarse como: sin memoria, mejorados y bastante mejorados
 - El estado socioeconómico: alta, media, baja
- **Escala de intervalos:** Se conoce la distancia entre dos mediciones cualesquiera, posee una distancia unitaria y un punto cero los cuales son arbitrarios
 - La diferencia entre una medida de 20 y 30 es equivalente a la de 40 y 30.
- **Escala de razones:** Posee un punto cero propio como origen, es decir, que el valor cero significa ausencia de la magnitud que estamos midiendo. Como la estatura, la edad.

1.2.1. Prueba Binomial para una sola muestra



Estadística II

Unidad 1. Estadística no paramétrica

En esta prueba el investigador busca comparar las frecuencias observadas de cada categoría de una variable dicotómica con la esperada en una población binomial y con ello poder hacer inferencia acerca de la población total

Datos

Los datos consisten de resultados dicotómicos provenientes de una distribución binomial con probabilidades constantes de éxito p en base a estos resultados podemos hacer inferencia sobre p .

Por ejemplo:

- Un analista de mercado quiere conocer la proporción de familias en una cierta región que tienen televisión de paga
- Un sociólogo quiere conocer la proporción de cabezas de familia que sean mujeres
- El político querrá conocer la proporción de simpatizantes hacia su partido en una cierta región

Suponemos que una población de tamaño n tienen sólo 2 elementos: Tipo A y Tipo B. La proporción del Tipo A se designa con p y $1 - p = q$ denota la proporción de elementos del Tipo B. Sea r el número de elementos Tipo A en la muestra

Supuestos:

- Los resultados en cada ensayo pueden ser clasificados como éxito o fracaso (Tipo A y Tipo B)
- La probabilidad de éxito, denotada por p , permanece constante de ensayo a ensayo
- Los n ensayos son independientes

Hipótesis:

A. $H_0: p = p_0$	vs	$H_1: p \neq p_0$
B. $H_0: p \leq p_0$	vs	$H_1: p > p_0$
C. $H_0: p \geq p_0$	vs	$H_1: p < p_0$

Estadístico de prueba:

Como se busca que los resultados sean éxitos, entonces, el estadístico de prueba será:

$$T = r$$

con r número de éxitos, es decir, r denota los elementos Tipo A en la muestra. Entonces la distribución de r es $Bin(p_0, n)$.



Estadística II

Unidad 1. Estadística no paramétrica

Regla de decisión:

- A. Para valores suficientemente grandes o valores suficientemente pequeños de T la región crítica bajo H_0 es:

$$p(r \leq t_2) \approx \alpha/2 \quad \text{y} \quad p(r > t_1) \approx \alpha/2$$

Por lo tanto rechazamos H_0 si $T \leq t_2$ ó $T > t_1$.

- B. Para valores muy grandes de T significa que H_0 es falsa. La región crítica consiste en todos los valores de T mayores a t_1 , en términos probabilísticos la región de rechazo es aquella que cumple

$$p(r > t_1) = \alpha$$

Por lo tanto, rechazamos H_0 al nivel de significancia α si: $T > t_1$

- C. Para valores muy pequeños de T significa que H_0 es falsa. La región crítica es:

$$p(r \leq t_2) = \alpha$$

Por lo tanto, rechazamos H_0 al nivel de significancia α si: $T \leq t_2$

Aproximación a una distribución Normal

La distribución exacta de T puede ser obtenida de la siguiente ecuación:

$$T = \sum_{j=1}^n \Psi_i$$

Donde:

$$\Psi_i = \begin{cases} 1, & \text{si el } i - \text{ésimo ensayo Bernoulli es éxito} \\ 0, & \text{si el } i - \text{ésimo ensayo Bernoulli es fracaso} \end{cases}$$

Cuando H_0 es cierta

$$E_{p_0}(T) = np_0$$

Y usando el hecho de que $\Psi_1, \Psi_2, \dots, \Psi_n$ son independientes

$$Var_{p_0}(T) = np_0(1 - p_0)$$

Si ahora utilizamos el Teorema Central del Límite cuando $n \rightarrow \infty$



Estadística II

Unidad 1. Estadística no paramétrica

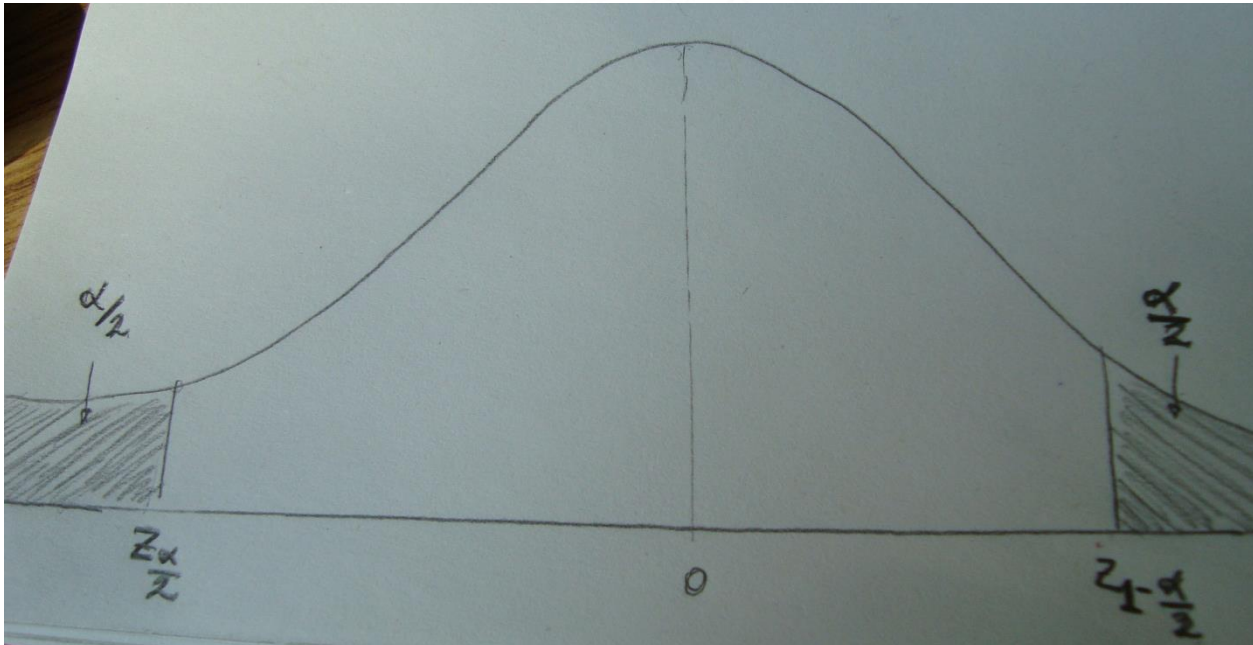
$$T^* = \frac{T - np_0}{[np_0(1 - p_0)]^{1/2}} \sim N(0,1),$$

Si Z_α denota el percentil superior de una $N(0,1)$. La aproximación normal para las reglas de decisión es:

- A. Rechaza H_0 si $|T^*| \geq Z_{\alpha/2}$
- B. Rechaza H_0 si $T^* \geq Z_\alpha$
- C. Rechaza H_0 si $T^* \leq Z_\alpha$

Intervalos de confianza

Sea $Z_{1-\alpha/2}$ el cuantil de una $N(0,1)$ y tenemos que $Z_{\alpha/2} = Z_{1-\alpha/2}$



Nombre gráfica: Cuantiles $Z_{\alpha/2}$ y $Z_{1-\alpha/2}$ de una distribución $N(0,1)$

Construimos el intervalo de confianza

$$1 - \alpha = P\left(-Z_{1-\frac{\alpha}{2}} < \frac{T - np}{\sqrt{npq}} < Z_{1-\frac{\alpha}{2}}\right)$$

Despejando a p



Estadística II

Unidad 1. Estadística no paramétrica

$$1 - \alpha = P\left(\frac{T}{n} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} < p < \frac{T}{n} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{pq}{n}}\right)$$

Ejemplo

El dueño de la pequeña empresa X de instalación de boilers afirma que instala más del 65% en las casas de una cierta colonia. Se muestrean 12 casas y se les pregunta el nombre de la empresa que instaló el boiler en su casa. En 10 casas coinciden con la instalación de la empresa X. En base a esta evidencia ¿Estaría de acuerdo con la afirmación del dueño con un nivel de significancia $\alpha = 0.05$?

Hipótesis:

$$H_0: p \leq 0.65 \quad vs \quad H_1: p > 0.65$$

Estadístico de prueba:

Se tiene que 10 casas poseen la característica de interés, $T = 10$

Bajo $H_0, T \sim \text{Bin}(0.65, 12)$

Regla de decisión:

De acuerdo a nuestra regla de decisión B rechazamos H_0 si $T > t_1$ donde t_1 es elegida para hacer el error tipo I igual a α . Por lo tanto necesitamos encontrar el cuantil t_1 de una distribución $\text{Bin}(0.65, 12)$ tal que $P_{0.65}(r > t_1) = 0.05$

Buscamos en la tabla de la distribución normal acumulada con $n = 12$ y $p = 0.65$ y sustituyendo los valores de x se tiene que:

T	$P_{0.65}(T \leq t)$	$1 - P_{0.65}(T \leq t)$
0	0	1
1	0.0001	0.9999
2	0.0008	0.9992
3	0.0056	0.9944
4	0.0255	0.9745
5	0.0846	0.9154
6	0.2127	0.7873
7	0.4167	0.5833
8	0.6533	0.3467
9	0.8487	0.1513
10	0.9576	0.0424
11	0.9943	0.0057

Como puedes observar no encontramos un cuantil t_1 que nos dé un nivel exacto de $\alpha = 0.05$,



Estadística II

Unidad 1. Estadística no paramétrica

esto es, por la peculiaridad de que la distribución Binomial que solo toma valores en los enteros.

Pero podemos tomar un nivel de significancia $\alpha = 0.0424$ que es lo más cercano a lo buscado con región de rechazo $\{11, 10\}$. Para este caso concluimos:

Como $T = 10 \not\geq t_i = 11$ no existe evidencia estadística suficiente para rechazar H_0 al nivel $\alpha = 0.0424$. Entonces, la empresa X no instala más del %65 de boilers en dicha colonia.

Para ello deberás utilizar la **tabla de la binomial acumulada** ubicada en la pestaña de **Material de apoyo**

Ejemplo

Continuando con el ejemplo anterior, supongamos que la muestra es de 110 casas en las que se encontró que en 85 la empresa X había instalado el boiler.

Ahora n es suficientemente grande como para aproximar T con una distribución normal.

Hipótesis:

$$H_0: p \leq 0.65 \quad vs \quad H_1: p > 0.65$$

Estadístico de prueba:

Se tiene que 85 casas poseen la característica de interés, $T = 85$

Regla de decisión:

La región de rechazo es aquella donde $T^* \geq t_1$. Donde t_1 se elige de tal manera que $P_{0.65}(T > Z_{0.05}) \cong 0.05$. Entonces bajo H_0 tenemos que:

$$P\left(\frac{T - np_0}{[np_0(1 - p_0)]^{1/2}} > Z_{0.05}\right) \cong 0.05$$

Entonces,

$$t_1 = np_0 + Z_{0.05}[np_0(1 - p_0)]^{1/2}$$
$$t_1 = (110)(0.65) + (1.64)[(110)(0.65)(1 - (0.65))]^{1/2} = 79.70$$



Estadística II

Unidad 1. Estadística no paramétrica

Recordemos que $Z_{0.05} = 1.64$

Como $T = 110 > t_i = 79.70$ rechazamos H_0 . Por lo tanto, hay evidencia estadística suficiente para suponer que la empresa X instaló el 65% de los boilers de cierta colonia.

1.2.2. Prueba de la tendencia Cox Stuart

Este test es una alternativa al test paramétrico para $H_0: \beta = 0$ en el modelo de regresión lineal $y = \alpha + \beta x + \varepsilon$. La hipótesis nula en esta prueba implica que la pendiente de la recta es 0.

La prueba de Cox Stuart se basa en variables aleatorias binomiales y permite contrastar la presencia de tendencias.

Contrasta la hipótesis de ausencia de tendencia contra la hipótesis alternativa de tendencia monótona

Recordemos que una tendencia es monótona si la variable dependiente crece cuando crece la variable independiente (monótona creciente) o decrece cuando crece la variable independiente (monótona decreciente)

Datos:

Tenemos una muestra aleatoria $x_1, x_2, \dots, x_n$.
La escala de medida es al menos ordinal

Estadístico de prueba

Formamos los grupos de variables

$(x_1, x_{1+c}), (x_2, x_{2+c}), \dots, (x_{n'-c}, x_{n'})$.

Donde:

$$c = \begin{cases} \frac{n'}{2}, & \text{si } n' \text{ es par} \\ \frac{n' + 1}{2}, & \text{si } n' \text{ es impar} \end{cases}$$

n' es el número de parejas



Estadística II

Unidad 1. Estadística no paramétrica

Asignamos signos a las parejas

$$x_i < x_{i+c} \Rightarrow + \quad \text{y si } x_i > x_{i+c} \Rightarrow -$$

Y se eliminan todas las parejas iguales.

$S = \text{es el número de } +\text{'s}$

Que bajo H_0 $S \sim \text{Bin}(n, 1/2)$. Si se tienen valores muy grandes de T se sugiere una tendencia creciente y si se encuentran valores de T bajos se sugiere una tendencia decreciente.

Hipótesis

A. H_0 : No existe tendencia

- En este caso $P(x_i < x_{i+c}) = P(x_i > x_{i+c}) = 1/2 \quad \forall i$
- También podemos escribir de manera abreviada $p_i = 1/2$

H_1 : Existe una tendencia creciente o decreciente

- En este caso $P(x_i < x_{i+c}) \neq P(x_i > x_{i+c}) \neq 1/2 \quad \forall i \quad / \quad p_i \neq 1/2$

B. H_0 : No existe tendencia creciente

En este caso $P(x_i < x_{i+c}) \leq P(x_i > x_{i+c}) \leq 1/2 \quad \forall i \quad / \quad p_i \leq 1/2$

H_1 : Existe una tendencia creciente o decreciente

En este caso $P(x_i < x_{i+c}) > P(x_i > x_{i+c}) > 1/2 \quad \forall i \quad / \quad p_i > 1/2$

C. H_0 : No existe tendencia decreciente

En este caso $P(x_i < x_{i+c}) \geq P(x_i > x_{i+c}) \geq 1/2 \quad \forall i \quad / \quad p_i \geq 1/2$

H_1 : Existe una tendencia creciente o decreciente

En este caso $P(x_i < x_{i+c}) < P(x_i > x_{i+c}) < 1/2 \quad \forall i \quad / \quad p_i < 1/2$

Regla de decisión:

A. Para valores suficientemente grandes o valores suficientemente pequeños de S la región crítica bajo H_0 es:

$$p(r \leq t_2) \approx \alpha/2 \quad \text{y } p(r > t_1) \approx \alpha/2$$

Por lo tanto rechazamos H_0 si $S \leq t_2$ ó $S > t_1$.



Estadística II

Unidad 1. Estadística no paramétrica

- B. Para valores muy grandes de S significa que H_0 es falsa. La región crítica consiste en todos los valores de S mayores a t_1 , en términos probabilísticos la región de rechazo es aquella que cumple

$$p(r > t_1) = \alpha$$

Por lo tanto, rechazamos H_0 al nivel de significancia α si: $T > t_1$

- C. Para valores muy pequeños de S significa que H_0 es falsa. La región crítica es:

$$p(r \leq t_2) = \alpha$$

Por lo tanto, rechazamos H_0 al nivel de significancia α si: $T \leq t_2$

Ejemplo 1.2.1 El Banco de México registra en su página el Índice de producción industrial en Construcción de manera mensual de 1994 al 2011. Nosotros tomaremos el promedio de cada año para construir un índice anual. Se obtienen los siguientes datos:

1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
12.66	-25.36	10.85	14.66	6.94	5.54	5.54	5.93	-3.43	2.15
2004	2005	2006	2007	2008	2009	2010	2011	2012	
3.46	5.38	3.90	7.84	4.38	3.17	-7.30	-0.01	4.86	

Fuente: Banco de México. (2012). *Índice de volumen de la producción industrial en construcción* (Base 2003=100).

Retrieved from Período: Ene 1994-Sep 2012, Mensual, Sin Unidad. website:

<http://www.banxico.org.mx/SieInternet/consultarDirectorioInternetAction.do?accion=consultarCuadro&iidCuadro=CR100&or=2&locale=es>

Observamos la gráfica de serie de tiempo para darnos una idea si existe tendencia en los datos.





Estadística II

Unidad 1. Estadística no paramétrica

A simple vista no observamos una tendencia en los datos. Realizaremos la prueba de Cox Stuart para comprobar si existe o no dicha tendencia.

Hipótesis:

$$H_0: \text{No existe tendencia} / p_i = 1/2 \quad \text{vs} \quad H_1: \text{Existe una tendencia} / p_i \neq 1/2$$

Estadístico de prueba:

En este caso $n' = 19$ por lo que $c = \frac{n'+1}{2} = \frac{19+1}{2} = \frac{20}{2} = 10$

Para formar los pares eliminamos la observación central. En nuestro ejemplo es la correspondiente al año 2003. Los pares resultantes quedan como:

1	(12.66,3.46)	-
2	(-25.36,5.38)	+
3	(10.85,3.90)	-
4	(14.66,7.84)	-
5	(6.94,4.38)	-
6	(5.54,3.17)	-
7	(5.54,-7.30)	-
8	(5.93,-0.01)	-
9	(-3.43,4.86)	+

Tenemos que

$$S = \text{es el numero de +'s} = 2 \quad \text{y} \quad n = 9$$

Entonces $S \sim \text{Bin}(9, 1/2)$

Regla de decisión:

Tomando un nivel de significancia $\alpha = 0.05$ la región crítica bajo H_0 es:

$$p(r \leq t_2) \approx 0.05/2 = 0.025 \quad \text{y} \quad p(r > t_1) = 1 - p(r \leq t_1) \approx 0.05/2 = 0.025$$

Buscando en la Tabla de la Binomial Acumulada con los parámetros $p = 1/2$ y $n = 9$

Se tienen los siguientes valores

r	$p(r \leq t_2)$	$1 - p(r \leq t_1)$
0	0.002	0.998
1	0.0195	0.9805



Estadística II

Unidad 1. Estadística no paramétrica

2	0.0898	0.9102
3	0.2539	0.7461
4	0.5	0.5
5	0.7461	0.2539
6	0.9102	0.0898
7	0.9805	0.0195
8	0.998	0.002

Por lo tanto rechazamos H_0 si $2 \leq 1$ ó $2 > 7$.

Como ninguno de lo anterior se cumple entonces rechazamos H_0 y por lo tanto no existe tendencia en los datos, lo que se reafirma al observar la gráfica de serie de tiempo del índice.



Estadística II

Unidad 1. Estadística no paramétrica

1.3. Pruebas para dos poblaciones independientes

1.3.1. Prueba U de Mann-Whitney

La prueba de U de Mann-Whitney está diseñada para determinar si dos muestras han sido extraídas de la misma población. Sirve como alternativa a la prueba t cuando el supuesto poblacional con varianzas iguales no se puede verificar. Los datos deben estar medidos al menos en una escala ordinal, haciendo que esta prueba sea útil para datos ordinales o categóricos.

Datos:

Se tienen dos poblaciones

$$x_1, x_2, \dots, x_{n_1} \text{ y } y_1, y_2, \dots, y_{n_2}$$

de tamaño n_1 y n_2 respectivamente. Las muestras se han tomado aleatoriamente y en forma independiente, no solamente entre los grupos considerados, sino además dentro de cada grupo.

Sea:

$F(t)$ es la función de distribución de probabilidad de X

$G(t)$ es la función de distribución de probabilidad de Y

Hipótesis

A. $H_0: F(t) = G(t) \quad \forall t$	vs	$H_1: G(t) \neq F(t) \text{ para alguna } t$
B. $H_0: F(t) = G(t) \quad \forall t$	vs	$H_1: G(t) > F(t) \text{ para alguna } t$
C. $H_0: F(t) = G(t) \quad \forall t$	vs	$H_1: G(t) < F(t) \text{ para alguna } t$

La hipótesis nula prueba que las dos distribuciones son iguales, mientras que las hipótesis alternativas nos dicen si la distribución de Y tiende a ser más grande o más pequeña que X o diferente.

Estadístico de prueba:

Se ordenan las dos muestras combinando los $n = n_1 + n_2$ valores de X y Y de menor a mayor.

s_1 denota el rango de y_1
 s_2 denota el rango de y_2



Estadística II

Unidad 1. Estadística no paramétrica

$\vdots$
 s_n denota el rango de y_n

Calculamos:

$$R_1 = \sum_{j=1}^{n_1} s_j$$

$$R_2 = \sum_{j=1}^{n_2} s_j$$

Donde:

R_1 = Es la suma de los rangos asignados al grupo cuyo tamaño muestral es n_1

R_2 = Es la suma de los rangos asignados al grupo cuyo tamaño muestral es n_2

En el caso de empates se acostumbra asignar el promedio de los rangos correspondientes a las observaciones ligadas.

El estadístico está dado por:

$$U_x = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_y = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

Estos índices satisfacen la propiedad de que

$$U_x + U_y = n_1 n_2$$

El estadístico de prueba será

$$U_m = \min(U_x, U_y)$$

Región de rechazo

A. Debe tomarse una región crítica de dos colas, formada por los valores de U tales que:

$$\begin{aligned} U_x &\leq k_1 \\ U_y &\geq k_2 \end{aligned}$$

siendo la región de aceptación la que verifica la igualdad bajo H_0 :



Estadística II

Unidad 1. Estadística no paramétrica

$$P(k_1 < U_m < k_2) = 1 - \alpha$$

donde α es el nivel de significación.

En la tabla U Mann Whitney se recogen los valores de las probabilidades, puedes visualizarla en la sección **“Material de apoyo”**

$$P(U < U_m) = p$$

estas probabilidades son iguales a

$$P(U \geq n_1 n_2 - U_m) = p$$

Si $2p < \alpha$ se rechaza la hipótesis nula de igualdad de distribuciones poblacionales.

Aproximación a la distribución normal:

B. Si la probabilidad p obtenida en la tabla U Mann Whitney es tal que

$$p < \alpha$$

se rechaza la hipótesis nula H_0 .

C. Si la probabilidad p obtenida en la tabla U Mann Whitney es tal que

$$p > \alpha$$

se rechaza la hipótesis nula H_0 .

Aproximación a la normal

Apoyándose en H_0 , la media y la varianza de U_m se puede calcular a partir de las siguientes expresiones:

$$E(U_m) = \frac{n_1 n_2}{2}$$
$$V(U_m) = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)$$

Los resultados anteriores son de gran utilidad en el caso de muestras grandes, ya que con el Teorema del Límite Central se tiene que la variable z expresa por:

$$z = \frac{U_m - \frac{n_1 n_2}{2}}{\sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)}}$$



Estadística II

Unidad 1. Estadística no paramétrica

Se distribuye como una normal estándar ó $N(0,1)$

En este caso la región de rechazo será:

- A. Rechaza H_0 al nivel de significancia α si $|Z| \geq Z_{\alpha/2}$
- B. Rechaza H_0 al nivel de significancia α si $Z \geq Z_{\alpha}$
- C. Rechaza H_0 al nivel de significancia α si $Z \leq Z_{\alpha}$

Ejemplo

Se aplicaron cuestionarios socioeconómicos a empleados de dos departamentos de una empresa. Obteniéndose los siguientes ingresos mensuales:

Departamento	1	2	3	4	5	6	7	8
D1	17000	4250	5800	5720	18500	1800	5400	1200
D2	3400	3680	5500	13500	3000	7500		

Se desea saber si los empleados pertenecen al mismo nivel socioeconómico. Con un nivel de significancia del 5%.

Hipótesis:

H_0 : Ambos grupos de empleados pertenecen al mismo nivel socioeconómico

$H_0: F(t) = G(t)$

H_1 : Los grupos de empleados pertenecen a distinto nivel socioeconómico

$H_1: G(t) \neq F(t)$

Procedimiento de cálculo

Ordenar la sucesión mezclada e identificada



Estadística II

Unidad 1. Estadística no paramétrica

Rango	1	2	3	4	5	6	7
	1200	1800	3000	3400	3680	4250	5400
	D1	D1	D2	D2	D2	D1	D1

Rango	8	9	10	11	12	13	14
	5500	5720	5800	7500	13500	17000	18500
	D2	D1	D1	D2	D2	D1	D1

Calcular el número de puntaje

Calculamos la suma de los rangos de por ser la de menor tamaño

$$R_{D1} = 1 + 2 + 6 + 7 + 9 + 10 + 13 + 14 = 62$$

$$U_{D1} = (8)(6) + \frac{8(8+1)}{2} - 62 = 22$$

$$U_{D1} = 228$$

Por otro lado

$$R_{D2} = 3 + 4 + 5 + 8 + 11 + 12 = 43$$

$$U_{D2} = (8)(6) + \frac{6(6+1)}{2} - 43 = 26$$

siendo

$$U_m = \min(U_{D1}, U_{D2}) = \min(22, 26) = 22$$

En la tabla del estadístico U Mann Whitney para $n_1 = 8$ y $n_2 = 6$ se obtiene que

$$P(U < U_m) = P(U < U_m) = 0.066$$

con lo cual

$$2p = 2 * 0.066 = 0.132 > \alpha = 0.05$$

no rechazándose la hipótesis nula de que ambas muestras puedan proceder de una misma población, es decir, los empleados de los dos departamentos comparten mismo nivel socioeconómico.



Estadística II

Unidad 1. Estadística no paramétrica

1.3.2. La prueba de la mediana

Este test tiene como finalidad verificar si dos muestras independientes proceden de poblaciones con la misma mediana. Es de utilidad cuando no se pueda verificar el supuesto de normalidad requerido para la prueba *t* Student para dos muestras independientes. Si no puede mantenerse esta hipótesis, las dos muestras corresponderán a poblaciones con tendencia central diferente.

Datos

Se tienen dos muestras aleatorias:

$$x_1, x_2, \dots, x_{n_1} \text{ y } y_1, y_2, \dots, y_{n_2}$$

De tamaño n_1 y n_2 que además cumplen con los siguientes supuestos:

- Las dos muestras se han tomado de forma independiente, solamente entre los grupos considerados, sino además dentro de cada grupo
- Las mediciones consideradas alcanzan al menos el nivel ordinal

Y se ordenan de menor a mayor la muestra conjunta, donde se combinan las observaciones x_i e y_i entre sí, y se determina la mediana muestral de la muestra combinada (Me).

Sea:

$F(t)$ es la función de distribución de probabilidad de X

$G(t)$ es la función de distribución de probabilidad de Y

Hipótesis

$$H_0: \text{Mediana de } F(t) = \text{Mediana de } G(t) \quad \forall t$$

$$H_1: \text{Mediana de } F(t) \neq \text{Mediana de } G(t) \text{ para alguna } t$$

Estadístico de prueba

Las observaciones se comparan con la mediana combinada para obtener las frecuencias de observaciones de ambas muestras que exceden a la mediana. Esas observaciones se arreglan en una tabla de contingencia (2×2):



Estadística II

Unidad 1. Estadística no paramétrica

	Muestra x_i	Muestra y_i	Totales marginales
Número de observaciones mayores a la mediana muestral	A	B	A+B
Número de observaciones inferiores a la mediana muestral	C	D	C+D
Tamaños de las muestras	A+C	B+D	n

La distribución muestral bajo H_0 es hipergeométrica.

$$P(A, B) = \frac{\binom{A+C}{A} \binom{B+D}{B}}{\binom{N}{A+B}}$$

Si el número de casos es pequeño $n < 30$, con frecuencia se utiliza la prueba exacta de Fisher, la cual se basa en el cálculo de la expresión anterior. Para $n > 30$ se puede utilizar la aproximación de una χ^2 con 1 grado de libertad.

$$W = \frac{n(|AD - BC| - n/2)^2}{(A+B)(C+D)(A+C)(B+D)}$$

Regla de decisión:

Rechazamos H_0 al nivel de significancia α si: $W > \chi^2_{\alpha,1}$

Ejemplo

Se aplicó una escala de satisfacción sobre la dotación de servicios públicos a dos grupos de ciudadanos de un municipio. Determine si existen diferencias entre uno y otro grupo considerando los siguientes datos con un nivel de significación de $\alpha = 0.05$.

Municipio	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	3	4	3	3	4	2	4	4	4	3	3	2	3	2	3	4	1	2	4	3	4
2	4	3	2	4	3	1	4	2	2	1	3	3	2	2	2	1	1	3			

Con la siguiente descripción en la escala de media:



Estadística II

Unidad 1. Estadística no paramétrica

Valor	Descripción
1	Muy insatisfecho
2	Insatisfecho
3	Satisfecho
4	Muy satisfecho

Hipótesis:

H_0 : No existen diferencias entre la satisfacción de ambos municipios

$$H_0: F(t) = G(t)$$

H_1 : Existen diferencias entre la satisfacción de ambos municipios

$$H_1: G(t) \neq F(t)$$

Procedimiento de cálculo

La mediana combinada de los dos grupos es 3

	Municipio		Totales
	1	2	Marginales
Mayores de la mediana	8	3	11
Menores o iguales a la mediana	13	15	28
Tamaños de las muestras	21	18	39

Cálculo de la estadística de prueba:

$$W = \frac{39(|8 * 15 - 3 * 13| - 39/2)^2}{(8 + 3)(13 + 15)(8 + 13)(3 + 15)} = 1.26$$

El valor de tablas de una χ^2 con un grado de libertad y una significancia de 5% es 3.84. Como $1.26 \ngtr 3.84$ la hipótesis H_0 no se rechaza. Existe evidencia estadística suficiente para suponer que no existen diferencias entre la satisfacción de ambos municipios.

Hipótesis:



Estadística II

Unidad 1. Estadística no paramétrica

1.3.3. Prueba de rachas Wald-Wolfowitz

El objetivo de este test es el de verificar que dos muestras independientes proceden de poblaciones con distribuciones continuas idénticas.

Definimos una **racha** como una sucesión de símbolos de la misma clase limitada por símbolos de clase distinta. El caso más simple es aquel en donde solo se tienen dos tipos de símbolos A y B. Consideremos la siguiente secuencia:

AA BBBBBB AAAAAA BB

La secuencia mostrada presenta 4 rachas.

Si las dos clases de observaciones A y B, proceden aleatoriamente de una misma población, entonces los símbolos A y B aparecerán bien mezclados en la secuencia y por lo tanto el número de rachas será grande. Mientras, que si por el contrario, las observaciones A y B no aparecen aleatoriamente, el número de rachas tenderá a dos.

Datos

Se tienen dos muestras independientes

$$x_1, x_2, \dots, x_{n_1}$$

$$y_1, y_2, \dots, y_{n_2}$$

Hipótesis

Se plantean los tres contrastes posibles, aunque generalmente solo se utiliza el contraste bilateral, que es con el que trabajaremos.

A. H_0 = El patrón de ocurrencia de las dos muestras es determinado por un proceso aleatorio

H_1 = El patrón de ocurrencia no es aleatorio

B. H_0 = El patrón de ocurrencia de las dos muestras es determinado por un proceso aleatorio

H_1 = El patrón de ocurrencia no es aleatorio (debido a la presencia de pocas rachas)

C. H_0 = El patrón de ocurrencia de las dos muestras es determinado por un proceso aleatorio

H_1 = El patrón de ocurrencia no es aleatorio (debido a la presencia de muchas rachas)

Estadístico de prueba

Cuando n_1 y n_2 sean menos a 20



Estadística II

Unidad 1. Estadística no paramétrica

Se combinan las $n = n_1 + n_2$ observaciones de menor a mayor y se calcula:

R = El número de rachas

Región de rechazo

A. Rechazamos H_0 al nivel de significancia α si:

$$R \leq R_{\text{mínimo}} = R_{n_1, n_2, \alpha/2} \quad \text{ó cuando} \quad R \geq R_{\text{máximo}} = R_{n_1, n_2, 1-\alpha/2}$$

B. Rechazamos H_0 al nivel de significancia α si:

$$R \leq R_{\text{mínimo}} = R_{n_1, n_2, \alpha/2}$$

C. Rechazamos H_0 al nivel de significancia α si:

$$R \geq R_{\text{máximo}} = R_{n_1, n_2, \alpha/2}$$

El valor crítico $R_{\text{mínimo}}$ se busca en la tabla M1 y $R_{\text{máximo}}$ en la tabla M2 de la sección de tablas de rachas cuando se tiene un nivel de significancia del 0.25%, la tabla M1 y M2, la puedes visualizar en la pestaña “**Material de apoyo**”

Aproximación a la normal

Cuando n_1 y n_2 son mayores a 20 se utiliza una aproximación normal. Se sabe que:

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$V(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2}$$

Y utilizando el Teorema del Límite Central se tiene que la variable z expresa por:

$$z = \frac{U_m - \frac{2n_1n_2}{n_1 + n_2} + 1}{\sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2}}}$$

Se distribuye como una normal estándar ó $N(0,1)$

Con región rechazo:



Estadística II

Unidad 1. Estadística no paramétrica

- A. Rechaza H_0 al nivel de significancia α si $|Z| \geq Z_{\alpha/2}$
- B. Rechaza H_0 al nivel de significancia α si $Z \geq Z_{\alpha}$
- C. Rechaza H_0 al nivel de significancia α si $Z \leq Z_{\alpha}$

Ejemplo

El director de una escuela desea saber si los niños son más agresivos que las niñas, por lo que realizó un estudio a 12 niños y 12 niñas de preescolar en grupos separados y en tiempos de 30 min. cada grupo.

Se registraron las incidencias por grados de agresión obteniéndose los siguientes resultados:

Género	1	2	3	4	5	6	7	8	9	10	11	12
Niños	75	34	34	53	91	58	97	42	20	47	8	66
Niñas	33	60	35	59	60	16	5	66	67	14	49	77

Hipótesis

H_0 = El género no influye en el patrón de agresiones de los niños, sino es un proceso aleatorio

H_1 = El patrón de ocurrencia no es aleatorio e influye el género de los niños

Procedimiento de cálculo

Ordenamos las muestras de menor a mayor diferenciando el grupo de procedencia y contamos el número de rachas

Niñas	Niños	Niñas	Niñas	Niños	Niñas	Niños	Niños	Niñas	Niños	Niños	Niñas
5	8	14	16	20	33	34	34	35	42	47	49
1 racha				2 rachas				3 rachas			
Niños	Niños	Niñas	Niñas	Niñas	Niños	Niñas	Niñas	Niños	Niñas	Niños	Niños
53	58	59	60	60	66	66	67	75	77	91	97
4 rachas				5 rachas				6 rachas			

Por lo que

$$R = \text{El número de rachas} = 6$$

Se buscan los valores críticos en las tablas M1 y M2 y se tiene que para la desigualdad se cumple para:



Estadística II

Unidad 1. Estadística no paramétrica

$$R = 6 \leq R_{\text{mínimo}} = 7$$

Por lo tanto rechazamos H_0 al nivel de significancia $\alpha = 0.025$. Existe evidencia estadística para suponer que las agresiones de los niños se deben a un factor de género y no son totalmente aleatorias.

1.3.4. Prueba de Mac Nemar

La prueba es famosa porque es muy utilizada en pruebas donde existe un antes y un después, por ejemplo, cuando se quiere decidir si puede o no aceptarse que determinado “tratamiento” induce un cambio en la respuesta dicotómica de los elementos sometidos al mismo, y es aplicable a los diseños del tipo “antes-después” en los que cada elemento actúa como su propio control.

Datos

Los datos consisten de observaciones bivariadas aleatorias $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. La escala de medida de x_i y de y_i es nominal con 2 categorías las cuales llamaremos “0” y “1”, esto es, los valores de (x_i, y_i) son $(0,0), (0,1), (1,0), (1,1)$.

Las muestras cumplen los siguientes supuestos:

- Los pares (x_i, y_i) son mutuamente independientes
- La escala de medida es nominal con 2 categorías para x_i y y_i

Hipótesis

A H_0 : El “tratamiento” no induce cambios significativos en la respuesta, es decir, los campos observados en la muestra se deben al azar; de forma que es igualmente probable un cambio de x_i a y_i que un cambio de y_i a x_i . Matemáticamente se puede escribir como:

$$P(x_i = 0, y_i = 1) = P(x_i = 1, y_i = 0) \quad \forall i \quad / \quad p_1 = p_2$$

H_1 : El “tratamiento” induce cambios

$$P(x_i = 0, y_i = 1) \neq P(x_i = 1, y_i = 0) \quad \forall i \quad / \quad p_1 \neq p_2$$

B H_0 : La característica de interés bajo la condición 1 es mayor que bajo la condición 2

$$P(x_i = 0, y_i = 1) \leq P(x_i = 1, y_i = 0) \quad \forall i \quad / \quad p_1 \leq p_2$$

H_1 : La característica de interés bajo la condición 1 no es mayor que bajo la condición 2



Estadística II

Unidad 1. Estadística no paramétrica

$$P(x_i = 0, y_i = 1) > P(x_i = 1, y_i = 0) \quad \forall i / \quad p_1 > p_2$$

$C \quad H_0$: La característica de interés bajo la condición 1 es menor que bajo la condición 2

$$P(x_i = 0, y_i = 1) \geq P(x_i = 1, y_i = 0) \quad \forall i / \quad p_1 \geq p_2$$

H_1 : La característica de interés bajo la condición 1 no es menor que bajo la condición 2

$$P(x_i = 0, y_i = 1) < P(x_i = 1, y_i = 0) \quad \forall i / \quad p_1 < p_2$$

Estadístico de prueba

Construimos la tabla de contingencia 2×2

	$y_i = 0$	$y_i = 1$	Total
$x_i = 0$	A	B	A+B
$x_i = 1$	C	D	C+D
Total	A+C	B+D	N

En A y en D se mantiene la misma respuesta, pero $B + A$ es el número total de respuestas que ha cambiado.

Tenemos que el número total de respuestas que ha cambiado es $B + C$. De acuerdo a H_0 se espera que $\frac{B+C}{2}$ sean las respuestas que hayan cambiado de lugar. Esto porque H_0 nos dice que no hay cambio, por lo tanto, los cambios que se han realizado se deben al azar, en otras palabras, H_0 es la frecuencia esperada en las correspondientes celdas. El estadístico de prueba que permite contrastar si existen diferencias significativas entre las frecuencias esperadas y las observadas es:

$$T = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Donde:

K = Número de celdas

O_i = Frecuencia observada en la i -ésima celda

E_i = Frecuencia esperada en la i -ésima celda

Como solo nos interesan las celdas que recogen cambios el estadístico puede expresarse como:



Estadística II

Unidad 1. Estadística no paramétrica

$$T = \frac{\left(B - \frac{B+C}{2}\right)^2}{\frac{B+C}{2}} + \frac{\left(C - \frac{B+C}{2}\right)^2}{\frac{B+C}{2}} = \frac{(B-C)^2}{B+C}$$

Bajo H_0 el estadístico T tiene una distribución χ^2 con un grado de libertad.

Para trabajar bajo muestras pequeñas se puede aplicar la corrección de Yates, en ese caso se tiene que:

$$T = \frac{(|B - C| - 1)^2}{B + C}$$

Regla de decisión

- Rechaza H_0 al nivel de significancia α si $T \geq \chi^2_{1,\alpha}$. Donde $\chi^2_{1,\alpha}$ es cuantil de una distribución χ^2 con un grado de libertad y probabilidad α
- Rechaza H_0 al nivel de significancia α si $T \geq Z_\alpha$. Donde es el cuantil de una distribución normal con probabilidad α
- Rechaza H_0 al nivel de significancia α si $T \leq -Z_\alpha$

Ejemplo

El encargado de campaña de un candidato a la presidencia desea saber el cambio de opinión que causa un debate entre todos los candidatos. Por lo que toma una muestra de 78 votantes elegidos de manera aleatoria y registro la preferencia hacia su candidato, inmediatamente después del debate, volvió a registrar la preferencia del candidato. Los resultados se muestran a continuación:

Antes del Debate	Después del debate		
	Desacuerdo (0)	Acuerdo (1)	Total
Desacuerdo (0)	24	18	42
Acuerdo (1)	6	30	36
Total	30	48	78

Hipótesis

H_0 : El debate produjo un cambio en la opinión de los votantes / $p_1 = p_2$

H_1 : El debate no produjo un cambio en la opinión de los votantes / $p_1 \neq p_2$

Estadístico de prueba:



Estadística II

Unidad 1. Estadística no paramétrica

$$T = \frac{(|B - C| - 1)^2}{B + C} = \frac{(|18 - 6| - 1)^2}{18 + 6} = 5.04$$

Regla de decisión

Rechazamos H_0 a nivel $\alpha = 0.05$ si $T = 5.04 \geq \chi^2_{1,0.05} = 3.841$. Dado que se cumple la condición, entonces, rechazamos H_0 y por lo tanto existe evidencia estadística suficiente para suponer que el debate no produjo un cambio en la opinión de los votantes.

Utiliza la tabla de la “**ji cuadrada**”, ubicada en la pestaña de material de apoyo

1.4. Pruebas para dos poblaciones independientes

1.4.1. Prueba de signos

La prueba de signos es la más vieja de las pruebas no paramétricas. *John Arbuthnot* presentó un documento a la *Royal Society* en 1710 discutiendo el ligero exceso de nacimientos de varones que de nacimientos femeninos en los años 1629 y 1710. Este trabajo, publicado en la *Philosophical Transsantion*, es tal vez la primera aplicación a la estadística social.

La prueba de signos es actualmente igual a la binomial con $p_0 = 1/2 = 1 - p_0$. Es una prueba con mucha versatilidad porque ayuda a probar si cualesquiera dos poblaciones tienen la misma mediana y también permite indicar la existencia de tendencias.

Datos

Los datos consisten de observaciones bivariadas aleatorias $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

Las muestras cumplen los siguientes supuestos:

- Variables aleatorias bivariadas mutuamente independientes
- La escala de medida es al menos ordinal dentro de cada par

Hipótesis

A. La mediana de $x_i =$ La mediana de $y_i \quad \forall i$



Estadística II

Unidad 1. Estadística no paramétrica

$$H_0: P(x_i < y_i) = P(x_i > y_i) \quad \forall i$$

La mediana de $x_i \neq$ La mediana de $y_i \quad \forall i$

$$H_1: P(x_i < y_i) < P(x_i > y_i) \quad \forall i \quad \text{ó} \quad H_1: P(x_i < y_i) > P(x_i > y_i) \quad \forall i$$

B. La mediana de $x_i \geq$ La mediana de $y_i \quad \forall i$

$$H_0: P(x_i < y_i) \leq P(x_i > y_i) \quad \forall i$$

La mediana de $x_i <$ La mediana de $y_i \quad \forall i$

$$H_1: P(x_i < y_i) > P(x_i > y_i) \quad \forall i$$

C. La mediana de $x_i \leq$ La mediana de $y_i \quad \forall i$

$$H_0: P(x_i < y_i) \geq P(x_i > y_i) \quad \forall i$$

La mediana de $x_i >$ La mediana de $y_i \quad \forall i$

$$H_1: P(x_i < y_i) < P(x_i > y_i) \quad \forall i$$

Estadístico de prueba

Dentro de cada par se puede hacer la siguiente comparación:

- Un par es clasificado por "+" si $x_1 < y_1$
- Un par es clasificado por "-" si $x_1 > y_1$
- Un par es clasificado por "0" si $x_1 = y_1$

$$T = \text{Total de +'s}$$

Se ignoran los "0", es decir, las igualdades en donde $x_i = y_i$

$$n = \text{total de de +'s y -'s}$$

Regla de decisión

Para $n \leq 25$ se cumple que

$$t = \frac{1}{2}(n + 1)$$

Rechazamos H_0 al nivel de significancia α si:

$$T \geq n - t$$



Estadística II

Unidad 1. Estadística no paramétrica

t Es el cuantil de una distribución $Bin(n, p_0)$ al tamaño α .

B. Valores grandes de T indican que los "+" son mas probables que los "-". Por lo tanto la región crítica corresponde a los valores de T más grandes o iguales $n - t$

C. Valores muy pequeños de T indican que "-" es más probable que "+". La región crítica de tamaño α corresponde a los valores de $T \leq t$.

Por lo que rechazamos H_0 si $T \leq t$ al nivel de significancia α .

Cuando $n \geq 25$ se puede utilizar la distribución normal y como esta es simétrica es igual a probar la media. Por consiguiente, la prueba de signo puede emplearse para probar hipótesis sobre la media de la población.

1.4.2. Prueba de Wilcoxon

Esta prueba se utiliza para comparar las distribuciones de probabilidad que no son normales. Es un equivalente a la prueba $t - Student$ y se aplica cuando el tipo de medición no cumpla con los requisitos que la $t - Student$ exige. La prueba *Wilcoxon* no solo toma en cuenta el signo, además considera las magnitudes de diferencias entre los valores asociados, es una prueba más sensible que la de signos.

Determinar el signo de la diferencia nos ayuda a saber cual miembro del par es "mas grande que" y establecer rangos en las diferencias en orden de tamaño absoluto ayuda a establecer juicios de "mayor que" entre los valores de cualquier par.

Supuestos:

- Variables aleatorias bivariadas $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ mutuamente independientes y con distribución simétrica y continua
- Las diferencias son mutuamente independientes
- Se utiliza una escala de medida de intervalos. Esto nos ayuda a saber cuál de los dos miembros del par es más grande y podemos ordenar las diferencias sin tener en cuenta su signo (valor absoluto).
- Las diferencias representan observaciones en una variable continua
- La distribución de la población de diferencias es simétrica alrededor de la mediana θ

Hipótesis

- A. $H_0: \theta = 0$ vs $H_0: \theta \neq 0$
- B. $H_0: \theta \leq 0$ vs $H_0: \theta > 0$
- C. $H_0: \theta \geq 0$ vs $H_0: \theta < 0$



Estadística II

Unidad 1. Estadística no paramétrica

Estadístico de prueba

Denotamos D el estadístico de prueba definido como:

$$D = \sum_{i=1} R_i$$

Donde:

R_i = Suma de los rangos asignados a las parejas (x_i, y_i) con el signo menos frecuente

Los valores de D con diferentes tamaños de muestra y niveles de significancia para pruebas de una o dos colas fueron tabulados por Wilcoxon. Checar la tabla M1 y M2 ubicada en la sección “material de apoyo”

Regla de decisión

- A. Buscamos el cuantil t_α en las tabla de Wilcoxon y rechazamos H_0 al nivel de significancia α si:

$$D \geq t_\alpha \text{ ó } D \leq -t_\alpha$$

- B. Buscamos el cuantil t_α en las tabla de Wilcoxon y rechazamos H_0 si:

$$D \geq t_\alpha$$

- C. Buscamos el cuantil t_α y rechazamos H_0 si:

$$D \leq -t_\alpha$$

Aproximación a la Normal

Cuando $n > 25$ se puede utilizar la aproximación normal.

Se tiene que:

$$E(D) = \frac{n(n+1)}{4}$$

$$Var(T^+) = \frac{n(n+1)(2n+1)}{24}$$

Bajo H_0 y utilizando el Teorema Central del Límite:



Estadística II

Unidad 1. Estadística no paramétrica

$$Z = \frac{D - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$$

Regla de decisión

- A. Rechazamos H_0 si $|D| \geq Z_{\frac{\alpha}{2}}$
- B. Rechazamos H_0 si $D \geq Z_{\alpha}$
- C. Rechazamos H_0 si $D \leq Z_{\alpha}$

Ejemplo 1

Con el fin de comprobar si la asistencia al jardín de niños tiene algún efecto en la capacidad de percepción social el psicólogo de una escuela realiza un experimento en el que forma parejas de actitudes similares como sexo, edad, calificación de la medición y durante la hora del recreo realiza una medición en total forma 10 parejas y solo somete al experimento a un integrante de cada pareja. Los resultados se muestran a continuación.

Hipótesis

H_0 = La percepción social de los niños que se sometieron al experimento es igual que la de los niños que no se sometieron

H_1 = La percepción social de los niños que se sometieron al experimento es diferente que la de los niños que no se sometieron

Puntaje niños asignados al experimento	Puntaje niños no asignados al experimento	Diferencias	Absoluto de las diferencias	Rango de las diferencias	Rango de signos menos frecuentes
56	36	20	20	8	
54	49	5	5	3	
87	72	15	15	6	
98	67	31	31	10	
12	41	-29	29	-9	9
34	50	-16	16	-7	7
54	53	1	1	1	
43	47	-4	4	-2	2
67	77	-10	10	-4	4
67	54	13	13	5	



Estadística II

Unidad 1. Estadística no paramétrica

Observa que el rango de las se toman en valor absoluto.

El estadístico de prueba es $D = 22$

Consultamos la tabla de Wilcoxon con $n = 10$ y $D = 22$ y con un $\alpha = 0.05$ para una cola y tenemos que $t_{0.05} = 75$

$$D = 22 \not\geq t_{0.05} = 75$$

No rechazamos H_0

1.5. Prueba de independencia y homogeneidad

Es común que en ocasiones los elementos de una muestra deban ser categorizados de acuerdo a dos o más criterios de clasificación. El uso de una tabla de contingencia será de ayuda en estos casos.

Resulta conveniente aclarar que las hipótesis a probar mediante tablas de contingencia, aun cuando los procedimientos de cálculo son los mismos, tienen básicamente dos sentidos diferentes.

- a) Como hipótesis de igualdad de proporciones en los diferentes niveles de cierta clasificación, cuando las observaciones provienen de 2 o más poblaciones
- b) Como hipótesis de independencia entre 2 criterios de clasificación aplicable a los elementos de una misma población

Como se mencionó, ambos casos son tratados idénticamente desde el punto de vista de los cálculos estadísticos, pero las diferencias básicas entre las dos aplicaciones justifican discusiones separadas.

1.5.1. Tablas de contingencia

Suponga que se tienen r poblaciones y que se extraen muestras aleatorias de cada una de ellas. El tamaño de cada muestra lo denotamos por n_i ($i = 1, 2, \dots, r$). Cada observación de las r muestras puede ser clasificada en una de c diferentes categorías. Se denotará por O_{ij} el número de observaciones de la i -ésima categoría en la j -ésima muestra. Denotamos además por c_i que es el total de observaciones pertenecientes a todas las muestras que quedan contenidas en la i -ésima categoría.



Estadística II

Unidad 1. Estadística no paramétrica

La información se dispone en forma tabular de la siguiente manera en la siguiente tabla de contingencia

$$r \times c$$

	Clase 1	Clase 2	...	Clase c	Total
Población 1	O_{11}		...	O_{1c}	n_1
Población 2	O_{21}		...	O_{2c}	n_2
...	...	...	...	...	...
Población r	O_{r1}	O_{r2}	...	O_{rc}	n_r
Total	c_1	c_2	...	c_c	n

En la tabla se puede verificar lo siguiente:

$$\sum_{j=1}^c O_{ij} = n_j$$
$$\sum_{i=1}^r c_i = \sum_{j=1}^c n_j = n$$

Se consideran los siguientes supuestos básicos en el planteamiento de hipótesis

- Las r muestras son aleatorias
- Los resultados de las diferentes muestras son mutuamente independientes
- Cada observación puede ser categorizada en una y solo una de las c diferentes categorías

Hipótesis

Sea p_{ij} la probabilidad de que un elemento de la j -ésima población seleccionado al azar, quede clasificado en la i -ésima categoría



Estadística II

Unidad 1. Estadística no paramétrica

H_0 : La probabilidad de pertenecer a cualquiera de las c clases es la misma para cualquier elemento de la j -ésima muestra

$$p_{1j} = p_{2j} = \dots = p_{rj} \quad \forall j$$

H_1 : La probabilidad de pertenecer a cualquiera de las c clases es diferente para al menos una clase

$p_{ij} \neq p_{kj}$ para al menos una pareja $j \neq k$

Estadístico de prueba

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Donde:

$$E_{ij} = \frac{n_i c_j}{n}$$

El término O_{ij} representa los valores observados en la celda (i, j) , y el término E_{ij} representa el número esperado de observaciones en la celda (i, j) , cuando H_0 es cierta.

Regla de decisión

Rechazamos H_0 al nivel de significancia α si T excede el cuantil de una χ^2 con probabilidad $1 - \alpha$ y $(r - 1)(c - 1)$ grados de libertad, matemáticamente lo podemos expresar como:

$$T > \chi^2_{\alpha, (r-1)(c-1)}$$

Ejemplo

En una encuesta telefónica se preguntó a los participantes hasta que grado estaban de acuerdo con la proposición: “se debe prohibir fumar en lugares públicos”. Los resultados son los siguientes:

Sexo	Grado en el que se está de acuerdo				En total desacuerdo	Total
	Muy de acuerdo	De acuerdo	Neutral	En desacuerdo		
Mujer	41	16	28	27	31	143
Varón	22	40	14	39	41	156
Total	63	56	42	66	72	299



Estadística II

Unidad 1. Estadística no paramétrica

Con base en los datos recabados se desea saber si existen diferencias significativas en el grado en el que están de acuerdo hombres y mujeres con respecto a prohibir fumar en lugares públicos.

Procedimiento de cálculo

Se calculan los valores E_{ij}

Columna	1	2	3	4	5
Fila 1	30.1	26.8	20.1	31.6	34.4
Fila 2	32.9	29.2	21.9	34.4	37.6

Para la fila 1 en la columna 1 tenemos que:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(41 - 30.1)^2}{30.1} = 3.2$$

Un cálculo similar es echo para cada celda y sumando todo se tiene que el estadístico de prueba es:

$$T = 3.9 + 4.3 + 3.1 + 0.7 + 0.3 + 3.6 + 4.0 + 2.9 + 0.6 + 0.3 = 23.7$$

Si utilizamos $\alpha = 0.05$ comparamos con una $\chi^2_{0.05, (2-1)(5-1)} = \chi^2_{0.05, 4} = 9.488$

Como $T = 23.7 \neq \chi^2_{0.95, 4} = 9.488$ no rechazamos H_0 y no existen diferencias significativas para suponer que el grado de opinión con respecto a si fumar en lugares públicos esté relacionado con el género.

1.5.2. Prueba de independencia con Ji-Cuadrada

Suponga que se dispone de una muestra aleatoria de tamaño n y que las observaciones de la muestra pueden clasificarse de acuerdo a dos criterios. Al usar el primer criterio cada observación puede asociarse con uno de los r filas y al usar el segundo criterio la observación puede asociarse con una de las c columna.

La disposición de las observaciones es igual que en 1.5.1 con la excepción de que en este caso, las n 's no se establecen previamente, sino que son aleatorias:



Estadística II

Unidad 1. Estadística no paramétrica

Los supuestos para este caso son los siguientes:

- Cada observación tiene la misma probabilidad de ser clasificada en el i -ésimo renglón y en la j -ésima columna, independientemente de cualquier otra observación
- Las observaciones pueden ser clasificadas en una de las c diferentes categorías de acuerdo al segundo criterio

Hipótesis

H_0 : El evento “la observación pertenece al i -ésimo renglón” es independiente del evento “la misma observación pertenece a la j -ésima columna” para toda i y j

La proposición anterior puede traducirse en términos probabilísticos de la siguiente forma
Sea p_i la probabilidad de pertenecer al i -ésimo renglón y p_j la probabilidad de pertenecer a la j -ésima columna

$$H_0: p_{ij} = p_i * p_j$$

$$H_1: p_{ij} \neq p_i * p_j$$

Estadística de prueba

La estadística coincide con 1.5.1

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Donde:

$$E_{ij} = \frac{n_i c_j}{n}$$

Regla de decisión

Rechazamos H_0 al nivel de significancia α si T excede el cuantil de una χ^2 con probabilidad $1 - \alpha$ y $(r - 1)(c - 1)$ grados de libertad, matemáticamente lo podemos expresar como:

$$T > \chi^2_{\alpha, (r-1)(c-1)}$$



Estadística II

Unidad 1. Estadística no paramétrica

Ejemplo 2

El propósito de un estudio era investigar la hipótesis de que las mujeres con leucemia que también están infectadas con VIH, tienen más probabilidades de tener anomalías citológicas cervicales que las mujeres con uno de los dos virus mencionados. Se pretende saber si es posible concluir que existe relación entre el estado de leucemia y la etapa de infección por VIH.

Leucemia	VIH			Total
	Seropositivo, sintomático	Seropositivo, asintomático	Seronegativo	
Positivo	20	31	39	90
Negativo	32	51	32	115
Total	52	82	71	205

Hipótesis

H_0 : El estado de leucemia y la etapa de infección por VIH son independientes

H_1 : Las dos variables no son independientes

Procedimiento de Cálculo

Se calculan los valores E_{ij}

Columna	1	2	3
Fila 1	22.8	36.0	31.2
Fila 2	29.2	46.0	39.8

Para la fila 1 en la columna 1 tenemos que:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(20 - 22.8)^2}{22.8} = 0.35$$

Un cálculo similar es echo para cada celda y sumando todo se tiene que el estadístico de prueba es:

$$T = 0.35 + 0.69 + 1.97 + 0.27 + 0.54 + 1.54 = 5.37$$

Si utilizamos $\alpha = 0.05$ comparamos con una $\chi^2_{0.05, (2-1)(3-1)} = \chi^2_{0.05, 2} = 5.991$



Estadística II

Unidad 1. Estadística no paramétrica

Como $T = 5.37 \nless \chi^2_{0.95,4} = 5.991$ no rechazamos H_0 y existen diferencias significativas para suponer que el estado de leucemia y la etapa de infección por VIH son independientes.

1.6. Prueba de tres o más poblaciones independientes

1.6.1. Extensión de la prueba de la mediana

Es la extensión de la prueba de la mediana para más de 2 poblaciones y tiene como propósito verificar si de k muestras independientes con igual o diferente tamaño de muestra proceden de la misma población o de poblaciones con medianas iguales.

Se tienen las muestras

$$\{x_1, x_2, \dots, x_{n_1}\}, \{y_1, y_2, \dots, y_{n_2}\}, \dots, \{z_1, z_2, \dots, z_{n_k}\}$$

de tal manera que

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

Supuestos:

- Las dos muestras se han tomado de forma independiente, solamente entre los grupos considerados, sino además dentro de cada grupo
- Las mediciones consideradas alcanzan al menos el nivel ordinal

Sea

Hipótesis

H_0 : Las k muestras tienen la misma mediana

H_1 : Al menos dos muestras son diferentes

Estadístico de prueba



Estadística II

Unidad 1. Estadística no paramétrica

Llamemos δ a la mediana común de los n elementos. Ahora definimos U_i al número de observaciones en la muestra i los cuales son menores que δ y sea t el número total de observaciones menores que δ .

De existir observaciones que son exactamente igual que el valor de la mediana y estos son muchos, se puede colocar uno por encima y otro por debajo del valor de la mediana, hasta agotarlos. Si son pocos los casos en esta situación, es decir, si el tamaño de n no se reduce grandemente, se pueden eliminar del análisis, modificando tanto el tamaño total como los tamaños marginales.

Se ordenan los cálculos en la siguiente tabla

	Muestra 1	Muestra 2	Muestra K	Total
$< \delta$	U_1	U_2	U_k	t
$\geq \delta$	$n_1 - U_1$	$n_2 - U_2$	$N_k - U_k$	$n - t$
Total	n_1	n_2	N_k	n

El estadístico de prueba es:

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

Regla de decisión

Rechazo H_0 al nivel de significancia α si

$$Q > \chi^2_{\alpha, k-1}$$

Ejemplo1

La siguiente tabla indica las calificaciones obtenidas por 10 estudiantes de la carrera de biología seleccionados al azar en los exámenes finales de tres materias. Las calificaciones se observan en la siguiente tabla



Estadística II

Unidad 1. Estadística no paramétrica

Estudiante	Materia		
	Química	Plantas	Animales
1	81	55	100
2	98	82	56
3	53	87	99
4	62	88	94
5	99	71	79
6	71	75	62
7	82	61	65
8	50	95	83
9	61	74	96
10	74	80	92

Pruebe

H_0 : Los estudiantes tienen el mismo aprovechamiento en las tres materias

H_1 : El aprovechamiento es mejor en alguna de las materias

Procedimiento de cálculo

La mediana común de las observaciones es 79.5

Grupo	1	2	3
<79.5	4	5	6
≥ 79.5	6	5	4

Tenemos $n = 30$, $t = 15$ y $n_{it}/n = 5$

$$Q = \frac{30^2}{15(30 - 15)} (0.1 + 2.5 + 3.6) = 24.8$$

Utilizamos $\alpha = 0.05$

$$\chi_{0.05,3-1}^2 = \chi_{0.05,2}^2 = 5.991$$

Se cumple que $Q > \chi_{0.05,2}^2$ por lo tanto rechazamos H_0 y no podemos suponer que el aprovechamiento de los estudiantes es el mismo en las tres materias.



Estadística II

Unidad 1. Estadística no paramétrica

1.6.2. Comparación de varias poblaciones Kruskal-Wallis

La prueba *Kruskal-Wallis* es útil para probar los resultados de k muestras que vienen de poblaciones diferentes.

Los datos consisten diferentes k muestras aleatorias que pueden tener distintos tamaños.

Muestra 1	Muestra 2	Muestra K
$X_{1,1}$	$X_{2,1}$	$X_{k,1}$
$X_{1,2}$	$X_{2,2}$	$X_{k,2}$
$X_{1,n1}$	$X_{2,n2}$	$X_{k,nk}$

De tal manera que

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

Supuestos:

- Las dos muestras se han tomado de forma independiente, solamente entre los grupos considerados, sino además dentro de cada grupo
- La escala de medida es al menos ordinal (un número moderado de casos repetidos se considera tolerable)

Hipótesis

H_0 : Las k muestras vienen de la misma población o de poblaciones cuyo promedio de rangos son idénticos

H_1 : Al menos dos muestras son diferentes

Estadístico de prueba

Tenemos

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

Ordenamos las n observaciones y les asignamos el rango correspondiente de menor a mayor, después se calcula

$$R_i = \text{La suma de los rangos asignados a la muestra } i \quad \forall i$$



Estadística II

Unidad 1. Estadística no paramétrica

La estadística de prueba se calcula así

$$H = \frac{12}{n(n+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(n+1)$$

Regla de decisión

Rechazo H_0 al nivel de significancia α si

$$H > \chi_{\alpha, k-1}^2$$

Ejemplo

En tres muestras de animales experimentales se estudió el tiempo de reacción de un medicamento. La tercera muestra sirvió como control al medicamento, a la primera muestra se les aplicó el medicamento A y a la segunda el medicamento B. Los tiempos de reacción se muestran en la siguiente tabla:

Muestra		
I	II	II
33	17	28
26	23	34
8	11	5
23	30	10
25	18	33
2	38	15
19		26
30		
32		

¿Es posible concluir que las tres poblaciones representadas por las tres muestras difieren con respecto al tiempo de reacción?

Hipótesis

H_0 : Las distribuciones de las poblaciones son idénticas

H_1 : Al menos una de ellas tiende a mostrar valores mayores que al menos una de las demás

Procedimiento del cálculo

Se combinan las tres muestras en una sola serie y los valores se clasifican por rangos. Recordemos que cuando los rangos se repiten se toma el promedio de ellos.



Estadística II

Unidad 1. Estadística no paramétrica

Muestra			
I	II	II	
19.5	7	15	
13.5	10.5	21	
3	5	2	
10.5	16.5	4	
12	8	19.5	
1	22	6	
9		13.5	
16.5			
18			
Suma Rangos	103	69	81

Se construye la estadística de prueba con $n_1 = 9$, $n_2 = 6$, $n_3 = 7$, $n = 22$

$$H = \frac{12}{22(22+1)} \left[\frac{103^2}{9} + \frac{69^2}{6} + \frac{81^2}{7} \right] - 3(22+1) = 0.0015057$$

Utilizamos $\alpha = 0.05$ y buscamos en tablas el cuantil $\chi^2_{0.05,2} = 5.931$

Como $H \not\geq \chi^2_{0.05,2}$ no rechazamos H_0 y por lo tanto hay evidencia estadística suficiente para suponer que las muestras provienen de la misma población. Por lo que ninguno de los dos tratamientos tiene un efecto en los tiempos de reacción.

1.7. Prueba de Bondad de Ajuste

Una prueba de bondad y ajuste es conveniente cuando se quiere decidir si existe incompatibilidad entre la distribución de frecuencias observadas y alguna distribución predeterminada o hipotética. En estadística es común realizar análisis basados en el hecho de cierta distribución de datos por lo que resulta importante corroborar la procedencia de estos para evitar la violación de algún supuesto.



Estadística II

Unidad 1. Estadística no paramétrica

1.7.1. Prueba de bondad y ajuste basada en Ji-Cuadrada

Los datos consisten de N observaciones independientes de una v.a. X que se agrupan en C clases o grupos. La escala de medida de las categorías es al menos de tipo nominal. Podemos presentar las categorías ordenadas en la siguiente tabla:

Clase	1	2	...	C	Total
Frecuencia	O_1	O_2	...	O_c	n

Donde

$$\sum_{i=1}^c O_i = n$$

Hipótesis

Sea $F(x)$ la f.d.p de x , y sea $F^*(x)$ alguna función específica

$$H_0 : F(x) = F^*(x) \quad \forall x \quad \text{vs} \quad H_1 : F(x) \neq F^*(x) \quad \text{al menos un valor de } x$$

Estadístico de prueba

Sea p_j^* la probabilidad de una observación aleatoria en x en la clase j , bajo el supuesto de que $F^*(x)$

es la función de distribución de x . Entonces definimos el número esperado de observaciones en la clase j cuando H_0 es cierta, E_j , como:

$$E_i = p_i^* n, \quad j = 1, 2, 3, \dots, c$$

El estadístico de prueba está dado por:

$$T = \sum_{i=1}^c \frac{(O_i - E_i)^2}{E_i}$$

Regla de decisión



Estadística II

Unidad 1. Estadística no paramétrica

Valores muy altos de T reflejan una incompatibilidad entre los observados y las frecuencias relativas esperadas. La distribución de T es difícil de calcular. Para muestras largas se tiene que:

$$T \sim \chi^2_{(\alpha, c-1)}$$

Rechazamos H_0 si $T > \chi^2_{(\alpha, c-1)}$

Ejemplo

Se lanza un dado 600 veces y se obtienen los siguientes resultados

Caras del dado	n_i
1	180
2	72
3	150
4	62
5	40
6	96
n	600

Se desea verificar al 5% de nivel de significancia la hipótesis de que el dado está bien construido.

Hipótesis

La hipótesis de que el dado está bien construido equivale a que la muestra de 600 lanzamientos procede de una población uniforme discreta con probabilidad igual a $1/6$ para cada cara del dado.

Entonces, bajo H_0 la probabilidad de ocurrencia es de $p_i = 1/6$.

H_0 = El dado sigue una distribución uniforme $1/6$

H_1 = El dado no sigue una distribución uniforme $1/6$

Procedimiento de cálculo

En primer lugar para realizar el contraste se determinan las frecuencias observadas:

$$E_j = p_j^* n = \frac{1}{6} * 600 = 100, \quad i = 1, 2, 3, 4, 5, 6$$



Estadística II

Unidad 1. Estadística no paramétrica

El valor muestral del estadístico es

$$T = \frac{(180 - 100)^2}{100} + \frac{(72 - 100)^2}{100} + \frac{(150 - 100)^2}{100} + \frac{(62 - 100)^2}{100} + \frac{(40 - 100)^2}{100} + \frac{(96 - 100)^2}{100} = 147.44$$

Buscamos el cuantil en tablas de una distribución $\chi^2_{(0.05, 6-1)} = \chi^2_{(0.05, 5)} = 11.070$

Como $T > \chi^2_{(0.05, 5)}$ rechazamos H_0 por lo que el dado no se ajusta a una distribución uniforme 1/6 y existe evidencia estadística suficiente para suponer que el dado está cargado.

1.7.2. Prueba de Kolmogorov-Smirnov para una muestra

Datos

Los datos consisten de una muestra aleatoria $x_1, x_2, \dots, x_n$ de tamaño n asociada a una distribución desconocida que denotamos por $F(x)$.

Supuestos

- La muestra es aleatoria
- La distribución hipotética $F(x)$ es continua

Sea $F^*(x)$ una función de distribución completamente especificada que toma valores

Hipótesis

- A. $H_0 : F(x) = F^*(x) \quad \forall x, \text{ de } -\infty \text{ a } \infty$
 $H_1 : F(x) \neq F^*(x) \quad \text{al menos un valor de } x$
- B. $H_0 : F(x) \geq F^*(x) \quad \forall x, \text{ de } -\infty \text{ a } \infty$
 $H_1 : F(x) < F^*(x) \quad \text{al menos un valor de } x$
- C. $H_0 : F(x) \leq F^*(x) \quad \forall x, \text{ de } -\infty \text{ a } \infty$
 $H_1 : F(x) > F^*(x) \quad \text{al menos un valor de } x$

Estadístico de prueba

La función de distribución empírica de una muestra se calcula como:



Estadística II

Unidad 1. Estadística no paramétrica

$$S(x) = \frac{n_i}{i}$$

A. Sea el estadístico D la mayor distancia vertical entre $S(x)$ y $F^*(x)$

$$D = \sup_x |F^*(x) - S(x)|$$

B. Sea el estadístico D^+ igual a la mayor distancia vertical de $F^*(x)$ por encima de $S(x)$

$$D^+ = \sup_x |F^*(x) - S(x)|$$

C. Sea el estadístico D^- definida como la mayor distancia vertical de $S(x)$ por encima de $F^*(x)$

$$D^- = \sup_x |F^*(x) - S(x)|$$

Regla de decisión:

Rechaza H_0 al nivel α si:

$$D, D^+, D^- > W_{1-\alpha}$$

Donde:

$W_{1-\alpha}$ Es el cuantil de una Kolmogorov-Smirnov

Ejemplo

Se efectuaron mediciones del nivel de glucosa en la sangre a 30 pacientes en ayuno, hombres, no obesos y aparentemente sanos.

Concentraciones de glucosa (mg/100 ml)											
93	100	88	91	98	67	87	77	72	95		
63	91	75	67	88	59	83	64	80	68		
90	92	52	85	85	98	60	62	59	100		

Se pretende saber si es posible concluir que tales datos no pertenecen a una población que sigue una distribución normal, con media 80 y desviación estándar de 6.

Hipótesis

$$H_0 : F(x) = F^*(x) \quad \forall x, \text{ de } -\infty \text{ a } \infty$$

$$H_1 : F(x) \neq F^*(x) \quad \text{al menos un valor de } x$$

Procedimiento del cálculo



Estadística II

Unidad 1. Estadística no paramétrica

El primer paso es calcular los valores $S(x)$ como se muestra en la siguiente tabla.

x	Frecuencia	Frecuencia acumulada	S(x)
52	1	1	0.033
59	2	3	0.100
60	1	4	0.133
62	1	5	0.167
63	1	6	0.200
64	1	7	0.233
67	2	9	0.300
68	1	10	0.333
72	1	11	0.367
75	1	12	0.400
77	1	13	0.433
80	1	14	0.467
83	1	15	0.500
85	2	17	0.567
87	1	18	0.600
88	2	20	0.667
90	1	21	0.700
91	2	23	0.767
92	1	24	0.800
93	1	25	0.833
95	1	26	0.867
98	2	28	0.933
100	2	30	1.000
30			

Los valores de $F(x)$ se obtienen al convertir cada valor observado de x en un valor de la normal estándar se observa a continuación



Estadística II

Unidad 1. Estadística no paramétrica

x	$z=(x-80)/6$	F(x)	S(x)	F(x)-S(x)
52	-4.67	0.000002	0.000000	0.000001480
59	-3.50	0.000233	0.000008	0.000224875
60	-3.33	0.000429	0.000014	0.000414758
62	-3.00	0.001350	0.000045	0.001304901
63	-2.83	0.002303	0.000077	0.002226491
64	-2.67	0.003830	0.000128	0.003702701
67	-2.17	0.015130	0.000504	0.014625802
68	-2.00	0.022750	0.000758	0.021991794
72	-1.33	0.091211	0.003040	0.088170846
75	-0.83	0.202328	0.006744	0.195584102
77	-0.50	0.308538	0.010285	0.298252954
80	0.00	0.500000	0.016667	0.483333333
83	0.50	0.691462	0.023049	0.668413713
85	0.83	0.797672	0.026589	0.771082565
87	1.17	0.878327	0.029278	0.849049912
88	1.33	0.908789	0.030293	0.878495821
90	1.67	0.952210	0.031740	0.920469326
91	1.83	0.966623	0.032221	0.934402709
92	2.00	0.977250	0.032575	0.944674872
93	2.17	0.984870	0.032829	0.952040865
95	2.50	0.993790	0.033126	0.96066399
98	3.00	0.998650	0.033288	0.965361765
100	3.33	0.999571	0.033319	0.966251908

El estadístico $D = 0.966251908$ por ser el máximo de las diferencias absolutas.

Con $\alpha = 0.05$ buscamos el cuantil en la tabla de la Kolmogorov-Smirnov ubicada en la pestaña de **“Material de apoyo”**

Como se cumple la condición:

$$D = 0.966251908 > W_{1-0.05} = 0.24170$$

Entonces rechazamos H_0 y por lo tanto los niveles de glucosa no siguen una distribución normal.



Estadística II

Unidad 1. Estadística no paramétrica

1.7.3. Prueba de Kolmogorov-Smirnov para dos muestras

El test quiere probar si dos muestras independientes provienen de la misma población, la diferencia con los test vistos anteriormente como la mediana, la prueba de signos, la *U Mann-Whitney* es que solo toman en cuenta información como la media o la mediana y desperdician otro tipo de información importante como es la variabilidad entre las observaciones.

Datos

Se tienen dos

$$x_1, x_2, \dots, x_{n_1}$$
$$y_1, y_2, \dots, y_{n_2}$$

De tamaño n_1 la primera de ellas y n_2 la segunda.

Supuestos:

- Las muestras son aleatorias
- Las muestras son independientes
- La escala de medida es al menos ordinal
- Se supone que las variables provienen de una función de probabilidad continua

Llamamos:

$F_1(x)$ f. d. p. continua de la primera muestra

$F_2(x)$ f. d. p. continua de la segunda muestra

Hipótesis

A. $H_0 : F_1(x) = F_2(x) \quad \forall x, \text{ de } -\infty \text{ a } \infty$
 $H_1 : F_1(x) \neq F_2(x) \quad \text{al menos un valor de } x$

B. $H_0 : F_1(x) \leq F_2(x) \quad \forall x, \text{ de } -\infty \text{ a } \infty$
 $H_1 : F_1(x) > F_2(x) \quad \text{al menos un valor de } x$

C. $H_0 : F_1(x) \geq F_2(x) \quad \forall x, \text{ de } -\infty \text{ a } \infty$
 $H_1 : F_1(x) < F_2(x) \quad \text{al menos un valor de } x$

Estadístico de prueba



Estadística II

Unidad 1. Estadística no paramétrica

Sean:

$S_1(x)$ la función de distribución empírica de la muestra $x_1, x_2, \dots, x_{n_1}$

$S_2(x)$ la función de distribución empírica de la muestra $y_1, y_2, \dots, y_{n_2}$

El estadístico está definido para las diferentes hipótesis como:

D. Sea el estadístico D_1 la mayor distancia vertical entre $S_1(x)$ y $S_2(x)$

$$D_1 = \sup_x |S_1(x) - S_2(x)|$$

E. Sea el estadístico D_1^+ igual a la mayor distancia vertical de $S_1(x)$ por encima de $S_2(x)$

$$D_1^+ = \sup_x |S_1(x) - S_2(x)|$$

F. Sea el estadístico D_1^- definida como la mayor distancia vertical de $S_1(x)$ por encima de $S_2(x)$

$$D_1^- = \sup_x |S_1(x) - S_2(x)|$$

Regla de decisión

Rechaza H_0 al nivel α si:

$$D_1, D_1^+, D_1^- > W_{1-\alpha}$$

Donde:

$W_{1-\alpha}$ es el cuantil de una Kolmogorov-Smirnov

Utiliza la tabla de inferencia ubicada en la pestaña de “**Material de apoyo**”

Si $n_1 = n_2$ se utiliza la tabla 12 de la **tabla de inferencia** ubicada en la pestaña de “**Material de apoyo**”

Si $n_1 \neq n_2$ se utiliza la tabla 13 de **tabla de inferencia** ubicada en la pestaña de “**Material de apoyo**”

Ejemplo

Se tienen dos muestras aleatorias de tamaño 12 y 10 respectivamente. Se desea probar que ambas muestras provienen de la misma distribución de probabilidad.



Estadística II

Unidad 1. Estadística no paramétrica

Hipótesis

$$H_0 : F_1(x) = F_2(x) \quad \forall x, \text{ de } -\infty \text{ a } \infty$$

$$H_1 : F_1(x) \neq F_2(x) \quad \text{al menos un valor de } x$$

Procedimiento de cálculo

Las dos muestras son ordenadas de menor a mayor por conveniencia y se calculan las funciones empíricas como se muestra a continuación

x_i	y_i	$S_1(x)$	$S_2(x)$	$S_1(x) - S_2(x)$	$ S_1(x) - S_2(x) $
	0.07	0	1/10	0-1/10	0.10
	0.50	0	2/10	0-2/10	0.20
0.62		1/12	2/10	1/12-2/10	0.12
	1.08	1/12	3/10	1/12-3/10	0.22
1.50		2/12	3/10	2/12-3/10	0.13
	1.58	2/12	4/10	2/12-4/10	0.23
2.32		3/12	4/10	3/12-4/10	0.15
2.46		4/12	4/10	4/12-4/10	0.07
	2.48	4/12	5/10	4/12-5/10	0.17
3.00		5/12	5/10	5/12-5/10	0.08
3.18		6/12	5/10	6/12-5/10	0.00
3.95		7/12	5/10	7/12-5/10	0.08
	5.83	7/12	6/10	7/12-6/10	0.02
5.46		8/12	6/10	8/12-6/10	0.07
	5.91	8/12	7/10	8/12-7/10	0.03
	6.68	8/12	8/10	8/12-8/10	0.13
6.78		9/12	8/10	9/12-8/10	0.05
6.90		10/12	8/10	10/12-8/10	0.03
8.56		11/12	8/10	11/12-8/10	0.12
10.35		1	8/10	1-8/10	0.20
	12.03	1	9/10	1-9/10	0.10
	12.04	1	1	1-1	0.00

El estadístico de prueba es $D_1 = 0.23$ por ser el máximo de las diferencias absolutas.

$$D_1, D_1^+, D_1^- > W_{1-\alpha}$$



Estadística II

Unidad 1. Estadística no paramétrica

Buscamos en la **tabla de Kolmogorov–Smirnov** para dos muestras de diferentes tamaños el cuantil con $n_1 = 10$ y $n_2 = 12$, este valor queda incorporado cuando tomamos $n_2 = 15$

$$W_{1-0.05} = W_{0.95} = \frac{17}{36} = 0.47$$

Como $D_1 \not> W_{0.95}$ no rechazamos H_0 y por lo tanto existe evidencia para suponer que las muestras provienen de la misma población.

1.7.4. Otras pruebas de bondad y ajuste

Las pruebas vistas anteriormente son aquellas que se utilizan con mayor frecuencia y son fáciles de localizar en los paquetes estadísticos. Por ejemplo, la prueba de Rao-Scott es una corrección a la prueba Ji-Cuadrada que se realiza cuando se toma en cuenta el diseño muestral.

En particular para la prueba Kolmogorov-Smirnov existen las variantes como la prueba **Anderson Darling** que da mayor peso a las colas de la distribución. La prueba de **Cramér-Von Mises** en donde además de tomar la mayor distancia vertical entre $S_1(x)$ y $S_2(x)$ realiza una corrección dependiendo el tamaño de las muestras.

En el caso de tener múltiples muestras se puede revisar la prueba que propone Birnbaum y Hall. Sin embargo, el cálculo de las pruebas se dificulta a medida que se tienen más de dos poblaciones, por lo que es necesario un paquete estadístico.

Ejemplo 1

Con los datos de glucosa se requiere probar si los datos provienen de una distribución normal con media 80 y desviación estándar de 6 utilizando la prueba Anderson Darling.

Hipótesis

$$\begin{aligned} H_0 : F(x) &= F^*(x) & \forall x, \text{ de } -\infty \text{ a } \infty \\ H_1 : F(x) &\neq F^*(x) & \text{al menos un valor de } x \end{aligned}$$

Procedimiento del cálculo

Acomodamos en orden las observaciones, estandarizamos y obtenemos los valores de $F(x)$ Correspondientes a una distribución normal estándar. Todo esto se había obtenido en el ejercicio anterior. Solo que ahora se realizan unos cálculos extras que se muestran en la siguiente tabla.



Estadística II

Unidad 1. Estadística no paramétrica

i	x	F(xi)	F(xn+1-i)	ln F(xi)	ln F(xn+1-i)	(2i-1)/n*[ln F(xi) - ln F(xn+1-i)]
1	52	0.000002	0.999571	-0.000429	-0.0004292	-0.0435156
2	59	0.000233	0.998650	-0.001351	-0.0013508	-0.1307872
3	60	0.000429	0.993790	-0.006229	-0.0062290	-0.2200996
4	62	0.001350	0.984870	-0.015246	-0.0152458	-0.3136279
5	63	0.002303	0.977250	-0.023013	-0.0230129	-0.4093145
6	64	0.003830	0.966623	-0.033946	-0.0339462	-0.5107312
7	67	0.015130	0.952210	-0.048970	-0.0489701	-0.6205748
8	68	0.022750	0.908789	-0.095643	-0.0956426	-0.7769251
9	72	0.091211	0.878327	-0.129736	-0.1297358	-0.9309137
10	75	0.202328	0.797672	-0.226058	-0.2260583	-1.1995745
11	77	0.308538	0.691462	-0.368946	-0.3689464	-1.5867717
12	80	0.500000	0.500000	-0.693147	-0.6931472	-2.3862944
13	83	0.691462	0.308538	-1.175912	-1.1759118	-3.6432864
14	85	0.797672	0.202328	-1.597863	-1.5978633	-4.9254181
15	87	0.878327	0.091211	-2.394577	-2.3945774	-7.2993690
16	88	0.908789	0.022750	-3.783184	-3.7831843	-11.5459752
17	90	0.952210	0.015130	-4.191066	-4.1910665	-13.4613213
18	91	0.966623	0.003830	-5.564791	-5.5647911	-18.4580599
19	92	0.977250	0.002303	-6.073427	-6.0734271	-21.1492872
20	93	0.984870	0.001350	-6.607726	-6.6077262	-24.1044628
21	95	0.993790	0.000429	-7.753913	-7.7539130	-29.4269942
22	98	0.998650	0.000233	-8.366065	-8.3660653	-33.1513746
23	100	0.999571	0.000002	-13.389833	-13.3898333	-54.3515215
Suma						-230.646200

El estadístico Anderson-Darling es:

$$A^2 = -N - \sum_{i=1}^n \frac{(2i-1)}{n} [\ln F(x_i) - \ln(1 - F(x_{n+1-i}))]$$

$$A^2 = -30 - (-230.646200) = 200.6462$$

El valor crítico con $\alpha = 0.05$ es 0.751 que se puede consultar en la **tabla valores críticos** ubicado en la pestaña “**Material de apoyo**”

Como el valor calculado $A^2 = 200.6462$ es mucho mayor se rechaza la hipótesis nula. Por lo tanto, no existe evidencia estadística suficiente para suponer que los datos siguen una distribución normal. La conclusión coincide con obtenida con la prueba Kolmogorov-Smirnov.



Estadística II

Unidad 1. Estadística no paramétrica

Cierre de la unidad

Durante la unidad 1 aprendiste pruebas que te ayudarán a comparar igualdad de distribuciones, tendencia, independencia de los datos sin necesidad de utilizar supuestos distribucionales y con la oportunidad de poder utilizar variables que sean al menos de tipo ordinal.

Con ayuda de la distribución Ji-Cuadrada podemos comparar poblaciones que están separadas por un antes y n después. En realidad, se trata de la misma población, pero medida en diferentes tiempos.

Finalmente aprendiste técnicas de Bondad de Ajuste para verificar un supuesto distribucional sobre los datos.

En Estadística I y en esta unidad has aprendido pruebas que te ayudarán a contrastar distintas hipótesis con diferentes escalas de medida. En la Unidad 2 desarrollarás modelos con variables correlacionadas, donde una sea la variable a explicar y las demás las variables que expliquen. Te ayudarás de algunas de las pruebas vistas anteriormente para poder hacer inferencia del modelo.

Para saber más

Te recomiendo los siguientes recursos para utilizar el paquete estadístico R en pruebas no paramétricas:

- R-Tutor. (s.f.). Non-Parametric Methods. R-Tutor. <http://www.r-tutor.com/elementary-statistics/non-parametric-methods>
- Chi, Y. (n.d.). *R tutorial, an introduction to statistics*. <http://www.r-tutor.com/elementary-statistics/non-parametric-methods>

Fuentes de consulta

Básicas

- Corder, G. y Foreman, D. (1980). *Noparametric Statistics: A Step-by-Step Approach*. (2.^a ed.). New Jersey: Wiley & Sons.
- Daniel, W. (1990). *Applied Nonparametric Statistics*. (2.^a ed.). Boston: PWS Kent.
- Gibbons, J. y Chakraborti, S. (2003). *Nonparametric Statistical Inference*. (4.^a ed.). New York: Marcel Dekker.



Estadística II

Unidad 1. Estadística no paramétrica

- González, M. y Pérez, A. (2005). *Estadística aplicada, una visión instrumental: teoría y más de 500 problemas resueltos o propuestos con solución*. España: Díaz de Santos.
- Hollander, M. (1999). *Nonparametric Statistical Methods*. New York: J. Wiley.