T.6 CONTRASTES NO PARAMÉTRICOS

1. Introducción
2 .Contraste de bondad de ajuste: correcciones de continuidad y Yates. *ejemplo 1 , ejemplo 2 , ejemplo3*
3. Test de Kolmogorov-Smirnov. *ejemplo 4*
4. Contraste de Independencia. *ejemplo 5*
5. Contraste de Homogeneidad. *ejemplo 6*
6. Test U de Wilcoxon, Mann y Whitney para la comparación de dos muestras independientes
7. Test de Kruskal Wallis para la comparación de varias muestras independientes

---

1. Introducción

En lo estudiado hasta ahora hemos realizado inferencias (estimaciones y contrastes de hipótesis) sobre características desconocidas de la población que resultaban ser parámetros de la distribución de la población. En las argumentaciones teóricas y con vistas a las aplicaciones prácticas era necesario conocer la distribución de la población y postular su dependencia de uno o más parámetros. De esta forma, a partir de las distribuciones muestrales podíamos diseñar métodos para estimar y contrastar los valores de los parámetros, según la distribución de la población de la que se tratara.

Parece lógico que la inferencia estadística no se circunscriba únicamente al tratamiento "decisorio" de los valores de los parámetros de la población, existen características intrínsecas o no de la población que son lo suficientemente importantes y necesarias de conocer que hacen imprescindible su estudio inferencial. Características como : el tipo de distribución de la población , si existe o no independencia de esta respecto a otra , la presumible homogeneidad de comportamiento frente a diversos valores de un factor , la existencia de rachas, etc. hacen necesario su estudio mediante inferencias que por no hacer referencia a los parámetros de la población denominamos "inferencia NO paramétrica.

Los contrastes no paramétricos serán así, y por tanto, una metodología de trabajo idéntica a la estudiada para el caso paramétrico pero con la diferencia evidente de que las hipótesis planteadas no harán referencia al valor o relaciones de valor de los desconocidos parámetros o ratios de la población.

En los contrastes no paramétricos las "pruebas" no están determinadas por condiciones férreas para su aplicación, a diferencia clara de lo que ocurría en los "si" paramétricos en los que condiciones como la normalidad o pseudo-normalidad (por muestral tamaño grande) eran cuasi-imprescindibles.

Durante algún tiempo los contrastes no paramétricos han sido, en cierto modo, denostados cuando en realidad poseen unas cualidades que los hacen imprescindibles para el adecuado tratamiento de la información que muchas llega a nosotros para su estudio. Constituyéndose para muchos como una "estadística No paramétrica". Prueba

de ello son las características que de este tipo de contrastes enumera Bradley (1968), y que pasamos a enumerar:

- Simplicidad de deducción: Los contrastes no paramétricos son más sencillos matemáticamente que los paramétricos y se deducen, claro está, de expresiones más sencillas.
- Rapidez y simplicidad de manejo: Los no paramétricos son más sencillos de realizar, las operaciones necesarias son más simples .Esta argumentación se mantiene (está realizada en 1968) incluso en la era de la informática.
- Campo de aplicación: Las hipótesis de aplicación en un no paramétrico son menos detalladas y restrictivas
- Sensibilidad respecto a la violación de las hipótesis: Dado que las hipótesis de aplicación son menos restrictivas y numerosas es más fácil percatarse de su no cumplimiento o violación
- Tipos de medición exigida: Los no paramétricos requieren datos ordinales o nominales incluso en escala de intervalo, mientras que los paramétricos requieren escala de razón o intervalos.
- Tamaño de la muestra: Con tamaños muestrales inferiores a 10 en los contrates paramétricos al no poder aplicar normalidad se cometen violaciones de las hipótesis de aplicación (normalidad), por lo que son más adecuados los tests-pruebas-contrastes no paramétricos; que disminuyen su eficiencia precisamente cuanto mayor se el tamaño muestral.

Cabe añadir a las características enumeradas otra planteada por J. Etxeverría, L.Joaristi y L.Lizasoain (1990) según la cual : la potencia de los contrastes no paramétricos es superior o por lo menos igual a la de los paramétricos siempre que se mantengan las hipótesis de aplicación en los "no" y no lo hagan en los paramétricos ; cabe recordar a este respecto que los paramétricos son más "sensibles" a la posible violación de las hipótesis de aplicación.

Por lo expuesto es evidente la gran importancia que poseen este tipo de contrastes. Es evidente que no podemos desarrollar la gran cantidad de pruebas (tests) no paramétricas que se han desarrollado en los últimos tiempos, si bien plantearemos en estas páginas las más conocidas y clásicas.

Como hemos dicho la "filosofía" (metodología de actuación) de estos tests se basa en la expuesta para el caso paramétrico, con hipótesis nula, alternativa, nivel de significación, hipótesis de aplicación, creación de estadístico, comparación, decisión. Dicho lo cual pasamos a enumerar una posible clasificación de este tipo de tests y que no es otra que la utilizada en el conocido "paquete" informático-estadístico "SPSS/pc", y que en esquema resultaría el siguiente:

MUESTRA	*ESCALA NOMINAL*	*ESCALA ORDINAL*
UNA MUESTRA	χ^2 RACHAS BINOMIAL	KOLMOGOROV-SMIRNOV
DOSMUESTRAS REALACIONADAS	McNEMAR	SIGNOS WILKOXON
K MUESTRAS *RELACIONADAS*	Q DE COCHRAN	FRIEDMAN KENDALL
DOS MUESTRAS *INDEPENDIENTES*		MEDIANA MANN-WHITNEY KOLMOGOROV-SMIRNOV WALD-WOLFOWITZ MOSES
K MUESTRAS *INDEPENDIENTES*		MEDIANA KRUSKAL-WALLIS

En el esquema se puede observar que los tests abarcan el estudio de cualidades no paramétricas en base al conocimiento lógico de la muestra y que el tipo de información (tipo de datos) genera la existencia de diversos tests específicas .En la clasificación no se explicita la hipótesis a contrastar pues algunos tests pueden ser utilizados para el contraste varias hipótesis distintas. Siendo, por tanto, este esquema un marco de actuación primario al respecto del punto de partida muestral (una, dos, k muestras relacionadas con la misma población o no).

Dicho esto, pasamos a desarrollar algunos tests, haciéndolo al respecto de la hipótesis a contrastar y basándonos, claro está, en unas ciertas características muestrales.

2. Contraste de la bondad de un ajuste

También se le conoce como contraste de adherencia a un ajuste, o como contraste de la χ^2.

La hipótesis a contrastar es el hecho (nótese que no hay valor de un parámetro) de que la muestra proviene de una distribución determinada y planteada de probabilidad, frente a la alternativa de que esto no es así.

Se parte de una sola muestra (lógico) normalmente en datos en forma de escala nominal, de ahí que este test se encuentre ubicado donde los está en la tabla resumen que antes

presentamos
A través de este contraste, y partiendo de los datos muestrales, se obtiene un criterio de decisión sobre la hipótesis de que la población de la que se ha extraído la muestra se distribuya (se ajuste bien, se adhiera), o no, según algún modelo teórico determinado y planteado a priori. Así:

H_0 : la muestra proviene(ajusta ,adhiere) a una población cuya función es (F(x))
H_1 : la muestra NO proviene(ajusta ,adhiere) a una población cuya función es (F(x))

y trabajando con un determinado nivel de significación plantearíamos que si: Las observaciones muestrales podremos considerarlas y disponerlas como una distribución de frecuencias ;frecuencias , claro está ,***observadas.***

x_i	$n_{observadas.i}$
x_1	$n_{o,1}$
x_2	$n_{o,2}$
.	.
x_m	$n_{o,m}$

Si la población sigue un determinado modelo teórico de distribución de probabilidad cada posible valor de la variable x_i tendrá asociada una determinada probabilidad, según ese modelo teórico.

Para cada uno de los valores muestrales podremos construir su distribución de probabilidad:

x_i	$P(x_i)$
x_1	P_1
x_2	P_2
.	.
x_m	P_m

Si multiplicamos, para cada x_i, su probabilidad, P_i, por el número total de observaciones, n , obtendremos las frecuencias que teóricamente debían corresponder a cada valor de la variable, según el modelo, ($P_i \cdot n = n_{teórica,i}$).

Y, así, podremos construir una distribución de frecuencias teóricas:

x_i	$n_{teóricas,i}$
x_1	$n_{t,1}$
x_2	$n_{t,2}$
.	.
x_m	$n_{t,m}$

A partir de la distribución de frecuencias observadas y de la de frecuencias teóricas puede construirse el siguiente estadístico:

$$\chi^2 = \sum_{i=1}^{m} \frac{\left(n_{O,i} - n_{t,i}\right)^2}{n_{t,i}}$$

donde m es el número de valores de la variable que se han muestreado (valores distintos)

Puede demostrarse que si la distribución de la población es efectivamente la utilizada para construir las frecuencias teóricas, el estadístico anterior se distribuye como una

es decir una chi-dos con m-k-1 grados de libertad, donde k es el número de paramétros estimados a partir de los datos muestrales y necesarios para la construcción de la tabla de frecuencias de la distribución teórica. La pérdida de k+1 grados de libertad se debe precisamente a que con los m datos se calculan k parámetros (habrá, por tanto, k ecuaciones que liguen los m datos) y, además, la suma de los m datos debe dar el número total de observaciones n (lo que supone una nueva ligadura):

m grados de libertad iniciales - (k + 1) ligaduras = m-k-1 grados de libertad finales

Debe observarse que este estadístico se distribuye siempre como una sea cual fuere el modelo teórico (binomial, Poisson ,normal ,exponencial ,cualquiera de los estudiados, u otro diseñado "ad hoc"), siempre y cuando la población se distribuya, efectivamente, según ese modelo.

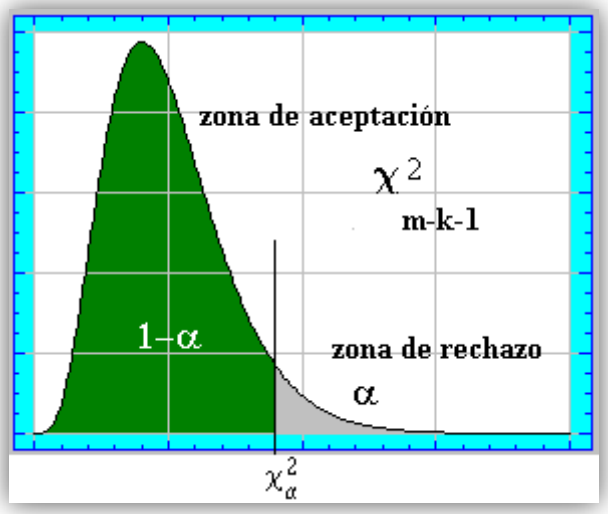
Teniendo en cuenta esto ,si queremos contrastar la hipótesis de que una cierta población sigue un modelo determinado, con un nivel de significación a , habrá que diseñar una región crítica según la cual si los datos muestrales nos conducen (bajo el supuesto de la hipótesis) a un estadístico χ^2 que pertenezca a ella rechazaremos la hipótesis.

Según la definición de nivel de significación a se habrá de cumplir que:

$$P\left[\chi^2 \in R_1 / \text{modelo es cieto}\right] = \alpha$$

Donde

χ^2 es el estadistico y R_1 es la region critica



Como siempre, de todas las posibles regiones que cumplen esa condición, escogeremos aquella que tenga mayor amplitud (lo que supone mayor amplitud de la zona de rechazo y, en consecuencia menor amplitud de la zona de aceptación) para poder realizar un contraste severo.

Teniendo en cuenta que el estadístico sigue una distribución , la región crítica de mayor amplitud será la cola de la derecha.

Así pues, una vez calculado el estadístico si:

$$\chi^2 < \chi^2_{\alpha(m-k-1)}$$

no rechazaremos la hipótesis de que la población sigue el modelo de probabilidad planteado ; mientras que si:

$$\chi^2 \geq \chi^2_{\alpha(m-k-1)}$$

rechazaremos la hipótesis

Por último quedan hacer dos observaciones finales sobre este contraste:

El estadístico χ^2 se calcula a partir de conteos discretos de las frecuencias para cada posible valor de la variable y, como es bien sabido, la distribución χ^2 es una distribución de variable continua. Si las frecuencias esperadas para todos los valores de la variable $n_{teóricas,i}$ $\forall$ i son grandes , este hecho no plantea problemas.

Pero si alguna de las frecuencias teóricas es inferior a 5 será necesario subsanar este inconveniente agrupando las observaciones adyacentes.

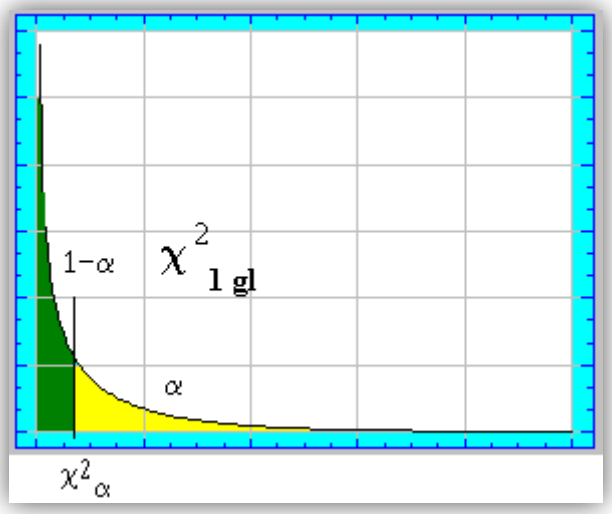
A modo de ejemplo:

en esta tabla existen frecuencias inferiores a 5

x	$n_{observadas,i}$	$n_{teóricas,i}$
..	..	..
x_i	6	4
x_{i+1}	1	2
..		

lo que resolveríamos de la siguiente forma

x	$n_{observadas,i}$	$n_{teóricas,i}$
..	..	..
x_i+x_{i+1}	6+1=7	4+2=6
..		



Por otro lado, cuando resulta que los grados de libertad implicados son sólo 1 (es decir, m-k-1 = 1) el estadístico χ^2 toma un sobrevalor que "infla" la rechazabilidad de la hipótesis, dado que la distribución chi-dos con un solo grado de libertad se eleva de forma evidente en la parte próxima a cero haciendo que el valor crítico, que divide las zonas, sea muy próximo a éste, primando, por ello, la rechazabilidad de la hipótesis. Para paliar esto, el americano Yates probó que es conveniente cuando m-k-1 = 1 utilizar como estadístico el siguiente:

$$\chi'^2 = \sum_{i=1}^{m} \frac{\left(|n_{o,i} - n_{t,i}| - 0{,}5\right)^2}{n_{t,i}}$$

Estas dos puntualizaciones deberán tenerse presentes a la hora de realizar los contrastes de adherencia a un ajuste (hipótesis: Población sigue cierto modelo), así como al realizar los contrastes de independencia (tablas de contingencia que estudiaremos después, cuando sean oportunas - son lo que se conoce como correcciones de continuidad de la prueba o test de la χ^2.

ejemplo 1

En 60 lanzamientos de una moneda se observaron 37 caras y 23 cruces. Contrastar la hipótesis de que la moneda no está trucada, con un nivel de significación del 1%.

H_o : la moneda no está trucada : luego P (c) =0,5 =P(+)

la información muestral y teórica quedaría establecida como:

x	$n_{0,i}$	$n_{t,i}=P_i \cdot n$
c	37	0,5·60=30
+	23	0,5·60=30
	60=n	60

el número de observaciones distintas es 2 (c y +) luego los grados de libertad para la chi-cuadrado serían m-k-1=2-0-1=1 , por lo que es necesaria la corrección de Yates , así :

$\lvert n_{o,i} - n_{t,i} \rvert$	$\left(\lvert n_{o,i} - n_{t,i} \rvert - 0{,}5\right)^2$	$\dfrac{\left(\lvert n_{o,i} - n_{t,i} \rvert - 0{,}5\right)^2}{n_{t,i}}$	variable
7	$(6,5)^2$	1,40833	c
7	$(6,5)^2$	1,40833	+
		χ^2=2,81667	

el valor crítico para un nivel de significación del 1% será : (ir a tabla de la χ^2)

$$\chi^2_{\alpha=0.01}(1\ \text{g.l.}) = 6{,}63$$

dado que χ^2 = 2,81667 < valor crítico=6,63

No podemos rechazar que la moneda no esté trucada.

ejemplo 2

Contrastar con un nivel de significación del 1 % la hipótesis de que la mitad de los accidentes de tráfico se producen los fines de semana (y en igual proporción el sábado y el domingo), mientras que la otra mitad se distribuye uniformemente a lo largo de los restantes cinco días de la semana, si se dispone de los siguientes datos:

Dado que el número de valores de la variable es 7 los grados de libertad serían m-k-1=7-0-1=6 no es de aplicación Yates

día	número de accidentes
lunes	12
martes	13
miércoles	12
jueves	8
viernes	14
sábado	25
domingo	16
total =n	100

la distribución teórica sería mitad de accidentes el fin de semana luego 100 a razón de 50 sábado y 50 domingo y el resto otros 100 a repartir uniformemente; 10 cada día de entre semana. Así

día	$n_{0,i}$	$n_{t,i}$	$(n_{o,i}-n_{t,i})^2$	$\frac{(n_{o,i}-n_{t,i})^2}{n_{t,i}}$
L	12	10	4	0,4
M	13	10	9	0,9
X	12	10	4	0,4
J	8	10	4	0,4
V	14	10	16	1,6
S	16	25	0	0
D	25	25	81	3,24
total	100	100		6,94

luego el estadístico $\chi^2 = 6{,}94$ (ir a tabla de la χ^2)

$$\chi^2_{\alpha=0.01}(6 \text{ g.l.}) = 16{,}8$$

dado que $\chi^2 = 6{,}94 <$ valor crítico=6,94

No podemos rechazar la hipótesis de distribución de probabilidad del número de accidentes por día de la semana establecida.

ejemplo 3

Se realiza una muestra de 1500 observaciones y resulta que obtenemos que los valores de x observados han sido:

x	observaciones
0	543
1	560
2	280
3	90
4	25
5	2

Contrastar con un nivel de significación del 1 % la hipótesis de que la población sigue una distribución de Poisson.

Para determinar la distribución teórica de frecuencias, será necesario primero estimar el valor del parámetro l de la distribución de Poisson que se supone que sigue. Habrá que estimarlo a partir de los datos muestrales.

El estimador máximo-verosímil del parámetro λ de una población de Poisson es la media muestral . Si calculamos la media de la muestra, ésta será:

Ahora calcularemos las probabilidades de cada valor de la variable en una distribución de Poisson con $\lambda = 1$: aplicando la función de cuantía de la Poisson obtendríamos la tabla:

x	$P(x_i)=\frac{e^{-1}\cdot 1^x}{x_i!}$	$n_{0,i}$	$n_{t,i}=n\cdot P(x_i)$	$(n_{o,i}-n_{t,i})^2$	$\frac{(n_{o,i}-n_{t,i})^2}{n_{t,i}}$
0	0,367879	534	551,82	77,79	0,14
1	0,367879	560	551.82	66,91	0,12
2	0,183940	280	275,91	16,73	0.06
3	0,061313	90	91,97	3,98	0,04
4	0,015328	25	22,99	4,41	0,19
5	0,003660	2	5,49	12,18	2,22
	≅ 1	1500	1500		2,78=χ^2

dado que m=6 que son los valores distintos de la variable

k=1 dado que hemos utilizado la muestra para conseguir estimar el parámetro λ

la chi cuadrado necesaria para el cálculo del valor crítico seria de m-k-1=6-1-1=4 grados de libertad lo que daría un valor critico para α =0,01 de

$\chi^2_{\alpha=0,01}(4\ g.l.)=13,3$ (ir a tabla de la χ^2)

no siendo necesarias las correcciones de continuidad. Dado que el valor del estadístico (2,78) es menor que 13,3 (valor crítico) no podemos rechazar la hipótesis de que la muestra proceda, se ajuste, o se adhiera a una población con distribución de Poisson de λ =1

3. Test de Kolmogorov-Smirnov

La hipótesis nula a contrastar es similar al caso anterior; se trata, por tanto, de comprobar si la muestra se ajusta o proviene de una población con una determinada distribución de probabilidad. Como se planteó en el esquema el test de K-S es más adecuado cuando la muestra viene planteada en escala ordinal.

El procedimiento es similar al del test de la chi-2, se trata de comparar la distribución muestral observada con la resultante de dar por cierta la hipotética distribución de la población .En el caso de la chi-2 se comparaban frecuencias absolutas observadas con sus homónimas teóricas, en el caso del test de Kolmogorov-Smirnov las frecuencias a comparar serán las frecuencias relativas acumuladas $F(x_i)$ de las dos distribuciones; observada y teórica. De ahí su utilidad para aquellas ocasiones en las que los datos se encuentren en forma de escala ordinal.

Escuetamente el procedimiento consiste en establecer las frecuencias relativas acumuladas referentes a la información muestral. $F_o(x_i)$. Establecer, también, en base a la distribución de probabilidad hipotética las frecuencias relativas acumuladas $F_t(x_i)$.

Compararemos ambas frecuencias creando el estadístico

$$D = max\left|F_t(x_i) - F_o(x_i)\right|$$

es decir el valor máximo de entre todas las diferencias entre frecuencias relativas acumuladas teóricas y observadas para los mismos valores o intervalos de la variable.

Dicho estadístico D se comparará con el correspondiente de la tabla del tests de K-S (ir a tabla de K-S) en base al nivel de significación establecido y el tamaño muestral; de manera que si

$D<D(tabla,n,\alpha)$ no rechazaremos la hipótesis de que la muestra procede de la hipotética población con distribución establecida , mientras que si $D>D(tabla,n,\alpha)$ rechazaremos dicha hipótesis.

ejemplo 4

Se ha realizado una muestra a 178 municipios al respecto del porcentaje de población activa dedicada a la venta de ordenadores resultando los siguientes valores:

porcentaje	nº de municipios
menos del 5%	18
entre el 5 y 10 %	14
entre 10 y 15%	13
entre 15 y 20%	16
entre 20 y 25 %	18
entre 25 y 30 %	17
entre 30 y 35 %	19
entre 35 y 40 %	24

entre 40 y 45 %	21
mas de 45%	18

queremos contrastar que el porcentaje de municipios para cada grupo establecido se distribuye uniformemente con un nivel de significación del 5%.

Bajo la hipótesis nula cada grupo debiera de estar compuesto por el 10% de la población dado que existen diez grupos. Así podemos establecer la tabla

grupos - variable	$n_{0,i}$	$F_0(x_i)$	$n_{t,i}=n\cdot P(x_i)$	$F_0(x_i)$	$D = max\left\|F_t(x_i) - F_o(x_i)\right\|$
menos del 5%	18	18/178=0,1011	17.8	17.8/178=0,1	0.0011
entre el 5y10 %	14	32/178=0,1798	17.8	35.6/178=02	0,0202
entre 10 y 15%	13	0,2584	17.8	0,3	0,0416
entre 15 y 20%	16	0,3427	17.8	0,4	0,0573
entre 20 y 25 %	18	0,4439	17.8	0,5	0,0561
entre 25 y 30 %	17	0,5393	17.8	0,6	***0,0607 max***
entre 30 y 35 %	19	0,6461	17.8	0,7	0,0539
entre 35 y 40 %	24	0,7809	17.8	0,8	0,0191
entre 40 y 45 %	21	0,8989	17.8	0,9	0,0011
mas de 45%	18	1	17.8	1	0

siendo la máxima diferencia y por tanto el estadístico de K-S que compararemos con el establecido en la tabla que será para un nivel de significación de 5% y una muestra de 178 (ir a tabla K-S)

$$D_{0.05,178} = \frac{1,36}{\sqrt{n}} = \frac{1,36}{\sqrt{178}} = 0,1019$$

dado que el estadístico es menor (0,0607) que el valor de la tabla (0,1019) no rechazamos la hipótesis de comportamiento uniforme de los grupos establecidos al respecto de la población activa dedicada a la venta de ordenadores.

4. Contraste de Independencia (Tablas de Contingencia)

A través de este contraste pretendemos probar si existe independencia entre dos variables o atributos (en el conjunto de la población) a partir de las observaciones de las dos característica (en una muestra).Se trata, en realidad, de un caso particular del contraste de adherencia a un ajuste, en el que el modelo teórico sujeto a contraste es el de una distribución bidimensional con variables independientes.

Las frecuencias observadas las podemos disponer en una tabla de contingencia:

X\Y	y_1	y_2	y_j	y_m	
x_1	$n_{1,1}$	$n_{1,2}$	·	·	$\mathbf{n_{1,*}}$
x_2	$n_{2,1}$	$n_{2,2}$	·	·	$\mathbf{n_{2,*}}$
x_i	·	·	$n_{i,j}$	·	$\mathbf{n_{i,*}}$
·	·	·	·	·	
x_n	·	·	·	$n_{n,m}$	$\mathbf{n_{n,*}}$
	$\mathbf{n_{*,1}}$	$\mathbf{n_{*,2}}$	$\mathbf{n_{*,j}}$	$\mathbf{n_{*,m}}$	**N**

Donde: $n_{i,j}$ es la frecuencia conjunta

$n_{i,*}$ es la frecuencia marginal de x

$n_{*,j}$ es la frecuencia marginal de y

Si la hipótesis de independencia se cumple, y por el teorema de caracterización, se deberá cumplir que todas las frecuencias relativas conjuntas sean iguales al producto de las respectivas frecuencias relativas marginales:

$$\frac{n_{i,j}}{N} = \frac{n_{i,\cdot}}{N} \cdot \frac{n_{\cdot,j}}{N} \quad \forall i, \forall j$$

luego en el caso de independencia cada una de las ij frecuencias conjuntas teóricas serán:

$$n_{i,j}^{T} = \frac{n_{i,\bullet} * n_{\bullet,j}}{N}$$

si establecemos el mismo método del test de la chi-2 crearemos el estadístico

$$\chi^2 = \sum_{\forall i}\sum_{\forall j} \frac{\left(n_{i,j} - n_{i,j}^{T}\right)^2}{n_{i,j}^{T}}$$

hay que puntualizar que el citado estadístico se distribuirá con una distribución χ^2 con (m-1)(n-1) grados de libertad.

Las frecuencias conjuntas debe verificar siempre :

para cada fila

$$\sum_j n_{ij} = n_{i\bullet} \rightarrow \text{lo que supone n ecuaciones o ligaduras}$$

para cada columna

$$\sum_j n_{ij} = n_{\bullet j} \rightarrow \text{lo que supone m ecuaciones o ligaduras}$$

pero además :

$$\sum_i \sum_j n_{ij} = N = \sum_i n_{i\bullet} = \sum_j n_{\bullet j}$$

una de las m + n ecuaciones anteriores será combinación lineal de las otras m+n-1.

De manera que de los m.n sumandos que constituyen el estadístico (m.n celdas de la tabla) ,m+n-1 están determinados por los demás y quedan por lo tanto:

m·n -(m+n-1) libres = m·n - m - n + 1 = (m-1).(n-1).

Como no estima ningún parámetro el número de grados de libertad será el número de sumandos (variables) libres (independientes): por tanto el estadístico seguirá

$$\chi^2 \rightarrow \chi^2_{(m-1)*(n-1)}$$

ejemplo 5

Se dispone de las observaciones del color del pelo y de los ojos de 400 individuos según la siguiente tabla:

	ojos azules	ojos negros	ojos pardos	
rubios	120	20	20	160
castaños	50	30	60	140
morenos	50	10	40	100
	220	60	120	400

Contrastar con un nivel de significación del 1 % la independencia de estos atributos.

Construyamos primero la tabla de frecuencias teóricas: aplicando para cada valor la expresión

$$n_{i,j}^{T} = \frac{n_{i,\bullet} \cdot n_{\bullet,j}}{N}$$

construimos la tabla de contingencia de frecuencias teóricas

	ojos azules	ojos negros	ojos pardos	
rubios	88	24	48	160
castaños	77	21	42	140
morenos	55	15	30	100
	220	60	120	400

Construimos el estadístico

$$\chi^2 = \sum_{\forall i} \sum_{\forall j} \frac{\left(n_{i,j} - n_{i,j}^{T}\right)^2}{n_{i,j}^{T}}$$

que tomará el valor 55,13

dado que los grados de libertad serán: g. l.=(3-1)(3-1) = 4

y el valor crítico para α =0.01 y g. l .= 4 es 13.3 (ir a tabla de la χ^2) ;de modo que dado que el estadístico es mayor que el valor de la tabla 55,13>13,3 rechazamos la hipótesis planteada. En consecuencia podemos concluir que existe dependencia entre el color de los ojos y el del pelo.

5. Contraste de Homogeneidad

A través de este contraste pretendemos determinar si varias poblaciones distintas (m) tienen una estructura similar o, por decirlo de otro modo, si son o no homogéneas en lo que se refiere a la forma de distribuirse en ellas una cierta variable o atributo que puede tomar un conjunto de n valores o tipos diferentes (en todas las poblaciones). Para ello partiremos de la información de m muestras de las m poblaciones y trabajaremos con las frecuencias que en cada población tiene cada posible valor de la variable (o tipo del atributo).

Si llamamos $n_{i,j}$ a la frecuencia con que se da el valor o tipo i en la muestra j, podemos construir una tabla con los datos similar a la que utilizábamos en el contraste de independencia.

La hipótesis que queremos contrastar es que la distribución de la variable (o atributo) es homogénea en las j poblaciones, por lo tanto la frecuencia teórica con que se dará el valor o el tipo x_i deberá ser tal que la proporción de observaciones (frecuencias relativas) del valor o tipo x_i deberá ser la misma en todas las muestras

Por lo que se cumplirá que:

$$\frac{n_{i1}}{n_{\bullet 1}} = \frac{n_{i2}}{n_{\bullet 2}} = \bullet\bullet\bullet\bullet\bullet = \frac{n_{im}}{n_{\bullet m}}$$

para i= 1,2,3,….n

y será obviamente la misma que la proporción de observaciones de ese tipo que hay en el total , que será:

$$\frac{n_{i\bullet}}{N}$$

para cada x_i siendo N el total de observaciones

De manera que las frecuencias teóricas deberán verificar:

$$n_{i,j}^{T} = \frac{n_{i,\bullet} * n_{\bullet,j}}{N}$$

así construida la tabla de contingencia de frecuencias teóricas y comparada con la de observadas , crearemos el estadístico ya conocido

$$\chi^2 = \sum_{\forall i}\sum_{\forall j} \frac{\left(n_{i,j} - n_{i,j}^{T}\right)^2}{n_{i,j}^{T}}$$

que como en el caso d e contraste de independencia seguirá el modelo

$$\chi^2 \rightarrow \chi^2_{(m-1)(n-1)}$$

ejemplo 6

Para intentar mejorar la posición en el mercado de cierto producto se llevaron a cabo tres campañas de promoción entre los minoristas distribuidores en otras tantas localidades: A,B,C. Se desea contrastar si las tres campañas son homogéneas respecto a los resultados en el incremento de las ventas en las tiendas, con un nivel de significación del 5 % .Para ello se han recogido los siguientes datos:

	localidad A	localidad B	localidad C	$n_{i,\bullet}$
tiendas aumentan ventas	165	141	152	458
tiendas no aumentan ventas	256	142	98	496
$n_{\bullet j}$	421	283	250	954

si calculamos las frecuencias teóricas mediante $n_{ij}^T = \frac{n_{i,\bullet} \cdot n_{\bullet j}}{N}$ tendremos:

	localidad A	localidad B	localidad C	$n_{i,\bullet}$
tiendas aumentan ventas	202,1153	135,86373	120,02096	458
tiendas no aumentan ventas	218,8847	147,13626	129,97903	496
$n_{\bullet j}$	421	283	250	954

aplicando la expresión del estadístico

$$\chi^2 = \sum_{\forall i}\sum_{\forall j} \frac{\left(n_{i,j} - n_{i,j}^T\right)^2}{n_{ij}^T}$$

quedará su valor establecido en 29.871131

el estadístico tendrá una distribución chi-dos con (m-1)·(n-1) g.l. es decir (3-1)·(2-1)=2 (ir a tabla de la χ^2)

el valor crítico para dicha distribución con nivel de significación α =0.05 será 5.99;

dado que el estadístico es mayor rechazamos la hipótesis de que las tres campañas de promoción sean homogéneas.

6. Test U de Wilcoxon, Mann y Whitney para la comparación de dos muestras independientes

Este test debido a Mann y Whitney (1947) y basado en el Wilcoxon para muestras independientes es en cierto modo el equivalente no paramétrico del test t para la comparación de medias de dos distribuciones .Es seguramente una de las pruebas más potentes de entre las no paramétricas.La aplicación de la prueba exige que los datos de ambas muestras vengan medidos, al menos en escala ordinal, y su correcta ejecución requiere que las distribuciones muestrales tengan la misma forma (asimetría y curtósis).

En estas circunstancias, para contrastar si el comportamiento de ambas poblaciones es semejante se contrasta la hipótesis nula de que "*la probabilidad de que una observación aleatoria de la primera población supera a una observación aleatoria de la segunda población es 0.5*" frente a la alternativa de que está probabilidad es distinta a 0.5 (pudiéndose plantear bilateral o unilateralmente)

La prueba parte de N_1 valores aleatorios de la primera población y de otros N_2 valores aleatorios de la segunda. Por ejemplo:

Poblacion 1			Población 2
16			12
20			18
11			17
15			

Para llevarla a cabo se ordenan de forma creciente la totalidad de las observaciones en una sola serie especificando la población de origen. Por ejemplo:

Orden	1	2	3	4	5	6	7
Observación	11	12	15	16	17	18	20
Población	1	2	1	1	2	2	1

A partir de aquí se pueden obtener los estadísticos U_1 y U_2 definidos como:

$$U_1 = N_1 N_2 \frac{N_1(N_1+1)}{2} - R_1$$
$$U_2 = N_1 N_2 \frac{N_2(N_2+1)}{2} - R_2$$

Donde N_1 y N_2 son los tamaños muestrales de cada una de las dos muestras y R_1 y R_2 la suma de los rangos de cada una de las dos muestras:

En nuestro caso:

R_1= 1+3+4+7= 15

R_2= 2+5+6=13

Con lo que los dos estadísticos quedarían:

$$U_1 = (4\times3+(4\times5/2))-15=1.5$$

$$U_2 = (3\times4+(3\times4/2))-13=2.5$$

Finalmente se compara el estadístico U= mínimo de U_1 y U_2 con el correspondiente valor crítico, $U(N_1,N_2,a)$, tabulado. De forma que fijado un determinado nivel de significación, a si:

- $U<U(N_1,N_2,a)$ se aceptará (no se rechazará) la hipótesis nula, y por lo tanto mantendremos la hipótesis de que las dos muestran provienen de dos poblaciones de comportamiento semejante.
- $U> U(N_1,N_2,a)$ rechazaremos la hipótesis nula , y por tanto concluiremos que no se puede mantener la hipótesis de que ambas muestran provengan de poblaciones semejantes.

En nuestro ejemplo el valor crítico para a= 0.10 y un contraste bilateral: U(3,4,0.1)=1 , por lo que no podríamos rechazar la hipótesis nula con ese nivel de significación.

Alternativamente a la comparación con el valor tabulado se puede comparar el valor del estadístico con la siguiente aproximación válida para tamaños muestrales grandes (N_1+N_2 >60):

$$U(N_1,N_2,\alpha) = \frac{N_1N_2}{2} + Z_{\alpha}\sqrt{\frac{N_1N_2(N_1+N_2+1)}{12}} \quad \text{para el contraste unilateral}$$

$$U(N_1,N_2,\alpha) = \frac{N_1N_2}{2} + Z_{\alpha/2}\sqrt{\frac{N_1N_2(N_1+N_2+1)}{12}} \quad \text{para el contraste bilateral}$$

Cuando no se pueda o no se quiera fijar de antemano un nivel de significación, o bien cuando no se disponga de valores tabulados para $U(N_1,N_2,a)$ se puede obtener el nivel de significación asociado (de una o dos colas) a partir de la siguiente aproximación (válida para N_1> 8 , N_2 > 8)

$$Z = \frac{\left|U - \frac{N_1N_2}{2}\right|}{\sqrt{\frac{N_1N_2(N_1+N_2+1)}{12}}}$$

7. Test de Kruskal Wallis para la comparación de varias muestras independientes

Este contraste, debido a Kruskal_y Wallis, viene a ser una generalización del test de Mann-Whitney para el caso de k muestras independientes. Se examina la hipótesis de que estas muestras provengan de la misma población o de poblaciones de idéntico comportamiento, frente a la alternativa de que no todas provienen de la misma población. En este sentido es una versión "no parámetrica" del Análisis de la Varianza. Este test, al igual que el de Mann-Whitney presenta una muy alta eficiencia (o potencia relativa). Siendo más potente que otras pruebas no paramétricas similares como la de la mediana.

La situación de aplicación es tal que disponemos de k muestras independientes de tamaños: $n_1,n_2,...,n_k$ con $n_1+ n_2+...+ n_k = n$ donde estas observaciones están medidas en, al menos escala ordinal.(ordenación)

Si llamamos R_i a la suma de los rangos de las observaciones de la i-sima muestra el estadístico H:

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

sigue aproximadamente una distribución χ^2 con k-1 grados de libertad:

De forma que, dado un nivel de significación, α, si el estadístico evaluado, H, supera el valor crítico, $\chi^2_\alpha(\kappa-1)$ rechazaremos la hipótesis nula de que las muestras provienen de poblaciones iguales.

ir script de realización para tres muestras
ir script de realización para cuatro muestras