TEMA 9. Contrastes no paramétricos y bondad de ajuste

9.1 Al finalizar el tema el alumno debe conocer........

- ✓ Diferencia entre un contraste parámetrico y uno no paramétrico
- ✓ Características de la estimación utilizando los contrastes o test de hipótesis no paramétricos.
- ✓ Conocer los tipos principales de contrastes no paramétricos,
- ✓ Utilizar el contraste no paramétrico adecuado.
- ✓ Interpretar el resultado del contraste.

9.2 Contrastes no paramétricos.

Previamente ya hemos planteado la diferencia entre contraste paramétrico y no paramétrico. Un contraste paramétrico es aquel en el que las hipótesis planteadas se refieren a un valor desconocido de la población y un contraste no paramétrico es aquel que se refiere a otras características de la población (forma de la distribución, localización…)
En este tema se estudiaran los contrastes no paramétricos.

En estos contrastes, no conoceremos la forma funcional de la distribución como antes, pero existen medios alternativos, que estudiaremos y que permitirán hacer inferencias sobre la población. Estos medios alternativos son los denominados contrastes no paramétricos o de libre distribución.

Al no conocer la distribución de partida, se utilizarán una serie de estadísticos que serán independientes de la distribución de partida.

Las pruebas no paramétricas tienen varias ventajas sobre las pruebas paramétricas:

1.- Por lo general, son fáciles de usar y entender.
2.- Eliminan la necesidad de suposiciones restrictivas de las pruebas paramétricas.
3.- Se pueden usar con muestras pequeñas.
4.- Se pueden usar con datos cualitativos.

También las pruebas no paramétricas tienen desventajas:

1.-Pueden, ignorar, desperdiciar o incluso prescindir de información.
2.- No son tan eficientes (potentes) como las paramétricas.

Los contrastes no paramétricos que vamos a estudiar se clasifican en los siguientes grupos:

- ✓ **Contrastes de aleatoriedad**: contrastan hipótesis sobre la aleatoriedad de las muestras
- ✓ **Contrastes de localización**: Contrastan hipótesis sobre medidas de posición
- ✓ **Contrastes de comparación de poblaciones**: contrastan hipótesis sobre las distribuciones poblacionales

9.3 Contrastes de aleatoriedad.

En ocasiones hemos supuesto que la muestra de partida ha sido seleccionada mediante un muestreo aleatorio simple. Con los siguientes contrastes comprobaremos si este supuesto es cierto.

Concepto de racha: Sea una sucesión en la que intervienen dos tipos de símbolos entonces definimos una racha como una sucesión de uno o más símbolos idénticos, que están precedidos o seguidos por un símbolo diferente o por ninguno, siendo la longitud de una racha el número de símbolos iguales que incluye.

9.3.1 Contraste de rachas de Wald-Wolfowitz:

Supongamos una población con función de distribución desconocida, y sea X una variable aleatoria asociada, que únicamente puede tomar dos valores posibles. Entonces si se obtiene una muestra aleatoria simple se pueden plantear las siguientes hipótesis.

H_0: la muestra es aleatoria.
H_1: la muestra es no aleatoria.

Y se procede como sigue;

1) Se calcula el valor del estadístico a utilizar que será, el número total de rachas (R). Bajo la hipótesis nula, R está distribuida como sigue:

$$P(R=r)\begin{cases} \dfrac{2\binom{n_1-1}{\frac{r}{2}-1}\binom{n_2-1}{\frac{r}{2}-1}}{\binom{n_1+n_2}{n_1}} & \text{r par} \\[2ex] \dfrac{2\binom{n_1-1}{\frac{r}{2}-\frac{3}{2}}\binom{n_2-1}{\frac{r}{2}-\frac{1}{2}}+\binom{n_1-1}{\frac{r}{2}-\frac{1}{2}}\binom{n_2-1}{\frac{r}{2}-\frac{3}{2}}}{\binom{n_1+n_2}{n_1}} & \text{r impar} \end{cases}$$

En la práctica estos valores se encuentran tabulados para todas las combinaciones $n_1<11$ y $n_2<11$.

Una vez que se conoce la distribución del estadístico se calcula la región crítica, que vendrá dada por los valores $k_{\alpha/2}$ y $k'_{\alpha/2}$ que cumplan:

$$P(R\leq k'_{\alpha/2})\leq\frac{\alpha}{2}$$

$$P(R\geq k_{\alpha/2})\leq\frac{\alpha}{2}$$

<u>Aproximación en muestras grandes:</u>

Para muestras grandes, la distribución R tiende a una normal a medida que n_1 y n_2 se van haciendo más grandes. Esta aproximación es bastante buena si $n_1> 10$ y $n_2>10$; de tal manera que:

$$R\rightarrow N\left(E\left[R\right],\sqrt{Var\left[R\right]}\right)$$

siendo:

$$E\left[R\right]=\frac{2n_1n_2}{n_1+n_2}+1 \qquad\qquad Var\left[R\right]=\frac{2n_1n_2\left(2n_1n_2-n_1-n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)}$$

Con estos valores, podemos tipificar la variable aleatoria R,

$$Z=\frac{R-E[R]}{V[R]}=\frac{R-\dfrac{2n_1n_2}{n_1+n_2}+1}{\dfrac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

Para una muestra concreta:

$$Z_{exp} = \frac{\hat{R} - \dfrac{2n_1n_2}{n_1+n_2} + 1}{\sqrt{\dfrac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}}$$

Que se distribuye según una N(0,1) y donde $\hat{R}$ es el número total de rachas observadas en la muestra. La región de aceptación para la hipótesis nula será entonces:

$$-z_{\alpha/2} < z_{\exp} < z_{\alpha/2}$$

Y El valor $z_{\alpha/2}$ se obtiene en la tabla de la N (0,1), de manera que

$$P(Z_1 \leq -z_{\alpha/2}) = P(Z_1 \geq z_{\alpha/2}) = \frac{\alpha}{2}$$

Este test se puede utilizar también en las siguientes situaciones:

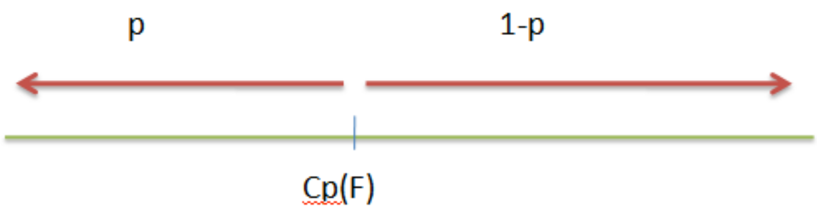
1) Cuando los datos son cuantitativos, se pueden obtener las rachas asignando un signo + o – según sean mayor o igual que la mediana. Aquellos valores que coinciden con la mediana se anulan.
2) En las series temporales, este test permite detectar la aleatoriedad de medidas cuantitativas a lo largo del tiempo detectando la no aleatoriedad en la serie, como consecuencia de la existencia de tendencias o variaciones estacionales.
3) En la comparación de dos poblaciones como se verá posteriormente.

9.4 Constrastes de localización.

Con la ayuda de estos test, podremos contrastar el valor de alguna medida de posición o localización de la distribución que sigue la población considerada, de tal manera que nos ayude a identificar estadísticamente la distribución.

Para realizar el test vamos a partir de una muestra aleatoria simple de tamaño n; $(X_1 \ldots X_n)$, procedente de una población X con función de distribución continua F(X) pero desconocida.

Nuestra hipótesis es que, para un determinado nivel de confianza, el cuantil de orden p, Cp(F), toma un determinado valor k_o , es decir el 100p% de las observaciones muestrales son inferiores a k_0 y el 100(1-p) son superiores a k_0.



Por ejemplo, para el caso particular de la mediana, se contrastaría, para un determinado nivel de confianza, la hipótesis nula de que la mediana de la variable aleatoria X de la población toma un valor concreto m. Es decir, que aproximadamente la mitad de los elementos de la población están por encima de ese valor m y por lo tanto, la otra mitad están por debajo del mismo.

Los tres test que se pueden plantear son los siguientes:

I. $$\begin{array}{ll} H_0 = C_p(F) = k_0 & \Rightarrow P(X < k_0) = p \\ H_1 = C_p(F) \neq k_0 & \Rightarrow P(X < k_0) \neq p \end{array}$$

II. $$\begin{array}{ll} H_0 = C_p(F) \leq k_0 & \Rightarrow P(X < k_0) \geq p \\ H_1 = C_p(F) > k_0 & \Rightarrow P(X < k_0) > 1 - p \end{array}$$

III. $$\begin{array}{ll} H_0 = C_p(F) \geq k_0 & \Rightarrow P(X < k_0) \leq p \\ H_1 = C_p(F) < k_0 & \Rightarrow P(X < k_0) < 1 - p \end{array}$$

Estos son algunos de los contrastes de localización más importantes:

9.4.1 Contrastes de signos:

Se contemplan tres casos distintos: caso de contraste de signos bilateral, caso de contraste de signos unilateral derecha y caso contrastes unilateral izquierdo.

Veamos a modo de ejemplo el caso de contraste de signos bilateral:

1. Planteamiento hipótesis: (contraste bilateral):

$$\begin{array}{ll} H_0 = C_p(F) = k_0 & \Rightarrow P(X < k_0) = p \\ H_1 = C_p(F) \neq k_0 & \Rightarrow P(X \geq k_0) \neq 1 - p = q \\ & \quad P(X < k_0) \neq p \end{array}$$

2. Calculamos la variable $D_{i,}$ que viene definida como sigue:

$$D_i = X_i - k_0 > 0 \Rightarrow \text{si } X_i > k_0 \quad \text{y asignamos a } D_i \text{ el signo } +$$

$$D_i = X_i - k_0 < 0 \Rightarrow \text{si } X_i > k_0 \quad \text{y asignamos a } D_i \text{ el signo } -$$

Es decir se reemplaza cada valor de la muestra por un signo + o – dependiendo de si es mayor o menor que el valor k_0 previamente establecido. De esta manera no se tienen en cuenta aquellos valores tales que $X_i=k_0$

3. Calculamos el estadístico que se define como S+=”número de signos positivos que aparecen en la muestra”, es decir número de elementos mayores que el k_0 previamente establecido.

4. Como S+, se distribuye B(n,1-p), determinamos los k y k’ que identifican la región crítica

$$P(S^+ \leq k'_{\alpha/2}) = \sum_{i=0}^{k'_{\alpha/2}} \binom{n}{i} q^i p^{n-i} \leq \frac{\alpha}{2}$$
$$P(S^+ \geq k_{\alpha/2}) = \sum_{i=k_{\alpha/2}}^{n} \binom{n}{i} q^i p^{n-i} \leq \frac{\alpha}{2}$$

5. Rechazamos la hipótesis nula si el valor de $\hat{S}^+$, es decir el valor que tomaría el estadístico S+ al sustituir los datos de la muestra aleatoria seleccionada se encuentra en la región crítica.

En el caso de los demás contrastes:

Caso contraste de signos unilateral derecha

1. Planteamiento hipótesis: (contraste unilateral derecha)

$$\begin{array}{ll} H_0 = C_p(F) \leq k_0 & \Rightarrow P(X < k_0) \geq p \\ H_1 = C_p(F) > k_0 & \Rightarrow P(X \geq k_0) > 1 - p = q \end{array}$$

2. Calculamos Di y S+, el estadístico, de forma análoga a como se hacía en el caso bilateral.

4. Como S^+->B(n,1-p), se obtienen los k y k' tales que.

$$P(S^+ \geq k_\alpha) = \sum_{i=k_{\alpha/2}}^{n} \binom{n}{i} q^i p^{n-i} \leq \alpha$$

5. Se rechaza la hipótesis nula si el valor experimental $\hat{S}^+$ se encuentra en la región crítica.

Caso contraste de signos unilateral izquierda

1. Planteamiento hipótesis: (contraste unilateral derecha)

$$\begin{array}{ll} H_0 = C_p(F) \geq k_0 & \Rightarrow P(X < k_0) \leq p \\ H_1 = C_p(F) < k_0 & \Rightarrow P(X < k_0) < 1 - p \end{array}$$

2. Calculamos Di y S+, el estadístico, de forma análoga a como se hacía en el caso bilateral.

4. Como S^+->B(n,1-p), se obtienen los k y k' tales que.

$$P(S^+ \leq k'_\alpha) = \sum_{i=k_{\alpha/2}}^{n} \binom{n}{i} q^i p^{n-i} \leq \alpha$$

9.4.2 Caso contraste de signos de la mediana

Este contraste es un caso particular del anterior, en donde el parámetro de localización es la mediana (Me). El procedimiento es por tanto el mismo que el explicado anteriormente.

- Aproximación a la normal: Si el tamaño de la muestra es grande, la distribución S+, se aproximará a una distribución normal :

$$S^{+} \rightarrow B\left(n, \frac{1}{2}\right) \rightarrow N\left(\frac{n}{2}, \sqrt{\frac{n}{4}}\right)$$

Esta aproximación obtiene resultados satisfactorios si np<5, lo que ocurre para muestras de tamaño n>10.

9.4.3 Contrastes de rangos-signos de Wilcoxon para una muestra.

Este test es una modificación del test de los signos con el fin de tener en cuenta las magnitudes de las diferencias. Sólo se puede aplicar si la distribución es simétrica y continua. Se procede como sigue:

1. Calculamos las diferencias respecto a la mediana poblacional

$$D_i = X_i - m > 0 \Rightarrow \text{si } X_i > m \;\; \text{y asignamos a } D_i \text{ el signo } +$$

$$D_i = X_i - m < 0 \Rightarrow \text{si } X_i > m \;\; \text{y asignamos a } D_i \text{ el signo } -$$

   Si D_i=0 despreciamos la observación, reduciendo el tamaño muestral en tantas unidades como valores D_i=0.

2. Obtenemos los D_i. Se obtienen los valores absolutos $|D_i|$ y se le asigna un valor que corresponde con el orden que ocupa una vez se han ordenado todos los $|D_i|$ de menor a mayor.

   En el caso de que algún $|D_i|$ esté repetido en la secuencia, entonces se asigna a cada $|D_i|$ un rango correspondiente al promedio de rangos que tendría cada uno de ellos si no hubieran sido iguales.

3. Construimos los estadísticos de rangos-signos de Wilcoxon: Se construyen los estadísticos siguientes:

   T+: suma de los rangos de las D_i^{+}

   T-: suma de los rangos de las D_i^{-}

Para ello primero es necesario definir para, la variable $z_{i,}$ que tomará el valor 1 siempre que $D_i>0$ y cero en caso de que $D_i<0$

A continuación se definen los estadísticos:

$$T+ = \sum_{i=1}^{n} Z_i r(|Di|)$$
$$T- = \sum_{i=1}^{n} (1 - Z_i) r(|Di|)$$

$$(T+) - (T-) = 2\sum_{i=1}^{n} Z_i r(|Di|) - \frac{n(n+1)}{2}$$

4. Calculamos la región crítica del contraste. Puede demostrarse que los tres estadísticos anteriores están relacionados, por lo que nos bastaría utilizar uno solo de ellos para realizar el contraste.
   Supongamos que el contraste es de dos colas:

$$H_0 : Me = m$$
$$H_1 : Me \neq m$$

Entonces la región crítica viene dada por valores pequeños o grandes de T+. Es decir, se rechaza si

$$\hat{T}+ \leq k'_{\alpha/2}$$
$$\hat{T}+ \geq k'_{\alpha/2}$$

Donde $k'_{\alpha/2}$ y $k_{\alpha/2}$, son respectivamente el mayor y menor enteros tal que

$$P(T+ \leq k'_{\alpha/2} / H_0) \leq \alpha/2$$
$$P(T+ \geq k'_{\alpha/2} / H_0) \leq \alpha/2$$

<u>Caso aproximación normal.</u>

Cuando n es grande n>15, el estadístico T+ se distribuye asintóticamente según una N(0,1)

$$T+ \longrightarrow N\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$$

9.5 Contrastes de comparación de dos poblaciones

A partir de ahora supondremos que ambas poblaciones tienen la misma distribución pero desfasada en una cierta cantidad k desconocida, Diremos entonces que las distribuciones difieren en ubicaciones.

En lo que sigue utilizaremos muestras aleatorias de tamaño n_1 y n_2 para las poblaciones X e Y, con funciones de distribución F(x), G(y) respectivamente.

9.5.1 Contraste de la mediana

Este contraste permite concluir si dadas dos distribuciones F(x), G(y) de las que proceden las muestras $(X_1,\ldots X_{n1})$ y $(Y_1\ldots Y_{n2})$ respectivamente son iguales, es decir a aceptar o rechazar

H_0: F(z)=G(Z)

Aunque sus gráficas sean distintas, siempre y cuando las medianas sean iguales .

El conjunto de hipótesis que podemos plantear son:

I. $$\begin{array}{ll} H_0: F(z) \le G(z) & \Rightarrow H_0: M_{ex} \ge M_{ey} \\ H_1: F(z) > G(z) & \Rightarrow H_1: M_{ex} < M_{ey} \end{array}$$

II. $$\begin{array}{ll} H_0: F(z) \ge G(z) & \Rightarrow H_0: M_{ex} \le M_{ey} \\ H_1: F(z) < G(z) & \Rightarrow H_1: M_{ex} > M_{ey} \end{array}$$

III. $$\begin{array}{ll} H_0: F(z) = G(z) & \Rightarrow H_0: M_{ex} = M_{ey} \\ H_1: F(z) \ne G(z) & \Rightarrow H_1: M_{ex} \ne M_{ey} \end{array}$$

Para realizar el contraste I procedemos como sigue:

1) Ordenamos las dos muestras independientes de tamaños n_1 y n_2 de las poblaciones X, Y.
2) Calculamos la mediana.
3) Definimos V, la variable aleatoria que nos indicará el número de valores observados de X que son menores o iguales que la mediana de la muestra combinada de n_1+n_2 elementos.
4) Se calcula la región crítica; en el caso de que V sea muy grande quiere decir que los valores de X son pequeños por lo que su mediana debe ser menor que la de Y. Se

rechaza la H_0 si:

$$\hat{V} \geq k_\alpha$$

Siendo k_α el menor entero tal que

$$P(V \geq k_\alpha / H_0) \leq \alpha$$

En la práctica para conocer las regiones críticas se necesita conocer la distribución de la variable aleatoria V. Si suponemos que $n_1>10$ y $n_2>10$, entonces V se distribuye asintóticamente según una normal;.

$$V \rightarrow N\left(k\frac{n_1}{n}, \sqrt{k\frac{n_1}{n}\frac{n_2}{n}\frac{n-k}{n-1}}\right)$$

Y en consecuencia resulta muy fácil trabajar con la distribución del estadístico V.

$$Z = \frac{V - k\frac{n_1}{n}}{\sqrt{k\frac{n_1}{n}\frac{n_2}{n}\frac{n-k}{n-1}}} \rightarrow N(0,1)$$

Los demás contraste se hacen de forma análoga.

9.5.2 Contraste de Wilcoxon Mann-Whitney.

Este test se utiliza para contrastar si dos muestras, extraídas independientemente, proceden de la misma población. El único supuesto preciso es que la población o poblaciones de las que se han extraído las muestras, sean de tipo continuo.

Supongamos que disponemos de dos muestras de tamaño n_1, y n_2, tomadas de dos poblaciones diferentes.

Las hipótesis que queremos contrastar serán entonces:

$$H_0 : F(z) = G(z)$$
$$H_1 : G(z) \neq G(z)$$

O lo que es lo mismo:

$$\begin{aligned} H_0 &: \mu_x \geq \mu_y \\ H_1 &: \mu_x \leq \mu_y \end{aligned}$$

El primer paso consiste en considerar ambas muestras como una muestra global de tamaño n1+n2, y ordenarla de menor a mayor. Posteriormente asignamos rangos a todos los elementos, resolviendo casos de igualdad de rangos del mismo modo que en los estadísticos previos (con los promedios). Después, calculamos las sumas de rangos, Wx y Wy, de ambas muestras.

$$Wx = \sum_{x_i} r_i \equiv \text{suma rangos correspondiente a la muestra } X$$

A continuación, calculamos el estadístico Ux de Mann-Whitney como:

$$Ux = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_x$$

Y cuya función de distribución, viene tabulada.
Rechazaremos Ho tanto para valores grandes de Ux, como para valores pequeños y por tanto la región crítica vendrá dada como:

$$\begin{aligned} P(U_x \leq u'_{\alpha/2} / H_0) \leq \alpha/2 \\ P(U_x \geq u_{\alpha/2} / H_0) \leq \alpha/2 \end{aligned}$$

Por lo que rechazaremos H0 si $\hat{U}_x$ es menor o igual a $u'_{\alpha/2}$ o mayor a $u_{\alpha/2}$

Aproximación para muestras grandes:

Para muestras grandes $n_1>10$, $n_2>10$ se demuestra que el estadístico U de Mann Whitney, en donde $U=U_x$, $U=U_y$ tiene como media y varianza:

$$E[U] = \frac{1}{2} n_1 n_2$$

$$V[U] = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Y se distribuye asintóticamente según una normal y en consecuencias para muestras suficientemente grandes, puede aplicarse:

$$Z = \frac{U - \frac{1}{2} n_1 n_2}{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \longrightarrow N(0,1)$$

9.6 Contrastes de comparación de más de una población

Estos contrastes no son más que una generalización de los test anteriores, pero a k-poblaciones.

9.6.1 Contraste de Kruskal-Wallis

Este contraste se considera una extensión del contraste Wilcoxon- Mann- Whitney. Trabajamos con más de 2 muestras independientes y se pretende contrastar la hipótesis nula de que todas ellas proceden de la misma población. El procedimiento es idéntico al seguido con el estadístico U,

1. Considerar las k muestras como una muestra conjunta y ordenar la muestra de menor a mayor.

2. Asignar un rango a cada una de ellas, si hay empate asignar el promedio, como ya se ha descrito previamente.

3. Calcular la suma de los rangos para cada muestra:

$$R_i = \sum_{j=1}^{n_i} r_{ij}$$

4. Cálculo de H, el estadístico para el constraste

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

Que para muestras grandes se distribuye según una chi-cuadrado con k-1 grados de libertad χ^2_{k-1}

5. Comparar el valor observado, con el valor real, de las tablas de tal forma que rechazaremos la hipótesis nula H_0 si le valor observado es mayor, es decir si se cumple:

$$\hat{H} \geq h_\alpha$$

Ha de tenerse en cuenta que cuando los tamaños de muestras son menores de 6, la aproximación de H como una chi-cuadrado con k-1 grados de libertad χ^2_{k-1}, es errónea y deben utilizarse otras aproximaciones.

9.7 Contrastes de Bondad de Ajuste

Estos contrastes nos servirán para determinar si la muestra obtenida se ajusta a un determinado modelo o a una determinada distribución de probabilidad.

Estudiaremos cuatro de ellos.

9.7.1 Contraste X2 Pearson

Consiste en comparar las frecuencias observadas en la muestra, con las que debería haberse obtenido en una población que perteneciese a una distribución de probabilidad específica.

Consideraremos dos casos:

Caso A: En este caso todos los parámetros de la distribución de la población son conocidos. Se procede como sigue:

Dividimos el campo de variación de la variable aleatoria X en k clases excluyentes y se extrae una muestra aleatoria simple de tamaño n de la población. Las observaciones, podrán entonces incluirse en cada una de las distintas clases excluyentes. La frecuencia de cada una de las clases se definirá como n_i y se denominará frecuencia absoluta observada. La hipótesis nula y hipótesis alternativa se definen entonces como:

H_0: La muestra aleatoria procede de una población con función de distribución $F_0(x)$
H_1: La muestra aleatoria no procede de una población con función de distribución $F_0(x)$

Bajo la hipótesis nula H_0, la frecuencia esperada se define como:

$E_i=E[n_i]= np_i \; i=1\ldots K$

El estadístico se define como:

$$\chi^2_{\exp} = \sum_{i=1}^{k} \frac{(n_1 - np_i)^2}{np_i}$$

Que se distribuye como una chi-cuadrado con k-1 grados de libertad χ^2_{k-1}

La regla de decisión, será rechazar la hipótesis nula si:

$$\chi^2_{\exp} = \sum_{i=1}^{k} \frac{(n_1 - np_i)^2}{np_i} - > \chi^2_{1-\alpha}$$

$$P[\chi^2_{k-1} > \chi^2_{1-\alpha} / H_0] = \alpha$$

$$donde \;\; \chi^2_{k-1} \text{sigue una distribución con k} - 1 \text{ grados de libertad}$$

Caso B: Ahora conocemos la población, pero no conocemos algún parámetro de la misma .El contraste es similar al anterior:

H_0: La muestra aleatoria procede de una población con función de distribución $F_0(x;\theta_1\ldots \theta_h)$
H_1: La muestra aleatoria no procede de una población con función de distribución $F_0(x;\theta_1\ldots \theta_h)$

Tal que $\theta_1\ldots \theta_h$ son parámetros desconocidos.
El estadístico y su distribución son:

$$\chi^2_{\exp} = \sum_{i=1}^{k} \frac{(n_1 - np_i(\hat{\theta}_1..\hat{\theta}_h))^2}{np_i(\hat{\theta}_1..\hat{\theta}_h)}$$

que sigue una distribución χ^2 con k - h - 1 grados de libertad siendo h es el número de parámetros desconocido.

La regla de decisión por tanto, será rechazar la hipótesis nula H_0 si el valor experimental es mayor que el valor teórico, es decir si

$$\chi^2_{\exp} = \sum_{i=1}^{k} \frac{(n_1 - np_i(\hat{\theta}_1..\hat{\theta}_h))^2}{np_i(\hat{\theta}_1..\hat{\theta}_h)} > \chi^2_{1-\alpha}$$

Donde $\chi^2_{1-\alpha}$ se define como :

$$P[\chi^2_{k-h-1} > \chi^2_{1-\alpha} / H_0] = \alpha$$

donde χ^2_{k-h-1} sigue una distribución con $k-1$ grados de libertad

Para que la aproximación χ2 de la distribución del estadístico del contraste sea válida deben cumplirse una serie de hipótesis:

- El tamaño muestral debe ser suficientemente grande (n > 30).
- En caso de que haya que estimar parámetros, los parámetros deben estimarse por el procedimiento de máxima verosimilitud.
- Las frecuencias esperadas $e_i = np_i$ deberían ser todas > 5, en caso contrario se deben reagrupar las clases de la variable aleatoria poblacional.

9.7.2 Contraste Kolmogorov-Smirnov:

Utilizar el anterior test cuando se tiene una muestra pequeña, puede suponer que obtengamos resultados erróneos. Como alternativa se presenta el test que sigue a continuación.

Las hipótesis de este test para el caso bilateral son las que siguen;

H_0: La muestra aleatoria procede de una población con función de distribución $F_0(x;\theta_1\dots\theta_h)$
H_1: La muestra aleatoria no procede de una población con función de distribución $F_0(x;\theta_1\dots\theta_h)$

Es decir,

$$H_0 : F(x) = F_o(x)$$
$$H_1 : F(x) \neq F_o(x)$$

Para realizar este contraste necesitamos calcular la función de distribución empírica, que se usará como estimador de la función de distribución F(x) de la población. A continuación se calcula el estadístico del contraste como la diferencia máxima entre ambas funciones de distribución.

$$D_n = \max_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

La región crítica del test viene dado como

$$P[D_n > D_\alpha] = \alpha$$

El valor correspondiente a D_α, viene dado en las tablas y se rechazará H_0 si el valor observado de $D_{n,\,exp}$ es mayor que D_α

Para los contrastes unilaterales:

	Contraste Unilateral 1	Contraste Unilateral 2
Hipótesis	$H_0 : F(x) \geq F_o(x)$ $H_1 : F(x) < F_o(x)$	$H_0 : F(x) \leq F_o(x)$ $H_1 : F(x) > F_o(x)$
Estadístico	$D_n^+ = \max_{-\infty<x<\infty} \|F_o(x) - F_n(x)\|$	$D_n^- = \max_{-\infty<x<\infty} \|F_n(x) - F_0(x)\|$
Región crítica	$P[D_n^+ > D_\alpha / H_0] = \alpha$	$P[D_n^- > D_\alpha / H_0] = \alpha$
Regla de decisión	(Rechazamos) $D^+_{n,exp} > D_\alpha$	(Rechazamos) $D^-_{n,exp} > D_\alpha$

El principal inconveniente de este test es que requiere que la población de partida sea continua, a diferencia del test χ2 de bondad de ajuste .

Las ventajas de este test son:

i. No hay pérdida de información por agrupamiento, se utilizan directamente los datos observados
ii. Es válido para muestras pequeñas (para muestras intermedias es más potente)
iii. Permite calcular un intervalo de confianza p.
iv. Es más o igual de potente que el test χ2

9.7.3 Contraste normalidad Lilliefors

Este test es una modificación del contraste de Kolmogorov-Smirnov. Sirve para contrastar la normalidad aún no conociendo todos los parámetros. Las hipótesis de las que se parte, son las que siguen:

H_0: La muestra procede de una población normal, con media y varianza desconocida.
H_1: La muestra no procede de una población normal.

El estadístico a utilizar, no es más que el estadístico de Kolmogorov-Smirnov pero construido sobre los valores normalizados de las observaciones iniciales, es decir la función de distribución empírica se obtiene a partir de la muestra normalizada.

$$D`_n = \max_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

La región crítica vendrá entonces dada por:

$$P[D`_n > D`_\alpha / H_0] = \alpha$$

Los valores $D`_\alpha$ están tabulados, por tanto bastará comparar el valor experimental obtenido con el tabulado y si el primero es mayor, se rechazará la H0, en caso contrario se aceptará la H1.

9.7.4 Contraste Kolmogorov-Smirnov para dos muestras.

Este test sirve para contrastar si dadas dos muestras de dos poblaciones con funciones de distribución asociada, proceden de la misma población.
Las hipótesis en el caso bilateral son por tanto:

$$H_0 : F(x) = G(x)$$
$$H_1 : F(x) \neq G(x)$$

El estadístico se define con ayuda de las dos funciones de distribución empíricas de la muestras $F_{n1}(x)$ y $G_{n2}(x)$,

$$D_{n1,n2} = \max_{-\infty<x<\infty} \left| F_{n1}(x) - G_{n2}(x) \right|$$

La región crítica viene dada entonces por $D_{n1,n2;\alpha}$

$$P[D_{n1,n2} > D_{n1,n2;\alpha} / H_0] = \alpha$$

De igual forma a como hemos realizado previamente, rechazaremos la hipótesis nula si el valor experimental es mayor que el valor $D_{n1,n2;\alpha}$

Para el caso unilateral:

	Contraste Unilateral 1	Contraste Unilateral 1
Hipótesis	$H_0 : F(x) \leq G(x)$ $H_1 : F(x) > G(x)$	$H_0 : F(x) \geq G(x)$ $H_1 : F(x) < G(x)$
Estadístico	$D^+_{n1,n2} = \max_x \left\| F_{n1}(x) - G_{n2}(x) \right\|$	$D^-_n = \max_x \left\| G_{n2}(x) - F_{n1}(x) \right\|$
Región crítica	$P[D^+_{n1,n2} > D_\alpha / H_0] = \alpha$	$P[D^-_{n1,n2} > D_\alpha / H_0] = \alpha$
Regla de decisión	(Rechazamos) $D^+_{n1,n2} > D_\alpha$	(Rechazamos) $D^-_{n1,n2} > D_\alpha$

9.8 Resumen y preguntas frecuentes.

- ✓ Diferencia entre un contraste paramétrico y uno no paramétrico
- ✓ Características de la estimación utilizando los contrastes o test de hipótesis no paramétricos.
- ✓ Tipos contrastes no paramétricos.
- ✓ Diferencia entre los distintos contrastes no paramétricos
- ✓ Utilización de los diferentes contrastes