UNIVERSIDAD
DE MÁLAGA

Verificación de hipótesis no paramétricas

Mª Isabel Aguilar, Eugenia Cruces y Bárbara Díaz

UNIVERSIDAD DE MÁLAGA
Departamento de Economía Aplicada (Estadística y Econometría)

Parcialmente financiado a través del PIE13-024 (UMA)

Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- Contrastes no paramétricos de localización para una población
- Contrastes no paramétricos de localización para dos poblaciones
- Test de rachas (aleatoriedad)
- Contrastes generales de bondad de ajuste
- Contrastes específicos de normalidad
- Contraste de independencia de dos variables aleatorias
- Contraste de homogeneidad

Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- Contrastes no paramétricos de localización para una población
- Contrastes no paramétricos de localización para dos poblaciones
- Test de rachas (aleatoriedad)
- Contrastes generales de bondad de ajuste
- Contrastes específicos de normalidad
- Contraste de independencia de dos variables aleatorias
- Contraste de homogeneidad

Introducción

UNIVERSIDAD DE MÁLAGA

Hipótesis paramétrica: Suposición o conjetura sobre algún parámetro desconocido de la distribución de una variable aleatoria. Se conoce la forma de f(x)

Hipótesis no paramétrica: Suposición o conjetura sobre cualquier otra característica distinta al valor del parámetro:

- Valor de la mediana (localización)
- Comparación de las medianas de varias poblaciones
- Distribución de la población (bondad de ajuste)
- Aleatoriedad de la muestra
- Independencia de dos variables
- Homogeneidad de varias poblaciones

Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- **Contrastes no paramétricos de localización para una población**
- Contrastes no paramétricos de localización para dos poblaciones
- Test de rachas (aleatoriedad)
- Contrastes generales de bondad de ajuste
- Contrastes específicos de normalidad
- Contraste de independencia de dos variables aleatorias
- Contraste de homogeneidad

Contrastes no paramétricos de localización

- **Objetivo**: verificar el valor de alguna medida de posición o localización de una población de distribución desconocida
- **Alternativa a los contrastes sobre la media** en poblaciones normales:
  - La distribución del estadístico de prueba se obtiene con independencia de la que siga la población de partida (también se denominan **contrastes de libre distribución**)
  - Útiles si no se cumple la hipótesis de normalidad, de la que se partía en los contrastes paramétricos sobre la media
  - **Se basan en la mediana**, permitiendo localizar distribuciones fuertemente asimétricas o con presencia de valores atípicos (*outliers)*
  - **Algunos** de ellos permiten contrastar hipótesis acerca de variables cualitativas con **escala ordinal**

Principal inconveniente: son menos potentes que los test paramétricos sobre la media

Si se cumple la hipótesis de normalidad, siempre elegiremos un test paramétrico para localizar una población frente a la alternativa no paramétrica

Localización de una población (mediana)

UNIVERSIDAD DE MÁLAGA

Localización de una población

(contrastes sobre la mediana)

- Contraste del signo
- Contraste del rango signado (Wilcoxon)

Contraste del signo (Fisher)

- **Objetivo:** localizar una población de distribución continua y desconocida

- Se basa en la **mediana** o valor central de la distribución, tal que:

$$P(X \leq Me) = P(X \geq Me) = 1/2$$

- **Hipótesis** a verificar:

$$\boxed{\begin{array}{l} H_0: Me = Me_0 \\ H_1: Me > Me_0 \end{array}} \qquad \boxed{\begin{array}{l} H_0: Me = Me_0 \\ H_1: Me < Me_0 \end{array}} \qquad \boxed{\begin{array}{l} H_0: Me = Me_0 \\ H_1: Me \neq Me_0 \end{array}}$$

Contraste del signo (Fisher)

Definición del estadístico de prueba:

Se asigna un signo positivo a las observaciones muestrales mayores al valor especificado para la mediana en H_0 (y un signo negativo a las inferiores a este valor):

$$\begin{cases} x_i > Me_0 \Rightarrow \text{signo } + \\ x_i < Me_0 \Rightarrow \text{signo } - \end{cases}$$

Si alguna observación coincide con Me_0, se elimina y se ajusta el tamaño de la muestra

$$\boxed{\begin{array}{c} S^{+}\text{: número de signos + en la muestra} \\ \\ S^{+} \sim B(n, p=1/2) \end{array}}$$

UNIVERSIDAD DE MÁLAGA

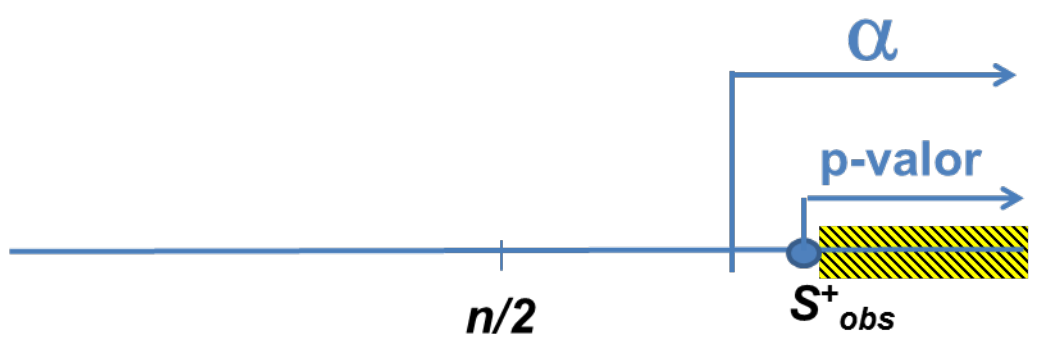
Región crítica del test:

Si H_0 es cierta ($Me = Me_0$), S^+ debería tomar valores próximos a $n/2$. Discrepancias grandes con respecto a ese valor, llevarían a rechazar dicha hipótesis nula, de manera que:

- Si $H_1 : Me > Me_0$ → Región Crítica a la derecha:

  Se calcula **p-valor=$P(S^+ \geq S^+_{obs})$**

  Si **p-valor < α** → rechazar H_0

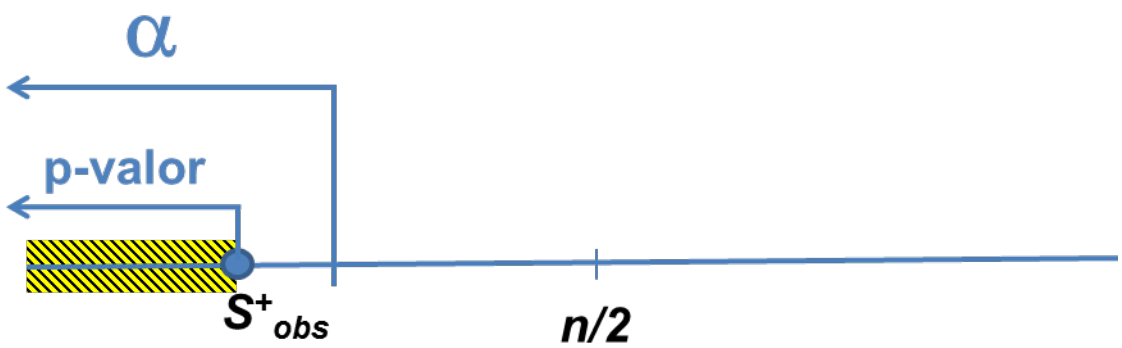


Contraste del signo (Fisher)

- Si $H_1 : Me < Me_0$ → Región Crítica a la izquierda:

  Se calcula $p\text{-valor}=P(S^+ \leq S^+_{obs})$

  Si $p\text{-valor} < \alpha$ → rechazar H_0



Contraste del signo (Fisher)

- Si $H_1 : Me \neq Me_0$ → Región Crítica bilateral

Si $S^+_{obs} > n/2$ → Se calcula el p-valor multiplicando por 2 la probabilidad que el valor observado deja a la derecha (menor cola).

$$p\text{-valor} = 2 \times P\left(S^+ \geq S^+_{obs}\right)$$

Si $S^+_{obs} < n/2$ → Se calcula el p-valor multiplicando por 2 la probabilidad que el valor observado deja a la izquierda (menor cola).

$$p\text{-valor} = 2 \times P\left(S^+ \leq S^+_{obs}\right)$$

$p\text{-valor} < \alpha$ → rechazar H_0

UNIVERSIDAD DE MÁLAGA

Aproximación a la normal:

$$S^{+} \sim B\left(n, p = \frac{1}{2}\right)$$

Al ser *p=1/2,* se puede aproximar la Binomial a la Normal a partir de 10 observaciones:

$$n \geq 10 \quad \Rightarrow \quad S^{+} \rightarrow N\left(\frac{n}{2}, \sqrt{\frac{n}{4}}\right)$$

Tipificando y aplicando la corrección de continuidad:

$$\mathrm{Z} = \frac{S^{+} \pm \frac{1}{2} - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \longrightarrow \mathrm{N}(0,1)$$

Contraste del rango signado (Wilcoxon)

- **Objetivo:** localizar una población con distribución desconocida (continua y simétrica)

- **Extensión del contraste del signo:** incorpora la cuantía de las diferencias entre cada valor muestral y la mediana especificada en H_0

- Contraste **más potente que el del signo,** pero sólo se puede aplicar a **variables cuantitativas**

- **Hipótesis** a verificar sobre la mediana:

$$\boxed{\begin{array}{l} H_0: Me = Me_0 \\ H_1: Me > Me_0 \end{array}} \quad \boxed{\begin{array}{l} H_0: Me = Me_0 \\ H_1: Me < Me_0 \end{array}} \quad \boxed{\begin{array}{l} H_0: Me = Me_0 \\ H_1: Me \neq Me_0 \end{array}}$$

Contraste del rango signado (Wilcoxon)

UNIVERSIDAD DE MÁLAGA

- **Procedimiento para obtener el estadístico de prueba:**
  - Calcular la diferencia entre cada valor de la muestra y la mediana propuesta en H_0, $d_i = x_i - Me_0$ $(i=1, 2, ..., n)$
  - Ordenar $|d_i|$ y asignarles un rango
    - Si hay un grupo de diferencias iguales, se les asigna la media de sus rangos
    - Si $d_i = 0$, se omite y se ajusta el tamaño de la muestra
  - Incorporar a cada rango el signo original de las diferencias (**rango signado**)
  - Obtener las sumas de los rangos con signos positivos (T^+) y con signos negativos (T^-) y elegir la menor de las dos

Contraste del rango signado (Wilcoxon)

- **Estadístico de prueba:**

T_{menor} (la menor de las dos sumas (T^{+} y T^{-})

- **Definición de la Región Crítica:**

Si H_0 es cierta ($Me = Me_0$), las sumas de los rangos positivos y negativos deberían estar equilibradas, aproximándose al valor esperado (la mitad de la suma del total de los rangos, es decir, $n(n+1)/4$)

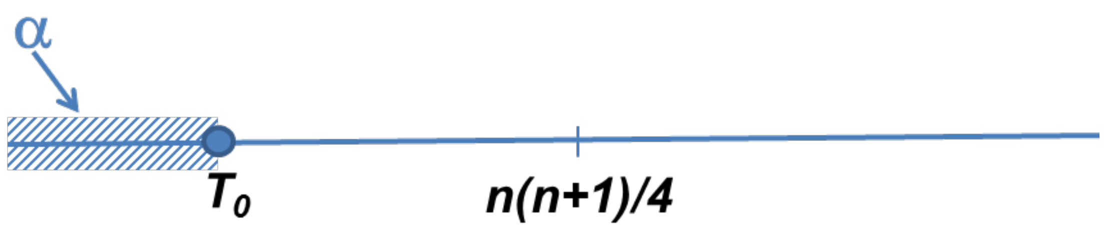
Por el contrario, si una de las sumas es muy pequeña en relación a la otra, habrá evidencia de que la mediana no coincide con la propuesta bajo H_0

Contraste del rango signado (Wilcoxon)

- Bajo este razonamiento, la **Región crítica** viene dada por la **cola izquierda** de la distribución del estadístico de prueba:

$$\boxed{C:\ T_{obs} \leq T_0} \longrightarrow$$

Valor crítico (se busca en la tabla elaborada por Wilcoxon en función n y α)



- Si $n \geq 15$ se puede aproximar T^+ a la Normal:

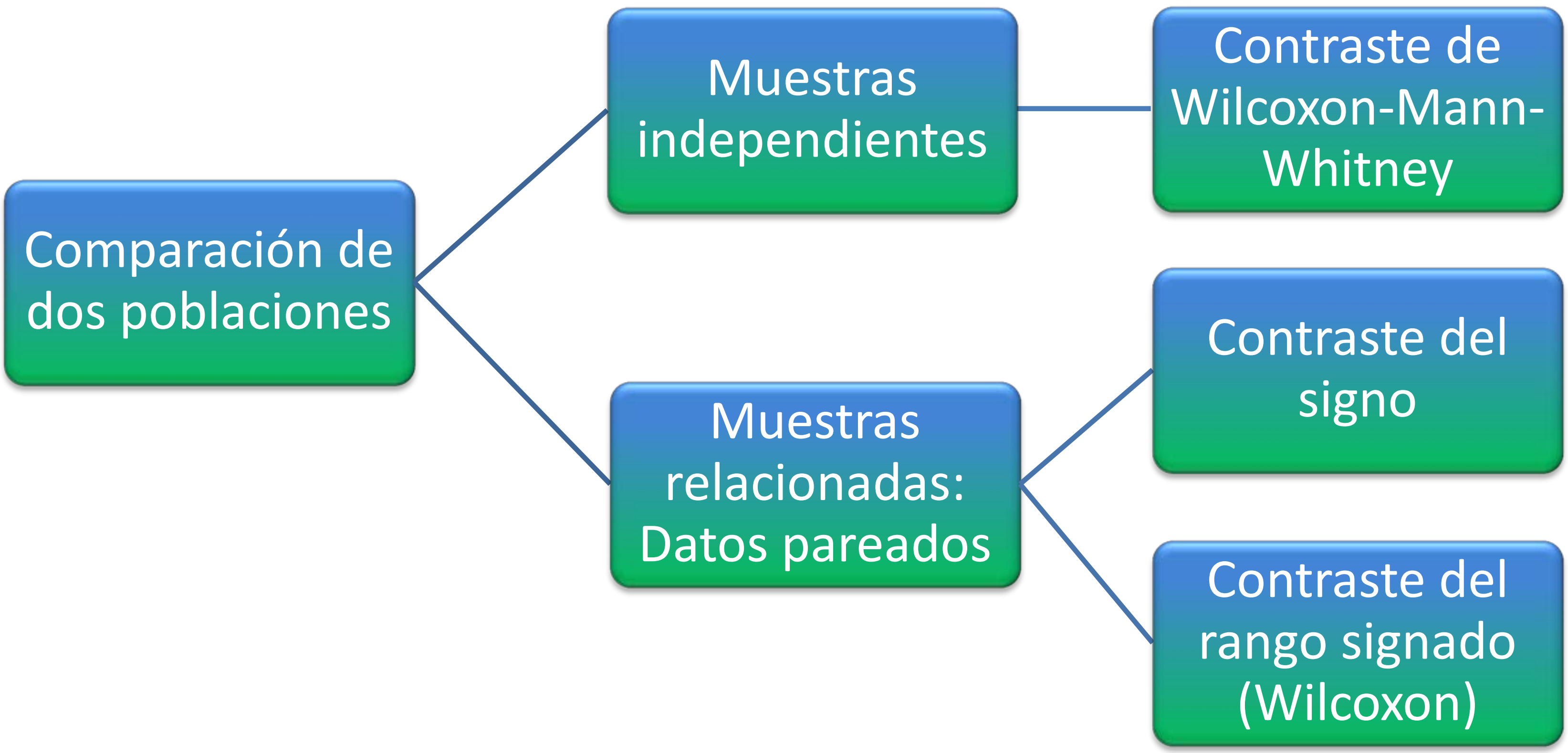
$$Z = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \rightarrow N(0,1)$$

Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- Contrastes no paramétricos de localización para una población
- **Contrastes no paramétricos de localización para dos poblaciones**
- Test de rachas (aleatoriedad)
- Contrastes generales de bondad de ajuste
- Contrastes específicos de normalidad
- Contraste de independencia de dos variables aleatorias
- Contraste de homogeneidad

Contrastes no paramétricos de localización de dos poblaciones



Muestras independientes: Contraste de Wilcoxon-Mann-Whitney

- **Objetivo:** comparar la localización de dos distribuciones desconocidas que se suponen continuas y con la misma forma
- Equivalente no paramétrico al test de comparación de medias en poblaciones normales
- Muestras **independientes**
- Aplicable a datos ordinales y a cuantitativos (se basa en rangos)

Muestras independientes: Contraste de Wilcoxon-Mann-Whitney

- **Hipótesis a verificar:**

$H_0: Me_1 = Me_2$	$H_0: Me_1 = Me_2$	$H_0: Me_1 = Me_2$
$H_1: Me_1 > Me_2$	$H_1: Me_1 < Me_2$	$H_1: Me_1 \neq Me_2$

Las hipótesis se plantean tomando como primer parámetro de referencia (Me_1) el de la muestra más pequeña ($n_1 < n_2$). Si $n_1 = n_2$ podrá usarse cualquiera de los dos

Si H_0 es cierta, las dos poblaciones serían iguales, por lo que ambas muestras pueden considerarse como una única muestra de tamaño $n = n_1 + n_2$

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

Muestras independientes: Contraste de Wilcoxon-Mann-Whitney

UNIVERSIDAD DE MÁLAGA

Obtención del estadístico de prueba y de la Región Crítica:

- Se ordenan las observaciones de la muestra conjunta y se les asigna un rango (desde 1 hasta n)

  (Nota: si hay un grupo de valores iguales, se les asigna la media de los rangos correspondientes)

- Se definen las sumas de rangos de cada muestra:

  T_1: suma de rangos en la muestra de menor tamaño ($n_1 \leq n_2$)

  T_2: suma de rangos en la otra muestra

  La suma total de rangos será:

$$T_1 + T_2 = n(n+1)/2, \quad con \;\; n = n_1 + n_2$$

Muestras independientes: Contraste de Wilcoxon-Mann-Whitney

- A partir de T_1, Wilcoxon desarrolló un test para contrastar la igualdad de medianas de dos poblaciones independientes. Posteriormente, Mann y Whitney propusieron otro contraste y demostraron que era equivalente al de Wilcoxon
- El test de Mann y Whitney ha prevalecido en la práctica (disponible en la mayoría de los paquetes informáticos estadísticos), de manera que sólo desarrollaremos esta versión basada en los siguientes estadísticos:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1 \quad \Rightarrow \quad 0 \leq \mathrm{U}_1 \leq n_1 n_2$$

$$U_1 + U_2 = n_1 \cdot n_2$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2 \quad \Rightarrow \quad 0 \leq \mathrm{U}_2 \leq n_1 n_2$$

Muestras independientes: Contraste de Wilcoxon-Mann-Whitney

UNIVERSIDAD DE MÁLAGA

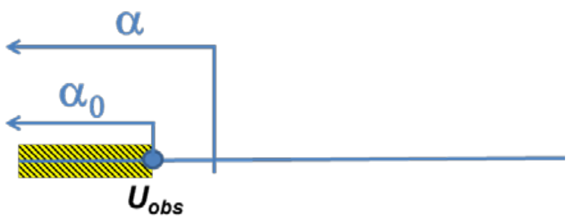
- La aplicación del contraste se basa en el estadístico de menor valor (***U_{menor}***)
- Para cualquier tipo de hipótesis alternativa (bilateral o unilateral), la región crítica viene definida por la cola izquierda de la distribución de este estadístico
- A partir de las tablas elaboradas por Mann y Whitney calculamos la probabilidad acumulada hasta el valor observado:

$$\alpha_0 = P(U \leq U_{obs.})$$

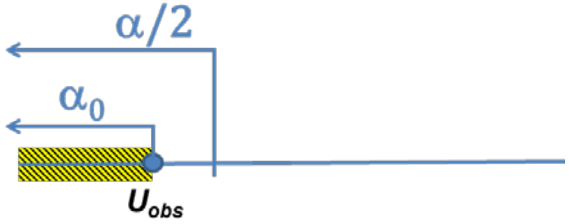
Muestras independientes: Contraste de Wilcoxon-Mann-Whitney

Regiones Críticas

- ***H_1 unilateral*** *(izquierda o derecha): se rechaza H_0 si* $\alpha_0 \leq \alpha$



- ***H_1 bilateral:*** *se rechaza H_0 si* $\alpha_0 \leq \alpha/2$



ORDENADOR: $p\text{-valor} \leq \alpha$ **se rechaza H_0**

Test unilateral $p\text{-valor} = \alpha_0$ *Test bilateral* $p\text{-valor} = 2\alpha_0$

Datos pareados: Contraste del signo

- **Objetivo:** comparar la localización de dos distribuciones desconocidas que se suponen continuas y con la misma forma

- **Datos pareados**: $(\boldsymbol{X_1}, \boldsymbol{Y_1}), (\boldsymbol{X_2}, \boldsymbol{Y_2}), \ldots, (\boldsymbol{X_n}, \boldsymbol{Y_n})$

- **Hipótesis** a verificar (comparación de medianas):

$$\boxed{\begin{array}{l} H_0: Me_1 = Me_2 \\ H_1: Me_1 > Me_2 \end{array}} \quad \boxed{\begin{array}{l} H_0: Me_1 = Me_2 \\ H_1: Me_1 < Me_2 \end{array}} \quad \boxed{\begin{array}{l} H_0: Me_1 = Me_2 \\ H_1: Me_1 \neq Me_2 \end{array}}$$

- Contraste: **extensión del test del signo** para una población

UNIVERSIDAD DE MÁLAGA

Estadístico de prueba:

Se asigna un signo positivo a las observaciones muestrales de X mayores a las de Y (y un signo negativo en caso contrario):

$$\begin{cases} x_i > y_i \Rightarrow \textit{signo } + \\ x_i < y_i \Rightarrow \textit{signo } - \end{cases}$$

Si $x_i = y_i$ se elimina la pareja correspondiente y se ajusta el tamaño de la muestra

$\boxed{S^+\text{: número de signos } + \qquad S^+ \sim B(n, p=1/2)}$

El resto del contraste (regiones críticas y aproximación a la normal) es idéntico al ya descrito para la localización de una población

Datos pareados: Contraste del rango signado

- **Objetivo:** comparar la localización de dos distribuciones desconocidas que se suponen continuas y con la misma forma

- **Datos pareados**: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

- **Hipótesis** a verificar (comparación de medianas):

$$\boxed{\begin{array}{l} H_0: Me_1 = Me_2 \\ H_1: Me_1 > Me_2 \end{array}} \qquad \boxed{\begin{array}{l} H_0: Me_1 = Me_2 \\ H_1: Me_1 < Me_2 \end{array}} \qquad \boxed{\begin{array}{l} H_0: Me_1 = Me_2 \\ H_1: Me_1 \neq Me_2 \end{array}}$$

- **Extensión del test del rango signado** para una población

Datos pareados: Contraste del rango signado

- Tener en cuenta que el test de rango signado es **más potente** que el test del signo, pero que sólo puede aplicarse a **variables cuantitativas**

- **Estadístico de prueba:** T_{menor} (entre T^{+} y T^{-})

Similar al del test de una población, pero ahora las diferencias d_i se obtienen comparando los valores de cada pareja:

$$d_i = x_i - y_i$$

El resto del contraste (regiones críticas y aproximación a la normal) es idéntico al ya descrito para la localización de una población

Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- Contrastes no paramétricos de localización para una población
- Contrastes no paramétricos de localización para dos poblaciones
- **Test de rachas (aleatoriedad)**
- Contrastes generales de bondad de ajuste
- Contrastes específicos de normalidad
- Contraste de independencia de dos variables aleatorias
- Contraste de homogeneidad

Test de rachas

- **Objetivo:** Verificar la aleatoriedad de la muestra (hipótesis básica de la inferencia estadística)

- **Hipótesis a verificar:**

H_0: la muestra es aleatoria

H_1: la muestra no es aleatoria

- **Estadístico de prueba:** Número de **rachas** (r)

Una **racha** es una sucesión de valores situados por encima (racha positiva) o por debajo de la mediana de la muestra (racha negativa). El número de elementos de una racha se llama longitud

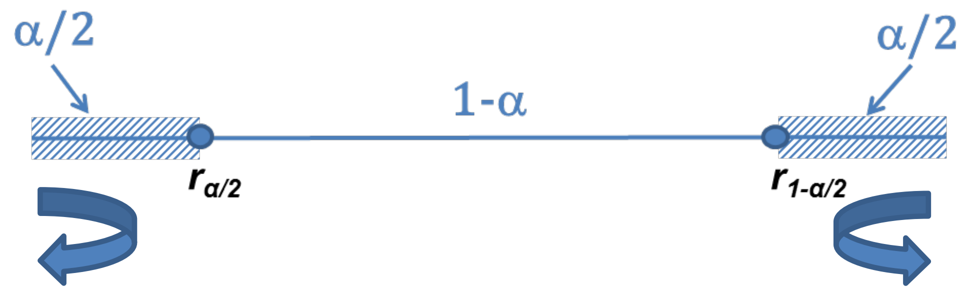
Test de rachas

Cálculo del número de rachas:

- Se ordenan los elementos de la muestra de menor a mayor y se calcula la mediana
- En la muestra original (sin ordenar) se asigna un signo + a los valores mayores que la mediana y un signo – a los inferiores (los valores iguales a la mediana no se tienen en cuenta, ajustándose el tamaño de la muestra)
- Se computa el número de rachas en la muestra (r_{obs}) y el número de signos positivos (k):

$$\oplus\ \oplus\ \oplus\ \oplus \quad \ominus\ \ominus \quad \oplus \quad \ominus\ \ominus\ \ominus \qquad \left\{ \begin{array}{l} n=10 \\ r_{obs}=4 \\ k=5 \end{array} \right.$$

- **Región crítica bilateral (dos colas):**



r_{obs} pequeño (indicativo de relación positiva entre valores muestrales; a un valor bajo extraído le sigue otro valor bajo y viceversa)

r_{obs} grande (indicativo de relación negativa entre valores muestrales; a un valor bajo extraído le sigue otro valor alto y al contrario)

$$\boldsymbol{r_{obs} \leq r_{\alpha/2} \quad \text{ó} \quad r_{obs} \geq r_{1-\alpha/2}}$$

→ Valores críticos (tabulados en función de n, α y k)

Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- Contrastes no paramétricos de localización para una población
- Contrastes no paramétricos de localización para dos poblaciones
- Test de rachas (aleatoriedad)
- **Contrastes generales de bondad de ajuste**
- Contrastes específicos de normalidad
- Contraste de independencia de dos variables aleatorias
- Contraste de homogeneidad

Contrastes de bondad de ajuste

UNIVERSIDAD DE MÁLAGA

- Contrastes de bondad de ajuste
  - Contrastes generales
    - Contraste χ^2 de Pearson
    - Contraste de Kolmogorov-Smirnov
  - Contrastes específicos de normalidad
    - Contraste de Kolmogorov-Smirnov-Lilliefors
    - Contraste de Jarque-Bera
    - Contraste de Shapiro-Wilk

Contrastes de bondad de ajuste

UNIVERSIDAD DE MÁLAGA

- **Contrastes de bondad de ajuste**
  - **Contrastes generales**
    - **Contraste χ^2 de Pearson**
    - **Contraste de Kolmogorov-Smirnov**
  - **Contrastes específicos de normalidad**
    - **Contraste de Kolmogorov-Smirnov-Lilliefors**
    - **Contraste de Jarque-Bera**
    - **Contraste de Shapiro-Wilk**

Contraste χ^2 de Pearson

UNIVERSIDAD DE MÁLAGA

- **Prueba de bondad de ajuste** más antigua. Aplicable a distribuciones discretas y continuas (en este último caso, agrupadas en intervalos), e incluso a atributos.

$$H_0: f(x) = f_0(x)$$
$$H_1: H_0 \text{ es falsa}$$

$\longrightarrow$ Los datos proceden de una población que sigue un determinado modelo de probabilidad

$f_0(x)$ puede estar completamente especificado (en forma y parámetros) o incluir algún parámetro desconocido que habrá que estimar por máxima verosimilitud

- El test se basa en la comparación de las **frecuencias observadas** en la muestra con las **esperadas** según el modelo propuesto (ver tabla de cálculos)

Contraste χ^2 de Pearson

UNIVERSIDAD DE MÁLAGA

Estadístico de prueba:

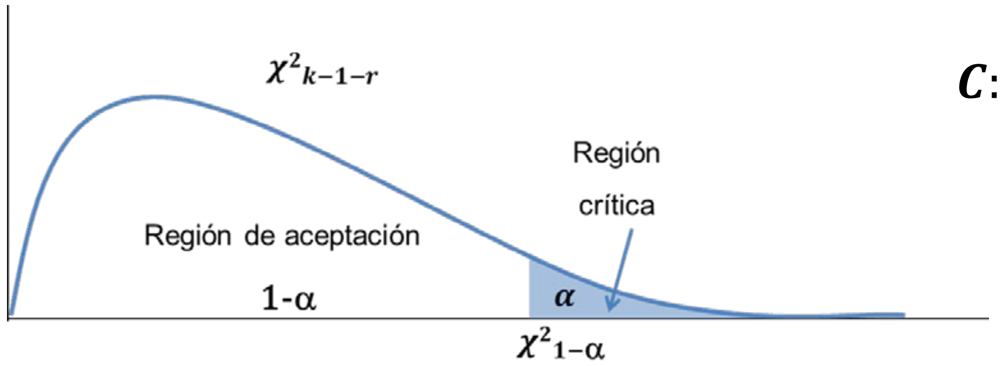
$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \xrightarrow[\substack{n\to\infty \\ E_i \geq 5}]{} \chi^2_{k-1-r}$$

k= nº de clases o intervalos
r=nº de parámetros estimados

OJO: Test asintótico (n grande)

Si $E_i=np_i<5$ es necesario **agrupar** las clases contiguas afectadas para que la aproximación a la χ^2 sea válida

Región crítica:



$$C:\ \chi^2_{obs} \geq \chi^2_{(1-\alpha)}$$

UNIVERSIDAD DE MÁLAGA

Cálculos para aplicar el contraste χ^2 de Pearson

Clases	Frecuencias observadas $O_i = n_i$	Probabilidades bajo H_0 p_i / H_0	Frecuencias esperadas $E_i = np_i$
S_1	$O_1 = n_1$	p_1	$e_1 = np_1$
S_2	$O_2 = n_2$	p_2	$e_2 = np_2$
..	..	..	..
S_i	$O_i = n_i$	p_i	$e_i = np_i$
..	..	..	..
S_k	$O_k = n_k$	p_n	$e_k = np_k$
	$\sum O_i = n$	$\sum p_i = 1$	$\sum E_i = n$

Volver a contraste χ^2

Contrastes de bondad de ajuste

- **Contrastes de bondad de ajuste**
  - **Contrastes generales**
    - Contraste χ^2 de Pearson
    - **Contraste de Kolmogorov-Smirnov**
  - Contrastes específicos de normalidad
    - Contraste de Kolmogorov-Smirnov-Lilliefors
    - Contraste de Jarque-Bera
    - Contraste de Shapiro-Wilk

Contraste de Kolmogorov-Smirnov

- **Prueba de bondad de ajuste** aplicable sólo a distribuciones continuas.

$$H_0: F(x) = F_0(x)$$
$$H_1: H_0 \text{ es falsa}$$

$\longrightarrow$ Los datos proceden de una población que sigue un determinado modelo de probabilidad continuo

El modelo debe estar totalmente especificado (en forma y parámetros)

- El test se basa en la comparación de la **función de distribución empírica** (frecuencias relativas acumuladas) con la función de distribución teórica (probabilidades acumuladas según el modelo propuesto)

Ver tabla de cálculos

Contraste de Kolmogorov-Smirnov

UNIVERSIDAD DE MÁLAGA

Cálculos para aplicar el contraste de Kolmogorov-Smirnov

Observaciones ordenadas x_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas $F_n(x_i)=N_i / n$	Probabilidad acumulada según H_0 $F_0(x_i)$	Distancia desde $F_n(x_{i-1})$ $D_1(x_i)$	Distancia desde $F_n(x_i)$ $D_2(x_i)$
x_1	N_1	$F_n(x_1)$	$F_0(x_1)$	$D_1(x_1)$	$D_2(x_1)$
x_2	N_2	$F_n(x_2)$	$F_0(x_2)$	$D_1(x_2)$	$D_2(x_2)$
..	..	..	..	..	..
x_i	N_i	$F_n(x_i)$	$F_0(x_i)$	$D_1(x_i)$	$D_2(x_i)$
..	..	..	..	..	..
x_n	N_n	$F_n(x_n)$	$F_0(x_n)$	$D_1(x_n)$	$D_2(x_n)$

UNIVERSIDAD DE MÁLAGA

Estadístico de prueba:

$$D_n = \max\{D_1(x_i) \cup D_2(x_i)\} \qquad i=1,2....n$$

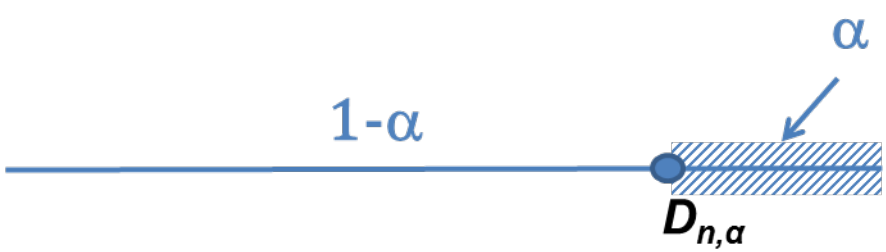
$$D_1(x_i) = |F_n(x_{i-1}) - F_0(x_i)| \qquad D_2(x_i) = |F_n(x_i) - F_0(x_i)|$$

Para cada x_i se calcula la **distancia entre la función de distribución empírica y la teórica**. Como la función empírica es un diagrama escalonado, su distancia a la distribución teórica puede medirse a partir de la ordenada $\boldsymbol{F_n(x_{i-1})}$ o de la $\boldsymbol{F_n(x_i)}$

Contraste de Kolmogorov-Smirnov

- **Región crítica:** cola derecha de la distribución del estadístico de prueba

$$C: D_{obs} \geq D_{n,\alpha}$$ ⟶ El valor crítico se busca en la tabla elaborada por Kolmogorov-Smirnov (en función n y α)

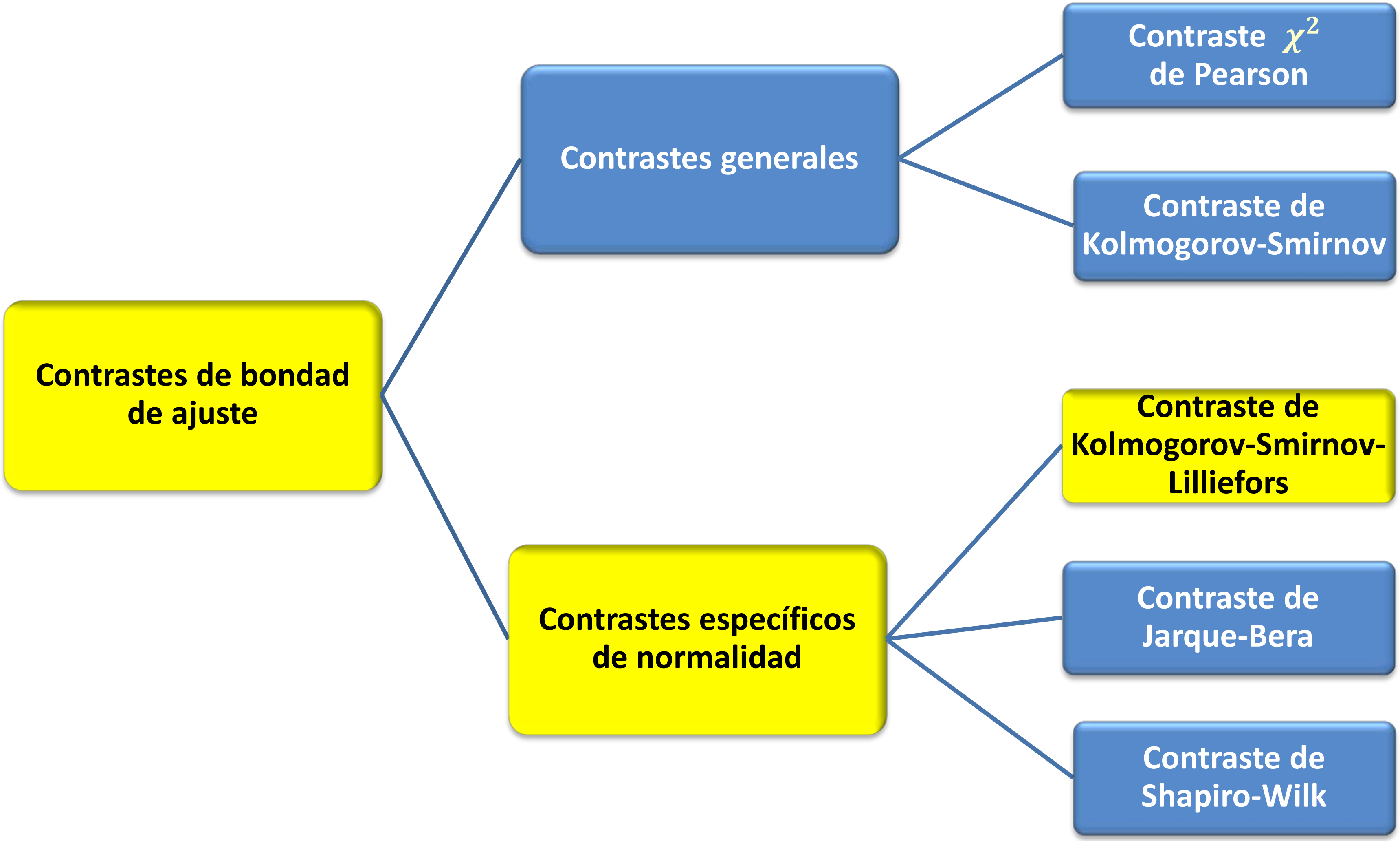


Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- Contrastes no paramétricos de localización para una población
- Contrastes no paramétricos de localización para dos poblaciones
- Test de rachas (aleatoriedad)
- Contrastes generales de bondad de ajuste
- **Contrastes específicos de normalidad**
- Contraste de independencia de dos variables aleatorias
- Contraste de homogeneidad

Contrastes de bondad de ajuste



Contraste de Kolmogorov-Smirnov-Lilliefors

- **Prueba de bondad de ajuste específica de normalidad**

H_0: $X \sim N\,(\mu, \sigma)$ *(normalidad)*
H_1: H_0 *es falsa*

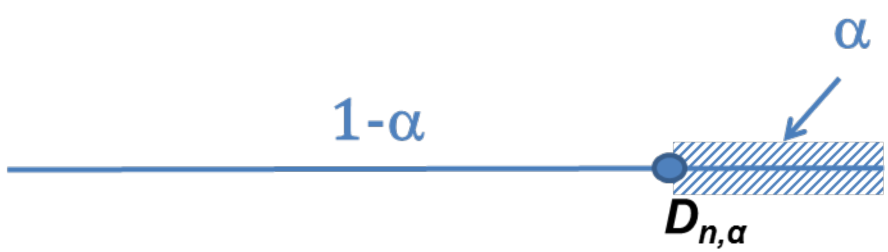
$\longrightarrow$ Los datos proceden de una población normal

No se especifican los parámetros

- Partiendo del test de Kolmogorov-Smirnov, Lilliefors tabuló la distribución del estadístico D_n para el contraste de normalidad, cuando μ **y** σ^2 **no se especifican** y se estiman en la muestra (mediante $\bar{X}$ y $\hat{S}^2$, respectivamente)

Contraste de Kolmogorov-Smirnov-Lilliefors

- Una vez obtenidos $\bar{X}$ y $\hat{S}^2$, el desarrollo del contraste es igual que el de Kolmogorov-Smirnov, aunque la tabla utilizada para la obtención de los valores críticos de D_n es diferente

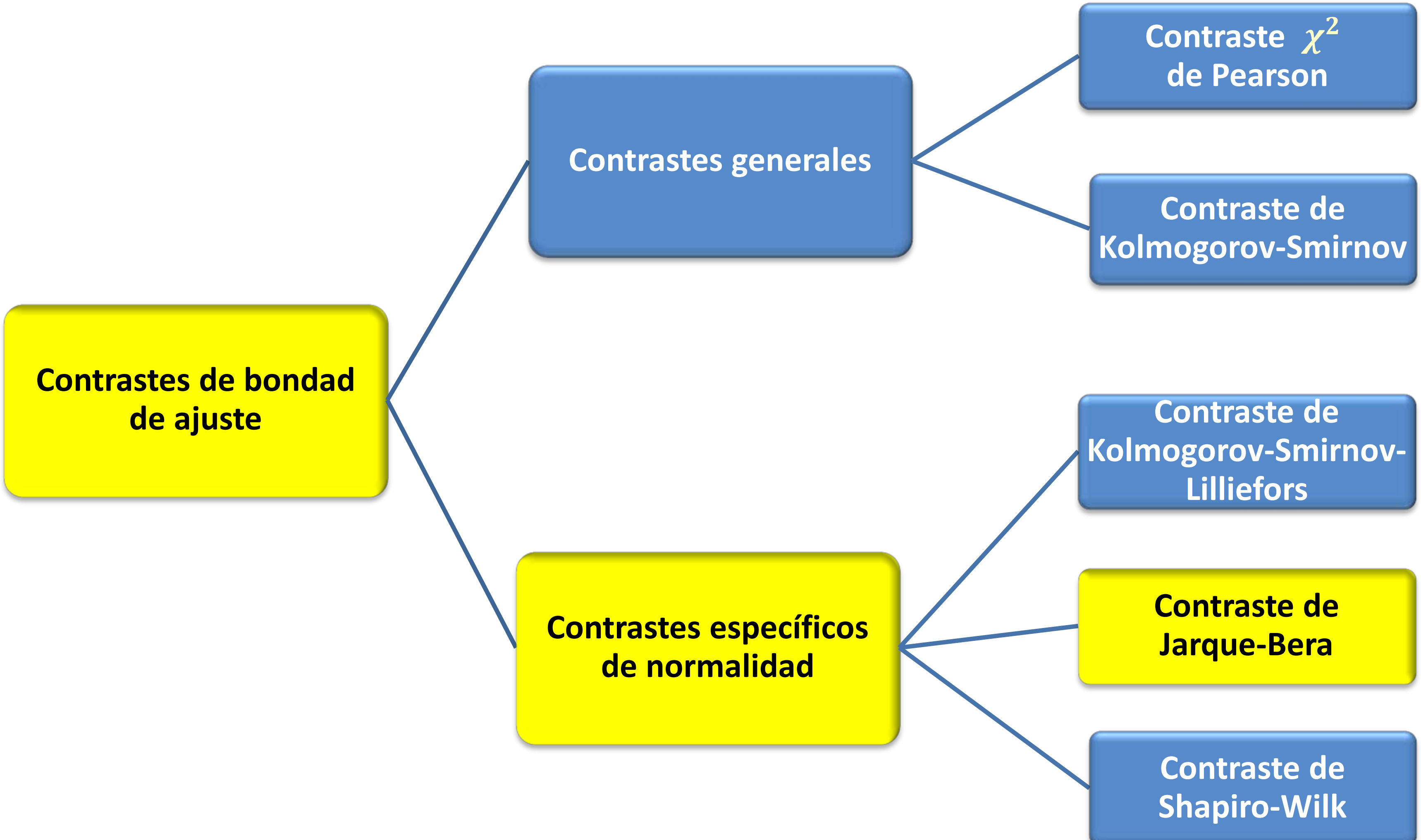


$$C: D_{obs} \geq D_{n,\alpha}$$

Valor crítico (se busca en la **tabla de Lilliefors**, en función de n y α)

- El contraste sólo debe aplicarse a **muestras grandes** ($n \geq 100$) porque en caso contrario sería poco potente

Contrastes de bondad de ajuste



Contraste de Jarque y Bera

- **Contraste específico de normalidad** basado en el análisis de las medidas de forma de una distribución: la **asimetría** y la **curtosis** o **apuntamiento**

H_0: $X \sim N(\mu, \sigma)$ (normalidad)

H_1: H_0 es falsa

→ Los datos proceden de una población normal

No se especifican los parámetros

- Coeficiente de asimetría: $\alpha_1 = \frac{\sum (X_i - \bar{X})^3}{nS^3}$

- Coeficiente de apuntamiento: $\alpha_2 = \frac{\sum (X_i - \bar{X})^4}{nS^4}$

Contraste de Jarque y Bera

Estadístico de prueba:

- Bajo la hipótesis de normalidad:

$$n \geq 50 \qquad \alpha_1 \to N\left(0, \sqrt{{}^{6}/_{n}}\right)$$

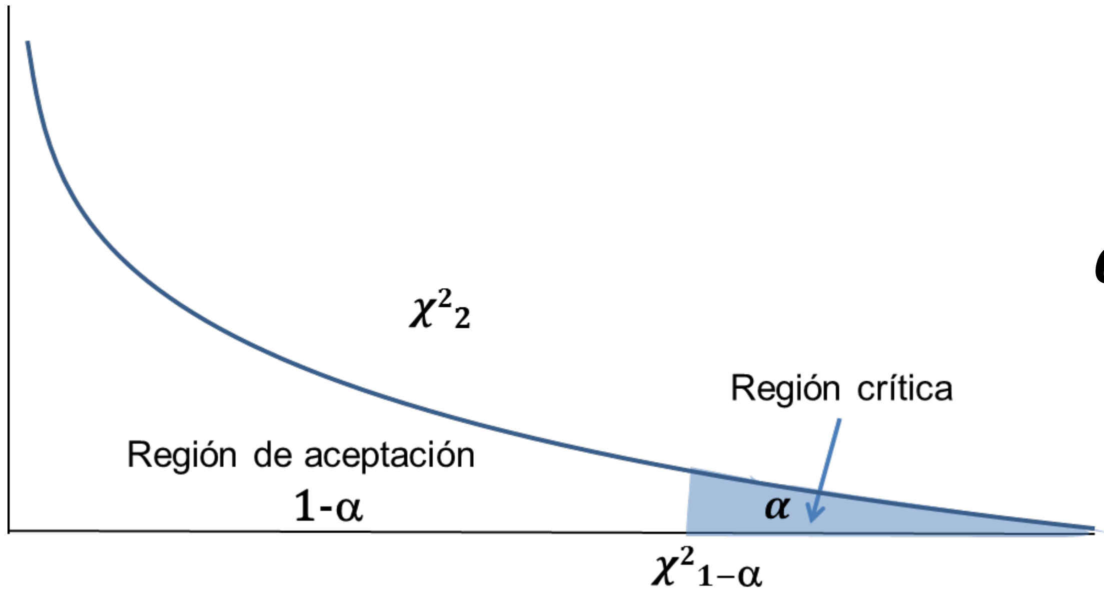
$$n \geq 200 \qquad \alpha_2 \to N\left(3, \sqrt{{}^{24}/_{n}}\right)$$

- Tipificando cada una de estas variables, elevando al cuadrado y sumándolas, se obtiene el estadístico de prueba del test

$$\boxed{\left(\frac{\alpha_1 - 0}{\sqrt{\frac{6}{n}}}\right)^2 + \left(\frac{\alpha_2 - 3}{\sqrt{\frac{24}{n}}}\right)^2 = \frac{n}{6}\left({\alpha_1}^2 + \frac{(\alpha_2 - 3)^2}{4}\right) \xrightarrow[n \geq 200]{} \chi_2^2}$$

Contraste de Jarque y Bera

- **Región crítica:**



$$C: \chi^2_{obs} \geq \chi^2_{1-\alpha}$$

- OJO: El contraste sólo debe aplicarse a **muestras grandes** (*n≥200*) para que la aproximación a la χ^2 sea adecuada

Contrastes de bondad de ajuste

UNIVERSIDAD DE MÁLAGA

- **Contrastes de bondad de ajuste**
  - Contrastes generales
    - Contraste χ^2 de Pearson
    - Contraste de Kolmogorov-Smirnov
  - **Contrastes específicos de normalidad**
    - Contraste de Kolmogorov-Smirnov-Lilliefors
    - Contraste de Jarque-Bera
    - **Contraste de Shapiro-Wilk**

Contraste de Shapiro-Wilk

UNIVERSIDAD DE MÁLAGA

- **Contraste específico de normalidad**

H_0: $X \sim N(\mu, \sigma)$ (normalidad)

H_1: H_0 es falsa

$\longrightarrow$

Los datos proceden de una población normal

No se especifican los parámetros

- El test se basa en el estudio del ajuste de los datos observados en la muestra a la **recta probabilística normal**
  - Si H_0 fuera cierta, los valores muestrales ordenados y tipificados constituirían una muestra ordenada de una N(0,1)

$$\frac{x_{(1)}-\mu}{\sigma}, \frac{x_{(2)}-\mu}{\sigma}, \ldots, \frac{x_{(n)}-\mu}{\sigma}$$

Contraste de Shapiro-Wilk

- El valor esperado del cuantil *i-ésimo* en esta muestra viene dado por:

$$E\left(\frac{x_{(i)} - \mu}{\sigma}\right) = c_{(i),n}$$

$$\boldsymbol{E(x_{(i)}) = \mu + \sigma c_{(i),n}}$$ → **Recta Probabilística Normal**

Si Ho es cierta los valores muestrales deberían estar próximos a esta recta

- **Gráficos Q-Q**: representación de los cuantiles de la muestra en relación a los valores esperados para una distribución normal

- **Contraste Shapiro-Wilk**: define un estadístico de prueba que cuantifica la bondad del ajuste (coeficiente de determinación)

Contraste de Shapiro-Wilk

UNIVERSIDAD DE MÁLAGA

- **Estadístico de prueba:**

$$W = \frac{1}{nS^2}\left[\sum_{i=1}^{h} a_{(i,n)}\left(x_{(n-i+1)} - x_{(i)}\right)\right]^2$$

$h = n/2$ (si n es par) ó $h=(n-1)/2$ (si n es impar)

S^2: Varianza muestral

$a_{(i,n)}$: Coeficientes tabulados

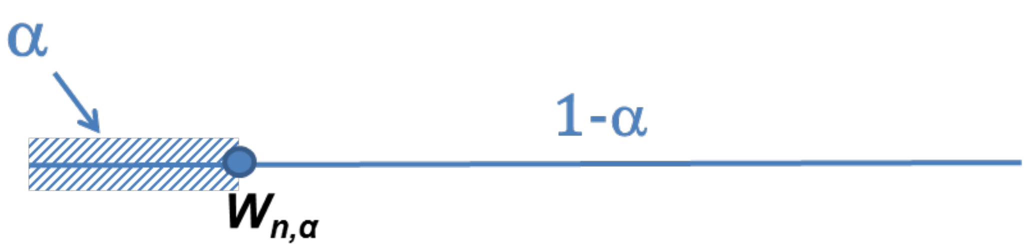
$x_{(i)}$: Valor que ocupa la posición i en la muestra ordenada

Contraste de Shapiro-Wilk

- **Región crítica:** cola izquierda de la distribución del estadístico de prueba (los valores bajos indican que el ajuste a la normal no es bueno)

$$C: W_{obs} \le W_{n,\alpha}$$

Valor crítico (se busca en la **tabla de Sapiro-Wilks**, en función de n y α)



Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- Contrastes no paramétricos de localización para una población
- Contrastes no paramétricos de localización para dos poblaciones
- Test de rachas (aleatoriedad)
- Contrastes generales de bondad de ajuste
- Contrastes específicos de normalidad
- **Contraste de independencia de dos variables aleatorias**
- Contraste de homogeneidad

Contraste de independencia

- **Objetivo:** verificar si existe asociación o dependencia entre dos características X e Y (cuantitativas o cualitativas) de una población

- **Hipótesis:**

H_0: *independencia entre X e Y* $\longrightarrow$ $p_{ij} = p_{i.}\ p_{.j}$
H_1: H_0 *es falsa*

- El test se basa en la comparación de las **frecuencias bidimensionales observadas** en la muestra $(\boldsymbol{O_{ij}} = \boldsymbol{n_{ij}})$ con las **esperadas** $(\boldsymbol{E_{ij}})$ según el modelo propuesto

- Las frecuencias observadas se recogen en una tabla de doble entrada denominada **tabla de correlación** (variables cuantitativas) o **de contingencia** (variables cualitativas)

Contraste de independencia: tabla de correlación o contingencia

Y / X	y_1	y_2	…	y_j	…	y_s	Total fila
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2.}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	$n_{i.}$
…	…	…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	$n_{r.}$
Total columna	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.s}$	n

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = n$$

Contraste de independencia

- **Estadístico de prueba:**

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}\xrightarrow[\substack{n\to\infty\\ E_i\geq 5}]{}\chi^2_{(r-1)(s-1)}$$

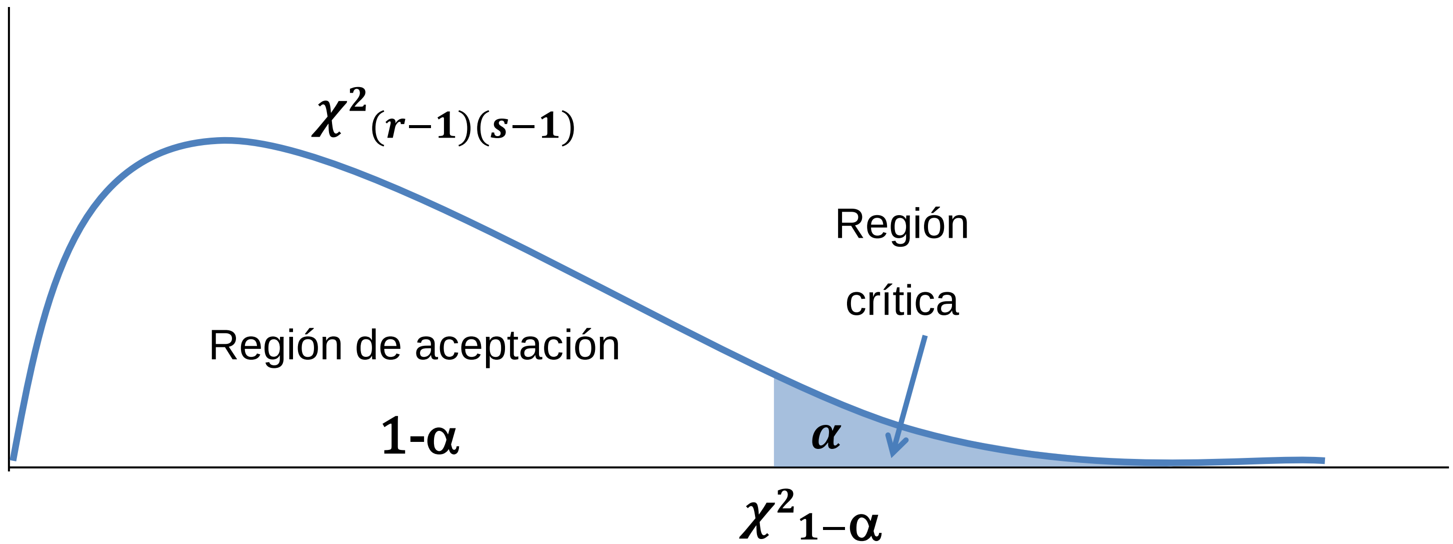
$$\boldsymbol{O_{ij}=n_{ij}} \qquad \boldsymbol{E_{ij}=\frac{n_{i.}n_{.j}}{n}}$$

OJO: - Test asintótico (n grande)

- Si alguna frecuencia esperada es menor que 5 ($E_{ij}<5$) es necesario **agrupar** celdas en la tabla (filas y/o columnas) para que la aproximación a la χ^2 sea válida

UNIVERSIDAD DE MÁLAGA

- **Región crítica:**



$$C:\ \chi^2_{obs} \geq \chi^2_{(1-\alpha)}$$

Verificación de hipótesis no paramétricas

UNIVERSIDAD DE MÁLAGA

- Introducción
- Contrastes no paramétricos de localización para una población
- Contrastes no paramétricos de localización para dos poblaciones
- Test de rachas (aleatoriedad)
- Contrastes generales de bondad de ajuste
- Contrastes específicos de normalidad
- Contraste de independencia de dos variables aleatorias
- **Contraste de homogeneidad**

Contraste de homogeneidad

- **Objetivo:** verificar si varias poblaciones son homogénas con respecto a una característica de clasificación X (cuantitativa o cualitativa). Es decir, si hay evidencia estadística de que varias muestras seleccionadas al azar provienen de la misma población (son resultados independientes del mismo experimento aleatorio)

- **Hipótesis:**

$$H_0: \text{homogeneidad} \longrightarrow p_{i1} = p_{i2} = \ldots = p_{ij} = \ldots = p_{is} = p_{i.}\,,\ \forall i$$

$$H_1: H_0 \text{ es falsa}$$

- El test se basa en la comparación de las **frecuencias observadas** en las muestras $(\boldsymbol{O_{ij}} = \boldsymbol{n_{ij}})$ con las **esperadas** $(\boldsymbol{E_{ij}})$ según el modelo propuesto

Distribución de frecuencias de X en las s muestras

X \ Muestras	1	2	…	j	…	s	Total fila
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2.}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	$n_{i.}$
…	…	…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	$n_{r.}$
Total columna	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.s}$	n

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = n$$

Contraste de homogeneidad

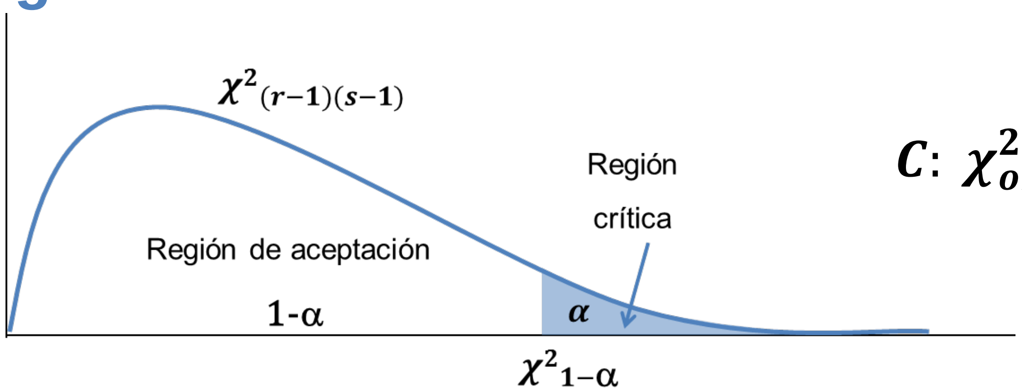
Estadístico de prueba:

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}\xrightarrow[E_{ij}\geq 5]{n\to\infty}\ \chi^2_{(r-1)(s-1)}$$

OJO: - Test asintótico (n grande)

- Si $E_{ij}<5$ hay que **agrupar** para que la aproximación a la χ^2 sea válida (preferiblemente, reducir categorías de la variable para no perder información de ninguna muestra)

Región crítica



$$C:\ \chi^2_{obs}\geq\chi^2_{(1-\alpha)}$$

Verificación de hipótesis no paramétricas

Mª Isabel Aguilar, Eugenia Cruces y Bárbara Díaz

UNIVERSIDAD DE MÁLAGA
Departamento de Economía Aplicada (Estadística y Econometría)
Parcialmente financiado a través del PIE13-024 (UMA)