

Tema 8. CONTRASTES PARAMÉTRICOS.

Objetivos

Conceptos:

- ✚ Comprender el método general del contraste de hipótesis para obtener información a partir de una muestra y tomar decisiones.
- ✚ Comprender el significado del p-valor de un contraste.

Saber hacer:

- ✚ **Dada una situación en la que se quiere estudiar la verosimilitud de una afirmación a partir de los datos de una muestra:**
 - **Plantear** el contraste eligiendo adecuadamente las hipótesis
 - Elegir el **pivote** adecuado (contrastos paramétricos) o el **test** adecuado (contrastos no paramétricos)
 - Hallar el **p-valor** del contraste (o por lo menos si es mayor o menor del nivel de significación α elegido)
 - Tomar una **decisión** a partir de los resultados obtenidos y el nivel de significación α elegido
 - bien a partir del valor del pivote (si está o no en la región crítica)
 - bien a partir del p-valor (si es mayor o menor que α)

NOTA (1)

Problemas de exámenes (web):

Contrastes Paramétricos:

- | | |
|---|---|
| SIN: Febrero 2006: Problema 3
Junio 2006: Problema 3 d)
Septiembre 2006: Problema 3 b)
Septiembre 2005: Problema 3 a)
Febrero 2004: Problema 3
Junio 2004: Problema 3 a) b)
Febrero 2003: Problema 3
Junio 2003. Problema 3 | CON: Febrero 2006 Parcial: (modelos A y B) 7. a 10.
Febrero 2006: Modelo A 3
Junio 2006: g) h)
Septiembre 2006: Modelos A y B f)
Febrero 2005: Dos primeros problemas
Septiembre 2005: g) |
|---|---|

Contrastes no paramétricos:

- SIN:** Febrero 2006: Problema 4 d)
 Junio 2006: Problema 3 c)
 Febrero 2005: Problema 3 a)
 Junio 2005: Problema 2 c)
 Septiembre 2004: Problema 3: 1.2.
 Septiembre 2003: Problema 2 c)

¹ En este curso habrá que saber plantear y resolver contrastes para estudiar:

- Media, varianza, diferencia de medias y cociente de varianzas, de poblaciones normales
- Media, proporción y diferencia de medias, de poblaciones cualesquiera
- Modelo de distribución de v.a. discretas y continuas

8.1.- Introducción. Tipos de contrastes.

En el tema 7 (intervalos de confianza) comentamos algunas de las conclusiones que se podían deducir de los resultados obtenidos, y que nos podían ayudar a tomar decisiones. En este tema, veremos otra técnica, el **contraste de hipótesis**, que también se utilizará para tomar decisiones (o sacar conclusiones) a partir de los resultados de una muestra.

Primero presentamos la **metodología general de los contrastes de hipótesis** y, posteriormente, su aplicación para:

- obtener información relativa a los parámetros de una variable (**contrastos paramétricos**) y
- para estudiar el tipo de modelo que sigue una determinada variable (**contrastos no paramétricos**).

8.2.- Elementos de un contraste

Pretendemos construir un modelo matemático que nos permita, a partir de los datos de una muestra:

- **decidir si una** determinada **hipótesis** sobre una población **es rechazable o no**,
- **medir el nivel de confianza** con el que aceptamos o rechazamos dicha hipótesis.

8.2.1. Planteamiento del problema. Hipótesis nula e hipótesis alternativa.

Para plantear un problema de *contraste de hipótesis* hemos de determinar **dos** hipótesis:

Hipótesis nula (H_0):

Es la hipótesis que mantendremos a no ser que los datos muestren de forma evidente su falsedad.

(Principio de la *presunción de inocencia*: toda persona es inocente mientras que no se demuestre lo contrario.)

Hipótesis alternativa (H_1):

Es la hipótesis que refleja la situación que nos haría rechazar la hipótesis nula. Para **rechazar H_0** frente a H_1 ha de haber una **fuerte evidencia** estadística (a partir de los datos de la muestra).

Ejemplos:

- 1) Decidir la inocencia o culpabilidad de una persona en un estado en donde siguen el principio de *presunción de inocencia*.

Como se quiere evitar condenar a una persona inocente, sólo se hará cuando haya una fuerte evidencia de su culpabilidad, cuando esté demostrada ésta. En caso de duda, se primará la inocencia frente a la culpabilidad. Por tanto, en la terminología propuesta sería:

$$\begin{cases} H_0 : \text{Inocente} \\ H_1 : \text{Culpable} \end{cases}$$

- 2) Decidir si un alumno sabe o no la asignatura de Estadística, y por tanto aprueba o suspende la asignatura.

Desde el punto de vista del profesorado, un estudiante no sabe la asignatura mientras no demuestre lo contrario; es decir, el examen ha de presentar pruebas suficientes de que conoce la asignatura. En general, en caso de duda o de falta de datos, se primará el suspenso frente al aprobado. Por tanto, en la terminología propuesta sería:

$$\begin{cases} H_0 : \text{El estudiante NO sabe la asignatura (Suspenso)} \\ H_1 : \text{El estudiante SI sabe la asignatura (Aprobado)} \end{cases}$$

3) Decidir si los vertidos de una fábrica están afectando o no el ecosistema de un río.

Se sabe que el ecosistema de determinada zona de un río se mantiene estable siempre que la temperatura del agua sea inferior a 18°C. Una asociación ecologista de la zona tiene sospechas de que los vertidos de una fábrica están afectando a la temperatura del agua y, ante sus denuncias, el ayuntamiento decide estudiar el tema a partir de los datos de la temperatura del agua en los últimos meses.

Para tomar una decisión, la empresa propone el siguiente contraste de hipótesis, siendo T la temperatura media del agua en esa zona del río.

$$\begin{cases} H_0 : \text{Los vertidos NO afectan al ecosistema } (T \leq 18) \\ H_1 : \text{Los vertidos SI afectan al ecosistema } (T > 18) \end{cases}$$

Con ese planteamiento, se concluiría que la empresa no afecta el ecosistema a no ser que haya una fuerte evidencia de que la temperatura es superior a 18°C.

Sin embargo, la asociación ecologista no está de acuerdo con ese planteamiento y defiende que la que tiene que demostrar su "inocencia" es la empresa y propone este otro:

$$\begin{cases} H_0 : \text{Los vertidos SI afectan al ecosistema } (T \geq 18) \\ H_1 : \text{Los vertidos NO afectan al ecosistema } (T < 18) \end{cases}$$

Con este otro planteamiento, se concluiría que la empresa no afecta el ecosistema cuando haya una fuerte evidencia de que así sea, es decir que esté "estadísticamente justificado" que la temperatura es inferior a 18°C.

Observación:

La metodología del *contraste de hipótesis*:

- No sirve para demostrar H_0 .
- Sirve para decir que, a partir de los datos de la muestra:
 - no puede rechazarse H_0
 - es aceptable suponer que H_0 es cierta
- Sirve para "demostrar" H_1 , en el sentido de que, a partir de los datos de la muestra, hay una fuerte evidencia de que H_1 es cierta en comparación con H_0 .

8.2.2. Errores.

Cuando trabajamos con el método del contraste de hipótesis podemos cometer dos tipos de errores:

	Rechazar H_0	No rechazar H_0
H_0 cierta	Error de tipo I	Decisión correcta
H_1 cierta	Decisión correcta	Error de tipo II

Ejemplos:

- 1) Decidir la inocencia o culpabilidad de una persona en un estado en donde siguen el principio de *presunción de inocencia*.

$\left\{ \begin{array}{l} H_0 : \text{Inocente} \\ H_1 : \text{Culpable} \end{array} \right.$		Rechazar H_0	No rechazar H_0
	H_0 cierta	Error de tipo I: Se condena a una persona que es inocente	Se absuelve a una persona que es inocente
	H_1 cierta	Se condena a una persona que es culpable	Error de tipo II: Se absuelve a una persona que es culpable

- 2) Decidir si un alumno sabe o no la asignatura de Estadística, y por tanto aprueba o suspende la asignatura.

$\left\{ \begin{array}{l} H_0 : \text{No sabe la asignatura} \\ H_1 : \text{Sí sabe la asignatura} \end{array} \right.$		Rechazar H_0	No rechazar H_0
	H_0 cierta	Error de tipo I: Se aprueba a un estudiante que NO sabe la asignatura	Se suspende a un estudiante que no sabe la asignatura
	H_1 cierta	Se aprueba a un estudiante que sí sabe la asignatura	Error de tipo II: Se suspende a un estudiante que SÍ sabe la asignatura

3) Decidir si los vertidos de una fábrica están afectando o no el ecosistema de un río.

Con el contraste de hipótesis propuesto por la empresa:

$\begin{cases} H_0 : \text{NO se afecta el ecosistema} \\ H_1 : \text{SI se afecta el ecosistema} \end{cases}$		Rechazar H_0	No rechazar H_0
	H_0 cierta	Error de tipo I: Se culpa a la empresa cuando en verdad no está afectando al ecosistema	Se exculpa a la empresa pues no está afectando al ecosistema
	H_1 cierta	Se culpa a la empresa pues sí está afectando al ecosistema	Error de tipo II: Se exculpa a la empresa cuando en verdad SÍ está afectando al ecosistema

8.2.3. Nivel de significación de un contraste. Potencia de un contraste.

Se definen:

- **Nivel de significación** del contraste: $\alpha = P$ (error de tipo I) = P (rechazar H_0 / H_0 cierta)
- **Potencia del contraste:** $1-\beta$, siendo $\beta = P$ (error de tipo II) = P (no rechazar H_0 / H_1 cierta)

Observación:

Lo ideal sería minimizar α y β . Pero esto no puede hacerse simultáneamente pues si se varía el criterio de rechazo para disminuir α entonces aumenta β , y viceversa. La única forma de disminuir ambos a la vez es incrementar el tamaño de la muestra.

Si un examen es muy exigente se disminuye $\alpha = P$ (error de tipo I), es decir la probabilidad de que apruebe un alumno que no sabe; sin embargo, aumenta β , la probabilidad de suspender a un estudiante que sí sabe la asignatura. Por el contrario, si el examen es poco exigente, disminuye la probabilidad de suspender a un estudiante que sí sabe la asignatura (β), pero aumenta α , la probabilidad de aprobar a estudiantes que no saben lo suficiente.

Para disminuir los dos, se puede incrementar "el tamaño de la muestra" (preguntar muchas cosas, para tener muchos datos sobre lo que no sabe o sí sabe el estudiante).

En general, por el carácter conservador de la metodología del *contraste de hipótesis*, **se fija un nivel de significación α** que asegure un error de tipo I admisible (haciendo mínima la probabilidad de "condenar a un inocente"). De entre los contrastes con un mismo nivel de significación, será preferible el contraste con mayor potencia².

En el ejemplo del ecosistema del río, el planteamiento de la empresa pretende minimizar la posibilidad de ser declarada culpable cuando en verdad no está afectando al ecosistema.

Por el contrario, el planteamiento de los ecologistas

² El estudio de la potencia se escapa de los objetivos de este curso. Los contrastes que presentamos cumplen la condición mencionada: son los de mayor potencia para un mismo nivel de significación.

$$\begin{cases} H_0 : \text{Los vertidos SI afectan al ecosistema } (T \geq 18) \\ H_1 : \text{Los vertidos NO afectan al ecosistema } (T < 18) \end{cases}$$

pretende minimizar la posibilidad de que la empresa sea exculpada cuando en verdad sí esté afectando al ecosistema.

8.3.- Metodología general de un contraste

Podemos distinguir las siguientes etapas para realizar un contraste de hipótesis:

- i) **Planteamiento:** se definen las hipótesis H_0 y H_1 .
- ii) Elegir un **estadístico** (X^*) como **medida de discrepancia** entre los datos muestrales y la hipótesis H_0 , comprobando que se cumplen las **condiciones necesarias**.
- iii) Decidir el **criterio para rechazar o no H_0** .
- iv) **Calcular el valor que toma X^*** a partir de los datos de la muestra y **aplicar el criterio** de decisión.

Veamos con un ejemplo en qué consiste cada una de ellas.

Ejemplo:

En general, los profesores nos quejamos mucho del rendimiento de nuestros estudiantes y a lo mejor no siempre con razón. Algunos consideran que un rendimiento razonable en una ingeniería es aprobar el 60% de las asignaturas matriculadas por curso. Un profesor muy optimista respecto al rendimiento de sus estudiantes de la EUI, afirma que a partir del 2º curso, ese rendimiento es incluso mayor. Sus compañeros le tachan de iluso pero él se empeña en demostrarlo utilizando la técnica del contraste de hipótesis.

- i) **Planteamiento:** se definen las hipótesis H_0 y H_1 .

Como se quiere ver si *hay evidencia estadística* o tener la *seguridad de que es cierta* la afirmación del "profesor optimista", según lo explicado en 8.2, haremos el siguiente planteamiento:

$$\begin{cases} H_0 : \mu = 60 \\ H_1 : \mu > 60 \end{cases}$$

siendo μ la media del porcentaje de asignaturas aprobadas respecto al número de asignaturas matriculadas.

Con este planteamiento, consideramos como hipótesis nula (H_0) la opción más *conservadora* (no hay un rendimiento extraordinario en nuestros estudiantes) y como hipótesis alternativa (H_1) la opción de la se quiere tener *certeza o pruebas estadísticamente significativas*.

- ii) Elegir un **estadístico** (X^*) como **medida de discrepancia** entre los datos muestrales y la hipótesis H_0 , comprobando que se cumplen las **condiciones necesarias**.

La decisión se tomará a partir de los datos obtenidos en una muestra y necesitamos tener alguna "herramienta" que nos permita medir si esos datos apoyan más a una hipótesis o a otra, siempre teniendo en cuenta que "favoreceremos a H_0 mientras no se demuestre que H_1 es más verosímil".

En este caso, la "herramienta de medida" será un **estadístico** (o pivote como lo denominamos en el tema 7), que nos servirá para medir:

- si los datos de la muestra apoyan que H_0 es cierta,
- o
- si dan más credibilidad a H_1 .

Es lo que denominamos **medida de discrepancia**.

Al igual que en tema 7, este estadístico depende de los datos de la **muestra** y tendrá una **distribución de probabilidad conocida**.

Para que esto se dé y su aplicación sea correcta, habrá que comprobar que se cumplen algunas **condiciones necesarias**:

- que $X \sim N(\mu, \sigma)$ para estudiar medias y varianzas de poblaciones para cualquier tamaño de la muestra,
- que el tamaño de la muestra sea suficientemente grande para estudiar y comparar medias de poblaciones cualesquiera,
- distinguir entre variables discretas y continuas, y algunas otras condiciones en el caso de contrastes sobre el modelo de distribución.

En el ejemplo, la hipótesis es relativa a la **media** de una variable (*porcentaje de asignaturas aprobadas respecto al número de asignaturas matriculadas*) y en el tema anterior vimos que para estudiarla se utiliza el **estadístico**:

$$X^* = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Eso es cierto siempre que la variable siga una distribución normal o que el tamaño de la muestra sea suficientemente grande.

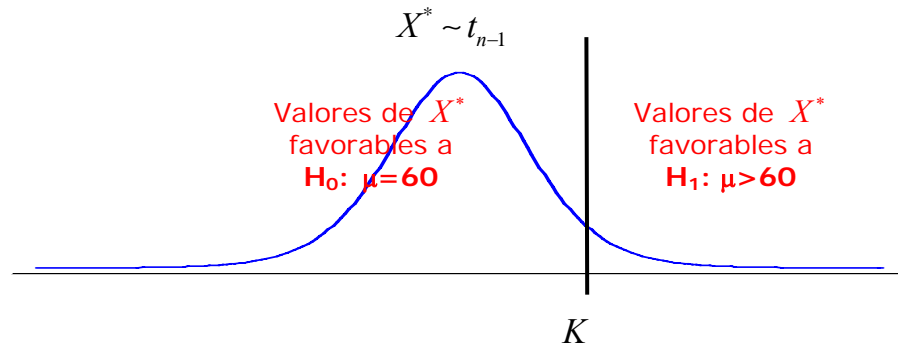
En este caso, se puede suponer que sigue una distribución **normal** (aproximate p-value: 0.69932, para la muestra obtenida con los estudiantes del SM22 del curso 2006/07).

Como la opción que vamos a "favorecer a priori" es H_0 , y queremos medir la discrepancia entre la muestra y H_0 , se trabajará con el estadístico X^* **bajo el supuesto de que H_0 es cierta**. Se tiene que:

$$X^* = \frac{\bar{X} - 60}{S/\sqrt{n}} \sim t_{n-1}$$

En este caso, si H_0 es cierta ($\mu = 60$) el valor que tome X^* a partir de los datos de la muestra ha de estar entorno al 0 (se supone que el valor obtenido en la muestra para \bar{X} será cercano a 60).

Por el contrario, si el valor que toma X^* es mucho mayor que 0, tendrá más credibilidad H_1 (pues significa que el valor obtenido en la muestra para \bar{X} es bastante mayor que 60).



iii) Decidir el **criterio para rechazar H_0** .

Una vez que ya tenemos una *medida de discrepancia* nos planteamos las siguientes preguntas:

- o ¿Cómo valoramos si es "mucho mayor" que cero?
- o ¿Cuándo será "suficientemente grande" para afirmar que H_1 es más verosímil que H_0 ?
- o ¿Con qué criterio hallamos el valor de K a partir del cual rechazamos H_0 frente a H_1 ?

Para determinar el criterio con el que rechazaremos o no H_0 :

- **se fija un nivel de significación α** (controlando y minimizando la probabilidad del error de tipo I)
- en función de él **se determinan qué valores del estadístico X^*** son favorables a rechazar H_0 y cuáles no.

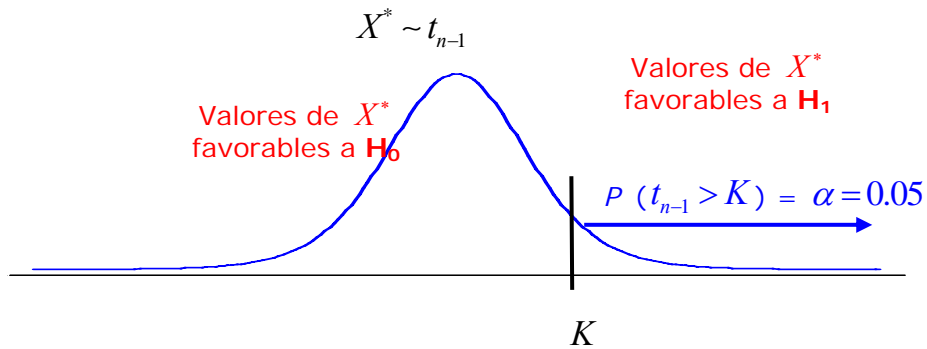
$$\begin{aligned} \text{Como } \alpha &= P(\text{error de tipo I}) = \\ &= P(\text{rechazar } H_0 / H_0 \text{ cierta}) = \end{aligned}$$

$$= P(X^* \text{ tome valores "favorables" a } H_1 / X^* \text{ sigue un modelo de distribución que supone } H_0 \text{ cierta})$$

En nuestro ejemplo: $\alpha = P \left(\underbrace{X^* > K}_{\text{Rechazamos } H_0} / \underbrace{X^* = \frac{\bar{X} - 60}{S/\sqrt{n}} \sim t_{n-1}}_{\text{Definición y distribución del pivote suponiendo cierta } H_0: \mu=60} \right)$.

Si fijamos un *nivel de significación* $\alpha=0.05$, tendríamos que hallar el valor de K tal que

$$P(t_{n-1} > K) = 0.05$$



Por tanto, un **criterio de rechazo** será el siguiente:

Siendo x^* el valor de X^* a partir de los datos de la muestra y $P(t_{n-1} > K) = 0.05$:

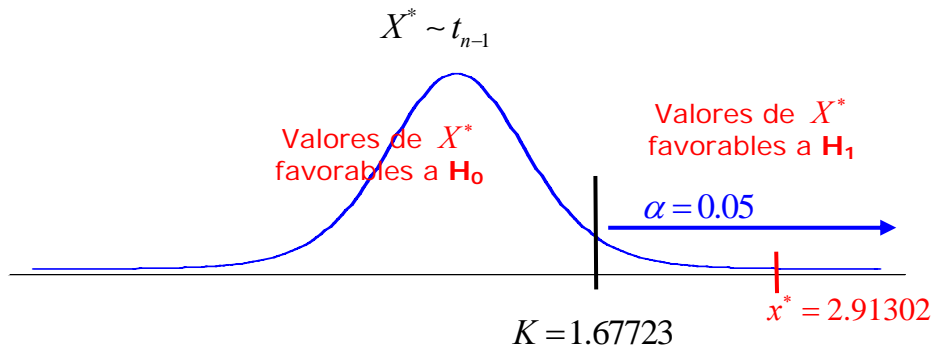
- o Rechazamos H_0 si $x^* > K$
- o No Rechazamos H_0 si $x^* < K$

iii) **Calcular el valor de X^*** a partir de la muestra y **aplicar el criterio** de decisión.

A partir de los valores obtenidos en la muestra (grupo SM22, curso 2006/06):

- o Hallamos x^* : $x^* = \frac{\bar{x} - 60}{s/\sqrt{n}} = \frac{68.7781 - 60}{21.0937/\sqrt{49}} = 2.91302$
- o Hallamos K : $P(t_{48} > K) = 0.05 \Rightarrow K = 1.67723$
Statgraphics

Como $x^* > K$, **rechazamos H_0** . Por tanto, **hay evidencia estadística ($\alpha=0.05$) de que H_1 es cierta**.

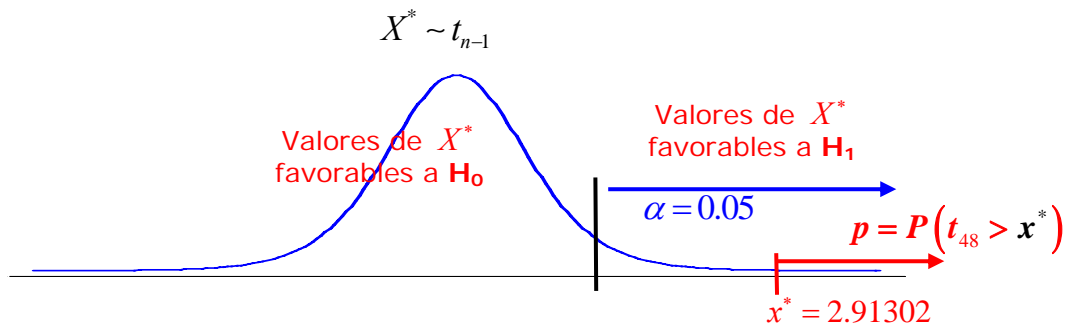


En el contexto del problema: podemos asegurar ($\alpha=0.05$) que el porcentaje de asignaturas que aprueban los estudiantes de la EUI respecto al número de asignaturas matriculadas es mayor que el 60%.

Observación:

Otra forma de saber si el valor x^* obtenido en la muestra está en la región favorable a H_0 o en la región favorable a H_1 , sería calcular la probabilidad $p = P(t_{48} > x^*)$.

- o Si $p < \alpha = 0.05$, significa que el valor x^* obtenido en la muestra está en la región favorable a H_1 (Rechazamos H_0)
- o Si $p > \alpha = 0.05$, significa que el valor x^* obtenido en la muestra está en la región favorable a H_0 (No rechazamos H_0)



En este caso $p = P(t_{48} > x^*) = P(t_{48} > 2.91302) = 0.0027 < 0.05$. Por tanto, rechazamos H_0 .
Statgraphics

En adelante, a esta probabilidad p , la denominaremos **p-valor del contraste**.

Conocer el **p-valor de un contraste** tiene la ventaja de que nos da más información. Puede interpretarse como **la probabilidad del equivocarnos si rechazamos H_0 a partir de los datos obtenidos en la muestra.**

Por tanto:

- o **cuanto más pequeño sea, rechazaremos H_0** con mayor seguridad (los datos de la muestra son más favorables a H_1)
- o **cuanto mayor sea, aceptaremos H_0** con mayor confianza, (los datos de la muestra son más favorables a H_0)

Los paquetes estadísticos suelen dar los resultados de los contrastes de hipótesis en función del p-valor. Por ejemplo, para los datos de este ejemplo, Statgraphics nos da la siguiente información:

Hypothesis Tests for AprobaMatricula

t-test

Null hypothesis: mean = 60,0

Alternative: greater than

Computed t statistic = 2,91302 (Valor de x^*)

P-Value = 0,00270972

Reject the null hypothesis for alpha = 0,05.

The StatAdvisor

[...]Since the P-value for this test is less than 0,05, we can reject the null hypothesis at the 95,0% confidence level.

8.4.- Criterios de rechazo.

Generalizamos ahora los criterios de rechazo, para todo tipo de contrastes.

8.4.1 Criterio de la región crítica

Hemos visto que para determinar los criterios de rechazo nos basaremos en la definición del nivel de significación:

$$\alpha = P(\text{rechazar } H_0 / H_0 \text{ cierta}) =$$

$$= P(X^* \text{ tome valores "favorables" a } H_1 / X^* \text{ sigue un modelo de distribución que supone } H_0 \text{ cierta})$$

Denominaremos **región crítica (RC)** a aquellos valores de X^* con los que rechazaríamos H_0 con un nivel de significación α (error de tipo I).

El criterio de rechazo será:

- o **Rechazamos H_0** cuando $x^* \in RC$.
- o **No rechazamos H_0** cuando $x^* \notin RC$.

siendo x^* es el valor del estadístico X^* a partir de los datos de la muestra.

En función de la hipótesis alternativa elegida, la región crítica será diferente:

Hipótesis alternativa	Región crítica	
$H_1: \theta > \theta_0$	<p>RC: $(K, +\infty)$</p> <p>siendo</p> $P(X^* > K / H_0 \text{ cierta}) = \alpha$	
$H_1: \theta < \theta_0$	<p>RC: $(-\infty, K)$</p> <p>siendo</p> $P(X^* < K / H_0 \text{ cierta}) = \alpha$	

$H_1: \theta \neq \theta_0$	$RC: (-\infty, K_1) \cup (K_2, +\infty)$ siendo $P\left(X^* < K_1 / H_0 \text{ cierta}\right) = \frac{\alpha}{2}$ $P\left(X^* > K_2 / H_0 \text{ cierta}\right) = \frac{\alpha}{2}$	
-----------------------------	--	--

En general, para facilitar los cálculos y tener una definición clara del estadístico X^* y de su modelo de distribución, H_0 se suele elegir del tipo $\theta = \theta_0$, hipótesis simple (el parámetro sólo toma un valor)³.

Por otra parte, H_1 suele ser una hipótesis compuesta (el parámetro puede tomar más de un valor): $\theta > \theta_0, \theta < \theta_0, \theta \neq \theta_0$.

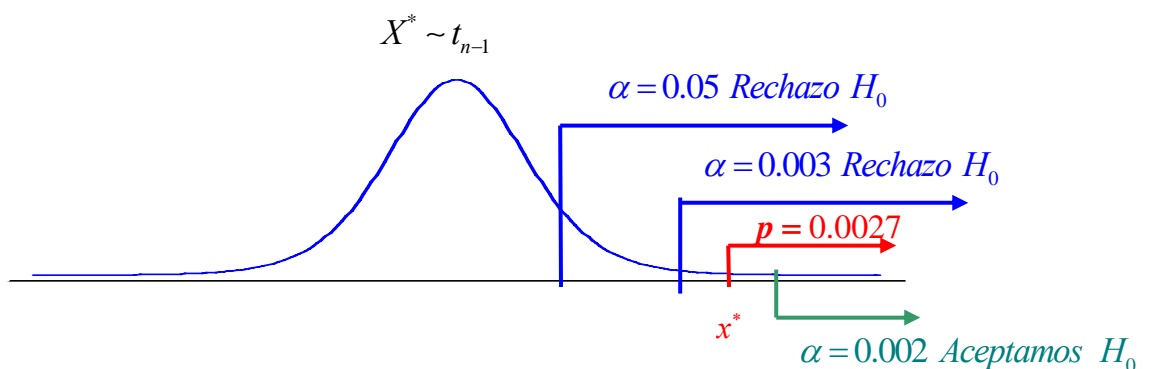
8.4.2.- El p-valor de un contraste. Criterio de rechazo

Ya hemos visto que **p-valor de un contraste** puede interpretarse como *la probabilidad del equivocarnos si rechazamos H_0 a partir de los datos obtenidos en la muestra*.

Es decir, el p-valor es la probabilidad de cometer un error de tipo I a partir de una muestra concreta, por lo que viene a ser un *nivel de significación*.

La **definición rigurosa del p-valor** de un contraste es: *el mínimo nivel de significación para el que se rechazaría H_0 con los datos de una muestra concreta*.

En el ejemplo del rendimiento de los estudiantes de ingeniería (porcentaje de asignaturas aprobadas respecto al número de asignaturas matriculadas), con el contraste $\begin{cases} H_0: \mu = 60 \\ H_1: \mu > 60 \end{cases}$, segúnelijamos un nivel de significación u otro la decisión puede cambiar:



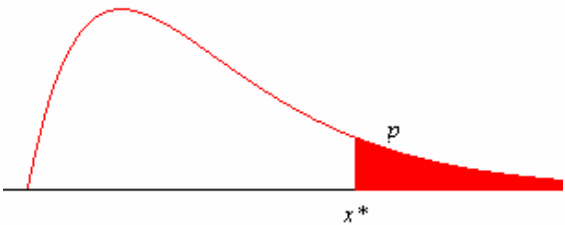
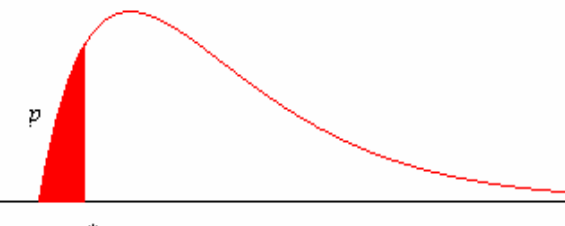
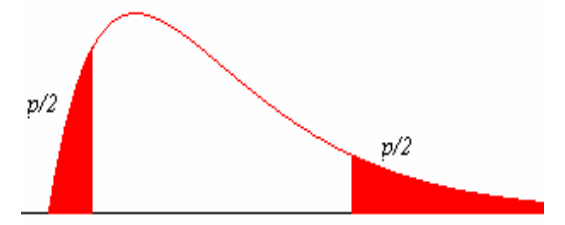
- o si fijamos $\alpha = 0.05$, se tiene que $p = 0.0027 < \alpha$, luego se rechaza H_0 .
- o si fijamos $\alpha = 0.003$, se tiene que $p = 0.0027 < \alpha$, también se rechaza H_0 .
- o si fijamos $\alpha = 0.002$, se tiene que $p = 0.0027 > \alpha$, y en este caso **no podemos rechazar H_0** , pues el valor de x^* cae dentro de la región que consideramos favorable a H_0 .

³ También nos basamos en el *principio de simplicidad científica*: sólo abandonaremos un modelo simple a favor de otro más complejo cuando la evidencia de este último sea fuerte. Además, se puede demostrar que de esta forma se obtienen los criterios de rechazo que optimizan la potencia del contraste para el valor de significación α elegido.

El valor $\alpha = p = 0.0027$, sería el nivel de significación "frontera" para rechazar o aceptar H_0 con los datos obtenidos en esa muestra.

Cálculo del p-valor:

Dependerá de la hipótesis alternativa elegida:

Hipótesis alternativa	p-valor <i>(x^* es el valor del estadístico X^* a partir de los datos de la muestra)</i>	
$H_1: \theta > \theta_0$	$p = P\left(X^* > x^* / H_0 \text{ cierta}\right)$	
$H_1: \theta < \theta_0$	$p = P\left(X^* < x^* / H_0 \text{ cierta}\right)$	
$H_1: \theta \neq \theta_0$	$p = 2 \cdot P\left(X^* > x^* / H_0 \text{ cierta}\right)$ <p style="text-align: center;">ó</p> $p = 2 \cdot P\left(X^* < x^* / H_0 \text{ cierta}\right)$	

Criterios de decisión en función del p-valor:

- Si hemos fijado un nivel de significación α para el contraste ($\alpha = P(\text{error de tipo I})$):
 - Rechazamos H_0 cuando **p-valor** $< \alpha$.
 - No rechazamos H_0 cuando **p-valor** $> \alpha$.

- Si no hemos fijado un nivel de significación α para el contraste:
 - Rechazamos H_0 cuando **p-valor** < 0.01 .
 - No rechazamos H_0 cuando **p-valor** > 0.3 .
 - Lo "pensamos con cuidado" cuando **0.01** $< \text{p-valor} < 0.3$.

Ejemplo:

Siguiendo con el ejemplo del rendimiento de los estudiantes de ingeniería (porcentaje de asignaturas aprobadas respecto al número de asignaturas matriculadas), nuestro "profesor optimista" quiere justificar estadísticamente que el rendimiento de los estudiantes de la EUI se puede suponer mayor o igual al 70% de las asignaturas matriculadas.

i) Para ello, **plantea** el siguiente contraste:

$$\begin{cases} H_0 : \mu \geq 70 \\ H_1 : \mu < 70 \end{cases}$$

siendo μ la media del porcentaje de asignaturas aprobadas respecto al número de asignaturas matriculadas.

Con este planteamiento, considera como hipótesis nula (H_0) su propia hipótesis y como hipótesis alternativa (H_1) la opción contraria. Es decir, él seguirá manteniendo su opinión de que el rendimiento podría suponerse mayor o igual al 70% siempre que no haya una fuerte evidencia estadística a favor de que dicho rendimiento es menor del 70%.

ii) El **estadístico** elegido es

$$X^* = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

pues es un contraste relativo a la media de una población normal.

iii) **Criterio de rechazo:**

Región crítica: Como la hipótesis alternativa es del tipo $H_1 : \mu < 70$, la **región crítica** es del tipo **RC:** $(-\infty, K)$, siendo $P\left(X^* < K / H_0 \text{ cierta}\right) = \alpha$.

Para facilitar los cálculos, suponemos que H_0 es una hipótesis simple ($H_0 : \mu = 70$). Por tanto, si

$$H_0 \text{ es cierta, } X^* = \frac{\bar{X} - 70}{S/\sqrt{n}} \sim t_{n-1}.$$

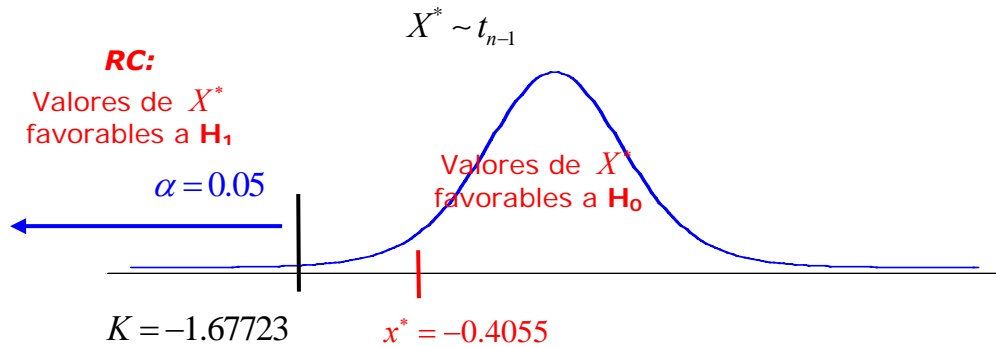
Si **fijamos un nivel de significación** $\alpha=0.05$, el criterio sería:

Rechazamos H_0 si $x^* \in RC$ ($x^* < K$), siendo $P(t_{n-1} < K) = 0.05$

o Hallamos x^* : $x^* = \frac{\bar{x} - 70}{s/\sqrt{n}} = \frac{68.7781 - 70}{21.0937/\sqrt{49}} = -0.4055$

o Hallamos K : $P(t_{48} < K) = 0.05 \Rightarrow K = -1.67723 \Rightarrow RC : (-\infty, -1.67723)$
Statgraphics

Como $x^* = -0.4055 \notin RC$, **no rechazamos H_0** .



Criterio del p-valor:

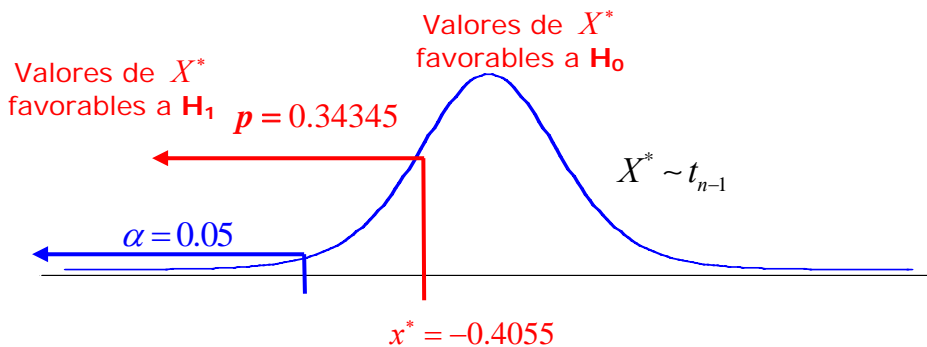
Como en este caso la hipótesis alternativa es del tipo $H_1 : \mu < 70$, se tiene que:

$$p\text{-valor} = P(X^* < x^* / H_0 \text{ cierta})$$

Por tanto, $p\text{-valor} = P(t_{n-1} < x^*)$.

o Hallamos x^* : $x^* = \frac{\bar{x} - 70}{s/\sqrt{n}} = \frac{68.7781 - 70}{21.0937/\sqrt{49}} = -0.4055$

o Hallamos el $p\text{-valor}$: $p = P(t_{48} < -0.4055) = 0.34345$
Statgraphics



A partir de ese valor, utilizamos los **criterios de decisión** vistos anteriormente:

- Si fijamos un nivel de significación $\alpha=0.05$, tenemos que $p > \alpha$, luego **no rechazamos H_0** .
- Si no fijamos un nivel de significación tenemos que $p > 0.3$, luego **no rechazamos H_0** .

Conclusión: Por tanto, tanto por un criterio como por otro, el resultado es claro y no podemos rechazar H_0 . Es decir, es aceptable suponer que el rendimiento de los estudiantes de la EUI es del 70% (porcentaje de asignaturas aprobadas respecto de las matriculadas), pues no hay evidencia estadística de que la media sea menor que 70.

8.5.- Contrastes paramétricos.

8.5.1. Poblaciones normales y para la media en poblaciones no normales.

Para cada caso, se elige el mismo estadístico pivote que para hallar el intervalo de confianza y se sigue el método explicado anteriormente (con criterio de rechazo del p-valor).

8.5.2.- El caso particular de la proporción.

El modelo matemático para estudiar la proporción es el mismo que se ha visto en el tema 7 (7.3.2).

Sin embargo, en el caso de un contraste de hipótesis para la proporción, con $H_0 : p=p_0$, se toma

como estadístico
$$X^* = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1).$$

Esto es porque, si suponemos cierta $H_0 : p=p_0$, se tiene que $V(X) = p_0(1-p_0)$, pues $X \sim B(1, p_0)$. Es decir, conocemos el valor de la varianza y no es necesario utilizar la cuasivarianza muestral. De esta forma, se puede aplicar directamente el Teorema Central del Limite y aproximamos directamente por una $N(0,1)$.

Por tanto, la única condición que hemos de pedir es que el **tamaño de la muestra** sea **suficientemente grande**.

Ejemplo:

Uno de los factores del bajo rendimiento académico es una predisposición emotiva hacia las materias que se estudian. Nuestra experiencia nos hacía estimar que, más o menos, la mitad de los estudiantes tenían una actitud negativa ante las matemáticas. Pero ante los resultados de los últimos años, algunos profesores queremos estudiar si hay evidencia estadística para afirmar que *más del 50% de los estudiantes de la EUI tienen sensaciones negativas ante las matemáticas*. Usaremos para ello la técnica del contraste de hipótesis ($\alpha=0.05$).

i) **Planteamiento:** se definen las hipótesis H_0 y H_1 .

Nuestra hipótesis es que *más del 50% de los estudiantes de la EUI tienen sensaciones negativas ante las matemáticas*. Como queremos ver si *hay evidencia* o tener la *seguridad de que es cierta*, según lo explicado antes, hacemos el siguiente planteamiento:

$$\begin{cases} H_0 : p = 0.5 \\ H_1 : p > 0.5 \end{cases}$$

siendo p la proporción de estudiantes de la EUI con sensaciones negativas ante las matemáticas.

Con este planteamiento, consideramos como hipótesis nula (H_0) la opción más *conservadora* (no hay cambios significativos en los últimos años) y como hipótesis alternativa (H_1) la opción de la queremos tener certeza antes de afirmarla categóricamente o tomar alguna decisión pertinente.

ii) Elección del **estadístico**, y comprobación de las **condiciones**:

En este caso, la hipótesis es relativa a la **proporción** de individuos de una población (*estudiantes de la EUI*) que cumplen una determinada característica (*tienen sensaciones negativas ante las matemáticas*). Suponiendo un **tamaño de la muestra suficientemente grande**, elegimos el estadístico:

$$X^* = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx N(0,1)$$

iii) Elegir **criterio de rechazo**:

Utilizaremos el del **p-valor**. Como la hipótesis alternativa es del tipo $H_1 : p > 0.5$, se tiene que:

$$p\text{-valor} = P\left(X^* > x^* / H_0 \text{ cierta}\right)$$

Por tanto, si H_0 es cierta ($p=0.5$):

$$\begin{aligned} \circ X^* &= \frac{\hat{P} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}} \approx N(0,1) \\ \circ p\text{-valor} &= P(N(0,1) > x^*) \end{aligned}$$

iv) **Calcular el valor de X^*** a partir de la muestra y **aplicar el criterio** de decisión:

Consideramos la misma muestra que en el tema 7. Con tamaño $n=86$, tenemos que hay 51 estudiantes que tenían sensaciones que se pueden considerar negativas ante las matemáticas. Con esos datos, una **estimación puntual de la proporción** sería $\hat{P} = \frac{51}{86} = 0.5930$.

$$\begin{aligned} \circ \text{Hallamos } x^* : x^* &= \frac{0.5930 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{86}}} = 1.7249 \\ \circ \text{Hallamos el } p\text{-valor: } p &= P(N(0,1) > 1.7249) = 1 - P(N(0,1) < 1.7249) = 1 - 0.958 = 0.042 \end{aligned}$$

tablas
aprox

Utilizamos los **criterios de decisión** vistos anteriormente:

- Si fijamos un nivel de significación $\alpha=0.05$, tenemos:

$$p=0.042 < \alpha, \text{ luego rechazamos } H_0.$$

Por tanto, de este estudio se deduce que **hay evidencia estadística** ($\alpha=0.05$) **de que la proporción** de estudiantes de la EUI con sensaciones negativas ante las matemáticas **es mayor del 50%**.

8.5.3.- Relación entre contrastes bilaterales e intervalos de confianza.

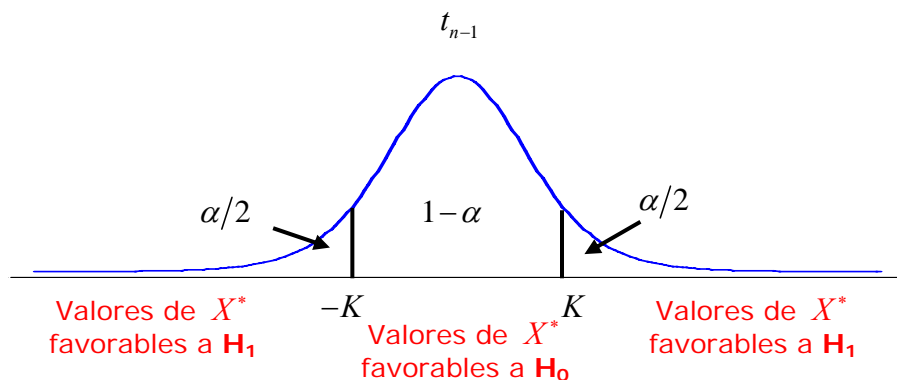
Por la metodología seguida para construir intervalos de confianza y los criterios de rechazo de un contraste, se verifica que si un contraste es del tipo: $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$, entonces son equivalentes:

- **Rechazar H_0** con nivel de significación $\alpha \Leftrightarrow \theta_0 \notin \mathbf{IC}(\theta)$ con nivel de confianza $1-\alpha$
- **No rechazar H_0** con nivel de significación $\alpha \Leftrightarrow \theta_0 \in \mathbf{IC}(\theta)$ con nivel de confianza $1-\alpha$

Lo justificamos en el caso del estudio de la media de una población (las ideas son las mismas):

Elegimos el estadístico $X^* = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$.

No rechazamos H_0 con nivel de significación α cuando x^* cae en la región "favorable a H_0 ", es decir, $-K < x^* < K$, siendo K tal que $P(t_{n-1} < -K) = \frac{\alpha}{2}$ y $P(t_{n-1} > K) = \frac{\alpha}{2}$:



Tenemos por tanto $-K < x^* < K \Leftrightarrow -K < \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < K \Leftrightarrow \bar{x} - K \cdot \frac{s}{\sqrt{n}} < \mu_0 < \bar{x} + K \cdot \frac{s}{\sqrt{n}}$, que es la expresión del intervalo de confianza para la media. Es decir, $\mu_0 \in \mathbf{IC}(\mu)$ con nivel de confianza $1-\alpha$.

En este tipo de contrastes, denominados **bilaterales**, podemos utilizar el intervalo de confianza para tomar la decisión.

Ejemplo:

En el tema 7 estudiamos las diferencias de la variable *Estatura* entre chicos (E_X) y chicas (E_Y) de la EUI a partir de intervalos de confianza (95% de confianza). Comparamos ahora las medias y varianzas utilizando contrastes de hipótesis bilaterales y llegaremos a resultados idénticos. Para ello, también tenemos que seguir los pasos propuestos en el tema 7:

a) Comprobar que ambas variables siguen distribuciones **normales**.

Ya se vió en el tema 7 que lo eran.

b) Estudiamos si las varianzas se pueden considerar iguales o no ($\alpha=0.05$).

i) **Planteamos** el contraste:
$$\begin{cases} H_0 : \sigma_X^2 = \sigma_Y^2 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 \end{cases} \equiv \begin{cases} H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1 \\ H_1 : \frac{\sigma_X^2}{\sigma_Y^2} \neq 1 \end{cases}$$

ii) Elegimos el **estadístico** $X^* = \frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim F_{n_X-1, n_Y-1}$ (válido pues las poblaciones son normales)

Cuando H_0 es cierta ($H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = 1$), se tiene $X^* = \frac{S_X^2/S_Y^2}{1} = \frac{S_X^2}{S_Y^2} \sim F_{n_X-1, n_Y-1}$.

iii) Elegimos como **criterio de rechazo** el de la **región crítica**.⁴

Como es bilateral ($H_1 : \frac{\sigma_X^2}{\sigma_Y^2} \neq 1$) la región crítica es de la forma:

RC: $(-\infty, K_1) \cup (K_2, +\infty)$ siendo $P\left(X^* < K_1 / H_0 \text{ cierta}\right) = \frac{\alpha}{2}$ y $P\left(X^* > K_2 / H_0 \text{ cierta}\right) = \frac{\alpha}{2}$.

iv) **Calcular el valor de X^*** a partir de la muestra y **aplicar el criterio** de decisión:

Como $n_X=39$ y $n_Y=10$, buscamos K_1 y K_2 tales que:

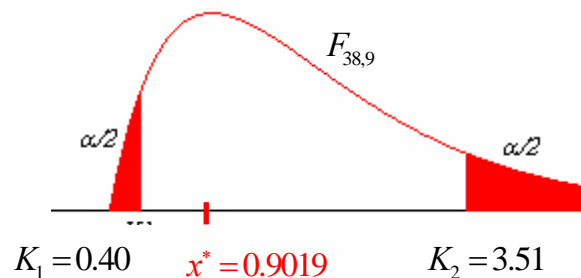
$$P(F_{38,9} < K_1) = 0.025 \text{ y } P(F_{38,9} > K_2) = 0.025.$$

y obtenemos $K_1 = 0,404701$, $K_2 = 3,51423$ (Statgraphics).

Tenemos que $x^* = \frac{s_X^2}{s_Y^2} = \frac{39.12}{43.38} = 0.9019$. Como:

$$K_1 = 0,404701 < x^* = 0.9019 < K_2 = 3,51423$$

No rechazamos H_0 , y podemos suponer que las varianzas son iguales.



Si hallamos el **intervalo de confianza** para el cociente de varianzas, hemos de despejar σ_X^2/σ_Y^2 de la expresión $K_1 < \frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} < K_2$, siendo **K_1 y K_2 los mismos que hemos obtenido antes**. Se obtiene un intervalo:

⁴ Se recomienda usar este criterio para la comparación de varianzas, sobre todo cuando la decisión que queremos tomar sea sólo si se pueden considerar iguales o no, y no necesitamos los matices que aporta el p-valor.

$$\left(\frac{S_X^2/S_Y^2}{K_2}, \frac{S_X^2/S_Y^2}{K_1} \right) \stackrel{\substack{\text{Sustituimos} \\ \text{con datos} \\ \text{de la muestra}}}{=} \left(\frac{0.9019}{3,51423}, \frac{0.9019}{0,404701} \right) = [0,256645; 2,22858]$$

Como $1 \in [0,256645; 2,22858]$, llegamos a la misma conclusión que con el contraste.

c) Comparamos las medias.

Para estudiar si las medias son o no iguales mediante un contraste:

i) **Planteamos** el contraste: $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases} \equiv \begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y \neq 0 \end{cases}$

ii) Elegimos el **estadístico** $X^* = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X+n_Y-2}$ con $S_p = \sqrt{\frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}}$

pues las varianzas se pueden suponer iguales y las poblaciones son normales.

Cuando H_0 es cierta ($H_0 : \mu_X - \mu_Y = 0$), se tiene: $X^* = \frac{(\bar{X} - \bar{Y}) - 0}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X+n_Y-2}$

iii) Elegimos como **criterio de rechazo** el **p-valor**.

Como es bilateral ($H_1 : \mu_X - \mu_Y \neq 0$) el p-valor se define:

$$p = 2 \cdot P\left(X^* > x^* / H_0 \text{ cierta}\right) \text{ ó } p = 2 \cdot P\left(X^* < x^* / H_0 \text{ cierta}\right)$$

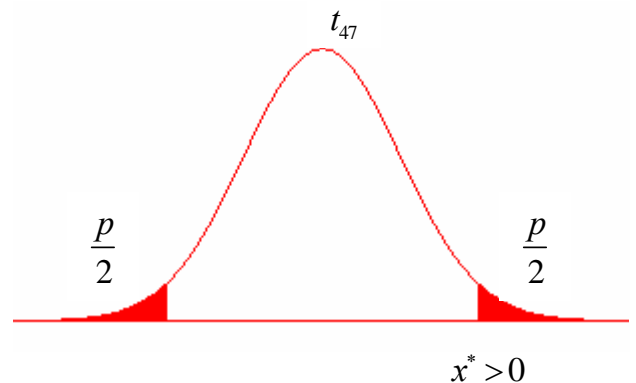
En este caso, por lo visto en ii), será $p = 2 \cdot P(t_{n_X+n_Y-2} > x^*)$ ó $p = 2 \cdot P(t_{n_X+n_Y-2} < x^*)$.

iv) **Calcular el valor de X^*** a partir de la muestra y **aplicar el criterio** de decisión:

Como $n_X=39$ y $n_Y=10$, será

$$p = 2 \cdot P(t_{47} > x^*) \text{ ó } p = 2 \cdot P(t_{47} < x^*).$$

$$\begin{aligned} \text{Calculamos } x^* &= \frac{(\bar{x} - \bar{y})}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} = \\ &= \frac{(179.67 - 163.4)}{\sqrt{\frac{(39-1)6.25^2 + (10-1)6.59^2}{39+10-2}} \sqrt{\frac{1}{39} + \frac{1}{10}}} = 7.26 \end{aligned}$$



Como $x^* = 7.26 > 0$, y el p-valor es $p = 2 \cdot P(t_{47} > 7.26) \stackrel{\text{Statgraphics}}{=} 2 \cdot 1.65 \cdot 10^{-9} \approx 0$

Por tanto, como $p \approx 0 < 0.05 = \alpha$, rechazamos H_0 , y concluimos que hay evidencia estadística de que las medias son diferentes.

Cuando hicimos el intervalo, obtuvimos **(11.76, 20.77)**, que no contiene a 0, por lo que **no podemos considerar iguales las medias**.

En el tema 7 nos preguntábamos:

- o ¿Se puede afirmar que la diferencia entre las estaturas medias es de menos de 20 cm?

Con un 95% de confianza, no se puede garantizar que la diferencia sea de menos de 20 cm, pues el posible valor de dicha diferencia está en el intervalo **(11.76, 20.77)**, que incluye valores mayores a 20, por lo que dicha diferencia podría ser mayor que 20.

¿Cómo se estudia eso mediante contraste de hipótesis?

De forma análoga a lo anterior, tendríamos que estudiar:

- Normalidad de las variables** (ya hecho)
- Igualdad o no de las varianzas** (ya hecho)
- Comparación de medias:**

- Planteamos** el contraste:
$$\begin{cases} H_0 : \mu_X - \mu_Y \geq 20 \\ H_1 : \mu_X - \mu_Y < 20 \end{cases}$$

- Elegimos el **estadístico**
$$X^* = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X + n_Y - 2} \quad \text{con} \quad S_p = \sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}}$$

pues las varianzas se pueden suponer iguales y las poblaciones son normales.

Cuando H_0 es cierta ($H_0 : \mu_X - \mu_Y = 20$)⁵, se tiene:
$$X^* = \frac{(\bar{X} - \bar{Y}) - 20}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X + n_Y - 2}$$

- Elegimos como **criterio de rechazo** el **p-valor**.

Como $H_1 : \mu_X - \mu_Y < 20$ el p-valor se define
$$p = P\left(X^* < x^* / H_0 \text{ cierta}\right).$$

En este caso, por lo visto en ii), será
$$p = P\left(t_{n_X + n_Y - 2} < x^*\right).$$

- Calcular el valor de X^*** a partir de la muestra y **aplicar el criterio** de decisión:

Como $n_X = 39$ y $n_Y = 10$, será
$$p = P\left(t_{47} < x^*\right).$$

Tenemos que
$$x^* = \frac{(\bar{x} - \bar{y}) - 20}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} = \frac{(179.67 - 163.4) - 20}{\sqrt{\frac{(39-1)6.25^2 + (10-1)6.59^2}{39+10-2}} \sqrt{\frac{1}{39} + \frac{1}{10}}} = -1.66$$

Por tanto:
$$p = P\left(t_{47} < -1.66\right) = 0.0517 \approx 0.05 = \alpha.$$

⁵ Consideramos H_0 como hipótesis simple para que X^* esté claramente definido.

De forma estricta, $p > \alpha$, y no rechazamos H_0 . Sin embargo, vemos que $p \approx \alpha$, por lo que estamos justo en límite entre la región de rechazo o de aceptación de H_0 . En estos casos, ($0.01 < p < 0.3$) la decisión depende del analista, que tendrá que tener en cuenta los riesgos de cada decisión o podrá plantearse tomar nuevas muestras con un tamaño mayor para ganar en precisión.

8.6.- Contrastes NO paramétricos.

Este apartado cierra lo que vamos a ver en este curso relativo a la **inferencia estadística**, cuyo objetivo es estudiar es alguna característica de una población determinada y **encontrar un modelo probabilístico** para ella **a partir de unos datos concretos**.

Hasta ahora hemos visto cómo obtener **información sobre los parámetros** (θ) del mismo, mediante:

- estimación puntual (tema 6),
- estimación por intervalo (tema 7) y
- contraste de hipótesis paramétricos (tema 8).

Ahora vamos a ver algunos métodos de *contraste* para obtener, a partir de los datos de la muestra, **información acerca del propio modelo**. De hecho, ya los hemos utilizado en los temas 4 y 5 cuando ajustábamos una variable estadística a un modelo teórico, y en los temas 7 y 8 para verificar la hipótesis de normalidad.

El método utilizado es análogo al de los contrastes paramétricos:

i) Planteamiento

La hipótesis es relativa al modelo de distribución que sigue una v.a. X y, por tanto, el planteamiento es de la forma: $\begin{cases} H_0 : X \sim F_0 \\ H_1 : X \neq F_0 \end{cases}$, donde F_0 es el modelo de distribución que queremos comprobar que sigue X .

ii) Elección de un estadístico D como medida de discrepancia y comprobación de las condiciones de aplicación.

Los estadísticos utilizados serán diferentes en función del tipo de variable a estudiar. Los veremos en cada caso. Los denotaremos por D .

iii) Elección del criterio de rechazo

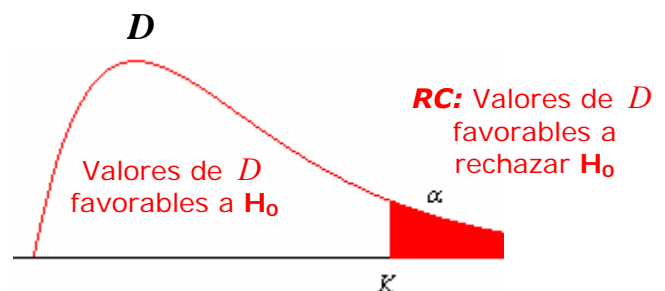
Son análogos a los vistos para contrastes paramétricos:

Región crítica:

Fijado α , buscamos un **valor $K \in \mathbb{R}$, tal que**

$$P\left(D > K / H_0 \text{ cierta}\right) = \alpha$$

La región crítica **RC** es el intervalo $(K, +\infty)$



Si denotamos por d el valor de D obtenido a partir de la muestra:

- **aceptaremos H_0** si $d \notin RC$ ($d < K$).
- **rechazaremos H_0** si $d \in RC$ ($d > K$).

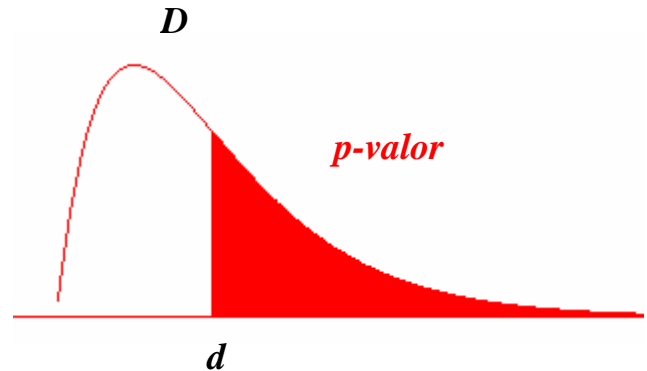
Criterio del P-valor:

En este caso:

$$p = P(D > d / H_0 \text{ cierta})$$

Siendo d el valor de D obtenido a partir de la muestra.

Se decide según los criterios habituales del p-valor.



iv) Calcular el valor de D a partir de la muestra y aplicar el criterio de decisión.

Hasta ahora, cuando hemos ajustado datos a un modelo de probabilidad, lo hemos hecho utilizando Statgraphics, "saltándonos" el paso ii) y dejando que el programa aplicara automáticamente el contraste en el paso iv).

Por ejemplo, en el tema 4, para estudiar qué modelo de probabilidad podría seguir la variable X : *Número de llamadas diarias que se hacen por teléfono móvil*, dimos los siguientes pasos:

i) Planteamiento

A partir de la definición de la variable (número de "éxitos" en un "intervalo"), la representación gráfica de los datos y los valores de la media y la varianza, inferimos que puede tratarse de un modelo de Poisson. Y a partir de la inferencia paramétrica, suponemos que el valor del parámetro puede ser $\lambda=1.25$ (la media).

Por tanto, nuestra hipótesis es que $X \sim P(1.25)$.

Como contraste lo planteamos:
$$\begin{cases} H_0 : X \sim P(1.25) \\ H_1 : X \neq P(1.25) \end{cases}$$

En principio, aceptamos como bueno el modelo propuesto a no ser que los datos de la muestra indiquen lo contrario.

ii) Nos lo "saltamos".

iii) Elección del criterio de rechazo

Como criterio de rechazo proponíamos el del **p-valor**:

Contrastes no paramétricos con Statgraphics (Distribution Fitting):

p-valor > 0.3 para aceptar la hipótesis

(cuanto mayor sea, con más confianza se acepta el modelo propuesto)

Con ese criterio, nos interesa tener confianza en que el modelo es el adecuado y somos "más exigentes" con H_0 , aceptando el modelo sólo cuando la probabilidad de equivocarnos si lo rechazamos (p-valor) es suficientemente alto (0.3).

iv) Calcular el valor de D a partir de la muestra y **aplicar el criterio de decisión**.

Al utilizar la opción Distribution Fitting de Statgraphics, obteníamos:

Goodness-of-Fit Tests for Llamadas diarias

Fitted Poisson distribution:
mean = 1,25641

Chi-Square Test
Chi-Square = 0,555238 with 2 d.f. P-Value = 0,757586

El valor de D a partir de la muestra es Chi-Square = 0,555238 . A partir de él, Statgraphics calcula el p-valor y podemos **aplicar el criterio de decisión**:

Como el p-valor=0.757586 > 0.3, aceptamos la hipótesis de que X sigue una distribución de Poisson (no rechazamos H_0).

Lo que vamos a ver en este apartado es el paso ii), es decir **qué estadístico D** elegimos como medida de discrepancia, qué **condiciones** le pedimos para que sea válido y **qué distribución sigue** cuando H_0 es cierta, de forma que podamos hallar el p-valor o la región crítica.

Veamos a continuación los dos tipos de estadísticos más utilizados.

8.6.1- Test de la χ^2 .

Este método:

- se utiliza para v.a. **discretas** (aunque también es válido para v.a. continuas)
- requiere que $n \geq 30$ (siendo n el tamaño de la muestra)
- requiere una **agrupación** con determinadas condiciones

La **medida de discrepancia D** mide la **diferencia** entre:

- la **frecuencia esperada** (suponiendo cierta H_0) y
- la **frecuencia obtenida en la muestra** para los valores que puede tomar la v.a.

Se define:

$$D = \sum_{i=1}^k \frac{(n_i - n p_i)^2}{n p_i}, \text{ siendo:}$$

k : número de valores distintos que puede tomar la variable (o número de clases en que hemos agrupado los valores que puede tomar la variable).

p_i : probabilidad del valor (clase) i -ésimo suponiendo cierta $H_0 : X \sim F_0$.

n : tamaño de la muestra

$n p_i$: **frecuencia esperada** del valor (clase) i -ésimo suponiendo cierta $H_0 : X \sim F_0$

n_i : **frecuencia obtenida en la muestra** del valor (clase) i -ésimo.

Suponiendo que H_0 es cierta, y cuando se verifica que **$n \geq 30$ y $n p_i \geq 5$** , para todo i , se tiene que:

$$D \approx \chi_{k-r-1}^2$$

siendo **r** el número de parámetros estimados a partir de la muestra.

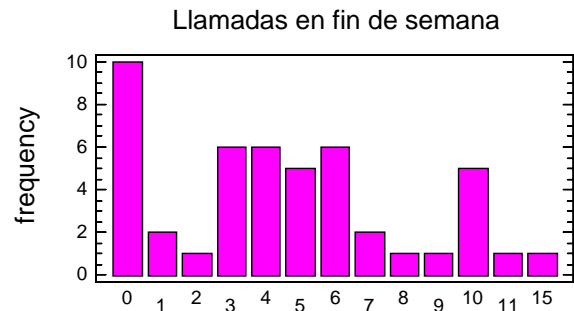
Ejemplo 1:

En el curso 2006/07 se preguntó a estudiantes de la EUI sobre el número de llamadas efectuadas en un fin de semana, y se quiere estudiar si la variable Y : *número de llamadas realizadas por el móvil en un fin de semana* sigue también una distribución de Poisson.

i) Planteamiento

La definición de la variable (número de “éxitos-llamadas” en un “intervalo-fin de semana”) nos hace pensar que puede seguir un modelo de Poisson.

Sin embargo, la representación gráfica de las frecuencias, y el valor de la varianza (13.3) mucho mayor que el de la media (4.55), nos hace dudar de que la hipótesis sea cierta.



Planteamos el contraste para salir de dudas.

El valor de la media muestral es 4.55319. Para facilitar la búsqueda en las tablas de las que disponemos, planteamos la hipótesis con $\lambda=4.6$:

$$\begin{cases} H_0 : X \sim P(4.6) \\ H_1 : X \neq P(4.6) \end{cases}$$

ii) Elección de un estadístico D como medida de discrepancia y comprobación de las condiciones de aplicación.

Como la variable es **discreta** y $n \geq 30$, el caso es propicio para utilizar el test de la Chi-cuadrado.

iii) Elección del criterio de rechazo

Fijamos $\alpha = 0.25$ y estudiaremos los dos criterios de rechazo: región crítica y p-valor.

(Como la gráfica y los valores de la varianza nos hacen desconfiar de la hipótesis, elegimos un nivel de significación más “exigente” para aceptar que sigue un modelo de Poisson.)

iv) Calcular el valor de D a partir de la muestra y **aplicar el criterio de decisión**.

A partir de los datos de la muestra, construimos la tabla de frecuencias (n_i) y de frecuencias esperadas (np_i).

Para hallar las frecuencias esperadas, hemos de obtener los valores p_i (probabilidad de cada valor o clase suponiendo H_0 cierta). Como $H_0: X \sim P(\lambda)$, para calcular dichas probabilidades hemos de tener un valor para λ .

Para ello, estimamos $\lambda \approx \bar{X}$ (que es el estimador tanto por el método de los momentos como por máxima verosimilitud).

En este caso $\bar{X} = 4.55319$ y para facilitar la búsqueda en las tablas de las que disponemos, consideramos $\lambda = 4.6$ ⁽⁶⁾.

Las frecuencias esperadas se calculan como $n \cdot p_i = 47 \cdot p_i$ ($n=47$ es el tamaño de la muestra, que puede obtenerse sumando todas las frecuencias).

Tenemos la siguiente **tabla**, en donde tenemos que incluir **todos los valores** que la variable puede tomar, aunque no los haya tomado en la muestra ⁽⁷⁾:

Valores	ni: Frecuencia	pi: Probabilidad Hocierta	npi: Frecuencia esperada
0	10	0,0101	0,4747
1	2	0,0462	2,1714
2	1	0,1063	4,9961
3	6	0,1631	7,6657
4	6	0,1875	8,8125
5	5	0,1726	8,1122
6	6	0,1237	5,8139
7	2	0,0849	3,9903
8	1	0,0518	2,4346
9	1	0,0285	1,3395
10	5	0,0143	0,6721
11	1	0,0065	0,3055
12	0	0,0045	0,2115
13	0	0,0007	0,0329
14	0	0,0002	0,0094
15	1	0,0001	0,0047
n=	47	1	47

En **rojo** hemos señalado los valores cuya **frecuencia esperada (np_i)** es **menor que 5**, y hemos de construir una nueva tabla en donde agrupemos los valores de forma que se cumpla la condición:

$$np_i \geq 5, \text{ para todo } i,$$

que es necesaria para asegurar que el estadístico **D** sigue una distribución Chi-Cuadrado.

Agrupamos adecuadamente los datos y obtenemos una nueva tabla. En la última columna, en azul, hemos calculado el valor del estadístico $D = \sum_{i=1}^k \frac{(n_i - n p_i)^2}{n p_i}$ y obtenemos **d=6.6599**.

Ahora, ya se verifican las condiciones en las que $D \approx \chi_{k-r-1}^2$. En este caso:

- o **k=6**, pues es el número de clases en las que finalmente agrupamos
- o **r=1**, pues hemos obtenido de la muestra una estimación del parámetro λ (**un** parámetro).

Por tanto, si H_0 es cierta, se tiene que $D \approx \chi_4^2$.

⁶ Otra forma de calcular p_i es utilizar la fórmula de la probabilidad de una distribución de Poisson: $p_i = P(X = x_i) = e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$.

⁷ En este caso, la probabilidad de los valores >15 es prácticamente nula, por lo que no los tenemos en cuenta.

Clases	Frecuencia	Probabilidad	Frecuencia esperada	di
[0,3)	13	0,1626	7,6422	3,7563
3	6	0,1631	7,6657	0,3619
4	6	0,1875	8,8125	0,8976
5	5	0,1726	8,1122	1,1940
6	6	0,1237	5,8139	0,0060
[7,inf)	11	0,1915	9,0005	0,4442
			d=	6,6599

▪ **Criterio región crítica:**

Como $\alpha=0.25$, buscamos un valor $K \in \mathbb{R}$, tal que

$$P\left(D > K / H_0 \text{ cierta}\right) = P\left(\chi_4^2 > K\right) = 0.25 \Rightarrow K = 5.3853$$

TABLAS

Como $d=6.6599 > K=5.3853$, rechazamos H_0 .

• **Criterio del p-valor:**

Calculamos el p-valor, según su definición:

$$p = P\left(D > d / H_0 \text{ cierta}\right) = P\left(\chi_4^2 > 6.6599\right) \stackrel{\text{Statgraphics}}{=} 0.156$$

Como $p=0.156 < 0.25$, rechazamos H_0 .

Conclusión: No puede considerarse que el número de llamadas por el móvil en un fin de semana siga una distribución de Poisson con $\lambda=4.6$.

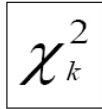
Comentarios al ejemplo:

A partir de las **tablas**, también podemos obtener información sobre el p-valor: buscamos en las tablas los valores entre los que esté el valor obtenido **d**, y calculamos la probabilidad que dejan a su **derecha** (las tablas dan el valor que dejan a su izquierda).

El p-valor estará comprendido entre dichas probabilidades.

Distribución CHI cuadrado acumulada

Valores x tales que $P(X_k \leq x) = p$



k es el nº de grados de libertad

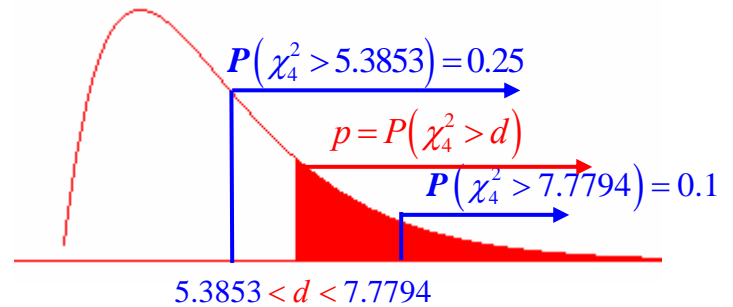
Valores con redondeo en el 5º decimal

k	p	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975	0,99	0,995
1		0,0000	0,0002	0,0010	0,0039	0,0158	0,1015	0,4549	1,3233	2,7055	3,8415	5,0239	6,6349	7,8794
2		0,0100	0,0201	0,0506	0,1026	0,2107	0,5754	1,3863	2,7726	4,6052	5,9915	7,3778	9,2104	10,5965
3		0,0717	0,1148	0,2158	0,3518	0,5844	1,2125	2,3660	4,1082	6,2514	7,8147	9,3484	11,3449	12,8381
4		0,2070	0,2971	0,4844	0,7107	1,0636	1,9226	3,3567	5,3853	7,7794	9,4877	11,1433	13,2767	14,8602

d=6.6599

En este caso: $0.1 < p < 0.25$.

En general, esa información es suficiente para tomar la decisión.



Observación:

Si la v.a. es continua tenemos muchas formas distintas de agrupar en intervalos, y se obtendría para cada caso un p-valor distinto. Por ello, el test de la χ^2 se utiliza para v.a. discretas y para las v.a. continuas se utilizará el que vemos a continuación.

8.6.2- Test de Kolmogorov-Smirnov.

Este método:

- es válido únicamente para v.a. **continuas**
- admite tamaños de la muestra pequeños.

En este caso la **medida de discrepancia D** mide la **diferencia** entre:

- la **función de distribución esperada** (F_0 , función de distribución del modelo a contrastar) y
- la **función de distribución empírica** (F_n , función de distribución obtenida a partir de la muestra).

y se define como

$$D_n = \max_x |F_n(x) - F_0(x)|.$$

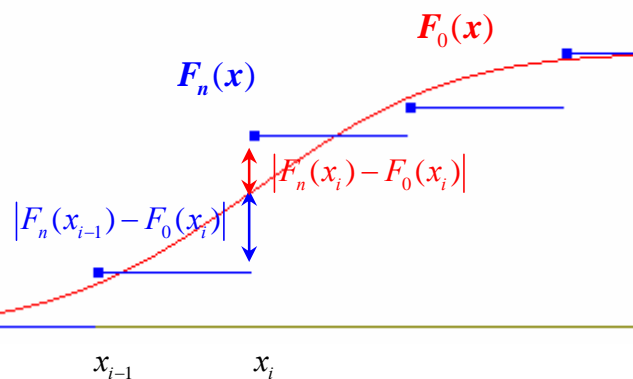
El estadístico D_n tiene una distribución conocida que está tabulada (ver tablas).

Para hallar el valor de D_n a partir de los datos de una muestra, seguiremos los siguientes pasos:

- (a) Se ordenan los valores de la muestra de menor a mayor: $x_1 \leq x_2 \leq \dots \leq x_n$
- (b) Se halla F_n , la **función de distribución empírica**, asignando a cada valor de la muestra le asigna la misma probabilidad $\frac{1}{n}$:

$$F_n(x) = \begin{cases} 0 & x < x_1 \\ \frac{1}{n} & x_1 \leq x < x_2 \\ \frac{2}{n} & x_2 \leq x < x_3 \\ \vdots & \\ \frac{n-1}{n} & x_{n-1} \leq x < x_n \\ 1 & x \geq x_n \end{cases}$$

(c) Se halla F_0 , la **función de distribución esperada**, que es la función de distribución del modelo a contrastar.



(d) Se calcula $D_n =$

Para calcularlo, se tiene en cuenta que al ser F_n una función escalonada, sólo hay probabilidad en los puntos de salto, por lo que la diferencia máxima se encuentra en los extremos de los escalones:

$$D_n = \max_i D_n(x_i), \text{ siendo}$$

$$D_n(x_i) = \max \left\{ |F_n(x_{i-1}) - F_0(x_i)|, |F_n(x_i) - F_0(x_i)| \right\} = \max \left\{ \left| \frac{i-1}{n} - F_0(x_i) \right|, \left| \frac{i}{n} - F_0(x_i) \right| \right\}.$$

Ejemplo:

Se han recogido datos sobre el tiempo que tardan los estudiantes en llegar a la EUI y se quiere contrastar si puede suponerse que dicha variable sigue un modelo exponencial, de media 60 minutos.

i) Planteamiento

$$\begin{cases} H_0 : X \sim \text{exp}(1/60) \\ H_1 : X \neq \text{exp}(1/60) \end{cases}$$

ii) Elección de un estadístico D como medida de discrepancia y comprobación de las condiciones de aplicación.

Como el modelo que se quiere contrastar es **continuo**, utilizamos el test de **Kolmogorov-Smirnof (K-S)**, y su estadístico:

$$D_n = \max_x |F_n(x) - F_0(x)|.$$

iii) Elección del criterio de rechazo

Fijamos $\alpha = 0.2$ y estudiaremos los dos criterios de rechazo: región crítica y p-valor.

iv) Calcular el valor de D a partir de la muestra y aplicar el criterio de decisión.

Seguimos los siguientes pasos:

- (a) Se ordenan los valores de la muestra de menor a mayor: $x_1 \leq x_2 \leq \dots \leq x_n$

En este ejemplo consideramos un amuestra de sólo 7 datos (para facilitar la comprensión del método): [45, 30, 70, 105, 90, 60, 15].

Ordenados: [15, 30, 45, 60, 70, 90, 105].

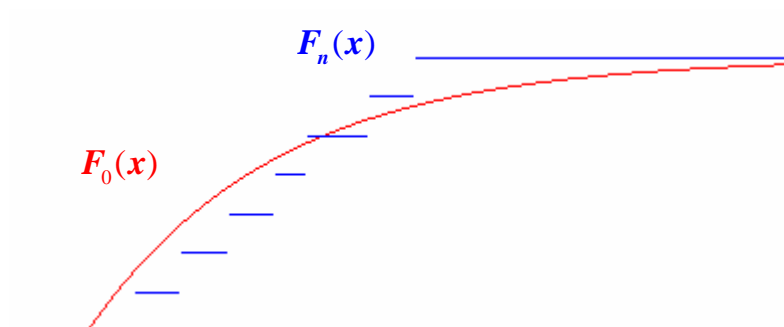
- (b) Se halla F_n , la **función de distribución empírica** asignando a cada valor de la muestra le asigna la misma probabilidad $\frac{1}{n} = \frac{1}{7}$:

$$F_n(x) = \begin{cases} 0 & x < 15 \\ \frac{1}{7} & 15 \leq x < 30 \\ \frac{2}{7} & 30 \leq x < 45 \\ \frac{3}{7} & 45 \leq x < 60 \\ \frac{4}{7} & 60 \leq x < 70 \\ \frac{5}{7} & 70 \leq x < 90 \\ \frac{6}{7} & 90 \leq x < 105 \\ 1 & x \geq 105 \end{cases}$$

- (c) Se halla F_0 , la **función de distribución esperada**, que en este caso es la función de distribución de la $\text{exp}(1/60)$.

La densidad de la $\text{exp}(1/60)$ es $f(x) = (1/60)e^{-x/60}$, si $x > 0$, y la función de distribución:

$$F_0(x) = \int_{-\infty}^x f(x)dx = \begin{cases} \int_{-\infty}^x 0 dx & \text{si } x < 0 \\ \int_{-\infty}^0 0 dx + \int_0^x (1/60)e^{-x/60} dx & \text{si } x \geq 0 \end{cases} = \begin{cases} 0 & \text{si } x < 0 \\ 1 - e^{-x/60} & \text{si } x \geq 0 \end{cases}$$



(d) Se calcula el valor del estadístico $D_n = \max_i D_n(x_i)$.

Para ello:

$$D_n(x_i) = \max \left\{ \left| F_n(x_{i-1}) - F_0(x_i) \right|, \left| F_n(x_i) - F_0(x_i) \right| \right\} = \max \left\{ \left| \frac{i-1}{n} - F_0(x_i) \right|, \left| \frac{i}{n} - F_0(x_i) \right| \right\}.$$

Para el cálculo podemos apoyarnos en la siguiente tabla:

x_i	$F_n(x_i)$	$F_0(x_i) = 1 - e^{-x_i/60}$	$\left \frac{i-1}{n} - F_0(x_i) \right $	$\left \frac{i}{n} - F_0(x_i) \right $	$d_n(x_i)$
15	$\frac{1}{7}$	0.2211	$ 0 - 0.2211 = \mathbf{0.2211}$	$ \frac{1}{7} - 0.2211 = 0.0782$	0.2211
30	$\frac{2}{7}$	0.3934	$ \frac{1}{7} - 0.3934 = \mathbf{0.2505}$	$ \frac{2}{7} - 0.3934 = 0.1076$	0.2505
45	$\frac{3}{7}$	0.5276	... = 0.2418	... = 0.09902	0.2418
60	$\frac{4}{7}$	0.6321	... = 0.2035	... = 0.06067	0.2035
70	$\frac{5}{7}$	0.6885	... = 0.1170	... = 0.02578	0.1170
90	$\frac{6}{7}$	0.7768	... = 0.06251	... = 0.08034	0.08034
105	1	0.8262	$ \frac{6}{7} - 0.8262 = 0.03094$	$ 1 - 0.8262 = \mathbf{0.1738}$	0.1738
$d = \max_i d_n(x_i) =$					0.2505

Para tomar la decisión:

• **Criterio región crítica:**

Como $\alpha = 0.2$, buscamos un valor $K \in \mathbb{R}$, tal que

$$P(D > K / H_0 \text{ cierta}) = P(D_7 > K) = 0.2 \stackrel{\text{TABLAS}}{\Leftrightarrow} K = 0.3815$$

Como $d = 0.2505 < K = 0.3815$, no rechazamos H_0 .

• **Criterio del p-valor:**

Contraste de Kolmogorov-Smirnov

$$K - S$$

n	Valores x tales que $P(D \leq x) = p$						
	1-p	0,8	0,85	0,9	0,95	0,98	0,99
1		0,9000	0,9250	0,9500	0,9750	0,9900	0,9950
2		0,6838	0,7260	0,7764	0,8419	0,9000	0,9293
3		0,5648	0,5970	0,6860	0,7076	0,7846	0,8290
4		0,4927	0,5250	0,5652	0,6239	0,6889	0,7342
5		0,4470	0,4740	0,5095	0,5633	0,6272	0,6685
6		0,4104	0,4360	0,4680	0,5193	0,5774	0,6166
7		0,3815	0,4050	0,4361	0,4834	0,5384	0,5758

$d = 0.2505$

Calculamos el p-valor, según su definición: $p = P\left(D > d / H_0 \text{ cierta}\right) = P(D_n > 0.2505)$

En la tabla vemos que $1-p < 0.8$, es decir $p > 0.2$, es decir $p > \alpha = 0.2$ y **no rechazamos H_0** .

Conclusión: Como no rechazamos H_0 , podemos suponer que la variable puede seguir un modelo exponencial.

Observación. ¿Cómo hace Statgraphics este test?

Si utilizamos Statgraphics (Distribution Fitting) y seleccionamos el modelo exponencial, el programa hace el test K-S realizando los cálculos con la estimación del parámetro a partir de los datos de la muestra. En este caso, como la media muestral es 59.2857, los realiza para $\beta = 1/59.2857$.

Los resultados son:

Estimated Kolmogorov statistic DPLUS = 0,170148
 Estimated Kolmogorov statistic DMINUS = 0,254255
 Estimated overall statistic DN = 0,254255 ($d = \max_i d_n(x_i)$)
 Approximate P-Value = 0,756075 (**p-valor**).

Puede verse que son muy similares a los obtenidos "a mano" con $\beta = 1/60$. La decisión sería la misma.

Ahora bien, ¿quiénes son los estadísticos DPLUS Y DMINUS?

El algoritmo que utiliza Statgraphics optimiza el número de operaciones que tiene que realizar teniendo en cuenta que siempre se va a verificar que

$$\frac{i-1}{n} - F_0(x_i) < \frac{i}{n} - F_0(x_i)$$

x_i	$F_n(x_i)$	$F_0(x_i) = 1 - e^{-x_i/60}$	$\frac{i-1}{n} - F_0(x_i)$	$\frac{i}{n} - F_0(x_i)$
15	$\frac{1}{7}$	0.2211	-0.2211	-0.0782
30	$\frac{2}{7}$	0.3934	-0.2505	-0.1076
45	$\frac{3}{7}$	0.5276	-0.2418	-0.09902
60	$\frac{4}{7}$	0.6321	-0.2035	-0.06067
70	$\frac{5}{7}$	0.6885	-0.1170	+0.02578
90	$\frac{6}{7}$	0.7768	-0.06251	+0.08034
105	1	0.8262	+0.03094	+0.1738
			DMINUS=0.2505	DPLUS=0.1738
			DN=0.2505	

Por tanto el valor negativo con mayor valor absoluto vendrá de una operación del tipo $\frac{i-1}{n} - F_0(x_i)$, y el valor positivo con mayor valor absoluto vendrá de una operación del tipo $\frac{i}{n} - F_0(x_i)$.

A la hora de buscar el máximo, buscará sólo entre los valores negativos de la columna $\frac{i-1}{n} - F_0(x_i)$ (y obtiene DMINUS, en valor absoluto) y entre los valores positivos de la columna $\frac{i}{n} - F_0(x_i)$ (y obtiene DPLUS, también en valor absoluto). El máximo de los dos será DN.

De esta forma, se ahorra bastantes comparaciones.