

T.6 CONTRASTES DE HIPÓTESIS .

6.1 Contraste estadístico de hipótesis: conceptos básicos. Hipótesis nula y alternativa. Significación, potencia y p-valor

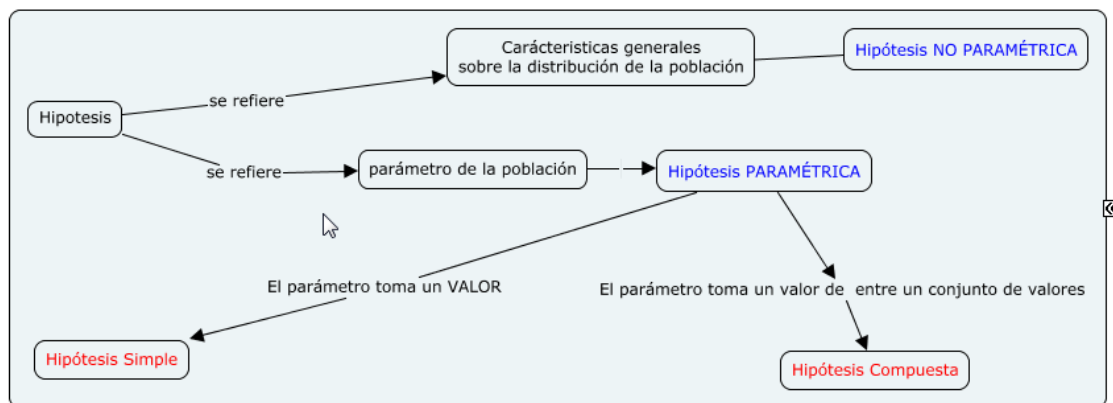
6.2. Principales contrastes paramétricos

6.3. Principales contrastes no paramétricos. Bondad de un ajuste

6.4. Significatividad estadística y significatividad práctica

6.1. Contraste estadístico de hipótesis: conceptos básicos. Hipótesis nula y alternativa. Significación, potencia y p-valor

El problema del contraste de hipótesis consiste, básicamente, en comprobar cotejar, decidir, en definitiva, sobre la veracidad de una hipótesis prefijada previamente como “supuestamente” cierta. En términos estadísticos, la o las hipótesis que formulamos serán hipótesis sobre la población; bien afectando a algún parámetro de ésta, lo que da origen a los **contrastos paramétricos**, o bien, a otras características de las mismas que no sean estrictamente sus parámetros, lo que origina los llamados **contrastos "no" paramétricos**. Aunque para introducir los conceptos básicos y el procedimiento general, “pensaremos” por defecto en contrastes paramétricos esta introducción puede considerarse común a ambos tipos de contrastes.



La solución estadística del problema de contrastación o contraste se basará en los datos muestrales y la base estadística (probabilística) de la que arrancará el contraste va a ser la distribución de algún estadístico muestral.

Supongamos que deseamos hacer un contraste acerca de un parámetro θ , de la población. Para llevarlo a cabo consideraremos la distribución de algún estadístico muestral que de alguna manera se corresponda y se relacione (en su distribución) con el parámetro θ ; designemos en general a este estadístico como T . Si con los datos muestrales obtenemos un valor concreto para T tal que pertenezca a una *determinada región* del campo de variación de T optaremos por **no rechazar** la hipótesis y en caso contrario por **rechazarla**. *Obviamente la clave del problema será delimitar la región del campo de variación de T que consideraremos como zona de aceptación de la hipótesis (y su complementaria de rechazo)*. Esto se resolverá por un criterio probabilístico a partir de la distribución muestral de T .

Esquema básico del contraste de hipótesis según el **criterio de decisión**:

- 1.- La hipótesis que se somete a prueba (inicialmente supuesta como aceptable) se llama hipótesis nula H_0 .
- 2.- Se trabaja con un cierto estadístico T (cuya distribución se conoce)

3.- La decisión se toma en función del valor que tome ese estadístico con los datos muestrales disponibles.

- 3.a. Si $T \in R_0$ se optará por NO RECHAZAR la hipótesis nula H_0 . A la región R_0 se la conoce como región de aceptación.
- 3.b Si $T \in R_1$ se optará por RECHAZAR la hipótesis nula H_0 . A la región R_1 se la conoce como región crítica o de rechazo.

En este esquema hay que puntualizar algunas cuestiones adicionales.

1.- El criterio de decisión debe ser exhaustivo: Como la hipótesis nula debe rechazarse o no rechazarse (aceptarse) las regiones de R_0 y R_1 deben ser complementarias y sin zonas comunes (constituir una partición del recorrido del estadístico T)

2.- Si se opta por rechazar la hipótesis nula se debe aceptar una **hipótesis alternativa** H_1 , como más asumible. Esta hipótesis alternativa jugará un gran papel en el diseño del contraste. Y, además, hay que considerar que será esta hipótesis alternativa la que acabará “constituyendo” el “nuevo conocimiento que queramos establecer” (Lo veremos un poco más adelante)

Una situación práctica para entender lo que queremos decir podría ser la siguiente: Sospechamos que la estatura media de los españoles “ya no es la que solía ser hace algún tiempo” (por ejemplo, 170 cm). La estatura media de los españoles es 170 cm sería la hipótesis nula.

Presumiblemente sería razonable trabajar con un estadístico como la media de estaturas de una muestra aleatoria.

Si esta media muestral toma un valor que pertenece a cierto rango (R_1) rechazaremos H_0 , pero si pertenece a su complementario (R_0) no rechazaremos H_0

Es evidente que la especificación de R_0 y R_1 no podrá ser la misma si rechazar que la media sea 170 cm (H_0) implica aceptar como hipótesis alternativa que la media sea mayor, que si la alternativa es que sea menor o que si la alternativa es que sea simplemente diferente. Un valor muestral de 195 cm, por ejemplo, podría hacer que nos decantáramos por rechazar H_0 en favor de la alternativa “la estatura media es mayor”, (Y también frente a la alternativa “es diferente”) pero **no debería** empujarnos a rechazar H_0 si ello conlleva como alternativa “la estatura media es menor”.

Llegados a este punto debe quedar claro que el problema del contraste estadístico de hipótesis consiste en establecer ese “criterio de decisión del que hemos hablado. En definitiva establecer las dos zonas pertinentes (de aceptación y de rechazo). La cuestión será hacerlo de forma que minimicemos de alguna manera la probabilidad de equivocarnos en nuestra decisión.

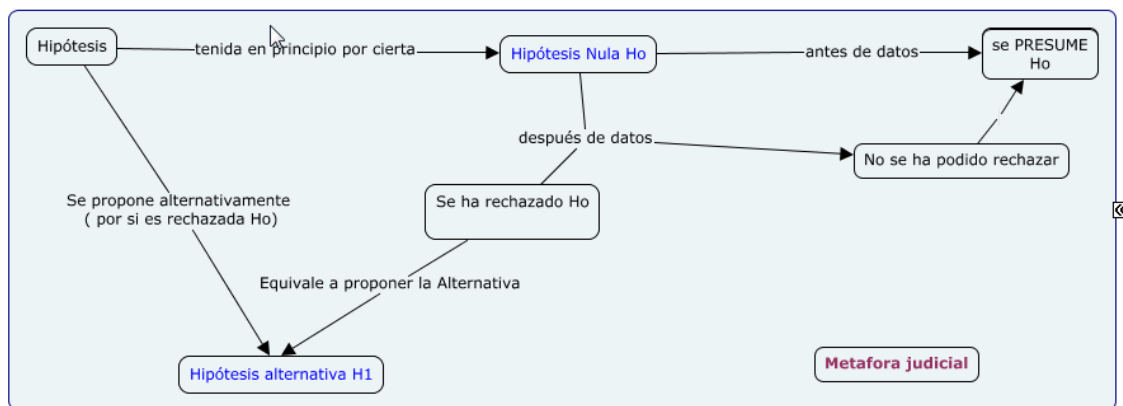
Antes de plantear **la matriz de decisión** sus **riesgos** (posibles errores) y de vislumbrar la estrategia a seguir, consideremos algunas definiciones con precisión.

Hipótesis nula (H_0): Es la hipótesis que se somete a prueba. Antes del contraste se ha dado por supuesta provisionalmente. Después de incorporar la información muestral quizá se siga manteniendo (no se rechaza) o se acabe rechazando. En la metáfora “judicial” (muy adecuada para entender la situación), la hipótesis nula juega el papel que en un juicio juega la “inocencia” del encausado. El principio de presunción de inocencia establece la inocencia del acusado “mientras no se demuestre lo contrario”.

La carga de la prueba corre a cuenta del “ministerio fiscal” que deberá probar “más allá de toda duda razonable” la culpabilidad.

Hipótesis alternativa (H_1): Es la hipótesis que quedará establecida (alternativamente) si se llega a rechazar la hipótesis nula. En la metáfora judicial, la hipótesis alternativa juega el papel de la culpabilidad, que deberá establecerse con “hechos probados”.

La idea es: en principio se supone (H_0) y sólo si los datos “tozudamente” se empeñan en hacer altísimamente inverosímil esta hipótesis nula, la descartaremos y estableceremos la alternativa. Con este modo de actuación, al igual que ocurre en los procesos judiciales, no tendrá las mismas garantías de certidumbre la aceptación de una u otra hipótesis. La culpabilidad (y la hipótesis alternativa) se establece más allá de “toda duda razonable”. La inocencia (y la hipótesis nula) se establece porque no se ha podido probar la culpabilidad. (En el caso de las hipótesis, añadiríamos que provisionalmente)



Estadístico del contraste, T , será el estadístico o función de los datos muestrales que utilizaremos para llevar a cabo el contraste y decidir si aceptamos o no la hipótesis nula.

Criterio de decisión: El establecimiento de una partición en el campo de variación del estadístico del contraste que permita tomar la decisión de aceptar o no H_0 , según el valor que tome T .

Región de aceptación (R_0): la región del campo de variación del estadístico que determinará la aceptación de H_0 si T pertenece a ella.

Región de rechazo o región crítica (R_1): la región del campo de variación del estadístico que determinará el rechazo de H_0 y, la consecuente, aceptación de la alternativa H_1 .

La clave de todo contraste de hipótesis es establecer el criterio de decisión que, como en la metáfora judicial, debe hacerse de forma que se minimicen las probabilidades de cometer una decisión errónea.

En el contraste de hipótesis y en un proceso judicial hay dos posibles decisiones erróneas:

- Rechazar H_0 siendo cierta (Condenar a un inocente)
- Aceptar H_0 siendo falsa (Absolver a un culpable)

Ambas deben minimizarse pero es evidente que en, ambos casos (contraste y juicio), establecer un procedimiento que se obstine en reducir casi a cero un tipo de error traerá como consecuencia aumentar el riesgo del otro error.

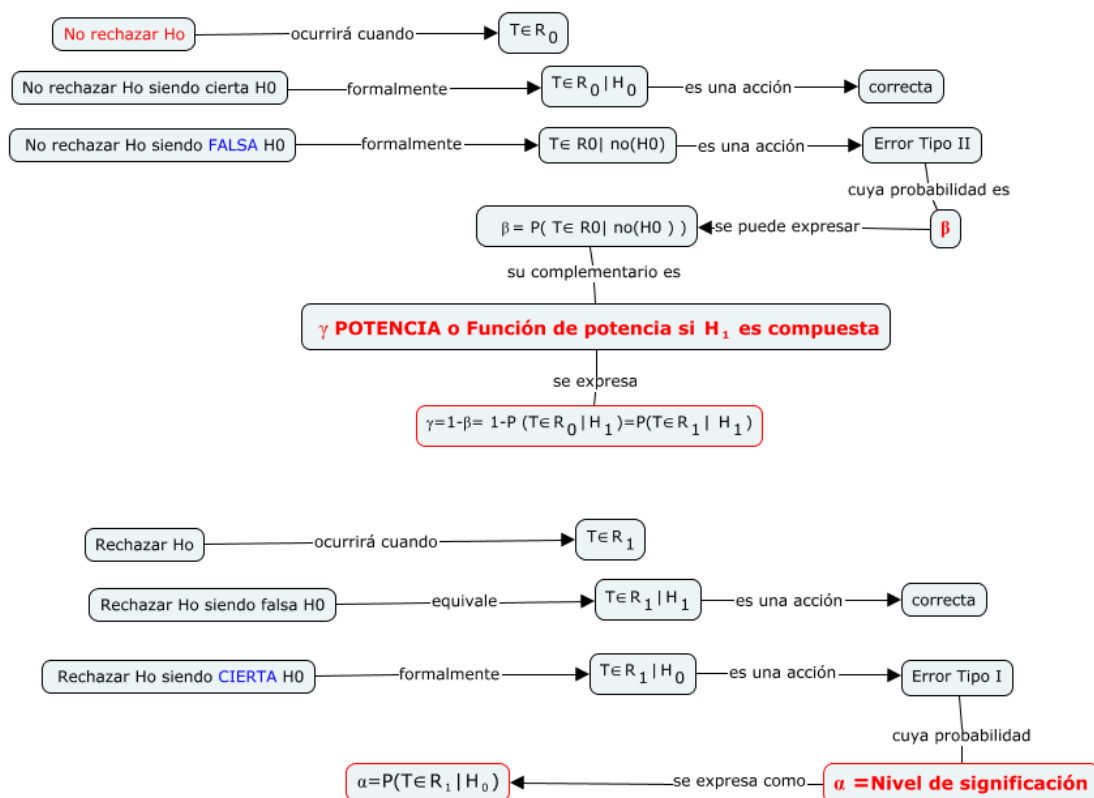
	Acción	Aceptar H_0	Rechazar H_0
Realidad			
H_0 es cierta		Acción correcta	Error tipo I
H_0 es falsa		Error tipo II	Acción correcta

Llamaremos **nivel de significación**, α , a la probabilidad de cometer en un contraste de hipótesis un error de tipo I.

Designaremos por β a la probabilidad de cometer un error tipo II. Y llamaremos **potencia del contraste**, γ , al complementario de esta probabilidad.

El test o contraste ideal debería minimizar tanto α como β . Pero ambas cosas no pueden llevarse a cabo a su extremo. Hay que priorizar pero sin descuidar lo no priorizado. La opción será fijar una de las dos en un valor bajo o muy bajo (que podamos consentir) e intentar minimizar la otra.

El procedimiento opta por fijar el nivel de significación, α y optar un criterio de decisión que conlleve $P(\text{error Tipo I}) = \alpha$ y $P(\text{error tipo II})$ lo más pequeña posible, aunque, a menudo, no podamos precisarla.



La estrategia para obtener las regiones R_0 y R_1 (el criterio de decisión) para llevar a cabo un contraste de hipótesis de la hipótesis nula H_0 , con hipótesis alternativa H_1 , y con un nivel de significación prefijado α

Formalmente : $\begin{cases} H_0 : \dots \\ H_1 : \dots \end{cases}$ con nivel de significación α será la siguiente:

Escoger R_0 (R_1 será su región complementaria) de forma que :

- 1) $P(T \in R_1 | H_0) = \alpha$ es decir $P(T \in R_0 | H_0) = 1 - \alpha$
- 2) Y se minimice β , la probabilidad de error tipo dos .Esto es que la potencia sea lo más alta posible .

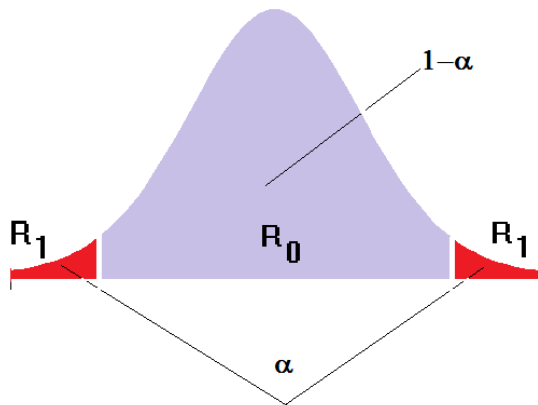
Aunque si la hipótesis alternativa es simple (sólo propone un único valor) podemos calcular la potencia exacta para las distintas regiones R_1 posibles, lo más habitual es que la hipótesis alternativa sea compuesta del tipo $\theta \neq \theta_0$, $\theta > \theta_0$ ó $\theta < \theta_0$

Si la hipótesis alternativa es compuesta, la segunda condición supondrá, en la práctica, que la amplitud de la región R_0 sea lo más pequeña posible (R_1 lo más grande posible) al tiempo que resulte **coherente** con la hipótesis alternativa considerada en el contraste.

Esto es: dos zonas de rechazo si la alternativa es bilateral y una zona (a la derecha o la izquierda) si la alternativa es unilateral (por la derecha o por la izquierda)

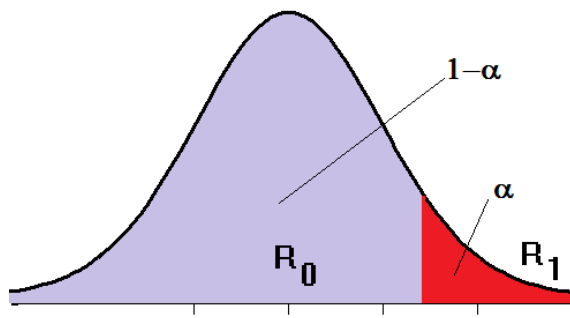
1: Contraste bilateral o de dos colas

$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$ con nivel de significación α se debe cumplir 1) con R_1 lo más amplia posible y por ambos lados :



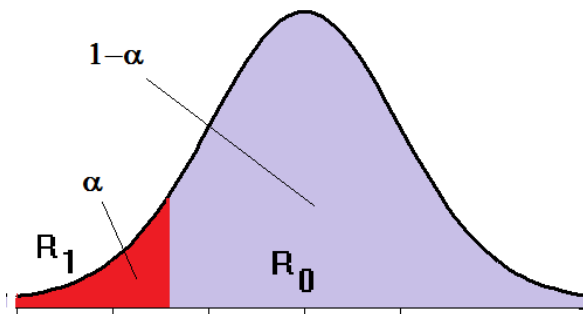
2.: Contraste unilateral (una cola) por la derecha

$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$ con nivel de significación α se debe cumplir 1) con R_1 lo más amplia posible por el lado derecho :

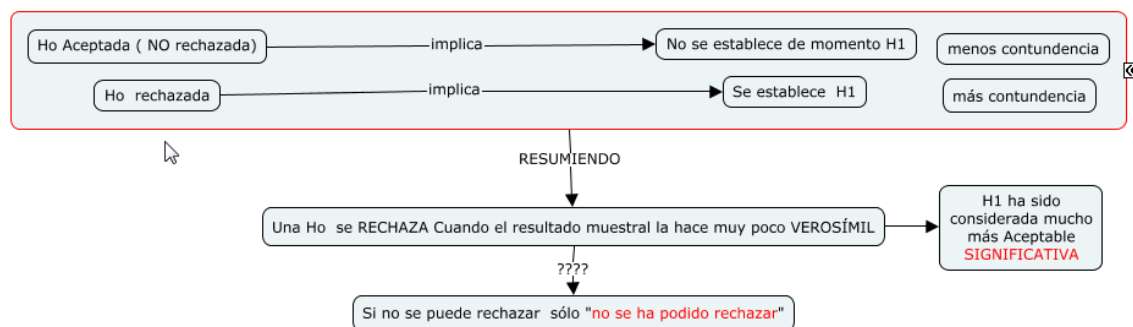


3.: Contraste unilateral (una cola) por la izquierda

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$
 con nivel de significación α se debe cumplir 1) con R_1 lo más amplia posible por el lado izquierdo :



Llegados aquí, aseguramos un procedimiento que permite optar por la hipótesis nula (provisionalmente) o por la hipótesis alternativa (**significativamente**) según el esquema:



6.2. Principales contrastes paramétricos

Para estudiar los principales contrastes paramétricos vamos a ir agrupándolos según el parámetro sobre el que se plantee la hipótesis , considerando, a la vez, los tres posibles distintos casos para la hipótesis alternativa (bilateral y unilateral por la izquierda y la derecha).

6.2.1. Contrastes para medias de poblaciones normales con varianza conocida

6.2.1.1 Contraste sobre la media de una población normal.(varianza conocida)

(Puede usarse `contraste.media {estadística}` o `z.test {TeachingDemos}`)

Contrastamos, con un nivel de significación α , la hipótesis nula de que la media de una población normal, μ , toma un determinado valor , μ_0 . Frente a la hipótesis alternativa de que toma (1) un valor distinto, (2) un valor inferior, o (3) un valor mayor que μ_0 . Para

realizar el contraste tomamos una MAS de tamaño n y consideramos la varianza poblacional conocida (n grande)

$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \begin{cases} \neq \mu_0 \\ < \mu_0 \\ > \mu_0 \end{cases} \end{array} \right\} \text{ con nivel de significación } \alpha \text{ a partir de un MAS de tamaño n } \text{Pob. } N(\mu, \sigma \text{ conocida})$$

$$\text{El estadístico del contraste será : } \bar{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1)$$

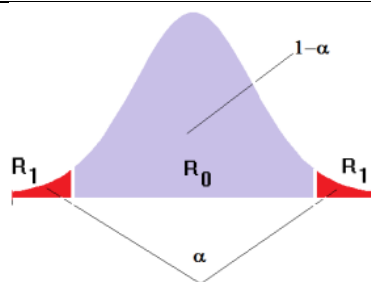
Para determinar las zonas de aceptación y de rechazo trabajaremos con la distribución del estadístico *supuesta la hipótesis nula cierta* (o condicionada a la hipótesis nula) , ya que R_0 y R_1 deben cumplir que :

$$P(T \in R_1 | H_0) = \alpha \text{ que en nuestro caso será: } P\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in R_1 | H_0\right) = \alpha$$

Adicionalmente para que el contraste sea lo más potente posible debe cumplir que R_1 sea lo más extensa posible (R_0 la más pequeña posible) .Esto nos conduce a que la zona de rechazo debe situarse en la (o las colas) Si toma una (y cuál) o las dos colas dependerá de la hipótesis alternativa.

De forma que el criterio de decisión será:

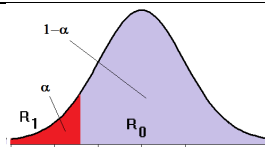
$$\left. \begin{array}{l} \text{Contraste con alternativa bilateral (dos colas)} \\ H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array} \right\} \text{ con n.s. } = \alpha$$



$$\left\{ \begin{array}{l} \text{si } \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in [-z_{\alpha/2}, z_{\alpha/2}] \Rightarrow \text{aceptaremos } H_0 \text{ (No la rechazaremos)} \\ \text{si } \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \notin [-z_{\alpha/2}, z_{\alpha/2}] \Rightarrow \text{rechazaremos } H_0 \text{ y aceptaremos } H_1 \text{ (} \mu \text{ es significativamente distinta de } \mu_0 \text{)} \end{array} \right.$$

Contraste con alternativa unilateral (cola de la izquierda)

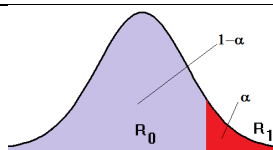
$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{array} \right\} \text{ con n.s.} = \alpha$$



$$\left\{ \begin{array}{l} \text{si } \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > -z_{\alpha} \Rightarrow \text{aceptaremos } H_0 \text{ (No la rechazaremos)} \\ \text{si } \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < -z_{\alpha} \Rightarrow \text{rechazaremos } H_0 \text{ y aceptaremos } H_1 \text{ (} \mu \text{ es significativamente menor que } \mu_0 \text{)} \end{array} \right.$$

Contraste con alternativa unilateral (cola de la derecha)

$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{array} \right\} \text{ con n.s.} = \alpha$$



$$\left\{ \begin{array}{l} \text{si } \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha} \Rightarrow \text{aceptaremos } H_0 \text{ (No la rechazaremos)} \\ \text{si } \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{\alpha} \Rightarrow \text{rechazaremos } H_0 \text{ y aceptaremos } H_1 \text{ (} \mu \text{ es significativamente mayor que } \mu_0 \text{)} \end{array} \right.$$

Automatización (paquete estadística): en la ayuda de la función `contraste.media {estadística}` encontramos su uso (usage):

```
contraste.media(x,
  variable = NULL,
  introducir = FALSE,
  var_pob = c("conocida","desconocida"),
  hipotesis_nula = NULL,
  tipo_contraste = c("bilateral","cola derecha","cola izquierda"),
  alfa = 0.05,
  grafico = FALSE)
```

Ejemplo1: Descargamos y cargamos el archivo de datos `alumnos.sav`.

Consideramos que las edades de estos alumnos constituyen una muestra (MAS de n=240 de una Pob.normal)

Contrastamos que la media del alumnado sea de 20 años (en la población) con el 5% de significación, frente a la alternativa de que es distinta.

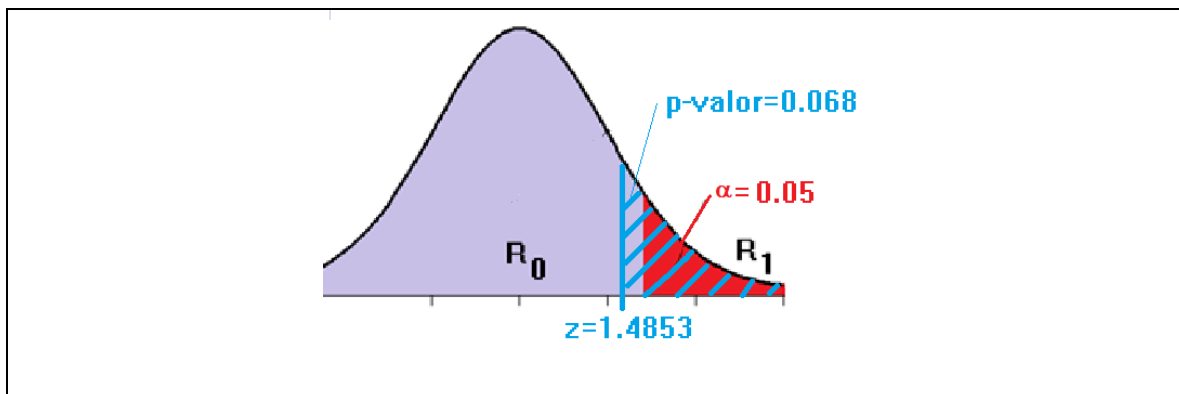
```
library(haven)
alumnos <-
```

- Cargamos base de datos.
- Consideramos la muestra de

<pre>read_sav("http://www.uv.es/mlejarza/azar/alumnos.sav") muestra<- alumnos\$edad library(estadistica) varianza(muestra) contraste.media(muestra,hipotesis_nula = 20)</pre>	las edades de 240 alumnos elegidos al azar (población Normal) •Cargamos el paquete estadística •La varianza de la muestra la tomaremos como varianza de la población supuesta conocida •Por defecto: var. conocida, bilateral, signif=0.05
<pre>> contraste.media(muestra,hipotesis_nula = 20) Introducir el valor de la varianza poblacional: 13.4479 [1] "Se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo [-1.96 , 1.96]" [1] "El valor del estadístico de prueba (o valor experimental) no se encuentra dentro de la región de aceptación" \$Estadístico Hipótesis nula estadístico de prueba p-valor 1 20 10.75495 5.61651e-27 \$`Intervalo de la media muestral (supuesta H0 cierta)` limite_inferior limite_superior [1,] 19.53604 20.46396</pre>	

Ejemplo2: Utilizamos los datos ya calculados : sustituimos los valores de media y varianza muestral(cambiamos intencionadamente el tamaño de la muestra a n=100).Y planteamos el contraste de que la media sea de 22 años frente a la alternativa de que sea mayor de 22 (mismo nivel de significación)

<pre>#necesitamos conocer la media y la varianza de la muestra para introducirlas media(muestra) ;varianza(muestra) #la opción introducir =TRUE para introducir los datos, hipt.nula (mu0) ,cola de la derecha, significación por defecto contraste.media(introducir = T ,hipotesis_nula = 22,tipo_contraste="cola derecha")</pre>	Contrastamos ahora (sustituyendo los datos por sus valores ya calculados) la hipótesis de que la edad media sea 22 años frente a la alternativa de que sea mayor. Como el dato muestral es 22.54 , la pregunta es ¿hay base suficiente para decir que la media es significativamente mayor a los 22 años? El p-valor es 0.068 > que alfa (0.05) Por lo tanto aceptamos H_0 La media NO es significativamente mayor de 22 años
<pre>> media(muestra);varianza(muestra) media_x 1 22.54583 x 1 13.4479 > contraste.media(introducir = T ,hipotesis_nula = 22,tipo_contraste="cola derecha") Introducir el tamaño de la muestra: 100 Introducir el valor de la media muestral: 22.54583 Introducir el valor de la varianza poblacional: 13.4479 [1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo]-Inf , 1.6449]" [1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación" \$Estadístico Hipótesis nula estadístico de prueba p-valor 1 22 1.488436 0.06831797 \$`Intervalo de la media muestral (supuesta H0 cierta)` limite_inferior limite_superior [1,] -Inf 22.60321</pre>	



Puede comprobarse que si utilizamos el auténtico valor de $n=240$ el resultado sería el contrario: Concluiríamos que la edad es significativamente mayor de 22 años.

Alternativamente podemos usar la función `z.test` del paquete `{TeachingDemos}`. Para este caso (triple, dependiendo de la hipótesis alternativa) la sintaxis en R será:

```
z.test(x,mu=0,stdev,alternative=c("two.sided", "less", "greater"), sd=stdev, n=length(x), conf.level= 0.95)
```

Donde:

`x` = el vector "muestra" o el valor "media de la muestra"

`mu`= es el valor que según la hipótesis nula toma la media de la población (μ_0): por defecto es 0

`stdev`= el valor conocido de la varianza de la población.

`alternative`= es la dirección de la hipótesis alternativa, por defecto es bilateral (`two.sided`)

`sd`= otra forma de escribir el argumento `stdev`

`n`= tamaño de la muestra (si `x` ha sido la media muestral)

`conf.level` = nivel de confianza = 1-nivel de significación.

El valor de la función es un objeto de la clase **htest** que contiene los resultados:

```
data: los datos
z = valor del estadístico z , n = tamaño de la muestra,
Std. Dev. = desviación típica de la población,
Std. Dev. of the sample mean = Desviación típica de la media muestral
p-value = p-valor (cola/0 dos colas) correspondiente al estadístico supue
sta la hipótesis nula cierta
alternative hypothesis: hipótesis alternativa que se ha considerado
% percent confidence interval: intervalo de confianza para 1-alfa
sample estimates:
mean of muestra= estimación: media de la muestra
```

Los ejemplos anterior quedarían

Ejemplo1:

```
library(haven)
alumnos <-
read_sav("http://www.uv.es/mlejarza/azar/alumnos.sav")
muestra<- alumnos$edad
z<-z.test(muestra, mu=20,sd=sd(muestra), conf.level=0.95)
z
```

One Sample z-test

```
data: muestra
z = 10.733, n = 240.00000, Std. Dev. = 3.67
480, Std. Dev. of the sample mean =
0.23721, p-value < 2.2e-16
```

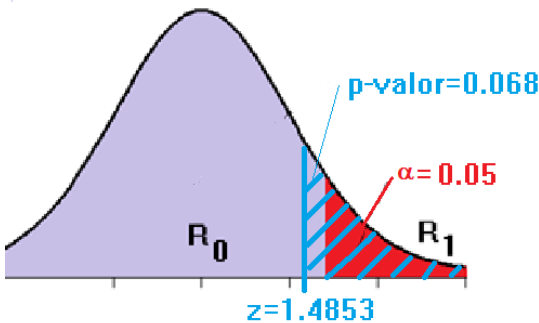
Descargamos y cargamos el archivo de datos `alumnos.sav`. Consideramos que las edades de estos alumnos constituyen una muestra (MAS de $n=240$ de una Pob.normal) Contrastamos que la media del alumnado sea de 20 años (en la población) con el 5% de significación, frente a la alternativa de que es distinta.

Análisis de resultados:

El p -valor =0 -> inferior a α

<pre>alternative hypothesis: true mean is not equal to 20 95 percent confidence interval: 22.08092 23.01075 sample estimates: mean of muestra 22.54583</pre>	El intervalo de conf. NO incluye el valor de la hipótesis ($\mu=20$) Conclusión: rechazamos la hipótesis nula (20 años) concluyendo que la edad media del alumnado es SIGNIFICATIVAMENTE distinta de 20 años.
--	---

Ejemplo2: Utilizamos los datos ya calculados : sustituimos los valores de media y des.tip.muestral(cambiamos intencionadamente el tamaño de la muestra a $n=100$).Y planteamos el contraste de que la media sea de 22 años frente a la alternativa de que sea mayor de 22 (mismo nivel de significación)

<pre>z2<-z.test(22.54583,mu=22, sd=3.67480, n=100,alternative="greater",conf.level=0.95) z2</pre> One Sample z-test <pre>data: 22.54583 z = 1.4853, n = 100.00000, Std. Dev. = 3.67480, Std. Dev. of the sample mean = 0.36748, p-value = 0.06873 alternative hypothesis: true mean is greater than 22 95 percent confidence interval: 21.94138 Inf sample estimates: mean of 22.54583 22.54583</pre>	Contrastamos ahora (sustituyendo los datos por sus valores ya calculados) la hipótesis de que la edad media sea 22 años frente a la alternativa de que sea mayor. Como el dato muestral es 22.54 , la pregunta es ¿hay base suficiente para decir que la media es significativamente mayor a los 22 años? El p-valor es 0.068 > que alfa (0.05) Por lo tanto aceptamos H_0 La media NO es significativamente mayor de 22 años 
--	--

6.2.1.2 Contraste sobre la diferencia de dos media de dos poblaciones normales. (varianzas conocidas)

Contrastamos, con un nivel de significación α , la hipótesis nula de que la diferencia de dos medias de dos poblaciones normal, $\mu_x - \mu_y$, toma un determinado valor, μ_0 . Frente a la hipótesis alternativa de que toma (1) un valor distinto, (2) un valor inferior, o (3) un valor mayor que μ_0 . Para realizar el contraste tomamos dos MAS de tamaño n_x y n_y y consideramos las varianzas poblacionales conocidas (n_x y n_y grandes)

$$\left. \begin{array}{l} H_0 : \mu_x - \mu_y = \mu_0 \\ H_1 : \mu_x - \mu_y \begin{cases} \neq \mu_0 \\ < \mu_0 \\ > \mu_0 \end{cases} \end{array} \right\} \text{ con nivel de significación } \alpha$$

a partir de dos MAS de tamaño n_x y n_y de dos poblaciones $N(\mu_x, \sigma_x); N(\mu_y, \sigma_y)$

El estadístico del contraste será :

$$\bar{X} - \bar{Y} \rightarrow N(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}) \Rightarrow \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \rightarrow N(0,1)$$

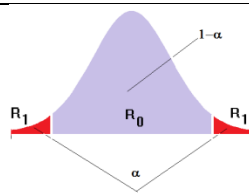
Para determinar las zonas de aceptación y de rechazo trabajaremos con la distribución del estadístico supuesta la hipótesis nula cierta (o condicionada a la hipótesis nula) , ya que R_0 y R_1 deben cumplir que :

$$P(T \in R_1 | H_0) = \alpha \text{ que en nuestro caso será: } P\left(\frac{\bar{X} - \bar{Y} - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \in R_1 | H_0\right) = \alpha$$

Adicionalmente para que el contraste sea lo más potente posible debe cumplir que R_1 sea lo más extensa posible (R_0 la más pequeña posible) .Esto nos conduce a que la zona de rechazo debe situarse en la (o las colas) Si toma una (y cuál) o las dos colas dependerá de la hipótesis alternativa.

Contraste con alternativa bilateral (dos colas)

$$\left. \begin{array}{l} H_0 : \mu_x - \mu_y = \mu_0 \\ H_1 : \mu_x - \mu_y \neq \mu_0 \end{array} \right\} \text{ con n.s. } = \alpha$$

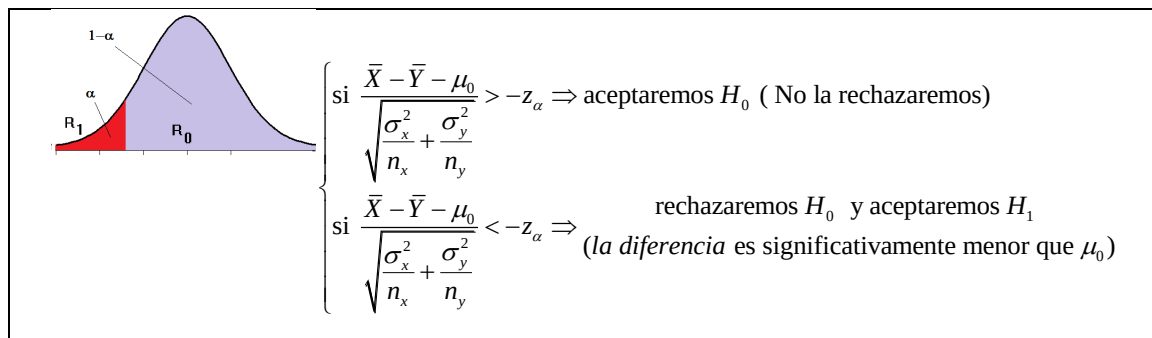


$$\left\{ \begin{array}{l} \text{si } \frac{\bar{X} - \bar{Y} - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \in [-z_{\alpha/2}, z_{\alpha/2}] \Rightarrow \text{aceptaremos } H_0 \text{ (No la rechazaremos)} \\ \text{si } \frac{\bar{X} - \bar{Y} - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \notin [-z_{\alpha/2}, z_{\alpha/2}] \Rightarrow \text{rechazaremos } H_0 \text{ y aceptaremos} \end{array} \right.$$

H_1 ($(\mu_x - \mu_y)$ es significativamente distinta de μ_0)

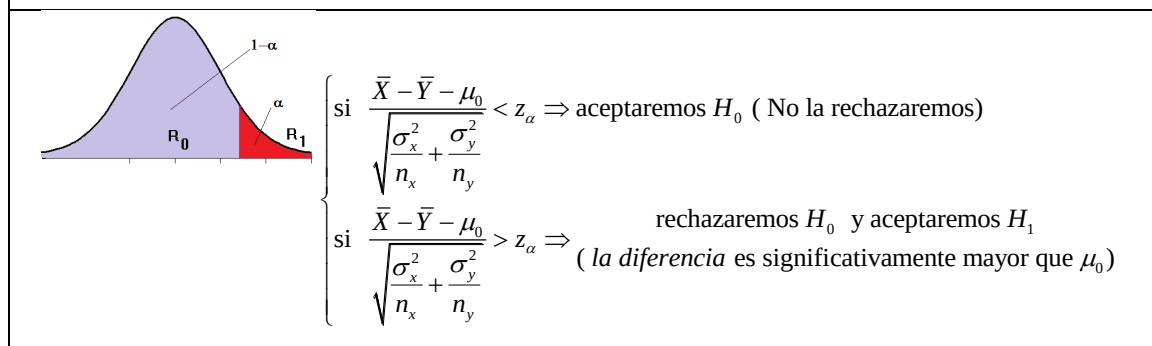
Contraste con alternativa unilateral (cola de la izquierda)

$$\left. \begin{array}{l} H_0 : \mu_x - \mu_y = \mu_0 \\ H_1 : \mu_x - \mu_y < \mu_0 \end{array} \right\} \text{ con n.s. } = \alpha$$



Contraste con alternativa unilateral (cola de la derecha)

$$\left. \begin{array}{l} H_0 : \mu_x - \mu_y = \mu_0 \\ H_1 : \mu_x - \mu_y > \mu_0 \end{array} \right\} \text{ con n.s.} = \alpha$$



Automatización (paquete estadística): en la ayuda de la función `contraste.diferencia.medias` {estadística} encontramos su uso (usage):

Usage

```
contraste.diferencia.medias(x,
  variable = NULL,
  introducir = FALSE,
  var_pob = c("conocida","desconocida"),
  iguales = FALSE,
  hipotesis_nula = 0,
  tipo_contraste = c("bilateral","cola derecha","cola izquierda"),
  alfa = 0.05,
  grafico = FALSE)
```

En los casos que nos ocupan en este apartado, el argumento `var_pob` será conocida, que es la opción por defecto. El argumento `iguales` hace referencia a si las varianzas de ambas poblaciones son iguales o no y es de aplicación en el caso de que las varianzas sean desconocidas (más adelante) el resto de argumentos son similares a los vistos en el epígrafe anterior.

Ejemplo 3

Trabajamos con las edades de los chicos y de las chicas consideradas como muestras aleatorias. Obtenemos sus medias y varianzas

a) Contrastaremos la hipótesis de que sus edades medias sean iguales $\mu=0$ frente a la alternativa bilateral. (Significación=5 %)

b) Contrataremos que la diferencia sea de un año(a favor de los chicos) frente a la alternativa mayor que 1 año.

c) Contrastaremos que son iguales, frente la alternativa bilateral pero con significación del 1 % .

a)

<pre>chicos<-subset(alumnos,sexo==0) chicas<-subset(alumnos,sexo==1) m1<-chicos\$edad m2<-chicas\$edad nrow(chicos);nrow(chicas) media1=media(m1) media2=media(m2) v1=varianza(m1) v2=varianza(m2) contraste.diferencia.medias(introducir=T)</pre>	<pre>> nrow(chicos);nrow(chicas) [1] 116 [1] 124 > media1;media2 media_x 1 23.14655 media_x 1 21.98387 > v1#varianza de la edad de los chicos x 1 14.14232 > v2#varianza de x 1 12.1449</pre>
<pre>> contraste.diferencia.medias(introducir= T) Introducir el tamaño de la muestra 1: 116 Introducir el valor de la media muestral 1: 124 Introducir el tamaño de la muestra 2: 23.14655 Introducir el valor de la media muestral 2: 21.98387 Introducir el valor de la varianza poblacional 1: 14.14232 Introducir el valor de la varianza poblacional 2: 12.1449 [1] "Se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo [-1.96 , 1.96]" [1] "El valor del estadístico de prueba (o valor experimental) no se encuentra dentro de la región de aceptación" \$Estadístico Hipótesis nula estadístico de prueba p-valor 1 0 126.8665 0 \$Intervalo de la diferencia de medias muestrales (supuesta H0 cierta)` limite inferior limite superior [1,] -1.576079 1.576079</pre>	

b)

<pre>contraste.diferencia.medias(introducir=T, hipotesis_nula=1,tipo_contraste="cola derecha")</pre>	Modificamos la H_0 y el tipo de contraste (H_1)
<pre>> contraste.diferencia.medias(introducir=T, hipotesis_nula=1,tipo_contraste="cola derecha") Introducir el tamaño de la muestra 1: 116 Introducir el valor de la media muestral 1: 23.14655 Introducir el tamaño de la muestra 2: 124 Introducir el valor de la media muestral 2: 21.98387 Introducir el valor de la varianza poblacional 1: 14.14232 Introducir el valor de la varianza poblacional 2: 12.1449 [1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo]-Inf , 1.6449]" [1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación" \$Estadístico Hipótesis nula estadístico de prueba p-valor 1 1 0.3469459 0.364316 \$Intervalo de la diferencia de medias muestrales (supuesta H0 cierta)` limite inferior limite superior [1,] -Inf 1.77128</pre>	

c)

<pre>contraste.diferencia.medias(introducir= T,alfa=0.01)</pre>
<pre>> contraste.diferencia.medias(introducir= T,alfa=0.01) Introducir el tamaño de la muestra 1: 116 Introducir el valor de la media muestral 1: 23.14655 Introducir el tamaño de la muestra 2: 124 Introducir el valor de la media muestral 2: 21.98387 Introducir el valor de la varianza poblacional 1: 14.14232 Introducir el valor de la varianza poblacional 2: 12.1449 [1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo [-2.5758 , 2.5758]" [1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación" \$Estadístico Hipótesis nula estadístico de prueba p-valor 1 0 2.479635 0.01315169 \$Intervalo de la diferencia de medias muestrales (supuesta H0 cierta)` limite inferior limite superior [1,] -1.207771 1.207771</pre>

6.2.2. Contrastes para medias de poblaciones normales con varianza desconocida. Procedimiento t.test.

La situación habitual es que no conozcamos la(s) varianza poblacional(es) si estamos realizando inferencias sobre las medias. Esta situación es la general y la del epígrafe anterior es una simplificación académica o una aproximación cuando la(s) muestra(s) es(son) grande(s).

6.2.2.1 Contraste sobre la media de una población normal. (varianza desconocida)

Contrastamos, con un nivel de significación α , la hipótesis nula de que la media de una población normal, μ , toma un determinado valor, μ_0 . Frente a la hipótesis alternativa de que toma (1) un valor distinto, (2) un valor inferior, o (3) un valor mayor que μ_0 . Para realizar el contraste tomamos una MAS de tamaño n

$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \begin{cases} \neq \mu_0 \\ < \mu_0 \\ > \mu_0 \end{cases} \end{array} \right\} \text{ con nivel de significación } \alpha \text{ a partir de un MAS de tamaño } n \text{ Pob. } N(\mu, \sigma)$$

El estadístico del contraste será :

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n-1}}} \rightarrow t_{n-1}$$

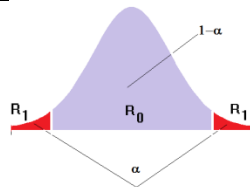
Para determinar las zonas de aceptación y de rechazo trabajaremos con la distribución del estadístico supuesta la hipótesis nula cierta (o condicionada a la hipótesis nula), ya que R_0 y R_1 deben cumplir que :

$$P(T \in R_1 | H_0) = \alpha \text{ que en nuestro caso será: } P\left(\frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n-1}}} \in R_1 | H_0\right) = \alpha$$

Adicionalmente para que el contraste sea lo más potente posible debe cumplir que R_1 sea lo más extensa posible (R_0 la más pequeña posible). Esto nos conduce a que la zona de rechazo debe situarse en la (o las colas) Si toma una (y cuál) o las dos colas dependerá de la hipótesis alternativa.

De forma que el criterio de decisión será:

$$\left. \begin{array}{l} \text{Contraste con alternativa bilateral (dos colas)} \\ H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array} \right\} \text{ con n.s.} = \alpha$$

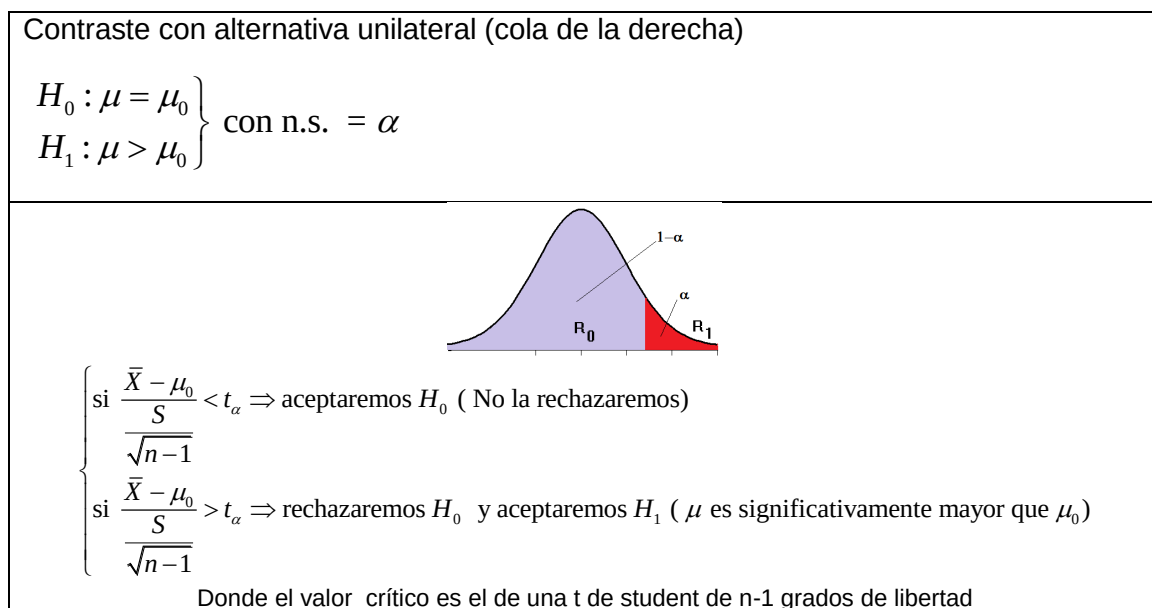
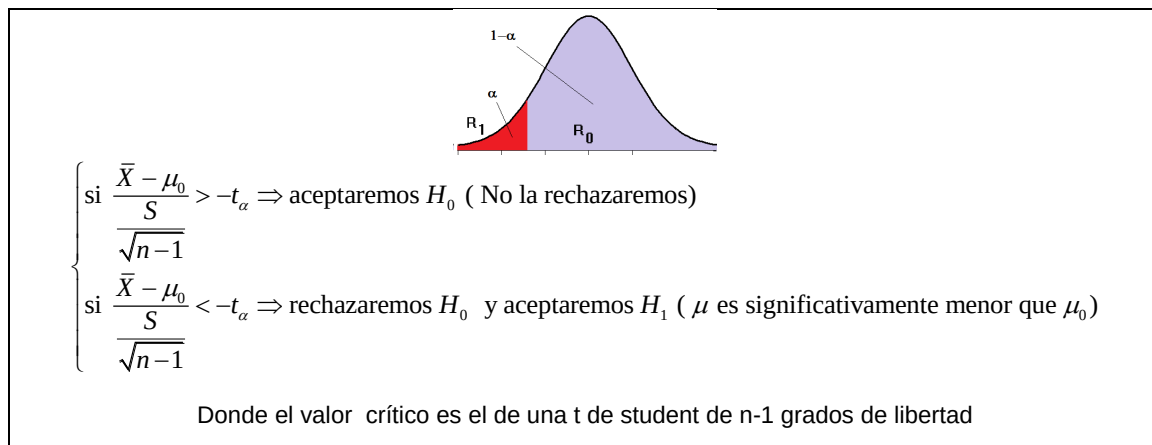


$$\left\{ \begin{array}{l} \text{si } \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n-1}}} \in [-t_{\alpha/2}, t_{\alpha/2}] \Rightarrow \text{aceptaremos } H_0 \text{ (No la rechazaremos)} \\ \text{si } \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n-1}}} \notin [-t_{\alpha/2}, t_{\alpha/2}] \Rightarrow \text{rechazaremos } H_0 \text{ y aceptaremos } H_1 \text{ (} \mu \text{ es significativamente distinta de } \mu_0 \text{)} \end{array} \right.$$

Donde los valores críticos son los de una t de student de $n-1$ grados de libertad

Contraste con alternativa unilateral (cola de la izquierda)

$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{array} \right\} \text{ con n.s.} = \alpha$$



Automatización (paquete estadística). Cambiando a la opción “desconocida” el argumento del atributo var_pob de la función contraste.media (ya vista más arriba) pueden llevarse a cabo contraste sobre la media como “**contraste t**”

Ejemplo 4.

Sobre la base de datos anterior, ahora hemos obtenido una muestra de tamaño 20 sobre la hacemos varios contrastes sobre la media de edad de los alumnos.Y contrastamos que la media de la edad de la población de alumnos sea de 20 años

```
set.seed(444)
muestra<-sample(alumnos$edad,20)
media(muestra)
varianza(muestra)
contraste.media(muestra,var_pob="desconocida",hipotesis_nula = 20,grafico=T)
```



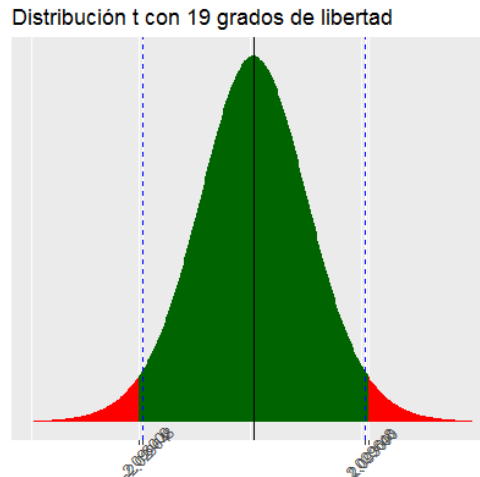
```

> contraste.media(muestra,var_pob="desconocida",hipotesis_nula = 20,grafico=T)
Selecciona el valor que quieres utilizar:
1. "Varianza muestral"
2. "Covarianza muestral"
1
[1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo [-2.093 , 2.093]"
[1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación"
$estadístico
  Hipótesis nula estadístico de prueba    p-valor
1              20              2.029648 0.05663467

$Intervalo de la media muestral (supuesta H0 cierta)
  limite_inferior limite_superior
[1,]           18.5563           21.4437

$graficos
$graficos[[1]]
$graficos[[2]]
NULL

```



Ejemplo 4.continuación.

Considerando esa misma muestra ¿podremos concluir con una confianza del 5 % que la media de edad es mayor de 20? ¿Y con una confianza del 1% ¿Y mayor de 21?

Contrastamos la hipótesis nula de que la media de edad es de 20 años, frente a la alternativa unilateral (mayor de 20 años)

```

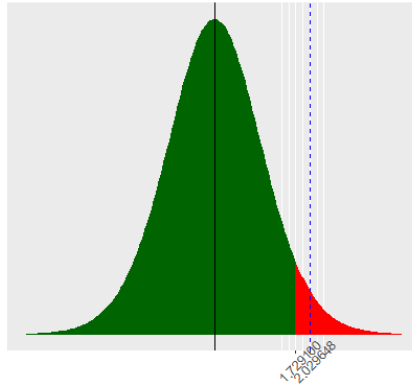
contraste.media(muestra,var_pob="desconocida",hipotesis_nula = 20,grafico=T,tipo_contraste="cola
derecha")
> contraste.media(muestra,var_pob="desconocida",hipotesis_nula = 20,grafico=T,tipo_contraste="cola derecha")
Selecciona el valor que quieres utilizar:
1. "Varianza muestral"
2. "Covarianza muestral"
1
[1] "Se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo ]-Inf , 1.7291]"
[1] "El valor del estadístico de prueba (o valor experimental) no se encuentra dentro de la región de aceptación"
$estadístico
  Hipótesis nula estadístico de prueba    p-valor
1              20              2.029648 0.02831733

$Intervalo de la media muestral (supuesta H0 cierta)
  limite_inferior limite_superior
[1,]           -Inf           21.19269

$graficos
$graficos[[1]]
$graficos[[2]]
NULL

```

Distribución t con 19 grados de libertad



Rechazaríamos la hipótesis nula en favor de la alternativa “ la edad es mayor que 20”
 $p\text{-valor} = 0.02831 < \alpha = 0.05$ ($t = 2.029 > t_{\alpha} = 1.729$)

#con una confianza del 1% , cambiaríamos alfa

`contraste.media(muestra,var_pob="desconocida",hipotesis_nula = 20,grafico=T,tipo_contraste="cola derecha",alfa=0.01)`

`> contraste.media(muestra,var_pob="desconocida",hipotesis_nula = 20,grafico=T,tipo_contraste="cola derecha",alfa=0.01)`

Selecciona el valor que quieres utilizar:

1. "Varianza muestral"
2. "Covarianza muestral"

1

[1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo $]-\infty, 2.5395]$ "

[1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación"

\$Estadistico

Hipótesis nula	estadístico de prueba	p-valor
1	20	2.029648 0.02831733

\$Intervalo de la media muestral (supuesta H0 cierta)`

limite_inferior	limite_superior
[1,]	-Inf 21.75168

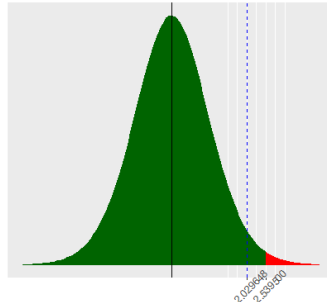
\$Graficos

\$Graficos[[1]]

\$Graficos[[2]]

NULL

Distribución t con 19 grados de libertad



La edad media es significativamente mayor de 20 años incluso con una significación del 1%

`contraste.media(muestra,var_pob="desconocida",hipotesis_nula = 21,grafico=T,tipo_contraste="cola derecha",alfa=0.05)`

```

> contraste.media(muestra,var_pob="desconocida",hipotesis_nula = 21,grafico=T,tipo_contraste="cola derecha",alfa=0.05)
Selecciona el valor que quieres utilizar:
1. "Varianza muestral"
2. "Covarianza muestral"
1
[1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo ]-Inf , 1.7291]"
[1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación"
$Estadístico
  Hipótesis nula estadístico de prueba   p-valor
1             21              0.5798993 0.2843981

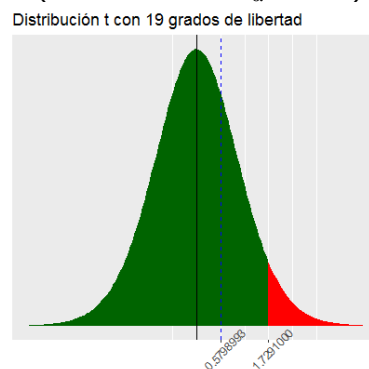
$Intervalo de la media muestral (supuesta H0 cierta)`
  limite_inferior limite_superior
[1,]           -Inf          22.19269

$Graficos
$Graficos[[1]]

$Graficos[[2]]
NULL

```

¿podríamos concluir que la edad es mayor que 21, con el 5% de significación?
 No. En este caso la hipótesis de que la edad media es igual 21 No puede rechazarse frente a la alternativa de que sea mayor :
 $p\text{-valor} = 0.2843981 > \alpha = 0.05$ ($t = 0.5798993 > t_{\alpha} = 1.729$)



Como no se ha podido rechazar la hipótesis nula $\mu = 21$ frente a la alternativa $\mu > 21$ con una significación del 5 % , tampoco se podrá a una significación más baja (del 1%)

Aprovechamos los resultados del ejemplo para aclarar dos cosas sobre las conclusiones que habitualmente se obtienen en los contrastes de hipótesis.

- 1.- La hipótesis aceptada sólo es concluyentemente ("significativamente") aceptada cuando juega el papel de **hipótesis alternativa**. La hipótesis nula si se acepta, en realidad, sólo podemos decir que "no ha sido rechazada". Una vez rechazada la hipótesis nula diremos que la hipótesis alternativa es "significativa"
- 2.- Cuanto menor sea el nivel de significación al que se rechaza la hipótesis nula más significativa es la hipótesis alternativa.

En el ejemplo anterior la hipótesis $\mu = 20$ ha sido rechazada frente a la alternativa $\mu \neq 20$, con un nivel de significación de 0.05, pero no se podría rechazar con el 0.01 (puede comprobarse: el p-valor era $0.02831 > 0.01$)

En cambio, la hipótesis $\mu = 20$ ha sido rechazada frente a la alternativa $\mu > 20$, tanto al 5% como al 1%.

Podríamos concluir que la media es MUY SIGNIFICATIVAMENTE ($\alpha = 0.01$) mayor que 20. Pero sólo SIGNIFICATIVAMENTE ($\alpha = 0.05$) distinta de 20

De forma muy similar se puede usar también la función `t.test` del paquete básico de R {stats}

Para este (triple) caso la sintaxis de la función `t.test` en R será:

```
t.test(x,alternative,mu,conf.level, ...)
```

Cuyos argumentos básicos son:

x= el vector formado por los datos muestrales

alternative = la dirección de la hipótesis alternativa puede ser "two.sided" (por defecto), "greater" (mayor que) o "less" (menor que). También se puede escribir "g" o "l".

mu= el valor de la media según la hipótesis nula .Por defecto es cero

conf.level= nivel de confianza = $1 - \alpha$ (complementario de la significación)

Existen otros argumentos adicionales para la realización de l contraste t sobre 2 muestras.

El valor de la función es un objeto de la clase **htest** que contiene los resultados:

data: los datos

t = valor del estadístico z , df = grados de libertad de la t (n-1),

p-value = p-valor (cola/0 dos colas) correspondiente al estadístico supuesta la hipótesis nula cierta

alternative hypothesis: hipótesis alternativa que se ha considerado

% percent confidence interval: intervalo de confianza para 1-alfa

sample estimates:

mean of x= estimación: media de la muestra

Ejemplo 4. Sobre la base de datos anterior, ahora hemos obtenido una muestra de tamaño 20 sobre la hacemos varios contrastes sobre la media de edad de los alumnos.

```
set.seed(444)
```

```
muestra<-sample(alumnos$edad,20)
```

```
t.test(muestra,mu=20)
```

```
> t.test(muestra,mu=20)
```

One Sample t-test

data: muestra

t = 2.0296, df = 19, p-value = 0.05663

alternative hypothesis: true

mean is not equal to 20

95 percent confidence

interval:

19.95628 22.84372

sample estimates:

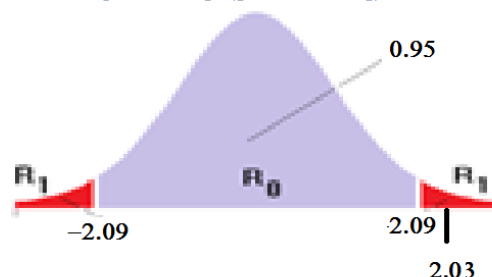
mean of x
21.4

Fijamos la semilla aleatoria

Tomamos una muestra de tamaño 20.

Contrastamos la hipótesis nula de que la media de edad es de 20 años, frente a la alternativa bilateral (por defecto) con un nivel de significación de 0.05 (por defecto)

No se rechazaría la hipótesis nula ya que el p-valor de $t=2.0296$ es inferior a $\alpha=0.05$
 $t=2.6285 \notin [-t_{\alpha/2}, t_{\alpha/2}]$ ($[-2.09, 2.09]$)



(Otra forma de comprobarlo el valor $\mu=20$ está en el I.C.)

Ejemplo 4. continuación. Considerando esa misma muestra ¿podremos concluir con una confianza del 5 % que la media de edad es mayor de 20? ¿Y con una confianza del 1% ¿Y mayor de 21?

```
t.test(muestra,mu=20,alternative="g")
```

```
t.alfa<-qt(0.05,19,lower.tail=FALSE)
```

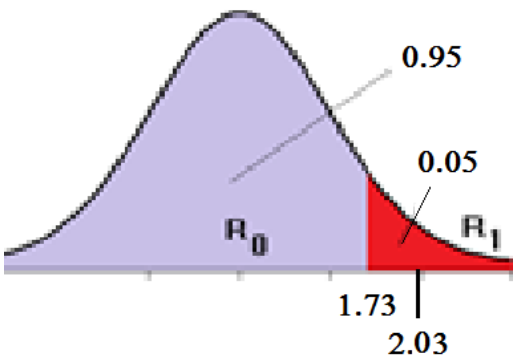
```
t.alfa
```

```
[1] 1.729133
```

One Sample t-test

data: muestra

Contrastamos la hipótesis nula de que la media de edad es de 20 años, frente a la alternativa unilateral (mayor de 20 años) con un nivel de significación de 0.05 (por defecto) Rechazaríamos la hipótesis nula en favor de la alternativa " la edad es mayor que 20"
p-valor=0.0082 < $\alpha=0.05$

$t = 2.0296$, $df = 19$, $p\text{-value} = 0.02832$ alternative hypothesis: true mean is greater than 20 95 percent confidence interval: 20.20729 Inf sample estimates: mean of x 21.4	$(t=2.0296 > t_{\alpha}=1.73)$  La edad media es significativamente mayor de 20 años (con un n.s. del 5%)
---	--

<pre>t.test(muestra,mu=20,alternative="g", conf.level = 0.99)</pre> One Sample t-test <pre>data: muestra t = 2.0296, df = 19, p-value = 0.02832 alternative hypothesis: true mean is greater than 20 99 percent confidence interval: 19.64833 Inf sample estimates: mean of x 21.4</pre>	Repetimos el contraste cambiando la significación al 0.01 En realidad no sería necesario hacerlo , ya que el p-valor asociado a los resultados ya lo conocemos y ya sabemos que es: $p\text{-valor}=0.02832 > \alpha=0.01$. De forma que no rechazamos la hipótesis nula La edad media no es “muy” significativamente mayor de 20 años con una significación del 1%) $(t=2.0296 < t_{\alpha}=2.539)$
---	--

<pre>t.test(muestra,mu=21,alternative="g")</pre> One Sample t-test <pre>data: muestra t = 1.3143, df = 19, p-value = 0.1022 alternative hypothesis: true mean is greater than 21 95 percent confidence interval: 20.68433 Inf sample estimates: mean of x 22</pre>	¿podríamos concluir que la edad es mayor que 21, con el 5% de significación? No. En este caso la hipótesis de que la edad media es igual 21 NO puede rechazarse frente a la alternativa de que sea mayor $p\text{-valor} = 0.1022 > \alpha=0.05$
---	--

6.2.2.2 Contraste sobre la diferencia de dos medias de dos poblaciones normales. (varianzas desconocidas)

En este sub-epígrafe desarrollamos el caso en el que las dos varianzas son desconocidas pero (supuestamente) iguales (homoscedasticidad). Este es el caso que

se basa en la distribución muestral t para la diferencia de medias que vimos en el tema 4 y que también dio origen al correspondiente IC en el tema 5. Sin embargo, el procedimiento t.test de R admite el caso heteroscedástico (varianzas distintas) e incluso el caso en que el muestreo haya sido “emparejado”.

Contrastamos, con un nivel de significación α , la hipótesis nula de que la diferencia de dos medias de dos poblaciones normal, $\mu_x - \mu_y$, toma un determinado valor, μ_0 . Frente a la hipótesis alternativa de que toma (1) un valor distinto, (2) un valor inferior, o (3) un valor mayor que μ_0 . Para realizar el contraste tomamos dos MAS de tamaño n_x y n_y y consideramos las varianzas poblacionales son iguales aunque desconocidas

$$\left. \begin{array}{l} H_0 : \mu_x - \mu_y = \mu_0 \\ H_1 : \mu_x - \mu_y \begin{cases} \neq \mu_0 \\ < \mu_0 \\ > \mu_0 \end{cases} \end{array} \right\} \text{ con nivel de significación } \alpha$$

a partir de dos MAS de tamaño n_x y n_y de dos poblaciones $N(\mu_x, \sigma_x); N(\mu_y, \sigma_y)$

y $\sigma_x = \sigma_y$.

El estadístico del contraste será :

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{n_x + n_y} \cdot \sqrt{n_x S_x^2 + n_y S_y^2}} \sqrt{n_x \cdot n_y} \cdot \sqrt{n_x + n_y - 2} \rightarrow t_{n_x + n_y - 2}$$

Para determinar las zonas de aceptación y de rechazo trabajaremos con la distribución del estadístico supuesta la hipótesis nula cierta (o condicionada a la hipótesis nula), ya que R_0 y R_1 deben cumplir que :

$P(T \in R_1 | H_0) = \alpha$ que en nuestro caso será:

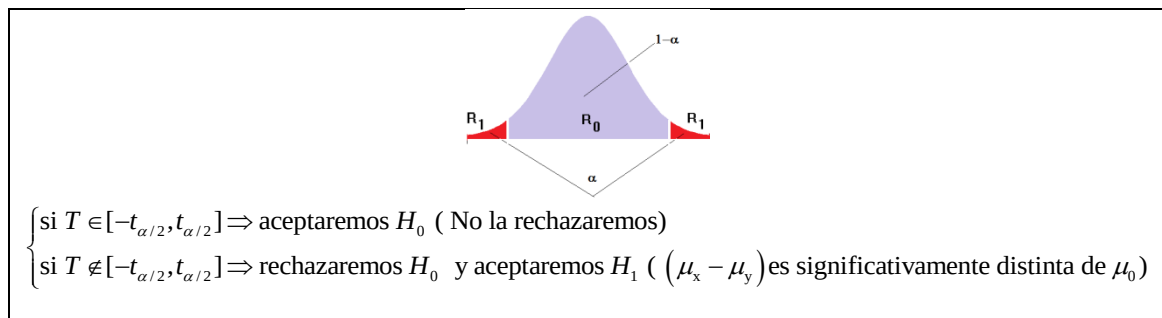
$$P\left(\frac{\bar{X} - \bar{Y} - \mu_0}{\sqrt{n_x + n_y} \cdot \sqrt{n_x S_x^2 + n_y S_y^2}} \sqrt{n_x \cdot n_y} \cdot \sqrt{n_x + n_y - 2} \in R_1 | H_0\right) = \alpha$$

En adelante, para simplificar la exposición a este estadístico bajo el supuesto de que la H_0 es cierta le llamaremos , simplemente “T”

$$T = \frac{\bar{X} - \bar{Y} - \mu_0}{\sqrt{n_x + n_y} \cdot \sqrt{n_x S_x^2 + n_y S_y^2}} \sqrt{n_x \cdot n_y} \cdot \sqrt{n_x + n_y - 2}$$

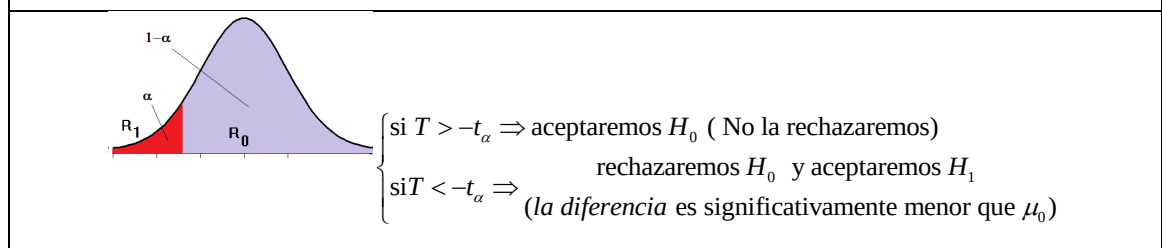
Adicionalmente para que el contraste sea lo más potente posible debe cumplir que R_1 sea lo más extensa posible (R_0 la más pequeña posible) . Esto nos conduce a que la zona de rechazo debe situarse en la (o las colas) Si toma una (y cuál) o las dos colas dependerá de la hipótesis alternativa.

Contraste con alternativa bilateral (dos colas)	$\left. \begin{array}{l} H_0 : \mu_x - \mu_y = \mu_0 \\ H_1 : \mu_x - \mu_y \neq \mu_0 \end{array} \right\} \text{ con n.s.} = \alpha$
--	--



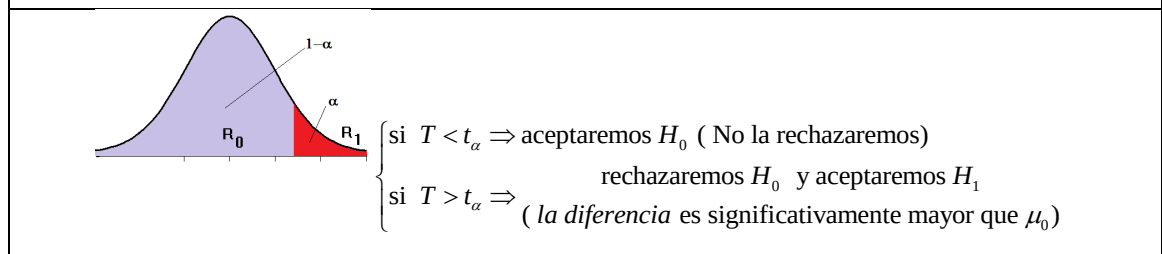
Contraste con alternativa unilateral (cola de la izquierda)

$$\left. \begin{array}{l} H_0 : \mu_x - \mu_y = \mu_0 \\ H_1 : \mu_x - \mu_y < \mu_0 \end{array} \right\} \text{ con n.s.} = \alpha$$



Contraste con alternativa unilateral (cola de la derecha)

$$\left. \begin{array}{l} H_0 : \mu_x - \mu_y = \mu_0 \\ H_1 : \mu_x - \mu_y > \mu_0 \end{array} \right\} \text{ con n.s} = \alpha$$



Para resolver este tipo de contrastes podemos utilizar la función `contraste.diferencia.medias {estadística}` con la opción del argumento `var_pob` como "desconocida". Los planteamientos, por lo demás, serán semejantes al caso anterior. Sólo considerar que, por defecto, el argumento `iguales` es FALSE por lo que es necesario cambiarlo a TRUE para considerar la hipótesis adicional habitual de que las varianzas poblacionales sean iguales.

Ejemplo 5.

Con los datos de "alumnos". A partir de dos MAS de las y los alumnos de tamaños 25 y 30 respectivamente podremos admitir que sus media de edad son iguales con una significación del 2% (Suponemos que las varianzas poblacionales de las edades de chicos y chicas son iguales)

<pre>set.seed(144) x<-sample(chicas\$edad,25) y<-sample(chicos\$edad,30)</pre>	Contrastamos que la diferencia es 0 (por defecto), con un nivel de significación del
--	---

media(x);media(y);varianza(x);varianza(y)	0.02
contraste.diferencia.medias(introducir=T,var_pob="desconocida",iguales=T,alfa=0.02,grafico=T)	(el contraste es bilateral (por defecto) suponemos varianza poblacionales iguales)

```

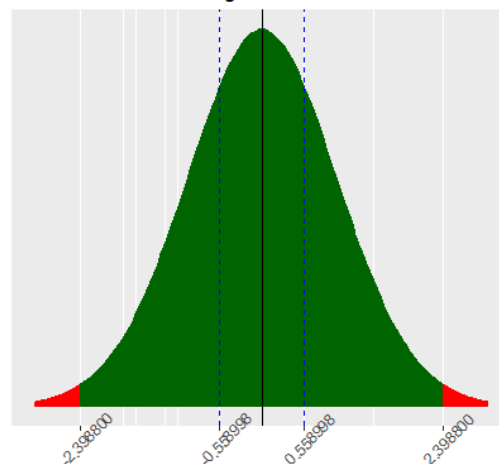
media_x
1 22.92
media_x
1 23.56667
x
1 24.9536
x
1 11.44556
>
> contraste.diferencia.medias(introducir=T,var_pob="desconocida",iguales=T,alfa=0.02,grafico=T)
Introducir el tamaño de la muestra 1: 25
Introducir el valor de la media muestral 1: 22.92
Introducir el tamaño de la muestra 2: 30
Introducir el valor de la media muestral 2: 23.56667
Selecciona el valor que quieres utilizar?
1. "Varianza muestral"
2. "Cuasivarianza muestral"
1
Introducir el valor de la varianza muestral 1: 24.9536
Introducir el valor de la varianza muestral 2: 11.44556
[1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo [-2.3988 , 2.3988]"
[1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación"
$estadístico
  Hipótesis nula estadístico de prueba  p-valor
1              0              -0.558998 0.5785175

$Intervalo de la diferencia de medias muestrales (supuesta H0 cierta)`
  limite inferior limite superior
[1,]          -2.775022           2.775022

$Graficos
$Graficos[[1]]
$Graficos[[2]]

```

Distribución t con 53 grados de libertad



¿ cómo resultaría este contraste si no supusiéramos la homoscedasticidad?

contraste.diferencia.medias(introducir=T,var_pob="desconocida",alfa=0.02)	Suprimimos el argumento iguales (por defecto es FALSE
---	--

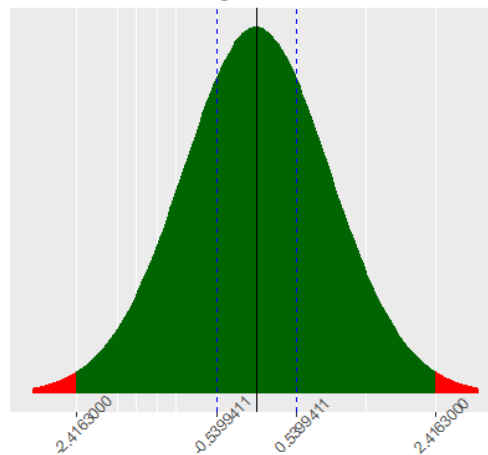

```

> contraste.diferencia.medias(introducir=T,var_pob="desconocida",alfa=0.02,grafico=T)
Introducir el tamaño de la muestra 1: 25
Introducir el valor de la media muestral 1: 22.92
Introducir el tamaño de la muestra 2: 30
Introducir el valor de la media muestral 2: 23.56667
Selecciona el valor que quieres utilizar?
1. "Varianza muestral"
2. "Covarianza muestral"
1
Introducir el valor de la varianza muestral 1: 24.9536
Introducir el valor de la varianza muestral 2: 11.44556
[1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo [-2.4163 , 2.4163]"
[1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación"
$estadístico
  Hipótesis nula estadístico de prueba   p-valor
1              0             -0.5399411 0.5920232

$`Intervalo de la diferencia de medias muestrales (supuesta H0 cierta)`
  limite inferior limite superior
[1,]      -2.893924       2.893924

```

Distribución t con 43 grados de libertad



Cambia ligeramente el valor del estadístico (recordemos que no hemos estudiado su expresión) , cambia los grados de libertad (que son aproximados) y ,por lo tanto, cambia el p-valor. Pero este p-valor sigue siendo muy superior 0.02 de modo que la conclusión sería que la edad media de las mujeres y los hombres no es significativamente diferente.

Para resolver este tipo de contrastes también podremos utilizar el procedimiento `t.test` simplemente incorporando dos argumentos (`x` y `y`) para las dos muestras (en lugar de uno sólo) y , eventualmente incorporando un argumento para especificar que las varianzas son iguales (`var.equal=TRUE` , por defecto R considera que no lo son) y otro por si el muestreo ha sido emparejado (`paired`).

```

t.test(x, y, alternative, mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)

```

Aplicado al ejemplo anterior (Ejemplo 5). Con los datos de “alumnos”. A partir de dos MAS de las y los alumnos de tamaños 25 y 30 respectivamente podremos admitir que sus media de edad son iguales con una significación del 2% (Suponemos que las varianzas poblacionales de las edades de chicos y chicas son iguales)

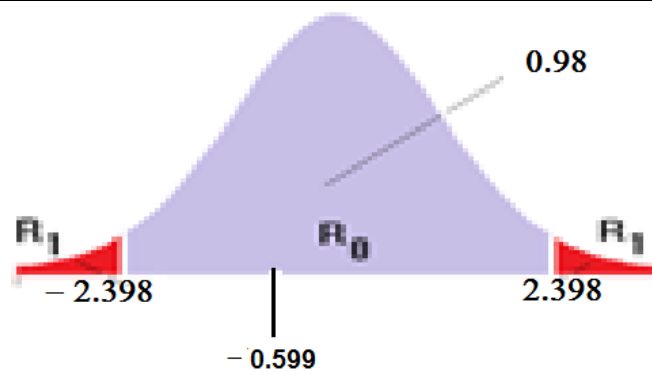
```
set.seed(144)
x<-sample(chicas$edad,25)
y<-sample(chicos$edad,30)

t.test(x,y,var.equal = TRUE,conf.level =
0.98)
```

Two Sample t-test

data: x and y
 $t = -0.559$, $df = 53$, $p\text{-value} = 0.5785$
 alternative hypothesis: true difference in means is not equal to 0
 98 percent confidence interval:
 -3.421677 2.128344
 sample estimates:
 mean of x mean of y
 22.92000 23.5666

```
t.0.01<-qt(0.01,53,lower.tail = FALSE)
t.0.01
[1] 2.39879
```



Trabajando al 2 % de significación podemos admitir que la media de edad es igual

E igualmente si no suponemos homoscedasticidad

```
t.test(x,y,alternative="l",conf.level=0.95)
```

welch Two Sample t-test

data: x and y
 $t = -2.1265$, $df = 52.945$, $p\text{-value} = 0.01907$
 alternative hypothesis: true difference in means is less than 0
 95 percent confidence interval:
 -Inf -0.3800366
 sample estimates:
 mean of x mean of y
 21.28000 23.06667

Ejercicio: Si cambiamos la muestras ,haciendo `set.seed(44)` compruébese que el contraste bilateral conduce ,también a aceptar la igual de medias al 5 y al 2% pero que si planteamos un contraste unilateral el resultado es que la media de las mujeres es significativamente inferior a la de los hombres ,al 5 % de significación.

6.2.3. Contrastes para proporciones.

Desarrollamos, ahora los contrastes sobre una proporción y sobre la diferencia de dos proporciones.

6.2.3.1.Contrates sobre una proporción

Contrastamos, con un nivel de significación α , la hipótesis nula de que la proporción de cierta característica , p , toma un determinado valor , p_0 . Frente a la hipótesis alternativa de que toma (1) un valor distinto, (2) un valor inferior, o (3) un valor mayor que p_0 . Para

realizar el contraste tomamos una MAS de tamaño n.

$$\left. \begin{array}{l} H_0 : p = p_0 \\ H_1 : p \begin{cases} \neq p_0 \\ < p_0 \\ > p_0 \end{cases} \end{array} \right\} \text{ con nivel de significación } \alpha \text{ a partir de un MAS de tamaño } n$$

$$\text{El estadístico del contraste será : } \hat{p} \rightarrow N(p, \sqrt{\frac{pq}{n}}) \Rightarrow \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \rightarrow N(0,1)$$

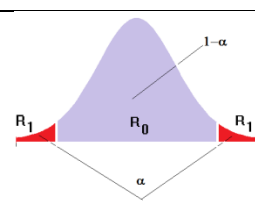
Para determinar las zonas de aceptación y de rechazo trabajaremos con la distribución del estadístico supuesta la hipótesis nula cierta (o condicionada a la hipótesis nula) , ya que R_0 y R_1 deben cumplir que :

$$P(T \in R_1 | H_0) = \alpha \text{ que en nuestro caso será: } P\left(\frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \in R_1 | H_0\right) = \alpha$$

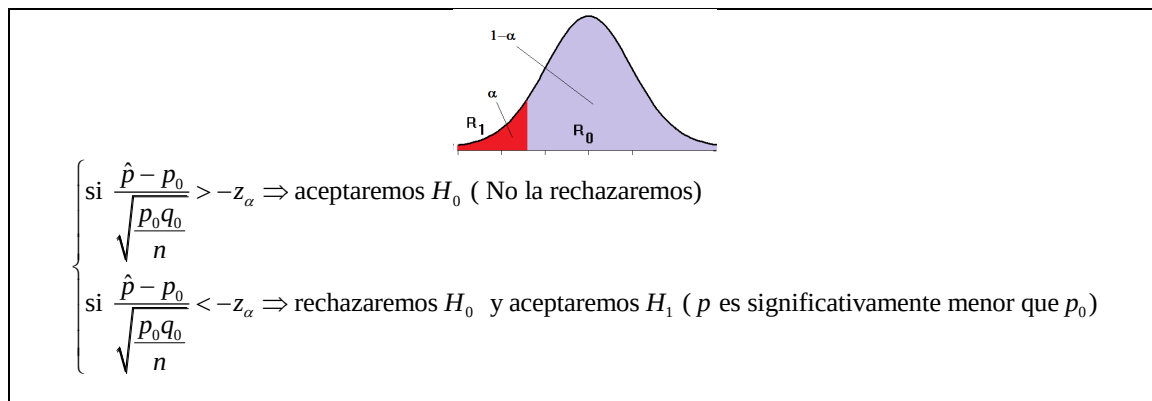
Notemos que si la hipótesis nula se supone cierta $p=p_0$ y $q=q_0=1-p_0$ (aquí no caben otras consideraciones)

Adicionalmente para que el contraste sea lo más potente posible debe cumplir que R_1 sea lo más extensa posible (R_0 la más pequeña posible) .Esto nos conduce a que la zona de rechazo debe situarse en la (o las colas) Si toma una (y cuál) o las dos colas dependerá de la hipótesis alternativa.

De forma que el criterio de decisión será:

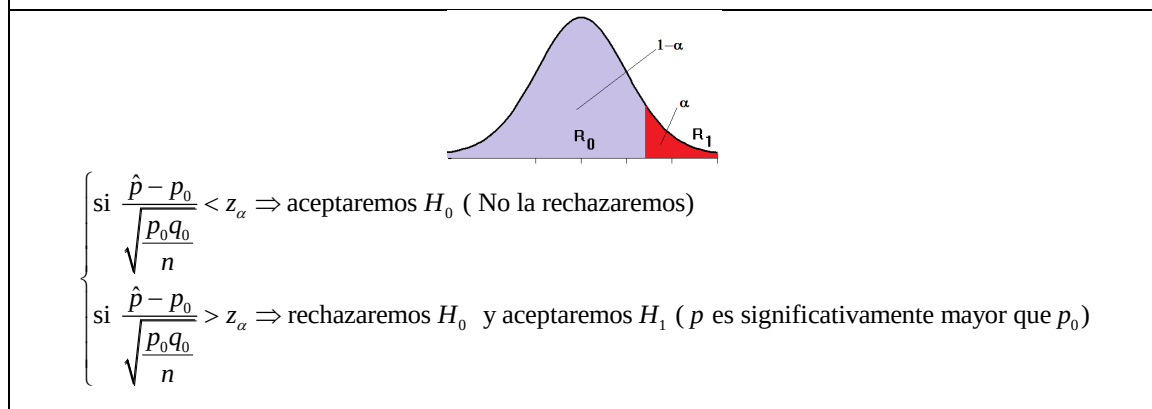
Contraste con alternativa bilateral (dos colas) $\left. \begin{array}{l} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{array} \right\} \text{ con n.s.} = \alpha$

$\left\{ \begin{array}{l} \text{si } \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \in [-z_{\alpha/2}, z_{\alpha/2}] \Rightarrow \text{aceptaremos } H_0 \text{ (No la rechazaremos)} \\ \text{si } \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \notin [-z_{\alpha/2}, z_{\alpha/2}] \Rightarrow \text{rechazaremos } H_0 \text{ y aceptaremos } H_1 \text{ (} p \text{ es significativamente distinta de } p_0 \text{)} \end{array} \right.$

Contraste con alternativa unilateral (cola de la izquierda) $\left. \begin{array}{l} H_0 : p = p_0 \\ H_1 : p < p_0 \end{array} \right\} \text{ con n.s.} = \alpha$
--



Contraste con alternativa unilateral (cola de la derecha)

$$\left. \begin{array}{l} H_0 : p = p_0 \\ H_1 : p > p_0 \end{array} \right\} \text{ con n.c.} = \alpha$$



La función `prop.test`, ya incorporada en R, también permite llevar a cabo un contraste sobre una proporción pero utilizando otro estadístico basado en la distribución chi-dos.

Para llevar a cabo el contraste tal como lo hemos planteado (estadístico, basado en la normal) podemos usar la función `contraste.proporcion` del paquete {estadística} cuyo “usage” es :

```
contraste.proporcion(x,
  variable = NULL,
  introducir = FALSE,
  hipotesis_nula = NULL,
  tipo_contraste = c("bilateral", "cola derecha", "cola izquierda"),
  alfa = 0.05,
  grafico = FALSE)
```

También se puede usar una función **diferente** ad-hoc, que hemos llamado “`p.test1`” que se muestra en un anexo.

Ejemplo 6. Con los datos de la base de datos “alumnos” seleccionamos, al azar, 50 alumnos. Con el resultado obtenido contrastar que el porcentaje poblacional de mujeres es del 60% al 0.01 de significación.

```

set.seed(144)
seleccion<-sample(240,50)

muestra<-alumnos[seleccion,]

table(muestra$sexo)#frecuencias de hombres(0) y mujeres (1)
p<-table(muestra$sexo)[2]/50 # proporción muestral de mujeres
contraste.proporcion(introducir=T,hipotesis_nula= 0.6,alfa=0.01)

#usando la función #p.test1(p,50,0.6,sign=0.01)

contraste.proporcion(introducir=T,hipotesis_nula= 0.6,alfa=0.01)
Introducir el tamaño de la muestra: 50
Introducir el valor de la proporción muestral: 0.56
[1] "No se rechaza la hipótesis nula. La región de aceptación viene dada
[-2.5758 , 2.5758]"
[1] "El valor del estadístico de prueba (o valor experimental) se encuen
egión de aceptación"
$Estadistico
  Hipótesis nula estadístico de prueba  p-valor
1          0.6          -0.57735 0.563703

$`Intervalo de la proporción muestral (supuesta H0 cierta)`
  limite_inferior limite_superior
[1,]          0.4215433          0.7784567

```

Puede comprobarse que si cambiamos el nivel de significación al 5 % tampoco podríamos rechazar la hipótesis. Y que tampoco la rechazaríamos, incluso, aunque la alternativa fuera unilateral (menor que 0.6).

6.2.3.2 Contraste sobre la diferencia de dos proporciones .Contrastamos, con un nivel de significación α , la hipótesis nula de que la diferencia de dos proporciones de dos características, $p_x - p_y$, toma un determinado valor, p_0 . Frente a la hipótesis alternativa de que toma (1) un valor distinto, (2) un valor inferior, o (3) un valor mayor que μ_0 . Para realizar el contraste tomamos dos MAS de tamaño n_x y n_y

$$\begin{aligned}
 &H_0 : p_x - p_y = p_0 \\
 &H_1 : p_x - p_y \begin{cases} \neq p_0 \\ < p_0 \\ > p_0 \end{cases} \quad \text{con nivel de significación } \alpha
 \end{aligned}$$

a partir de dos MAS de tamaño n_x y n_y

El estadístico del contraste será :

$$\hat{p}_x - \hat{p}_y \rightarrow N\left(p_x - p_y, \sqrt{\frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}}\right) \Rightarrow \frac{\hat{p}_x - \hat{p}_y - (p_x - p_y)}{\sqrt{\frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}}} \rightarrow N(0,1)$$

Para determinar las zonas de aceptación y de rechazo trabajaremos con la distribución del estadístico supuesta la hipótesis nula cierta (o condicionada a la hipótesis nula), ya que R_0 y R_1 deben cumplir que :

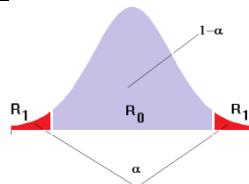
$$P(T \in R_1 | H_0) = \alpha \text{ que en nuestro caso será: } P \left(\frac{\hat{p}_x - \hat{p}_y - p_0}{\sqrt{\frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}}} \in R_1 | H_0 \right) = \alpha$$

En este caso, aunque sustituyamos la diferencia por el valor que toma según la hipótesis (p_0) en el numerador, los valores que tomen los parámetros poblacionales del denominador no pueden ser conocidos. Debemos hacer una consideración adicional: o tomar los valores muestrales o preferiblemente tomar $p_x=p_y=q_x=q_y=0.5$.

$$\text{quedando, entonces : } P \left(\frac{\hat{p}_x - \hat{p}_y - p_0}{\sqrt{\frac{0.5 \cdot 0.5}{n_x} + \frac{0.5 \cdot 0.5}{n_y}}} \in R_1 | H_0 \right) = \alpha$$

Adicionalmente para que el contraste sea lo más potente posible debe cumplir que R_1 sea lo más extensa posible (R_0 la más pequeña posible). Esto nos conduce a que la zona de rechazo debe situarse en la (o las colas) Si toma una (y cuál) o las dos colas dependerá de la hipótesis alternativa.

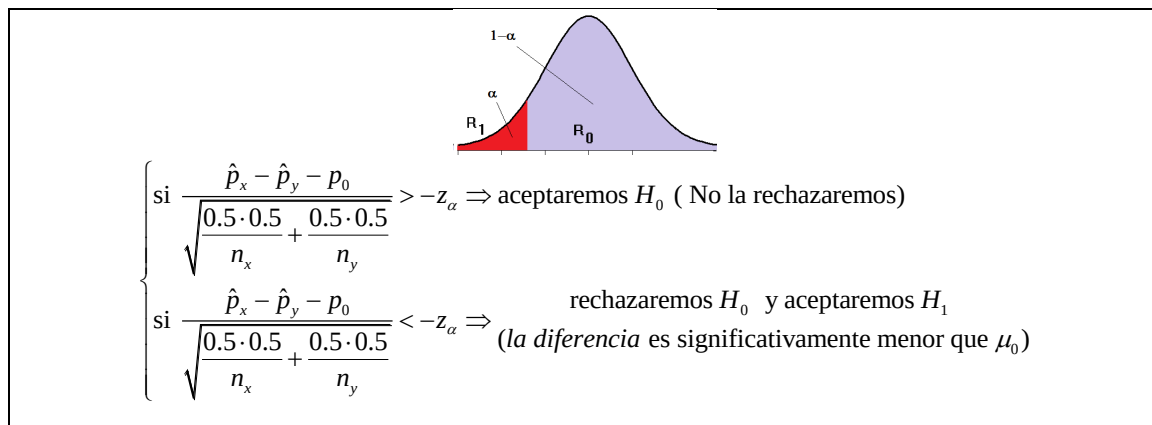
$$\text{Contraste con alternativa bilateral (dos colas) } \left. \begin{array}{l} H_0 : p_x - p_y = p_0 \\ H_1 : p_x - p_y \neq p_0 \end{array} \right\} \text{ con n.s.} = \alpha$$



$$\left\{ \begin{array}{l} \text{si } \frac{\hat{p}_x - \hat{p}_y - p_0}{\sqrt{\frac{0.5 \cdot 0.5}{n_x} + \frac{0.5 \cdot 0.5}{n_y}}} \in [-z_{\alpha/2}, z_{\alpha/2}] \Rightarrow \text{aceptaremos } H_0 \text{ (No la rechazaremos)} \\ \text{si } \frac{\hat{p}_x - \hat{p}_y - p_0}{\sqrt{\frac{0.5 \cdot 0.5}{n_x} + \frac{0.5 \cdot 0.5}{n_y}}} \notin [-z_{\alpha/2}, z_{\alpha/2}] \Rightarrow \text{rechazaremos } H_0 \text{ y aceptaremos } H_1 \text{ (} (\mu_x - \mu_y) \text{ es significativamente distinta de } \mu_0 \text{)} \end{array} \right.$$

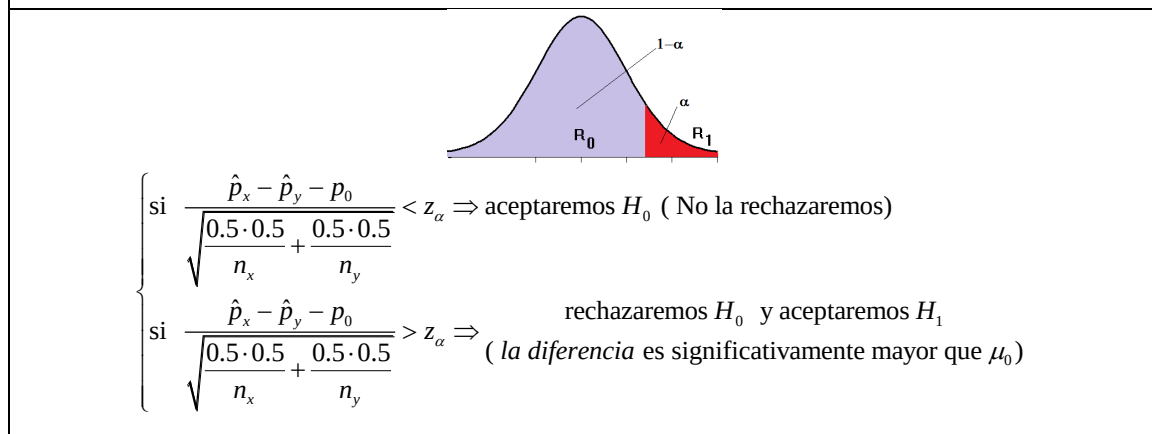
Contraste con alternativa unilateral (cola de la izquierda)

$$\left. \begin{array}{l} H_0 : p_x - p_y = p_0 \\ H_1 : p_x - p_y < p_0 \end{array} \right\} \text{ con n.s.} = \alpha$$



Contraste con alternativa unilateral (cola de la derecha)

$$\left. \begin{array}{l} H_0 : p_x - p_y = p_0 \\ H_1 : p_x - p_y > p_0 \end{array} \right\} \text{ con n.s.} = \alpha$$



Utilizando una función ad-hoc, podemos automatizar el contraste para la diferencia de dos proporciones con R según el esquema anterior basado en la distribución normal. En un anexo podemos ver cómo podemos construir una función ad-hoc que hemos llamado “p.test2”, para este contraste

La función `contraste.diferencia.proporciones` del paquete `{estadística}` permite también automatizar este contraste en R su “usage” es :

```
contraste.diferencia.proporciones(x,
  variable = NULL,
  introducir = FALSE,
  hipotesis_nula = 0,
  tipo_contraste = c("bilateral", "cola derecha", "cola izquierda"),
  alfa = 0.05,
  grafico = FALSE)
```

Con todo, esta función, que es fácil de usar, al igual que las anteriores, calcula el estadístico z del contraste considerando como aproximación de la proporciones en el denominador, no el valor de holgura máxima (0.5), como ocurre arriba, sino o bien los

valores de las proporciones muestrales o bien una media ponderada de ambas . La función ad-hoc del anexo considera la opción “0,5”

Ejemplo 7. Sobre la base de datos anterior (alumnos) suponemos que los datos constituyen ya una muestra . De hecho podemos considerar dos muestras una de las mujeres de tamaño 124 y otras de hombres de tamaño 116. A partir de estas muestras contrastar que la proporción de mujeres que trabajan es igual a la de hombres. (nivel de significación del 2%)

```
p1<-table(chicas$trabajo)[2]/124
p2<-table(chicos$trabajo)[2]/116
p1;p2
contraste.diferencia.proporciones(introducir=T,alfa=0.02)

> p1;p2
      1
0.2822581
      1
0.4224138
> contraste.diferencia.proporciones(introducir=T,alfa=0.02)
[1] "Primero vas a introducir los datos de la muestra 1 y a continuación introducirás los datos de la muestra 2"
[1] "Si los datos provienen de encuestas realizadas antes y después de una determinada acción, introduce primero los datos de la encuesta realizada después de dicha acción"
Introducir el tamaño de la muestra 1: 124
Introducir el valor de la proporción muestral 1: 0.2822581
Introducir el tamaño de la muestra 2: 116
Introducir el valor de la proporción muestral 2: 0.4224138
Selecciona el valor que quieres utilizar para el error típico bajo la H0:
1. "Estimar p como media ponderada de las proporciones muestrales"
2. "Utilizar las proporciones muestrales"
2
[1] "No se rechaza la hipótesis nula. La región de aceptación viene dada por el intervalo [-2.3263 , 2.3263]"
[1] "El valor del estadístico de prueba (o valor experimental) se encuentra dentro de la región de aceptación"
$Estadístico
  Hipótesis nula estadístico de prueba      p-valor
1              0          -2.29269 0.02186586

$Intervalo de la proporción muestral (supuesta H0 cierta)
  limite_inferior limite_superior
[1,]          -0.1422101          0.1422101
```

Vemos, que aunque no se puede rechazar (al 2% de significación) que sean iguales las proporciones, sí podemos concluir que la proporción que trabaja ente las mujeres es significativamente menor (también al 2 %)

El procedimiento prop.test.

Como ya hemos comentado, más arriba el prop.test permite realizar contrastes sobre proporciones (una o diferencia de dos) pero utilizando un estadístico diferente. Los resultados obtenidos serán diferentes y las conclusiones finales de aceptación/ rechazo puede ser distintas en los casos “fronterizos”.

```
prop.test(x, n, p, alternative, conf.level, correct)
```

con los argumentos:

x= número de éxitos (también puede ser un vector con el número de éxitos en dos categorías cuando se vaya a contrastar la diferencia de dos proporciones).

n= tamaño de la muestra (vector de los tamaños de las dos muestras en el caso del contraste de la diferencia

p = valor de la proporción muestral según la H_0 (cero, por defecto)

alternative= dirección de la hipótesis nula (“two.sided”, “greater”,o “less”) Bilateral, por defecto. También se puede usar “g” o “l”

conf.level= nivel de confianza (1- alfa) Por defecto 0.95

Existen otros contrastes paramétricos de interés. Algunos de ellos particulares de modelos estadísticos concretos como pueden ser los contrastes sobre los coeficientes de regresión y algunos otros de parámetros poblacionales general que no hemos

estudiados. Entre éstos uno de los más importantes es el contraste de incorrelación que admite dos formas:

Un contraste F en el que se contrasta la incorrelación exclusivamente frente a la alternativa bilateral.

Y Un contraste t en el que es posible hacer un diseño bilateral o unilateral para constatar en su caso la eventual presencia de correlación significativamente positiva o negativa.

Puede verse en ceaces: <https://www.uv.es/ceaces/txt/6%20para/cincorrela.htm>

6.3. Principales contrastes no paramétricos.

En lo estudiado hasta ahora hemos realizado inferencias (estimaciones y contrastes de hipótesis) sobre características desconocidas de la población que resultaban ser parámetros de la distribución de la población. Era necesario conocer la distribución de la población y postular su dependencia de uno o más parámetros. De esta forma, a partir de las distribuciones muestrales, podíamos diseñar métodos para estimar y contrastar los valores de los parámetros, según la distribución de la población de la que se tratara. Sin embargo, existen características de la población que son lo suficientemente importantes que hacen imprescindible la inferencia sobre ellas. Es el caso de: el tipo de distribución de la población, la existencia o no independencia de esta respecto a otra, la (presumible o descartable) homogeneidad de comportamiento frente a diversos valores de cierto factor, la existencia de rachas, la procedencia aleatoria de los datos etc.

Los contrastes no paramétricos constituyen una metodología de trabajo idéntica a la estudiada para el caso paramétrico pero con la diferencia evidente de que las hipótesis planteadas no harán referencia al valor o relaciones de valor de los (desconocidos) parámetros la población. Sin embargo las cuestiones y conceptos de hipótesis nula, alternativa, probabilidades de error, nivel de significación, hipótesis básicas para aplicación, obtención del estadístico, comparación, decisión jugarán idéntico papel.

6.3.1. Contraste de la χ^2 (de bondad del ajuste)

La hipótesis a contrastar es el hecho (nótese que no hay valor de un parámetro) de que la muestra proviene de una distribución (poblacional) determinada frente a la alternativa de que esto no es así (o de que proviene de otra diferente no especificada).

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

H_0 : la muestra proviene de (se ajusta a) una población cuya función de distribución es $F(x)$

H_1 : la muestra NO proviene de una población cuya función de distribución es $F(x)$

trabajando, claro está, con un determinado nivel de significación, α .

A partir de la distribución de frecuencias de la muestra y de la distribución de frecuencias esperadas (o frecuencias teóricas) según el modelo o la distribución que postula la hipótesis nula, se puede construir un **estadístico** que mide la discrepancia entre los resultados (las frecuencias, en realidad) observados y los teóricos.

Este estadístico suele denominarse estadístico “CHI-DOS” (ya que, si la hipótesis nula es cierta, tendrá aproximadamente esta distribución):

$$\chi^2 = \sum_{i=1}^m \frac{(n_i^0 - n_i^t)^2}{n_i^t} \quad \text{donde : } \begin{cases} m = \text{número de valores (niveles) diferentes de la muestra} \\ n_i^0 = \text{frecuencia observada del i-simo valor o nivel} \\ n_i^t = \text{frecuencia teórica del i-simo valor o nivel según la } H_0 \end{cases}$$

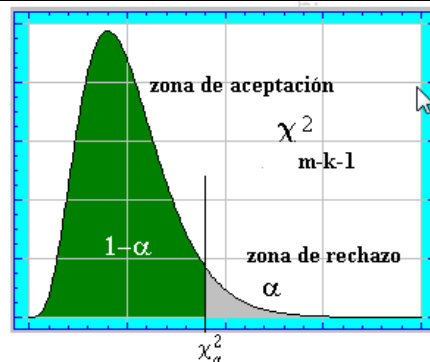
Por puro azar las frecuencias observadas diferirán de las teóricas pero si la hipótesis nula se cumple, cabe esperar que no difieran demasiado y el estadístico no tomará valor excesivamente altos. En cambio si el estadístico tomara un valor muy muy alto parece razonable pensar que quizá la hipótesis de que los datos provienen de la distribución considerada no sea cierta y deba rechazarse.

Puede probarse que si la hipótesis nula es cierta este estadístico se distribuye con una distribución chi-dos de Pearson con $m-k-1$ grados de libertad; siendo m el número de valores diferentes y k el número de parámetros que haya sido necesario estimar para calcular las frecuencias teóricas.

Fijado un nivel de significación α la decisión será:

$$\chi^2 = \sum_{i=1}^m \frac{(n_i^0 - n_i^t)^2}{n_i^t} < \chi_{\alpha}^2 \rightarrow \text{aceptaremos } H_0$$

$$\chi^2 = \sum_{i=1}^m \frac{(n_i^0 - n_i^t)^2}{n_i^t} > \chi_{\alpha}^2 \rightarrow \text{rechazaremos } H_0$$



Alternativamente si el p-valor del estadístico es inferior al nivel de significación rechazaremos. (Si no es así, no lo haremos)

Ejemplo 8: Queremos comprobar que un dado de póker no está trucado (cada cara tiene una probabilidad de 1/6 de salir) Para ello lo lanzamos 60 veces saliendo los siguientes resultados:

Cara	Frecuencia
As	8
K	9
Q	7
J	13
Negro	12
Rojo	11
Total	60

¿Puede admitirse que el dado es correcto con un 5 % de significación?

$$\left. \begin{aligned} H_0 : P(as) = P(K) = P(Q) = \dots = P(negro) = 1/6 \\ H_1 : \text{las probabilidades son diferentes (} H_0 \text{ no es cierta)} \end{aligned} \right\} \alpha = 0.05$$

Las frecuencias teóricas será todas iguales a 10 ya que

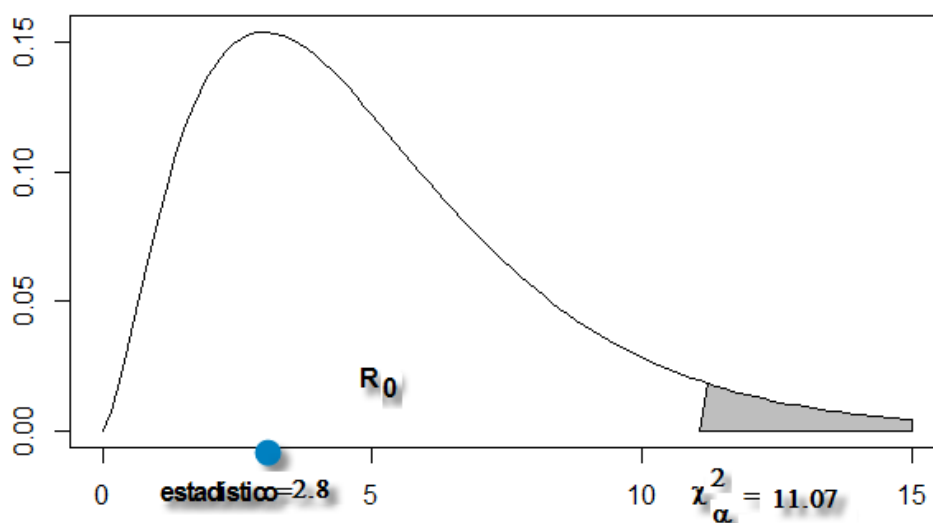
$n_i^t = P(i) \cdot n = 1/6 \cdot 60 = 10 \quad \forall i$ y tendremos que estadístico Chi-dos se podrá calcular como:

Cara	n_i^o	n_i^t	$(n_i^o - n_i^t)^2$	$\frac{(n_i^o - n_i^t)^2}{n_i^t}$

As	8	10	4	0.4
K	9	10	1	0.1
Q	7	10	9	0.9
J	13	10	9	0.9
Negro	12	10	4	0.4
Rojo	11	10	1	0.1
Total	60		CHI DOS=	2.8

Los grados de libertad de la distribución del estadístico serán $m-k-1=6-0-1=5$
Y el valor crítico para una chi-dos de 5 grados de libertad y $\alpha=0.05$ es 11.07
`qchisq(0.05,5,lower.tail = FALSE)`
[1] 11.0705

De modo que aceptaremos la hipótesis nula (no podemos descartar que el dado sea legal)



La función `chisq.test` de `r` automatiza el contraste de la chi-dos, aunque tiene otras opciones con sus tres argumentos principales tendríamos:

`chisq.test(x, (y), (p))`

x = vector o matriz numérica. Puede ser también un factor. Habitualmente las frecuencias observadas

y = un vector numérico.(También puede ser un factor) Se ignora si x es una matriz. Si x e y son factores tienen que tener la misma longitud

p = el vector de probabilidades según la hipótesis nula.Tiene la misma longitud que x y por defecto es $1/m$.

`chisq.test(c(8,9,7,13,12,11))`

Chi-squared test for given probabilities

data: c(8, 9, 7, 13, 12, 11)
X-squared = 2.8, df = 5,
p-value = 0.7308

No introducimos las probabilidades porque son todas $1/6$.
Da como el resultado:el estadístico CHI-dos, los grados de libertad y el p-valor
Con p-valor de 0.7320 se acepta la Hipótesis a todos los N.S. habituales (1%,5%,10% e ¡incluso 70%!)

Ejemplo 9 Se realiza una muestra de 1500 observaciones y resulta que obtenemos que los valores de x observados han sido:

x	observaciones
-----	---------------

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

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

Este ejemplo está resuelto en CEACES

(<https://www.uv.es/ceaces/tex1t/7%20no%20para/ejemplo3.htm>)

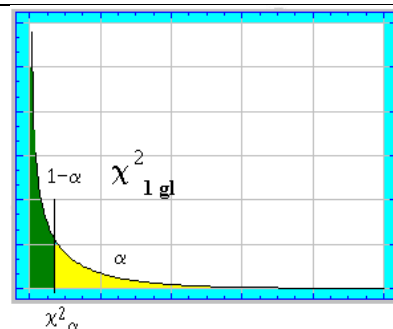
Pero nos interesa llamar la atención aquí en que para calcular las frecuencias teóricas es necesario conocer las probabilidades que supuesta una distribución de Poisson requerirán estimar el parámetro λ . Perderemos un grado de libertad. La estimación siempre que sea necesaria, como ahora será por el método de máxima verosimilitud. (estimaremos λ por el valor de la media muestral)

Este tipo de contrastes de la chi-dos donde se trata de ver si los datos se ajustan a una distribución específica que requiere que se estime alguno de sus parámetros, no es aconsejable llevarlos a cabo en R con la instrucción `chisq.test`.

Corrección de Yates.

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

$$\chi^2 = \sum_{i=1}^m \frac{(|n_{o,i} - n_{t,i}| - 0,5)^2}{n_{t,i}}$$



Ejemplo 10. En 60 lanzamientos de una moneda se observaron 39 caras y 21 cruces. Contrastar la hipótesis de que la moneda no está trucada, con un nivel de significación del 1%.

En este caso tenemos $m=2$ (cara y cruz) y $k=0$ (no estimamos ningún parámetro. Por lo tanto los g.l van a ser 1 por lo que tendremos que hacer la corrección de Yates.

	$n_{0,i}$	$n_{t,i}=P_i \cdot n$	$ n_{o,i} - n_{t,i} $	$\frac{(n_{o,i} - n_{t,i} - 0,5)^2}{n_{t,i}}$
c	39	$0,5 \cdot 60 = 30$	9	2.408
+	21	$0,5 \cdot 60 = 30$	9	2.408
	$60=n$	60		Chidos= 4.816

El valor crítico para 1 % de significación y un grado de libertad es 6.63 por lo que no podemos rechazar que la moneda sea legal.

6.3.2. Contraste de independencia (tabla de contingencia)

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

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

Frec.Observadas

X\Y	y ₁	y ₂	...	y _j	...	y _m	
x ₁	n _{1,1}	n _{1,2}	•	•	•	n _{1,m}	n_{1,*}
x ₂	n _{2,1}	n _{2,2}	•	•	•	•	n_{2,*}
•	•	•	•	•	•	•	
x _i	•	•	...	n _{i,j}	...	•	n_{i,*}
•	•	•	•	•	•	•	
x _n	n _{n,1}	•	•	•	•	n _{n,m}	n_{n,*}
	n_{*,1}	n_{*,2}		n_{*,j}		n_{*,m}	N

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

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

luego en el caso de independencia, cada una de las celdas “ij”, las frecuencias conjuntas teóricas serán :

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

Y usando el procedimiento del contraste de la chi-dos, podemos considerar el estadístico

$$\chi^2 = \sum_i \sum_j \frac{(n_{i,j} - n_{i,j}^T)^2}{n_{i,j}^T}$$

Sólo hay que puntualizar que el citado estadístico se distribuirá con una distribución χ^2 con $(m-1)(n-1)$ grados de libertad ; ya que de los $m \cdot n$ sumandos que constituyen el estadístico ($m \cdot n$ celdas de la tabla) , $m+n-1$ están determinados por los demás y quedan por lo tanto: $m \cdot n - (m+n-1)$ libres = $m \cdot n - m - n + 1 = (m-1) \cdot (n-1)$.

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

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

Ejemplo 11 Se dispone de las observaciones del color del pelo y de los ojos de 400 individuos según la siguiente tabla:

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

Contrastar con un nivel de significación del 1 % la independencia de estos atributos. ([ir a script de realización](#) en CAEST)

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

expresión $n_{ij}^r = \frac{n_{i \cdot} \cdot n_{\cdot j}}{N}$ construimos la tabla de contingencia de frecuencias teóricas

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

Construimos el estadístico :

$$\chi^2 = \sum_i \sum_j \frac{(n_{i,j} - n_{i,j}^r)^2}{n_{i,j}^r}$$

que acaba tomando el valor CHI-dos= 55.13

Para llevar a cabo el contraste los grados de libertad de la distribución serán:

$(3-1).(3-1)=4$ y el valor crítico será, por lo tanto:

$$\chi^2_{\alpha}(4g.l.) = qchisq(0.01, 4, lower.tail = FALSE) = 13.2767$$

Al ser el estadístico mayor que el valor crítico descartamos la hipótesis nula (descartamos la independencia) . Existe una relación significativa (al 1%)

Podemos llevar a cabo este contraste con R introduciendo la tabla de frecuencias observadas como argumento (en forma de matriz) de la función `chisq.test` (Cuando el argumento es una matriz (o una tabla) R entiende que es un contraste de independencia.

```
tabla<-as.table(matrix(data= c( 120,50,50,20,30,10,20,60,40),nrow=3,ncol=3))
chisq.test(tabla)
>tabla
  A  B  C
A 120 20 20
B  50 30 60
C  50 10 40
> chisq.test(tabla)

      Pearson's Chi-squared test

data:  tabla
X-squared = 55.13, df = 4, p-value = 3.052e-11
```

A partir de un `data.frame` si antes generamos la tabla de frecuencias de doble entrada (tabla de contingencia) también podemos llevar a cabo el contraste de forma similar

```
# cargamos la base de datos alumnos
library(haven)
alumnos <- read_sav("AzarProbabilidadInferencia/datos/alumnos.sav")
#convertimos en factores con sus etiquetas de nivel las variables estudios de la madre y estud. del padre
estudimadre<-factor(alumnos$estud.madre,labels=c("sin estudios", "primarios","secundarios", "técnicos",
"superiores"))
estudipadre<-factor(alumnos$estudi.padre,labels=c("sin estudios","primarios","secundarios","técnicos",
"superiores"))
#construimos la tabla de frecuencias ( de contingencia)
tabla<-table(estudimadre,estudipadre)
tabla
>tabla
      estudipadre
estudimadre sin estudios primarios secundarios técnicos superiores
sin estudios      9         14           0           0           0
primarios         1        109          36           8           8
secundarios       0         6          20           6           5
técnicos          0         1           3           7           3
superiores       0         0           1           1           2
# hacemos el contraste de la chi-dos sobre la tabla
chisq.test(tabla)

      Pearson's Chi-squared test

data:  tabla
X-squared = 168.56, df = 16, p-value < 2.2e-16
# con ese p-valor : DESCARTARÍAMOS LA INDEPENDENCIA
```

Prueba exacta de Fisher .

El contraste de independencia utilizando la prueba Chi-dos es ,en realidad una aproximación que funciona bastante bien cuando el tamaño de la muestra es grande. Sin embargo existe una prueba exacta (no aproximada) que será preferible si el tamaño de la muestra es pequeña. Esta prueba, que se debe a Ronald Fisher se basa en un

estadístico que sigue una distribución hipergeométrica. Puede usarse con R a través de la función `fisher.test` tanto con la información de la tabla de contingencia como con los datos de los factores directamente considerados como dos vectores de datos (columnas del `data.frame`).

A partir de los datos de la base de datos `alumnos` consideramos las variables (factores, en realidad) `estudios de la madre` y `trabajo` y nos planteamos contrastar su independencia con la prueba de Fisher.

<pre> estudimadre<-factor(alumnos\$estud.madre, labels = c("sinestudios", "primarios", "secundarios", "técnicos", "superiores")) trabajo<-factor(alumnos\$trabajo, labels=c("no trabaja", "trabaja")) contingencia1<-fisher.test(estudimadre,trabajo) contingencia1 tabla<-table(estudimadre,trabajo) tabla contingencia2<-fisher.test(tabla) ----- tabla trabajo estudimadre no trabaja trabaja sin estudios 12 11 primarios 109 53 secundarios 26 11 técnicos 7 7 superiores 2 2 > fisher.test(tabla) Fisher's Exact Test for Count Data data: tabla p-value = 0.3272 alternative hypothesis: two.sided </pre>	Primero convertimos en factores las dos variables consideradas y les asignamos etiquetas a los niveles. Podemos aplicar el test de Fisher a los dos vectores O bien construir la tabla de contingencia y aplicar el test al objeto tabla de contingencia. Los objetos "contingencia1" y "contingencia2" son idénticos. Como el p-valor es mayor que los niveles de significación más habituales concluiríamos que el factor de trabajo (si los estudiantes trabajan o no) no depende del nivel de estudios de la madre.
--	--

6.3.3. Contraste de homogeneidad

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

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

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

$$\frac{n_{i1}}{n_{\cdot 1}} = \frac{n_{i2}}{n_{\cdot 2}} = \dots = \frac{n_{im}}{n_{\cdot m}} \quad \text{para } i= 1,2,3,\dots,n \quad \text{y será obviamente la misma que la proporción de observaciones de ese tipo que hay en el total :}$$

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

para cada x_i siendo N el total de observaciones

manera que las frecuencias teóricas deberán verificar:

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

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

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - n_{ij}^T)^2}{n_{ij}^T}$$

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

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

Aunque se contrastan hipótesis diferentes el contraste es técnicamente idéntico al de independencia.

6.3.4. Contraste de Kolmogorov-Smirnov.

El contraste de Kolmogorov-Smirnov es un contraste de bondad del ajuste alternativo al de la chi-dos que pretende dar solución a algunos de sus inconvenientes.

El contraste de la Chi-dos puede resultar inadecuado si la muestra no es demasiado grande ya que se basa en una distribución aproximada del estadístico. Por otro lado si el modelo que queremos testar es continuo el cálculo de las frecuencias teóricas habrá que llevarlo a cabo por intervalos y la consideración de unos u otros intervalos siempre será algo subjetivo.

El contraste de Kolmogorov-Smirnov se basa en el resultado siguiente si la muestra observada proviene de una población cuya distribución es la de la hipótesis que contrastamos, la función de distribución empírica (las frecuencias relativas acumuladas) no debe alejarse mucho de la función de distribución que propone la hipótesis.

A partir de esta idea se trabaja con un estadístico que mide esta diferencia en valor absoluto:

$$D = \max_{i \in \{1, 2, \dots, n\}} \left\{ \left| F_i^o - F_i^T \right| \right\}$$

Si este estadístico supera un cierto valor crítico (dependiente del nivel de significación escogido) rechazaremos la hipótesis nula.

Los valores críticos suelen consignarse en tablas ad-hoc o, en otras ocasiones, se aproximan tras una transformación sobre la distribución normal. En los paquetes informáticos, (SPSS, R, etc), lo habitual es que nos dé el resultado del estadístico calculado y el correspondiente p-valor. Si el p-valor es inferior al nivel de significación que consideremos rechazaremos la hipótesis nula.

En R los contrastes de Kolmogorov-Smirnov se llevan a cabo con la instrucción `ks.test`:

```
ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"),
        exact = NULL)
```

Donde:

`x` es el vector de datos muestrales

`y` es o bien un vector de valores o el nombre de una función de distribución (por ejemplo `pnorm` , `pbinom`, `pgamma`, etc.

(los puntos suspensivos indican que quizá tengamos/queroamos introducir los valores de los parámetros de la f.de distribución :Por ejemplo:
`lambda=3`)

`alternative` juega el papel habitual

por ejemplo

```
k<- ks.test(alumnos$edad, pnorm, mean(alumnos$edad), sd(alumnos$edad))
```

```
k
```

One-sample Kolmogorov-Smirnov test

data: alumnos\$edad

D = 0.13415, p-value = 0.0003543

alternative hypothesis: two-sided

```
k3<-ks.test(alumnos$edad, pgamma, shape=mean(alumnos$edad)^2/var(alumnos$edad),
            scale=var(alumnos$edad)/mean(alumnos$edad))
```

```
k3
```

One-sample Kolmogorov-Smirnov test

data: alumnos\$edad

D = 0.11307, p-value = 0.004324

alternative hypothesis: two-sided

#en ambos casos a hay que descartar que la distribución de las edades se ajuste a la
distribución propuesta (normal y gamma) a los N.S. habituales.

6.3.5. Otros contrastes no paramétricos

Además de los vistos hasta aquí hay muchísimos más contrastes no paramétricos. Dar cuenta de todos ellos aquí sería tarea imposible. Baste decir que más allá de que unos trabajen con una, dos, o varias muestras, y que lo hagan con muestras independientes o emparejada, todos participan de la misma filosofía y metodología y por ello su aplicación no debe suponer ningún problema una vez conocidos los que hemos estudiado aquí.

6.4. Significatividad estadística y significatividad práctica

Una nota final al respecto de la significatividad. En la salidas de resultados por ordenador es habitual que aparezca junto con el valor de cada estadístico calculado, su p-valor o su significación. En función de este p-valor suele hablarse de la significación o significatividad de un resultado experimental o de un estadístico.

La significatividad estadística es siempre inversa al valor del p-valor. Cuanto más pequeño el p-valor más significativo es el resultado. El sentido de esto es que con ese resultado obtenido podría rechazarse la hipótesis nula con un nivel de significación igual a ese p-valor (también mayor, pero no menor). Habitualmente la hipótesis nula considerada es que el parámetro (media, correlación etc.) es cero, o que la diferencia de parámetros es cero. En ese sentido, el valor del estadístico se aparta del de la hipótesis nula (que es también nula en ese sentido de ser cero) tan significativamente

como que la probabilidad de que haya salido ese valor (por puro azar) siendo cierta la hipótesis nula es tan baja como lo que dice ese nivel de significación.

La significación estadística da una idea de la relevancia o la contundencia de un resultado empírico, es cierto, pero en este sentido preciso que, a menudo, puede llevar a confusión si se extraen conclusiones de forma no crítica.

Supuesto que un parámetro vale cero, si tras tomar unos pocos datos su estimación (estadístico) resulta ser por ejemplo 20 puede resultar un valor muy “contundentemente diferente de cero” (sobre todo si hablamos de cosas que no toman valores muy grandes, digamos del orden de decenas o quizá poco más). Pero puede que no sea significativo estadísticamente, porque al haber tomado pocos datos la probabilidad de que el resultado salga 20 por puro azar puede ser grande.

La significatividad estadística ayuda a poner las cosas en su lugar ante “aparentes” discrepancias de los datos respecto a las hipótesis nulas.

Pero, pecar de basar cualquier conclusión en el p-valor, también, es, a menudo una fuente de errores. En ocasiones, la significatividad estadística no lo es todo. En situaciones en las que la variabilidad natural de la variable que estudiamos sea muy grande, o los datos no se hayan podido seleccionar por procedimientos “estrictamente” aleatorios, o cuando los tamaños de la muestra o el diseño del contraste no sea los adecuados un resultado significativo puede no ser suficiente incluso más allá de la “duda razonable” que deja “tan enrasada” el p-valor obtenido.

Por eso es tan fundamental considerar la provisionalidad de todo conocimiento empírico el permanente sentido crítico de la ciencia y su empeñamiento en la repetitividad (y la repetición efectiva) de las pruebas.

ANEXO1 FUNCIÓN AD-HOC z.test.2m

(El procedimiento de R “z.test” incluye el contraste para la media de una población normal con varianza poblacional conocida pero (a diferencia del procedimiento t.test) no el contraste para la media de **dos** poblaciones normales con las varianzas poblacionales conocidas aunque puede llevarse a cabo con un procedimiento similar)

Para realizar estos contrastes **no** puede utilizarse el procedimiento z.test pero podemos construir ad-hoc una función similar:

```
#####  
##función para el contraste z de diferencia de medias  
## z.test.2m=function(a,b,var.a,var.b,mu.0=0,alternativa.1="distinta",sign=0.05)  
## a= muestra1  
## b= muestra2  
## var.a= varianza pob1  
## var.b= varianza pob2  
## mu.0= valor de la diferencia según la Hip.nula ( 0, por defecto)  
## alternativa.1 = "distinta" / "mayor" / "menor" ( distinta ,por defecto)  
## sign= nivel de signif. <<alfa >> para el contraste ( 0.05, por defecto)  
  
z.2m = function(a, b, var.a, var.b, mu.0=0){
```

```

n.a = length(a)
n.b = length(b)
zeta = (mean(a) - mean(b)-mu.0) / (sqrt(var.a/n.a + var.b/n.b))
return(zeta)
}
p.v=function(a,b,var.a,var.b,mu.0,alternativa.1="distinta"){
z0=z.2m(a,b,var.a,var.b,mu.0)
if(alternativa.1=="mayor")
p=pnorm(z0,0,1,lower.tail = FALSE)
else
if (alternativa.1=="menor")
p=pnorm(z0,0,1)
else
p=pnorm(abs(z0),0,1,lower.tail=FALSE)*2
return(p)
}
z.test.2m=function(a,b,var.a,var.b,mu.0,alternativa.1="distinta",sign=0.05)
{
z=z.2m(a,b,var.a,var.b,mu.0)
p=p.v(a,b,var.a,var.b,mu.0,alternativa.1)
print("Contraste z para la diferencia de dos medias")
print("=====")
print(paste("H0 : la diferencia =",mu.0))
print (paste ("H1:", alternativa.1,"que/de",mu.0))
print("=====")
print("estadístico")
print (paste ("z=",z))
print("=====")
print(paste("p-valor=",p))
print("conclusión:")
if (p<sign) print(paste("la diferencia es significativamente",alternativa.1,
"que/de",mu.0,"a un nivel del ",
sign*100,"% de significación"))
else
print(paste("la diferencia NO es significativamente",alternativa.1,
"que/de",mu.0,"a un nivel del ",
sign*100,"% de significación"))
}

```

Ejemplo 3 .Usando tal función y trabajando sobre la base de datos anterior:

<pre> chicas<-subset(alumnos,sexo==1) chicos<-subset(alumnos,sexo==0) chicas<-subset(alumnos,sexo==1) chicos<-subset(alumnos,sexo==0) m1<-chicos\$edad m2<-chics\$edad v1=var(m1) v2=var(m2) z.test.2m(m1,m2,v1,v2,0) z.test.2m(m1,m2,v1,v2,1,"mayor") z.test.2m(m1,m2,v1,v2,0,sign=0.01) </pre>	Trabajamos con las edades de los chicos y de las chicas consideradas como muestras aleatorias. Obtenemos sus medias y varianzas Contrastamos la hipótesis de que sus edades medias sean iguales $\mu=0$ frente a la alternativa bilateral. Contratamos que la diferencia sea de un año(a favor de los chicos) frente
--	--

	a la alternativa mayor que 1 año. Contrastamos que son iguales, frente la alternativa bilateral pero con significación del 1 % .
"Contraste z para la diferencia de dos medias" ===== [1] "H0 : la diferencia = 0" [1] "H1: distinta que/de 0" ===== [1] "estadístico" [1] "z= 2.46923411637166" "===== [1] "p-valor= 0.0135402599797826" [1] "conclusión:" [1] "la diferencia es significativamente distinta que/de 0 a un nivel del 5 % de significación" "Contraste z para la diferencia de dos medias" ===== [1] "H0 : la diferencia = 1" [1] "H1: mayor que/de 1" ===== [1] "estadístico" [1] "z= 0.345491977535859" ===== " [1] "p-valor= 0.364862270207122" [1] "conclusión:" [1] "la diferencia NO es significativamente mayor que/de 1 a un nivel del 5 % de significación" "Contraste z para la diferencia de dos medias" ===== [1] "H0 : la diferencia = 0" [1] "H1: distinta que/de 0" ===== [1] "estadístico" [1] "z= 2.46923411637166" [1] "===== [1] "p-valor= 0.0135402599797826" [1] "conclusión:" [1] "la diferencia NO es significativamente distinta que/de 0 a un nivel del 1 % de significación"	

ANEXO2 : FUNCIÓN p.test1

Utilizando una función **diferente** ad-hoc, podemos automatizar el contraste para una proporción con R según el esquema anterior basado en la distribución normal. Veamos cómo podemos construir una función ad-hoc que hemos llamado "p.test1", para este contraste,

```
##función para el contraste z sobre una proporción
## p.test1= function(p, n, p0,alternativa.1="distinta",sign=0.05 )
## p= proporción muestral
## n= tamaño muestral
## p0= valor de la proporción según la Hip.nula
## alternativa.1 = "distinta" / "mayor" / "menor" ( distinta ,por defecto)
```

```

## sign= nivel de signif. para el contraste ( 0.05, por defecto)
#####
p.test1= function(p, n, p0,alternativa.1="distinta",sign=0.5 ){
  zeta = ((p - p0) / 0.5)*sqrt(n)
  if(alternativa.1=="mayor")
    pv=pnorm(zeta,0,1,lower.tail = FALSE)
  else
    if (alternativa.1=="menor")
      pv=pnorm(zeta,0,1)
    else
      pv=pnorm(abs(zeta),0,1,lower.tail=FALSE)*2
  print("Contraste z para una proporción")
  print("=====")
  print(paste("H0 : p =",p0))
  print (paste ("H1: p es", alternativa.1,"que/de",p0))
  print("=====")
  print(paste("proporción muestral=",p))
  print("estadístico")
  print (paste ("z=",zeta))
  print("===== ")
  print(paste("p-valor=",pv))
  print("conclusión:")
  if (pv<sign) print(paste("la proporción es significativamente",alternativa.1,
    "que/de",p0,"a un nivel del ",
    sign*100,"% de significación"))
  else
    print(paste("proporción NO es significativamente",alternativa.1,
    "que/de",p0,"a un nivel del ",
    sign*100,"% de significación"))
}

```

Ejemplo 6. Con los datos de la base de datos “alumnos” seleccionamos, al azar, 50 alumnos. Con el resultado obtenido contrastar que el porcentaje poblacional de mujeres es del 60% al 0.01 de significación.

```

set.seed(144)
seleccion<-sample(240,50)

muestra<-alumnos[seleccion,]

table(muestra$sexo)#frecuencias de hombres(0) y mujeres (1)
p<-table(muestra$sexo)[2]/50 # proporción muestral de mujeres
p.test1(p,50,0.6,sign=0.01)

"Contraste z para una proporción"
[1] "===== "
[1] "H0 : p = 0.6"
[1] "H1: p es distinta que/de 0.6"
[1] "===== "
[1] "proporción muestral= 0.52"
[1] "estadístico"
[1] "z= -1.13137084989848"
[1] "===== "
[1] "p-valor= 0.25789903529234"
[1] "conclusión:"
[1] "proporción NO es significativamente distinta que/de 0.6 a un nivel del 1 % de significación".

```

Anexo 3 la función p.test2

Utilizando una función ad-hoc, podemos automatizar el contraste para la diferencia de dos proporciones con R según el esquema anterior basado en la distribución normal. Veamos, cómo podemos construir una función ad-hoc que hemos llamado “p.test2”, para este contraste:

```
##función para el contraste z sobre la diferencia de dos proporciones
## p.test2= function(p1, p2,n1,n2, p0,alternativa.1="distinta",sign=0.05 )
## p1= proporción muestral 1 ## n1= tamaño muestral 1
## p2=proporción muestral 2 ## n2= tamaño muestral 2
## p0= valor de la diferencia de proporciones según la Hip.nula
## alternativa.1 = "distinta" / "mayor" / "menor" ( distinta ,por defecto)
## sign= nivel de signif. para el contraste ( 0.05, por defecto)

p.test2= function(p1, p2,n1,n2, p0=0,alternativa.1="distinta",sign=0.5 ){
  zeta = ((p1-p2) - p0) / sqrt((0.25/n1)+(0.25/n2))
  if(alternativa.1=="mayor")
    pv=pnorm(zeta,0,1,lower.tail = FALSE)
  else
    if (alternativa.1=="menor")
      pv=pnorm(zeta,0,1)
    else
      pv=pnorm(abs(zeta),0,1,lower.tail=FALSE)*2
  print("Contraste z para la diferencia de dos proporciones")
  print("=====")
  print(paste("H0 : p1-p2 =",p0))
  print (paste ("H1: p1-p2 es", alternativa.1,"que/de",p0))
  print("=====")
  print(paste("la diferencia muestral=",p1-p2))
  print("estadístico")
  print (paste ("z=",zeta))
  print("=====")
  print(paste("p-valor=",pv))
  print("conclusión:")
  if (pv<sign) print(paste("la diferencia es significativamente",alternativa.1,
    "que/de",p0,"a un nivel del ",
    sign*100,"% de significación"))
  else
    print(paste("diferencia NO es significativamente",alternativa.1,
    "que/de",p0,"a un nivel del ",
    sign*100,"% de significación"))
}
```

Ejemplo 7. Sobre la base de datos anterior (alumnos) suponemos que los datos constituyen ya una muestra . De hecho podemos considerar dos muestras una de las mujeres de tamaño 124 y otras de hombres de tamaño 116. A partir de estas muestras contrastar que la proporción de mujeres que trabajan es igual a la de hombres. (nivel de significación del 2%)

```
p1<-table(chicas$trabajo)[2]/124
```

```

p2<-table(chicos$trabajo)[2]/116
p.test2(p1,p2,124,116,sign=0.02)
"Contraste z para la diferencia de dos proporciones"
[1] "===== "
[1] "H0 : p1-p2 = 0"
[1] "H1: p1-p2 es distinta que/de 0"
[1] "===== "
[1] "la diferencia muestral= -0.140155728587319"
[1] "estadístico"
[1] "z= -2.17007660706573"
[1] "===== "
[1] "p-valor= 0.0300010430065497"
[1] "conclusión:"
[1] "diferencia NO es significativamente distinta que/de 0 a un nivel
del 2 % de significación"

p.test2(p1,p2,124,116,alternativa.1 = "menor", sign=0.02)

[1] "Contraste z para la diferencia de dos proporciones"
[1] "===== "
[1] "H0 : p1-p2 = 0"
[1] "H1: p1-p2 es menor que/de 0"
[1] "===== "
[1] "la diferencia muestral= -0.140155728587319"
[1] "estadístico"
[1] "z= -2.17007660706573"
[1] "===== "
[1] "p-valor= 0.0150005215032748"
[1] "conclusión:"
[1] "la diferencia es significativamente menor que/de 0 a un nivel de
1 2 % de significación"

```

Vemos, que aunque no se puede rechazar (al 2% de significación) que sean iguales las proporciones, sí podemos concluir que la proporción que trabaja ente las mujeres es significativamente menor (también al 2 %)