Tutorial 09: Inferencia sobre dos poblaciones.

Atención:

- Este documento pdf lleva adjuntos algunos de los ficheros de datos necesarios. Y está pensado para trabajar con él directamente en tu ordenador. Al usarlo en la pantalla, si es necesario, puedes aumentar alguna de las figuras para ver los detalles. Antes de imprimirlo, piensa si es necesario. Los árboles y nosotros te lo agradeceremos.
- Fecha: 10 de septiembre de 2015. Si este fichero tiene más de un año, puede resultar obsoleto. Busca si existe una versión más reciente.

Índice

Introducción.

Este tutorial contiene la parte práctica del Capítulo 9 del libro. Como hemos dicho en la introducción de ese capítulo, las ideas básicas (intervalos de confianza y contrastes) resultan ya conocidas, y aquí de lo que se trata es de aplicarlas al caso en el que estudiamos una misma variable aleatoria en dos poblaciones independientes. Las operaciones necesarias, en la práctica, son muy parecidas a las que hemos visto en anteriores capítulos y tutoriales para el caso de una población. Para usar el ordenador en estos problemas sólo necesitamos, en principio, ser capaces de resolver problemas de probabilidad (directos e inversos) en las distribuciones correspondientes: la normal Z, la t de Student, la χ^2 y, como novedad de este capítulo, la F de Fisher. Cualquiera de las herramientas con las que nos hemos familiarizado podría ser suficiente para este trabajo: R, por supuesto, o GeoGebra, Wolfram Alpha, incluso Calc permita calcular los valores de cualquiera de esas distribuciones. Pero, como ya sabe el lector, algunas herramientas son mucho más cómodas y fiables que otras. En este tutorial se incluye, en la Tabla 1 (pág. 44), una colección de *plantillas* de R que, junto con los de los anteriores tutoriales, cubren todos los casos que aparecen en las Tablas del Apéndice B del libro. Ilustraremos el uso de cada una de estas plantillas con un ejemplo detallado y aprovecharemos algunos de esos ejemplos para abordar el uso de otras herramientas, como la *Calculadora de Probabilidades* de GeoGebra.

1. Diferencia de proporciones en dos poblaciones.

Ver la Sección 9.1 del libro (pág. 296).

Usando la plantilla de R.

Vamos a usar las plantillas de R de la la Tabla 1 para obtener los resultados que aparecen en el Ejemplo 9.1.1 del libro (pág. 300). Recordemos que en ese ejemplo se trata de contrastar la hipótesis nula:

$$H_0 = \{p_1 = p_2\},$$

Y que para ello se han obtenido dos muestras independientes de tamaños $n_1 = 456$ y $n_2 = 512$ en las que los respectivos números de *"éxitos"* eran 139 y 184, con lo que las proporciones muestrales de éxitos son:

$$\hat{p}_1 = \frac{139}{456} \approx 0.3048, \quad \hat{p}_2 = \frac{184}{512} \approx 0.3594,$$

mientras que las proporciones de *"fracasos"* son:

$$\hat{q}_1 \approx 0.6952, \quad \hat{q}_2 \approx 0.6406.$$

El cálculo del p-valor de este contraste se obtiene muy fácilmente con el fichero plantilla:

Tut09-Contraste-2Pob-DifProporciones-UsandoZ.R

de la la Tabla 1. Incluimos aquí sólo la parte inicial del fichero, en la que hemos introducido los datos de este ejemplo. Fíjate especialmente en que las proporciones muestrales se introducen como cocientes, no mediante el número de éxitos. Esto se ha hecho así por si, en algún caso, el enunciado del problema contiene directamente la proporción sin mencionar explícitamente el número de éxitos:

```
# PRIMERA MUESTRA
# Numero de elementos
(n1 = 456)

## [1] 456

# proporcion muestral
(pMuestral1 = 139/456)

## [1] 0.30482

# SEGUNDA MUESTRA
# Numero de elementos
(n2 =  512)

## [1] 512

# proporcion muestral
(pMuestral2 = 184/512)

## [1] 0.35938

# ¿Que tipo de contraste estamos haciendo?
# Escribe 1 si la HIP. ALTERNATIVA es p1 > p2, 2 si es p1 < p2, 3 si es bilateral
TipoContraste = 3
  #Nivel de significacion
  (nSig = 0.95)

## [1] 0.95
```

El final del fichero plantilla contiene las instrucciones que producen los resultados del contraste (no incluimos la región de rechazo porque no la vamos a usar):

```
pValor(Estadistico,TipoContraste)

## [1] "El p-Valor es 0.0723854663297254"

Estadistico

## [1] -1.7967
```

Como puede verse, el p-valor coincide con lo que aparece en ese ejemplo.

Usando la función `prop.test`

Esta función, que ya conocimos en el Tutorial08, sirve también para este tipo de contrastes. Para el Ejemplo 9.1.1 del libro que acabamos de calcular, el comando a ejecutar sería:

```
prop.test(c(139, 184), c(456, 512), correct=FALSE,
    alternative="two.sided", conf.level=0.95)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(139, 184) out of c(456, 512)
## X-squared = 3.23, df = 1, p-value = 0.072
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1138167  0.0047159
## sample estimates:
##  prop 1  prop 2
## 0.30482 0.35938
```

Como ves:

- Se introducen dos vectores conteniendo cada uno de ellos, respectivamente, los éxitos y los tamaños muestrales. ¡Cuidado con este formato!
- La hipótesis alternativa se indica, como en otros casos, eligiendo entre `less` para $H_a = \{p_1 < p_2\}$, `greater` para $H_a = \{p_1 > p_2\}$ y `two.sided` para $H_a = \{p_1 \neq p_2\}$.
- Es necesario incluir la opción `correct=FALSE` si queremos obtener el mismo resultado que con la plantilla. De lo contrario, R aplica una *corrección de continuidad* para mejorar la aproximación de la binomial por la normal.
- Por último, como producto secundario del contraste bilateral obtenemos un intervalo de confianza para $p_1 - p_2$, al nivel de confianza que hayamos indicado.

Vamos a usar ese intervalo de confianza como excusa para proponerte un ejercicio.

Ejercicio 1.

1. *Usa el fichero plantilla de R de la Tabla 1 (pág. 44) para obtener este mismo intervalo de confianza.*
2. *Haz lo mismo usando la pestaña* Estadísticas *de la* Calculadora de Probabilidades de GeoGebra. *La opción que tienes que usar tiene un nombre poco claro: se llama* Z estimada, diferencia de proporciones. *Luego usa el comando:*

   ```
   IntervaloProporcionesZ[ <Proporción (muestra 1)>, <Tamaño (muestra 1)>,
   <Proporción (muestra 2)>, <Tamaño (muestra 2)>, <Nivel> ]
   ```

para hacer la misma cuenta directamente.

3. *En Wolfram Alpha puedes teclear* `two proportion confidence interval` *para llegar a una interfaz web en la que hacer este tipo de cálculos.*

□

2. Diferencia de medias en dos poblaciones, muestras grandes.

Para ilustrar este tipo de situaciones, vamos a usar un ejemplo relacionado con el que abría el Capítulo 7 del libro.

Los dos laboratorios han seguido trabajando, y ahora tenemos dos tratamientos de segunda generación para aliviar la depresión en los canguros, el *Saltaplus Extraforte* y el *Pildorín con Ginseng*. Para establecer cuál de los dos tratamientos es superior, los hemos usado para tratar a los canguros deprimidos de dos muestras independientes, midiendo la altura media de sus saltos en metros. Llamando μ_1 a la altura media (en metros) de los canguros tratados con *Saltaplus* y μ_2 a la altura media de los tratados con *Pildorín*, queremos contrastar la hipótesis (alternativa):

$$H_a = \{\mu_1 < \mu_2\},$$

que sostiene que la nueva versión de *Pildorín* es mejor que el *Saltaplus* renovado. Los datos muestrales son estos (la muestra 1 corresponde a *Saltaplus*, la 2 a *Pildorín*):

$$\begin{cases} n_1 = 245 \\ \bar{X}_1 = 2.73 \\ s_1 = 0.4 \end{cases} \qquad \begin{cases} n_2 = 252 \\ \bar{X}_2 = 2.81 \\ s_2 = 0.3 \end{cases}$$

Como las dos muestras son grandes, para hacer este contraste podemos usar la plantilla

Tut09-Contraste-2Pob-DifMedias-UsandoZ.R

Incluimos los datos del problema en las primeras líneas de este fichero, como se muestra aquí. Fíjate en que hemos usado, descomentándolas, las líneas de s_1 y s_2:

```
# PRIMERA MUESTRA
# Numero de elementos
    (n1 = 245)

## [1] 245

# Media muestral
    (xbar1 = 2.73)

## [1] 2.73

# Cuasidesviacion tipica muestral o sigma (descomenta el que uses)
    (s1 = 0.4)

## [1] 0.4

    #(sigma1 = )


# SEGUNDA MUESTRA
# Numero de elementos
    (n2 = 252)

## [1] 252
```

```
# Media muestral
    (xbar2 = 2.81)

## [1] 2.81

# Cuasidesviacion tipica muestral o sigma (descomenta el que uses)
    (s2 = 0.3)

## [1] 0.3

    #(sigma2 = )

# ¿Que tipo de contraste estamos haciendo?
    # Escribe 1 si la HIP. ALTERNATIVA es mu1 > mu2, 2 si es mu1 < mu2, 3 si es mu1 distinto de mu2
    TipoContraste = 2
#Nivel de significacion
    (nSig = 0.95)

## [1] 0.95
```

Los resultados de la ejecución del fichero son (de nuevo, excluimos la región de rechazo):

```
    pValor(Estadistico, TipoContraste)

## [1] "El p-Valor es 0.00591772613290591"

    Estadistico

## [1] -2.517
```

Con ese p-valor, rechazaríamos la hipótesis nula, de forma que no hay base experimental para creer que los canguros tratados con *Saltaplus* saltan más que los tratados con *Pildorín*.

Vamos a aprovechar este ejemplo para explorar otras herramientas con las que puedes hacer este tipo de contrastes y los intervalos de confianza asociados:

Ejercicio 2.

1. *Usa el fichero plantilla de R*

   Tut09-IntConf-2Pob-DifMedias-UsandoZ.R

   de la la Tabla 1 (pág. 44) para obtener un intervalo de confianza al 95 % *para la diferencia* $\mu_1 - \mu_2$.

2. *Haz lo mismo con la* Calculadora de Probabilidades de GeoGebra. *En este caso debes usar* Z estimada, diferencia de medias. *También puedes hacerlo directamente con el comando:*

   ```
   IntervaloMediasZ[ <Media (muestra 1)>, <s1>, <Tamaño (muestra 1)>,
       <Media (muestra 2)>, <s2>, <Tamaño (muestra 2)>, <Nivel> ]
   ```

3. *Volviendo al contraste de hipótesis, en Wolfram Alpha puedes teclear* `hypothesis test for the difference between two means` *para llegar a una interfaz web con la que hacer contrastes de diferencias de medias usando* Z. *Si usas* `confidence interval for the difference between two means` *podrás calcular intervalos de confianza para* $\mu_1 - \mu_2$ *usando* Z.

4. *Usa cualquiera de estos métodos (aún mejor, varios de ellos) para comprobar las cuentas del Ejemplo 9.2.1 del libro (pág. 305). A pesar de que en ese ejemplo disponemos de los datos, se trata de que uses los valores* $n_1, n_2, \bar{X}_1, \bar{X}_2, s_1, s_2$ *que aparecen en el texto del ejemplo. Más adelante en el tutorial volveremos sobre el cálculo a partir de los* datos en bruto.

Soluciones en la página 30. □

¿Y el caso de *datos en bruto*? Advertencia sobre `data.frames`

No hemos incluido ficheros plantilla para el caso de datos en bruto. ¿Por qué? Bueno, una posibilidad sería cargar los datos de cada una de las muestras desde un fichero `csv`, uno para cada muestra. Pero eso resultaría muy forzado y artificioso. La práctica habitual (y recomendable) en estadística es usar para esto un único fichero con dos columnas. Cada fila de ese fichero corresponde a una observación. Una de las columnas contiene los valores de la variable X. La otra es un factor F con dos niveles que identifica a cuál de las poblaciones pertenece esa observación. Por ejemplo, el comienzo del fichero podría tener un aspecto similar al de esta tabla:

X	F
7.35	A
8.23	A
7.75	B
⋮	⋮

La primera columna contiene los valores de X, mientras que la segunda permite conocer a cuál de las dos poblaciones pertenece ese valor (en este ejemplo, identificadas respectivamente por los niveles A y B del factor F). La estructura de datos natural para trabajar con este tipo de ficheros en R es el `data frame` del que hemos hablado por primera vez en el Tutorial04. Y para gestionar de forma adecuada un `data.frame` que contenga un fichero como el que estamos describiendo, es preciso usar factores de R, de los que hemos hablado en la Sección **??** del Tutorial08 (pág. **??**). Por otra parte, en el Capítulo 11, al hablar del Anova unifactorial, nos vamos a encontrar con una generalización natural de los problemas que estamos tratando en este capítulo. Así que podemos posponer parte de la discusión sobre la mejor forma de gestionar esos datos hasta ese capítulo. Pero no es menos cierto que R incluye algunas funciones interesantes para trabajar con datos *en bruto*, específicamente dedicadas a los problemas de este capítulo, los de dos poblaciones. Por eso vamos a incluir en la Sección 5 de este tutorial (pág. 22) la discusión de esas funciones. Advertencia: el lector que no haya practicado el uso de *data.frames* en los tutoriales anteriores tendrá algunos problemas para entender el código que se usa con esa funciones. En cualquier caso, recuerda que usando un editor de texto (como el *Bloc de Notas*) y una hoja de cálculo como *Calc*) puedes manipular los ficheros y en la mayoría de los casos extraer así la información necesaria.

2.1. El caso de datos emparejados.

El caso de datos emparejados se describe en la Sección 9.2.2 del libro (pág. 312). En este apartado sólo queremos destacar que, como hemos dicho allí, no hay nada nuevo en realidad en esa situación, porque en realidad se trata de un contraste en una única población, como los que hemos aprendido a realizar en el Capítulo 7 y en el tutorial que lo acompaña. Para evidenciar esto vamos a realizar los cálculos necesarios para el Ejemplo 9.2.3 del libro y usaremos una plantilla del Tutorial07. Concretamente, la plantilla titulada

`Tut07-Contraste-Media-UsandoT-DatosEnBruto.R`

en la que únicamente es necesario hacer una pequeña modificación para acomodar el hecho de que ahora tenemos datos antes y después del tratamiento. El código de esa plantilla, con los datos necesarios, aparece a continuación. Fíjate en que hemos añadido dos líneas al bloque inicial para definir los vectores `antes` y `despues`, y que los hemos usado para obtener los valores del vector Y del libro mediante

```
(muestra = despues - antes)
```

En particular ten en cuenta que lo que en libro se denomina $\bar{Y}$, en el código será `xbar`. El resto de las adaptaciones del código deberían resultar evidentes. Revisa el código, cotejando los valores que se obtienen con los que aparecen en el libro.

```
#####################################################
# www.postdata-statistics.com
# POSTDATA. Introducción a la Estadísitica
# Tutorial-07.
```

```r
#
# Fichero de instrucciones R para calcular
# un contraste de hipotesis para la media de una
# poblacion normal N(mu,sigma), a partir de
# un fichero con una muestra de esa poblacion.
#
# El fichero no funcionara si no introduces todos los datos.
# Además tendrás que descomentar algunas lineas para elegir
# la forma en la que lees los datos.
#
#####################################################################


#####################################################################
# CASO:  sigma desconocida, muestra pequeña  n<30.
#####################################################################

rm(list = ls())

antes = c(1.80, 2.48, 2.33, 3.28, 1.24, 2.49, 2.44, 2.54, 2.59, 3.90)
despues = c(3.31, 2.33, 2.65, 2.16, 1.62, 3.15, 2.14, 4.01, 2.42, 2.91)

# Una posibilidad es que tengas la muestra como un vector.

(muestra = despues - antes)

##  [1]  1.51 -0.15  0.32 -1.12  0.38  0.66 -0.30  1.47 -0.17 -0.99

# Si lees la muestra de un fichero csv:

# 1. Recuerda seleccionar el directorio de trabajo.

# 2. Ahora introduce entre las comillas el nombre del fichero, y el tipo de separador, etc.

#muestra = scan(file="",sep=" ",dec=".")


# Valor a contrastar de la media (aparece en la hipotesis nula)

(mu0 = 0)

## [1] 0

# ¿Que tipo de contraste estamos haciendo?
# Escribe 1 si la HIP. ALTERNATIVA es mu > mu0, 2 si es mu < mu0, 3 si es mu distinto de mu0

(TipoContraste = 1)

## [1] 1

##Nivel de significacion

(nSig = 0.95)

## [1] 0.95

################################################
# NO CAMBIES NADA DE AQUÍ PARA ABAJO
```

```
################################################
```

```
(alfa = 1 - nSig)
```

```
## [1] 0.05
```

```
# Numero de elementos en la muestra
```

```
(n = length(muestra))
```

```
## [1] 10
```

```
# Grados de libertad
```

```
(k = n - 1)
```

```
## [1] 9
```

```
# Media muestral
```

```
(xbar = mean(muestra))
```

```
## [1] 0.161
```

```
# Cuasidesviacion tipica muestral
```

```
(s = sd(muestra))
```

```
## [1] 0.89691
```

```
# Calculo del estadistico del contraste
```

```
(Estadistico = (xbar - mu0) / (s/sqrt(n)))
```

```
## [1] 0.56764
```

```
# Funcion para el calculo del p-valor
```

```
 pValor = function(EstadCon, tipoCon){
      if(tipoCon == 1){
        (pV = 1 - pt(EstadCon, df=k ))
      }
      if(tipoCon == 2){
        (pV = pt(EstadCon, df=k ))
      }
      if(tipoCon == 3){
        pV = 2 * (1 - pt(abs(EstadCon), df=k ))
      }
      return(paste0("El p-Valor es ", pV, collapse=""))
    }
```

```
# Funcion para el calculo del límite de la región de rechazo
```

```
    RegionRechazo = function(alfa, tipoCon){
      if(tipoCon == 1){
        (regionRech = paste("mayores que ",
                            qt(1 - alfa, df=k)))
```

```
      }
      if(tipoCon == 2){
        (regionRech = paste("menores que ",
                            qt(alfa, df=k)))
      }
      if(tipoCon == 3){
        (regionRech = paste("mas alejados del origen que ",
                            qt(1 - (alfa/2), df=k)))
      }
      regionRech = paste0("La region de rechazo la forman los valores del Estadistico ",
                          regionRech, collapse="")
      return(regionRech)
    }

# Y ahora se aplican ambas funciones para mostrar los resultados

pValor(Estadistico, TipoContraste)

## [1] "El p-Valor es 0.292078879999332"

paste0("El valor del estadístico es ", Estadistico, collapse = "")

## [1] "El valor del estadístico es 0.56764281922141"

RegionRechazo(alfa, TipoContraste)

## [1] "La region de rechazo la forman los valores del Estadistico mayores que  1.8331129326562
```

3. Cociente de varianzas en dos poblaciones normales. Distribución F de Fisher-Snedecor.

Como hemos discutido en la Sección 9.2 del libro (pág. 303), cuando las muestras son pequeñas (y, como suele ocurrir, las varianzas poblacionales son desconocidas), el contraste de diferencias de las medias nos conduce a un contraste de cociente de varianzas como paso previo para decidir si estamos en el caso (c) o en el caso (d) de los casos que aparecen en esa Sección.

Vamos, por tanto, a aprender primero a hacer un contraste sobre el cociente de varianzas, antes de retornar a los contrastes de diferencia de medias. Y para eso tenemos que aprender más sobre la forma de trabajar con la distribución de Fisher en el ordenador.

3.1. La distribución F de Fisher.

En R.

Muy brevemente: en R disponemos de las funciones pf y qf, con el comportamiento esperable. La única novedad es que para trabajar con la distribución $F_{k_1;k;2}$ debemos indicarlo mediante los argumentos opcionales `df1` y `df2` de esas funciones de R. Por ejemplo, para calcular la probabilidad

$$P(F_{13;8} > 3)$$

haríamos

```
1 - pf(3, df1=13, df2=8)

## [1] 0.062372
```

o también

```r
pf(3, df1=13, df2=8, lower.tail=FALSE)

## [1] 0.062372
```

Y para calcular el valor K tal que

$$P(F_{7;9} < K) = 0.975$$

haríamos

```r
qf(0.975, df1=7, df2=9)

## [1] 4.197
```

¡Es muy importante recordar que no podemos cambiar el orden de los valores de `df1` y `df2`! Las distribuciones de Fisher $F_{k1;k2}$ y $F_{k2;k1}$, aunque relacionadas, son distintas.

En GeoGebra.

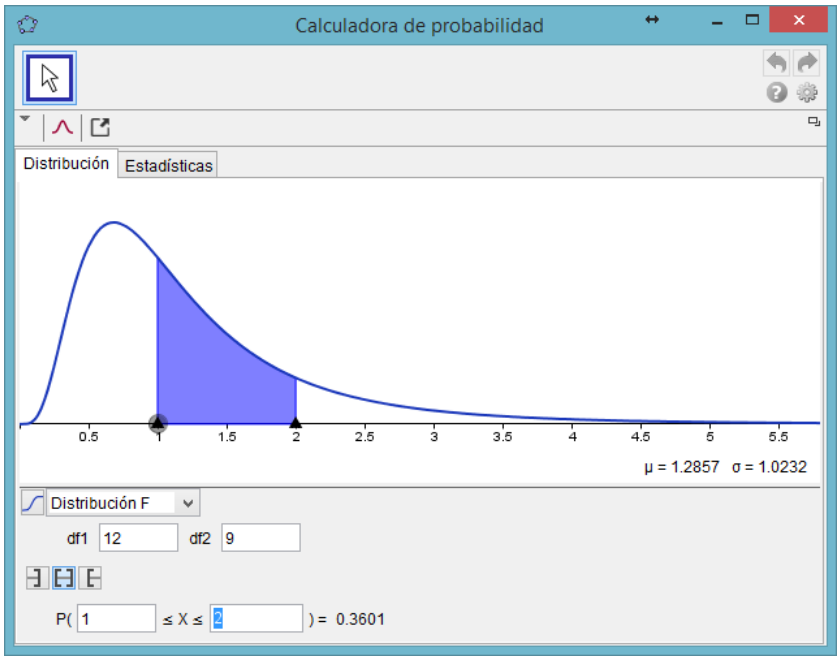
Para trabajar con la distribución de Fisher en GeoGebra podemos usar los comandos `DistribuciónF` y `DistribuciónFInversa` que, como sugieren los nombres, permiten resolver, respectivamente, problemas directos e inversos de probabilidad que involucren a la F de Fisher. Por ejemplo, para resolver el problema

$$P(1 < F_{12;9} < 2)$$

basta con ejecutar

```
DistribuciónF[12, 9, 2] - DistribuciónF[12, 9, 1]
```

y se obtiene, aproximadamente, 0.3601. Naturalmente, también podemos usar la *Calculadora de Probabilidades*, como se muestra en la siguiente figura que ilustra ese mismo cálculo de la probabilidad.



Ejercicio 3. *Repite con GeoGebra los cálculos de probabilidades (directas e inversas) que hemos hecho antes con R.* *Solución en la página 36.* □

En Wolfram Alpha y Calc.

Para trabajar en Wolfram Alpha puedes usar comandos como los de estos dos ejemplos que, con ligeras modificaciones, cubren todas nuestras necesidades. Para un problema directo usamos algo como esto:

```
P(X > 3) for X ~ F(13,8)
```

y para un problema inverso, por ejemplo para calcular el valor K tal que

$$P(F_{12;16} < K) = 0.975$$

usaríamos este comando:

```
97.5th percentile for F(12, 16)
```

¡Ten en cuenta que la probabilidad se ha traducido en percentiles!

Y, finalmente, no queremos dejar de mencionar las funciones `DISTR.F` y `DISTR.F.INV` de Calc, que permiten trabajar con esta distribución en la hoja de cálculo.

3.2. Contrastes e intervalos de confianza sobre cocientes de varianzas.

Ahora que ya sabemos cómo trabajar con la distribución F de Fisher podemos usarla para hacer contrastes de hipótesis e intervalos de confianza relativos al cociente de varianzas. Recuerda que el estadístico adecuado para esos contrastes es:

$$\Xi = \frac{s_1^2}{s_2^2},$$

y que en la Tabla B.4 del libro (pág. 580) tienes la información necesaria para saber cómo usar el valor del estadístico Ξ^2 para calcular el p-valor del contraste.

Antes de hacer algunos ejemplos, unas observaciones genéricas sobre las herramientas de las que disponemos:

- A nuestro juicio, y para las versiones actuales del software que usamos, la opción más ventajosa para hacer este tipo de contrastes con la menor cantidad de errores es usar la plantilla de R que hemos incluido en la Tabla 1 de este tutorial (pág. 44).
- Siguiendo con R, la función `var.test` es especialmente interesante si trabajamos con muestras *en bruto*.
- En GeoGebra, la *Calculadora de Probabilidades* no permite hacer este tipo de contrastes y tampoco hay un comando que se pueda usar directamente en la *Línea de Entrada* o el panel de *Cálculo Simbólico*. A fecha de hoy, la única forma de hacer este contraste es calculando directamente el p-valor mediante un problema directo de probabilidad con la F de Fisher. En Wolfram Alpha, hasta donde sabemos, sucede algo similar: no hay una herramienta específica para este tipo de contrastes.

Un ejemplo básico de contrastes de cocientes de varianzas.

Vamos a supone que estamos estudiando una variable X en dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ y queremos contrastar la hipótesis alternativa bilateral:

$$H_a = \{\sigma_1^2 = \sigma_2^2\}.$$

Para ello hemos tomado muestras aleatorias independientes en cada una de las poblaciones y hemos obtenido estos valores muestrales:

$$\begin{cases} n_1 = 59 \\ s_1 = 3.1 \end{cases} \qquad \begin{cases} n_2 = 64 \\ s_2 = 4.5 \end{cases}$$

Para hacer este contraste de la forma más rápida posible, lo más recomendable es usar la plantilla de R de la Tabla 1. Incluimos aquí las primeras líneas de esa plantilla con los datos que debes introducir:

```
# PRIMERA MUESTRA
# Numero de elementos
(n1 = 59)

## [1] 59

# Cuasidesviacion tipica muestral
(s1 = 3.1)

## [1] 3.1

# SEGUNDA MUESTRA
# Numero de elementos
(n2 =  64)

## [1] 64

# Cuasidesviacion tipica  muestral
(s2 = 4.5)

## [1] 4.5

# TIPO DE CONTRASTE:
# Escribe 1 si la HIP. ALTERNATIVA es sigma > sigma2,
#         2 si es sigma1 < sigma2,
#         3 si es bilateral.
TipoContraste = 3

#NIVEL DE SIGNIFICACION:
(nSig = 0.95)

## [1] 0.95
```

Y los resultados que se obtienen al ejecutar el fichero son:

```
    pValor(Estadistico,TipoContraste)

## [1] "El p-Valor es 0.00459021398523596"

    Estadistico

## [1] 0.47457
```

Así que, por ejemplo, para un nivel de significación del 99 % rechazaríamos la hipótesis nula y concluiríamos que los datos no permiten afirmar que las varianzas sean iguales.

Y un intervalo de confianza.

Análogamente, la forma más rápida de obtener elintervalo de confianza es usando la plantilla que aparece al final de este tutorial, en la Tabla 1. Vamos a usarla para calcular un intervalo de confianza al 95 % para los mismos datos que acabamos de usar para el contraste. El código de la plantilla para ese ejemplo es este:

```
#####################################################
# www.postdata-statistics.com
# POSTDATA. Introducción a la Estadísitica
# Tutorial-09.
#
# Fichero de instrucciones R para calcular un
# INTERVALO DE CONFIANZA PARA EL COCIENTE DE VARIANZAS
# al nivel (1-alfa)   en dos poblaciones normales.
#
# El fichero no funcionara si no introduces todos los datos.
#####################################################


# Introducimos los valores de las desviaciones tipicas muestrales,
s1 = 3.1
s2 = 4.5


# los tamaños de las muestras,
n1 = 59
n2 = 64

# y el nivel de confianza deseado.
nc = 0.95

## --- NO CAMBIES NADA DE AQUI PARA ABAJO

(alfa = 1 - nc)

## [1] 0.05

# Calculamos los valor criticos necesarios:

(f_alfamedios = qf(alfa/2, df1=n1 - 1, df2=n2 - 1))

## [1] 0.59935

(f_unomenosalfamedios = qf(1 - alfa/2, df1=n1 - 1, df2= n2-1))

## [1] 1.6594

# El intervalo de confianza para el cociente de varianzas es este:
(intervalo = c( (1/f_unomenosalfamedios), (1/f_alfamedios)) * (s1^2/s2^2))

## [1] 0.28598 0.79180
```

Podemos aprovechar este cálculo para confirmar las conclusiones del contraste: puesto que el intervalo no contiene al 1, estamos en condiciones de rechazar H_0 al 95 %.

4. Diferencia de medias en dos poblaciones, muestras pequeñas.

4.1. Los contrastes de los ejemplos de la Sección 9.3.1 del libro.

Vamos a empezar mostrando como comprobar los datos de esos ejemplos usando R. En todos los casos es necesario realizar un contraste previo de varianzas, para luego pasar al contraste de

diferencia de medias. La forma más rápida de proceder es usando las plantillas de R. Concretamente, usaremos la plantilla

`Tut09-Contraste-2Pob-CocienteVarianzas.R`

para los contrastes sobre cocientes de varianzas y después usaremos una de las plantillas

`Tut09-Contraste-2Pob-DifMedias-UsandoT-VarDistintas.R`
`Tut09-Contraste-2Pob-DifMedias-UsandoT-VarIguales.R`

Ejemplo 9.3.1

Empezamos por este ejemplo, que aparece en la página 319 del libro. Allí puedes ver los valores necesarios, así que sólo mostraremos el principio del código de la plantilla que usamos para el contraste de varianzas. Ten en cuenta que puede haber pequeños discrepancias con respecto a los valores del libro debidos al redondeo, porque aquí no estamos tomando como partida los datos en bruto que aparecen en el ejemplo:

```
 # PRIMERA MUESTRA
 # Numero de elementos
 (n1 = 10)

## [1] 10

 # Cuasidesviacion tipica muestral
 (s1 = 2.098)

## [1] 2.098

 # SEGUNDA MUESTRA
 # Numero de elementos
 (n2 = 10)

## [1] 10

 # Cuasidesviacion tipica  muestral
 (s2 = 2.111)

## [1] 2.111

 # TIPO DE CONTRASTE:
 # Escribe 1 si la HIP. ALTERNATIVA es sigma > sigma2,
 #         2 si es sigma1 < sigma2,
 #         3 si es bilateral.
 TipoContraste = 3

 #NIVEL DE SIGNIFICACION:
 (nSig = 0.95)

## [1] 0.95
```

Y los resultados que obtenemos:

```
 # Y ahora se aplican ambas funciones para mostrar los resultados
     pValor(Estadistico,TipoContraste)

## [1] "El p-Valor es 0.985618870598065"
```

```
     Estadistico
```

```
## [1] 0.98772
```

Como puedes ver y salvo la pequeña discrepancia numérica, confirmamos la conclusión que aparece en el texto: no tenemos razones para pensar que las varianzas sean distintas. Así que de las dos posibles usamos la plantilla `Tut09-Contraste-2Pob-DifMedias-UsandoT-VarIguales.R`. Vamos a ver la parte inicial del código de esa plantilla, con los datos del problema. Ten en cuenta, insistimos, que puede haber pequeñas discrepancias numéricas con los valores que aparecen en el libro. Además en este ejemplo estamos llamando μ_t, μ_b a lo que normalmente llamamos $\{\mu_1, \mu_2\}$. Ten presente esto a la hora de elegir el tipo de contraste.

```
 # PRIMERA MUESTRA # Numero de elementos
 (n1 = 10)
```

```
## [1] 10
```

```
 # Media muestral
 (xbar1 = 94.2)
```

```
## [1] 94.2
```

```
 # Cuasidesviacion tipica muestral
 (s1 = 2.098)
```

```
## [1] 2.098
```

```
 # SEGUNDA MUESTRA
 # Numero de elementos
 (n2 = 10)
```

```
## [1] 10
```

```
 # Media muestral
 (xbar2 = 97.7)
```

```
## [1] 97.7
```

```
 # Cuasidesviacion tipica  muestral
 (s2 = 2.111)
```

```
## [1] 2.111
```

```
 # ¿Que tipo de contraste estamos haciendo?
 # Escribe 1 si la HIP. ALTERNATIVA es mu1 > mu2, 2 si es mu1 < mu2, 3 si es mu1 distinto de mu
     TipoContraste = 2
 #Nivel de significacion
     (nSig = 0.95)
```

```
## [1] 0.95
```

Los resultados son:

```
     pValor(Estadistico, TipoContraste)
```

```
## [1] "El p-Valor es 0.000785741251043506"
```

```
     Estadistico
```

```
## [1] -3.7188
```

```
     RegionRechazo(alfa, TipoContraste)
```

```
## [1] "La region de rechazo la forman los Valores del Estadistico menores que  -1.7340636066175
```

respaldando las conclusiones que hemos obtenido en este ejemplo.

Ejemplo 9.3.1

Este ejemplo aparece en la pág. 9.3.2 del libro. Como en el anterior, empezamos con el código necesario para el contraste de varianzas. El comienzo de la plantilla sería así:

```
 # PRIMERA MUESTRA
 # Numero de elementos
 (n1 = 12)
```

```
## [1] 12
```

```
 # Cuasidesviacion tipica muestral
 (s1 = 0.4216)
```

```
## [1] 0.4216
```

```
 # SEGUNDA MUESTRA
 # Numero de elementos
 (n2 = 12)
```

```
## [1] 12
```

```
 # Cuasidesviacion tipica  muestral
 (s2 = 0.1740)
```

```
## [1] 0.174
```

```
 # TIPO DE CONTRASTE:
 # Escribe 1 si la HIP. ALTERNATIVA es sigma > sigma2,
 #         2 si es sigma1 < sigma2,
 #         3 si es bilateral.
 TipoContraste = 3

 #NIVEL DE SIGNIFICACION:
 (nSig = 0.95)
```

```
## [1] 0.95
```

Y los resultados que obtenemos:

```
 # Y ahora se aplican ambas funciones para mostrar los resultados
     pValor(Estadistico,TipoContraste)
```

```
## [1] "El p-Valor es 0.00666781125885452"
```

```
     Estadistico
```

```
## [1] 5.8709
```

En este caso, como el punto de partida son los propios valores que se han usado en el libro, no hay errores de redondeo apreciables. La conclusión, como se explica en el libro, es que rechazamos la hipótesis nula de igualdad de varianzas.

Por tanto, de vuelta al contraste de medias, vamos a usar la plantilla de la Tabla 1 titulada

`Tut09-Contraste-2Pob-DifMedias-UsandoT-VarIguales.R`

Ten en cuenta además la notación $H_a = \{\mu_2 - \mu_3\}$ que se ha usado en este ejemplo, a la hora de seleccionar el tipo de contraste. Con los datos del ejemplo, la primera parte de esa plantilla quedaría así:

```
# PRIMERA MUESTRA # Numero de elementos
(n1 = 12)
```

```
## [1] 12
```

```
# Media muestral
(xbar1 = 1.914)
```

```
## [1] 1.914
```

```
# Cuasidesviacion tipica muestral
(s1 = 0.4216)
```

```
## [1] 0.4216
```

```
# SEGUNDA MUESTRA
# Numero de elementos
(n2 = 12)
```

```
## [1] 12
```

```
# Media muestral
(xbar2 = 2.344)
```

```
## [1] 2.344
```

```
# Cuasidesviacion tipica  muestral
(s2 = 0.1740)
```

```
## [1] 0.174
```

```
# ¿Que tipo de contraste estamos haciendo?
# Escribe:    1 si la HIP. ALTERNATIVA es mu1 > mu2,
#             2 si es mu1 < mu2,
#             3 si es mu1 distinto de mu2
TipoContraste = 2

#Nivel de significacion
(nSig = 0.95)
```

```
## [1] 0.95
```

En este caso vamos a mostrar el número de grados de libertad que se obtienen usando la aproximación de Welch:

```
# Grados de libertad, aproximacion de Welch
(k = (s1^2/n1 + s2^2/n2)^2 / ((s1^4/(n1^2 * (n1 - 1))) + (s2^4 / (n2^2 * (n2 - 1)))))
```

```
## [1] 14.642
```

Los resultados son:

```
    pValor(Estadistico, TipoContraste)
```

```
## [1] "El p-Valor es 0.002676528260678"
```

```
    Estadistico
```

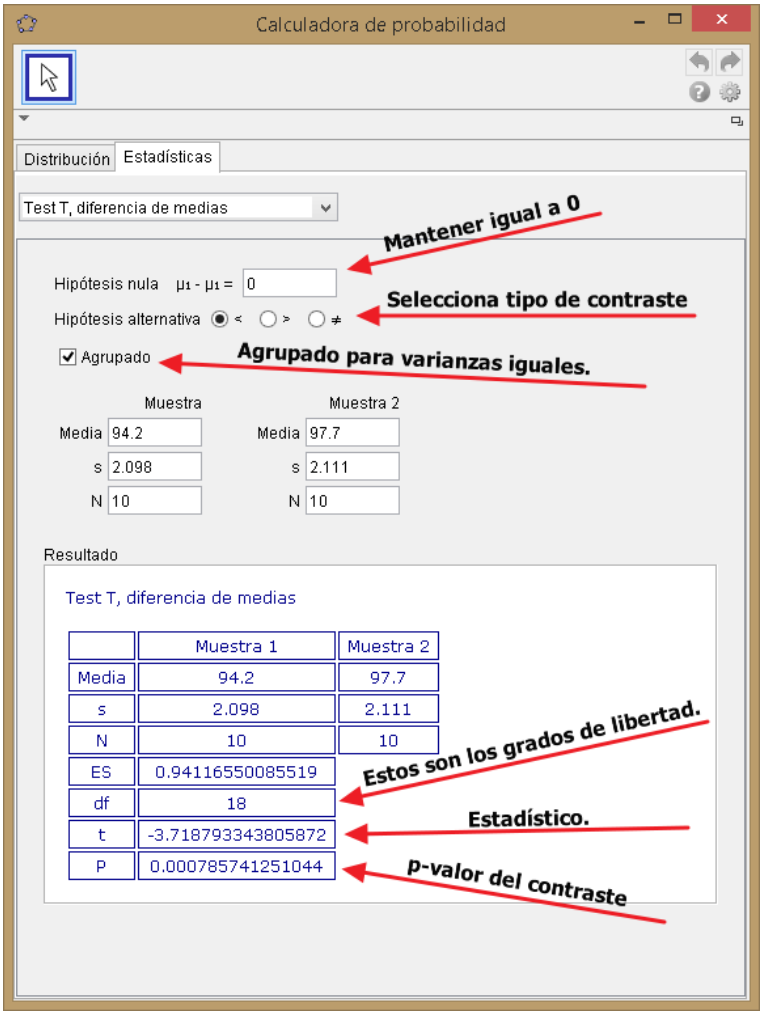
```
## [1] -3.2659
```

```
    RegionRechazo(alfa, TipoContraste)
```

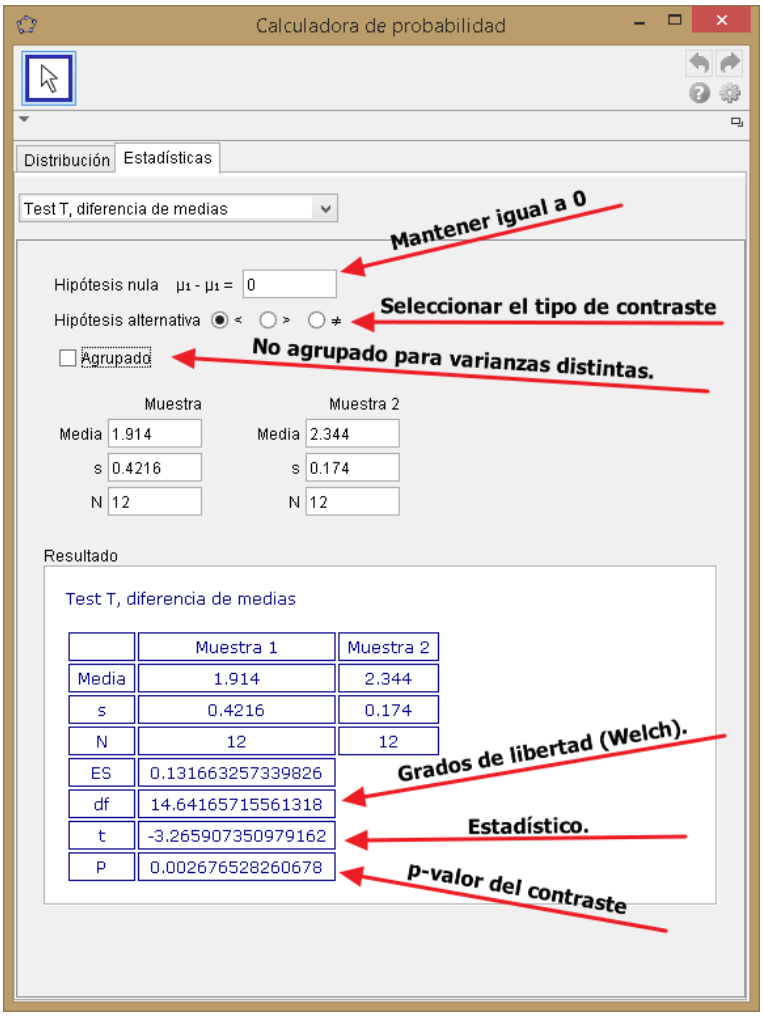
```
## [1] "La region de rechazo la forman los valores del Estadistico menores que  -1.75587212046059"
```

Contrastes de diferencia de medias con GeoGebra en el caso de muestras pequeñas.

La *Calculadora de Probabilidades* de GeoGebra incluye, en la pestaña titulada *Estadísticas*, la opción de calcular estos contrastes de diferencia de medias, introduciendo los valores muestrales en los campos del formulario que se muestra. Para revisar el funcionamiento de esta herramienta vamos a usar los datos de los dos ejemplos que hemos hecho antes con las plantillas de R y luego comentaremos algunos aspectos particulares. En esta primera figura se ilustra la forma de obtener el contraste del Ejemplo 9.3.1 del libro.



Mientras que para el Ejemplo 9.3.2 del libro debemos proceder como se muestra en esta figura:



Vamos a comentar algunos aspectos reseñables de esta herramienta:

- Aunque GeoGebra es un programa que las más de las veces resulta intuitivo y fácil de usar, esta interfaz no es, tal vez, de las más conseguidas. En la versión actual se ha colado además una errata que hace que en la hipótesis nula aparezca la fórmula $\mu_1 - \mu_1$ donde debería decir $\mu_1 - \mu_2$. Esta diferencia aparece igualada inicialmente a 0, aunque ese valor puede modificarse, para dar cabida a posibles hipótesis nulas como, por ejemplo (también podría ser con $\geq$ o $=$):
$$H_0 = \{(\mu_1 - \mu_2) \leq \Delta\mu_0\}$$
donde $\Delta\mu_0$ es una cantidad dada, en el mismo sentido que hemos discutido, para el caso de proporciones, en la Sección 9.1.1 del libro (pág. 297). En particular eso significa que **en la mayoría de las ocasiones queremos mantener el valor** $\mu_1 - \mu_2 = 0$.

- Los programadores de GeoGebra usan descripciones de la hipótesis nula que podemos resumir en la forma:
$$H_a = \{\mu_1 - \mu_2 \bigstar 0\}$$
donde $\bigstar$ es un símbolo que puede ser $<$, $>$ 0 $\neq$. Pero hay que tener en cuenta que, por ejemplo
$$H_a = \{\mu_1 - \mu_2 < 0\} = \{\mu_1 < \mu_2\}.$$
Así que decir que $\mu_1 - \mu_2 \bigstar 0$ es lo mismo que decir $\mu_1 \bigstar \mu_2$ *sea cual sea la interpretación del símbolo* $\bigstar$ de entre las tres posibles.

- Para elegir entre el caso en que asumimos varianzas iguales y el caso de varianzas distintas, debemos usar la casilla titulada *Agrupado*. Como hemos indicado en las figuras, marcamos esa casilla para el caso de varianzas iguales y la dejamos sin marcar en el caso de varianzas distintas.

4.2. Intervalos de confianza para la diferencia de medias con R.

Vamos a calcular intervalos de confianza al 95 % para la diferencia $\mu_1 - \mu_2$ en los Ejemplos 9.3.1 y 9.3.2 del libro que estamos usando en estos últimos apartados. Para ello usaremos los dos ficheros plantilla de la Tabla 1.

Para el Ejemplo 9.3.1 usamos el fichero `Tut09-IntConf-2Pob-DifMedias-UsandoT-VarianzasIguales.R`. El código con los datos del ejemplo sería así:

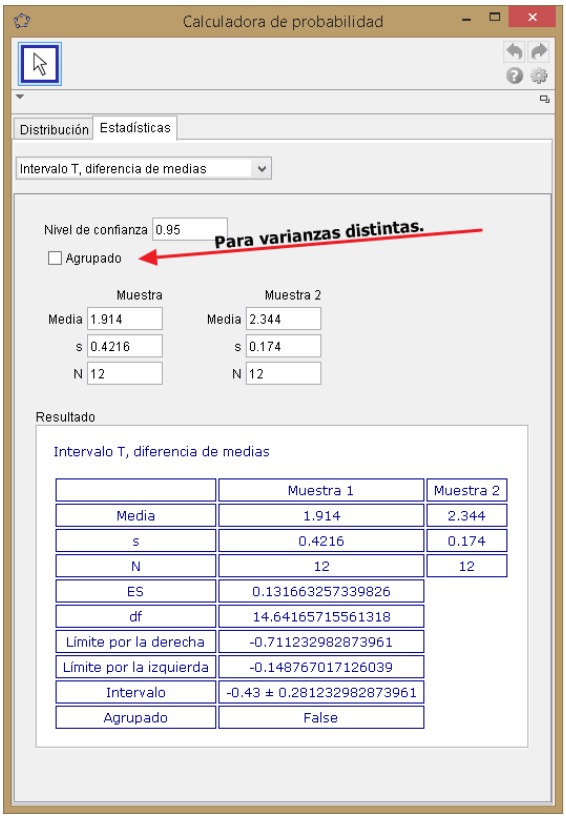
```
#########################################################
# www.postdata-statistics.com
# POSTDATA. Introducción a la Estadísitica
# Tutorial-09.
#
# Fichero de instrucciones R para calcular
# un intervalo de confianza para la
# DIFERENCIA DE MEDIAS DE 2 POBLACIONES NORMALES
# Es el caso de
# MUESTRAS PEQUEÑAS
# bajo la hipotesis de
# VARIANZAS IGUALES.
#################################################################

# Introducimos los tamaños de las muestras:
n1 = 10
n2 = 10
# Medias muestrales:
barX1 = 94.2
barX2 = 97.7
# Cuasidesviaciones tipicas muestrales:
s1 = 2.098
s2 = 2.111

# Nivel de confianza deseado:
nc = 0.95

#################################################
#NO CAMBIES NADA DE AQUI PARA ABAJO
#################################################
# Calculamos los grados de libertad:
(k = n1 + n2 - 2)

## [1] 18

# Calculamos el valor critico:
(alfa = 1 - nc)

## [1] 0.05

(t_alfa2 = qt(1 - alfa/2, df=k))

## [1] 2.1009

# La semianchura del intervalo es
(semianchura = t_alfa2 * sqrt((((n1 - 1) * s1^2 + (n2 - 1) * s2^2) /k) * (1/n1 + 1/n2)))

## [1] 1.9773

# Intervalo de confianza
(intervalo = (barX1 - barX2) + c(-1, 1) * semianchura)

## [1] -5.4773 -1.5227
```

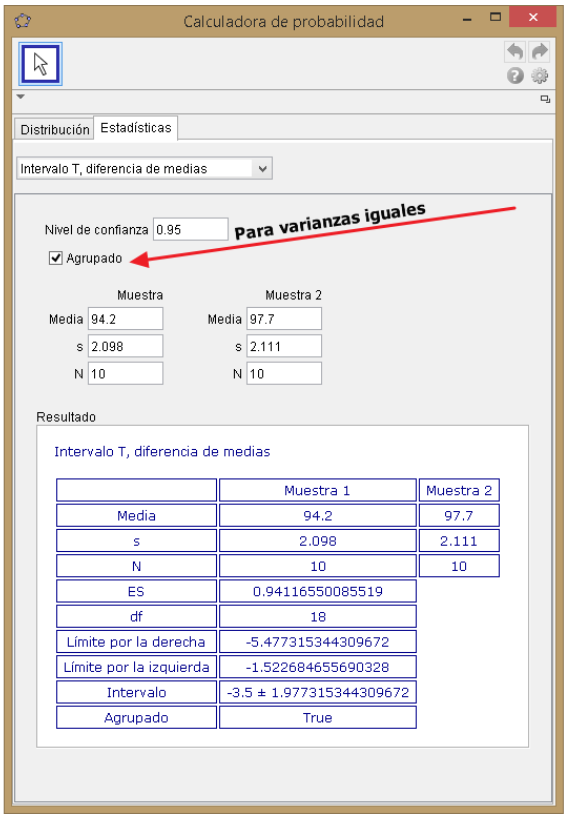
Para el Ejemplo 9.3.2 usaremos el fichero `Tut09-IntConf-2Pob-DifMedias-UsandoT-VarianzasDistintas.R`. Con los datos del Ejemplo el código quedaría así:

```
#####################################################
# www.postdata-statistics.com
# POSTDATA. Introducción a la Estadísitica
# Tutorial-09.
#
# Fichero de instrucciones R para calcular
# un intervalo de confianza para la
# DIFERENCIA DE MEDIAS DE 2 POBLACIONES NORMALES
# Es el caso de
# MUESTRAS PEQUEÑAS
# bajo la hipotesis de
# VARIANZAS IGUALES.
#####################################################################

# Introducimos los tamaños de las muestras:
n1 = 12
n2 = 12
# Medias muestrales:
barX1 = 1.914
barX2 = 2.344
# Cuasidesviaciones tipicas muestrales:
s1 = 0.4216
s2 = 0.1740

# Nivel de confianza deseado:
nc = 0.95

#################################################
#NO CAMBIES NADA DE AQUI PARA ABAJO
#################################################

# Calculamos los grados de libertad usando la aprox. de Welch
(k = (s1^2/n1 + s2^2/n2)^2 / ((s1^4/(n1^2 * (n1 - 1))) + (s2^4 / (n2^2 * (n2 - 1)))))

## [1] 14.642

# Calculamos el valor critico:
(alfa = 1 - nc)

## [1] 0.05

(t_alfa2 = qt(1-alfa/2, df=k))

## [1] 2.136

#La semianchura del intervalo es:
(semianchura = t_alfa2 * sqrt(s1^2/n1 + s2^2/n2))

## [1] 0.28123

# El intervalo de confianza es:
(intervalo = (barX1 - barX2) + c(-1, 1) * semianchura  )

## [1] -0.71123 -0.14877
```

Con GeoGebra.

En la *Calculadora de Probabilidades* de GeoGebra, podemos usar la opción *Intervalo T, diferencia de Medias*. Las siguientes figuras muestran el cálculo de los intervalos que hemos obtenido previamente con R.





5. Datos *en bruto* con R.

Opcional: esta sección puede omitirse en una primera lectura. De hecho, para leer esta sección, es necesario haber aprendido previamente a manejar los `data.frame` de R. Se recomienda en particular la lectura de la Sección 2 (pág. 9) del Tutorial04.

Vamos a dedicar esta sección a profundizar en el uso de varias funciones de R que son especialmente útiles para realizar contrastes entre parámetros de dos poblaciones. Las funciones son:

- `prop.test`
- `z.test`
- `t.test`
- `var.test`

Ya hemos discutido la función `prop.test` en la Sección 1 (pág. 3). Y la función `t.test` ha aparecido en Tutoriales previos. La función `var.test` está disponible por defecto en la instalación estándar de R, mientras que la función `z.test` se puede obtener instalando la librería `BSDA`. Esta librería, cuyo autor es Alan T. Arnholt, contiene numerosos conjuntos de datos relacionados con el libro *Basic Statistics and Data Analysis* de Larry J. Kitchens[1]. Puedes encontrar más información en este enlace:

cran.r-project.org/web/packages/BSDA/BSDA.pdf

[1] Kitchens, L. J. (2003) Basic Statistics and Data Analysis. Duxbury. ISBN: 978-0534384654

Hemos visto, en el Tutorial07, otra función llamada igualmente `z.test`, incluida en . Puede suceder que librerías distintas, a menudo escritas por diferentes autores, contengan funciones con el mismo nombre. En cualquier caso, si alguna vez necesitas las dos funciones, puedes referirte a ellas sin ambigüedad usando nombres como

```
BSDA::z.test
TeachingDemos::z.test
```

Como ves, la inclusión del nombre de la librería elimina las posibles confusiones.

Vamos a empezar instalando la librería `BSDA`. Puedes hacerlo desde RStudio, o también, simplemente, ejecutando este comando en R:

```
install.package(BSDA)
```

Una vez instalada la librería, la cargamos mediante

```
library(BSDA)
```

```
## Warning: package 'BSDA' was built under R version 3.2.2
```

```
## Loading required package: e1071
## Loading required package: lattice
##
## Attaching package: 'BSDA'
##
## The following object is masked from 'package:datasets':
##
##     Orange
```

Un contraste de igualdad de medias con muestras pequeñas: las funciones `t.test` y `var.test`

Como hemos dicho, esa librería incluye, además de la función `z.test`, numerosos conjuntos de datos, almacenados en `data.frames` de R. Vamos a usar uno de ellos para empezar nuestro trabajo. Concretamente, vamos a usar un conjunto de datos llamado `Statisti`. Para empezar a trabajar con ese conjunto de datos escribimos:

```
data(Statisti)
```

y para verlo, puedes usar este comando, que en RStudio abrirá un nuevo panel en el que puedes inspeccionar los datos:

```
View(Statisti)
```

Cuando se abra esa pestaña, verás que el `data.frame Statisti` contiene una tabla de datos con dos columnas, llamadas `Class1` y `Class2`. Cada columna representa las puntuaciones obtenidas por los alumnos de dos grupos de un curso de Estadística. Además, si te desplazas hacia la parte inferior de la tabla, verás que el número de alumnos de los dos grupos es distinto, y que la columna `Class2` contiene varias observaciones cuyo valor es `NA` (recuerda, *not available*, no disponible). Esta es la situación más común cuando trabajamos con muestras de tamaños distintos.

Recuerda también que para acceder a los datos de cada uno de los grupos por separado puedes usar una notación matricial, como en:

```
Statisti[ ,1]
```

```
##  [1] 81 73 86 90 75 80 75 81 85 87 83 75 70 65 80 76 64 74 86 80 83 67 82
## [24] 78 76 83 71 90 77 81 82
```

o también la notación `$` combinada con el nombre de la variable (columna), como en:

```
Statisti$Class1

##  [1] 81 73 86 90 75 80 75 81 85 87 83 75 70 65 80 76 64 74 86 80 83 67 82
## [24] 78 76 83 71 90 77 81 82
```

Vamos a suponer que las poblaciones muestreadas son normales y que las muestras son independientes. Llamamos μ_1 y μ_2 respectivamente a las puntuaciones medias de ambos grupos y usaremos esas dos muestras para contrastar la hipótesis nula:

$$H_0 = \{\mu_1 \neq \mu_2\}$$

Si tratas de usar `length` para hallar los tamaños de ambas muestras

```
length(Statisti$Class1)

## [1] 31

length(Statisti$Class2)

## [1] 31
```

comprobarás que R incluye los valores `NA` de `Class2` en ese recuento de la longitud. Y es razonable que así sea, porque es la opción menos problemática en la mayoría de los casos. Cuando trabajamos con `data.frames` y queremos saber si hay datos ausentes, una buena opción es usar la función `complete.cases`, que devuelve un vector de valores lógicos, iguales a `TRUE` cuando la fila correspondiente del `data.frame` no contiene valores ausentes e igual a `FALSE` en caso contrario. Para nuestro conjunto de datos:

```
(noAusentes = complete.cases(Statisti))

##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [12]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [23]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
```

Usando `complete.cases` junto con `which` y otros métodos que hemos visto en tutoriales previos (por ejemplo, la suma de valores lógicos) se puede gestionar de forma my eficaz la presencia de valores `NA` en un `data.frame` de R.

Pero para el trabajo que nos ocupa, no es necesario hacer nada complicado. Aunque hemos dicho varias veces a lo largo del curso que las muestras de más de 30 elementos pueden considerarse grandes, en este caso estamos al filo de ese tamaño y, de hecho, a causa de los datos ausentes, una de las muestras es de un tamaño menor que 30. Así que vamos a usar la distribución t para este contraste. Eso implica, com ya sabemos, que debemos empezar haciendo el contraste de la hipótesis nula de igualdad de varianzas:

$$H_0 = \{\sigma_1^2 = \sigma_2^2\}.$$

Para hacer este contraste vamos a recurrir a la función `var.test`. Simplemente escribimos:

```
var.test(Statisti$Class1, Statisti$Class2, alternative = "two.sided", conf.level = 0.95)

##
##  F test to compare two variances
##
## data:  Statisti$Class1 and Statisti$Class2
## F = 0.551, num df = 30, denom df = 26, p-value = 0.12
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
```

```
##  0.25541 1.16350
## sample estimates:
## ratio of variances
##             0.5508
```

Fíjate en que hemos usado `two.sided` para obtener el contraste bilateral que buscábamos. Como ves, el p-valor permite rechazar la hipótesis alternativa y, por tanto, seguir trabajando bajo la hipótesis de que las varianzas de ambos grupos son iguales. No queremos dejar pasar sin mencionarlo que además hemos obtenido un intervalo de confianza para el valor del cociente de varianzas.

Teniendo en cuenta este resultado, podemos volver al contraste de diferencia de medias, usando ahora la función `t.test`. Es tan simple como hacer:

```
t.test(Statisti$Class1, Statisti$Class2,
    alternative = "two.sided", conf.level = 0.95, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Statisti$Class1 and Statisti$Class2
## t = -1.07, df = 56, p-value = 0.29
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.3993  1.9310
## sample estimates:
## mean of x mean of y
##    78.581    80.815
```

Fíjate en que la opción `var.equal` nos permite ajustar el método que usa `t.test` al resultado del contraste de igualdad de varianzas que hemos hecho antes. Y, como ves, el p-valor permite rechazar H_a, para concluir que no hay base empírica para creer que las medias de los dos grupos son distintas.

Como ves, el uso combinado de `var.test` y `t.test` hace que los contrastes de igualdad de medias sean muy fáciles de llevar a cabo.

Sobre el formato del `data.frame` de este ejemplo. Datos con `read.table`.

A pesar de la facilidad con la que hemos trabajado en el apartado anterior, no podemos tampoco dejar pasar el hecho de que el formato del conjunto de datos que hemos usado en este ejemplo no es el recomendable. En el Tutorial11 volveremos sobre esto, pero queremos avanzar la idea básica para que el lector se vaya acostumbrando a oírla. Una tabla de datos en el formato correcto debe tener **una variable por columna y una observación por fila**. Hemos creado una nueva versión del `data.frame Statisti` en este formato correcto y la hemos almacenado en el fichero:

Tut09-Statisti2.csv

Descarga este fichero y guárdalo en tu carpeta `datos`. Antes de continuar, inspecciónalo con un editor de textos, como el *Bloc de Notas*. Vamos a aprovechar esta oportunidad para refrescar lo que sabemos del uso de la función `read.table`. Para leer el fichero y almacenarlo en un `data.frame` llamado `Statisti2` hacemos:

```
Statisti2 = read.table("./datos/Tut09-Statisti2.csv", header = TRUE, sep = "")
```

Y para ver que todo ha ido bien usamos `head` y `tail` así:

```
head(Statisti2)
```

```
##   scores group
## 1     81     1
```

```
## 2     73     1
## 3     86     1
## 4     90     1
## 5     75     1
## 6     80     1
```

```
tail(Statisti2)
```

```
##    scores group
## 53     74     2
## 54     77     2
## 55     87     2
## 56     69     2
## 57     96     2
## 58     65     2
```

Como ves, `Statisti2` contiene también dos columnas, pero ahora la primera, llamada `scores` (puntuaciones, en inglés) contiene las puntuaciones de ambos grupos, mientras que la segunda, llamada `group` es un factor que identifica el grupo al que pertenece esa puntuación. Como sucede muchas veces, los factores sirven para clasificar en grupos. Y de esta forma, el } respeta el principio de *una variable por columna, una observación por fila.*

¿Qué ocurre ahora con los contrastes de hipótesis? Pues que son igual de fáciles, pero debemos cambiar ligeramente la forma en que usamos la función, para explicarle a R que `group` es un factor que agrupa las observaciones de `scores` en grupos o niveles. Primero hacemos el contraste de igualdad de varianzas con `var.test`

```
var.test(scores ~ group, data = Statisti2, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  scores by group
## F = 0.551, num df = 30, denom df = 26, p-value = 0.12
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.25541 1.16350
## sample estimates:
## ratio of variances
##             0.5508
```

El resultado es, desde luego, exactamente el mismo que cuando usábamos el otro formato. Y prácticamente con la misma forma, hacemos el contraste para las medias:

```
t.test(scores ~ group, data = Statisti2,
    alternative = "two.sided", conf.level = 0.95, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  scores by group
## t = -1.07, df = 56, p-value = 0.29
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.3993  1.9310
## sample estimates:
## mean in group 1 mean in group 2
##          78.581          80.815
```

que, de nuevo, es idéntico al que hicimos con anterioridad.

Vamos a proponerte un ejercicio para que practiques estas ideas.

Ejercicio 4. *El fichero adjunto*

Tut09-Ejemplos-ContrasteMedias-01.csv

contiene muestras de una variable X en dos poblaciones normales, que llamamos población A y población B. Usa esos datos para contrastar la hipótesis nula:

$$H_0 = \{\mu_A = \mu_B\}.$$

Asegúrate de explorar primero los datos del fichero. Solución en la página 36. □

La función `z.test` de la librería BSDA.

En el caso de muestras grandes, en lugar de `t.test` podemos usar la función `z.test` de la librería `BSDA` para hacer los contrastes e intervalos de confianza correspondientes a ese tipo de problemas.

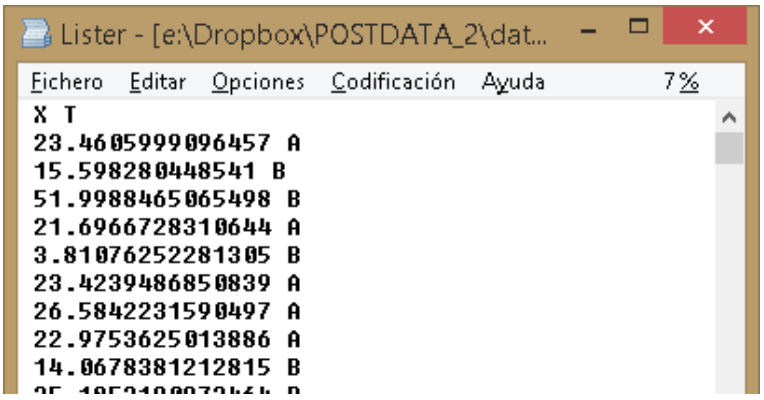
Para practicar esto vamos a usar los datos del fichero adjunto:

Tut09-Ejemplos-ContrasteMedias-02.csv

Este fichero contiene, de forma análoga a lo que sucedía en el Ejercicio 4, muestras de una variable X en dos poblaciones normales, que llamamos población A y población B. Y de nuevo, vamos a usar esos datos para contrastar la hipótesis nula:

$$H_0 = \{\mu_A = \mu_B\}.$$

La principal diferencia, como vamos a comprobar enseguida, es que ahora las muestras son de tamaño grande. Recuerda que la primera tarea consiste siempre en explorar el fichero de datos. Al abrirlo en un editor de texto verás algo como esto:



Para leer los datos del fichero usamos `read.table` y comprobamos que la lectura ha sido correcta con `head` así:

```
datos = read.table("./datos/Tut09-Ejemplos-ContrasteMedias-02.csv", header = TRUE, sep = "")
head(datos)
```

```
##          X T
## 1 23.4606 A
## 2 15.5983 B
## 3 51.9988 B
## 4 21.6967 A
## 5  3.8108 B
## 6 23.4239 A
```

La función `z-test` de la librería `BSDA` no es tan cómoda como las funciones `t.test` o `var.test`. En particular, con esta función no podemos usar una fórmula como `X ~ T` para describir lo que queremos hacer. Así que vamos a hacer algo mucho más *"manual"*. Definimos dos vectores que contienen los valores de X para cada uno de los grupos (niveles) definidos por el factor `T`:

```
XA = datos$X[datos$T=="A"]
XB = datos$X[datos$T=="B"]
```

Y ahora aplicamos así la función:

```
z.test(x = XA, y = XB, alternative = "two.sided", sigma.x = sd(XA), sigma.y = sd(XB))
```

```
##
##  Two-sample z-Test
##
## data:  XA and XB
## z = -3.22, p-value = 0.0013
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.8238 -1.1762
## sample estimates:
## mean of x mean of y
##        23        26
```

Fíjate que además debemos incluir las cuasidesviaciones típicas (calculadas con `sd`), porque de lo contrario se produce un error, ya que la función no las calcula por defecto.

Con esto, hemos obtenido el p-valor del contraste. Es posible que te pregunte qué sucedería si, en lugar de `z.test`, usáramos `t.test` en este caso de muestras grandes. Y si la usamos, ¿debemos usar la opción de varianzas iguales o distintas?

Ejercicio 5. *Usa la función* `t.test` *para realizar este contraste. Prueba las dos opciones posibles sobre las varianzas. ¿Cuál de ellas produce un resultado más parecido al que hemos obtenido con* `z.test`*? ¿Qué sucede si, al usar* `t.test`*, no indicas ninguna opción sobre la igualdad de las varianzas? Es decir, ¿cuál es el comportamiento* por defecto *de R? Solución en la página 37.* □

La función `t.test` para datos emparejados.

En la Sección 9.2.2 del libro (pág. 312) y también en este mismo tutorial, en la Sección 2.1 (pág. 6) hemos discutido el caso de los datos emparejados. Este tipo de contrastes, cuando disponemos de los datos *en bruto*, se llevan a cabo con mucha comodidad usando `t.test`, con la opción `paired=TRUE`.

Veamos un ejemplo. La librería `BSDA`, que hemos usado antes, contiene un conjunto de datos, llamado `Fitness`. Este conjunto de datos representa el número de un cierto tipo de flexiones que un grupo de sujetos podían hacer antes (en la columna *Before*) y después (columna *After*) de someterse a un programa de entrenamiento deportivo. Vamos a cargar ese conjunto de datos y a explorar su estructura:

```
library(BSDA)
data(Fitness)
head(Fitness)
```

```
##   Before After
## 1     28    32
## 2     31    33
## 3     17    19
## 4     22    26
## 5     12    17
## 6     32    30
```

```
str(Fitness)
```

```
## 'data.frame': 9 obs. of  2 variables:
##  $ Before: int  28 31 17 22 12 32 24 18 25
##  $ After : int  32 33 19 26 17 30 26 19 25
```

Además de `head` hemos usado la función `str`, que puede ser de mucha utilidad en este tipo de exploraciones preliminares. Como ves, el conjunto de datos contiene 5 observaciones, dos para cada individuo que se sometió al programa de entrenamiento. Por eso es un ejemplo típico de las situaciones que englobamos bajo esta etiqueta de *datos emparejados*. Llamando μ_a a la media antes del entrenamiento y μ_d a la media después del entrenamiento, queremos usar los datos para contrastar la hipótesis alternativa unilateral

$$H_a = \{\mu_a < \mu_d\}.$$

Y para hacer esto basta con usar `t.test` así:

```
t.test(Fitness$Before, Fitness$After,
    alternative = "less", paired = TRUE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  Fitness$Before and Fitness$After
## t = -2.75, df = 8, p-value = 0.012
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -0.64907
## sample estimates:
## mean of the differences
##                      -2
```

La clave, por supuesto, es la opción `paired=TRUE`. Fíjate, aparte de esto, en que el conjunto de datos no cumple el principio deseable de *una variable por columna, una observación por fila*. Por eso hemos usado la notación `$` para acceder a las columnas `Before` y `After`. La conclusión, es que al 95 % rechazamos H_0, pero no al 99 %. Con una muestra tan pequeña, eso significaría en la práctica, casi siempre, que los datos no son concluyentes. Se necesitan más datos: más potencia en el contraste, en el sentido que hemos discutido en el Capítulo 7.

6. Ejercicios adicionales y soluciones.

Ejercicios adicionales.

Hemos usado R en todos los casos para obtener las soluciones de los siguientes ejercicios. Pero es recomendable que pruebes alguna de las otras herramientas a tu disposición, al menos en alguno de estos ejercicios.

Ejercicio 6. *Para hacer un contraste de proporciones en dos poblaciones disponemos de estos datos muestrales, procedentes de dos muestras aleatorias independientes tomadas respectivamente de cada una de esas dos poblaciones:*

$$\begin{cases} n1 = 532, \textit{número de éxitos en la primera muestra} = 197. \\ n2 = 486, \textit{número de éxitos en la segunda muestra} = 151. \end{cases}$$

Usa estos datos para contrastar la hipótesis nula $H_0 = \{p_1 = p_2\}$. □

Ejercicio 7. *Para hacer un contraste de diferencia de medias de la variable X entre dos poblaciones normales disponemos de estos datos muestrales, procedentes de dos muestras aleatorias independientes tomadas respectivamente de cada una de esas dos poblaciones:*

$$\begin{cases} n_1 = 286, \\ \bar{X}_1 = 137.5 \\ s_1 = 2.2 \end{cases} \qquad \begin{cases} n_2 = 331, \\ \bar{X}_2 = 142.4 \\ s_2 = 15.6 \end{cases}$$

Usa estos datos para contrastar la hipótesis nula $H_0 = \{\mu_1 \geq \mu_2\}$. *Solución en la página 39.* □

Ejercicio 8. *De nuevo, para hacer un contraste de diferencia de medias de la variable X entre dos poblaciones normales disponemos de estos datos muestrales, procedentes de dos muestras aleatorias independientes tomadas respectivamente de cada una de esas dos poblaciones:*

$$\begin{cases} n_1 = 12, \\ \bar{X}_1 = 45.3 \\ s_1 = 3.7 \end{cases} \qquad \begin{cases} n_2 = 14, \\ \bar{X}_2 = 40.4 \\ s_2 = 3.9 \end{cases}$$

Usa estos datos para contrastar la hipótesis nula $H_0 = \{\mu_1 \leq \mu_2\}$. Solución en la página 40. □

Ejercicio 9. *Y por último, para hacer un contraste de diferencia de medias de la variable X entre dos poblaciones normales disponemos de estos datos muestrales, procedentes de dos muestras aleatorias independientes tomadas respectivamente de cada una de esas dos poblaciones:*

$$\begin{cases} n_1 = 7, \\ \bar{X}_1 = 0.9 \\ s_1 = 0.96 \end{cases} \qquad \begin{cases} n_2 = 7, \\ \bar{X}_2 = 1.2 \\ s_2 = 0.27 \end{cases}$$

Usa estos datos para contrastar la hipótesis nula $H_0 = \{\mu_1 \geq \mu_2\}$. Solución en la página 42. □

Soluciones de algunos ejercicios.

- **Ejercicio 2, pág. 5**

1. El código del fichero con los datos de este ejercicio aparece a continuación. Hemos descomentado las líneas donde aparecen los valores de s_1 y s_2 :

```
    ########################################################
    # www.postdata-statistics.com
    # POSTDATA. Introducción a la Estadísitica
    # Tutorial-09.
    #
    # Fichero de instrucciones R para calcular
    # un intervalo de confianza para la
    # DIFERENCIA DE MEDIAS DE 2 POBLACIONES NORMALES
    # usando la distribución Z
    # Es el caso de MUESTRAS GRANDES o (poco frecuente)
    # de varianzas poblacionales conocidas.
    ########################################################


    rm(list=ls())

    # PRIMERA MUESTRA
    # Numero de elementos
    (n1 = 245)

## [1] 245

    # Media muestral
    (xbar1 = 2.73)

## [1] 2.73

    # Cuasidesviacion tipica muestral o sigma (descomenta el que uses)
    (s1 = 0.4)

## [1] 0.4
```

```r
#(sigma1 = )


# SEGUNDA MUESTRA
# Numero de elementos
(n2 = 252)
```

```
## [1] 252
```

```r
# Media muestral
(xbar2 = 2.81)
```

```
## [1] 2.81
```

```r
# Cuasidesviacion tipica muestral o sigma (descomenta el que uses)
(s2 = 0.3)
```

```
## [1] 0.3
```

```r
#(sigma2 = )

# Nivel de confianza deseado:
nc = 0.95

################################################
#NO CAMBIES NADA DE AQUI PARA ABAJO
################################################

(alfa = 1 - nc)
```

```
## [1] 0.05
```

```r
# Calculamos el valor critico:
(z_alfa2 = qnorm( 1 - alfa / 2))
```

```
## [1] 1.96
```

```r
# La diferencia de las medias muestrales es:

(xbar1 - xbar2)
```

```
## [1] -0.08
```

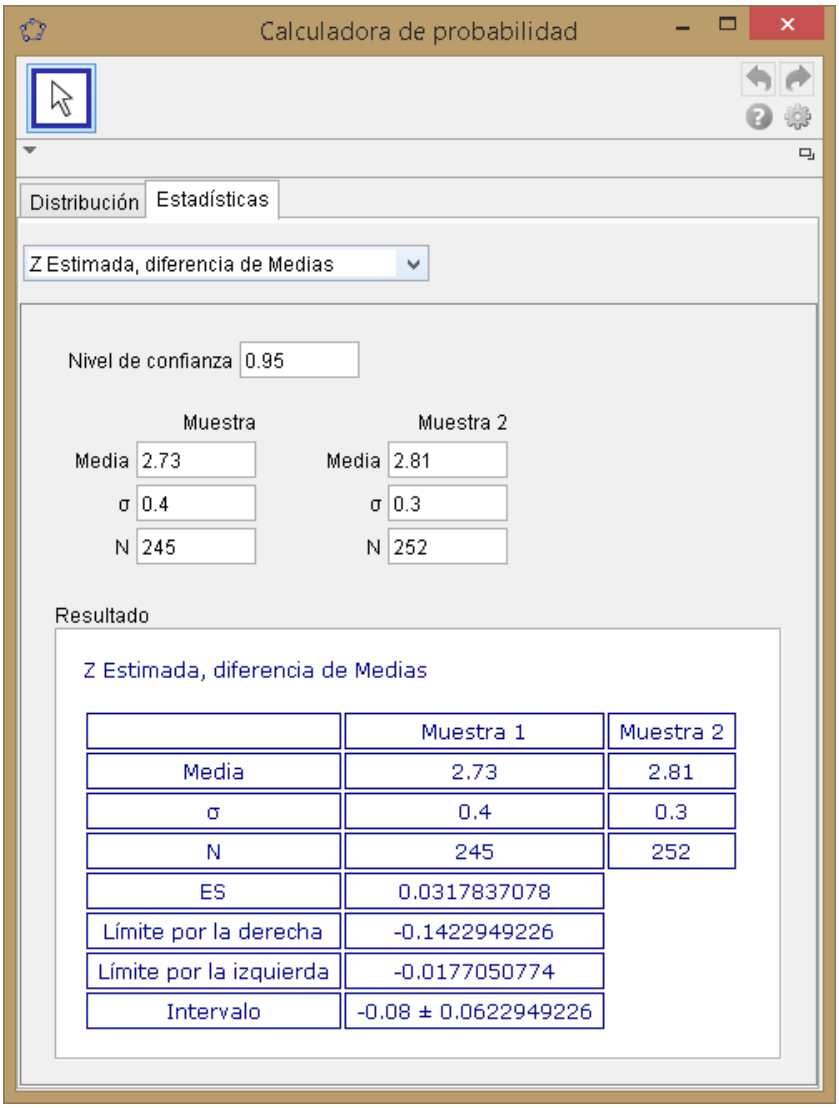
```r
# Comprobamos si se ha usado sigma como sustituto de s.

if(exists("sigma1")){s1 = sigma1}
if(exists("sigma2")){s2 = sigma2}

# La semianchura del intervalo es:
(semianchura = z_alfa2 * sqrt(s1^2/n1 + s2^2/n2))
```

```
## [1] 0.062295
```

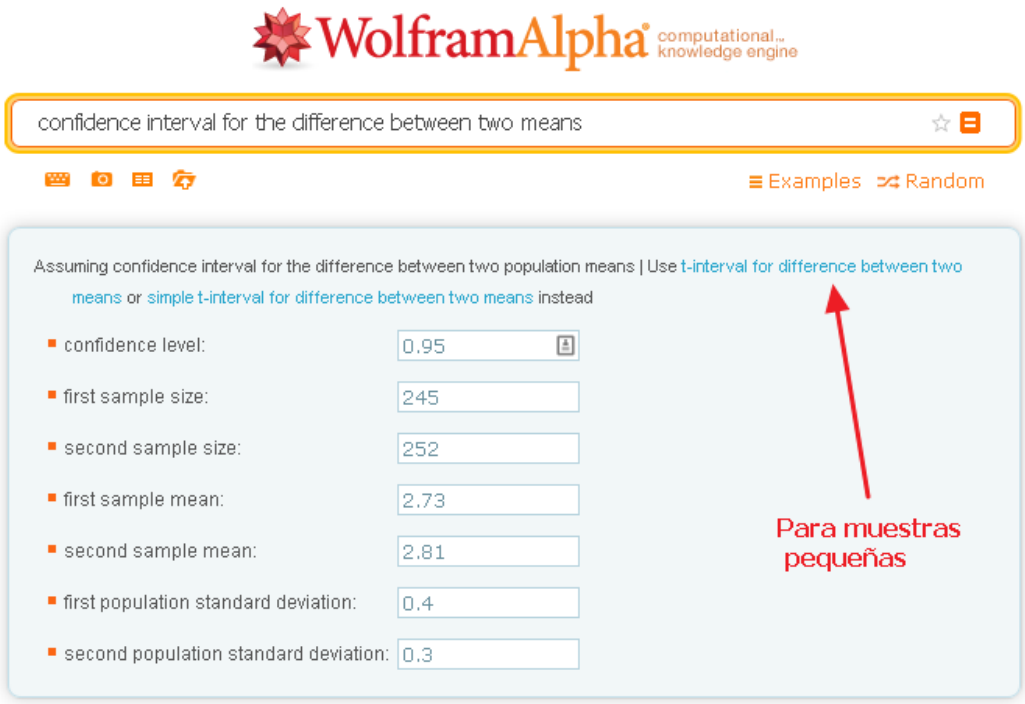
```r
# El intervalo de confianza es este:

(intervalo = xbar1 - xbar2 + c(-1, 1) * semianchura )
```

```
## [1] -0.142295 -0.017705
```

2. Esta es la forma de usar la *Calculadora de Probabilidades*:



3. En la siguiente figura se muestra como introducir ls datos para este ejercicio. Observa la forma de elegir entre muestras grandes y pequeñas, como indica la flecha roja.



Y en esta figura puedes ver la salida de Wolfram Alpha:

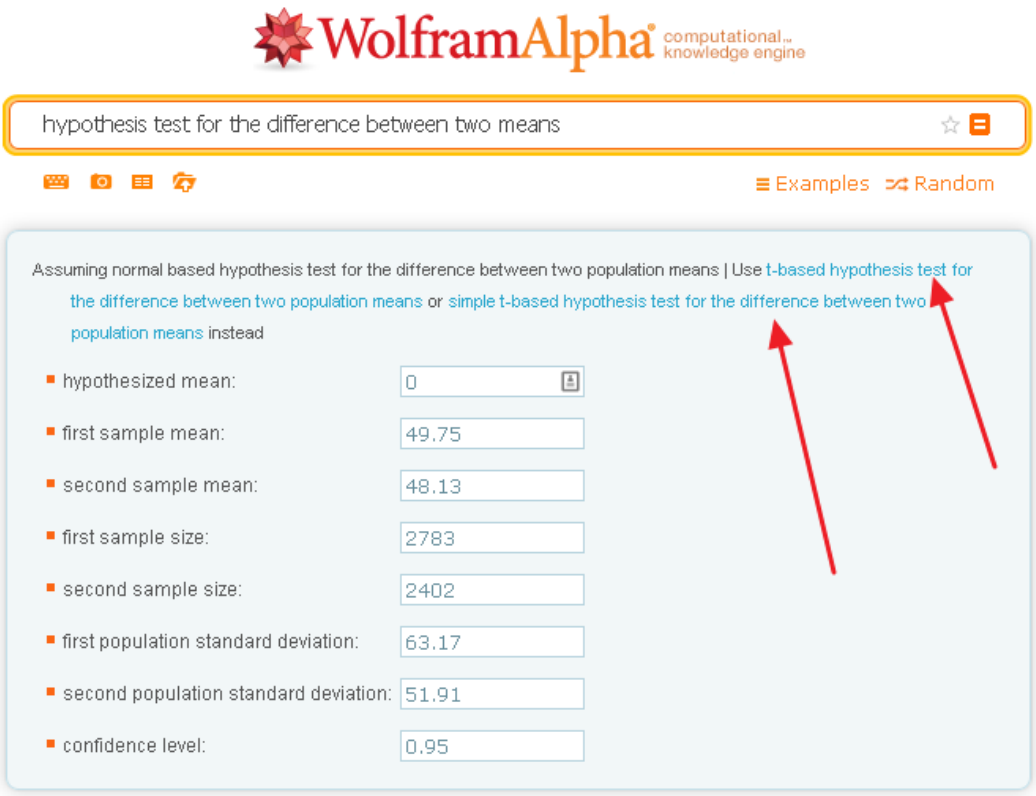
Input information:

confidence interval for the difference between two population means	
confidence level	0.95
first sample size	245
second sample size	252
first sample mean	2.73
second sample mean	2.81
first population standard deviation	0.4
second population standard deviation	0.3

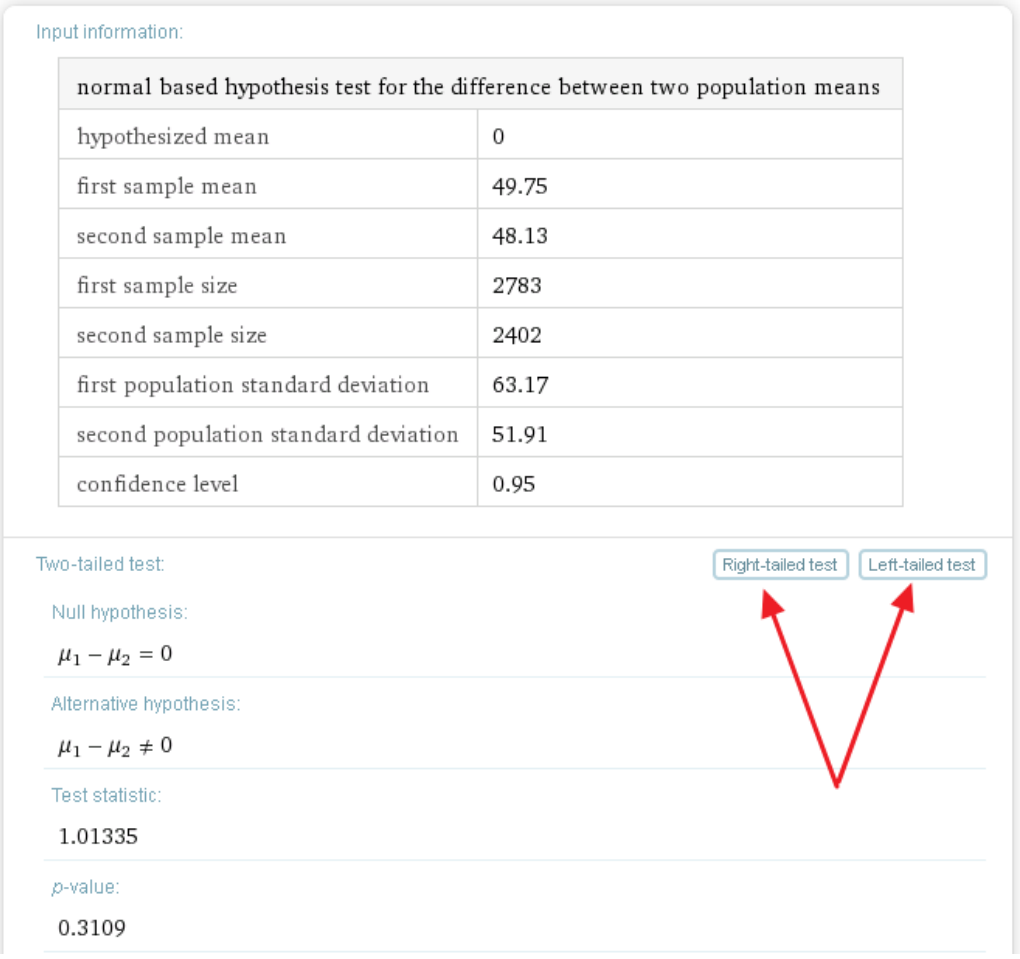
95% confidence interval:

-0.1423 to -0.01771

4. Introducimos los datos para el contraste en Wolfram Alpha como se muestra en la figura. Fíjate en las opciones que te permiten trabajar con muestras pequeñas, que hemos destacado con las flechas rojas:



La respuesta que se obtiene es esta. Fíjate de nuevo en las opciones disponibles para usar contrastes unilaterales o bilaterales.



Para hacer el mismo contraste usando la plantilla de R llamada

introducimos los datos del ejemplo al principio del código. Recuerda descomentar las líneas de s_1 y s_2:

```
  # PRIMERA MUESTRA
  # Numero de elementos
   (n1 = 2783)

## [1] 2783

  # Media muestral
   (xbar1 = 49.75)

## [1] 49.75

  # Cuasidesviacion tipica muestral o sigma (descomenta el que uses)
   (s1 = 63.17)

## [1] 63.17

   #(sigma1 = )


  # SEGUNDA MUESTRA
  # Numero de elementos
   (n2 = 2402)

## [1] 2402

  # Media muestral
   (xbar2 = 48.13)

## [1] 48.13

  # Cuasidesviacion tipica muestral o sigma (descomenta el que uses)
   (s2 = 51.91)

## [1] 51.91

   #(sigma2 = )

  # ¿Que tipo de contraste estamos haciendo?
  # Escribe 1 si la HIP. ALTERNATIVA es mu1 > mu2,
  #         2 si es mu1 < mu2,
  #         3 si es mu1 distinto de mu2
  TipoContraste = 3

  #Nivel de significacion
   (nSig = 0.95)

## [1] 0.95
```

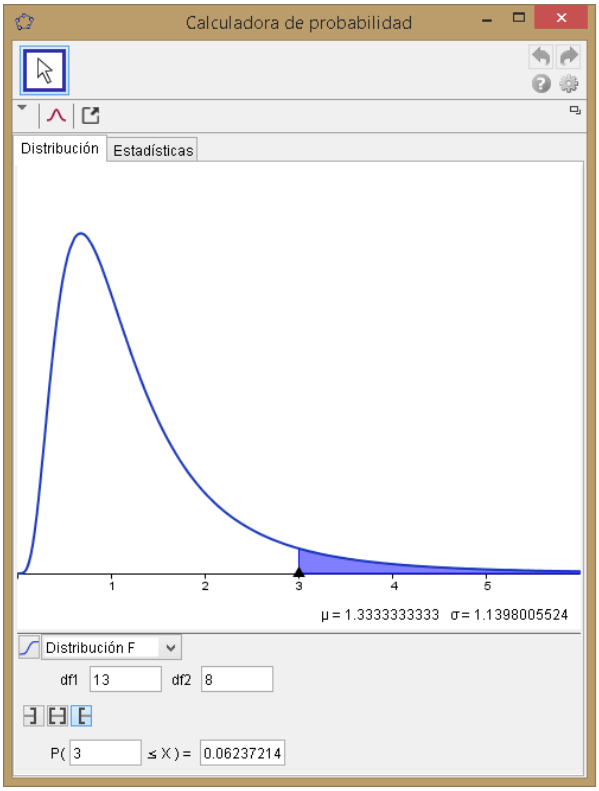
Y los resultados que se obtienen coinciden, como cabía esperar con los de Wolfram Alpha.

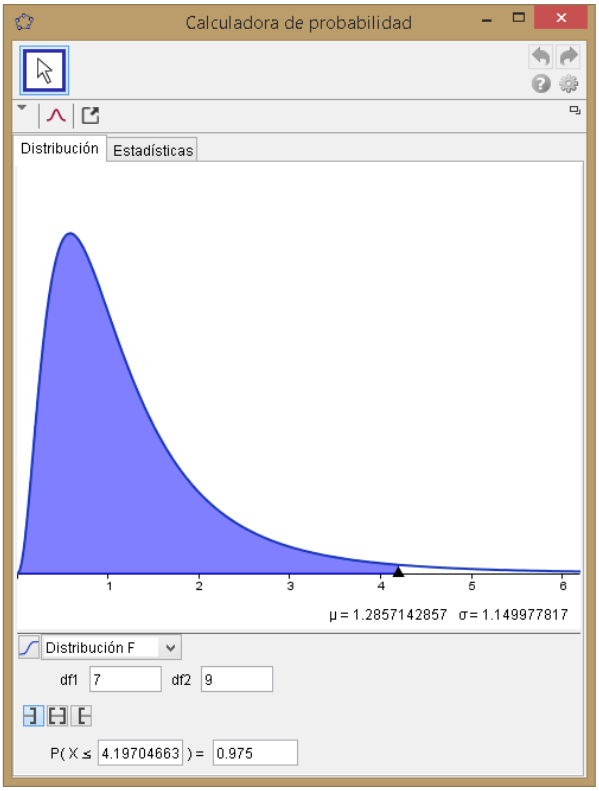
```
      pValor(Estadistico, TipoContraste)

## [1] "El p-Valor es 0.31089244301084"
```

```
      Estadistico

## [1] 1.0134

      RegionRechazo(alfa, TipoContraste)

## [1] "La region de rechazo la forman los Valores del Estadistico mas alejados del origen
```

- **Ejercicio 3, pág. 10**

Las siguientes figuras muestran la solución de ambos problemas de probabilidad:





- **Ejercicio 4, pág. 27**

El código R para leer el fichero es:

```
 datos = read.table("./datos/Tut09-Ejemplos-ContrasteMedias-01.csv", header = TRUE, sep = "")
 head(datos)

##         X T
## 1 4.3056 A
## 2 6.5297 A
## 3 6.0386 A
## 4 9.1185 A
## 5 2.4946 A
## 6 6.5334 A

 tail(datos)

##           X T
```

```
## 23 10.87338 B
## 24 -6.60762 B
## 25 -2.71845 B
## 26 21.50246 B
## 27 17.35569 B
## 28 -0.18161 B
```

Ahora podemos hacer el contraste de igualdad de varianzas en una sola línea de código:

```
var.test(X ~ T, data = datos, alternative = "two.sided", conf.level = 0.95)

##
##  F test to compare two variances
##
## data:  X by T
## F = 0.056, num df = 11, denom df = 15, p-value = 0.000027
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.018605 0.186344
## sample estimates:
## ratio of variances
##            0.05596
```

El p-valor obtenido nos lleva a rechazar la hipótesis nula de varianzas iguales. Así que podemos hacer el contraste de igualdad de medias, teniendo en cuenta este resultado para elegir el valor de la opción `var.equal` de `t.test`:

```
t.test(X ~ T, data = datos,
    alternative = "two.sided", conf.level = 0.95, var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data:  X by T
## t = 1.58, df = 17.2, p-value = 0.13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2807  8.8807
## sample estimates:
## mean in group A mean in group B
##             6.7             2.9
```

El p-valor que hemos obtenido indica que debemos rechazar la hipótesis alternativa,y concluir que no hay evidencia basada en los datos para creer que las medias de ambas poblaciones sean distintas.

- **Ejercicio 5, pág. 28**

Vamos a recordar primero el contraste con Z:

```
datos = read.table("./datos/Tut09-Ejemplos-ContrasteMedias-02.csv", header = TRUE, sep = "")
XA = datos$X[datos$T=="A"]
XB = datos$X[datos$T=="B"]
z.test(x = XA, y = XB, alternative = "two.sided", sigma.x = sd(XA), sigma.y = sd(XB))

##
##  Two-sample z-Test
##
## data:  XA and XB
```

```
## z = -3.22, p-value = 0.0013
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.8238 -1.1762
## sample estimates:
## mean of x mean of y
##        23        26
```

Y ahora veamos las tres posibilidades con t:

```
t.test(x = XA, y = XB, alternative = "two.sided", var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data:  XA and XB
## t = -3.22, df = 295, p-value = 0.0014
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.8313 -1.1687
## sample estimates:
## mean of x mean of y
##        23        26

t.test(x = XA, y = XB, alternative = "two.sided", var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  XA and XB
## t = -3.42, df = 607, p-value = 0.00067
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.7235 -1.2765
## sample estimates:
## mean of x mean of y
##        23        26

t.test(x = XA, y = XB, alternative = "two.sided")

##
##  Welch Two Sample t-test
##
## data:  XA and XB
## t = -3.22, df = 295, p-value = 0.0014
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.8313 -1.1687
## sample estimates:
## mean of x mean of y
##        23        26
```

Como ves, la más parecida es aquella en la primera, en la que suponemos que las varianzas son distintas y que es, además, la opción por defecto que usa R.

- **Ejercicio 6, pág. 29**

Podemos usar así la función `prop.test`:

```r
prop.test(c(197,151),n=c(532,486),alternative="two.sided",conf.level=0.95,correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(197, 151) out of c(532, 486)
## X-squared = 4.01, df = 1, p-value = 0.045
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0014931 0.1177092
## sample estimates:
## prop 1 prop 2
## 0.3703 0.3107
```

Como puedes ver, hemos usado la opción `correct=FALSE`, para evitar que R use una corrección de continuidad en la aproximación normal a la binomial. De esa forma, y aunque perdamos un poco de precisión, tratamos de obtener los resultados a los que conduce el estadístico que aparece en la Ecuación 9.2 (pág. 297) del Capítulo 9 del libro.

- **Ejercicio 7, pág. 29**

Este es el código de la plantilla de R con los datos del ejercicio:

```r
  # PRIMERA MUESTRA
  # Numero de elementos
      (n1 = 286)
```

```
## [1] 286
```

```r
  # Media muestral
      (xbar1 = 137.5)
```

```
## [1] 137.5
```

```r
  # Cuasidesviacion tipica muestral o sigma (descomenta el que uses)
      (s1 = 15.6)
```

```
## [1] 15.6
```

```r
      #(sigma1 = )


  # SEGUNDA MUESTRA
  # Numero de elementos
      (n2 = 331)
```

```
## [1] 331
```

```r
  # Media muestral
      (xbar2 = 142.4)
```

```
## [1] 142.4
```

```r
  # Cuasidesviacion tipica muestral o sigma (descomenta el que uses)
      (s2 = 19.4)
```

```
## [1] 19.4
```

```
      #(sigma2 = )

# ¿Que tipo de contraste estamos haciendo?
# Escribe 1 si la HIP. ALTERNATIVA es mu1 > mu2,
#          2 si es mu1 < mu2,
#          3 si es mu1 distinto de mu2
      TipoContraste = 2
  #Nivel de significacion
      (nSig = 0.95)
```

```
## [1] 0.95
```

Y los resultados que se obtienen son:

```
      pValor(Estadistico, TipoContraste)
```

```
## [1] "El p-Valor es 0.000255131809259936"
```

```
      Estadistico
```

```
## [1] -3.4753
```

- **Ejercicio 8, pág. 30**

Al tratarse de un contraste de diferencia de medias con muestras pequeñas debemos usar la t de Student y, previamente, para ello debemos hacer un contraste de la hipótesis nula de igualdad de varianzas:

$$H_0 = \{\sigma_1^2 = \sigma_2^2\}$$

El estadístico de este contraste es

```
(EstadisticoVar = s1^2/s2^2)
```

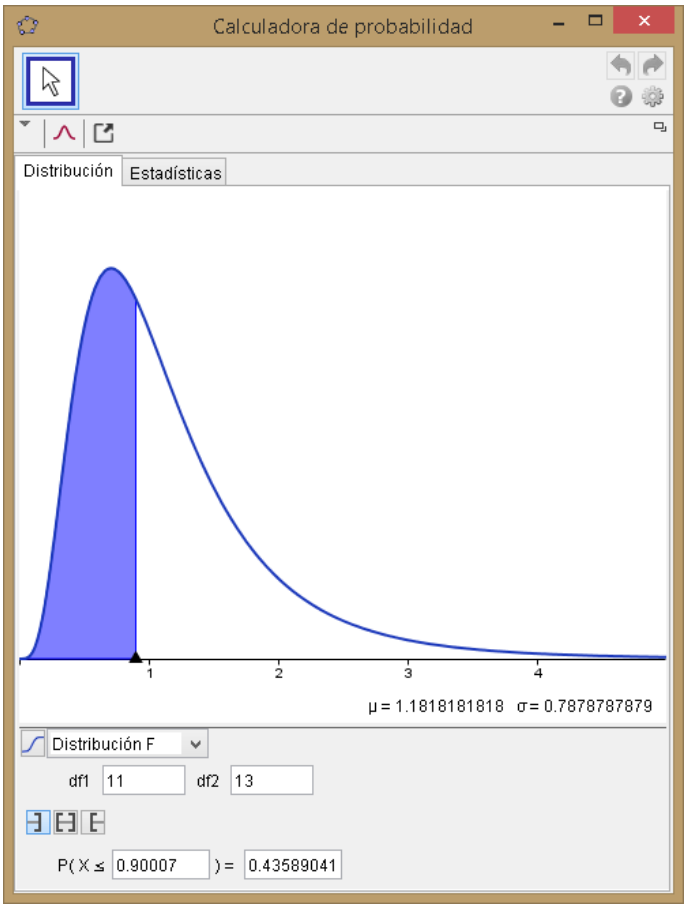
```
## [1] 0.90007
```

Y, puesto que este estadístico es menor que 1, usamos la cola izquierda de la distribución de Fisher para calcular el p-valor:

```
(pValorVar = pf(EstadisticoVar, df1 = n1 - 1, df2 = n2 - 1))
```

```
## [1] 0.43589
```

Puedes calcularlo igualmente con la *Calculadora de Probabilidades* de GeoGebra, como en la figura:



Con este p-valor rechazamos la hipótesis alternativa de que las varianzas sean distintas. Teniendo esto en cuenta, volvamos al contraste sobre la diferencia de medias. Esta es la parte inicial del código de la plantilla de R

`Tut09-Contraste-2Pob-DifMedias-UsandoT-VarIguales.R`

con los datos del ejercicio:

```
# PRIMERA MUESTRA # Numero de elementos
(n1 = 12)

## [1] 12

# Media muestral
(xbar1 = 45.3)

## [1] 45.3

# Cuasidesviacion tipica muestral
(s1 = 3.7)

## [1] 3.7

# SEGUNDA MUESTRA
# Numero de elementos
(n2 = 14)

## [1] 14

# Media muestral
(xbar2 = 40.4)
```

```
## [1] 40.4
```

```
 # Cuasidesviacion tipica  muestral
 (s2 = 3.9)
```

```
## [1] 3.9
```

```
 # ¿Que tipo de contraste estamos haciendo?
 # Escribe 1 si la HIP. ALTERNATIVA es mu1 > mu2, 2 si es mu1 < mu2, 3 si es mu1 distinto de mu
   TipoContraste = 1
 #Nivel de significacion
   (nSig = 0.95)
```

```
## [1] 0.95
```

Y los resultados que se obtienen son:

```
     pValor(Estadistico, TipoContraste)
```

```
## [1] "El p-Valor es 0.0015847637376516"
```

```
     Estadistico
```

```
## [1] 3.2833
```

La conclusión es que rechazamos la hipótesis nula: los datos no permiten afirmar que sea $\mu_1 \geq \mu_2$

- **Ejercicio 9, pág. 30**

De nuevo, puesto que las muestras son pequeñas debemos usar la t de Student y eso nos lleva a empezar con un contraste de la hipótesis nula de igualdad de varianzas:

$$H_0 = \{\sigma_1^2 = \sigma_2^2\}$$

El estadístico de este contraste vale, en este caso

```
(EstadisticoVar = s1^2/s2^2)
```

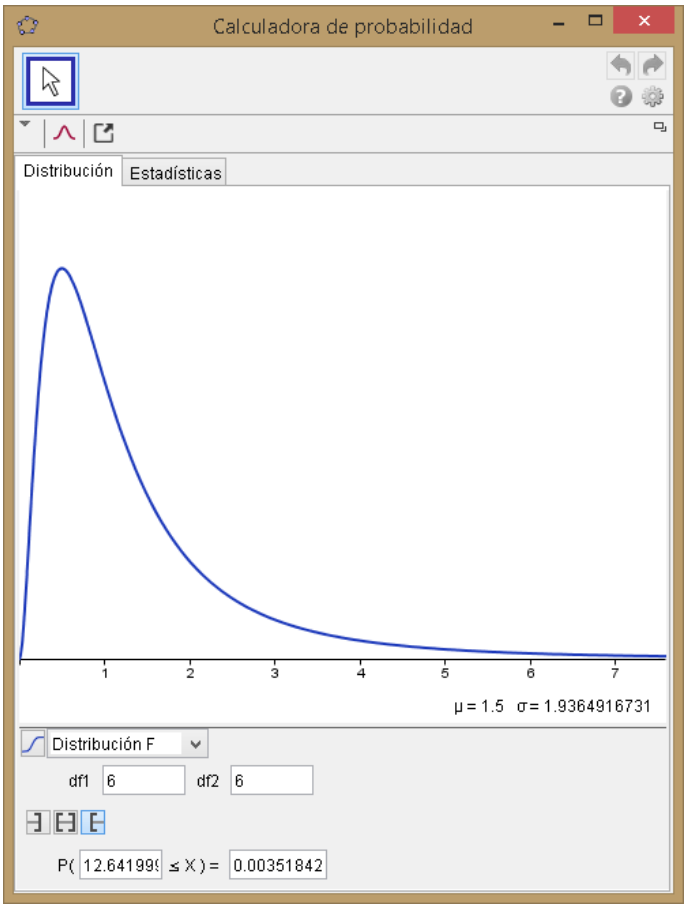
```
## [1] 12.642
```

Y, puesto que este estadístico es mayor que 1, usamos la cola derecha de la distribución de Fisher para calcular el p-valor:

```
(pValorVar = 1 - pf(EstadisticoVar, df1 = n1 - 1, df2 = n2 - 1))
```

```
## [1] 0.0035184
```

También puedes calcularlo con GeoGebra, desde luego:



Con este p-valor rechazamos la hipótesis nula de que las varianzas sean iguales. Usamos esto para decidir lo que hay que hacer en el contraste sobre la diferencia de medias. Este es el código de la plantilla de R

`Tut09-Contraste-2Pob-DifMedias-UsandoT-VarDistintas.R`

con los datos del ejercicio:

Y los resultados que se obtienen son:

```
     pValor(Estadistico, TipoContraste)

## [1] "El p-Valor es 0.22621403141095"

     Estadistico

## [1] -0.79592
```

La conclusión es que rechazamos la hipótesis alternativa: los datos no permiten afirmar que sea $\mu_1 < \mu_2$

Plantillas de R para contrastes e intervalos de confianza.

- **Diferencia medias.**
  - Usando Z.
    Tut09-Contraste-2Pob-DifMedias-UsandoZ.R
    Tut09-IntConf-2Pob-DifMedias-UsandoZ.R
  - Usando la t de Student.
    - Varianzas desconocidas pero iguales.
      Tut09-Contraste-2Pob-DifMedias-UsandoT-VarIguales.R
      Tut09-IntConf-2Pob-DifMedias-UsandoT-VarianzasIguales.R
    - Varianzas desconocidas pero distintas.
      Tut09-Contraste-2Pob-DifMedias-UsandoT-VarDistintas.R
      Tut09-IntConf-2Pob-DifMedias-UsandoT-VarianzasDistintas.R
- **Cociente varianzas.**
  Tut09-Contraste-2Pob-CocienteVarianzas.R
  Tut09-IntConf-2Pob-CocienteVarianzas.R
- **Diferencia proporciones.**
  Tut09-Contraste-2Pob-DifProporciones-UsandoZ.R
  Tut09-IntConf-2Pob-DifProporciones-UsandoZ.R

Tabla 1: Ficheros para los contrastes de hipótesis e intervalos de confianza en dos poblaciones independientes.

Fin del Tutorial09. ¡Gracias por la atención!