

Distribuciones de probabilidad e inferencia estadística con R-Commander

Daniel Liviano Solís

Maria Pujol Jover

PID_00208274

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, ya sea este eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.

Índice

Introducción	5
Objetivos	6
1. Distribuciones de probabilidad	7
1.1. Introducción	7
1.2. Distribuciones de probabilidad discretas	8
1.2.1. Distribución binomial	8
1.2.2. Distribución geométrica	12
1.2.3. Distribución hipergeométrica	14
1.2.4. Distribución de Poisson	15
1.3. Distribuciones de probabilidad continuas	16
1.3.1. Distribución uniforme	16
1.3.2. Distribución exponencial	18
1.3.3. Distribución normal	20
1.3.4. Distribución t de Student	25
1.3.5. Teorema del límite central	26
2. Inferencia estadística	29
2.1. Introducción	29
2.2. Inferencia sobre la media con varianza poblacional desconocida	29
2.3. Inferencia sobre la media con varianza poblacional conocida	32
2.4. Inferencia sobre la proporción	34
2.5. Inferencia sobre la varianza	35
2.6. Inferencia sobre la diferencia de medias con muestras independientes	37
2.6.1. Varianzas poblacionales desconocidas pero iguales	37
2.6.2. Varianzas poblacionales conocidas	38
2.7. Inferencia sobre la diferencia de medias con muestras apareadas	39
2.8. Inferencia sobre la diferencia de proporciones	41
2.9. Inferencia sobre el cociente de varianzas	43
Bibliografía	47

Introducción

En este módulo, se aborda toda la parte de la inferencia estadística mediante el uso de R-Commander. En un primer apartado se aprenderá a trabajar con variables discretas, sus distribuciones asociadas y el cálculo de sus probabilidades. Posteriormente, se hará lo mismo con variables continuas. De esta manera, se estudiarán las principales distribuciones de probabilidad tanto discretas como continuas que se utilizan en la mayoría de las asignaturas de estadística de la UOC.

En concreto, se profundizará en las siguientes distribuciones de probabilidad discretas:

- La distribución binomial.
- La distribución geométrica.
- La distribución hipergeométrica.
- La distribución de Poisson.

Y en las distribuciones de probabilidad continuas que se citan a continuación:

- La distribución uniforme.
- La distribución exponencial.
- La distribución normal.
- La distribución t de Student.
- La distribución chi-cuadrado.
- La distribución F de Snedecor.

Como será posible denotar al final del primer apartado, casi todas las distribuciones podrán aproximarse a una normal gracias al teorema del límite central. Este aspecto será muy importante para obtener las distribuciones de la media muestral y de la proporción a partir de una población, aunque esta no necesariamente se distribuya según una normal.

El segundo apartado se dedicará de manera íntegra a la inferencia estadística sobre los principales parámetros de las distribuciones: la media poblacional (μ), la varianza (σ^2) y la proporción (π).

Objetivos

El estudiante ha de ser capaz de utilizar R-Commander para:

1. Calcular los valores críticos y las probabilidades asociadas a una distribución de probabilidad discreta.
2. Calcular los valores críticos y las probabilidades asociadas a una distribución de probabilidad continua.
3. Calcular intervalos de confianza y llevar a cabo contrastes de hipótesis (unilaterales y bilaterales) para la media poblacional (μ) con varianza poblacional (σ^2) conocida.
4. Calcular intervalos de confianza y efectuar contrastes de hipótesis (unilaterales y bilaterales) para la media poblacional (μ) con varianza poblacional (σ^2) desconocida.
5. Calcular intervalos de confianza y efectuar contrastes de hipótesis (unilaterales y bilaterales) para la varianza poblacional (σ^2).
6. Calcular intervalos de confianza y llevar a cabo contrastes de hipótesis (unilaterales y bilaterales) para la proporción poblacional (π).
7. Calcular intervalos de confianza y efectuar contrastes de hipótesis (unilaterales y bilaterales) para la diferencia de medias poblacionales (μ_1 y μ_2) con varianzas (σ_1^2 y σ_2^2) conocidas.
8. Calcular intervalos de confianza y llevar a cabo contrastes de hipótesis (unilaterales y bilaterales) para la diferencia de medias poblacionales (μ_1 y μ_2) con varianzas desconocidas pero iguales (σ^2).
9. Calcular intervalos de confianza y llevar a cabo contrastes de hipótesis (unilaterales y bilaterales) para el cociente de varianzas poblacionales (σ_1^2 y σ_2^2).
10. Calcular intervalos de confianza y efectuar contrastes de hipótesis (unilaterales y bilaterales) para la diferencia de proporciones poblacionales (π_1 y π_2).

1. Distribuciones de probabilidad

1.1. Introducción

Una parte fundamental de la estadística es el análisis de variables aleatorias, las cuales se pueden distribuir según diferentes funciones de probabilidad. R-Commander ofrece la posibilidad de analizar estas distribuciones de probabilidad de varios modos. De manera específica, para cada distribución de probabilidad encontramos cuatro opciones de cálculo:

1) Cuantiles. La función cuantil es la inversa de la función de distribución. Es decir, un cuantil será el valor x tal que $P(X \leq x) = p$, siendo p la probabilidad que como tal se encuentra en el intervalo $[0, 1]$.

Los cuantiles nos serán muy útiles para calcular los valores críticos de las distribuciones.

2) Probabilidades. La probabilidad será aquel valor p tal que $P(X \leq x) = p$, es decir, se trata de la función inversa a la función cuantil.

3) Gráficas. Cada distribución de probabilidad ofrece la posibilidad de calcular los gráficos de su función de densidad o de cuantía asociados, así como su función de distribución acumulada.

La función de cuantía solo existe para las variables aleatorias discretas y la de densidad, para las continuas. Además, el área bajo la curva de todas las funciones de distribución siempre será la unidad, ya que incluye toda la probabilidad acumulada de una variable aleatoria.

4) Muestras aleatorias. Para cada distribución se pueden generar muestras de n valores aleatorios, dados unos parámetros iniciales determinados.

En lo que respecta al cálculo de cuantiles y probabilidades, hay que especificar que R ofrece la posibilidad de hacer cálculos para las dos colas (izquierda y derecha), es decir, tanto $P(X \leq x) = p$ como $P(X > x) = p$. Es fundamental subrayar el símbolo \leq para la cola izquierda y el símbolo $>$ para la cola derecha, para evitar confusiones en el cálculo de probabilidades. Además, fuera de R-Commander, también es posible calcular mediante funciones en código R el valor de la **función de densidad** (para variables continuas) y de la **función de cuantía** (para variables discretas).

Es importante denotar que para variables continuas no tiene sentido calcular $P(X = x) = p$. Por lo tanto, para este tipo de variables es más correcto calcular $P(X < x) = p$ para obtener la probabilidad de la cola de la izquierda y $P(X > x) = p$ para la de la derecha.

A continuación, ofrecemos una relación de las funciones en código R de las distribuciones de probabilidad incorporadas en la distribución básica de R y que están disponibles en el menú desplegable de R-Commander. En la tabla 1, se incluyen las distribuciones de probabilidad discretas:

Tabla 1. Distribuciones de probabilidad discretas

Distribución	Función de cuantía	Probabilidades	Cuantiles	Muestra aleatoria
Binomial	dbinom	pbinom	qbinom	rbinom
Poisson	dpois	ppois	qpois	rpois
Geométrica	dgeom	pgeom	qgeom	rgeom
Hipergeométrica	dhyper	phyper	qhyper	rhyper
Binomial negativa	dnbinom	pnbinom	qnbinom	rnbinom

De igual manera, en la tabla 2 se ofrecen las distribuciones de probabilidad continuas y las funciones de R asociadas:

Tabla 2. Distribuciones de probabilidad continuas

Distribución	Función de densidad	Probabilidades	Cuantiles	Muestra aleatoria
Normal	dnorm	pnorm	qnorm	rnorm
t-Student	dt	pt	qt	rt
Chi-cuadrado	dchisq	pchisq	qchisq	rchisq
F-Snedecor	df	pf	qf	rf
Exponencial	dexp	pexp	qexp	rexp
Uniforme	dunif	punif	qunif	runif
Beta	dbeta	pbeta	qbeta	rbeta
Cauchy	dcauchy	pcauchy	qcauchy	rcauchy
Logística	dlogis	plogis	qlogis	rlogis
Log-normal	dlnorm	plnorm	qlnorm	rlnorm
Gamma	dgamma	pgamma	qgamma	rgamma
Weibull	dweibull	pweibull	qweibull	rweibull

A continuación, ofrecemos varios ejemplos prácticos de análisis de variables aleatorias con diferentes distribuciones.

1.2. Distribuciones de probabilidad discretas

1.2.1. Distribución binomial

La distribución binomial es una de las principales distribuciones de probabilidad discretas. De manera específica, esta mide el número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí, y con una probabilidad p de éxito entre los ensayos. Es decir, definimos formalmente una variable aleatoria X que se distribuye como una binomial con parámetros n y p :

$$X \sim B(n, p)$$

La función de probabilidad tiene la forma siguiente:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

La distribución de Bernoulli

Las variables aleatorias dicotómicas que toman únicamente el valor 1 cuando un ensayo o suceso se ha llevado a cabo con éxito y un valor 0 en caso contrario se distribuyen bajo una distribución de Bernoulli. Por tanto, la distribución binomial será una generalización de una distribución de Bernoulli.

Recordad que para variables binomiales, $E(X) = np$ y $V(X) = np(1 - p)$.

Siendo $\binom{n}{x}$ el cociente binomial, que define el número de subconjuntos de x elementos elegidos de un conjunto con n elementos. Este toma la expresión siguiente:

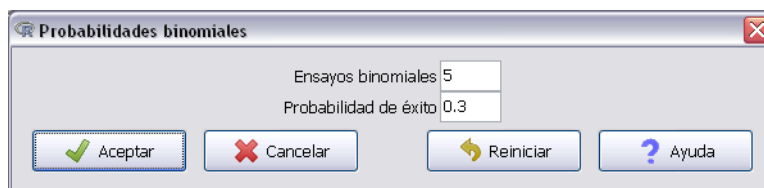
$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Un cociente binomial no es más que las combinaciones de n elementos tomados de x en x .

Para ver cómo trabajamos con variables aleatorias distribuidas como una binomial, lo haremos con un ejemplo ficticio. Sabemos que la probabilidad que tiene un jugador de baloncesto determinado de encestar un triple es del 30 %, es decir, $p = 0,3$, y queremos calcular las probabilidades de que enceste 1, 2, ... hasta n triples, siendo n el número de lanzamientos (ensayos). Si suponemos que este jugador lanza 5 veces, $n = 5$, haremos nuestro cálculo con R-Commander de la manera siguiente:

Distribuciones / Distribuciones discretas / Distribución binomial / Probabilidades binomiales

Veremos que aparecerá una ventana en la que introduciremos la probabilidad de éxito y el número de ensayos:



Al pulsar en *Aceptar*, obtendremos el resultado siguiente.

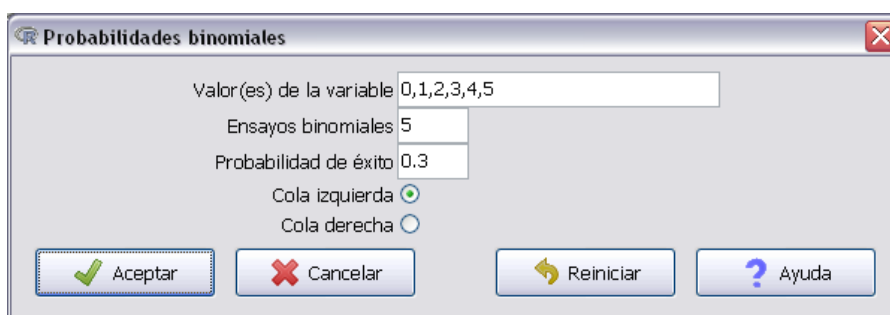
```
> .Table <- data.frame(Pr=dbinom(0:5, size=5, prob=0.3))
> rownames(.Table) <- 0:5
> .Table
      Pr
0 0.16807
1 0.36015
2 0.30870
3 0.13230
4 0.02835
5 0.00243
```

Sin embargo, ¿cómo se interpreta este resultado? Si definimos X como el número de éxitos, tenemos que $P(X = 0) = 0,168$. Esto no es más que la probabilidad de que en 5 ensayos nuestro jugador no acierte ningún triple. En el otro extremo, la probabilidad de que acierte 5 triples en 5 intentos es de $P(X = 5) = 0,002$. Resulta importante observar que la suma de estas cinco probabilidades es igual a uno.

Si quisiéramos calcular las probabilidades acumuladas (también denominadas de la cola), la ruta que será necesario seguir en el menú desplegable es esta:

Distribuciones / Distribuciones discretas / Distribución binomial / Probabilidades binomiales acumuladas

El cuadro de diálogo que nos aparecerá es el siguiente, donde en los valor(es) de la variable tenemos que introducir los valores de X que nos interesan (en este caso los queremos todos), y en las dos siguientes opciones introduciremos el número de ensayos y la probabilidad de éxito, de manera respectiva. Además, tenemos que especificar la cola de la distribución que nos interesa. Si activamos *Cola izquierda*, estamos calculando $P(X \leq x)$:



Por el contrario, si deseáramos calcular $P(X > x)$, tenemos que activar la opción *Cola derecha*. Si seleccionamos primero *Cola izquierda* y después *Cola derecha* obtenemos los resultados siguientes, de manera respectiva:

```
> pbinom(c(0,1,2,3,4,5), size=5, prob=0.3,
+ lower.tail=TRUE)
[1] 0.16807 0.52822 0.83692 0.96922 0.99757 1.00000

> pbinom(c(0,1,2,3,4,5), size=5, prob=0.3,
+ lower.tail=FALSE)
[1] 0.83193 0.47178 0.16308 0.03078 0.00243 0.00000
```

Observad que, si sumamos las probabilidades obtenidas por columnas, la suma de cada par de valores siempre es uno. Esto sucede porque, por definición, siempre se cumple la igualdad siguiente:

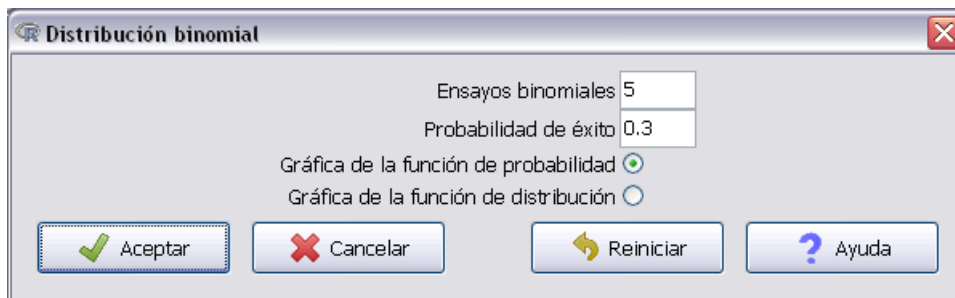
$$P(X \leq x) + P(X > x) = 1.$$

En el ejemplo anterior, ¿cuál sería la probabilidad de encestar más de tres triples? La respuesta es $P(X > 3) = 0,03078$. ¿Y la probabilidad de encestar menos de dos triples? La respuesta es $P(X \leq 1) = 0,52822$.

Un último aspecto que veremos de la distribución binomial es el resultado gráfico de las probabilidades calculadas. Tenemos que acceder a la ruta siguiente:

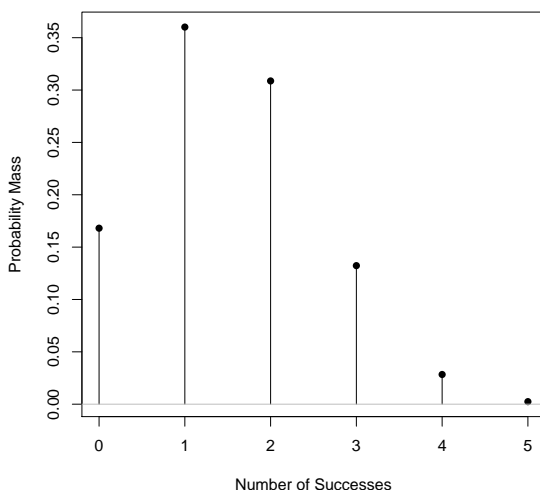
Distribuciones / Distribuciones discretas / Distribución binomial / Gráfica de la distribución binomial

En el cuadro de diálogo resultante, R-Commander nos da la opción de obtener dos tipos de gráficos: la *función de probabilidad* y la *función de distribución*:

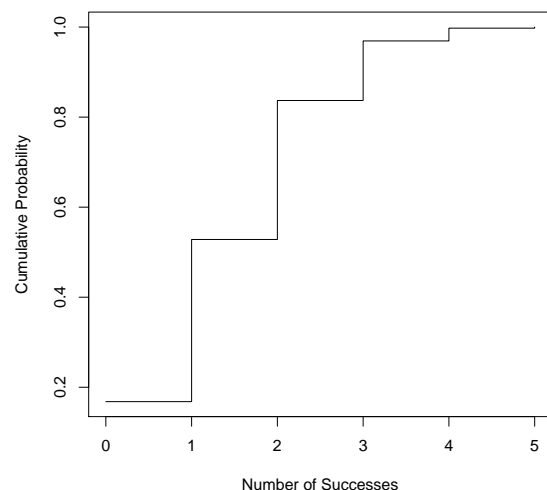


Los gráficos obtenidos siempre se mostrarán en la consola de R, con lo que, para acceder a los mismos, tendremos que ir a la barra de tareas y cambiar la ventana activa a la de R. Si activamos primero la *Gráfica de la función de probabilidad* y después la *Gráfica de la función de distribución*, obtendríamos los siguientes gráficos (por separado):

Binomial Distribution: Binomial trials=5, Probability of success=



Binomial Distribution: Binomial trials=5, Probability of success=



Para acabar, destacaremos que en el menú desplegado al seleccionar distribución binomial aparecen más opciones disponibles, como por ejemplo el cálculo de los cuantiles y la obtención de muestras aleatorias.

Más adelante, ya veremos la utilidad de estas dos opciones.



1.2.2. Distribución geométrica

La distribución geométrica está asociada a la distribución binomial que acabamos de ver. De manera específica, esta distribución describe la probabilidad del número x de ensayos de Bernoulli necesarios para obtener un éxito. Si la probabilidad de éxito en cada ensayo es p , entonces la probabilidad de que x ensayos sean necesarios para obtener un éxito viene definida por la expresión siguiente:

$$P(X = x) = (1 - p)^{x-1}p.$$

Para $x = 1, 2, 3, \dots$. La secuencia de probabilidades obtenida se denomina una progresión geométrica. Así pues, podemos definir la función de distribución de la manera siguiente.

$$P(X \leq x) = F(x) = 1 - (1 - 0,3)^x.$$

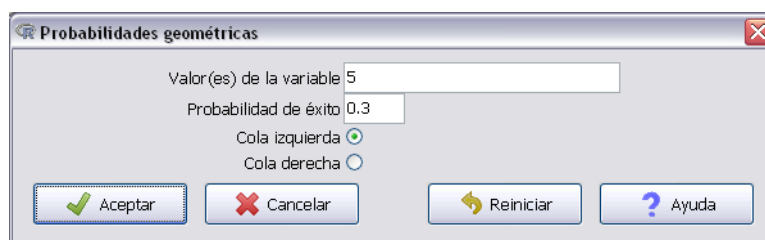
Si retomamos el ejemplo anterior del jugador de baloncesto que intenta encestar triples, en este caso podemos definir una nueva variable aleatoria X , que será el número del intento en el que el jugador encesta el primer triple, es decir, el número de ensayos necesarios hasta encestar. Formalmente, decimos que X sigue una distribución geométrica con parámetro $p = 0,3$, que se corresponde con la probabilidad de encestar un triple, como vimos en el ejemplo anterior. Si queremos saber la probabilidad de que enceste el primer triple dentro de los 6 primeros lanzamientos a canasta, haremos el cálculo siguiente.

$$P(Y \leq 6) = F(6) = 1 - (1 - 0,3)^6 = 0,8823.$$

Para hacer este cálculo con R-Commander, deberemos seguir esta ruta:

Distribuciones / Distribuciones discretas / Distribución geométrica / Probabilidades geométricas acumuladas

Obtendremos el cuadro de diálogo siguiente, donde introduciremos los datos de nuestro ejemplo:



Hay que tener en cuenta dos aspectos. El primero es que, en el espacio *Valor(es) de la variable*, debemos introducir **el número de fallos** antes del primer acierto, es decir,

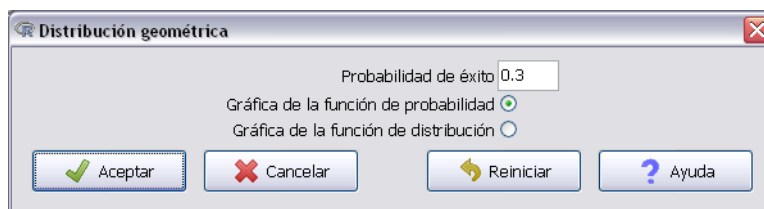
$x - 1$. En nuestro ejemplo, esto es $6 - 1 = 5$. El segundo aspecto que hay que tener en cuenta es que, ya que estamos calculando $P(Y \leq x)$, es necesario activar la opción *Cola izquierda*. En la ventana de resultados obtenemos lo siguiente:

```
> pgeom(c(5), prob=0.3, lower.tail=TRUE)
[1] 0.882351
```

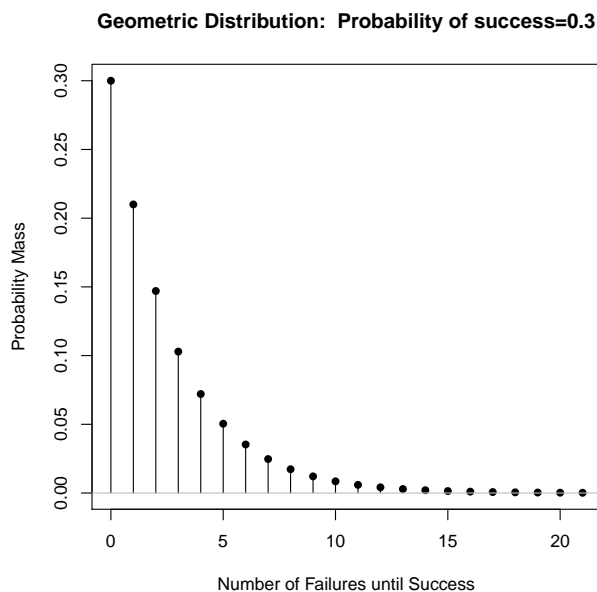
En lo que respecta a la distribución geométrica, es muy interesante ver el gráfico de su función de probabilidad. Esto se hace accediendo a esta ruta del menú desplegable:

Distribuciones / Distribuciones discretas / Distribución geométrica / Gráfica de la distribución geométrica

Nos aparecerá el cuadro de diálogo que se muestra a continuación, y en el que seleccionamos la función de probabilidad:



La interpretación del gráfico resultante es muy intuitiva: muestra, en el eje horizontal, el número de fallos hasta encestar el primer triple, y en el eje vertical, la probabilidad asociada. El primer valor coincide con $p = 0,3$, que es la probabilidad de encestar un triple, y de ahí va decreciendo hasta cero a medida que incrementa el número de intentos.



1.2.3. Distribución hipergeométrica

Hasta ahora hemos considerado que las pruebas u observaciones hechas eran independientes, pero con frecuencia no se da esta condición y el uso del modelo binomial puede resultar equivocado cuando las poblaciones estudiadas son pequeñas. De manera específica, supongamos que tenemos una muestra de N bolas, de las cuales N_1 son verdes y N_2 son rojas, de modo que $N_1 + N_2 = N$. Ante una extracción de n bolas de este conjunto (sin retorno), la variable X será el número de bolas verdes obtenidas. Decimos entonces que X sigue una distribución hipergeométrica tal que:

$$P(X = x) = \frac{\binom{N_1}{x} \binom{N_2}{n-x}}{\binom{N}{n}}.$$

Supongamos un ejemplo empírico. Un estudiante de oposiciones ha preparado doce de las dieciocho lecciones de las que consta un programa. El examen consiste en tres lecciones elegidas de manera aleatoria. ¿Qué probabilidad tiene el estudiante de conocer los tres temas? Sabemos que $N_1 = 12$, $N_2 = 6$ y $N = 18$. Además, la extracción es $n = 3$. Acudimos al menú desplegable y accedemos a la ruta siguiente:

Distribuciones / Distribuciones discretas / Distribución hipergeométrica / Probabilidades hipergeométricas

Nos aparecerá el cuadro de diálogo siguiente, donde introduciremos la información de nuestro ejemplo:

El resultado obtenido es el siguiente:

```
> .Table <- data.frame(Pr=dhyper(0:3 , m = 12 , n = 6
, k = 3))

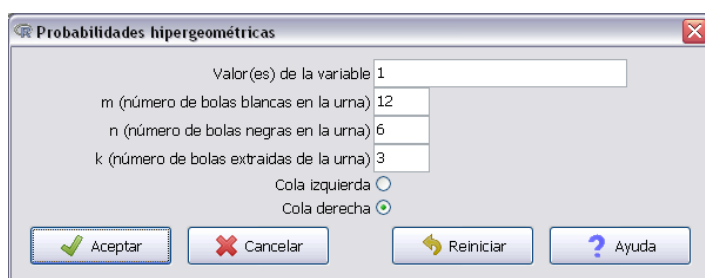
> rownames(.Table) <- 0:3

> .Table
      Pr
0 0.0245098
1 0.2205882
2 0.4852941
3 0.2696078
```

Como vemos, la probabilidad de que los 3 temas que ha estudiado entren en el examen es del 26,9 %. La probabilidad más alta es $P(X = 2) = 0,485$, es decir, que haya estudiado dos de los tres temas. Si nos preguntamos, por ejemplo, cuál es la probabilidad de que sepa como mínimo dos temas ($P(X \geq 2)$), necesitamos la probabilidad acumulada:

Distribuciones / Distribuciones discretas / Distribución hipergeométrica / Probabilidades hipergeométricas acumuladas

Es importante destacar que, en este caso, necesitamos la cola derecha. Aquí, R calcula $P(X > x)$, de manera que tendremos que introducir en el cuadro de diálogo el valor $x = 1$, ya que $P(X > 1) = P(X \geq 2)$:



De esta manera, obtenemos el resultado siguiente.

```
> phyper(c(1), m=12, n=6, k=3, lower.tail=FALSE)
[1] 0.754902
```

Es decir, hay una probabilidad del 75 % de que conozca dos o tres temas en el examen.

1.2.4. Distribución de Poisson

Otra de las distribuciones discretas importantes es la de Poisson, también derivada de la distribución binomial. Vamos a tomar el ejemplo de una variable aleatoria X , definida como el tiempo de espera del autobús en minutos, y que tiene como único parámetro λ , que representa tanto la media como la varianza de la variable aleatoria.

Supongamos que el tiempo medio de espera es de 3 minutos. Entonces $\lambda = 3$ minutos y, por tanto, $X \sim \text{Poiss}(3)$ o, de manera alternativa, $X \sim P(3)$. Para saber la probabilidad de que el autobús tarde 5 minutos, el cálculo que hay que hacer es el siguiente:

$$P(X = 5) = \frac{\lambda^k}{k!} e^{-\lambda} = \frac{3^5}{5!} e^{-3} = 0,1008.$$

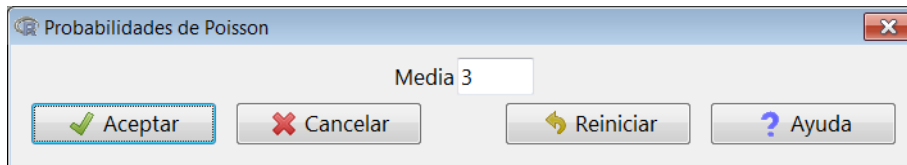
R-Commander nos facilita un resultado inmediato a este cálculo utilizando la ruta que se muestra a continuación del menú desplegable *Distribuciones*:

En una distribución de Poisson,
 $E(X) = V(X) = \lambda$.



Distribuciones / Distribuciones discretas / Distribución de Poisson / Probabilidades de Poisson

Como de costumbre, introduciremos los datos de nuestro ejemplo en el cuadro de diálogo que aparecerá:



Por defecto, R-Commander nos da los diez primeros valores de la probabilidad, es decir, desde $P(X = 0)$ hasta $P(X = 10)$:

```
> .Table <- data.frame(Pr=dpois(0:10 , lambda = 3))  
  
> rownames(.Table) <- 0:10  
  
> .Table  
      Pr  
0 0.0497870684  
1 0.1493612051  
2 0.2240418077  
3 0.2240418077  
4 0.1680313557  
5 0.1008188134  
6 0.0504094067  
7 0.0216040315  
8 0.0081015118  
9 0.0027005039  
10 0.0008101512
```

De manera similar a las anteriores distribuciones descritas en este manual, con la distribución de Poisson también se pueden calcular probabilidades acumuladas, gráficos y simulaciones muestrales.

1.3. Distribuciones de probabilidad continuas

1.3.1. Distribución uniforme

La distribución uniforme modela variables aleatorias continuas, de tal modo que todos los intervalos en el rango de la distribución tienen la misma longitud y son igualmente probables. El dominio de esta distribución está definido por los valores máximo y mínimo a y b , de manera respectiva. Una variable aleatoria X distribuida según una

uniforme se denota como $X \sim U(a, b)$. Su función de densidad viene dada por la expresión siguiente:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{para } a \leq x \leq b, \\ 0 & \text{para } x < a \text{ ó } x > b. \end{cases}$$

Mediante un ejemplo, veremos cómo trabajamos con variables aleatorias distribuidas según una uniforme. Supongamos que una mujer está dando a luz a un bebé, y la hora exacta del alumbramiento (X) sucederá en cualquier momento entre la hora 0 (ahora) y la hora 24 (la misma hora del día siguiente), de modo que se sigue una distribución uniforme, $X \sim U(0, 24)$.

La probabilidad de que la madre dé a luz dentro de las primeras cinco horas se calculará de la manera siguiente:

Distribuciones / Distribuciones continuas / Distribución uniforme / Probabilidades uniformes

En el cuadro de diálogo, introduciremos el valor de la variable de interés, el mínimo y el máximo:

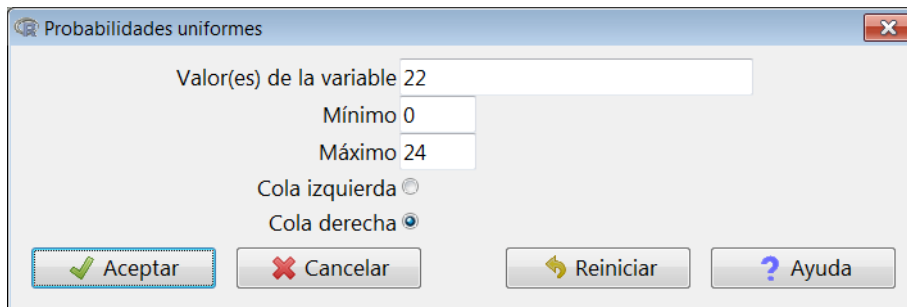
Tened presente que buscamos la probabilidad que queda en la cola de la izquierda, $P(X < 5)$, y esto deberemos indicarlo también en el cuadro de diálogo.

Obtendremos en la ventana de resultados la probabilidad siguiente:

```
> punif(c(5), min=0, max=24, lower.tail=TRUE)
[1] 0.2083333
```

¿Cómo se interpretan estos resultados? Hemos definido X como la hora del parto, y hemos obtenido que $P(X < 5) = 0,208$. Es decir, la probabilidad de que la madre dé a luz dentro de las primeras cinco horas es aproximadamente del 21 %. Es importante volver a insistir en que, cuando trabajamos con distribuciones continuas, $P(X < x)$ equivale a $P(X \leq x)$, ya que se calculan las áreas de la distribución y no puntos exactos. De esta manera, como ya hemos mencionado, en una distribución continua el cálculo de una probabilidad del tipo $P(X = x)$ será siempre nula, es decir, $P(X = x) = 0$.

Hay que recordar que podemos calcular probabilidades en las dos colas de la distribución. Esta opción está disponible en el cuadro de diálogo anterior, donde hemos tenido que especificar la cola de la distribución que nos interesa. Observad que hemos activado la opción *Cola izquierda*, ya que hemos calculado $P(X \leq x)$. Si ahora queremos calcular la probabilidad de que la madre dé a luz en las últimas dos horas (es decir, a partir de la hora 22), estaremos calculando $P(X > 22)$, y entonces deberemos activar la opción *Cola derecha*:



Con lo que obtenemos el resultado siguiente:

```
> punif(c(22), min=0, max=24, lower.tail=FALSE)
[1] 0.08333333
```

1.3.2. Distribución exponencial

La distribución exponencial está relacionada con la distribución de Poisson. Esta distribución modeliza el intervalo de tiempo que transcurre entre dos sucesos. Formalmente, tiene un único parámetro $\theta = \lambda > 0$, y **está definida para valores no negativos de la variable aleatoria**. Las funciones de densidad y de distribución toman la forma siguiente:

$$f(x) = \theta e^{-\theta x}$$

$$F(x) = P(X \leq x) = 1 - e^{-\theta x}$$

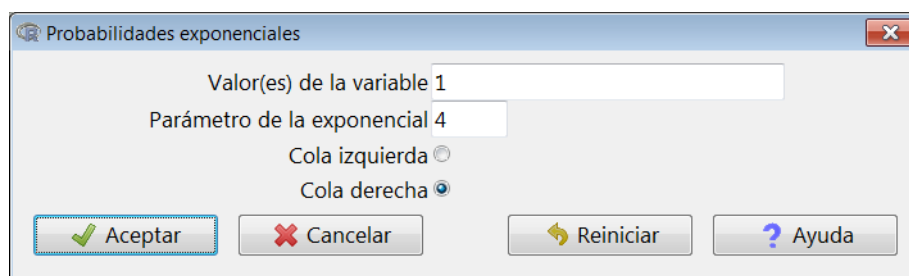
En la distribución exponencial,
 $E(X) = 1/\theta$ y $V(X) = 1/\theta^2$.



Veamos un ejemplo: en un gran hospital, el tiempo que transcurre entre dos partos (medido en horas) sigue una distribución exponencial con un parámetro $\theta = 4$. Esto significa que, ya que en esta distribución la esperanza viene definida como $E(X) = 1/\theta = 0,25$, de media transcurre un cuarto de hora ($1/4 = 0,25$) entre un parto y otro. Veamos la probabilidad de que transcurra una hora o más entre dos partos, es decir, queremos calcular $P(X > 1)$. La ruta que hay que seguir en R-Commander será:

Distribuciones / Distribuciones continuas / Distribución exponencial / Probabilidades exponenciales

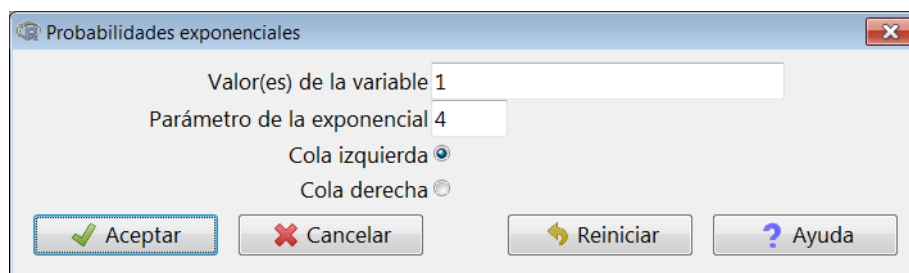
En el cuadro de diálogo emergente, introducimos la información siguiente:



El resultado será este:

```
> pext(c(1), rate=4, lower.tail=FALSE)
[1] 0.01831564
```

Es decir, $P(X > 1) = 0,018$. Esto indica que hay aproximadamente un 1,8 % de probabilidad de que transcurra una hora o más entre dos partos. De manera alternativa, podemos calcular la probabilidad complementaria, es decir, la probabilidad de que transcurra como máximo una hora entre dos partos. Para esto, será necesario seleccionar la opción *Cola izquierda* en el menú anterior:



El resultado obtenido será el siguiente:

```
> pext(c(1), rate=4, lower.tail=TRUE)
[1] 0.9816844
```

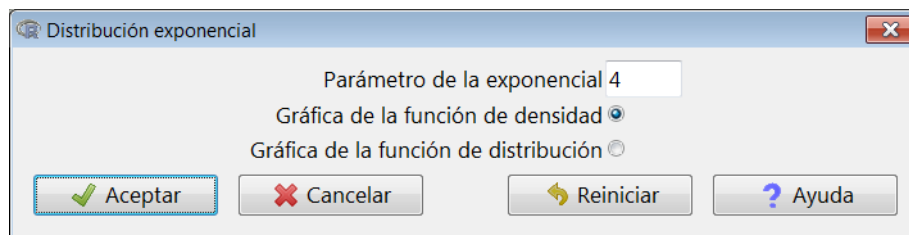
Este resultado no es de extrañar, ya que si sumamos las probabilidades obtenidas en las dos colas, la suma siempre es uno. Esto es así porque siempre se cumple que:

$$P(X \leq x) + P(X > x) = 1.$$

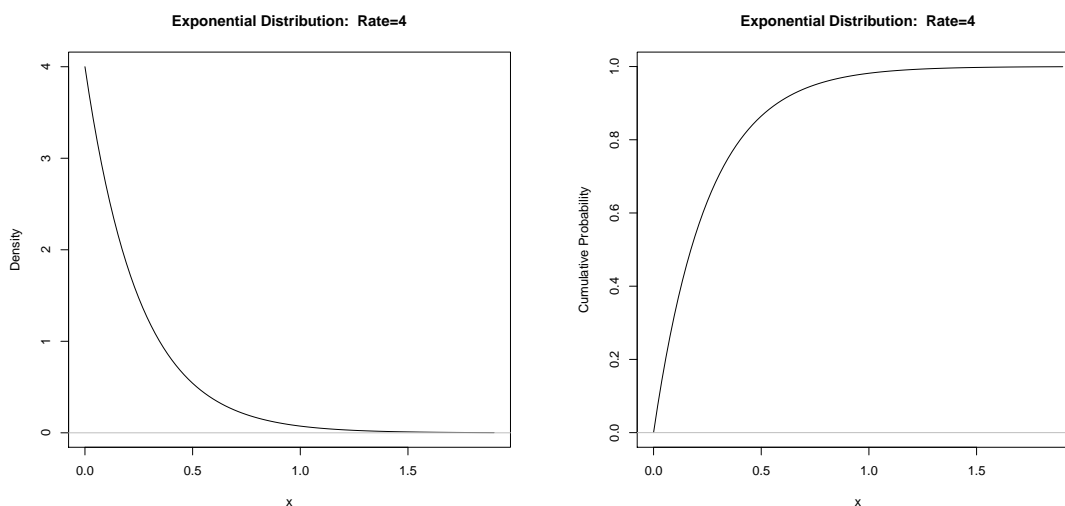
En este caso, $P(X \leq 1) = 1 - P(X > 1) = 1 - 0,018 = 0,982$. Un último aspecto que veremos de la distribución exponencial es el resultado gráfico de las probabilidades calculadas, mediante la ruta siguiente:

Distribuciones / Distribuciones continuas / Distribución exponencial / Gráfica de la distribución exponencial

Como con el resto de las distribuciones, podemos elegir entre la gráfica de la función de densidad y de la función de distribución:



Los gráficos obtenidos siempre se mostrarán en la consola de R, con lo que, para acceder a los mismos, tendremos que ir a la barra de tareas y cambiar la ventana activa a la de R. Si activamos primero la *Gráfica de la función de densidad* y después la *Gráfica de la función de distribución*, obtendríamos los siguientes gráficos (por separado):



Igual que con el resto de las distribuciones, en el menú desplegable aparecen más opciones disponibles, como por ejemplo el cálculo de los cuantiles y la obtención de muestras aleatorias.

1.3.3. Distribución normal

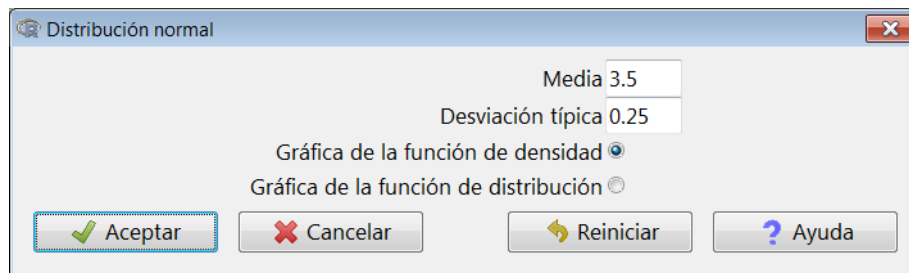
La distribución normal o gaussiana es la principal distribución de probabilidad en estadística, y se utiliza en infinidad de ámbitos. Esta distribución se caracteriza por tener dos parámetros: la media μ y la desviación típica o estándar σ . De este modo, se dice que una variable aleatoria X obedece a una ley normal si $X \sim N(\mu, \sigma)$. Su función de densidad viene definida por la expresión siguiente:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

Veamos un primer ejemplo. En el hospital del ejemplo anterior, el peso de los bebés al nacer sigue una distribución normal con media $\mu = 3,5$ y desviación típica $\sigma = 0,25$. Si queremos visualizar la forma de la función de densidad, seguiremos esta ruta:

Distribuciones / Distribuciones continuas / Distribución normal / Gráfica de la distribución normal

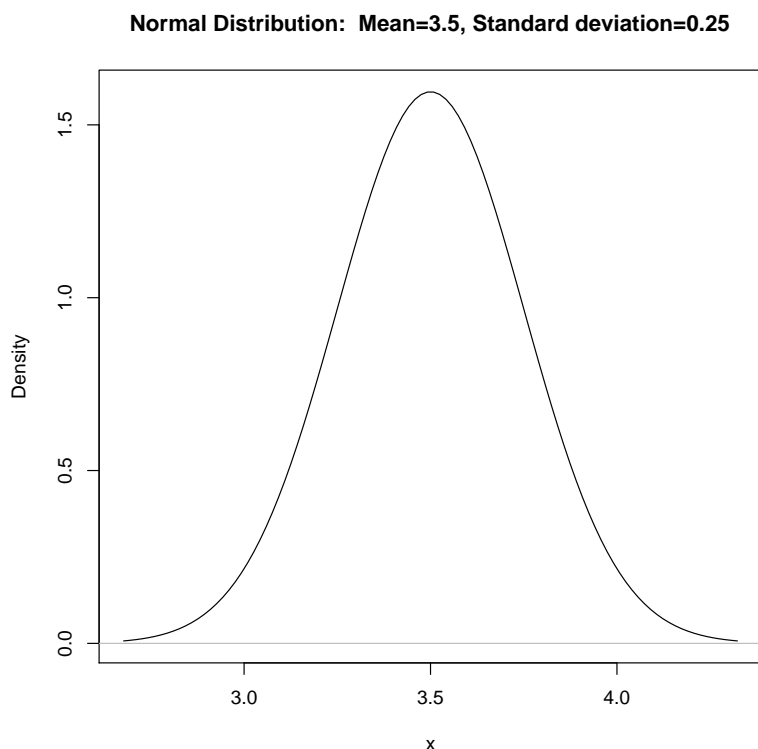
En el cuadro de diálogo resultante, introduciremos los parámetros de la distribución y seleccionaremos la opción *Gráfica de la función de densidad*:



La notación de la distribución Normal

Con frecuencia, se suele denotar como $X \sim N(\mu, \sigma)$ aquella variable que se distribuye según una normal. No obstante, no hay consenso sobre esta notación, ya que a veces se expresa la varianza (σ^2) en lugar de la desviación típica (dependiendo del libro o manual). Aquí utilizaremos siempre σ por defecto.

Con esto, obtenemos el gráfico de la función de densidad:



A simple vista, vemos cómo la mayor parte de los niños pesan entre 3 y 4 kilos al nacer. Calculemos ahora esta probabilidad ($P(3 \leq X \leq 4)$) con R-Commander accediendo a la ruta siguiente:

Distribuciones / Distribuciones continuas / Distribución normal / Probabilidades normales

El hecho de que sea una distribución simétrica nos ofrece diferentes opciones de calcular probabilidades. Si tomamos el gráfico de arriba, vemos que si trazamos una línea vertical en la media aritmética (3,5), en las dos partes de la distribución queda el 50 % de la masa probabilística. Esto mismo también se puede ver si nos fijamos en que $\mu = 3,5$ se sitúa en medio de los valores 3 y 4; entonces, el área que quedará a la izquierda de 3 y el área que quedará a la derecha de 4 serán iguales debido a la simetría de la distribución. Por tanto, si queremos calcular el área que queda entre 3 y 4, es decir, $P(3 \leq X \leq 4)$, una opción consiste en calcular uno menos las dos colas de los extremos, $P(X \leq 3)$ y $P(X \geq 4)$:

$$P(3 \leq X \leq 4) = 1 - P(X \leq 3) - P(X \geq 4)$$

Como sabemos que estos dos extremos han de tener el mismo área o valor, es decir, $P(X \leq 3) = P(X \geq 4)$, la formula anterior se puede simplificar, por ejemplo, de la manera siguiente.

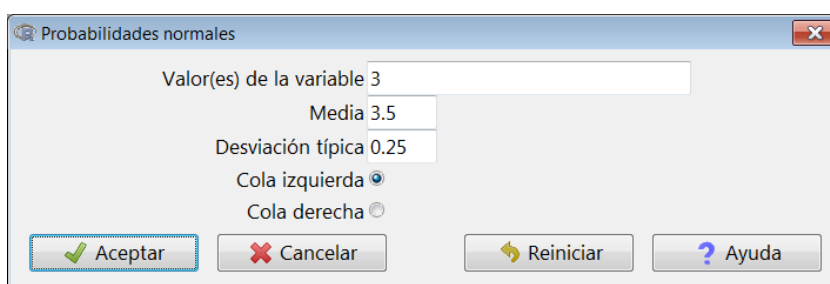
$$P(3 \leq X \leq 4) = 1 - 2 P(X \leq 3)$$

Otras maneras alternativas de obtener esta probabilidad son:

$$P(3 \leq X \leq 4) = 1 - 2 P(X \geq 4)$$

$$P(3 \leq X \leq 4) = P(X \leq 4) - P(X \leq 3)$$

Si seguimos la primera opción, la probabilidad $P(X \leq 3)$ con R-Commander se calcula así:



El resultado obtenido es el siguiente:

```
> pnorm(c(3), mean=3.5, sd=0.25, lower.tail=TRUE)
[1] 0.02275013
```

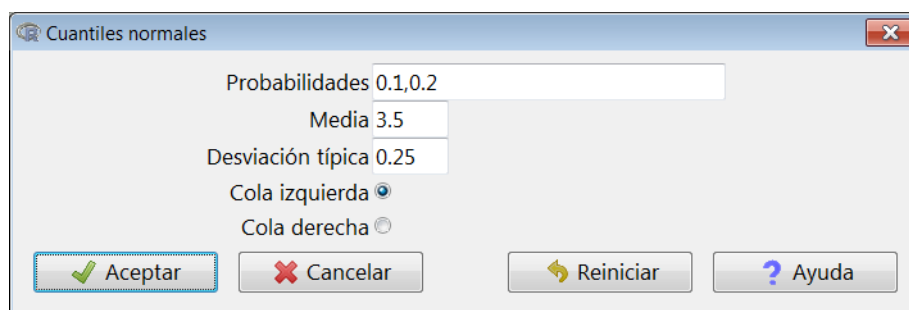
De esta manera, obtenemos que $P(3 \leq X \leq 4) = 1 - 2 \cdot 0,023 = 0,954$. Es decir, más del 95 % de los bebés nacerán dentro del intervalo de pesos [3,4]. **Consejo:** en el cuadro de diálogo anterior se puede introducir más de un valor, separando los valores por comas.

Os recomendamos que probéis el mismo cálculo en R-Commander siguiendo las maneras alternativas antes descritas.

Veamos ahora un ejemplo de cálculo de cuantiles. Supongamos que en este hospital, el jefe de pediatría quiere separar a los bebés en tres grupos: (a) muy poco peso, (b) poco peso y (c) peso normal, con el objetivo de hacer un uso eficiente de las incubadoras y del personal médico para la vigilancia de los bebés. El doctor decide que habrá un 10 % de los bebés en el primer grupo, un 20 % en el segundo y el resto (un 70 %) en el tercero. La pregunta es: ¿cuáles serán los puntos de corte entre los tres grupos? Para responder a esta pregunta, tenemos que calcular los percentiles 0,1 y 0,3, los cuales dividen el área total en las tres partes que nos interesan (es decir, 10 %, 20 % y 70 %, en este orden). Hagamos esta operación con R-Commander:

Distribuciones / Distribuciones continuas / Distribución normal / Cuantiles normales

Introducimos los cuantiles en el cuadro de diálogo:



Con esto, obtenemos los pesos que marcan el límite entre los tres grupos:

```
> qnorm(c(0.1,0.2), mean=3.5, sd=0.25, lower.tail=TRUE)
[1] 3.179612 3.289595
```

Estos percentiles indican que el primer grupo (bebés con muy poco peso) estará formado por los bebés que pesan menos de 3,18 kilos; el segundo (bebés con poco peso), por los que pesen entre 3,18 y 3,29 kilos; y el tercero (bebés con peso normal), por los que pesen más de 3,29 kilos.

Distribución de la media muestral

Un caso relevante en el estudio de la distribución normal es el de la **distribución de la media muestral** de una variable aleatoria. Para ilustrarlo, partamos del ejemplo anterior del hospital. Los doctores quieren hacer estadísticas de control del peso de los bebés que nacen. Interpretando los n bebés que nacen cada día como una muestra aleatoria independiente, se analiza para cada una de estas muestras la media muestral. Ya que cada día (es decir, cada muestra) se obtendrá una media muestral distinta, esta se puede analizar como una variable aleatoria denominada \bar{X} , que tiene su propia distribución. A la hora de analizar y calcular probabilidades sobre \bar{X} , es imprescindible distinguir si conocemos o no la dispersión de la población (σ) de la que se obtiene \bar{X} .

Si esta dispersión es conocida (como en el ejemplo anterior), \bar{X} se distribuye como una normal, tal que:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Si n es el tamaño de la muestra, vamos a suponer que disponemos de una muestra de 150 bebés ($n = 150$) y queremos calcular $P(\bar{X} \geq 3,75)$, sabiendo que $\mu = 3,5$ y asumiendo que σ es conocida e igual a 0,25.

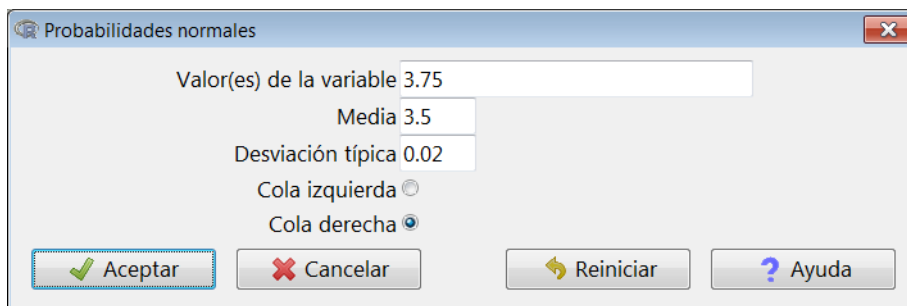
Antes de nada, debemos calcular cuál es la desviación típica de la variable media muestral \bar{X} :

$$\frac{\sigma}{\sqrt{n}} = \frac{0,25}{\sqrt{150}} = 0,02.$$

De este modo, si sabemos que $\bar{X} \sim N(3,5; 0,02)$, calcularemos esta probabilidad con R-Commander accediendo a la ruta siguiente:

Distribuciones / Distribuciones continuas / Distribución normal / Probabilidades normales

Tendremos que introducir los valores en el siguiente cuadro de diálogo:



El resultado obtenido es este:

```
> pnorm(c(3.75), mean=3.5, sd=0.02, lower.tail=FALSE)
[1] 3.732564e-36
```

Observamos que el resultado, en notación científica, es prácticamente cero. Este resultado indica que la probabilidad de que, en un día, la media muestral de los $n = 150$ bebés nacidos sea de 3,75 kilos ($P(\bar{X} \geq 3,75)$) es prácticamente nula.

1.3.4. Distribución t de Student

La distribución t de Student está asociada a la distribución normal que acabamos de ver. Se utiliza, entre otros casos, cuando desconocemos la dispersión de la variable que hay que analizar. Retomemos el ejemplo anterior del peso medio de los bebés que nacen en el hospital citado (\bar{X}). Supongamos que, como antes, tomamos una muestra de $n = 150$ bebés, pero en este caso solo sabemos que $\mu = 3,5$. Así pues, nos faltaría un parámetro para modelizar \bar{X} según una distribución normal.

En este caso, deberemos hacer una estimación de la desviación típica calculando la desviación típica muestral:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2},$$

de manera que en los cálculos del ejercicio anterior reemplazaremos σ por s . Entonces, la distribución muestral de la media ya no es una distribución normal como pasaba cuando conocíamos el verdadero valor de la desviación (σ).

Si σ es desconocida y n es el tamaño de la muestra, calcularemos el error estándar mediante este cociente:

$$\text{Error estándar} = \frac{s}{\sqrt{n}}$$

Así pues, si la variable que estudiamos sigue una distribución normal con media μ y desviación típica desconocida, entonces

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

sigue una distribución t de Student con $n - 1$ grados de libertad. En nuestro caso podemos definir una nueva variable aleatoria \bar{Y} , que representará el peso medio de estos bebés que nacen en el hospital citado. De manera formal, establecemos así esta distribución:

$$\bar{Y} \sim t_{n-1}$$

Supongamos que queremos saber la probabilidad de que el peso medio de los bebés sea inferior a 3 kilos, es decir, $P(\bar{X} \leq 3)$, sabiendo únicamente que $\mu = 3,5$, $n = 150$ y $s = 0,18$. El primer paso será transformar la variable \bar{X} en la variable \bar{Y} con la que podamos operar. En este caso, el objetivo será calcular $P(\bar{Y} \leq y)$. El valor de y lo obtenemos a partir de la expresión:

$$y = \frac{x - \mu}{\frac{s}{\sqrt{n}}} = \frac{3 - 3,5}{\frac{0,18}{\sqrt{150}}} = -34,02.$$

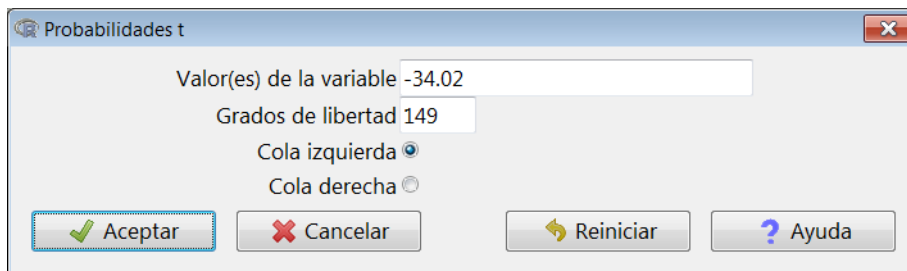
Muchas funciones de distribución se derivan de la normal estándar

A partir de la distribución normal tipificada, obtenemos la función chi-cuadrado $\chi_n^2 = \sum_{i=1}^n Z_i^2$; la t de Student $t_n = \frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$ y la F de Snedecor $F_{n,d} = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_d^2}{d}}$.

Una vez ya tenemos el valor de la nueva variable distribuida como una *t* de Student, para efectuar el cálculo de $P(Y \leq -34,02)$ con R-Commander utilizaremos la ruta siguiente:

Distribuciones / Distribuciones continuas / Distribución t-Student / Probabilidades t

El cuadro de diálogo que tenemos que rellenar es el siguiente:



Hay que considerar que hemos introducido el valor que nos ha dado de la nueva variable en *Valor(es) de la variable*, y que en *Grados de libertad* hemos introducido $n - 1 = 150 - 1 = 149$. Además, ya que estamos calculando $P(Y \leq y)$, hay que activar la opción *Cola izquierda*. El resultado obtenido es el siguiente:

```
> pt(c(-34.02), df=149, lower.tail=TRUE)
[1] 1.973326e-72
```

Así pues, hay una probabilidad nula de que el peso medio de los 150 bebés nacidos ese día sea inferior a los 3 kilos.

1.3.5. Teorema del límite central

El teorema del límite central (TLC) establece que si una muestra es lo bastante grande ($n > 30$), sea cual sea la distribución de la variable de interés, **la distribución de la media muestral** seguirá aproximadamente una distribución normal. Además, la media será la misma que la de la variable de interés, y la desviación típica de la media muestral será aproximadamente el error estándar. El TLC tiene algunas implicaciones, como que una variable aleatoria X distribuida según una binomial o una Bernoulli se puede aproximar a una normal. En el caso de la distribución binomial, cuando $n > 30$ y np y $n(1 - p)$ son mayores que 5, la variable X se aproxima a una normal con $E(X) = np$ y $Var(X) = np(1 - p)$.

El teorema de Moivre-Laplace

Según este teorema, una distribución binomial se puede aproximar a una distribución normal siempre que se cumpla que $np(1 - p) > 5$.

Retomemos el ejemplo visto en el caso de la distribución binomial, el del jugador de baloncesto que encestaba triples. Recordemos que $p = 0,3$, pero supongamos ahora que el número de intentos (ensayos) se eleva hasta $n = 50$. Queremos calcular la

probabilidad de que el jugador enceste más de 20 triples. Por el teorema de Moivre-Laplace y el TLC sabemos que:

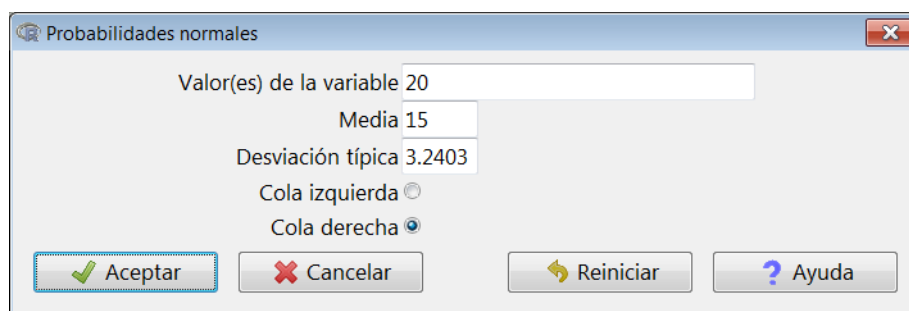
$$E(X) = np = 15$$

$$Var(X) = np(1 - p) = 10,5$$

La ruta que hay que seguir en R-Commander para calcular $P(X > 20)$ será la siguiente:

Distribuciones / Distribuciones continuas / Distribución normal / Probabilidades normales

Y el cuadro de diálogo que nos aparecerá será:



Lo primero que hay que destacar es que **no hay que introducir el valor de la varianza en el espacio de la desviación típica; primero debemos calcularla haciendo la raíz cuadrada**: es decir, $\sqrt{10,5} = 3,2403$. El resultado de esta probabilidad es el siguiente:

```
> pnorm(c(20), mean=15, sd=3.2403, lower.tail=FALSE)
[1] 0.06140726
```

Es decir, el jugador tiene un poco más del 6 % de probabilidades de encestar 20 triples en 50 intentos.

El segundo caso planteado hace referencia a las variables discretas dicotómicas que solo toman los valores 0 y 1, y que siguen una distribución de Bernoulli. Lógicamente estas variables están estrechamente vinculadas con las proporciones, que de hecho son la media de n variables aleatorias de Bernoulli de parámetro p , donde n es el tamaño de la muestra y p , la probabilidad de éxito de cada evento individual. Suponiendo la variable aleatoria dicotómica Y , cuando el tamaño de la muestra n sea grande, la distribución de la proporción Y será aproximadamente una distribución normal cuyos parámetros son $E(Y) = p$ y $Var(Y) = p(1 - p)/n$. Así pues, al calcular la desviación típica, siempre que hablemos de variables de Bernoulli esta será igual al error estándar de dicha variable.

En nuestro ejemplo, teníamos que $n = 50$ y $p = 0,3$, y queríamos averiguar la probabilidad de que el jugador de baloncesto encestrara más de 20 triples. Si suponemos que estamos ahora aproximando una distribución de Bernoulli a una normal, lo que estaremos especificando es la proporción de triples que efectúa el jugador en 50 intentos. Entonces, si queremos calcular la probabilidad de que esta proporción sea superior al 40 %, los parámetros de la distribución normal que deberemos utilizar serán los siguientes:

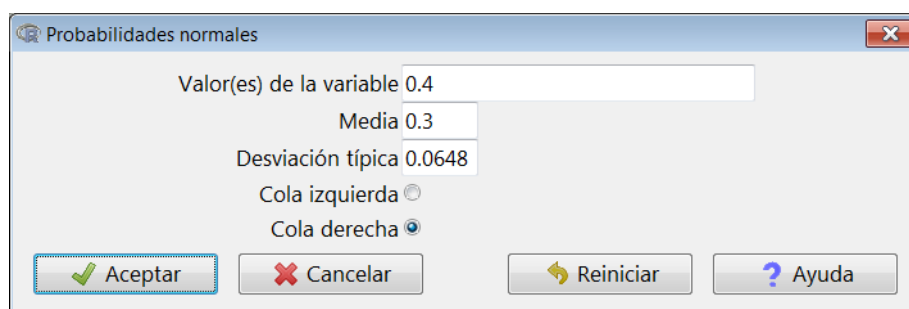
$$E(Y) = p = 0,3$$

$$Var(Y) = \frac{p(1-p)}{n} = 0,0042$$

Y la ruta que hay que seguir en el menú desplegable de R-Commander para calcular $P(Y > 0,4)$ será:

Distribuciones / Distribuciones continuas / Distribución normal / Probabilidades normales

Con el siguiente cuadro de diálogo:



Como antes, debemos destacar que **no hay que introducir el valor de la varianza en el espacio de la desviación típica**, esto es, hay que introducir $\sqrt{0,0042} = 0,06480$. Además, es preciso resaltar que el resultado de esta probabilidad es exactamente el mismo que el de la calculada en el ejemplo anterior. Esto resulta del todo lógico, ya que una distribución de Bernoulli no es más que un caso particular de la distribución binomial. En nuestro caso, el 40 % de triples equivale a encestar 20 triples en 50 intentos. El resultado obtenido en R-Commander confirma este apunte:

```
> pnorm(c(0.4), mean=0.3, sd=0.06480741, lower.tail=
  FALSE)
[1] 0.0614113
```

Es decir, el jugador tiene un poco más del 6 % de probabilidades de encestar el 40 % de los triples, o lo que es lo mismo, encestar 20 triples en 50 intentos.

2. Inferencia estadística

2.1. Introducción

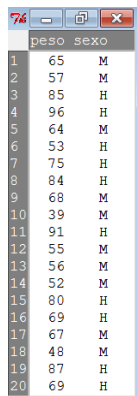
Este apartado está dedicado al estudio de las principales herramientas de la inferencia estadística: los intervalos de confianza (IC) y los contrastes de hipótesis (CH). Normalmente, disponemos de una serie de datos muestrales de los que ignoramos los verdaderos parámetros poblacionales que han generado esta muestra. Para esto, los tendremos que estimar. En concreto, aprenderemos a trabajar con IC y CH, y veremos cómo se calcula un IC y se resuelve un CH para la media aritmética, para la proporción y para la varianza. Además, trabajaremos la construcción de IC y la resolución de CH para la diferencia de medias tanto para muestras apareadas como independientes, la diferencia de proporciones y el cociente de varianzas.

2.2. Inferencia sobre la media con varianza poblacional desconocida

A la hora de plantear el cálculo de intervalos de confianza para el parámetro de la media poblacional μ , en el caso de desconocer la varianza poblacional (σ^2) tendremos que utilizar la varianza muestral s^2 y la desviación estándar muestral (s). Como consecuencia, al utilizar un parámetro estimado trabajaremos con la distribución t de Student. De este modo, con una muestra n y un nivel de significación α , el IC para el parámetro μ vendrá dado por la siguiente expresión que se obtiene a partir de la media muestral \bar{X} :

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}.$$

Veamos un ejemplo práctico de inferencia sobre el parámetro μ con varianza poblacional desconocida. Supongamos que queremos estudiar el peso de los estudiantes de una clase, discriminando por sexos. Los datos iniciales son los siguientes:



	peso	sexo
1	65	M
2	57	M
3	85	H
4	96	H
5	64	M
6	53	H
7	75	H
8	84	H
9	68	M
10	39	M
11	91	H
12	55	M
13	56	M
14	52	M
15	80	H
16	69	H
17	67	M
18	48	M
19	87	H
20	69	H

En total hay $n = 20$ estudiantes en la muestra, de los cuales 10 son chicos y 10 son chicas. Vemos que tenemos dos variables: peso y sexo. La primera es una variable numérica, mientras que la segunda es un factor que informa sobre si la observación es de un hombre (H) o una mujer (M).

Veamos, para empezar, la estadística descriptiva de estos datos accediendo a la ruta siguiente:

Estadísticos / Resúmenes / Conjunto de datos activo

Esta instrucción hace que aparezca este resultado:

```

      peso  sexo
Min.   : 39.00 H:10
1st Qu.:55.75 M:10
Median :67.50
Mean   :68.00
3rd Qu.:81.00
Max.   :96.00

```

Otra manera alternativa de obtener más estadísticos descriptivos, como vimos en el módulo 3, es mediante la ruta siguiente:

Estadísticos / Resúmenes / Resúmenes numéricos

Si pulsamos *Aceptar* en el menú que nos aparece, obtendremos la siguiente información referente a la variable numérica (peso):

```

mean      sd 0 %   25 %   50 %  75 % 100 %  n
68 15.55297 39 55.75 67.5  81   96 20

```

El primer paso será hacer inferencia sobre la media poblacional, para lo que utilizaremos tanto el CH como el IC. Como veremos, los dos cálculos están íntimamente relacionados. Queremos preguntarnos si la media de peso de la clase se corresponde con la de la escuela, que es, supongamos, $\mu_0 = 70$. Una primera aproximación consiste en hacer un contraste de hipótesis bilateral, mediante el cual plantearemos las hipótesis siguientes:

$$H_0 : \mu = 70$$

$$H_1 : \mu \neq 70$$

Los posibles resultados de este contraste son dos: o bien no se rechaza H_0 o bien se rechaza H_0 . Es importante destacar que el planteamiento de un contraste requiere un valor μ_0 y un nivel de confianza (o de manera alternativa, de significación). Además,

Las hipótesis nunca se aceptan

Es muy importante aclarar que, aunque parezca lo mismo, estadísticamente no tiene ningún sentido **aceptar** una hipótesis. Lo que hacemos es **no rechazarla** con la información de la que disponemos y al nivel de confianza dado. Por tanto, nuestros resultados pueden cambiar si nos modifican el nivel de confianza o la información con la que estamos trabajando.

la H_1 puede estar planteada a dos colas (bilateral) o a una cola (unilateral). En este último caso, puede ser por la izquierda ($H_1 : \mu < \mu_0$) o por la derecha ($H_1 : \mu > \mu_0$).

Un intervalo de confianza parte de un razonamiento similar. La principal diferencia es que no hay que proporcionar ningún valor μ_0 , sino que es preciso buscar aquellos dos valores de la variable aleatoria (en este caso, la media poblacional μ) que dejen en las dos colas un porcentaje α , que es el nivel de significación. Dicho de otro modo, la probabilidad de que el verdadero valor de μ esté dentro del intervalo calculado será igual al nivel de confianza $1 - \alpha$.

Veamos estos conceptos en nuestro ejemplo. Si tomamos la ruta siguiente:

Estadísticos / Medias / Test t para una muestra

Obtendremos el siguiente cuadro de diálogo. Aquí tenemos que introducir el valor μ_0 (que hemos fijado en 70), si queremos un contraste unilateral o bilateral, y el nivel de confianza, que es $(1 - \alpha) = 0,95$. Esto implica que trabajaremos con una significación de $\alpha = 0,05$.

El resultado es el siguiente:

```
One Sample t-test
data: datos$peso
t = -0.5751, df = 19, p-value = 0.572
alternative hypothesis: true mean is not equal to 70
95 percent confidence interval:
 60.72099 75.27901
sample estimates:
mean of x
      68
```

¿Cómo interpretamos este resultado? Por una parte, el contraste de hipótesis nos da un estadístico de $t = -0,5751$ y un p-valor de 0,572. Esto nos indica que el estadístico está situado en la región de aceptación, y no en la región crítica, con lo que no rechazamos la H_0 . Dicho de otro modo, la media muestral obtenida no es estadísticamente

Al hacer inferencia con R-Commander, el p-valor obtenido ya considera si el contraste es unilateral o bilateral. Por tanto, este p-valor **siempre se ha de comparar con α y no con $\alpha/2$** .

diferente de 70. En general, hemos de seguir la siguiente regla para la resolución de cualquier contraste:

$$\begin{array}{l} p\text{-valor} \leq \alpha \Rightarrow \text{Rechazo de } H_0 \\ p\text{-valor} > \alpha \Rightarrow \text{No rechazo de } H_0 \end{array}$$

En lo que respecta al intervalo de confianza, vemos que hemos obtenido el intervalo $[60, 72 ; 75, 28]$. Este resultado es coherente con el anterior, ya que el valor $\mu_0 = 70$ está incluido en el intervalo.

¿Qué pasaría si fijásemos un valor $\mu_0 = 80$? Vamos a hacer de nuevo el test con este dato (bilateral y con $\alpha = 0,05$). Los resultados que obtendremos serán:

```
One Sample t-test

data: datos$peso
t = -3.4505, df = 19, p-value = 0.00268
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
 60.72099 75.27901
sample estimates:
mean of x
68
```

Vemos que, en este caso, rechazamos la H_0 , ya que $p\text{-valor} < 0,05$, y esto coincide con el hecho de que 80 no está en el intervalo $[60, 72 ; 75, 28]$.

2.3. Inferencia sobre la media con varianza poblacional conocida

A diferencia del caso anterior, ahora sí conocemos el parámetro de la varianza poblacional (con desviación estándar σ). Utilizaremos entonces la distribución normal para el cálculo del intervalo para el parámetro μ . De manera específica, tendremos la siguiente expresión para el intervalo de confianza también basado en la media muestral de la que disponemos (\bar{x}):

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

En el ejemplo anterior hemos asumido que no conocemos la desviación típica poblacional, y para hacer inferencia (IC y CH) hemos tomado la desviación típica muestral, que es $s = 15,55$. Supongamos que conocemos la desviación típica poblacional, y que esta es $\sigma = 20$. En este caso, para el cálculo de IC y de CH tendremos que utilizar la distribución normal, y no la t de Student.

Si nos fijamos, esta opción no está configurada en el menú de R-Commander y por este motivo tendremos que introducirla manualmente. Previamente, necesitaremos instalar un paquete estadístico denominado PASWR (que significa *Probability And Statistics With R*). Recordad que todos los paquetes adicionales que necesitemos solo los tendremos que instalar una vez, y después ya podremos disponer de los mismos sin necesidad de instalarlos de nuevo. Para proceder a la instalación, deberemos introducir lo siguiente en la ventana de instrucciones:

```
> install.packages("PASWR")
```

Después, seleccionaremos esta línea y pulsaremos *Ejecutar*, con lo que nos saldrá un menú desplegable de repositorios, y elegiremos uno cualquiera. Una vez instalado el paquete PASWR, tendremos que cargarlo mediante la ruta siguiente:

Herramientas / Cargar paquete(s) / PASWR

Este segundo paso sí que deberemos hacerlo siempre que queramos utilizar la función `z.test`. Plantearemos el contraste siguiente:

$$H_0 : \mu = 80$$

$$H_1 : \mu \neq 80$$

Introduciremos en la ventana de instrucciones el test de la manera siguiente.

```
> z.test(datos$peso, alternative='two.sided', mu=80,
+ sigma.x=20, conf.level=.95)
```

Cuando tengamos escrita esta instrucción, deberemos seleccionarla y pulsar la opción *Ejecutar*. Además de la variable *peso*, colocada en primera posición, veamos qué otras instrucciones componen la función `z.test`:

alternative	'greater': unilateral por la derecha 'less': unilateral por la izquierda 'two.sided': bilateral
mu	Valor de μ_0
sigma.x	Valor de σ
conf.level	Nivel de confianza ($1 - \alpha$)

En la ventana de resultados, aparecerá lo siguiente:

```
One-sample z-Test

data: datos$peso
z = -2.6833, p-value = 0.00729
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
59.23477 76.76523
sample estimates:
mean of x
68
```

La conclusión que extraemos es que rechazamos H_0 , es decir, el parámetro μ es estadísticamente distinto de 80.

2.4. Inferencia sobre la proporción

En este caso, estamos interesados en hacer inferencia sobre el parámetro π , es decir, la proporción poblacional. Si \hat{p} es la proporción muestral y utilizamos la distribución normal, el intervalo de confianza vendrá dado por esta expresión:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Siguiendo con el ejemplo anterior, tenemos que la muestra de $n = 20$ observaciones está dividida entre hombres ($n_h = 10$) y mujeres ($n_m = 10$). Este factor nos permite hacer inferencia sobre las proporciones. Ya que los factores están ordenados de manera alfabética como H y M , la proporción por defecto será $p = n_h/n$, es decir, la proporción de hombres sobre el total. La proporción muestral es de $\hat{p} = 0,5$. Supongamos que deseamos contrastar si la proporción es estadísticamente diferente de $\pi = 0,6$. Al 95 % de confianza y planteado de modo bilateral, este contraste se plantea así:

$$H_0 : \pi = 0,6$$

$$H_1 : \pi \neq 0,6$$

Algunos autores y R-Commander denominan el parámetro de proporción poblacional p en lugar de π . Los dos son correctos: lo que ocurre es que el alfabeto griego se suele reservar para los parámetros referidos a la población y, por este motivo, nosotros hemos decidido utilizar π .

Intervalos de máxima holgura

En ocasiones, no disponemos de suficiente información como para utilizar la proporción muestral \hat{p} . En estos casos se utiliza $\hat{p} = 0,5$ ya que da lugar al IC más ancho posible, denominado de máxima holgura.

Para hacer el test, procederemos de la manera siguiente:

Estadísticos / Proporciones / Test de proporciones para una muestra

Obtenemos el siguiente cuadro de diálogo, similar a los casos anteriores:

El resultado del contraste será:

```
1-sample proportions test without continuity
correction

data: rbind(.Table), null probability 0.6
X-squared = 0.8333, df = 1, p-value = 0.3613
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.299298 0.700702
sample estimates:
 p
0.5
```

La interpretación de este resultado es análoga a los casos anteriores. El contraste indica un no rechazo de H_0 . Además, se comprueba que el intervalo de confianza es, aproximadamente, $[0,3 ; 0,7]$. Es importante destacar cómo el IC y el CH son dos caras de la misma moneda: si la proporción $\pi = 0,6$ hubiera estado fuera de este intervalo, habríamos rechazado H_0 .

2.5. Inferencia sobre la varianza

En esta sección, implementaremos la inferencia con un CH y un IC basados en la distribución χ^2 . De manera específica, queremos analizar si el valor de la varianza

poblacional (σ^2) es estadísticamente distinto a un valor predefinido σ_0^2 . Este contraste, planteado bilateralmente, toma la forma siguiente.

$$\begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{array}$$

En el ejemplo anterior, hemos visto que la distribución estándar muestral toma el valor $s = 15,55297$, con lo que la varianza muestral es $s^2 = 241,89$. Supongamos que deseamos hacer inferencia sobre el parámetro σ^2 , de manera que queremos calcular un intervalo al 95 % de confianza de este parámetro. Sabemos que el estadístico toma la expresión siguiente.

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

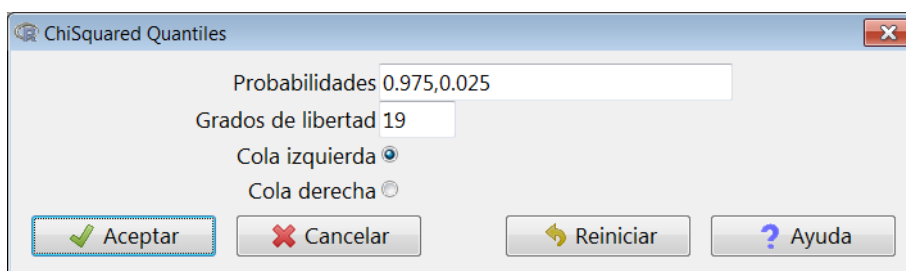
Por tanto, el intervalo que incluye el verdadero valor poblacional de σ^2 tomará la forma siguiente.

$$P\left(\frac{(n-1)s^2}{\chi_{1-\alpha/2; n-1}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{\alpha/2; n-1}^2}\right) = 1 - \alpha$$

El primer paso será calcular los valores críticos. Con un nivel de confianza de $\alpha = 0,05$ y una muestra de $n = 20$, estos valores serán $\chi_{0,975; 19}^2$ y $\chi_{0,025; 19}^2$, que equivalen a los cuantiles 0,975 y 0,025 de la distribución χ^2 con 19 grados de libertad, respectivamente. En R-Commander, lo hacemos de la manera siguiente.

Distribuciones / Distribuciones continuas / Distribución chi-cuadrado / Cuantiles chi-cuadrado

Obtendremos este cuadro de diálogo:



En algunos libros, podéis encontrar esta relación como $\frac{(n-1)s^2}{\sigma^2} \sim \chi_n$. En realidad no hay consenso pero, a medida que crece n , las diferencias entre uno y otro cálculo disminuyen.

Pensad que la distribución χ^2 no es simétrica y, por tanto, deberemos buscar los dos valores para construir el IC.

Con esto, obtenemos los siguientes valores de $\chi^2_{0,975; 19}$ y $\chi^2_{0,025; 19}$:

```
> qchisq(c(0.975, 0.025), df=19, lower.tail=TRUE)
[1] 32.852327 8.906516
```

Estos valores marcan los límites dentro de los cuales el estadístico cae en la región de no rechazo de H_0 . Para calcular los valores del intervalo para el parámetro σ^2 , tendremos que calcular los extremos del intervalo de la anterior expresión. Si efectuamos los cálculos manualmente, obtenemos este resultado:

```
> (20-1) * (15.55297^2) / c(32.852327, 8.906516)
[1] 139.8988 516.0270
```

Es decir, al 95 % de confianza, el parámetro poblacional σ^2 estará en el IC = [139,9 ; 516,0].

2.6. Inferencia sobre la diferencia de medias con muestras independientes

El objetivo de este análisis es comprobar si hay diferencias estadísticamente significativas entre la media de una misma variable (peso) extraída en dos muestras independientes diferenciadas por la variable (sexo). La primera muestra está compuesta por hombres (H) y la segunda, por mujeres (M). Así pues, para ver si hay diferencias en el peso de las dos muestras, se plantea el contraste siguiente:

$$H_0 : \mu_h = \mu_m$$

$$H_1 : \mu_h \neq \mu_m$$

De manera alternativa, también se puede expresar así:

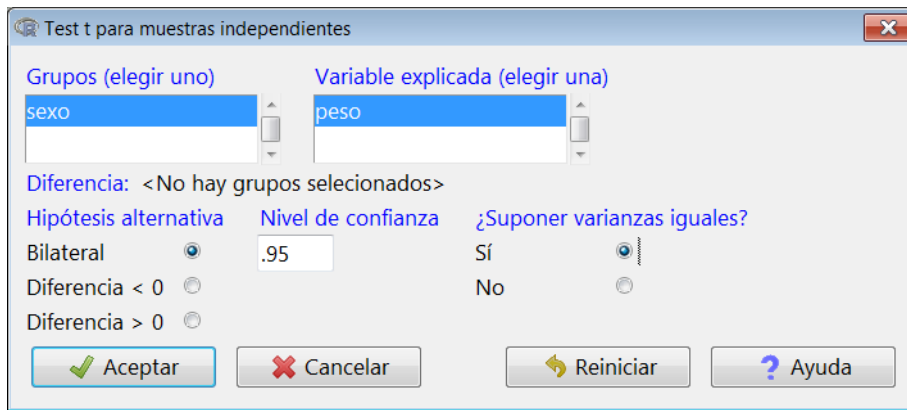
$$H_0 : \mu_h - \mu_m = 0$$

$$H_1 : \mu_h - \mu_m \neq 0$$

2.6.1. Varianzas poblacionales desconocidas pero iguales

En R-Commander, utilizaremos la siguiente ruta y obtendremos el cuadro de diálogo que se muestra a continuación:

Estadísticos / Medias / Test t para muestras independientes



Hay que destacar que este contraste (*Test t*) asume que las varianzas son desconocidas. El cuadro de diálogo nos ofrece la posibilidad de elegir si estas son o no iguales. En este caso, hemos supuesto que son iguales. El resultado es el siguiente:

```
Two Sample t-test

data: peso by sexo
t = 4.3896, df = 18, p-value = 0.0003535
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 11.36613 32.23387
sample estimates:
mean in group H mean in group M
      78.9      57.1
```

El resultado del contraste nos indica que la H_0 se rechaza, y que las medias son estadísticamente distintas. Llegamos a esta misma conclusión considerando el IC para el parámetro $\mu_h - \mu_m$, [11, 37 ; 32, 23]. Como este intervalo no contiene el valor cero, el peso medio de los hombres es estadísticamente distinto al de las mujeres. Es más, al ser los valores del IC positivos, podemos afirmar que la media del peso de los hombres es estadísticamente superior a la de las mujeres.

2.6.2. Varianzas poblacionales conocidas

Este contraste varía si conocemos los valores de las varianzas poblacionales de μ_h y μ_m . En este caso, tendremos que utilizar un contraste basado en la distribución normal y no en la distribución t de Student. De manera análoga al caso de trabajar con una sola muestra, R-Commander no dispone de este contraste en el menú y tendremos que introducirlo manualmente. En la ventana de instrucciones, primero crearemos dos variables diferenciadas: el peso de los hombres y el de las mujeres, cada una con 10 observaciones:

```
> peso_h<-Datos$peso[Datos$sexo=="H"]  
> peso_m<-Datos$peso[Datos$sexo=="M"]
```

Una vez creadas estas variables, utilizaremos de nuevo la función `z.test` del paquete PASWR, esta vez con dos variables, y especificaremos la desviación estándar de μ_h y μ_m (`sigma.x` y `sigma.y`, respectivamente):

```
> z.test(peso_h, peso_m, alternative='two.sided',  
+ sigma.x=4, sigma.y=3, conf.level=0.95)
```

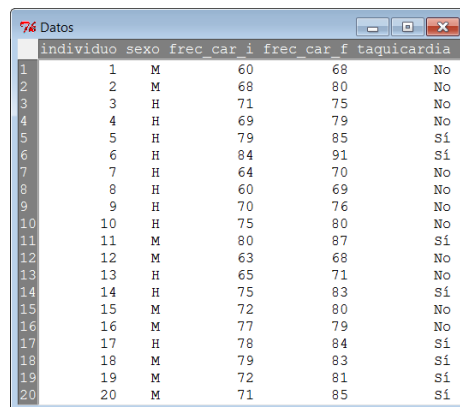
El resultado de este contraste que aparecerá en la ventana de resultados es:

```
Two-sample z-Test  
  
data: peso_h and peso_m  
z = 13.7875, p-value < 2.2e-16  
alternative hypothesis: true difference in means is not  
equal to 0  
95 percent confidence interval:  
18.70102 24.89898  
sample estimates:  
mean of x mean of y  
78.9      57.1
```

Observamos que, en este caso, el IC es mucho más preciso, es decir, el intervalo es más estrecho, debido a que conocemos los datos poblacionales y utilizamos la distribución normal en lugar de la distribución *t* de Student.

2.7. Inferencia sobre la diferencia de medias con muestras apareadas

Una muestra apareada implica que se toman dos o más observaciones de los mismos individuos. Para ilustrarlo, consideraremos unos datos ficticios sobre medicina. Supongamos que queremos averiguar si un determinado fármaco afecta a la frecuencia cardíaca. Para esto, se diseña un experimento médico con una muestra de $n = 20$ personas, entre las que hay hombres (*H*) y mujeres (*M*). El experimento consistió en registrar la frecuencia cardíaca antes (*frec_car_i*) y después (*frec_car_f*) de tomar este fármaco, además de registrar mediante una variable dicotómica si la persona tuvo taquicardia. Las variables de la muestra son las siguientes:



	individuo	sexo	frec_car_i	frec_car_f	taquicardia
1	1	M	60	68	No
2	2	M	68	80	No
3	3	H	71	75	No
4	4	H	69	79	No
5	5	H	79	85	Si
6	6	H	84	91	Si
7	7	H	64	70	No
8	8	H	60	69	No
9	9	H	70	76	No
10	10	H	75	80	No
11	11	M	80	87	Si
12	12	M	63	68	No
13	13	H	65	71	No
14	14	H	75	83	Si
15	15	M	72	80	No
16	16	M	77	79	No
17	17	H	78	84	Si
18	18	M	79	83	Si
19	19	M	72	81	Si
20	20	M	71	85	Si

El objetivo del test es comprobar si la media de la frecuencia cardíaca es estadísticamente distinta antes y después del experimento. Así pues, se plantea este contraste:

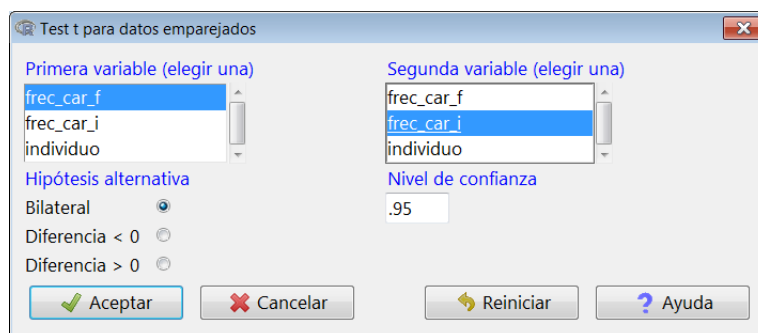
$$H_0 : \mu_{final} - \mu_{inicial} = 0$$

$$H_1 : \mu_{final} - \mu_{inicial} \neq 0$$

Para efectuar el test, accederemos a la ruta siguiente:

Estadísticos / Medias / Test t para datos relacionados

Obtendremos el cuadro de diálogo que se muestra a continuación, donde seleccionamos las dos variables:



El resultado obtenido es el siguiente:

```

Paired t-test

data: cardio$frec_car_f and cardio$frec_car_i
t = 11.3082, df = 19, p-value = 7.015e-10
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 5.78587 8.41413
sample estimates:
mean of the differences
      7.1

```


El resultado del contraste nos indica que la H_0 se rechaza, y que las medias son estadísticamente distintas. Otra manera de verlo consiste en considerar el intervalo de confianza para el parámetro $\mu_{final} - \mu_{inicial}$, que es $[5,79 ; 8,41]$. Puesto que este intervalo no contiene el valor cero, la media de la frecuencia cardíaca de los individuos al final del experimento es estadísticamente distinta de la media registrada al principio. Es más, al ser los valores del IC positivos, podemos afirmar que estadísticamente la media de la frecuencia cardíaca se ha incrementado con el fármaco. Resulta interesante mencionar que la conclusión sería exactamente la misma, pero con signo opuesto, si hubiéramos planteado el contraste de esta manera:

$$H_0 : \mu_{inicial} - \mu_{final} = 0$$

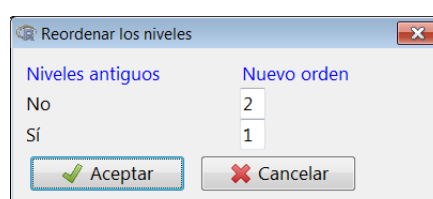
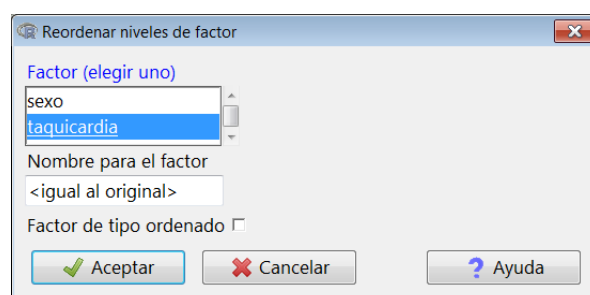
$$H_1 : \mu_{inicial} - \mu_{final} \neq 0$$

2.8. Inferencia sobre la diferencia de proporciones

En esta sección, contrastamos si dos proporciones distintas son estadísticamente iguales o no. Siguiendo el ejemplo anterior, contrastaremos si la proporción de taquicardias de los hombres (π_{th}) es igual a la de las mujeres (π_{tm}) sobre el total de individuos que participaron en el experimento. R-Commander calcula estas proporciones de manera automática y en orden alfabético a partir de los factores de la muestra. Por este motivo, primero tenemos que reordenar los factores, de manera que el orden del factor sexo sea H y M y el del factor taquicardia, $Sí$ y No . Esto lo haremos así:

Datos / Modificar variables del conjunto de datos activo / Reordenar niveles de factor

Aparecerán los dos cuadros de diálogo que se muestran a continuación, en los que se nos preguntará si queremos sobrescribir la variable, a lo que diremos que sí:



En este último cuadro de diálogo hemos invertido el orden de los niveles para que el primero fuera un sí, de manera que la proporción sea sí ha tenido taquicardia sobre el total de individuos. En general, haremos esto para todos los factores cuyo orden alfabético no coincida con nuestro interés.

Ahora, procedemos a hacer el contraste siguiente:

$$H_0 : \pi_{th} - \pi_{tm} = 0$$

$$H_1 : \pi_{th} - \pi_{tm} \neq 0$$

Para efectuar este contraste, seguiremos esta ruta:

Estadísticos / Proporciones / Test de proporciones para dos muestras

Y obtendremos el cuadro de diálogo que se muestra a continuación:

El resultado obtenido es el siguiente:

```

    taquicardia
sexo  Sí    No Total Count
  H 36.4 63.6   100     11
  M 44.4 55.6   100      9

2-sample test for equality of proportions without
continuity correction

data:  .Table
X-squared = 0.1347, df = 1, p-value = 0.7136
alternative hypothesis: two.sided
95 percent confidence interval:
-0.5123192 0.3507031

```

```
sample estimates:
  prop 1      prop 2
0.3636364 0.4444444
```

El resultado del contraste nos indica que la H_0 no se rechaza, y que no podemos afirmar que las proporciones de los hombres y las mujeres con taquicardia sean estadísticamente distintas. Ocurre igual si consideramos el IC para el parámetro $\pi_{th} - \pi_{tm}$, que es $[-0,51 ; 0,35]$. Dado que este intervalo contiene el valor cero, no podemos afirmar que la diferencia de proporciones sea estadísticamente diferente de cero.

2.9. Inferencia sobre el cociente de varianzas

Esta sección incluye la inferencia basada en la distribución F de Snedecor que se lleva a cabo para evaluar si las varianzas de dos muestras son iguales o no. Cabe mencionar que se supone que las dos muestras proceden de dos variables normales x_1 y x_2 que, además, son independientes e idénticamente distribuidas. El objetivo es comprobar si las varianzas de estas variables son iguales.

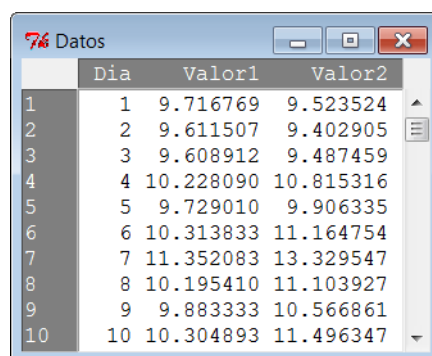
El contraste que hay que plantear es el siguiente:

$$\begin{array}{l} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{array}$$

O bien, si despejamos y tenemos en cuenta que cuando hablamos de igualdad de varianzas nos referimos a que el cociente de las mismas sea la unidad:

$$\begin{array}{l} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{array}$$

Veamos esto mediante un ejemplo. Supongamos que disponemos de datos sobre la cotización de dos valores bursátiles, V_1 y V_2 , cuya cotización diaria se asume de manera independiente una de otra. Los primeros valores de las dos variables son los siguientes.



	Dia	Valor1	Valor2
1	1	9.716769	9.523524
2	2	9.611507	9.402905
3	3	9.608912	9.487459
4	4	10.228090	10.815316
5	5	9.729010	9.906335
6	6	10.313833	11.164754
7	7	11.352083	13.329547
8	8	10.195410	11.103927
9	9	9.883333	10.566861
10	10	10.304893	11.496347

En este ejemplo específico, elaboraremos dos gráficos de cotización y los compararemos entre sí para tener una primera evidencia visual. Basándonos en la sintaxis de R

aprendida en estos materiales, elaboraremos los dos gráficos con un cierto nivel de detalle, incorporando título y bandas de fluctuación que marquen el mínimo y el máximo de las cotizaciones.

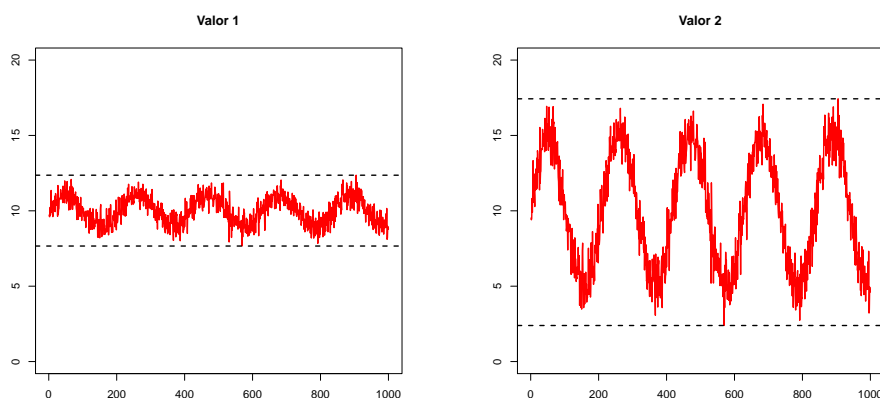
El gráfico para el primer valor se compone a partir de las instrucciones siguientes:

```
> v1 <- Datos$Valor1  
  
> plot(v1,type="l",lwd=2,col="red",main="Valor 1",  
+      xlab="",ylab="",ylim=c(0,20))  
> abline(h=min(v1),lwd=2,lty=2)  
> abline(h=max(v1),lwd=2,lty=2)
```

De manera análoga, para el segundo valor tenemos estas instrucciones:

```
> v2 <- Datos$Valor2  
  
> plot(v2,type="l",lwd=2,col="red",main="Valor 2",  
+      xlab="",ylab="",ylim=c(0,20))  
> abline(h=min(v2),lwd=2,lty=2)  
> abline(h=max(v2),lwd=2,lty=2)
```

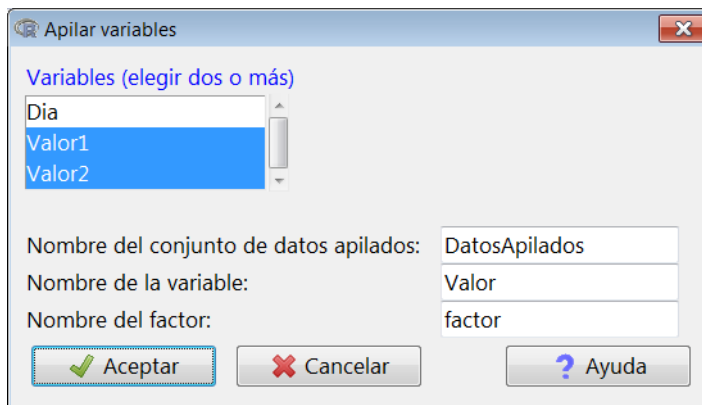
Si visualizamos los dos gráficos de manera conjunta, podemos observar una mayor variabilidad o dispersión en la cotización del *Valor 2*:



Para obtener una certeza estadística sobre la diferencia de varianzas entre las dos variables, tendremos que efectuar un test F de diferencia de varianzas. Previamente, tenemos que apilar las variables en una sola variable. Como se ha visto anteriormente, la ruta que hay que seguir es esta:

Datos / Conjunto de datos activo / Apilar variables del conjunto de datos activo

En el siguiente cuadro de diálogo, apilaremos las variables *Valor1* y *Valor2* en una sola variable, y crearemos además un factor que identifique, para cada observación, de qué valor se trata:



Si visualizamos el nuevo conjunto de datos, obtenemos lo siguiente:

	Valor	factor
1	9.716769	Valor1
2	9.611507	Valor1
3	9.608912	Valor1
4	10.228090	Valor1
5	9.729010	Valor1
6	10.313833	Valor1
7	11.352083	Valor1
8	10.195410	Valor1
9	9.883333	Valor1
10	10.304893	Valor1

El test F se basa en la ratio de varianzas muestrales, de manera que el estadístico toma la forma siguiente.

$$F = \frac{S_1^2}{S_2^2} \sim F_{n-1, m-1}.$$

Donde n y m son las dimensiones de las dos muestras, de manera respectiva. En nuestro ejemplo, ya que sospechamos que la varianza del segundo valor es superior, plantearemos este contraste:

$$\begin{array}{l} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1 \end{array}$$

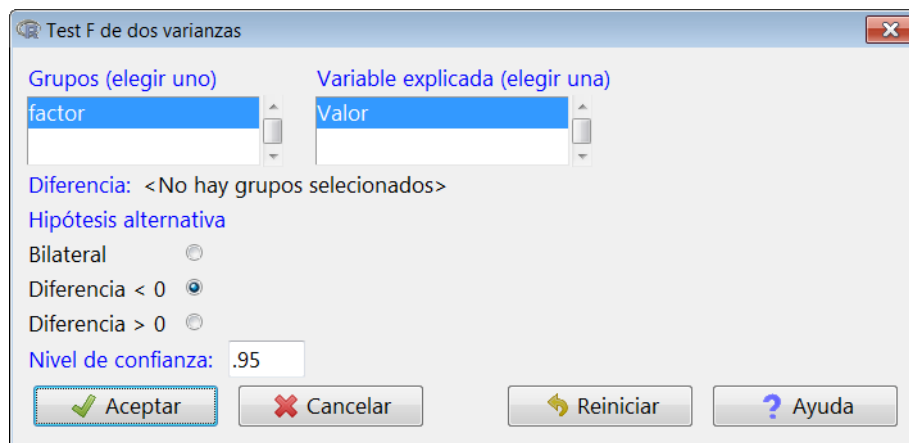
En R-Commander, seguiremos esta ruta para efectuar el contraste, asumiendo un nivel de significación de $\alpha = 0,05$:

Estadísticos / Varianzas / Test F para dos varianzas

La varianza muestral

Recordemos que, para n observaciones de una variable X , la varianza muestral se calcula como $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

En el cuadro de diálogo que aparecerá, especificaremos cuál es la hipótesis alternativa:



El resultado es este:

```
> tapply(DatosApilados$Valor, DatosApilados$factor, var,
+ na.rm=TRUE)
      Valor1      Valor2
0.7700158 13.6700790

> var.test(Valor ~ factor, alternative='less',
+ conf.level=.95, data=DatosApilados)

F test to compare two variances

data:  Valor by factor
F=0.0563, num df=999, denom df=999, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is less
than 1
95 percent confidence interval:
 0.0000000 0.0625104
sample estimates:
ratio of variances
      0.05632856
```

La interpretación es inmediata: rechazamos H_0 , es decir, claramente la varianza del *Valor 2* es superior a la del *Valor 1*. Observamos cómo el valor del estadístico es aproximadamente de $F = 0,05$, lo que nos indica que la varianza del segundo valor es más o menos 20 veces superior a la varianza del primer valor.

Bibliografía

Gibernans Bàguena, J.; Gil Estallo, À. J.; Rovira Escofet, C. (2009). *Estadística*.
Barcelona: Material didáctico UOC.

