Capítulo 3

Variables Aleatorias

3.1. Introducción

R y cualquier otro programa de software estadístico proporcionan una excelente manera de realizar cálculos asociados a las distribuciones de probabilidad más comunes.

Recordamos que según la naturaleza de la variable aleatoria pueden considerarse distribuciones de probabilidad discretas o continuas. Las principales distribuciones de probabilidad de variables discretas son: Bernoulli, Binomial, Geométrica, y de Poisson. Entre los modelos de variable continua destacan las distribuciones: Uniforme, Exponencial y Normal. Todas estas distribuciones y muchas más, están recogidas en R.

Incluso los modelos más sencillos, como el Binomial o el de Poisson, presentan dificultades en cuanto a lo tedioso que resulta realizar los cálculos. En otras distribuciones, como la Normal, el problema es que es imposible trabajar analíticamente, siendo absolutamente necesario la utilización de algún software matemático para obtener aproximaciones precisas de las probabilidades.

Utilizaremos 4 tipos de funciones:

1. Las funciones que R llama densidades, pero que en realidad son densidades (si la distribución es discreta) o funciones masa (si la distribución es continua). Todas estas funciones empiezan por la letra `d`.

2. Las funciones de distribución. Todas ellas empiezan por la letra `p`.

3. Las funciones cuantil, que empiezan por la letra `q`.

4. Las funciones de simulación de datos, que empiezan por la letra `r`.

Lo que sigue a esas letras que son el comienzo de cada función es la identificación de la distribución, por ejemplo

- `binom` para la binomial.
- `geom` para la geométrica.
- `pois` para la Poisson.
- `exp` para la exponencial.
- `norm` para la normal.
- etc.

Finalmente, los argumentos de cada una de las funciones tendrán que especificar, en cada caso, los parámetros concretos de la distribución, y determinar qué probabilidad (`prob`) o qué cuantil queremos, o cuántos datos simulados necesitamos (`m`). La siguiente tabla muestra algunos ejemplos de estas funciones, especificando los argumentos de cada función.

	F. Densidad	F. Distribución	F. Cuantil	Muestras aleatorias
Binom(n, p)	dbinom(x, n, p)	pbinom(x, n, p)	qbinom(prob, n, p)	rbinom(m, n, p)
Geom(p)	dgeom(x, p)	pgeom(x, p)	qgeom(prob, p)	rgeom(m, p)
Pois(λ)	dpois(x, λ)	ppois(x, λ)	qpois(prob, λ)	rpois(m, λ)
U(a, b)	dunif(x, a, b)	punif(x, a, b)	qunif(prob, a, b)	runif(m, a, b)
Beta(α, β)	dbeta(x, α, β)	pbeta(x, α, β)	qbeta(prob, α, β)	rbeta(m, α, β)
Exp(λ)	dexp(x, λ)	pexp(x, λ)	qexp(prob, λ)	rexp(m, λ)
N(μ, σ)	dnorm(x, μ, σ)	pnorm(x, μ, σ)	qnorm(prob, μ, σ)	rnorm(m, μ, σ)

Tabla 3.1: Resumen de las funciones asociadas a las distribuciones y sus argumentos

3.2. Cálculo de probabilidades

3.2.1. Distribuciones discretas

En el caso de las distribuciones discretas, estaremos, básicamente, interesados en el cálculo de dos tipos de probabilidades:

1. Las probabilidades que proporciona la función masa, que podríamos llamar probabilidades simples, del tipo $P(X = x)$.
2. Probabilidades acumuladas (dadas en términos de la función de distribución), del tipo $P(X \leq x)$.

Es evidente que las probabilidades acumuladas se pueden calcular a partir de las probabilidades simples, sin más que tener en cuenta que $P(X \leq x) = \sum_{xi \leq x} P(X = x_i)$.

Por ejemplo, supongamos que tenemos una distribución binomial de parámetros $n = 10$ y $p = 0.25$ y deseamos calcular $P(X < 4)$. Entonces, dado que

$$P(X < 4) = P(X = 0, 1, 2, 3) = \sum_{x=0}^{3} P(X = x),$$

sólo tenemos que calcular la probabilidad de 0, 1, 2 y 3 y sumarlas. Para ello podríamos ejecutar cualquiera de las ordenes:

```
sum(dbinom(0:3,10,0.25))
pbinom(3,10,0.25)
```

3.2.2. Distribuciones continuas

En el caso de las distribuciones de tipo continuo sabemos que los valores concretos de la variable tienen probabilidad cero o, dicho de otra forma, no tienen masa de probabilidad, sino densidad de probabilidad. En estas variables no tiene sentido preguntarse por probabilidades del tipo $P(X = x)$ porque todas son cero. En su lugar, lo que nos preguntamos es por las probabilidades

de que las variables proporcionen valores en intervalos, es decir, probabilidades del tipo $P(a < X < b)$, y donde las desigualdades pueden ser estrictas o no, ya que el resultado final no varía.

Recordamos además que las probabilidades del tipo $P(a < X < b)$ se calculan como

$$P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a),$$

donde $f(x)$ es la función de densidad de la variable y $F(x)$ es la función de distribución. En resumen, podremos calcular probabilidades del tipo $P(a < X < b)$ siempre que podamos obtener los valores de la función de distribución $F(x)$. Y recordemos que estas funciones de distribución en R son las que empiezan por la letra `p`.

Así, si tuviéramos una distribución normal de media 5 y desviación típica 2, y quisiéramos calcular $P(2 < X < 7.6)$ lo haríamos teniendo en cuenta que

$$P(2 < X < 7.6) = F(7.6) - F(2)$$

mediante

```
pnorm(7.6,5,2)-pnorm(2,5,2)
```

3.3. Cálculo de cuantiles

Recordemos que los cuantiles son medidas de posición relativas. El cuantil p, (siendo p un n^o entre 0 y 1), Q_p, de una distribución de probabilidad es aquél que deja por debajo de sí una probabilidad p. Si, hipotéticamente, tuviéramos 100 datos, el cuantil p es el que, ordenados todos los valores de menor a mayor, ocuparía la posición $100p$. Eso es lo que permite interpretarlos como medidas de posición relativas, ya que permite analizar si un dato es *alto, medio o bajo* en su distribución.

Desde el punto de vista del cálculo, observemos que en la sección anterior hemos aprendido a calcular los valores $p = P(X \leq x)$. Lo que ahora tenemos que hacer es justo lo contrario, es calcular los valores x tales que $P(X \leq x) = p$.

3.3.1. Distribuciones discretas

A la hora de hablar de cuantiles en distribuciones discretas hay que recordar que es posible que no podamos encontrar algunos cuantiles concretos, precisamente por el carácter discreto de la variable. Es por eso que en el caso de distribuciones discretas la definición de cuantil debe afinar un poco más. Hablamos del cuantil p como del mayor valor x tal que $P(X \leq x) \geq p$. Por ejemplo:

- Si $X \sim B(15, 0.65)$, $Q_{0.05} = 7$, lo que en R se consigue mediante

  ```
  qbinom(0.05,15,0.65)
  ```

- Si $X \sim Poisson(5.8)$, $Q_{50} = 6$, dado en R por

  ```
  qpois(0.50,5.8)
  ```

3.3.2. Distribuciones continuas

En el caso de las distribuciones continuas, dado $p \in (0, 1)$ siempre existe un valor x tal que $P(X \leq x)$ es exactamente igual a p, y ese valor es el percentil $100p$ o el cuantil p. A modo de ejemplos:

- Si $X \sim Exp(1/2)$, $Q_{0.05} = 0.1026$, dado por

  ```
  qexp(0.05,1/2)
  ```

- Si $X \sim N(0, 4.5)$, $Q_{0.95} = 7.402$, dado en R por

  ```
  qnorm(0.95,0,4.5)
  ```

3.4. Simulación de muestras

Una de las aplicaciones más comunes de la Estadística es la de proporcionar un marco teórico para poder realizar simulaciones de procesos más o menos complejos. En esas simulaciones existe la necesidad de contar con datos inventados, pero inventados según un modelo proporcionado por una distribución de probabilidad que se suponga adecuado para el fenómeno que estamos simulando. Es por ello que la mayoría de los paquetes de software estadístico, entre ellos R, facilitan la posibilidad de obtener muestras aleatorias simples de las distribuciones más usuales.

Supongamos que estamos interesados en simular datos de una distribución normal. Utilizamos entonces, la función `rnorm()`, cuya sintaxis es muy sencilla: tan sólo tenemos que especificarle el número de valores que queremos simular y los parámetros de la distribución. Por ejemplo, para generar 250 valores de una $N(\mu = 100, \sigma = 10)$, usamos

```
rnorm(250,100,10)
```

Vamos a comprobar que, en efecto, esta muestra procede de una $N(100, 10)$. Para ello podemos comparar el histograma de la muestra con la función de densidad de esa $N(100, 10)$. El resultado aparece en la Figura 3.1. Hay que decir que el parecido es notable, aunque sería aún más notable si hubiéramos elegido aún más datos (más de 250).

El código es el siguiente:

```
muestra<-sort(rnorm(250,100,10))
```

genera la muestra de tamaño 250 de la $N(100, 10)$ y la ordena mediante la función `sort()`.

```
hist(muestra,freq=FALSE)
```

grafica el histograma de los datos de la muestra en escala de densidad.

```
lines(muestra,dnorm(muestra,100,10))
```

añade al histograma la gráfica de la función de densidad evaluada en los valores de la muestra.

Observamos que la función `lines()` se trata de una función muy similar a `plot()` en su estructura y utilidad, cuya principal diferencia radica en que `plot()` es lo que en R se conoce

como una función gráfica de primer nivel y `lines()` lo es de segundo nivel. Que una función gráfica sea de primer nivel quiere decir que por sí misma abrirá una ventana gráfica y mostrará su resultado. Que una función sea de segundo nivel implica que por sí misma no tiene autoridad para abrir una ventana de gráficos, tan sólo puede añadirse a una ventana ya abierta.

`hist()` es una función gráfica de primer nivel, por lo que realiza el histograma y lo muestra en una ventana de gráfico. A esa ventana se añade la función de densidad graficada mediante `lines()`.

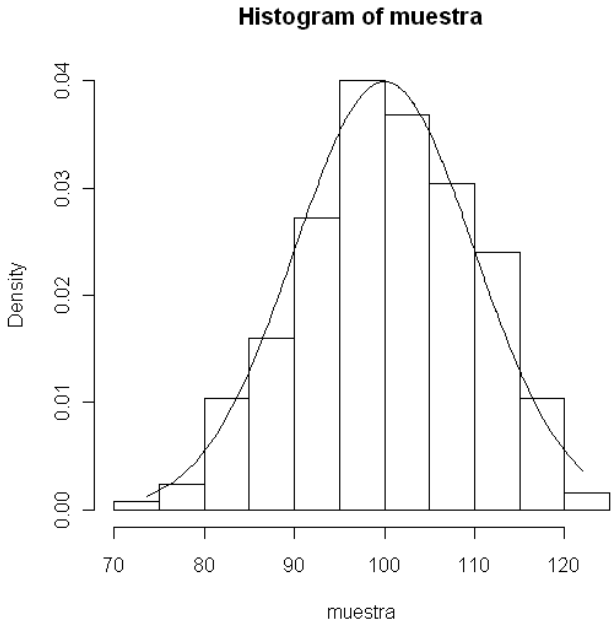

Figura 3.1: Histograma de una muestra junto con la función de densidad del modelo del que procede

.

Capítulo 4

Inferencia Estadística

En este capítulo se describe una manera de realizar estimaciones por intervalos de confianza y contrastes de hipótesis de algunos parámetros a través del código de R.

4.1. Estimación por Intervalos de confianza

Los intervalos de confianza guardan una relación biunívoca con los contrastes de hipótesis estadísticas. Ese es el motivo por el que R no tiene un comando especifico para la construcción de intervalos de confianza, sino que éstos son proporcionados como parte de los resultados vinculados a los contrastes de hipótesis.

Sin embargo, es casi trivial la posibilidad de usar R como una simple calculadora para aplicar las fórmulas de los intervalos de confianza conocidos y eso es lo que vamos a hacer aquí.

En esta sección vamos a plasmar con ejemplos cómo podemos obtener algunos intervalos de confianza bilaterales. Lo vamos a hacer exclusivamente con código y sin la ayuda de ningún paquete adicional, para lo cual recordamos la sintaxis de algunas funciones:

- `mean(datos)` devuelve la media muestral de datos.
- `sd(datos)` devuelve la cuasi–desviación típica muestral de datos.
- `sqrt(x)` devuelve $\sqrt{x}$.
- `qnorm(a)` devuelve z_a.

- `qt(a, v)` devuelve $t_{a,v}$.
- `qchisq(a, v)` devuelve $\chi^2_{a,v}$.

4.1.1. De la media de una distribución normal con varianza desconocida

Recordemos que si denotamos $x_1, \ldots, x_n$ a una muestra de una distribución $N(\mu, \sigma)$, ambas desconocidas, un intervalo de confianza $(1-\alpha)$ para μ viene dado por

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

donde $\bar{x}$ es la media muestral, s es la cuasi-varianza muestral, n es el número de observaciones y $t_{n-1;\frac{\alpha}{2}}$ es el percentil $\frac{\alpha}{2}$ de la distribución t-Student con $n-1$ grados de libertad.

Ejemplo: Los siguientes datos corresponden al nivel de glucosa en sangre, en ayunas, obtenidos de muestras de sangre de 30 pacientes diabéticos. Si asumimos que los datos siguen una distribución aproximadamente normal, ¿cómo podemos obtener un intervalo de confianza para la concentración media de glucosa en sangre, en ayunas para personas diabéticas, con $\alpha = 0.05$?

Tabla 4.1: Concentraciones de glucosa en sangre

97.68	105.82	104.58	121.84	107.77	95.79	111.19	98.47
106.05	101.05	96.83	104.17	94.20	94.42	104.17	97.13
103.55	101.05	98.40	107.68	97.13	102.79	100.56	98.95
103.57	105.14	96.19	102.55	106.45	113.16	103.47	107.28

Podemos cargar los datos, usando el comando `scan`

```
muestra <- scan()
 97.68 105.82 104.58 121.84 107.77  95.79 111.19  98.47
106.05 101.05  96.83 104.17  94.20  94.42 104.17  97.13
103.55 101.05  98.40 107.68  97.13 102.79 100.56  98.95
103.57 105.14  96.19 102.55 106.45 113.16 103.47 107.28
```

- Calculamos la cota inferior del intervalo, llamándola `ci`.

  ```
  ci<-mean(muestra)-qt(0.975,31)*sd(muestra)/sqrt(32)
  ```

- Calculamos la cota superior del intervalo, llamándola `cs`.

  ```
  cs<-mean(muestra)-qt(0.975,31)*sd(muestra)/sqrt(32)
  ```

- Unimos en un vector la cota inferior y la cota superior, haciéndolas aparecer en la ventana de resultados.

  ```
  c(ci,cs)
  ```

4.1.2. De la media de una distribución cualquiera, con muestras grandes

Ejemplo: Consideramos 300 datos correspondientes al tiempo hasta el fallo (en años) de unas determinadas componentes electrónicas usadas en los marcapasos. Los datos se encuentran en el fichero tiempos.fallo.rda*, en una hoja llamada* datos.tiempos.fallo

Al cargarlos, podemos ver que la variable se llama *años* y realizando un histograma, podemos ver la forma que tienen los datos.

```
load("tiempos.fallo.rda")
names(datos.tiempos.fallo)
hist(datos.tiempos.fallo$años)
```

Sabemos que, visto el histograma, no es admisible pensar que la variable sigue una distribución normal, pero tenemos 300 datos, suficientes para poder aplicar el resultado basado en el teorema central del límite que determina que un intervalo para μ a un nivel de confianza α es

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

siendo n el tamaño de la muestra. El código es el siguiente:

```
ci<-mean(datos.tiempos.fallo$años)-qnorm(0.975)*sd(datos.tiempos.fallo$años)/sqrt(300)
cs<-mean(datos.tiempos.fallo$años)+qnorm(0.975)*sd(datos.tiempos.fallo$años)/sqrt(300)
c(ci,cs)
```

4.1.3. De una proporción

Ejemplo: Supongamos que un estudio para detectar alergias en menores, se comprueba que en una muestra de 300 niños y niñas, 21 sufren algún tipo de alergia alimenticia. Obtener un intervalo de confianza al 95 % para el porcentaje de menores alérgicos.

El intervalo, para un nivel α viene dado por

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

donde $\hat{p}$ es la proporción muestral (en nuestro caso, $\frac{21}{300}$) y n es el tamaño de la muestra (en nuestro caso, 300).

El código para obtener el intervalo es el siguiente:

```
n=300
k=21
ci<-k/n-qnorm(0.975)*sqrt((k/n)*(1-k/n)/n)
cs<-k/n+qnorm(0.975)*sqrt((k/n)*(1-k/n)/n)
c(ci,cs)
```

4.1.4. De la varianza de una distribución normal

Ejemplo: Finalmente, vamos a obtener un intervalo de confianza para la varianza de la variable concentración de glucosa en sangre, en ayunas, en personas diabéticas.

Recordamos que dicho intervalo viene dado por

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}} \right).$$

Usamos el código:

```
n=32
ci<-(n-1)*var(muestra)/qchisq(0.975,n-1)
cs<-(n-1)*var(muestra)/qchisq(0.025,n-1)
c(ci,cs)
```

4.2. Contrastes de hipótesis

A lo largo de este tema vamos a abordar la realización de contrastes de hipótesis paramétricas a través de diversos ejemplos.

4.2.1. Contrastes sobre medias: La función t.test()

En esta sección veremos cómo resolver problemas que involucran a la media de una población comparándola con un valor hipotético o a la media de dos poblaciones (independientes o apareadas), comparándolas entre sí. Lo que tienen en común estas pruebas es que todas ellas se basan en un estadístico de contraste que sigue una distribución t de *Student*. En R, el código necesario para llevar a cabo estos contrastes se basa en la misma función, la función *t.test*(). Su sintaxis básica es la siguiente:

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0,
paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

donde

- `x` es un vector de datos correspondiente a una de las muestras o a la única muestra del problema. Si estamos haciendo un test sobre la media de una población, `x` contiene la única muestra. Si estamos realizando un test de comparación de medias, `x` será la primera de las dos muestras.
- `y` corresponde a la segunda muestra en un test de comparación de medias. Si no es el caso y estamos realizando un test sobre una sola población, simplemente no se incluye.
- `alternative` especifica la dirección de la hipótesis alternativa. Como puede verse, tiene 3 posibles valores, "two.sided"(bilateral), "less"(unilateral a la izquierda) y 'greater"(unilateral a la derecha).
- `mu` es el valor hipotético con el que se compara la media o la diferencia de medias en el contraste.
- `paired` especifica si las dos muestras x e y, en caso de que aparezcan, son apareadas o no.

- En el caso en el que aparecen dos muestras, `var.equal` especifica si podemos suponer varianzas iguales o no.
- `conf.level` es el nivel de confianza de los intervalos que se mostrarán asociados al test.

4.2.1.1. Para la media de una población

Ejemplo: Se sospecha que el nivel medio de glucosa en sangre, en ayunas, de las personas diabéticas es de 108 mg/100ml. Para contrastar esta suposición se utilizan análisis de sangre realizados a 45 personas diabéticas en ayunas. Los datos se encuentran en el fichero diabeticos.txt. *¿Es correcto asumir que la media del nivel de glucosa en ayunas es de 108 mg/100ml? (Utilícese un nivel de significación del 5%).*

Fijémonos que nos piden claramente que confirmemos una afirmación: que la media es $108mg/100ml$. Por lo tanto, si denotamos μ al nivel medio de glucosa en sangre, podemos plantear el contraste $H_0 : \mu = 108$ frente a $H1 : \mu \neq 108$.

Dado que el tamaño muestral es generoso (45, superior a 30), no necesitamos la hipótesis de normalidad de los datos. Dicho esto, vemos que se trata de un contraste sobre la media de una distribución normal, contraste bilateral.

Abrimos el fichero de datos `diabeticos.txt`, por ejemplo con el bloc de notas, y observamos que los decimales están separados por una coma y que la primera línea de , lo que debemos indicar al utilizar el comando `read.table`. Y una vez que hemos importado los datos debemos especificar, cómo definimos el test.

```
datos.diabeticos<-read.table("diabeticos.txt",header=TRUE,dec=",")
t.test(datos.diabeticos$glucosa,alternative="two.sided",mu=108)
```

El resultado es el siguiente:

```
One Sample t-test
data:  diabeticos$glucosa
t = -1.6771, df = 44, p-value = 0.1006
alternative hypothesis: true mean is not equal to 108
95 percent confidence interval:
 103.5965 108.4035
```

```
sample estimates:
mean of x
      106
```

Analicemos el resultado con detalle:

- En primer lugar, nos recuerda que estamos analizando la variable `diabeticos$glucosa`.
- A continuación nos informa del valor del estadístico de contraste (`t = -1.6771`), de los grados de libertad (`df = 44`) y del *p-valor* (`p-value = 0.1006`). Ya podemos, por tanto, concluir: Dado que el *p-valor* no es inferior a 0.05, no tenemos suficientes evidencias en los datos para rechazar la hipótesis nula ($\mu = 108$) en favor de la alternativa ($\mu \neq 108$). Es decir, con los datos de la muestra no tenemos suficientes evidencias de que el nivel medio de concentración de glucosa en sangre, para diabéticos sea distinto de 108.
- Nos recuerda cuál era la hipótesis nula que habíamos planteado:

  `alternative hypothesis: true mean is not equal to 108`.
- A continuación proporciona un intervalo de confianza bilateral, con un nivel de confianza del 95 %, para la media de la distribución normal que se le supone a los datos:

  ```
  95 percent confidence interval:
  103.5965 108.4035
  ```

  Es importante recordar la relación que guarda el intervalo de confianza con el contraste de hipótesis. Observamos que el valor hipotético que hemos considerado para la media, 108, está dentro de este intervalo, luego éste es un valor de confianza para μ. Es otra forma de concluir que no hay datos que avalen que la media de la variable es significativamente distinta de 108, ya que éste es un valor bastante plausible para esta media. Si los datos fueran tales que el intervalo de confianza para μ dejara fuera al valor 108, tendríamos razones para pensar que el valor de μ es significativamente distinto de 108, pero no es el caso.
- Finalmente, proporciona los estadísticos muestrales utilizados, en este caso, la media muestral:

  ```
  sample estimates:
  mean of x
  106
  ```

Por lo tanto, y a modo de conclusión, podemos decir que no hay evidencias de que la concentración de glucosa en sangre, en el caso de las personas diabéticas, no sea 108.

4.2.1.2. Para la diferencia de medias de poblaciones independientes

Ejemplo: Un ingeniero industrial ha sintetizado en el laboratorio una feromona con la que pretende luchar contra una plaga de insectos. La feromona se aplica en trampas donde caen los insectos masivamente. Hasta ahora se trabajaba introduciendo otro producto que se supone que atraía al insecto, por lo que el ingeniero desearía demostrar que su feromona sintetizada es más efectiva que dicho producto. Para probar si esto ocurre, prepara 100 trampas con el producto tradicional y 100 con su feromona y las distribuye, contabilizando el número de insectos atrapados en cada una de las 200 trampas. Con esos datos, ¿puede concluir el ingeniero que su feromona es más efectiva que el producto tradicional? Los datos se encuentran en el fichero feronoma.txt.

Vamos a llamar μ_V a la media de las capturas con el viejo producto y μ_N a la media de las capturas con la nueva feromona. Lo que nos piden en el enunciado es que contrastemos la hipótesis nula $H_0 : \mu_V \geq \mu_N$ frente a la alternativa $H_1 : \mu_V < \mu_N$:

- En primer lugar, podemos suponer que las muestras son independientes. Nada hace pensar que los datos de la muestra bajo el producto antiguo hayan tenido nada que ver en la muestra bajo el producto nuevo ni al contrario.
- Con respecto al tamaño muestral, debe preocuparnos la hipótesis de normalidad: recordemos que si el tamaño de la muestras es pequeño, éstas deberían proceder de una distribución normal, pero no es el caso: ambas muestras tienen tamaños superiores a 30.
- Finalmente, deberemos plantearnos si podemos suponer o no que las varianzas son iguales.

Lo primero que tenemos que hacer para importar los datos (que se encuentran en un fichero de tipo texto) es ver cómo están almacenados. Si lo abrimos, por ejemplo con el bloc de notas, vemos que están separados por tabulaciones y que los nombres de las variables están en la primera fila. La Figura 4.1 (a) muestra el bloc de notas que contiene a los datos.

Es importante observar que los datos de las dos muestras aparecen en dos columnas paralelas. Esta no es una forma correcta de especificarlas, ya que parece que cada dato de la primera

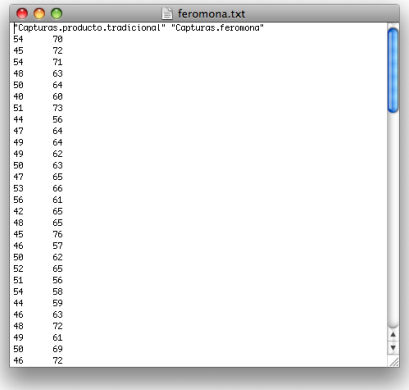

Figura 4.1: Importando los datos del fichero *feromona.txt*

muestra está relacionado con otro dato de la segunda muestra y, en realidad, las muestras son independientes (de hecho podrían tener distinto tamaño muestral). Por este motivo, tenemos que especificar este hecho, para que R entienda que se trata de dos muestras independientes.

Podemos usar el código:

```
feromona<-read.table("feromona.txt",header=TRUE)
t.test(x=feromona$capturas.feromona, y=feromona$capturas.producto.tradicional,
alternative="greater",mu=0)
```

Cuyo resultado es el siguiente:

```
Welch Two Sample t-test
data: capturas by producto
t = 20.4367, df = 197.952, p-value < 2.2e-16
alternative hypothesis: true difference in means
is greater than 0
95 percent confidence interval:
14.25580 Inf
sample estimates:
mean in group capturas.feromona
64.94
mean in group capturas.producto.tradicional
49.43
```

Vamos a analizarlo punto por punto:

- Especifica que se trata de un test t para la variable *capturas* separada por el factor *producto*.
- Proporciona el valor del estadístico de contraste (`t = 20.4367`), los grados de libertad (`df = 197.952`), y el $p-valor$ (`p-value < 2.2e-16`). Dado que el $p-valor$ es inferior a 0.05, ya podemos concluir que tenemos evidencias en los datos para afirmar con un 95 % de confianza que la media de las capturas con la feromona es superior a la de las capturas con el viejo producto.
- Especifica cuál es nuestra hipótesis alternativa.
- Proporciona un intervalo de confianza para la diferencia de las medias. En este caso, la probabilidad de que el intervalo (`14.25580 Inf`) contenga a la diferencia de las medias es del 95 %. El cero no es, por tanto, un valor bastante plausible y por ello hemos rechazado la hipótesis nula en favor de la alternativa.
- Proporciona las dos medias muestrales.

En resumen, hemos concluido que podemos afirmar con un 95 % de confianza que la feromona es más efectiva que el viejo producto al aumentar significativamente el promedio de capturas.

4.2.1.3. Para la diferencia de medias apareadas

Ejemplo: En un programa de Control de Enfermedades Crónicas, la hipertensión está incluida como la primera patología a controlar. 15 pacientes hipertensos son sometidos al programa y su tensión asistólica es controlada antes y después de 6 meses de tratamiento. Los datos son los siguientes:

inicio	180	200	160	170	180	190	190	180	190	160	170	190	200	210	150
fin	140	170	160	140	170	150	140	150	190	170	120	160	170	160	150

Si los datos siguen una distribución normal, ¿podemos afirmar que es efectivo el tratamiento?

Vamos a denotar por μ_D al promedio de las medidas de tensión asistólica después del tratamiento y por μ_A al mismo promedio antes del tratamiento. Nos piden contrastar $H_0 : \mu_D \geq \mu_A$ frente a $H1 : \mu_D < \mu_A$.

En este caso, el código sería el siguiente:

```
tension<-read.table("tension.txt",header=TRUE,dec=",")
t.test(x=tension$inicio, y=tension$fin, alternative="greater",mu=0,paired=TRUE)
```

```
Paired t-test
data:  tension$inicio and tension$fin
t = 4.8316, df = 14, p-value = 0.0001332
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 16.09828      Inf
sample estimates:
mean of the differences
              25.33333
```

Vamos a analizar los resultados con detalle:

- En las dos primeras líneas se nos informa que estamos realizando un test t para muestras apareadas sobre los datos relativos a las variables *fin* e *inicio* del conjunto de datos *tension*.
- En la siguiente línea aparece el valor del estadístico de contraste (`t = 4.8316`), de los grados de libertad (`df = 14`) y el $p-valor$ (`p-value = 0.0001332`). Visto el valor de éste podemos concluir que existen evidencias en los datos de la muestra de que el tratamiento disminuye el promedio de la tensión asistólica. Observamos que el $p-valor$ es muy bajo, luego las diferencias detectadas son muy significativas.
- A continuación aparece el tipo de hipótesis alternativa que hemos elegido.
- Posteriormente aparece un intervalo de confianza al 95 % para la diferencia de las medias. El hecho de que el cero no esté contenido en dicho intervalo es otra forma de ver que podemos rechazar la hipótesis nula en favor de la alternativa.
- Finalmente aparece el valor muestral de la diferencia de las medias.

En resumen, los datos muestran indicios de que el tratamiento disminuye la tensión asistólica en promedio.

4.2.2. Contrastes sobre proporciones: La función prop.test()

R realiza este tipo de contrastes de dos formas: mediante una prueba tipo χ^2 o mediante una prueba binomial exacta.

En el primero de los casos, el de las pruebas tipo χ^2, lo que se hace es comparar las frecuencias de casos favorables en la muestra de los datos (frecuencias observadas, O_i) con la muestra de casos favorables que habría en una muestra con el mismo número de datos si la hipótesis nula fuera cierta (frecuencias esperadas, E_i). El estadístico utilizado para este contraste es

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

que, bajo el supuesto de que ninguna frecuencia esperada E_i es inferior a 5, se distribuye según una distribución χ^2. En el caso de que alguna frecuencia esperada sea inferior a 5 se suele utilizar la corrección por continuidad de Yates, que da lugar al estadístico del contraste

$$\chi^2 = \sum_i \frac{(|O_i - E_i| - 0.5)^2}{E_i}.$$

La prueba binomial exacta parte del hecho de que, si la hipótesis nula fuera cierta, la distribución del número de casos favorables en la muestra sería binomial. Valorando el número observado de casos favorables dentro de la distribución binomial que se daría bajo la hipótesis nula, se obtiene el p-valor de la prueba.

Vamos a partir del hecho de que conocemos el número de éxitos y fracasos en la muestra. Si no es así y únicamente tenemos los datos en una hoja de datos, podemos rápidamente tabularla mediante la función `table()` a la que sólo hay que especificarle la hoja de datos a tabular y, si ésta tuviera más de una variable, cuál de ellas queremos tabular. La sintaxis de la función `prop.test` es la siguiente. Dicha sintaxis también nos servirá para los contrastes de comparación de dos proporciones:

```
prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95, correct = TRUE)
```

Comentamos cada uno de los argumentos de la función:

- `x` puede especificar dos cosas. O bien el número de éxitos, o bien, mediante una matriz

de dos columnas, el número de éxitos y de fracasos en cada muestra.

- `n` especifica el número de datos de la muestra en el caso en que x sea el número de éxitos, y es ignorado en el caso en que `x` proporcione también el número de fracasos.
- `p` es el vector de probabilidades de éxito bajo la hipótesis nula. Debe ser un vector de la misma dimensión que el número de elementos especificado en `x`.
- `alternative` especifica la dirección de la hipótesis alternativa, tomando los valores `"two.sided"`, `"greater"` o `"less"`.
- `conf.level` es el nivel de confianza de los intervalos que se muestran entre los resultados.
- `correct` especifica si se usa la corrección por continuidad de Yates. Observamos que la opción por defecto es que sí se use esta corrección.

4.2.2.1. Para la proporción en una población

Ejemplo: En un determinado servicio de odontología se espera que aproximadamente el 75% de las visitas no requieran una extracción dentaria inmediata. En cierto año, de 1225 visitas, 926 no necesitaron una extracción inmediata. ¿Se puede decir que el porcentaje de ese año fue significativamente superior al porcentaje esperado?

Si denotamos por p la proporción de visitas que no requieren extracción, se nos está pidiendo que contrastemos $H_0 : p = 0.75$ frente a $H_1 : p > 0.75$.

En este caso, podemos resolver el contraste usando:

```
prop.test(x=926,n=1225,p=0.75,alternative="greater",correct=FALSE)
```

cuyo resultado es:

```
1-sample proportions test without continuity correction
data:  926 out of 1225, null probability 0.75
X-squared = 0.2288, df = 1, p-value = 0.3162
alternative hypothesis: true p is greater than 0.75
95 percent confidence interval:
0.7351821 1.0000000
```

```
sample estimates:
p
0.7559184
```

Analizamos estos resultados:

- Especifica en primer lugar que se trata de un test para la proporción en una muestra.
- Especifica a continuación el valor hipotético en la hipótesis nula.
- Proporciona el valor del estadístico de contraste y, lo que más nos interesa, el $p-valor$. En este caso el $p-valor$ es bastante superior a 0.05, luego a la luz de estos datos no podemos rechazar la hipótesis nula en favor de la alternativa, es decir, no podemos concluir que el porcentaje de visitas que no requieren extracción esté por encima del 75 %.
- A continuación recuerda la hipótesis alternativa.
- Facilita un intervalo de confianza (en este caso unilateral a la derecha) para la proporción.
- Finalmente, muestra la proporción muestral.

Comentamos también que la prueba binomial exacta puede realizarse mediante la función `binom.test`, cuya sintaxis básica es similar a la de `prop.test`:

```
binom.test(x,n,p=0.5,alternative=c("two.sided","less","greater"), conf.level=0.95)
```

4.2.2.2. Para la diferencia de proporciones

Este tipo de contrastes se realiza también mediante la función `prop.test()`, pero previamente debemos comentar algo acerca de cómo introducir los datos. Los éxitos y los fracasos de cada muestra deben ir en dos filas de una matriz de dos columnas, es decir, con una estructura como la siguiente:

N^o de éxitos de la muestra 1 (n11)	N^o de fracasos de la muestra 1 (n12)
N^o de éxitos de la muestra 2 (n21)	N^o de fracasos de la muestra 2 (n22)

Para crear una matriz así se utiliza la función `matrix`, a la que tenemos que especificarle mediante un vector los elementos de la matriz, las dimensiones de la matriz y el sentido en el que vienen especificados los elementos. En el ejemplo nuestro caso sería

```
matrix(c(n11, n12, n21, n22),2,2,byrow=TRUE)
```

o bien

```
matrix(c(n11, n21, n12, n22),2,2,byrow=FALSE).
```

Las pruebas que R utiliza para este tipo de contrastes de nuevo se basan en el uso del estadístico χ^2, comparando las frecuencias observadas en ambas muestras con las que aparecerían bajo la hipótesis nula, o también en la conocida como prueba exacta de Fisher.

Vamos a trabajar sobre el siguiente enunciado:

Ejemplo: A raíz de la alarma creada entre la opinión pública por la repercusión que tuvo el caso de un bloque de edificios con un transformador en su planta baja y donde un gran porcentaje de vecinos sufrió cáncer, se decide realizar un estudio para tratar de encontrar relación entre la cercanía de un transformador eléctrico y la incidencia del cáncer.

Para ello, se eligió una muestra aleatoria de edificios con transformadores en su planta baja durante un periodo de más de 10 años, contabilizando el número de habitantes en ellos, 2150, y todos los casos de cáncer detectados en los 5 últimos años, 37. Por su parte, se recolectó otra muestra aleatoria de control con edificios que no tuvieran ningún transformador eléctrico cercano, contabilizando también el número de personas, 2200, y el número de casos de cáncer en ellos en los últimos 5 años, 33. En ambas muestras, los expertos procuraron eliminar la posibilidad de ruidos, es decir, la presencia de otros factores que pudieran incidir en variar la incidencia del cáncer en alguna de las muestras.

A la luz de los datos de este estudio, ¿podemos afirmar que la cercanía de un transformador eléctrico aumenta la proporción de casos de cáncer? (Utilícese un nivel de significación del 5%)

Si llamamos p_{CT} a la proporción de casos de cáncer en los edificios con transformador cercano y p_{ST} a la proporción análoga en los edificios sin transformador, nos piden que contrastemos $H_0 : p_{CT} = p_{ST}$ frente a $H_1 : p_{CT} > p_{ST}$.

En este caso usaríamos:

```
tabla<-matrix(c(37,2113,33,2167),2,2,byrow=TRUE)
```

Finalmente, la aplicación de la función `prop.test` sería la siguiente:

```
prop.test(tabla, alternative='greater', correct=FALSE)
```

Los resultados son los siguientes:

```
2-sample test for equality of proportions without continuity
correction

data:  tabla
X-squared = 0.3352, df = 1, p-value = 0.2813
alternative hypothesis: greater
95 percent confidence interval:
-0.004071904 1.000000000
sample estimates:
prop 1 prop 2
0.01720930 0.01500000
```

Lo que realmente nos interesa es el $p-$valor, que aparece en la tercera línea, 0.2813, y que indica que no se puede rechazar la hipótesis nula en favor de la alternativa, es decir, no podemos, con los datos existentes, asegurar con un 95 % de confianza que la proporción de casos de cáncer sea superior en los edificios que tengan un transformador eléctrico en su planta baja.

Comentamos, por último, que la prueba exacta de Fisher se realiza con la función `fisher.test`, que tiene una sintaxis parecida a la de `prop.test`, aunque admite muchos más argumentos en función de las características de la tabla de contingencia a analizar.

4.2.3. Contraste para la comparación de varianzas: La función var.test()

Ejemplo: Un grupo de personas participa en un estudio nutricional que trata de analizar los niveles asimilados de vitamina C en sangre de fumadores y no fumadores. Los resultados en mg/l se encuentran en el fichero vitaminac.txt. Si asumimos que los datos son normales, ¿se puede concluir que el nivel de vitamina C asimilado es superior en los no fumadores?

Con el fin de contestar a la pregunta, planteamos las hipótesis $H_0 : \mu_1 \geq \mu_2$ versus $H_1 : \mu_1 < \mu_2$ donde estamos asumiendo que la primera muestra corresponde a los fumadores y la segunda a los no fumadores.

Observamos que el enunciado no menciona nada relativo a la igualdad de varianzas en ambos grupos, por lo que vamos a plantear en primer lugar un contraste de igualdad de varianzas (o desviaciones típicas). Si denotamos σ_F a la desviación típica del nivel de vitamina C asimilado en fumadores y σ_{NF} a la desviación típica del nivel de vitamina C asimilado en no fumadores, se trata de contrastar $H_0 : \sigma_F = \sigma_{NF}$ frente a $H_1 : \sigma_F \neq \sigma_{NF}$.

La función *var.test*() puede utilizarse con una sintaxis muy parecida a las anteriores:

```
var.test(x, y, ratio = 1, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95)
```

1. `x` corresponde al vector de datos de la primera muestra.
2. `y` es el vector de datos de la segunda muestra.
3. `ratio` es el cociente hipotético con el que se compara. Habitualmente deseamos contrastar que las varianzas son distintas, o lo que es lo mismo, que su cociente es 1, así que la opción por defecto es precisamente `ratio=1`.
4. `alternative` especifica la hipótesis alternativa: `"two.sided"` para $H1 : \sigma_1 \neq \sigma_2$, `"less"` para $H1 : \sigma_1 < \sigma_2$ y `"greater"` para $H1 : \sigma_1 > \sigma_2$.
5. `conf.level` es el nivel de confianza del intervalo para el cociente de varianzas que se muestra en las salidas.

Para la resolución del ejemplo anterior, tendríamos el siguiente código:

```
datos.vitaminac<-read.table("vitaminac.txt",header=TRUE,sep=" ",dec=".")
datos.vitaminac

var.test(datos.vitaminac$nivel[datos.vitaminac$tabaco=="fumadores"],
datos.vitaminac$nivel[datos.vitaminac$tabaco=="nofumadores"], alternative="two.sided")
```

O alternativamente, podríamos haber utilizado la función `var.test()` de la forma

```
var.test(nivel ~ tabaco, alternative='two.sided', conf.level=.95,
data=datos.vitaminac)
```

Los resultados son los siguientes:

```
F test to compare two variances
data:  nivel by tabaco
F = 0.3131, num df = 11, denom df = 10, p-value = 0.06976
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.08544081 1.10400497
sample estimates:
ratio of variances
         0.3131332
```

En la tercera línea podemos ver que aparece el valor del estadístico F , los grados de libertad en el numerador y el denominador y el $p-$valor. Ese `p-value=0.06976` indica que no hay suficientes evidencias en los datos para rechazar la igualdad de varianzas. Y por lo tanto, a la vista de los resultados, realizaríamos el contraste de medias planteado asumiendo igualdad de varianzas.