Universidad de Cantabria
Curso 2011–2012

Práctica 4: Inferencia y contraste de hipótesis

En esta práctica vamos a utilizar las opciones de R para aplicar el análisis inferencial ya sea para calcular intervalos de confianza o contraste de hipótesis. Con este análisis, como ya hemos visto en clase, vamos a estudiar una muestra pequeña y representativa de la población, para extraer conclusiones que afectan a todos los miembros de dicha población. En particular nos vamos a centrar en obtener información de las principales características de la población como son la proporción, la media y la desviación típica.

En cuanto a la forma de calcular intervalos de confianza, en R no hay una funcion específica para calcularlos sino que forman parte de la informacion que nos da R al hacer un contraste de hipotesis. Como hemos visto en teoría, un intervalo de confianza es equivalente a un contraste de hipotesis bilateral. La opción del menú *Estadísticos* de R proporciona las funciones básicas para resolver un contraste de hipótesis y proporcionará la información que necesitamos del análisis inferencial para estos tres parámetros.

1. Significado del intervalo de confianza

Comenzaremos analizando el significado del intervalo de confianza. El objetivo del intervalo de confianza es dar una cierta garantía de la presencia del parámetro de la población dentro de un intervalo construido a partir de la muestra.

Podemos plantearnos la siguiente pregunta: ¿Es 0.5 la probabilidad de obtener cara al lanzar una moneda?

Para contestar esta cuestión vamos a realizar el siguiente experimento:

1. Lanzamos una moneda 20 veces, y estimamos ese valor con la proporción de caras obtenidas (p=pest). Simulando con R el lanzamiento de las monedas,
   `n<-20;  P<-0.5  pest<-rbinom(1,n,P)/n`

2. Repetimos el experimento, lanzando 20 monedas 50 veces (m=50) y calculamos la proporción de caras obtenidas para cada una de las muestras. `m<-50;  pest<-rbinom(m,n,P)/n`

3. Fijamos el nivel de confianza $1-\alpha = 0{,}90$ y calculamos el intervalo de confianza como: $p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
   ```
   alpha <- 0.10
   z <- qnorm(1-alpha/2)
   e <- z*sqrt(pest*(1-pest)/n)
   ```

4. Representamos los m = 50 intervalos con la opción
   ```
   matplot(rbind(pest-e, pest+e),rbind(1:m,1:m),type="l",lty=1,ylab="m",
   xlab="(pest-e, pest+e)")
   ```

5. Marcamos con una línea vertical el valor 0.5 ¿qué observas?
   `abline(v=P)`

6. Realizar ahora el lanzamiento de 20 monedas 10000 veces y comprobar con que probabilidad el valor de la proporción poblacional se encuentran dentro del intervalo de confianza calculado para cada muestra. Probar para diferentes niveles de confianza (90 % y 95 %).

7. ¿Cuántos intervalos contienen al valor 0.5? ¿Qué relación existe entre ese número y el nivel de confianza?

8. ¿El intervalo de confianza depende de la muestra elegida?

2. Contrastes de una población

2.1. Proporciones

En el caso de las proporciones, deberemos tener definida alguna columna de datos de tipo factor, para que podamos agrupar por ella. El fichero de datos *Pulso.rda* ya contiene varias columnas que son factores y nos permitirán calcular proporciones. En concreto, las columnas *Corre*, *Fuma* y *Sexo* nos permitirán distinguir a personas que tienen una cierta propiedad (haber corrido en la prueba, fumar, ser hombre o mujer) y podremos calcular la proporción de personas con esa propiedad en nuestra muestra (estimación puntual) y un intervalo de confianza para esa proporción en la población.

Vamos primero a calcular el intervalo de confianza de una proporción.

1. Calcular el estimador puntual de la proporción P de individuos que fuman.

2. Calcular el intervalo de confianza para la proporción P de individuos que fuman con una confianza del 95 % utilizando la fórmula vista en clase:
   $p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

3. Repetir el apartado anterior utilizando el menú que proporciona R *Estadísticos* → *Proporciones* → *Test de proporciones para una muestra*, dejando las opciones que vienen por defecto en la ventana emergente (Ver figura 1).
   Vemos como la forma de calcular el intervalo de confianza con R es a través del menú con el que se calculan los contrastes de hipótesis, suponiendo que se trata de un contraste bilateral. En este ejemplo el contraste de hipótesis se plantea suponiendo que la proporción de fumadores es igual a la de no fumadores (50 %), siendo por tanto la hipótesis nula P=0.5. La hipótesis alternativa vendrá dada como que la proporción de la población es por tanto distinta a ese valor como aparece indicado en la figura 1.

   Como se observa en la figura 2 el menú de R nos da gran cantidad de información como son el numero de individuos que fuman y no fuman, la estimación puntual de la proporción, el intervalo de confianza para un nivel de confianza dado y el p-valor entre otros.

   Notar que la función *prop.test* de R es la que genera el intervalo de confianza por lo que la podríamos escribir directamente como:

   ```
   > prop.test(c(28),c(92), alternative='two.sided', p=.5, conf.level=.95,
       correct=FALSE)
   ```

   sin utilizar el menú de R. En sucesivos ejemplos, no sería necesario utilizar el menú nuevamente, sino simplemente variar las opciones de esta función según sea necesario.

   Compara los resultados obtenidos con el menuú de R, con el intervalo calculado en el apartado anterior. La diferencia que observas es debida a que en R la fórmula implementada para el cálculo del intervalo de confianza de una prorporción viene dada por el intervalo de Wilson:

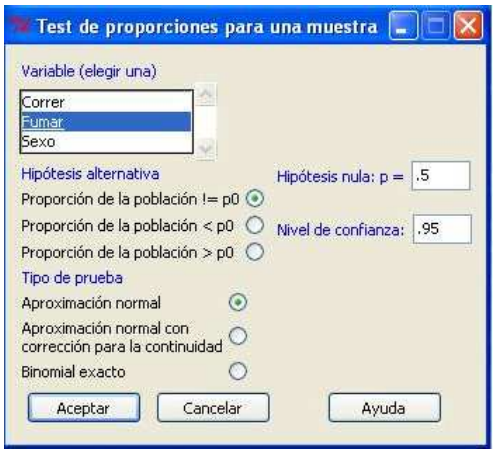

Figura 1: Test de proporciones para una muestra

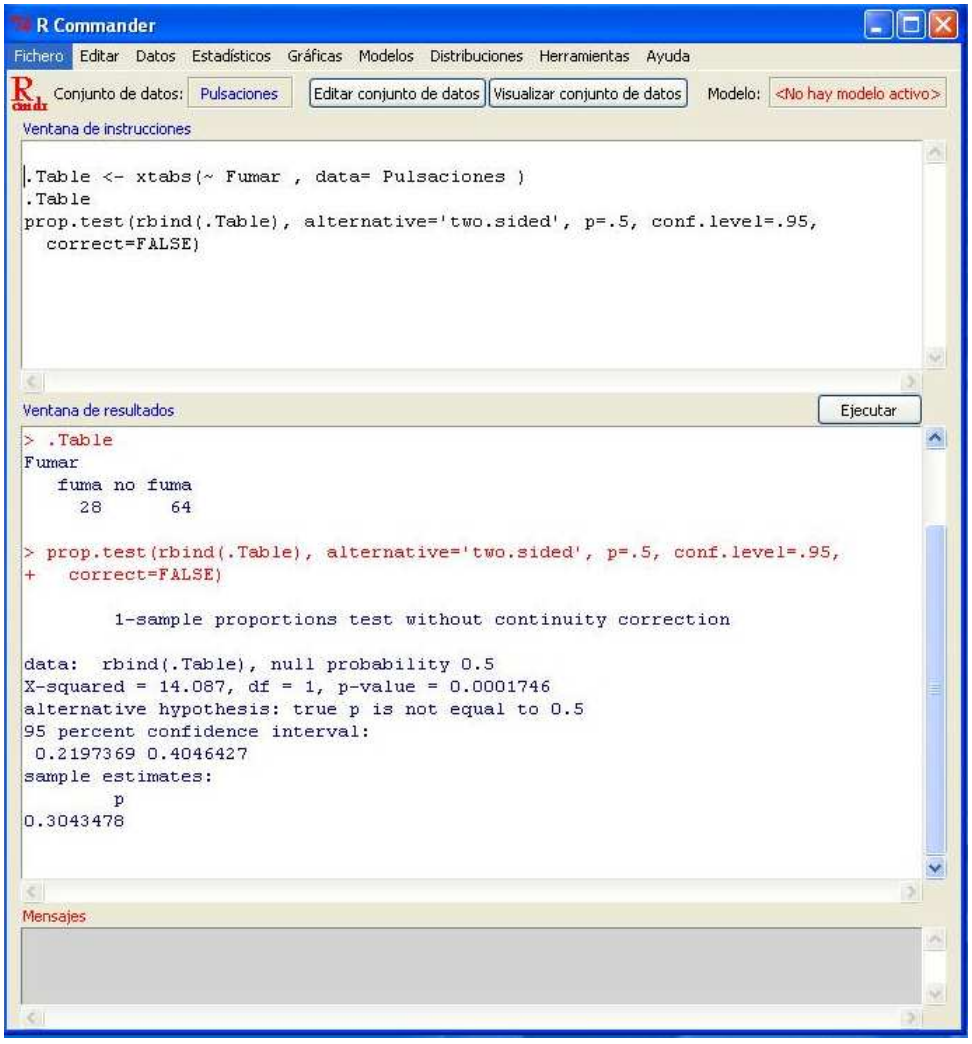

Figura 2: Salida del test de proporciones para una muestra

$$\frac{p+\frac{1}{2n}z^2_{\alpha/2} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}+\frac{z^2_{\alpha/2}}{4n^2}}}{1+\frac{1}{n}z^2_{\alpha/2}}$$

Esta es una versión mejorada de la aproximación normal que se considera en la fórmula del apartado anterior. Mientras que la estimación usada en ese apartado solo debe considerarse bajo ciertas condiciones vistas en clase, el intervalo de Wilson da buenos resultados incluso con tamaños de muestra pequeños.

Centrémonos ahora en el resultado obtenido para el p-valor. Según el contraste de hipótesis planteado en el ejemplo anterior, en el que se pone en duda si la proporción de fumadores de la población es igual a la de no fumadores ($H_0 : \mu == 0{,}5$, $H_1 : \mu \neq 0{,}5$), observamos que la muestra con la que hemos trabajado ha dado evidencias suficientes para rechazar la hipótesis nula con una confianza del 95 %, ya que el p-valor es menor que α. Es decir, el 50 % de la población no son fumadores. A esta misma conclusión llegamos si en vez de analizar el valor del p-valor consideramos el resultado obtenido para el intervalo de confianza, ya que en este ejemplo se plantea un contraste bilateral. Se observa que el intervalo de confianza obtenido no incluye al valor $P = 0{,}5$, luego se rechaza la hipótesis nula.

De la misma forma que hemos aplicado el contraste para este ejemplo en el que la hipótesis alternativa es de desigualdad, se podrían definir las hipótesis de un contrastes unilateral (por la izquierda o por la derecha), de acuerdo al problema analizado, sin más que definir de la forma correspondiente la hipótesis alternativa a través del menú de R (figura 1) y asignar el valor correspondiente a P.

Practica tú mismo

---

1) Con los datos del fichero *Pulso.rda* calcular:

1. El intervalo de confianza para la proporción de mujeres que fuman con una confianza del 95 %
2. Según los resultados obtenidos en el apartado anterior, se puede admitir como cierto que la proporción de mujeres que fuman es la misma que las que no fuman. Justifica la respuesta.
3. Calcular el intervalo de confianza para la proporción de individuos con el Pulso2 superando las 100 pulsaciones de entre los que corrieron, con una confianza del 95 %
4. Calcular el intervalo de confianza para la proporción de individuos con altura superior a 180 y peso superior a 85kg, con una confianza del 99 %

Practica tú mismo

---

2) Cierta medicina en tabletas ha sido comprobada eficaz en el alivio de una alergia en al menos el 60 % de los pacientes. El fabricante ha desarrollado una versión soluble del producto y desea comprobar si la medicina en esta forma es igual de eficaz. Se toma una muestra de 40 personas que tienen la alergia. El nuevo producto alivió a 19 de ellos. ¿Hay suficiente evidencia para sugerir que la introducción de la versión soluble ha alterado la eficacia de la medicina? Realiza el contraste usando $\alpha = 0{,}01$ y encuentra el nivel crítico del contraste (p-valor).

2.2. Media

El menú de R *Estadísticos* → *Medias* → *Test t para una muestra*, nos permite plantear contrastes de hipótesis para la media y obtener el intervalo de confianza. Para familiarizarnos con la función de R correspondiente, vamos a calcular el intervalo de confianza del peso medio de todos los individuos del fichero *Pulso.rda* con $\alpha = 0{,}05$. Seleccionamos la variable *Peso* en la ventana emergente del test t y marcamos el nivel de confianza correspondiente dejando el resto de opciones que vienen por defecto como aparecen en la figura 3.

En la ventana de resultados aparecen calculados el intervalo de confianza de la media poblacional, su estimador puntual (la media muestral) y el valor del estadístico de contraste calculado t. R calcula el intervalo de confianza mediante la expresión $\bar{x} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$ que es el más adecuado cuando no se conoce la varianza poblacional (caso más habitual). El cuantil $t_{n-1,\alpha/2}$ se puede calcular como `qt(1-alfa/2,n-1)`. Es decir, podríamos haber obtenido el mismo resultado mediante:

```
> x <- Altura
> n <- length(Altura)
> alfa <- 0.05
> t <- qt(1-alfa/2,n-1)
> x.min <- mean(x) - t*sd(x)/sqrt(n); x.min
> x.max <- mean(x) + t*sd(x)/sqrt(n); x.max
```

Si nos fijamos en la ventana de instrucciones, la orden que proporciona el intervalo de confianza es la siguiente:

```
> t.test(Pulsaciones$Peso, alternative='two.sided', mu=0.0, conf.level=.95)
```

por lo que en futuros cálculos del intervalo de confianza de una media podremos utilizar esa función directamente, sin utilizar el menú, cambiando únicamente los parámetros necesarios.

Figura 3: Test t para una muestra

Analicemos ahora un ejemplo para ver como trabajar con este menú de R para resolver contrastes de hipótesis de una media, a través del p-valor.

Estudios recientes afirman que la altura de esta población es superior a $\mu = 180$ cm. A la vista de la muestra recogida en el fichero *Pulso.rda* ¿podemos aceptar dicha hipótesis con un nivel de confianza del 99 %?

En este caso el contraste que se plantea es el siguiente:

$H_0 : \mu \geq 180$

$H_0 : \mu < 180$

En el menú utilizado antes para calcular el intervalo de confianza de la media seleccionamos ahora la opción de contraste unilateral por la izquierda e introducimos el valor de μ_0 y el nivel de confianza correspondientes. En la ventana de instrucciones aparece la orden que se ha ejecutado para calcular este contraste:

```
>t.test(Pulsaciones$Altura, alternative='less', mu=180, conf.level=.99)
```

La opción *alternative* se iguala ahora al valor *less* para indicar que se trata de un contraste unilateral por la izquierda. El valor *greater* indicaría un contraste unilateral por la derecha. Conociendo estas opciones, podríamos utilizar esa función modificando las opciones adecuadamente, para calcular cualquier otro contraste de medias.

En la ventana de resultados obtenenos el valor del estadístico de contraste (t), el número de grados de libertad y el p-valor entre otros. Del valor obtenido para el p-valor se puede concluir que la muestra analizada muestra evidencias suficientes para rechazar la hipótesis nula ya que obtenemos un p-valor menor que el nivel de significación considerado.

Practica tú mismo

---

3) Con los datos del fichero *Pulso.rda* calcular:

- el intervalo de confianza para el peso medio de las mujeres con $\alpha = 0{,}05$
- el intervalo de confianza para la media del incremento del pulso (Pulse2-Pulse1) para los individuos que corrieron $\alpha = 0{,}1$
- Estudios recientes afirman que la altura media de las mujeres de esta población es $\mu = 167$ cm. A la vista de estos datos, ¿podemos aceptar dicha hipótesis con un nivel de confianza del 99 %? Justificar el resultado.

Practica tú mismo

---

4) En cincuenta días lectivos consecutivos y a la misma hora se ha observado el número de terminales de una universidad conectados a internet. Los resultados están en el fichero terminales.dat. En base a esos datos,

1. Dar los intervalos de confianza al 95 % y 99.5 % para el número medio de terminales conectados a internet
2. Suponiendo que la población sigue una distribución normal, calcular intervalos de confianza al 90 % y 95 % para la varianza del número de terminales conectados a internet. Hacer una función que devuelva los valores de a y b dado un nivel de confianza.

Practica tú mismo

5) El pH del suelo es una variable importante cuando se diseñan estructuras que estarán en contacto con el terreno. El propietario de un solar posible lugar de construcción afirma que el pH del suelo no es superior a 6.5. Se han tomado 9 muestras del suelo del terreno, obteniéndose los resultados que se recogen en el archivo de datos *pH.txt*. Suponiendo que la variable pH sigue una distribución normal, responde a las siguientes cuestiones.

1. Hallar un intervalo de confianza para el pH medio con un nivel de significación del 10 %.
2. ¿Se acepta como verdadera la afirmación del propietario del solar con un riesgo de $\alpha = 0{,}05$?

3. Contrastes de dos poblaciones

En este caso en vez de trabajar con una sola muestra vamos a trabajar con dos ya que se quiere contrastar cierta información acerca de los parámetros (media o proporción) de dos poblaciones.

Analizaremos en detalle el caso de un contraste de medias de dos poblaciones, en el que se quiere determinar, utilizando los datos de fichero *Pulso.rda*, si hay diferencia significativa entre la altura media de hombres y mujeres con un nivel de significación $\alpha = 0{,}05$. Es decir, el contrastes que se plantea es el siguiente:
$H_0 : \mu_h - \mu_m = 0$
$H_1 : \mu_h - \mu_m \neq 0$

El menú de R *Estadísticos* $\rightarrow$ *Medias* $\rightarrow$ *Test t para muestras independientes*, nos permite calcular este contraste. En este caso seleccionamos el caso de muestras independientes ya que los datos no son apareados sino que hombres y mujeres son muestras independientes. En la ventana emergente selecionamos la variable categórica *Sexo* y como variable respuesta la variable *Altura*. En este caso el tipo de contraste es bilateral y como el enunciado no indica nada, no asumimos varianzas iguales. Tras darle al botón aceptar vemos que la función de R que calcula el contraste es:

```
t(Altura~Sexo, alternative='two.sided', conf.level=.95,var.equal=FALSE, data=Pulsaciones)
```

Que sería equivalente a escribir:

```
ah<-Altura[Sexo=='hombre']
am<-Altura[Sexo=='mujer']
t(ah, am, alternative='two.sided', conf.level=.95,var.equal=FALSE, data=Pulsaciones)
```

en vez de utilizar los menús de R.

En la ventana de resultados, como en los casos anteriores, obtenemos el valor del estadístico de contraste (t), y el p-valor entre otros. Del resultado del p-valor se puede concluir que la muestra analizada aporta evidencias suficientes para rechazar la hipótesis nula ya que como vemos, el p-valor es menor que α. Es decir, la media de la altura de los hombres no es igual que la de las mujeres.

El contraste de proporciones de dos poblaciones se haría de la misma manera que el de medias a través del menú *Estadísticos* → *Proporciones* → *Test de proporciones para dos muestras*. Seleccionando las variables correspondientes, el tipo de contraste e introduciendo el nivel de confianza determinado.

Practica tú mismo

---

6) Una empresa tiene en su poder dos dispositivos para mejorar la eficiencia de los sistemas de calefacción en los hogares. Uno de ellos funciona con energía eléctrica (sistema 1) y el otro con energía térmica (sistema 2). Se quiere estudiar si ambos dispositivos son igualmente efectivos, para lo cual se compara el consumo de energía en 90 hogares que tienen uno u otro sistema. Los datos del estudio se recogen en el archivo *energía.rda*. Responder a las siguientes preguntas considerando un nivel de confianza del 95 %:

1. Valor del estadistico de contraste
2. Valor del p-valor
3. ¿Se acepta H_0? Razona la respuesta.

Practica tú mismo

---

7) De estudios anteriores de conoce que la proporción de mujeres fumadoras es mayor que la de hombres fumadores. En los últimos años se sospecha que este dato ha cambiado. Utilizando los datos del fichero *Pulso.rda*, determinar con una confianza del 90 % si la proporción de mujeres fumadoras es menor que la de hombres fumadores.

Practica tú mismo

8) Con el fin de evaluar el proceso de aprendizaje en clase, a los alumnos de una clase de un colegio al comienzo del curso se les dicta un texto y se les corrige el número de faltas de ortografía cometidas. Al final del curso, se les vuelve a dictar el mismo texto y se les corrige de nuevo las faltas de ortografía. El número de faltas de ortografía en cada caso se recojen en la siguiente tabla.

Antes	24	20	24	28	30	20	24	22	18	18	24
Después	25	18	22	21	27	15	19	23	16	19	19

1. Determine, para un α de 0.05, si la metodología aplicada en las clases ayuda a los alumnos a disminuir los errores que cometen al escribir.

2. Determinar el intervalo de confianza al 99 % para la media de las faltas de ortografía obtenidas al comienzo del curso.