Práctica 4. Inferencia estadística

Grado de Ingeniería Civil - Universidad de Cantabria

En esta práctica vamos a utilizar las opciones de R para aplicar el análisis inferencial. 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.

No hay una funcion específica para calcular los intervalos de confianza, pero forman parte de la informacion que nos da R al hacer un contraste de hipotesis. Como sabemos, un intervalo de cofianza es equivalente a un contraste de hipotesis bilateral. El menú *Estadísticos* de R proporciona las funciones básicas para resolver un contraste de hipótesis y por tanto nos 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. Intervalo de confianza de una proporción

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 ahora a utilizar el menú de R para 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).
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.

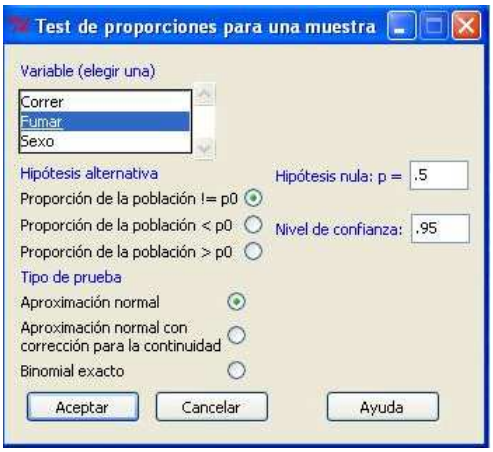

Figura 1: Test de proporciones para una muestra

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 y

otra serie de valores. De todo ello nos interesa la parte que hace referencia al intervalo de confianza.

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.

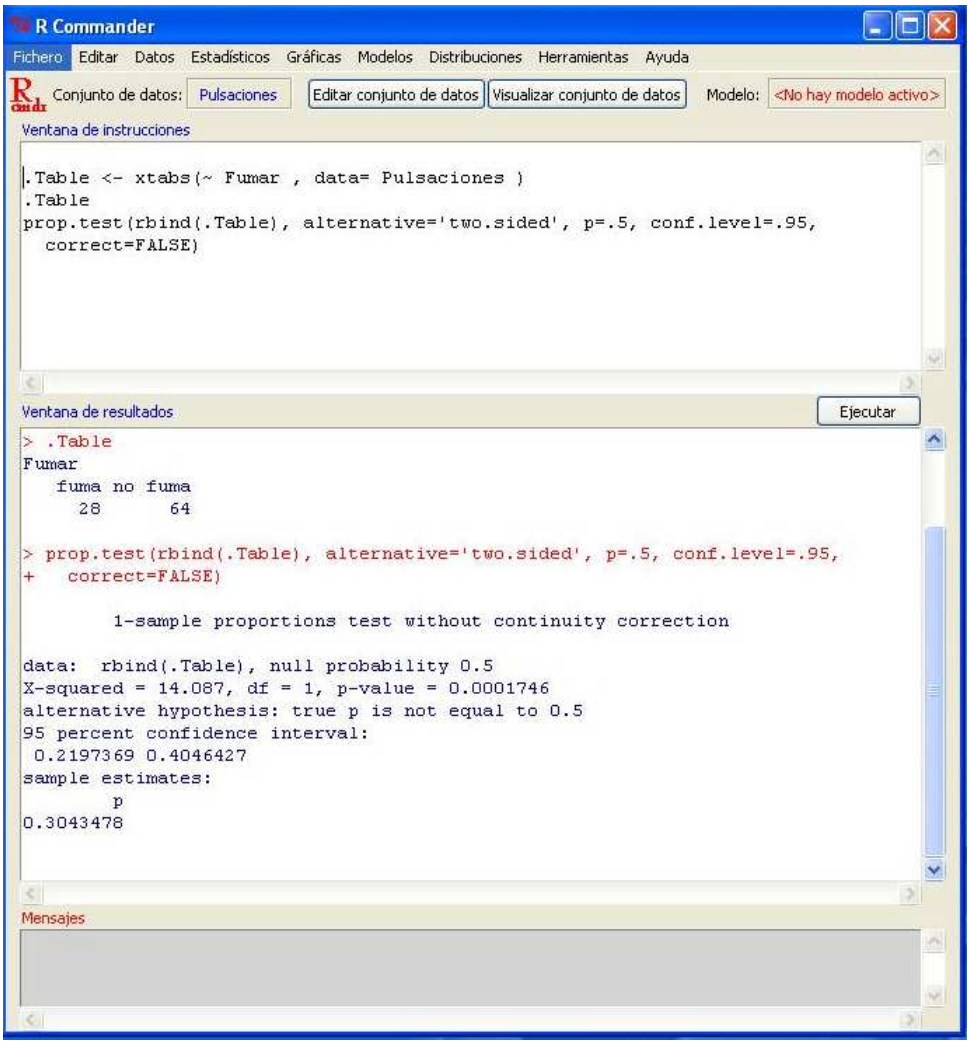

Figura 2: Salida del test de proporciones para una muestra

Compara los resultados con lo obtenido 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:

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

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.

Practica tu mismo:

2.1 Con los datos del fichero *Pulso.rda* calcula los siguientes intervalos de confianza:

1. Calcular el intervalo de confianza para la proporción de mujeres que fuman con una confianza del 95 %
2. 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 %
3. Calcular el intervalo de confianza para la proporción de individuos con altura superior a 180 y peso superior a 85kg, con una confianza del 95 %

3. Intervalo de confianza de una media

El menú de R *Estadísticos* → *Medias* → *Test t para una muestra*, nos permite obtener el intervalo de confianza de una media. 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

Practica tu mismo:

3.2 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 $\alpha = 0{,}1$

3.3 Se ha observado una muestra de 41 datos aleatorios del tiempo de respuesta de un servidor web. Los tiempos (medidos en segundos) obtenidos están en el fichero tiempos.dat. En base a esta muestra, calcular según la teoría vista en clase

1. Intervalos de confianza al 90 % y 85 % del tiempo medio de respuesta.
2. Calcular el apartado anterior mediante la opción de R
   t.test(data, alternative='two.sided', mu=0.0, conf.level=.90)
3. Cuando el servidor tarda más de 3.5 segundos se produce un fallo en el acceso. Estimar el intervalo de confianza del 90 % para la proporción de fallos.

3.4 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