Estadística I
Tema 7: Intervalos de Confianza

UNA POBLACIÓN NORMAL: ICs para la MEDIA

1) En este ejercicio con intervalos de confianza para la media muestral μ se necesita la instalación del paquete `RcmdrPlugin.TeachingDemos`. Cargamos el paquete mediante el menú de `Rcmdr`:
`Herramientas` $\rightarrow$ `Cargar plug-in(s) de Rcmdr`
En la ventana de diálogo, elige `RcmdrPlugin.TeachingDemos`. `R` preguntará sobre el permiso para rearrancar `Rcmdr` (`¿Reiniciar ahora?`), de modo que se pueda añadir la opción de `Demostraciones` a su menú principal.
Para ver cómo funcionan los intervalos de de confianza para la media poblacional:

a) Genera $I = 50$ muestras de tamaño $n = 9$ de una distribución normal $N(\mu = 100, \sigma^2 = 4^2)$ y para cada una de las I muestras construye un intervalo de confianza al 95% para μ (se supone que μ es desconocida) asumiendo que la desviación estándar poblacional es conocida:
`Demostraciones` $\rightarrow$ `Intervalos de confianza para la media...`
En la ventana de diálogo rellena los campos `Population mean, Population standard deviation, Sample size, Number of samples, Confidence level` con 100, 4, 9, 50, 0.95, respectivamente. Marca el campo `Population standard deviation known`. Pulsa OK.
Con un nivel de confianza del 95%, en teoría, sólo el 95% de los I intervalos de confianza cubrirán μ y los restantes 5% no lo harán.
¿Cómo funciona esta regla en la práctica? es decir, ¿cuál es el porcentaje de los $I = 50$ intervalos de confianza que no incluyen a μ?
¿Son simétricos los intervalos de confianza respecto a las correspondientes medias muestrales?

b) Abre una ventana gráfica nueva (ejecuta `x11()` en la ventana `Ventana de Instrucciones`). Repite a), asumiendo que la desviación estándar poblacional σ es desconocida.
¿Cuál es la distribución muestral de $\bar{x}$ en ese caso?

c) Abre una nueva ventana gráfica. Repite a) para $n = 100$ y con el resto haz lo mismo.
¿Los intervalos de confianza son más largos o más cortos comparados con los del apartado a)? ¿Por qué?

d) Abre una nueva ventana gráfica. Repite a) para $\sigma = 1$ y con el resto haz lo mismo.
¿Los intervalos de confianza son más largos o más cortos comparados con los del apartado a)? ¿Por qué?

e) Abre una nueva ventana gráfica. Repite a) para el `Confidence level` del 99% y con el resto haz lo mismo.
¿Los intervalos de confianza son más largos o más cortos comparados con los del apartado a)? ¿Por qué?

2) Carga el conjunto de datos `trees` del paquete `datasets` ejecutando:

```
data(trees, package = "datasets")
```

Además, si escribes en la `Ventana de Instrucciones`:

```
attach(trees)
```

Te puedes referir a las variables de este conjunto de datos por sus nombres sin necesidad de escribir `trees$` antes del nombre.

La variable de interés es `Height`.

a) Comprueba la suposición de normalidad de la población de `Heights` examinando el histograma de la muestra:
```
hist(Height)
```

b) Construye un intervalo de confianza al 90% para μ (media poblacional de la variable height):
```
t.test(x=Height,alternative="two.sided",conf.level=0.90)
```
(Esta función de `R` se aplica fundamentalmente en el "Tema 8: Contrastes de Hipótesis", pero la estudiaremos también en el "Tema 7".)
Se obtendrá $(74.05764, 77.94236)$. Interpreta el resultado.

c) Indica cuáles son las suposiciones necesarias para que sean legítimos los intervalos de confianza en b).

UNA POBLACIÓN NORMAL: ICs para la VARIANZA

3) En este ejercicio se harán gráficos similares a los del problema 1) (donde se usó el paquete `Teaching Demos`), aunque calculando los intervalos de confianza para σ^2.

Supongamos que $X \sim N(\mu = 0, \sigma^2 = 3^2)$.

Generamos $I = 50$ muestras de X, cada una con $n = 25$ observaciones (guardamos los datos como una matriz I por n, donde las muestras aparecen en las filas), y calculamos un intervalo de confianza al 95% para σ^2 (suponiendo que es desconocida) para cada muestra:

`Distribuciones` $\longrightarrow$ `Distribuciones continuas` $\longrightarrow$ `Distribución normal` $\longrightarrow$ `Muestra de una distribución normal...`

En la ventana de diálogo, `Introducir el nombre del conjunto de datos:` Ponemos como nombre `Muestras.normales`. Ajustamos los parámetros de la distribución si fuera necesario. En el campo `Número de muestras` ponemos '50' y en el de `Número de observaciones` '25', y marcamos el campo `Desviación estándar de cada muestra`. El conjunto activo será, entonces, `Muestras.normales`. Si marcamos en `Visualizar conjunto de datos` se ven las I desviaciones muestrales estándar en la última columna de la hoja de datos etiquetada como `sd`.

a) Calcula los cuantiles $\alpha/2$ y $(1 - \alpha/2)$ de la distribución chi-cuadrada con $\text{df} = n - 1 = 24$ (revisa el Laboratorio anterior que tiene instrucciones similares):
`Distribuciones` $\longrightarrow$ `Distribuciones continuas` $\longrightarrow$ `Distribución Chi-cuadrado` $\longrightarrow$ `Cuantiles Chi-cuadrado`
En la ventana de diálogo, especifica las `Probabilidades` como `c(0.025,0.975)`, los `Grados de libertad` como 24, y marca la `Cola derecha`.
Debes obtener 39.36 y 12.40, respectivamente.

b) Calcula los extremos inferiores de los intervalos:
   `Datos` → `Modificar variables del conjunto de datos activo` → `Calcular una nueva variable`
   En la ventana de diálogo, selecciona una de las `Variables actuales` (pulsando con el ratón). Introduce el `Nombre de la nueva variable`, por ejemplo, `ext.inferior` e introduce la `Expresión a calcular` como `24*sd^2/39.36`. Pulsa OK.

c) Calcula el extremo superior: Introduce la `Nombre de la nueva variable`, por ejemplo, `ext.superior` e introduce la `Expresión a calcular` como `24*sd^2/12.40`

Nota: Los pasos a)-c) se pueden hacer también ejecutando el siguiente código en la `Ventana de Instrucciones`:

```
ext.inferior.denom=qchisq(p=0.025,df=24,lower.tail=FALSE)
ext.superior.denom=qchisq(p=0.975,df=24,lower.tail=FALSE)
Muestras.normales$ext.inferior=with(Muestras.normales,24*sd^2/ext.inferior.denom)
Muestras.normales$ext.superior=with(Muestras.normales,24*sd^2/ext.superior.denom)
```

Después de esto, el conjunto de datos activo tendrá dos columnas más cada una de longitud I denominada `ext.inferior` y `ext.superior`.
Para dibujar los intervalos de confianza, ejecuta el siguiente código en la `Ventana de Instrucciones`:

```
I=50
sigma=3
Index=1:I
xlim=c(0,sigma^2+40)
ylim=c(1,I)
plot(x=xlim,y=ylim,type="n",xlab=expression(sigma^2),ylab="Index")
segments(Muestras.normales$ext.inferior,Index,
Muestras.normales$ext.superior,Index)
abline(v=sigma^2,col="red")
```

Con un nivel de confianza del 95%, en teoría, sólo el 95% de los I intervalos de confianza recubren a σ^2 y el restante 5% no lo hace.
¿Cómo funciona esta regla en la práctica? es decir, ¿qué porcentaje de los $I = 50$ intervalos de confianza no incluyen a σ^2?

4) Usa el mismo conjunto de datos que en 2).

a) Construye un intervalo de confianza al 90% para σ^2 (varianza poblacional de la variable `Height`):

```
n=length(Height)
sample.var=var(Height)
ext.inferior.denom=qchisq(p=0.05,df=n-1,lower.tail=FALSE)
ext.superior.denom=qchisq(p=0.95,df=n-1,lower.tail=FALSE)
c((n-1)*sample.var/ext.inferior.denom,
(n-1)*sample.var/ext.superior.denom)
```

Interpreta.

b) Construye un intervalo de confianza al 95% para σ (desviación estándar poblacional de la variable `Height`):

```
n=length(Height)
sample.var=var(Height)
ext.inferior.denom=qchisq(p=0.025,df=n-1,lower.tail=FALSE)
ext.superior.denom=qchisq(p=0.975,df=n-1,lower.tail=FALSE)
c(sqrt((n-1)*sample.var/ext.inferior.denom),
sqrt((n-1)*sample.var/ext.superior.denom))
```

Interpreta.

DOS POBLACIONES NORMALES: ICs para la DIFERENCIA entre DOS MEDIAS y para el COCIENTE (RAZÓN) de las VARIANZAS

5) Genera dos muestras aleatorias independientes de $X_1 \sim N(\mu_1 = 5, \sigma_1^2 = 4^2)$ de tamaño $n_1 = 15$ y de $X_2 \sim N(\mu_2 = 5, \sigma_2^2 = 5^2)$ de tamaño $n_2 = 20$ ejecutando en la `Ventana de Instrucciones`:

```
x1=rnorm(n=15,mean=5,sd=4)
x2=rnorm(n=20,mean=5,sd=5)
```

a) Calcula un intervalo de confianza al 90% para $\mu_1 - \mu_2$ asumiendo que las varianzas poblacionales son desconocidas y diferentes:

```
t.test(x=x1, y=x2, alternative = "two.sided", conf.level = 0.9,
var.equal=FALSE)
```

Interpreta.

b) Calcula un intervalo de confianza al 85% para σ_1^2/σ_2^2:

```
var.test(x=x1, y=x2, alternative="two.sided", conf.level=0.85)
```

Interpreta.

6) Carga los siguientes datos sobre tasas de inflación en países desarrollados (X_1) y en vías de desarrollo (X_2) en la `Ventana de Instrucciones` ejecutando:

```
developed=c(3.99, 4.07, 3.70, 1.79, 5.30, 3.47, 2.39, 3.33, 4.14, 3.11)
developing=c(4.73, 5.01, 5.07, 4.66, 4.49, 4.00, 4.33, 5.14, 3.15, 3.46)
```

a) Comprueba la normalidad de ambas poblaciones examinando los histogramas de las muestras (usamos el comando `par(mfrow=c(1,2))` para dibujar dos figuras en una misma ventana gráfica):

```
par(mfrow=c(1,2))
hist(developed,breaks=2)
hist(developing,breaks=2)
```

b) Calcula un intervalo de confianza al 95% para σ_1^2/σ_2^2:

```
var.test(x=developed,y=developing,alternative="two.sided",conf.level=0.95)
```

Interpreta.

Nota: Si el valor igual a 1 pertenece a este intervalo, se concluye que σ_1^2/σ_2^2 es igual a 1 (con una confianza del 95%). Así, $\sigma_1^2 = \sigma_2^2$ y se dice que las dos poblaciones tienen varianzas iguales (Leer “Tema 8” para más detalles). En caso contrario, se dice que las varianzas poblacionales son diferentes (con una confianza del 95%).

c) Calcula un intervalo de confianza al 95% para $\mu_1 - \mu_2$, teniendo en cuenta los resultados de la parte a) (fija `var.equal` en `TRUE` si las varianzas poblacionales son iguales y `FALSE` en otro caso):

```
t.test(x=developed, y=developing, alternative="two.sided",
conf.level=0.95, var.equal=...)
```

Interpreta.

d) Indica cuáles son las suposiciones necesarias para que sean legítimos los intervalos de confianza en b) y en c).

MUESTRAS GRANDES: UNA POBLACIÓN (ICs para la PROPORCIÓN), DOS POBLACIONES (ICs para la DIFERENCIA de las PROPORCIONES)

7) Genera dos muestras aleatorias independientes de $X_1 \sim Bernoulli(p_1 = 0.3)$ de tamaño $n_1 = 40$ y de $X_2 \sim Bernoulli(p_2 = 0.4)$ de tamaño $n_2 = 50$ ejecutando en la `Ventana de Instrucciones`:

```
x1=sample(x=c(0,1),size=40,replace=TRUE,prob=c(0.7,0.3))
x2=sample(x=c(0,1),size=50,replace=TRUE,prob=c(0.6,0.4))
```

a) Calcula un intervalo de confianza al 90% para p_1 (sin corrección de continuidad, `correct=FALSE`):

```
prop.test(sum(x1), n=length(x1), p=NULL, alternative="two.sided",
conf.level=0.90, correct=FALSE)
```

Alternativamente,

```
z.test(x=x1, stdev=sqrt(mean(x1)*(1-mean(x1))), alternative="two.sided",
conf.level=0.90)
```

Interpreta.

b) Calcula un intervalo de confianza al 95% para $p_1 - p_2$:

```
prop.test(c(sum(x1),sum(x2)),n=c(length(x1),length(x2)), p = NULL,
alternative = "two.sided",conf.level = 0.95, correct = FALSE)
```

Alternativamente,

```
n1=40
n2=50
p1=mean(x1)
p2=mean(x2)
v1=(p1*(1-p1))/n1
v2=(p2*(1-p2))/n2
z=qnorm(p=0.025,lower.tail=FALSE)
c(p1-p2-z*sqrt(v1+v2),p1-p2+z*sqrt(v1+v2))
```

Interpreta.

c) ¿Qué teorema permite calcular los intervalos de confianza en a) y b)? ¿Qué condiciones deben cumplir n_1 y n_2?