# Práctica 9, parte II: inferencia sobre dos poblaciones

### Objetivos

- Determinar las condiciones de aplicabilidad de los intervalos de confianza y el tipo de intervalo a utilizar en cada situación.
- Construir los intervalos de confianza para la media, la varianza y la proporción de variables aleatorias con distribución normal.
- Aprender a construir los intervalos de confianza para la diferencia de medias (datos pareados y no pareados), cociente de varianzas, y diferencia de proporciones de dos variables aleatorias normales.
- Interpretar los resultados obtenidos.

## Enunciado

Los datos que contiene el fichero ciudades.sf3 se refieren (a excepción de la última columna) a 41 ciudades de Estados Unidos y fueron extraídos de distintas revistas publicadas por el gobierno de este país durante los años 1969 y 1970. Las variables que aparecen son: CIUDAD (nombre de la ciudad), SO2 (contenido de SO2 en el aire medido en microgramos por metro cúbico), TEMP (promedio de la temperatura anual en grados Fahrenheit), MAN (número de empresas de manufacturación con 20 o más empleados), POP (tamaño demográfico en decenas de miles según el censo de 1970), VIENTO (promedio de la velocidad del viento anual en millas por hora), LLUVIA (promedio de la precipitación anual en pulgadas), DIASLLUVIA (promedio del número de días que llueve al año), CLIMA (variable categórica que especifica el tipo de clima) y SO2_*tr* (la variable SO2 transformada adecuadamente). En la última columna se añaden datos ficticios correspondientes a una posible evolución de la variable SO2, en el año 1975; el nombre que recibe esta variable, en el fichero, es SO2tr_75.

Puedes decargar los datos desde `https://goo.gl/en1f3r`. Para leerlos directamente de internet y guardarlos en la variable `ciudades`, copia en un script de R y ejecuta el siguiente código

```
web = "https://goo.gl/en1f3r"
ciudades = read.table(file = web, sep = ";", header = TRUE, dec = ".")
```

```
head(ciudades, 4)
```

```
##          CIUDAD SO2 TEMP MAN POP VIENTO LLUVIA DIASLLUVIA       CLIMA
## 1       Phoenix  10 70.3 213 582    6.0   7.05         36 continental
## 2    LittleRock  13 61.0  91 132    8.2  48.52        100 subtropical
## 3 San Francisco  12 56.7 453 716    8.7  20.66         67 subtropical
## 4        Denver  17 51.9 454 515    9.0  12.95         86 continental
##     SO2_tr SO2_tr75
## 1 2.302585  4.68068
## 2 2.564949  4.15766
## 3 2.484907  6.47724
## 4 2.833213  2.03278
```

1. ¿Se puede afirmar, con un nivel de confianza del 95%, que la varianza de la lluvia es la misma que la de días de lluvia en las ciudades de Estados Unidos?

   Se pide contrastar la hipótesis

   $$H_0 : \sigma_{lluvia} = \sigma_{diaslluvia} \qquad H_1 : \sigma_{lluvia} \neq \sigma_{diaslluvia}$$

   Es necesario que las muestras provengan de poblaciones normales, lo que supondremos cierto

   ```
   ## [1] "El p-Valor es 1.18398876506385e-06"
   ```

   a la vista del p-valor se rechaza que las varianzas sean iguales.

2. ¿Qué relación (numérica) hay entre las varianzas?

   Ahora hay que calcular el intervalo de confianza para la varianza. Usando la plantilla correspondientes (intervalo confianza para cociente de varianzas datos en bruto) se obtiene

   ```
   ## [1] 0.1052 0.3698
   ```

   es decir, $0.1052 < \sigma_1^2/\sigma_2^2 < 0.3698$ o, lo que es lo mismo:

   $$0.1052\sigma_2^2 < \sigma_1^2 < 0.3698\sigma_2^2$$

   de donde se deduce que $\sigma_1^2$ es entre 0.1052 y 0.3698 veces menor que $\sigma_2^2$.

3. ¿Qué cantidad de lluvia ha caído en Albuquerque?
   Cayeron

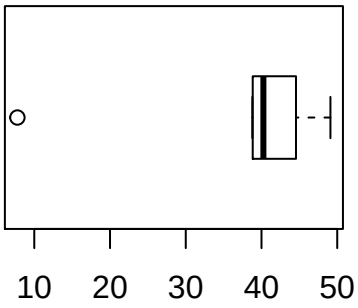
   ```
   ciudades$LLUVIA[which(ciudades$CIUDAD == "Albuquerque")]

   ## [1] 7.77
   ```

   pulgadas de agua.

   ¿Piensas que este dato se corresponde con el resto de ciudades de clima templado?

   ```
   boxplot(ciudades$LLUVIA[ciudades$CLIMA == "templado"],
           horizontal = TRUE)
   ```




En el diagrama de caja y bigotes se observa un dato atípico que corresponde a Albuquerque.

Descarta provisionalmente este dato para responder a los dos siguientes ejercicios.
Para eso

```
SinAlbqq = ciudades[ciudades$CIUDAD != "Albuquerque" , ]
```

4. Compara la lluvia caída en ciudades de clima templado y de clima continental. Supón normalidad donde sea necesario.

   (a) ¿Observas diferencias significativas, al 90% de confianza?
   Se pide contrastar la hipótesis

   $$H_0 : \mu_{templados} = \mu_{continental} \qquad H_1 : \mu_{templados} \neq \mu_{continental}$$

   para decidir si las medias pueden ser consideradas similares o no. Hay que las variables en dos grupos, que llamaremos `lluviaTemp` y `lluviaCont`, respectivamente:

   ```
   lluviaTemp = SinAlbqq$LLUVIA[SinAlbqq$CLIMA == "templado"]
   lluviaCont = SinAlbqq$LLUVIA[SinAlbqq$CLIMA == "continental"]
   ```

   Veamos el tamaño de cada muestra para decidir si se usa una normal o la t de Student

   ```
   length(lluviaTemp)

   ## [1] 6
   ```

```
length(lluviaCont)
```

```
## [1] 6
```

Como las muestras son pequeñas, para poder usar la t de Student habría que comprobar la normalidad de los datos (el enunciado dice que supongamos que es así). En primer lugar hay que determinar si las varianzas pueden ser consideradas como iguales o no. Con la plantilla correspondiente se obtiene

```
## [1] "El p-Valor es 0.0129405084960459"
```

es decir, las varianzas son diferentes al nivel de significación $\alpha = 0.1$. Usando la plantilla correspondiente para la t de Student, varianzas distintas y datos en bruto se obtiene, $(\mu_{templado} - \mu_{continental})$

```
## [1] "El p-Valor es 0.0616103724463573"
```

(b) En caso afirmativo, cuantifica dicha diferencia. En este caso se pide el intervalo de confianza; hay que usar la plantilla de diferencia de medias usando t con varianza distintas. El intervalo calculado es para `lluviaTemplados - lluviaContinental`:

```
## [1]  2.319 27.690
```

como la diferencia es positiva, la media de lluviaTemplados es mayor que la de lluviaCont. EN concreto, es entre 2.319 y 27.69 pulgadas mayor, al nivel de confianza del 90%.

5. Admitiendo como ciertos, para el valor de SO2_ tr en 1975 los datos de la variable SO2tr_ 75,

(a) ¿puede decirse, al 90% de confianza, que el nivel de SO2 se incrementó entre 1970 y 1975? En este caso los datos están pareados. Por tanto, hay que crear una nueva variable, que llamaremos `incremento`:

```
incremento = ciudades$SO2_tr75 - ciudades$SO2_tr
```

de nuevo habría que comprobar la normalidad de la nueva variable, pero se asume que lo es. A partir de la próxima práctica, que dispondrás de las herramientas adecuadas, habrá que hacer la comprobación, claro).

Hay que contrastar las hipótesis

$$H_0 : \mu_{incremento} \leq 0 \qquad H_1 : \mu_{incremento} > 0$$

```
## [1] 0.04310819
```

luego podemos afirmar que hubo un incremento.

(b) Cuantifica la evolución de la concentración de SO2 entre los años 1970 y 1975.

Para obtener el intervalo se puede usar la plantilla de intervalo de confianza de una variable usando t. El intervalo de confianza para la media al 90% es

```
## [1] 0.02544 1.15000
```

como el intervalo no contiene al cero, una de ellas supera a la otra y, como es positivo, efectivamente ha habido un incremento en la concentración de SO2 de entre 0.02544 y 1.15.

6. Hemos comprobado que, sobre esta muestra, el 56% de las ciudades con más de 1 millón de habitantes (POP>100) tienen un plan de sostenibilidad medioambiental. Este mismo porcentaje en Canadá, medido sobre una muestra de 35 ciudades, es del 72%. ¿Podemos admitir, al 99% de confianza, que no hay diferencias significativas a este respecto entre ambos países?

   De nuevo hay que contrastar hipótesis. En este caso sobre proporciones: las hipótesis

$$H_0 : p_{USA} = p_{Canada} \qquad H_1 : p_{USA} \neq p_{Canada}$$

```
## [1] "El p-Valor es 0.149012474619518"
```

   Por lo que las proporciones se pueden considerar iguales.

   Si se quisiera cuantificar entre qué valores está la diferencia, basta con calcular el intervalo de confianza para la diferencia de proporciones:

   es decir, $p_{USA} - p_{canada} \in (-0.4394, 0.1194)$ lo que equivale a

$$p_{canada} - 0.4394 < p_{USA} < p_{canada} + 0.1194$$