universidad de león

Inferencias sobre una media

Métodos Numéricos y Estadísticos
Grado en Ingeniería Informática / Mecánica

Curso 2021-2022

Eva María Mazcuñán Navarro

Contenidos

Página

Introducción **2**
Requisitos previos . . . 2
Flujo de trabajo . . . 2

1 Planteamiento del problema **2**

2 Estadística descriptiva **4**
2.1 Media muestral . . . 4
2.2 Varianza muestral y desviación típica muestral . . . 5
2.3 Cuantiles . . . 5
2.4 La función `summarize()` . . . 7
2.5 Histograma . . . 8
2.6 Diagrama de caja y bigotes . . . 10

3 Contraste de hipótesis **14**

4 Intervalo de confianza **16**

Introducción

En el tema **Inferencia Estadística** se estudiaron los fundamentos de las principales técnicas de inferencia estadística: los contrastes de hipótesis y los intervalos de confianza. Se presentaron los nuevos conceptos partiendo del caso particular de un problema de inferencia para una **proporción**.

En esta práctica se aplicarán las mencionadas técnicas para realizar inferencias sobre un nuevo parámetro: la **media** de una variable continua.

Requisitos previos

Antes de comenzar esta práctica, necesitas:

- Tener R y RStudio instalados en tu equipo (ver Instalación de R y RStudio).
- Haber estudiado la práctica Primeros pasos con R y RStudio.
- Haber estudiado el tema de inferencia estadística.

Flujo de trabajo

Documenta lo que vayas aprendiendo conforme leas la práctica usando un documento R Markdown. Puedes utilizar esta plantilla.

Se recomienda guardar el archivo R Markdown en una carpeta propia. En dicha carpeta se creará el archivo HTML resultante de la compilación y después añadiremos los archivos con los datos que usaremos a lo largo de la práctica.

Recuerda que para crear encabezados se utiliza la sintaxis `#` (nivel 1), `##` (nivel 2), ...; y que los bloques de código se crean con el atajo `Ctrl + Alt + I`.

Respecto al seccionado del documento, lo más práctico es que imites la estructura de este guión de prácticas.

1. Planteamiento del problema

> Una empresa de domótica comercializa un modelo de mandos a distancia para la apertura de puertas de garaje.

> Estamos interesados en investigar el radio de alcance medio de este modelo de mandos, siendo el radio de alcance de un mando la distancia máxima medida en metros a la que el mando consigue abrir la puerta.
>
> La empresa afirma que el radio de alcance medio de sus mandos es superior a 50 metros. Se tomó una muestra 100 mandos y se midió su radio de alcance, con el objetivo de contrastar si los datos recogidos resultan compatibles con la afirmación de la empresa.

Consideremos la variable aleatoria

$$X = \text{radio de alcance de un mando, en metros.}$$

Nuestro objetivo es investigar el valor del parámetro

$$\mu = E(X),$$

la media o valor esperado de la variable X.

En el archivo enlazado a continuación se registran los radios de alcance de los 100 mandos analizados: mandos.csv. En el resto de la práctica se asume que el archivo `mandos.csv` está ubicado en un directorio de nombre `data` dentro del directorio en el que se encuentre el archivo `.Rmd` con el código.

En esta práctica exploraremos qué conclusiones o inferencias podemos realizar sobre el parámetro desconocido μ, a partir de las 100 mediciones realizadas.

Comenzamos cargando el paquete `tidyverse` para tener acceso a todas las funciones que usaremos:

```
library("tidyverse")
```

Importamos los datos del problema con la instrucción

```
mandos <- read_csv("data/mandos.csv")
```

que creará la hoja de datos `mandos` con los contenidos del archivo `mandos.csv`.

Al visualizar el objeto `mandos` recién creado (desde la pestaña Environment) verás que tiene una única variable, de nombre `alcance`, en la que se registra el radio de alcance medido para cada uno de los 100 mandos analizados.

2. Estadística descriptiva

Como punto de partida de nuestro estudio, realizaremos un análisis exploratorio de nuestros datos, incluyendo:

- El cálculo de la media muestral $\bar{x}$, como estimación puntual de la media poblacional μ.
- El cálculo de otras medidas descriptivas, tales como la desviación típica muestral y los cuartiles.
- La representación gráfica de los datos mediante un histograma y un diagrama de caja y bigotes.

2.1. Media muestral

Una forma natural de estimar el valor de la media teórica μ es calcular la media aritmética o promedio del radio de alcance de los 100 mandos analizados, que calculamos con la instrucción:

```
mean(mandos$alcance)
```

```
## [1] 48.8852
```

La expresión `mandos$alcance` extrae la variable `alcance` de la hoja de datos `mandos` y la función `mean()` calcula el promedio del vector resultante.

En general, dadas n observaciones x_1, x_2, ..., x_n de una variable aleatoria continua X la cantidad

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

se llama **media muestral** y proporciona una estimación puntual de la **media poblacional** $\mu = E(X)$.

La media muestral $\overline{x} = 48.8852$ de nuestros datos nos proporciona una estimación de μ, pero hay que tener claro que el verdadero valor de μ sigue siendo desconocido:

- El verdadero valor de μ es la media o valor esperado de la variable aleatoria X. Según hemos estudiado en la teoría, identificamos μ con el valor medio teórico del radio de alcance de todos los mandos fabricados por la empresa, con el valor límite al que tendería el alcance promedio

observado para n mandos cuando n tiende a infinito.

- Por su parte, el valor $\overline{x} = 48.8852$ es tan sólo el radio de alcance promedio de los $n = 100$ mandos que hemos observado.

Podría ser que $\mu = 53$, en cuyo caso los 100 mandos observados tendrían un radio de alcance promedio inferior a la media teórica μ. Y de la misma forma podría ser $\mu = 45$, y que hayamos observado por azar 100 mandos con radio de alcance promedio superior a la media teórica μ.

2.2. Varianza muestral y desviación típica muestral

Dadas n observaciones x_1, x_2, ..., x_n de una variable aleatoria continua X, su **varianza muestral** se define como

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

y proporciona una estimación puntual de la **varianza poblacional** $\sigma^2 = Var(X)$.

Notar que se divide por $n - 1$ y no por n, por razones técnicas para las que por ahora podemos dar la siguiente explicación informal: En la fórmula de s^2 estamos utilizando la media muestral $\bar{x}$ de las n observaciones x_i, que representan una muestra de la población con media μ. Y conociendo la media muestral, solo $n - 1$ de los n valores x_i pueden variar libremente. Se demuestra que si para calcular s^2 dividiéramos por n, en lugar de $n - 1$, el valor obtenido tiende a subestimar el verdadero valor de la varianza poblacional σ^2.

La **desviación típica muestral** es la raíz cuadrada de la varianza muestral:

$$s = \sqrt{s^2}.$$

Para calcular la desviación típica muestral de nuestros datos usamos la función `sd()`:

```
sd(mandos$alcance)
```

```
## [1] 7.554216
```

2.3. Cuantiles

Dadas n observaciones x_1, x_2, ..., x_n de una variable aleatoria continua X y un valor α entre 0 y 1, se define el **cuantil de orden** α de dichas observaciones

como el valor que, en la secuencia ordenada de valores de menor a mayor, queda situado de forma que divide la muestra en dos partes, dejando por debajo una fracción o proporción α de los valores.

Por ejemplo, el cuantil de orden 0.1, posicionado en la secuencia ordenada de valores, dejaría al 10% de las observaciones por debajo y al 90% restante por arriba.

Como casos particulares de cuantiles tenemos los **cuartiles**:

- El **primer cuartil** es el cuantil 0.25, esto es, el valor que deja por debajo el 25% de los datos (y por arriba el 75% restante).
- El **segundo cuartil** o **mediana** es el cuantil 0.25, o lo que es lo mismo, el valor que deja por debajo el 50% de los datos (y por arriba el otro 50%).
- Y el **tercer cuartil** es el cuantil 0.75, es decir el valor que deja por debajo el 75% de los datos (y por arriba el 25% restante).

Calculamos el primer cuartil, la mediana, y el tercer cuartil de nuestra muestra con la siguiente instrucción:

```
quantile(mandos$alcance, probs = c(0.25, 0.5, 0.75))
```

```
##    25%    50%    75%
## 44.545 49.520 54.660
```

El primer cuartil resulta ser 44.545, la mediana 49.52, y el tercer cuartil 54.66. Veamos por ejemplo cómo se ha obtenido el valor 49.52 para la mediana, que deja el 50% de los datos por debajo y la otra mitad por arriba. En primer lugar ordenamos nuestros datos, de menor a mayor, con la orden:

```
sort(mandos$alcance)
```

```
##   [1] 28.26 29.78 35.32 36.08 36.39 36.74 36.82 37.01 37.13
37.24 38.65 39.07
##  [13] 39.88 40.85 40.85 41.10 41.25 41.37 42.36 42.60 43.15
43.26 43.38 44.04
##  [25] 44.47 44.57 44.58 44.67 44.82 45.19 45.23 45.37 45.39
45.68 45.76 45.90
##  [37] 45.99 46.04 46.07 46.50 46.56 46.99 47.12 48.06 48.14
48.32 48.79 48.88
##  [49] 49.35 49.36 49.68 49.94 49.96 49.97 50.12 50.32 50.52
50.53 50.56 50.58
```

```
##  [61] 50.76 50.83 51.04 51.30 51.74 52.19 52.37 53.50 53.58
53.94 54.22 54.31
##  [73] 54.42 54.59 54.59 54.87 55.05 55.06 55.23 55.39 55.52
55.58 56.12 56.68
##  [85] 56.79 56.90 56.98 56.99 57.01 57.35 57.59 58.78 58.96
60.17 60.18 60.57
##  [97] 60.85 62.72 65.00 66.24
```

En la salida del comando anterior verás que:

- El dato que ocupa la posición 50 en la secuencia de valores ordenados, es 49.36 (recuerda que tienes que fijarte en los números entre corchetes al principio de cada línea de la salida para saber qué posición ocupa cada dato). Este valor deja 49 valores por debajo, y 50 por arriba.
- Y el dato que ocupa la posición 51, es 49.68. Deja 50 valores por debajo y 49 por arriba.

Así, ni 49.36 (posición 50), ni 49.68 (posición 51), dejan justo la mitad de los datos por debajo y la otra mitad por arriba. La posición ideal para la mediana sería la posición 50.5. Por esta razón, el cálculo que hace R para obtener la mediana es

$$\frac{49.36 + 49.68}{2} = 49.52,$$

la media aritmética entre el dato de la posición 50 y el de la posición 51.

2.4. La función `summarize()`

La siguiente instrucción utiliza la función `summarize()` para obtener todas las medidas descriptivas que se han calculado de forma individual en los apartados anteriores:

```
summarize(
  .data = mandos,
  "Media" = mean(alcance),
  "Desv. estandar" = sd(alcance),
  "1er cuartil" = quantile(alcance, 0.25),
  "Mediana" = quantile(alcance, 0.5),
  "3er cuartil" = quantile(alcance, 0.75)
)
```

```
## # A tibble: 1 x 5
##   Media `Desv. estandar` `1er cuartil` Mediana `3er cuartil`
##   <dbl>            <dbl>         <dbl>   <dbl>         <dbl>
## 1  48.9             7.55          44.5    49.5          54.7
```

El valor `mandos` del primer argumento `.data` indica la hoja de datos en la que se encuentran las variables de interés. En el resto de argumentos se especifican los estadísticos para resumir los datos que se quieren calcular. Por ejemplo, el argumento `"Media" = mean(alcance)` en nuestro ejemplo solicita calcular la media de la variable `alcance` y le asigna el nombre `"Media"`. Las variables de la hoja de datos pueden usarse escribiendo solo su nombre (escribiendo solo `alcance` en lugar de `mandos$alcance`). La salida se organiza como una nueva hoja de datos, con una columna con el valor de cada estadístico pedido, rotulada con el nombre que se haya indicado.

Nota: La función `kable()` del paquete `knitr` imprime hojas de datos, como la que devuelve el comando `summarize()`, en un formato de tabla adecuado para el documento de salida. Aplicada a la salida de la instrucción anterior quedaría:

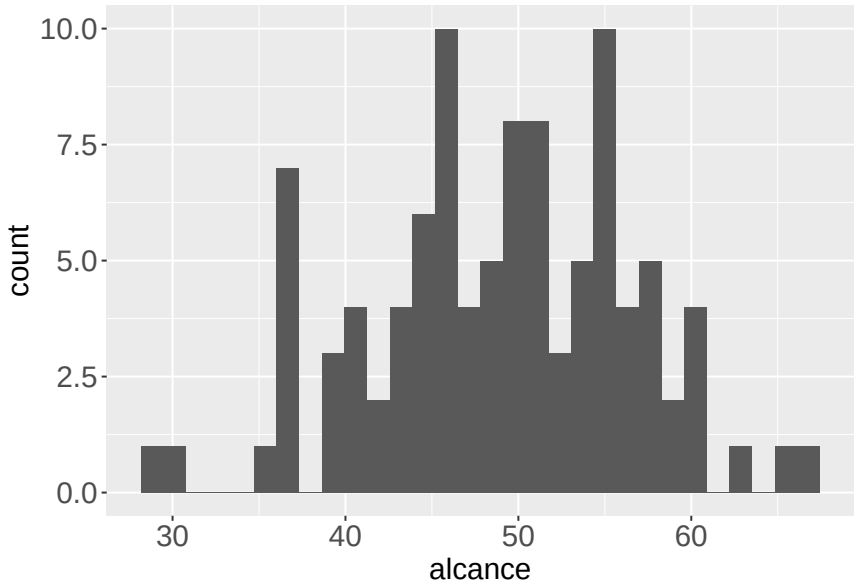
Media	Desv. estandar	1er cuartil	Mediana	3er cuartil
48.8852	7.554216	44.545	49.52	54.66

Si quieres reproducir la tabla anterior en tu documento R Markdown, guarda la salida de la función `summarize()` en la instrucción anterior en un objeto, pongamos de nombre `summary`, y escribe después `knitr::kable(summary)`.

2.5. Histograma

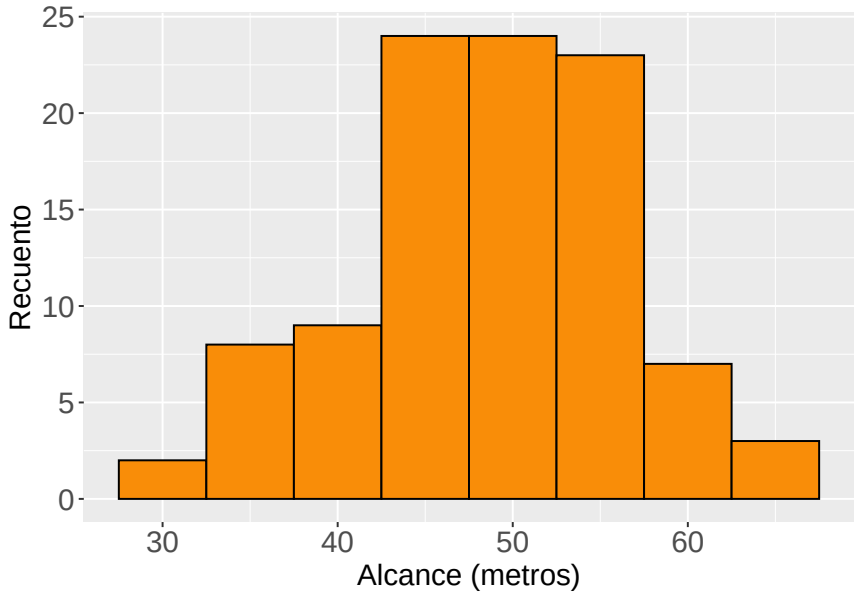
Los **histogramas** son uno de los gráficos más comunes para representar una variable continua. Para obtener un histograma del alcance de los mandos en nuestra muestra utilizamos la función `geom_histogram()`:

```
ggplot(
  data = mandos,
  mapping = aes(x = alcance)
) +
  geom_histogram()
```



Con el siguiente código se mejora el histograma anterior usando intervalos de anchura 5 (`binwidth=5`), personalizando el color de las barras y añadiendo rótulos a los ejes:

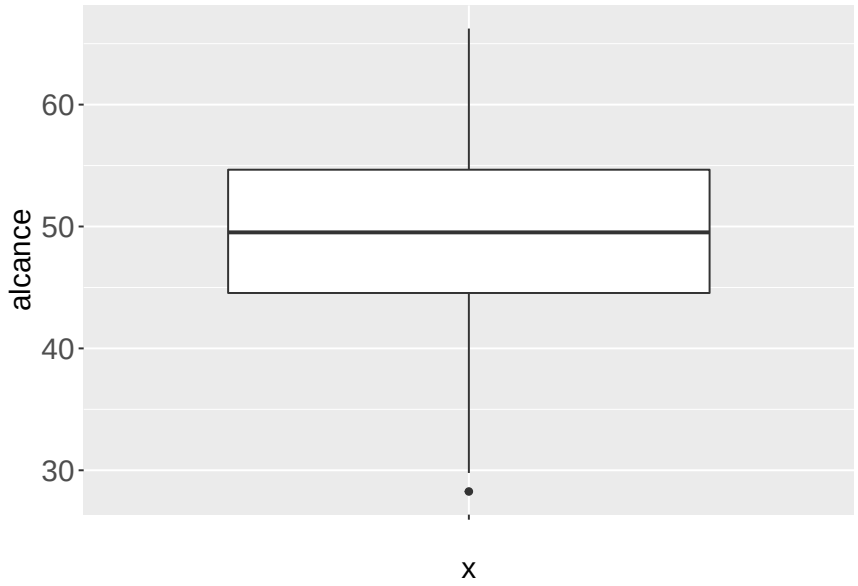
```
ggplot(
  data = mandos,
  mapping = aes(x = alcance)
) +
  geom_histogram(
    binwidth = 5,
    color = "black",
    fill = "#F98D08"
  ) +
  labs(
    x = "Alcance (metros)",
    y = "Recuento"
  )
```



2.6. Diagrama de caja y bigotes

Ahora vamos a representar las observaciones de la variable alcance mediante un gráfico que se denomina **diagrama de caja y bigotes**, usando la función `geom_boxplot()`:

```
ggplot(
  data = mandos,
  mapping = aes(
    x = "",
    y = alcance
  )
) +
  geom_boxplot()
```



Pasamos ahora a explicar cómo se construye e interpreta el diagrama de caja y bigotes que hemos dibujado.

2.6.1. Caja

La caja central del diagrama de caja y bigotes se construye con los valores de los cuartiles: la línea central se corresponde con la mediana y las líneas que delimitan la caja con el primer y tercer cuartil. De esta forma, la caja central contiene el 50% de los datos: el 25% en la mitad inferior de la caja, desde el primer cuartil hasta la mediana; y el 25% en la mitad superior, entre la mediana y el tercer cuartil.

2.6.2. Bigotes y outliers

Queda por explicar qué significan los **"bigotes"**, que son las líneas verticales por encima y debajo de la caja central, y por qué aparece resaltado un valor por debajo del bigote inferior.

El propósito de los bigotes es resaltar los datos extremos, muy pequeños, o muy grandes, que se denominan **outliers**. Datos por debajo del bigote inferior son catalogados como outliers por ser atípicamente pequeños, y datos por encima del bigote superior son considerados outliers por atípicamente grandes. Vemos que en nuestra muestra ha aparecido un oulier por debajo del bigote inferior, que, mirando la escala del eje Y, vemos que tiene un valor menor que 30. Concretamente, se trata del valor mínimo 28.26 que apareció en primer lugar

cuando ordenamos los datos de menor a mayor, y que podemos ver directamente calculando el mínimo:

```
min(mandos$alcance)
```

```
## [1] 28.26
```

Los datos extremos u outliers pueden ser datos erróneos (debidos a errores en las mediciones, en la transcripción de los datos al fichero ...). Pero también pueden ser datos correctos que, aun teniendo poca probabilidad de aparecer, han aparecido por azar en nuestra muestra. Aunque es una práctica frecuente desechar los outliers sistemáticamente, no es en absoluto una práctica recomendable. De hecho los outliers pueden ser lo más interesante de la muestra. La historia más famosa sobre las posibles consecuencias de la eliminación automática de outliers está relacionada con la detección del agujero de ozono. En 1985 se publicó el estudio mostrando que los niveles de ozono en la Antártida habían caído un 10% por debajo de lo normal. Se descubrió entonces que las mediciones inusualmente bajas del nivel de ozono ya habían sido registradas por el satélite Nimbus-7 de la NASA en 1976, pero que dichas mediciones fueron ignoradas al ser procesadas mediante un programa informático que descartaba automáticamente los valores excesivamente pequeños, como si se tratara de errores. De no ser por este tratamiento inadecuado de los outliers, el agujero de ozono podría haberse detectado casi una década antes.

Veamos cómo se dibujan los bigotes para decidir qué datos son outliers. En primer lugar se calcula la diferencia entre tercer cuartil y primer cuartil (que es la anchura de la caja y se denomina **rango intercuartílico**) y se multiplica por 1.5. En nuestro caso, esta cantidad queda

$$1.5(54.66 - 44.545) = 15.1725.$$

Se catalogan como outliers aquellos valores que disten de los bordes de la caja central más de la cantidad anterior. En nuestro caso, serán outliers los valores inferiores a

$$44.545 - 15.1725 = 29.3725$$

y los superiores a

$$54.66 + 15.1725 = 69.8325.$$

El bigote inferior se extiende hasta el menor dato que no es considerado outlier. Y el bigote superior se extiende hasta el mayor dato que no es considerado outlier. Los datos por debajo y por arriba de los bigotes se clasifican como outliers.

Miremos de nuevo los primeros y últimos valores en la secuencia ordenada para nuestros 100 datos:

```
sorted <- sort(mandos$alcance)
head(sorted)
tail(sorted)
```

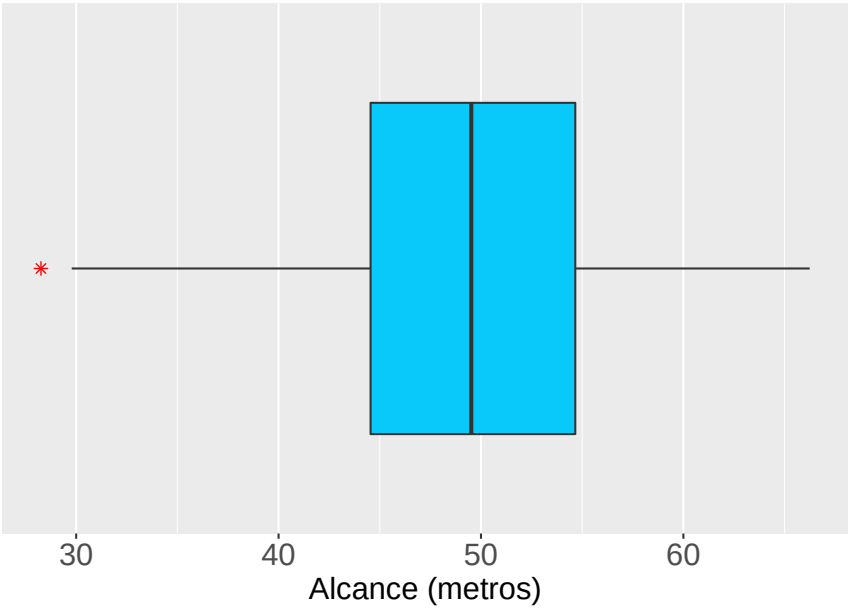
```
## [1] 28.26 29.78 35.32 36.08 36.39 36.74
## [1] 60.18 60.57 60.85 62.72 65.00 66.24
```

El único dato catalogado como outlier es 28.26, por eso el bigote izquierdo se extiende hasta 29.78 (el mínimo no outlier) y el bigote derecho hasta 66.24 (el máximo, no outlier).

2.6.3. Personalización

El siguiente código incluye varias opciones para personalizar el aspecto del diagrama de caja y bigotes creado al comienzo.

```
ggplot(
  data = mandos,
  mapping = aes(
    x = "",
    y = alcance
  )
) +
  geom_boxplot(
    fill = "#08C9F9",
    outlier.colour = "red",
    outlier.size = 2,
    outlier.shape = 8
  ) +
  labs(
    x = "",
    y = "Alcance (metros)"
  ) +
  scale_x_discrete(breaks = NULL) +
  coord_flip()
```



3. Contraste de hipótesis

Hasta ahora hemos descrito los datos numérica y gráficamente. Retomamos ahora el siguiente problema de inferencia estadística planteado al inicio: A nivel de significación 0.05 ¿puede afirmarse que los datos recogidos contradicen la afirmación de la empresa de que el radio de alcance medio de sus mandos es superior a 50 metros?

Según la empresa $\mu > 50$. Sin embargo, la estimación de la media poblacional μ que proporcionan nuestros datos es la media muestral

$$\overline{x} = 48.8852,$$

sugiriendo la hipótesis alternativa $\mu < 50$. Planteamos en consecuencia el siguiente contraste de hipótesis para la media μ:

$$\begin{cases} H_0: & \mu = 50 \\ H_1: & \mu < 50 \end{cases}$$

Como ya sabemos, la hipótesis H_0 se denomina hipótesis nula y la hipótesis H_1 hipótesis alternativa.

Conocemos la función `binom.test` para calcular el p-valor asociado a un contraste de hipótesis para una proporción. La función de R que usaremos

ahora para resolver nuestro contraste de hipótesis sobre una media y calcular el p-valor asociado es `t.test`. Concretamente, tenemos que ejecutar el siguiente código:

```
t.test(
  x = mandos$alcance,
  mu = 50,
  alternative = "less"
)
```

```
##
##  One Sample t-test
##
## data:  mandos$alcance
## t = -1.4757, df = 99, p-value = 0.07159
## alternative hypothesis: true mean is less than 50
## 95 percent confidence interval:
##     -Inf 50.1395
## sample estimates:
## mean of x
##   48.8852
```

En la salida anterior vemos que el p-valor de nuestro contraste es 0.07.

Recordemos que, en general, el p-valor de un contraste es la probabilidad de observar valores tan extraños como en la muestra si la hipótesis nula fuera cierta, entendiendo por valores extraños valores más a favor de la hipótesis alternativa que de la nula. En este caso, el p-valor 0.07 se interpreta como la probabilidad de observar una media muestral tan pequeña como $\overline{x} = 48.89$ si H_0 fuera cierta, es decir, si $\mu = 50$.

Recordemos también que el nivel de significación $\alpha = 0.05$ fija el valor umbral con el que comparar el p-valor para tomar la decisión final en el contraste de hipótesis, de acuerdo con la siguiente regla: Si el p-valor es inferior a 0.05 rechazamos la hipótesis nula en favor de la alternativa, concluyendo que los datos arrojan suficiente evidencia para afirmar que $\mu < 50$ y contradicen por tanto la afirmación de la empresa. Si por el contrario el p-valor es superior a 0.05 no rechazamos la hipótesis nula, y decimos que no hay suficiente evidencia para contradecir la afirmación de la empresa.

Como en nuestro caso hemos obtenido un p-valor de 0.07, mayor que $\alpha = 0.05$, no rechazamos la hipótesis nula, y concluimos que nuestros datos no arrojan

suficiente evidencia para contradecir la afirmación de la empresa.

4. Intervalo de confianza

En la salida del comando `t.test()` encontramos también un intervalo de confianza del 95%:

$$(-\infty, 50.1395).$$

Como sabemos, un intervalo de confianza del 95% para un parámetro, incluye los valores del parámetro que no serían rechazados en un contraste de hipótesis a nivel de significación 0.05.

En este caso, el extremo derecho del intervalo de confianza obtenido, 50.1395, nos da el mayor valor aceptable para μ a la vista de nuestro datos. Así, si bien no hemos encontrado suficiente evidencia en nuestros datos para afirmar, con un 95% de confianza, que $\mu < 50$, sí nos permitirían concluir que $\mu < 55$ (y que μ es menor que cualquier valor superior a 50.1395).