Ejercicio Práctico

Jose Antonio Lorencio Abril

9/12/2021

1 Introducción

En este documento voy a detallar el proceso de comparación entre dos muestras de datos, siguiendo las indicaciones de los guiones de prácticas. Las muestras a estudiar constan de la cantidad de peces pescados en el Mediterráneo por España y por Grecia, desde el año 2019 hasta el 2019. Los datos han sido obtenidos de la web de Eurostat, el organismo de estadísticas oficiales de la Unión Europea [1], concretamente de [2].

Así, vamos a analizar los siguientes datos:

```
##      Grecia    España
## 1  80048.30 105820.23
## 2  68817.69 100288.34
## 3  61757.76 103505.00
## 4  59590.06  78985.11
## 5  62733.05  82999.00
## 6  59589.56  78467.20
## 7  63706.35  76415.77
## 8  74588.32  81774.10
## 9  77348.93  86851.06
## 10 76771.92  87442.76
## 11 82232.46  75928.77
```

Estos datos corresponden a toneladas de peso vivo al ser recogido del mar, excluyendo cualquier producto que, por algún motivo, no se lleva a tierra.

Así, las variables sobre las que vamos a realizar inferencia son:

G = "Toneladas anuales de pesca en el Mediterráneo por barcos griegos"

```
##  [1] 80048.30 68817.69 61757.76 59590.06 62733.05 59589.56 63706.35 74588.32
##  [9] 77348.93 76771.92 82232.46
```

E = "Toneladas anuales de pesca en el Mediterráneo por barcos españoles"

```
##  [1] 105820.23 100288.34 103505.00  78985.11  82999.00  78467.20  76415.77
##  [8]  81774.10  86851.06  87442.76  75928.77
```

Estas dos variables, además, podemos considerarlas independientes, ya que, si bien es cierto que puede haber algunos factores que afecten a la cantidad de pesca de ambos países, es razonable pensar que lo determinante son las demandas internas de cada país.

2 Distribución de las muestras

Comenzamos haciendo inferencia sobre cada una de estas variables por separado.

2.1 Inferencia sobre G

Calculamos la media y la cuasi-varianza muestral:

```
muG = mean(G)
sG = var(G)
muG
```

```
## [1] 69744.04
```

```
sG
```

```
## [1] 74929415
```

Por lo que es

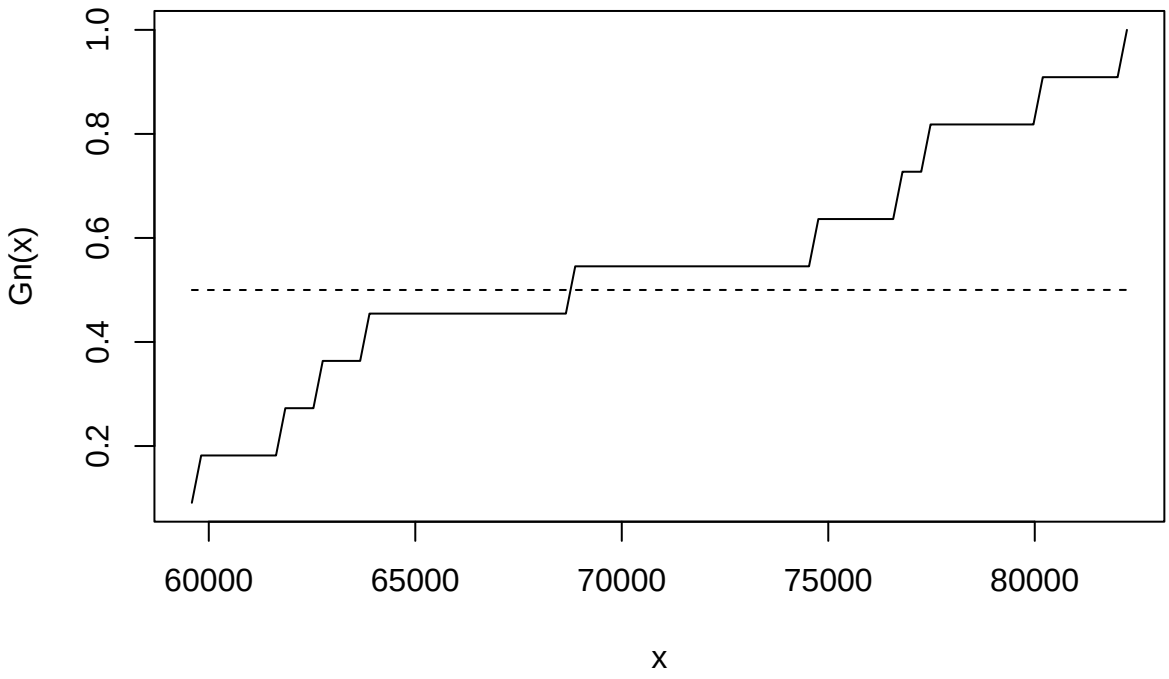
$$\overline{G} = 69744.04$$

$$S^2 = 74929415$$

Y podemos también ver su función de distribución empírica, así como sobreponer la distribución de una normal y visualizar el ajuste:

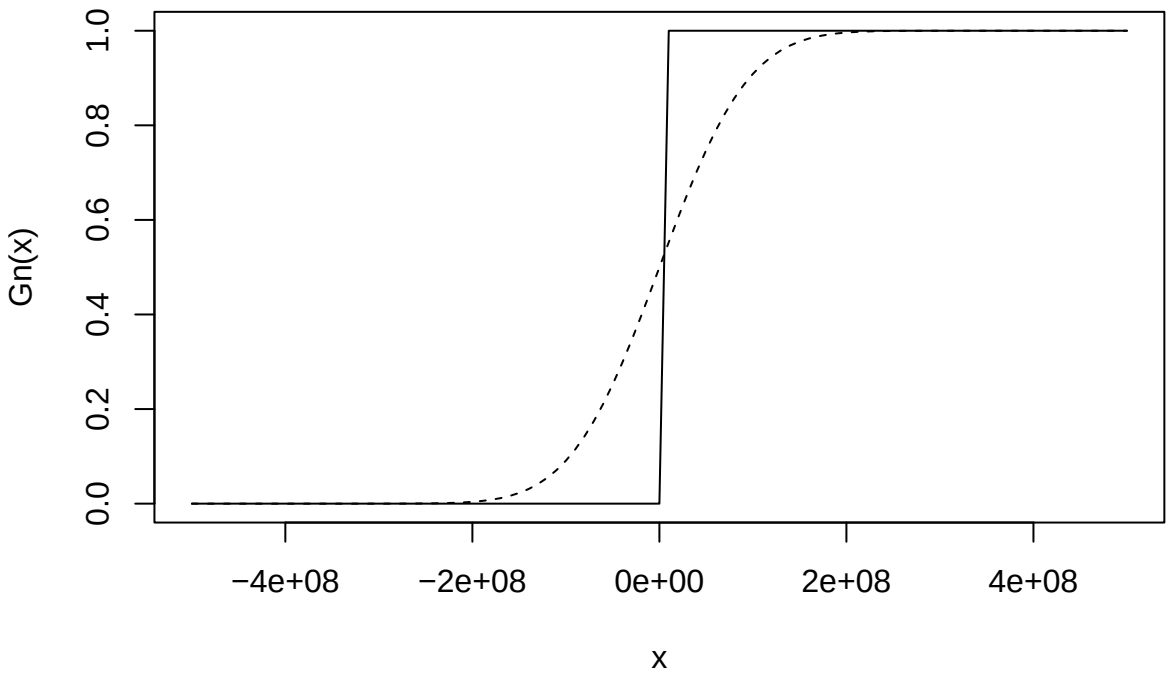
```
mG <- min(G)
MG <- max(G)
Gn <- ecdf(G)

curve(Gn(x), mG, MG)
curve(pnorm(x, m=muG, sd=sG), add=TRUE, lty=2)
```



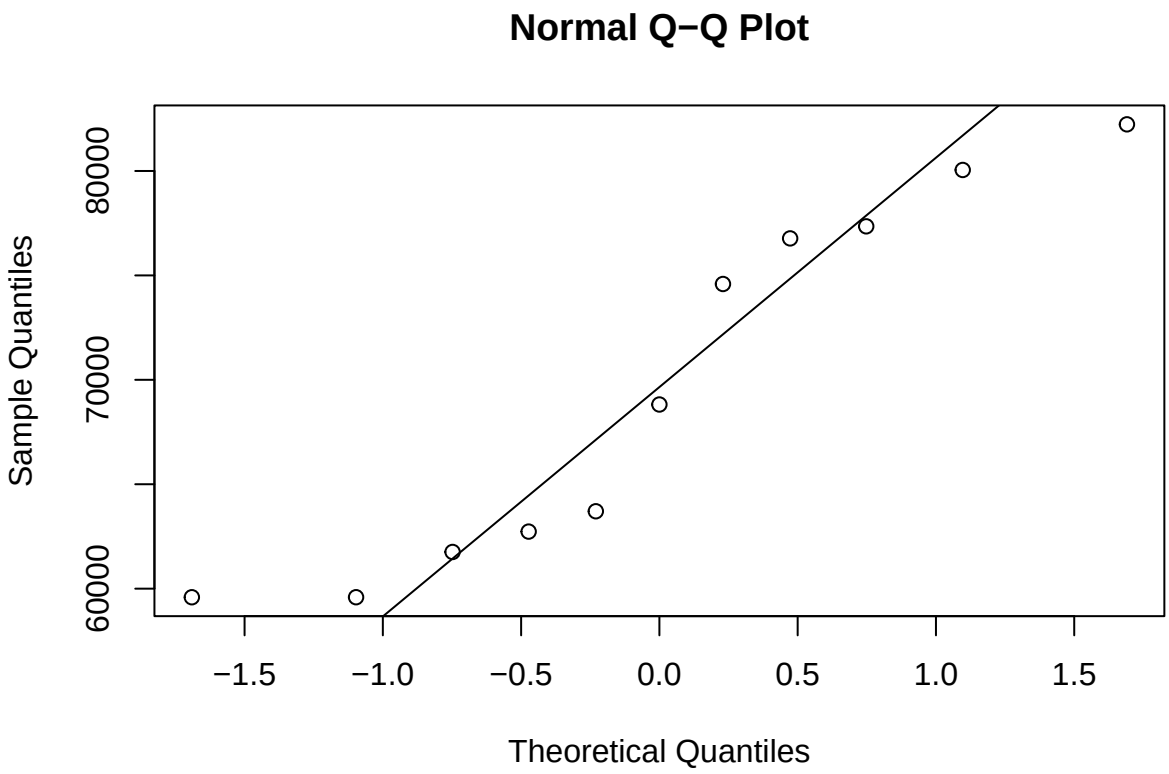
Viendo esta gráfica parece que la distribución no será normal. No obstante, vamos a agrandar el intervalo hasta poder ver la forma de la normal completa:

```
curve(Gn(x), -500000000, 500000000)
curve(pnorm(x, m=muG, sd=sG), add=TRUE, lty=2)
```



Ahora no parece tan descabellado pensar que es una normal, solo que la escala es muy grande. Vamos a visualizar el gráfico Q-Q para indagar un poco más en la normalidad de la variable:

```
qqnorm(G)
qqline(G)
```



Vemos como los cuartiles empíricos no se alejan demasiado de los teóricos, y que es posible que, finalmente, la distribución sí sea normal. Para asegurarnos, vamos a contrastarlo mediante el contraste de Shapiro-Wilk.

Definimos nuestras hipótesis:

$$\begin{aligned} H_0: &\quad G \textit{ sigue una distribución normal} \\ H_1: &\quad G \textit{ no sigue una distribución normal} \end{aligned}$$

Y hacemos el test:

```
shapiro.test(G)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  G
## W = 0.8854, p-value = 0.1218
```

Observamos un valor para el estadístico de $W = 008854$ y un p-valor $p = 0.1218 > 0.01$, por lo que aceptamos H_0 y, efectivamente, G sigue una distribución normal.

2.2 Inferencia sobre E

Vamos a repetir el mismo estudio sobre la variable E, que, en principio, debería ser también normal.

Calculamos media y cuasi-varianza muestrales:

```
muE <- mean(E)
sE <- sd(E)
muE
```

```
## [1] 87134.3
```

```
sE
```

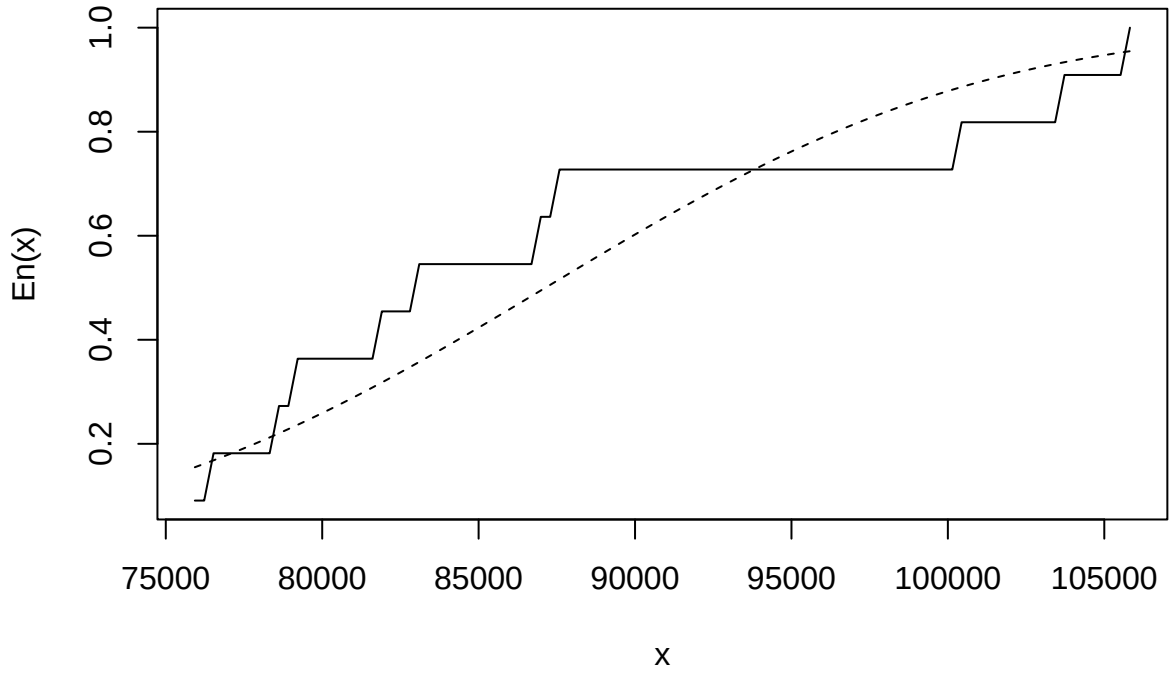
```
## [1] 11037.12
```

Por lo que

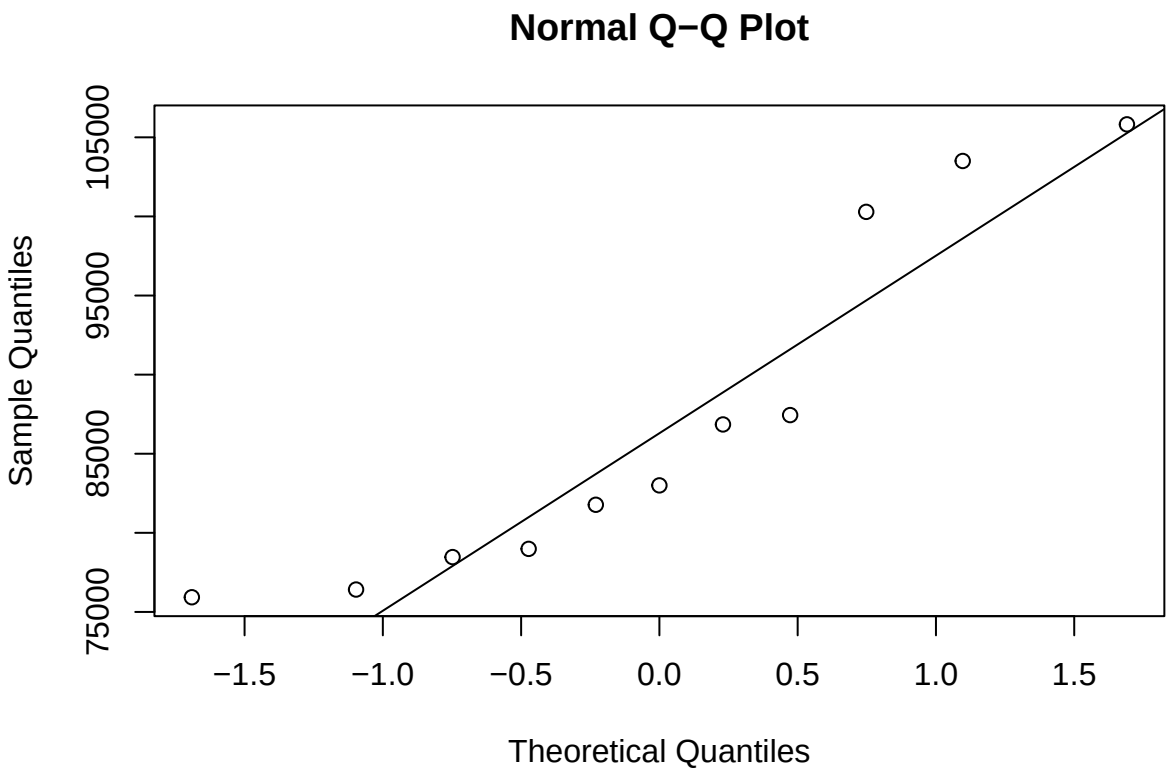
$$\overline{E} = 87134.3$$

$$S^2 = 11037.12$$

Realizamos los mismos plots para visualizar su distribución respecto a la normal con misma media y varianza:




```
qqnorm(E)
qqline(E)
```



Y esta vez se parece más a una normal. Realizamos, no obstante, el test de Shapiro-Wilk para el contraste:

$$H_0: \quad E \text{ sigue una distribución normal}$$
$$H_1: \quad E \text{ no sigue una distribución normal}$$

```
shapiro.test(E)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  E
## W = 0.85209, p-value = 0.04544
```

Obtenemos, así, que el estadístico del constraste vale $W = 085209$ con p-valor $p = 0.04544 > 0.01$, por lo que podríamos aceptar H0 con una significancia de 0.01, y E seguiría una distribución normal.

2.3 Yendo un poco más allá

Hemos obtenido la normalidad satisfactoriamente, pero dada la magnitud de los datos (especialmente llamativa es la varianza de G), podemos tratar de aplicar logaritmos y estudiar la log-normalidad.

Definmos, entonces

$$E_1 = \log E \qquad G_1 = \log G$$

```
G1 = log(G)
G1
```

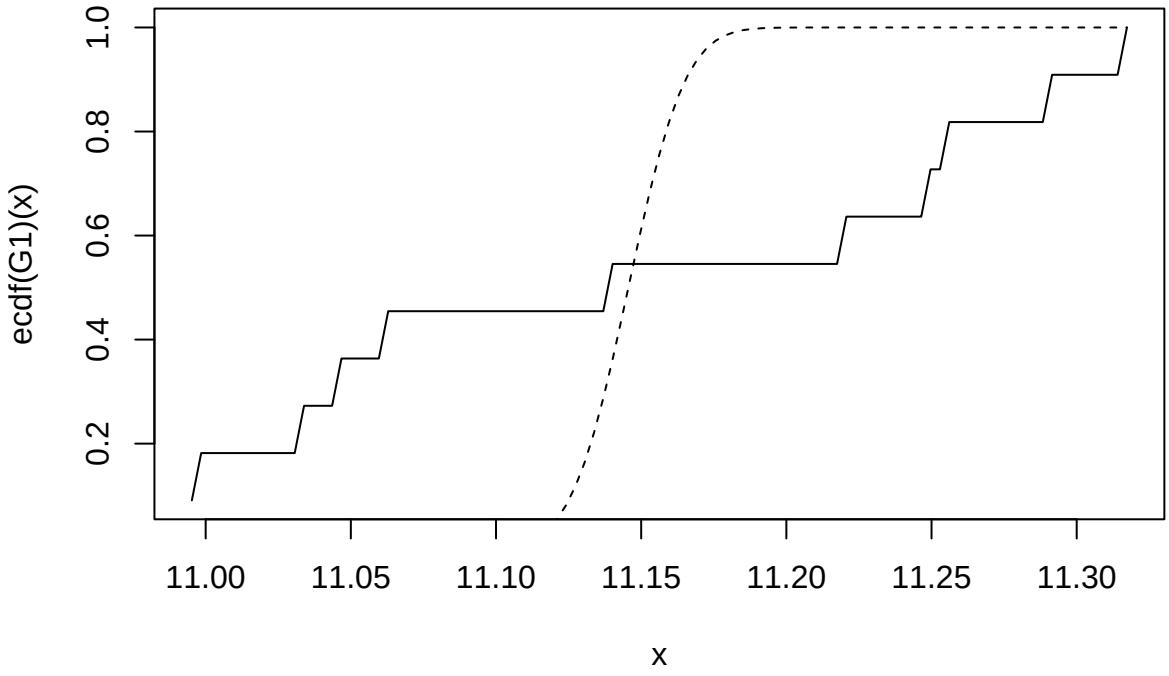
```
##  [1] 11.29039 11.13922 11.03097 10.99524 11.04664 10.99524 11.06204 11.21974
##  [9] 11.25608 11.24859 11.31731
```

```
E1 = log(E)
E1
```

```
##  [1] 11.56950 11.51580 11.54738 11.27701 11.32658 11.27044 11.24394 11.31172
##  [9] 11.37195 11.37874 11.23755
```

Y podemos ver la función de distribución empírica de G1:

```
curve(ecdf(G1)(x), min(G1), max(G1))
curve(pnorm(x, m=mean(G1), sd=var(G1)), add=TRUE, lty=2)
```



Así como el estadístico de Shapiro-Wilk ante el contraste:

$$H_0: \quad G_1 \textit{ sigue una distribución normal}$$
$$H_1: \quad G_1 \textit{ no sigue una distribución normal}$$

```
shapiro.test(G1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  G1
## W = 0.88469, p-value = 0.1193
```

Vemos como obtenemos un estadístico $W = 0.88469$ con un p-valor $p = 0.1193$ y aceptamos H0, de forma que G_1 es normal, y por tanto G es log-normal.

Por último, contrastamos:

$$H_0: \quad E_1 \text{ sigue una distribución normal}$$
$$H_1: \quad E_1 \text{ no sigue una distribución normal}$$

```
shapiro.test(E1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  E1
## W = 0.86838, p-value = 0.0739
```

Y de nuevo, aceptamos H0, y nos creemos que E_1 sigue una distribución normal, y, así, E sigue una log-normal.

A partir de ahora vamos a trabajar con las variables E_1 y G_1, pues son más cómodas y han superado, incluso con mayor significación, los test de normalidad.

3 Inferencia como variables independientes

Como he comentado en la introducción, estas variables son independientes, y así es como las voy a analizar.

En un principio, parece que España pesca más peces que Grecia, por lo que nuestra atención se centrará en ver si esto es cierto, o sea, queremos verificar que

$$\mu_E > \mu_G$$

Primero, contrastamos la igualdad de las varianzas de ambas variables:

$$H_0: \quad \sigma^2_{E_1} = \sigma^2_{G_1}$$
$$H_1: \quad \sigma^2_{E_1} \neq \sigma^2_{G_1}$$

Y lo haremos mediante el contraste de la $\mathcal{F}$ de Snédecor:

```
var.test(G1,E1)
```

```
##
##  F test to compare two variances
##
## data:  G1 and E1
## F = 1.0273, num df = 10, denom df = 10, p-value = 0.9669
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2763876 3.8181677
## sample estimates:
## ratio of variances
##           1.027275
```

Y observamos que nos da un estadístico $F = 1.0273$ con un p-valor $p = 0.9669 > 0.01$, por lo que aceptamos H0 y asumimos la igualdad de las varianzas.

Ahora podemos pasar a la comparación de las medias, teniendo en cuenta que nos encontramos en el caso de igualdad entre varianzas. Definimos el contraste:

$$H_0: \quad \mu_{E_1} \geq \mu_{G_1}$$
$$H_1: \quad \mu_{E_1} < \mu_{G_1}$$

Y hacemos el contraste de la t, indicando la igualdad de las varianzas:

```r
t.test(E1, G1, alternative = "less", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  E1 and G1
## t = 4.2369, df = 20, p-value = 0.9998
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.3132844
## sample estimates:
## mean of x mean of y
##  11.36824  11.14559
```

Nos da un valor para el estadístico $t = 4.2369$ y un p-valor $p = 0.9998 > 0.01$, por lo que aceptamos H0: $\mu_{E_1} \geq \mu G_1$

Por último, si queremos llegar a la conclusión que buscábamos, debemos verificar que no son iguales. Es decir, debemos de hacer el contraste:

$$\begin{aligned} H_0: &\quad \mu_{E_1} = \mu_{G_1} \\ H_1: &\quad \mu_{E_1} \neq \mu_{G_1} \end{aligned}$$

donde, para poder concluir lo que anunciábamos en un principio, debemos rechazar H0. Procedemos con el contraste de la t:

```r
t.test(E1,G1,alternative = "two.sided",var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  E1 and G1
## t = 4.2369, df = 20, p-value = 0.0004043
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1130325 0.3322679
## sample estimates:
## mean of x mean of y
##  11.36824  11.14559
```

Y... ¡estamos de enhorabuena! Hemos obtenido un estadístico con valor $t = 4.2369$ y con p-valor $0.0004043 < 0.01$. Por tanto, rechazamos la igualdad de las medias y, juntando esto con lo anterior, comprobamos que $\mu_{E_1} > \mu_{G_1}$, tal y como esperábamos.

Para terminar, podemos obtener el intervalo de confianza al 99% para la diferencia de las medias, que, según hemos visto, debería estar contenido en $\mathbb{R}^+$:

```r
t.test(E1, G1, conf.level = 0.99, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  E1 and G1
## t = 4.2369, df = 20, p-value = 0.0004043
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  0.07312712 0.37217322
```

```
## sample estimates:
## mean of x mean of y
##  11.36824  11.14559
```

El intervalo obtenido es

$$(0.07313, 0.37217) \subset \mathbb{R}^+$$

contenido en los positivos, como era de esperar.

4 Conclusión

Tras este estudio, podemos concluir que España realizó un mayor volumen de pesca que Grecia en el Mediterráneo durante el período 2009-2019. Podría ser interesante ampliar este estudio a cantidades per cápita, y ver si se mantienen las conclusiones, o la diferencia solo se debe a la mayor población española (aunque a este estudio deberíamos de añadir del lado de España la pesca en el Atlántico, pues el Mediterráneo constituye toda la zona de pesca de Grecia, pero no toda la de España).

5 Bibliografía

1. Eurostat. Official website [Internet]. Available from: https://ec.europa.eu/eurostat/web/main/home

2. Eurostat. Catches in the mediterranean [Internet]. Available from: https://ec.europa.eu/eurostat/databrowser/view/tag00081/default/table?lang=en