Actividad 2: Análisis descriptivo e inferencial

Stwart Porras

2024-10-29

Contents

1 Análsis descriptivo

Analizar gráficamente la variable TEA según:

1.1 Distribución de TEA según el continenete

Se lee el csv como en la práctica anterior. En este caso no es necesario, saltar la primera línea, ya que directamente están los cabeceros de las columnas en su sitio.

```
df <- read.csv("gem02-1.csv", header = TRUE, sep = ',', encoding = "UTF-8")
```

A continuación se muestra in gráfico boxplot, donde se ve la distribución de la variable TEA según continente, para ello, se calculará la media del valor TEA por cada país a lo largo de los años

```
head(df,5)
```

```
##   code       economy year   OPP    PC  FAIL    EI   TEA   OWN EMPL MOT FMTEA
## 1    1 United States 2023 53.81 48.99 44.55 12.09 14.71  6.74   NA  NA  0.84
## 2   27  South Africa 2023 64.10 69.21 59.51  7.45 11.11  5.92   NA  NA  0.76
## 3   30        Greece 2023 45.29 53.76 53.16  9.10  6.74 14.70   NA  NA  0.79
## 4   31   Netherlands 2023 67.37 46.01 40.29 16.02 13.69  6.92   NA  NA  0.80
## 5   33        France 2023 50.86 49.51 40.08 13.41 10.75  4.56   NA  NA  0.72
##   FMOD   JOB INNOV  BSER  STAT  CHOI
## 1   NA 28.80    NA 20.42 78.94 79.21
## 2   NA 21.42    NA  3.57 85.51 78.46
## 3   NA 12.12    NA 22.53 69.76 72.90
## 4   NA 15.13    NA 23.65    NA    NA
## 5   NA 23.17    NA 32.58 51.76 65.41
```

```
library(countrycode)
```

```
## Warning: package 'countrycode' was built under R version 4.4.2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

mean_TEA_data <- df %>%
  group_by(economy) %>%
  summarise(TEA_mean = mean(TEA))

mean_TEA_data <- mean_TEA_data %>%
  mutate(Continentes = countrycode(sourcevar = economy,
                                   origin = "country.name",
                                   destination = "continent"))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Continentes = countrycode(sourcevar = economy, origin =
##   "country.name", destination = "continent")`.
## Caused by warning:
## ! Some values were not matched unambiguously: Kosovo
```

Se comprueba que se hayan asignado correctamente los países.

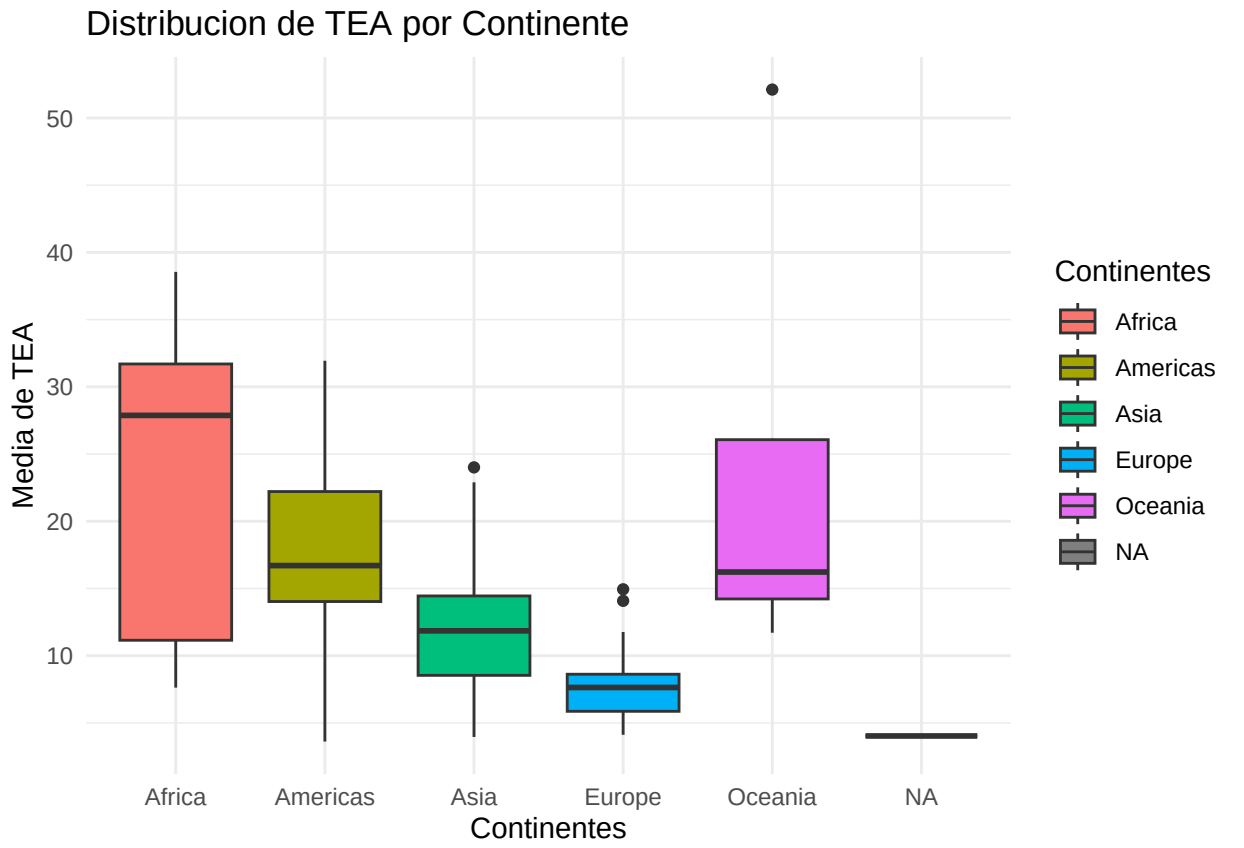
```r
head(mean_TEA_data,5)
```

```
## # A tibble: 5 x 3
##   economy   TEA_mean Continentes
##   <chr>        <dbl> <chr>
## 1 Algeria       9.90 Africa
## 2 Angola       31.6  Africa
## 3 Argentina    14.1  Americas
## 4 Armenia      21.0  Asia
## 5 Australia    11.7  Oceania
```

Se representa el boxplot:

```r
boxplot <- ggplot(mean_TEA_data, aes(x = Continentes, y = TEA_mean, fill =
                                       Continentes)) +
                                       geom_boxplot() + theme_minimal() +

                                labs(title = "Distribucion de TEA por Continente",
                                     x = "Continentes",
                                     y = "Media de TEA")
print(boxplot)
```



Para la representación se ha calculado la media de la variable TEA, (emprendedores entre 18 y 64 años) por continente, para ello se ha calculado la media de esta variable agrupado por países, como valor representativo. Se puede observar en el gráfico que los continentes con mayor TEA, son África y Oceanía, dado que sus rangos inter-cuartílicos son amplios, esto indica que hay una variedad alta en los valores de TEA entre los países. Por otro lado Europa y Asia tiene un valor bajo de media de TEA, pero sus rangos inter-cuartílicos

son más compactos, lo que quiere decir que los valores de TEA entre los países son más consistentes. Hay algunos valores anómalos, lo que quieren decir que dentro de dichos continentes hay algunos países con mayor cantidad de emprendedores. NOTA: nótese que al tener un rango inter-cuartílico menor la presencia de valores atípicos ocurre con mayor facilidad.

2 Análisis de la variabilidad

Se quiere analizar si la variabilidad en los valores promedios de TEA es la misma entre Europa y el resto de países. Pregunta: **¿Los paises de Europa tienen una varianza diferente al resto de países en la variable TEA ?**

Se calculan el sub set de Europa.

```
Europe <- mean_TEA_data %>%
  filter(Continentes == "Europe")

head(Europe,5)
```

```
## # A tibble: 5 x 3
##   economy                TEA_mean Continentes
##   <chr>                     <dbl> <chr>
## 1 Austria                    7.44 Europe
## 2 Belarus                    9.63 Europe
## 3 Belgium                    4.12 Europe
## 4 Bosnia And Herzegovina     7.35 Europe
## 5 Bulgaria                   4.5  Europe
```

Resto del mundo:

```
Rest_of_World <- mean_TEA_data %>%
  filter(Continentes != "Europe")

head(Rest_of_World,5)
```

```
## # A tibble: 5 x 3
##   economy   TEA_mean Continentes
##   <chr>        <dbl> <chr>
## 1 Algeria       9.90 Africa
## 2 Angola       31.6  Africa
## 3 Argentina    14.1  Americas
## 4 Armenia      21.0  Asia
## 5 Australia    11.7  Oceania
```

Ahora comprobamos la cantidad de datos que tenemos en cada subset, esto es debido a que, para que el test sea más fiable probablemente haya que hacer un under sampling, de el df Rest of World, debido a que lógicamente, habrán más datos.

```
print(nrow(Europe))
```

```
## [1] 35
```

```
print(nrow(Rest_of_World))
```

```
## [1] 79
```

Para que en cada ejecución, no nos de un resultado “diferente”, para realizar el under sampling, vamos a hacer uso de la funcion sample, y 35 valores, en otros casos sería neceserio utilizar una semilla, o seed, para los valores aleatorios

```
Rest_of_World <- Rest_of_World[sample(nrow(Rest_of_World), 35), ]
```

```
nrow(Rest_of_World)
```

```
## [1] 35
```

Ahora que tenemos el mismo número de datos para cada subset, se procede a aplicar el test de Levene.

2.1 Test de Levene

El test de Levene es una prueba de estadística inferencial, que se utiliza para evaluar la igualdad de las varianzas entre 2 o más grupos. Su uso es debido a que en algunos test como t-student, se asume que las varianzas poblacionales de las muestras son iguales. La hipótesis nula H_0 es que no hay diferencia entre las varianzas de los grupos, y la hipótesis alternativa H_1 indica que al menos en uno de los grupos la varianza es diferente.

En este caso la pregunta es directamente si la varianza entre los dos grupos es diferente, por lo que sólo se hará uso del test de Levene en esta ocasión

2.1.1 Definición del Test de Levene

El test de Levene se define según:

$$L = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} N_i (Z_i - Z..)^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Zi.)^2}$$

donde:

- L es el resultado de la prueba
- k es el número de diferentes grupos a los que pertencen los valores
- N es el número total de casos en todos los grupos
- N_i es el número de casos en el grupo i
- Y_{ij} es el valor de la variable medida para el jésimo caso del iésimo grupo
- Y_i^- es la media o la mediana del iésimo grupo
- Z.. $= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$ es la medida de Z_{ij}
- $Z_i. = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij}$ es la media de Z_{ij} para el grupo i.

A continuación se procede a hacer los cálculos pertinentes para el test. Se agrupan los datos en un solo dataframe de 70 valores, 35 europa, 35, del resto del mundo.

```
Europe_ROW <- data.frame(
  group = c(rep("Europe", nrow(Europe)), rep("Rest of the World",
                                             nrow(Rest_of_World))),
  values = c(Europe$TEA_mean, Rest_of_World$TEA_mean)
)
head(Europe_ROW,5)
```

```
##    group  values
## 1 Europe 7.44125
## 2 Europe 9.63000
## 3 Europe 4.11800
## 4 Europe 7.34750
## 5 Europe 4.50000
```

```
tail(Europe_ROW,5)
```

```
##                group    values
## 66 Rest of the World  8.618889
## 67 Rest of the World 14.730000
## 68 Rest of the World  9.600000
## 69 Rest of the World 38.550000
## 70 Rest of the World 11.709231
```

Mediana del conjunto de datos agrupados:

```
median_all <- tapply(Europe_ROW$values, Europe_ROW$group, median)
print(median_all)
```

```
##            Europe Rest of the World
##          7.636842         12.691304
```

Con esto ya se puede observar que los centros van a ser diferentes. Desviaciones absolutas.

```
Europe_ROW$abs_dev <- abs(Europe_ROW$values - median_all[Europe_ROW$group])
head(Europe_ROW)
```

```
##    group   values   abs_dev
## 1 Europe 7.441250 0.1955921
## 2 Europe 9.630000 1.9931579
## 3 Europe 4.118000 3.5188421
## 4 Europe 7.347500 0.2893421
## 5 Europe 4.500000 3.1368421
## 6 Europe 8.127273 0.4904306
```

Se calcula la media de los dos grupos Europa y el Resto del Mundo, y su media en conjunto

```
means_all <- tapply(Europe_ROW$abs_dev, Europe_ROW$group, mean)
overall_mean <- mean(Europe_ROW$abs_dev)
```

```
group_sizes <- table(Europe_ROW$group) # para ambos hay 35 valores
```

```
group_sizes
```

```
##
##            Europe Rest of the World
##                35                35
```

```
numerator <- sum(group_sizes * (means_all-overall_mean)^2)
denominator <- sum((Europe_ROW$abs_dev - means_all[Europe_ROW$group])^2)
```

Se aplica la fórmula

```
N <- nrow(Europe_ROW)
k <- length(unique(Europe_ROW$group))
L <- ((N-k)*numerator) / ((k-1) * denominator)

df1 <- k-1
df2 <- N-k
p_value <- 1-pf(L,df1,df2) # La función se usa para calcular la distribución
                           # acumulada de la distribción F

cat("L's value: ", L, "\n")
```

```
## L's value:  16.25763
```

```r
cat("p-value: ",p_value, "\n")
```

```
## p-value:  0.0001422295
```

En este caso el valor de p se calcularía según:

$$p - value = 1 - P(X \leq L)$$

donde $P(X \leq L$ la función de distribución acumulada. Más información

Ahora comprobamos que efectivamente obtenemos el mismo resultado, con la librería de Rstudio, para el test de Levene

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

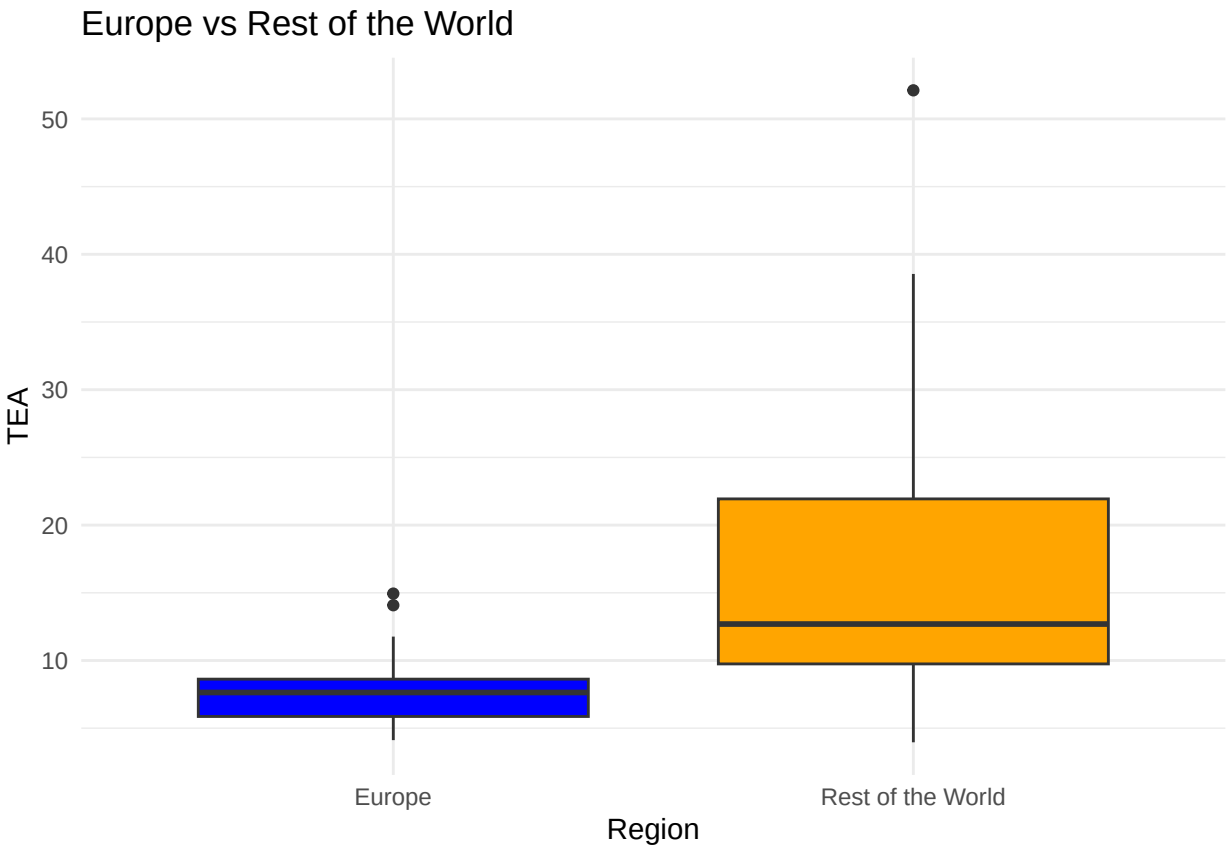
```r
leveneTest(values ~ group, data = Europe_ROW, center = median)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

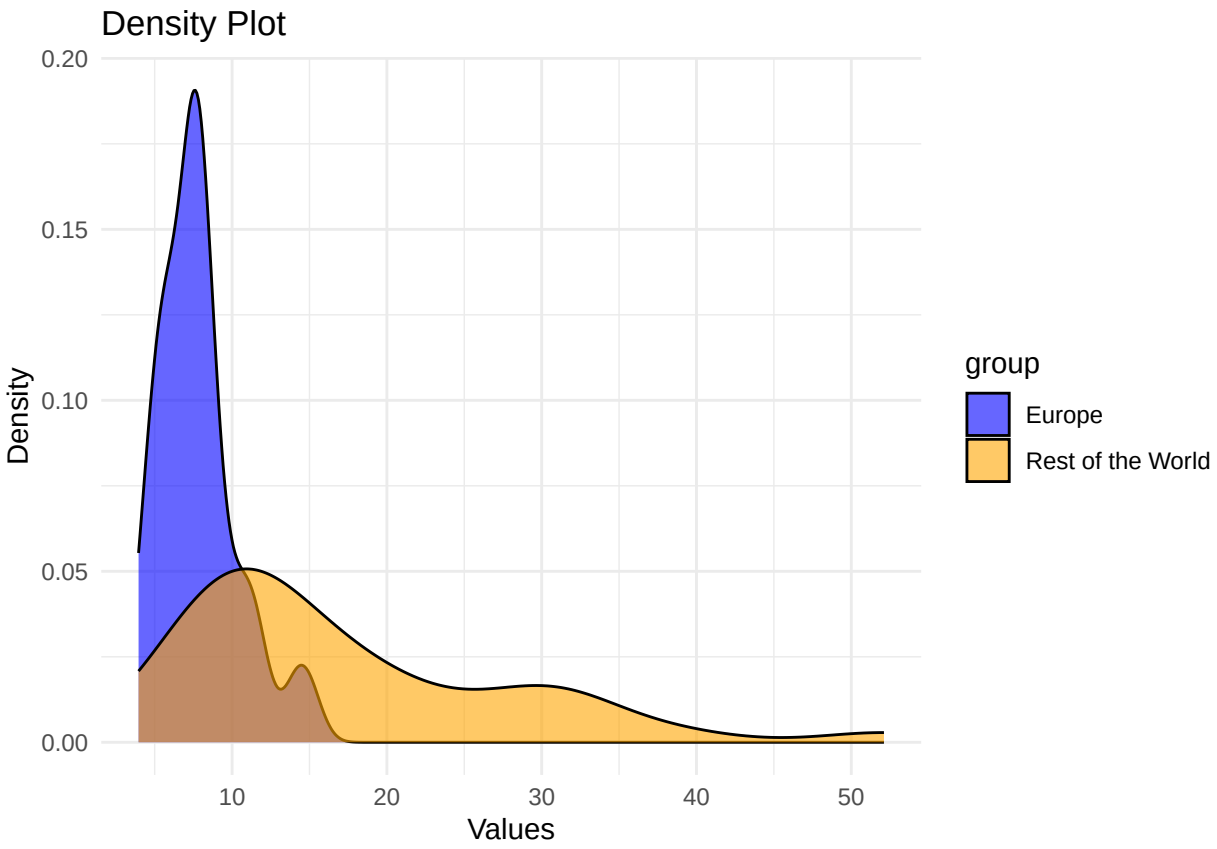
```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value    Pr(>F)
## group  1  16.258 0.0001422 ***
##       68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ya que el nivel de confianza es del 95%, es decir, $\alpha = 0.05$, como el valor de p es inferior 0.05, se rechaza la hipótesis nula, a favor de la hipótesis alternativa, por lo que la varianza de Europa, con respecto al resto del mundo no son iguales.

```r
ggplot(Europe_ROW, aes(x = group, y = values, fill = group)) +
  geom_boxplot() +
  scale_fill_manual(values = c("Europe" = "blue",
                               "Rest of the World" = "orange")) +
  labs(
    title = "Europe vs Rest of the World",
    x = "Region",
    y = "TEA"
  ) +
  theme_minimal() + theme(legend.position = "none")
```




```
ggplot(Europe_ROW, aes(x = values, fill = group)) +
 geom_density(alpha = 0.6) +
 scale_fill_manual(values = c("Europe" = "blue", "Rest of the World" = "orange")) +
 labs(
   title = "Density Plot",
   x = "Values",
   y = "Density"
 ) +
 theme_minimal()
```



3 Análisis de diferencias entre los países de Europa y el resto

En este apartado queremos analizar las diferencias entre los países de Europa y el resto de países en las variables: OPP, PC, FAIL, EI, TEA y OWN. La pregunta de investigación a responder es: ** ¿Los países de Europa tienen medias inferiores que el resto de países en los valores de las variables OPP, PC, EI TEA, y OWN?¿Y una media superior al resto en la variable FAIL? **

En este apartado se nos pide comparar las medias de las variables, no solo si es diferente si no si una comparación entre mayor y menor. Para ello se hará uso del test t-student que compara las medias entre varios grupos, pero, como se mecnionó anteriormente, t-student asume, que las varianzas son iguales, y que los conjunto de datos, siguen más o menos una distribución normal. En esta ocasión para simplificar el documento, se aplicará directacmente el test de t-student asumiendo lo mencionado anteriormente.

```
calculate_means <- function(df){

  df %>%
    group_by(economy, Continentes) %>%
    summarise(
      OPP_mean = mean(OPP, na.rm = TRUE),
      PC_mean = mean(PC, na.rm = TRUE),
      EI_mean = mean(EI, na.rm = TRUE),
      TEA_mean = mean(TEA, na.rm = TRUE),
      OWN_mean = mean(OWN, na.rm = TRUE),
      FAIL_mean = mean(FAIL, na.rm = TRUE),
      .groups = "drop"
    )
```

```
}
```

```
df_3 <- df %>%
  select(economy,OPP,PC,EI,TEA,OWN,FAIL)
```

```
df_3 <- df_3 %>%
  mutate(Continentes = case_when(
    economy == "Kosovo" ~ "Europe",
    TRUE ~ countrycode(sourcevar = economy,
                       origin = "country.name",
                       destination = "continent")
  ))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Continentes = case_when(...)`.
## Caused by warning:
## ! Some values were not matched unambiguously: Kosovo
```

```
head(df_3,10)
```

```
##          economy   OPP    PC    EI   TEA   OWN  FAIL Continentes
## 1  United States 53.81 48.99 12.09 14.71  6.74 44.55    Americas
## 2   South Africa 64.10 69.21  7.45 11.11  5.92 59.51      Africa
## 3         Greece 45.29 53.76  9.10  6.74 14.70 53.16      Europe
## 4    Netherlands 67.37 46.01 16.02 13.69  6.92 40.29      Europe
## 5         France 50.86 49.51 13.41 10.75  4.56 40.08      Europe
## 6          Spain 30.65 53.18  9.63  6.79  6.73 46.22      Europe
## 7        Hungary 28.22 38.29  8.16  9.88  7.38 34.41      Europe
## 8          Italy 33.75 50.85 10.39  8.33  7.75 48.48      Europe
## 9        Romania 55.67 52.62  5.79  5.85  5.09 58.07      Europe
## 10   Switzerland 52.49 44.89  9.96 10.29  5.83 36.35      Europe
```

```
df_3 <- calculate_means(df_3)
```

```
head(df_3,5)
```

```
## # A tibble: 5 x 8
##   economy   Continentes OPP_mean PC_mean EI_mean TEA_mean OWN_mean FAIL_mean
##   <chr>     <chr>          <dbl>   <dbl>   <dbl>    <dbl>    <dbl>     <dbl>
## 1 Algeria   Africa          52.6    55.3    30.3     9.90     4.13      35.6
## 2 Angola    Africa          69.1    67.6    56.0    31.6      8.75      38.8
## 3 Argentina Americas        45.2    59.2    20.7    14.1      9.18      30.9
## 4 Armenia   Asia            53.9    70      32.2    21.0      7.84      48.2
## 5 Australia Oceania         46.8    52.8    11.8    11.7     10.5       38.2
```

```
Europe_3 <- df_3 %>%
  filter(Continentes == "Europe")
ROW_3 <- df_3 %>%
  filter(Continentes != "Europe")
```

```
head(Europe_3)
```

```
## # A tibble: 6 x 8
##   economy       Continentes OPP_mean PC_mean EI_mean TEA_mean OWN_mean FAIL_mean
##   <chr>         <chr>          <dbl>   <dbl>   <dbl>    <dbl>    <dbl>     <dbl>
## 1 Austria       Europe          42.9    50.2    7.31     7.44     7.34      35.5
## 2 Belarus       Europe          27.3    47.2   15.3      9.63     4.13      47.0
```

```
## 3 Belgium       Europe          27.8    36.7    7.26     4.12     3.57      36.1
## 4 Bosnia And H~ Europe          26.9    52.4   18.0      7.35     5.34      27.5
## 5 Bulgaria      Europe          18.9    37.5    5.34     4.5      6.61      27.6
## 6 Croatia       Europe          34.6    55.4   14.7      8.13     3.71      36.0
```

```
head(ROW_3,5)
```

```
## # A tibble: 5 x 8
##   economy   Continentes OPP_mean PC_mean EI_mean TEA_mean OWN_mean FAIL_mean
##   <chr>     <chr>          <dbl>   <dbl>   <dbl>    <dbl>    <dbl>     <dbl>
## 1 Algeria   Africa          52.6    55.3    30.3     9.90     4.13      35.6
## 2 Angola    Africa          69.1    67.6    56.0    31.6      8.75      38.8
## 3 Argentina Americas        45.2    59.2    20.7    14.1      9.18      30.9
## 4 Armenia   Asia            53.9    70      32.2    21.0      7.84      48.2
## 5 Australia Oceania         46.8    52.8    11.8    11.7     10.5       38.2
```

```
ROW_3 <- ROW_3[sample(nrow(ROW_3), 36), ]
```

```
Europe_ROW_3 <- data.frame(
  group = c(rep("Europe", nrow(Europe_3)), rep("Rest of the World",
                                               nrow(ROW_3))),
  TEA_mean = c(Europe_3$TEA_mean, ROW_3$TEA_mean),
  OPP_mean = c(Europe_3$OPP_mean, ROW_3$OPP_mean),
  PC_mean = c(Europe_3$PC_mean, ROW_3$PC_mean),
  FAIL_mean = c(Europe_3$FAIL_mean, ROW_3$FAIL_mean),
  EI_mean = c(Europe_3$EI_mean, ROW_3$EI_mean),
  OWN_mean = c(Europe_3$OWN_mean, ROW_3$OWN_mean)
)
```

```
head(Europe_ROW_3,5)
```

```
##    group TEA_mean OPP_mean  PC_mean FAIL_mean   EI_mean OWN_mean
## 1 Europe  7.44125 42.89125 50.16125  35.48750  7.313750 7.338750
## 2 Europe  9.63000 27.27000 47.15000  46.98500 15.325000 4.130000
## 3 Europe  4.11800 27.81867 36.73333  36.09067  7.261429 3.574667
## 4 Europe  7.34750 26.89875 52.39500  27.45250 17.968750 5.345000
## 5 Europe  4.50000 18.89750 37.52000  27.55250  5.337500 6.610000
```

```
tail(Europe_ROW_3,5)
```

```
##                group TEA_mean OPP_mean  PC_mean FAIL_mean  EI_mean  OWN_mean
## 68 Rest of the World 11.85187 33.08375 59.03500  31.68000 30.34937 10.641250
## 69 Rest of the World 12.77000 64.43000 23.63000  72.01000 24.57000 11.600000
## 70 Rest of the World 13.65143 45.67857 55.62857  31.93571 31.97571  5.257143
## 71 Rest of the World 10.85750 57.44000 59.75688  44.86500 22.70933  8.375625
## 72 Rest of the World 25.74500 71.76500 73.97500  33.61500 48.73000  5.530000
```

Pregunta de investigación: ¿Los países de Europa tienen medias inferiores que el resto de países en los valores de las variables OPP, PC, EI TEA, y OWN?¿Y una media superior al resto en la variable FAIL? Hipótesis nula H_0 : Los países de Europa no tienen la media inferior que el resto de países en las variables mencionadas. Hiótesis alternativa H_1, los países de europa tiene medias inferiores en los valores de las variables mencionadas al resto del mundo.

$$H_0 : \mu_1 \geq \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

3.1 Nivel de confianza

- 95%
- $\alpha = 0.05$

3.2 Inferencia

$$t = \frac{\bar{x_1} - \bar{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t_{n_1+n_2-2}$$

$$S = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

donde: * $\bar{x_i}$ son las medias de los grupos 1, 2.. * S es la varianza respectivamente * n1,n2 son los tamaños de las muestras de cada grupo

Ahora que los datos están preparados como antes, se procederá a realizar el t-student. ## 3.1 T-Student La distribución de t-student es una distribución de estadística inferencial, que surge de estimar la media de una población **normalmente distribuida**. La distribución es sensible al tamaño de la muestra, por lo que se puede utilizar con muestras grandes o pequeñas. También es robusta ante las desviaci- ones de la normalidad, especialmente cuando el tamaño es grande.

Grados de libertad

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2-1}}$$

```
variables <- c("TEA_mean", "OPP_mean", "PC_mean","FAIL_mean", "EI_mean", "OWN_mean")

results <- data.frame(
  Variable = character(),
  t_statistic = numeric(),
  degrees_freedom = numeric(),
  p_value = numeric(),
  stringsAsFactors = FALSE
)
```

```
for(var in variables) {
  europe_values <- Europe_ROW_3[Europe_ROW_3$group == "Europe", var]
  row_values <- Europe_ROW_3[Europe_ROW_3$group == "Rest of the World", var]

  mean_europe <- mean(europe_values)
  mean_row <- mean(row_values)
  var_europe <- var(europe_values)
  var_row <- var(row_values)
  n_europe <- length(europe_values)
  n_row <- length(row_values)

  t_stat <- (mean_europe - mean_row) / sqrt((var_europe / n_europe) +
                                            (var_row / n_row))

  dgf <- ((var_europe / n_europe + var_row / n_row)^2) /
    (( (var_europe / n_europe)^2 / (n_europe-1)) + ((var_row / n_row)^2 /
```

```
                                                    (n_row-1)))
  p_value <- pt(t_stat, dgf, lower.tail = TRUE)

  results <- rbind(results, data.frame(
    Variable = var,
    t_statistic = t_stat,
    degrees_freedom = dgf,
    p_value = p_value
  ))
}

print(results)
```

```
##    Variable t_statistic degrees_freedom      p_value
## 1  TEA_mean  -6.7369210        40.94422 1.953034e-08
## 2  OPP_mean  -4.2353648        63.51672 3.753932e-05
## 3   PC_mean  -4.9806694        53.32211 3.513853e-06
## 4 FAIL_mean   0.5779715        51.03080 7.170874e-01
## 5   EI_mean  -6.6899823        45.13106 1.451277e-08
## 6  OWN_mean  -3.0026010        42.03684 2.246404e-03
```

Se comprueban los resultados con la función t.test de Rstudio.

```
for (var in variables) {
  t_test <- t.test(
    Europe_ROW_3[Europe_ROW_3$group == "Europe", var],
    Europe_ROW_3[Europe_ROW_3$group == "Rest of the World", var],
    alternative = "less",
    conf.level = 0.95
  )
}
```

3.3 Función y Contraste de hipótesis

Para el contraste de hipótesis se asumirá como en el apartado anterior que las varianzas son desconocidas y diferentes

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

```
tstudent_contrast <- function(group1,group2, var_igual = TRUE, alpha = 0.05,
                              type = "bilateral"){
  n1 <- length(group1)
  n2 <- length(group2)

  mean_1 <- mean(group1)
  mean_2 <- mean(group2)
  var_1 <- var(group1)
  var_2 <- var(group2)

  if (var_igual) {
```

```
    S_p <- sqrt(((n1 - 1) * var_1 + (n2 - 1) * var_2) / (n1 + n2 - 2))
    t_obs <- (mean_1 - mean_2) / (S_p * sqrt(1 / n1 + 1 / n2))
    gl <- n1 + n2 - 2
  }
  else{
    t_obs <- (mean_1 - mean_2) / sqrt((var_1 / n1) + (var_2 / n2))
    gl <- ((var_1 / n1 + var_2 / n2)^2) / (((var_1 / n1)^2 / (n1 - 1)) +
                                           ((var_2 / n2)^2 / (n2 - 1)))

  }


  if (type == "bilateral"){
    t_crit <- qt(1-alpha/2, df = gl)
    p_value <- 2 * (1-pt(abs(t_obs), df = gl))
    decision <- abs(t_obs) > t_crit
  } else if(type == "left tail"){
    t_crit <- qt(alpha, df = gl)
    p_value <- pt(t_obs, df = gl)
    decision <- t_obs < t_crit
   } else if(type == "right tail") {
     t_crit <- qt(1 - alpha, df = gl)
     p_Value <- 1-pt(t_obs, df = gl)
     decision <- p_value < alpha

   }
  results <- list(
    t_obs = t_obs,
    t_crit = t_crit,
    p_value = p_value,
    decision = ifelse(decision, "Rechazar H0","No rechazar H0"),
    degree_freedom = gl
  )

  return(results)
}
```

```
resultados <- list()
for (var in variables){
  group1 <- Europe_ROW_3[Europe_ROW_3$group == "Europe", var]
  group2 <- Europe_ROW_3[Europe_ROW_3$group == "Rest of the World", var]

  resultado <- tstudent_contrast(group1,group2, var_igual = FALSE,
                                 alpha = 0.05, type = "left tail")
  resultados[[var]] <- resultado
}

#for (var in variables) {
  #cat("\nVariable:", var, "\n")
  #print(resultados[[var]])
#}
```

Como se puede comprobar en las 3 ocasiones obtenemos los mismo resultados. Tabla resumen:

```r
table_results <- data.frame(
  Variable = character(),
  t_obs = numeric(),
  t_crit = numeric(),
  p_value = numeric(),
  degree_freedom = numeric(),
  Decision = character()
)
```

```r
for (var in variables) {
  resultado <- resultados[[var]]


  table_results <- rbind(table_results, data.frame(
    Variable = var,
    t_obs = round(resultado$t_obs, 3),
    crit_value = round(resultado$t_crit, 3),
    p_value = round(resultado$p_value, 5),
    degree_freedom = round(resultado$degree_freedom, 2),
    Decision = resultado$decision
  ))
}


print(table_results)
```

```
##    Variable  t_obs crit_value p_value degree_freedom       Decision
## 1  TEA_mean -6.737     -1.683 0.00000          40.94    Rechazar H0
## 2  OPP_mean -4.235     -1.669 0.00004          63.52    Rechazar H0
## 3   PC_mean -4.981     -1.674 0.00000          53.32    Rechazar H0
## 4 FAIL_mean  0.578     -1.675 0.71709          51.03 No rechazar H0
## 5   EI_mean -6.690     -1.679 0.00000          45.13    Rechazar H0
## 6  OWN_mean -3.003     -1.682 0.00225          42.04    Rechazar H0
```

En conclusión para las variables TEA, OPP, PC, EI, y OWN, se rechaza la hipóte- sis nula, en cuanto a la primera pregunta. En cuento a la segunda pregunta con la variable FAIL, para este análisis se aceptaría la hipótesis nula, lo que quiere decir, que FAIL de Europa tiene una media superior al resto. En todos los casos las diferencias son significativas si se rechaza la hipótesis nula, y no lo es al aceptarla.

4 Análisis longitudinal

En este apartado, se quiere analizar si ha habido una mejora significativa del porcentaje emprendedor en un período de cinco años. Por este motivo, escogeremos el período 2019-2023 para dar respuesta a la pregunta. Por tanto, la pregunta de investigación es: **PR3: ¿Los valores promedios de TEA de los diferentes países aumentan en cinco años?**

```r
head(df,5)
```

```
##   code       economy year   OPP    PC  FAIL    EI   TEA   OWN EMPL MOT FMTEA
## 1    1 United States 2023 53.81 48.99 44.55 12.09 14.71  6.74   NA  NA  0.84
## 2   27  South Africa 2023 64.10 69.21 59.51  7.45 11.11  5.92   NA  NA  0.76
## 3   30        Greece 2023 45.29 53.76 53.16  9.10  6.74 14.70   NA  NA  0.79
## 4   31   Netherlands 2023 67.37 46.01 40.29 16.02 13.69  6.92   NA  NA  0.80
## 5   33        France 2023 50.86 49.51 40.08 13.41 10.75  4.56   NA  NA  0.72
##   FMOD   JOB INNOV  BSER  STAT  CHOI
```

```
## 1   NA 28.80    NA 20.42 78.94 79.21
## 2   NA 21.42    NA  3.57 85.51 78.46
## 3   NA 12.12    NA 22.53 69.76 72.90
## 4   NA 15.13    NA 23.65    NA    NA
## 5   NA 23.17    NA 32.58 51.76 65.41
```

```
df_4 <- subset(df, year %in% c(2019,2023))
head(df_4,5)
```

```
##   code       economy year   OPP    PC  FAIL    EI   TEA   OWN EMPL MOT FMTEA
## 1    1 United States 2023 53.81 48.99 44.55 12.09 14.71  6.74   NA  NA  0.84
## 2   27  South Africa 2023 64.10 69.21 59.51  7.45 11.11  5.92   NA  NA  0.76
## 3   30        Greece 2023 45.29 53.76 53.16  9.10  6.74 14.70   NA  NA  0.79
## 4   31   Netherlands 2023 67.37 46.01 40.29 16.02 13.69  6.92   NA  NA  0.80
## 5   33        France 2023 50.86 49.51 40.08 13.41 10.75  4.56   NA  NA  0.72
##   FMOD   JOB INNOV  BSER  STAT  CHOI
## 1   NA 28.80    NA 20.42 78.94 79.21
## 2   NA 21.42    NA  3.57 85.51 78.46
## 3   NA 12.12    NA 22.53 69.76 72.90
## 4   NA 15.13    NA 23.65    NA    NA
## 5   NA 23.17    NA 32.58 51.76 65.41
```

```
tail(df_4,5)
```

```
##     code              economy year   OPP    PC  FAIL    EI   TEA  OWN EMPL MOT
## 230  966         Saudi Arabia 2019 73.80 83.00 41.78 32.32 13.96 5.35 3.19  NA
## 231  968                 Oman 2019 72.31 56.32 40.78 62.91  6.94 2.00 1.17  NA
## 232  971 United Arab Emirates 2019 66.10 62.16 41.73 38.50 16.41 6.96 8.24  NA
## 233  972               Israel 2019 46.00 43.34 55.36 21.20 12.69 5.45 5.75  NA
## 234  974                Qatar 2019 75.59 75.47 45.17 45.25 14.69 2.96 3.60  NA
##     FMTEA FMOD   JOB INNOV  BSER  STAT  CHOI
## 230  1.09   NA 61.90    NA  8.77 79.32 69.68
## 231  0.72   NA 23.97    NA  8.77 85.67 85.34
## 232  0.70   NA 65.98    NA 23.82 79.02 70.34
## 233  0.69   NA 21.73    NA 27.06 84.13 64.21
## 234  1.00   NA 58.28    NA 22.75 87.09 82.10
```

Se filtran los datos para los años 2019, y 2023, después se verifica que países tienen datos para dichos años y se filtra por ellos. Para finalizar se crean las muestras de la variable TEA para ambos años, y verificamos las dimensiones de cada sample, una vez más para analizar los mismos tamaños de muestra.

```
countries_2019_2023 <- df_4 %>%
  group_by(economy) %>%
  summarise(count_years = n_distinct(year)) %>%
  filter(count_years == 2) %>%
  pull(economy)
```

```
df_4 <- df_4 %>%
  filter(economy %in% countries_2019_2023)
```

```
TEA_2019 <- df_4 %>%
  filter(year ==2019) %>%
  arrange(economy) %>%
  pull(TEA)

TEA_2023 <- df_4 %>%
  filter(year == 2023) %>%
```

```
  arrange(economy) %>%
  pull(TEA)
```

```
length(TEA_2019)
```

```
## [1] 37
```

```
length(TEA_2023)
```

```
## [1] 37
```

4.1 Planteamiento y elección del Test

Para responder a la pregunta planteada se plantean las siguientes hipótesis: Hipótesis nula H_0 No ha habido un aumento significativo en los valroes promedia de TEA entre 2019 y 2023. Hipótesis alternativa H_1 ha habido un aumento significativo en los valores promedios de TEA entre 2019 y 2023.

$$H_0 : \mu_{2023} \leq \mu_{2019}$$

$$H_1 : \mu_{2023} > \mu_{2019}$$

Como se quiere comprobar si es mayor se usa un test unilateral a la derecha. Se va a hacer una adaptación de la función de tstudent del apartado anterior

```
result <- tstudent_contrast(TEA_2019,TEA_2023,var_igual = FALSE, alpha = 0.05,
                            type = "right tail")
print(result)
```

```
## $t_obs
## [1] -0.3210953
##
## $t_crit
## [1] 1.666375
##
## $p_value
## [1] 0.002246404
##
## $decision
## [1] "Rechazar H0"
##
## $degree_freedom
## [1] 71.73102
```

En este caso se rechaza la hipótesis nula, lo que quiere decir que durante los años ha habido un aumento entre las personas de entre 18 y 64 años que son emprendedoras.

5 Diferencias en TEA según la valoración del emprendimiento: En este apartado estudiaremos si existen diferencias en el emprendimiento (TEA) entre los países

según su ranking en STAT. Concretamen, nos preguntamos: **PR4: Existen diferencias significativas en los valores promedios de TEA entre los 10 países con mejor valoración de los emprendedores (STAT), en relación a los 10 países con peor valoración de STAT?**

```
head(df,5)
```

```
##   code       economy year   OPP    PC  FAIL    EI   TEA   OWN EMPL MOT FMTEA
## 1    1 United States 2023 53.81 48.99 44.55 12.09 14.71  6.74   NA  NA  0.84
## 2   27  South Africa 2023 64.10 69.21 59.51  7.45 11.11  5.92   NA  NA  0.76
## 3   30        Greece 2023 45.29 53.76 53.16  9.10  6.74 14.70   NA  NA  0.79
## 4   31   Netherlands 2023 67.37 46.01 40.29 16.02 13.69  6.92   NA  NA  0.80
## 5   33        France 2023 50.86 49.51 40.08 13.41 10.75  4.56   NA  NA  0.72
##   FMOD   JOB INNOV  BSER  STAT  CHOI
## 1   NA 28.80    NA 20.42 78.94 79.21
## 2   NA 21.42    NA  3.57 85.51 78.46
## 3   NA 12.12    NA 22.53 69.76 72.90
## 4   NA 15.13    NA 23.65    NA    NA
## 5   NA 23.17    NA 32.58 51.76 65.41
```

Para este apartado se va plantear la hipótesis nula H_0, no existe diferencia significativa en los valores de TEA entre los países con mayor y menor valor de STAT.

$$H?0 : \mu_{top10} = \mu_{bot10}$$

Y la Hipótesis alternativa

$$H_1 : \mu_{top10} \neq \mu_{bottom10}$$

existe una diferencia significativa en los valores de TEA entre los países con mayor y menor valor de STAT. Es decir sus medias no son iguales. En esta ocasión de nuevo se hará uso del test t-student ya que se está comparando dos grupos independientes asumiento que las varianzas puedan no ser iguales. Se preparan los datos y se hace uso de la función tstudent_contrast desarrollada anteriormente. Obteniendo los siguientes resultados

```r
df_5 <- df[complete.cases(df$TEA, df$STAT), ]

top10_stat <- df_5[order(-df_5$STAT), ][1:10, c("economy", "TEA", "STAT")]
bottom10_stat <- df_5[order(df_5$STAT), ][1:10, c("economy", "TEA", "STAT")]
```

TOP 10 países con mayor valor en STAT:

```r
top10_stat
```

```
##          economy   TEA   STAT
## 721   Bangladesh 12.77 100.00
## 833        Yemen 24.01  97.46
## 90  Saudi Arabia 19.24  96.73
## 137 Saudi Arabia 19.62  96.33
## 122        Sudan 33.58  95.39
## 42  Saudi Arabia 25.34  95.36
## 574       Uganda 25.21  95.29
## 180 Saudi Arabia 17.30  95.10
## 162         Iran  8.00  94.30
## 570        Ghana 25.82  94.08
```

TOP 10 países con menor valor en STAT

```r
bottom10_stat
```

```
##       economy   TEA  STAT
## 192     Italy  2.79 13.06
## 215   Ireland 12.41 20.69
## 881    Russia  2.67 31.47
## 971   Hungary  1.88 34.47
## 655   Croatia  8.27 41.73
## 449   Croatia  7.69 42.34
```

```
## 271   Croatia  9.61 42.97
## 586   Croatia  8.27 43.07
## 1039    Spain  6.65 44.23
## 555  Malaysia  6.60 44.98
```

```
top10 <- top10_stat$TEA
bot10 <- bottom10_stat$TEA

resultado <- tstudent_contrast(top10,bot10,var_igual = FALSE, alpha = 0.05,
                               type = "bilateral")

print(resultado)
```

```
## $t_obs
## [1] 5.653936
##
## $t_crit
## [1] 2.166736
##
## $p_value
## [1] 8.774693e-05
##
## $decision
## [1] "Rechazar H0"
##
## $degree_freedom
## [1] 12.63467
```

Se ha usado el tipo bilateral, debido a que al ser condición que sea diferente tenemos los valores menors y mayores, de las dos colas. P_value ha salido menor que 0.05 por lo que se rechaza la hipótesis nula a favor de la alternativa. Es decir, existe una diferencia significativa entre los valores de TEA, con mayor y menor valor de STAT.

6 Conclusiones

6.1 Apartado 1

En el apartado 1 se pedía crear un gráfico del tipo boxplot, para observar la distribución de la variable TEA, que está relacionada con el emprendimiento de las personas de entre 18 y 64 añós. Tras realizarlo, se pudo apreciar como en el gráfico que los continentes con mayor TEA, son África y Oceanía, dado que sus rangos inter-cuartílicos son amplios, esto indica que hay una variedad alta en los valores de TEA entre los países. Por otro lado Europa y Asia tiene un valor bajo de media de TEA, pero sus rangos inter-cuartílicos son más compactos, lo que quiere decir que los valores de TEA entre los países son más consistentes.

6.2 Apartado 2

Se pedía responder a la siguiente pregunta: ¿Los países de Europa tienen una varianza diferente al resto de países en la variable TEA?

Se concluye mediante el test de Levene, (específico para este tipo de pregunta= que la varianza de Europa, con respecto al resto del mundo no son iguales. Lo que tiene sentido, desde el punto de vista anaítico planteado en este documento, ya que en el apartado anterior se pudo observar como Europa y Asia, eran los contienentes con menor proporción de emprendedores.

6.3 Apartado 3

En este apartado se pedía lo siguiente, responder a la pregunta: PR2: ¿Los países de Europa tienen medias inferiores que el resto de países en los valores de las variables OPP, PC, EI, TEA y OWN? ¿Y una media superior al resto en la variable FAIL?

Para responder a esta pregunta se realizaron 3 comprobaciones: la primera consitió en realizar el contraste de hipótesis de manera manual para el caso en específico, además se realizó una función de contraste t-student para poder utilizarla de manera genérica aplicable a distintos casos, y para finalizar se verificaron que los datos coincidían con la función establecida de Rstudio, verificando así que los cálculos eran correctos en ambos casos. Respondiendo a la pregunta se comprobó que todas la variables, menos FAIL, es decir: OPP, PC, EI, TEA y OWN, se rechazaba la hipótesis nula, excepto en la variable FAIL que se aceptaba. Las diferencias son significativas si se rechaza la hipótesis nula ya que existe suficiente evidencia para rechazar la misma con un intervalo de confianza del 95%.

```
print(table_results)
```

```
##    Variable  t_obs crit_value p_value degree_freedom       Decision
## 1  TEA_mean -6.737     -1.683 0.00000          40.94    Rechazar H0
## 2  OPP_mean -4.235     -1.669 0.00004          63.52    Rechazar H0
## 3   PC_mean -4.981     -1.674 0.00000          53.32    Rechazar H0
## 4 FAIL_mean  0.578     -1.675 0.71709          51.03 No rechazar H0
## 5   EI_mean -6.690     -1.679 0.00000          45.13    Rechazar H0
## 6  OWN_mean -3.003     -1.682 0.00225          42.04    Rechazar H0
```

6.4 Apartado 4

En este apartado se pedia responder a la siguiente pregunta: PR3: ¿Los valores promedios de TEA de los diferentes países aumentan en cinco años?

Tras preparar los datos se llegó a la siguiente conclusión: se rechaza la hipótesis nula, lo que quiere decir que durante los años ha habido un aumento entre las personas de entre 18 y 64 años que son emprendedoras. Respuesta que tiene sentido lógico ya que se ha comprobado que a lo largo de los años ha estado habiendo un crecimiento significativo de las personas que deciden emprender con sus propios negocios

6.5 Apartado 5

Se pedía responder a la siguiente pregunta: PR4: Existen diferencias significativas en los valores promedios de TEA entre los 10 países con mejor valoración de los emprendedores (STAT), en relación a los 10 países con peor valoración de STAT?

Concluyendo que: se rechaza la hipótesis nula a favor de la alternativa. Es decir, existe una diferencia significativa entre los valores de TEA, con mayor y menor valor de STAT. El STAT se define como: Porcentaje de población de 18 a 64 años que está de acuerdo con la afirmación de que en su país, los emprendedores de éxito reciben un estatus elevado. Por lo que una vez más el resultado tiene sentido ya que lógicamente los países en los que sus habitantes están de acuerdo que los emprendedores de éxito reciben un estatus importante, decidirán emprender con mayor facilidad y asertividad, que las personas que piensen lo contrario.

6.6 Conclusión final

Para terminar, se confirma que los resultados obtenidos coinciden con lo que se esperaría lógicamente, sin hacer el estudio, en principio, no se ha apreciado ninguna contradicción, por lo que se considera que los test y los análisis se han realizado de manera satisfactoria.