Activitat 2: Anàlisi descriptiva i inferencial

Òscar Casals

Contents

```
library(tibble)
library(dplyr)
library(tidyr)
library(stringr)
library(purrr)
library(ggplot2)
library(pander)
library(kableExtra)
library(countrycode)

knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
```

1 Breu anàlisi descriptiva

Analitzeu gràficament la variable TEA segons s'indica a continuació.

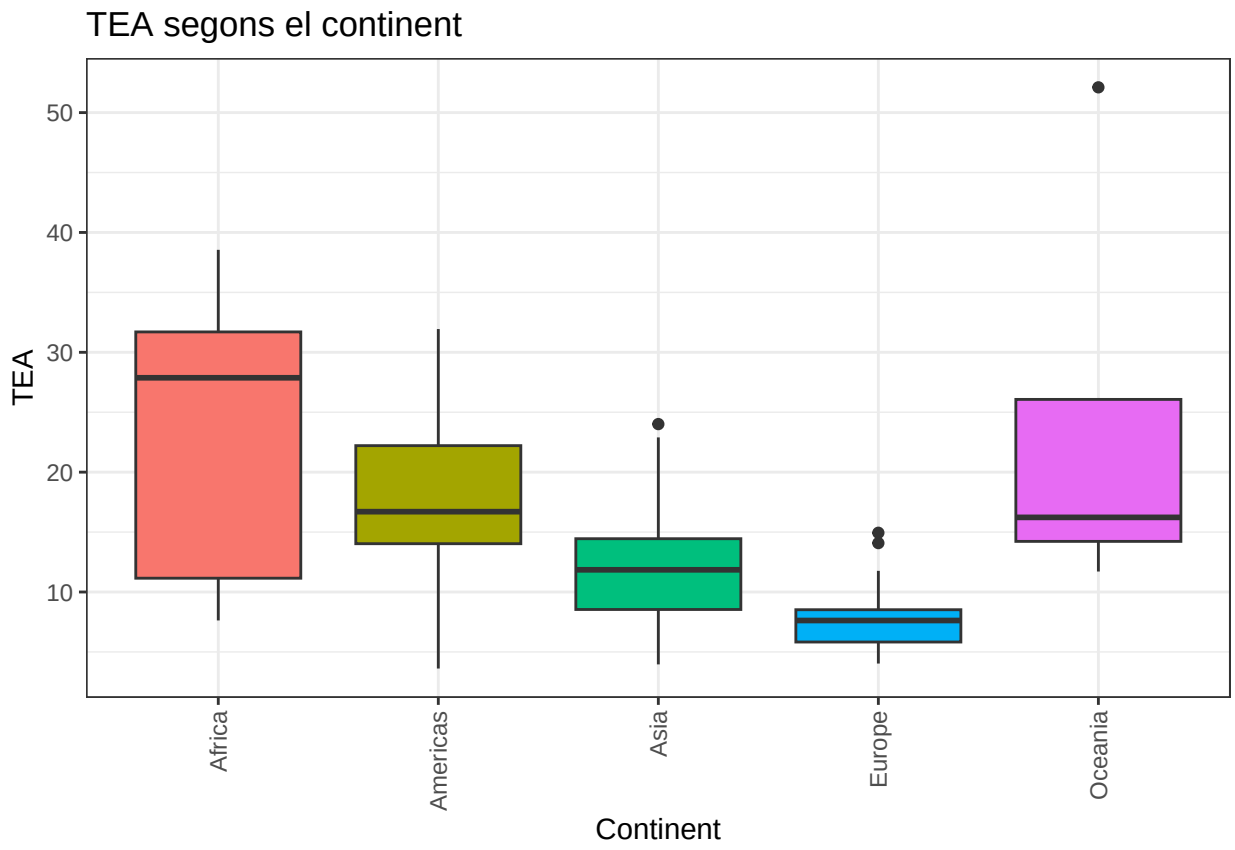
1.1 Distribució de TEA segons el continent

Mostreu un gràfic de tipus boxplot, on es vegi la distribució de la variable TEA segons el continent. Per a fer-ho, calculeu la mitjana del valor TEA per cada país al llarg dels anys. Interpreteu breument.

Nota: Podeu usar llibreries d'R per identificar a quin continent pertany cada país de la mostra.

```
# Carregant el data.
df <- read.csv("gem02.csv")
# Transformant les variables en els tipus adequats.
df$year <- as.factor(df$year)

df$economy <- as.factor(df$economy)
# Determinant a quin continent pertany cada país.
df$continent <- countrycode(sourcevar = df$economy, origin = "country.name", destination = "continent")
# Com Kosovo no està inclós en la funció countrycode, el continent per a aquest país s'afageix manualme
df[df$economy=="Kosovo","continent"] <- "Europe"
# Es converteix la columna continent a factor, per a que tingui el tipus correcte.
df$continent <- as.factor(df$continent)
# Es calcula la mitjana del valor TEA per cada país al llarg dels anys
SummaryData <- df %>% group_by(economy, continent) %>% summarize(Mean_TEA = mean(TEA))
# Es crea el boxplot
ggplot(data = SummaryData, aes(x = continent, y = Mean_TEA, fill = continent)) +
  geom_boxplot() +
```

```
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1), legend.position = "none") +
  labs(title = "TEA segons el continent", y = "TEA", x = "Continent")
```



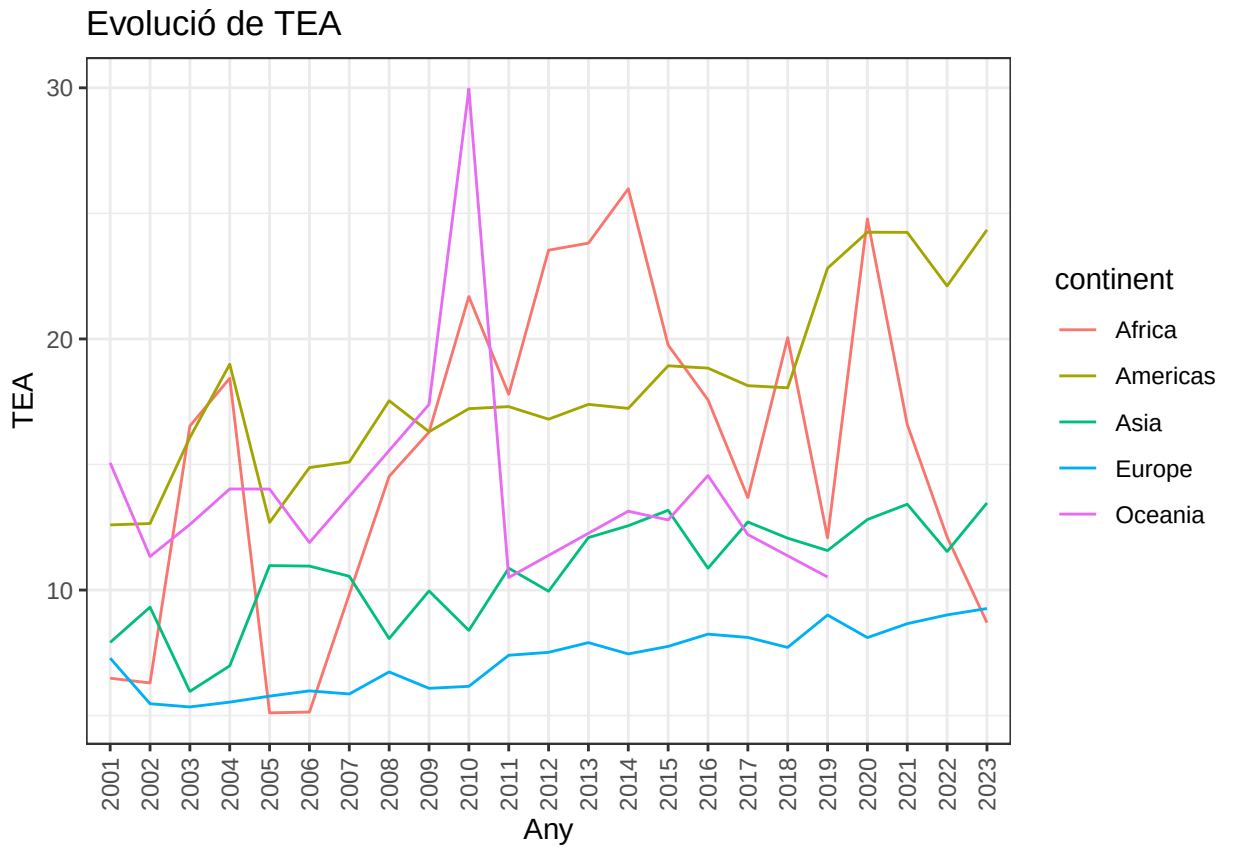
En el boxplot es pot veure que Àfrica és el continent on més percentatge de la població entre 18-64 són emprenedors o tenen un negoci propi, seguit d'Amèrica, Oceania, Àsia i Europa.

Tot i això, també s' observa que Àfrica i Oceània són els països amb més variancia en el TEA, el que significa que en aquests continents hi ha zones amb un TEA molt més baix o alt que la resta.

1.2 Evolució de TEA per continent

Mostreu l'evolució de TEA segons el continent i al llarg dels anys. Interpreteu breument.

```
# Es calcula la mitjana de TEA per continent i any.
SummaryData.Years <- df %>% group_by(continent, year) %>% summarize(Mean_TEA = mean(TEA))
# Es crea un lineplot.
ggplot(data = SummaryData.Years, aes(x = year, y = Mean_TEA, group = continent)) +
  geom_line(mapping = aes(colour = continent)) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title = "Evolució de TEA", y = "TEA", x = "Any")
```



Per cada continent es pot veure que:

- A Europa el TEA va creixent de forma lenta sense cap pic, puixada, o baixada important. Tot i així el seu TEA és menor al dels altres continents.

- Si bé Àsia té algun pic petit, la seva situació és d'anar augmentant poc a poc el TEA sense tenir grans baixades. El seu TEA és major al d'Europa.

- A Oceania es pot veure que el TEA va augmentar bastant de 2006 a 2010 i, posteriorment, aquest va disminuir dràsticament al 2011. Com el dataset no inclou dades d'Oceania posteriors a 2019 no es pot saber si el continent s'ha recuperat de la baixada de TEA o, per contra, ha continuat baixant.

- Amèrica després del pic entre 2002 i 2005 ha seguit creïxent sense augments importants fins al 2019 , quan el TEA augmenta bastant i es manté sense baixades dràstiques els posteriors anys. Dels continents que es mostren en el gràfic Amèrica és el que té el TEA amb creixement més alt i estable.

- A Àfrica l'evolució del TEA està conformada per pics, després d'una pujada summament gran vé una baixada dràstica i després d'aquesta, es dona un altre gran pic.

2 Anàlisi de la variabilitat

En primer lloc, volem analitzar si la variabilitat en els valors promitjos de TEA és la mateixa entre Europa i la resta de països.

La pregunta de recerca a respondre és:

PR1: Els països d'Europa tenen una variança diferent a la resta de països en la variable TEA?

Per a fer-ho, seguiu les indicacions que us donem a continuació.

Nota: Al llarg de tota l'activitat podeu assumir que el nivell de confiança és del 95%.

2.1 Preparació de les mostres

Prepareu les dues mostres, una amb els països d'Europa i l'altra amb la resta de països. Per a cada país, calculeu la mitjana de TEA en els diferents anys en què el país té dades. Si un país no té dades en tots els anys, calculeu igualment la mitjana en aquells anys que sí que hi ha dades.

```
# Es crean dos mostres, una amb els països d'Europa i l'altre amb els països que no
# són d'Europa.

df.Europe <- df %>% filter(continent == "Europe")

df.NoEurope <- df %>% filter(continent != "Europe")

# Es fa la mitjana de TEA per cada país per les dues mostres.

Means.Europe <- df.Europe %>% group_by(economy) %>% summarize(Mean_TEA = mean(TEA))

Means.NoEurope <- df.NoEurope %>% group_by(economy) %>% summarize(Mean_TEA = mean(TEA))

# Es mostren les taules

knitr::kable(head(Means.Europe), caption = "Means countries of Europe")
```

Table 1: Means countries of Europe

economy	Mean_TEA
Austria	7.441250
Belarus	9.630000
Belgium	4.118000
Bosnia And Herzegovina	7.347500
Bulgaria	4.500000
Croatia	8.127273

```
knitr::kable(head(Means.NoEurope), caption = "Means countries of Europe")
```

Table 2: Means countries of Europe

economy	Mean_TEA
Algeria	9.89500
Angola	31.60143
Argentina	14.08611
Armenia	20.97000
Australia	11.70923
Bangladesh	12.77000

2.2 Hipòtesis

Escriviu la hipòtesi nul · la i alternativa per aquest test.

La hipòtesi nul · la és que els països d'Europa no tenen una variança diferent a la resta de països en la variable TEA i la alternativa és que sí la tenen. Dit d'una altra manera:

- ***Hipòtesi nul · la:*** $H_0 = \sigma^2_{Europa} = \sigma^2_{NoEuropa}$

- ***Hipòtesi alternativa:*** $H_1 = \sigma^2_{Europa} \neq \sigma^2_{NoEuropa}$

2.3 Elecció del test

Escolliu i justifiqueu el test més adequat per donar resposta a la pregunta.

El test més adequat és el test F d'igualtat de variàncies, perquè aquest test compara dues variàncies per veure si són significativament diferents assumint una distribució normal.

Com la n de les dues mostres és superior a 30, es pot assumir que els resultats seran similars als de una distribució normal tal com indica el Teorema del límit central.

2.4 Desenvolupament del test

Implementeu el test usant el vostre propi codi.

Nota: No podeu usar funcions ja implementades de les llibreries d'R que calculin el contrast d'hipòtesis. Només podeu usar funcions com **qnorm, qt, pnorm, pt,** etcètera.

```
Ftest <- function(x, y, alfa = 0.05){
  # Es calcula la n de cada mostra.
  n.X <- length(x)
  n.Y <- length(y)
  # Es calcula la desviació estàndard de cada mostra.
  s.X <- sd(x)
  s.Y <- sd(y)
  # Calculem l'estadístic F observat.
  fobs <- (s.X^2) / (s.Y^2)
  # Es calculen els límits d'acceptació de la hipòtesis nul·la
  fcritL <- qf(alfa/2, df1=n.X-1, df2=n.Y-1)
  fcritU <- qf(1 - alfa/2, df1=n.X-1, df2=n.Y-1)
  # Es calcula el valor p
  pvalue <- min(pf( fobs, df1=n.X-1, df2=n.Y-1, lower.tail=FALSE ), pf( fobs, df1=n.X-1, df2=n.Y-1))*2
  # Es retorna el F observat, els límits d'acceptació de la hipòtesis nul·la, i el valor p
  return(list(
    Fobs = fobs,
    Limits = c(fcritL, fcritU),
    Pvalue = pvalue))
}

Result <- Ftest(Means.Europe$Mean_TEA, Means.NoEurope$Mean_TEA)
```

2.5 Resultat i interpretació

Mostreu el resultat del test i doneu resposta a la pregunta plantejada.

Els resultats del test són:

```
pander(cat("- **El valor observat és:", Result$Fobs, sep = " ", end = "**\n\n"))
```

- **El valor observat és: 0.07593518**

```
pander(cat("- **Els límits de la zona d'acceptació de la hipòtesi nul·la són:", Result$Limits[1], "i", R
```

- **Els límits de la zona d'acceptació de la hipòtesi nul · la són: 0.5473671 i 1.716101**

```
pander(cat("- **El valor _p_ és:", Result$Pvalue, sep = " ", end = "**\n\n"))
```

- **El valor *p* és: 8.406953e-13**

Com el valor observat es troba fora de límits de la zona d'acceptació de la hipòtesi nul · la i el valor p és menor a alpha (0.05), la variança dels països d'Europa és diferent a la resta de països en la variable TEA.

3 Anàlisi de diferències entre els països d'Europa i la resta

En aquest apartat volem analitzar les diferències entre els països d'Europa i la resta de països en les variables: OPP, PC, FAIL, EI, TEA i OWN. La pregunta de recerca a respondre és:

PR2: Els països d'Europa tenen mitjanes inferiors que la resta de països en els valors de les variables OPP, PC, EI, TEA i OWN? I una mitjana superior a la resta en la variable FAIL?

Per a fer-ho seguiu els passos que s'indiquen a continuació.

3.1 Preparació de les mostres

Prepareu les dues mostres, una amb els països d'Europa i l'altra amb la resta de països. Per a cada país, calculeu la mitjana de les variables indicades en els diferents anys en què el país té dades. Si un país no té dades en tots els anys, calculeu igualment la mitjana en aquells anys que sí que hi ha dades.

```
# Es crean dos mostres, una amb els països d'Europa i l'altre amb els països que no
# són d'Europa.

df.Europe <- df %>% filter(continent == "Europe")

df.NoEurope <- df %>% filter(continent != "Europe")

# Es fa la mitjana de cada variable que es té en compte.

Means.Europe <- df.Europe %>% group_by(economy) %>% summarize(Mean_OPP = mean(OPP, na.rm = TRUE), Mean_

Means.NoEurope <- df.NoEurope %>% group_by(economy) %>% summarize(Mean_OPP = mean(OPP, na.rm = TRUE), M

# Es mostren les taules

knitr::kable(head(Means.Europe), caption = "Means countries of Europe")
```

Table 3: Means countries of Europe

economy	Mean_OPP	Mean_PC	Mean_EI	Mean_TEA	Mean_OWN	Mean_FAIL
Austria	42.89125	50.16125	7.313750	7.441250	7.338750	35.48750
Belarus	27.27000	47.15000	15.325000	9.630000	4.130000	46.98500
Belgium	27.81867	36.73333	7.261429	4.118000	3.574667	36.09067
Bosnia And Herzegovina	26.89875	52.39500	17.968750	7.347500	5.345000	27.45250
Bulgaria	18.89750	37.52000	5.337500	4.500000	6.610000	27.55250
Croatia	34.58727	55.43636	14.662273	8.127273	3.706364	36.03773

```
knitr::kable(head(Means.NoEurope), caption = "Means countries of Europe")
```

Table 4: Means countries of Europe

economy	Mean_OPP	Mean_PC	Mean_EI	Mean_TEA	Mean_OWN	Mean_FAIL
Algeria	52.55250	55.31750	30.32750	9.89500	4.132500	35.59750
Angola	69.13857	67.58714	56.01000	31.60143	8.747143	38.83286
Argentina	45.22778	59.16167	20.71706	14.08611	9.180000	30.92111
Armenia	53.89000	70.00000	32.20000	20.97000	7.840000	48.22000
Australia	46.76385	52.80769	11.81167	11.70923	10.517692	38.19231
Bangladesh	64.43000	23.63000	24.57000	12.77000	11.600000	72.01000

3.2 Hipòtesis

Escriviu les hipòtesis nul · la i alternativa per aquest test.

La hipòtesi nul · la és que els països d'Europa tenen les mitjanes iguals que la resta de països en totes les variables proposades, mentre que la alternativa és que Europa té una mitjana inferior en les variables OPP, PC, EI, TEA i OWN, i una superior en la variable FAIL. Dit d'una altre manera:

- Hipòtesi nul · la: $H_o = \mu_{NoEuropa} = \mu_{Europa}$

- Hipòtesi alternativa per les variables OPP, PC, EI, TEA i OWN: $H_1 = \mu_{NoEuropa} > \mu_{Europa}$

- Hipòtesi alternativa per la variable FAIL: $H_1 = \mu_{NoEuropa} < \mu_{Europa}$

3.3 Elecció del test

Indiqueu quin test o tests aplicareu per donar resposta a la pregunta plantejada. Justifiqueu la vostra elecció.

El test més adequat és el t-test per a dues mostres independents degut a que s'utilitza per comparar les mitjanes d'una variable entre dos grups (en aquest cas s'hauria de fer un per a cada variable).

Per saber si convé aplicar el t-test de variàncies desconegudes diferents o el t-test de variàncies desconegudes iguals s'utilitzarà el test F d'igualtat de variàncies, perquè aquest test compara dues variàncies per veure si són significativament diferents assumint que segueixen una distribució normal. Com la n de les dues mostres és superior a 30, es pot assumir que els resultats seran similars als de una distribució normal tal com indica el Teorema del límit central.

3.4 Desenvolupament del test

Desenvolupeu una funció que permeti realitzar el contrast d'hipòtesis. Cal que desenvolupeu el codi complet. Calculeu el valor observat, el valor crític i el valor p.

Nota: No podeu usar funcions ja implementades de les llibreries d'R que calculin el contrast d'hipòtesis. Només podeu usar funcions com qnorm, qt, pnorm, pt, etcètera.

```
Ttest <- function(x, y, alternative="bilateral", alfa = 0.05, var.equal = FALSE){
  # Es calcula la mitjana de les mostres.
  mean.X <- mean(x)
  mean.Y <- mean(y)
  # Es calcula la desviació estandard de les mostres.
  s.X <- sd(x)
  s.Y <- sd(y)
  # Es calcula la quantitat de mostres.
  n.X <- length(x)
  n.Y <- length(y)
  # S'utilitza la formula adequada segons si la variancia desconeguda és diferent
  # o igual.
  if(var.equal == FALSE){
    tobs <- (mean.X - mean.Y) / sqrt(((s.X^2)/n.X) + ((s.Y^2)/n.Y))

    df <- ((((s.X^2)/n.X) + ((s.Y^2)/n.Y))^2) /
          (((((s.X^2)/n.X)^2) / (n.X-1)) + ((((s.Y^2)/n.Y)^2) / (n.Y-1)))
  }else{
    S <- sqrt(((n.X-1)*s.X^2 + (n.Y-1)*s.Y^2 ) / (n.X+n.Y-2))

    tobs <- (mean.X-mean.Y) / (S * sqrt(1/n.X + 1/n.Y))

    df <- n.X+n.Y-2
  }
  # Es calculen els valors crítics i el valor p segons la hipòtesis alternativa.
  switch(alternative,
      "bilateral" =
      {
      tcritL <- qt(alfa/2, df)
      tcritU <- qt(1-alfa/2, df)

      Pvalue <- pt(abs(tobs), df = df, lower.tail = FALSE) * 2
      },
      "greater" =
      {
        # tcritL <- qt(alfa, df)
        tcritL <- -Inf
        tcritU <- qt(1-alfa, df)
        Pvalue <- 1 - pt(tobs, df)
      },
      "lower" =
      {
        tcritL <- qt(alfa, df)
        # tcritU <- qt(1-alfa, df)
        tcritU <- Inf
        Pvalue <- pt(tobs, df)
```

```
      }
    )

  return(list(ValorObservat = tobs,
              ValorsCritics = c(tcritL, tcritU),
              pvalue = Pvalue
              ))
}
```

3.5 Càlcul del contrast

Apliqueu els contrastos per avaluar si hi ha diferències significatives entre els països en les variables indicades. Aquí heu d'usar la funció que heu implementat en la secció anterior.

```
Ftest.Ttest <- function(x, y, alternative="bilateral", alfa = 0.05){
  # Es realitza el test F per saber si la variança desconeguda és diferent o no en
  # les mostres.
  VarTest <- Ftest(x, y, alfa)
  # Si el valor p del test F és menor a 0.05, la variança desconeguda és diferent entre les
  # mostres, sinò és igual.
  if(VarTest$Pvalue >= 0.05){
    Result <- Ttest(x, y, alternative, alfa, var.equal = TRUE)
    return(Result)
  }
  Result <- Ttest(x, y, alternative, alfa, var.equal = FALSE)
  return(Result)
}

# Per cada columna es mira si la variança desconeguda és igual o diferent entre
# les mostres, i després es realitza el t test.

Conts <- map(colnames(Means.Europe)[-c(1, ncol(Means.Europe))], function(x){

  X <- Means.Europe[[x]]
  Y <- Means.NoEurope[[x]]

  R <- Ftest.Ttest(X, Y, alternative = "lower")

  R$ColumnName <- x

  return(R)
})

# Es fan el F i t test per la variable FAIL, la raó per la qual no es
# fa en el map anterior és perquè la hipòtesis alternativa és diferent a la resta.

ResFail <- Ftest.Ttest(Means.Europe$Mean_FAIL,
                       Means.NoEurope$Mean_FAIL,
                       alternative = "greater")

ResFail$ColumnName <- "Mean_FAIL"
```

3.6 Resultats

Un cop realitzats els càlculs, resumiu els resultats en una taula, de manera que cada fila correspongui a una variable i a les columnes s'indiqui el valor observat, el valor crític, el valor p obtingut i un breu comentari sobre si la diferència és significativa.

```
# Es crea una taula amb els resultats del t test per a cada variable.

Table <- tibble()

colnames(Table) <- c("Variable", "Valor Observat", "Valor Critic", "P valor", "Comentari")

AddRow <- function(x){
  Nrow <- tibble(Variable = stringr::str_split_i(x$ColumnName, "[_]", 2), `Valor Observat` = x$ValorObs
  # El comentari que es dona de cada variable és determinat per si el valor p és menor a alfa o no.
  Nrow$Comentari <- ifelse(x$pvalue < 0.05,
                           "Com el valor p és menor a alfa i el valor observat és fora dels límits de la
                           "Com el valor p és major a alfa i el valor observat és dins dels límits de la

  Table <<- rbind(Table, Nrow)
}

AddRows <- map(Conts, AddRow)

# Afegint la variable FAIL.

AddRow(ResFail)

Table$`P valor` <- format(Table$`P valor`, digits = 3)

# Mostrant la taula

knitr::kable(Table, caption = "Resultats del test d'hipòtesis ttest.") %>% kable_styling() %>% column_sp
```

Table 5: Resultats del test d'hipòtesis ttest.

Variable	Valor Observat	Valor Critic	P valor	Comentari
OPP	-5.085204	-1.66, Inf	7.36e-07	Com el valor p és menor a alfa i el valor observat és fora dels límits de la zona d'acceptació de la hipòtesis nul · la, la diferència és significativa.
PC	-6.665444	-1.66, Inf	6.03e-10	Com el valor p és menor a alfa i el valor observat és fora dels límits de la zona d'acceptació de la hipòtesis nul · la, la diferència és significativa.
EI	-9.743158	-1.66, Inf	7.47e-17	Com el valor p és menor a alfa i el valor observat és fora dels límits de la zona d'acceptació de la hipòtesis nul · la, la diferència és significativa.

TEA	-8.853516	-1.66, Inf	1.61e-14	Com el valor p és menor a alfa i el valor observat és fora dels límits de la zona d'acceptació de la hipòtesis nul · la, la diferència és significativa.
OWN	-4.066231	-1.66, Inf	4.55e-05	Com el valor p és menor a alfa i el valor observat és fora dels límits de la zona d'acceptació de la hipòtesis nul · la, la diferència és significativa.
FAIL	1.360273	-Inf, 1.66	8.83e-02	Com el valor p és major a alfa i el valor observat és dins dels límits de la zona d'acceptació de la hipòtesis nul · la, la diferència no és significativa.

3.7 Interpretació

Interpreteu els resultats obtinguts i doneu resposta a la pregunta plantejada.

Els resultats mostren que els països d'Europa tenen mitjanes inferiors a la resta de països en les variables OPP, PC, EI, TEA i OWN, degut a que totes aquestes tenen un valor observat negatiu fora dels límits de la zona d'acceptació de la hipòtesi nul · la i un valor p menor a alfa (0.05). Per contra, també mostren que la mitjana de FAIL no és significativament diferent a la resta de variables perquè el valor observat es troba dintre dels límits de la zona d'acceptació de la hipòtesi nul · la.

4 Anàlisi longitudinal

En aquest apartat, es vol analitzar si hi ha hagut una millora significativa del percentatge emprenedor en un període de cinc anys. Per aquest motiu, escollirem el període 2019-2023 per donar resposta a la pregunta. Per tant, la pregunta de recerca és:

PR3: Els valors promitjos de TEA dels diferents països augmenten en cinc anys?

Seguiu els passos que s'indiquen a continuació.

4.1 Selecció de les mostres

Prepareu les mostres per a l'anàlisi. Descarteu els països pels quals no tenim el valor de TEA en els dos anys de l'anàlisi.

```
# Mirant quins països no tenen NA TEA i poseeixen entrades pels anys 2019 i 2023.
Countries.With.All.Years <- df %>%
  filter(!(is.na(TEA))) %>%
  group_by(economy) %>%
  summarize(Number.Of.Years = sum(year %in% c(2019, 2023))) %>%
  filter(Number.Of.Years == 2)
# Filtrant els països que han entrat a la selecció anterior.
Df.Longitudinal <- df %>%
  filter((economy %in% Countries.With.All.Years$economy) & (year %in% c(2019, 2023))) %>% droplevels()
```

```r
# Creant dues mostres, una per l'any 2019 i l'altre a l'any 2023.
Df.2019 <- Df.Longitudinal %>% filter(year == 2019) %>% arrange(economy) %>% droplevels()

Df.2023 <- Df.Longitudinal %>% filter(year == 2023) %>% arrange(economy) %>% droplevels()
```

4.2 Hipòtesis

Escriviu les hipòtesis nul · la i alternativa per aquest test.

En aquest cas la hipòtesi nul · la és que els valors promitjos de TEA dels diferents països no augmenten en cinc anys, i la hipotèsi alternativa és que si ho fan. Dit d'una altre manera.

- ***Hipòtesi nu · la:*** $H_0 : \mu_{anys} = 0$
- ***Hipòtesi alternativa:*** $H_0 : \mu_{anys} > 0$

4.3 Elecció del test

Justifiqueu quin test aplicareu per donar resposta a la pregunta plantejada.

El test que aplicaré és el test t aparellat, degut a que estem tractant amb mostres aparellades amb variància desconeguda. Com la n és superior a 30, podem assumir normalitat segons el Teorema del límit central.

4.4 Desenvolupament del test

Desenvolupeu el test usant un codi propi.

Nota: No podeu usar funcions ja implementades de les llibreries d'R que calculin el test. Sí que podeu usar funcions com **qnorm, qt, pnorm, pt,** etcètera.

```r
Paired.T.Test <- function(x, y, alfa = 0.05, alternative = "bilateral"){
  # Calculant la diferència entre les dues mostres.
  d <- x - y
  # Calculant la mitjana de diferència.
  Mean.D <- mean(d)
  # Calculant la diferència estàndard de la diferència.
  SD.D <- sd(d)
  # Calculant la quantitat de mostres.
  N.D <- length(d)
  # Calculant el valor observat.
  tobs <- Mean.D / (SD.D/(sqrt(N.D)))
  # Graus de llibertat.
  df <- N.D - 1
  # Es calculen els valors crítics i el valor p segons la hipòtesis alternativa.
  switch(alternative,
      "bilateral" =
      {
      tcritL <- qt(alfa/2, df)
      tcritU <- qt(1-alfa/2, df)

      Pvalue <- pt(abs(tobs), df = df, lower.tail = FALSE) * 2
      },
```

```
      "greater" =
      {
        # tcritL <- qt(alfa, df)
        tcritL <- -Inf
        tcritU <- qt(1-alfa, df)
        Pvalue <- 1 - pt(tobs, df)
      },
      "lower" =
      {
        tcritL <- qt(alfa, df)
        # tcritU <- qt(1-alfa, df)
        tcritU <- Inf
        Pvalue <- pt(tobs, df)
      }
    )

  return(list(
    Tobs = tobs,
    Tcrit = c(tcritL, tcritU),
    pvalue = Pvalue
  ))
}
```

4.5 Resultat i interpretació

Mostreu el resultat del test i interpreteu el resultat.

```
# Es mostren els resultats en una taula.
ResultTest <- Paired.T.Test(Df.2023$TEA, Df.2019$TEA, alternative = "greater")

Res <- tibble(tobs = ResultTest$Tobs, `Valors Crítics` = paste(ResultTest$Tcrit, collapse = ", "), Pvalo

knitr::kable(Res, caption = "Results from paired t test.")
```

Table 6: Results from paired t test.

tobs	Valors Crítics	Pvalor
0.8437381	-Inf, 1.68829771411682	0.2021938

Els resultats indiquen que la hipòtesi nul · la és correcte i, per tant, els valors promitjos de TEA dels diferents països no augmenten significativament en cinc anys. Això degut a que el valor observat es troba dins dels límits de la zona d'acceptació de la hipòtesi nul · la i el valor p és superior a alfa (0.05)

5 Diferències en TEA segons la valoració de l'emprenedoria

En aquest apartat estudiarem si hi ha diferències en l'emprenedoria (TEA) entre els països segons el seu rànquing en STAT. Concretamen, ens preguntem:

PR4: Hi ha diferències significatives en els valors promitjos de TEA entre els 10 països amb millor valoració dels emprenedors (STAT), en relació als 10 països amb pitjor valoració d'STAT?

Per a respondre a aquesta pregunta, seguiu els passos següents.

5.1 Selecció de les mostres

Seleccionar els 10 països amb més STAT i els 10 països amb menys valor de STAT. Mostreu dues taules amb les dues mostres. En cada taula, mostreu els noms dels països i els valors de TEA i STAT de cada un.

Nota: no tingueu en compte els països dels quals no tenim tota la informació necessària.

```
# Es filtren les files que tenen NA a les variables STAT i TEA.
DataFrame <- df %>% filter(!is.na(STAT) & !is.na(TEA)) %>% arrange(desc(TEA))
# Es calculen les mitjanes de STAT i TEA de cada país i s'ordenen de més STAT a menys.
NoNA <- DataFrame %>% group_by(economy) %>%
  summarise(Mean_STAT = mean(STAT), Mean_TEA = mean(TEA)) %>%
  arrange(desc(Mean_STAT))
# Guarden els països amb 10 millors STAT i 10 menors STAT.
top10.Stat <- head(NoNA, n = 10)

worst10.Stat <- tail(NoNA, n = 10)

# Mostren les dades.

knitr:: kable(top10.Stat, caption = "10 països amb més STAT.")
```

Table 7: 10 països amb més STAT.

economy	Mean_STAT	Mean_TEA
Bangladesh	100.000	12.77000
Yemen	97.460	24.01000
Ghana	92.040	32.09667
Ethiopia	91.850	14.73000
Sudan	90.345	27.87500
Togo	89.695	28.50000
Syria	89.480	8.46000
Dominican Republic	88.800	24.14250
Uganda	88.062	30.12400
Tunisia	87.286	9.52000

```
knitr:: kable(worst10.Stat, caption = "10 països amb menys STAT.")
```

Table 8: 10 països amb menys STAT.

economy	Mean_STAT	Mean_TEA
Singapore	58.76000	7.806250
Malaysia	58.46800	7.411000
Mexico	58.03077	12.933077
Spain	55.98190	5.909524

economy	Mean_STAT	Mean_TEA
Belgium	55.11385	4.199231
Puerto Rico	54.18667	10.901111
Japan	53.60824	4.141765
Tonga	51.82000	17.390000
Croatia	49.17429	8.341905
Czech Republic	47.97333	7.606667

5.2 Hipòtesis

Escriviu les hipòtesis nul · la i alternativa per aquest test.

La hipòtesi nul · la és que no hi ha diferències significatives en els valors promitjos de TEA entre els 10 països amb millor STAT i l'alternativa és que sí hi han, és a dir:

- ***Hipòtesi nul · la:*** $H_0 : \mu_{10millorSTAT} = \mu_{10pitjorSTAT}$
- ***Hipòtesi alternativa:*** $H_0 : \mu_{10millorSTAT} \neq \mu_{10pitjorSTAT}$

5.3 Elecció del test

Justifiqueu quin test cal aplicar per donar resposta a la pregunta plantejada.

Com la n no és suficientment gran com per assumir normalitat, abans de determinar quin test utilitzar s'ha de mirar si les mostres segueixen una distribució normal mitjançant el test de shapiro i una qqplot:

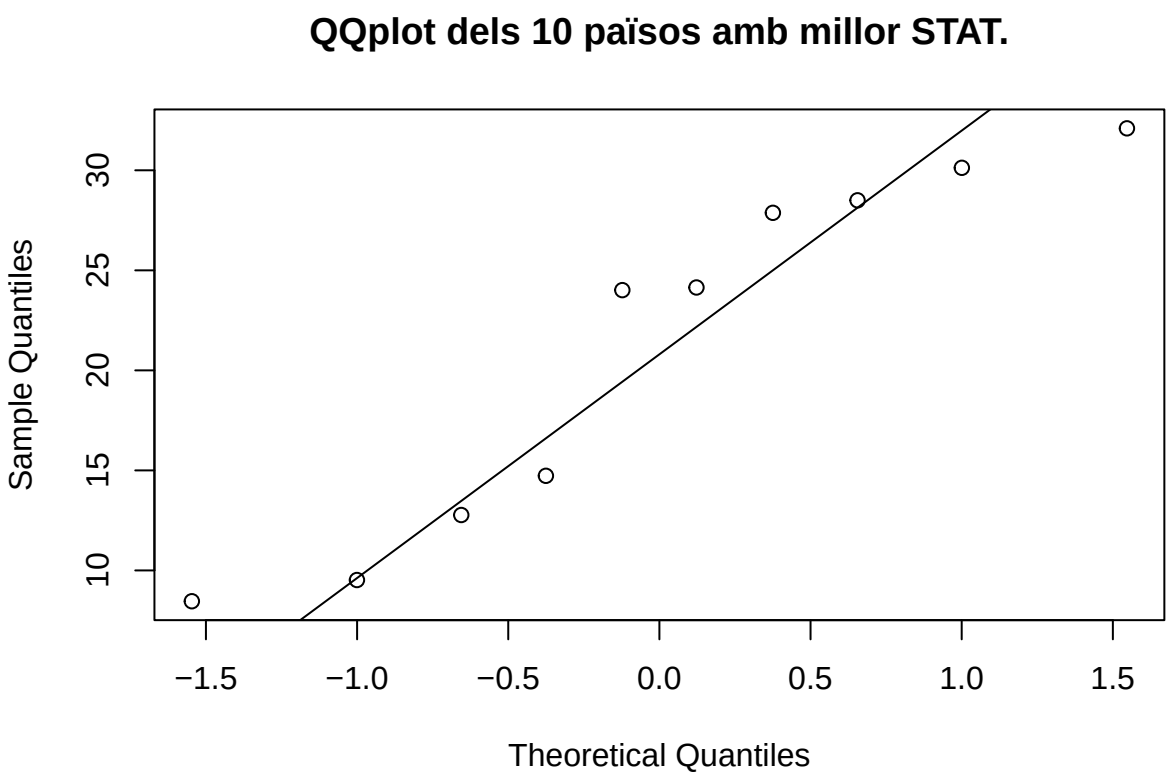
```
# Fent el test de shapiro
top10Sh <- shapiro.test(top10.Stat$Mean_TEA)
worst10Sh <-shapiro.test(worst10.Stat$Mean_TEA)

pander(cat("Shapiro test per la mostra amb els 10 millors STAT té p-valor de: ", top10Sh$p.value, end =
```

Shapiro test per la mostra amb els 10 millors STAT té p-valor de: 0.1514973

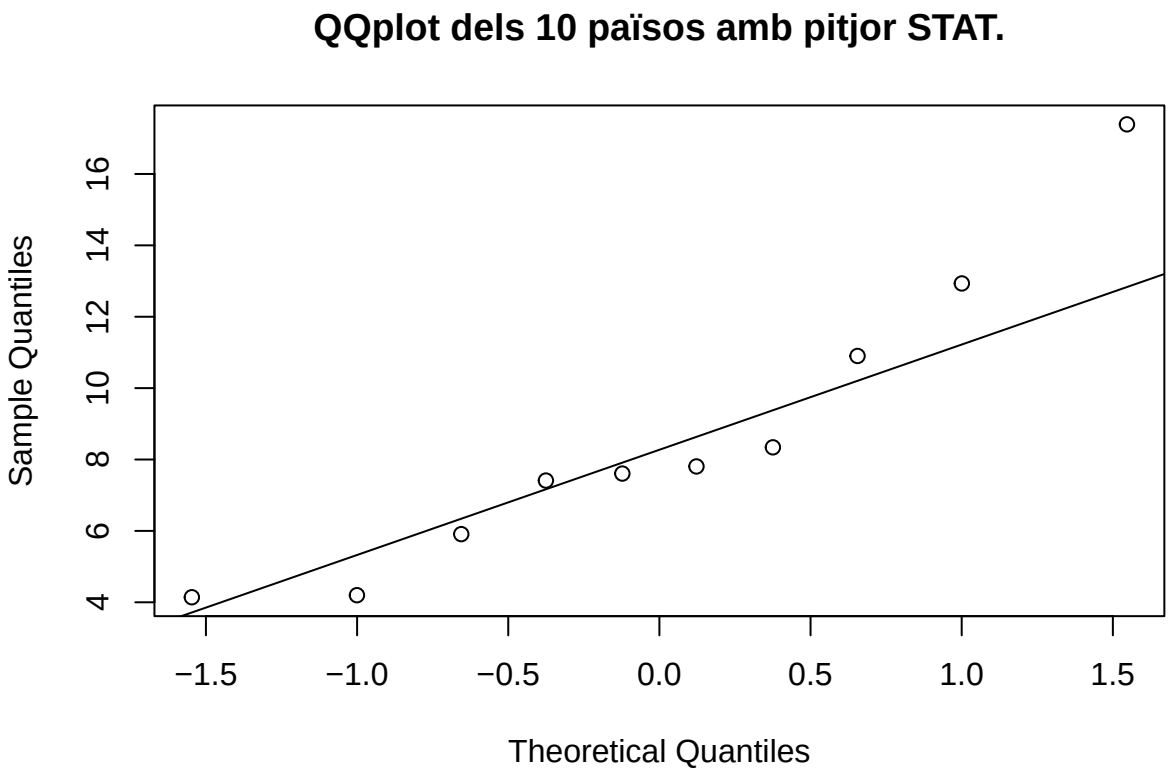
```
pander(cat("Shapiro test per la mostra amb els 10 pitjors STAT té p-valor de: ", worst10Sh$p.value, end
```

Shapiro test per la mostra amb els 10 pitjors STAT té p-valor de: 0.230884

```
# Creant les qqplot
qqnorm(top10.Stat$Mean_TEA, main = "QQplot dels 10 països amb millor STAT.")
qqline(top10.Stat$Mean_TEA)
```




```
qqnorm(worst10.Stat$Mean_TEA, main = "QQplot dels 10 països amb pitjor STAT.")
qqline(worst10.Stat$Mean_TEA)
```




```
# Fent el test F.
VarTest <- Ftest(top10.Stat$Mean_TEA, worst10.Stat$Mean_TEA)

pander(cat("El test F mostra un valor p de:", VarTest$Pvalue))
```

El test F mostra un valor p de: 0.0286282

Tant els test de Shapiro com les qqplot indiquen que la distribució de les dades és normal, i el test F mostra que les variancies desconegudes són diferents; per tant s'aplicarà el t-test de variàncies desconegudes diferents.

5.4 Desenvolupament del test

Desenvolupeu el test usant un codi propi.

Nota: No podeu usar funcions ja implementades de les llibreries d'R que calculin el test. Sí que podeu usar funcions com qnorm, qt, pnorm, pt, etcètera.

El t-test de variàncies desconegudes diferents ja ha sigut implementat a la pregunta 2. Per tant només es cridarà a la funció creada en aquell apartat.

```
Result <- Ttest(top10.Stat$Mean_TEA, worst10.Stat$Mean_TEA, alternative = "bilateral")
```

5.5 Resultat i interpretació

Mostreu el resultat i interpreteu-lo, tot donant resposta a la pregunta plantejada.

```r
Res <- tibble(`Valor Observat` = Result$ValorObservat,
              `Valors Critics` = paste(Result$ValorsCritics, collapse = ", "),
              Pvalor = Result$pvalue)

knitr::kable(Res, caption = "Resultat del t-test de variàncies desconegudes diferents")
```

Table 9: Resultat del t-test de variàncies desconegudes diferents

Valor Observat	Valors Critics	Pvalor
4.02708	-2.16748924196297, 2.16748924196297	0.0015278

Com el valor observat està fora dels límits d'acceptació de la hipòtesi nul · la i el valor p és menor a alfa (0.05), s'accepta la hipòtesi alternativa de que hi ha diferències significatives en els valors promitjos de TEA entre els 10 països amb millor valoració dels emprenedors i els 10 països amb pitjor valoració.

6 Conclusions

Escriviu breument les conclusions de l'anàlisi, resumint les respostes a les preguntes de recerca.

L'anàlisi mostra que:

- ***La variança del percentatge de la població entre 18-64 en Europa que són emprenedors o tenen un negoci propi és diferent a la que s'observa en els altres continents.***
- ***Les mitjanes de: OPP, PC, EI, TEA i OWN; són inferiors a Europa que a la resta de continents, per contra, la mitjana de FAIL és similar en tots els continents.***
- ***Els valors promitjos de TEA dels diferents països no han augmentat significativament en cinc anys.***
- **Hi ha diferències significatives en els valors promitjos de TEA entre els 10 països amb millor valoració dels emprenedors i els 10 països amb pitjor valoració.**