TRABALHO DE IAA004 – Estatística Aplicada I

Equipe 03:

- Gustavo Costa de Souza
- Marcos Vinicius de Melo
- Marcus Eneas Silveira Galvao do Rio Apa II
- Patricia Verdugo Pascoal
- Rodrigo de Araujo
- William de Souza Alencar

1 Gráficos e tabelas

a) Elaborar os gráficos box-plot e histograma das variáveis "age" (idade da esposa) e "husage" (idade do marido) e comparar os resultados

```
#install.packages("car")
#install.packages("fdth")
#install.packages("gt")
library(car)
```

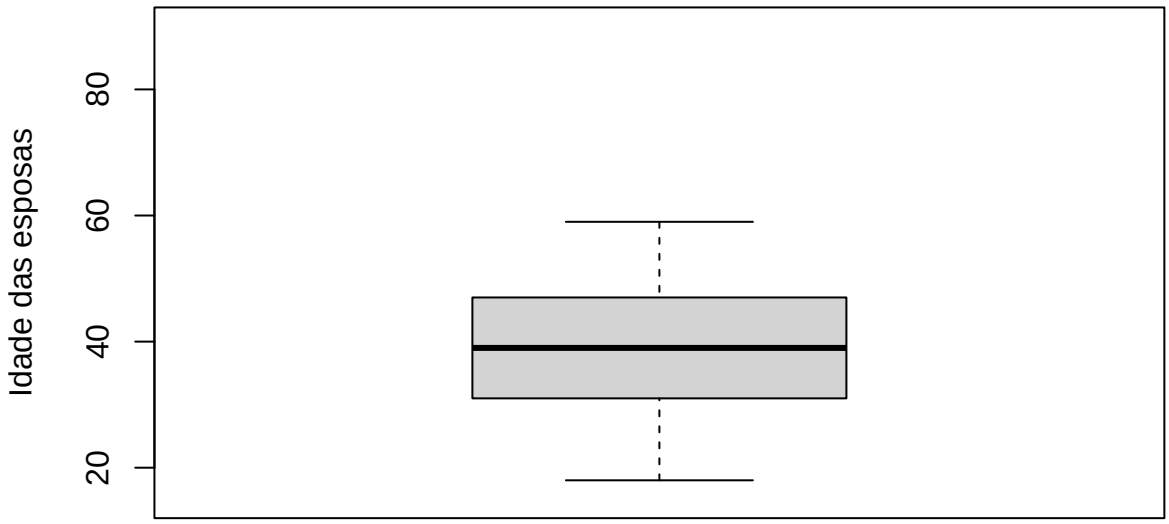
```
## Loading required package: carData
```

```
suppressPackageStartupMessages(library(fdth))

load("salarios.RData")

summary(salarios)
```

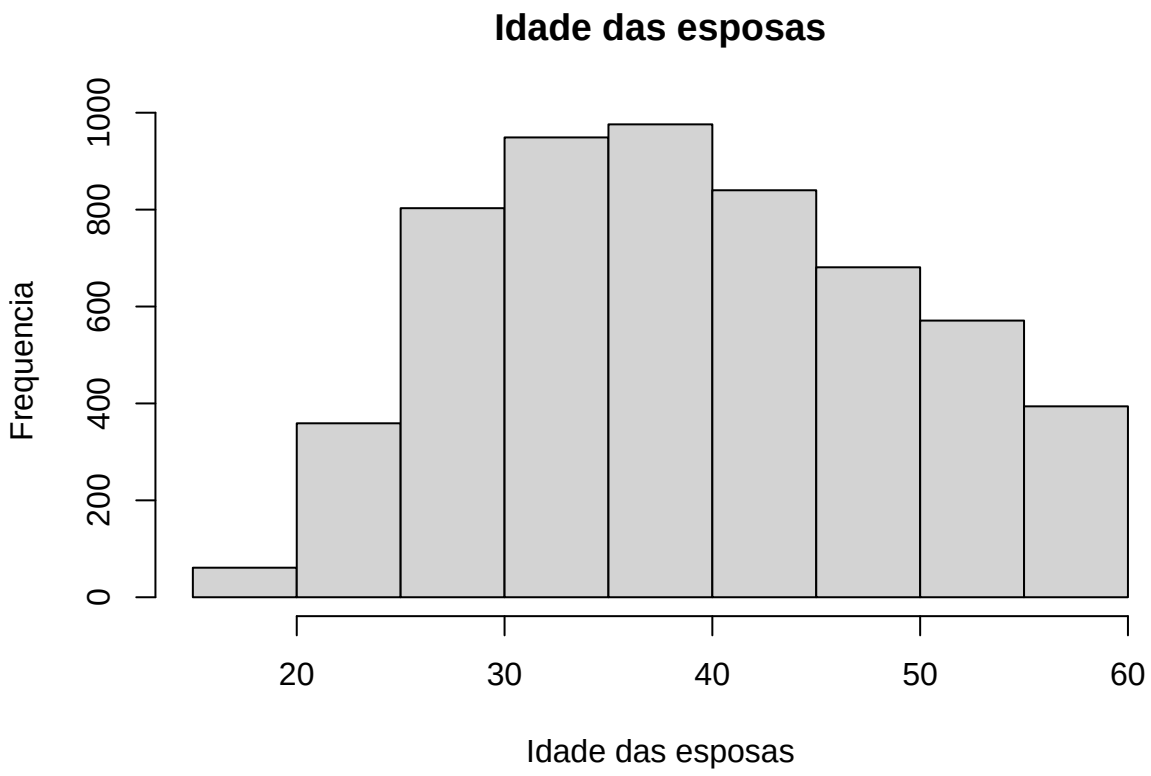
```
##      husage         husunion         husearns         huseduc
##  Min.   :19.00   Min.   :0.0000   Min.   :   0.0   Min.   : 0.00
##  1st Qu.:34.00   1st Qu.:0.0000   1st Qu.:   0.0   1st Qu.:12.00
##  Median :41.00   Median :0.0000   Median : 418.5   Median :12.00
##  Mean   :42.45   Mean   :0.2324   Mean   : 453.5   Mean   :13.15
##  3rd Qu.:50.00   3rd Qu.:0.0000   3rd Qu.: 675.0   3rd Qu.:16.00
##  Max.   :86.00   Max.   :1.0000   Max.   :1923.0   Max.   :18.00
##                  NA's   :1486
##     husblck           hushisp            hushrs          kidge6
##  Min.   :0.00000   Min.   :0.00000   Min.   : 0.00   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:37.00   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :40.00   Median :0.0000
##  Mean   :0.05946   Mean   :0.06621   Mean   :37.88   Mean   :0.3076
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:50.00   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :99.00   Max.   :1.0000
##
##      earns             age             black             educ
##  Min.   :   0.0   Min.   :18.00   Min.   :0.00000   Min.   : 0.00
##  1st Qu.:   0.0   1st Qu.:31.00   1st Qu.:0.00000   1st Qu.:12.00
##  Median : 185.0   Median :39.00   Median :0.00000   Median :12.00
```

```
##  Mean   : 232.8   Mean   :39.43   Mean   :0.05733   Mean   :12.98
##  3rd Qu.: 380.0   3rd Qu.:47.00   3rd Qu.:0.00000   3rd Qu.:15.00
##  Max.   :2884.5   Max.   :59.00   Max.   :1.00000   Max.   :18.00
##
##     hispanic           union            faminc           husexp
##  Min.   :0.00000   Min.   :0.0000   Min.   :     0   Min.   : 0.00
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 22500   1st Qu.:14.00
##  Median :0.00000   Median :0.0000   Median : 37500   Median :22.00
##  Mean   :0.07029   Mean   :0.1501   Mean   : 40993   Mean   :23.31
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 55000   3rd Qu.:32.00
##  Max.   :1.00000   Max.   :1.0000   Max.   :112500   Max.   :72.00
##                    NA's   :2076
##      exper           kidlt6           hours           nwifeinc
##  Min.   : 0.00   Min.   :0.0000   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:12.00   1st Qu.:0.0000   1st Qu.:  0.00   1st Qu.: 11.50
##  Median :19.00   Median :0.0000   Median : 24.00   Median : 24.20
##  Mean   :20.44   Mean   :0.2794   Mean   : 20.72   Mean   : 30.27
##  3rd Qu.:29.00   3rd Qu.:1.0000   3rd Qu.: 40.00   3rd Qu.: 40.17
##  Max.   :52.00   Max.   :1.0000   Max.   :120.00   Max.   :112.50
##
##       inlf            hrwage             lwage
##  Min.   :0.0000   Min.   :  0.0333   Min.   :-3.401
##  1st Qu.:0.0000   1st Qu.:  6.2500   1st Qu.: 1.833
##  Median :1.0000   Median :  8.7500   Median : 2.169
##  Mean   :0.5832   Mean   : 10.3672   Mean   : 2.196
##  3rd Qu.:1.0000   3rd Qu.: 12.5000   3rd Qu.: 2.526
##  Max.   :1.0000   Max.   :200.0000   Max.   : 5.298
##                   NA's   :2348       NA's   :2348
```

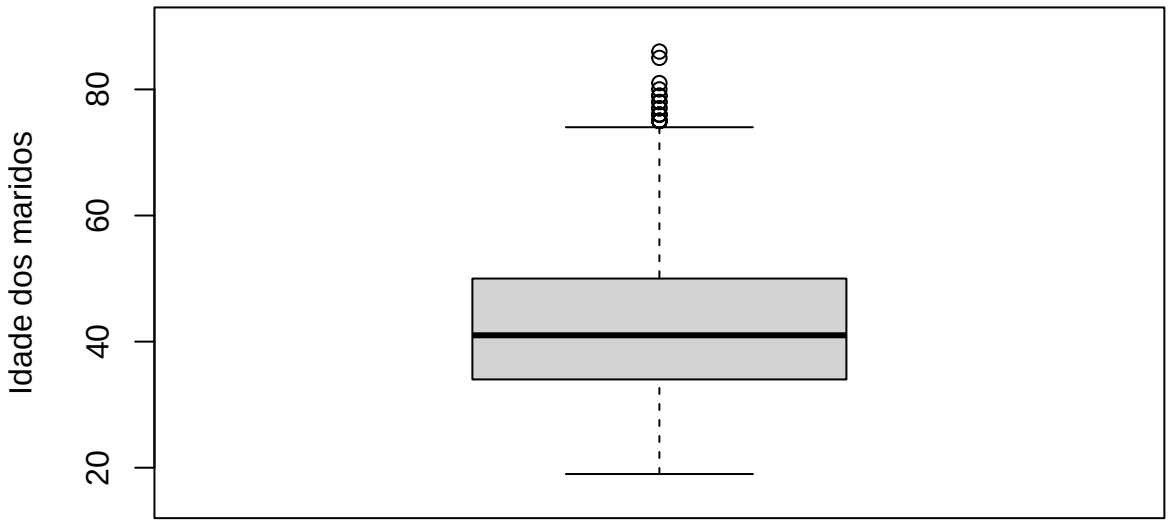
```
Boxplot( ~ age, data=salarios, id=list(method="y"),
         ylab="Idade das esposas",
         ylim = c(15, 90))
```



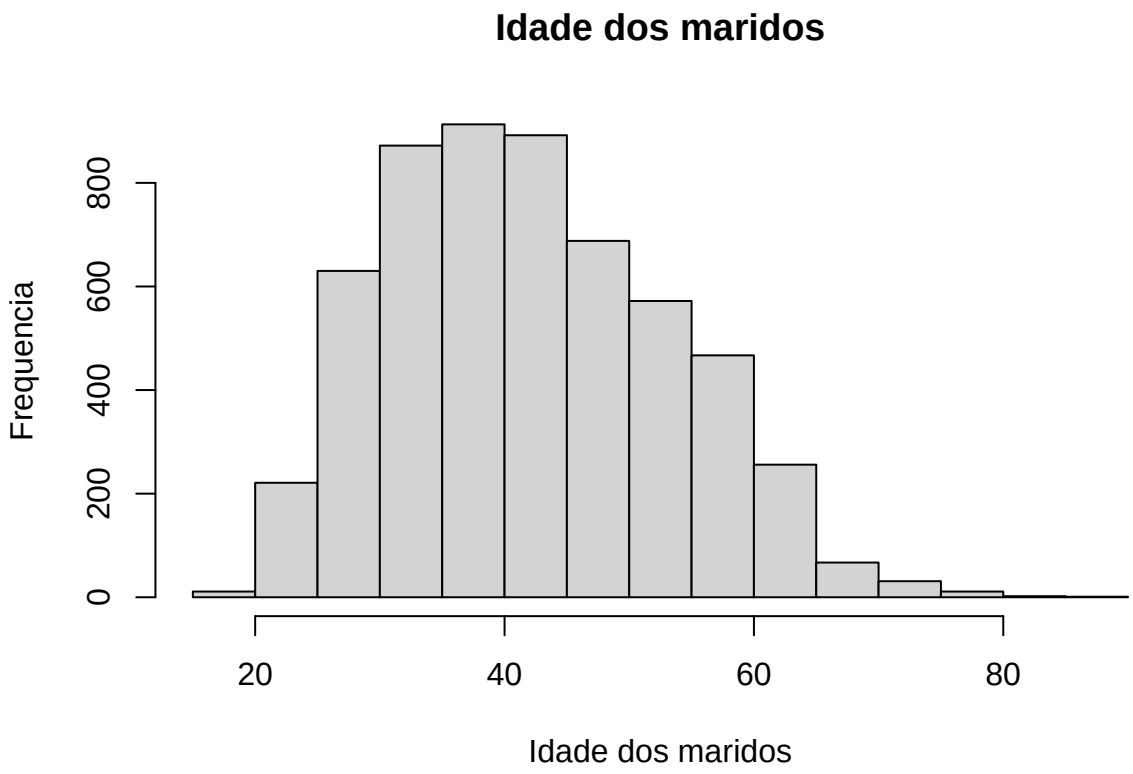

```
hist(salarios$age, main="Idade das esposas", breaks = 10, xlab="Idade das esposas",
     ylab = "Frequencia")
```




```
Boxplot( ~ husage, data=salarios, id=list(n=0),
         ylab="Idade dos maridos",
         ylim = c(15, 90))
```




```
hist(salarios$husage, main="Idade dos maridos", xlab="Idade dos maridos", ylab = "Frequencia", breaks =
```



As principais comparações do estudo são:

- A idade mediana das esposas é menor que a dos maridos.
- O boxplot da idade dos maridos apresentam uma maior dispersão dos dados.
- A distribuição de idades das esposas é mais simétrica que a dos maridos.
- No conjunto de idades dos maridos alguns outliers são apresentados.

b) Elaborar a tabela de frequencias das variáveis "age" (idade da esposa) e "husage" (idade do marido) e comparar os resultados

```
tabela_feq_idade_esposa <- fdt(salarios$age)
print("Tabela de frequencia das idades das esposas")
```

```
## [1] "Tabela de frequencia das idades das esposas"
```

```
print(tabela_feq_idade_esposa)
```

```
##     Class limits   f   rf rf(%)   cf  cf(%)
##   [17.82,20.804)  61 0.01  1.08   61   1.08
##  [20.804,23.787) 161 0.03  2.86  222   3.94
##  [23.787,26.771) 312 0.06  5.54  534   9.48
##  [26.771,29.754) 505 0.09  8.96 1039  18.44
##  [29.754,32.738) 562 0.10  9.98 1601  28.42
##  [32.738,35.721) 571 0.10 10.13 2172  38.55
##  [35.721,38.705) 624 0.11 11.08 2796  49.63
##  [38.705,41.689) 510 0.09  9.05 3306  58.68
##  [41.689,44.672) 542 0.10  9.62 3848  68.30
```

```
##  [44.672,47.656) 432 0.08  7.67 4280  75.97
##  [47.656,50.639) 389 0.07  6.90 4669  82.87
##  [50.639,53.623) 358 0.06  6.35 5027  89.23
##  [53.623,56.606) 304 0.05  5.40 5331  94.62
##   [56.606,59.59) 303 0.05  5.38 5634 100.00
```

```
tabela_feq_idade_marido <- fdt(salarios$husage)
print("Tabela de frequencia das idades dos maridos")
```

```
## [1] "Tabela de frequencia das idades dos maridos"
```

```
print(tabela_feq_idade_marido)
```

```
##     Class limits   f   rf rf(%)   cf  cf(%)
##   [18.81,23.671) 102 0.02  1.81  102   1.81
##  [23.671,28.531) 466 0.08  8.27  568  10.08
##  [28.531,33.392) 809 0.14 14.36 1377  24.44
##  [33.392,38.253) 895 0.16 15.89 2272  40.33
##  [38.253,43.114) 917 0.16 16.28 3189  56.60
##  [43.114,47.974) 629 0.11 11.16 3818  67.77
##  [47.974,52.835) 649 0.12 11.52 4467  79.29
##  [52.835,57.696) 541 0.10  9.60 5008  88.89
##  [57.696,62.556) 394 0.07  6.99 5402  95.88
##  [62.556,67.417) 152 0.03  2.70 5554  98.58
##  [67.417,72.278)  51 0.01  0.91 5605  99.49
##  [72.278,77.139)  21 0.00  0.37 5626  99.86
##  [77.139,81.999)   6 0.00  0.11 5632  99.96
##   [81.999,86.86)   2 0.00  0.04 5634 100.00
```

As principais conclusões do estudo são:

- A idade mais comum entre esposas está entre 35–39 anos, e entre maridos, em uma faixa mais alta 38–43 anos.
- As idades dos maridos possuem uma maior amplitude da distribuição, sendo 18 a 86 frente a 17 a 59 das esposas.
- As idades dos maridos possuem uma leve assimetria positiva (tendência simétrica com cauda à direita).
- A frequência acumulada mostra que 50% das esposas estão abaixo de aproximadamente 38 anos, enquanto para maridos esse ponto de mediana está mais próximo de 40–43 anos.
- O grupo de maridos apresenta alguns individuos muito idosos, o que não ocorre com as esposas.

---

2 Medidas de posição e dispersão

a) Calcular a média, mediana e moda das variáveis "age" (idade da esposa) e "husage" (idade do marido) e comparar os resultados

```
media_esposas <- mean(salarios$age)
cat("Media idade das esposas: ", media_esposas, "\n")
```

```
## Media idade das esposas:  39.42758
```

```
media_maridos <- mean(salarios$husage)
cat("Media idade dos maridos: ", media_maridos, "\n")
```

```
## Media idade dos maridos:  42.45296
```

```r
cat("Mediana idade das esposas: ", median(salarios$age), "\n")
```

```
## Mediana idade das esposas:  39
```

```r
cat("Mediana idade dos maridos: ", median(salarios$husage), "\n")
```

```
## Mediana idade dos maridos:  41
```

```r
moda_esposas <- subset(table(salarios$age),
       table(salarios$age) == max(table(salarios$age)))
cat("Moda idade das esposas: ", names(moda_esposas), "com ", moda_esposas[1], "pessoas.","\n")
```

```
## Moda idade das esposas:  37 com  217 pessoas.
```

```r
moda_maridos <- subset(table(salarios$husage),
       table(salarios$husage) == max(table(salarios$husage)))
cat("Moda idade dos maridos: ", names(moda_maridos), "com ", moda_maridos[1], "pessoas.","\n")
```

```
## Moda idade dos maridos:  44 com  201 pessoas.
```

As principais conclusões do estudo são:

- A idade média dos maridos é 7.67% maior que das esposas. Isso sugere que que os maridos são mais velhos que as esposas.
- A idade mediana dos maridos e das esposas é muito próximas, indicando simetria na distribuição de idades ou no máximo levemente inclinada.
- A moda da idade dos maridos é 19% maior que das esposas, indica que a faixa etária dos maridos está concentrada em uma idade superior das esposas.

b) Calcular a variância, desvio padrão e coeficiente de variação das variáveis "age" (idade da esposa) e "husage" (idade do marido) e comparar os resultados

```r
sd_esposas <- sd(salarios$age)
cat("Desvio padrao das idades das esposas: ", sd_esposas, "\n")
```

```
## Desvio padrao das idades das esposas:  9.98761
```

```r
sd_maridos <- sd(salarios$husage)
cat("Desvio padrao das idades dos maridos: ", sd_maridos, "\n")
```

```
## Desvio padrao das idades dos maridos:  11.22817
```

```r
cv_esposas <- (sd_esposas/media_esposas) * 100
cat("Coeficiente de variacao das das idades das esposas: ", sd_esposas, "\n")
```

```
## Coeficiente de variacao das das idades das esposas:  9.98761
```

```r
cv_maridos <- (sd_maridos/media_maridos) * 100
cat("Coeficiente de variacao das das idades dos maridos: ", cv_maridos, "\n")
```

```
## Coeficiente de variacao das das idades dos maridos:  26.44849
```

As principais conclusões do estudo são:

- O desvio padrão das idades dos maridos é 12.4% maior que das esposas, o que sugere uma maior variação nas idades dos maridos que das esposas.
- As idades das esposas variam pouco na amostra, já dos maridos tem uma média dispersão dentro da amostra.
- O coeficiente de variação dos maridos é maior que das esposas, o que indica uma maior variabilidade das idades dos maridos.

3 Testes paramétricos ou não paramétricos

a) Testar se as médias (se você escolher o teste paramétrico) ou as medianas (se você escolher o teste não paramétrico) das variáveis "age" (idade da esposa) e "husage" (idade do marido) são iguais, construir os intervalos de confiança e comparar os resultados.

Obs:

1) Você deve fazer os testes necessários (e mostra-los no documento pdf) para saber se você deve usar o unpaired test (paramétrico) ou o teste U de Mann-Whitney (não paramétrico), justifique sua resposta sobre a escolha.
2) Lembre-se de que os intervalos de confiança já são mostrados nos resultados dos testes citados no item 1 acima.

```
#install.packages("BSDA")
#install.packages("onewaytests")
#install.packages("sjPlot")
#install.packages("devtools")
#devtools::install_github("homerhanumat/tigerstats")
#install.packages("misty")
#install.packages("ggpubr")
#install.packages("dplyr")

suppressPackageStartupMessages(library("BSDA"))
suppressPackageStartupMessages(library("onewaytests"))
suppressPackageStartupMessages(library("sjPlot"))
suppressPackageStartupMessages(library("tigerstats"))
suppressPackageStartupMessages(library("misty"))
suppressPackageStartupMessages(library("ggpubr"))
suppressPackageStartupMessages(library("dplyr"))
```

Vamos realizar os testes necessários para utilizar o modelo paramétrico:

- Amostras independentes
- normalidade
- homogeneidade das variancias entre grupos

Premissa 1: As duas amostras sao independentes? Sim, pois os grupos de esposas e maridos nao estao relacionados. Nao se trata de uma amostra ou grupos emparelhados.

Premissa 2: Os dados de cada amostra/grupo possuem distribuicao normal? Vamos usar o teste Kolmogorov-Smirnov para descobrir se a idade das esposas segue uma distribuição normal.

obs: Não foi utilizado Shapiro-Wilk pois possui uma limitação de 5000 individuos na amostra.

teste de hipoteses:

- H0: os dados sao normalmente distribuidos
- Ha: os dados nao sao normalmente distribuidos

```
options(scipen = 999)

ks.test(salarios$age, "pnorm", mean = mean(salarios$age), sd = sd(salarios$age))
```

```
## Warning in ks.test.default(salarios$age, "pnorm", mean = mean(salarios$age), :
## ties should not be present for the one-sample Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
```

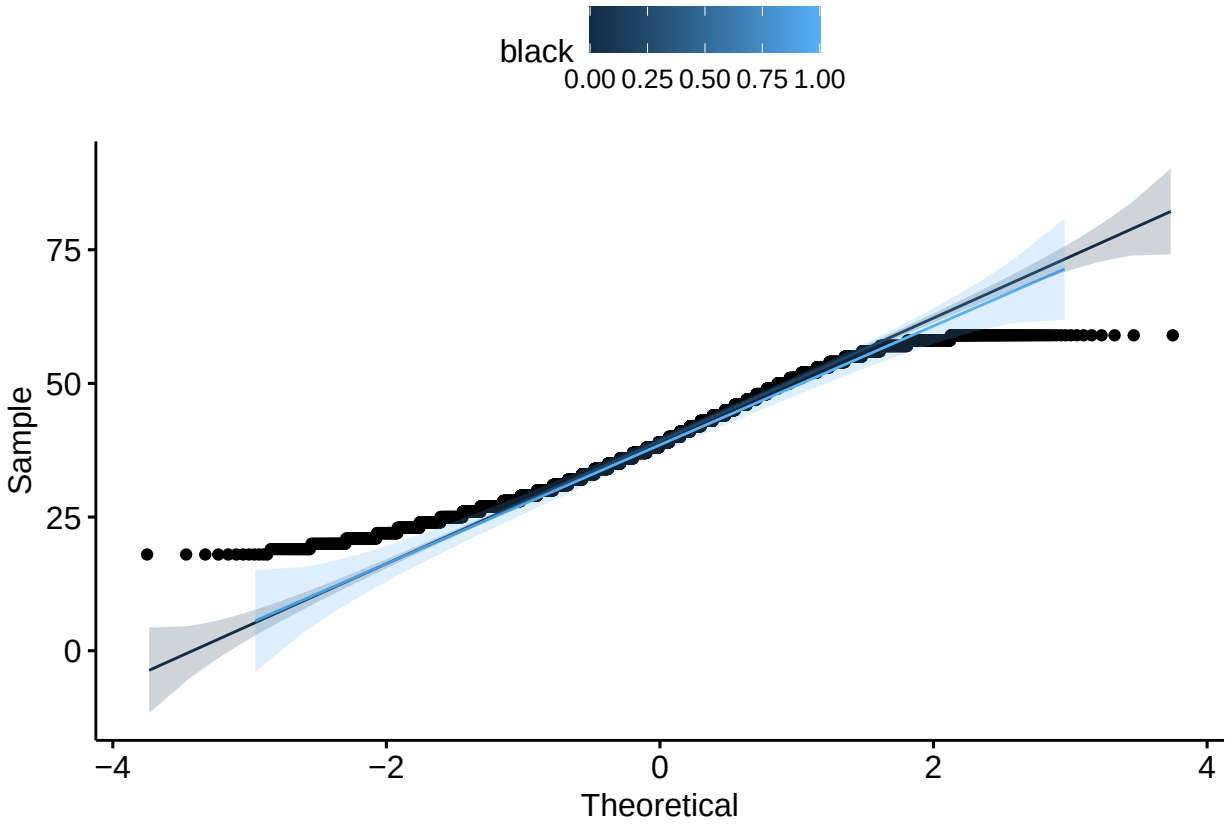
```
##
## data:  salarios$age
## D = 0.058909, p-value < 0.00000000000000022
## alternative hypothesis: two-sided
```

Como pode ser observado o resultado do teste obteve p-value = 0,00000000000000022 que é menor que 0,05 logo os dados de idade das esposas não está normalmente distribuido, portanto não satifaz as restrições para utilização de modelos paramétricos.

Vamos visualizar o resultado do teste no gráfico QQ normal.

```
ggqqplot(salarios, "age")
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



No gráfico QQ normal é possível ver que muitos pontos estão fora da linha de referência, logo a amostra não segue uma distribuição normal, portanto **não é possível aplicar os testes paramétricos**.

Teste de requisitos impedem o uso de modelos paramétricos, logo o modelo não paramétrico U de Mann-Whitney será utilizado para testar a hipótese abaixo.

a) Testar se as medianas das variáveis "age" (idade da esposa) e "husage" (idade do marido) são iguais, construir os intervalos de confiança e comparar os resultados.

```
# criando o data-frame longo.
idades <- data.frame(
  group = factor(c(rep("esposa", length(salarios$age)), rep("marido", length(salarios$husage)))),
  age = c(salarios$age, salarios$husage)
)

levels(idades$group)
```

```
## [1] "esposa" "marido"
```

```
# Reordenando os niveis
idades$group <- ordered(idades$group,
                        levels = c("esposa", "marido"))

summary(idades)
```

```
##     group           age
##  esposa:5634   Min.   :18.00
##  marido:5634   1st Qu.:33.00
##                Median :40.00
##                Mean   :40.94
##                3rd Qu.:49.00
##                Max.   :86.00
```

```
head(idades)
```

```
##    group age
## 1 esposa  43
## 2 esposa  26
## 3 esposa  49
## 4 esposa  35
## 5 esposa  43
## 6 esposa  58
```

Vamos calcular um sumário estatistico

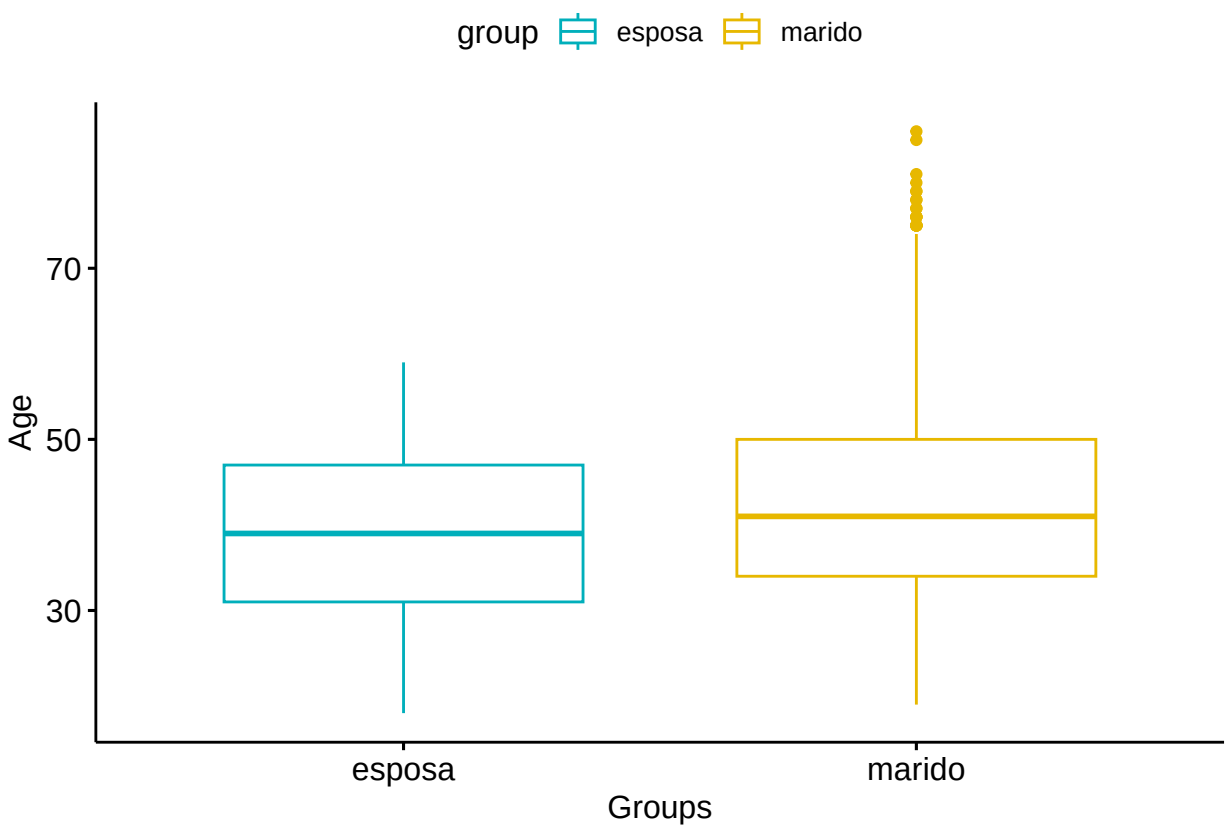
```
group_by(idades, group) %>%
  summarise(
    count = n(),
    median = median(age, na.rm = TRUE),
    IQR = IQR(age, na.rm = TRUE)
  )
```

```
## # A tibble: 2 x 4
##   group  count median   IQR
##   <ord>  <int>  <dbl> <dbl>
## 1 esposa  5634     39    16
## 2 marido  5634     41    16
```

Vamos visualizar os dados usando box-plots, plotaremos a "age" por "group"

```
ggboxplot(idades, x = "group", y = "age",
          color = "group", palette=c("#00AFBB", "#E7B800"),
```

```
          ylab = "Age", xlab = "Groups")
```



Vamos fazer o teste se a idade mediana das esposas eh igual a idade mediana dos maridos

Hipoteses do teste:

- H0: Nao existe diferenca entre as medianas dos grupos
- Ha: Existe diferenca entre as medianas dos grupos

```
res <- wilcox.test(age ~ group, data = idades, exact = FALSE, conf.int=TRUE)
res
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by group
## W = 13619912, p-value < 0.00000000000000022
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -3.000024 -2.000033
## sample estimates:
## difference in location
##              -2.999966
```

As principais conclusões do estudo são:

O p-value do teste eh 0,00000000000000022, que eh menor que o nivel de significancia 0,05, logo podemos concluir que a idade mediana das esposas e dos maridos eh estatisticamente diferente (rejeitamos H0). O intervalo de confianca da diferenca entre as medianas esta entre -3 e -2 com uma mediana de -2,99.