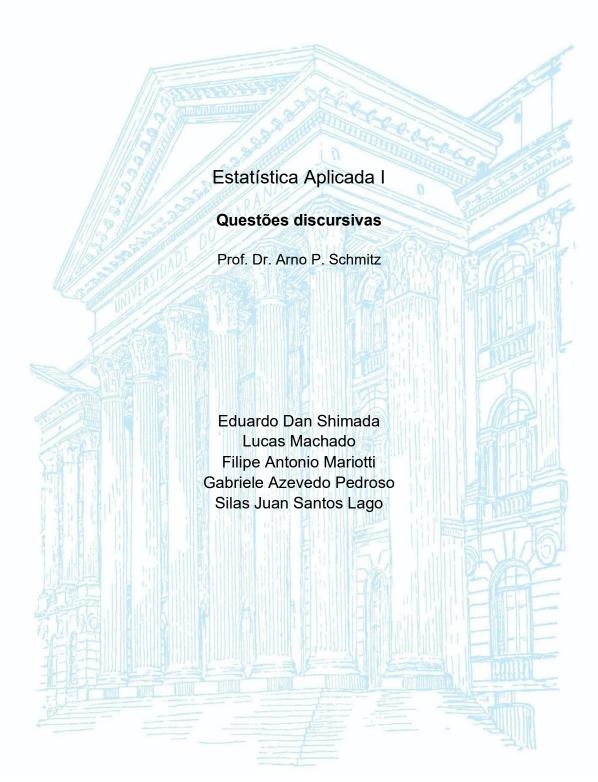
## UNIVERSIDADE FEDERAL DO PARANÁ



CURITIBA 2025





## TRABALHO DE IAA004 - Estatística Aplicada I

Este trabalho deve ser realizado em equipes de no mínimo 3 até no máximo 6 integrantes. Adicionar o NOME COMPLETO de todos os integrantes da equipe. Entregar como um documento PDF com todas as respostas. Para execução do trabalho usar a base de dados "salarios.RData"

## 1 Gráficos e tabelas

a) (15 pontos) Elaborar os gráficos box-plot e histograma das variáveis "age" (idade da esposa) e "husage" (idade do marido) e comparar os resultados

#### Código:

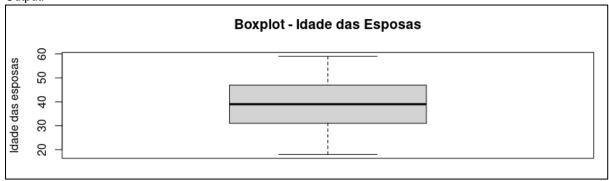
```
# a) Boxplot e Histograma para "age" e "husage"

# Instalando o pacote que gera o boxplot
install.packages("car")

# Carregando o pacote boxplot
library(car)

# Gerando boxplot das esposas
Boxplot(
    ~age,
    data = salarios,
    id = list(method = "y"),
    ylab = "Idade das esposas",
    main = "Boxplot - Idade das Esposas"
```

### Output:







```
# Gerando boxplot dos maridos
Boxplot(
   ~husage,
   data = salarios,
   id = list(method = "y"),
   ylab = "Idade dos maridos",
   main = "Boxplot - Idade dos Maridos"
)
```

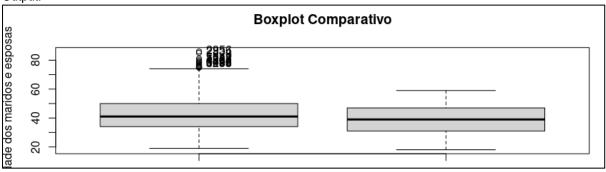
### Output:

```
Boxplot - Idade dos Maridos
```

#### Código:

```
# Gerando boxplot das esposas e maridos
Boxplot(
    ~husage + ~age,
    data = salarios,
    id = list(method = "y"),
    ylab = "Idade dos maridos e esposas",
    main = "Boxplot Comparativo"
)
```

#### Output:

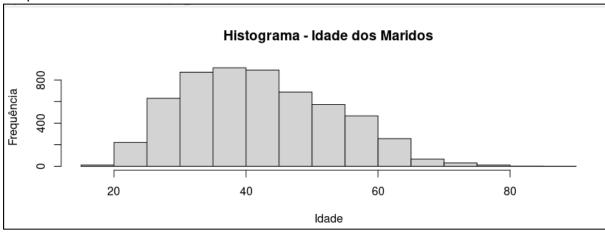






```
# Histograma - Idade dos Maridos
hist(
  salarios$husage,
  main = "Histograma - Idade dos Maridos",
  xlab = "Idade",
  ylab = "Frequência"
```

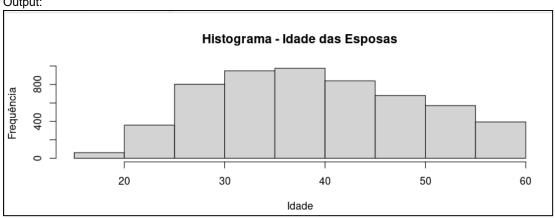
### Output:



### Código:

```
# Histograma - Idade das Esposas
hist(
 salarios$age,
 main = "Histograma - Idade das Esposas",
 xlab = "Idade",
  ylab = "Frequência"
```

#### Output:



### Comparação:

As idades dos maridos são mais altas em média.





Há maior dispersão entre as idades dos homens.

**b)** (15 pontos) Elaborar a tabela de frequências das variáveis "age" (idade da esposa) e "husage" (idade do marido) e comparar os resultados

#### Código:

```
# b) Tabela de Frequência para "age" e "husage"

# Instalar e carregar pacote
install.packages("fdth")
library(fdth)

# Tabela - Idade dos Maridos
freq_marido <- fdt(salarios$husage)
print(freq_marido)</pre>
```

#### Output:

```
> print(freq_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
 [33.392,38.253) 895 0.16 15.89 2272
                                    40.33
 [38.253,43.114) 917 0.16 16.28 3189
 [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
```

#### Código:

```
# Tabela - Idade das Esposas
freq_esposa <- fdt(salarios$age)
print(freq_esposa)</pre>
```

#### Output:





```
> print(freq_esposa)
   Class limits f
                      rf rf(%)
                                cf
                                    cf(%)
 [17.82,20.804) 61 0.01 1.08
                                     1.08
[20.804,23.787) 161 0.03 2.86
[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
[32.738,35.721) 571 0.10 10.13 2172
[35.721,38.705) 624 0.11 11.08 2796
[38.705,41.689) 510 0.09 9.05 3306
[41.689,44.672) 542 0.10 9.62 3848
                                    68.30
[44.672,47.656) 432 0.08 7.67 4280
[47.656,50.639) 389 0.07 6.90 4669
[50.639,53.623) 358 0.06 6.35 5027
[53.623,56.606) 304 0.05 5.40 5331 94.62
 [56.606,59.59) 303 0.05 5.38 5634 100.00
```

## Comparação:

Podemos verificar que a faixa etária dos homens é maior em relação às das esposas, onde a idade dos homens *varia entre 18 anos e 81 anos* enquanto das mulheres varia *entre 17 anos e 56 anos*.

Nota-se também que a moda dos homens e mulheres são quase da mesma idade, onde os homens apresentam maior frequência na idade de 38 anos, com valor  $\mathbf{f} = 917$ , já nas mulheres apresenta maior frequência na idade de 35 anos, com valor  $\mathbf{f} = 624$ .

## 2 Medidas de posição e dispersão

a) (15 pontos) Calcular a média, mediana e moda das variáveis "age" (idade da esposa) e "husage" (idade do marido) e comparar os resultados

```
# a) Média, Mediana e Moda

# Média
mean(salarios$husage) # Marido
mean(salarios$age) # Esposa

Output:
> # Média
```

```
Output:

> # Média

> mean(salarios$husage) # Marido

[1] 42.45296

> mean(salarios$age) # Esposa

[1] 39.42758

>
```





#### Código:

```
# Mediana
median(salarios$husage)
median(salarios$age)
```

#### Output:

```
> # Mediana
> median(salarios$husage)
[1] 41
> median(salarios$age)
[1] 39
> |
```

#### Código:

```
# Moda
table(salarios$husage)
subset(table(salarios$husage), table(salarios$husage) == max(table(salarios$husage)))

table(salarios$age)
subset(table(salarios$age), table(salarios$age) == max(table(salarios$age)))
```

#### Output:

#### Comparação:





```
# Cálculo de medidas
media_h <- mean(salarios$husage)</pre>
media_m <- mean(salarios$age)
mediana_h <- median(salarios$husage)</pre>
mediana m <- median(salarios$age)</pre>
moda_h <- as.numeric(names(which.max(table(salarios$husage))))</pre>
moda_m <- as.numeric(names(which.max(table(salarios$age))))</pre>
# Diferenças percentuais
media diff <- ((media h / media m) - 1) * 100
mediana_diff <- ((mediana_h / mediana_m) - 1) * 100
moda diff <- ((moda h / moda m) - 1) * 100
# Impressão formatada
cat("
--- COMPARAÇÃO ENTRE MARIDOS E ESPOSAS ---
cat("Média - Marido:", round(media_h, 2), "| Esposa:", round(media_m, 2),
   "| Diferença:", round(media_diff, 2), "%
"| Diferença:", round(mediana_diff, 2), "%
cat("Moda - Marido:", moda_h,
                                       "| Esposa:", moda_m,
   "| Diferença:", round(moda_diff, 2), "%
```

#### Output:

```
--- COMPARAÇÃO ENTRE MARIDOS E ESPOSAS ---
> cat("Média - Marido:", round(media_h, 2), "| Esposa:", round(media_m, 2),
+ "| Diferença:", round(media_diff, 2), "%
+ ")

Média - Marido: 42.45 | Esposa: 39.43 | Diferença: 7.67 %
> cat("Mediana - Marido:", mediana_h, "| Esposa:", mediana_m,
+ "| Diferença:", round(mediana_diff, 2), "%
+ ")

Mediana - Marido: 41 | Esposa: 39 | Diferença: 5.13 %
> cat("Moda - Marido:", moda_h, "| Esposa:", moda_m,
+ "| Diferença:", round(moda_diff, 2), "%
+ ")

Moda - Marido: 44 | Esposa: 37 | Diferença: 18.92 %
> |
```

**b)** (15 pontos) Calcular a variância, desvio padrão e coeficiente de variação das variáveis "age" (idade da esposa) e "husage" (idade do marido) e comparar os resultados.





#### Código:

```
# b) Variância, Desvio Padrão e Coeficiente de Variação

# Variância
var(salarios$husage)
var(salarios$age)
```

#### Output:

```
> # Variância
> var(salarios$husage)
[1] 126.0717
> var(salarios$age)
[1] 99.75234
```

#### Código:

```
# Desvio padrão
sd(salarios$husage)
sd(salarios$age)
```

#### Output:

```
> # Desvio padrāo
> sd(salarios$husage)
[1] 11.22817
> sd(salarios$age)
[1] 9.98761
```

### Código:

```
# Coeficiente de Variação
media_h <- mean(salarios$husage)
media_m <- mean(salarios$age)
dp_h <- sd(salarios$husage)
dp_m <- sd(salarios$age)

cv_h <- (dp_h / media_h) * 100
cv_m <- (dp_m / media_m) * 100

cv_h
cv_h</pre>
```

#### Output:

```
> cv_h
[1] 26.44849
> cv_m
[1] 25.33153
>
```

#### Comparação:





#### Código:

```
# Comparação automatizada entre maridos e esposas (Dispersão)
# Cálculo de variâncias
var_h <- var(salarios$husage)</pre>
var_m <- var(salarios$age)</pre>
var_diff <- ((var_h / var_m) - 1) * 100</pre>
# Cálculo de desvios padrão
sd_h <- sd(salarios$husage)
sd_m <- sd(salarios$age)
sd_diff <- ((sd_h / sd_m) - 1) * 100
# Cálculo de coeficientes de variação
cv_h <- (sd_h / media_h) * 100
cv_m <- (sd_m / media_m) * 100
cv_diff <- ((cv_h / cv_m) - 1) * 100
# Impressão formatada
cat("
--- COMPARAÇÃO DE DISPERSÃO ---
cat("Variância - Marido:", round(var_h, 2), "| Esposa:", round(var_m, 2),
     "| Diferença:", round(var_diff, 2), "%
cat("Desvio Padrão - Marido:", round(sd_h, 2), "| Esposa:", round(sd_m, 2),
    "| Diferença:", round(sd_diff, 2), "%
cat("Coef. Variação - Marido:", round(cv_h, 2), "% | Esposa:", round(cv_m, 2), "%",
    "| Diferença:", round(cv_diff, 2), "%
```

#### Output:

## 3 Testes paramétricos ou não paramétricos

a) (40 pontos) 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 mostrá-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.

#### Código:

```
# Instalar e carregar pacote
library(nortest)

tabela_aux_esposa <- data.frame(idade = salarios$age)
tabela_aux_esposa$group <- "esposa"

tabela_aux_marido <- data.frame(idade = salarios$husage)
tabela_aux_marido$group <- "marido"

nova_tabela <- rbind(tabela_aux_esposa, tabela_aux_marido)

# Cálculo da média da idade entre as esposas e maridos juntos
mean(nova_tabela$idade)

library(dplyr)
group_by(nova_tabela, group) %>% summarise(
    count = n(),
    mean = mean(idade, na.rm = TRUE),
    sd = sd(idade, na.rm = TRUE)
)
```

Output:





```
> # Instalar e carregar pacote
> library(nortest)
> tabela aux esposa <- data.frame(idade = salarios$age)</pre>
> tabela_aux_esposa$group <- "esposa"
> tabela_aux_marido <- data.frame(idade = salarios$husage)</pre>
> tabela_aux_marido$group <- "marido"
> nova_tabela <- rbind(tabela_aux_esposa, tabela_aux_marido)</pre>
> # Cálculo da média da idade entre as esposas e maridos juntos
> mean(nova_tabela$idade)
[1] 40.94027
> library(dplyr)
> group_by(nova_tabela, group) %>% summarise(
+ count = n(),
   mean = mean(idade, na.rm = TRUE),
  sd = sd(idade, na.rm = TRUE)
+ )
# A tibble: 2 × 4
 group count mean
 <chr> <int> <dbl> <dbl>
1 esposa <u>5</u>634 39.4 9.99
2 marido <u>5</u>634 42.5 11.2
```





```
# Teste de normalidade (Anderson-Darling)
cat("
--- TESTE DE NORMALIDADE (Anderson-Darling) ---
")

normalidade_h <- ad.test(salarios$husage)
normalidade_m <- ad.test(salarios$age)

cat("p-valor Marido:", format.pval(normalidade_h$p.value, digits = 4), "
")
cat("p-valor Esposa:", format.pval(normalidade_m$p.value, digits = 4), "
")

if (normalidade_h$p.value < 0.05 | normalidade_m$p.value < 0.05) {
   cat("Conclusão: Pelo menos uma das variáveis não segue distribuição normal.")
   cat("Utilizaremos o teste não-paramétrico de Mann-Whitney.
")
} else {
   cat("Conclusão: Ambas as variáveis seguem distribuição normal. Poderíamos usar o teste t.")
}</pre>
```

#### Output:

```
--- TESTE DE NORMALIDADE (Anderson-Darling) ---
> normalidade_h <- ad.test(salarios$husage)
> normalidade_m <- ad.test(salarios$age)</pre>
> cat("p-valor Marido:", format.pval(normalidade_h$p.value, digits = 4), "
+ ")
p-valor Marido: < 2.2e-16
> cat("p-valor Esposa:", format.pval(normalidade_m$p.value, digits = 4), "
+ ")
p-valor Esposa: < 2.2e-16
> if (normalidade_h$p.value < 0.05 | normalidade_m$p.value < 0.05) {
   cat("Conclusão: Pelo menos uma das variáveis não segue distribuição normal.
   cat("Utilizaremos o teste não-paramétrico de Mann-Whitney.
+ } else {
   cat("Conclusão: Ambas as variáveis seguem distribuição normal. Poderíamos usar o teste t.
+ }
Conclusão: Pelo menos uma das variáveis não segue distribuição normal.
Utilizaremos o teste não-paramétrico de Mann-Whitney.
```

H0: Ambas as variáveis seguem Distribuição Normal.

Ha: Uma, ou ambas, as variáveis não seguem Distribuição Normal.





Confirmando o que foi analisado pelos histogramas, as distribuições de idade **não seguem** uma distribuição Normal nos grupos. Seu p-valor sendo 2.2e-16, que é *estritamente menor* que 0.05. Dessa forma, **rejeitamos a hipótese H0** de que a distribuição é Normal e seguimos adiante com **Ha** e com o Teste Não Paramétrico "U" de Mann-Whitney.

Código:

**H0**: A mediana da idade de homens e mulheres é estatisticamente igual **Ha**: A mediana da idade de homens e mulheres **NÃO** é estatisticamente igual

```
# Teste de Mann-Whitney (não-paramétrico)
 mann_whitney <- wilcox.test(salarios$husage, salarios$age, alternative = "two.sided", conf.in
 --- RESULTADO TESTE DE MANN-WHITNEY ---
 cat("Estatística W:", round(mann_whitney$statistic, 2), "
  ")
 cat("Valor-p:", format.pval(mann_whitney$p.value, digits = 4), "
  ")
 cat("Intervalo de confiança para a diferença das medianas:
  ")
 print(mann_whitney$conf.int)
if (mann_whitney$p.value < 0.05) {</pre>
   cat("
 Conclusão: As idades dos maridos e esposas são significativamente diferentes (p < 0.05).
 ")
→ } else {
   cat("
 Conclusão: Não há diferença significativa entre as idades (p ≥ 0.05).
^ }
```

Output:





```
--- RESULTADO TESTE DE MANN-WHITNEY ---
> cat("Estatística W:", round(mann_whitney$statistic, 2), "
+ ")
Estatística W: 18122044
> cat("Valor-p:", format.pval(mann_whitney$p.value, digits = 4), "
+ ")
Valor-p: < 2.2e-16
> cat("Intervalo de confiança para a diferença das medianas:
+ ")
Intervalo de confiança para a diferença das medianas:
> print(mann_whitney$conf.int)
[1] 2.000033 3.000024
attr(,"conf.level")
[1] 0.95
>
```

Isto é, H0: *A mediana da idade de homens e mulheres é estatisticamente igual* é **rejeitada** pois seu p-valor é 2.2e-16, que é estritamente menor que 0.05. Com um intervalo de confiança de 95% - [2.000033, 3.000024].

Logo, **aceitamos** Ha: *A mediana da idade de homens e mulheres* **NÃO** é estatisticamente igual.