# 1 Gráficos e tabelas

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

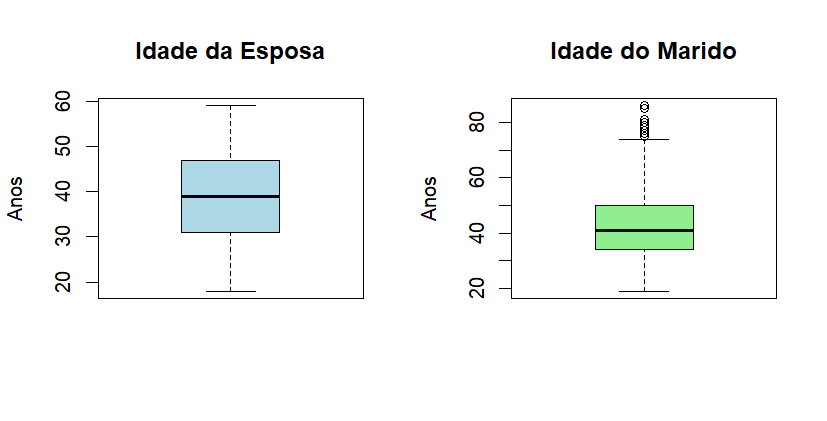
## Boxplot para idades

# Resetar configuração de plotagem

par(mfrow=c(1,2))

boxplot(salarios$age, main="Idade da Esposa", ylab="Anos", col="lightblue")

boxplot(salarios$husage, main="Idade do Marido", ylab="Anos", col="lightgreen")



cat("1.a) COMPARAÇÃO DOS BOXPLOTS-Amplitude total (diferença entre máximo e mínimo):\n")

cat(" Esposas:", diff(range(salarios$age)), "anos\n")

cat(" Maridos:", diff(range(salarios$husage)), "anos\n\n")

Comparação dos BOXPLOTS-Amplitude total (diferença entre máximo e mínimo):

Esposas: 41 anos

Maridos: 67 anos

cat("- Mediana (linha central no boxplot):\n")

cat(" Esposas:", median(salarios$age), "anos\n")

cat(" Maridos:", median(salarios$husage), "anos\n\n")

Esposas: 39 anos

Maridos: 41 anos

cat("- Intervalo interquartil (IQR, altura da caixa):\n")

cat(" Esposas:", IQR(salarios$age), "anos\n")

cat(" Maridos:", IQR(salarios$husage), "anos\n\n")

Intervalo interquartil (IQR, altura da caixa):

Esposas: 16 anos

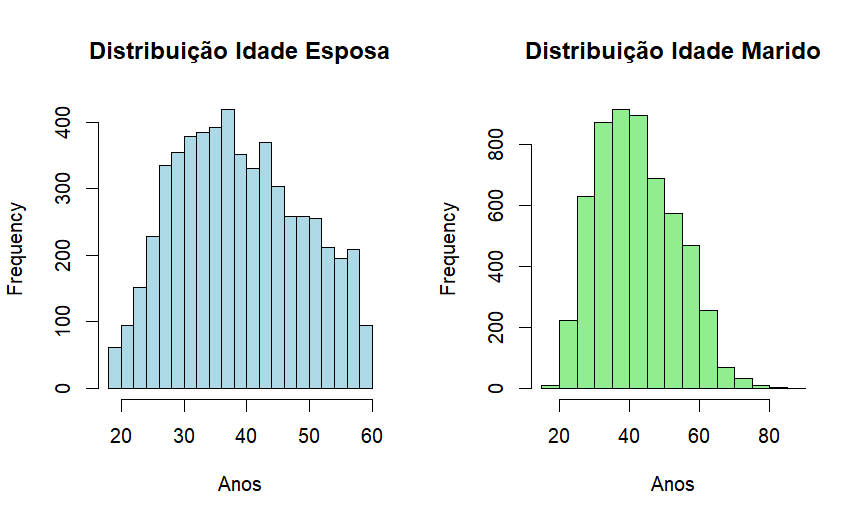
Maridos: 16 anos

## Histogramas

par(mfrow=c(1,2)) # Resetar configuração de plotagem

hist(salarios$age, main="Distribuição Idade Esposa", xlab="Anos", col="lightblue", breaks=15)

hist(salarios$husage, main="Distribuição Idade Marido", xlab="Anos", col="lightgreen", breaks=15)



Comparação dos histogramas

Forma de distribuição

age\_skew <- ifelse(mean(salarios$age) > median(salarios$age), "positiva", "negativa")

husage\_skew <- ifelse(mean(salarios$husage) > median(salarios$husage), "positiva", "negativa")

cat(" Esposas: Assimetria", age\_skew, "\n")

cat(" Maridos: Assimetria", husage\_skew, "\n\n")

Resultados:

Esposas: Assimetria positiva

Maridos: Assimetria positiva

Faixa etária mais frequente:

cat(" Esposas:", names(which.max(table(round(salarios$age)))), "anos\n")

cat(" Maridos:", names(which.max(table(round(salarios$husage)))), "anos\n")

Resultado:

Esposas: 37 anos

Maridos: 44 anos

**(15 pontos)** Elaborar a tabela de frequencias das variáveis “age” (idade da esposa) e “husage” (idade do marido) e comparar os resultados

freq\_age <- table(cut(salarios$age, breaks=seq(15, 75, by=5)))

freq\_husage <- table(cut(salarios$husage, breaks=seq(15, 75, by=5)))

# Exibir tabelas

print(freq\_age)

print(freq\_husage)

"Frequência Idade Esposa:"

> print(freq\_age)

Resultados:

Frequência Idade Esposa:

(15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75]

61 359 803 949 976 840 681 571 394 0 0 0

Frequência Idade Marido:

(15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75]

11 221 630 872 913 892 688 572 467 256 67 31

## Comparação dos resultados:

Faixa Etária Dominante:

Esposas: 35-40 anos (976)

Maridos: 35-40 anos (913) e 40-45 anos (892) quase equivalentes

Distribuição em Idades Jovens:

Esposas têm mais casos nas faixas 20-25 (359) e 25-30 (803)

Maridos têm menos casos jovens: 20-25 (221) e 25-30 (630)

Distribuição em Idades Avançadas:

Esposas: Nenhum caso acima de 60 anos

Maridos: 354 casos acima de 60 anos (256+67+31)

Formato da Distribuição:

Esposas: Curva mais assimétrica à esquerda

Maridos: Distribuição mais uniforme nas idades maduras

# 2 Medidas de posição e dispersão

* 1. **a) (15 pontos)** Calcular a média, mediana e moda das variáveis “age” (idade da esposa) e “husage” (idade do marido) e comparar os resultados
  2. # Para idade da esposa

media\_age <- mean(salarios$age)

mediana\_age <- median(salarios$age)

moda\_age <- as.numeric(names(sort(table(salarios$age), decreasing=TRUE)[1]))

* 1. # Para idade do marido

media\_husage <- mean(salarios$husage)

mediana\_husage <- median(salarios$husage)

moda\_husage <- as.numeric(names(sort(table(salarios$husage), decreasing=TRUE)[1]))

## Comparação dos resultados

cat("Idade da Esposa:\nMédia:", media\_age, "\nMediana:", mediana\_age, "\nModa:", moda\_age, "\n")

cat("\nIdade do Marido:\nMédia:", media\_husage, "\nMediana:", mediana\_husage, "\nModa:", moda\_husage, "\n")

Idade da Esposa:

Média: 39.42758

Mediana: 39

Moda: 37

Idade do Marido:

Média: 42.45296

Mediana: 41

Moda: 44

Média:

A média da idade dos maridos (42,45 anos) é superior à média da idade das esposas (39,43 anos), indicando que, em média, os maridos são aproximadamente 3 anos mais velhos do que as esposas.

Mediana:

A mediana da idade dos maridos (41 anos) também é maior do que a das esposas (39 anos), o que reforça a tendência central observada na média.

Moda:

A moda — que representa a idade mais frequente — é de 44 anos para os maridos e 37 anos para as esposas. Essa diferença mostra que, além da tendência central ser maior entre os maridos, o valor mais comum também é mais elevado.

* 1. **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
  2. Para idade da esposa

var\_age <- var(salarios$age)

sd\_age <- sd(salarios$age)

cv\_age <- (sd\_age/media\_age)\*100

* 1. Para idade do marido

var\_husage <- var(salarios$husage)

sd\_husage <- sd(salarios$husage)

cv\_husage <- (sd\_husage/media\_husage)\*100

cat("Idade da Esposa:\nVariância:", var\_age, "\nDesvio Padrão:", sd\_age, "\nCoef. Variação:", cv\_age, "%\n")

cat("\nIdade do Marido:\nVariância:", var\_husage, "\nDesvio Padrão:", sd\_husage, "\nCoef. Variação:", cv\_husage, "%\n")

## Comparação dos resultados

Idade da Esposa:

Variância: 99.75234

Desvio Padrão: 9.98761

Coef. Variação: 25.33153 %

Idade do Marido:

Variância: 126.0717

Desvio Padrão: 11.22817

Coef. Variação: 26.44849 %

Variância:

A variância da idade dos maridos (126,07) é maior do que a das esposas (99,75), o que indica que a dispersão das idades dos maridos em torno da média é mais ampla. Em outras palavras, há maior diversidade nas idades dos maridos em comparação às esposas.

Desvio padrão:

O desvio padrão, que é a raiz quadrada da variância e representa a dispersão média em unidades da variável, também é maior para os maridos (11,23 anos) do que para as esposas (9,99 anos). Isso reforça a observação anterior de que a idade dos maridos varia mais em relação à média.

Coeficiente de variação (CV):

O coeficiente de variação é uma medida relativa da dispersão, expressa em porcentagem da média. Ambos os grupos têm valores semelhantes:

Esposas: 25,33%

Maridos: 26,45%

Apesar da leve diferença, o CV dos maridos é um pouco maior, o que indica uma ligeiramente maior variabilidade relativa em relação à sua própria média, mesmo levando em conta que a média dos maridos é mais alta.

Conclusão

Tanto em termos absolutos (variância e desvio padrão) quanto relativos (coeficiente de variação), a idade dos maridos apresenta maior dispersão do que a das esposas. Isso sugere que há uma maior heterogeneidade nas idades dos maridos, enquanto a distribuição das idades das esposas é um pouco mais concentrada.

# 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 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.**

**# configurar o caminho do banco de dados**

**setwd("/base-dados")**

**# Carregar os dados**

**load("salarios.RData")**

**# pacote nortest**

**install.packages("nortest")**

**library(nortest)**

**Teste de normalidade com Anderson-Darling**

**# idade do esposo**

**ad.test(salarios$husage)**

**# idade da esposa**

**ad.test(salarios$age)**

**Ao realizar os testes de normalidade, os resultados foram:**

**Anderson-Darling normality test**

**data: salarios$husage**

**A = 28.176, p-value < 2.2e-16**

**> ad.test(salarios$age) # idade da esposa**

**Anderson-Darling normality test**

**data: salarios$age**

**A = 31.828, p-value < 2.2e-16**

**Hipóteses:**

* H0​: A amostra segue uma distribuição normal.
* H1​: A amostra **não** segue uma distribuição normal.

**Interpretação dos resultados:**

* **p-value < 2.2e-16** → Extremamente pequeno.
* Portanto, **rejeitamos H0​** para ambas as variáveis.
* Conclusão: Nem husage (idade do esposo) nem age (idade da esposa) são normalmente distribuídas.

**#Visualização da distribuição**

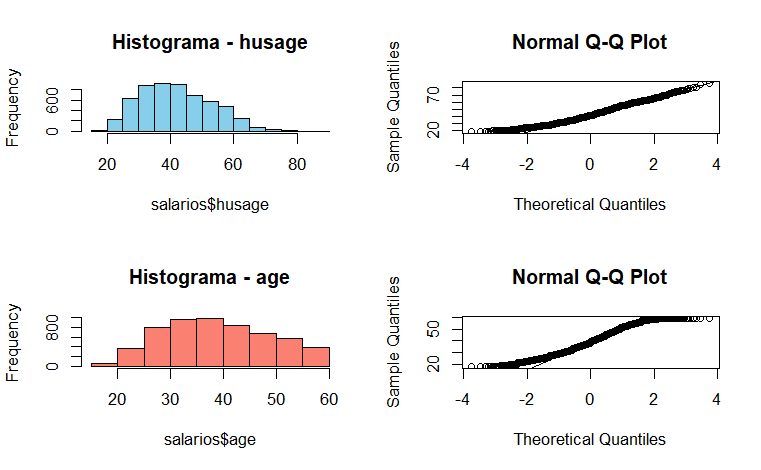
**par(mfrow = c(2, 2))**

**hist(salarios$husage, main = "Histograma - husage", col = "skyblue")**

**qqnorm(salarios$husage); qqline(salarios$husage)**

**hist(salarios$age, main = "Histograma - age", col = "salmon")**

**qqnorm(salarios$age); qqline(salarios$age)**



**Agora que sabemos que nem a idade do esposo e nem a idade da esposa são normalmente distribuídos. Vamos verificar se as variâncias são homogêneas (para o t-test)**

**var.test(salarios$husage, salarios$age) # Teste F de igualdade de variâncias**

**Resultado do teste F de igualdade de variáveis:**

**F test to compare two variances**

**data: salarios$husage and salarios$age**

**F = 1.2638, num df = 5633, denom df = 5633, p-value < 2.2e-16**

**alternative hypothesis: true ratio of variances is not equal to 1**

**95 percent confidence interval: 1.199526 1.331617**

**sample estimates: ratio of variances 1.263847**

**Interpretação dos resultados:**

**Hipóteses:**

* H0​: As variâncias populacionais de husage e age são **iguais**.
* H1​: As variâncias **são diferentes**.

**Explicação do Resultado:**

* Como o **p-value < 0.05**, você deve **rejeitar a hipótese nula H0**.
* Conclusão: Existe **diferença estatisticamente significativa nas variâncias**.
* Além disso, o **intervalo de confiança [1.20, 1.33]** não contém 1, o que reforça a rejeição de H0​.

Em resposta à questão sobre a escolha entre o teste t não pareado (paramétrico) e o teste de Mann-Whitney (não paramétrico), considerando que os dados não seguem uma distribuição normal e que há heterogeneidade das variâncias entre os grupos, conclui-se que o teste mais apropriado é o teste não paramétrico de Mann-Whitney.

# Caso não-paramétrico

teste <- wilcox.test(salarios$husage, salarios$age, paired = TRUE, conf.int = TRUE)

teste  
cat("\nTeste de Wilcoxon Pareado (dados não-normais)\n")

intervalo\_confianca <- teste$conf.int

limite\_inferior <- intervalo\_confianca[1]

limite\_superior <- intervalo\_confianca[2]

cat(sprintf("Intervalo de confiança: [%.2f, %.2f]\n", limite\_inferior, limite\_superior))

#resultado do teste  
Wilcoxon signed rank test with continuity correction

data: salarios$husage and salarios$age

V = 10551938, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

2.999961 3.000018

sample estimates:

(pseudo)median

2.999932

#Comparação dos resultados  
**Interpretação detalhada**

**p-value < 2.2e-16:**

* Extremamente pequeno → Rejeitamos H0​.
* Isso significa que há **diferença estatisticamente significativa** entre as idades de esposos e esposas.

**Pseudo-median = ~3 anos:**

* A **mediana da diferença** entre idades (esposo − esposa) é aproximadamente **3 anos**.

**Intervalo de confiança: [2.999961, 3.000018]**

* O IC de 95% está **totalmente acima de 0**, reforçando a evidência contra H0H\_0H0​.
* Isso confirma que, com 95% de confiança, os esposos são **em média 3 anos mais velhos** que as esposas.

**Conclusão final:**

* Há uma **diferença significativa** entre as idades dos esposos e esposas.
* Os **esposos são consistentemente mais velhos**, com uma **diferença mediana de 3 anos**.
* O uso de **teste pareado** foi correto, pois cada par de idades está logicamente relacionado.