# Testes de Hipóteses

**Prof.Vagner Antônio Marques**
vagner.marques@ufes.br
*Universidade Federal do Espírito Santo*

19/06/2023

# INTRODUÇÃO E AGENDA

- Terceira oficina de pesquisa do GECAT em 2023;
- Contextualizar o tema;
- Testes de normalidade;
- Testes de diferenças entre as médias (Paramétricos);
- Testes de diferenças entre as medianas (Não-paramétricos).

## Testes de Hipóteses na Pesquisa em Contabilidade

Os **testes de hipóteses** são úteis na pesquisa em contabilidade para obtermos evidências de que o fenômeno que estamos estudando seja estatisticamente significativo. Isso significa que estamos querendo verificar a veracidade (falsidade) quanto a uma hipótese formulada sobre:

- distribuição dos dados;
- diferenças de médias, medianas, variâncias entre grupos;
- associação entre variáveis;
- significâncias dos parâmetros;
- significância de um modelo;
- etc.

## Testes de Hipóteses na Pesquisa em Contabilidade

Na prática, estaremos interessados em entender:

- A variável apresenta uma distribuição normal?
- As médias de uma variável entre dois ou mais grupos são iguais?
- A correlação (associação ou independência) entre as variáveis é estatisticamente significativa?

# Formulação das hipóteses

Em geral, formulamos a hipótese nula ($H_0$), por exemplo:

$$\mu_i = \mu_t$$

e uma hipótese alternativa ($H_1$):

$$\mu_i \neq \mu_t$$

# Formulação das hipóteses

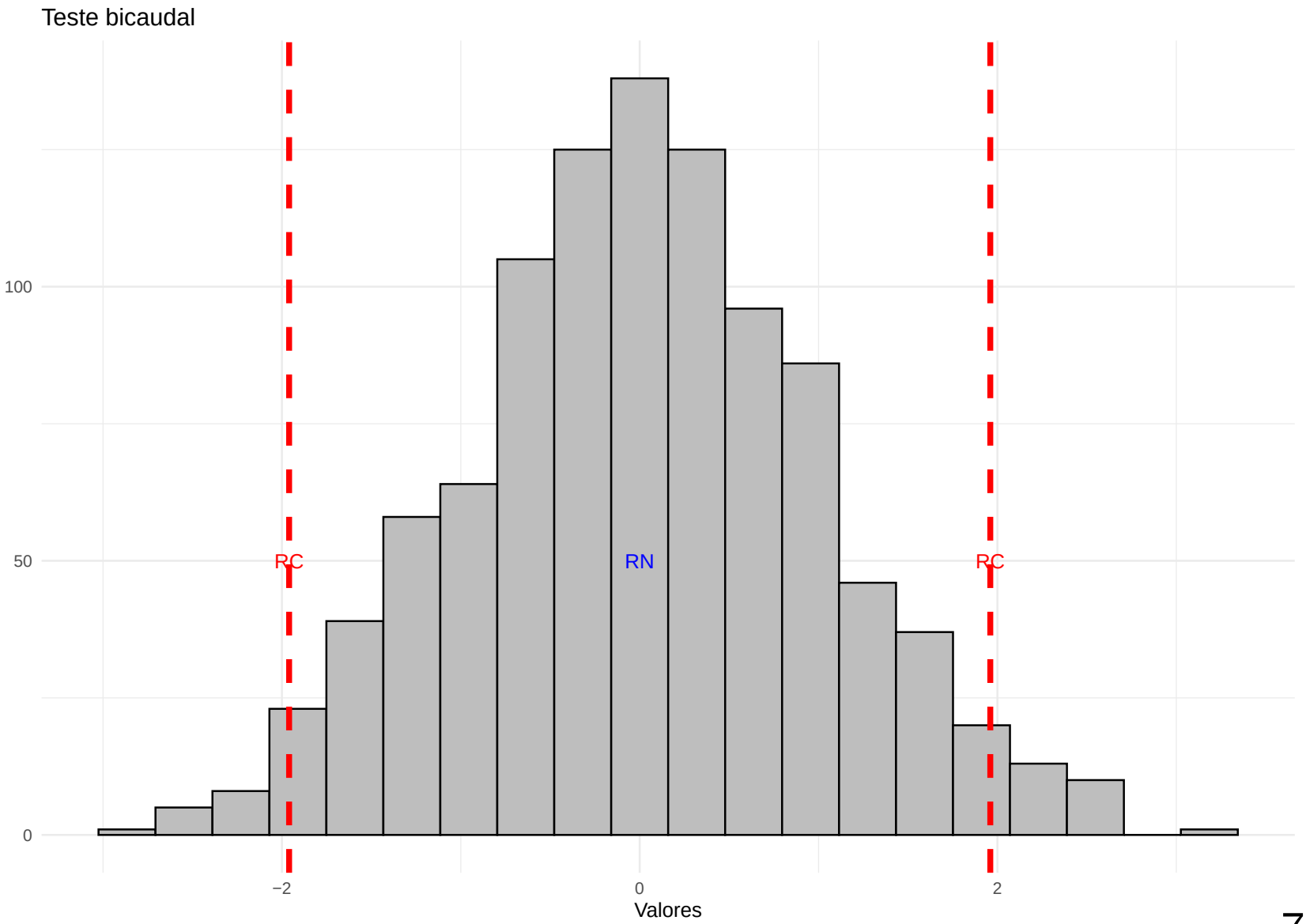
Podemos ainda, formular uma hipótese alternativa ($H_1$) unicaudal (a direita):

$$\mu_i > \mu_t$$

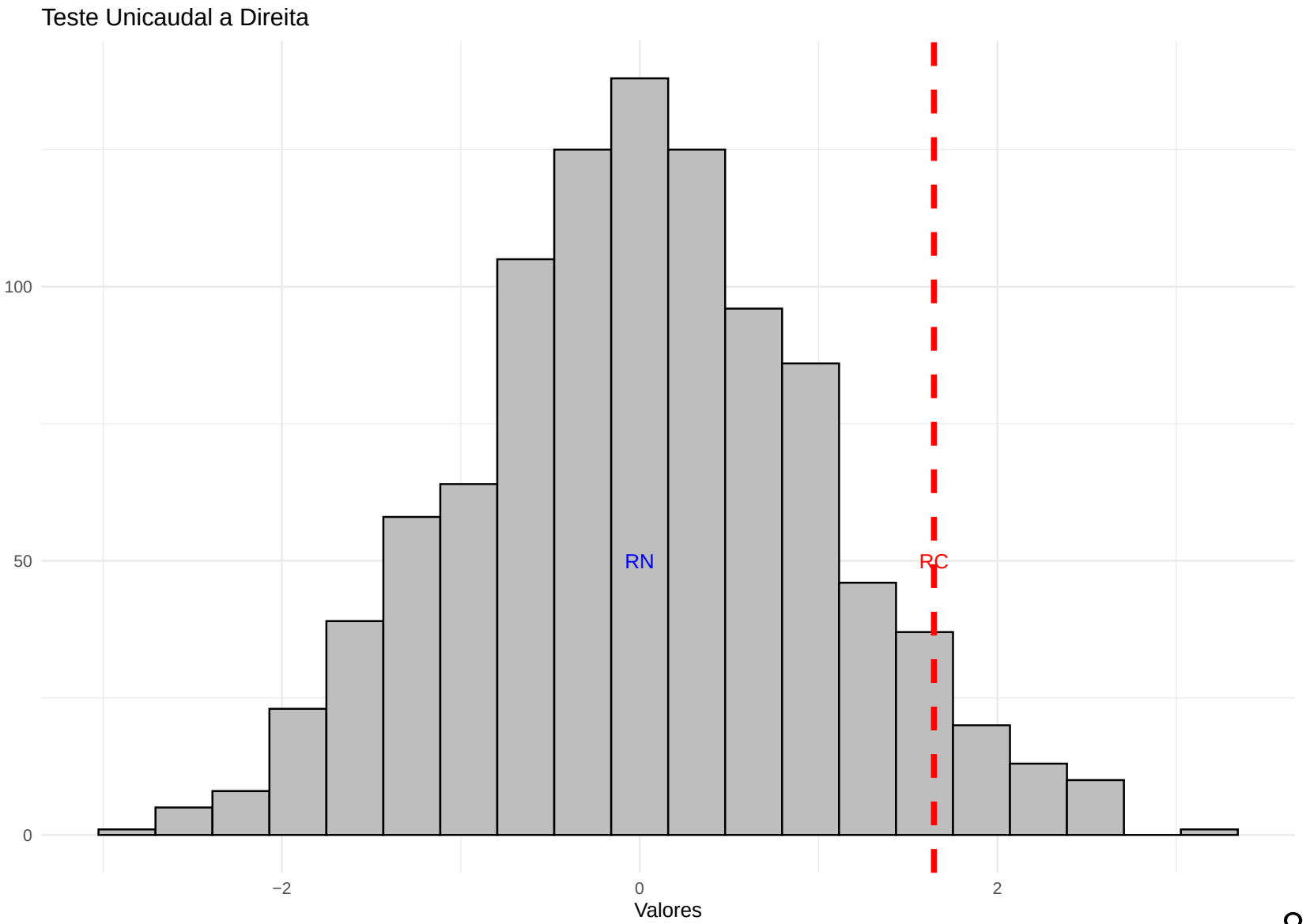
Ou uma hipótese alternativa ($H_1$) unicaudal (a esquerda):

$$\mu_i < \mu_t$$

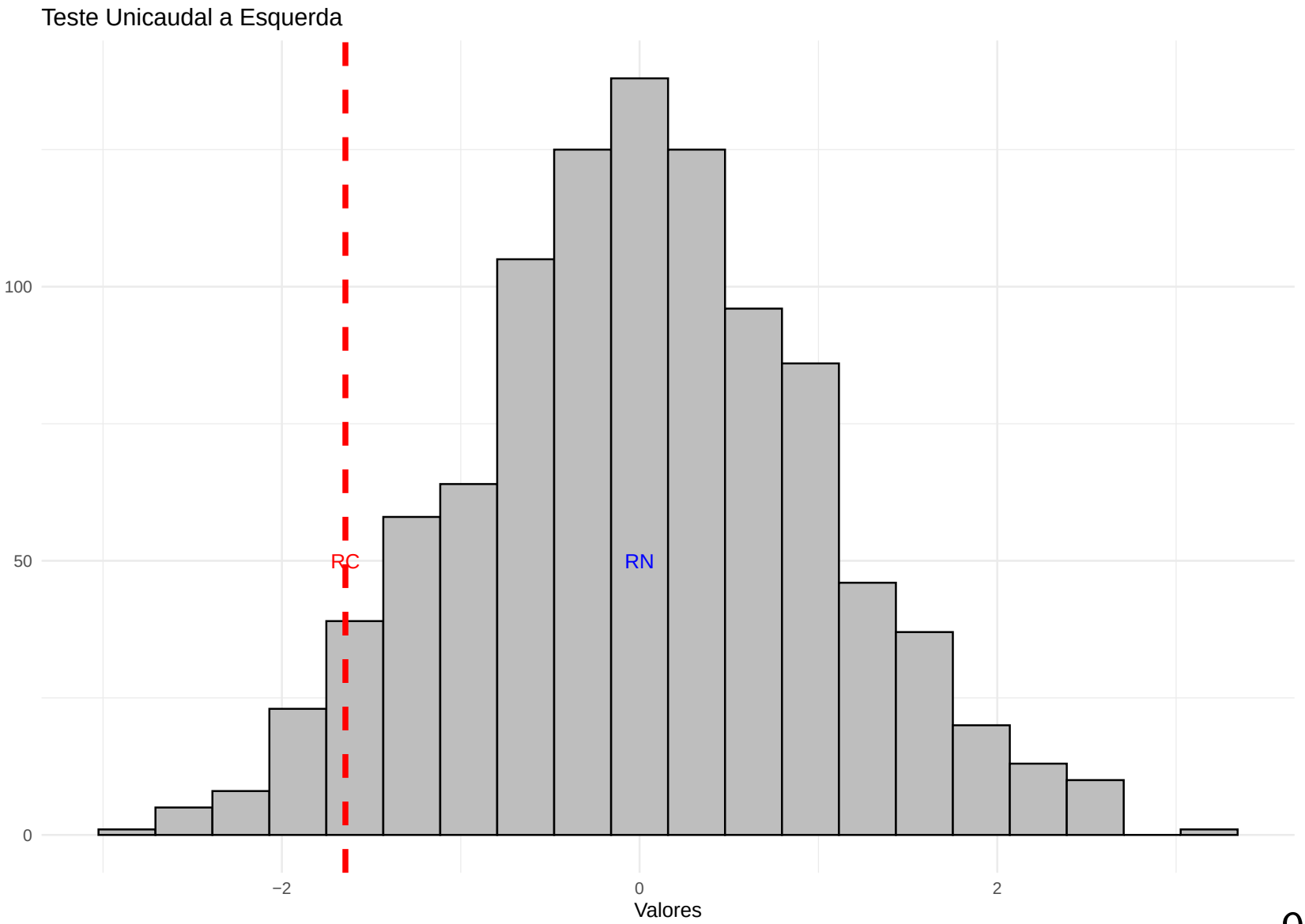
# Teste Bicaudal



# Teste Unicaudal à direita



# Teste Unicaudal à esquerda



## Tabela - Probabilidade Acumulada por quantidade de desvios padrões

```
##   Desvios_Padroes Probabilidade_Acumulada
## 1              -3                0.001349898
## 2              -2                0.022750132
## 3              -1                0.158655254
## 4               0                0.500000000
## 5               1                0.841344746
## 6               2                0.977249868
## 7               3                0.998650102
```

# Para a construção de uma hipótese devemos seguir os seguintes passos:

1. Escolher o teste estatístico adequado;
2. Declarar a hipótese nula $H_0$ e a hipótese alternativa $H_1$;
3. Definir o nível de significância $\alpha$ (em geral de $\alpha \leq 0{,}05$);
4. Calcular o valor observado da estatística do teste com base na amostra extraída da população;
5. Estabelecer a região crítica do teste em função do $\alpha$ fixado na etapa 3;
6. Tomar a decisão de não rejeitar $H_0$ ou rejeitá-la em detrimento de $H_1$.

Fávero & Belfiore (2017)

# Para a construção de uma hipótese devemos seguir os seguintes passos:

Em geral, define-se nível de significância ($\alpha$) e a partir dele, decide-se quanto à rejeição ($H_1$) ou não rejeição ($H_0$). A regra de decisão consiste em **rejeitar $H_0$ quando o p-valor for menor que $\alpha$**.

Fávero & Belfiore (2017)

## AVISO

Para cada teste de hipótese, busque entender:

(i) O que ele faz?
(ii) Quais os pressupostos?
(iii) Para que ele serve?
(iv) Como ele é operacionalizado?
(v) Qual a regra de decisão?

# Testes Paramétricos e Não Paramétricos

Existem diversos testes de hipóteses possíveis.

Uma primeira classificação os agrupam em **testes paramétricos** e **testes não paramétricos**. Os **testes paramétricos** são aqueles que exigem que se atendam algumas condições:

- observações independentes;
- distribuição normal;
- variâncias homogêneas;
- variáveis sejam **quantitativas (contínuas ou discretas)**.

Os **testes não paramétricos** não exigem essas suposições e são aplicáveis a **variáveis qualitativas**.

# Testes Paramétricos e Não Paramétricos

Considerando essa diversidade de testes, nós vamos focar em três tipos específicos, que são os mais utilizados na pesquisa em contabilidade:

1. Normalidade (Shapiro-Wilk; Kolmogorov-Smirnov; Shapiro-Francia);
2. Diferença de médias (Teste t para amostras independentes, Teste t emparelhado, ANOVA);
3. Diferença de medianas (Mann-Whitney, Wilcoxon, Kruskal-Wallis);

# Testes de Normalidade

Os testes de normalidade são necessários para **verificarmos se uma variável possui uma distribuição normal ou não**. Isso pode ser necessário para observarmos, por exemplo, se atende:

1. À suposição de um teste de hipótese paramétrico;
2. Ao pressuposto de normalidade dos resíduos (em análise de regressão).

**Tabela 1:** Tabela de Testes de Normalidade e Funções no R

| Teste | Função | Descrição |
|---|---|---|
| SW | shapiro.test() | para pequenas amostras |
| KS | ks.test() | para grandes amostras |
| SF | sf.test() | para grandes amostras |

# Teste de Shapiro-Wilk

## Quando usar?

- É utilizado para se avaliar a normalidade da distribuição de um conjunto de dados (amostrais).
- É utilizado quando se tem amostras com menos de 30 observações.

## Hipótese do teste

- $H_0$: a variável x na população possui uma distribuição normal $\mathrm{N}(\mu, \sigma)$
- $H_1$: a variável x na população não possui uma distribuição normal $\mathrm{N}(\mu, \sigma)$

# Teste de Shapiro-Wilk

## Definição do nível de significância

Na área de Ciências Sociais Aplicadas, em especial, na Contabilidade, *5% de nível de significância é o mais observado*.

# Teste de Shapiro-Wilk

## Como operacionalizá-lo?

No R, você pode utilizar o seguinte código:

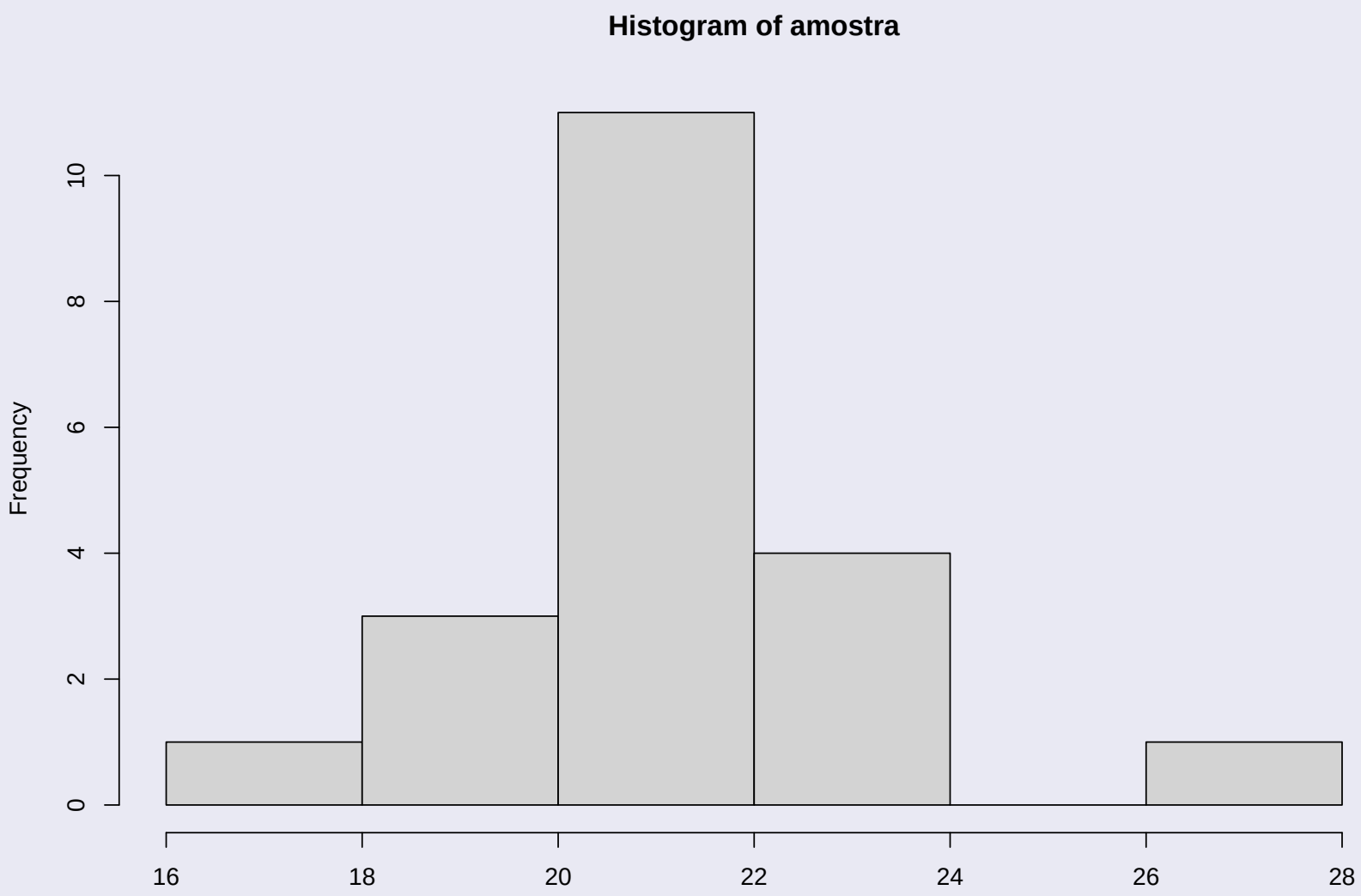
```
library(dplyr)

if (!exists("Base_R_Artigo_PEA")) {
  load("C:/Users/Vagne/OneDrive/DOCUME~1-LAPTOP-0CUBQV8E-28
}

# Selecionar uma amostra menor que trinta observações
amostra <- sample(Base_R_Artigo_PEA$TAM, size = 20)
```

# Teste de Shapiro-Wilk

## Histograma (Amostra pequena)



# Teste de Shapiro-Wilk

## Como operacionalizá-lo?

```r
# Realizar o teste de Shapiro-Wilk na amostra
result_sw <- shapiro.test(amostra)

# Imprimir o resultado
result_sw
```

```
##
##  Shapiro-Wilk normality test
##
## data:  amostra
## W = 0.91505, p-value = 0.07961
```

# Teste de Shapiro-Wilk

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, logo, os dados não apresentam uma distribuição normal.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, os dados apresentam uma distribuição normal.

# Teste Shapiro-Francia

## Quando usar?

- É utilizado para se avaliar a normalidade da distribuição de um conjunto de dados (amostrais).
- É utilizado quando se tem amostras com igual ou superior a 30 observações.

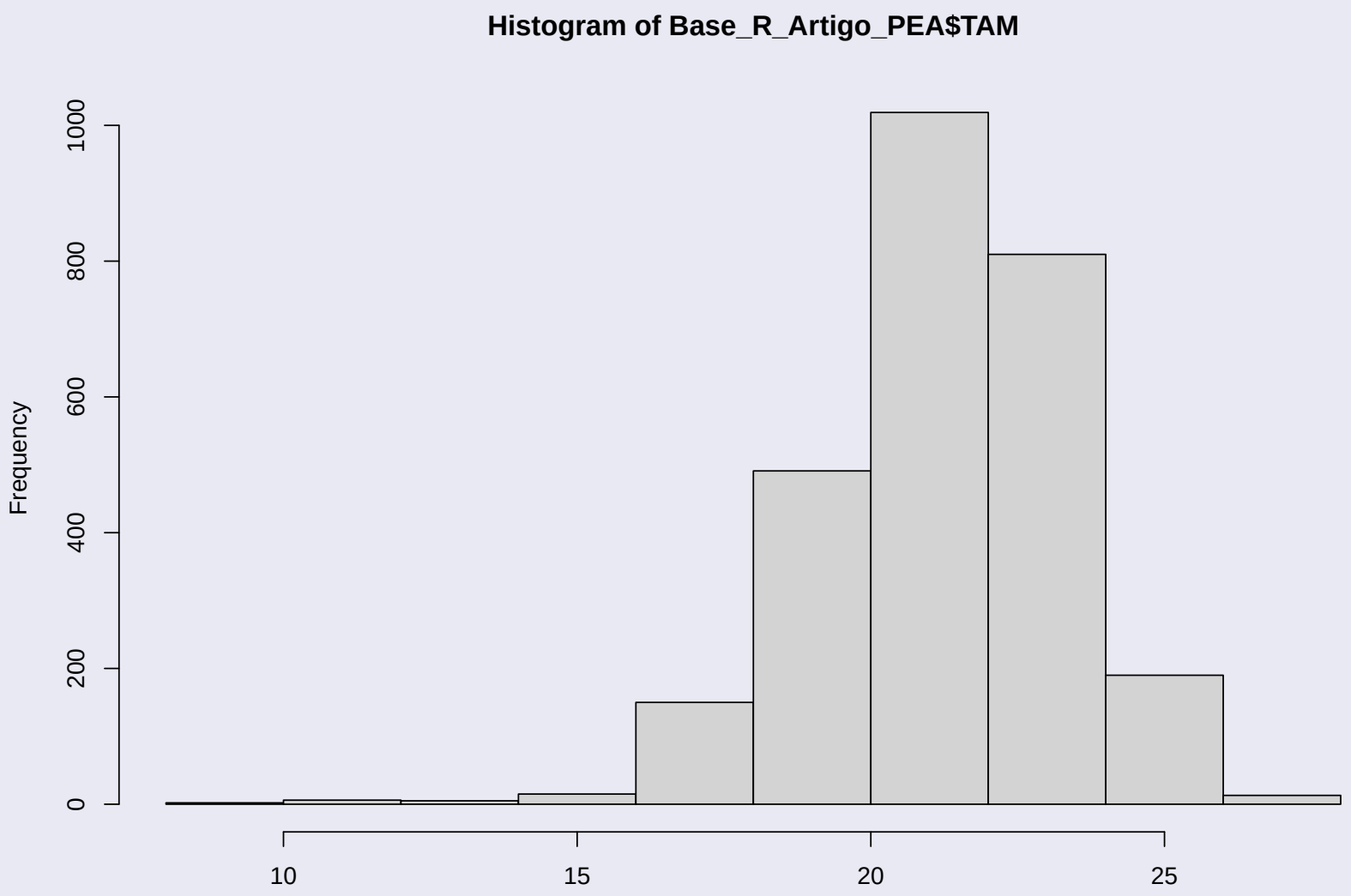
## Hipótese do teste

- $H_0$: a variável x na população possui uma distribuição normal $\mathrm{N}(\mu, \sigma)$

- $H_1$: a variável x na população não possui uma distribuição normal $\mathrm{N}(\mu, \sigma)$

# Teste Shapiro-Francia

## Histograma (com oultliers - amostra grande)



# Teste Shapiro-Francia

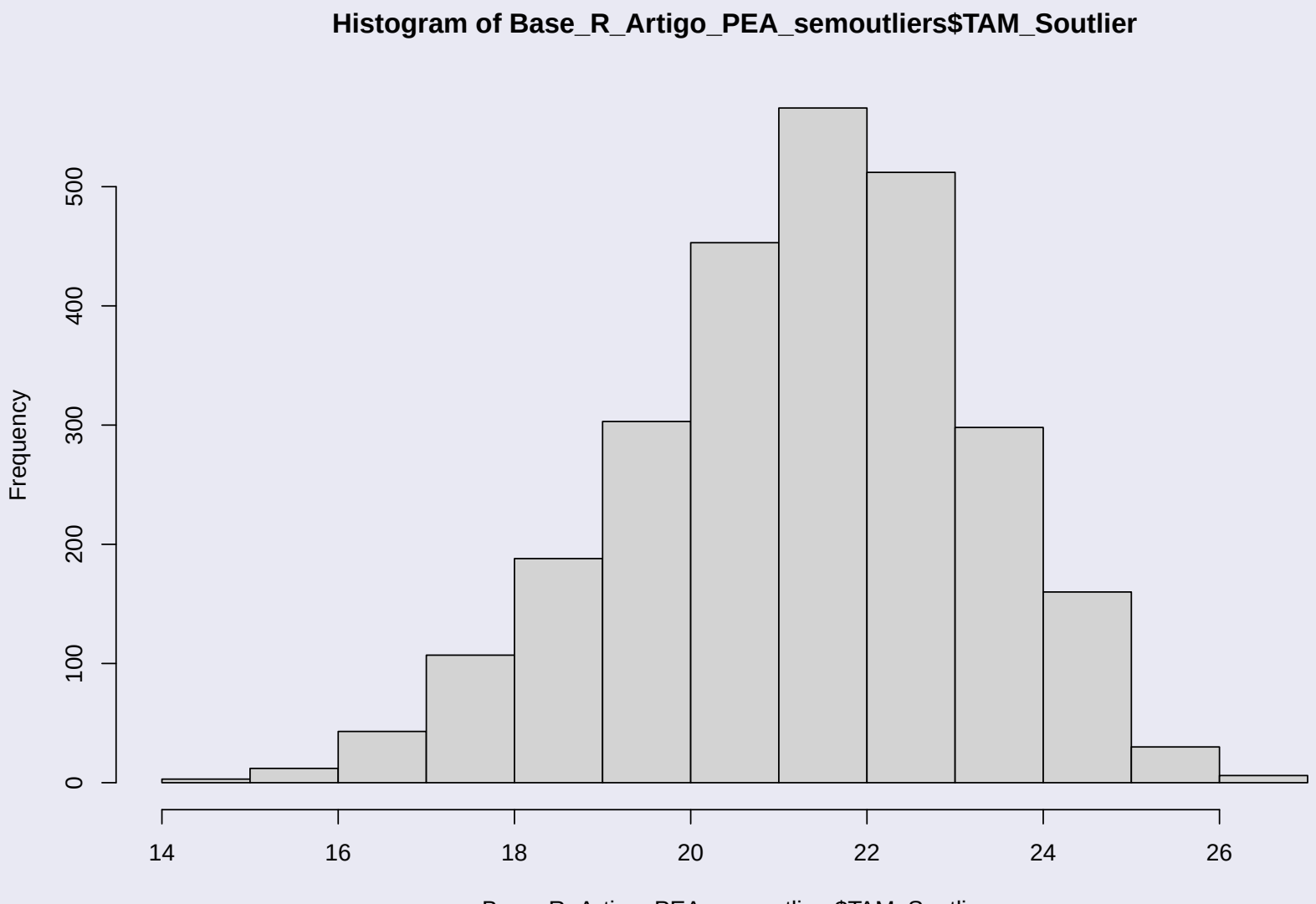
## Excluindo os valores extremos (outliers)

```
library(dplyr)
#Histograma sem outliers

Base_R_Artigo_PEA_semoutliers <- as.data.frame(Base_R_Artig
  dplyr:: mutate(TAM_Soutlier = TAM) %>%
  dplyr:: filter(TAM_Soutlier >= 14 & TAM_Soutlier <= 27 )
  dplyr:: filter(!is.na(TAM_Soutlier))
```

# Histograma (Sem outliers - amostra grande)

```
hist(Base_R_Artigo_PEA_semoutliers$TAM_Soutlier)
```



# Teste Shapiro-Francia

## Como operacionalizá-lo?

No R, você pode utilizar o seguinte código:

```
# Carregar o pacote nortest
library(nortest)

# Realizar o teste Shapiro-Francia
result_sfrancia <- sf.test(Base_R_Artigo_PEA$TAM)
result_sfrancia_soutlier <- sf.test(Base_R_Artigo_PEA_semou
```

# Teste Shapiro-Francia

## Como operacionalizá-lo?

```
# Imprimir o resultado
result_sfrancia
```

```
##
##  Shapiro-Francia normality test
##
## data:  Base_R_Artigo_PEA$TAM
## W = 0.97186, p-value < 2.2e-16
```

```
result_sfrancia_soutlier
```

```
##
##  Shapiro-Francia normality test
##
## data:  Base_R_Artigo_PEA_semoutliers$TAM_Soutlier
## W = 0.992, p-value = 1.944e-10
```

# Teste Shapiro-Francia

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, logo, os dados não apresentam uma distribuição normal.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, os dados apresentam uma distribuição normal.

# Teste Kolmogorov-Smirnov

## Quando usar?

- É utilizado para se avaliar a normalidade da distribuição de um conjunto de dados (amostrais).
- É utilizado quando se tem amostras com tamanho igual ou superior a 30 observações.

## Hipótese do teste

- $H_0$: a variável x na população segue uma distribuição normal $\mathrm{N}(\mu, \sigma)$

- $H_1$: a variável x na população não segue uma distribuição normal $\mathrm{N}(\mu, \sigma)$

# Teste Kolmogorov-Smirnov

## Como operacionalizá-lo?

No R, você pode utilizar o seguinte código:

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Base_R_Artigo_PEA$TAM
## D = 0.047263, p-value = 1.149e-05
## alternative hypothesis: two-sided
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Base_R_Artigo_PEA_semoutliers$TAM_Soutlier
## D = 0.038346, p-value = 0.0008011
## alternative hypothesis: two-sided
```

# Teste Kolmogorov-Smirnov

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, logo, os dados não apresentam uma distribuição normal.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, os dados apresentam uma distribuição normal.

# Testes de diferenças entre as médias (Testes Paramétricos)

Os testes de diferenças entre as médias são úteis na pesquisa em contabilidade, pois com muita frequência queremos:

- Comparar médias de diversas variáveis entre dois grupos independentes;

$$\mu_{reapr} = \mu_{Nreapr}$$

$$\mu_{Big4} = \mu_{NBig4}$$

- comparar médias de diversas variáveis antes e depois de um tratamento.

$$\mu_{Pré-IFRS} = \mu_{Pós-IFRS}$$

$$\mu_{Pré-reapr} = \mu_{Pós-reapr}$$

# Testes de diferenças entre as médias (Testes Paramétricos)

**Os testes paramétricos mais usados, sobretudo na pesquisa em contabilidade são:**

- Teste t de Student para duas amostras independentes (Tratamento e Controle);
- Teste t de Student para duas amostras emparelhadas (Pré e Pós evento);
- Teste ANOVA (Para mais de duas amostras).

# Teste t de Student para duas amostras independentes

## Quando utilizar?

- Quando o objetivo for analisar dados de amostras independentes (Empresas que reapresentaram e que não reapresentaram);
- Quando os dados das duas amostras tiverem uma distribuição normal;
- É preciso analisar também se a variância é homogênea (ou não).

# Teste t de Student para duas amostras independentes

## Hipótese do teste

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

# Teste t de Student para duas amostras independentes

## Como operacionalizar?

```
# Carregando a biblioteca
# Realizar o teste t para amostras independentes
result_ttest <- t.test(FEES ~ BIG, data = Base_R_Artigo_PEA
                       alternative = "two.sided",
                       var.equal = TRUE,
                       na.rm = TRUE)

# Extrair os resultados relevantes
p_value <- format(result_ttest$p.value, digits = 4)
estimate_diff <- format(result_ttest$estimate, digits = 4)
```

# Teste t de Student para duas amostras independentes

## Como operacionalizar?

```
# Imprimir os resultados
cat("Teste t para amostras independentes:\n")
```

```
## Teste t para amostras independentes:
```

```
cat("NBig4 vs. Big4\n")
```

```
## NBig4 vs. Big4
```

```
cat("P-valor:", p_value, "\n")
```

```
## P-valor: 3.34e-165
```

```
cat("Diferença de médias:", estimate_diff, "\n")
```

```
## Diferença de médias: 11.72 13.26
```

# Teste t de Student para duas amostras independentes

## Como operacionalizar?

```
## Teste t para amostras independentes:
## NBig4 vs. Big4
## P-valor: 1.567e-150
## Diferença de médias: 11.72 13.26
```

# Teste t de Student para duas amostras independentes

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, existe diferença entre as médias.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, não existe diferença entre as médias.

# Teste t de Student para duas amostras emparelhadas

## Quando utilizar?

- Quando o objetivo for analisar dados de amostras emparelhadas (antes e depois de um tratamento);
- Quando os dados das duas amostras tiverem uma distribuição normal;
- Quando os dados tiverem variância homogênea.

# Teste t de Student para duas amostras emparelhadas

## Hipótese do teste

$$H_0 : \mu_d = 0, \mu_d = \mu_{pré} - \mu_{pós}$$

$$H_1 : \mu_d \neq 0$$

# Teste t de Student para duas amostras emparelhadas

## Como operacionalizar?

```
base_fee_pre <- Base_R_Artigo_PEA_semoutliers %>%
  dplyr::mutate(Pos_PAA = ifelse(Data < 2017, 0, 1)) %>%
  dplyr::select("Ticker", "Data", "FEES", "Pos_PAA") %>%
  dplyr::filter(Data < 2017) %>%
  dplyr::slice_sample(n = 773)

base_fee_pos <- Base_R_Artigo_PEA_semoutliers %>%
  dplyr::mutate(Pos_PAA = ifelse(Data < 2017, 0, 1)) %>%
  dplyr::select("Ticker", "Data", "FEES", "Pos_PAA") %>%
  dplyr::filter(Data > 2016)
```

# Teste t de Student para duas amostras emparelhadas

## Como operacionalizar?

```
# Realizar o left join pelo campo "Ticker"
base_ttest_par <- dplyr::left_join(base_fee_pre, base_fee_p

base_ttest_par <- base_ttest_par %>%
  dplyr::rename(FEES_pre = "FEES.x",
                FEES_pos = "FEES.y")
```

# Teste t de Student para duas amostras emparelhadas

## Como operacionalizar?

```
# Realizar o teste t pareado

attach(base_ttest_par)
result <- t.test(FEES_pre,
                 FEES_pos,
                 paired = TRUE)
```

# Teste t de Student para duas amostras emparelhadas

## Como operacionalizar?

```
# Exibir o resultado do teste t pareado
print(result)
```

```
##
##  Paired t-test
##
## data:  FEES_pre and FEES_pos
## t = -11.573, df = 2054, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equa
## 95 percent confidence interval:
##  -0.2655350 -0.1885843
## sample estimates:
## mean difference
##      -0.2270596
```

# Teste t de Student para duas amostras emparelhadas

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, existe diferença entre as médias.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, não existe diferença entre as médias.

# Teste ANOVA (Análise de Variância)

## Quando utilizar?

- Quando o objetivo for comparar médias de três ou mais amostras independentes, por meio da variância; (Segmentos Econômicos, Estágios do Ciclo de Vida, Firmas de Auditoria);
- Quando os dados das duas amostras tiverem uma distribuição normal;
- Quando as variâncias forem homogêneas.

# Teste ANOVA (Análise de Variância)

## Hipótese do teste

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots \mu_k$$

$$H_1 : \exists_{(i,j)} \mu_i \neq \mu_j$$

# Teste ANOVA (Análise de Variância)

## Como operacionalizar?

```
anova_teste <- aov(FEES ~ factor(SetorL)*factor(ECV),
                   data = Base_R_Artigo_PEA_semoutliers)
```

# Teste ANOVA (Análise de Variância)

## Como operacionalizar?

```
##                                 Df Sum Sq Mean Sq F value
## factor(SetorL)                   8    101   12.61   7.091 2
## factor(ECV)                      4    157   39.35  22.130
## factor(SetorL):factor(ECV)      32    168    5.25   2.950 6
## Residuals                     2531   4501    1.78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
## 84 observations deleted due to missingness
```

# Teste ANOVA (Análise de Variância)

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, existe diferença entre as médias.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, não existe diferença entre as médias.

# Testes de diferenças entre as medianas (Testes Não-Paramétricos)

**Tabela 2:** Testes Paramétricos e Não Paramétricos

| Não.Paramétrico | Paramétrico |
|---|---|
| Teste de Wilcoxon | Teste t de Student para amostras independentes |
| Teste de Mann-Whitney | Teste t para amostras emparelhadas |
| Teste de Kruskal-Wallis | Análise de Variância |

# Teste de Wilcoxon

## Quando utilizar ?

Substituto do Teste T para amostras independentes, quando uma ou mais suposições do equivalente paramétrico forem violadas.

## Hipótese do teste

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$

# Teste de Wilcoxon

## Como operacionalizar ?

```
library(rstatix)

wilcox_test(FEES ~ BIG,
            data = Base_R_Artigo_PEA_semoutliers)
```

```
## # A tibble: 1 x 9
##   .y.   group1 group2    n1    n2 statistic          p
## * <chr> <chr>  <chr>  <int> <int>     <dbl>      <dbl>
## 1 FEES  0      1        764  1838   227992. 1.64e-162 1.
```

# Teste de Wilcoxon

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, existe diferença entre as médias.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, não existe diferença entre as médias.

# Teste de Mann-Whitney

## Quando utilizar ?

Substituto do Teste T para amostras emparelhadas, quando uma ou mais suposições do equivalente paramétrico forem violadas.

## Hipótese do teste

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$

# Teste de Mann-Whitney

## Como operacionalizar ?

```
# Teste U de Mann-Whitney (Wilcoxon)
wilcox.test(base_ttest_par$FEES_pre,
            base_ttest_par$FEES_pos)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  base_ttest_par$FEES_pre and base_ttest_par$FEES_p
## W = 1996493, p-value = 3.798e-08
## alternative hypothesis: true location shift is not equal
```

# Teste de Mann-Whitney

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, existe diferença entre as médias.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, não existe diferença entre as médias.

# Teste Kruskal-Wallis

## Quando utilizar ?

Substituto do ANOVA, quando uma ou mais suposições do equivalente paramétrico forem violadas.

## Hipótese do teste

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots \mu_k$$

$$H_1 : \exists_{(i,j)} \mu_i \neq \mu_j$$

# Teste Kruskal-Wallis

## Como operacionalizar ?

```
# Executar o teste de Kruskal-Wallis com interação
kruskal.test(FEES ~ ECV,
             data = Base_R_Artigo_PEA_semoutliers)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  FEES by ECV
## Kruskal-Wallis chi-squared = 79.794, df = 4, p-value < 2
```

# Teste Kruskal-Wallis

## Regra de decisão

- Se o p-valor $\leq 0{,}05$ $(\alpha)$ implica em rejeição da hipótese nula $(H_0)$, existe diferença entre as médias.
- Se o p-valor $> \alpha$ implica em não rejeitar $H_0$, logo, não existe diferença entre as médias.

# Referências


- Fávero, L.P.; Belfiore, P. **Análise de Dados: Estatística e Modelagem Multivariada com Excel, SPSS e Stata**. (2017). 1 ed. Rio de Janeiro: Elsevier.
- Fávero, L.P.; Belfiore, P.; & Souza, R.F. (2023). **Data Science, Analytics and Machine Learning With R**. London: Elsevier.
- Morettin, P.A.; Singer, J.M. **Estatística e Ciência de Dados**.(2022) 1.ed. Rio de Janeiro: LTC.

**Obrigado pela sua participação!**

Contamos com a sua presença nas próximas oficinas!

Para saber mais sobre as oficinas, siga o perfil **@gecat.ufes** no Instagram. ☺

Acesse o nosso canal no YouTube, faça sua inscrição e ative as notificações: https://www.youtube.com/@GECAT_UFES