Exercício de Análise Estatística de Ensaio Clínico

Métodos Laboratoriais e Modelos Experimentais Aplicados à Pesquisa

Guilherme Camargo Brito

April 02, 2023

Contents

Introdução

> Esse relatório é referente a análise estatística de um ensaio clínico para a disciplina Métodos laboratoriais e modelos experimentais aplicados à pesquisa do curso de pós-graduação em Biologia Celular e Molecular da PUCRS.

Foram utilizados os seguintes métodos estatísticos:

- Análise descritiva: média, mediana, desvio padrão, erro padrão e intervalo interquartil.
- Teste de normalidade: Shapiro-Wilk para definição de uso de testes paramétricos ou não-paramétricos.
- Análise de frequência: frequência absoluta e relativa.
- Análise de diferenças entre grupos: teste t de duas amostras de Welch (variáveis paramétricas) ou Mann-Whitney U (variáveis não-paramétricas).
- Análise de diferenças entre grupos de idade (>=20 vs <20 anos): Teste t de duas amostras de Welch (variáveis paramétricas) ou Mann-Whitney U (variáveis não-paramétricas).
- Cálculo de tamanho de efeito: Cohen's d (variáveis paramétricas) ou Coeficiente de correlação ponto-bisserial (variáveis não-paramétricas).
- Análise de correlações entre variáveis: correlação produto-momento de Pearson (variáveis paramétricas) ou correlação de Spearman (variáveis não-paramétricas).
- Todas as análises consideraram um nível alfa de 5% (p.value = 0.05).

Acesso ao código da análise e aos dados

Esse projeto, com o código R da análise, o documento em formato rmarkdown que gera o pdf, assim como os dados associados podem ser acessado no seguinte link:

Exercícios

Abaixo seguem as tabelas com os resultados das análises descritas.

Questão 1

Estatísticas da Amostra

Variável	n	missing	normality	mean	median	sd	se	iqr
Altura_cm	16	0	parametric	158.9	163.0	19.1	4.8	21.2
CVFpercprevist	16	0	parametric	65.1	53.3	25.4	6.3	42.5
DNA_corrigido15x	16	0	parametric	305.3	334.0	167.1	41.8	252.9
Dias_ATB	16	2	parametric	53.3	49.0	30.3	7.6	43.0
Dias_internacao	16	0	non-parametric	10.2	0.0	14.4	3.6	15.8
IMC_absoluto	16	0	parametric	19.8	19.6	2.8	0.7	2.8
Idade	16	0	parametric	17.4	19.8	7.1	1.8	12.7
Peso_kg	16	0	parametric	51.3	55.0	14.9	3.7	17.5
ReservaVentilatoria	16	0	non-parametric	36.8	43.7	19.9	5.0	15.1
SpO2final	16	0	parametric	90.6	91.0	6.3	1.6	12.0
VEF1percprevist	16	0	non-parametric	51.8	35.7	28.1	7.0	45.9
VEabsolutofinal	16	0	parametric	44.2	45.0	15.9	4.0	28.5
VO2mLkgminfinal	16	0	parametric	32.8	32.7	5.2	1.3	6.6

Estatísticas por Grupo

Variável	Grupo	n	missing	normality	mean	median	sd	se	iqr
Altura_cm	1	8	0	parametric	147.1	148.8	19.2	6.8	34.5
Altura_cm	2	8	0	parametric	170.8	172.4	9.6	3.4	13.4
CVFpercprevist	1	8	0	parametric	76.6	83.4	24.8	8.8	38.7
CVFpercprevist	2	8	0	parametric	53.5	45.6	21.5	7.6	20.2
DNA_corrigido15x	1	8	0	parametric	221.8	177.5	132.3	46.8	236.1
DNA_corrigido15x	2	8	0	parametric	388.8	366.5	162.3	57.4	121.0
Dias_ATB	1	8	1	parametric	51.9	42.0	32.0	11.3	21.0
Dias_ATB	2	8	1	parametric	54.7	56.0	31.0	11.0	52.0
Dias_internacao	1	8	0	non-parametric	8.2	0.0	12.0	4.2	16.5
Dias_internacao	2	8	0	non-parametric	12.2	7.0	17.0	6.0	15.8
IMC_absoluto	1	8	0	parametric	19.6	17.8	4.0	1.4	4.4
IMC_absoluto	2	8	0	parametric	20.0	19.9	0.8	0.3	1.2
Idade	1	8	0	parametric	15.5	13.2	7.3	2.6	11.4
Idade	2	8	0	parametric	19.3	21.1	6.8	2.4	6.3
Peso_kg	1	8	0	parametric	44.1	42.6	17.6	6.2	29.9
Peso_kg	2	8	0	parametric	58.4	57.9	7.2	2.5	9.9
ReservaVentilatoria	1	8	0	non-parametric	30.9	45.5	25.8	9.1	24.8
ReservaVentilatoria	2	8	0	parametric	42.7	41.3	10.0	3.5	15.2
SpO2final	1	8	0	non-parametric	93.8	97.0	6.5	2.3	4.8
SpO2final	2	8	0	parametric	87.5	88.5	4.6	1.6	6.2
VEF1percprevist	1	8	0	parametric	65.8	67.6	29.2	10.3	52.2
VEF1percprevist	2	8	0	non-parametric	37.7	29.0	19.8	7.0	10.7
VEabsolutofinal	1	8	0	parametric	40.5	37.0	17.8	6.3	27.9
VEabsolutofinal	2	8	0	parametric	47.9	48.0	13.8	4.9	20.3
VO2mLkgminfinal	1	8	0	parametric	32.6	32.7	3.9	1.4	6.0
VO2mLkgminfinal	2	8	0	parametric	32.9	34.0	6.6	2.3	6.6

Questão 2

Análise de Frequências

Variável	Grupo	Valor	Freq_abs	Freq_rel
DNAcat243	1	1	5	0.62
DNAcat243	1	2	3	0.38
DNAcat243	2	1	1	0.12
DNAcat243	2	2	7	0.88
Internacao	1	1	3	0.38
Internacao	1	2	5	0.62
Internacao	2	1	4	0.50
Internacao	2	2	4	0.50
Mutacao	1	1	2	0.25
Mutacao	1	2	3	0.38
Mutacao	1	3	3	0.38
Mutacao	2	1	6	0.75
Mutacao	2	3	2	0.25
Pseudomonas_cronico	1	1	2	0.25
Pseudomonas_cronico	1	2	6	0.75
Pseudomonas_cronico	2	1	3	0.38
Pseudomonas_cronico	2	2	5	0.62
SPO2cat	1	1	2	0.25
SPO2cat	1	2	6	0.75
SPO2cat	2	1	5	0.62
SPO2cat	2	2	3	0.38
Sexo	1	1	6	0.75
Sexo	1	2	2	0.25
Sexo	2	1	4	0.50
Sexo	2	2	4	0.50
VO2categ	1	1	3	0.38
VO2categ	1	2	5	0.62
VO2categ	2	1	2	0.25
VO2categ	2	2	6	0.75

Questão 3

Diferenças entre Grupos

Variável	Teste	df	Stat	p_value	Sign	Effect	CI_low	CI_up
VEF1percprevist	Wilcoxon rank sum test	NA	56.00	0.01	*	0.38	5.09	60.38
DNA_corrigido15x	Welch Two Sample t-test	13.45	-2.26	0.04	*	-1.13	-326.42	-7.57
SpO2final	Welch Two Sample t-test	12.63	2.23	0.04	*	1.11	0.17	12.33
Idade	Welch Two Sample t-test	13.94	-1.08	0.30	-	-0.54	-11.41	3.75
Dias_internacao	Wilcoxon rank sum test	NA	29.00	0.77	-	-0.05	-14.00	14.00
IMC_absoluto	Welch Two Sample t-test	7.57	-0.27	0.79	-	-0.13	-3.77	2.99
VO2mLkgminfinal	Welch Two Sample t-test	11.32	-0.11	0.91	-	-0.06	-6.23	5.63

Questão 4

Correlações entre Variáveis

Variável1	Variável2	Teste	df	Corr	p_value	Sign	CI_low	CI_up
DNA_corrigido15x	Idade	Pearson	14	0.19	0.49	-	-0.34	0.62
DNA_corrigido15x	Dias_internacao	Spearman	NA	0.27	0.31	-	NA	NA
DNA_corrigido15x	IMC_absoluto	Pearson	14	0.00	0.99	-	-0.50	0.49
DNA_corrigido15x	VEF1percprevist	Spearman	NA	-0.46	0.08	-	NA	NA
DNA_corrigido15x	VO2mLkgminfinal	Pearson	14	-0.19	0.48	-	-0.63	0.34
DNA_corrigido15x	ReservaVentilatoria	Spearman	NA	0.20	0.45	-	NA	NA
Idade	Dias_internacao	Spearman	NA	0.21	0.43	-	NA	NA
Idade	IMC_absoluto	Pearson	14	0.25	0.36	-	-0.28	0.66
Idade	VEF1percprevist	Spearman	NA	-0.35	0.18	-	NA	NA
Idade	VO2mLkgminfinal	Pearson	14	-0.01	0.98	-	-0.50	0.49
Idade	ReservaVentilatoria	Spearman	NA	0.13	0.62	-	NA	NA
Dias_internacao	IMC_absoluto	Spearman	NA	-0.02	0.95	-	NA	NA
Dias_internacao	VEF1percprevist	Spearman	NA	-0.61	0.01	*	NA	NA
Dias_internacao	VO2mLkgminfinal	Spearman	NA	-0.58	0.02	*	NA	NA
Dias_internacao	ReservaVentilatoria	Spearman	NA	-0.47	0.06	-	NA	NA
IMC_absoluto	VEF1percprevist	Spearman	NA	-0.13	0.62	-	NA	NA
IMC_absoluto	VO2mLkgminfinal	Pearson	14	0.18	0.50	-	-0.35	0.62
IMC_absoluto	ReservaVentilatoria	Spearman	NA	0.02	0.94	-	NA	NA
VEF1percprevist	VO2mLkgminfinal	Spearman	NA	0.24	0.36	-	NA	NA
VEF1percprevist	ReservaVentilatoria	Spearman	NA	0.30	0.26	-	NA	NA
VO2mLkgminfinal	ReservaVentilatoria	Spearman	NA	0.03	0.92	-	NA	NA

Questão 5

Diferenças pelo Efeito da Idade

Variável	Teste	df	Stat	p_value	Sign	Effect	CI_low	CI_up
VEF1percprevist	Wilcoxon rank sum test	NA	45.00	0.19	-	0.20	-8.79	57.67
DNA_corrigido15x	Welch Two Sample t-test	12.10	-0.83	0.42	-	-0.42	-254.01	113.40
IMC_absoluto	Welch Two Sample t-test	9.18	-0.67	0.52	-	-0.34	-4.19	2.27
VO2mLkgminfinal	Welch Two Sample t-test	12.04	0.37	0.72	-	0.19	-4.86	6.86
Dias_internacao	Wilcoxon rank sum test	NA	30.00	0.86	-	-0.03	-14.00	14.00

Outros Resultados

Teste de Normalidade

Variável	n	shapiro_w	shapiro_p	normality
Altura_cm	16	0.912	0.128	parametric
CVFpercprevist	16	0.899	0.076	parametric
DNA_corrigido15x	16	0.931	0.251	parametric
Dias_ATB	16	0.903	0.125	parametric
Dias_internacao	16	0.753	0.001	non-parametric
IMC_absoluto	16	0.909	0.112	parametric
Idade	16	0.894	0.066	parametric
Peso_kg	16	0.906	0.100	parametric
ReservaVentilatoria	16	0.773	0.001	non-parametric
SpO2final	16	0.936	0.307	parametric
VEF1percprevist	16	0.862	0.021	non-parametric
VEabsolutofinal	16	0.952	0.524	parametric
VO2mLkgminfinal	16	0.970	0.845	parametric

Métodos

A análise estatística foi realizada utilizando o software R, versão 4.23 (R Core Team, 2023). São necessarios instalados os packages tidyverse, rio, broom, openxlsx e kableExtra.

Importação e Pre-processamentos dos Dados

Essa etapa consiste na importação dos dados e na codificação dos caracteres especiais para ASCII (remove acentuação, ç e outros que causam erros ao executar o código).

- Os dados do estudo são importados de uma tabela de excel "data.xlsx", que deve conter cada unidade observacional em uma linha e cada variável em um coluna.
- As variáves de interesse devem ser listadas em arquivos de texto abaixo na pasta "input".
- Os arquivos determinam quais variaveis serão incluidas em cada teste estatistico:
  - "stats.txt" = análise descritiva
  - "freqs.txt" = análise de frequência
  - "diffs.txt" = análise de diferenças entre grupos
  - "corrs.txt" = análise de correlações entre variáveis
  - "age_diffs.txt" = análise de diferenças pelo efeito da idade, >= ou < de 20 anos

```
library(tidyverse)
library(rio)
library(broom)
library(openxlsx)
library(kableExtra)
p.value <- 0.05
input <- paste0(
  "input/",
  c(
    "stats.txt",
    "freqs.txt",
    "diffs.txt",
    "corrs.txt",
    "age_diffs.txt"
  )
)
read <- function(path) {
  names(import(path, setclass = "tibble"))
}
encoding <- function(data) {
  lapply(data, function(x) {
    iconv(x, from = "UTF-8", to = "ASCII//TRANSLIT")
  })
}
targets <- lapply(input, read) %>% lapply(encoding)
names(targets) <- c("stats", "freqs", "diffs", "corrs", "age_diffs")
data <- import("input/data.xlsx", setclass = "tibble", na = "NA") %>%
  setNames(iconv(colnames(.),
    from = "UTF-8",
    to = "ASCII//TRANSLIT"
  )) %>%
  mutate(Categoria_Idade = if_else(
    Idade <= 20, "20_ou_menos", "acima_de_20"
  ))
```

Teste de Normalidade

Aqui define-se uma função que aplica o teste de Shapiro-Wilk para verificar se os dados seguem uma distribuição normal. A partir desses resultados, define-se a aplicação de testes paramétricos ou não-paramétricos nas análises subsequentes.

```
shapiro <- function(data, vars) {
  data %>%
    select(all_of(vars)) %>%
    gather(Variável, Valor) %>%
    group_by(Variável) %>%
    summarise(
      n = n(),
      shapiro_w = shapiro.test(Valor)$statistic,
      shapiro_p = shapiro.test(Valor)$p.value,
      normality = if_else(shapiro.test(Valor)$p.value > p.value,
        "parametric", "non-parametric"
      )
    )
}
```

Estatísticas Descritivas

Aqui define-se a função que calcula as estatísticas descritivas para as variáveis de interesse. Nesse caso Idade, Dias_ATB, Dias_internacao,Altura_cm ,Peso_kg, IMC_absoluto, DNA_corrigido15x, VEF1percprevist, CVFpercprevist, SpO2final,VO2mLkgminfinal, VEabsolutofinal, ReservaVentilatória.

```
calc_stats <- function(data, vars, group_by_var = NULL, overall = TRUE) {
  data %>%
    select(all_of(vars), if (!overall) all_of(group_by_var)) %>%
    gather(Variável, Valor, -if (!overall) group_by_var) %>%
    {
      if (overall) {
        group_by(., Variável)
      } else {
        group_by(., Variável, !!sym(group_by_var))
      }
    } %>%
    summarise(
      n = n(),
      missing = sum(is.na(Valor)),
      normality = if_else(shapiro.test(Valor)$p.value > p.value,
        "parametric", "non-parametric"
      ),
      mean = mean(Valor, na.rm = TRUE),
      median = median(Valor, na.rm = TRUE),
      sd = sd(Valor, na.rm = TRUE),
      se = sd(Valor, na.rm = TRUE) / sqrt(length(Valor)),
      iqr = IQR(Valor, na.rm = TRUE)
    )
}
```

Análise de Frequências para Variáveis Categóricas

Aqui calcula-se as frequências absoluta e relativa das variáveis categóricas por grupo. As variáveis de interesse são: Sexo, Mutação, Pseudomonas_cronico, Internação, DNAcat243,SPO2cat, VO2categ.

```
calc_freqs <- function(data, vars) {
  data %>%
    select(Grupo, all_of(vars)) %>%
    gather(Variável, Valor, -Grupo) %>%
    count(Grupo, Variável, Valor) %>%
    group_by(Variável, Grupo) %>%
    mutate(
      Freq_rel = n / sum(n),
      Freq_abs = n
    ) %>%
    select(Variável, Grupo, Valor, Freq_abs, Freq_rel) %>%
    arrange(Variável, Grupo, Valor)
}
```

Diferenças entre Grupos e pelo Efeito da Idade

Aqui calcula-se as diferenças entre grupos e diferenças pelo efeito da idade, >= ou < de 20 anos para cada uma das seguintes variáveis de interesse: Idade, Dias_internacao, IMC_absoluto, DNA_corrigido15x, VEF1percprevist, SpO2final, VO2mLkgminfinal.

```
calc_diffs <- function(data, vars, normality, p.value, group_var) {
  map_df(vars, function(var) {
    normality <- filter(normality, Variável == var)
    is_normal <- normality$shapiro_p > p.value
    test_result <- if (is_normal) {
      t.test(data[[var]] ~ data[[group_var]])
    } else {
      wilcox.test(data[[var]] ~ data[[group_var]], conf.int = TRUE)
    }
    effect_size <- if (is_normal) {
      cohen_d(data, var, group_var)
    } else {
      biserial(data, var, group_var)
    }
    tibble(
      Variável = var,
      Teste = if (is_normal) {
        "Welch Two Sample t-test"
      } else {
        "Wilcoxon rank sum test"
      },
      df = ifelse(is_normal, test_result$parameter, NA),
      Stat = test_result$statistic,
      p_value = test_result$p.value,
      Sign = ifelse(test_result$p.value < p.value, "*", "-"),
      Effect = effect_size,
      CI_low = test_result$conf.int[1],
      CI_up = test_result$conf.int[2]
    )
  }) %>%
    arrange(p_value)
}
```

Cálculo de Tamanho do Efeito

Aqui calcula-se o tamanho do efeito para as diferenças entre grupos e diferenças pelo efeito da idade, >= ou < de 20 anos para cada uma das seguintes variáveis de interesse: Idade, Dias_internacao, IMC_absoluto, DNA_corrigido15x, VEF1percprevist, SpO2final, VO2mLkgminfinal.

```
cohen_d <- function(data, var, group_var) {
  group1 <- data %>%
    filter(data[[group_var]] == unique(data[[group_var]])[1]) %>%
    pull(var)
  group2 <- data %>%
    filter(data[[group_var]] == unique(data[[group_var]])[2]) %>%
    pull(var)
  mean_diff <- mean(group1, na.rm = TRUE) - mean(group2, na.rm = TRUE)
  pooled_sd <- sqrt(((length(group1) - 1) * var(group1, na.rm = TRUE) +
    (length(group2) - 1) * var(group2, na.rm = TRUE))
  / (length(group1) + length(group2) - 2))
  cohen_d <- mean_diff / pooled_sd
  return(cohen_d)
}
biserial <- function(data, var, group_var) {
  data <- data %>% mutate(Rank = rank(data[[var]],
    na.last = "keep",
    ties.method = "average"
  ))
  group1 <- data %>%
    filter(data[[group_var]] == unique(data[[group_var]])[1]) %>%
    pull(Rank)
  group2 <- data %>%
    filter(data[[group_var]] == unique(data[[group_var]])[2]) %>%
    pull(Rank)
  rbc <- (sum(group1) / length(group1) - sum(group2) /
    length(group2)) / length(data[[var]])
  return(rbc)
}
```

Correlações entre Variáveis

Aqui calcula-se as correlações entre as variáveis de interesse: Idade, Dias_internacao, IMC_absoluto, DNA_corrigido15x, VEF1percprevist, SpO2final, VO2mLkgminfinal. Ainda, extrai-se a variável DNA_corrigido15x para análise isolada.

```
calc_corrs <- function(data, vars, norms, p.value) {
  combn(vars, 2, simplify = FALSE) %>%
    map_df(function(pair) {
      norms1 <- filter(norms, Variável == pair[1])
      norms2 <- filter(norms, Variável == pair[2])
      is_normal <- norms1$shapiro_p > p.value &
        norms2$shapiro_p > p.value
      corr_result <- if (is_normal) {
        corr_result <- cor.test(data[[pair[1]]], data[[pair[2]]],
          method = "pearson"
        )
      } else {
        corr_result <- cor.test(data[[pair[1]]], data[[pair[2]]],
          method = "spearman"
        )
      }
      tibble(
        Variável1 = pair[1],
        Variável2 = pair[2],
        Teste = if (is_normal) "Pearson" else "Spearman",
        df = corr_result$parameter,
        Corr = corr_result$estimate,
        p_value = corr_result$p.value,
        Sign = ifelse(
          corr_result$p.value < p.value, "*", "-"
        ),
        CI_low = corr_result$conf.int[1],
        CI_up = corr_result$conf.int[2]
      )
    })
}
```

Apresentação de dados

Define uma função de formatação dos resultados em formato de tabela.

```
custom_kable <- function(table_data, ...) {
  table_output <- kable(table_data, ...)
  return(table_output)
}
```

Exportação dos Resultados

Os resultados foram exportados para arquivos CSV e excel para uso posterior. Os arquivos são salvos na pasta `output`. A tabela principal com os resultados finais está no arquivo "results.xlsx", com cada análise em uma aba da planilha.

Conclusão

Limitações:

A análise ainda carece de toda a parte de visualização dos resultados e dos dados, a qual deveria incluir gráficos de barra, boxplot, matriz de correlação, entre outros.