Universität Zürich[UZH]

Master Program in Biostatistics
www.biostat.uzh.ch

STA490: Statistical Practice in Clinical Research

Prof. Ulrike Held

EBPI, Department of Biostatistics

<**Project Title**>

Analysis for <Client Name>, <Clinic Name>

Supervision by <Supervisor Name>

<Your name> (<*Your email adress*>)

Version of September 19, 2024

1 Abstract

A report of statistical results does not always contain an abstract. You are required to write an abstract here for training purposes.

2 Introduction

Write a short description of the background such that the research questions become clear. This requires explanation of the medical background related to the problem and may also require explanation of the statistical methods to be used (if the methods are not familiar or are complicated). In this report we use Holm (1979); use this as an example of how to cite references.

3 Research Questions

State the questions one by one.

1. What is the median survival time for patients within the study population?
2. Are there differences in survival between patients receiving therapy A and those receiving therapy B?
3. Do certain factors affect the probability of survival while receiving therapy B?

4 Methods

This section should describe how the analysis is conducted, but should not include results (no data!).

Study Design

Discuss details relevant to the design of the study (if applicable in your case), such as:

Type of study

Study population

Describe the composition including inclusion/exclusion criteria

Data collection

Describe specific issues in data collection, variables names, definitions and range.

Primary and secondary outcomes

Statistical Analysis

Describe the statistical methods you use; the more non-standard they are, the longer the description should be. Please describe the following:

Data Preparation

Descriptive Statistics and Simple Methods

Visualization Methods

If you use non-standard plot types, e.g. fan plots, forest plots, etc. briefly describe what they show.

Imputation Methods

Description of advanced statistical methods

Give a short and clear summary of what the purpose of the advanced method is and if possible the main principle how it works (e.g. likelihood estimation or shrinkage).

Implementation

All analyses were performed in the R programming language (R Core Team, 2024) using base packages and the following analysis-specific packages: `lme4` to fit linear mixed effects models (Bates et al., 2015) and `lmtest` to perform diagnostic checks of the resulting models (Zeileis and Hothorn, 2002).

Cite packages which are important in your analysis. You can obtain a bibtex entry by using `citation(package = "packagename")` in an R-session.

5 Results

In this section you write about the results of your data analysis. Make sure that all figures and tables have captions that are complete and self-explaining. Start with the basics, such as a "Table 1" summarizing the subjects under investigation or a flowchart for participant inclusion. Here, Table 1 gives a short overview of the iris dataset using the function `CreateTableOne` from the package `tableone`. The package also contains

functions specific to nominal or continuous variables, but these can usually be automatically detected. Notice the statement clarifying IQR in the caption. The `nonnormal` argument in the `print` statement allows to display median and IQR instead of mean and standard deviation for continuous variables.

Table 1: Descriptive statistics of iris data (n=150). Mean and standard deviation (SD) should be reported for continuous variables with approximate normal distribution, median [first quartile, third quartile] should be reported for skewed continuous or ordinal variables and frequency (%) for categorical variables. The column *Missing (%)* shows % of missing values.

Variable	Overall	setosa	versicolor	virginica	Missing (%)
n	150	50	50	50	
Sepal.Length (mean (SD))	5.82 (0.82)	5.01 (0.35)	5.94 (0.52)	6.56 (0.63)	1.3
Sepal.Width (mean (SD))	3.06 (0.44)	3.43 (0.38)	2.77 (0.31)	2.97 (0.32)	0.0
Petal.Length (mean (SD))	3.76 (1.77)	1.46 (0.17)	4.26 (0.47)	5.55 (0.55)	0.0
Petal.Width (median [IQR])	1.30 [0.30, 1.80]	0.20 [0.20, 0.30]	1.30 [1.20, 1.50]	2.00 [1.80, 2.30]	0.0
Sepal.Length.Cat (%)					1.3
length <=5 cm	32 (21.3)	28 (56.0)	3 (6.0)	1 (2.0)	
length >5 cm	116 (77.3)	22 (44.0)	47 (94.0)	47 (94.0)	
Missing	2 (1.3)	0 (0.0)	0 (0.0)	2 (4.0)	

Consider that you can also extract data into your text with `\Sexpr{}`, e.g., the `Species` variable contained 50 records each. In Figure 1 we show a pairs plot of the Iris data.

The data description should be followed by results per research question. Make clear which results answer research questions and which results come from additional analyses. These additional results should be presented after the main results.

Median Survival for Patients with Disease A

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc nec tincidunt leo. Sed eleifend ex nunc, ut suscipit quam luctus et. Praesent et faucibus ligula, eu facilisis dolor. Praesent non nunc ipsum. Curabitur iaculis justo eget eros dignissim, ornare luctus leo fringilla. Nunc vel metus blandit, blandit metus ac, fringilla nibh.

Table 2: Linear regression model choosing some columns.

	Coefficient	95% confidence interval	p-value
Intercept	2.60	from 2.03 to 3.17	$<$ 0.0001
Sepal.Width	0.34	from 0.16 to 0.52	0.0003
Petal.Length	0.79	from 0.66 to 0.93	$<$ 0.0001
Petal.Width	-0.26	from -0.54 to 0.03	0.079
Speciesversicolor	-0.95	from -1.43 to -0.48	0.0001
Speciesvirginica	-1.24	from -1.89 to -0.60	0.0002
Sepal.Length.Catlength $>$5 cm	0.32	from 0.16 to 0.49	0.0002

Table 3: Linear regression model with adapted row names.

	Coefficient	95%-confidence interval	p-value
Intercept	6.46	from 5.52 to 7.39	$<$ 0.0001
Width Sepal	-0.21	from -0.51 to 0.10	0.18

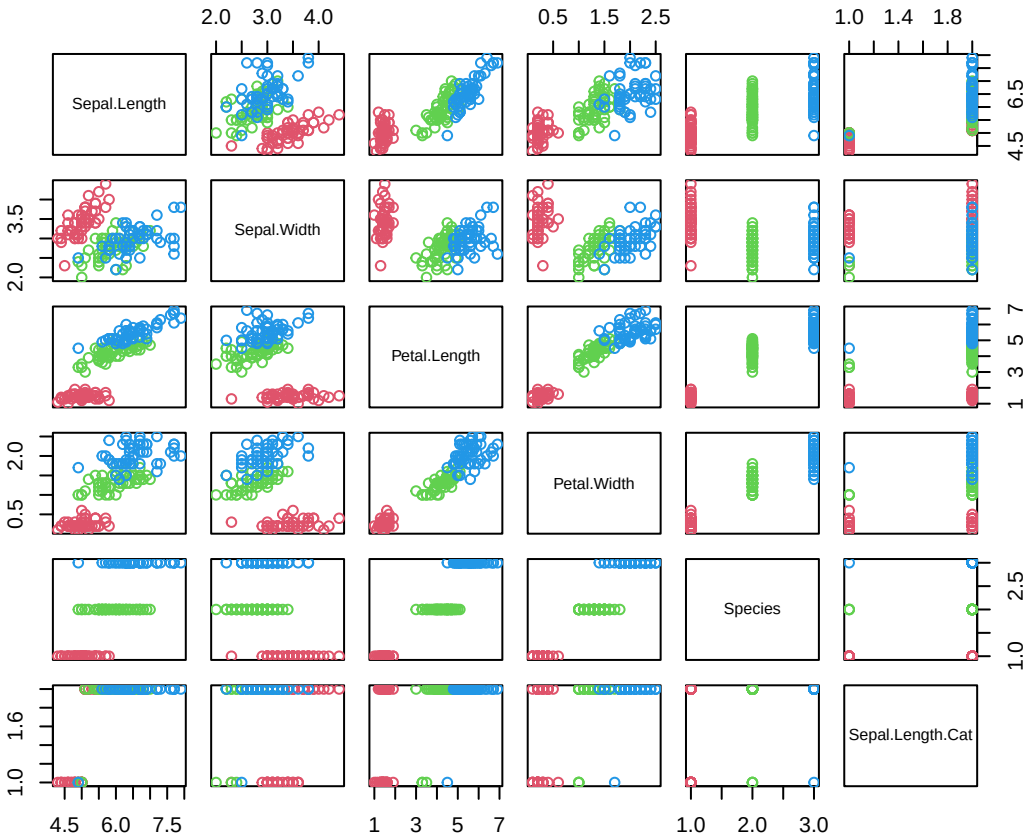

Figure 1: Pairs plot for iris data.

Effect of Therapy on Survival

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc nec tincidunt leo. Sed eleifend ex nunc, ut suscipit quam luctus et. Praesent et faucibus ligula, eu facilisis dolor. Praesent non nunc ipsum. Curabitur iaculis justo eget eros dignissim, ornare luctus leo fringilla. Nunc vel metus blandit, blandit metus ac, fringilla nibh.

Table 4: First six records of the iris dataset.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Sepal.Length.Cat
5.10	3.50	1.40	0.20	setosa	length >5 cm
4.90	3.00	1.40	0.20	setosa	length <=5 cm
4.70	3.20	1.30	0.20	setosa	length <=5 cm
4.60	3.10	1.50	0.20	setosa	length <=5 cm
5.00	3.60	1.40	0.20	setosa	length <=5 cm
5.40	3.90	1.70	0.40	setosa	length >5 cm

Table 5: Contingency table for iris data.

Species	Petal.Width < 1.5	
	FALSE	TRUE
setosa	0	50
versicolor	15	35
virginica	49	1

Factors Affecting Survival of Patients Receiving Therapy B

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc nec tincidunt leo. Sed eleifend ex nunc, ut suscipit quam luctus et. Praesent et faucibus ligula, eu facilisis dolor. Praesent non nunc ipsum. Curabitur iaculis justo eget eros dignissim, ornare luctus leo fringilla. Nunc vel metus blandit, blandit metus ac, fringilla nibh.

6 Conclusion

Summarize your conclusions regarding each research question. Each question listed in Section 3 should be addressed. Hard numbers belong in Section 5. Your conclusions should provide interpretations of your results in words.

The median survival time found in this study supports the growing literature on disease prognosis for patients with Disease A. Our results suggest that Therapy B is just as effective as Therapy A at fighting Disease A, but with less side effects. Similar to other studies, we found that Factor Y is an important predictor of long-term mortality. We also found Factor Z to be predictive of long-term mortality. This may be due to Reason X. Possible interactions between Factor Z and Therapy B require further research.

Limitations of this study include Q, R, and S.

References

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67** 1–48.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6** 65–70.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News* **2** 7–10.
URL `https://CRAN.R-project.org/doc/Rnews/`

7 Appendix

7.1 Computational Details

This document was generated on September 19, 2024 at 12:14. R version and packages used to generate this report:

```
## R version 4.2.0 (2022-04-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur/Monterey 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] stringr_1.5.1      ggplot2_3.4.4      biostatUZH_2.2.4   MASS_7.3-56
## [5] survival_3.5-7     xtable_1.8-4       tableone_0.13.2    RColorBrewer_1.1-3
## [9] knitr_1.45
##
## loaded via a namespace (and not attached):
##  [1] zoo_1.8-12             tidyselect_1.2.0       xfun_0.41
##  [4] listenv_0.8.0          mitools_2.4            haven_2.5.4
##  [7] splines_4.2.0          lattice_0.20-45        labelled_2.9.1
## [10] generics_0.1.3         colorspace_2.1-1       vctrs_0.6.5
## [13] utf8_1.2.4             prodlim_2019.11.13     rlang_1.1.2
## [16] e1071_1.7-14           nloptr_2.0.3           pillar_1.9.0
## [19] withr_3.0.0            glue_1.6.2             DBI_1.1.3
## [22] lifecycle_1.0.4        lava_1.7.0             munsell_0.5.0
## [25] gtable_0.3.4           future_1.29.0          codetools_0.2-18
## [28] evaluate_0.23          forcats_1.0.0          psy_1.2
## [31] class_7.3-20           parallel_4.2.0         fansi_1.0.6
## [34] highr_0.10             ReplicationSuccess_1.2 Rcpp_1.0.11
## [37] scales_1.3.0           cmprsk_2.2-11          parallelly_1.32.1
## [40] lme4_1.1-35.1          hms_1.1.3              digest_0.6.33
## [43] stringi_1.8.3          dplyr_1.1.4            survey_4.1-1
## [46] grid_4.2.0             cli_3.6.2              tools_4.2.0
## [49] magrittr_2.0.3         proxy_0.4-27           tibble_3.2.1
## [52] future.apply_1.10.0    pkgconfig_2.0.3        Matrix_1.6-4
## [55] minqa_1.2.6            rstudioapi_0.15.0      R6_2.5.1
## [58] globals_0.16.2         boot_1.3-28            nlme_3.1-157
## [61] compiler_4.2.0
```

7.2 Code

Please provide ALL code with which you produced this report by code chunk reuse, i.e. name all your chunks and display them here by typing «`chunkname`». You can also display code from external scripts, see the example for data preparation below. Try to format your code such that it fits in the lines. include comments to indicate which section the chunks were used in.

```
##############################################################################
# code for packages, settings
##############################################################################
```

```r
## Import external functions
## ------------------------------------------------------------------------

## Packages
## ------------------------------------------------------------------------
library(RColorBrewer) # colors for plots
library(tableone) # for Table 1 functions
library(xtable) # formatting tables and generating the tex code
library(biostatUZH) # EBPI-written package, if not installed, uncomment code below
#devtools::install\_github(repo = "felix-hof/biostatUZH")
library(ggplot2) # customizable plots
library(stringr) # to prettify tables
### include project-specific packages here as well (e.g., lme4 for linear mixed effects models)
### if possible do not load libraries in chunks further below or in scripts that you source

################################################
### only load libraries that you really use!####
################################################

## Additional settings
## ------------------------------------------------------------------------
cols <- brewer.pal(3, "Set1")
options(width = 85, digits = 4, show.signif.stars = FALSE)
```

```r
##############################################################################
# code for data preparation from external script
##############################################################################
dat.raw <- read.csv("../data/iris_data_20121206.csv")


## Final Dataset
## ------------------------------------------------------------------------
dat <- dat.raw


## Data Preparation
## ------------------------------------------------------------------------

# for teaching purposes, a few numbers are set to missing
obs.to.missing <- sample(x=length(iris$Sepal.Length), size=2)
dat$Sepal.Length[obs.to.missing] <- NA

# to include a categorical variable, we create one here
dat$Sepal.Length.Cat <- cut(dat$Sepal.Length, breaks=c(0, 5, 8),
                            labels=c("length <=5 cm", "length >5 cm"))
```

```r
##############################################################################
# code for results: descriptive results
##############################################################################
plot(dat, col = as.numeric(factor(dat$Species)) + 1)

##############################################################################
# code for results: first research question
##############################################################################
mod.lm <- lm(Sepal.Length ~ Sepal.Width, data = dat)
mod.lm1 <- lm(Sepal.Length ~ ., data = dat)
```

```
## choosing columns
tableRegression(mod.lm1, stats = c("estimate", "ci.95", "p.value"),
                col.nam = c("Coefficient", "95\\% confidence interval", "$p$-value"),
                caption = "Linear regression model choosing some columns.",
                caption.placement = "top",
                label = "tab:regmod1", booktabs = TRUE)

## adapt row names
tableRegression(mod.lm, row.nam = c("Intercept", "Width Sepal"),
                stats = c("estimate", "ci.95", "p.value"),
                col.nam = c("Coefficient", "95\\%-confidence interval", "$p$-value"),
                caption = "Linear regression model with adapted row names.",
                caption.placement = "top",
                label = "tab:regmod2", booktabs = TRUE)

################################################################################
# code for results: second research question
################################################################################
## important for tables: set 'results = "asis"' in the knitr chunk options

## single header
## -----------------------------------------------------------------------------
mat <- head(dat)

mat.xtab <- xtable(mat, align = "lrccccc",
                   caption = "First six records of the iris dataset.", label = "tbl:head")
print(mat.xtab, size = "footnotesize", table.placement = "!ht",
      caption.placement = "top", include.rownames = FALSE,
      hline = c(-1,0, nrow(mat.xtab)), sanitize.text.function = function(x){x},
      booktabs = TRUE)


## with additional header
## -----------------------------------------------------------------------------
mat <- table(dat$Species, dat$Petal.Width < 1.5)


addtorow <- list()
addtorow$pos <- list()
addtorow$pos[[1]] <- -1
addtorow$command <- c('\\hline Species & \\multicolumn{2}{c}{Petal.Width $<$ 1.5} \\\\')

mat.xtab <- xtable(mat, align = "r|cc",
                   caption = "Contingency table for iris data.",
                   caption.placement = "top", label = "tbl:cont")
print(mat.xtab, size = "footnotesize", table.placement = "!ht",
      caption.placement = "top", include.rownames = TRUE,
      hline = c(0,nrow(mat.xtab)), add.to.row = addtorow,
      sanitize.text.function = function(x){x})
```