09-Which fruits are the biggest? An ANOVA example

2024-12-16

Preface

The following template is intended as a starting point for a short report of an ANOVA. It is recommended to follow the general outline (Introduction, Methods, Results, Discussion), but to adapt the content to the specific needs of the analysis.

It is good practice to concentrate on the main findings and their discussion and to avoid unnecessary technical detail. The report should be illustrated with meaningful (and only the most important) tables and graphs.

The data set consists of different samples of clementine fruits from different brands and different shops. The data sets can be downloaded from

https://tpetzoldt.github.io/datasets/.

A description is found in the file clementines_info.txt.

1 Introduction

Describe briefly the setup of the experiment. Which data sets are to be compared by your group?

What are the Null hypothesis (H_0) and the alternative hypothesis (H_A) in this experiment?

2 Material and Methods

2.1 Data

- write something about the data
  - describe the different samples and how they were obtained
  - weight determination with a scale, length determination with a caliper (see below)
- read the data in:

```
brands <- read.csv("clementines2022-brands.csv")
fruits <- read.csv("clementines2022-fruits.csv")
```

Show the structure of the data. Use the data explorer of RStudio or the `head`-function, that shows the first lines of each data set. **Please do not forget to report the sample size of your data!**

```
cat("sample size:", nrow(fruits), "\n")
head(fruits, n=3)
```

Look also at the structure of `brands`.

Then **join** the two tables and convert the brand column into a factor variable.

```
dat <- left_join(fruits, brands, by="brand")
dat$brand <- factor(dat$brand)
```

2.2 Data analysis

Mention R (R Core Team, 2024) in a single sentence, cite special packages. Write that an ANOVA was performed and which methods were used in addition.

3 Results

- Visualize the data first, show box plots, either the full data set **or** selected samples.
- explain briefly what you see in the box plot
- perform an ANOVA:
  - the ANOVA
  - check variance homogeneity and approximate normality
  - perform a post-hoc test
- give a short explanation of the ANOVA results and report effect sizes

Creative part: You can also be somewhat creative and analyse further properties of the data, e.g. length, width, weight, … Select appropriate test methods, e.g. t-tests, ANOVA, correlation or regression, etc. You may also use **tidyverse** methods (e.g. `group_by` and `summarize`) to summarize the data (Wickham et al., 2019). It can also be possible to try a two-way ANOVA with a subset of the data. In that case, make sure to have a balanced data set or use `Anova` function (with capital letter "A") from package **car**.

4 Discussion and Conclusions

The discussion should contain three parts:

1. The "scientific" toy problem: Summary of the outcome of the ANOVA experiment (focus on the fruits).
2. Technical aspects about the ANOVA or other methods:
   - do the results look plausible and sound? Can things be made better?
   - are there any comments to the method?
   - Which results were obtained from additional analyses from the creative part?

3. A conclusion, how can the results can be interpreted. Which clementines would you buy? Why?

5 Appendix

5.1 Hypotheses

It is important to formulate clear hypotheses. Here are a few **examples**, related to the data set from 2019. Please think about it and define your own hypotheses, related to the **current** data set.

- $H_{0,1}$: The mean weight of the fruits is the same in all samples.
- $H_{A,1}$: The weight is different in any of the samples.
- $H_{A,2}$: Which brand is the smallest? The mean weight of smallest sample is significantly smaller than of the 2nd samplest sample.
- $H_{A,3}$: The fruits from the premium brands are bigger than corresponding basic brands.

Note: Hypotheses $H_{A,2}$ and $H_{A,3}$ have their own, different H_0.

A hypothesis like $H_{A,3}$ is more difficult and optional. It requires a two-way ANOVA and can only be applied to a subset of the data, where different brands from the same shops are available.

5.2 Measurement methods

- **Weight**: was determined with a kitchen scale in gramm (g).
- **Height and Width**: were measured with a **caliper** (Fig. 1)

Figure 1: Size measurement with a digital caliper.

5.3 R example code

- The following code is intended as a starting example. It is recommended to adapt the script to analyse the data from another year.
- Don't forget that it is an exercise, not a serious analysis, so feel free to create your own story.
- Don't make your report too technical, concentrate on your message.

```r
brands <- read.csv("https://tpetzoldt.github.io/datasets/data/clementines2019-brands.csv")
fruits <- read.csv("https://tpetzoldt.github.io/datasets/data/clementines2019-fruits.csv")

dat <- left_join(fruits, brands) # merge tables by their common colum 'Brand'
dat$brand <- factor(dat$brand)

boxplot(weight ~ brand, data=dat)

## the ANOVA
m <- lm(weight ~ brand, data=dat)
anova(m)

## posthoc test
TukeyHSD(aov(m))

## graphical display of Tukey's test
plot(TukeyHSD(aov(m)), las=1, cex.axis=0.5)

## graphical and numerical checks of variance homogeneity
plot(m, which=1)
fligner.test(weight ~ brand, data=dat)

## approximate normality of residuals
plot(m, which=2)

## optional: special one-way anova alternative if variances are unequal
oneway.test(weight ~ brand, data=dat)
```

5.4 Text processing

The report can, in principle, be written in any text processing software, e.g. Microsoft Word or LibreOffice Write, but I recommend to try Quarto. It needs only a little extra learning, but is an extremely efficient way to combine text and analysis with **R** in one document and write it directly in RStudio. A comprehensive documentation can be found on https://quarto.org/.

5.5 Questions and literature research

Please post questions, comments and ideas to the chat group. As the exercise is a toy example, it may be difficult to find relevant citeable literature. However, we don't limit creativity.

References

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686