Lab 2: ANOVA

David Issa Mattos - (davidis@chalmers.se)

25 Nov 2020

Aim

The goals of this first lab in the Empirical Software Engineering course are:

- Familiarize with conducting experimental design analysis in R statistical, providing descriptive statistics, regression models and performing simple hypothesis testing.
- These exercises are meant to be a basis for discussions within the groups.
- Do the exercises but stop and reflect along the way why you did what you did, and the usefulness of it.

Hand-in

- The report must be written in English.
- Hand in the answers as a pdf through Canvas (One hand-in per group).
- Write concise and direct answers to the questions. For most questions 2-3 lines is enough.
- Questions (outside the lab time) should be sent to the group supervisor.
- Unless it is a graphic/plot DO NOT add print screen from the R output or copy the whole output (I know what I should look for, but I want to be sure you know which parts are relevant as well). If adding a figure or a table, briefly explain what it is about and reference it in the text. I suggest using Rmarkdown to generate the report from the code automatically. See notes on Canvas and in the Help! section.
- You should include the code that you have written as well as the results and the discussion points. Only showing the code is not enough.
- All reports go through plagiarism check
- **You must also include a statement of contribution in your report for each of your group members.**
- The contribution report must have
- How each member participated
- If they attended the lab session or not
- It is not enough to write that everyone did everything. Deadline for submission: **Dec 4, 2020**

Pre-requisites

We assume you have already have basic knowledge with R by working through lab 1 (including statistical tests, import data and create a linear regression model).

Ideally you will also have gone through the very short descripition of the basic statistical modeling in: https://www.statmethods.net/stats/index.html

We provide code snippets and suggested functions in the report.

Suggested libraries

```
library(tidyverse)
library(car)
library(knitr)
library(pwr)
library(psych)
```

In particular, for importing, data manipulation, transformation and visualization we recommend the libraries from `tidyverse`.

A/B/n testing

A Gothenburg-based startup specialized in streaming on Scandinavian film and TV series, Götaflix, recently saw a blog post from its largest competitor (https://medium.com/netflix-techblog/selecting-the-best-artwork-for-videos-through-a-b-testing-f6155c4595f6) that discusses the impact that the movie artwork has on capturing the audience and increasing customer engagement. Impressed by the results, they decided to run their own experiments to understand better how their customers interact and engage with their content. For their first experiment, they selected a Scandinavian-noir crime TV series and created 4 art covers and to be tested against the original art cover from the studio. The experiment is planned to be run for a full month

They designed a one-way ANOVA experiment varying 5 levels (original cover + 4 new ones) of 1 factor. They designed a randomized balanced experiment for all levels (20% allocation for each level and users were randomly allocated to an experimental group) and looked at their customer engagement metric for this specific series.

The studio cover (the control version) had an average engagement of 0.162. Based on the blog post and the relation between their revenue and customer engagement, they decided that having at least 8% improvement with a confidence of 95% compared to the control and a power of 90% was sufficient for them to invest more resources in developing better art covers and run future experiments.

Questions

a

Before running an experiment, they need to plan the minimum number of users that will be randomized to each art cover, and therefore the time the experiment is going to take. As you might have guessed, this planning is done by conducting a power analysis. What is the minimum number of customers in each group for the experiment to have this power? How many monthly customers must the company have to be able to run this experiment in 1 month? What happens with the minimum number of customers needed if the effect size decreases (and power and significance level remain constant)? What happens if the effect size increases?

- Suggested function: `pwr::pwr.anova.test`

b

In one month, the company collected the data (in long format) in the file gotaflix-abn.csv. Import this data to R and generate descriptive statistics for each group.

- Tip: see lab1
- Don't forget to convert the independent variable into a factor

c

Generate a linear model to fit this experiment data. Write the equation that represents this model. What does the intercept mean? What is the value the model gives when only the coefficient for Cover C equals to 1 and the rest of the coefficients equals to 0?

d

One of the assumptions of a linear model is 'normality'. What needs to be normal? Analyze the normality of the model using a qqplot and then use a Shapiro test. Does the Shapiro-Wilk test agree with the qqplot? Which one do you choose to justify the normality?

- You can use the plot function on the `lm` model or use the `qqPlot` function from the `car` package on the `lm` model as well
- To access the residuals of a `lm` model you can do `mylmmodel$residuals`
- Also read: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

e

The other assumptions of the one-way ANOVA are homoscedasticity of the residuals and independence of the data. Create a scatter plot for to analyze the homogeneity of the variances and them conduct a Levene's test.

- The scatter plot is generated automatically when you run the function plot on the fitted linear model
- You can run the Levene's test with the function leveneTest from the car package with the lm model as argument
- A comment on the use of the Levene's test with large sample size: As with any other statistical test for the assumptions, if you have a large sample size you can detect very small differences, sometimes these differences are not relevant and sometimes is just a difference on the sample. Don't rely on tests for assumptions only. **Always look at other techniques to make a decision**

f

Discuss the independence assumption. How can we verify it? Is it part of the design of the experiment or the analysis?

g

Suppose you have received the data of the gotaflix-abn-modified.csv. Does the data have homoscedasticity?

h

Analyze the data (gotaflix-abn.csv) and discuss which art cover had a better engagement? First, run an Anova test and see if the model is statistically significant. Second, run the Tukey HSD post-hoc test and analyze the results?

- The `summary(myLinearModel)` function already gives the F statistics for the model. But if you want to see the ANOVA table that we usually present in class you can use the function `Anova(myLinearModel)` from the `car` package
- Note. You don't need to describe what is each type of Sum of Squares but you need to comment which you used. R has a default function for ANOVA called `aov`. For balanced designs (same number of measurements in each experimental group) the `aov` and the `Anova` functions are equivalent. However, for unbalanced designs you often want to specify which type of sum of squares you want to perform. The `aov` performs type I or sequential, while the Anova can perform type II (default) and type III. The different choices for types of sum of squares give different results. In general, we want sum of squares type II (more powerful for main effects). However, if we are using sum of squares type III we need to define the contrasts of the factors as a sum-to-zero (using the function `contr.sum`) For more information see: https://www.r-bloggers.com/anova-%E2%80%93-type-iiiiii-ss-explained/ and http://md.psych.bio.uni-goettingen.de/mv/unit/lm_cat/lm_cat_unbal_ss_explained.html
- The Tukey's post-hoc test can be run with the function `TukeyHSD`, but this function receives an `aov` object. E.g. `TukeyHSD(aov(myLinearModel))`

Full factorial experiment

After this successful first experiment, Götaflix decided to evaluate how their users engage with the combination of new art covers and new summary of the movies. They selected 50 random movies among the top 200 most watched and created an experiment where they varied the art covers and the summaries. They carefully designed for each movie two art covers: one that focuses on the main characters, and one the focuses on the genre. For the same movies they also carefully designed two summaries. One that focuses on the genre of the movie and one that focuses on qualities of the main characters. The users are randomly assigned for each one of the experimental groups.

Questions

a

How many and what are the experimental groups?

b

The design of this experiment can be seen as a two-way ANOVA. Load the data (gotaflix-2wayANOVA.csv) and fit a linear model. Write the equation of this linear model. What does it mean the intercept value?

- In R, we can represent interactions between A and B in different ways. A and B are independent variables and C is the dependent variable
  - Consider only the main effects C ~ A + B
  - Consider the main effects plus the interaction C~ A + B + A:B
  - Represent the main effects and the interactions C~ A*B
  - Use the `lm` function and fit. You might want to try the last two ways just to verify that they are equivalent

c

Analyze the assumptions of this two-way ANOVA. Independence, normality and homoscedasticity

d

Present the ANOVA table. What is statistically significant? What are the effect sizes? What conclusion can you make about this experiment?

- You might want to run a post hoc test and look at the descriptive statistics carefully to understand the result in more details

Help!

Decimal points.

- You can set a global option for number of digits in r `options(digits=2)`
- But many functions, have the digits option as a parameter. Look at the manual/help for each function
- If you want descriptive statistics for each group you can group the variables in a list with the `psych::describeBy`, e.g: `describeBy(df2$Engagement,list(df2$Cover,df2$Summary),mat=T)`, the option mat=T returns a matrix with more decimal points than is usually printed

Tables ready from lm models

- There are many packages that generates publication ready tables (in Latex) directly from R models (including descriptive statistics). One good example is the `stargazer` package that has many configurations options. https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

- See example below. If you are using Rmarkdown, remember to set the results in the chunck to 'asis' so the latex table is properly rendered

```
fakedataset<-data.frame(a=rnorm(10, mean=1, sd=2),
                        b=rnorm(10, mean=1.3, sd=2),
                        c=2+10*rnorm(10, mean=1, sd=2) + 8*rnorm(10, mean=1.3, sd=2))
```

```
stargazer::stargazer(lm(c ~ a + b, data = fakedataset), header = F)
```

Table 1:

	Dependent variable:
	c
a	−2.258
	(7.909)
b	1.683
	(5.696)
Constant	27.540*
	(13.697)
Observations	10
R^2	0.027
Adjusted R^2	−0.252
Residual Std. Error	33.137 (df = 7)
F Statistic	0.095 (df = 2; 7)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
stargazer::stargazer(fakedataset,summary = T, header = F)
```

Table 2:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
a	10	1.059	1.406	−2.491	0.967	1.943	2.459
b	10	0.347	1.952	−3.706	−0.364	1.834	2.604
c	10	25.734	29.620	−14.780	4.932	51.939	67.100