Data Analysis and Visualization in R (IN2339)

Exercise Session 6 - Graphically supported hypotheses

Daniela Klaproth-Andrade Felix Brechtmann Dominik Eichhorn
Pedro Tomaz da Silva Julien Gagneur

Quizzes (from the lecture)

The following quizzes will be solved orally by the students and the professor during the lecture.

1. Which of the following claims could be an example of reversing cause and effect?
    1. Healthier diets increase blood pressure.
    2. Low social status leads to a higher risk of schizophrenia.
    3. The number of fire engines on a fire gives rise to higher damages.
    4. Entering an intensive care unit increases your chances of dying.
2. A study conducted in Datavizland by Woman's magazine analyzed the dating preference of women.
    - A study conducted in Datavizland by Woman's magazine analyzed the dating preference of women. For each previous or current partner, **women** were asked to **evaluate men on two parameters from 0-10: How handsome and how fun they are**.
    - The study concluded: " **We report a negative correlation between how handsome and how fun a man is** , therefore we conclude that handsome men are boring".
    - In the same week a study conducted on the **whole population of men** by Men's magazine reported **no correlation between how handsome and how fun a man is.**

How can you explain this apparent paradox? Which kind of causal diagram describes this scenario?

1. Common cause
2. Indirect cause
3. Common consequence

3. The ministry of health of Datavizland observed a positive correlation between the liters of water drunk per day and sunburns. Which kind of causal diagram may best describe this scenario?

1. Common cause
2. Indirect cause
3. Common consequence

4. Which item is no chart junk?
    1. A bright red plot border
    2. Light grey major grid lines
    3. Bold labels and grid lines
    4. Data labels in Batik Gangster font

Tutorial

The following exercises will be solved during the tutorial sessions.

Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)
```

Section 01 - Color guidelines

What are best practices when using color for data visualizations? Select all that apply.

1. Avoid having too many colors for categorical data.
2. Use color only when it actually adds meaning to the plot.
3. Use divergent color scales for categorical data types.

Section 02 - Correlation and Causation

Read the following statements. Decide for each statement separately whether it is true or false and give an explanation.

1. The concept of reverse causality states that whenever A causes B, B also causes A.
2. If A causes B and A causes C, then B also causes C.
3. If A and B correlate and A happens before B, then A causes B.
4. Causation implies linear association.

Section 03 - Effect of a third variable

Investigate the file `coffee_sim.csv` (simulated dataset) by first loading it as a `data.table`.

```
coffee_dt <- fread("./extdata/coffee_sim.csv")
coffee_dt
summary(coffee_dt)
```

1. Suggest an appropriate visualization and implement it with ggplot2 to display a possible association between coffee consumption and "datavizitis" disease risk, measured in deaths per 1000 individuals. Does this plot by itself seem consistent with a causal effect of coffee on datavizitis?

2. Investigate the full dataset. Do you see evidence for a third variable influencing association? Support your statement with an appropriate plot. Draw a graph with the potential causal relationships you find consistent with the data. Relate it to one of the situations from the lecture script's figure 6.3 or Simpson's paradox.

Section 04 - General guidelines in data visualization

Below is a graph taken from one published paper. Read the figure legend.

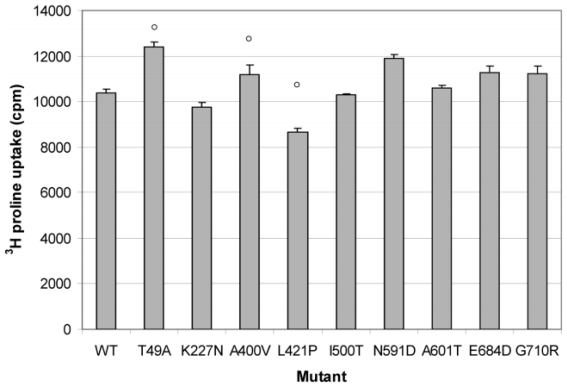

Figure 2. Maximal ^{3}H proline uptake of wildtype (WT) and all tested mutants. The maximum in uptake was measured in the presence of 3 µM cold L-proline. Data are expressed as means ± standard deviation (SD) obtained from triplicate samples. Mutants with a circle were tested in a second independent experiment.
doi:10.1371/journal.pone.0068645.g002

In biology, a wild type strain (WT) is a strain whose genome has not been artificially modified. In contrast, a mutant is a strain whose genome has been artificially modified.

1. Discuss in groups what could be better representations. Use pen and paper. There can be many options.

2. Implement the solution proposed by the tutor.

```
# simulate data
dt <- data.table(pro_uptake = c(rnorm(3, 10100, 300), rnorm(4, 12100, 300),
                                rnorm(3, 9850, 300), rnorm(4, 11100, 300),
                                rnorm(4,8300, 300), rnorm(3,10050, 300),
                                rnorm(3, 12000, 300), rnorm(3, 10020, 300),
                                rnorm(3, 10080, 300), rnorm(3, 10070, 300)  ),
                 mutants = c(rep("WT",3), rep("T49A",4), rep("K227N",3), rep("A400V",4),
                             rep("L421P",4), rep("I500T",3), rep("N591D",3),
                             rep("A601T",3), rep("E684D",3), rep("G710R",3) )          )
```

Homework

Please solve the exercises below at home. The solutions will be discussed in the central exercise.

Section 05 - Covid19 fatality rate in Belgium

Consider the following dataset which contains the fatality rate (percentage of deaths among infected) by COVID 19 of different age groups in Belgium in June 2020 (taken from De Smet, D. (2020). Is corona erger dan de griep? De Standaard, 22 June 2020). The overall rate (`all_ages`) suggests that being a female increases the risk of dying from COVID-19 upon infection. Using appropriate plots discuss the validity of this hypothesis and draw a graph with the potential causal relationships you find consistent with the data. Relate it to one of the situations from the lecture script's figure 6.3 or Simpson's paradox.

```
fatality_dt <- fread('extdata/belgium_infection_fatality_rate_june2020.csv')
fatality_dt
```

Section 06 - Smoking and Datavizitis severity

Consider the following dataset contains a population of 2000 individuals who got datavizitis. It consists of the number of cigaretes each individual smokes per day, the severity of their datavizitis and if they were hospitalized or not.

```
datavizitis_smoking_dt <- fread("./extdata/datavizitis_smoking.csv")
datavizitis_smoking_dt
```

1. Visualize the relationship between the number of cigarettes smoked per day and datavizitis severity among hospitalized individuals. Make use of `geom_smooth(method="lm")` to highlight the general trend.

2. Visualize the relationship between datavizitis severity and cigarettes smoked per day among all population. Make use of `geom_smooth(method="lm")` to highlight the general trend.

3. Visualize the same relationship distinguishing between hospitalized and all individuals.

4. Recent studies have looked at hospitalized patients who tested positive for Covid19 and their smoking status. They propose smoking may provide a lower risk of developing severe Covid19 based on a negative association between Covid19 severity and smoking status. Considering the previous results on datavizitis can you come up with a different explanation? Draw a graph with the potential causal relationships you find consistent with the data. Relate it to one of the situations from the lecture script's figure 6.3 or Simpson's paradox.

Section 07 - Supporting hypotheses with visualizations

1. Read the `titanic.csv` file into a `data.table`. You can read the description of the dataset on kaggle: https://www.kaggle.com/c/titanic/data.

2. Inspect the data table and make a summary of the variables in the dataset. What is the overall passenger survival rate?

3. Does age associate with survival? Make a plot showing the distribution of age per survival outcome.

4. Visualize the relationship between passenger class and survival rate.

5. How is age distributed in each passenger class?

6. Considering the passenger class, do age and survival outcome associate? Given the findings on question 4, comment on the results. Draw a graph with the potential causal relationships you find consistent with the data. Relate it to one of the situations from the lecture script's figure 6.3 or Simpson's paradox.