Statistics Tutorials by Lisa

Winter Semester 2018/2019

My philosophy is to search for the “common ground”, the starting point in each topic. I try to explain the new topic at the earliest point the students will understand to then progress through the topic. I highly value clear visuals, and that’s why **my slides are the absolute go-to for my lecture**. The visuals will tell you everything you need, and this document is mainly an abbreviation of the topics I covered.

Also, (almost) everything I wrote on the **blackboard** is also in the slides and marked as such.

# Day 1

* Introduction
* **Hans Rosling’s Statistics Quiz** 
  + I took it in class using a poll made with <https://pollev.com/> by Heike, the original quiz can be found here: [factfulnessquiz.com](http://factfulnessquiz.com)
  + I discussed the results of the quiz, why we score as bad as we do and why we need statistics
  + I was done pretty early

# Day 2 - Data types

* I prepared **flipchart** paper with the following questions:
  + 1) Where was your starting point this morning in relation to this class room (N,S,W,E)?
  + 2) What distance did you travel to get here? write down the km.
  + 3) How are you feeling this morning?
  + 4) How did you get here? on foot, bike, car, train, other
* I used the flipchart as a warm-up exercise. If you are planning on doing it, make sure to give each student an ID, otherwise you won’t be able to do any tests with the data. A filled out flipchart could look like this:
  + Categories 1-5 as a table, the students enter their ID (eg for question 3)
  + Student writes ID + his or her value, eg 13: 254 km; 14: 25km…
* I then introduced **data types** based on the experiences with the flipchart exercise. How do you reply to such a question? What different types are possible?
* Data types are mainly **discrete** vs. **continuous**. I had trouble with the distinction interval-ratio and would most likely not introduce it again.
* Data types **exercise** where each student replied to one question
* R introduction: what’s R? How do you set a working directory (**I highly recommend looking at my slides**)
* *02\_Basics.R*
  + I would skip my input on packages as we agreed to not teach them R (for which you’d need packages) but what the commands in R are
* **Homework**: swirl

# Day 3 - Distributions

* I wasn’t in Lüneburg on that day
* my students went to Heike’s and Olli’s tutorials
* I nevertheless prepared basic slides, they are uploaded in the drive, introducing:
  + **Probability distributions**: discrete ones yield probability mass functions, continuous yield probability density functions
  + Introduction to the **normal distribution**
  + Is my data normally distributed - examples
  + Why do we need distributions?
  + Introduction to **other types of distributions**, eg poisson
  + **Mean, median and mode** and what they have to do with distributions
  + **Homework**: Swirl

# Day 4 - Simple tests

* Video by Sal Khan on significance: <https://www.youtube.com/watch?v=5D1gV37bKXY>
* The falsification principle, **Null and Alternative Hypothesis**
* Run-through through the **most common tests**, always with the idea of rejecting - accepting the null hypothesis
* **exercise** on how to do a t-test
  + my problem: I used the formula for the general case which is why I had to explain our two samples case again. If you use the slides, better just introduce the correct formula straight away

# Day 5 - Continuing simple tests

* I finished the slides I didn’t get done and clarified on the t-test formula.
* *04\_tests.R* covers **simple tests in R**
* In the end, I tried to run *survey\_analysis.R* (my analysis of the results with the flipchart exercise) with them and **my takeaways for basic R are**:
  + though I oppose plain run-through scripts without having to type because they don’t learn that way, I still had to find a working compromise for me
  + Having them type only the easiest things instead of everything I type is the only way this works
  + *04\_tests.R* was very easy without having to type. That’s why I would write future survey analysis scripts differently - more filled out beforehand, less to type. However, as I’d suggest doing the survey with IDs anyways next time (see above), we will have to come up with a new analysis next year anyways

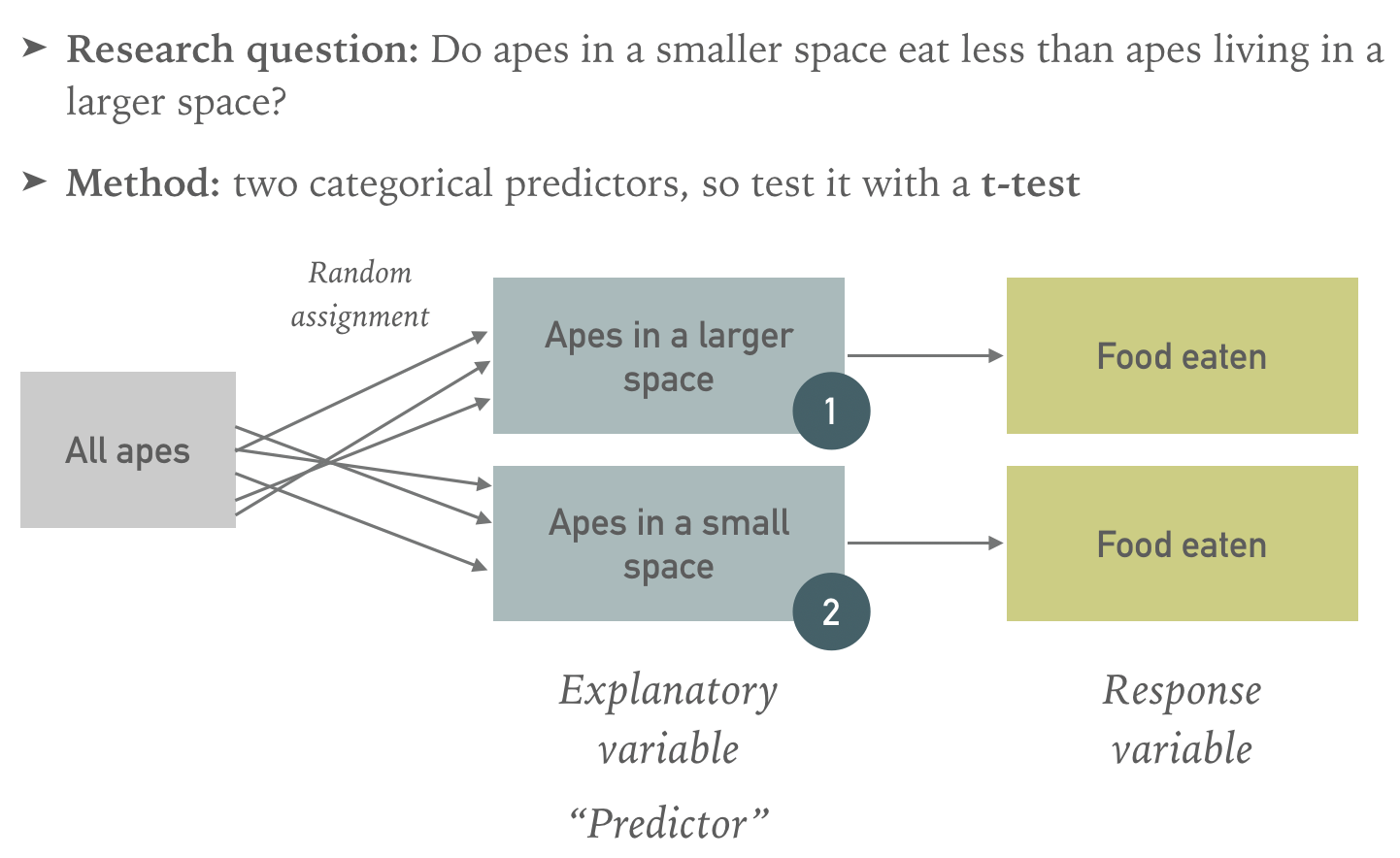
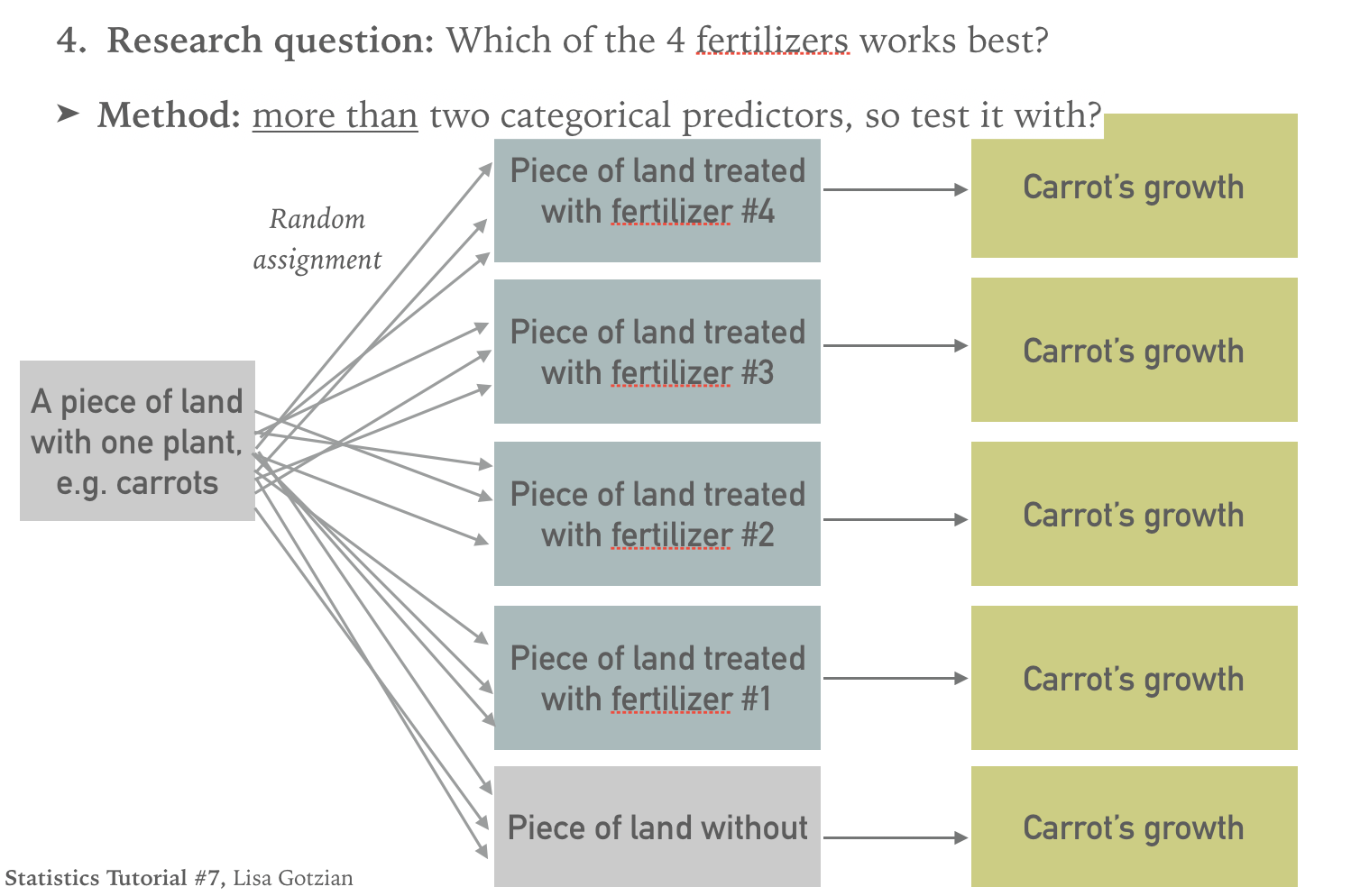
# Day 6 - Correlations

* What’s a p-value? MrNystrom aka APStatsguy <https://www.youtube.com/watch?v=-MKT3yLDkqk>
* explain **correlations, strong & weak, negative & positive, scatterplots**
* play [guessthecorrelation.com/](http://guessthecorrelation.com/)
* introduce **correlation tests**

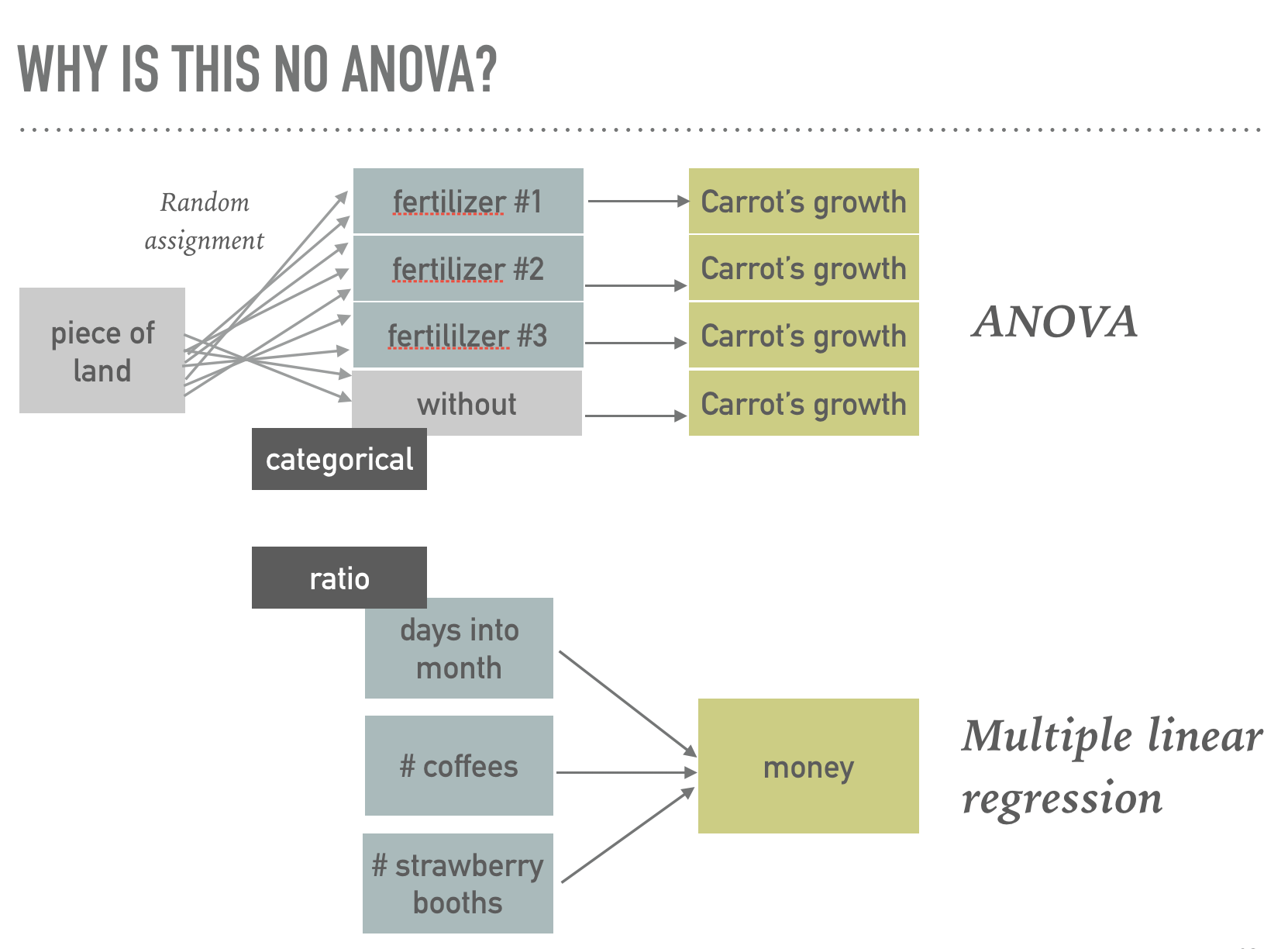
# Day 7 - In the slides: Tutorial #6 Linear Models

* explain what a **model** is: an abbreviation of reality, some kind of mathematical relationship
* introduce **linear models** with the example of the **women dataset** (which is ridiculously highly correlated)
* *06\_linear\_model.R* to show how a linear model works, what residuals are and what histograms look like. Only problem: reading in the blood\_pressure.txt
* <https://www.youtube.com/watch?v=IMjrEeeDB-Y> - MrNystrom on **R Squared**
* **extrapolating & intrapolating**

# Day 8 - In the slides: Tutorial #7 Analysis of Variance

* Introduction to **scientific experiments** and my way of visualizing it, left is a t-test, right an anova. I will use this visualization from now on.
* my idea is to always count the number of predictors/response variables to determine if we need an anova or not. Distinction to linear models will follow.
* **deriving how an Anova is calculated in 3 steps**
  + 1a) The distance between points and the mean squared is the Sum of Squares
  + 1b) We will split the overall Sum of Squares
  + 2) Then, we’ll calculate the ratio of the two…
  + 3) that will give us the p-value.
* Introduction to **different variations of Anovas:**
  + **One-way/Two-/Three-way, Factorial design** with the example of coffee
* **Post-hoc tests**
* The **process** of an Anova
  + 1. Normal distribution of all variables, 2. F-Test, 3. Anova, 4. Post-hoc test
* <https://www.youtube.com/watch?v=09Cx7xuIXig> - Khan Academy on drawing boxplots

# Day 9 - In the slides: Tutorial #8a Scientific Papers

* Introduction to **scientific papers** and the way of communication as a scientist
* **Structure** of scientific papers
* Discussion of **one scientific paper and its method**
  + I didn’t do the Phalan paper as it was too complicated (even for its criticism).
  + To show a working example as some kind of “Musterbeispiel” (“very good example”), I went for the Bullock AND the Klein paper. One I discussed in class, the other one I assigned as group homework (and discussed it in class afterwards). **Check out my slides if you are having trouble with the papers!**
  + **Klein paper**: the three hypothesis piece by piece including its method and results
  + **Multiple linear regression** vs. Anova in a diagram
  + Structure of the paper
  + Assign the same to do with the **Bullock paper** as homework

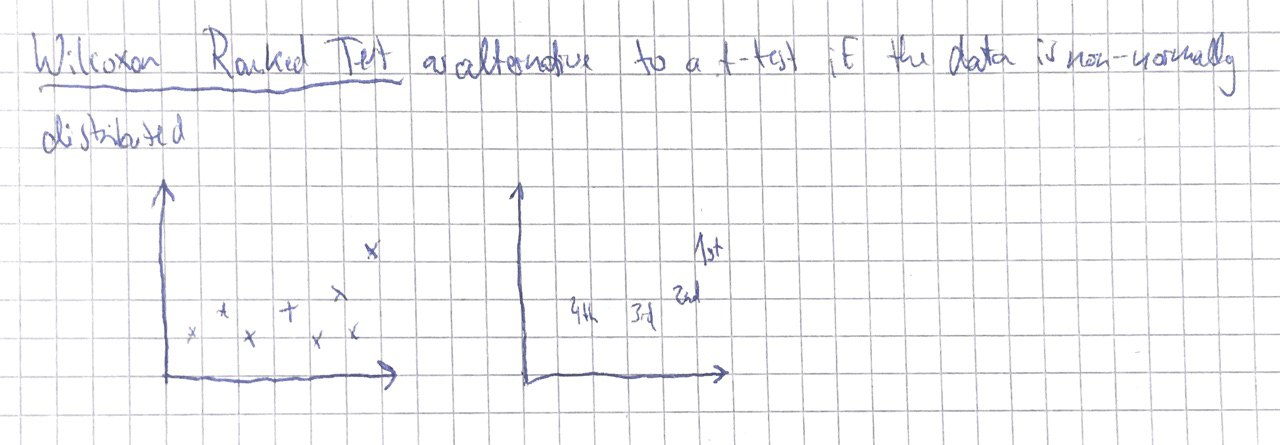
# Day 10 - In the slides: Tutorial #8b Scientific Papers

* presentation of the Klein paper
* discussion of my take on the Klein paper
* **exercise**: design an experiment yourself
* **homework**: from Prabesh’s *Mathematics\_\_\_Statistics\_Problem\_Set\_1.pdf* exercise #5

# Day 11 - Recap of R

* I did *Recap.R*, though the students received a version they had to fill out, *Recap\_class.R*
* On the blackboard, I provided an overview to the script, it was present the entire time and I even re-wrote it on the board for the next session. In brackets is the type of the predictor/response. See diagrams above why I labeled t-tests and Anovas as cat/cont.
* I also explained why the Wilcoxon-test is an alternative to non-normally distributed data on the board:

# Day 12 - Recap

* I finished my *Recap.R*

**Tests**

* correlations (cont/cont)
* t-tests (cat/cont)
* chi-square test (cat/cat)
* linear models (cont/cont)
* Anova (cat/cont)

**Data Analysis Approach**

1) visual inspection

2) independent histograms of the data & shapiro

3) the test

3b) optional depending on the model: residuals of the model, plotting the model etc.

* I then talked about all the things I felt were important but haven’t mentioned so far: **Natural, lab** and **field** experiments for example
* I explained my solutions to the **homework** from day 10
* I asked for **feedback** as I as a student assistant wasn’t evaluated
* questions on the **exam**

# Points for our meeting:

* data types
* survey analysis
* distributions: I feel like it’s too early to discuss distributions in session 3. Normal distributions, ok, but next time I’ll opt for a later additional meeting on distributions where I also explain the “pile analogy” by the one and only philipp: In a t-test, you mainly test if the data comes from one big pile or two big piles.
* I’d suggest we take more time for simple tests (we have time in the end)
* Phalan paper