Dirk Metzler SS 2023

STATISTICS FOR EES — EXERCISE SHEET 2

1. An experiment was carried out in which 19 test persons estimated the size of a blue area (relative to a total area of 100) in five different types of plots. Download the R file http://evol.bio.lmu.de/_statgen/StatEES/17SS/bluearea_first_analysis.R that contains the commands for downloading the data and visualizing certain aspects of the data.

(a) Understand what this R file does and how it does it by

- trying it out, also parts of it,
- using the online help system, and
- changing some commands and options and check how this affects the results.
- adapting and applying it the blue area experiment data of your cohort

(b) Add similar R commands to the file to visualize how the estimation errors depend on the type of the plot **and** on the true value. Try to do this in just one plot, and explore several possibilities for this.

(c) What (preliminary) conclusions would you draw from your new plots?

(d) Discuss the experimental design. Should the experiment be repeated in a different way? Discuss the advantages and disadvantages of several possible approaches.

2. Repeat exercise 1 with various alternative error measures, e.g. squared error, relative error, ...

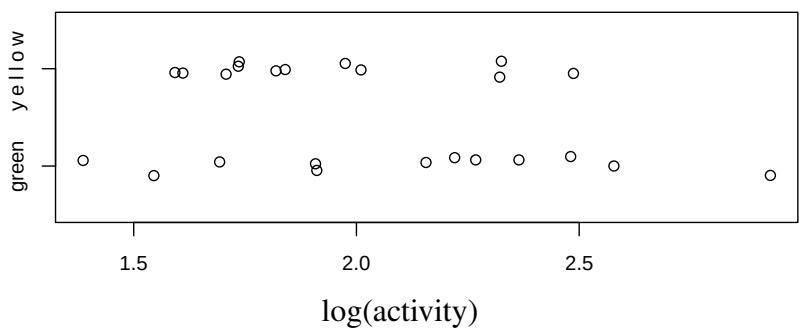
(a) Discuss and explore with the data how sensitive these error measures are for outliers. Discuss how sensitive they *should be* for outliers.

(b) Discuss and explore with the data how the different error measures depend on the true value of the blue area.

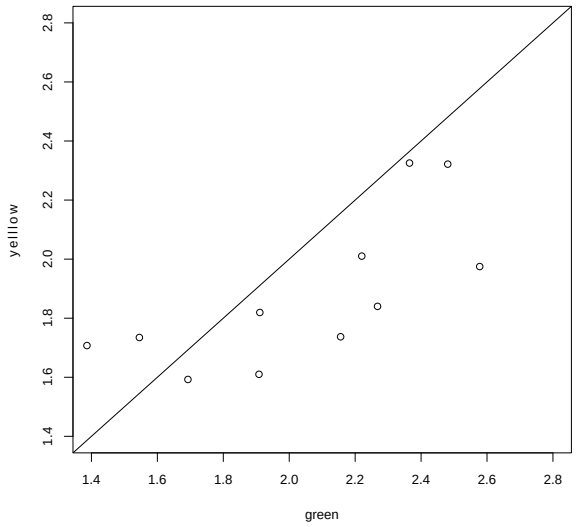
(c) Search by visualization for further possible effects in the data.

(d) Explore this: If the errors of a certain visualization method is below average in this dataset according to a certain way of measuring errors, does this also hold for any other reasonable way of measuring errors?

3. **a)** Can the following values be true or are some of them obviously wrong?
$\overline{x}$(yellow) = 1.93, σ(yellow) = 0.30, $\overline{x}$(green) = 2.12 and σ(green) = 0.45



b) Adam claims that the activity with green light is significantly higher. Do you agree?



c) Eve responds: None of the means is significantly different from 2.0. Is she right?

The data come from 12 birds. Each bird gave a activity value for green and one for yellow. The paired samples are shown in the second graph.

d) How do you describe the activities corresponding to the two colors?

e) Compute the t-statistic to test if the means are equal. The standard deviation of the differences is s=0.27.

f) Summarise the result of the test.

4. Student's classical hypnotics dataset: Two hypnotics were tested with 10 test persons. The sleep time was measured in hours relative to the average in a control group. Thus, a negative value indicates that the test person slept less than the average in the control group. You get the data in R in a data frame `sleep` after typing `data(sleep)`. The column `group` shows which hypnotic was used, and row n refers to the same test person as test person $n + 10$. Perform an appropriate t-test to compare the efficacy of the two hypnotics. Do this first without using the R command t.test. Then apply this command to check your results.

5. The R script sem_cll.R contains commands to perform the following steps:

(a) Simulates a population of 1,000,000 values of some variable x.

(b) Compute the mean μ and the standard deviation σ of all values x.

(c) Draw 1,000 samples of size $n = 10$ from the population of values x.

(d) Compute the sample mean $\overline{x}$ and the standard deviation s for each of the 1,000 samples.

(e) Determine for what fraction of the samples the interval between $\overline{x} - s/\sqrt{n}$ and $\overline{x} + s/\sqrt{n}$ contained the population mean μ.

(f) Visually compare the distribution of sample means to the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

At the end of the R script another population of values y is simulated. Perform the steps listed above also for the population y with various values for the sample size n. For which n is the normal distribution a good approximation for the distribution of sample means?