Übung Ultimate attainment in SLA

Statistics 02

Laura Becker

FAU Erlangen-Nürnberg

SS 2020

The cookbook approach in statistics

There are 2 main approaches to introducing statistics:

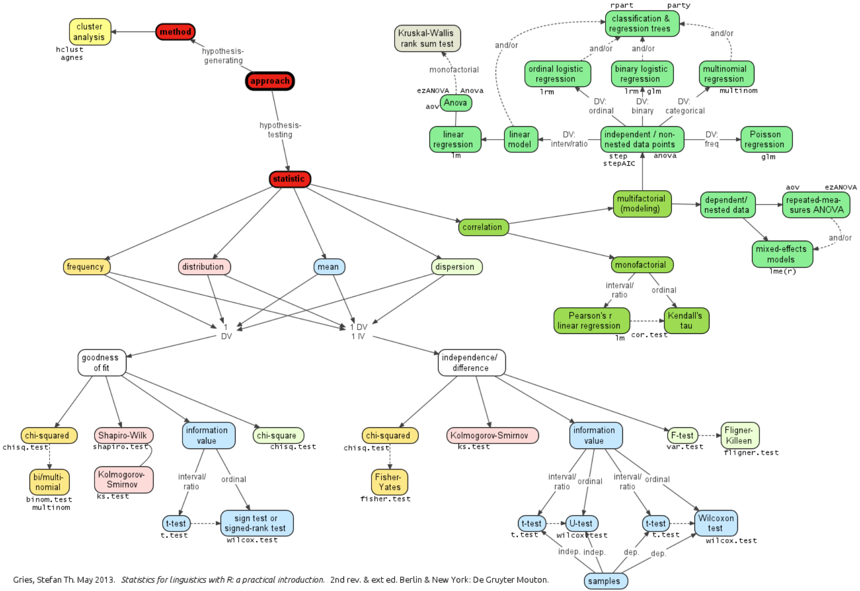
The conceptual approach

- ☺ gives you the big picture and an overview
- ☺ helps to understand what statistics can and cannot do (conceptually)
- ☹ does not tell you anything about **how** to actually apply statistics to your data

The "cookbook" approach

- ☺ tells you which test to apply for which type of data and how to compute the statistic
- ☺ helps you to **apply** statistical testing to your data
- ☹ you may not understand how and when the test works, without understanding what your result actually means and implies

The cookbook approach



Gries, Stefan Th. May 2013. *Statistics for linguistics with R: a practical introduction.* 2nd rev. & ext ed. Berlin & New York: De Gruyter Mouton.

Some tests for numeric measures

These are some traditional tests that can be applied if your **dependent variable** (outcome) is **numeric**. (count data = categorical and ≠ numeric‼)

- **t-test** (*t* because it assumes a t-distribution)
  1 numeric variable from 1 or 2 groups (e.g. reaction times of L1 vs. L2 speakers)
- **correlation**
  2+ numeric variables from 1+ groups (correlation is a standardized measure of co-variation, reaction times of L2 and N years lived in country of L2)
- **ANOVA (analysis of variance)**
  1 numeric variable from n categorical predictor variables (groups, conditions, etc.)
  - independent one-way ANOVA (between-group design)
    comparison of 3+ groups, with independent/unpaired observations
  - independent factorial ANOVA (between-group design)
    2+ categorical predictors (e.g. groups, conditions), interactions can be taking into account
  - repeated-measures and mixed ANOVA (within-group design)
    as above, but with dependent/paired observations (more than 1 measure from the same participant, or the same item presented more than once)

Other (traditional) tests

This is a traditional test that can be applied if your **dependent variable** (outcome) is **categorical**.

- ***χ*-square test** (χ^2 because it assumes a χ^2 distribution)
  1 categorical variable from 1 or 2+ groups (e.g. correct answers of L1 vs. L2 speakers)

More recent or other methods are

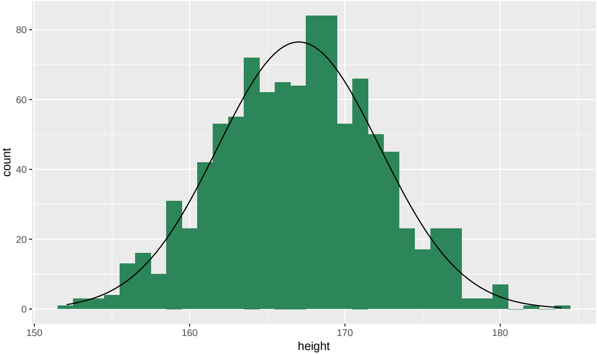
- various types of **regression** (e.g. linear for numeric outcomes, logistic for categorical outcomes)
- if you have input and output data, i.e. independent and dependent variables, you will use **supervised** methods
- **unsupervised** methods:
  e.g. **classification** or **clustering** methods (you have data complex data that you want to arrange into classes of similar observations)
- ...

Distributions & probability

Normal distribution

- Let's imagine we ask 1000 adults about their height.
- We can plot the 1000 measures according how often they occur in our sample.

☞ A **distribution** is a collection of values of a variable (height measures in cm).

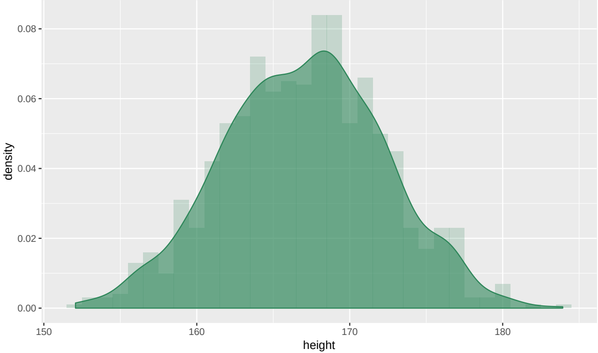


- Large collections of values of a variable (samples) will eventually follow this distribution.

☞ This bell-shaped distribution is called a **Gaussian** or **normal distribution**.

Probability

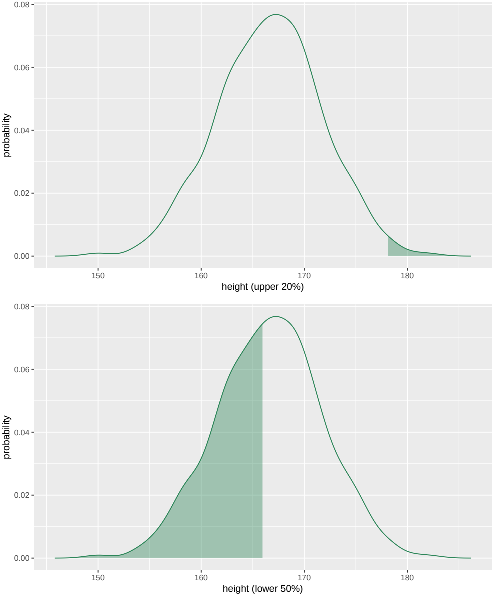
- Instead of plotting the numbers of each height value in our sample (histogram), we can represent the distribution as a **probability density function (PDF)**:
- The x-axis shows all height values in the sample.
- The y-axis now shows the probability of each height value in our sample.



☞ The probability function shows us how likely each value of the sample is to occur.

- We could ask another person about their height, and we can be relatively certain it will be between 165 and 170 cm.

Probability



Statistical models:

The mean as a simple model

The mean as a model

- Let's imagine a toy experiment in which we measure reaction times (ms) for 4 items of L1 and L2 participants.

- We get the following results:

	item1	item2	item3	item4
$L1_1$	1240	1344	2423	1231
$L1_2$	1034	3243	2583	4384
$L2_1$	5439	3923	3845	5654
$L2_2$	2432	5835	3434	4227

The mean as a model

- Let's forget for a moment that we have 2 groups, and just calculate the mean of all reaction times:

$$
\begin{aligned}
m_{RT} &= \frac{RT_1 + RT_2 + RT_3 + \ldots + RT_{16}}{N_{RT}} \\
&= \frac{1240ms + 1344ms + 2423ms + \ldots + 4227ms}{16} \\
&= 3266.938ms
\end{aligned}
$$

- This is a model because it compresses our data to 1 figure, showing a relevant property of the data.
- It is also a model because it allows us to make predictions about future participants: we can assume that it is very likely that they will have RTs around 3267 ms.

? How well does the mean actually represent the variance in our data?

Part 1: How well does the model represent the sample?

- We use statistics to fit a model to our data; a model is a compressed representation of the data and allows for predictions.

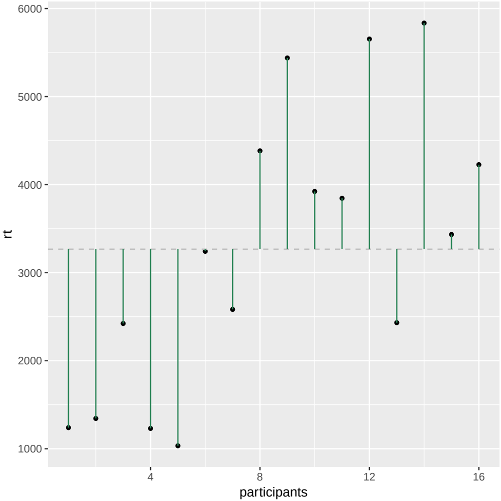
- To work with the model, we need to know how much variance in our data the model can explain, and how much variance is left unexplained.

- This is where the test statistic comes in:

$$test\ statistic = \frac{variance\ explained\ by\ the\ model}{variance\ not\ explained\ by\ the\ model} = \frac{effect\,(part1)}{error\,(part2)}$$

Part 1 continued: Sums of squares

- To assess how well the mean represents our observed RTs, we can:
- calculate the deviance from the mean for each datapoint, and then sum up all deviances.



Part 1 continued: Sums of squares

- If all RTs are centered closely around the mean, deviances will be small and the mean is a good summary.
- If the RTs are spread out more, deviances will be larger and the mean may not be such a good representation of our data.

⚠ Some deviances are positive (RT > mean), some are negative (RT < mean), so they can cancel each other out!

☞ We can square each deviance so that it always has a positive value.

- The sum of the squared deviances from the mean of all datapoints is called **Sums of squares (*SS*)**.

$$
\begin{aligned}
SS &= (RT_1 - m)^2 + (RT_2 - m)^2 + \ldots + (RT_{16} - m)^2 \\
&= 38787811 ms^2
\end{aligned}
$$

☹ The more observations we have, the larger the *SS*.

Part 1 continued: Variance

- We can divide the *SS* value by our sample size (actually, N-1, which are the degrees of freedom) to avoid that our measure of how representative the mean is grows infinitely large.

☞ This is the **variance (s^2)**, the average error between the mean and the observations (RTs).

$$s^2 = \frac{SS}{N-1} = \frac{(RT_1 - m)^2 + (RT_2 - m)^2 + \ldots + (RT_{16} - m)^2}{N-1}$$

$$= \frac{38787811 ms^2}{15} = 2585854 ms^2$$

Part 1 continued: Standard deviation

☹ But how should we interpret a variance of 2585854 ms^2?

☞ We can take the square root of the variance s^2,
this is the **standard deviation *s***:

$$s = \sqrt{\frac{SS}{N-1}} = \sqrt{\frac{(RT_1 - m)^2 + (RT_2 - m)^2 + \ldots + (RT_{16} - m)^2}{N-1}}$$

$$= \sqrt{2585854ms^2} = 1608.059ms$$

- A standard deviation of about 1608ms tells us that most RTs of our sample lie between $mean \pm s$, i.e. 3266 ± 1608 ms.

Part 2: How well does the model represent the population?

- We saw that the mean RT of our sample was 3266ms, with a standard deviation of 1608ms.

? How representative are those values for the entire population?

? In other words: How certain can we be that the mean and standard deviation of the sample also hold for the population?

- What we have calculated so far and what is missing:

$$\textit{test statistic} = \frac{\textit{variance explained by the model}}{\textit{variance not explained by the model}} = \frac{\textit{effect}}{\textit{error}}$$

Part 2 continued: Standard error

- The standard deviation (*s*) showed us how well the mean represented the data in our sample.
- We want to quantify how well the mean of our sample represents the mean of the entire population.
- If we took several, say 5 samples, we could compare their means:
  e.g. 2564ms, 4352ms, 3347ms, 3754ms, 2684ms

☞ **sampling variation**

- Imagine that we repeated the experiment many, many times:
  we would find that the sample means follow a normal distribution

☞ **sampling distribution**

- We cannot exactly calculate the sampling distribution, but we can approximate it (because a large number of samples will end up normally distributed)

  `https://www.youtube.com/watch?v=6YDHBFVIvIs`

- The **standard error** is the approximated deviation between samples of the population.

$$SE = \frac{s}{\sqrt{N}} = \frac{variability\ of\ the\ data}{sample\ size} = \frac{1608.059ms}{16} = 100.5037ms$$

Part 2 continued: Standard error and confidence intervals

- The standard error tells us about how big of a difference to expect of we repeated the experiment.
- If *N* is large (lots of data) or *s* is small (little variation)

☞ the standard error *SE* is small as well, and you can be (more) confident in your test statistic (because you can be confident that the next experiment would show similar results)

- We can use the standard error to calculate **confidence intervals** (CIs).
- Imagine we repeat our experiment an infinite number of times:
- what CIs show:
  If we repeated the experiment an infinite number of times, the real population mean (or other statistic) is included in the CIs of 95% of the experiment

⚠ But it would not be included in 5% of the cases.

⚠ Unless you have superhuman time and money resources, you will never repeat your study an infinite number of times!

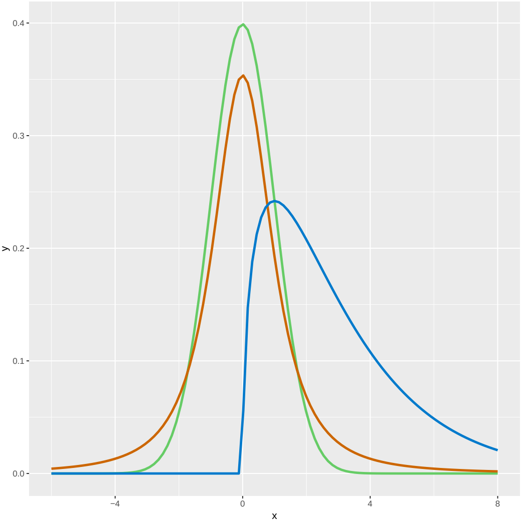
⚠ From **a single** experiment with **a single** CI, we cannot conclude much!

✗ Especially **not** that **a single** CI shows where the population mean (statistic) lies with a probability of 95%!

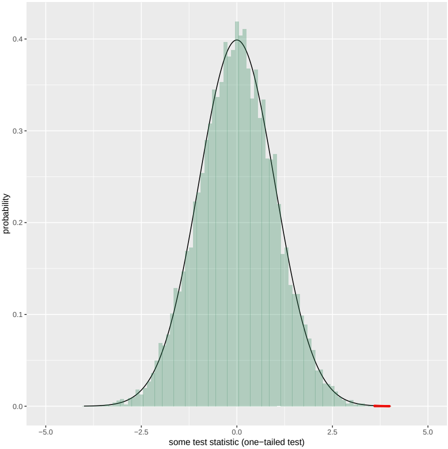
Some more relevant notions and concepts

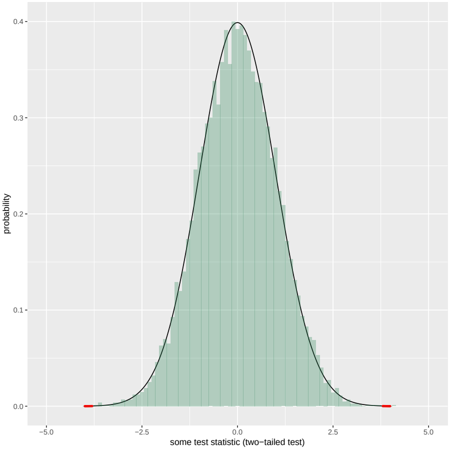
Distributions of common test statistics

normal distribution **t-distribution** χ^2-distribution



One-tailed vs. two-tailed test





Parametric vs. non-parametric tests

For all statistic tests, your data need to meet certain criteria so that you can apply the tests with an interpretable result.

- **Parametric** tests assume your data to be normally distributed (or follow another certain distribution, e.g. t-distribution, F-distribution, χ^2-distribution)

⚠ Even if your data does not meet these criteria, you can still calculate the test statistic, but the result will be meaningless.

- **Non-parametric** tests do not assume any specific distribution of your data.

☞ If you know that your data violates the criteria of a parametric test, you should use a non-parametric test.

Degrees of freedom (df)

- The degrees of freedom depend on the number of observations in your sample: $df = N - 1$ (usually).
- They indicate how many observations are free to vary.

? Why $N - 1$?

- In our sample of 16 observations, these 16 RTs are free to vary in any possible way.
- However, if we use this sample of 16 observations to calculate the e.g. standard error, we have to use the sample mean as an estimate of the population's mean.

☞ We hold one parameter (the mean) constant.

☞ If we are holding the mean constant, only $N - 1 = 15$ values are free to vary.

Let’s do a t-test step-by-step.

Step 1: Data and hypothesis

- Back to our toy experiment with 1 numeric dependent variable (RT) and a categorical independent variable (speaker group)

	item1	item2	item3	item4
L1	1240	1344	2423	1231
L1	1034	3243	2583	4384
L2	5439	3923	3845	5654
L2	2432	5835	3434	4227

- We can calculate the mean RTs for both groups separately to see whether we find a difference:
- $m_1 = 2185.25$ and $m_2 = 4348.625$

? We see a difference, but how certain can we be that the population of L1 speakers is faster than the one of L2 speakers?

☞ We will use the **t-test** to quantify the probability of our data under the assumption of H_0 (no difference of the 2 means)

$H_0 = \mu_1 = \mu_2$ (no difference in their means)

$H_1 = \mu_1 \neq \mu_2$ (there is a difference in their means)

Step 2: The idea of the t-test

1. We calculate a t-value from out 2 group means and sample sizes.

2. For t-values, we assume an underlying t-distribution.

3. We can associate each t-value in the t-distribution with a probability value.

☞ For each t-value, we can estimate how likely it is.

4. We check the probability of the t-value obtained from our sample:

   This is our p-value:

   - If $p > 0.05$, the t-value of our data is not very unlikely (under H_0), so no significant result.
   - if $p < 0.05$, the t-value of our data is very unlikely (under H_0), so we have a significant result.

Step 3: The formula to calculate t

- We calculate the t-value the following way:

$$t = \frac{(m_1 - m_2) - (m_1 - m_2)_{expected}}{SE} = \frac{(m_1 - m_2)}{SE_1 + SE_2}$$

- Variation captured by the model divided by random fluctuations between means of the same population.
  The difference between groups divided by the difference within groups.
- The **smaller** the t-value, the smaller the real difference between groups, making H_0 **more likely**.
- The **larger** the t-value, the larger the real group differences, making H_0 **less likely**.

- m_1 and m_2 are the two group means
- As $(m_1 - m_2)_{expected} = 0$ under the null hypothesis, we can leave this term out.
- SE is the standard error and equal to: $\frac{s}{\sqrt{N}}$, so we need the standard deviation s and the sample size N to calculate SE.

Step 3: The formula to calculate t

- The formula t can be formulated like this:

$$t = \frac{m_1 - m_2}{\frac{s_1+s_2}{\sqrt{N_1+N_2}}}$$

- There are some variations of this formula that we will not go into, depending on the specific assumptions of the two samples (dependent, independent, same, different variance, etc.).
- For an actual analysis, you need to check first if your data meet the assumptions of the test statistic!

- Let's solve the equation for our toy example, with:
  - $m_1 = 2185.25$, $m_2 = 4348.625$
  - $s_1 = 1195.541$, $s_2 = 1198.917$
  - $N_1 = 8$, $N_1 = 8$

Step 4: Calculating the standard deviation s_1 of L1

By hand:

$$\begin{aligned} s_1 &= \sqrt{\frac{SS}{N-1}} \\ &= \sqrt{\frac{(1240-2185.25)^2+(1344-2185.25)^2+\ldots+(4384-2185.25)^2}{8-1}} \\ &= \sqrt{\frac{10005236}{7}} \\ &= 1195.541 \end{aligned}$$

In R:

```
> l1 <- c(1240, 1344, 2423, 1231, 1034, 3243, 2583, 4384)
> sd(l1)
[1] 1195.541
```

Step 5: Calculating the standard deviation s_2 of L2

…

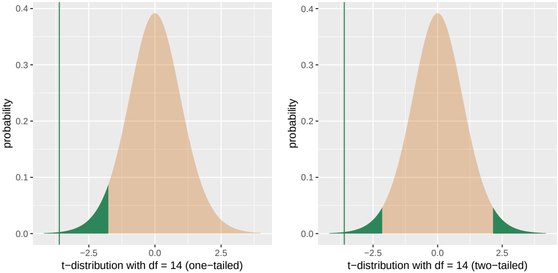
Step 6: Inserting all values and calculating t

By hand:

$$
\begin{aligned}
t &= \frac{m_1 - m_2}{\frac{s_1+s_2}{\sqrt{N_1+N_2}}} \\
&= \frac{2185.25 - 4348.625}{\frac{1195.541+1198.917}{\sqrt{8+8}}} \\
&= -3.613967
\end{aligned}
$$

Step 7: Checking t's probability

- We can now look up the probability p of our t-value of $t = 3.614$ and check if it is $p < 0.05$.
- There are tables that list p-values for various t-values for t-distributions of various degrees of freedom.



- **one-tailed test**: if $t \geq 1.761$ then $p < 0.05$
- **two-tailed test**: if $t \geq 2.145$ then $p < 0.05$

☞ The difference of our toy experiment is statistically significant.

- In an actual experiment, according to the hypthesis, one needs to decide in advance if a one or two-tailed test is used!

Step 7: Performing the t-test in R

```
> l1 <- c(1240, 1344, 2423, 1231, 1034, 3243, 2583, 4384)
> l2 <- c(5439, 3923, 3845, 5654, 2432, 5835, 3434, 4227)
```

One-tailed t-test in R:

```
> t.test(l1, l2)


data:  l1 and l2
t = -3.614, df = 14, p-value = 0.002819
alternative hypothesis: true difference in means is not equal to 0
sample estimates:
mean of x mean of y
 2185.250  4348.625
```

Two-tailed t-test in R:

```
> t.test(l1, l2, alternative = c("less"))
...
data:  l1 and l2
t = -3.614, df = 14, p-value = 0.00141
alternative hypothesis: true difference in means is less than 0
...
```