Chapter 11

Statistical Tests

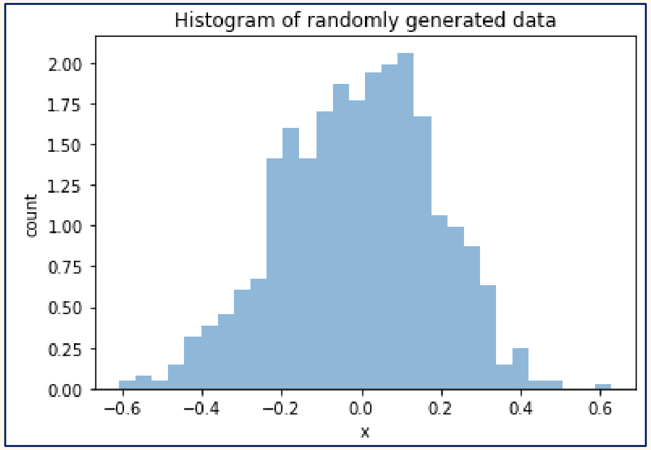
Dr. Steffen Herbold

herbold@cs.uni-goettingen.de

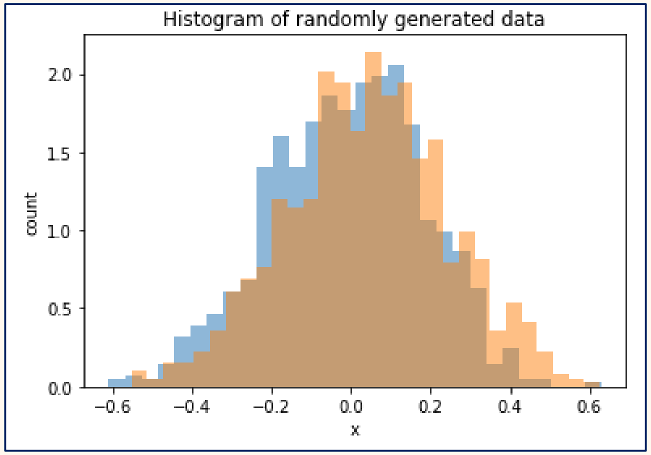
Outline

- Hypothesis Testing
- Effect sizes
- Confidence Intervals
- Summary

Reasons for Hypothesis Testing



Is this data normally distributed?



Do both populations have the same central tendency and/or variance?

Null and Alternative Hypothesis

- Hypothesis testing evaluates assumptions about data
  - Assumption == Hypothesis

- Null Hypothesis (H_0)
  - Assumption of the test holds and is failed to be rejected at some level of significance.
- Alternative Hypothesis (H_1, H_a)
  - Assumption of the test does not hold and is rejected at some level of significance.

- Most important questions:
  - What is the assumption of a test?
  - What does „rejected at some level of significance“ mean?

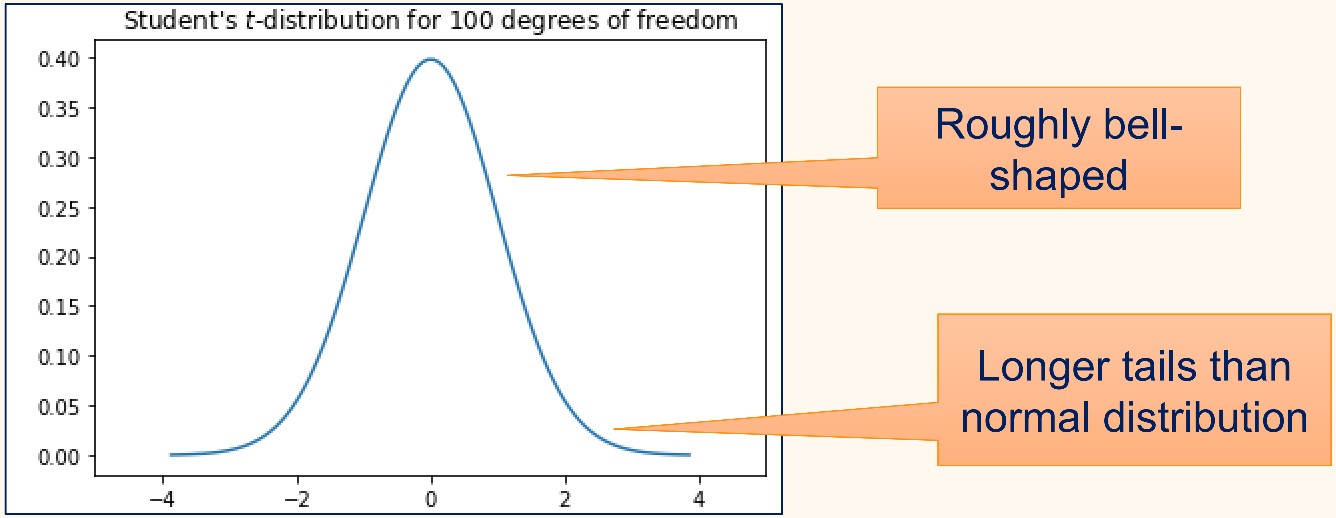
P-Values

- The probability of the observed data or more extreme data, given the null hypothesis is true.
  - Given the hypothesis, how likely is the data?
  - Not the same as given my data, how likely is the hypothesis!
    - Never use p-values as scores!

- Calculated using the probability density function of a test statistic
  - Given the hypothesis, how likely is this statistical value about the data?

Students t-Distribution

- Test statistic used for estimating mean values for normally distributed data
- Probability density function for the location of the deviation of a sample mean value from the real mean value

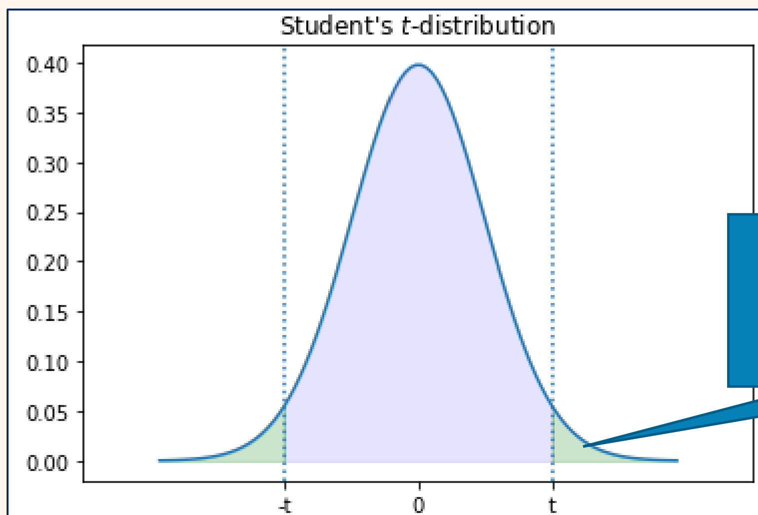


Example: Welch‘s t-Test

Also requires calculation of the degrees of freedom, which we omit here

- Null Hypothesis
  - The means of two normally distributed populations X_1 and X_2 are equal.

- t statistic for Welch’s test:
  - $t = \frac{mean(X_1) - mean(X_2)}{\sqrt{\frac{sd(X_1)^2}{N_1} + \frac{sd(X_2)^2}{N_2}}}$



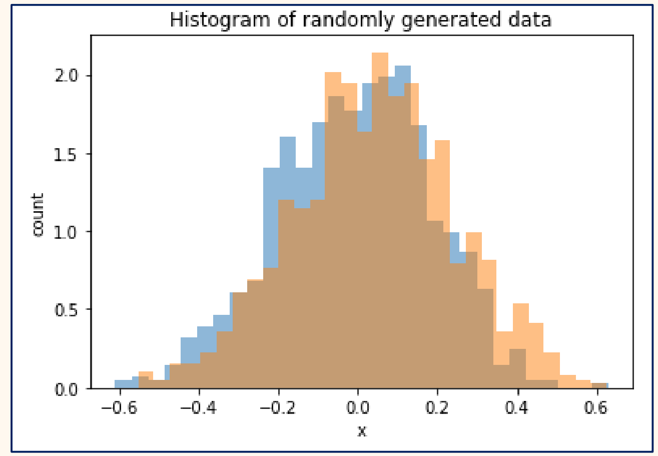
Probability of the observed or more extreme data assuming the null hypothesis is true

Null hypotheses of important tests

- Welch‘s t-test
  - The means of two normally distributed populations are equal
- Shapiro-Wilk test
  - A population of 3 to 5000 independent samples is normally distributed
- Kolmogorov-Smirnoff test
  - Two populations have the same probability distribution
- Mann-Whitney-U Test / Wilcoxon-Ranksum-test
  - The values from one population dominate the values of another population (more or less difference of means/medians)
- Levene‘s test
  - The variances of a group of populations are equal
- ANOVA
  - The mean values of a group of populations are equal

Combinations of Tests

- Example:
  - Are the mean values different?



Shapiro-Wilk test for normality

Both normal

At least one not normal

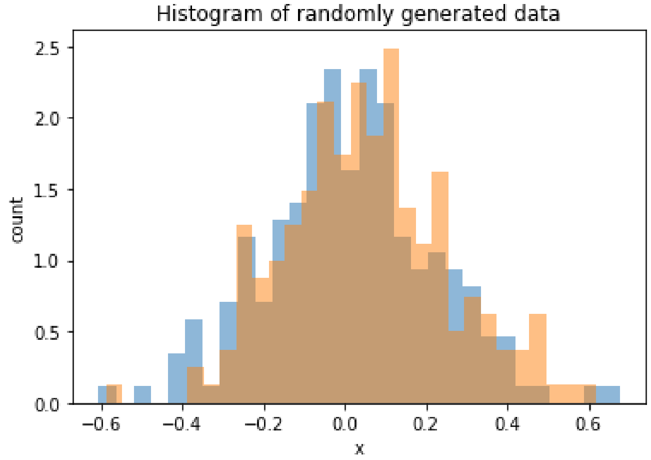
Welch's t-test

Mann-Whitney U-test

Significance and Confidence

- Significance level α
  - Probability of rejecting the null hypothesis, given that it is true
  - Depends on domain
  - Common values:
    - 0.05 (de-facto standard)
    - 0.005 (currently proposed newer standard to reduce false positives)
- Confidence level
  - Probability of not rejecting the null hypothesis, given that it is true
  - $1 - \alpha$
- Used to evaluate tests
  - If $p - value > alpha$ fail to reject null hypothesis
  - If $p - value \leq alpha$ reject the null hypothesis → significant result

Example for running tests




```
p-value of Shapiro-Wilk test for "blue" data:   0.9292
The test found that the data sample was normal, failing to reject the null hypothesis at significance level alpha=0.005

p-value of Shapiro-Wilk test for "orange" data: 0.3986
The test found that the data sample was normal, failing to reject the null hypothesis at significance level alpha=0.005

Both populations normal. Using Welch's t-test.

p-value of Welch's t-tests: 0.048920
The test found that the population means are equal, failing to reject the null hypothesis at significance level alpha=0.005
```

"fail to reject" instead of "accept"

Problems with hypothesis testing

- No binary results!
  - Results are probabilistic, i.e., you can reject/fail to reject the null hypothesis with a certain probability, but you cannot make binary claims.

- Using p-values for scoring
  - p-values describe the likelihood of the data, given the hypothesis
  - Not the same as likelihood of hypothesis, given data!

  → Scoring makes no sense!

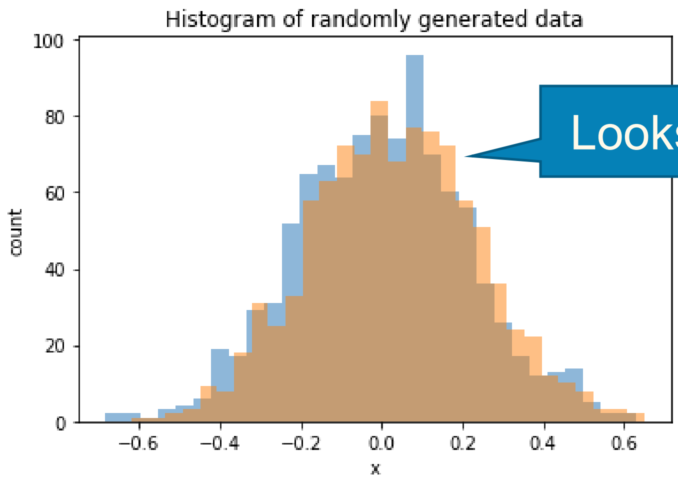
- p-hacking
  - Re-running tests with different data until a desired result is found
  - Often inadvertent, e.g., due to subgroup analysis

Outline

- Hypothesis Testing
- **Effect sizes**
- Confidence Intervals
- Summary

Significant ≠ Important

Can we estimate the size of the difference?



Looks equal

```
p-value of Shapiro-Wilk test for "blue" data:   0.4635
The test found that the data sample was normal, failing to rejec                              gnificance level alpha=0.005

p-value of Shapiro-Wilk test for "orange" data: 0.8631
The test found that the data sample was normal, failing to reject the null hypothesis at significance level alpha=0.005

Both populations normal. Using Welch's t-test.

p-value of Welch's t-tests: 0.003965
The test found that the population means are not equal, rejecting the null hypothesis at significance level alpha=0.005
```

Is probably different

Effect size

There are also other measures like Hedge's g^*, Glass Δ that all use a similar concept

- Measures the distance between central tendency with respect to the variance

- Cohen's d
  - Difference of means relative to the standard deviation
    - $d = \frac{mean(X_1) - mean(X_2)}{s}$
    - $d = 1$ means that the difference of means is "one standard deviation"
  - s is the pooled standard deviation
    - $s = \sqrt{\frac{(N_1-1) \cdot sd(X_1)^2 + (N_2-1) \cdot sd(X_2)^2}{N_1+N_2-2}}$

Square root of the weighted mean of the variances

```
Effect size (Cohen's d): -0.129 - small effect
```

How do we know this is small?

Interpretation of Effect Sizes

- According to Cohen and Sawilowsky

Cohen's d	Effect size
$\lvert d \rvert < 0.01$	Very small
$\lvert d \rvert < 0.2$	Small
$\lvert d \rvert < 0.5$	Medium
$\lvert d \rvert < 0.8$	Large
$\lvert d \rvert < 1.2$	Very Large
$\lvert d \rvert < 2.0$	Huge

- Designed for social sciences
- Should be used with care, but is broadly used in many domains

Outline

- Hypothesis Testing
- Effect sizes
- **Confidence Intervals**
- Summary

How accurate are estimations?

- Example:
  - You have twenty different data sources.
  - You train on five of them and test on the other fifteen.
  - The mean value of the 15 test values is 0.83, the standard deviation is 0.13.

How accurate is the estimated mean value?

Confidence Intervals

We are assuming that everything is normally distributed. Similar formulas are available for non-normal data

- Definition:
  - A $C\%$ confidence interval θ for some parameter p is an interval that is expected with probability $C\%$ to contain p.
  - C is also called the confidence level

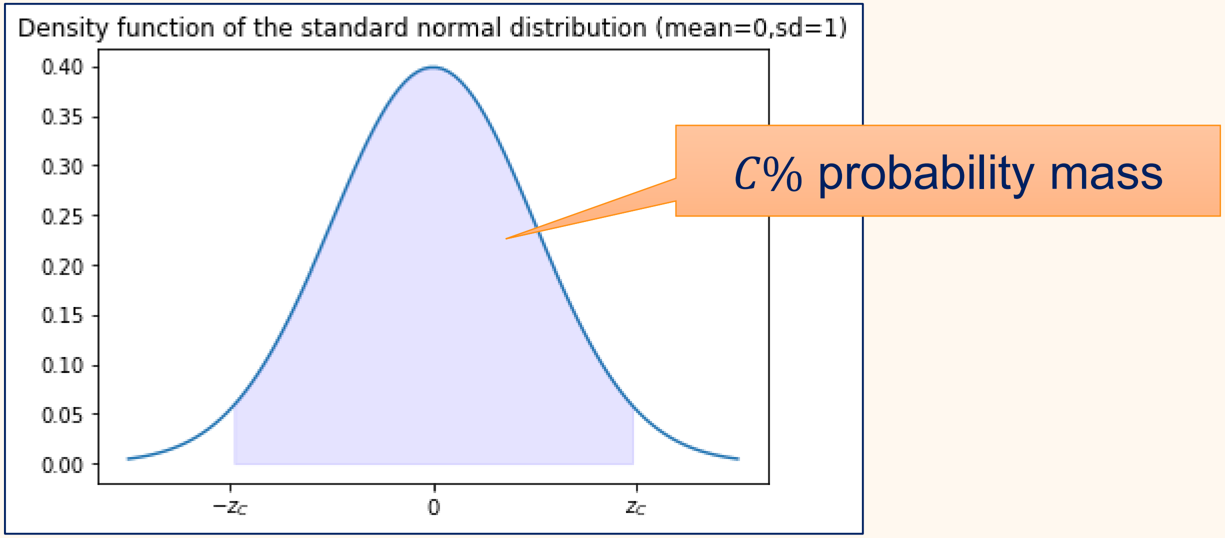
p

- Confidence interval for the mean value of normally distributed data with known standard deviation
  - $\theta = \left[mean(X) - z_C \frac{sd(X)}{\sqrt{n}}, mean(X) + z_C \frac{sd(X)}{\sqrt{n}}\right]$

  → $\pm z_C \frac{sd(X)}{\sqrt{n}}$ uncertainty about the actual mean value

Explanation of z_C

- Choosen such that $[-z_C, z_C]$ is the $C\%$ confidence interval of the standard normal distribution



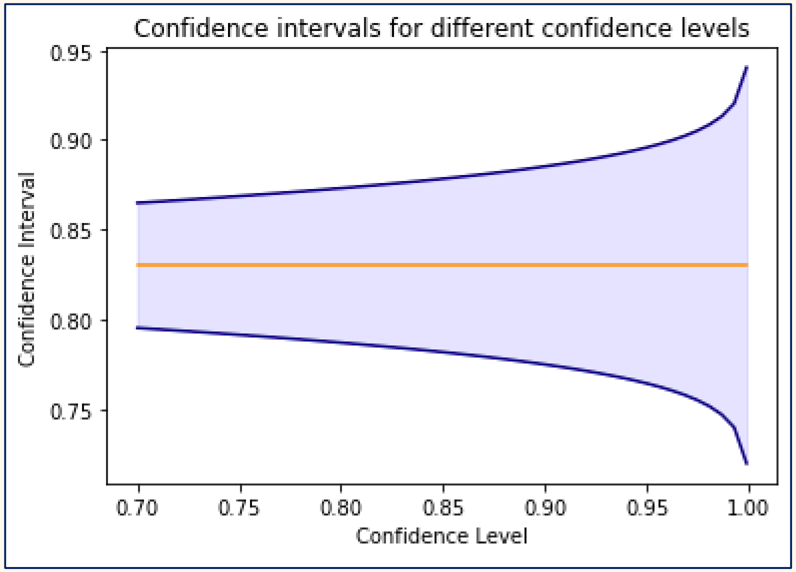
C	z_C
90%	1.645
95%	1.96
99%	2.58
99.5%	2.807
99.9%	3.291

- In practice: look it up in a table

Example for Confidence Intervals

- Example (repeated):
  - The mean value of the 15 test values is 0.83, the standard deviation is 0.13.

C	z_C	θ
90%	1.645	[0.775, 0.885]
95%	1.96	[0.764, 0.896]
99%	2.58	[0.743, 0.916]
99.5%	2.807	[0.736, 0.924]
99.9%	3.291	[0.720, 0.940]



Interpretation of Confidence Intervals

- Correct interpretation of a $C\%$ confidence interval θ
  - $C\%$ chance that results of future replications fall into θ
  - No statistical difference from estimated parameter with $C\%$ confidence
  - There is a $1 - C\%$ probability of the observed data, if the true value for the estimate is outside of θ

- Wrong interpretation
  - The true value lies with $C\%$ probability in θ
  - $C\%$ of the observed data is in θ

Outline

- Hypothesis Testing
- Effect sizes
- Confidence Intervals
- **Summary**

Summary

- Hypothesis testing to evaluate significance of differences
  - Not significant → can be explained by random effects
  - Test results are probabilities, not binary true/false statements!

- Effect sizes to evaluate strengths of significant differences

- Confidence intervals to estimate accuracy of results
  - How stable are the results, if the experiment is repeated?

- All of the above are often misused or interpreted wrongly!