# Recap

- Collocations
  - Conventionally used word combinations
  - Meaning beyond its composition
- Collocation discovery
  - Raw frequency: Not helpful, because of Zipf
  - Pointwise Mutual Information (PMI)
    - Ratio between expected and actual relative frequency

# Inferential Statistics

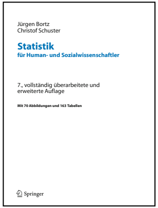
## Sprachverarbeitung (VL + Ü)

Nils Reiter

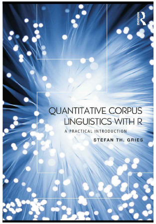
April 27, 2023

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

# Literature




Jürgen Bortz/Christof Schuster (2010). *Statistik für Human- und Sozialwissenschaftler*. 7th ed. Berlin, Heidelberg: Springer




Stefan Gries (2009). *Quantitative Corpus Linguistics with R*. Routledge

Section 1

# Introduction

# Inferential Statistics

- Statements about general populations, *inferred* from a sample
- Hypothesis testing, falsification of the opposite

# Inferential Statistics

- Statements about general populations, *inferred* from a sample
- Hypothesis testing, falsification of the opposite
- Linguistics:
  - Population: Language ›in general‹
    - We may make assumptions about text type, modality, …
  - Sample: Concrete corpus

# Inferential Statistics

- Statements about general populations, *inferred* from a sample
- Hypothesis testing, falsification of the opposite
- Linguistics:
  - Population: Language ›in general‹
    - We may make assumptions about text type, modality, …
  - Sample: Concrete corpus

## Example (»Verwirrter Professor« is a collocation)

- Sample: Wikipedia
- Population: Text that is edited, properly spelled, non-fictional, contemporary, …
  - Satire from the 19th century ⊘
  - Post in a gaming forum ⊘
  - Article in a newspaper ✔

# Section 2

## Hypothesis Testing

# Example

## Gries (2009)

- Players $A$ and $B$ toss a coin $100$ times
  - Heads: $A$ wins / Tails: $B$ wins
- Are they playing fair?

# Example

## Gries (2009)


- Players $A$ and $B$ toss a coin $100$ times
  - Heads: $A$ wins / Tails: $B$ wins
- Are they playing fair?
  - Which results make you suspicious and lets you belief someone cheated?
  - This is the core goal of hypothesis testing

# Example

## Gries (2009)


- Players $A$ and $B$ toss a coin $100$ times
  - Heads: $A$ wins / Tails: $B$ wins
- Are they playing fair?
  - Which results make you suspicious and lets you belief someone cheated?
  - This is the core goal of hypothesis testing
- Hypotheses ($W_X$: Number of wins for player $X$)
  - $H_1$: $A$ more often than 50 times
    expected frequencies of my $A$-wins is higher than that of $B$-wins: $W_A > W_B$

# Example

## Gries (2009)


- Players $A$ and $B$ toss a coin $100$ times
  - Heads: $A$ wins / Tails: $B$ wins
- Are they playing fair?
  - Which results make you suspicious and lets you belief someone cheated?
  - This is the core goal of hypothesis testing
- Hypotheses ($W_X$: Number of wins for player $X$)
  - $H_1$: $A$ more often than 50 times
    expected frequencies of my $A$-wins is higher than that of $B$-wins: $W_A > W_B$
  - $H_0$: $A$ and $B$ win the same number of times
    expected frequencies of wins are $W_A = W_B = 50$
  - »Falsification«: To accept $H_1$, we show that $H_0$ cannot be true

# Example

Gries (2009)


- Players $A$ and $B$ toss a coin $100$ times
  - Heads: $A$ wins / Tails: $B$ wins
- Are they playing fair?
  - Which results make you suspicious and lets you belief someone cheated?
  - This is the core goal of hypothesis testing
- Hypotheses ($W_X$: Number of wins for player $X$)
  - $H_1$: $A$ more often than 50 times
    expected frequencies of my $A$-wins is higher than that of $B$-wins: $W_A > W_B$
  - $H_0$: $A$ and $B$ win the same number of times
    expected frequencies of wins are $W_A = W_B = 50$
  - »Falsification«: To accept $H_1$, we show that $H_0$ cannot be true
- »Bernoulli trial«: A sequence of (independent) binary outcomes
  - In each toss, the probabilities are the same

# Intuition

- Hypotheses

  $H_0$ $W_A = W_B$

  $H_1$ $W_A > W_B$

  ⚠ Not strict opposites – disregard this for the moment

- How often does $A$ need to win, so that we believe they cheat?

# Intuition

- Hypotheses

  $H_0$ $W_A = W_B$

  $H_1$ $W_A > W_B$

  ⚠ Not strict opposites – disregard this for the moment
- How often does $A$ need to win, so that we believe they cheat?

## Recipe

- We play the game and observe $W_A$ (e.g., 15)
- If the probability that $W_A$ takes place gets very small without cheating ($=$ if $H_0$ is true)
  - We assume $H_0$ and see how probable the observed results are under this assumption

# Intuition

- Hypotheses
  - $H_0$ $W_A = W_B$
  - $H_1$ $W_A > W_B$
  - ⚠ Not strict opposites – disregard this for the moment
- How often does $A$ need to win, so that we believe they cheat?

## Recipe

- We play the game and observe $W_A$ (e.g., 15)
- If the probability that $W_A$ takes place gets very small without cheating ($=$ if $H_0$ is true)
  - We assume $H_0$ and see how probable the observed results are under this assumption
  - »very small«: E.g., $0.05$ ($=$ significance level)
  - But: Our decision! Conventions: $0.005$, $0.01$, $0.05$

# How probable are $x$ wins, with three tosses?

A: Kopf B: Zahl

| 1 | 2 | 3 | $W_A$ | $W_B$ | p |
|---|---|---|---|---|---|
| K | K | K | 3 | 0 | 0,125 = $\frac{1}{8}$ |
| K | K | Z | 2 | 1 | 0,125 = $\frac{1}{8}$ |
| K | Z | K | 2 | 1 | 0,125 |
| Z | K | K | 2 | 1 | 0,125 |
| K | Z | Z | 1 | 2 | 0,125 |
| Z | K | Z | 1 | 2 | 0,125 |
| Z | Z | K | 1 | 2 | 0,125 |
| Z | Z | Z | 0 | 3 | 0,125 |

$p(W_A = 3) = \frac{1}{8} = 0{,}125$

$p(W_A = 2) = \frac{3}{8}$

$p(W_A \geq 2) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2} = 0{,}5$

$p(Z) \cdot p(Z) \cdot p(Z) = 0{,}5^3 = 0{,}125$

# How probable are $x$ wins, with three tosses?

| Tosses | | | $W_A$ | $W_B$ | p |
|---|---|---|---|---|---|
| H | H | H | $3$ | $0$ | $0.125$ |
| H | H | T | $2$ | $1$ | $0.125$ |
| H | T | H | $2$ | $1$ | $0.125$ |
| T | H | H | $2$ | $1$ | $0.125$ |
| H | T | T | $1$ | $2$ | $0.125$ |
| T | T | H | $1$ | $2$ | $0.125$ |
| T | H | T | $1$ | $2$ | $0.125$ |
| T | T | T | $0$ | $3$ | $0.125$ |

# How probable are $x$ wins, with three tosses?

| Tosses | | | $W_A$ | $W_B$ | p |
|---|---|---|---|---|---|
| H | H | H | 3 | 0 | 0.125 |
| H | H | T | 2 | 1 | 0.125 |
| H | T | H | 2 | 1 | 0.125 |
| T | H | H | 2 | 1 | 0.125 |
| H | T | T | 1 | 2 | 0.125 |
| T | T | H | 1 | 2 | 0.125 |
| T | H | T | 1 | 2 | 0.125 |
| T | T | T | 0 | 3 | 0.125 |

$$p(W_A = 3) = 0.125$$
$$p(W_A = 2) =$$

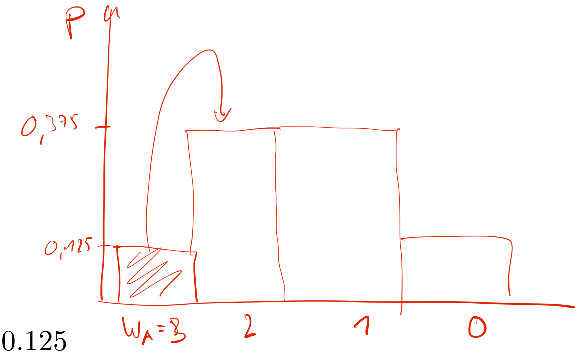
# How probable are $x$ wins, with three tosses?

| Tosses | | | $W_A$ | $W_B$ | p |
|---|---|---|---|---|---|
| H | H | H | 3 | 0 | 0.125 |
| H | H | T | 2 | 1 | 0.125 |
| H | T | H | 2 | 1 | 0.125 |
| T | H | H | 2 | 1 | 0.125 |
| H | T | T | 1 | 2 | 0.125 |
| T | T | H | 1 | 2 | 0.125 |
| T | H | T | 1 | 2 | 0.125 |
| T | T | T | 0 | 3 | 0.125 |

$$
\begin{aligned}
p(\,W_A = 3) &= 0.125 \\
p(\,W_A = 2) &= 0.125 + 0.125 + 0.125 = 0.375 \\
p(\,W_A \geq 2) &=
\end{aligned}
$$

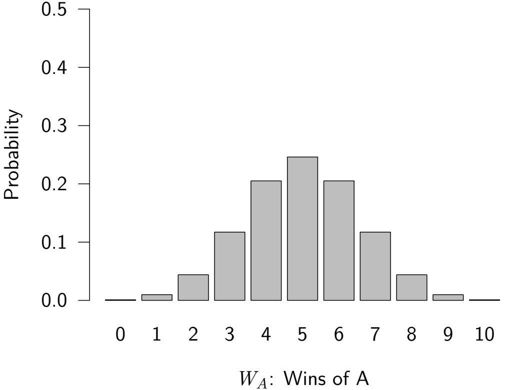
# How probable are $x$ wins, with three tosses?

| Tosses | | | $W_A$ | $W_B$ | p |
|---|---|---|---|---|---|
| H | H | H | 3 | 0 | 0.125 |
| H | H | T | 2 | 1 | 0.125 |
| H | T | H | 2 | 1 | 0.125 |
| T | H | H | 2 | 1 | 0.125 |
| H | T | T | 1 | 2 | 0.125 |
| T | T | H | 1 | 2 | 0.125 |
| T | H | T | 1 | 2 | 0.125 |
| T | T | T | 0 | 3 | 0.125 |




$$
\begin{aligned}
p(W_A = 3) &= 0.125 \\
p(W_A = 2) &= 0.125 + 0.125 + 0.125 = 0.375 \\
p(W_A \geq 2) &= p(W_A = 2) + p(W_A = 3) = 0.5
\end{aligned}
$$

# How probable are $x$ wins, with ten tosses?

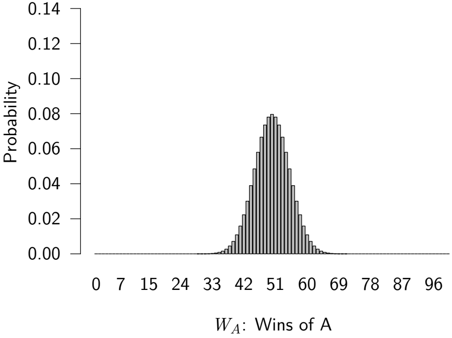



| $W_A$ | $W_B$ | P |
|---|---|---|
| 5 | 5 | 0.246 |
| 6 | 4 | 0.205 |
| 7 | 3 | 0.117 |
| 8 | 2 | 0.043 |
| 9 | 1 | 0.0097 |
| 10 | 0 | 0.000 97 |

Table: Probabilities for $W_A \geq 5$ wins with 10 tosses

Using a fair coin, $p(W_A \geq 9)$ is $0.0097 + 0.000\,97 = 0.01$

# How probable are $x$ wins, with 100 tosses?




| $W_A$ | P |
|---|---|
| 50 | 0.080 |
| 51 | 0.078 |
| 52 | 0.074 |
| 53 | 0.067 |
| 54 | 0.058 |
| 55 | 0.048 |
| 56 | 0.039 |
| 57 | 0.030 |
| 58 | 0.022 |
| 59 | 0.016 |
| 60 | 0.011 |
| 61 | 0.007 |

Table: Probabilities for $50 \leq W_A \leq 61$ with 100 tosses

$$\begin{aligned} p(W_A \geq 55) &= p(W_A = 55) + p(W_A = 56) + \dots \\ &= \sum_{i=55}^{100} p(W_A = i) \end{aligned}$$

# Interpretation

- Above: Situation under $H_0$ (fair coin)

# Interpretation

- Above: Situation under $H_0$ (fair coin)
- Interpretation with respect to a concrete situation
  - E.g., $A$ has won 55 times: $p(W_A = 55) = 0.184$
  - E.g., $A$ has won 60 times: $p(W_A = 60) = 0.028$

# Interpretation

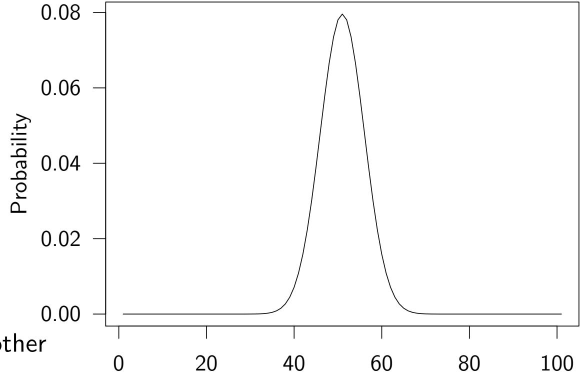
- Above: Situation under $H_0$ (fair coin)
- Interpretation with respect to a concrete situation
  - E.g., $A$ has won 55 times: $p(W_A = 55) = 0.184$
  - E.g., $A$ has won 60 times: $p(W_A = 60) = 0.028$
- If $p < 0.05$: Reject $H_0$, accept $H_1$

# Interpretation

- Above: Situation under $H_0$ (fair coin)
- Interpretation with respect to a concrete situation
  - E.g., $A$ has won 55 times: $p(W_A = 55) = 0.184$
  - E.g., $A$ has won 60 times: $p(W_A = 60) = 0.028$
- If $p < 0.05$: Reject $H_0$, accept $H_1$
- How to calculate this probability (= »p-value«)
  - Statistical tests!

# Binomial Test

- The above: Binomial test
- Binomial distribution $B(n, p)$
  - Not (exactly) the same shape as a normal distribution
  - $p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
- Assumptions
  - One sample
  - Two nominal items
  - Sample items are independent of each other



Section 3

# Application to Collocations

## Why at all?

Observations of linguistic expressions (= corpora) exhibit a randomness similar to random variables like in the game above

## Why at all?

Observations of linguistic expressions (= corpora) exhibit a randomness similar to random variables like in the game above

- Some words/word types appear more often than others
- Choice of words is influenced by a huge number of factors (topic, author, style, creativity, ...)

# Collocation Discovery

- Given two words $w_1, w_2$
- Hypotheses

  $H_0$ $w_1$ and $w_2$ are not collocated (i.e., if they appear together, it's by chance)

  $H_1$ $w_1$ and $w_2$ form a collocation

# Collocation Discovery

- Given two words $w_1, w_2$
- Hypotheses
  - $H_0$ $w_1$ and $w_2$ are not collocated (i.e., if they appear together, it's by chance)
  - $H_1$ $w_1$ and $w_2$ form a collocation
- Our corpus: A sequence of $n$ bigrams
  - Under $H_0$, how many of these bigrams are $w_1 w_2$?
    - Formally: $p(w_1 w_2) = p(w_1) \times p(w_2)$ ?

# Collocation Discovery

- Given two words $w_1, w_2$
- Hypotheses
  - $H_0$ $w_1$ and $w_2$ are not collocated (i.e., if they appear together, it's by chance)
  - $H_1$ $w_1$ and $w_2$ form a collocation
- Our corpus: A sequence of $n$ bigrams
  - Under $H_0$, how many of these bigrams are $w_1 w_2$?
    - Formally: $p(w_1 w_2) = p(w_1) \times p(w_2)$ ?
- Sequence of bigrams: Bernoulli trial
  - Established mathematical framework
  - Sequence of 0/1 decisions with associated probability

# Collocation Discovery

- Given two words $w_1, w_2$
- Hypotheses

  $H_0$ $w_1$ and $w_2$ are not collocated (i.e., if they appear together, it's by chance)

  $H_1$ $w_1$ and $w_2$ form a collocation
- Our corpus: A sequence of $n$ bigrams
  - Under $H_0$, how many of these bigrams are $w_1 w_2$?
    - Formally: $p(w_1 w_2) = p(w_1) \times p(w_2)$ ?
- Sequence of bigrams: Bernoulli trial
  - Established mathematical framework
  - Sequence of 0/1 decisions with associated probability
  - But: Individual ›tosses‹ are not independent ⚠
- We need a different test – it's not a Bernoulli trial!

# $\chi^2$-Test

- Comparison of expected and observed frequencies
- How likely are the observed frequencies if $H_0$ (independence) holds?

# $\chi^2$- Test

- Comparison of expected and observed frequencies
- How likely are the observed frequencies if $H_0$ (independence) holds?
- Steps
  1. Decide significance level 5%
  2. Extract contingency table from corpus
  3. Calculate $\chi^2$-value
  4. Lookup $\chi^2$-value to get to $p$-value

# $\chi^2$-Test

- Comparison of expected and observed frequencies
- How likely are the observed frequencies if $H_0$ (independence) holds?
- Steps
  1. Decide significance level
  2. Extract contingency table from corpus
  3. Calculate $\chi^2$-value
  4. Lookup $\chi^2$-value to get to $p$-value

| | $w_1 =$ »Film« | $w_1 \neq$ »Film« |
|---|---|---|
| $w_2 =$ »Festival« | $24$ | $701$ |
| $w_2 \neq$ »Festival« | $88$ | $1\,880\,208$ |

Table: Contingency table $O$ for »Film Festival«

# $\chi^2$- Test
## Calculate $\chi^2$-value

Simplification for 2x2-matrix

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \qquad \left(\begin{array}{c|c} O_{11} & O_{12} \\ \hline O_{21} & O_{22} \end{array}\right)$$

# $\chi^2$- Test

## Calculate $\chi^2$-value

### Simplification for 2x2-matrix

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \qquad \left(\begin{array}{c|c} O_{11} & O_{12} \\ \hline O_{21} & O_{22} \end{array}\right)$$

| | $w_1$ | $\neg w_1$ |
|---|---|---|
| $w_2$ | 24 | 701 |
| $\neg w_2$ | 88 | 1 880 208 |

# $\chi^2$- Test

## Calculate $\chi^2$-value

Simplification for 2x2-matrix

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \qquad \left(\begin{array}{c|c} O_{11} & O_{12} \\ \hline O_{21} & O_{22} \end{array}\right)$$

| | $w_1$ | $\neg w_1$ |
|---|---|---|
| $w_2$ | 24 | 701 |
| $\neg w_2$ | 88 | 1 880 208 |

$$\chi^2 = \frac{1880232 \times (24 \times 1880208 - 701 \times 88)^2}{(24+701)(24+88)(701+1880208)(88+1880208)}$$

# $\chi^2$- Test

## Calculate $\chi^2$-value

### Simplification for 2x2-matrix

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

$$\left(\begin{array}{c|c} O_{11} & O_{12} \\ \hline O_{21} & O_{22} \end{array}\right)$$

| | $w_1$ | $\neg w_1$ |
|---|---|---|
| $w_2$ | 24 | 701 |
| $\neg w_2$ | 88 | 1 880 208 |

$$\begin{aligned}\chi^2 &= \frac{1880232 \times (24 \times 1880208 - 701 \times 88)^2}{(24+701)(24+88)(701+1880208)(88+1880208)} \\ &= \frac{1880232 \times (45124992 - 61688)^2}{725 \times 112 \times 1880909 \times 1880296} \\ &= \frac{1880232 \times 45063304^2}{2.871\,772\,52 \times 10^{17}}\end{aligned}$$

# $\chi^2$- Test

## Calculate $\chi^2$-value

### Simplification for 2x2-matrix

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \qquad \left(\begin{array}{c|c} O_{11} & O_{12} \\ \hline O_{21} & O_{22} \end{array}\right)$$

| | $w_1$ | $\neg w_1$ |
|---|---|---|
| $w_2$ | 24 | 701 |
| $\neg w_2$ | 88 | 1 880 208 |

$$\begin{aligned}
\chi^2 &= \frac{1880232 \times (24 \times 1880208 - 701 \times 88)^2}{(24+701)(24+88)(701+1880208)(88+1880208)} \\
&= \frac{1880232 \times (45124992 - 61688)^2}{725 \times 112 \times 1880909 \times 1880296} \\
&= \frac{1880232 \times 45063304^2}{2.871\,772\,52 \times 10^{17}} \\
&= \frac{3.818\,189\,69 \times 10^{21}}{2.871\,772\,52 \times 10^{17}} \\
&= 13\,295.59
\end{aligned}$$

# $\chi^2$- Test

## Lookup $\chi^2$-value to get to $p$-value

- In reality: Use a library/program to calculate and get $p$-value
  - Python: `scipy.stats.chi2`
  - R: `chisq.test`

```
m <- matrix( c(24,88,701,1880208), nrow=2)
chisq.test(m)
```

# $\chi^2$- Test

## Lookup $\chi^2$-value to get to $p$-value

- In reality: Use a library/program to calculate and get $p$-value
  - Python: `scipy.stats.chi2`
  - R: `chisq.test`

```r
m <- matrix( c(24,88,701,1880208), nrow=2)
chisq.test(m)
```

  - Java: `org.apache.commons.math3.stat.inference.ChiSquareTest`
  - ...

# $\chi^2$- Test

## Lookup $\chi^2$-value to get to $p$-value

- In reality: Use a library/program to calculate and get $p$-value
  - Python: `scipy.stats.chi2`
  - R: `chisq.test`

```
m <- matrix( c(24,88,701,1880208), nrow=2)
chisq.test(m)
```

  - Java: `org.apache.commons.math3.stat.inference.ChiSquareTest`
  - ...
- Historically
  - Computation of $p$-values is complicated
  - Collections of »critical values« have been published for different levels of significance
    - Critical value for $\alpha = 0.05$: $3.841$

# $\chi^2$- Test

## Lookup $\chi^2$-value to get to $p$-value

- In reality: Use a library/program to calculate and get $p$-value
  - Python: `scipy.stats.chi2`
  - R: `chisq.test`

```r
m <- matrix( c(24,88,701,1880208), nrow=2)
chisq.test(m)
```

  - Java: `org.apache.commons.math3.stat.inference.ChiSquareTest`
  - ...
- Historically
  - Computation of $p$-values is complicated
  - Collections of »critical values« have been published for different levels of significance
    - Critical value for $\alpha = 0.05$: $3.841$
- Since $\chi^2 > 3.841$: reject $H_0$
  (tables often do not give you exact $p$-values)

*demo*

# $\chi^2$ vs. Mutual Information

- Both can be applied to collocation discovery
- Tools for different questions
  - $\chi^2$: Are these two a collocation?
  - PMI: How much information does one word give to the other?

# Interpretation and Pitfalls

- Statistical significance $\neq$ practical significance
- Statistical significance $\neq$ theoretical significance

# Interpretation and Pitfalls

- Statistical significance $\neq$ practical significance
- Statistical significance $\neq$ theoretical significance
- Significance: It's unlikely that the outcomes were achieved under $H_0$
- Important questions:
  - Are $H_0$ and $H_1$ really opposites?
  - Is $H_1$ really what I want to show?
  - What's the ›population‹?
  - Is the sample representative of it?

# $p$-Hacking

- Practice of ›pushing‹ the p-value below 5%
- Consequence of publication preferences by journals and conferences

# $p$-Hacking

- Practice of ›pushing‹ the p-value below 5%
- Consequence of publication preferences by journals and conferences
- Reminder: We allow for 5% error probability!
- If we do 100 significance tests, 5 of them will have false results

# Summary

Inferential statistics

- Hypothesis testing
  - We have made some observations
  - How probably are the observations we have seen under different assumptions?
  - If the result is very unlikely under one assumption, the other must be true
- Not an idiot-proof tool though – think when interpreting results

# Summary

Inferential statistics

- Hypothesis testing
    - We have made some observations
    - How probably are the observations we have seen under different assumptions?
    - If the result is very unlikely under one assumption, the other must be true
- Not an idiot-proof tool though – think when interpreting results

Next: Language Modelling

- Predict the next word, given some history
- It's what your phone does every day!

# References I


Bortz, Jürgen/Christof Schuster (2010). *Statistik für Human- und Sozialwissenschaftler*. 7th ed. Berlin, Heidelberg: Springer.

Gries, Stefan (2009). *Quantitative Corpus Linguistics with R*. Routledge.