

# **Basiswissen Statistik & Schnelleinstieg in R**

# Schnelleinstieg in R

# Datenstrukturen



- Vektoren



- Matrizen

Einnahmen	Ausgaben
375	897
480	390
7209	10978

- Dataframes

Person	Burger/Tag
Stefan	0.03
Moni	0.04
Michael	0.8

# Datenstrukturen



- Vektoren

```
myLuckyNumbers <- c(1, 7, 1, 5, 3, 9)
```

my  
Vector



- Matrizen

```
b <-  
matrix(data=myVector,  
nrow=2, ncol=3)
```

Einnahmen	Ausgaben
375	897
480	390
7209	10978

- Dataframes

```
as.data.frame(b)
```

Person	Burger/Tag
Stefan	0.03
Moni	0.04
Michael	0.8

# Datenstrukturen



- Vektoren



- Matrizen

numeric

Einnahmen	Ausgaben
375	897
480	390
7209	10978

- Dataframes

character /  
factor

Person	Burger/Tag
Stefan	0.03
Moni	0.04
Michael	0.8

# Numeric vs. character

---

```
c(1, 2, 3, 4)
```

```
> [1] 1 2 3 4
```

```
c(94, 95, "Gesundheit", 97)
```

```
> [1] "94" "95" "Gesundheit"  
"97"
```

# Numeric vs. character vs. factor

```
a <- c(1, 2, 3, 4)
```



```
as.factor(a)
```

```
[1] 1 2 3 4
```

```
Levels: 1 2 3 4
```

```
c(94, 95, "Gesundheit")
```

```
> [1] "94" "95" "Gesundheit"  
"97"
```



# Datenstrukturen

Liste



- Vektoren



- Matrizen

numeric

Einnahmen	Ausgaben
375	897
480	390
7209	10978

- Dataframes

character /  
factor

Person	Kekse/Tag
A	0.03
B	0.04
C	0.8



# Listen

---

```
> myList <- list(a,b,c)
> myList [[1]] [1] 1 2 3 4
[[2]]
Col_1 Col_2
Row_1 1 3
Row_2 2 4
[[3]] [1] "94" "95" "Gesundheit" "97"

unlist(myList)
```

# Benutzung zur Textverarbeitung

---

Relevante Funktionen u.a.

- grep (zum Suchen – auch mit regex)
- gsub (zum Ersetzen)
- strplit (zum Aufsplitten)
- paste (zum Zusammensetzen)

# Plan für heute

---

- Denken wie ein/e Statistiker/in
- Skalenniveaus
- Einfache statistische Testverfahren

# Denken wie ein/e Statistiker/in

DID THE SUN JUST EXPLODE?  
(IT'S NIGHT, SO WE'RE NOT SURE.)

THIS NEUTRINO DETECTOR MEASURES  
WHETHER THE SUN HAS GONE NOVA.

THEN, IT ROLLS TWO DICE. IF THEY  
BOTH COME UP SIX, IT LIES TO US.  
OTHERWISE, IT TELLS THE TRUTH.

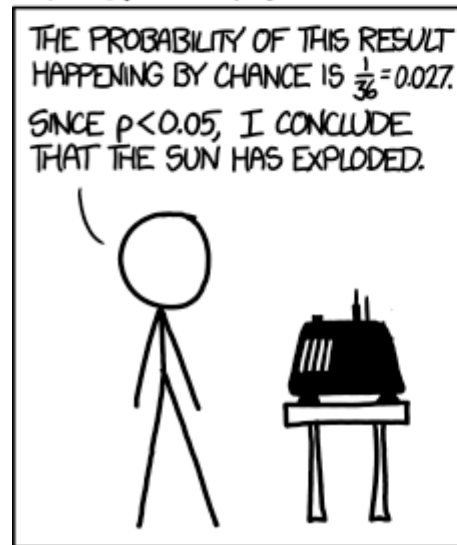
LET'S TRY.  
DETECTOR! HAS THE  
SUN GONE NOVA?

(ROLL)  
YES.



FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT  
HAPPENING BY CHANCE IS  $\frac{1}{36} = 0.027$ .  
SINCE  $p < 0.05$ , I CONCLUDE  
THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50  
IT HASN'T.



# Denken wie ein/e Statistiker/in

---

Falsifikationistischer Ansatz:

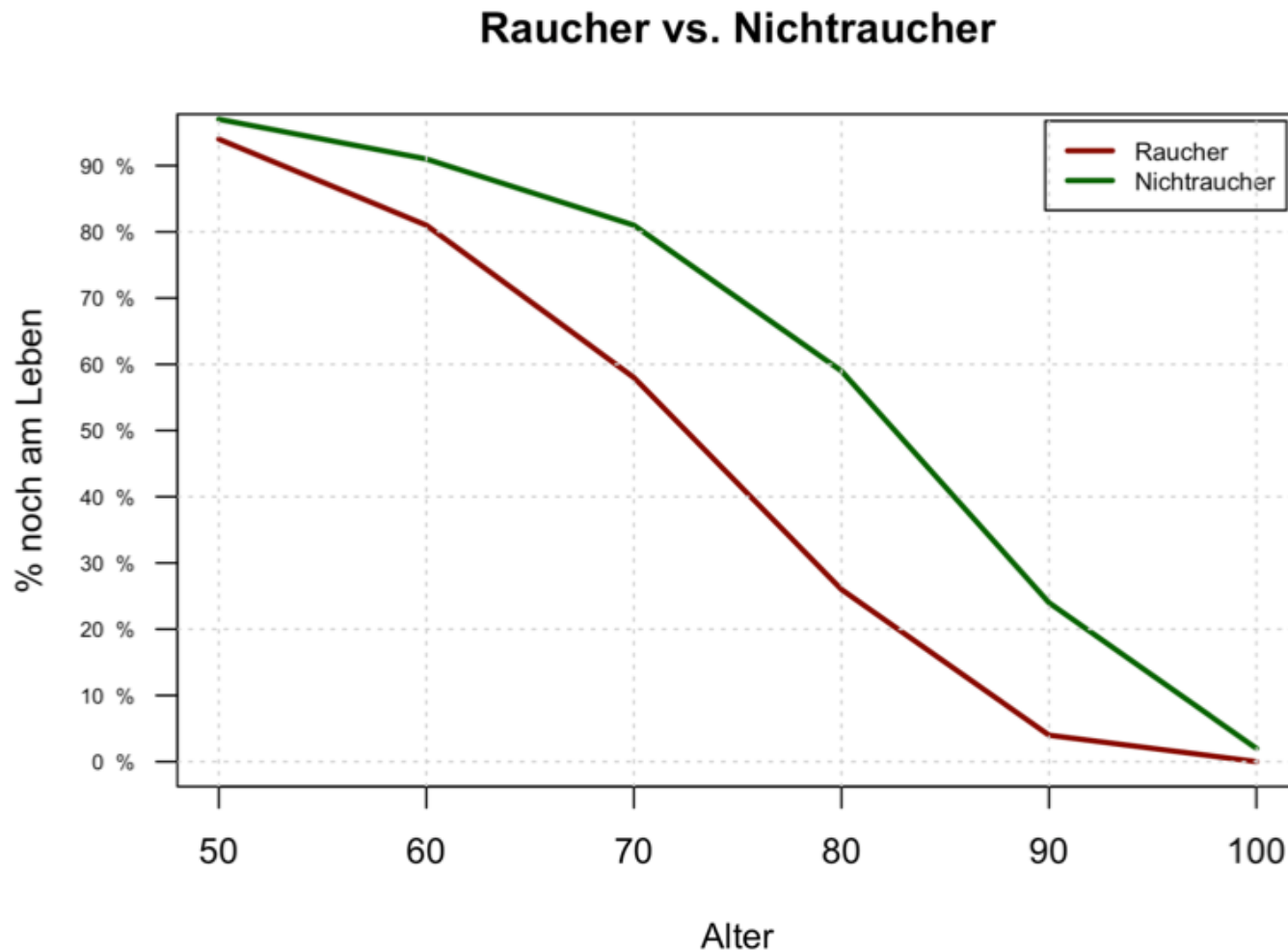
- Wir beginnen mit einer Fragestellung, z.B.: Sterben Raucher früher?
- Wir stellen eine Hypothese auf, z.B.: Raucher sterben früher...
- ...und überprüfen die **Nullhypothese**: Raucher sterben nicht früher.

# Denken wie ein/e Statistiker/in

---

- Wir tun dies, indem wir **Daten erheben...**
- ...diese Daten **auswerten...**
- und fragen, wie wahrscheinlich es ist, dass die beobachtete Verteilung **durch Zufall** zustandekommt.

# Visuelle Inspektion der Daten



# Skalenniveaus

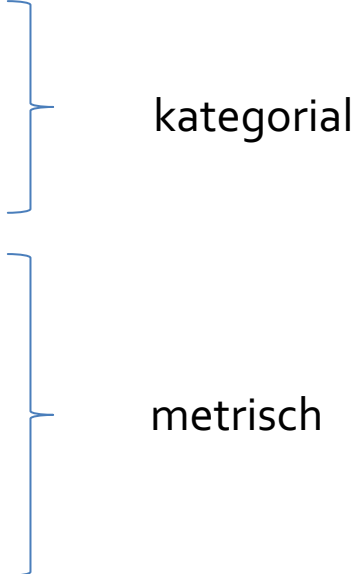
- Variablen sind unterschiedlich **skaliert**
- Beispiel: Familienstand vs. Schulnoten vs. Körpergröße
- Wodurch unterscheiden sich diese drei Variablen?





# Skalenniveaus

---

- Nominalskala
  - Ordinalskala
  - Intervallskala
  - Verhältnisskala
  - Absolutskala
- 
- kategorial
- metrisch

# Nominalskala

---

- Klassifizierung ohne Ordnungsrelation
- z.B. Familienstand: ledig, verheiratet, verwitwet, geschieden
- keine "Hierarchie": ledig ist nicht "besser" oder "größer" als verheiratet, geschweige denn "halb so gut" oder "doppelt so gut" o.ä.
- Auch **binäre** Variablen sind nominalskaliert, z.B. "lebendig" vs. "tot"

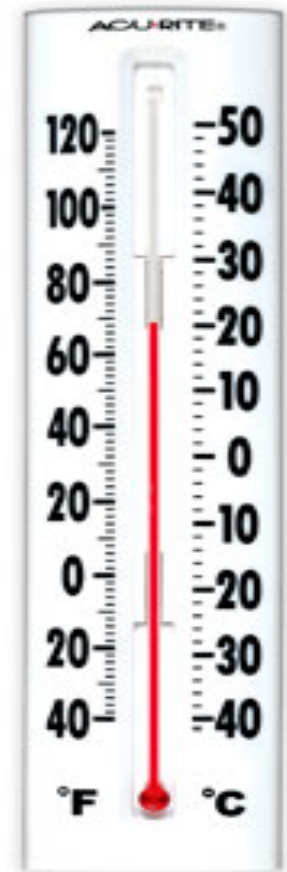
# Ordinalskala

- Ordnungsrelation: z.B. Gold  $>$  Silber  $>$  Bronze
- keine Angaben, **um wie viel** Gold "besser" ist als Silber, Silber besser als Bronze etc.



# Intervallskala

- metrische Skala: Intervalle zwischen einzelnen Punkten sind gleich groß
- z.B. Temperatur: Abstand zwischen  $10^{\circ}\text{C}$  und  $20^{\circ}\text{C}$  so groß wie zwischen  $20^{\circ}\text{C}$  und  $30^{\circ}\text{C}$
- Jedoch:  $40^{\circ}\text{C}$  ist **nicht** doppelt so warm wie  $20^{\circ}\text{C}$  und  $20^{\circ}\text{C}$  **nicht** viermal so warm wie  $5^{\circ}\text{C}$ !
- Grund: Nullpunkt willkürlich festgelegt



# Verhältnisskala

---

- alle intervallskalierten Variablen mit **natürlichem Nullpunkt**
- z.B. Lebensalter, Temperatur in Kelvin (beginnt mit  $-273,15^{\circ}\text{C}$ , dem absoluten Nullpunkt)

# Absolutskala

---

- genauer Wert einer Merkmalsausprägung
- nur natürliche Zahlen möglich
- lässt sich paraphrasieren mit " $n$  Stück", z.B. 5 Stück, 10 Stück, 1000 Stück etc.
- z.B. Zahl der Schüler/innen in einer Klasse, Anzahl der Todesfälle im Jahr

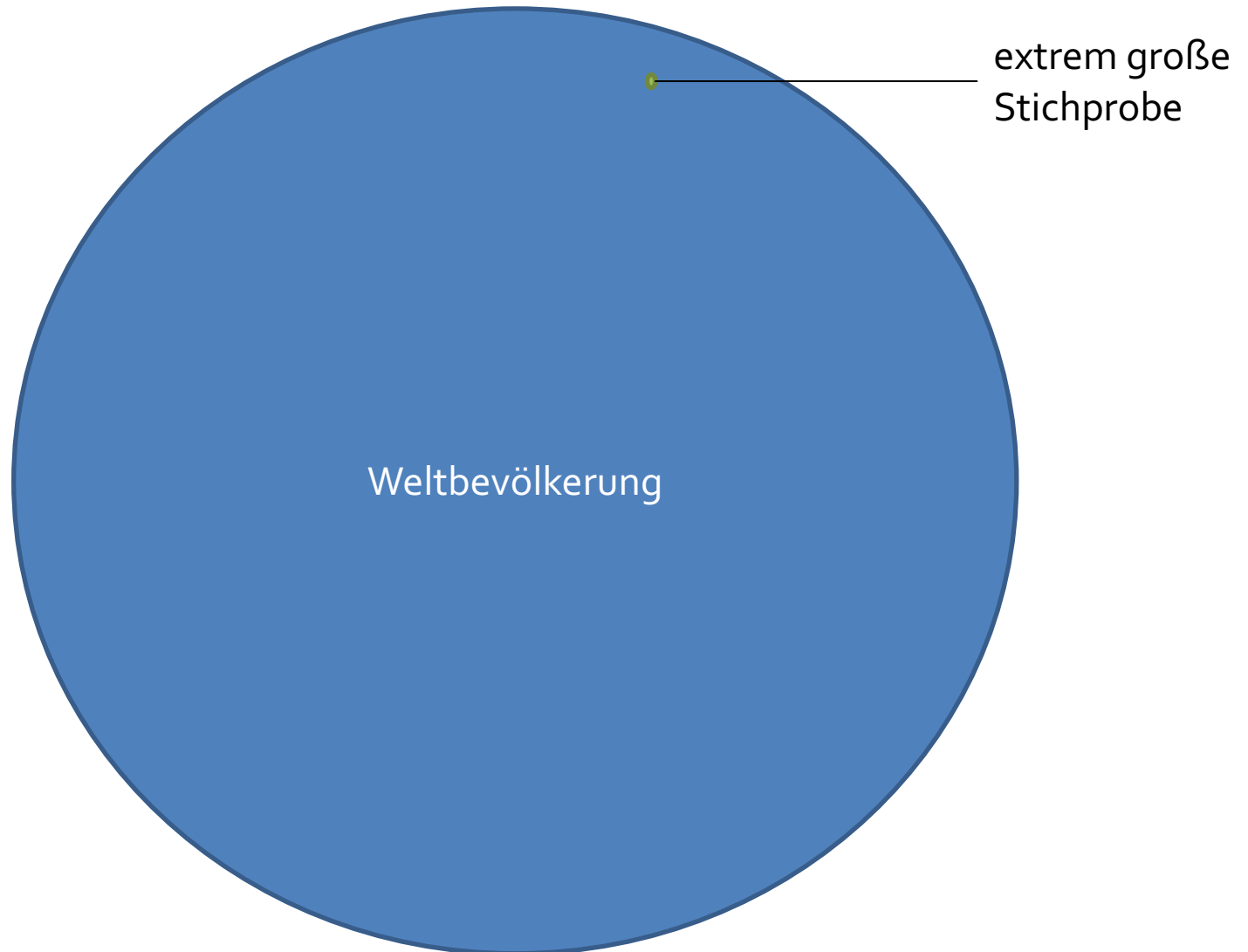
# Übungsaufgabe

---

- Affixart (Präfix vs. Suffix)
- Schulnote
- Alter von ProbandInnen in der Form "unter 18", "18-49", "49 und älter"
- Dauer einer Veranstaltung
- Körpergröße
- Likert-Skala (z.B.: Bewerten Sie x auf einer Skala von 1-5)

# Grundgesamtheit und Stichprobe

---





# Sind meine Daten repräsentativ?

---

- Nur in den seltensten Fällen können wir die gesamte **Population** untersuchen
- Deshalb ziehen wir eine **Stichprobe**
- Faustregel: Je größer die Stichprobe, desto besser
- Beispiel: Münzwurf

# Ist meine Münze gezinkt?

---

- Wahrscheinlichkeit für Kopf und Zahl ist für gewöhnlich 50:50
- Natürlich ist es dennoch möglich, bei 10 Würfeln 10mal Kopf zu bekommen...
- ...aber nicht sehr wahrscheinlich!

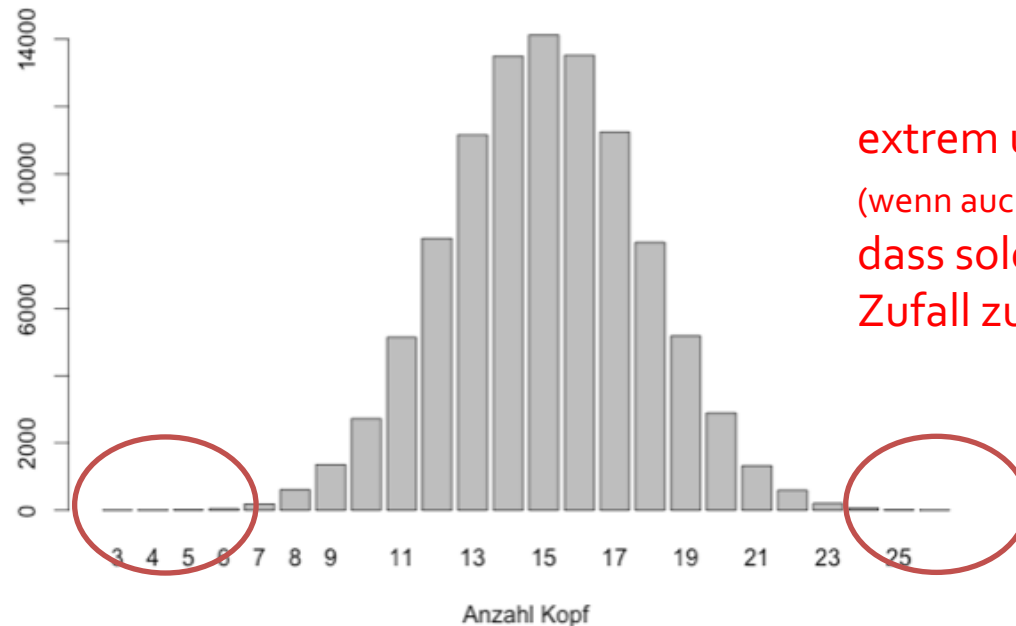
# Gerichtete und ungerichtete Hypothesen

## Gerichtet:

- Die Münze ist so gezinkt, dass öfter Kopf erscheint.
- Die Münze ist so gezinkt, dass öfter Zahl erscheint.

## Ungerichtet:

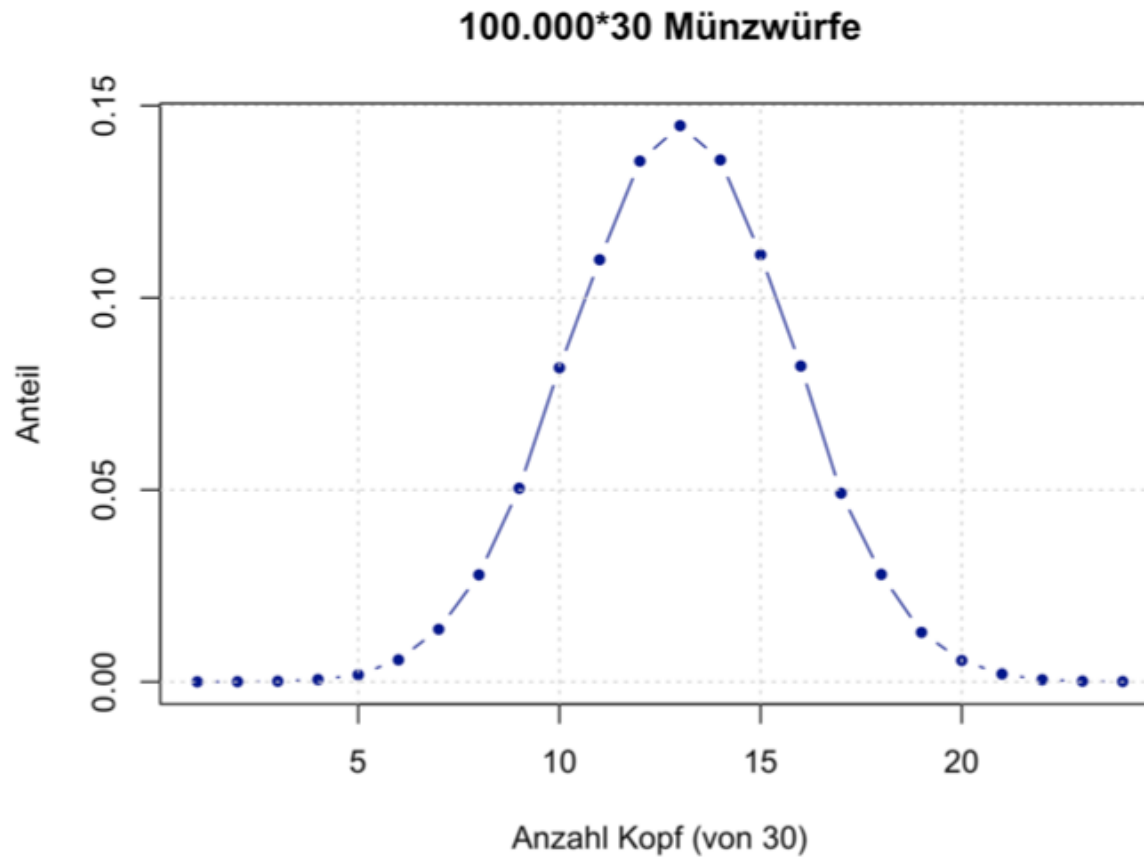
- Die Münze ist irgendwie gezinkt.



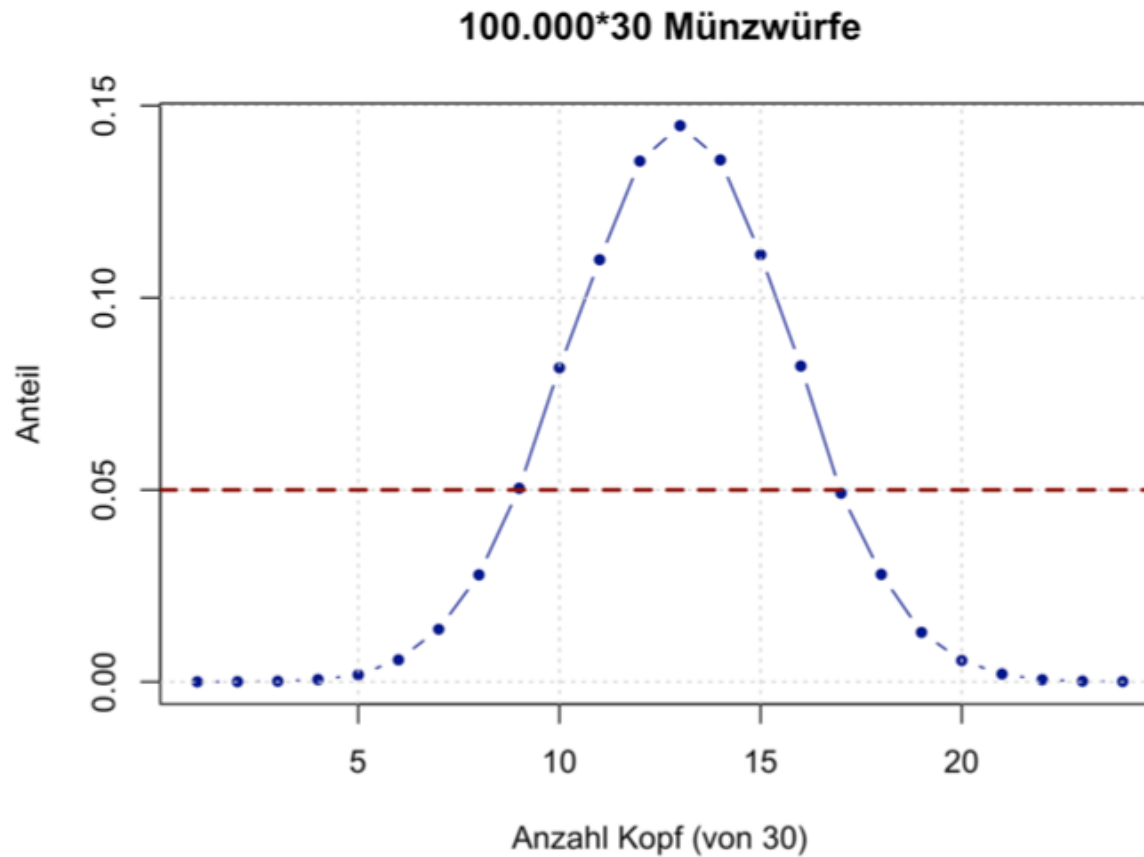
extrem unwahrscheinlich  
(wenn auch prinzipiell möglich),  
dass solche Werte durch  
Zufall zustandekommen.

Grafik: 100.000\*30  
Würfe einer Münze  
(Simulation)

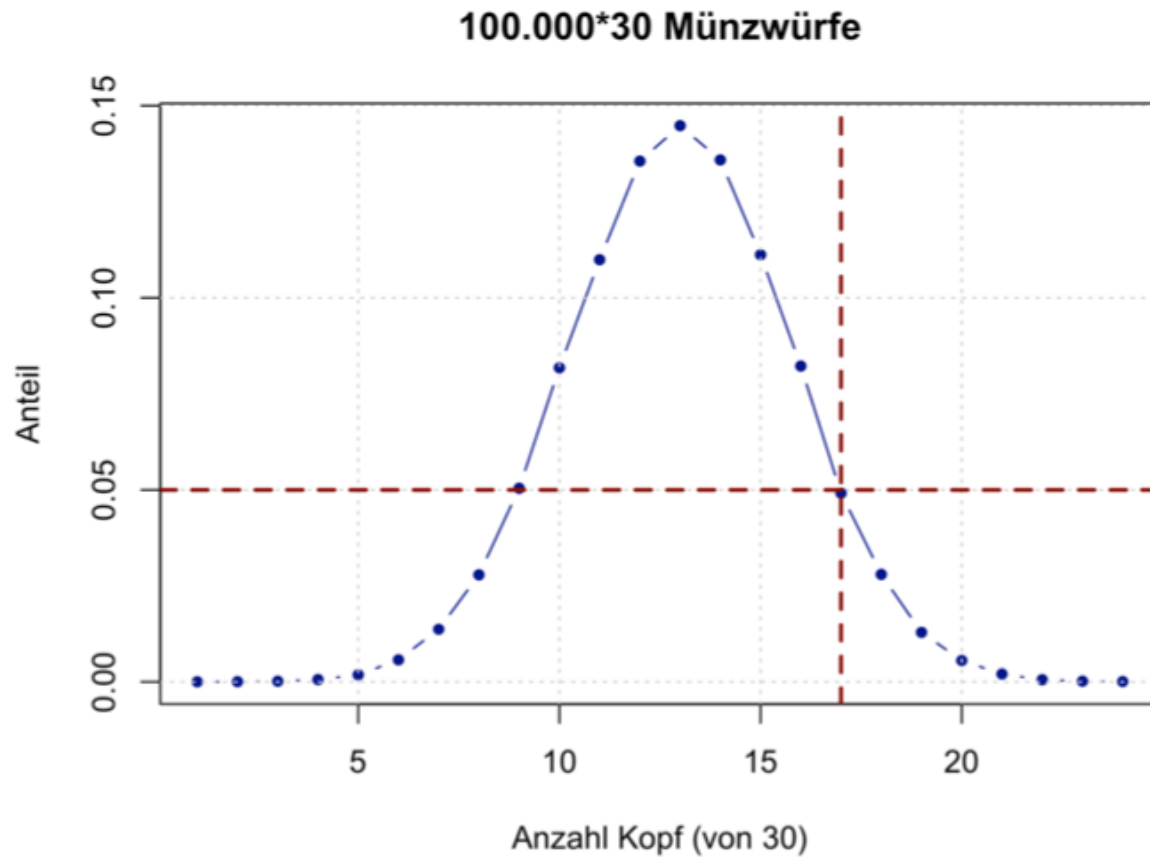
# Signifikanzschwelle



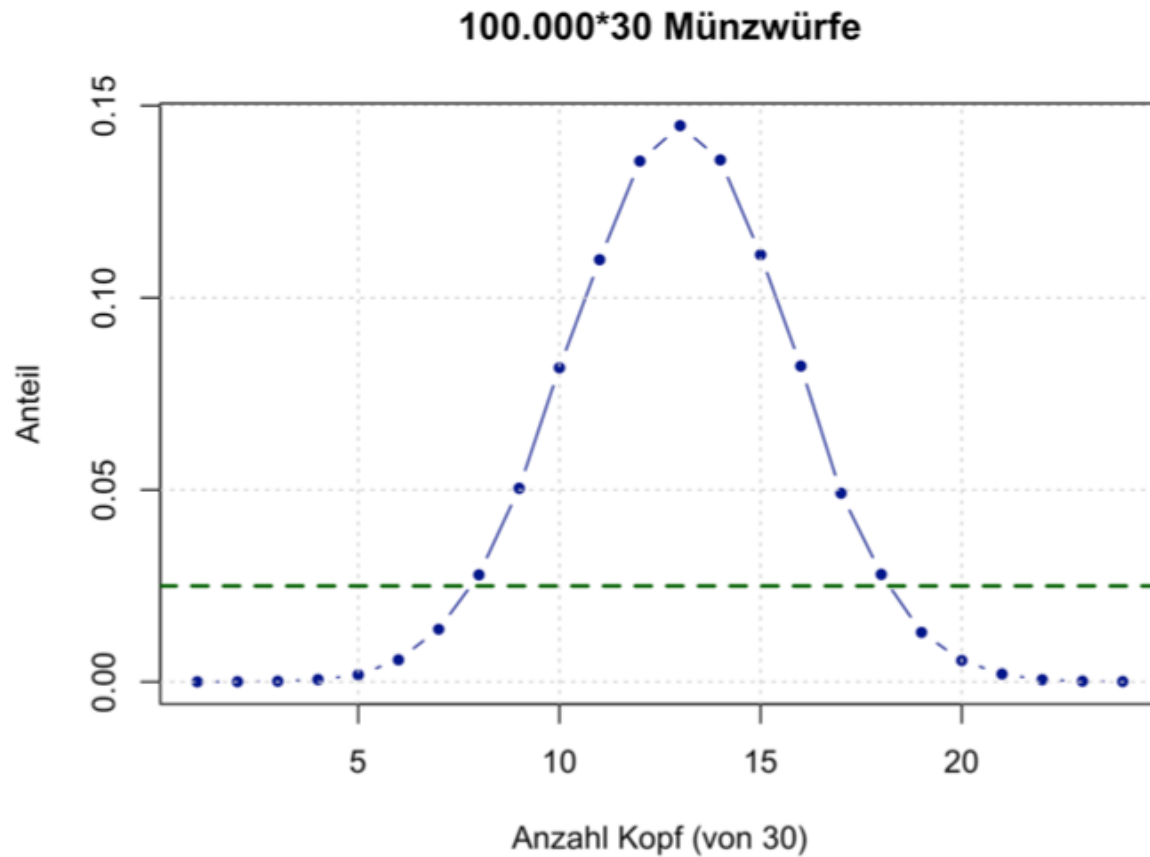
# Signifikanzschwelle



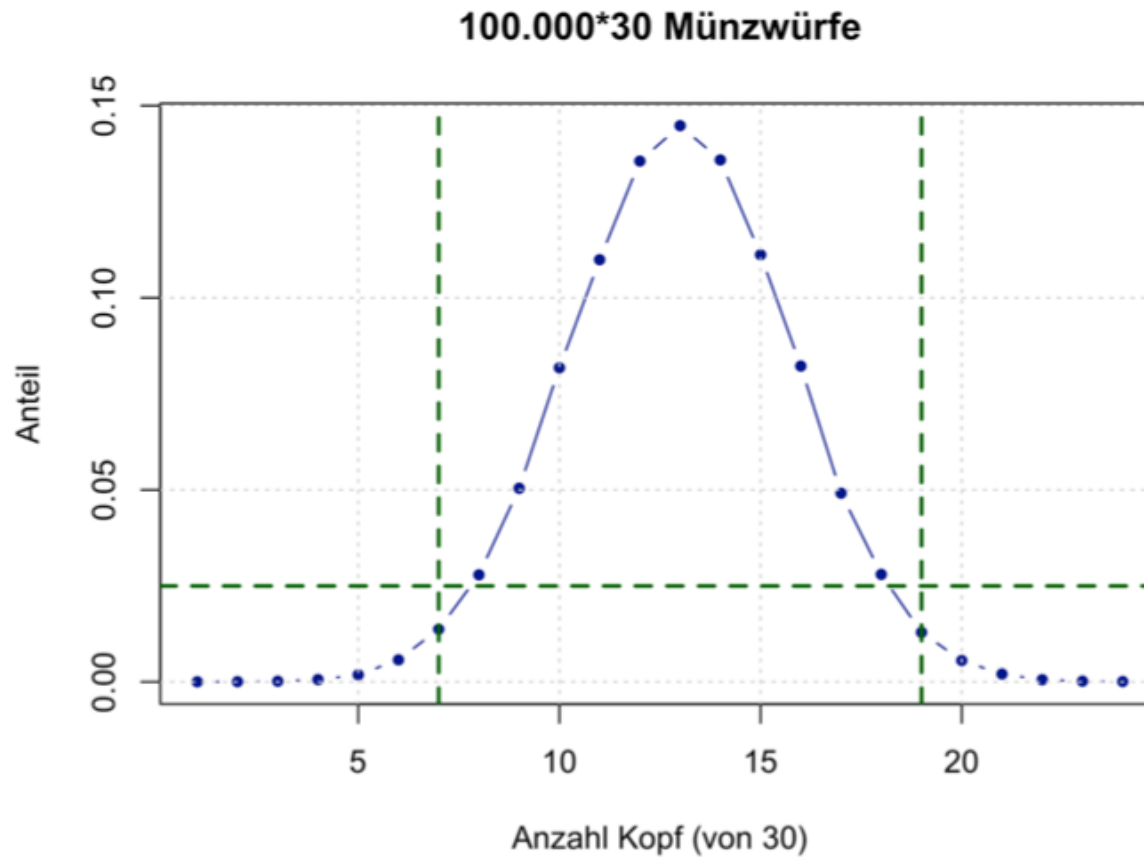
# Signifikanzschwelle



# Signifikanzschwelle



# Signifikanzschwelle





# Zentraler Grenzwertsatz

---

- Zieht man theoretisch unendlich viele Stichproben aus einer Grundgesamtheit, geht die Verteilung mit wachsendem Stichprobenumfang in eine **Normalverteilung** über
- Grenzwert:  $n \geq 30$
- ab einer Stichprobengröße von 30 kann man statistische Verfahren heranziehen, die auf der Normalverteilung beruhen.

# Worauf testen wir?

---

- Isolierte Daten machen in der Regel keinen Sinn
- "Studierende trinken häufig Bier" ist keine statistisch wirklich sinnvolle Aussage
- "Studierende trinken häufiger Bier als Grundschüler" hingegen ist eine falsifizierbare Hypothese.
- Ebenso: "Mit zunehmendem Alter steigt bei Akademikern der Alkoholkonsum"

# Was ist eine wissenschaftliche Hypothese?

---

1. Eine wissenschaftliche Hypothese macht eine **allgemeine** Aussage, die sich auf **mehr als ein einzelnes Ereignis** bezieht.
2. Diese Aussage muss sich in der Form **wenn..., dann** oder **je..., desto** paraphrasieren lassen.
3. Sie ist **potentiell falsifizierbar**.

# Wie testen wir?

---

- Wir testen **Nullhypothesen**...
- ...um durch deren Zurückweisung die **Alternativhypothese** zu untermauern.

## Beispiel:

$H_1$ : Die Münze, die ich werfe, ist gezinkt.

Wie muss  $H_0$  lauten?

$H_0$ : Die Münze, die ich werfe, ist nicht gezinkt.

# Wie testen wir?

---

- Wir testen **Nullhypothesen**...
- ...um durch deren Zurückweisung die **Alternativhypothese** zu untermauern.

## Beispiel:

$H_1$ : Kopf vs. Zahl  $\sim 50:50$

Wie muss  $H_0$  lauten?

$H_0$ : Kopf vs. Zahl  $\neq 50:50$

# Beispiel Münzwurf

Kopf	Zahl
20	0
0	20
7	13

klarer Fall: Kopf häufiger

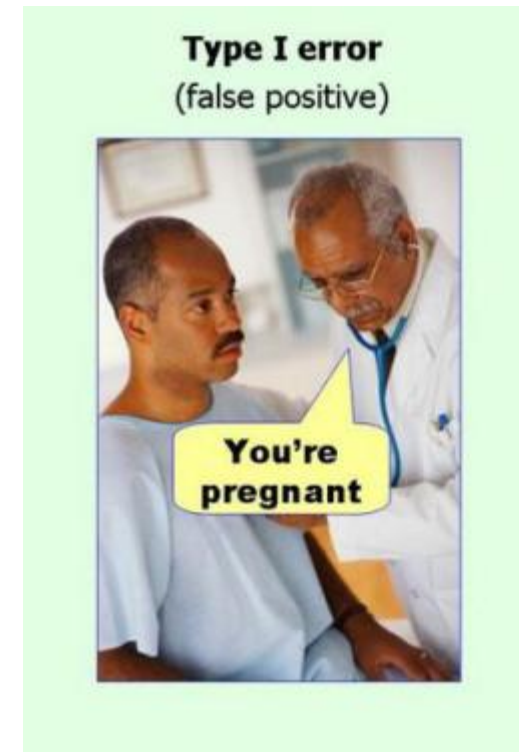
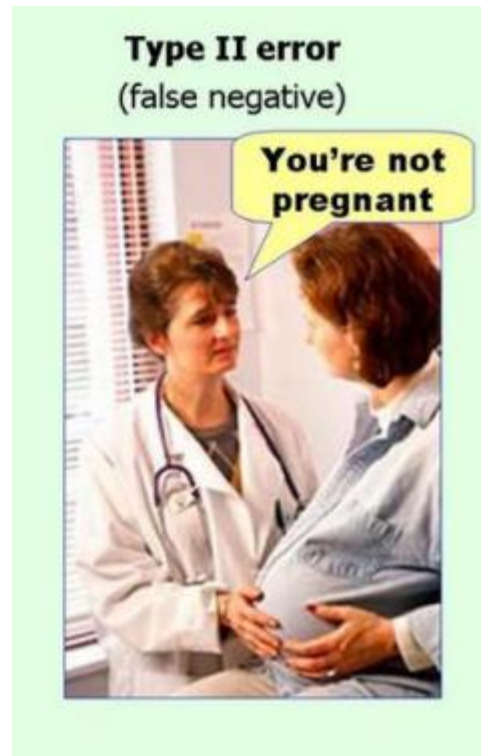
klarer Fall: Zahl häufiger

???

- Können wir bei dieser Verteilung (7:13) die Nullhypothese zurückweisen?

# Fehler erster und zweiter Art

- **Fehler erster Art:** Die Nullhypothese trifft zu, wird aber abgelehnt.
- **Fehler zweiter Art:** Die Nullhypothese ist falsch, wird aber nicht abgelehnt.



Grafik:  
<https://chemicalstatistician.wordpress.com/2014/05/12/applied-statistics-lesson-of-the-day-type-i-error-false-positive-and-type-2-error-false-negative/>

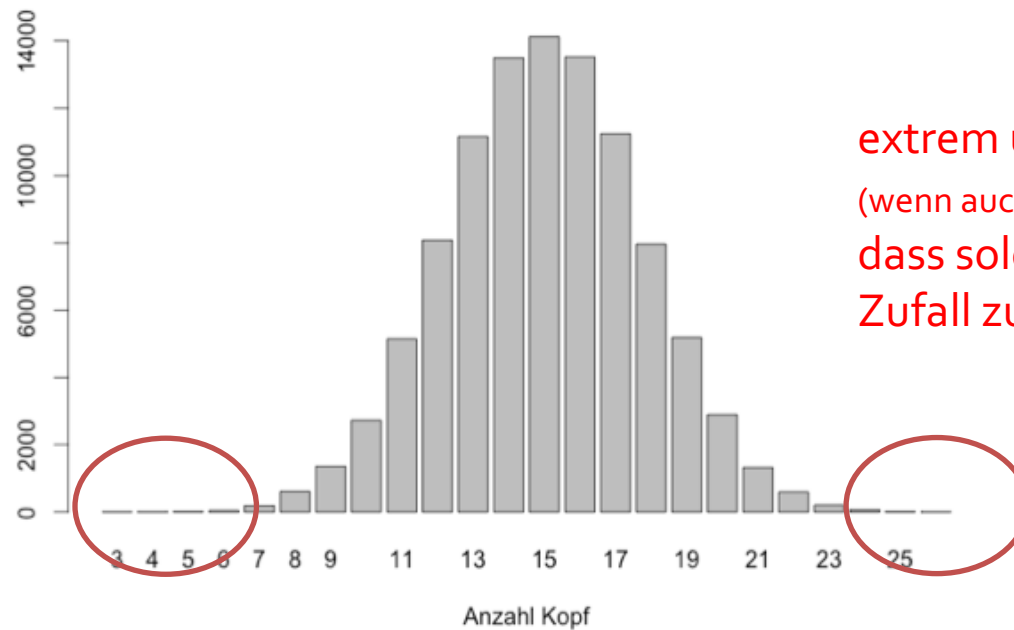
# Was ist Signifikanz?

---

- Wir definieren (im Voraus!) ein **Kriterium**, um  $H_0$  zurückzuweisen
- Das Risiko der **falschen Zurückweisung** (Fehler erster Art) soll möglichst gering sein
- Daher: **Signifikanzschwelle**



Grafik: 100.000\*30  
Würfe einer Münze  
(Simulation)

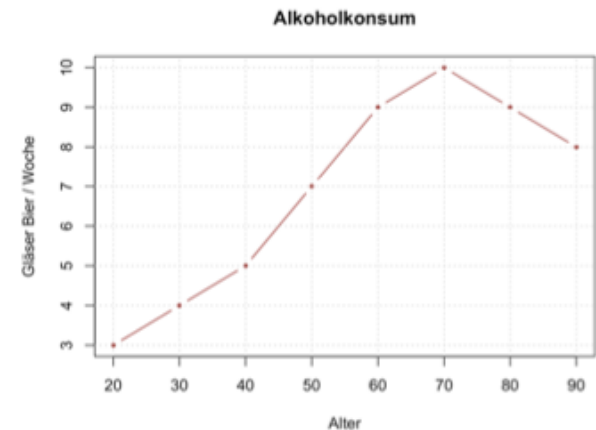
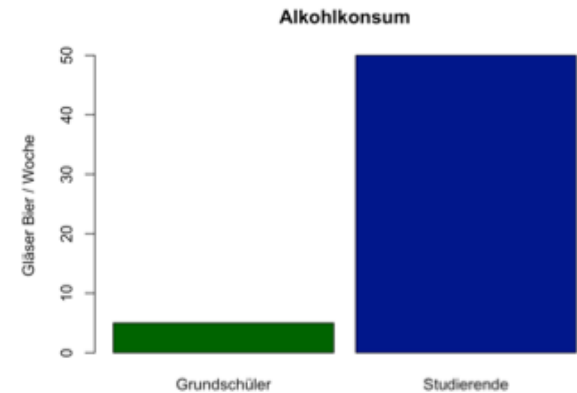


extrem unwahrscheinlich  
(wenn auch prinzipiell möglich),  
dass solche Werte durch  
Zufall zustandekommen.

# Worauf testen wir

d.h. wir testen auf...

- ...signifikante **Unterschiede**
- ...signifikante **Zusammenhänge**



# Wie testen wir?

---

Parametrische vs. non-parametrische Tests:

- Parametrischen Tests liegt eine bestimmte **Verteilung** – meist die Normalverteilung – zugrunde.
- Non-parametrische Tests sind **verteilungsfrei** und können auf Variablen aller Skalenniveaus angewandt werden.

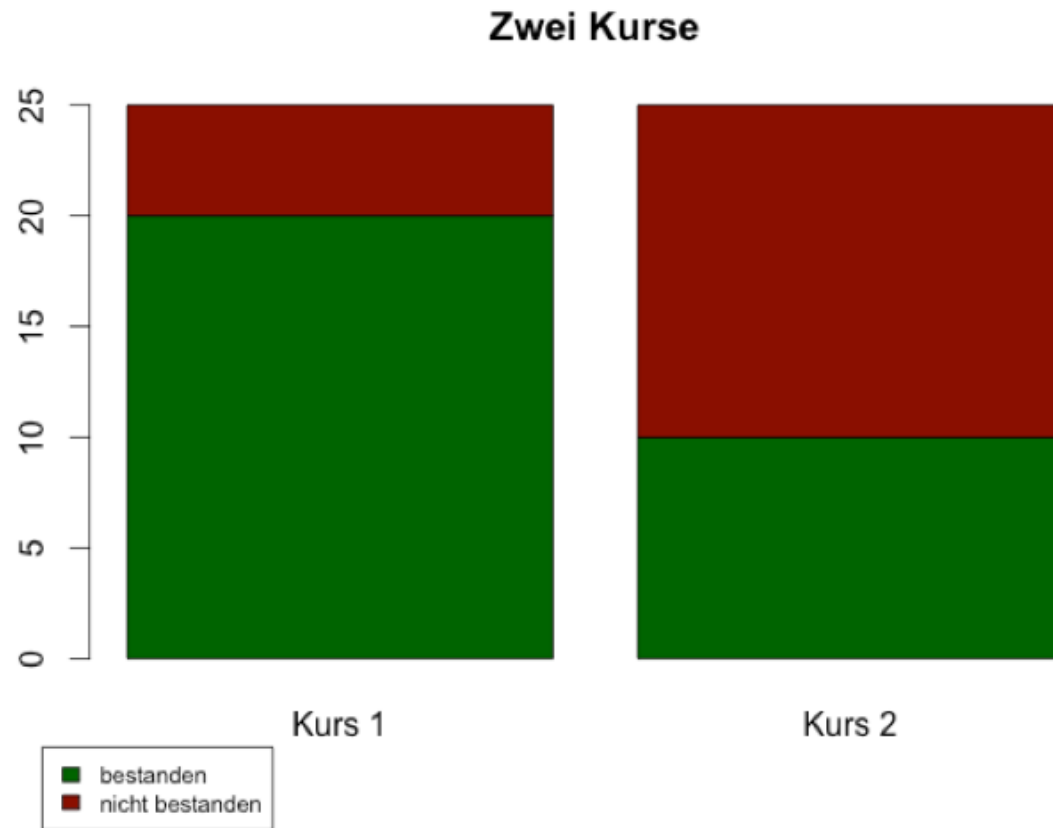
# Beispiel: Chi-Quadrat

## (verteilungsfrei - Nominaldaten)

---

- Beispiel: zwei Kurse in historischer Sprachwissenschaft
- einer geleitet von Dozent A, der andere von Dozent B
- Rahmenbedingungen genau gleich
- alle machen am Ende genau die gleiche Prüfung

# Ergebnisse



# Beispiel: Chi-Quadrat

- Prüfungsergebnisse:

Dozent	bestanden	durchgefallen
Dozent A	20	10
Dozent B	10	20

- Was wäre eigentlich zu erwarten?

Dozent	bestanden	durchgefallen
Dozent A	15	15
Dozent B	15	15

# Beispiel: Chi-Quadrat

$$\sum_{i=1}^n \frac{(\textit{beobachtete Freq.} - \textit{erwartete Freq.})^2}{\textit{erwartete Freq.}}$$

Dozent	bestanden	durchgefallen
Dozent A	20	10
Dozent B	10	20

Dozent	bestanden	durchgefallen
Dozent A	15	15
Dozent B	15	15

# Chi-Quadrat: Voraussetzungen

---

Jeder Test kann nur unter bestimmten **Voraussetzungen** angewandt werden. Bei Chi-Quadrat sind diese:

- Alle Beobachtungen sind unabhängig voneinander
- mind. 80% der beobachteten Werte sind  $\geq 5$
- alle erwarteten Frequenzen sind  $> 1$



# Übungsbeispiel

- Groß- und Kleinschreibung in Hexenverhörprotokollen: belebt vs. unbelebt

beob. Werte		unbelebt	belebt
	klein	1201	451
	groß	576	594

- Erwartete Werte werden errechnet mit:  
$$(\text{Zeilensumme} * \text{Spaltensumme}) / N$$

# Übungsbeispiel

- Groß- und Kleinschreibung in Hexenverhörprotokollen: belebt vs. unbelebt

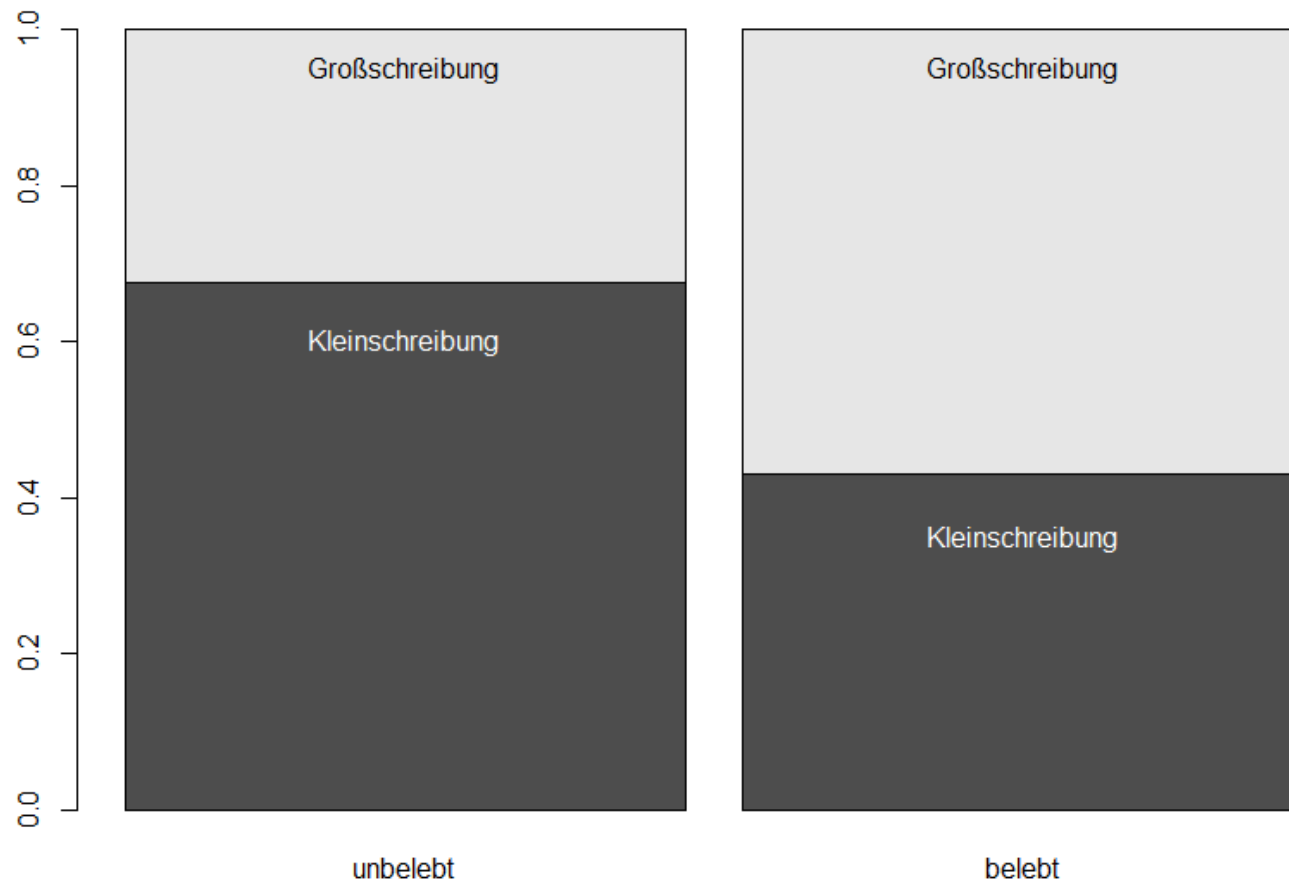
beob.  
Werte

	unbelebt	belebt
klein	1201	451
groß	576	594

erw.  
Werte

	unbelebt	belebt
klein	1040.26	611.74
groß	736.74	233.26

# Übungsbeispiel



# Ergebnis

---

- X-squared = 160.78, df = 1, p-value < 2.2e-16
- **Aber:** p-Wert ist nicht alles!
- Beim Chi-Quadrat-Test (und auch bei anderen Vierfeldertests wie Fisher Exact Test) ist er von der **Stichprobengröße** abhängig.
- Er sagt nichts darüber aus, wie groß der **Effekt** wirklich ist.

# Signifikanz und Effektstärke

---

- Signifikanz gibt an, wie wahrscheinlich es ist, dass eine Verteilung beobachtet werden kann, wenn die Nullhypothese gilt.
- Davon zu unterscheiden ist die **Effektstärke**.

# Effektstärke

---

- berechnet Stärke der beobachteten Korrelation unabhängig von der Stichprobengröße
- bei Chi-Quadrat:  $\phi$  bzw. Cramér's V.

$$V / \phi = \sqrt{\frac{\chi^2}{n \cdot (k-1)}}$$

k = kleinere Anzahl der Zeilen/Spalten

# Phi-Koeffizient

---

- Chi-Quadrat = 160.78
- N = 2822

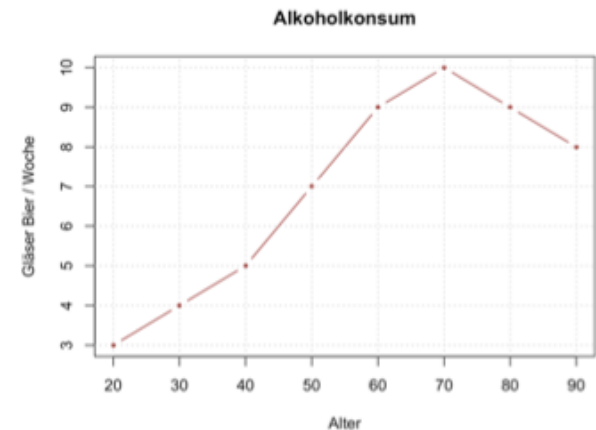
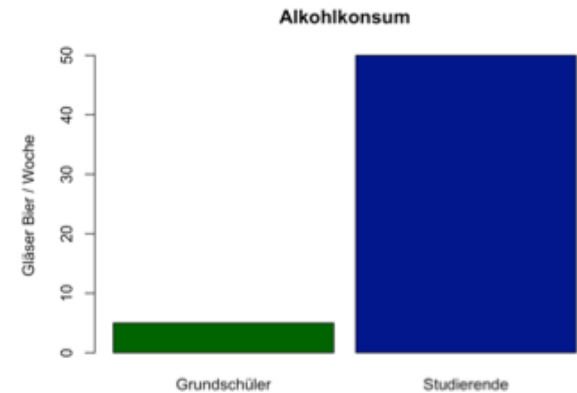
$$\phi = \sqrt{\frac{160.78}{2822}}$$

$$\phi = 0.24$$

# Worauf testen wir

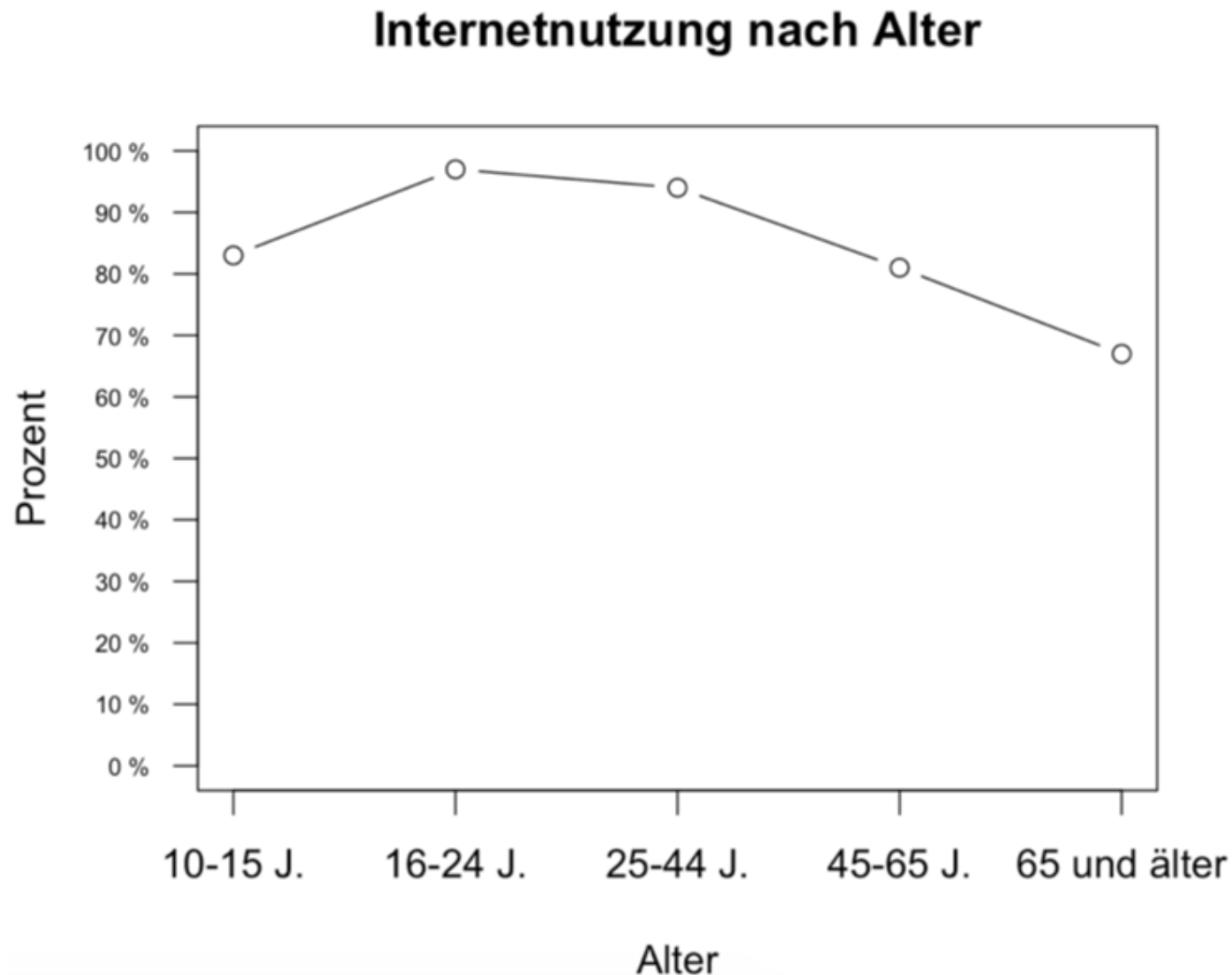
d.h. wir testen auf...

- ...signifikante  
**Unterschiede**
- ...signifikante  
**Zusammenhänge**





# Signifikante Zusammenhänge



# Rangkorrelationskoeffizienten

---

- z.B. Spearman's Rho, Kendall's Tau
- vergleicht die Rangfolge der unabhängigen Variable mit der Rangfolge der abhängigen Variable
- Beispiel (aus Howell 2010): durchschnittliche Ausgaben für Alkohol und Tabak

Howell, David C. 2010. *Statistical Methods for Psychology*. 7th ed. Belmont: Wadsworth.

# Beispiel: Alkohol und Tabak

---

Region	Alkohol	Tabak
Nordirland	4,02	4,56
East Anglia	4,52	2,92
Südwesten	4,79	4,79
East Midlands	4,89	4,89
Wales	5,27	3,53
West Midlands	5,63	3,47
Südosten	5,89	3,2
Schottland	6,08	4,51
Yorkshire	6,13	3,76
Nordosten	6,19	3,77
Norden	6,47	4,02

# Beispiel: Alkohol und Tabak

Region	Alkohol	Tabak	Rang Alkohol
Nordirland	4,02	4,56	1
East Anglia	4,52	2,92	2
Südwesten	4,79	4,79	3
East Midlands	4,89	4,89	4
Wales	5,27	3,53	5
West Midlands	5,63	3,47	6
Südosten	5,89	3,2	7
Schottland	6,08	4,51	8
Yorkshire	6,13	3,76	9
Nordosten	6,19	3,77	10
Norden	6,47	4,02	11

# Beispiel: Alkohol und Tabak

Region	Alkohol	Tabak	Rang Alkohol	Rang Tabak
Nordirland	4,02	4,56	1	11
East Anglia	4,52	2,92	2	2
Südwesten	4,79	4,79	3	1
East Midlands	4,89	4,89	4	4
Wales	5,27	3,53	5	6
West Midlands	5,63	3,47	6	5
Südosten	5,89	3,2	7	3
Schottland	6,08	4,51	8	10
Yorkshire	6,13	3,76	9	7
Nordosten	6,19	3,77	10	8
Norden	6,47	4,02	11	9

# Beispiel: Alkohol und Tabak

Region	Alkohol	Tabak	Rang Alkohol	Rang Tabak	Inversionen
Nordirland	4,02	4,56	1	11	10
East Anglia	4,52	2,92	2	2	1
Südwesten	4,79	4,79	3	1	0
East Midlands	4,89	4,89	4	4	1
Wales	5,27	3,53	5	6	3
West Midlands	5,63	3,47	6	5	1
Südosten	5,89	3,2	7	3	0
Schottland	6,08	4,51	8	10	3
Yorkshire	6,13	3,76	9	7	0
Nordosten	6,19	3,77	10	8	0
Norden	6,47	4,02	11	9	0

# Kendall's Tau

---

$$\tau = 1 - \frac{2 \times (\text{Anzahl der Inversionen})}{\text{Anzahl an Objektpaaren}}$$

- Wie viele Objektpaare?

→  $n(n-1)/2 = 11(10)/2 = 55$

- Wie viele Inversionen?

→ Summe der letzten Tabellenspalte

# Beispiel: Alkohol und Tabak

Region	Alkohol	Tabak	Rang Alkohol	Rang Tabak	Inversionen
Nordirland	4,02	4,56	1	11	10
East Anglia	4,52	2,92	2	2	1
Südwesten	4,79	4,79	3	1	0
East Midlands	4,89	4,89	4	4	1
Wales	5,27	3,53	5	6	3
West Midlands	5,63	3,47	6	5	1
Südosten	5,89	3,2	7	3	0
Schottland	6,08	4,51	8	10	3
Yorkshire	6,13	3,76	9	7	0
Nordosten	6,19	3,77	10	8	0
Norden	6,47	4,02	11	9	0

Summe:  
18



# Kendall's Tau

---

$$\tau = 1 - \frac{2 \times (\text{Anzahl der Inversionen})}{\text{Anzahl an Objektpaaren}}$$

- $1 - 2(18)/55 = 0.345$

# Vergleich von Koeffizienten

---


verbreitete Rangkorrelationskoeffizienten:

- Pearson's  $r$
- Spearman's  $Rho$
- Kendall's  $Tau$

# Vergleich von Koeffizienten

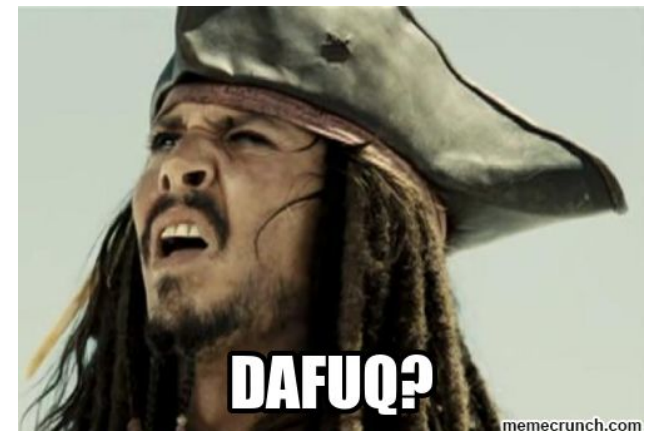
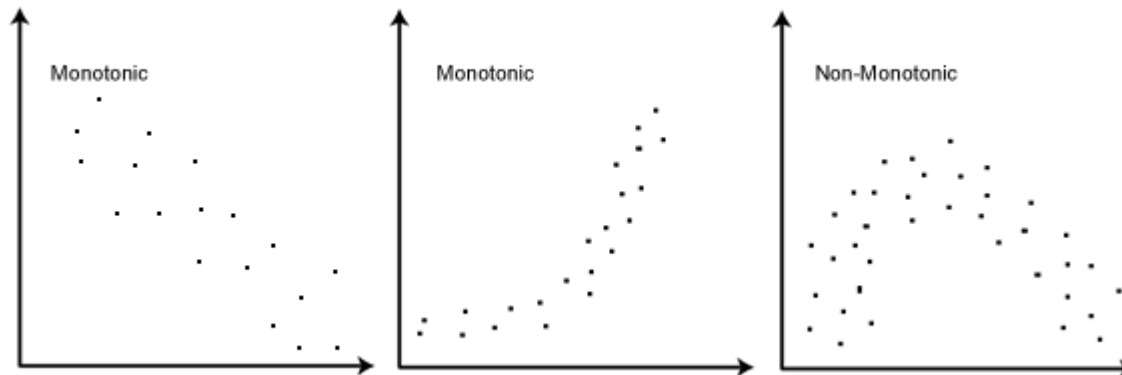
---

verbreitete Rangkorrelationskoeffizienten:

- Pearson's  $r$
  - Spearman's  $Rho$
  - Kendall's  $Tau$
- 
- parametrisch
- nicht-parametrisch

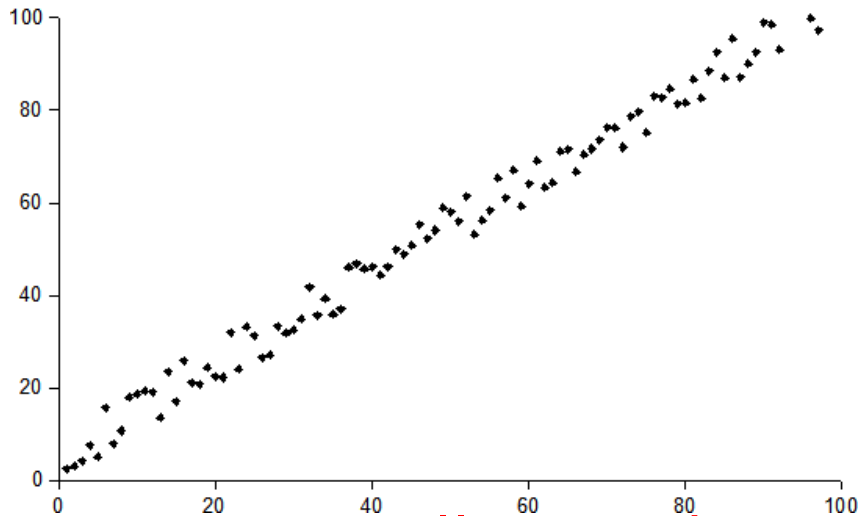
# Vergleich von Koeffizienten

- Die drei Koeffizienten haben unterschiedliche **Voraussetzungen:**
- Pearson's r: bivariate Normalverteilung und/oder  $>30$  Beobachtungen; lineares und monotonen Verhältnis zwischen unabh. und abh. Variable; beide Variablen mindestens intervallskaliert; Homoskedastizität der Residuenvarianz; keine Autokorrelation

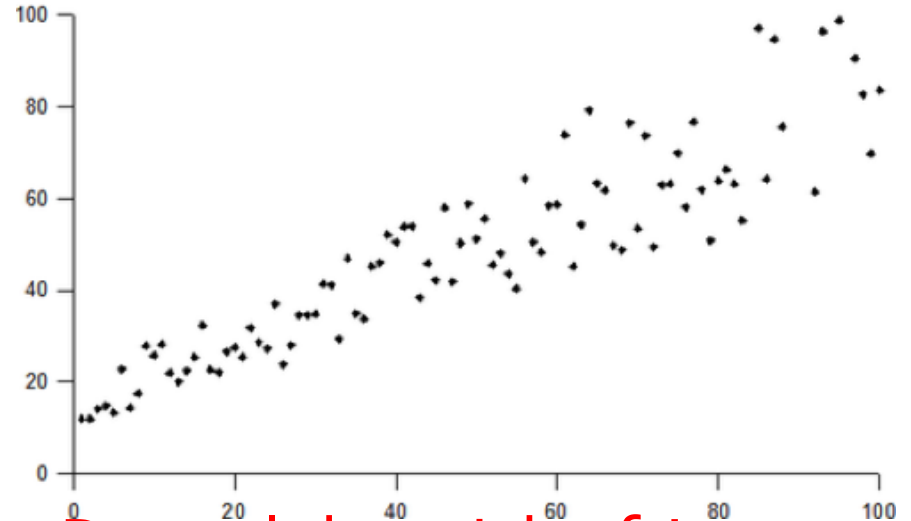


# Homo-/Heteroskedastizität

Homoscedasticity



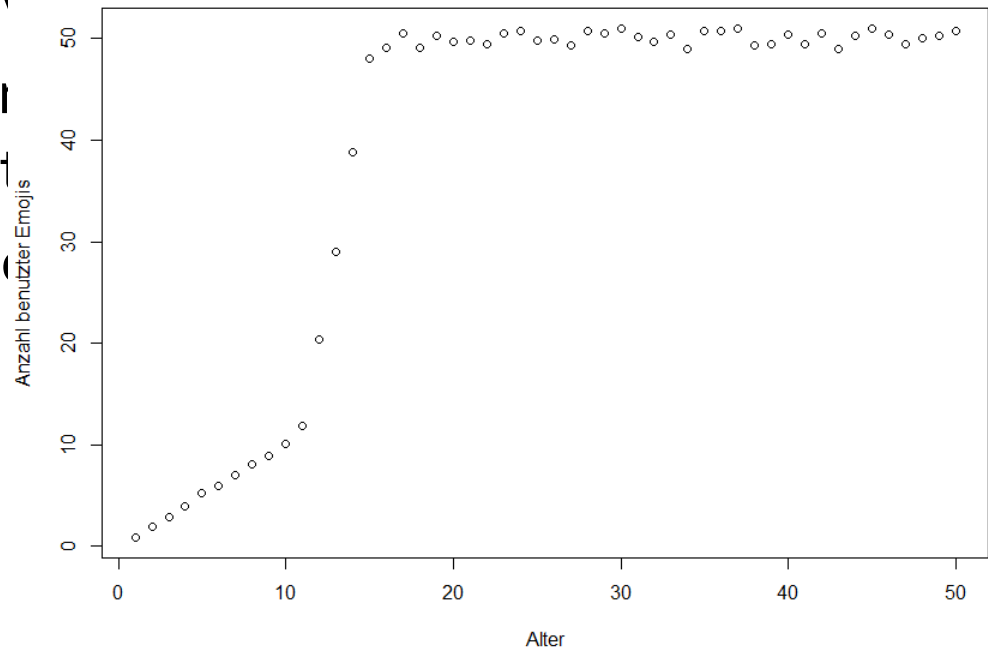
Heteroscedasticity



Visuelle Inspektion der Daten lohnt sich oft!

# Vergleich von Koeffizienten

- Die drei Koeffizienten haben unterschiedliche **Voraussetzungen:**
- Pearson's r: bivariate Normalverteilung und/oder  $>30$  Beobachtungen; lineares  $\rho$  zwischen unabh. und abh. Variable; mindestens intervallskalierte Variable; Residuenvarianz; keine Autokorrelation
- Spearman Rank Test: monotonen Zusammenhang zwischen unabh. und abh. Variable



# Vergleich von Koeffizienten

---

- Die drei Koeffizienten haben unterschiedliche **Voraussetzungen:**
- Pearson's r: bivariate Normalverteilung und/oder  $>30$  Beobachtungen; lineares und monotonen Verhältnis zwischen unabh. und abh. Variable; beide Variablen mindestens intervallskaliert; Homoskedastizität der Residuenvarianz; keine Autokorrelation
- Spearman Rank Test: mindestens intervallskalierte Daten; monotonen Verhältnis zwischen abh. und unabh. Variable
- Kendall's Tau: monotonen Verhältnis zwischen abh. und unabh. Variable

# Vergleich von Koeffizienten

---

x: 1,2,3,4,5,6,7,8,9,10

y: 1,3,7,9,17,19,20,21,23,24

Spearman:

$S = 3.6637e-14$

$\rho = 1$

Pearson:

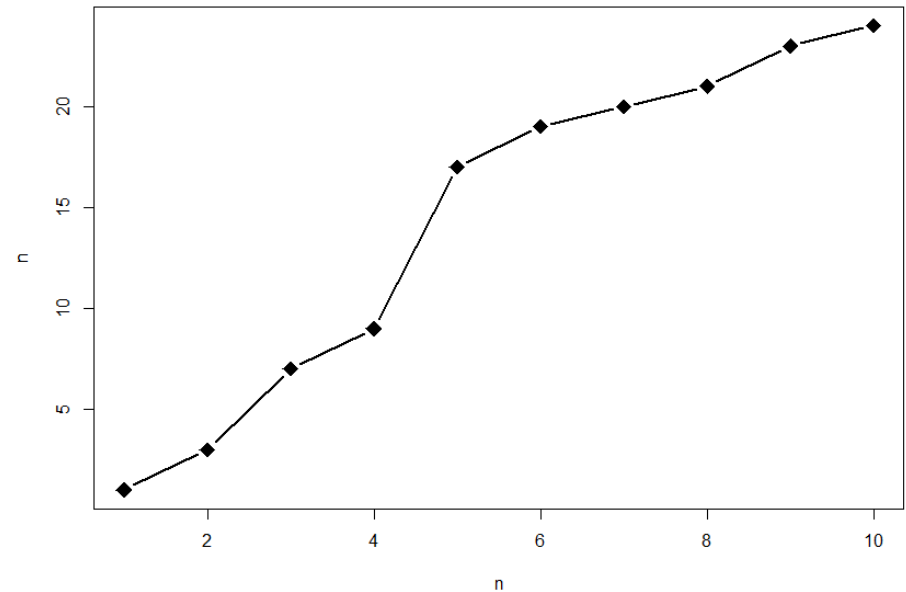
$t = 10.70$

$\text{cor} = 0.67$

Kendall:

$T = 45$

$\tau = 1$





# Vergleich von Koeffizienten

---

X: 1, 11, 21, 31, 41, 51, 61, 71, 81, 91

Y: 1,3,7,9,17,19,20,21,23,24

Spearman:

$S = 3.6637e-14$

$\rho=1$

Pearson:

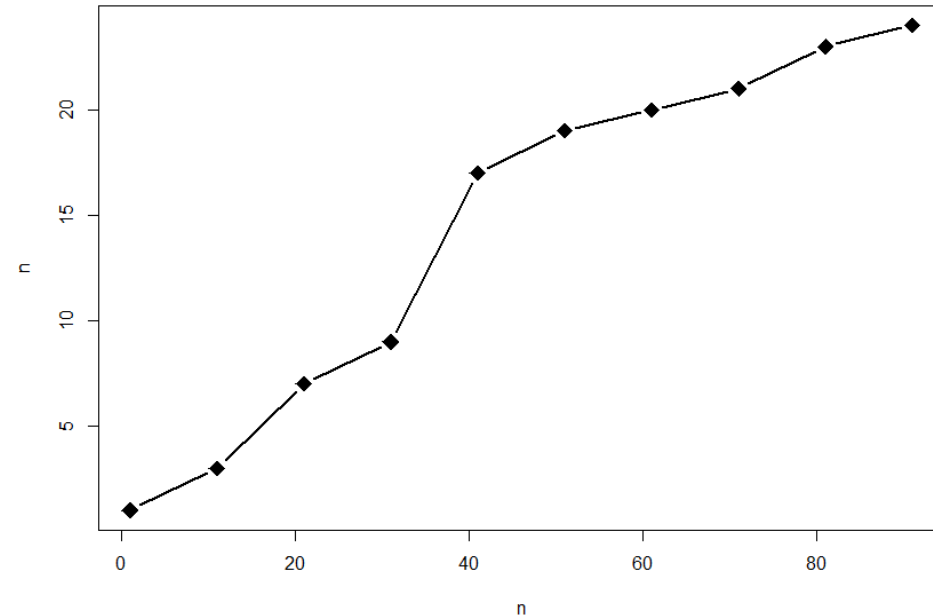
$t=10.70$

$\text{cor}=0.67$

Kendall:

$T=45$

$\tau=1$



# Vergleich von Koeffizienten

X: 1,3,5,10,17,19,22,100,120,140

Y: 1,3,7,9,17,19,20,21,23,24

Spearman:

$S = 3.6637e-14$

$\rho = 1$

Pearson:

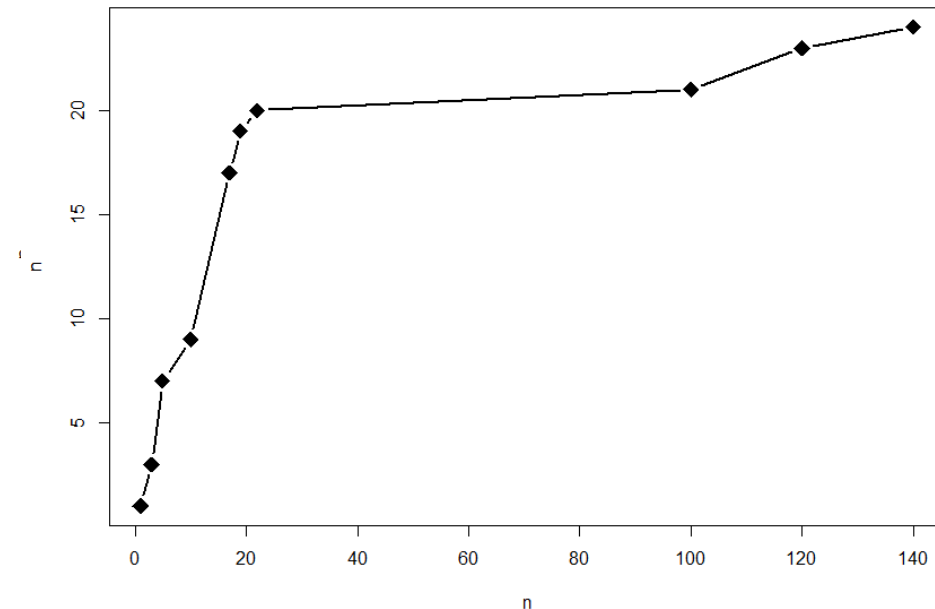
$t = 3.28$

$\text{cor} = 0.76, p = 0.01$

Kendall:

$T = 45$

$\tau = 1$



# Wie berechne ich Koeffizienten?

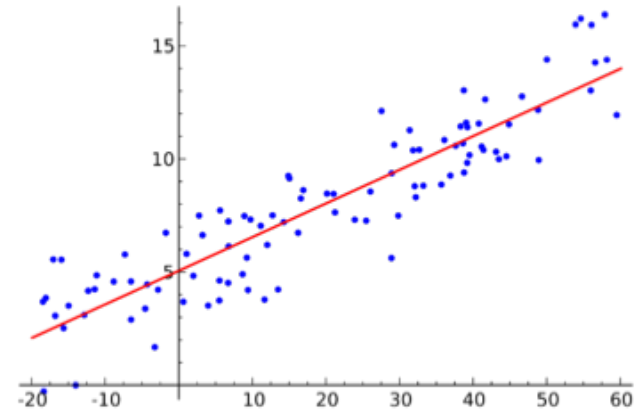
---

- Pearsons Produkt-Moment-Korrelation ist über die Excel-Funktion KORREL verfügbar
- Spearman lässt sich z.B. über <http://vassarstats.net> errechnen
- alle Koeffizienten lassen sich mit etwas Basiswissen gut in R errechnen.

# Was gibt es sonst noch?

- **Regression**

- linear (metrische Daten) oder logistisch (kategoriale Daten)
- Grundformel:  
$$\text{outcome} = \text{predictor}_1 + \text{predictor}_2 + \dots + \text{error}$$



- **Bayessche Statistik**

- Fokus auf Erwartungen / Vorhersagen auf der Basis bereits bekannter Informationen

# Zum Weiterlesen

---

## Lineare Modelle und lineare gemischte Modelle

- Tutorials auf [www.bodowinter.com](http://www.bodowinter.com)

## Bayessche Statistik

- McElreath, Richard. 2016. *Statistical Rethinking. A Bayesian Course with R and Stan*. Boca Raton: CRC Press.