WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Einführung in die Quantitative Datenanalyse

Sitzung 5: Lineare Regression III und Deskriptive Statistik

Proseminar an der Freien Universität Berlin
2.05.2017 - Marcus Spittler

Inhalt der 5. Sitzung

- **Univariate Häufigkeitsverteilungen**
  - Absolute Häufigkeit
  - Relative Häufigkeit
  - Kumulierte Häufigkeit
- Grafische Darstellung
  - Balkendiagramm
  - Histogramm
  - Dichteverteilung
  - Boxplot
- **Lineare Regression III**
  - Modellbildung
  - Interpretation

Univariate Häufigkeitsverteilungen

Annahme: nicht-häufbare Merkmale, das heißt: *eine Untersuchungseinheit kann stets nur eine Ausprägung bei einem Merkmal haben*. D.h. die Antwortmöglichkeiten sind **diskjunkt**, es gibt keine Schnittmenge zwischen ihnen.

Ordnet man den Merkmalsausprägungen Häufigkeiten zu, so erhält man eine Häufigkeitsverteilung.

Absolute Häufigkeit

Absolute Häufigkeit h_{x_j} oder einfacher h_j: Die Anzahl der Merkmalsträger mit der Ausprägung x_j.

Eigenschaften:

$$0 <= h_j <= 1$$

$$\sum_{j=1}^{k}(h_j) = n$$

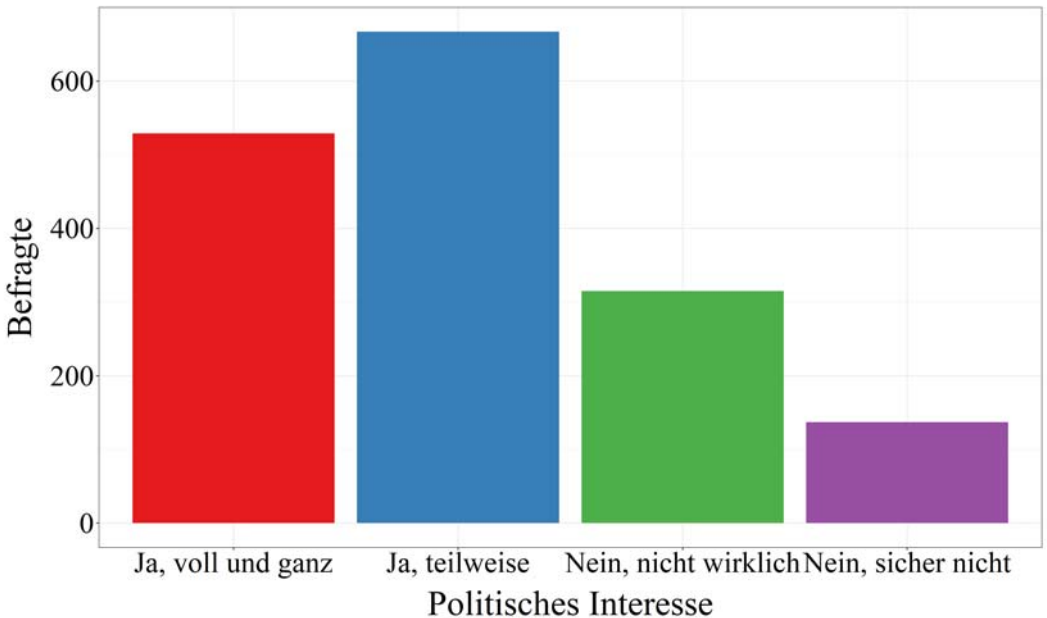
Absolute Häufigkeit

Beispiel: In der Wahlstudie zur Europawahl 2014 wurde gefragt: *Bitte sagen Sie mir für jede der folgenden Aussagen , inwieweit diese Ihrer Ansicht oder Meinung entspricht bzw. nicht entspricht: Sie sind sehr an Politik interessiert.*

Variable	Absolute H.
Ja, voll und ganz	529
Ja, teilweise	667
Nein, nicht wirklich	315
Nein, sicher nicht	137
Summe:	1648

Quelle: Schmitt, Hermann; Popa, Sebastian Adrian; Devinger, Felix (2015): *European Parliament Election Study 2014*, Voter Study, SVoter Study, Supplementary Study. GESIS Data Archive, Cologne. ZA5161 Data file Version 1.0.0, doi:10.4232/1.5161

Balkendiagramm (*Barchart*) mit absoluten Häufigkeiten



Relative Häufigkeit

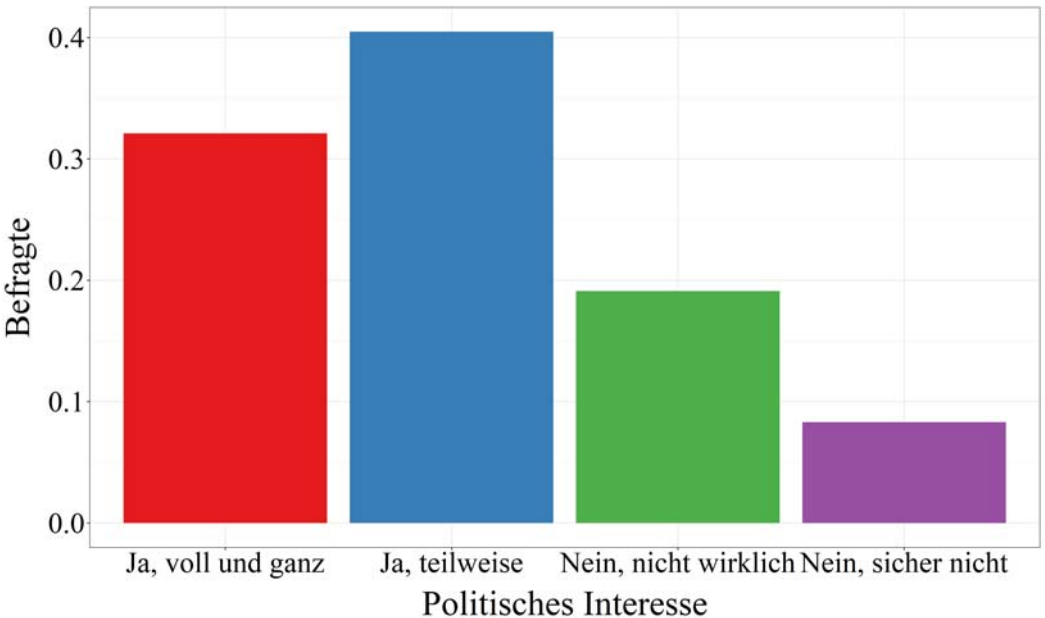
Vergleich zweier absoluten Häufigkeitsverteilungen für ein Merkmal schwierig, v.a. wenn n unterschiedlich, daher Verwendung der relativen Häufigkeit.

$$f_{x_j} = \frac{h_{x_j}}{n}$$

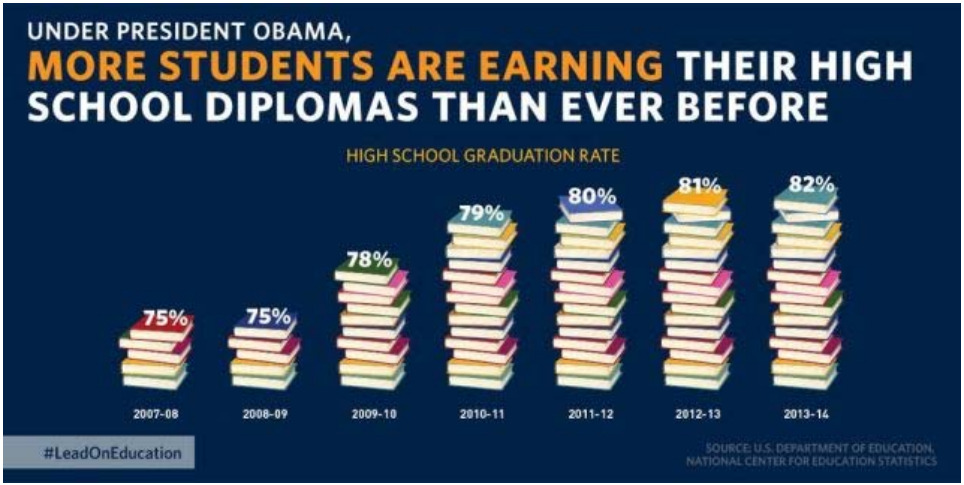
$f_j * 100$ ergibt die prozentualen Anteile

Variable	**Absolute H.**	**Relative H.**
Ja, voll und ganz	529	0.32
Ja, teilweise	667	0.40
Nein, nicht wirklich	315	0.19
Nein, sicher nicht	137	0.08
Summe	1648	1
Variable	**Absolute H.**	**Relative H.**

Balkendiagramm (*Barchart*) mit relativen Häufigkeiten

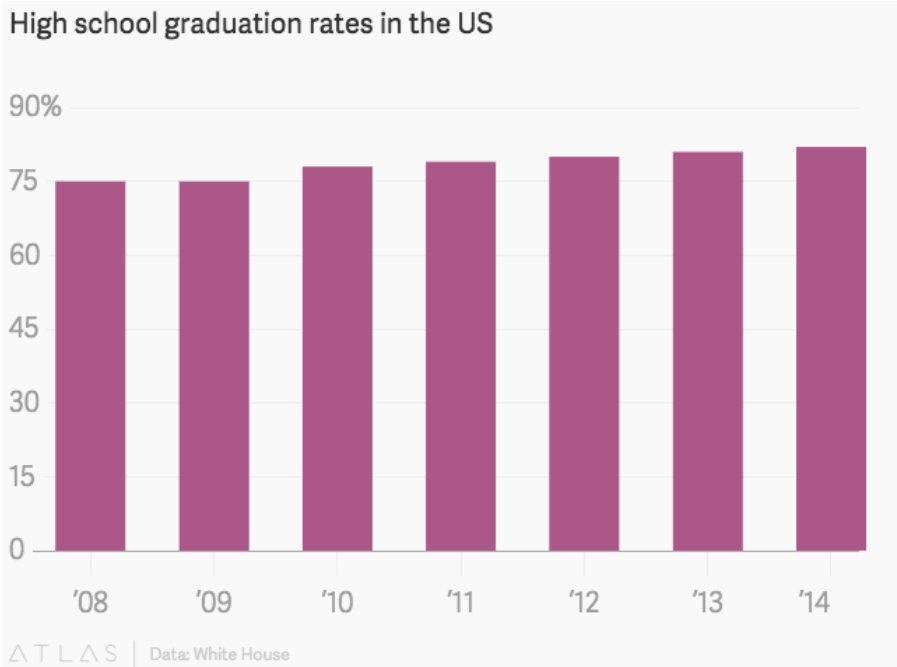


Irreführende Grafiken: Balkendiagramm



Balkendiagramme sollten immer am Nullpunkt beginnen. Falls nicht, sollten sie das sehr deutlich dokumentieren.

Irreführende Grafiken: Balkendiagramm



Die gleichen Daten in einer korrigierten Grafik

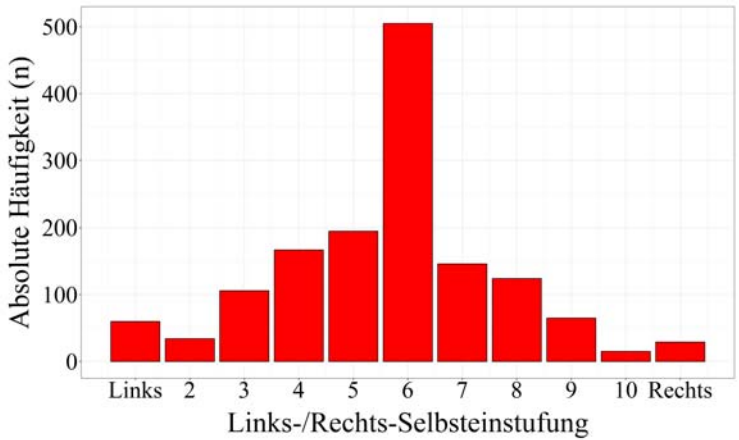
Kumulierte Häufigkeit

Mit den relativen Summenhäufigkeiten lässt sich die Summenhäufigkeitsfunktion F_{x_j} definieren (*empirische Verteilungsfunktion*). Sie gibt zu jeder Merkmalsausprägung den Anteil der Untersuchungseinheiten an, die kleiner oder höchstens gleich einer Ausprägung sind. Die Summenhäufigkeitsfunktion hat (insbesondere bei nur wenigen Ausprägungen) das Bild einer Treppenfunktion.

Variable	Absolute H.	Relative H.	Kumulierte H.
Ja, voll und ganz	529	0.32	0.32
Ja, teilweise	667	0.40	0.72
Nein, nicht wirklich	315	0.19	0.91
Nein, sicher nicht	137	0.08	0.99
Summe:	1648	1	-

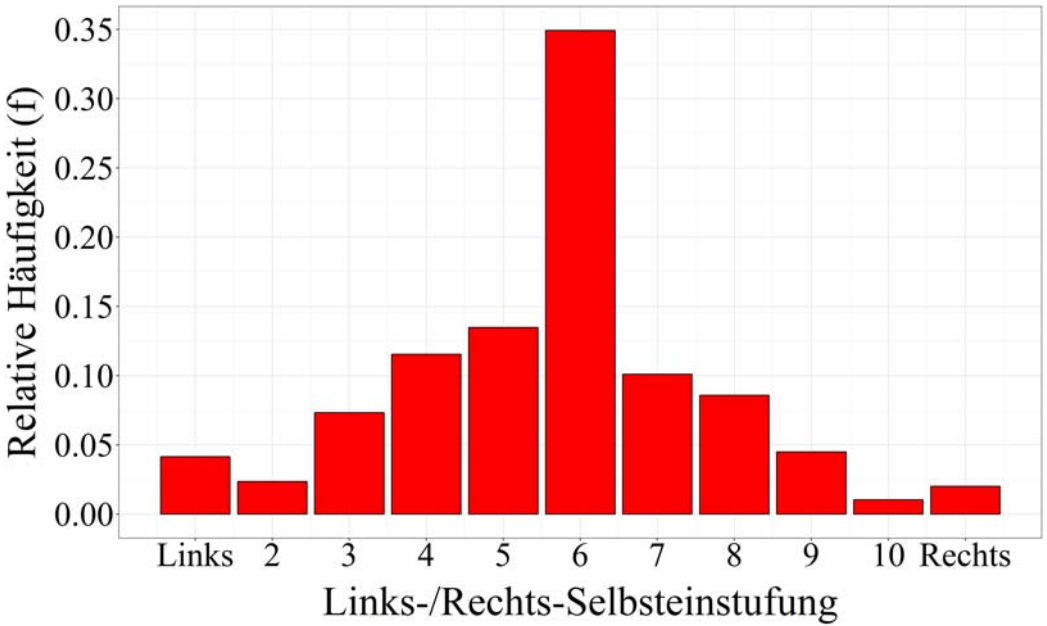
Balkendiagramm mit linearem Merkmal

In der Politik spricht man von links und rechts. Welche Position haben Sie? Bitte geben Sie Ihren persönlichen Standpunkt auf einer Skala von 0 bis 10 an. 0 bedeutet □links□ und 10 bedeutet □rechts□. Welche Zahl gibt am besten Ihren Standpunkt wider?

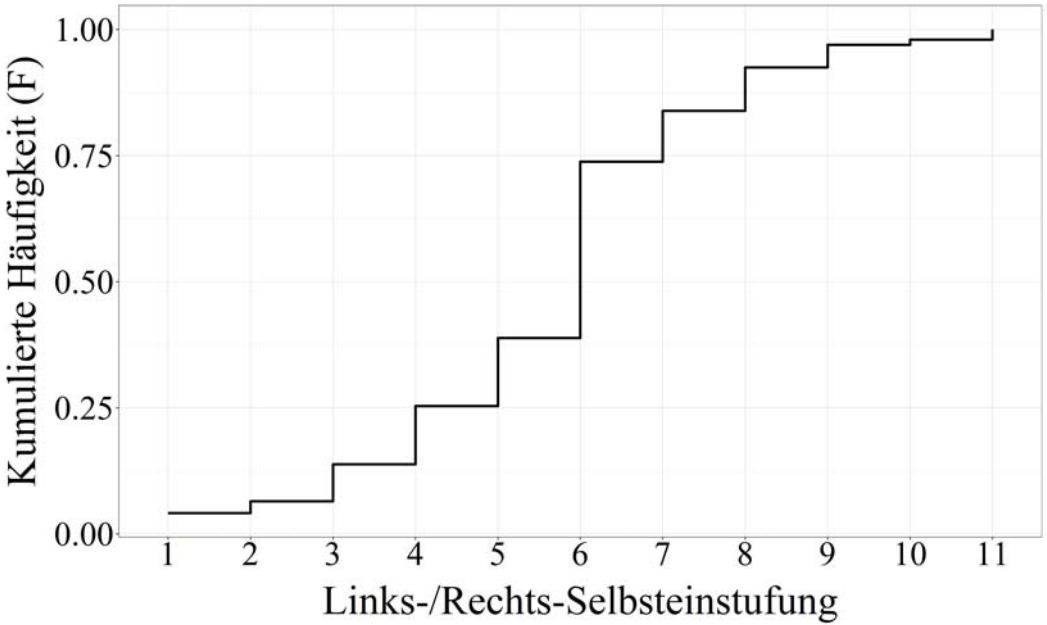


Quelle: EES Voter Study 2014

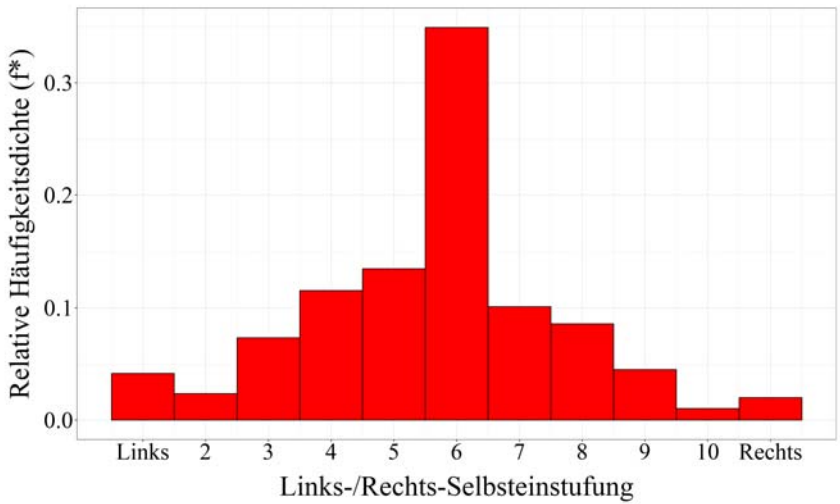
Balkendiagramm mit relativen Häufigkeiten



Treppenfunktion

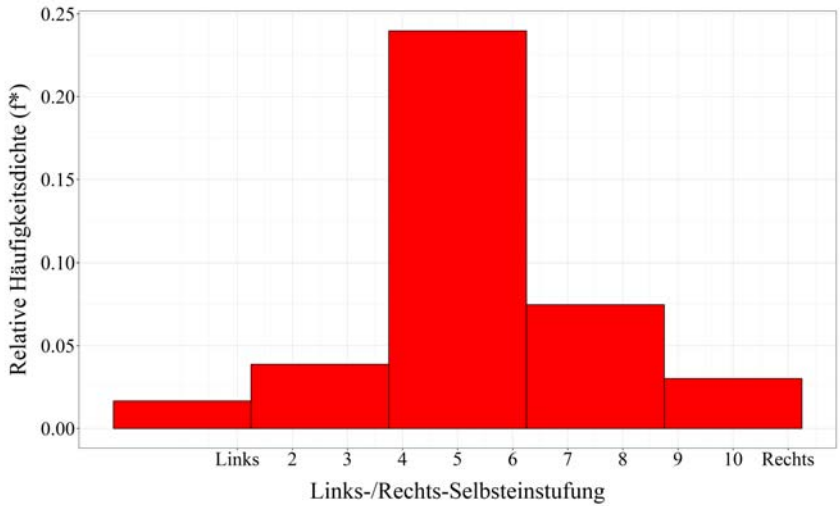


Histogramm (*Histogram*)



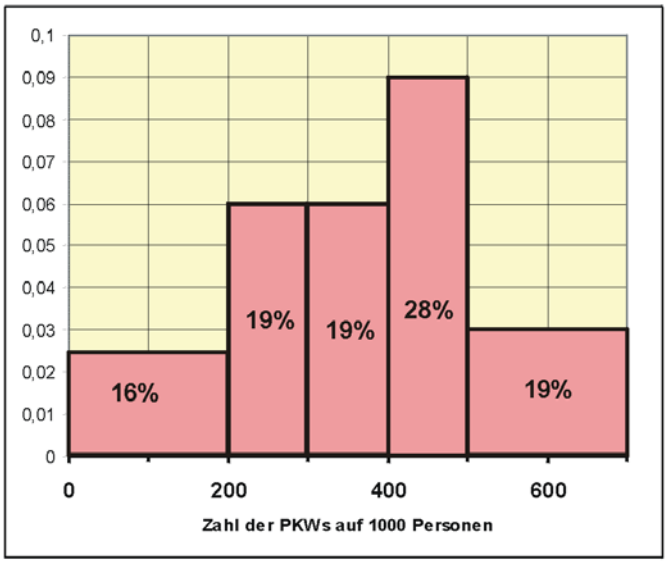
Im Unterschied zum Balkendiagramm sind hier die Flächen interpretierbar.
Metrische Merkmale werden in Klassen eingeteilt (engl. *bins*) mit konstanter oder variabler **Klassenbreite**.

Histogramm



Histogramm mit alternativer Klasseneinteilung

Histogramm

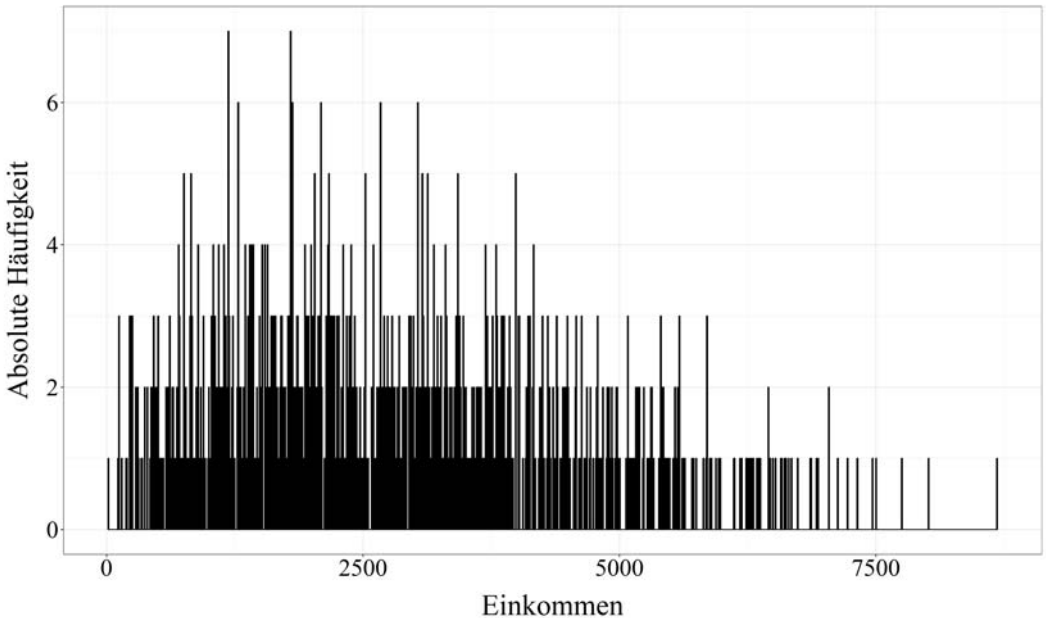


Hier ist ein Histogramm hilfreich, da die **Klassen** unterschiedlich **breit** sind.

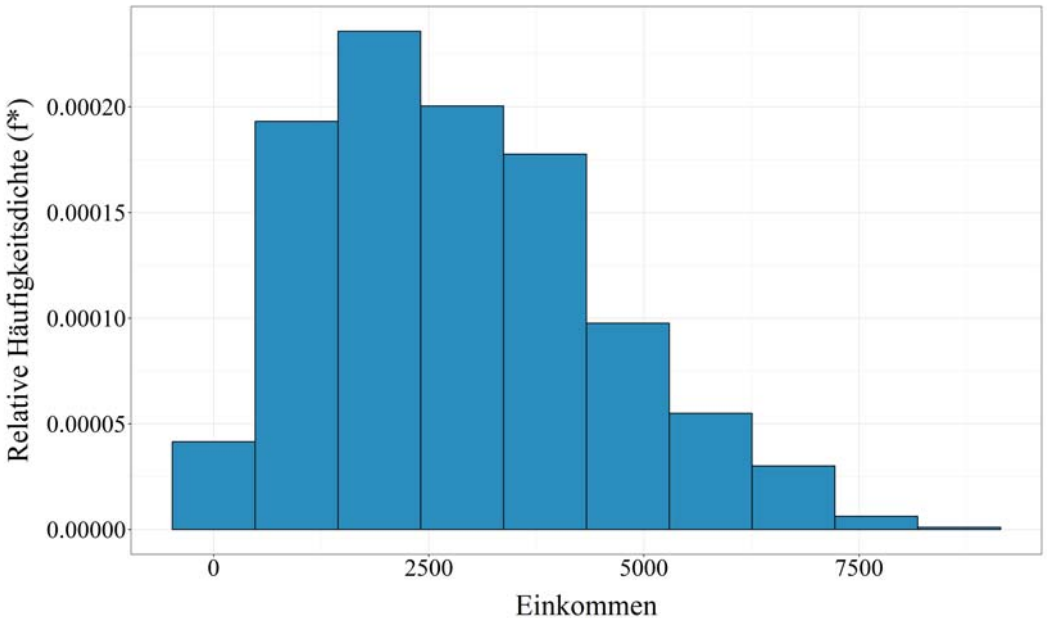
Klassierte Daten

- Es gibt zwei Gründe, klassierte Daten zu betrachten:
  - Es gibt bei einer Befragung sehr viele unterschiedliche Merkmalswerte, so dass die empirische Verteilungsfunktion zu nahezu keiner Informationsverdichtung führt: **Nachträgliche Klassenbildung.**
  - Es sind in einer sekundärstatistischen Analyse nur klassierte Häufigkeitstabellen verfügbar: **Rechnen mit vorgegebenen Klassengrenzen.**
- Die **Gestaltungsparameter** der Klassierung sind **Anzahl** und **Breite** der Klassen.

Histogramm mit nachträglicher Klassenbildung



Histogramm mit nachträglicher Klassenbildung



Klassierte Daten

Beispiel: Dauer von Arbeitslosigkeit

Klasse	Dauer (in Monaten)	Klassenbreite	Klassenmitte	Anzahl
1	0 bis 1	1	0.5	19
2	über 1 bis 2	1	1.5	12
3	über 2 bis 3	1	2.5	24
4	über 3 bis 6	3	4.5	28
5	über 6 bis 12	6	9.0	31
6	über 12 bis 24	12	18.0	6
Summe:				120

Häufigkeitsdichte

Die absolute Häufigkeit gibt an, wie viele der Beobachtungen in eine Klasse fallen. Wird mit **ungleichen Klassenbreiten** gearbeitet, so ist neben der absoluten bzw. relativen Häufigkeit auch die **Häufigkeitsdichte** interessant.

Zweck der **Häufigkeitsdichte** ist es bei ungleichen Klassenbreiten die tatsächlichen Häufigkeiten durch die jeweilige Klassenbreite zu relativieren.

Häufigkeitsdichte ist definiert als:

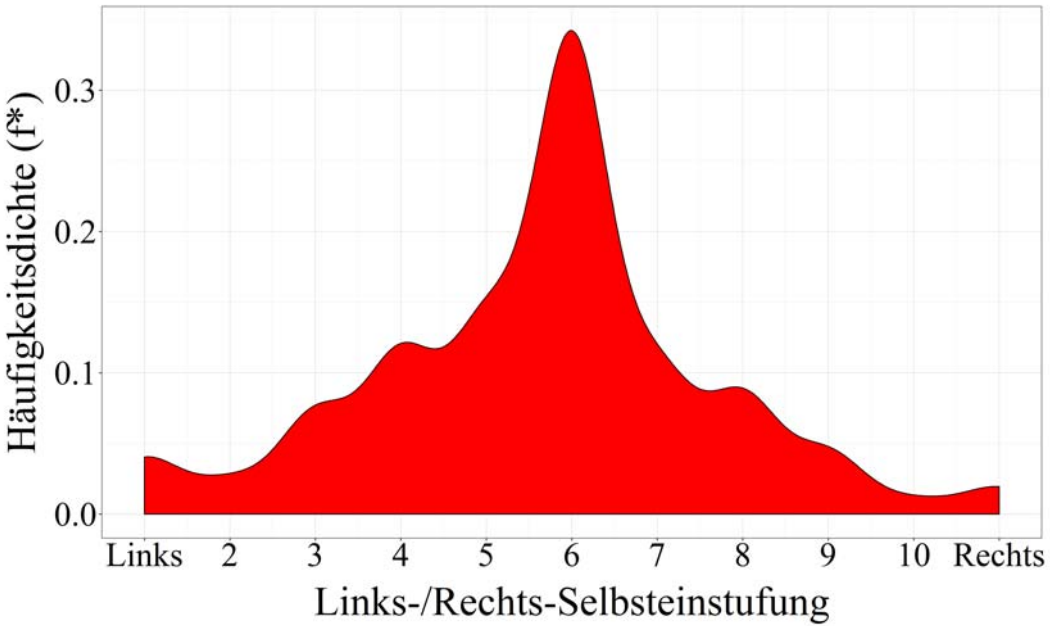
$$f^*(h_{x_j}) = \frac{f_j}{\Delta x_j} = \frac{Relative\ H.}{Klassenbreite}$$

Häufigkeitsdichte

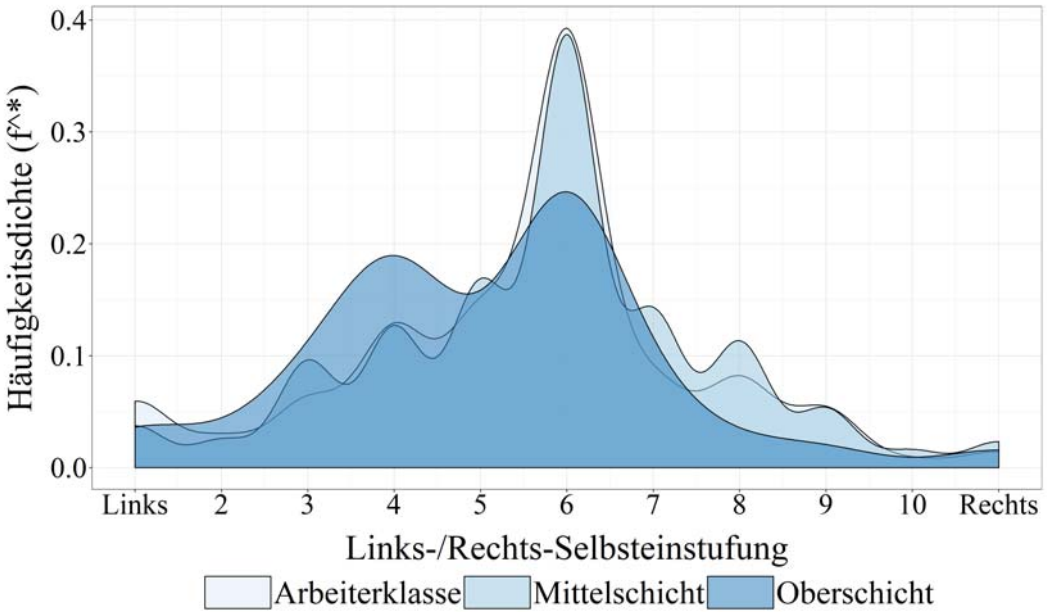
Beispiel: Dauer von Arbeitslosigkeit

j	$(\tilde{x}_{j-1}; \tilde{x}_j]$	Δx_j	h_j	f_j	F_j	f_j^*
1	0-1	1	19	0.16	0.16	0.160
2	1-2	1	12	0.10	0.26	0.100
3	2-3	1	24	0.20	0.46	0.200
4	3-6	3	28	0.23	0.69	0.077
5	6-12	6	31	0.26	0.95	0.043
6	12-24	12	6	0.05	1.00	0.004
Summe			120	1		

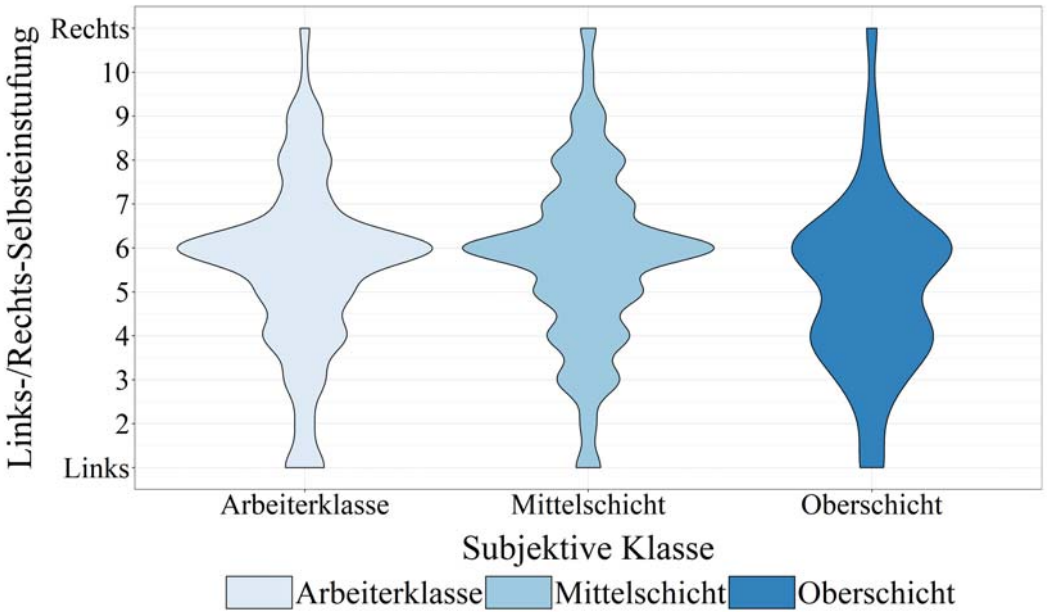
Dichteverteilung



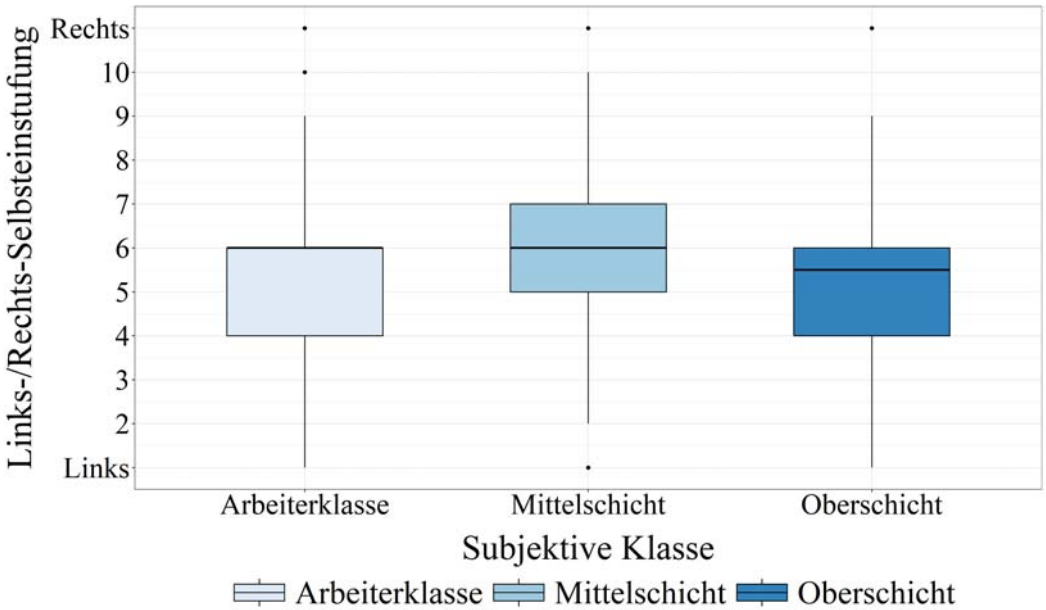
Dichteverteilung



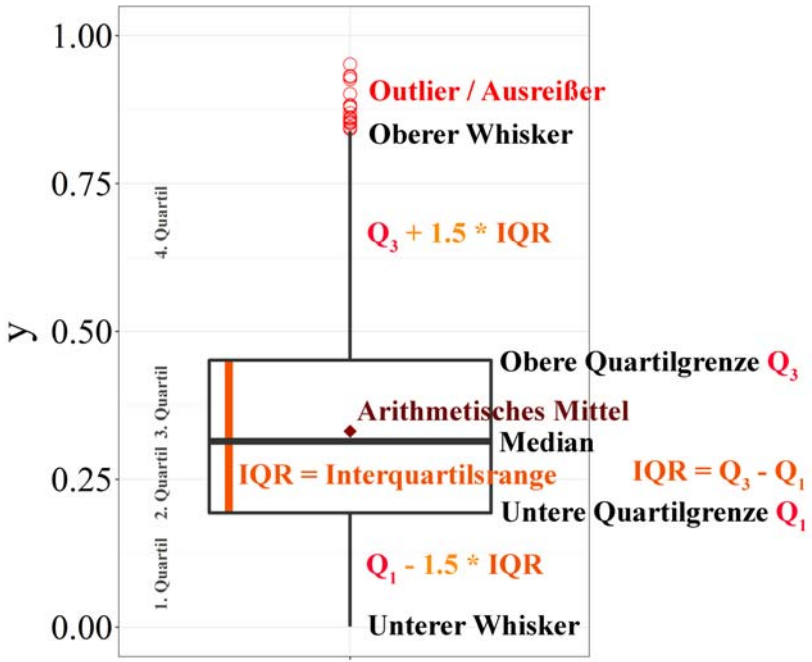
Violin plot


Links-/Rechts-Selbsteinstufung
Rechts
10
9
8
7
6
5
4
3
2
Links
Arbeiterklasse
Mittelschicht
Oberschicht
Subjektive Klasse
Arbeiterklasse
Mittelschicht
Oberschicht

Boxplot

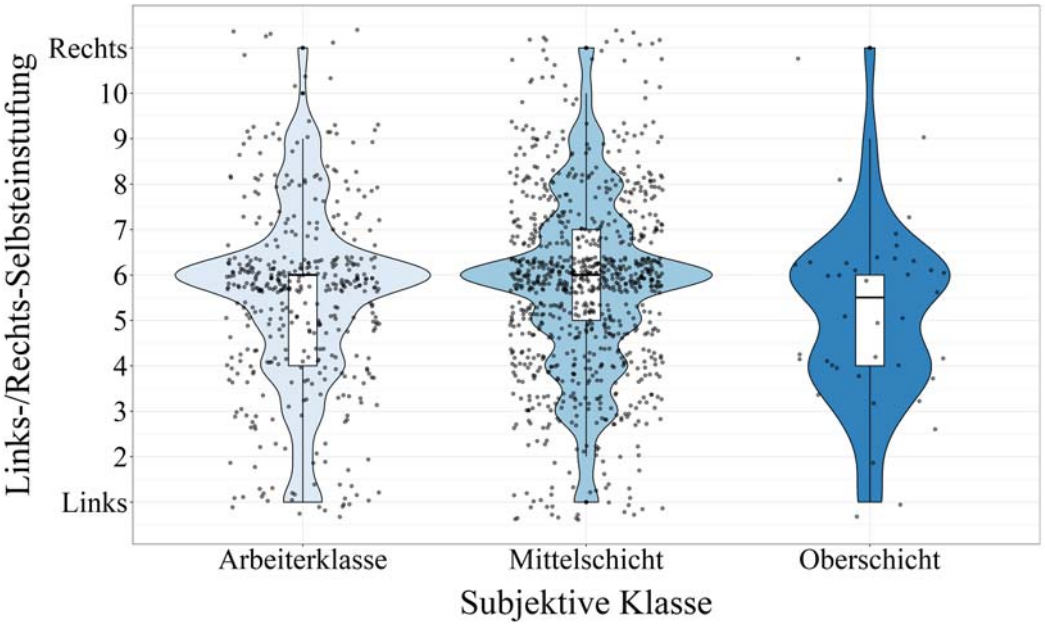

Links-/Rechts-Selbsteinstufung
Rechts
10
9
8
7
6
5
4
3
2
Links
Arbeiterklasse
Mittelschicht
Oberschicht
Subjektive Klasse
Arbeiterklasse
Mittelschicht
Oberschicht

Boxplot

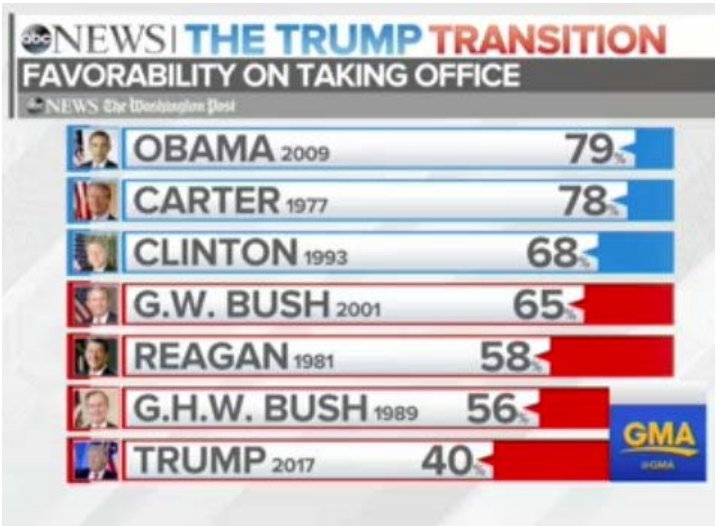


Weitere Erläuterung

Boxplot

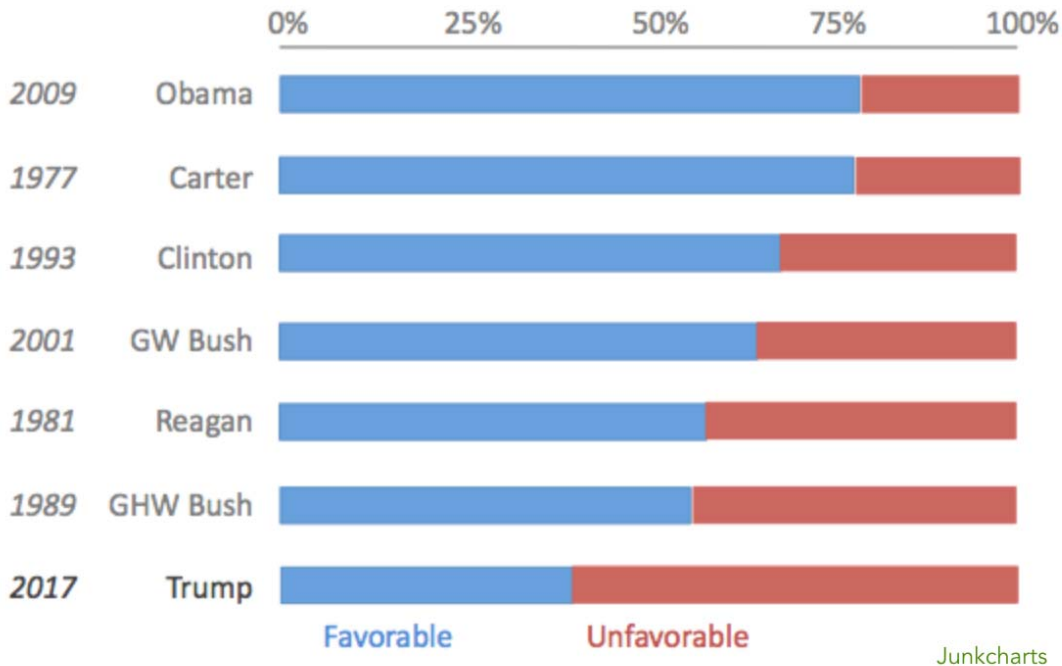

Links-/Rechts-Selbsteinstufung
Rechts
10
9
8
7
6
5
4
3
2
Links
Arbeiterklasse
Mittelschicht
Oberschicht
Subjektive Klasse

Irreführende Grafiken: ABC News



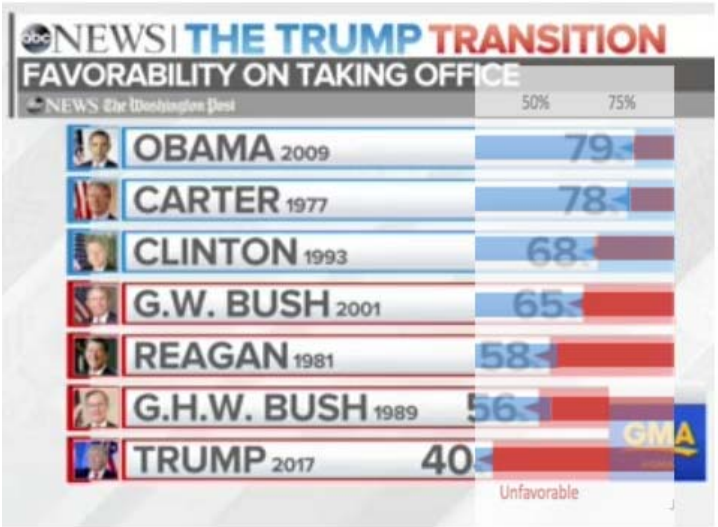
Quelle: http://junkcharts.typepad.com/

Irreführende Grafiken: ABC News

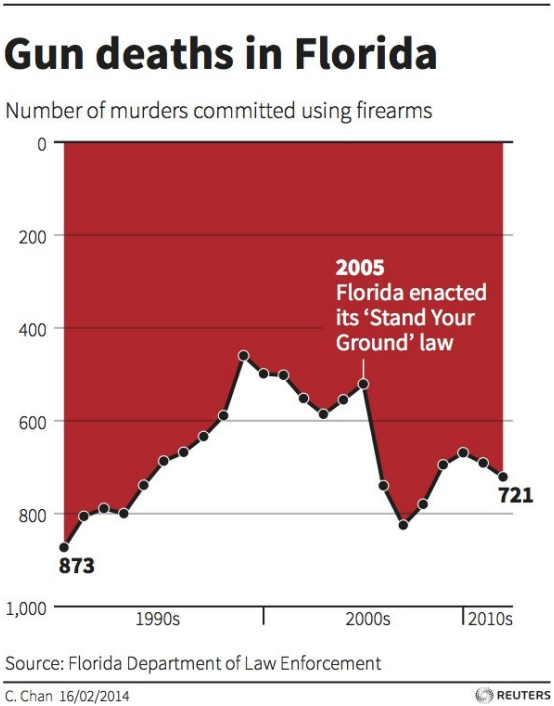


Irreführende Grafiken: ABC News

Irreführende Grafiken: ABC News



Irreführende Grafiken: Tote durch Schusswaffen



Quelle: Ravi Parikh - Heap Analytics

Modellinterpretation

- Für das **Gesamtmodell** interessiert uns die Modellgüte. Diese lesen wir am Bestimmtheitsmaß R^2 ab.

- Für die **einzelnen Erklärungsfaktoren** des Modells prüfen wir:
  - Effektstärke (hier: *unstandardisierter Regressionskoeffizient*)
  - Richtung des Zusammenhangs
  - Signifikanz

```
lm(ptv.spd ~ europe.unification + left.right + gender, data=A) %>%
  summary()
```

```
##
## Call:
## lm(formula = ptv.spd ~ europe.unification + left.right + gender,
##     data = A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5280 -2.3478  0.0403  2.7652  6.1385
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.22588    0.35116  20.577  < 2e-16 ***
## europe.unification  0.12773    0.03084   4.142 3.66e-05 ***
## left.right         -0.31746    0.04391  -7.229 8.08e-13 ***
## genderWeiblich      0.21462    0.17726   1.211    0.226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.265 on 1355 degrees of freedom
##   (289 observations deleted due to missingness)
## Multiple R-squared:  0.05526,    Adjusted R-squared:  0.05317
## F-statistic: 26.42 on 3 and 1355 DF,  p-value: < 2.2e-16
```

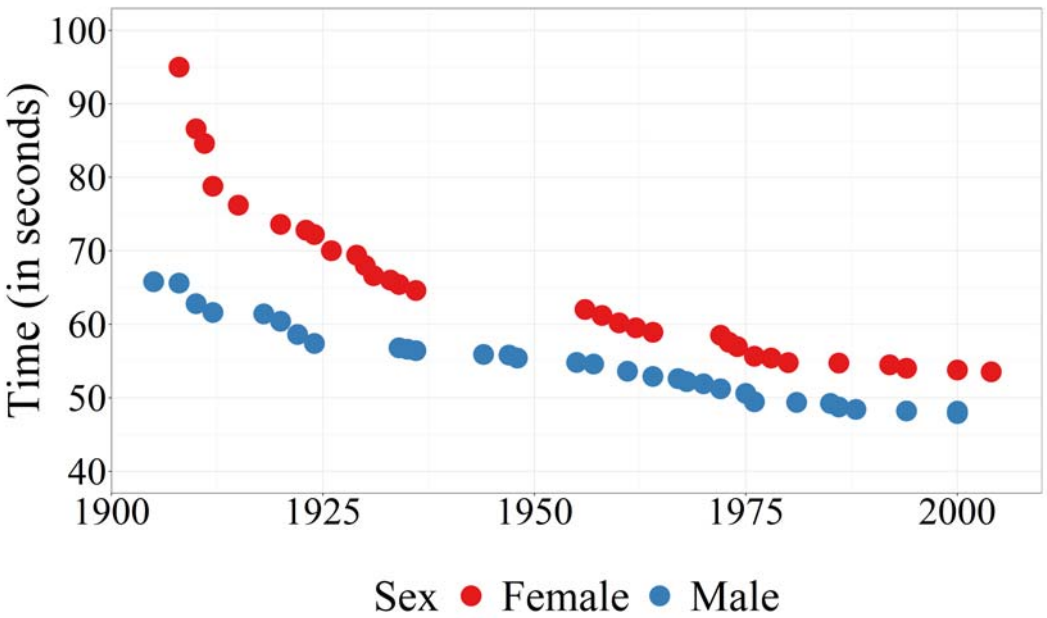
```
lm(ptv.spd ~ europe.unification + left.right + gender, data=A) %>%
  summary()
```

```
##
## Call:
## lm(formula = ptv.spd ~ europe.unification + left.right + gender,
##     data = A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5280 -2.3478  0.0403  2.7652  6.1385
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.22588    0.35116  20.577  < 2e-16 ***
## europe.unification  0.12773    0.03084   4.142 3.66e-05 ***
## left.right         -0.31746    0.04391  -7.229 8.08e-13 ***
## genderWeiblich      0.21462    0.17726   1.211    0.226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.265 on 1355 degrees of freedom
##   (289 observations deleted due to missingness)
## Multiple R-squared:  0.05526,    Adjusted R-squared:  0.05317
## F-statistic: 26.42 on 3 and 1355 DF,  p-value: < 2.2e-16
```

Modellinterpretation

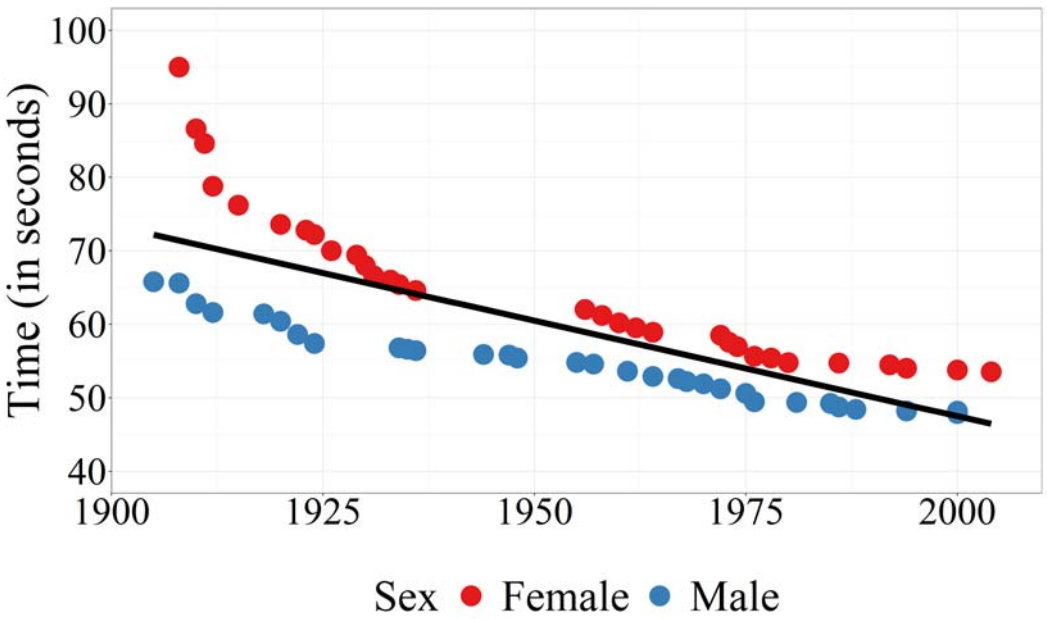
- **Beispiel:**
  - Das Modell erklärt die subj. Wahrscheinlichkeit SPD WählerIn zu sein. Die Erklärungskraft des Modells ist mit einem Bestimmtheitsmaß R^2 von 0.055 nur schwach, da nur ca. 5,5% der Varianz erklärt werden.
  - Die Zustimmung zu einer weiteren EU-Integration hängt **positiv** mit der Wahrscheinlichkeit SPD zu wählen zusammen. Mit jedem Punkt Zunahme auf der Skala EU-Integration nimmt die Wahrscheinlichkeit der SPD-Wahl um 0.12-Punkt zu. Dieser Zusammenhang ist *signifikant*.
  - Die Links/Rechts Skala hängt **negativ** mit der SPD-Wahl zusammen. Je linker eine Person eingestellt ist, desto wahrscheinlicher ist die SPD-Wahl. Mit jedem Skalenpunkt nach rechts nimmt die Wahrscheinlichkeit SPD zu wählen um 0.31-Punkte ab. Dieser Zusammenhang ist *signifikant*.
  - Der Effekt für Geschlecht ist *nicht signifikant*.

Modellbildung



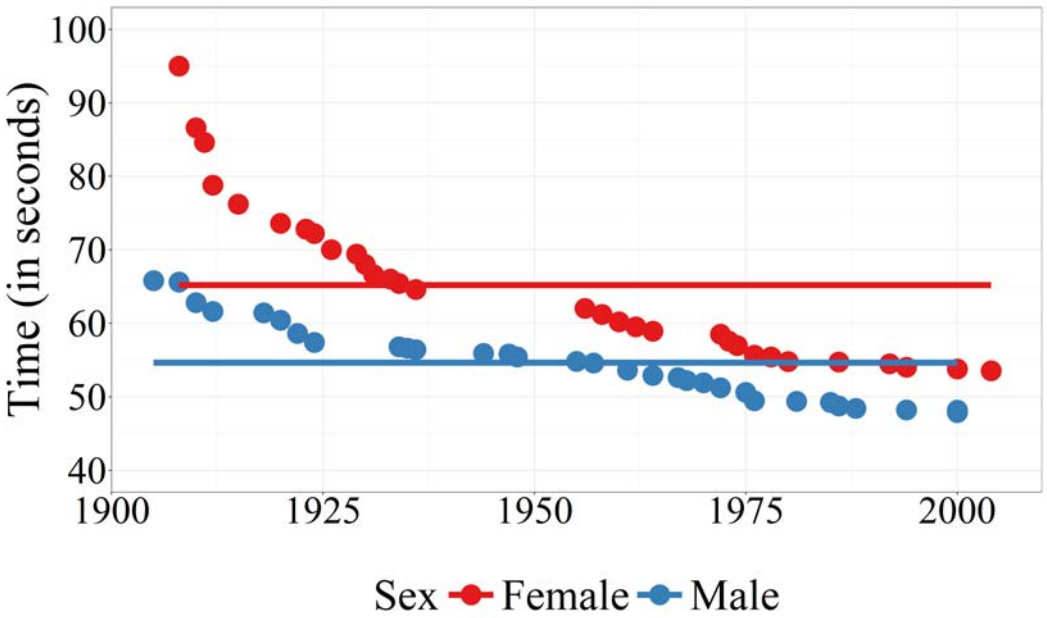
Erklärung durch den jährlichen Fortschritt

```
lm(racetime ~ year, data = Swim)
```



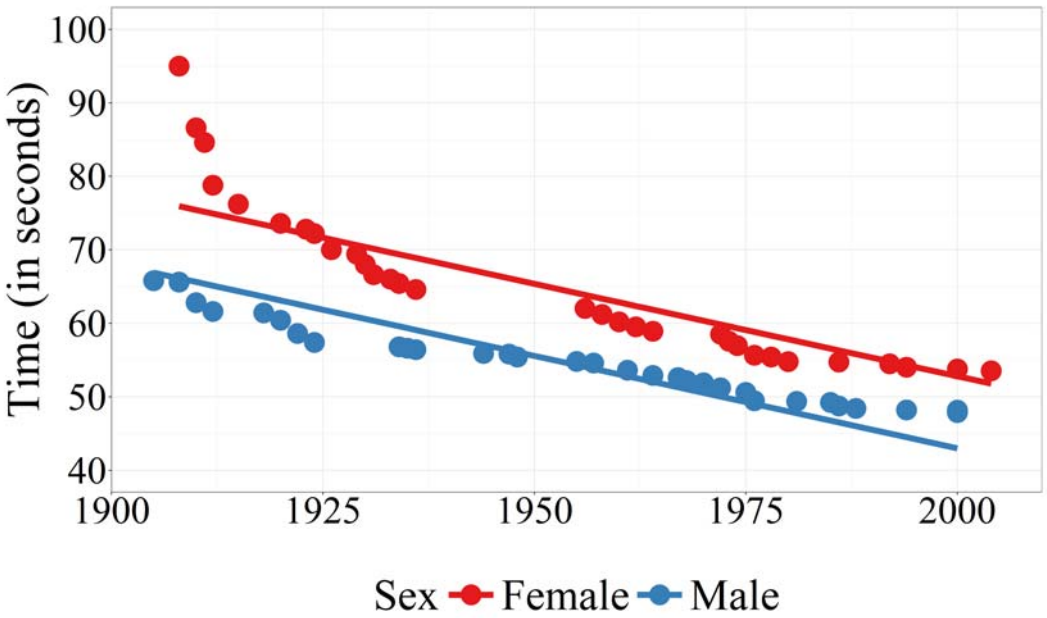
Erklärung durch das Geschlecht

```
lm(racetime ~ sex, data = Swim)
```



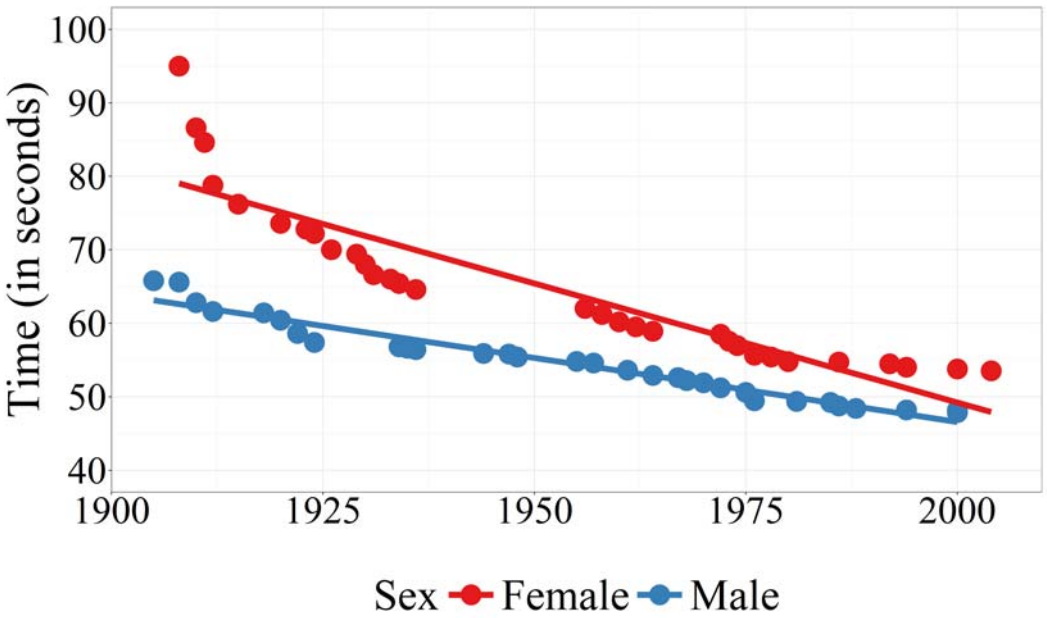
Erklärung durch Jahr und Geschlecht

```
lm(racetime ~ year + sex, data = Swim)
```



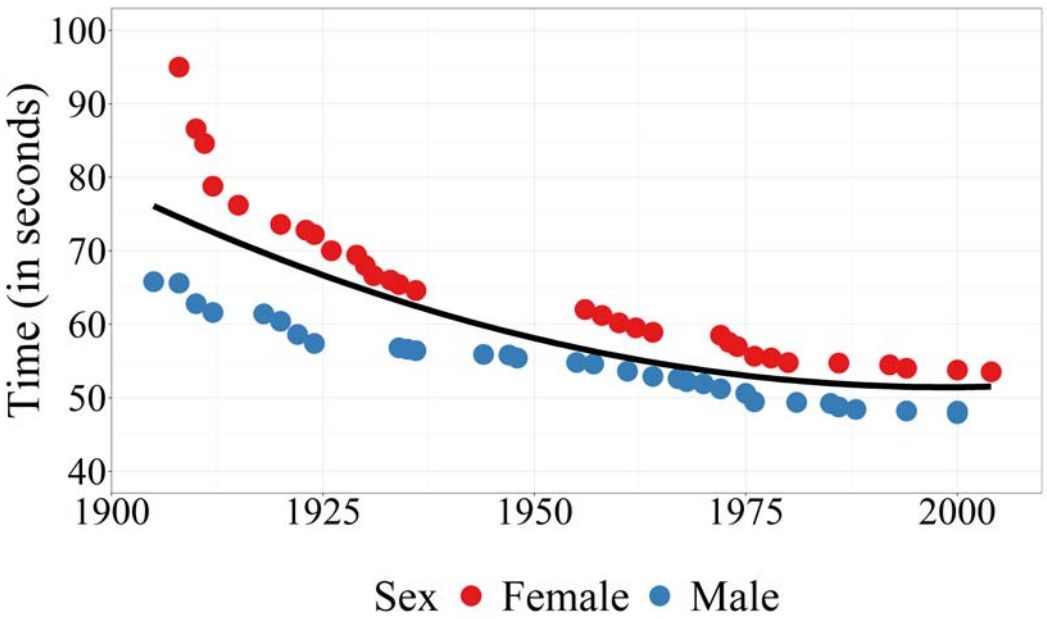
Erklärung durch Jahr und Geschlecht, als **Interaktionsterm**

```
lm(racetime ~ year + sex + year:sex, data = Swim)
```



Erklärung durch Jahr Polynom

```
lm(racetime ~ poly(year,2), data = Swim)
```



Erklärung durch Jahr Polynom und Geschlecht, beide interagiert miteinander

```
lm(racetime ~ sex:poly(year,2), data = Swim)
```