

Grundlagen  
Häufigkeiten  
Lagemaße  
Streuung  
Inferenzstatistik  
Kreuztabellen  
Gruppen-  
unterschiede  
Kovarianz/  
Korrelation  
Lineare  
Regression

# Herzlich willkommen zur Vorlesung Statistik

Zusammenhänge zwischen  
nominalen (und/oder ordinalen)  
Merkmalen:  
**Kreuztabellenanalyse und  
Assoziationsmaße**

Grundlagen  
Häufigkeiten  
Lagemaße  
Streuung  
Inferenzstatistik  
Kreuztabellen  
**Allgemeines**  
Gestaltung  
Graphiken  
Differenzen und  
Proportionen  
Chi-Quadrat-Test  
Größere Tabellen  
Assoziations-  
maße  
Gruppen-  
unterschiede  
Kovarianz/  
Korrelation  
Lineare  
Regression

## Allgemeines

- Kreuztabellen eignen sich zur Darstellung und Analyse von Zusammenhängen zwischen nominal- und ordinalskalierten Merkmalen (in allen Kombinationen).
- Metrische Merkmale (mit vielen Ausprägungen) müssen für eine Präsentation in Kreuztabellen in geeigneter Weise zusammengefasst werden.
- Dies sollte – wenn überhaupt – nur zum Zweck der übersichtlichen Präsentation von Daten geschehen, auf keinen Fall für die statistische Analyse (Informationsverlust!).

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung**
  - Graphiken
  - Differenzen und Proportionen
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Anordnung der Merkmale in der Kreuztabelle

Wenn möglich: Unabhängige (erklärende) Variable als Spaltenvariable, abhängige Variable als Zeilenvariable. Prozentuierung in den Spalten!

Beispiel (fiktiv, aber nicht ganz unrealistisch): Beruf des Vaters und Schulbesuch des Kindes (Spaltenprozent)

	Arbeiter	Angestellter	n
Hauptschule	90	45	630
RS/Gymnasium	10	55	370
n	400	600	1000

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung**
  - Graphiken
  - Differenzen und Proportionen
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Schlechte Gestaltung der Tabelle

SPSS nimmt die Sache mit den „Zellen“ einer Kreuztabelle zu wörtlich: Die Tabellen sehen aus wie ein Fenstergitter („Knast-Look“). Bitte nicht in eigene Arbeiten übernehmen (Punktabzug!).

	Arbeiter	Angestellter	gesamt
Hauptschule	360 90,0%	270 45,0%	630 66,0%
RS/Gymnasium	40 10,0%	330 55,0%	370 34,0%
gesamt	400	600	1000

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung**
  - Graphiken
  - Differenzen und Proportionen
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

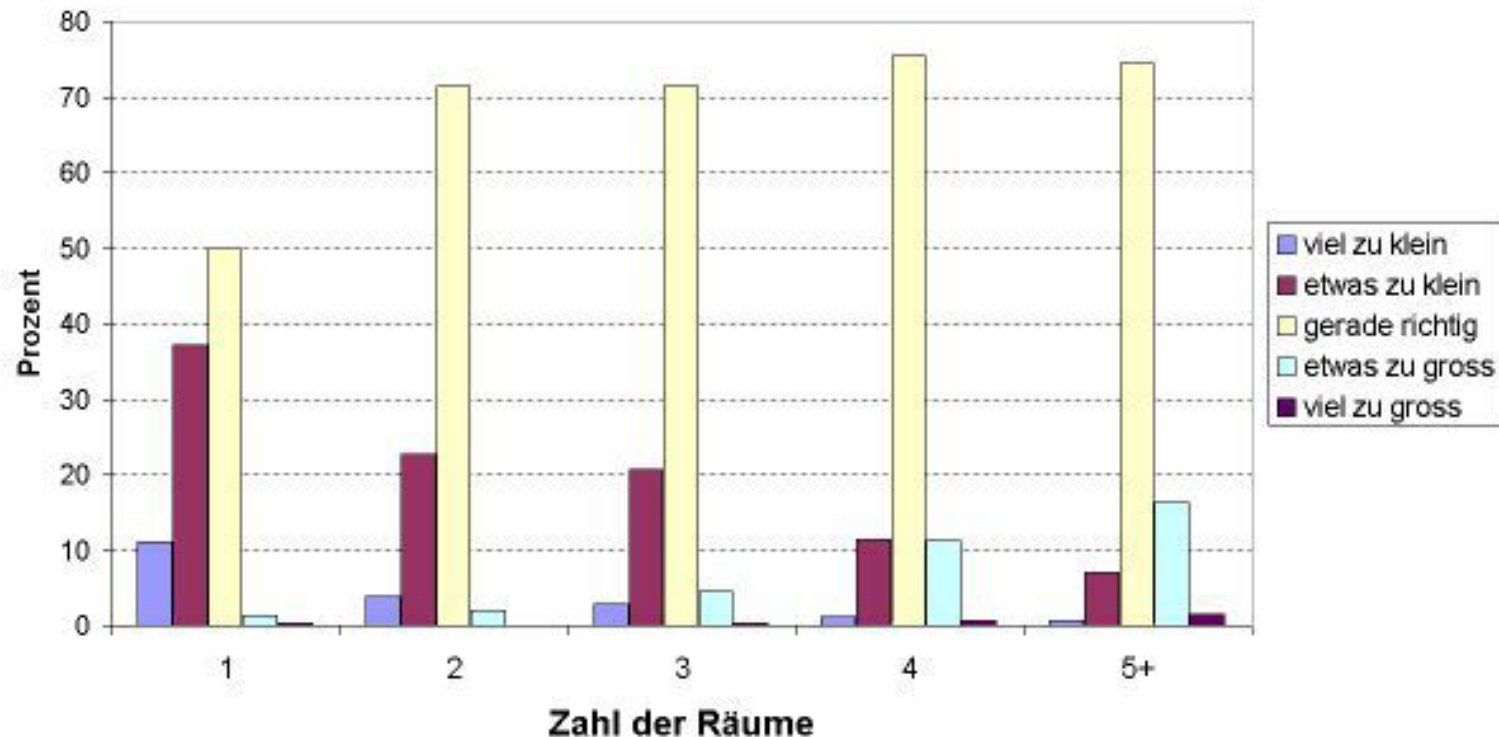
## Weitere Hinweise zur Gestaltung

- Normalerweise (vor allem bei ohnehin fehlerbehafteten Stichprobendaten) ist die Angabe der Prozentwerte mit Nachkommastellen überflüssig!
- Auch genügt es, die absoluten Zahlen in den Randverteilungen auszuweisen.
- Ebenso wenig muss das Prozentzeichen hinter jeder einzelnen Prozentzahl stehen.
- Dafür sollte möglichst in der Tabellenüberschrift angegeben sein, ob es sich um Spalten- oder Zeilenprozent (oder evtl. beides) handelt.

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken**
  - Differenzen und Proportionen
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Graphische Darstellung: Gruppiertes Säulendiagramm

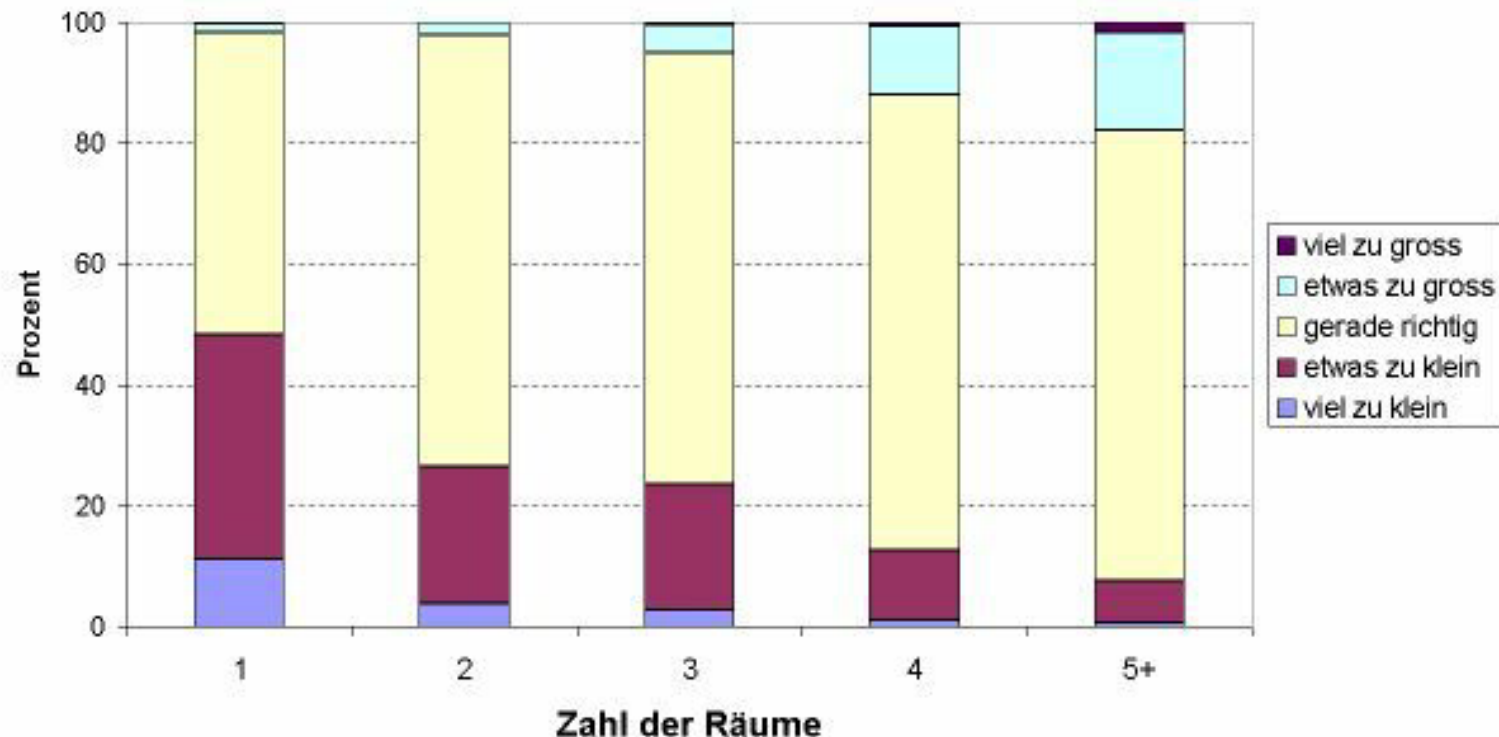
Beurteilung der Wohnungsgröße nach Zahl der Räume  
(SOEP, Welle 1997, n= 6.549)



Grundlagen  
Häufigkeiten  
Lagemaße  
Streuung  
Inferenzstatistik  
Kreuztabellen  
Allgemeines  
Gestaltung  
**Graphiken**  
Differenzen und  
Proportionen  
Chi-Quadrat-Test  
Größere Tabellen  
Assoziations-  
maße  
Gruppen-  
unterschiede  
Kovarianz/  
Korrelation  
Lineare  
Regression

## Graphische Darstellung: Gestapeltes Säulendiagramm

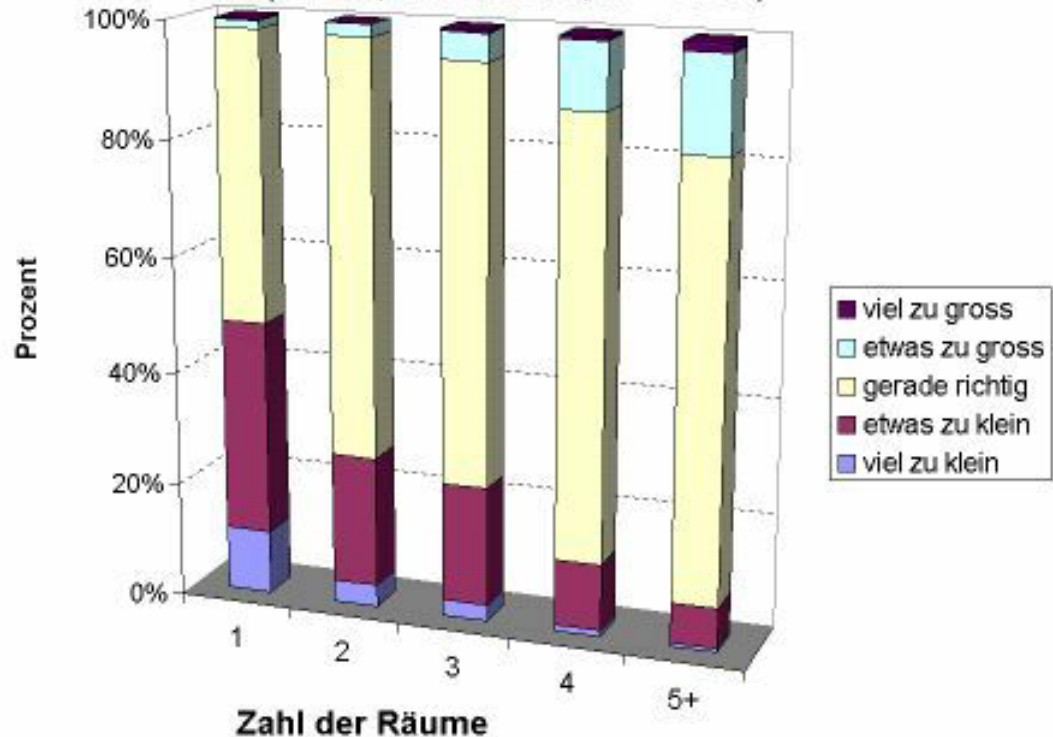
Beurteilung der Wohnungsgröße nach Zahl der Räume  
(SOEP, Welle 1997, n = 6.549)



Grundlagen  
 Häufigkeiten  
 Lagemaße  
 Streuung  
 Inferenzstatistik  
 Kreuztabellen  
 Allgemeines  
 Gestaltung  
**Graphiken**  
 Differenzen und  
 Proportionen  
 Chi-Quadrat-Test  
 Größere Tabellen  
 Assoziations-  
 maße  
 Gruppen-  
 unterschiede  
 Kovarianz/  
 Korrelation  
 Lineare  
 Regression

## Achtung: Chart Junk!

Beurteilung der Wohnungsgröße nach Zahl der Räume  
 (SOEP, Welle 1997, n = 6.549)





- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/  
Korrelation
- Lineare  
Regression

## Formale Notation zur Bezeichnung der Zellen (Beispiel: Vierfelder- Tabelle)

	X=1	X=2	$\Sigma$
Y=1	$n_{11}$	$n_{12}$	$n_{1\cdot}$
Y=2	$n_{21}$	$n_{22}$	$n_{2\cdot}$
$\Sigma$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot \cdot}$

Die Indices – hier für die Häufigkeiten angegeben – gelten in analoger Weise auch für die Anteilswerte  $p_{ij}$

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Formale Notation zur Bezeichnung der Zellen: bedingte Anteilswerte

	X=1	X=2
Y=1	$p_{1 X=1}$	$p_{1 X=2}$
Y=2	$p_{2 X=1}$	$p_{2 X=2}$

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/  
Korrelation
- Lineare  
Regression

## Anteilswerte/bedingte Anteilswerte: Hier: Anteilswerte

	X=1	X=2	$\Sigma$
Y=1	$p_{11}$	$p_{12}$	$p_{1\cdot}$
Y=2	$p_{21}$	$p_{22}$	$p_{2\cdot}$
$\Sigma$	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot \cdot}$
	X=1	X=2	$\Sigma$
Y=1	36 %	27 %	63 %
Y=2	4 %	33 %	37 %
$\Sigma$	40 %	60 %	100 %

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Anteilswerte/bedingte Anteilswerte:

### Hier: bedingte Anteilswerte

	X=1	X=2
Y=1	$p_{1 X=1}$	$p_{1 X=2}$
Y=2	$p_{2 X=1}$	$p_{2 X=2}$
	X=1	X=2
Y=1	90	45
Y=2	10	55

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen**
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Prozentsatzdifferenz

„Bei spaltenbezogener Prozentuierung kann die Differenz der Spaltenprozentwerte innerhalb einer Zeile als Maß für die Stärke des Zusammenhangs herangezogen werden.“  
(K&K, S. 317)

$$d_{XY} \% = 100 \cdot \left( \frac{n_{11}}{n_{\bullet 1}} - \frac{n_{12}}{n_{\bullet 2}} \right) = 100 \cdot (p_{1|X=1} - p_{1|X=2})$$

Spaltenprozent<sub>11</sub> – Spaltenprozent<sub>12</sub>

Gilt in analoger Weise für Zeilenprozent, wenn erklärende Variable als Zeilenvariable angeordnet ist.

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen**
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Relatives Risiko

Das relative Risiko ist das Verhältnis zweier interessierender Anteilswerte. Im Beispiel: Das Risiko von Arbeiterkindern, nur die Hauptschule zu besuchen ist mit  $90\% / 45\% = 2$  doppelt so groß wie das von Angestelltenkindern. Allgemein:

$$rr = \frac{p_{1|X=1}}{p_{1|X=2}} \quad \text{wenn die unabhängige Variable als Spaltenvariable abgetragen ist}$$

$$rr = \frac{p_{1|X=1}}{p_{2|X=1}} \quad \text{wenn die unabhängige Variable als Zeilenvariable abgetragen ist}$$

Statt der Prozentwerte in der ersten Zeile bzw. der ersten Spalte können auch die Werte aus anderen Zeilen oder Spalten zu einander in Beziehung gesetzt werden.

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen**
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/  
Korrelation
- Lineare  
Regression

## Odds Ratio (Kreuzproduktverhältnis)

Odds = „Gewinnchancen“; Ratio= Verhältnis.

Die Odds Ratio (im folgenden mit  $\gamma$  bezeichnet) ist also das Verhältnis der „Chancen“ von Arbeiterkindern auf einen Hauptschulbesuch zu den „Chancen“ von Angestelltenkindern auf einen Hauptschulbesuch.

$$\gamma(1,2 | \text{Arbeiter}) = \frac{360}{40} = 9,0$$

$$\gamma(1,2 | \text{Angestellter}) = \frac{270}{330} = 0,818$$

$$\gamma = \frac{360/40}{270/330} = \frac{360 \cdot 330}{40 \cdot 270} = \frac{9}{0,818} = 11$$

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen**
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/  
Korrelation
- Lineare  
Regression

## Odds Ratio (Kreuzproduktverhältnis)

Allgemein formuliert:

$$\gamma(1,2 | X = 1) = \frac{n_{11}}{n_{21}} = \frac{p_{1|X=1}}{p_{2|X=1}}$$

$$\gamma(1,2 | X = 2) = \frac{n_{12}}{n_{22}} = \frac{p_{1|X=2}}{p_{2|X=2}}$$

$$\gamma = \frac{n_{11} / n_{21}}{n_{12} / n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

Praktisch ist die Berechnung anhand von  $n$  vorzuziehen, da die Anteilswerte meist gerundet vorliegen.



- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen**
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Prozentsatzdifferenz, RR oder OR?

$$d_{XY} \%_{(1960)} = 35; \quad rr_{(1960)} = \frac{95}{60} = 1,58; \quad \gamma_{(1960)} = \frac{95/5}{60/40} = 12,66$$

$$d_{XY} \%_{(2000)} = 45; \quad rr_{(2000)} = \frac{90}{45} = 2; \quad \gamma_{(2000)} = \frac{90/10}{45/55} = 11$$

<i>Spaltenprozent</i>	1960		2000	
Väter:	Arbeiter	Ange-	Arbeiter	Ange-
Kinder:		stellte		stellte
Hauptschule	95	60	90	45
RS/Gymnasium	5	40	10	55

- Grundlagen
- Häufigkeiten
- Lagemaße
- Streuung
- Inferenzstatistik
- Kreuztabellen
  - Allgemeines
  - Gestaltung
  - Graphiken
  - Differenzen und Proportionen**
  - Chi-Quadrat-Test
  - Größere Tabellen
  - Assoziationsmaße
- Gruppenunterschiede
- Kovarianz/Korrelation
- Lineare Regression

## Prozentsatzdifferenz, RR oder OR?

### Erste vs. zweite Zeile (nur 2000)

$$d_{XY} \% = 45; \quad rr = \frac{90}{45} = 2; \quad \gamma = \frac{90/10}{45/55} = 11$$

$$d_{XY} \% = -45; \quad rr = \frac{10}{55} = 0.1818; \quad \gamma = \frac{10/90}{55/45} = 0,09 = 1/11$$

<i>Spaltenprozent</i>	1960		2000	
Väter:	Arbeiter	Ange-	Arbeiter	Ange-
Kinder:		stellte		stellte
Hauptschule	95	60	90	45
RS/Gymnasium	5	40	10	55