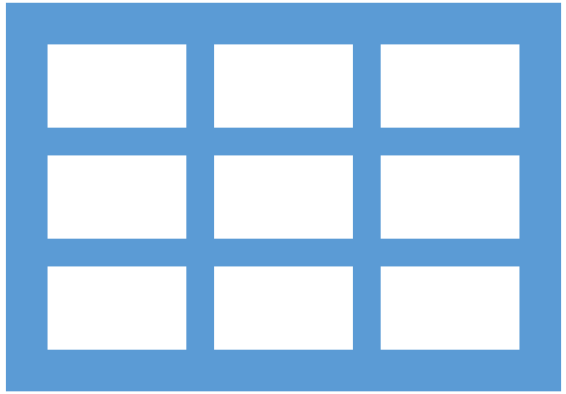
Statistik 1 – Tutorate
Sitzung 9: Tabellenanalyse

Marco Giesselmann, Norma De Min, Mara Moos, Lea Elina Hofer, Rémy Blum

Kreuztabelle mit R



1 Kreuztabellen

- Welchen Zusammenhang vermutet ihr zwischen den Merkmalen ***Geschlecht*** und ***Rauchverhalten****?* Greift auch die Zusammenhangsform (Asymmetrie?, Tendenz?) mit in der Vermutung auf.

- Ladet die Kursdaten (Achtung neue Version auf HP und OLAT!) nach R

- Inspiziert die zu den Merkmalen korrespondierenden Variablen ***geschlecht*** und ***rauchen_aktuell*** (*attributes*, *table*, *class*)

- Sinnvoll da es sich um kategoriale Variablen handelt: Faktorisierungen per ***as_factor****.*

geschlecht	**rauchen_aktuell** letzte Woche geraucht?
maennlich	0
weiblich	0
maennlich	1
weiblich	0
weiblich	1
weiblich	0
maennlich	1
weiblich	1
maennlich	0
maennlich	0
maennlich	0
maennlich	0

1 Kreuztabellen

```
[1] "letzte Woche geraucht?"

$format.stata
[1] "%12.0g"

$class
[1] "haven_labelled" "vctrs_vctr"     "double"

$labels
Keine Angabe         nein           ja
         -99            0            1


-99   0   1
  1  51  24
```

Inspektion z.B. per attributes: «rauchen_aktuell» enthält einen nicht korrekt codierten fehlenden Wert!

daher…

```
kursdata_anon$rauchen_aktuell[kursdata_anon$rauchen_aktuell==-99]<-NA
table(kursdata_anon$rauchen_aktuell)
```

Faktorisierung:

```
kursdata_anon$geschlecht <- as_factor(kursdata_anon$geschlecht)
kursdata_anon$rauchen_aktuell <- as_factor(kursdata_anon$rauchen_aktuell)
```

Achtung: Wird die Faktorisierung per «as_factor» vor der Bereinigung vorgenommen, ändert sich der Name der zu bereinigenden Kategorie *('-99'-> 'Keine Angabe')*

geschlecht	rauchen_aktuell letzte Woche geraucht?
maennlich	0
weiblich	0
maennlich	1
weiblich	0
weiblich	1
weiblich	0
maennlich	1
weiblich	1
maennlich	0
maennlich	0
maennlich	0
maennlich	0

1.1 Kreuztabellen

Über den **tab_xtab()** Befehl aus dem „**sjPlot**“ Package lassen sich Einigermassen ansehnliche Kreuztabellen erstellen.

```
tab_xtab(var.row = kursdata_anon$rauchen_aktuell,
         var.col = kursdata_anon$geschlecht,
         show.col.prc = TRUE,
         show.obs = TRUE)
```

letzte Woche geraucht?	*geschlecht*		***Total***
	maennlich	weiblich	
Keine Angabe	0 0 %	0 0 %	0 0 %
nein	15 55.6 %	36 75 %	51 68 %
ja	12 44.4 %	12 25 %	24 32 %
Total	27 100 %	48 100 %	75 100 %

Problem:

Durch die Faktorisierung über «as_factor» wurde möglicherweise ein nicht verwertetes Variablenlabel als (leere) Kategorie neu angelegt und stört nun die Darstellung von Tabellen und Abbildungen.

Löschung der Phantomkategorie durch:

```
library (forcats)
kursdata_anon$rauchen_aktuell<-fct_drop(kursdata_anon$rauchen_aktuell)
```

1.1 Kreuztabellen

Über den **tab_xtab()** Befehl aus dem „**sjPlot**“ Package lassen sich Einigermassen ansehnliche Kreuztabellen erstellen.

```
tab_xtab(var.row = kursdata_anon$rauchen_aktuell,
         var.col = kursdata_anon$geschlecht,
         show.col.prc = TRUE,
         show.obs = TRUE)
```

letzte Woche geraucht?	*geschlecht* maennlich	weiblich	***Total***
nein	15 55.6 %	36 75 %	51 68 %
ja	12 44.4 %	12 25 %	24 32 %
Total	27 100 %	48 100 %	75 100 %

- Beschreibt die einzelnen Elemente des Befehls
- Beschreibt den Tabellenaufbau
- Wie viele Befragungspersonen rauchen aktuell?
- Wie gross ist deren Anteil?
- Was sagt der Prozentwert im Feld unten links („ja“ & „männlich“) aus?
- Unterscheidet sich der Anteil aktuell Rauchender zwischen den Geschlechtern?
- Produziert eine Tabelle mit Zeilen- statt Spaltenprozenten mit dem Befehl

1.1 Kreuztabellen NEUE FOLIE

Über den **tab_xtab()** Befehl aus dem „**sjPlot**“ Package lassen sich formatierte Kreuztabellen erstellen.

```
tab_xtab(var.row = kursdata_anon$rauchen_aktuell,
         var.col = kursdata_anon$geschlecht,
         show.row.prc = TRUE,
         show.obs = TRUE)
```

- Was sagt der Prozentwert im Feld unten links („ja“ & „männlich“) ***nun*** aus?
- Was sagt der Prozentwert im Feld unten rechts („weiblich“ & „nein“) aus?
- Lässt der Differenzwert der Zellen [1;1] und [2;1] einen Rückschluss auf den Zusammenhang zwischen den beiden Variablen zu?
- Warum ist dieser Differenzwert trotzdem nicht die *richtige* Prozentsatzdifferenz des Zusammenhangs?

letzte Woche geraucht?	*geschlecht*		***Total***
	maennlich	weiblich	
nein	15 29.4 %	36 70.6 %	51 100 %
ja	12 50 %	12 50 %	24 100 %
Total	27 36 %	48 64 %	75 100 %

1.1 Kreuztabellen NEUE FOLIE

Über den **tab_xtab()** Befehl aus dem „**sjPlot**“ Package lassen sich formatierte Kreuztabellen erstellen.

```
tab_xtab(var.row = kursdata_anon$rauchen_aktuell,
         var.col = kursdata_anon$geschlecht,
         show.row.prc = TRUE,
         show.obs = TRUE)
```

- Was sagt der Prozentwert im Feld unten links („ja“ & „männlich“) ***nun*** aus?
- Was sagt der Prozentwert im Feld unten rechts („weiblich“ & „nein“) aus?
- Lässt der Differenzwert der Zellen [1;1] und [2;1] einen Rückschluss auf den Zusammenhang zwischen den beiden Variablen zu?
- Warum ist dieser Differenzwert trotzdem nicht die *richtige* Prozentsatzdifferenz des Zusammenhangs?
- Welche der beiden Tabellen veröffentlichen?

letzte Woche geraucht?	*geschlecht* maennlich	weiblich	***Total***
nein	15 29.4 %	36 70.6 %	51 100 %
ja	12 50 %	12 50 %	24 100 %
Total	27 36 %	48 64 %	75 100 %

letzte Woche geraucht?	*geschlecht* maennlich	weiblich	***Total***
nein	15 55.6 %	36 75 %	51 68 %
ja	12 44.4 %	12 25 %	24 32 %
Total	27 100 %	48 100 %	75 100 %

1.1 Kreuztabellen

Ist die Tabelle in dieser Form vollständig und publikationswürdig?

letzte Woche geraucht?	*geschlecht* maennlich	weiblich	***Total***
nein	15 55.6 %	36 75 %	51 68 %
ja	12 44.4 %	12 25 %	24 32 %
Total	27 100 %	48 100 %	75 100 %

Weitere Bearbeitungsschritte zur Publikation:

- Titel, Untertitel, Datenquelle
- Generelle Formatierungsarbeiten, Schriftgrösse?
- Kann z.T. über Suboptionen innerhalb des Befehls spezifiziert werden, grundsätzlich aber extern (z.B. Word oder Powerpoint)

Externe Weiterverarbeitung / Export:

- Die Tabelle wird automatisch im „Viewer“-Tab der R-Studio Konsole (rechts unten) angezeigt.
- Einfach per select/copy/paste in andere Dokumente bzw. Formate einfügen

1.1 Kreuztabellen

Ist die Tabelle in dieser Form vollständig und publikationswürdig?

Kreuztabelle: Rauchstatus nach Geschlecht

letzte Woche geraucht?	*Geschlecht* männlich	weiblich	***Total***
nein	15 55.6 %	36 75 %	51 68 %
ja	12 44.4 %	12 25 %	24 32 %
Total	27 100 %	48 100 %	75 100 %

Daten: Kursdatensatz 2023, n=75

Assoziationsmass Lambda?

Grafische Darstellung kreuztabellarischer Zusammenhänge

1.2 Visualisierung von Kreuztabellen

Achtung: Anders als Tabellenkommandos integrieren ggplot-Befehle Fehlende Werte (NAs) in die Darstellung. Das ist meistens schlecht – siehe HP

Daher vorab:

```
kursdata_rauchplot <- filter(kursdata_anon, !is.na(geschlecht) & !is.na(rauchen_aktuell))
```

Analysespezifischer Datensatz

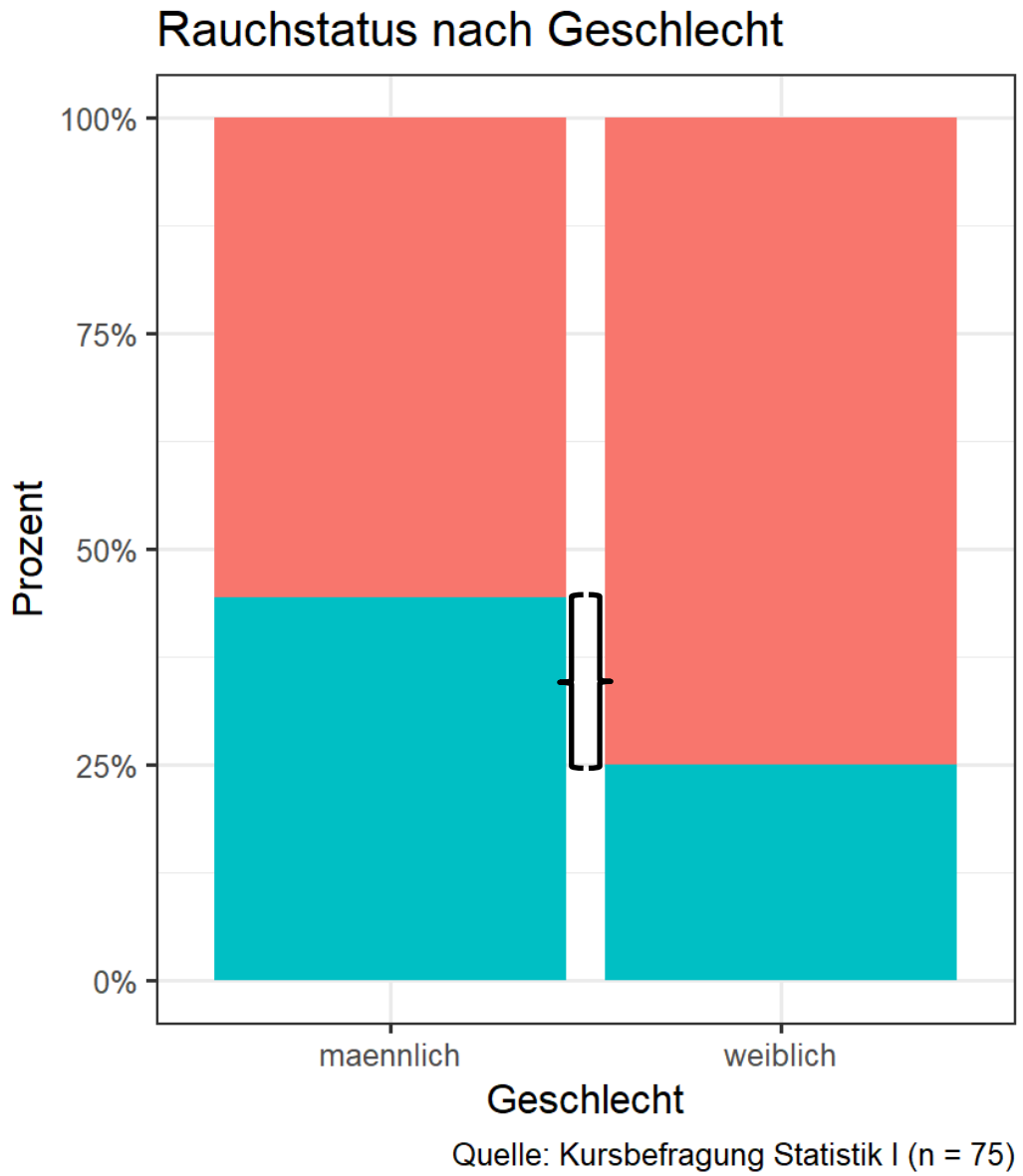
Achtung:

Funktioniert (natürlich) nur dann, wenn fehlende Werte korrekt als «NA» definiert wurden. Ggf. nochmal checken!

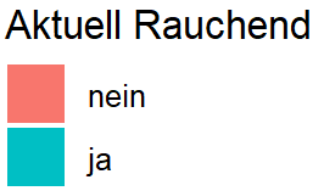
1.2 Stacked Barplot: Visualisierung gemeinsamer Verteilung

```
ggplot(kursdata_rauchplot, aes(x = geschlecht,
                               fill = rauchen_aktuell)) +
  geom_bar(position = "fill") +
  labs(title = "Rauchstatus nach Geschlecht",
       x = "Geschlecht",
       y = "Prozent",
       fill="Aktuell Rauchend",
       caption="Quelle: Kursbefragung Statistik I (n = 75)") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```

1.2 Stacked Barplot: Visualisierung gemeinsamer Verteilung

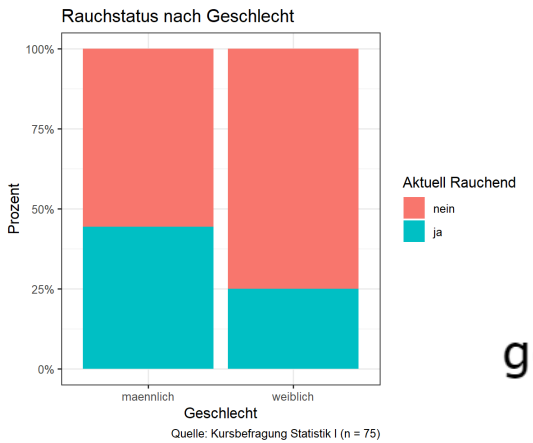



```r
ggplot(kursdata_rauchplot, aes(x = geschlecht,
                               fill = rauchen_aktuell)) +
  geom_bar(position = "fill") +
  labs(title = "Rauchstatus nach Geschlecht",
       x = "Geschlecht",
       y = "Prozent",
       fill="Aktuell Rauchend",
       caption="Quelle: Kursbefragung Statistik I (n = 75)") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```



- Wofür stehen hier jeweils die beiden Säulen?
- Repräsentieren die linken 100% gleich viele Personen wie die rechten 100%?
- Was kennzeichnet jeweils die rote Fläche?
- Wo wird in dieser Abbildung die Prozentsatzdifferenz visualisiert?

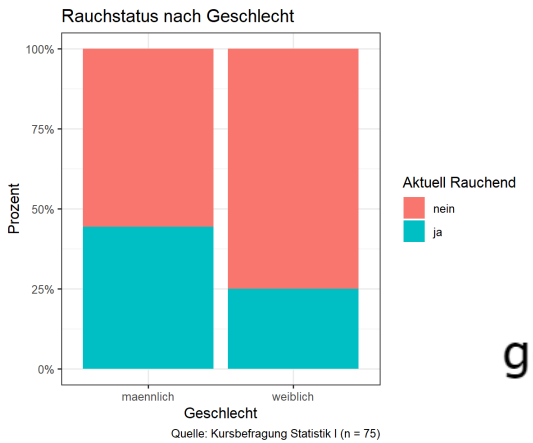
1.2 Stacked Barplot: Visualisierung gemeinsamer Verteilung



Welche Rolle spielen die rot markierten Teilbefehle? Versucht es durch Modifikation und Auslassen herauszufinden

```
ggplot(kursdata_rauchplot, aes(x = geschlecht,
                               fill = rauchen_aktuell)) +
  geom_bar(position = "fill") +
  labs(title = "Rauchstatus nach Geschlecht",
       x = "Geschlecht",
       y = "Prozent",
       fill="Aktuell Rauchend",
       caption="Quelle: Kursbefragung Statistik I (n = 75)") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```

1.2 Stacked Barplot: Visualisierung gemeinsamer Verteilung



Bilde ein Säulendiagramm ab

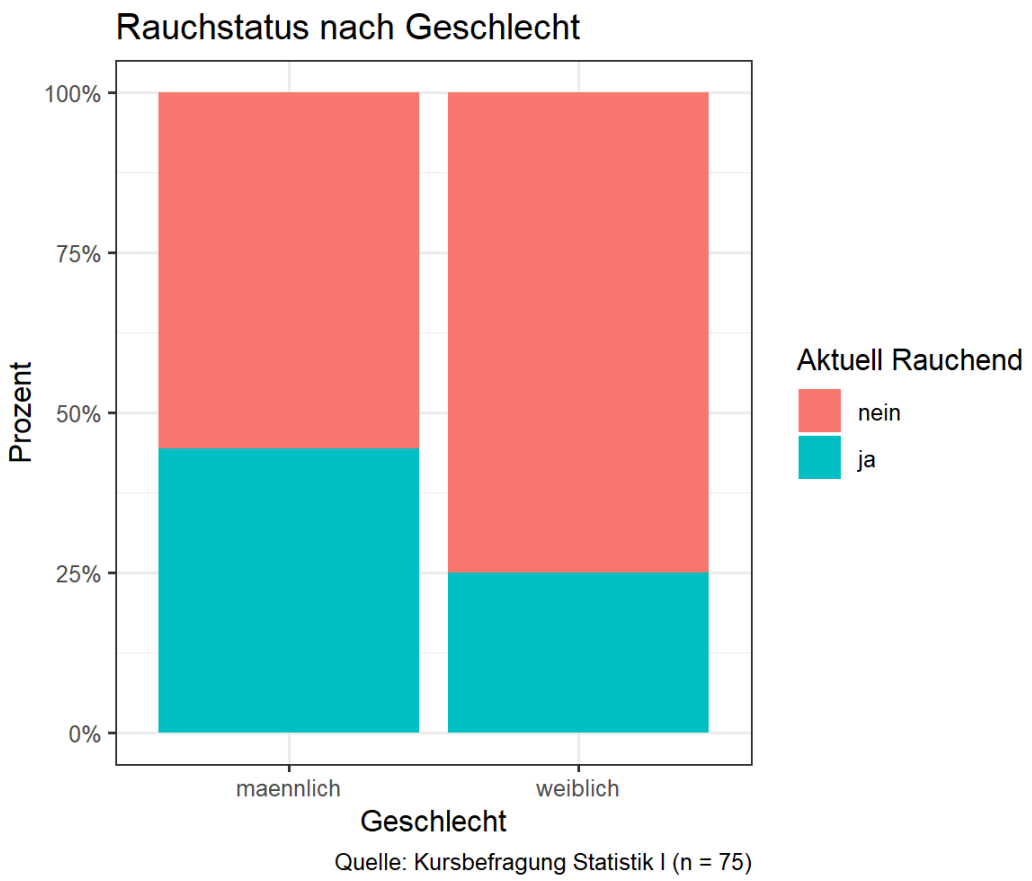
Fülle die «geoms» nicht mit einer Farbe, sondern jeweils entsprechend der Verteilung der «rauchen»-Variable

```
ggplot(kursdata_rauchplot, aes(x = geschlecht,
                               fill = rauchen_aktuell)) +
  geom_bar(position = "fill") +
  labs(title = "Rauchstatus nach Geschlecht",
       x = "Geschlecht",
       y = "Prozent",
       fill="Aktuell Raucher
       caption="Quelle: Kursbefragung Statistik I (n = 75)") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```

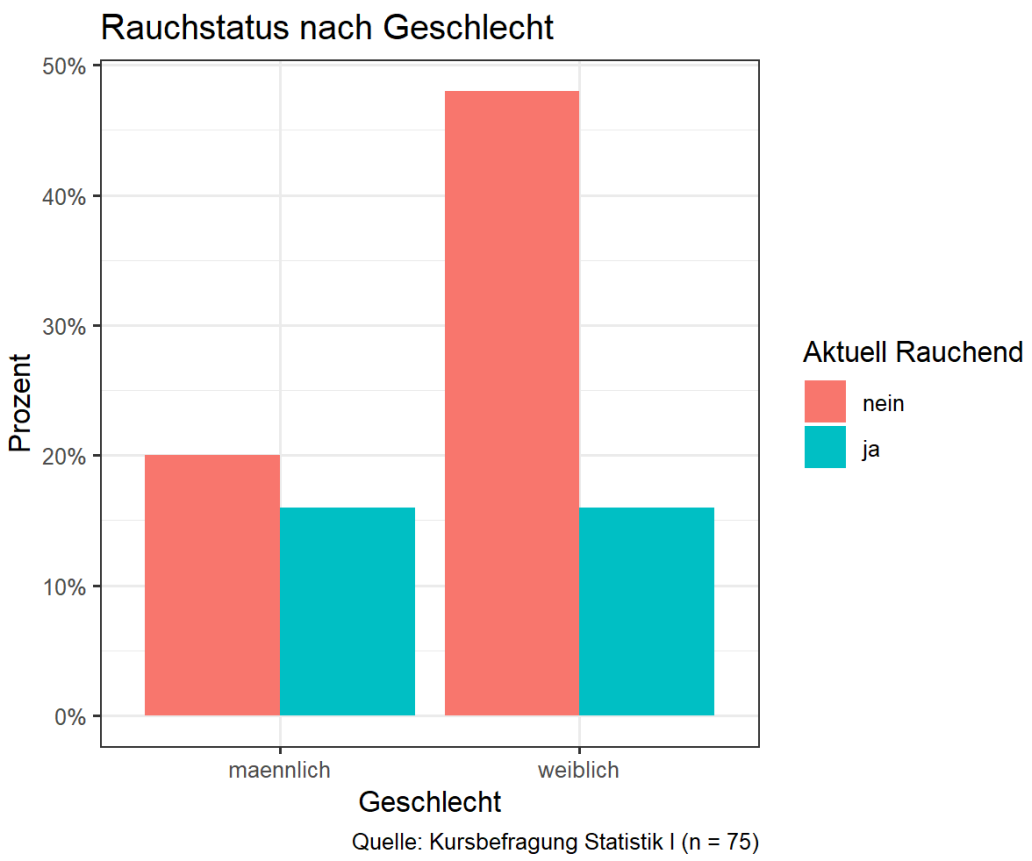
- Alle Säulen sollen die gleiche Höhe haben
- die Kategotieren der Füllvariable sollen dabei als Anteilswerte dargestellt werden

Multipliziere die y-Werte mit 100 und stelle sie mit Prozentzeichen dar

1.2 Alternative „Dodge“-Plot – Unterschiede in der Darstellung?

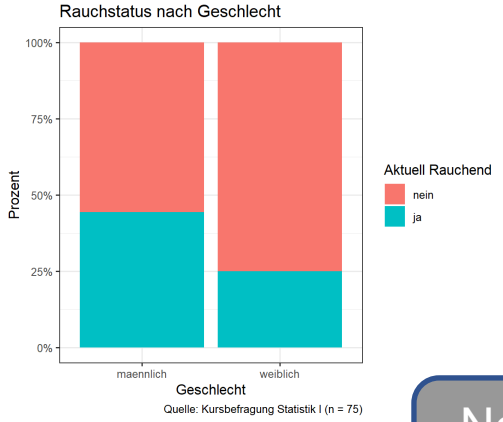



```
ggplot(kursdata_rauchplot, aes(x = geschlecht,
                               fill = rauchen_aktuell)) +
  geom_bar(position = "fill") +
  labs(title = "Rauchstatus nach Geschlecht",
       x = "Geschlecht",
       y = "Prozent",
       fill="Aktuell Rauchend",
       caption="Quelle: Kursbefragung Statistik I (n = 75)") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```



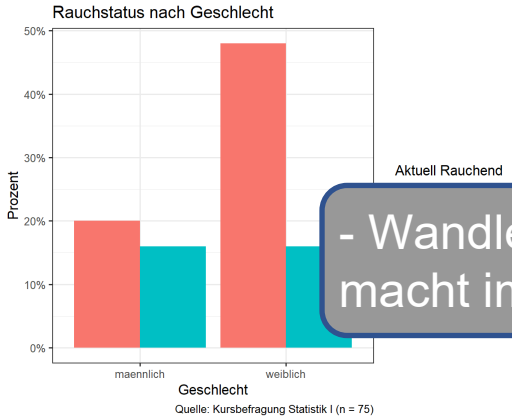

```
ggplot(kursdata_rauchplot, aes(x = geschlecht,
                               fill = rauchen_aktuell)) +
  geom_bar(position="dodge") +
  aes(y = after_stat(count / sum(count)))+
  labs(title = "Rauchstatus nach Geschlecht",
       x = "Geschlecht",
       y="Prozent",
       fill="Aktuell Rauchend",
       caption="Quelle: Kursbefragung Statistik I (n = 75)") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```

1.2 Alternative „Dodge“-Plot – Unterschiede im Code?




```
ggplot(kursdata_rauchplot, aes(x = geschlecht,
                               fill = rauchen_aktuell)) +
  geom_bar(position = "fill") +
  labs(title = "Rauchstatus nach Geschlecht",
       x = "Geschlecht",
       y = "Prozent",
       fill="Aktuell Rauchend",
       caption="Quelle: Kursbefragung Statistik I (n = 75)") +
                                        ::percent) +
```

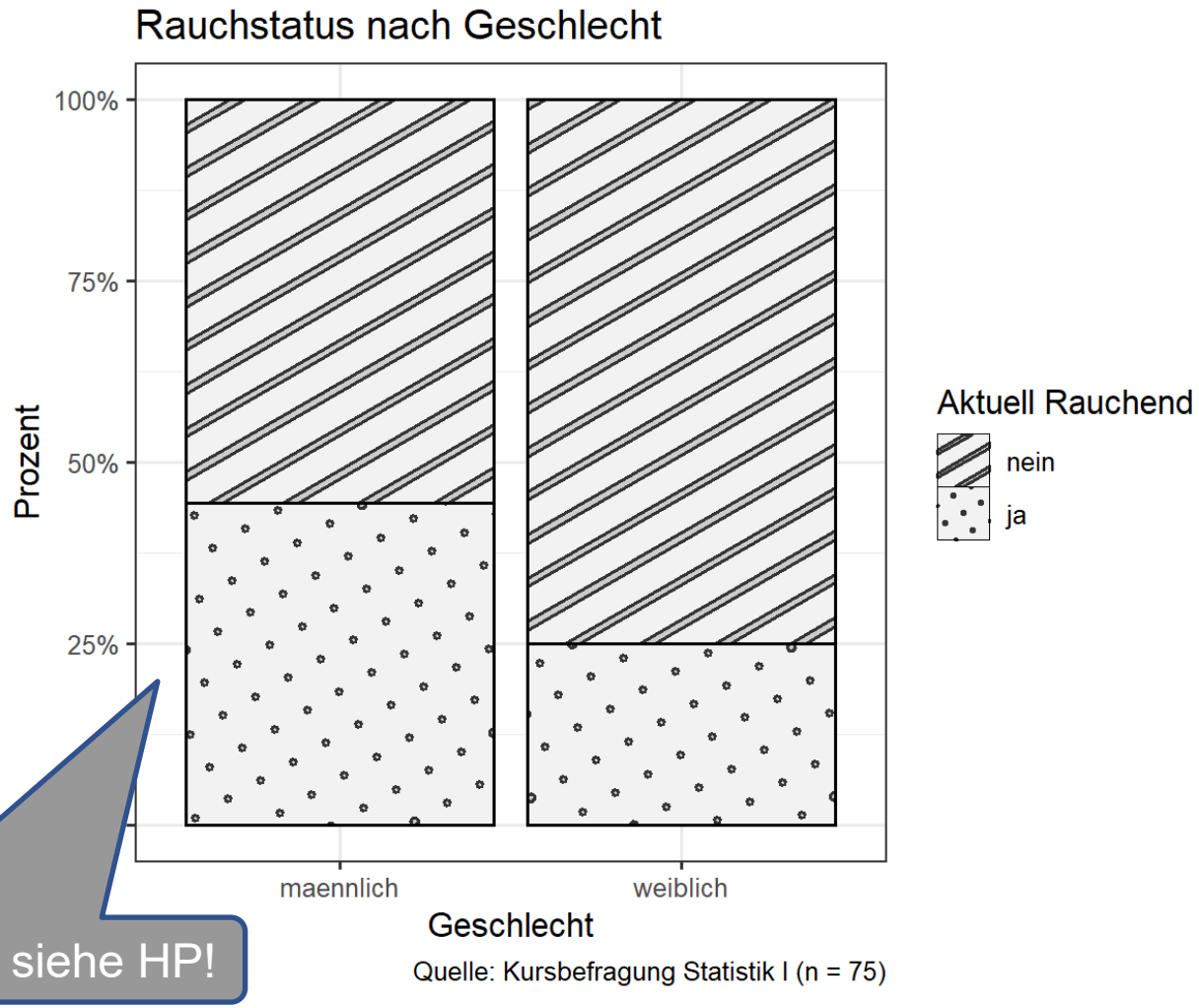
- Nehme die «Geschlechter»-Balken jeweils nach Kategorien der «fill»-Variable auseinander.




```
ggplot(kursdata_ra      t, aes(x = geschlecht,
                               fill = rauchen_aktuell)) +
  geom_bar(position="dodge") +
  aes(y = after_stat(count / sum(count)))+
  labs(title = "Rauchstatus nach Geschlecht",
       x = "Geschlech
                                                k I (n = 75)") +
  scale_y_continuous(labels = scales::percent) +
  theme_bw()
```

- Wandle absolute Häufigkeiten in Anteilswerte um (dies macht im Befehl oben «position=«fill» per Default)

1.2 Stacked Barplot: Was tun wenn Schwarz/Weiss Abbildungen gefordert sind?



Zusatzpackage «ggpattern», siehe HP!

2. Kreuztabelle: Weiteres Beispiel aus der Kursbefragung

Wir wollen prüfen, in welchem Zusammenhang die *Elterliche Bildung* und das *allgemeine Vertrauen* innerhalb des Kurses stehen. Dazu verwenden wir die Variablen **akback und** zusätzlich **trustkat**.

I. Inspiziert die Variable **trustkat**. In welchem Verhältnis steht diese zur (Originalvariable) **trust**?

II. Formuliert und begründet eine **Hypothese** zu den beiden Variablen

III. Erstellt eine **Kreuztabelle** welche die gemeinsame Verteilung der beiden Variablen sinnvoll (im Sinne der formulierten Hypothese) abbildet.

IV. Wertet die Tabelle in einem inhaltlich gehaltvollen Antwortsatz aus (**Prozentsatzdifferenz**!).

V. Visualisiert den Zusammenhang

VI. Stützt Eure Auswertung durch Berechnung und Bericht des Assoziationsmasses Lambda

VII. Stützt Eure Auswertung durch Berechnung und Bericht eines Chi-Quadrat basierten Korrelationsmasses

VIII. Stützt Eure Auswertung durch Einbindung der Test-Statistik des Chi-Quadrat Tests

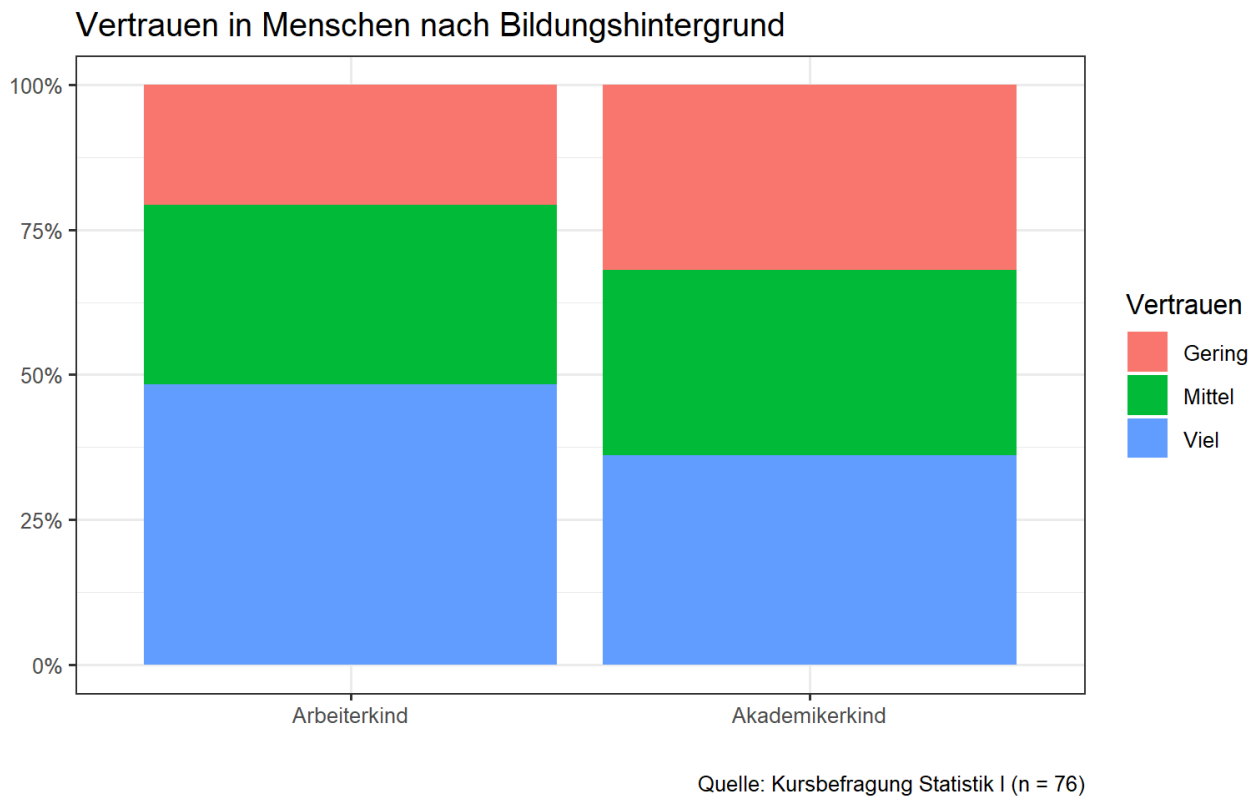
2 *trust und trustkat*

id	trust Kann man Menschen im Allg. vertrauen? (5-volle Zustimmung, 1-v...	trustkat Allg. Vertrauen (kat.)
77	4	Viel
76	2	Gering
78	5	Viel
49	3	Mittel
79	3	Mittel
81	2	Gering
82	3	Mittel
80	1	Gering
86	3	Mittel

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %

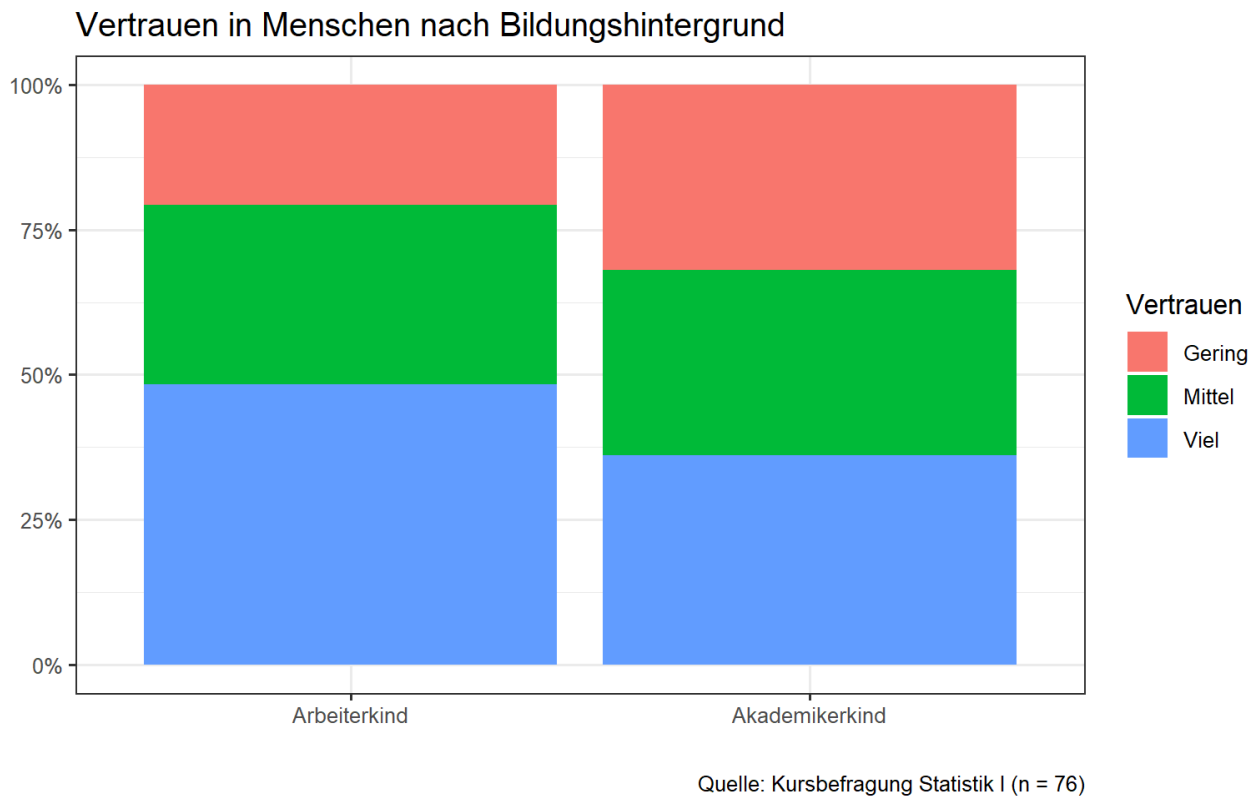


Auswertung:

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %



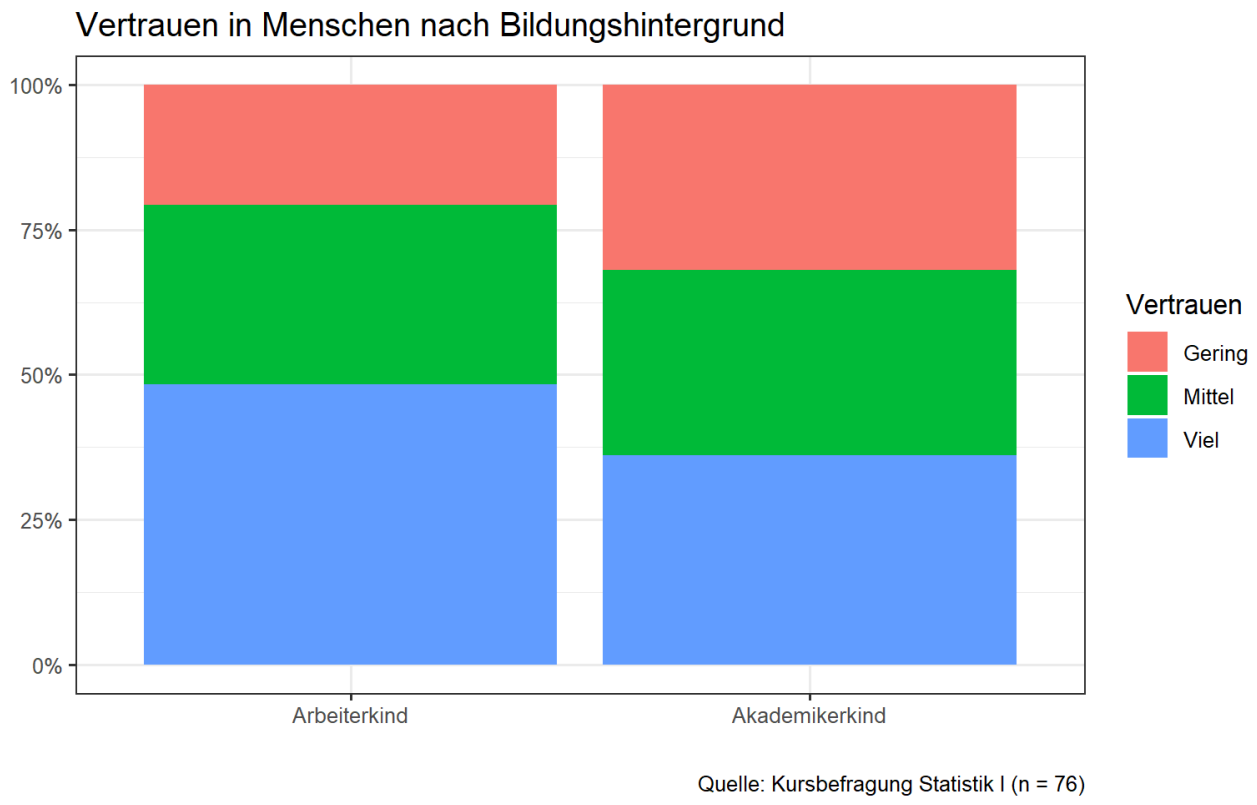
Auswertung:

Der Anteil vertrauensvoller Personen ist unter Arbeiterkindern gut 12 ppt. grösser als unter Akademikerkindern. Unter Akademikerkindern ist dagegen der Anteil der Personen mit geringem Vertrauen etwa 11 ppt. grösser als unter Akademikerkindern. Der dargelegte Zusammenhang lässt sich *deskriptiv* im Sinne der Hypothese interpretieren, dass Arbeiterkinder vertrauensvoller sind als Akademikerkinder.

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %

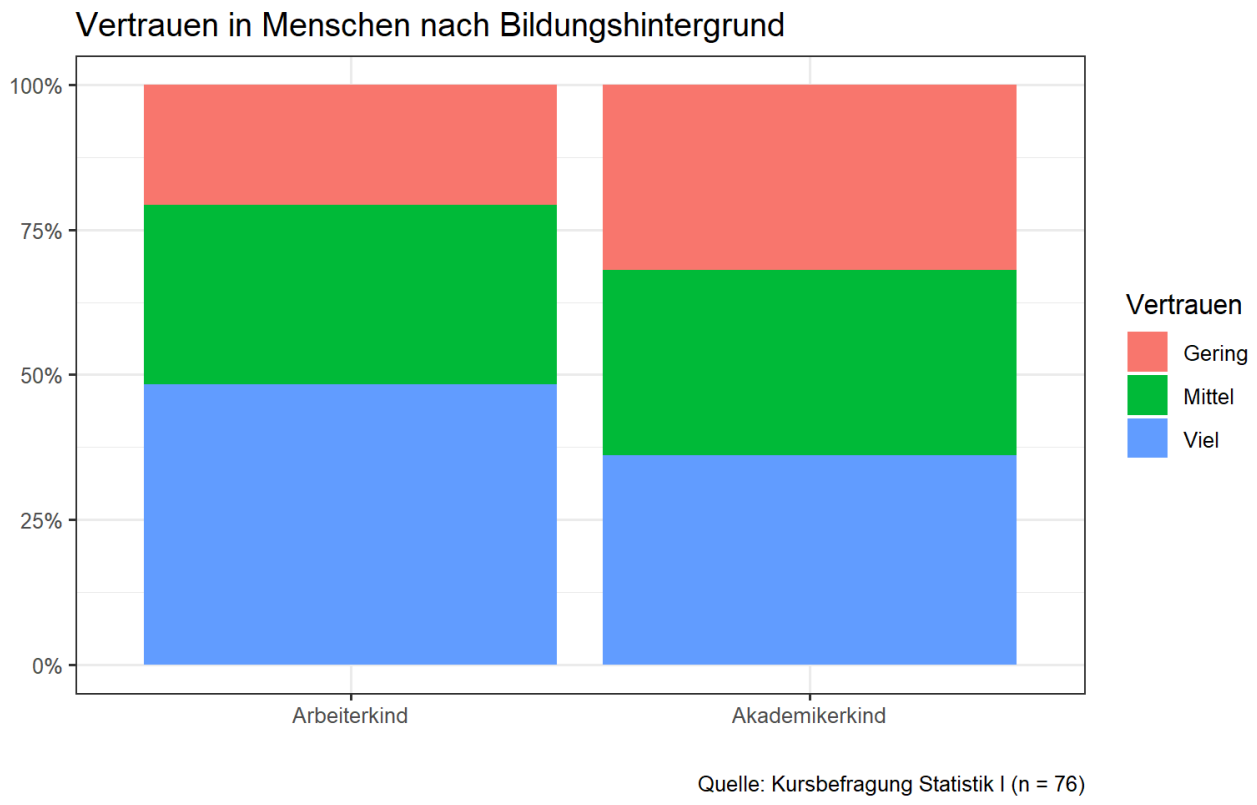


Lambda:?

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %

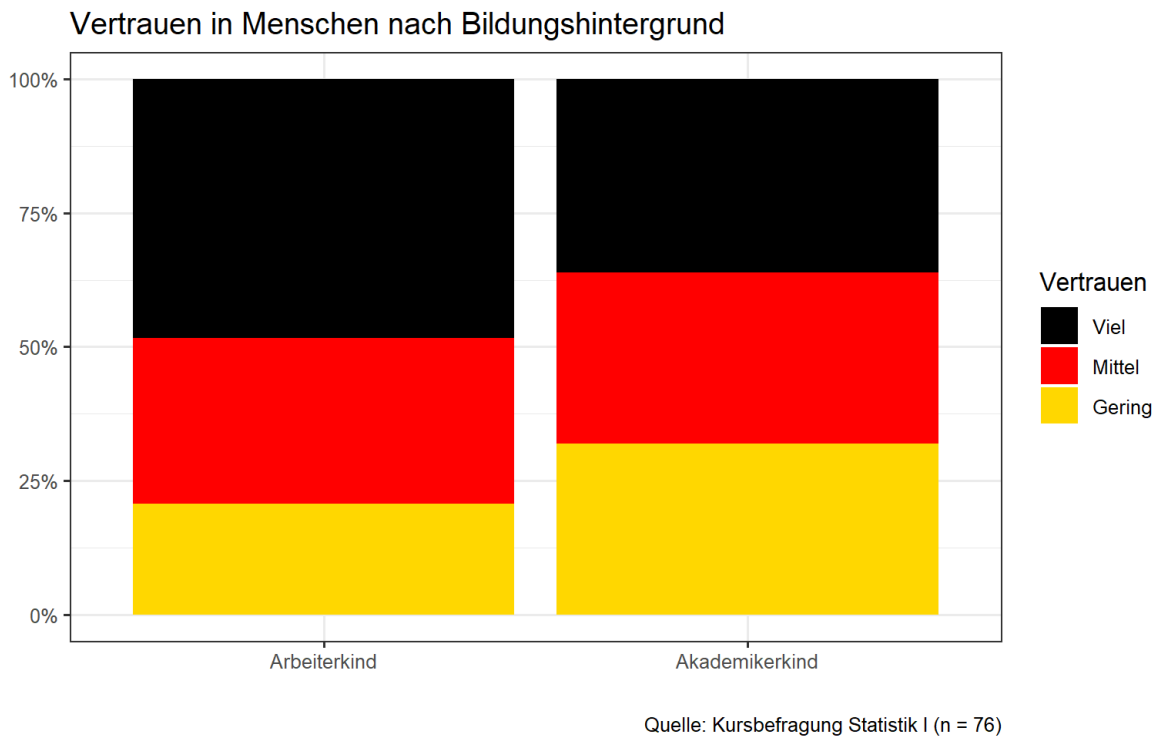


Lambda:?

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %



Zusatzaufgabe:

- Ordnet die Kategorien in der Abbildung neu und intuitiver: Niedriges Vertrauen unten, hohes Vertrauen oben

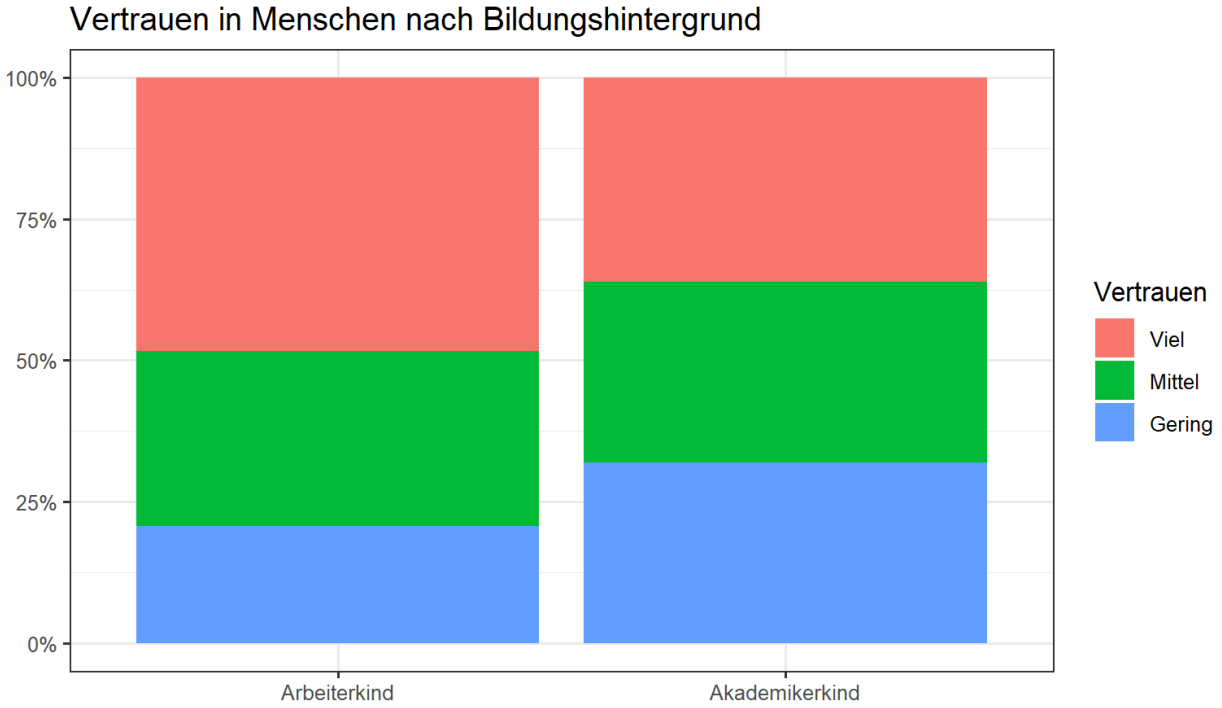
- Eventuell per ChatGPT: Ändert die Farben

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %

$\chi^2=1.467 \cdot df=2 \cdot Cramer's\ V=0.139 \cdot p=0.480$



Quelle: Kursbefragung Statistik I (n = 76)

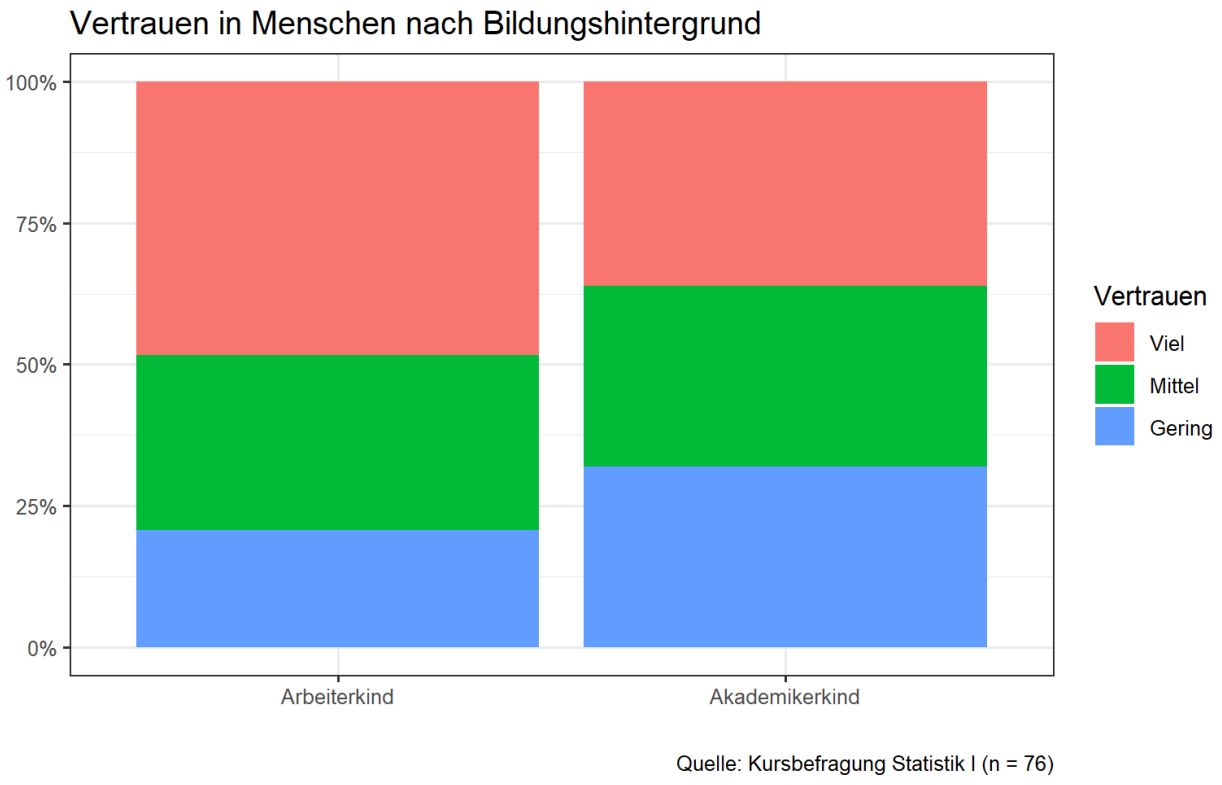
Cramer's V:

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %

$\chi^2=1.467 \cdot df=2 \cdot Cramer's\ V=0.139 \cdot p=0.480$



Cramer's V:

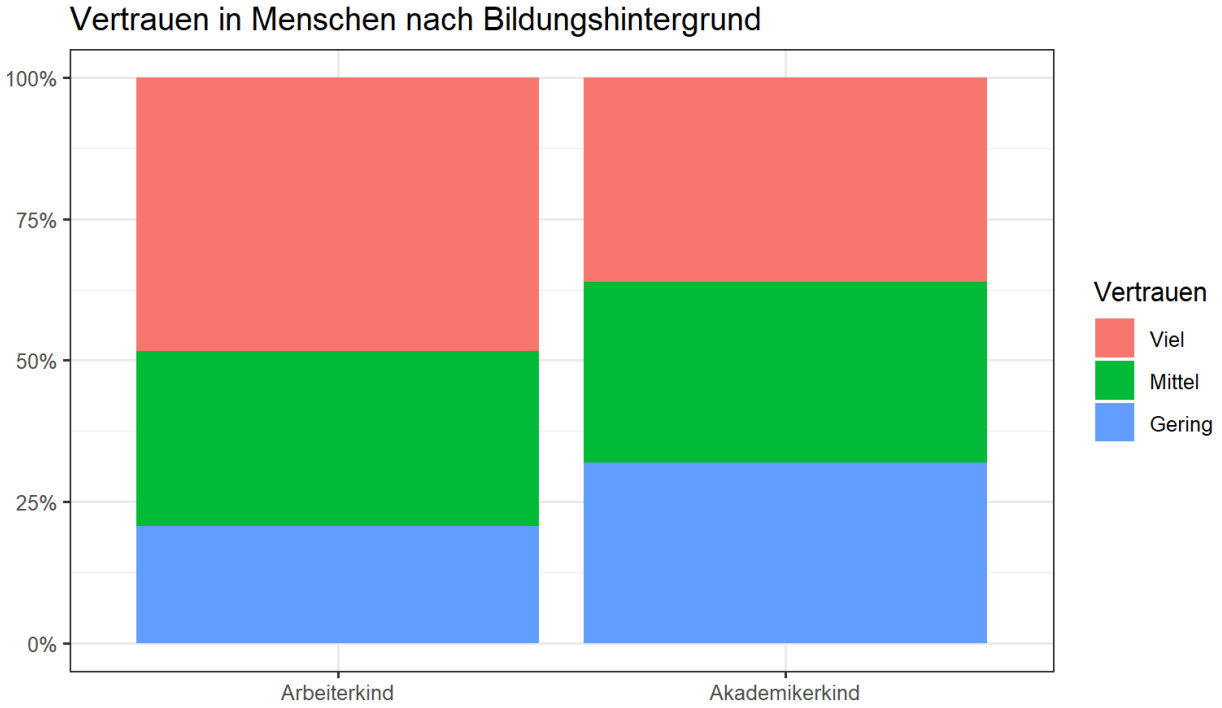
Cramer's V (0,14) drückt eine bestehende Abhängigkeit zwischen den beiden Merkmalen aus, die sich nach gängigen Klassifikationen (siehe Vorlesung) als schwacher Zusammenhang deuten lässt.

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %

$\chi^2=1.467 \cdot df=2 \cdot Cramer's\ V=0.139 \cdot p=0.480$



Quelle: Kursbefragung Statistik I (n = 76)

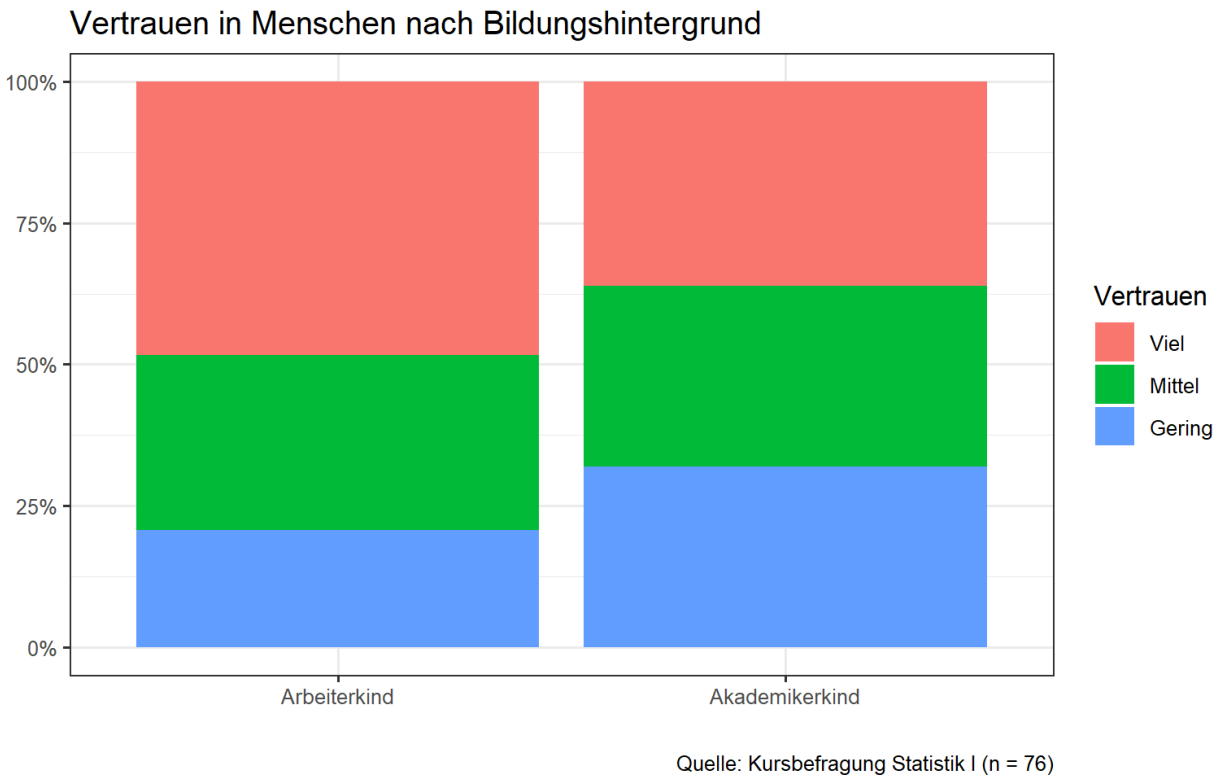
Inferenzstatistische Hypothesenbewertung:

2 Kreuztabelle und Abbildung

Kreuztabelle: Vertrauen nach Bildungshintergrund

Vertrauen in Mitmenschen	*Akademikerkind?* nein	ja	***Total***
Gering	6 20.7 %	15 31.9 %	21 27.6 %
Mittel	9 31 %	15 31.9 %	24 31.6 %
Viel	14 48.3 %	17 36.2 %	31 40.8 %
Total	29 100 %	47 100 %	76 100 %

$\chi^2=1.467 \cdot df=2 \cdot Cramer's\ V=0.139 \cdot p=0.480$



Inferenzstatistische Hypothesenbewertung:

Die Nullhypothese, dass **in der Population** Unabhängigkeit zwischen dem Bildungshintergrund und dem Vertrauen besteht, kann auf Basis des Stichprobenergebnisses nicht abgelehnt werden (chi2=1.5, p>0,05). Gleichwohl ist die der Analyse zugrunde liegende, einseitige Hypothese mit der nominalen Logik des Chi-Quadrat Unabhängigkeitstest nicht vereinbar und somit auch nicht exakt testbar. Dazu müsste die Tabelle zunächst auf 2*2 Felder vereinfacht werden.

Weitere Übung

- http://www.suz.uzh.ch/dataforstat/