05 Daten zusammenfassen

Einführung in die quantitativen Forschungsmethoden

Heute

- Quiz zu letzter Sitzung
- Heutige Variablen
- Mittelwerte
- Verteilungen
- Zusammenhänge zwischen Variablen
- Übung an den ESS Daten

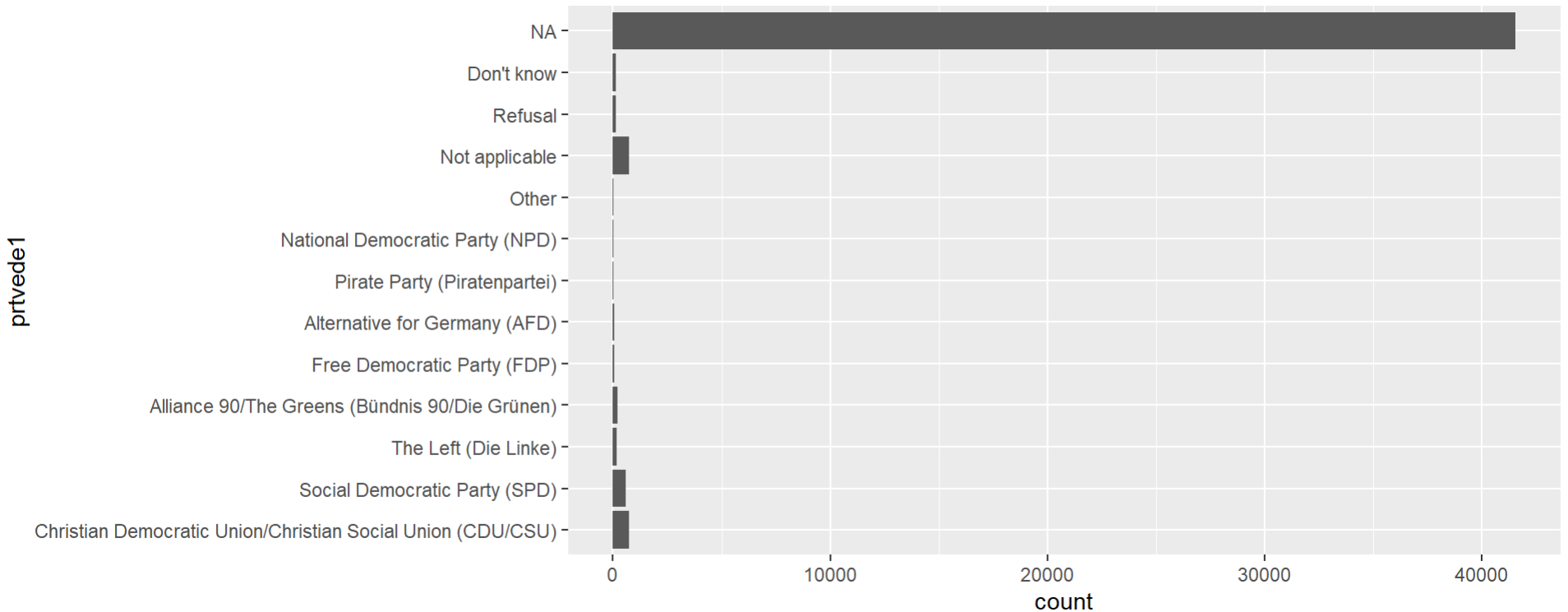
Quiz zu letzter Sitzung

Quiz Link

Heutige Variablen

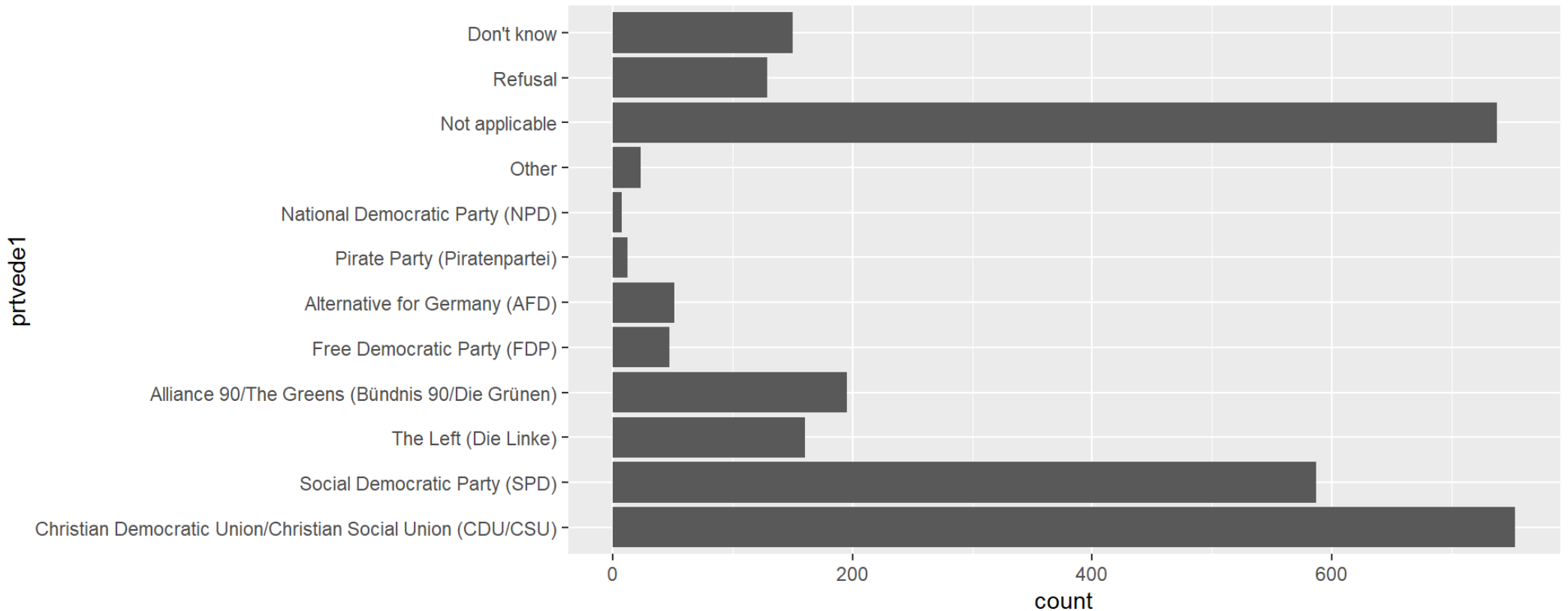
prtvede1

“Bei der Bundestagswahl konnten Sie ja zwei Stimmen vergeben. Die Erststimme für einen Kandidaten aus Ihrem Wahlkreis, die Zweitstimme für eine Partei. Welchem Kandidaten haben Sie Ihre Erststimme gegeben?”



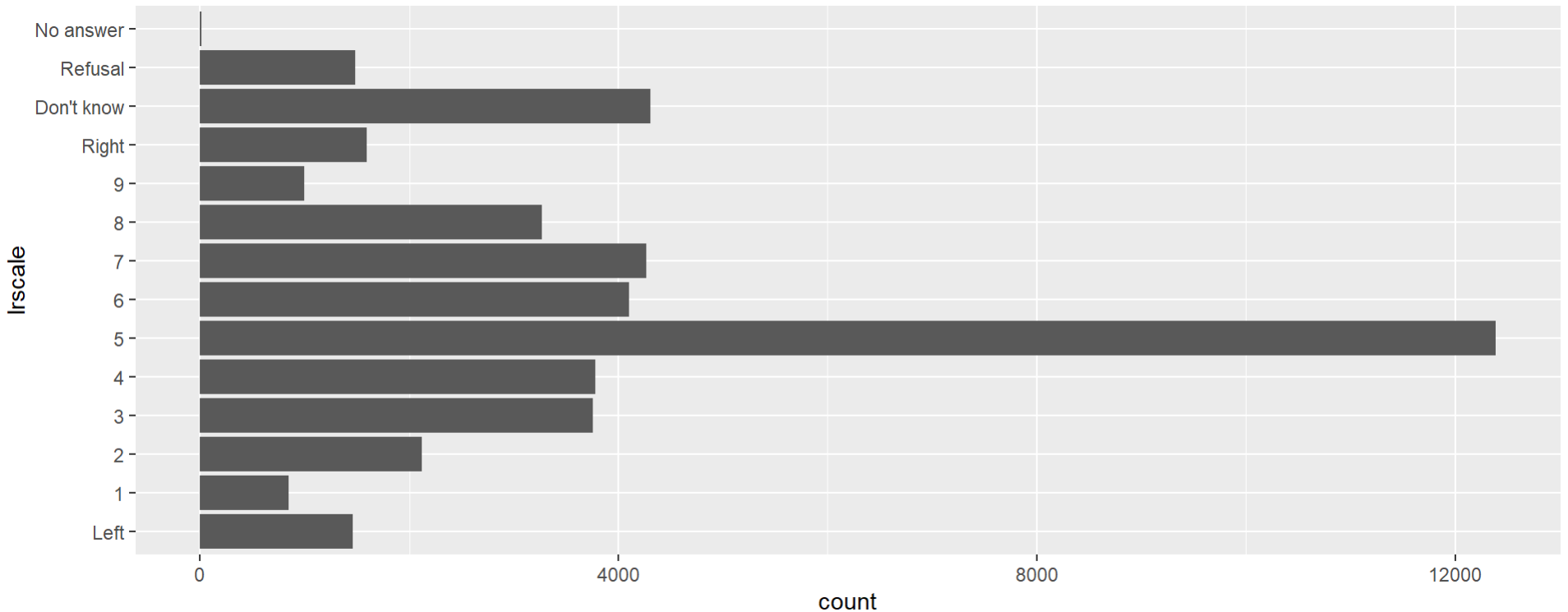
prtvede1 (DE)

“Bei der Bundestagswahl konnten Sie ja zwei Stimmen vergeben. Die Erststimme für einen Kandidaten aus Ihrem Wahlkreis, die Zweitstimme für eine Partei. Welchem Kandidaten haben Sie Ihre Erststimme gegeben?”



lrscale

In der Politik spricht man manchmal von ‘links’ und ‘rechts’. Wo auf dieser Skala würden Sie sich selbst einstufen, wenn 0 für links steht und 10 für rechts?



lrscale

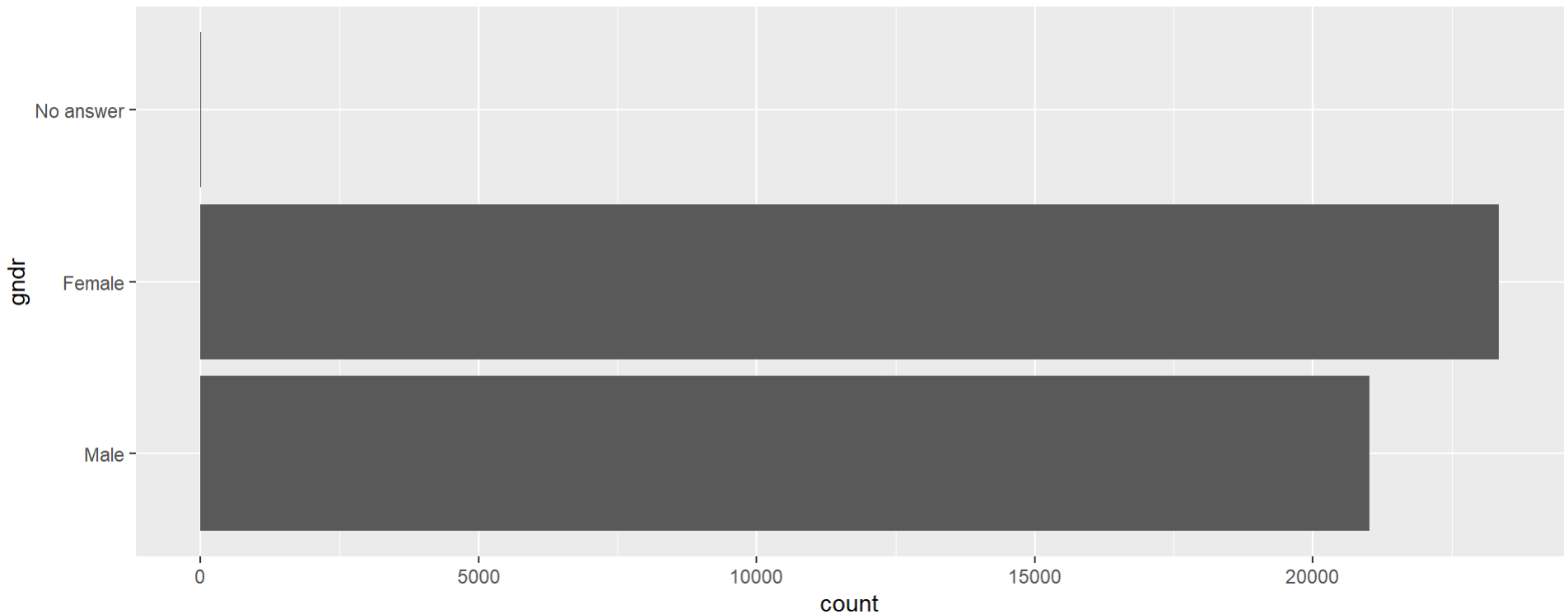
- sog. Likert-Skalen wie die Links-Rechts Skala werden manchmal als metrisch, manchmal als ordinal betrachtet
  - metrisch: diskrete Zahlen
  - ordinal: Bedeutung

→ Wir sprechen am Ende des Semesters noch einmal über Vor- und Nachteile

→ auch in R können wir die Variable `as.numeric()` oder `haven::as_factor()` behandeln

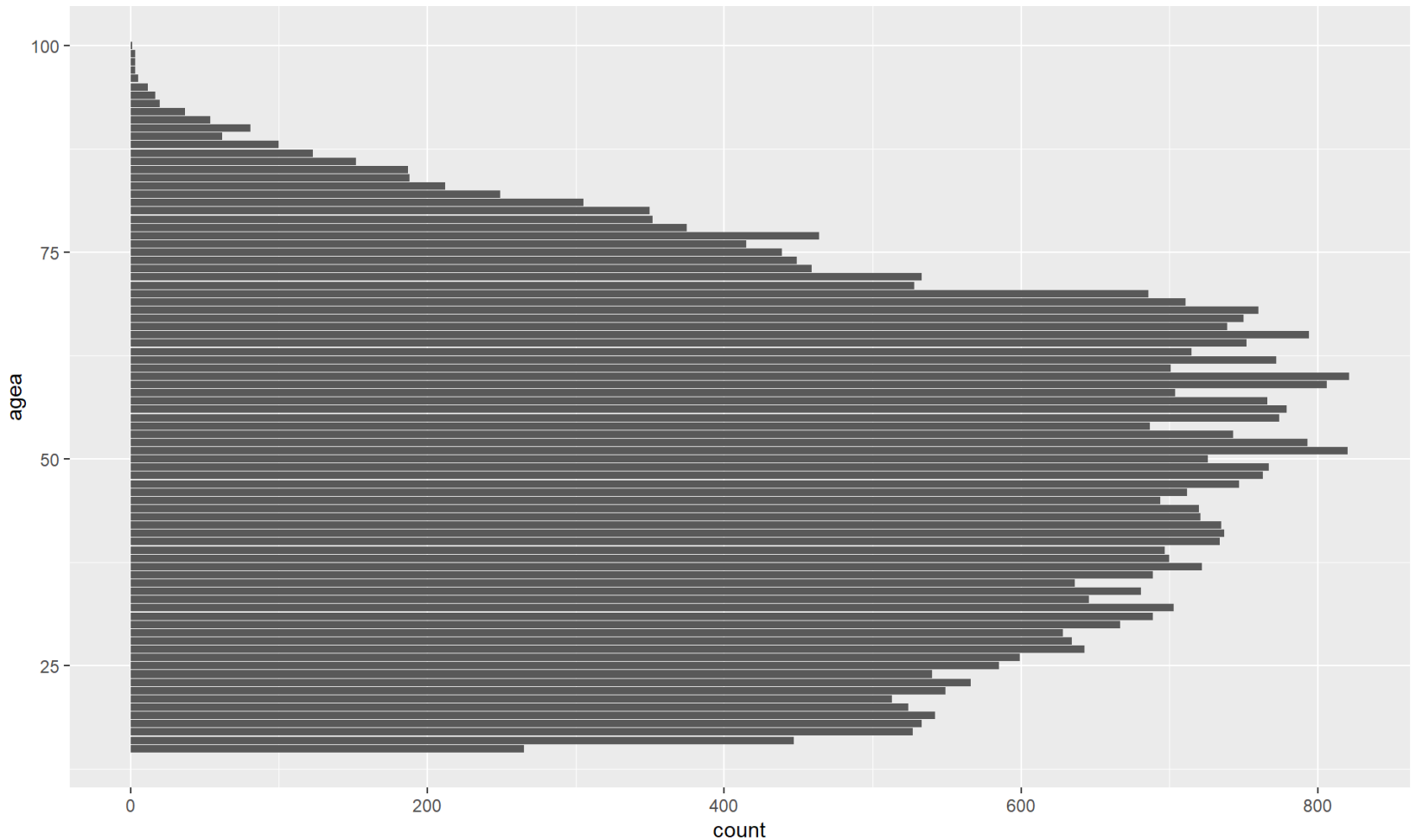
gndr

Welches Geschlecht haben Sie?



agea

Alter, kalkuliert aus Geburtsjahr



Unsere Umfrage

Ich arbeite auf den Folien - um das ganze übersichtlicher zu gestalten - mit den Ergebnissen unserer Umfrage aus der ersten Sitzung.

Mittelwerte

Mittelwerte

→ Wie können wir einen charakteristischen Wert für eine Verteilung angeben?

z.B. für Wahlpräferenz, Geschlecht & links-rechts-Einordnung?

→ Und was ist eigentlich ein Mittelwert?

Modus

Was ist der häufigste Wert?

```
1 head(kurs$prtvede1, n=21)
```

```
 [1] NA                       "SPD"                    "Die Linke"
 [4] "Bündnis 90/Die Grünen" "Bündnis 90/Die Grünen" "Die Linke"
 [7] "Bündnis 90/Die Grünen" "Bündnis 90/Die Grünen" "Die Linke"
[10] "CDU/CSU"               "Bündnis 90/Die Grünen" "Die Linke"
[13] "SPD"                   "FDP"                    "Die Linke"
[16] "Bündnis 90/Die Grünen" "Die Linke"             "SPD"
[19] "SPD"                   "Andere Partei"          "SPD"
```

Modus

Was ist der häufigste Wert?

```
1  table(kurs$prtvede1)
```

```
Andere Partei Bündnis 90/Die Grünen                 CDU/CSU
            1                       6                       1
   Die Linke                     FDP                     SPD
            6                       1                       5
```

→ Dieser sog. **Modus** lässt sich unabhängig vom Skalenniveau berechnen und ist auch aus Häufigkeitstabellen ablesbar

Durchschnitt

Was ist der durchschnittliche Wert?

Häufig interessiert uns aber der Durchschnitt - z.B. bei Bewertungen

```
1  mean(kurs$lrscale,na.rm=T)
```

```
[1] 3.047619
```

→ das funktioniert aber nur bei *metrischen Skalen*!

Durchschnitt

Idee: Kombination von Summe & Zahl der Observationen

```
1 sum(kurs$lrscale,na.rm=T)
```

```
[1] 64
```

```
1 length(kurs$lrscale)
```

```
[1] 21
```

aufgepasst: mögliche fehlende Beobachtungen!

→ `mean()` ist die sicherere Option!

Median

Welcher Wert liegt ‘in der Mitte’?

→ Der Median teilt die Werte der Variablen in zwei gleiche Teile

Im Gegensatz zum Mittelwert ist der Median **weniger anfällig für extreme Observationen**

z.B. Studi A verdient 400€ als stud. Hilfskraft, Studi B 500€ als Barkeeper, Studi C 3.000€ mit Mieteinnahmen

Median

Beispiel: Links-rechts Einordnung im Kurs

```
1 sort(kurs$lrscale)
```

```
 [1] 0 0 1 1 1 2 2 2 2 2 2 3 3 4 5 5 5 5 5 6 8
```

```
1 median(kurs$lrscale)
```

```
[1] 2
```

Median

Beispiel: Einkommen aus dem sozio-ökonomischen Panel (Umfrage)

Vergleiche über Gruppen hinweg

Vergleiche zwischen diskreten Gruppen können wir anstellen, indem wir deren Mittelwerte vergleichen

z.B. Sind Frauen oder Männer linker?

```
1 kurs_gender <- group_by(kurs,gndr)
2 summarize(kurs_gender,lr=mean(lrscale))
```

```
# A tibble: 2 × 2
  gndr         lr
  <chr>     <dbl>
1 männlich   2.7
2 weiblich   3.36
```

Vergleiche über Gruppen hinweg

Alternative Berechnung (siehe Llaudet & Imai)

```r
# Männer
mean(kurs$lrscale[kurs$gndr=="männlich"],na.rm=T)
```

```
[1] 2.7
```

```r
# Frauen
mean(kurs$lrscale[kurs$gndr=="weiblich"],na.rm=T)
```

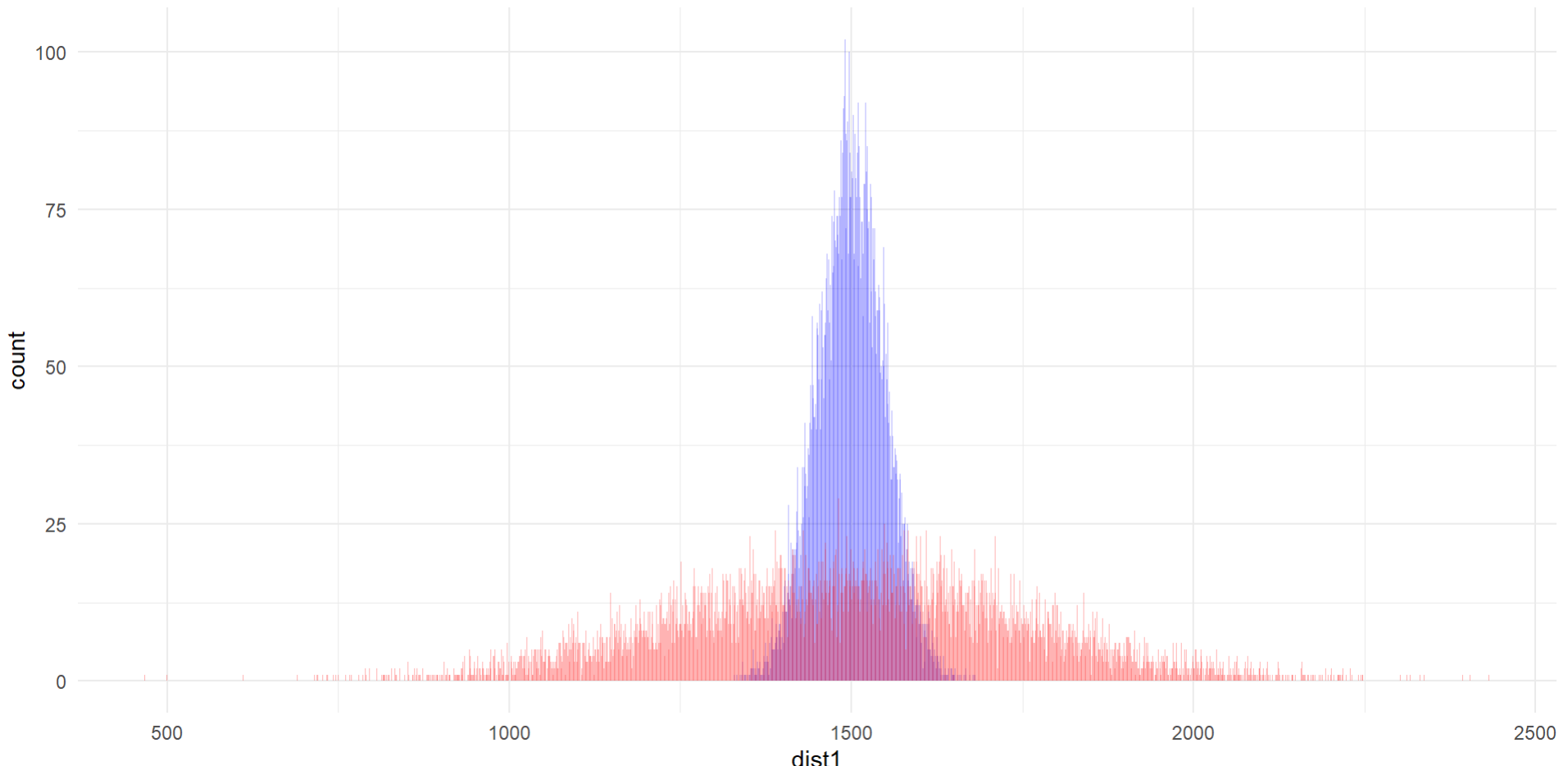
```
[1] 3.363636
```

Verteilungen

Verteilungen

Der Informationswert von Mittelwerten ist begrenzt

- z.B. bei Gehaltsverteilungen



Quartile

Detaillierter sind **Quartile**, die den Datensatz in vier gleich umfangreiche Teile teilen

```
1  quantile(kurs$lrscale,na.rm=T)
```

```
  0%  25%  50%  75% 100% 
   0    2    2    5    8 
```

→ der Abstand zwischen den Quantilen gibt einen Überblick über die Verteilung (‘Quartilsabstand’)

Standardabweichung

- **Standardabweichung** als **Verteilung der Variablen**: durchschnittlicher Abstand der Beobachtungen zum Durchschnitt

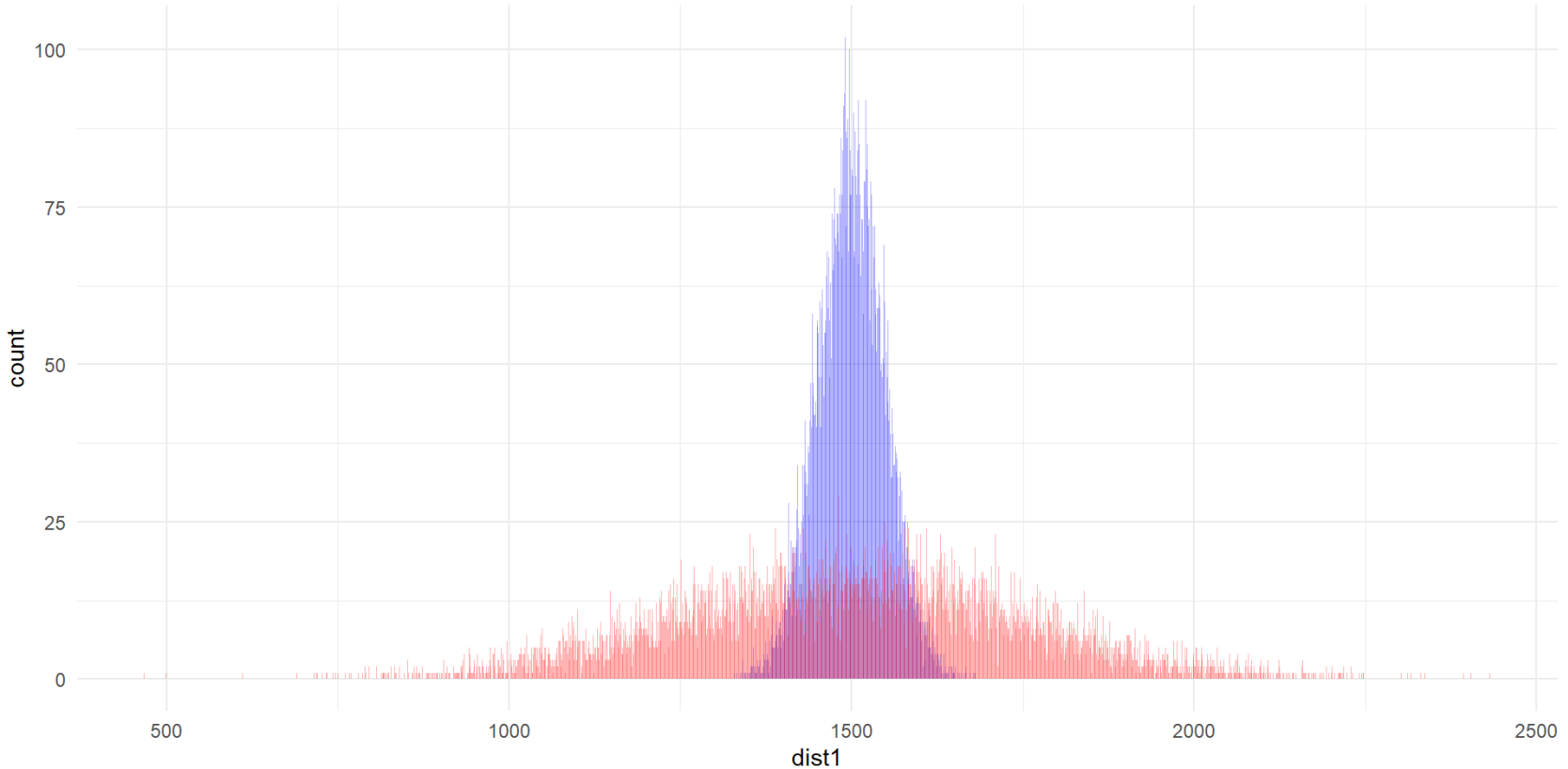
$$sd(X) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}$$

```
1  sd(kurs$lrscale,na.rm=T)
```

```
[1] 2.132515
```

→ Die Standardabweichung zeigt wie breit die Variable um den Mittelwert gestreut ist; bei einer kleineren Standardabweichung sind die Werte näher am Mittelwert

Standardabweichung



Blau: Durchschnitt 1500.3 & Standardabweichung 49.4

Rot: Durchschnitt 1498.1 & Standardabweichung 251.5

Varianz

- Die sogenannte **Varianz** ist ein weiteres Streuungsmaß, das eng mit der Standardabweichung verbunden ist:
  - $sd(X) = \sqrt{var(X)}$

```
1 var(kurs$lrscale,na.rm=T)
```

```
[1] 4.547619
```

Box plots

- Häufig visualisieren wir Verteilungen mit sogenannten ‘Box-Plots’
  - Median & Quartile
  - Maximalwerte und Ausreißer

```
1  boxplot(kurs$lrscale,na.rm=T)
```

Zusammenhänge zwischen Variablen

Zusammenhänge zwischen Variablen

- Um die Welt zu verstehen müssen wir auch **Zusammenhänge zwischen Variablen** verstehen

→ Zusammenhänge oft interessanter als der Mittelwert oder die Verteilung einer Variablen

→ Ziel: Vorhersage und Erklärung

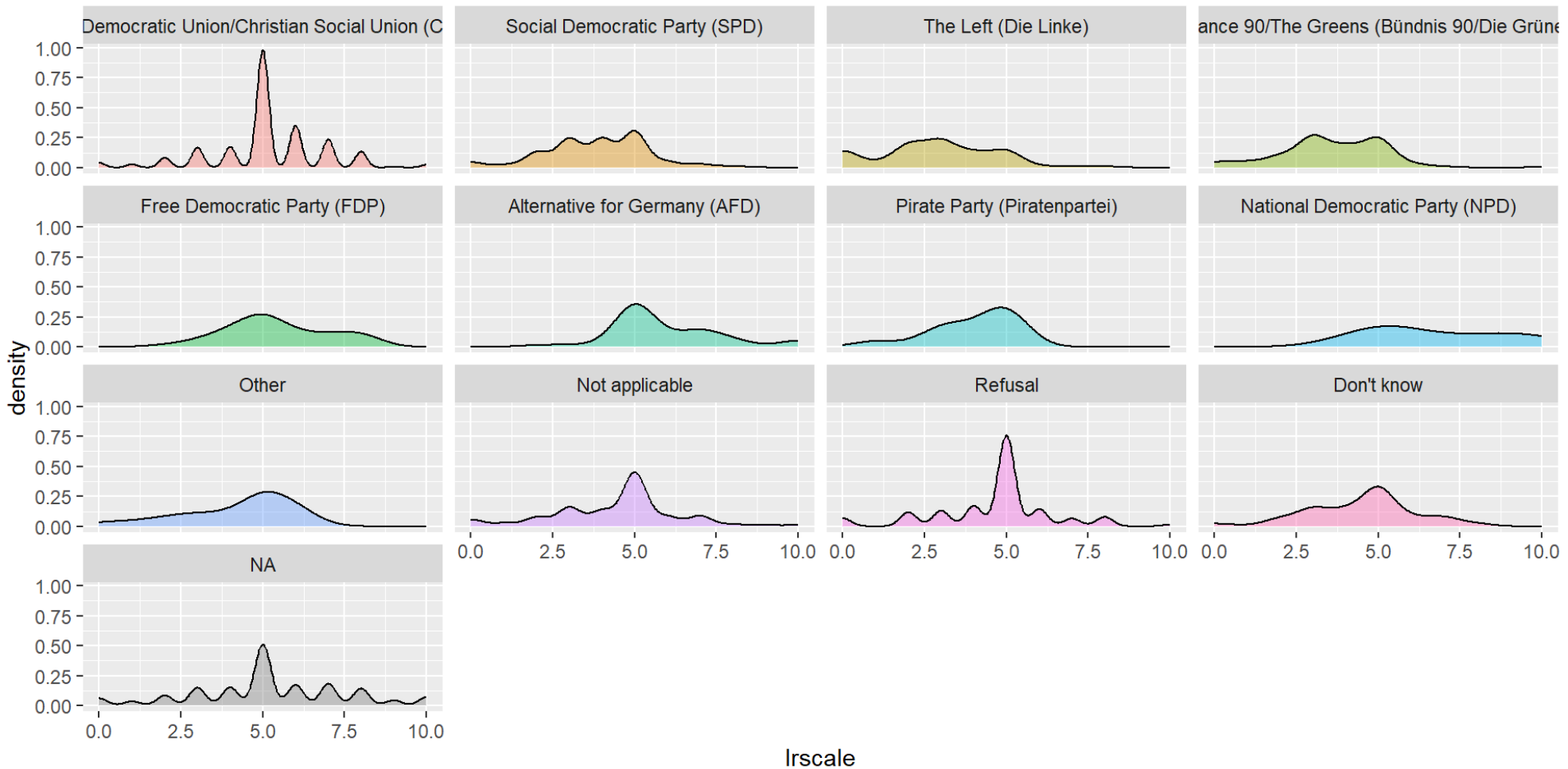
Zusammenhänge zwischen Variablen

- bei **kontinuierlichen Variablen** (z.B. Alter) wäre eine Einteilung in Gruppen willkürlich
  - auch ordinale Variablen mit vielen Werten behandeln wir oft als nazu-kontinuierlich (z.B. links-rechts Skala)

- stattdessen: Darstellung des Zusammenhangs
  - Betrachten der Verteilung (Scatterplots und Verteilungsgrafiken → Sitzung 11 & Lektüre)
  - Zusammenhangsmaße wie Korrelationen

Verteilungsgrafiken

Beispiel aus der ESS: lrscale nach prtvede1 Kategorien



Korrelationen

Häufig wollen wir den Zusammenhang in Zahlen darstellen

→ **Korrelation**

Korrelationen

Korrelationen (nach Pearson's Korrelationskoeffizient) berechnen sich aus der Kovarianz (einem Zusammenhangsmaß), geteilt durch das Produkt der Standardabweichungen

$$r = \frac{\text{cov}(X,Y)}{sd(x)sd(y)}$$

Korrelationen

- **Korrelationskoeffizient r** (-1 bis 1) als Maß für **Stärke und Richtung des linearen Zusammenhangs zweier Variablen**
  - positiv oder negativ: Richtung des Zusammenhangs
  - $|r|$: je näher an 1, desto stärker
- Spiel: Guess the correlation

Korrelationen

In R: `cor()`

z.B.: Zusammenhang zwischen Alter und links-rechts Orientierung

```
1  cor(ess8$agea,ess8$lrscale, use="complete.obs")
```

```
[1] 0.03053121
```

`use="complete.obs"` schließt fehlende Werte aus (ähnlich wie `na.rm=T`)

Korrelationen

Aber: **Correlation does not imply causation**

Source

→ Wir werden später im Kurs diskutieren, wie und ob wir Kausalität in Zusammenhängen zeigen können. Dabei lernen wir auch komplexere Analysemethoden kennen. (Beispiele)

Übung an den ESS Daten

Daten laden

```
1 library(tidyverse)
2 library(haven)
3 ess8 <- read_dta("../data/ESS8e02_2.dta")
```

Faktor vs. numerisch

Die ESS Variablen sind (wegen des Datensatz-Formats) als Variablen des Typs `haven_labelled` gespeichert

→ wir können sie als numerische oder ordinale Variablen behandeln

```
1 head(ess8$lrscale)
```

```
<labelled<double>[6]>: Placement on left right scale
[1] 0 1 5 0 5 5

Labels:
 value      label
     0       Left
     1          1
     2          2
     3          3
     4          4
     5          5
     6          6
     7          7
     8          8
     9          9
    10      Right
 NA(a)    Refusal
 NA(b) Don't know
 NA(c)  No answer
```

Faktor vs. numerisch

```
ess8$lrscale_num <- as.numeric(ess8$lrscale)
head(ess8$lrscale_num)
```

```
[1] 0 1 5 0 5 5
```

```
ess8$lrscale_fact <- as_factor(ess8$lrscale)
head(ess8$lrscale_fact)
```

```
[1] Left 1    5    Left 5    5
Levels: Left 1 2 3 4 5 6 7 8 9 Right Don't know Refusal No answer
```

Aufgabe

Lösen Sie die Übungsaufgaben im .r-Skript zur heutigen Stunde

Aufgabe

Erinnern Sie sich an die Fragen aus der ESS, die Sie in der vorletzten Stunde herausgesucht haben. Welche Maße können Sie zur Zusammenfassung verwenden?

Erstellen Sie eine kurze Übersicht und wenden Sie die Maße an.

Laden Sie ein Dokument mit der Übersicht und einem aussagekräftigen Wert für jede Variable (z.B. Mittelwert oder Verteilung) und ein paar Gedanken zur Interpretation auf Moodle hoch.

ca. 1 Seite, Abgabe 31.05.

Kleine Hausaufgabe

Bringen Sie eine Datenvisualisierung mit (z.B. aus der Zeitung, einem Kurs, …) - überlegen Sie sich, was Sie an dieser Darstellung gut oder schlecht finden.

Nächste Woche: Visualisierung

- Kieran Healy Data Visualization: A Practical Introduction (Princeton, NJ: Princeton University Press, 2018)., Kapitel 2
- optional für R: Garrett Grolemund and Hadley Wickham R for Data Science, 2017. (2nd edition), Kapitel 2 & 11
- optional: Video Data Visualization
- optional: Hadley Wickham ggplot2: Elegant Graphics for Data Analysis, 2nd ed. 2016 Edition (New York, NY: Springer, 2016).