Computergestützte Datenanalyse: DATA-Übung mit R

Tag 3 – 24.07.2025

UNSER PLAN

- **Tag 1**
  - Einführung in R und RStudio
  - „Basics“:
  - Coding Konventionen
  - Objekte, Datenimport & Co
- **Tag 2**
  - Skalenniveau
  - Troubleshooting
  - Datenaufbereitung
  - Datenvisualisierung
  - Deskriptive Statistik
- **Tag 3**
  - Bivariate Analyse
- **Tag 4**
  - Inferenzstatistik
  - Abschluss

Genereller Ablauf

- Vier Tage geblockt
- Mischung aus Input- und Übungssessions
- Anwesenheitsabfrage alle 90 Minuten

Heute

- Zwei 15 Minuten Pause
- Eine Mittagspause

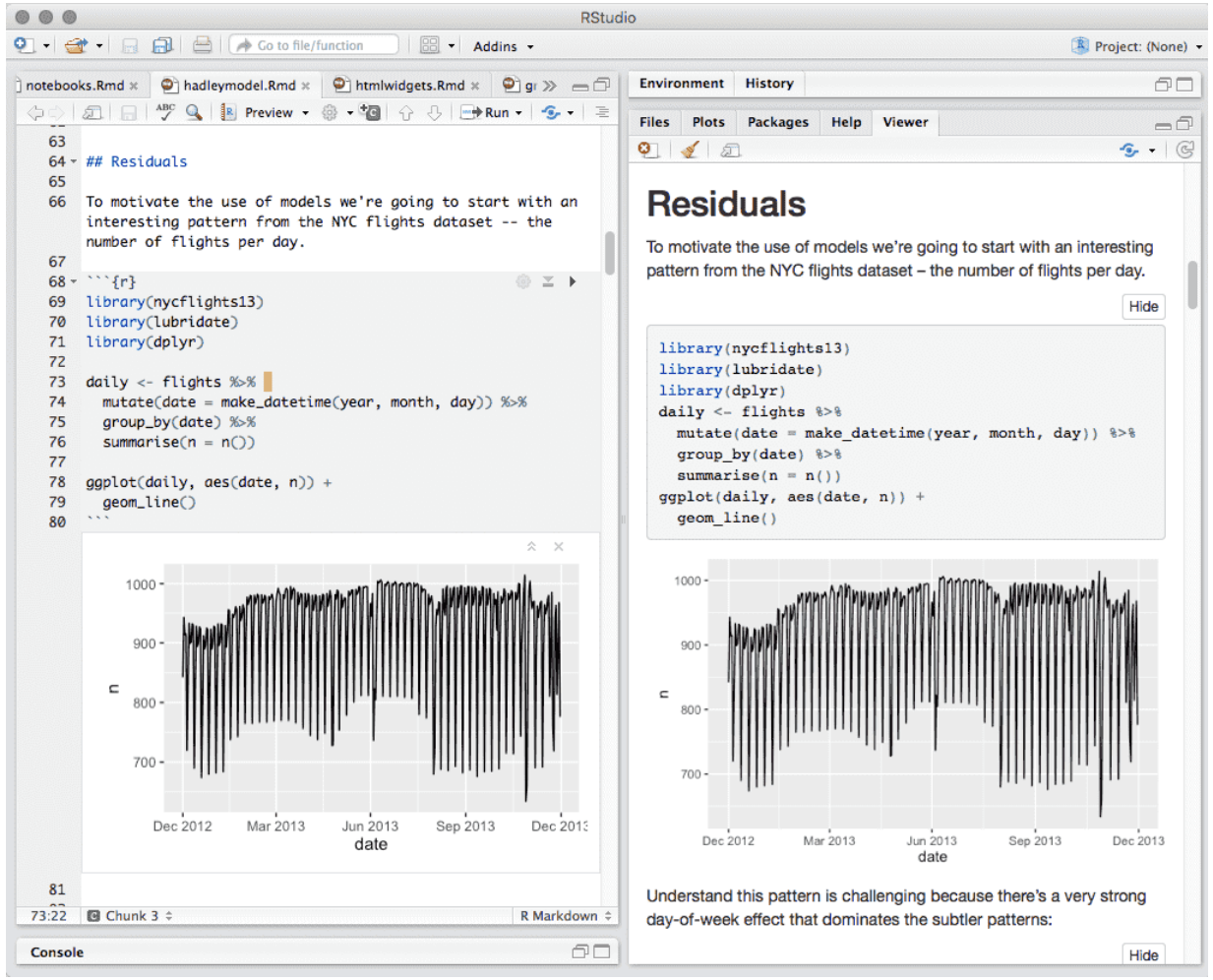
REFRESHER

- Was sind Funktionen?
- Was sind die vier Skalenniveaus?
- Welche Datentypen/-klassen gibt es?
- Was tun, wenn Error Meldungen auftauchen?
- Was ist der Unterschied zwischen Subsetting und Rekodieren?

  Welche Packages, Funktionen?
- Warum visualisieren wir Daten? Welches Package hilft dabei?

R Markdown

- Kombination aus Code und Text
- Gut zum…
  - Teilen
  - Veranschaulichen



Getting started

1. Rmarkdown file erstellen und einstellen
2. Packages installieren und/ oder aktivieren: rio und tidyverse
3. Import der ALLBUS 2018 Daten
4. In einem **neuem** Dataframe werden die Variablen pt01 bis pt20 gespeichert. Nutzt dafür den Befehl select() und recherchiert, wie das funktioniert
5. Um welche Variablen handelt es sich? (s. im Variablen Report nach und kommentiert euer Rmd)
6. Berechnet Mittelwert und Standardabweichung für die Variablen pt01, pt02 und pt03, in Textform reporten! Beispiel: pt01 hat einen *M*=X und *SD*=Y.
7. Untersucht die Verteilungen, ergo absolute und relative Häufigkeiten für pt03, in Textform reporten!
8. Erstellt einen barplot für pt03 mit Beschriftung

Exkurs: Boxplot erstellen

Was ist ein Boxplot?

```
ggplot(data = ?, aes(y = pt01)) +
  geom_boxplot() +
  labs(title = "?",
       y = "?")
```

Welche Datenanalysemethoden sind Ihnen bekannt?

Welche Datenanalyse würden Sie gerne mit R durchführen, wenn Ihren Daten und Programmierkenntnissen keine Grenzen gesetzt sind?

5 Minuten untereinander diskutieren, kurz im Plenum vorstellen

Bivariate Analysen

```
allbus$westost <- factor(allbus$eastwest,
                         labels = c("Westdeutschland", "Ostdeutschland"))


allbus <- allbus %>%
  mutate(schulabschluss = case_when(
                         educ == 2 ~ "Hauptschulabschluss",
                         educ == 3 ~ "Mittlere Reife",
                         educ %in% c(4,5) ~ "(Fach-)Abitur"))
```

Kreuztabellen und Zusammenhangsmaße

- Typische Frage: Gibt es Unterschiede in der Verteilung zwischen zwei Variablen?
- Anzahl der Ausprägungen beachten

Aufgabe:

Was ist zu sehen?

```
tabelle <- table(df$westost, df$schulabschluss)
tabelle_sum <- addmargins(tabelle) #summenwerte
```

```
spaltenprozente <- round(100*prop.table(tabelle, 2),2)
spaltenprozente #spaltenprozente
```

Kreuztabellen und Zusammenhangsmaße

- Typische Frage: Gibt es Unterschiede in der Verteilung zwischen zwei Variablen?
- Anzahl der Ausprägungen beachten

Aufgabe:

```
{r eval = FALSE}
install.packages(“gmodels“)
library(gmodels)
CrossTable(allbus$westost, allbus$schulabschluss,
format="SPSS")
```

Bivariate Analyse mittels Kreuztabelle: Test auf statistische Unabhängigkeit

- Untersucht den statistischen Zusammenhang zwischen zwei Variablen
  - Gibt es einen überzufälligen Zusammenhang?
  - Falls ein überzufälliger Zusammenhang besteht: Wie stark ist der Zusammenhang?
  - Wie können wir den Zusammenhang interpretieren?
- Gängiges Werkzeug zur Analyse von Daten auf nominalem oder ordinalem Messniveau
  - Zum Beispiel: Chi-Quadrat, Cramer‘s V oder Phi (bei 2x2-Tabellen sind letztere identisch)

Zusammenhangsmaß

- Gibt an, wie stark der Zusammenhang zwischen zwei Variablen ist

- Das Zusammenhangsmaß ist abhängig vom Skalenniveau:
  - Zwei dichotome Variablen: phi (ϕ)
  - Eine dichotome und eine nominale Variable, zwei nominale Variablen, eine ordinale und eine dichotome oder nominale Variable: Cramer's V
  - Zwei mindestens ordinalskalierte Variablen oder eine ordinale und eine mindestens intervallskalierte Variable: Spearman's rho (ρ) oder Kendall's tau-b (τ -b) oder tau-c(τ -c)
  - Zwei mindestens intervallskalierte Variablen: Pearson's r
- Zusammenhangsmaße liegen zwischen 0 und 1 beziehungsweise zwischen -1 und +1
  - > 0.2 interpretierbarer Zusammenhang
  - > 0.5 starker Zusammenhang
  - *1 perfekter Zusammenhang*

Chi-Quadrat (χ^2)

- Test auf statistische Unabhängigkeit
- Je größer χ^2, desto stärker der Zusammenhang von zwei Variablen (0=kein Zusammenhang)

Aufgabe

Kreuztabelle erstellen:

```
tabelle <- table(df$variable, df$variable)
```

Chi^2-Test durchführen mithilfe der Kreuztabelle

```
chisq.test(tabelle)
```

```
Pearson's Chi-squared test
data: tabelle
X-squared = 131.13, df = 2, p-value < 2.2e-16
```

Und was sagt uns das Ergebnis?

Chi-Quadrat (χ^2)

- Test auf statistische Unabhängigkeit
- Je größer χ^2, desto stärker der Zusammenhang von zwei Variablen (0=kein Zusammenhang)
- Problem bei der **Interpretation**:
  - Kann sehr große Werte annehmen
  - Zusammenhang hängt von Tabellenformat und Fallzahl ab

- Lösung: Normierung

Cramer‘s V

- Normiert einen χ^2-Wert von 0 bis 1
  - Unempfindlich gegenüber Tabellenformat und Fallzahl
- 0 bedeutet kein Zusammenhang
- 1 bedeutet perfekter Zusammenhang

Aufgabe

chisq.test(tabelle)

install.packages("vcd")
library(vcd)

assocstats(tabelle)

Schreibt eine kurze Interpretation

Zur Erinnerung:

0 bedeutet kein Zusammenhang
1 bedeutet perfekter Zusammenhang

Pearson‘s R

- Korrelationen zwischen zwei (metrischen) Variablen: Pearson‘s R

- Pearson‘s R lässt sich auch berechnen zwischen
  - Einer Dummy und einer metrischen Variablen
  - Zwei Dummy Variablen (als Punktbiseriale Korrelation)

Aufgabe

- Korrelation zwischen Vertrauen in den Bundestag und in Parteien
- Welche Variablen sind das? Verteilung anschauen
- table(allbus$pt03) vs. table(allbus$pt03, useNA = “ifany“)
- Streudiagramm erstellen:

```
ggplot(allbus, aes(x= pt15, y = pt03)) +
geom_point(position = "jitter", alpha = 0.3) +
labs(x = "Vertrauen in politische Parteien",
     y = "Vertrauen in den Bundestag")
```

Aufgabe

- Korrelation zwischen Vertrauen in den Bundestag und in Parteien
- cor.test(allbus$pt15, allbus$pt03)

Interpretation Korrelationskoeffizient

- Wie stark eine Variable mit der anderen Variable zusammenhängt, wird in der Höhe des Koeffizienten angegeben. Je höher dieser Wert ist, desto stärker wird eine Variable durch die andere Variable bestimmt. Der Wert kann zwischen 0 und +-1 liegen

- Richtung des Zusammenhangs: Hier kommt es auf das Vorzeichen des Koeffizienten an. Bei einem + sprechen wir von einem positiven Zusammenhang, während wir bei einem - von einem negativen Zusammenhang sprechen

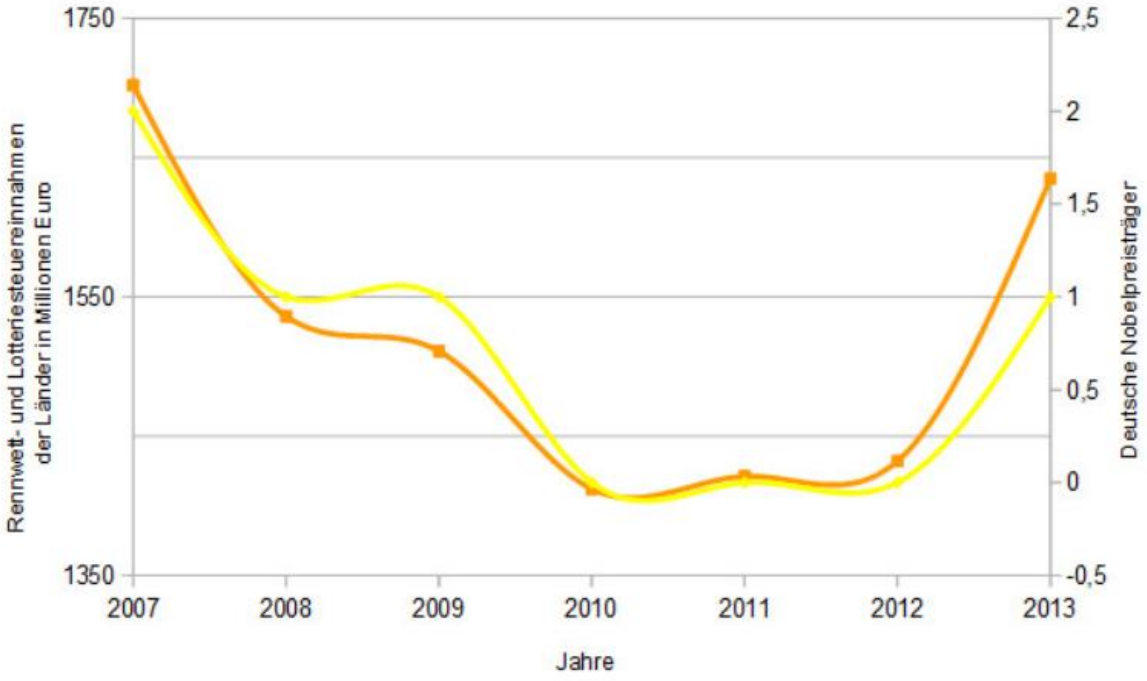
- Signifikanztest: Hier wird überprüft, ob wir auch in der Grundgesamtheit von einem Zusammenhang zwischen den beiden Variablen ausgehen können. Wenn der p-value kleiner als 0.05 ist, können wir davon ausgehen, dass ein Zusammenhang zwischen den beiden Variablen auch in der Grundgesamtheit vorliegt

Korrelation ist keine Kausalität!

- Korrelation ist eine Assoziation zwischen Variablen, sagt aber nicht über einen Ursache-Wirkungsmechanismus aus

- In den SoWi arbeiten wir stets theoriegeleitet, leiten Hypothesen über spezifische Mechanismen/Zusammenhänge/Pfade ab

Scheinkorrelation

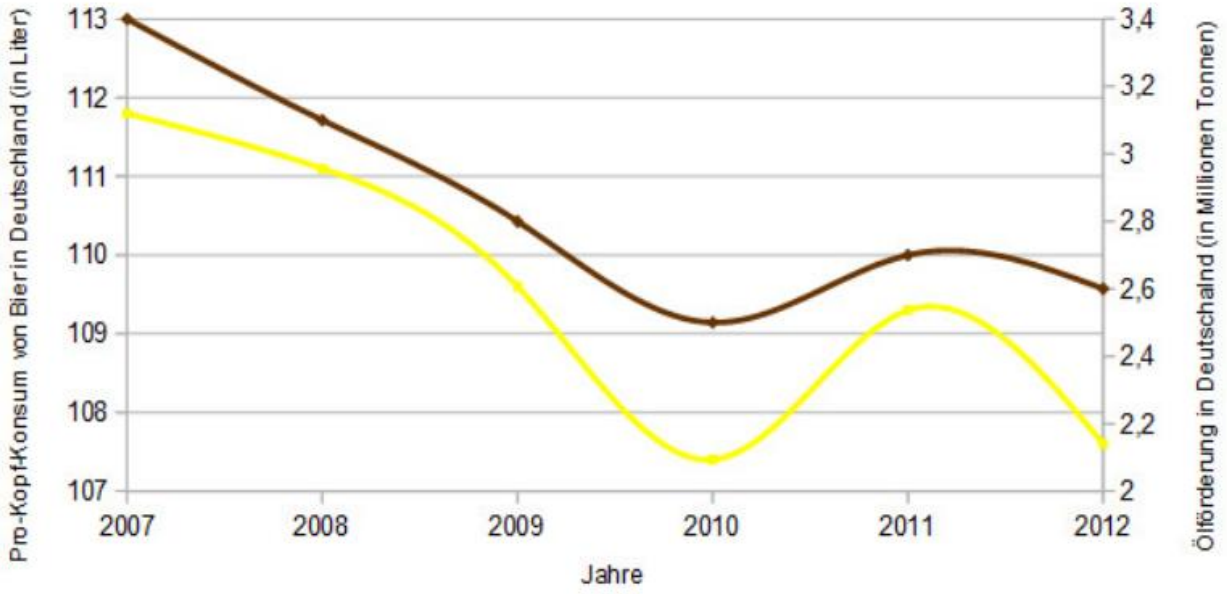


Rennwett- und Lotteriesteuereinnahmen der Länder in Deutschland (orange) und Zahl der deutschen Nobelpreisträger (goldgelb)

Korrelation: 0,9402

Quelle: Statista und Wikipedia
Grafik/Berechnung: mit OpenOffice Calc

Quelle: https://scheinkorrelation.jimdofree.com/



Deutscher Pro-Kopf-Konsum von Bier (gelb) und Ölförderung in Deutschland (braun)

Korrelation: 0,9587

Quelle: Statistia& Wikipedia
Grafik/Berechnung: mit OpenOffice Calc

Quelle: https://scheinkorrelation.jimdofree.com/

Spearman‘s rho und Kendall‘s tau

- Zusammenhang zwischen zwei Variablen, für die wir höchstens ordinales Skalenniveau annehmen können

- Interpretation zur Stärke und Richtung des Zusammenhangs sowie Signifikanztest

Aufgabe

Spearman‘s rho:

```
cor.test(allbus$pt15,
         allbus$pt03,
         method = "spearman")
```

Kendall‘s tau

```
cor.test(allbus$pt15,
         allbus$pt03,
         method = "kendall")
```

Und was sagt uns das Ergebnis?

Aufgabe

Korrelationsmatrix:

```
vertrauen <- subset(allbus, select = pt03:pt15) # Datensatz zum pol. Vertrauen
cor(vertrauen,
        use = "pairwise.complete.obs",
        method = "pearson")
```

Was seht ihr?

REFRESHER

- Was ist eine Kreuztabelle?
- Warum wollen wir Zusammenhangsmaße lieber normiert haben?
- Was ist der Unterschied zwischen Cramer V und Pearsons R?
- Was ist die Nullhypothese?
- Was ist eine Scheinkorrelation?

FRAGEN, UNKLARHEITEN, FEEDBACK?

VIELEN DANK ☺

UND BIS NÄCHSTE WOCHE!

ahrabhi.kathirgamalingam@cais-research.de