ulm university universität ulm

Wirtschaftsstatistik

Sommersemester 2019

Dr. Tobias Bluhmki

Institut für Statistik, Universität Ulm

25. April 2019

Die Vorlesungsfolien sind nur zum internen Gebrauch für Veranstaltungsteilnehmer/innen bestimmt. Mein besonderer Dank gilt Prof. Dr. Jan Beyersmann, Prof. Dr. Markus Pauly, sowie Dr. Marco Oesting für die Bereitstellung ihrer Vorlesungsmaterialien.

Organisatorisches

Organisatorisches

Dozent: Dr. Tobias Bluhmki
Büro: HeHo 20/E.45
Sprechstunde: Nach Vereinbarung
E-Mail: tobias.bluhmki@uni-ulm.de

Übungsleiter: Jan Feifel
Büro: HeHo 20/E.45
Sprechstunde: Nach Vereinbarung
E-Mail: jan.feifel@uni-ulm.de

Zeiten und Orte

- Vorlesung: Mo, 12:15-13:45 Uhr, Hörsaal Innere Medizin (O23/2619)
- Übung: Do, 12-13 Uhr, c.t., O25/H2 ⇒ ggf. längere Übung?
  → **Wöchentlich!**
  → **Start:** nächste Woche (2. Mai)

Homepage:

`https://www.uni-ulm.de/mawi/statistics/courseslehre/summer-semester-2019/wirtschaftsstatistik/`

Organisatorisches

Moodle (!):

- bitte anmelden (Einschreibeschlüssel: `lm(y~x)`)
- Informationen, Materialien zur Vorlesung, Übungsblätter, `R`-Lösungen, etc.
- Vorlesungs-/Übungsplan beachten
- Vorlesungfolien werden Woche für Woche **vor** der Vorlesung hochgeladen (Sonntags?)
- **Achtung:** Es wird auch gesonderte Tafelanschriebe geben!

Übungsaufgaben:

- Abgabe jeweils am Donnerstag **vor** Beginn der Übung, d.h., bis spätestens 12:15 (Bearbeitungszeit: eine Woche)
  → **handschriftlich, mit dokumentenechtem Stift geschrieben und getackert**
- korrigierte Rückgabe unmittelbar vor der darauffolgenden Übung
- gemeinsame Abgaben (2 Personen) erlaubt + **empfohlen** (Austausch in Kleingruppen!)

Organisatorisches

- Zu den Aufgaben werden **auch** die Interpretation von Grafiken & R-Output gehören
- R-Programmieraufgaben, eigenständige Implementierung von R-Code
  - **ausgedruckt** mit den anderen Lösungen abgeben **und** digitale **(!)** Abgabe über Moodle-Upload
  - Alle Namen der Leute aus der Abgabegruppe bitte als Kommentar oben im R-Script einfügen
- **Bonusaufgaben werden extra gekennzeichnet!**

Zulassung zur Klausur: 50% der regulären Übungspunkte

Organisatorisches

Prüfung:

- **Hauptklausur:** Mi. 31.07.19, 8:00 – 10:00Uhr ☺
- Schriftlich (ohne MC); **mit** Interpretation von R-Code/Output, ggf. Wissen von ganz einfachen R-Befehlen
- **Rechtzeitige** Anmeldung im Hochschulportal (nach Erreichen der Vorleistung) notwendig
- **Nachklausur:** Fr. 27.09.19 **(offen!)**, **vermutlich** 14:30 – 16:00
- Hilfsmittel: Ein **handbeschriebenes** DIN A4 Blatt (Vorder- und Rückseite) sowie ein **nichtprogrammierbarer** Taschenrechner

Prüfungsvorbereitung:

- Teilnahme an und aktive Beteiligung in der Vorlesung und Übung
- Vor-/Nachbereitung des Vorlesungs- & Übungsstoffes
- Regelmässige Bearbeitung und Abgabe der Übungsblätter
- Diskussionen in Kleingruppen
- Fragen, Fragen, Fragen!
- **Probeklausur (?)**

Vorwissen

Aus den Vorlesungen *Mathematische Grundlagen der Ökonomie I+II* sowie *Stochastik für WiWis* sollten Sie bereits kennen:

- Ereignisse und Wahrscheinlichkeiten
- Zufallsvariablen und ihre Charakteristiken
- Gesetze der großen Zahlen
- Grenzwertsätze
- Parameterschätzung
- Konfidenzintervalle
- Statistische Tests

Oder?!

Bitte diese Dinge ggf. zu Hause wiederholen ☺

(Statistik-)Software

- Unausweichlich für umfangreichere Berechnungen (→ **Big Data/Data Science!)** und **Erstellen von Grafiken**

Wir verwenden die **open source** Software `R`:

- Programmiersprache (orientiert sich an S), die mittels eines Editors **komfortabel und intuitiv** verwendet werden kann
- Sehr weit verbreitet an Hochschulen und in der Industrie
- Grafische/Interaktive Benutzeroberfläche: **RStudio**

Kommerzielle Programme mit umfangreichen Funktionen: SPSS, SAS, STATA
→ `https://www.uni-ulm.de/einrichtungen/kiz/service-katalog/software/softwareliste/`

R-Einführung

Installationsanleitung:

1. Gehe zu `https://www.r-project.org`
2. Reiter 'Download CRAN'
3. 'Germany' suchen und einen der Links auswählen, z.B. `https://cran.uni-muenster.de/`
4. R herunterladen für Linux, Mac oder Windows (bei MacOS darauf achten, dass es verschiedene Pakete (pkg) für unterschiedliche MacOS Versionen gibt; bei Windows auf 'Install R for the first time')
5. R installieren
6. Danach auf `https://www.rstudio.com/products/rstudio/download/` RStudio herunterladen und installieren

Alternativ: Vorinstalliert in allen PC-Pools!

'Blatt 0':

- Nachbereitung der R-Einführung auf Moodle bis nächste Woche (**Wichtig!**)
- Genügend Tutorials sind im Internet zu finden ☺
- Bei Problemen bitte fragen!

Literatur

Die folgende Liste umfasst lediglich eine kleine Auswahl von Texten, die neben dem Vorlesungsskripts für ein ergänzendes und vertiefendes Studium empfohlen werden können:

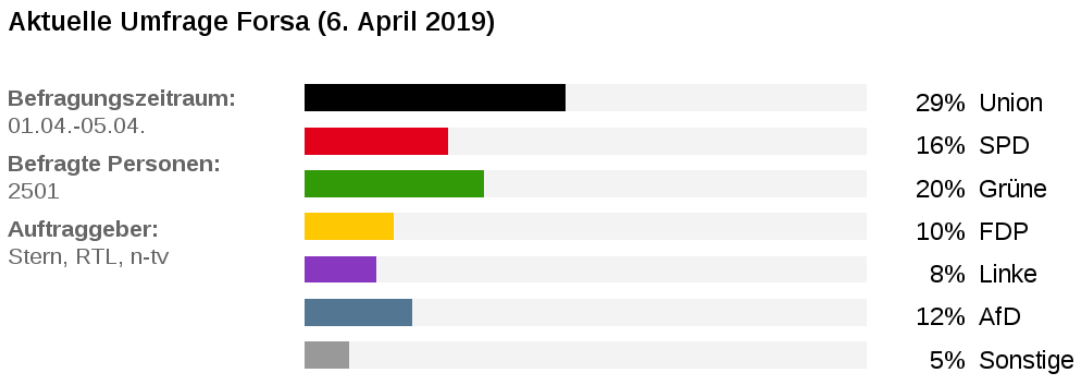
- Fahrmeir, L., Künstler, R., Pigeot, I. und Tutz, G. (2007). Statistik - Der Weg zur Datenanalyse, Springer.
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2014). An Introduction to Statistical Learning – with Applications in R, Springer
- Haug, S. (2013). Statistik für Betriebswirtschaftslehre (Einführung mit R), Skript, TU München.
- Kunze, M. (2014). Wirtschaftsstatistik, Skript Universität Ulm.
- Fahrmeir, L., Kneib, T., und Lang, S. (2007). Regression. Springer.
- Heiss, F. (2016). Using R for Introductory Econometrics.
- Kleiber, C. und Zeileis, A. (2008). Applied econometrics with R. Springer.
- ...

... Noch Fragen?

- Sommerfest? ☺

Kapitel 0:
Die Antwort auf das ‘Warum?’

Motivierende Beispiele aus der Praxis – Wahljahr 2019



*Quelle: `www.spiegel.de`

- Wie entstehen diese repräsentativen Umfragen?
- Wie valide sind diese?
- ...

Querschnittstudie: An einer bestimmten Anzahl von Objekten (hier: Wähler/innen) wird/werden an einem bestimmten Zeitpunkt ein oder mehrere Merkmal(e) (hier: Parteipräferenz) erfasst.

Motivierende Beispiele aus der Praxis – Brexit und Pfund

*Quelle: `https://boersen.manager-magazin.de`

- Systematische Struktur erkennbar?
- Einfluss Brexitchaos?
- ...

Längsschnittstudie: An einer bestimmten Anzahl von Objekten wird/werden an mehreren Zeitpunkten ein oder mehrere Merkmal(e) (hier: Wechselkurs) erfasst.

Motivierende Beispiele aus der Praxis – Abschlussarbeit

- Abschlussarbeit aus dem Bereich Wirtschaftswissenschaften (2018)
- Statistische Auswertung einiger Daten des Bloomberg Systems
- Auszug aus einer an mich addressierten Email:

  '[...] Habe nun meinen Datensatz soweit zusammen gebaut, dass es eine Treatment group und eine control group gibt (inkl. den notwendigen Firmendaten). Allerdings komme ich einfach nicht weiter, was den Aufbau des Regressionsmodells angeht. [...] Wäre es möglich, dass wir uns das noch einmal anschauen bzw. das Modell durchsprechen? [...]'

$\Rightarrow$ Techniken aus der Vorlesungs 'Wirtschaftsstatistik' ☺

Worum geht es?

Wirtschafts-Statistik:

Lehre vom Umgang mit Daten motiviert aus den Wirtschaftswissenschaften

Idealisierter Ablauf einer statistischen Untersuchung:

1. **Fragestellung, Versuchsplanung und Datenerhebung:**
   z.B. Fragebögen, longitudinale Beobachtung einer Aktie im DAX, etc.
2. **Deskriptive (beschreibene) Statistik:**
   - Datenaufbereitung und einfache/verständliche -darstellung (z.B. durch Diagramme, Verlaufskurven, Tabellen, etc.)
   - Datenauswertung (z.B. durch Berechnung von Maßzahlen wie Häufigkeiten, Mittelwert oder Streuung $\Rightarrow$ Auffinden von Strukturen, Hypothesen, etc.)
3. **Induktive oder schließende oder beurteilende Statistik:**
   - weitergehende statistische Analyse (z.B. durch Schätzen / Testen in statistischen Modellen) und Bewertung (Entscheidung unter Unsicherheit)
   - Verwendung von mathematischer Stochastik

Idealerweise sollten Punkt 1 und 3 Hand in Hand gehen (Datenplausibilität, wird auf ÜB 1 behandelt)

Kapitel 1: Versuchsplanung und Datenerhebung

Datenerhebung

Definition 1.1.

In der Datenerhebung werden, z.B. durch Beobachten/Befragen/Messung, die Ausprägungen/Werte von verschiedenen Merkmalen/Variablen in einer Grundgesamtheit / Population (gerne mit Ω bezeichnet) oder einem Teil der Population (Stichprobe) erfasst.

Beispiel 1.1. Erhebung von Studierendendaten in WiStat SS 2019:

Am Anfang werden z.B. Geschlecht, Abiturnote, Alter, ... von allen Teilnehmern erfragt

- **Statistische Einheit:** jede(r) einzelne(r) Teilnehmer(in)
- **Grundgesamtheit:** alle Teilnehmer/innen der Vorlesung
- **Stichprobe:** tatsächlich untersuchte Teilmenge der Grundgesamtheit (nur, die die heute da sind)
- **Merkmale:** Erhobene Variablen wie das Alter, Geschlecht, ...
- **Ausprägung:** konkreter Wert des Merkmals für eine(n) bestimmte(n) Teilnehmer(in)

Urliste (simulierte Daten, 50 Teilnehmer/innen)

```r
set.seed(2) #Setze Zufallsgenerator fest (Reproduzierbarkeit!)
urliste<-data.frame(StudiID=rep(1:50),
                    Sex=factor(rbinom(50,size=1,prob=0.5),
                               levels = c(0,1),
                               labels = c("w","m")),
                    Grade=round(runif(50,1,3),2),
                    Age=round(rep(17,50)+rexp(50,1/4),2),
                    HairColor=sample(
          x=c("black","brown","red","blond"),size=50,replace=T))

print(head(urliste,n=4),row.names = F)
##  StudiID Sex Grade   Age HairColor
##        1   w  1.01 24.98     brown
##        2   m  1.03 22.07     black
##        3   m  2.37 20.36       red
##        4   w  2.86 17.77       red
```

Datenerhebung

Datenerhebung erfolgt durch Beobachtung bzw. Messung von Realisierungen interessierender Merkmale bzw. Sachverhalte.
Man unterscheidet einerseits zwischen

- einer primärstatistischen Datenerhebung: Die Daten werden explizit für eine vorliegende Fragestellung neu erhoben.
- einer sekundärstatistischen Erhebung: Die Daten sind bereits vorhanden und werden nur entnommen (z.B. aus Datenbanken, Registern).

und andererseits zwischen

- Vollerhebung: Messung des interessierenden Merkmals für jede statistische Einheit der Grundgesamtheit
- Teilerhebung: Messung des interessierenden Merkmals für eine Teilmenge (Stichprobe) der Grundgesamtheit.

Bemerke: Für viele Fragestellungen werden aus Zeit- oder Kostengründen 'nur' Teilerhebungen durchgeführt (Wahlumfragen, Qualitätskontrolle,...).

Frage: Lassen sich die Ergebnisse auf die Grundgesamtheit übertragen?

Datenerhebung

Beispiel 1.2. (Durchschnittsnote Klausur WiStat 2015–2018):

Schätzung der erreichten Durchschnittsnote in Wirtschaftsstatistik

- Kann man auf Daten des Studiensekretariats zurückgreifen
  $\Rightarrow$ sekundärstatistisch
- Führt man z.B. eine Umfrage unter (zufällig ausgewählten) Studierenden vor der Mensa durch
  $\Rightarrow$ primärstatistisch

Im letzteren Fall kann man i.d.R. 'nur' eine Teilerhebung durchführen.
Weiteres Problem: fehlende Daten (z.B. keine Antwort/weiss ich nicht)

Aufgabe 1.1.

Definieren Sie in diesem Beispiel die statistischen Einheiten, die Grundgesamtheit, die Merkmale und die Ausprägungen.

Statistische Versuchsplanung

Im Vorfeld sollte im Rahmen einer statistischen Versuchsplanung entschieden werden

- welches Ziel erreicht werden soll (unabdingbar: möglichst präzise und eindeutig formulierte Fragestellung)
- welche Größen dafür relevant sind (Zielvariablen) und durch welche Faktoren diese beeinflusst werden (Einflussfaktoren)
- Datenerhebung:
  - primärstatistisch: wie und in welchem Umfang man welche Daten/Merkmale erhebt
  - sekundärstatistisch: Daten/Merkmale vorhanden? Qualität?
- welche statistischen Methoden geeignet sind

Dabei gehen Fachwissen zur Fragestellung und statistische Kenntnisse Hand in Hand!

Beispiel 1.3. (Dozent und Lernerfolg)

- **Frage von Interesse:** Hat der Dozent einen Einfluss auf den Lernerfolg der Studierenden?
- Man sollte also in jedem Fall die Zielgröße 'Lernerfolg' (z.B. Klausurergebnis) und den Einflussfaktor 'Dozent' (wie?) erheben
- Allerdings: Obige Größen werden u.U. auch von anderen Faktoren beeinflusst (z.B. Erfahrung des Dozenten, gegenseitige Sympathie, Vorkenntnisse/Mitarbeit/Vorbereitung/Aufwand der Studierenden, etc.)
- Die Identifizierung aller relevanten Merkmale, die erhoben werden sollten, benötigt spezifisches Fachwissen
- Wichtig: Wurden zentrale Merkmale vergessen, so kann es zu Fehlinterpretationen kommen (Beispiel: Simpson's Paradoxon, mehr dazu später)
- Aus statistischer Sicht muss geklärt werden, wie viele Daten gesammelt werden müssen, um statistisch gesicherte Aussagen treffen zu können.
  Typisch dabei: Der benötigte Datenumfang wird größer, je mehr Faktoren erhoben werden und je mehr diese untereinander interagieren, sich verstärken oder abschwächen. Beispiel für Interaktion: Vorkenntnisse und Aufwand

Merkmalstypen: diskret vs. stetig

Man unterscheidet zwischen

- diskreten Merkmalen: Besitzen eine abzählbare Anzahl von möglichen Ausprägungen (ggf. ohne obere Schranke); treten z.B. bei Anzahlen oder Kategorisierungen auf.
  Beispiele: Geschlecht (m/w/d), Anzahl Geschwister, Anzahl Studierende in WiStat SS19, etc.
- stetigen Merkmalen: Können beliebig (überabzählbar) viele Ausprägungen annehmen.
  Beispiele: Alter, Körpergröße/-gewicht, Blutdruck, Einkommen, Geschwindigkeit, etc.

Bemerkung

Auch wenn viele Merkmale theoretisch beliebig genau bestimmt werden können, so werden sie häufig nur mit einer gewissen Genauigkeit gemessen (Größe: cm, Gewicht: kg, Alter: Jahre).
→ Man spricht hier deshalb auch von quasi-stetigen Merkmalen

Merkmalstypen nach Skalenniveau

Merkmale werden auf unterschiedlichen Skalen gemessen, die unterschiedliche mathematische Operationen zulassen:

- Nominalskala: Es gibt weder eine natürliche Ordnung der Ausprägungen noch ist es möglich, sinnvoll Abstände zu messen.
  Erlaubte Operationen: $=, \neq$
  z.B. Geschlecht, Farbe, Familienstand, Blutgruppe, etc.
- Ordinalskala: Es gibt eine natürliche Ordnung, aber Abstände lassen sich nicht sinnvoll interpretieren.
  Erlaubte Operationen: $=, \neq, <, >$
  z.B. Noten, Altersgruppe, Güteklassen, militärischer Rang, etc.

Merkmalstypen nach Skalenniveau

- Intervallskala: Ausprägungen (die Zahlen sind) lassen sich ordnen ihre Abstände haben sinnvolle Interpretation, aber i.A. willkürlich festgelegter Nullpunkt
  Erlaubte Operationen: $=, \neq, <, >, +, -$
  z.B. Kalenderdatum, IQ-Werte, etc.
- Verhältnisskala: Intervallskala mit festem, nicht willkürlichem Nullpunkt; Verhältnisse sind sinnvoll.
  Erlaubte Operationen: $=, \neq, <, >, +, -, *, /$
  z.B. Temperatur (Kelvin), Lebensalter, Nettomiete, Fläche, Volumen, etc.

Bemerkung:

Intervall- und Verhältnisskalenniveau werden häufig zum sog. Kardinalskalenniveau/metrischen Skalenniveau zusammengefasst.

Aufgabe 1.2.

Auf welcher Skala wird die Temperatur gemessen in $°C$ und $°F$ erhoben?

Merkmalstypen: qualitativ vs. quantitativ

- qualitativ: Ausprägung beschreibt ‘Qualität(stufe)’
  - endlich viele Ausprägungen
  - nominal-, höchstens ordinalskaliert
- quantitativ: Ausprägung beschreibt ‘Ausmaß/Intensität’
  - in der Regel kardinalskaliert
  - darunter fallen alle ‘Messungen’ im herkömmlichen Sinn (Werte sind Zahlen)

Bemerkung

Ordinalskalierte Daten haben häufig einen qualitativen Aspekt (Anordnung!), i.d.R. dominiert aber der qualitative Charakter → ‘Zwitterstellung’

Wie geht es weiter?

1. statistische Versuchsplanung ✓ + Datenerhebung ✓
2. Datenauswertung und -darstellung (deskriptiv)
   - Suche nach Kennzahlen zur 'Bewertung' der erhobenen Merkmale
   - unterschiedliche Kennzahlen für unterschiedliche Datentypen
3. Datenanalyse und -interpretation (induktiv)

- Statt formaler mathematischer Beweise wollen wir hier Statistik eher anwendungsorientiert gestalten:
  - praktische Beispiele
  - eigene Datenanalysen, z.B. anhand von selbst simulierten & realen Daten
- 'Spielen' Sie z.B. mit simulierten Daten 'herum', um ein Gefühl für praktische Datenanalysen zu bekommen!
- Während die reine Beschreibung der Daten keine Annahmen über die Datengewinnung erfordert, ist für die Analyse und -interpretation von großer Bedeutung, ob die Datenerhebung als ein Zufallsexperiment angesehen kann (z.B. Stichprobe der Studierendendaten enthält nur weibliche Teilnehmer → Systematik!)

Zusammenfassung

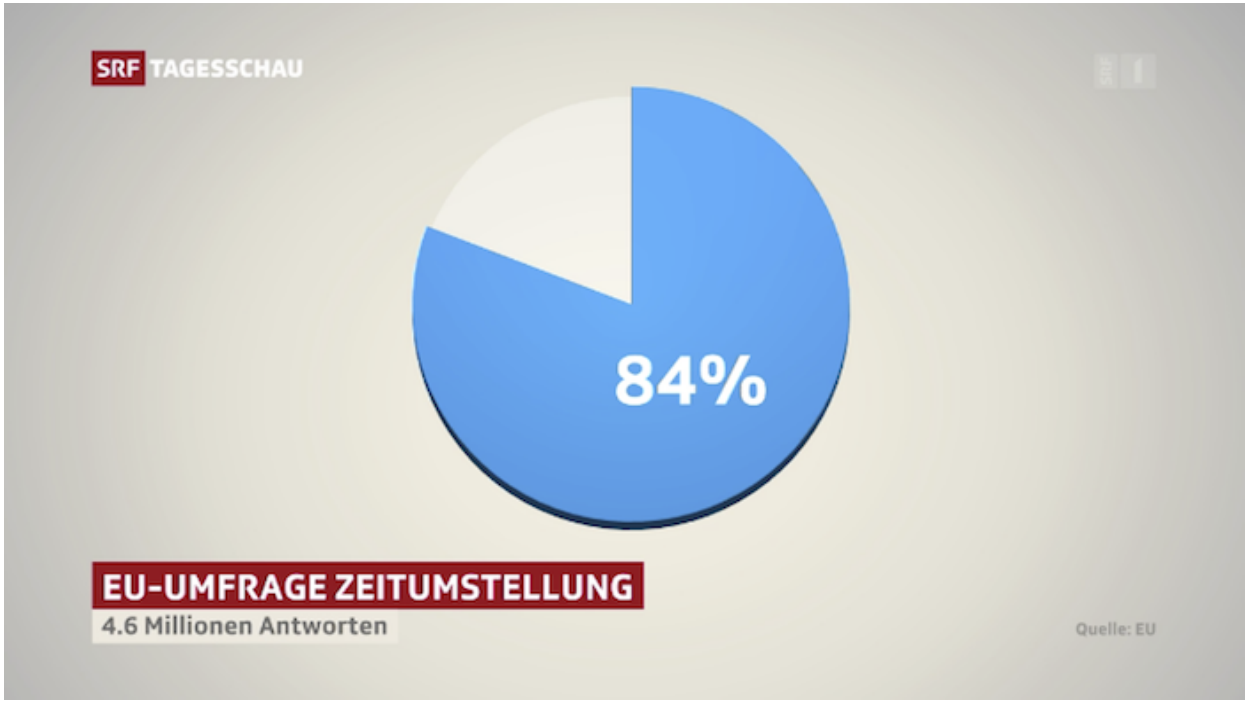
Unterschiedliche Skalenniveaus von Merkmalen

- Nominalskala: keine natürlich Ordnung
- Ordinalskala: Ordnung, aber keine Abstände
- Kardinalskala: Abstände, evtl. auch Verhältnisse

Unterschiedliche Ausprägungen von Merkmalen

- Qualitativ
- Quantitativ

‘Unstatistik’: 84% der EU-Bürger gegen Zeitumstellung!



*Quelle: `https://www.infosperber.ch`

Warum ist diese Interpretation ‘unseriös’?