Inferenzstatistik

Crashkurs

Inhaltsverzeichnis

Version vom 28. April 2022

1 Begriffe

Statistische Objekte, Merkmale und Ausprägungen

Die zu untersuchenden *statistische Objekte* (Personen, Pflanzen, Texte, ...) haben *Merkmale* mit jeweils unterschiedlichen *Ausprägungen*. Abhängig von den mathematischen Eigenschaften der Ausprägungen unterscheidet man zwischen *kategorialen* und *metrischen* Skalenniveaus.

Objekt	Merkmal	Ausprägung	Skalenniveau
Person	Blutgruppe	AB	kategorial
Person	Reaktionsgeschwindigkeit	1.5 s	metrisch
Stadt	mittlerer Ozonwert	$110\,\mu/m^3$	metrisch
Baum	Pilzbefall	schwach	kategorial

Kategoriale und metrische Skalen lassen sich jeweils noch in zwei weitere Skalenniveaus differenzieren aber für unsere Zwecke genügen die beiden oben genannten Typen.

Grundgesamtheit

„Als *Grundgesamtheit* bezeichnen wir allgemein alle potenziell untersuchbaren Einheiten [...], die ein gemeinsames Merkmal (oder eine gemeinsame Merkmalskombination) aufweisen." (Schuster Ch., S. 79)

- die Leserinnen einer bestimmten Zeitung
- die Menge aller dreisilbigen deutschen Substantive
- alle linkshändigen Schülerinnen und Schüler der Schweiz

Stichproben

- Eine *einfache Zufallsstichprobe* ist eine aus n Elementen bestehende Teilmenge der Grundgesamtheit, bei der jedes Element die gleichen Chance hat, in die Teilmenge aufgenommen zu werden.
- Eine *Klumpenstichprobe* besteht aus *allen* Untersuchungsobjekten, die sich in k zufällig ausgewählten Klumpen befinden.
- Sind die Faktoren bekannt, welche die Verteilung eines Merkmals beeinflussen, dann kann man versuchen, die Objekte (zufällig) so auszuwählen, dass der Anteil jedes Einflussfaktors in der Stichprobe mit dem in der Population übereinstimmt. Man spricht dann von einer *geschichteten Stichprobe*.
- Eine *Ad-hoc-Stichprobe* ist eine bereits bestehende Gruppe von Objekten (Schulklasse, Familie, Freundeskreis, ...) und ist für eine inferenzstatistische Auswertung ungeeignet.

Beispiel 1

Die Stichprobe S_1 in Abbildung 1 bildet die Ausprägungen eines Merkmals in der Population G gut ab, während die Stichprobe S_2 gewisse Ausprägungen überproportional und andere gar nicht berücksichtigt.

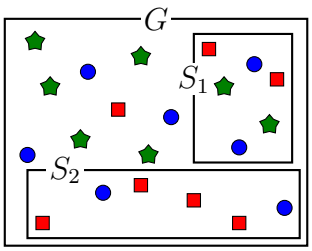


Abbildung 1: verschiedene Stichproben

Beschreibende (deskriptive) Statistik

Die *deskriptive Statistik* hat die Aufgabe, das Datenmaterial

- übersichtlich darzustellen (Tabellen),
- durch Kennzahlen zu charakterisieren und
- mittels Grafiken zu veranschaulichen.

Schliessende (induktive) Statistik

Die *induktive Statistik* hat zum Ziel, die aus Stichproben gewonnenen Beobachtungen auf die Grundgesamtheit zu verallgemeinern. Zwei Methoden stehen dabei im Vordergrund:

- die Überprüfung von Hypothesen durch Tests
- die Schätzung von Parametern

Grundgesamtheit und Stichprobe

Die folgende Grafik zeigt die Beziehung zwischen Grundgesamtheit und Stichprobe. Ist es vom Aufwand her möglich und vertretbar, die Grundgesamtheit zu untersuchen, so genügen die Methoden der deskriptiven Statistik, um die Population zu analysieren.

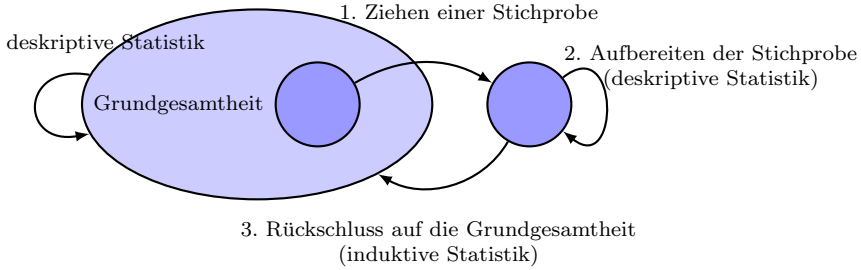


Abbildung 2: Vorgehen bei der statistische Analyse

2 Deskriptive Statistik

Populations- und Stichprobenparameter

- *Modus:* der am häufigsten auftretende Wert D
- *arithmetisches Mittel:* $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- *Varianz:* $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ und $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

  Die Stichprobenvarianz s^2 wird anders berechnet, damit sie die Populationsvarianz σ^2 besser schätzt.
- *Median:* Wert $\tilde{x} = x_{0.5}$, der die sortierte Liste in zwei gleich grosse Hälften teilt.
- *1. Quartil:* Wert $x_{0.25}$, der die unteren 50% der sortierten Daten in zwei gleich grosse Hälften teilt.
- *3. Quartil:* Wert $x_{0.75}$, der die oberen 50% der sortierten Daten in zwei gleich grosse Hälften teilt.
- *Interquartilsabstand:* IQR $= x_{0.75} - x_{0.25}$

Beispiel 2

Stichprobenwerte: 2, 9, 8, 4, 2

Ordnungsstatistik: 2, 2, 4, 8, 9

empirischer Mittelwert: $\overline{x} = \frac{2+2+4+8+9}{5} = 5$

empirische Varianz: $s^2 = \frac{(2-5)^2 + \cdots + (9-5)^2}{4} = 11$

1. Quartil: $x_{0.25} = 2$

Median: $\tilde{x} = 4$

3. Quartil: $x_{0.75} = 8.5$

Interquartilsabstand: IQR $= 6.5$

Modus: 2

Beispiel 3

Punktzahlen in einer Prüfung

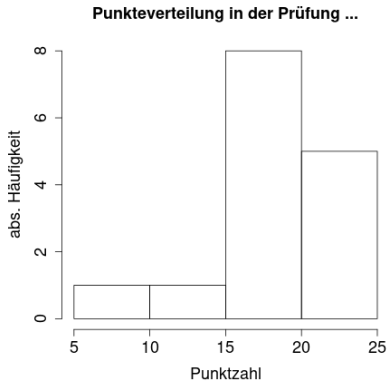
Punkte	m/f	Punkte	m/f	Punkte	m/f
18.5	f	16.0	m	20.0	f
7.5	m	19.0	f	21.0	f
16.0	m	12.5	f	17.0	f
22.5	f	19.0	m	24.5	m
22.5	m	17.0	m	21.0	m

Balkendiagramm

Die Häufigkeiten der Ausprägungen eines kategorialen Merkmals werden durch Rechtecke gleicher Breite über einer diskreten Skala dargestellt.

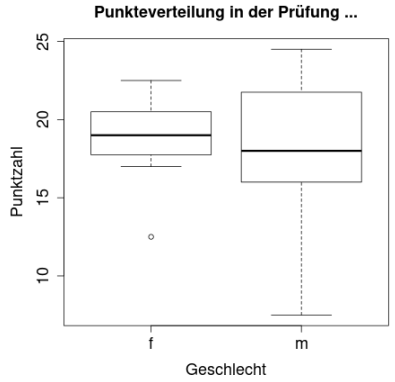
Histogramm

Stellt die Häufigkeiten der Ausprägungen eines metrisch skalierten Merkmals in Klassen (meist gleicher Breite) über einer kontinuierlichen Skala dar.



Box-and-Whiskers-Plot

Stellt metrisch oder geordnet kategorial skalierte Merkmale durch eine Box (die mittleren 50% um den Median) und Whiskers (die äussersten 25%) dar.



Werte, die mehr als das 1.5-fache des Interquartilsabstands (die Boxhöhe) vom 1. oder 3. Quartil entfernt sind, werden als *Ausreisser* bezeichnet.

3 Induktive Statistik

Einschränkung

Wir befassen uns hier mit einer sehr kleinen Auswahl statistischer Tests, bei denen die Stichprobendaten nur wenige Voraussetzungen erfüllen müssen.

Wissenschaftliche Hypothesen

Aussagen oder Schlussfolgerungen, die aus allgemeinen Theorien abgeleitet sind, werden als *Hypothesen* bezeichnet. *Beispiele:*

- Blinde haben überdurchschnittliche Fähigkeiten zur akustischen Reizdiskriminierung.
- Die mittlere „Milchleistung“ von Kühen verändert sich durch Beschallung mit leiser klassischer Musik.

Wichtig: die Hypothesen müssen sich empirisch überprüfen lassen.

Statistische Hypothesen

Um eine wissenschaftlichen Hypothese zu überprüfen, muss sie als statistische Hypothese formuliert werden. *Beispiel:*

wissenschaftliche Hypothese: Neue Erkenntnisse der Gehirnforschung lassen vermuten, dass Jugendliche eine Fremdsprache schneller lernen, wenn sie mit einer entsprechenden Lehrmethode unterrichtet werden.

statistische Alternativhypothese H_1: Werden Jugendliche nach der neuen Lehrmethode in einer Fremdsprache unterrichtet, so sind ihre durchschnittlichen Unterrichtsleistungen besser als bei Jugendlichen, die mit der herkömmlichen Methode unterrichtet werden.

Bezeichnen wir den Populationsmittelwert der bisherigen Unterrichtsleistungen mit μ_0 und den (unbekannten) Mittelwert bei der neuen Methode mit μ, so lautet die Kurzform von H_1: $\mu > \mu_0$

Gerichtete und ungerichtete Hypothesen

Postuliert eine Hypothese eine Veränderung in einer bestimmten Richtung (grösser oder kleiner) spricht man von einer *gerichteten* (einseitigen) Hypothese.

Ist nur von Interesse, ob ein Effekt eintritt (aber nicht in welche Richtung), spricht man von einer *ungerichteten* (zweiseitigen) Hypothese.

Wenn die der Hypothese zugrunde liegende Theorie einen gerichteten systematischen Effekt vorhersagt, dann sollte sich dies auch in der Formulierung widerspiegeln. Gibt es dazu keinen Anhaltspunkt, sollte grundsätzlich *zweiseitig* getestet werden.

Da ein systematischer Effekt bei einer einseitigen Hypothesen leichter nachzuweisen ist, als bei einer zweiseitigen Hypothese, ist es nicht erlaubt, eine zweiseitige Hypothese nachträglich in eine einseitige umzuwandeln.

Die Nullhypothese

In der Regel wird durch die Alternativhypothese ein (gerichteter oder ungerichteter) Unterschied gegenüber dem Status Quo ausgedrückt.

Die *Nullhypothese* postuliert nun, dass dieser Unterschied *nicht* vorhanden ist.

H_1: $\mu > \mu_0$ $\quad H_0$: $\mu = \mu_0$
H_1: $\mu < \mu_0$ $\quad H_0$: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$ $\quad H_0$: $\mu = \mu_0$

Hypothesentests

Ein Hypothesentest liefert immer eine Antwort auf die Frage, ob die experimentellen Daten mit der Nullhypothese verträglich sind. Wenn ja, wird die Nullhypothese *beibehalten*; andernfalls wird sie *verworfen*. Diese Entscheidung erfolgt auf der Grundlage der Wahrscheinlichkeit mit der die empirischen Daten im Einklang mit der Nullhypothese sind.

Fehler

Das Ergebnis eines Hypothesentest muss nicht korrekt sein. Die Stichprobe kann zu viele Objekte enthalten, welche für (oder gegen) die Alternativhypothese sprechen.

	Entscheidung für H_0	Entscheidung für H_1
in Population gilt H_0	Entscheidung richtig	Fehler 1. Art (α)
in Population gilt H_1	Fehler 2. Art (β)	Entscheidung richtig

Die Bedeutung der Fehlerarten

Im Beispiel mit der Lernmethode würde der Fehler 1. Art bedeuten, dass die Lernmethode nicht wirksamer ist als die herkömmliche und es würden unnötige Kosten für Umschulungen und neue Lehrmittel entstehen.

Ein Fehler 2. Art bedeutet, dass die neue Lehrmethode effektiver wäre, was aufgrund der Stichprobe aber nicht erkannt wird. In diesem Fall vergibt man eine Chance, den Unterricht zu verbessern.

Standardwerte für die Fehlerarten

Üblicherweise wird für α der Wert 0.05 gewählt. Das bedeutet, dass im Mittel 5 von 100 Stichproben zufällig ein so extremes Resultat zeigen, so dass die Nullhypothese verworfen wird, obwohl sie richtig ist.

Leider kann man diesen Fehler nicht beliebig klein wählen, weil dies die Wahrscheinlichkeit vergrössert, dass die Nullhypothese beibehalten wird, obwohl dies falsch ist (β-Fehler).

Der p-Wert

Vor dem grossflächigen Einsatz von Computern, musste man für jeden Test eine Prüfgrösse aus den Stichprobendaten berechnen. Anschliessend konnte man in einer Tabelle nachsehen, wie (un)wahrscheinlich es ist, einen solchen Wert zu erhalten, wenn man die Nullhypothese als gültig voraussetzt. Lag die betreffende Wahrscheinlichkeit unter dem Signifikanzniveau α, wurde die Nullhypothese verworfen, andernfalls beibehalten.

Aktuelle Computerprogramme berechnen die Prüfgrösse automatisch aber liefern uns die für die Entscheidung notwendige Information in einer anderen Form: *Sie sagen uns, wie wahrscheinlich es ist, dass das in der Stichprobe beobachtete Resultat (oder ein extremeres in Richtung der Hypothese) beobachtet wird, wenn man von der Gültigkeit der Nullhypothese ausgeht.* Diese Wahrscheinlichkeit wird *p-Wert* genannt.

Der Wilcoxon-Rangsummentest für zwei unabhängige Stichproben

Dieser Test prüft, ob sich die Rangdaten der Stichproben aus zwei Populationen systematisch unterscheiden. *Voraussetzungen:*

- die Stichprobenwerte lassen sich ordnen (rangieren)
- die Stichproben werden unabhängig gezogen
- für die Stichprobenumfänge gilt $n_1 \geq 3$ und $n_2 \geq 9$ (Schuster Ch., S. 599f.)

Beispiel 4

In einem Experiment wurden frisch geschlüpfte Hühnerküken zufällig in Gruppen eingeteilt und jede Gruppe mit einem anderen Futter ernährt (Kasein, Sojabohnen). Nach 6 Wochen wurde das Gewicht in Gramm gemessen.

(a) Stelle die Kükengewichte in einem Boxplot dar.

(b) Formuliere eine Hypothese in Bezug auf die Wirkung des Futters

(c) Prüfe die Hypothese mit dem Wilcoxon-Rangsummentest ($\alpha = 0.05$).

R-Code für Beispiel 4

```
# https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/chickwts.html

kasein <- c(359, 332, 222, 404, 368, 216, 260, 352, 379, 318)
soja <- c(158, 243, 267, 193, 199, 316, 248, 327, 271, 248)

boxplot(kasein, soja) # noch ohne korrekte Beschriftung

wilcox.test(kasein, soja, alternative="greater")
```

Eine typische Aussage zum Beispiel 4

Das Gewicht der Küken ist nach 6 Wochen signifkant höher, wenn sie mit Kasein anstelle von Leinsamen gefüttert werden (Wilcoxon-Rangsummentest, $n_1 = n_2 = 10$, $W = 81$, $p = 0.01054$, $\alpha = 0.05$).

Der Wilcoxon-Test für zwei verbundene Stichproben

Dieser Test berechnet zunächst für jedes Objekt die Differenz der beiden Stichprobenwerte (z. B. vorher − nachher). Anschliessend wird geprüft, ob sich die Ränge der positiven Differenzen und die Ränge der Absolutbeträge der negativen Differenzen systematisch unterscheiden. *Voraussetzungen:*

- Die Stichprobenwerte erlauben die Bildung von Differenzen.
- Die Stichproben*paare* sind voneinander unabhängig.
- Für die Anzahl Paare sollte $n \geq 6$ gelten (Schuster Ch., S. 500).

Beispiel 5

Eine Lehrerin hat nach einer Klausur den Eindruck, dass die Klasse den geprüften Stoff noch nicht gut genug beherrscht. Daher repetiert sie das Thema und lässt eine zweite Prüfung darüber schreiben.

(a) Analysiere die Notenverteilung beider Prüfungen mit der Funktion `summary()`.

(b) Stelle jeden Schüler als Punkt in einem xy-Diagramm dar, wobei x die Note der ersten Prüfung und y die Note der zweiten Prüfung ist.

(c) Untersuche mit dem Wilcoxon-Test für verbundene Stichproben, ob die Schüler in der zweiten Prüfung signifikant besser waren als in der ersten ($\alpha = 0.05$).

R-Code für Beispiel 5

```
p1 <- c(3.9, 4.1, 5.0, 4.5, 4.5, 4.3, 4.9, 3.1, 5.3, 3.9, 3.9, 4.3)
p2 <- c(4.2, 4.5, 5.8, 5.4, 5.0, 4.0, 4.7, 4.0, 6.0, 4.9, 3.3, 4.5)

summary(p1)
summary(p2)

plot(p1, p2)
abline(a=0,b=1) # Gerade y=b*x+a

wilcox.test(p1, p2, alternative="less", paired=TRUE)
```

Eine typische Aussage zum Beispiel 5

Die Leistungen der Schülerinnen und Schüler haben sich durch die Repetition signifkant verbessert (Wilcoxon-Test, $n = 12$, $V = 11.5$, $p = 0.01705$, $\alpha = 0.05$).

Beispiel 6

Der Verhaltensforscher Onur Güntürkün hat im Jahr 2003 einen Artikel in der Fachzeitschrift *Nature* veröffentlicht, in dem er die Hypothese äussert, dass eine leichte Tendenz, den Kopf nach rechts zu bewegen, die bei Säuglingen während einer kurzen Phase vor und nach der Geburt beobachtet wird, später beim Küssen eines Liebespartners wieder erkennbar werde.

Alternativhypothese H_1: Küssen sich Liebespaare, so ist dabei eine Tendenz zu beobachten, den Kopf nach rechts zu drehen.

Nullhypothese H_0: Küssen sich Liebespaare, so ist dabei keine Tendenz zu beobachten, den Kopf in eine bestimmte Richtung zu drehen.

Das Experiment

„... I observed kissing couples in public places (international airports, large railway stations, beaches and parks) in the United States, Germany and Turkey. The headturning behaviour of each couple was recorded for a single kiss, with only the first being counted in instances of multiple kissing. The following criteria had to be met to qualify: lip contact, face-to-face positioning, no hand-held objects (as these might induce a side preference), and an obvious head-turning direction during kissing. Subjects' ages ranged from about 13–70 years. Of 124 kissing pairs, 80 (64.5%) turned their heads to the right and 44 (35.5%) turned to the left. ...“ (Güntürkün, 2003).

Der Binomialtest für eine Proportion

Hier wird das Modell zugrunde gelegt, dass sich jedes Objekt (das sich küssende Paar) unabhängig von anderen mit einer festen (aber unbekannten) Wahrscheinlichkeit π „entscheidet“ den Kopf nach rechts zu drehen.

Sofern keine anderen Daten verfügbar sind, werdendet man für die Nullhypothese H_0 die Wahrscheinlichkeit $\pi_0 = 0.5$.

Mit dem Test wird geprüft, ob die in der Stichprobe beobachtete relative Häufigkeit mit der in der Nullhypothese postulierten Proportion (Wahrscheinlichkeit) verträglich ist. Dafür muss das Merkmal genau zwei Ausprägungen haben (Schwarz J.).

Der R-Code für Beispiel 6

```
binom.test(80, 124, p=0.5, alt="greater")

# Achtung: p und p-Value sind nicht dasselbe!
```

Eine typische Aussage zum Beispiel 6

Die beobachteten Paare drehen sich beim Küssen signifikant häufiger nach rechts als nach links (Binomialtest, $n = 124$, $x = 80$, $p = 0.00078$, $\alpha = 0.05$).

Bemerkungen

Die Hypothese von Beispiel 6 ist seither Gegenstand kontroverser Diskussionen (siehe z. B Jennifer R. Sedgewick).

Da Wiederholungen des Experiments in unterschiedlichen Regionen der Erde zu unterschiedlichen Ergebnissen geführt haben, vermuten einige Autoren auch einen Zusammenhang mit kulturellen Gepflogenheiten, wie beispielsweise der Schreibrichtung.

Die Statistik-Software `R`

`R` ist ein freies Softwarepaket für statistische Berechnungen und Graphiken und für alle gängigen Betriebssysteme verfügbar (R Core Team).

Literatur

[Gün03] O. Güntürkün. "Human behaviour: Adult persistence of head-turning asymmetry". In: *Nature* 421 (2003), S. 421.

[Jen19] Lorin J. Elias Jennifer R. Sedgewick Abby Holtslander. "Kissing Right? Absence of Rightward Directional Turning Bias During First Kiss Encounters Among Strangers". In: *Journal of Nonverbal Behavior* 43 (2019), S. 271–282.

[R C22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: `https://www.R-project.org`.

[Sch10] Bortz J. und Schuster Ch. *Statistik für Human- und Sozialwissenschaftler*. 7. Auflage. Springer, 2010.

[Sch22] Bruderer Enzler H. Schwarz J. Käch W. *Methodenberatung*. März 2022. URL: `https://www.methodenberatung.uzh.ch/de.html`.