Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Stichprobentheorie

Zufallsstichprobe

Eine Population wird durch Annahmen über Zufallsvariablen bzw. deren Verteilung und Parameter (Modell) charakterisiert

Damit man mit Hilfe von empirisch erhobenen Werten Aussagen über die Population machen kann, müssen die Werte durch Zufallsexperimente gewonnen werden, mit denen die Zufallsvariablen eindeutig festgelegt werden

Die derartig erhobenen Werte $x_1, \ldots, x_n$ heißen *Zufallsstichprobe* bzw. *Stichprobe*

Mit n wird der *Umfang* der Stichprobe bezeichnet

Die Zufallsvariable, die bei Durchführung des Zufallsexperiments den Stichprobenwert x_i (Realisierung) liefert, bezeichnen wir mit X_i $(i = 1, \ldots, n)$

Zufallsstichprobe

Eine Stichprobe $x_1, \ldots, x_n$ heißt *unabhängig*, wenn die Zufallsvariablen $X_1, X_2, \ldots, X_n$ stochastisch unabhängig sind

Eine Stichprobe heißt *einfach*, wenn die Zufallsvariablen $X_1, X_2, \ldots, X_n$ stochastisch unabhängig sind und alle dieselbe Verteilung besitzen

- Man spricht in diesem Fall auch von unabhängig identisch verteilten Zufallsvariablen (independent and identically distributed, i.i.d.)

Beispiel

- Die Indikatorvariable X_i kennzeichne eine "Ja"- bzw. "Nein"-Antwort bei der wiederholten Präsentation einer Reizsituation (x, s) im Rahmen der Konstanzmethode
- Die Erfüllung der Bedingungen einer einfachen Stichprobe ist durch versuchsplanerische Maßnahmen sicher zu stellen

Verteilung von Stichprobenkennwerten

Die Zufallsvariablen $X_1, X_2, \ldots, X_n$ seien unabhängig identisch verteilt mit Erwartungswert μ und Varianz σ^2

Wir betrachten das arithmetische Mittel

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

Dann kann man unmittelbar den Erwartungswert und die Varianz des arithmetischen Mittels $\bar{X}$ ableiten

Stichprobenverteilung des Mittelwerts

Erwartungswert von $\bar{X}$

$$
\begin{aligned}
E(\bar{X}) &= E\left(\frac{X_1 + \ldots + X_n}{n}\right) \\
&= \frac{1}{n} \cdot [E(X_1) + \ldots + E(X_n)] \\
&= \frac{1}{n} \cdot [\underbrace{\mu + \ldots + \mu}_{n\text{-mal}}] \\
&= \frac{1}{n} \cdot n \cdot \mu \\
&= \mu
\end{aligned}
$$

Stichprobenverteilung des Mittelwerts

Varianz von $\bar{X}$

$$
\begin{aligned}
Var(\bar{X}) &= Var\left(\frac{X_1 + \ldots + X_n}{n}\right) \\
&= \frac{1}{n^2} \cdot [Var(X_1) + \ldots + Var(X_n)] \\
&= \frac{1}{n^2} \cdot [\underbrace{\sigma^2 + \ldots + \sigma^2}_{n\text{-mal}}] \\
&= \frac{1}{n^2} \cdot n \cdot \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Stichprobenverteilung des Mittelwerts

Es gilt also

$$E(\bar{X}) = \mu \quad \text{und} \quad Var(\bar{X}) = \frac{\sigma^2}{n} \quad \text{bzw.} \quad SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Die Standardabweichung eines Stichprobenkennwertes wird auch als *Standardfehler* dieser Statistik bezeichnet

Sind Zufallsvariablen $X_1, X_2, \ldots, X_n$ unabhängig identisch normalverteilt gemäß $N(\mu, \sigma^2)$, dann ist auch das arithmetische Mittel $\bar{X}$ normalverteilt (Statistik I), und zwar gemäß $N(\mu, \sigma^2/n)$

Standardfehler des Mittelwerts in R

Wir ziehen (100-fach wiederholt) aus einer Normalverteilung mit $\mu = 0$ und $\sigma = 10$ insgesamt 25 Werte und weisen den hieraus berechneten Mittelwert dem Vektor `xq` zu (besteht aus 100 Elementen)

```
xq <- replicate(100,
                rnorm(n = 25, mean = 0, sd = 10) |> mean()
)
```

Ausgabe von Mittelwert und Standardabweichung der erhaltenen Mittelwerte

```
mean(xq)
sd(xq)
```

Stichprobenverteilung der empirischen Varianz einer normalverteilten Population

Wir betrachten für unabhängige und identisch $N(\mu, \sigma^2)$-verteilte Zufallsvariablen $X_1, \ldots, X_n$ die empirische Varianz

$$\tilde{S}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Dann ist

$$\frac{n \cdot \tilde{S}^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

$\chi^2(n-1)$-verteilt

Die Stichprobenverteilung der empirischen Varianz (normiert bezüglich σ^2 und n) einer normalverteilten Population ist also eine Chi-Quadrat-Verteilung mit $n-1$ Freiheitsgraden

Regression zur Mitte

Als "Regression zur Mitte" (eigentlich genauer "Regression zur größten Dichte") wird ein statistisches Artefakt bezeichnet, das häufig als inhaltlicher Effekt fehlinterpretiert wird

Phänomen

- Es werden zwei verbundene Messungen x und y betrachtet (z. B. wiederholte Testung eines Probanden, Erhebung von Körpergröße, Intelligenz, ... bei Eltern-Kind-Paaren)
- Liegt der Wert x über (bzw. unter) dem Mittelwert, dann wird der Wert y wahrscheinlich näher zum Mittelwert liegen als x
- Das Ausmaß dieser Regression zur Mitte wird durch eine große Abweichung des Werts x vom Mittelwert und einen geringen Zusammenhang beider Messungen begünstigt

Regression zur Mitte

Der Begriff geht auf Francis Galton (1822-1911) zurück, der 1877 die Größe einer bestimmten Sorte von Erbsen untersuchte

- Der Durchmesser der Erbsen erwies sich als normalverteilt
- Kategorisierung der Erbsen und Bestimmung des Durchmessers der Nachkommengeneration lieferte folgendes Ergebnis (Einheit: 1/100 Zoll)

Eltern	15	16	17	18	19	20	21
Nachkommen (Mittelwert)	15.4	15.7	16.0	16.3	16.6	17.0	17.3

- Galton fand heraus, dass sich die gegeneinander abgetragenen Mittelwerte beider Generationen durch eine Gerade verbinden ließen: die erste Regressionsgerade

Regression zur Mitte

Galton beschrieb das Phänomen folgendermaßen

> *Reversion is the tendency of the ideal mean filial type to depart from the parental type, reverting to what may be roughly and perhaps fairly described as the average ancestral type*

Regression zur Mitte konnte Galton auch nachweisen bei

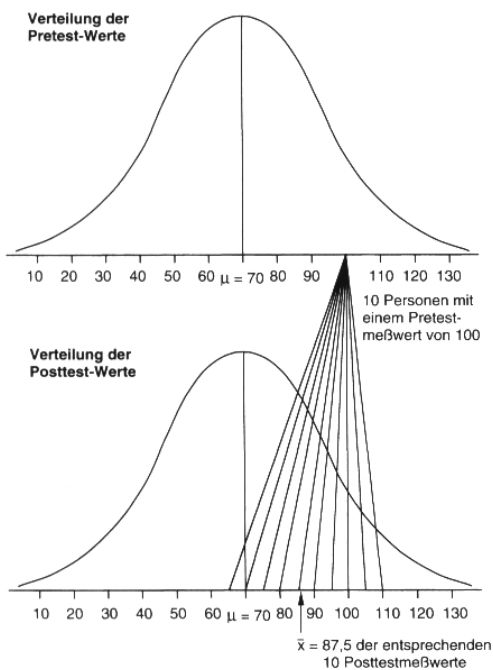
- Verteilung der Körpergröße bei Eltern und ihren Kindern
- Begabung von Genies und ihren Nachkommen

Man beachte, dass das Phänomen unabhängig von einer zeitlichen Orientierung auftritt

- Die Körpergröße der Eltern extrem großer Söhne/Töchter liegt tendenziell näher zum Mittelwert

Beispiel

Pretest-Postest
bzw.
Test-Retest



Regression zur Mitte: Erklärung

Seien X und Y zwei gemeinsam normalverteilte Zufallsvariablen mit $\mu_x = \mu_y = 0$ und $\sigma_x^2 = \sigma_y^2$, deren bivariate Korrelation ρ beträgt

Für normalverteilte Zufallsvariablen ist der (bedingte) Erwartungswert von Y gegeben einen Wert von X eine lineare Funktion von X, es gilt nämlich

$$E(Y \mid X) = \rho \cdot X$$

Wegen $|\rho| \leq 1$ liegt der erwartete Wert für Y im Vergleich zum beobachteten Wert X näher am Erwartungswert 0

Ein entsprechendes Ergebnis kann man auch für allgemeinere Klassen von Verteilungen ableiten

Regression zur Mitte: Beispiel 1

The Massachusetts Dept. of Education committed what appears to be a howling statistical blunder yesterday. It would be funny if not for the millions of dollars, thousands of hours of work, and thousands of students' lives that could be affected. Massachusetts has implemented a state-wide mandatory student testing program, called the MCAS. Students in the 4th, 8th and 10th grades are being tested and next year 12th grade students must pass the MCAS to graduate. The effectiveness of school districts is being assessed using average student MCAS scores. Based on the 1998 MCAS scores, districts were placed in one of 6 categories: very high, high, moderate, low, very low, or critically low. Schools were given improvement targets based on the 1998 scores, with schools in the highest two categories were expected to increase their average MCAS scores by 1 to 2 points, while schools in the lowest two categories were expected to improve their scores by 4-7 points. Based on the average of 1999 and 2000 scores, each district was evaluated yesterday on whether they had met their goals.

Regression zur Mitte: Beispiel 1

The report was posted on the MA Dept. of Education web site: Those familiar with "regression to the mean" know what's coming next. The poor schools, many in urban centers like Boston, met their improvement "targets", while most of the state's top school districts failed to meet their improvement targets. (Dr. Eugene D. Gallagher, ECOS, UMASS/Boston)

Regression zur Mitte: Beispiel 2

From Daniel Kahneman's speech when he won the Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel 2002

> *I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning. When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand and made his own short speech, which began by conceding that positive reinforcement might be good for the birds, but went on to deny that it was optimal for flight cadets. He said, "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not, because the opposite is the case."*

Regression zur Mitte: Beispiel 2

From Daniel Kahneman's speech when he won the Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel 2002

> *This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them. . . .*
>
> *. . . I immediately arranged a demonstration in which each participant tossed two coins at a target behind his back, without any feedback. We measured the distances from the target and could see that those who had done best the first time had mostly deteriorated on their second try, and vice versa. But I knew that this demonstration would not undo the effects of lifelong exposure to a perverse contingency.*

Parameterschätzung

Übersicht

- Punktschätzung
- Güte von Schätzfunktionen
- Verfahren zur Parameterschätzung
- Intervallschätzung

Punktschätzung

Schätzverfahren zielen darauf ab, aus Stichprobenkennwerten Informationen über die als zugrunde liegend angenommene Population zu gewinnen

Insbesondere will man hieraus möglichst präzise Näherungswerte ableiten für Populationsparameter

- Kennwerte der angenommenen Verteilung
  z. B. Erwartungswert, Varianz, ...
- Spezifische Parameter der angenommenen Verteilung
  - Parameter μ und σ^2 der Normalverteilung
  - Parameter p der Binomialverteilung
  - Parameter λ der Exponentialverteilung

Ausgangspunkt

Die Population wird durch ein statistisches (daten-erzeugendes) Modell repräsentiert, das im Allgemeinen Annahmen zur Verteilung des betrachteten Merkmals in der Grundgesamtheit macht

- Wir machen die vereinfachende Annahme, dass die Verteilungsfunktion einer Zufallsvariablen X von einem *Parameter* ϑ abhängt

Über die Durchführung eines Zufallsexperiments wird eine (einfache) Stichprobe $x = (x_1, x_2, \ldots, x_n)$ gewonnen

- Aus den n Stichprobenwerten $x_1, x_2, \ldots, x_n$ soll dann durch eine geeignete Formel ein Schätzwert $\hat{\vartheta}$ für den unbekannten Parameter ϑ berechnet werden

Schätzwert und Schätzfunktion

Dieser *Schätzwert* ist eine Funktion t_n der Werte $x_1, x_2, \ldots, x_n$, d. h.

$$\hat{\vartheta} = t_n(x_1, x_2, \ldots, x_n)$$

Der Funktionswert $t_n(x_1, x_2, \ldots, x_n)$ wird dabei als Realisierung einer Zufallsvariablen

$$T_n = t_n(X_1, X_2, \ldots, X_n)$$

betrachtet, die als *Schätzfunktion* bezeichnet wird

Die Verteilungsfunktion der Zufallsvariablen T_n (sampling distribution) hängt über die $X_1, X_2, \ldots, X_n$ dann ebenfalls vom Parameter ϑ ab

Beispiele für Schätzfunktionen

Das arithmetische Mittel

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

als Schätzer für den Erwartungswert $\mu = E(X)$

Für $X_i(\Omega) = \{0, 1\}$ die relative Häufigkeit

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

als Schätzer für die Auftretenswahrscheinlichkeit $\pi = P(X = 1)$

Beispiele für Schätzfunktionen

Die empirische Varianz

$$\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

als Schätzer für die Varianz $\sigma^2 = Var(X)$

Die Stichprobenvarianz

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

als Schätzer für die Varianz $\sigma^2 = Var(X)$

Güte von Schätzfunktionen

Obwohl intuitiv einleuchtend ist, dass $\bar{X}$ eine geeignete Schätzfunktion für den Erwartungswert ist, gilt es klare Kriterien für die Güte von Schätzfunktionen bereit zu stellen

Zu diesem Zweck betrachten wir folgende Eigenschaften

- Erwartungstreue
- Konsistenz
- Effizienz
- Suffizienz

Erwartungstreue

Eine Schätzfunktion sollte tendenziell den richtigen Wert liefern, d. h. der Parameterwert sollte weder systematisch über- noch unterschätzt werden

Diese Eigenschaft wird als *Erwartungstreue* bezeichnet

Eine Schätzfunktion $T_n = t_n(X_1, X_2, \ldots, X_n)$ für den Parameter ϑ heißt *erwartungstreu* (unverzerrt, engl. unbiased), wenn für ihren Erwartungswert gilt

$$E(T_n) = E(t_n(X_1, X_2, \ldots, X_n)) = \vartheta$$

Bestimmt man also den Erwartungswert von T_n unter der Bedingung, dass der Parameter ϑ zugrunde liegt, ergibt sich bei einer erwartungstreuen Schätzfunktion als Erwartungswert gerade ϑ

Beispiele

Haben die Zufallsvariablen $X_1, \ldots, X_n$ alle denselben Erwartungswert $\mu = E(X_i)$ für alle $i = 1, \ldots, n$, dann ist die Zufallsvariable

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

eine erwartungstreue Schätzfunktion für den Parameter μ, denn

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\, \mu = \mu$$

Beispiele

Haben die unabhängigen Zufallsvariablen $X_1, \ldots, X_n$ identische Erwartungswerte $E(X_i) = \mu$ und Varianzen $Var(X_i) = \sigma^2$ für alle $i = 1, \ldots, n$, dann ist die empirische Varianz

$$\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

keine erwartungstreue Schätzfunktion für den Parameter σ^2

Hierzu betrachten wir

$$E(\tilde{S}^2) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 \right)$$

Wegen stochastischer Unabhängigkeit erhalten wir

$$E(\tilde{S}^2) = \frac{1}{n} \sum_{i=1}^{n} E(X_i^2) - E(\bar{X}^2)$$

Beispiele

Darüberhinaus folgt aus der Definition der Varianz

$$E(X_i^2) = \sigma^2 + \mu^2 \quad \text{für alle } i = 1, \ldots, n$$

$$E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2$$

Eingesetzt ergibt sich dann

$$\begin{aligned} E(\tilde{S}^2) &= \frac{1}{n} \sum_{i=1}^{n} (\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2) \\ &= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 \\ &= \frac{n-1}{n} \sigma^2 \end{aligned}$$

Damit gilt

$$E(\tilde{S}^2) \neq \sigma^2$$

Bias

Eine systematische Über- bzw. Unterschätzung des Parameters durch eine Schätzfunktion wird durch den Begriff des *Bias* erfasst, der definiert ist durch

$$Bias(T_n) = E(T_n) - \vartheta$$

Beispiel

- Mit

$$E(\tilde{S}^2) = \frac{n-1}{n}\,\sigma^2$$

folgt für den Bias der empirischen Varianz

$$Bias(\tilde{S}^2) = E(\tilde{S}^2) - \sigma^2 = \frac{n-1}{n}\,\sigma^2 - \sigma^2 = -\frac{1}{n}\,\sigma^2$$

Bias

Eine Korrektur der systematischen Verzerrung ist in diesem Fall einfach zu erreichen, denn aus

$$E(\tilde{S}^2) = \frac{n-1}{n}\,\sigma^2$$

folgt

$$E(\frac{n}{n-1}\,\tilde{S}^2) = \sigma^2$$

Als erwartungstreue Schätzfunktion für die Varianz σ^2 erhält man also gerade die Stichprobenvarianz

$$\frac{n}{n-1}\,\tilde{S}^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = S^2$$

Asymptotische Erwartungstreue

Eine abgeschwächte Forderung an die Schätzfunktion resultiert, wenn Erwartungstreue nur für hinreichend große Stichprobenumfänge gelten soll

- Eine Schätzfunktion T_n heißt *asymptotisch erwartungstreu*, wenn gilt

$$\lim_{n\to\infty} E(T_n) = \vartheta$$

Beispiel

- Wegen $\lim_{n\to\infty} \frac{n-1}{n} = 1$ ist die empirische Varianz asymptotisch erwartungstreu, denn es folgt

$$\lim_{n\to\infty} E(\tilde{S}^2) = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$$

Bias der empirischen Varianz

Der Bias der empirischen Varianz $\tilde{S}^2$ als Schätzfunktion für die Varianz σ^2 wirkt sich also insbesondere bei vergleichsweise kleinen Stichprobenumfängen n aus

Bei hinreichend großen Stichproben vermindert sich das Ausmaß der systematischen Unterschätzung zunehmend

Man beachte, dass die Standardabweichung $S = \sqrt{S^2}$ *keine* erwartungstreue Schätzfunktion für σ ist

Der Faktor zur Korrektur des Bias ist abhängig von der Verteilung und lautet für eine normalverteilte Population und hinreichend großem Stichprobenumfang n

$$\left(1 + \frac{1}{4\,(n-1)}\right) \cdot S$$

Beispiel: Bias der empirischen Varianz

Es wird ein Vektor `v` angelegt, der die berechneten Schätzwerte aufnimmt

```
v <- numeric(100)
```

100-fach wiederholte Ziehung einer Stichprobe vom Umfang $n = 10$ aus einer $N(0, 10^2)$-Verteilung und Zuweisung der berechneten empirischen Varianz auf die Komponente `v[i]` des Vektors (`i` läuft von 1 bis 100)

```
for(i in seq_along(v)) {
  x <- rnorm(n = 10, mean = 0, sd = 10)
  v[i] <- sum((x - mean(x))^2)/10
}
```

Beispiel: Bias der empirischen Varianz

Ausgabe des Mittelwerts der erhaltenen Schätzwerte

```
mean(v)
```

Korrektur des Bias der empirischen Varianz durch Faktor $\frac{n}{n-1}$

```
(10/9)*mean(v)
```

Bias der empirischen Varianz für Stichprobenumfänge $n = 1, \ldots, 50$

```
curve(-(1/n)*10^2, 1, 50, xname = "n")
```