4: Schätzstatistik

Statistische Verfahren in der Geographie

Till Straube <straube@geo.uni-frankfurt.de>
Institut für Humangeographie
Goethe-Universität Frankfurt

1 Lernziele dieser Sitzung

Sie können...

- eine Punktschätzung für μ und σ durchführen.
- den Standardfehler der Stichprobenverteilung von \bar{x} bestimmen.
- eine Intervallschätzung für μ durchführen.

2 Stichprobenverteilung

Die Stichprobenverteilung ist eine theoretische Verteilung, welche die möglichen Ausprägungen eines statistischen Kennwertes (z.B. \bar{x}) sowie deren Auftretenswahrscheinlichkeit beim Ziehen von Zufallsstichproben des Umfanges n beschreibt. (Bortz und Schuster 2010: 83)

Hier ist zunächst die theoretische Verteilung des Mittelwerts einer Stichprobe relevant. Insbesondere interessiert uns, wie sich die theoretische Verteilung des Mittelwerts abhängig von der Stichprobengröße verhält.

2.1 Szenario 1: Normalverteilte Grundgesamtheit

Die Grundgesamtheit (Population) einer Variable x sei normalverteilt mit $\mu=50$ und $\sigma^2=25$. Wir können also schreiben:

$$x \sim N(50,25)$$

Die Standardabweichung der Population beträgt entsprechend:

$$\sigma = \sqrt{\sigma^2}$$
$$= \sqrt{25} = 5$$

Graphisch ist die Dichtefunktion der Verteilung in Abbildung 1 veranschaulicht.

Wenn eine einzelne Stichprobe der Größe n=3 aus dieser Verteilung gezogen würde, hätte sie drei konkrete Werte $(x_1, x_2 \text{ und } x_3)$ sowie ein konkretes arithmetisches Mittel (\bar{x}) .

Stand: 11. Mai 2019 1/11

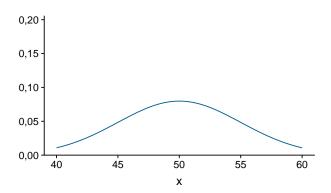


Abbildung 1: Dichtefunktion der Grundgesamtheit

Es lässt sich jedoch auch eine Wahrscheinlichkeitsdichtefunktion der Mittelwerte aller theoretisch möglichen Stichproben der Größe n=3 (und zusätzlich der Größe n=6) zeichnen (s. Abbildung 2).

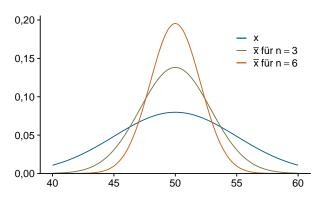


Abbildung 2: Dichtefunktionen der Stichprobenverteilungen

2.1.1 Erwartungswert

Es fällt auf, dass die Stichprobenverteilungen für \bar{x} normalverteilt sind und um das arithmetische Mittel der Grundgesamtheit (μ) symmetrisch sind.

Das arithmetische Mittel der Stichprobenverteilung $\mu_{\bar{x}}$ wird auch als **Erwartungswert** (engl. *expected value*) von \bar{x} bezeichnet. Es gilt:

$$\mu_{\bar{x}} = \mu \tag{1}$$

Wir können auch sagen: \bar{x} ist ein "ertwartungstreuer" Schätzparameter für μ ; nicht weil er in der Empirie zwangsläufig identisch mit μ wäre, sondern weil er mit zunehmender Stichprobengröße immer stärker zu μ tendiert.

2.1.2 Standardfehler

Zusätzlich fällt in Abbildung 2 auf: Je größer die Stichprobe, desto gestauchter die Dichtekurve der Stichprobenverteilung: Die theoretische Verteilung von \bar{x} bei n=6 weist eine kleinere Varianz auf als

Stand: 11. Mai 2019 2/11

bei n=3. Das ist einigermaßen intuitiv, denn wir können uns vorstellen, dass das arithmetische Mittel \bar{x} bei steigender Stichprobengröße ein immer präziserer Schätzwert für μ wird.

Die Varianz der Stichprobenverteilung für \bar{x} bezeichnen wir mit σ_x^2 . Sie hängt von der Varianz der Population ab und ist invers proportional zur Stichprobengröße. Es gilt:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \tag{2}$$

Die Standardabweichung der Stichprobenverteilung ($\sigma_{\bar{x}}$) wird auch Standardfehler (engl. *standard error*) genannt. Durch Wurzelziehen ergibt sich:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{3}$$

Zusammenfassend lässt sich sagen:

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$
 für $x \sim N(\mu, \sigma^2)$ (4)

2.2 Szenario 2: Nicht normalverteilte Grundgesamtheit

Die Gleichungen 1, 2 und 3 gelten uneingeschränkt auch für die Stichprobenverteilungen von nicht normalverteilten Populationen. Nur die Normalverteilung der Stichprobenverteilung (Gleichung 4) ist bei nicht normalverteilten Grundgesamtheiten grundsätzlich nicht gegeben.

Das zentrale Grenzwerttheorem (engl. central limit theorem) besagt jedoch:

Die Verteilung von Mittelwerten aus Stichproben des Umfangs n, die derselben Grundgesamtheit entnommen wurden, geht mit wachsendem Stichprobenumfang in eine Normalverteilung über. (Bortz und Schuster 2010: 86)

Abbildung 3 veranschaulicht diesen Effekt für eine nicht normalverteilte Grundgesamtheit.

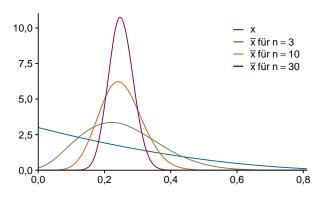


Abbildung 3: Stichprobenverteilung bei nicht normalverteilter Population

In der Praxis gilt die Faustregel: Ab einer Stichprobengröße von n=30 können wir statistische Verfahren anwenden, die von einer theoretischen Normalverteilung von \bar{x} ausgehen – und zwar *unabhängig* von der Verteilung der Grundgesamtheit.

Stand: 11. Mai 2019 3/11

3 Punktschätzung

Bei statistischen Untersuchungen geht es oft darum, ausgehend von der empirischen Verteilung einer Stichprobe auf Parameter der Grundgesamtheit zu schließen.

Die Punktschätzung (engl. *point estimation*) ist dabei eine vergleichsweise einfache und intuitive Vorgehensweise.

3.1 Punktschätzung des arithmetischen Mittels

Wenn eine Stichprobe vorliegt, dann ist ihr arithmetisches Mittel (\bar{x}) als erwartungstreuer Punktschätzer der wahrscheinlichste Wert für das arithmetische Mittel der Grundgesamtheit (μ).

Beispiel:

- Zehn Studierende der Humangeographie werden zufällig ausgewählt, um ihre Pendelzeit zum IG-Farben-Campus zu erfassen.
- Die Angaben in Minuten lauten: 22 26 12 23 48 31 15 71 17 35
- Das arithmetische Mittel der Messreihe lässt sich wie in Sitzung 2 ausführlich besprochen berechnen: $\bar{x}=30$
- Da es sich um eine erwartungstreue Schätzgröße (und eine valide Zufallsstichprobe) handelt, kann die durchschnittliche Pendelzeit aller Studierenden der Humangeographie auf $\mu=30$ Minuten geschätzt werden.

Gleichzeitig wissen wir jedoch, dass diese Punktschätzung des arithmetischen Mittels vermutlich nicht ganz präzise ist, sondern einem Standardfehler ($\sigma_{\bar{x}}$) unterliegt. Woher wissen wir, wie groß dieser Standardfehler ist (und wie unpräzise damit unsere Schätzung)?

3.2 Schätzung des Standardfehlers

Wir führen das obige Beispiel fort:

- Die Varianz der Stichprobe können wir berechnen: $s^2 \approx 319.78$ (s. Sitzung 2).
- Dabei handelt es sich ebenfalls um einen erwartungstreuen Punktschätzer für die Varianz der Grundgesamtheit (σ^2).
- Die Varianz der Grundgesamtheit kann also auch auf $\sigma^2=s^2\approx 319{,}78$ geschätzt werden.
- Analog können wir die Standardabweichung der Population auf $\sigma=s\approx 17{,}88$ schätzen.
- Den Standardfehler können wir mit diesem Schätzwert anhand Gleichung 3 berechnen. Allerdings benutzen wir statt $\sigma_{\bar{x}}$ das Symbol $s_{\bar{x}}$, da es sich um einen Schätzwert handelt:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$\approx \frac{17,88}{\sqrt{10}} \approx 5,65$$

Je größer die Stichprobe, desto genauer lassen sich also Parameter der Population schätzen. Die statistische Antwort auf die Frage, wie groß die Stichprobe denn sein müsse, lautet demnach zunächst immer: Möglichst groß!

Bemerkernswert ist jedoch, dass dabei die Größe der Grundgesamtheit (N, im Beispiel die Anzahl aller Studierenden der Humangeographie) bei diesen Überlegungen überhaupt keine Rolle spielt.

Stand: 11. Mai 2019 4/11

4 Intervallschätzung

Um eine Intervallschätzung durchführen zu können, muss:

- die Standardabweichung der Grundgesamtheit σ bekannt und
- die theoretische Verteilung von \bar{x} normalverteilt sein. Das bedeutet:
 - Entweder es ist bekannt, dass die Grundgesamtheit normalverteilt ist
 - *Und/oder* die Stichprobengröße ist $n \ge 30$

Für das obige Beispiel der Pendelzeiten wissen wir nicht, wie die Verteilung der Grundgesamtheit aussieht, und die Stichprobengröße (n=10) ist kleiner als 30. Eine Intervallschätzung können wir hier also nicht durchführen!

Auch bei der Intervallschätzung (engl. *interval estimation*) geht es darum, das arithmetische Mittel der Population (μ) zu schätzen. Allerdings geben wir nicht einfach nur den wahrscheinlichsten Wert an, sondern einen Bereich (ein *Intervall*), in dem μ mit einer bestimmten Wahrscheinlichkeit liegt.

Die Grundüberlegung ist dabei folgende:

- Wir haben eine *empirische* Stichprobe vorliegen (und können ihren Mittelwert \bar{x} und ihre Standardabweichung s berechnen).
- Wir wissen dass die *theoretische* Verteilung aller möglichen Stichproben normalverteilt ist, und um den gesuchten Wert μ symmetrisch ist.
- Den Mittelwert unserer empirischen Stichprobe \bar{x} können wir uns als zufälligen Wert der theoretischen Stichprobenverteilung von \bar{x} vorstellen.
- Wo genau in dieser theoretischen Verteilung wir mit unserem empirischen Wert "gelandet" sind, wissen wir nicht.
- Wenn wir den Wert μ kennen würden, könnten wir (mit den Methoden aus Sitzung 3) die Wahrscheinlichkeit für einen beliebeigen Bereich angeben, in den ein zufälliges \bar{x} fällt.
- Der entscheidende Trick: Weil die Normalverteilung symmetrisch ist, sind diese Wahrscheinlichkeiten analog anzuwenden auf die Bereiche einer konstruierten Verteilung mit gleichem $\sigma_{\bar{x}}$ um unser \bar{x} , in die der wirkliche Wert μ fällt. (s. Abbildung 4).

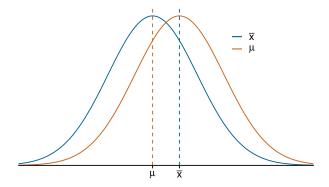


Abbildung 4: Konstruierte Verteilung um \bar{x}

Dabei heißt der Bereich Konfidenzintervall (engl. *confidence interval*), und seine Breite wird mit KIB abgekürzt. Die Wahrscheinlichkeit, dass wir mit unserer Schätzung *außerhalb* des Konfidenzintervalls liegen wird mit α gekennzeichnet. Ein 95%-Konfidenzintervall hat also ein α von 0,05 (s. Abbildung 5).

Ein Beispiel soll dies verdeutlichen: Wir wissen, dass die jährliche Niederschlagsmenge in Hessen normalverteilt ist mit $\sigma=10{,}23$. Wir haben die Messwerte in Tabelle 1 erhoben und möchten den Mittelwert (μ) per Intervallschätzung angeben.

Stand: 11. Mai 2019 5/11

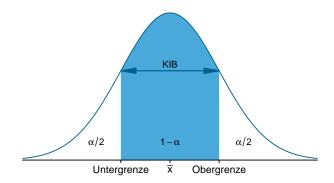


Abbildung 5: Konfidenzintervall

Tabelle 1: Jahresniederschlag in Hessen

Jahr	Niederschlag (l/m²)
2011	855,3
2012	839,5
2013	850,6
2014	873,1
2015	858,3
2016	857,1
2017	861,4

Zunächst errechnen wir den Mittelwert unserer empirischen Stichprobe:

$$\bar{x} \approx 856,47$$

Dann errechnen wir anhand Gleichung 3 den Standardfehler der theoretischen Verteilung von \bar{x} :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\approx \frac{10,23}{\sqrt{7}} \approx 3,86$$

4.1 Gesuchtes α

Nun könnte eine Fragerichtung lauten: Wie groß ist die Wahrscheinlichkeit, dass der Mittelwert der Population μ in einem Korridor von \pm 5 l/m² um \bar{x} liegt?¹

Gesucht ist bei einer Konfidenzintervallbreite von $\it KIB=10$ also die Wahrscheinlichkeit:

$$1 - \alpha \approx P(851,47 < \mu < 861,47)$$

Stand: 11. Mai 2019 6/11

 $^{^1}$ Genau genommen ist das nicht ganz korrekt, "denn tatsächlich kann der Parameter nur innerhalb oder außerhalb des gefundenen Bereichs liegen. Die Wahrscheinlichkeit, dass ein Parameter in einen bestimmten Bereich fällt, ist damit entweder 0 oder 1." (Bortz und Schuster 2010: 93). Mathematisch korrekt müsste es heißen: "Die Wahrscheinlichkeit, dass \bar{x} zu einer Population gehört, deren Parameter μ in diesem Bereich liegt..."

Generalisierend lässt sich schreiben:

$$1 - \alpha = P(x_{\alpha/2} < \mu < x_{(1-\alpha/2)})$$

...wobei $x_{\alpha/2}$ die Untergrenze darstellt und $x_{(1-\alpha/2)}$ die Obergrenze.

In z-Werten ausgedrückt:

$$1 - \alpha = P(z_{\alpha/2} < z_{\mu} < z_{(1-\alpha/2)}) \tag{5}$$

In Sitzung 3 haben wir bereits gelernt, wie diese Wahrscheinlichkeit berechnet werden kann. Im Folgenden wird der Rechenweg noch einmal am Beispiel dargelegt.

4.1.1 Die umständliche Variante

Zunächst müssen wir die Intervallgrenzen inz-Werte umwandeln, um die Unter- bzw. Überschreitungswahrscheinlichkeiten ermitteln zu können. Die z-Transformation muss hier jedoch anhand des Standardfehlers $\sigma_{\bar{x}}$ geschehen, da wir ja an der Stichprobenverteilung interessiert sind. Durch z-Transformation mit \bar{x} und dem Standardfehler $\sigma_{\bar{x}}$ erhalten wir die standardisierten Intervallgrenzen.

Untergrenze:

$$z_{\alpha/2} = \frac{x_{\alpha/2} - \bar{x}}{\sigma_{\bar{x}}}$$

$$\approx \frac{851,47 - 856,47}{3,86} \approx -1,30$$

Obergrenze:

$$z_{(1-\alpha/2)} = \frac{x_{(1-\alpha/2)} - \bar{x}}{\sigma_{\bar{x}}}$$

$$\approx \frac{861,47 - 856,47}{3.86} \approx 1,30$$

Es ist wenig überraschend, dass die z-transformierten Werte symmetrisch sind. Wir setzen in Gleichung 5 ein:

$$1 - \alpha \approx P(-1.30 < z_u < 1.30)$$

Dies lässt sich umformen in:

$$1 - \alpha \approx P(z_u < 1.08) - P(z_u < -1.08)$$

Stand: 11. Mai 2019 7/11

Die jeweiligen Wahrscheinlichkeiten lassen sich in der Tabelle für p-Werte der Normalverteilung nachschauen (bzw. für den negativen z-Wert errechnen, s. Sitzung 3):

$$1 - \alpha \approx 0,9032 - 0,0968$$
$$= 0.8064$$

Die Wahrscheinlichkeit, dass μ im Konfidenzintervalls 856,47 ± 5 l/m² liegt, beträgt also 80,64%.

4.1.2 Die schnelle Variante

Wir können den z-Wert für die Obergrenze des Konfidenzintervalls ganz einfach ausrechnen, weil wir wissen, dass die Obergrenze um 5 größer ist als \bar{x} und dass $z_{\bar{x}}=0$:

$$z_{(1-\alpha/2)} = \frac{5}{\sigma_{\bar{x}}}$$

$$\approx \frac{5}{3,86}$$

$$\approx 1,30$$

Oberhalb dieses Werts liegt bekanntermaßen der Anteil $\frac{\alpha}{2}$, woraus sich mit Blick auf die Tabelle ergibt:

$$\frac{\alpha}{2} = 1 - 0,9032$$
 $\alpha = 0,1936$

4.2 Gesuchtes Konfidenzintervall

Eine weitere Möglichkeit der Fragestellung lautet: In welchem Bereich liegt das arithmetische Mittel μ mit einer Wahscheinlichkeit von 90%?

Vorgegeben ist also $\alpha=0.1$, und gesucht sind die Unter- und die Obergrenze des Konfidenzintervalls.

Wir setzen ein:

$$1 - \alpha = P(z_{\alpha/2} < z_{\mu} < z_{(1-\alpha/2)})$$
$$0.9 = P(z_{5\%} < z_{\mu} < z_{95\%})$$

Die entsprechenden z-Werte der Intervallgrenzen lassen sich (in umgekehrter Suchrichtung) aus der Tabelle ablesen:

$$z_{5\%} \approx -1.64$$

 $z_{95\%} \approx 1.64$

Durch umgekehrte z-Transformation – auch hier weider mit \bar{x} und $\sigma_{\bar{x}}$ – ergeben sich die Intervallgrenzen.

Stand: 11. Mai 2019 8/11

Untergrenze:

$$x_{5\%} = z_{5\%} \cdot \sigma_{\bar{x}} + \bar{x}$$

 $\approx -1.64 \cdot 3.86 + 856.47$
 ≈ 850.14

Obergrenze:

$$x_{95\%} = z_{95\%} \cdot \sigma_{\bar{x}} + \bar{x}$$

 $\approx 1,64 \cdot 3,86 + 856,47$
 ≈ 862.80

Auch hier gibt es wieder eine kleine Abkürzung: Aufgrund der Symmetrie unserer theoretischen Verteilung gilt für die Konfidenzintervallbreite generell:

$$\frac{KIB}{2} = z_{(1-\alpha/2)} \cdot \sigma_{\bar{x}} \tag{6}$$

Wir setzen einfach unsere Werte ein:

$$\frac{KIB}{2} = z_{95\%} \cdot s_{\bar{x}}$$

$$\approx 1,64 \cdot 3,86$$

$$\approx 6.33$$

Die Intervallgrenzen ergeben sich dann trivial aus $ar{x} \pm \frac{K\!I\!B}{2}.$

4.3 Gesuchtes n

Eine letzte Fragerichtung lautet: Wie viele Messwerte müssten vorliegen, um den durchschnittlichen Niederschlag mit einem Konfidenzniveau von 99% und einer Genauigkeit von \pm 5 l/m² schätzen zu können?

Gegeben sind also das Konfidenzintervall und $\alpha=0,01$, gesucht wird n. Wir wissen, dass die Stichprobengröße n den Standardfehler $\sigma_{\bar{x}}$ bestimmt. Also benutzen wir zunächst Gleichung 6 und formen um:

$$\frac{\mathit{KIB}}{2} = z_{(1-\alpha/2)} \cdot \sigma_{\bar{x}}$$

$$\sigma_{\bar{x}} = \frac{\mathit{KIB}}{2 \cdot z_{(1-\alpha/2)}}$$

Stand: 11. Mai 2019 9/11

Durch Einsetzen und mit Blick auf die Tabelle erhalten wir:

$$\sigma_{\bar{x}} = \frac{10}{2 \cdot z_{99,5\%}}$$

$$\approx \frac{10}{2 \cdot 2,58}$$

$$\approx 1,94$$

Dieser Standardfehler $\sigma_{\bar{x}}\approx 1{,}94$ würde unseren Anforderungen genügen. Welches n ist nötig, um diesen Standardfehler zu erreichen? Wir formen Gleichung 3 um...

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{\sigma}{\sigma_{\bar{x}}}\right)^2$$

...und setzen den angestrebten Standardfehler sowie die Standardabweichung der Population ($\sigma=10{,}23$) ein:

$$n = \left(\frac{\sigma}{\sigma_{\bar{x}}}\right)^2$$
$$n \approx \left(\frac{10,23}{1,94}\right)^2$$
$$\approx 27,80$$

Wir müssten also 28 Stichproben vorliegen haben.

5 Aufgaben

Die folgenden Aufgaben sind zur eigenständigen Überprüfung Ihrer Lernleistung gedacht (als Vor- oder Nachbereitung der Vorlesung, oder als Klausurübung) und nicht etwa als Hausaufgabe.

5.1 Aufgabe 1

Eine Messreihe habe die Werte: 165 173 155 179 158 142

- a) Führen Sie eine Punktschätzung für μ und σ der Grundgesamtheit durch.
- b) Welcher Standardfehler für \bar{x} ist zu erwarten?

5.2 Aufgabe 2

Die Sonnenstunden auf einer Ferieninsel (pro Tag, im Jahresdurschnitt) sind annähernd normalverteilt mit einer Standardabweichung von vier Minuten. Der Mittelwert μ ist unbekannt, es liegen neun Messwerte vor.

Stand: 11. Mai 2019 10/11

- a) Welcher Standardfehler für \bar{x} ist zu erwarten?
- b) Welche Konfidenzintervallbreite korrespondiert mit einem Konfidenzniveau von 95%?
- c) Mit welchem Konfidenzniveau lässt sich μ "auf die Minute genau" (\pm 30 Sekunden) schätzen?
- d) Welche Stichprobengröße ist nötig um den Mittelwert mit einer Konfidenzintervallbreite von zwei Minuten und -niveau von 90% zu schätzen?

5.3 Aufgabe 3

Sie intressieren sich für das Durchschnittseinkommen (in EUR) der Haushalte eines Stadtteils. Die Varianz ist mit $\sigma^2=4096$ bekannt. Eine Zufallsstichprobe von 40 befragten Haushalten weist einen Mittelwert von $\bar{x}=2650$ auf.

- a) Wie lautet das 90%-Konfidenzintervall?
- b) Mit welcher Wahrscheinlichkeit liegt das Durchschnittseinkommen zwischen 2640 und 2660 EUR?

6 Tipps zur Vertiefung

- YouTube-Kanal "Kurzes Tutorium Statistik": Intervallschätzungen Konfidenzintervalle
- Kapitel 8 in Klemm (2002)
- Kapitel 6.2–6.4 in Bortz und Schuster (2010)
- Kapitel 5.3.1 in Bahrenberg, Giese und Nipper (2010)

Quellen

Bahrenberg, Gerhard, Ernst Giese und Josef Nipper. 2010. *Statistische Methoden in der Geographie*. 5. Aufl. Bd. 1. Univariate und bivariate Statistik. Stuttgart: Bornträger.

Bortz, Jürgen und Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin: Springer.

Klemm, Elmar. 2002. *Einführung in die Statistik. Für die Sozialwissenschaften*. Wiesbaden: Westdeutscher Verlag.

Stand: 11. Mai 2019