Willkommen zur Vorlesung Statistik (Master)

Thema dieser Vorlesung:
Punkt- und Intervallschätzung

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Punkt- und Intervallschätzung

Punkt- und Intervallschätzung

Aus Stichproben errechnen wir einen oder mehrere verschiedene Werte, die Schätzwerte für die Grundgesamtheit darstellen sollen. Man spricht hier von Punktschätzer, da eben jeweils genau ein Wert (Anteils-, Mittelwert oder andere Größe, z. B. Regressionskoeffizient) geschätzt wird.

Der Punktschätzer ist der beste Schätzer für den Parameter. Dennoch ist es recht unwahrscheinlich, dass der Punktschätzer genau dem Parameter entspricht. Bsp.: Bei $n = 1\,000$ und einem Anteilswert von 0,4 für Merkmalsausprägung X_i beträgt die Wahrscheinlichkeit, genau 400 mal X_i zu bekommen, 0,0257.

Daher sollte man die Punktschätzung durch eine Intervallschätzung ergänzen, die eine größere Wahrscheinlichkeit aufweist – um den Preis einer größeren Bandbreite.

Zunächst aber ein paar Worte zur Punktschätzung ...

Eigenschaften von Schätzern

Wünschenswerte Eigenschaften von Schätzern sind:

- Erwartungstreue: Der Erwartungswert (Mittelwert) der Kennwerteverteilung soll dem wahren Parameter in der GG entsprechen.
- Effizienz: Die Streuung des Schätzers soll möglichst klein sein (d. h., die Schätzwerte sollen möglichst häufig möglichst nahe am wahren Wert liegen)
- Konsistenz: Mit zunehmendem Stichprobenumfang sollen Abweichungen vom wahren Wert geringer werden.

Schätzung von Anteilswerten

Wir erinnern uns: Der Erwartungswert einer binomialverteilten Zufallsvariablen ist $E(X) = n \cdot \pi$.

Der Erwartungswert für einen Anteilswert ist daher

$$E(X_{Anteil}) = \frac{n \cdot \pi}{n} = \pi$$

p, der Anteilswert in der Stichprobe, ist also ein erwartungstreuer (und konsistenter) Schätzer für den Anteilswert π der Grundgesamtheit.

Schätzung von Stichprobenmittelwerten

Erwartungswert einer Summe von n Zufallsvariablen:

$$\begin{aligned} E(X_1 + X_2 + \cdots + X_n) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= \mu_1 + \mu_2 + \cdots + \mu_n \end{aligned}$$

Wenn aber alle X aus der gleichen Grundgesamtheit stammen (i. i. d.-Bedingung), gilt $\mu_1 = \mu_2 = \cdots = \mu_n$ und somit:

$$E(X_1 + X_2 + \cdots + X_n) = n \cdot \mu$$

Der Erwartungswert des Stichprobenmittelwertes ist dann

$$E(\bar{X}) = \frac{1}{n} \cdot (n \cdot \mu) = \mu$$

Der Mittelwert der Stichprobe $\bar{x}$ ist also ein erwartungstreuer (und konsistenter) Schätzer für den Mittelwert der Grundgesamtheit μ.

Schätzung von Varianzen

Wir erinnern uns (aus Statistik im Bachelor): Die für die Stichprobe errechnete Varianz

$$s_x^2 = \frac{1}{n} \sum_{n=1}^{i} (x_i - \bar{x})^2$$

ist kein erwartungstreuer Schätzer für die Varianz der Grundgesamtheit. Ein erwartungstreuer Schätzer lautet:

$$\widehat{\sigma}_x^2 = \frac{1}{n-1} \sum_{n=1}^{i} (x_i - \bar{x})^2 = s_x^2 \cdot \frac{n}{n-1}$$

Die Varianz in der Stichprobe s^2 ist asymptotisch erwartungstreu: Geht n gegen unendlich, strebt $n/(n-1)$ gegen 1.

Hinweis: Statistik-Software berechnet in aller Regel automatisch $\widehat{\sigma}_x^2$ (meist muss s^2 sogar von Hand aus $\widehat{\sigma}_x^2$ errechnet werden).

Intervallschätzung

Die Intervallschätzung zielt darauf ab, einen Bereich anzugeben, der mit einer gewissen (von der Forscherin gewählten) Wahrscheinlichkeit den wahren Wert enthält (überdeckt). Dieser Bereich heißt „Konfidenzintervall“.

Die Wahrscheinlichkeit, mit der das Intervall den wahren Wert enthält, sollte in der Regel möglichst hoch sein. Der trade-off: Je größer die gewählte Wahrscheinlichkeit, desto breiter das resultierende Intervall.

Die Überlegungen hierzu beruhen auf Annahmen über die Streuung von Stichprobenkennwerten. Wir greifen also den Faden der letzten Vorlesungen wieder auf.

Der Standardfehler von Anteilswerten

Wir erinnern uns: Für die Binomialverteilung gilt:
$Var(X) = \sigma^2 = n \cdot \pi \cdot (1 - \pi)$

Die Standardabweichung der Stichprobenkennwerte (aka Standardfehler) ist also

$$\sigma = \sqrt{\sigma^2} = \sqrt{n \cdot \pi \cdot (1 - \pi)}$$

Für Anteilswerte müssen wir diesen Ausdruck durch n dividieren:

$$\sigma_{\text{Anteil}} = \frac{\sqrt{n \cdot \pi \cdot (1 - \pi)}}{n} = \sqrt{\frac{n \cdot \pi \cdot (1 - \pi)}{n^2}} = \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

Wenn π (wie meist) unbekannt ist, wird es anhand von p geschätzt, und es gilt:

$$\widehat{\sigma}_{\text{Anteil}} = \sqrt{\frac{p \cdot (1 - p)}{n}}$$

Varianz von Mittelwerten

Die Varianz der Mittelwerte von Stichproben des Umfangs n, die die i. i. d.-Bedingung erfüllen, ergibt sich (wegen $Var(aX + b) = a^2 Var(X)$) wie folgt:

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\ &= \frac{1}{n^2} Var(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n^2}\left(Var\left(X_1\right) + Var\left(X_2\right) + \cdots + Var\left(X_n\right)\right) \\ &= \frac{1}{n^2}(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Die Varianz der Stichprobenmittelwerte entspricht also der Varianz der Variablen in der Grundgesamtheit, dividiert durch n.

Der Standardfehler von Mittelwerten

Der Standardfehler (=Standardabweichung des Schätzers) kann bei bekannter Varianz der Grundgesamtheit berechnet werden als:

$$S.E. = \sigma_{\bar{x}} = \sqrt{\frac{\sigma_x^2}{n}} = \frac{\sigma_x}{\sqrt{n}}$$

Wenn die Varianz in der Population unbekannt, kann der Standardfehler aus der Stichprobe geschätzt werden:

$$S.E. = \widehat{\sigma}_{\bar{x}} = \frac{\widehat{\sigma}_x}{\sqrt{n}} = \frac{s_x}{\sqrt{n-1}}$$

Konfidenzintervalle I

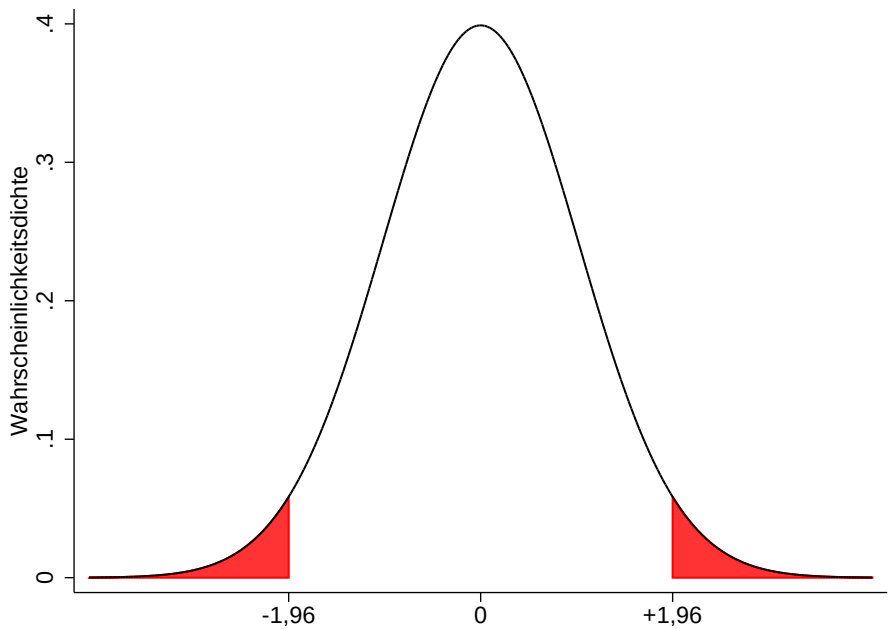
Für das gesuchte Intervall muss zunächst eine Wahrscheinlichkeit festgelegt werden, mit der das Intervall den wahren Wert enthalten soll. Diese wird manchmal als Konfidenzwahrscheinlichkeit oder Konfidenzniveau bezeichnet (manchmal auch: Vertrauenswahrscheinlichkeit). Man spricht daher von einem Konfidenzintervall.

Diese Wahrscheinlichkeit soll ziemlich groß sein; in der Regel wählt man eine Wahrscheinlichkeit von 95 Prozent (oder 0,95).

Die verbleibende Wahrscheinlichkeit heißt „Irrtumswahrscheinlichkeit“; im genannten Regelfall beträgt sie also 5 Prozent oder 0,05. Sie wird mit α (alpha) bezeichnet, das Konfidenzniveau also mit $1 - \alpha$.

Konfidenzintervalle II

Für die Bestimmung des Konfidenzintervalls gehen wir aus von der Frage, in welchem Wertebereich ein Anteil von $1 - \alpha$ der Stichprobenkennwerte liegen würde. Der Wertebereich wird in der Regel so bestimmt, dass wir $\alpha/2$ der Werte am extremen linken und $\alpha/2$ am extremen rechten Rand der Verteilung ausschließen (nachfolgend am Beispiel einer SNV und $\alpha = 0{,}05$).



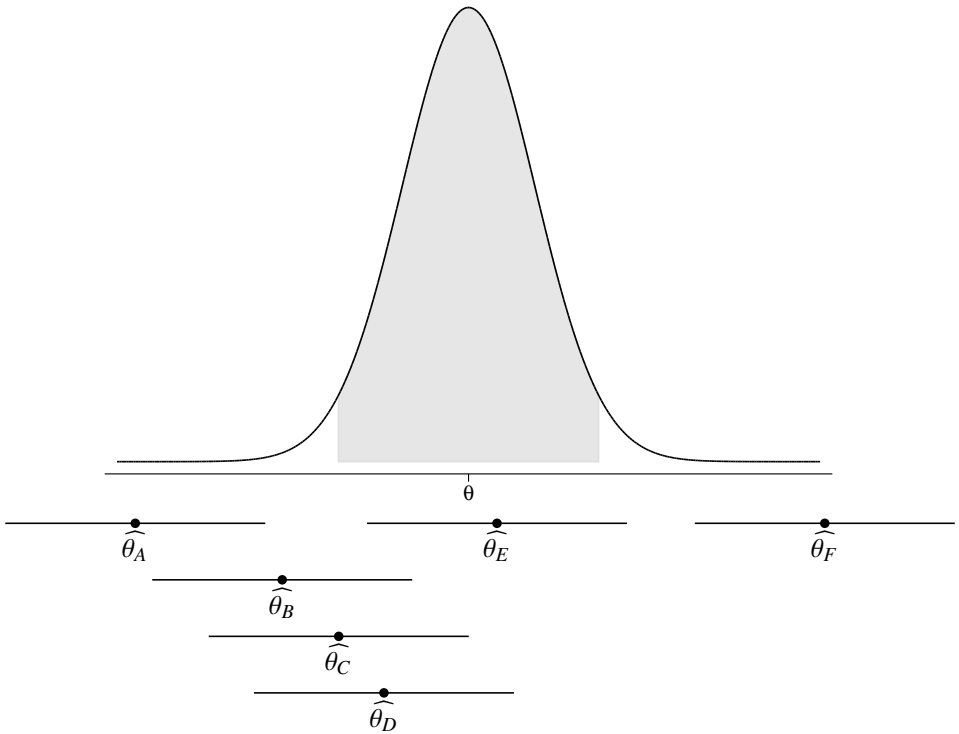
Von der Schätzfunktion zum Konfidenzintervall

Allgemein gilt: Der aus einer Stichprobe gewonnene Schätzwert $\widehat{\theta}$ liegt mit einer Wahrscheinlichkeit von $1 - \alpha$ im Bereich

$$\theta \pm (1 - \alpha/2)\text{-Quantilwert} \cdot \text{Standardfehler}$$

um den wahren Wert (Parameter) der GG. Dann können wir aber auch sagen: Mit einer Wahrscheinlichkeit von $1 - \alpha$ enthält ein Intervall der gleichen Breite (also das Intervall
$\widehat{\theta} \pm (1 - \alpha/2)$-Quantilwert $\cdot$ Standardfehler) den wahren Wert der Grundgesamtheit.

Konfidenzintervalle visualisiert



Die Intervalle um $\widehat{\theta}_A$, $\widehat{\theta}_B$ und $\widehat{\theta}_F$ schließen den wahren Wert θ nicht ein. $\widehat{\theta}_A$, $\widehat{\theta}_B$ und $\widehat{\theta}_F$ gehören zu den 5 Prozent der „unwahrscheinlichsten“ (am weitesten vom wahren Wert entfernten) Stichprobenkennwerte.

Konfidenzintervalle III

Unter Berücksichtigung des Wissens, dass die Stichprobenkennwerte (bei ausreichend großen Fallzahlen) einer Normalverteilung folgen, können wir nun die Streuung von Stichprobenkennwerten durch Standardisierung auf die Standardnormalverteilung zurückführen.

Standardisierung heißt hier: Wir ziehen (gedanklich) von den Stichprobenkennwerten deren Mittelwert (=Erwartungswert) ab und dividieren durch die Standardabweichung (=den Standardfehler).

Streuung von Mittelwerten: σ bekannt

Ist die Varianz der Grundgesamtheit bekannt, können wir zur Bestimmung des Konfidenzintervalls direkt auf die SNV zurückgreifen.

$$P\left(-z_{1-\alpha/2} < \frac{\bar{x}-\mu_x}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha/2}\right) = 1-\alpha \quad \Big| \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow P\left(-z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \bar{x}-\mu_x < z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \quad |\cdot -1$$

$$\Rightarrow P\left(z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} > \mu_x-\bar{x} > -z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \quad |\text{Umformen:}$$

$$\Rightarrow P\left(\bar{x}-z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu_x < \bar{x}+z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$\bar{x}-z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ ist also die Untergrenze, $\bar{x}+z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ die Obergrenze des Konfidenzintervalls, das mit $P = 1-\alpha$ den wahren Wert μ umschließt.

$z_{1-\alpha/2}$: 1-α/2-Quantilwert der SNV

Streuung von Mittelwerten: σ geschätzt

Schätzen wir die Streuung der untersuchten Variablen aus der Stichprobe, wird die entstehende Ungenauigkeit durch die t-Verteilung berücksichtigt.

$$P\left(-t_{1-\alpha/2,n-1} < \frac{\bar{x}-\mu_x}{\frac{\widehat{\sigma}}{\sqrt{n}}} < t_{1-\alpha/2,n-1}\right) = 1-\alpha$$

$$\Rightarrow P\left(-t_{1-\alpha/2,n-1}\cdot\frac{\widehat{\sigma}}{\sqrt{n}} < \bar{x}-\mu_x < t_{1-\alpha/2,n-1}\cdot\frac{\widehat{\sigma}}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow P\left(t_{1-\alpha/2,n-1}\cdot\frac{\widehat{\sigma}}{\sqrt{n}} > \mu_x-\bar{x} > -t_{1-\alpha/2,n-1}\cdot\frac{\widehat{\sigma}}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\bar{x}-t_{1-\alpha/2,n-1}\cdot\frac{\widehat{\sigma}}{\sqrt{n}} < \mu_x < \bar{x}+t_{1-\alpha/2,n-1}\cdot\frac{\widehat{\sigma}}{\sqrt{n}}\right) = 1-\alpha$$

Das Konfidenzintervall zum Niveau $1-\alpha$ reicht also von $\bar{x}-t_{1-\alpha/2,n-1}\cdot\frac{\sigma}{\sqrt{n}}$ bis $\bar{x}+t_{1-\alpha/2,n-1}\cdot\frac{\sigma}{\sqrt{n}}$.

$t_{1-\alpha/2,n-1}$: $1-\alpha/2$-Quantilwert der t-Verteilung mit $n-1$ Freiheitsgraden

Streuung von Anteilswerten

Wenn gilt: $n \cdot \pi \cdot (1 - \pi) > 9$, ist der Anteilswert p approximativ normalverteilt:

$$P\left(-z_{1-\alpha/2} < \frac{p - \pi}{\sqrt{p \cdot (1-p)/n}} < z_{1-\alpha/2}\right) \approx 1 - \alpha$$

$$\vdots$$

$$\vdots$$

$$P\left(p - z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} < \pi < p + z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}\right) \approx 1 - \alpha$$

Das Konfidenzintervall zum Niveau $1 - \alpha$ reicht also von $p - z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$ bis $p + z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$.

$z_{1-\alpha/2}$: 1-α/2-Quantilwert der SNV

Eine Faustregel für 95-%-Konfidenzintervalle

Für eine grobe („quick and not so dirty after all“) Abschätzung des Konfidenzintervalls um einen normalverteilten Schätzwert $\widehat{\theta}$ kann man bei 5-prozentiger Irrtumswahrscheinlichkeit das Intervall

$$\widehat{\theta} - 2 \cdot S.E. \leqslant \widehat{\theta} \leqslant \widehat{\theta} + 2 \cdot S.E.$$

verwenden.

Diese Faustregel kann auch für Größen herangezogen werden, deren Stichprobenverteilung der t-Verteilung folgt, wenn die Stichprobe ausreichend groß ist.

Interpretation von Konfidenzintervallen

Was heißt es, ein Konfidenzintervall berechnet zu haben?

„Ich habe ein Verfahren eingesetzt, das mit 95-prozentiger Wahrscheinlichkeit ein Intervall ergibt, welches den wahren Wert der Grundgesamtheit einschließt. Daher habe ich ein gewisses Vertrauen (Hoffnung), dass das vorliegende Intervall zu den Intervallen gehört, auf die das zutrifft.“

Fehlinterpretationen z.B.:

- „95 Prozent unserer befragten Personen erhalten pro Monat zwischen 124,13 Euro bis 181,05 Euro mehr Gehalt pro Monat pro Bildungsjahr, das sie absolviert haben.“
- „ . . . dass 95 Prozent der Stichproben zwischen 124,13 und 181,06 mehr Nettoeinkommen pro Monat pro weiteres Bildungsjahr liegen.“

Konfidenzintervalle im Beispiel I: Mittelwert, großes n

Daten zum Alter von 100 Befragungspersonen (siehe Vorlesung Statistik I): $\bar{x} = 42, \widehat{\sigma} = 11{,}1, \alpha = 0{,}05$

Untere Grenze des Konfidenzintervalls:

$$42 - 1{,}980 \cdot \frac{11{,}1}{\sqrt{100}} \approx 39{,}8$$

Obere Grenze des Konfidenzintervalls:

$$42 + 1{,}980 \cdot \frac{11{,}1}{\sqrt{100}} \approx 44{,}2$$

(Werte der t-Verteilung; wegen $n = 100$ könnte man auch die Faustregel oder die Standardnormalverteilung (z-Wert: 1,96) verwenden.)

Konfidenzintervalle im Beispiel II: Mittelwert, kleines n

Bruttolöhne (in DM); $\bar{x} = 5\,627, \widehat{\sigma} = 2\,862, \alpha = 0{,}05, n = 20$

Untere Grenze des Konfidenzintervalls:

$$5\,627 - 2{,}093 \cdot \frac{2\,862}{\sqrt{20}} \approx 4\,333$$

Obere Grenze des Konfidenzintervalls:

$$5\,627 + 2{,}093 \cdot \frac{2\,862}{\sqrt{20}} \approx 7\,011$$

Konfidenzintervalle im Beispiel III: Anteilswert

Das Intervall $\pm 1{,}96 \cdot$ S.E. um den geschätzten Anteilswert (Bsp.: 0,4) enthält mit 95-prozentiger Wahrscheinlichkeit den wahren Wert.

n	S.E.	Konfidenzintervall
96	0,0500	0,302 – 0,498
192	0,0354	0,330 – 0,469
384	0,0250	0,351 – 0,449
768	0,0177	0,365 – 0,435
1 536	0,0125	0,376 – 0,425

Weitere Intervallschätzer I: Median

Für $n > 50$ gelten (nach Jann 2002, S. 139) folgende Grenzen des Konfidenzintervalls für den Median:

Untergrenze: $x_{(k)}$; Obergrenze: $x_{(n-k+1)}$

mit $k = \frac{1}{2}\left(n - 1 - z_{1-\alpha/2} \cdot \sqrt{n}\right)$

Indizes: k-ter bzw. n-k+1-ter Wert der geordneten Datenreihe

Einige Beispiele bei $\alpha = 0{,}05$ (Werte gerundet):

n	$x_{(k)}$	$x_{(n-k+1)}$
100	40	61
200	86	115
500	228	273
1 000	469	532

In der Literatur finden sich leicht unterschiedliche Formeln!

Weitere Intervallschätzer II: Varianz

Untere Grenze:

$$\frac{(n-1)\cdot\widehat{\sigma}^2}{\chi_{1-\alpha/2;\ \text{d.f.}=n-1}}$$

Obere Grenze:

$$\frac{(n-1)\cdot\widehat{\sigma}^2}{\chi_{\alpha/2;\ \text{d.f.}=n-1}}$$

χ: Entsprechender Wert der χ^2-Verteilung

Weitere Intervallschätzer II: Varianz

Beispiel: Altersdaten (siehe oben, 2. Beispiel Mittelwert)

$\widehat{\sigma} = 11{,}1 \rightarrow \widehat{\sigma}^2 = 123{,}21$

Problem: χ^2 (99 d.f.) nicht tabelliert.

Lösung: Approximation durch $\chi^2_\alpha = \frac{1}{2}\left(z_\alpha + \sqrt{2n-1}\right)^2$
(wobei α hier: das jeweils benötigte Quantil, also $1 - \alpha/2$ bzw. $\alpha/2$)

Mit exakten Werten:

Untere Grenze: $(99 \cdot 123{,}21)/128{,}42 = 94{,}98$

Obere Grenze: $(99 \cdot 123{,}21)/73{,}36 = 166{,}27$