Vorlesung: Statistik II für Wirtschaftswissenschaft

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

Sommersemester 2017

Intervallschätzung: Motivation

Annahme: Der wahre Anteil der CDU/CSU - Wähler 2017 liegt bei genau 40.0%. Wie groß ist die Wahrscheinlichkeit, in einer Zufallsstichprobe von 1000 Personen genau einen relativen Anteil von 40.0% von CDU/CSU Wählern zu erhalten?

$$
\begin{aligned}
X_i &= \begin{cases} 1, \text{CDU/CSU} \\ 0, \text{sonst} \end{cases} \\
P(X_i = 1) &= \pi = 0.4 \\
X &= \sum_{i=1}^{n} X_i \sim B(n, \pi) \text{ mit } n = 1000 \\
\hat{\pi} &= \frac{X}{n}
\end{aligned}
$$

Punktschätzer

$$
\begin{aligned}
P(X = 400) &= \binom{n}{x} \cdot \pi^x \cdot (1-\pi)^{n-x} \\
&= \binom{1000}{400} \cdot 0.4^{400} \cdot (1-0.4)^{600} \\
&= 0.026
\end{aligned}
$$

Mit Wahrscheinlichkeit von etwa 97.4%, verfehlt der Schätzer den wahren Wert.
Beim Runden auf ganze Prozente muss der Anteil der CDU/CSU - Wähler in der Stichprobe zwischen 395 und 404 liegen, um 40% zu erhalten:

$$P(395 \leq X \leq 404) = 0.25$$

Auch beim Runden auf ganze Prozente ergibt sich mit Wahrscheinlichkeit 75% ein falscher Wert.

Konsequenzen

- Vorsicht bei der Interpretation, insbesondere bei „knappen Ergebnissen“
- Angabe der Genauigkeit
- Geeignete Wahl des Stichprobenumfangs
- Es ist häufig nicht sinnvoll, sich genau auf einen Wert festzulegen. Oft ist die Angabe eines Intervalls, von dem man hofft, dass es den wahren Wert überdeckt, vorzuziehen: *Intervallschätzung*

Schätzgenauigkeit

Anteilschätzer: Schätzung des Anteils in der Grundgesamtheit (bzw. die Erfolgswahrscheinlichkeit) π durch relative Häufigkeit in der Stichprobe.
Gegeben: iid Stichprobe $X_1, \ldots X_n$ mit $X_i \in \{0, 1\}$

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Dann kann die Schätzgenauigkeit durch die Standardabweichung von $\hat{\pi}$ charakterisiert werden:

$$SE(\hat{\pi}) = \sqrt{\frac{\pi \cdot (1 - \pi)}{n}}$$

Die Standardabweichung eines Schätzers wird auch häufig als **Standardfehler**, **engl. standard error**, bezeichnet.

Berechnung des Standardfehlers

Standardfehler für verschieden Stichprobenumfänge n und (wahre) Erfolgswahrscheinlichkeiten n: Angaben in Prozentpunkten.

n	$\pi = 10\%$	$\pi = 40\%$	$\pi = 50\%$
20	6.71	10.95	11.18
100	3.00	4.90	5.00
1000	0.95	1.55	1.58
2000	0.67	1.10	1.12
5000	0.42	0.69	0.71

Beachte: π unbekannt.
Höchste Werte für $\pi = 0.5$. Daher können diese Werte als obere Grenze verwendet werden. Bei einem Stichprobenumfang von n =1000 liegt der Standardfehler (SE) also unter 1.58%.

Mittelwertsschätzung

Schätzung des Mittelwertes in der Grundgesamtheit (bzw. des Erwartungswertes μ bei einem Experiment) durch den Mittwelwert $\bar{X}$ in der Stichprobe.
Gegeben: iid Stichprobe $X_1, \ldots X_n$ mit $E(X_i) = \mu$

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

Dann kann die Schätzgenauigkeit durch die Standardabweichung von $\hat{\mu}$ charakterisiert werden:

$$SE(\hat{\mu}) = \sqrt{\frac{\sigma^2}{n}} = SEM$$

Die Standardabweichung wird auch häufig als **Standardfehler**, **engl. standard error oder standard error of the mean (SEM)**, bezeichnet.

Beispiel: Schätzgenauigkeit bei Umsatz von Kunden

Big Data Anwendung: Eine Firma möchte die durchschnittliche Dauer von der Internetnutzung ihrer 1 Million Kunden schätzen.
Konservative, d.h. eher zu hohe Schätzung der Standardabweichung:
$\sigma = 120$ Minuten
Berechnung des Standardfehlers bei verschiedenen Stichprobengrößen:

n	SE (Minuten)
20	26.83
100	12.00
1000	3.79
2000	2.68
5000	1.70

Es ist also nicht immer nötig, die Daten von allen Kunden auszuwerten.
Man kann sich oft auf eine Zufallsstichprobe beschränken.

Standardfehler und Angabe von Schwankungsbreiten

- Standardfehler wichtiges Kriterium, aber eher schwer zu kommunizieren
- Alternative: Schwankungsbreiten und Unsicherheit
- Benutze asymptotische Normalverteilung

Die Schätzer $\hat{\pi}$ und $\hat{\mu}$ sind asymptotisch normalverteilt. Ist der Standardfehler des Schätzer gegeben, so gilt

$$\begin{aligned} P(\hat{\pi} \in [\pi - 2 \cdot SE(\hat{\pi}); \pi + 2SE(\hat{\pi})]) &= 0.95 \\ P(\hat{\mu} \in [\mu - 2 \cdot SE(\hat{\pi}); \mu + 2SE(\hat{\mu})]) &= 0.95 \end{aligned}$$

Illustration mit R

Symmetrische Intervallschätzung

Allgemeiner Ansatz: Basierend auf einer Schätzfunktion $T = g(X_1, \ldots, X_n)$ sucht man:

$$I(T) = [T - a, T + a]$$

„Trade off“ bei der Wahl von a:

- Je größer man a wählt, also je breiter man das Intervall $I(T)$ macht,
- umso größer ist die Wahrscheinlichkeit, dass $I(T)$ den wahren Wert überdeckt,
- *aber* umso weniger aussagekräftig ist dann die Schätzung.

Extremfall im Wahlbeispiel: $I(T) = [0, 100\%]$ überdeckt sicher π, macht aber eine wertlose Aussage

Typisches Vorgehen

- Man gebe sich durch inhaltliche Überlegungen einen Sicherheitsgrad (*Konfidenzniveau*) γ vor.
- Dann konstruiert man das Intervall so, dass es mindestens mit der Wahrscheinlichkeit γ den wahren Parameter überdeckt.

Definition von Konfidenzintervallen

Definition

Gegeben sei eine i.i.d. Stichprobe $X_1, \ldots, X_n$ zur Schätzung eines Parameters ϑ und eine Zahl $\gamma \in (0; 1)$. Ein zufälliges Intervall $\mathcal{C}(X_1, \ldots, X_n)$ heißt *Konfidenzintervall* zum *Sicherheitsgrad* γ (Konfidenzniveau γ), falls für jedes ϑ gilt:

$$P_\vartheta(\vartheta \in \underbrace{\mathcal{C}(X_1, \ldots, X_n)}_{\text{zufälliges Intervall}}) \geq \gamma.$$

Die Wahrscheinlichkeitsaussage bezieht sich auf das Ereignis, dass das zufällige Intervall den festen, wahren Parameter überdeckt. Streng genommen darf man im objektivistischen Verständnis von Wahrscheinlichkeit nicht von der *Wahrscheinlichkeit* sprechen, „dass ϑ in dem Intervall liegt“, da ϑ nicht zufällig ist und somit keine Wahrscheinlichkeitsverteilung besitzt.

Konstruktion von Konfidenzintervallen

Praktische Vorgehensweise: Suche Zufallsvariable Z_ϑ, die

- den gesuchten Parameter ϑ enthält und
- deren Verteilung aber nicht mehr von dem Parameter abhängt, („*Pivotgröße*“, dt. Angelpunkt).
- Dann wähle den Bereich C_Z so, dass $P_\vartheta(Z_\vartheta \in C_Z) = \gamma$ und
- löse nach ϑ auf.

Konfidenzintervall für den Mittelwert (normalverteiltes Merkmal, Varianz bekannt)

$X_1, \ldots, X_n$ i.i.d. Stichprobe gemäß $X_i \sim \mathrm{N}(\mu, \sigma^2)$, wobei σ^2 bekannt sei.

1. Starte mit der Verteilung von $\bar{X}$:

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n).$$

2. Dann erfüllt

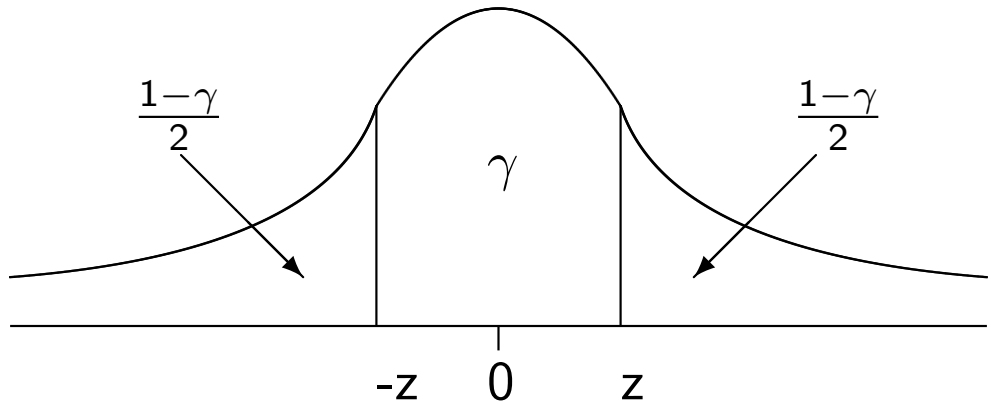
$$Z = \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} \sim \mathrm{N}(0; 1)$$

   die obigen Bedingungen an eine Pivotgröße.

3. Bestimme jetzt einen Bereich $[-z, z]$, wobei z so gewählt sei, dass

$$P(Z \in [-z; z]) = \gamma$$

KI-Bestimmung Strategie



Bestimmung von z:

$$P(Z \in [-z; z]) = \gamma \quad \Longleftrightarrow \quad P(Z \geq z) = \frac{1-\gamma}{2}$$

beziehungsweise

$$P(Z \leq z) = 1 - \frac{1-\gamma}{2} = \frac{2-1+\gamma}{2} = \frac{1+\gamma}{2}.$$

Wichtige Quantile der NV

Die Größe z heißt das $\frac{1+\gamma}{2}$–Quantil und wird mit $z_{\frac{1+\gamma}{2}}$ bezeichnet.

$$\gamma = 90\% \quad \frac{1+\gamma}{2} = 95\% \quad z_{0.95} = 1.65$$

$$\gamma = 95\% \quad \frac{1+\gamma}{2} = 97.5\% \quad z_{0.975} = 1.96$$

$$\gamma = 99\% \quad \frac{1+\gamma}{2} = 99.5\% \quad z_{0.995} = 2.58$$

Herleitung KI

$$P\left(-z_{\frac{1+\gamma}{2}} \leq Z_\mu \leq z_{\frac{1+\gamma}{2}}\right) = P\left(-z_{\frac{1+\gamma}{2}} \leq \frac{\bar{X}-\mu}{\sigma} \cdot \sqrt{n} \leq z_{\frac{1+\gamma}{2}}\right) = \gamma$$

Jetzt nach μ auflösen (Ziel: $P(\ldots \leq \mu \leq \ldots)$):

$$\begin{aligned}
\gamma &= P\left(-\frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}\right) \\
&= P\left(-\bar{X} - \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{X} - \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}\right)
\end{aligned}$$

KI für Mittelwert (NV mit bekanntem σ)

Damit ergibt sich:

Konfidenzintervall für μ bei bekannter Varianz

$$\left[\bar{X} - \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}\right] = \left[\bar{X} \pm \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}\right]$$

Eigenschaften

- Je größer σ, desto größer das Intervall!
  (Größeres $\sigma \Rightarrow$ Grundgesamtheit bezüglich des betrachteten Merkmals heterogener, also größere Streuung von $\bar{X} \Rightarrow$ ungenauere Aussagen.)
- Je größer γ, desto größer $z_{\frac{1+\gamma}{2}}$
  (Je mehr Sicherheit/Vorsicht, desto breiter das Intervall)
- Je größer n und damit $\sqrt{n}$, desto schmaler ist das Intervall
  (Je größer der Stichprobenumfang ist, desto genauer!)
  Aufpassen, die Genauigkeit nimmt nur mit $\sqrt{n}$ zu. Halbierung des Intervalls, Vervierfachung des Stichprobenumfangs.

  Kann man zur *Stichprobenplanung* verwenden!

Konfidenzintervall für den Mittelwert (normalverteiltes Merkmal, Varianz unbekannt)

Neben dem Erwartungswert ist auch σ^2 unbekannt und muss entsprechend durch

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

(mit $S = \sqrt{S^2}$) geschätzt werden. Allerdings ist

$$Z = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

jetzt nicht mehr normalverteilt, denn S ist zufällig.

$\longrightarrow$ Wir benötigen die t-Verteilung

Eigenschaften t-Verteilung

- Je größer ν ist, umso ähnlicher sind sich die $t(\nu)$-Verteilung und die Standardnormalverteilung.
  - Für $\nu \to \infty$ sind sie gleich.
  - Ab $\nu = 30$ gilt der Unterschied als vernachlässigbar.

- Je größer n, desto geringer ist der Unterschied zwischen S^2 und σ^2 und damit zwischen $\frac{\bar{X}-\mu}{S}\sqrt{n}$ und $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$.

Konfidenzintervall zum Konfidenzniveau

Ausgehend von

$$P\left(-t_{\frac{1+\gamma}{2}}^{(n-1)} \leq \frac{\bar{X}-\mu}{S} \cdot \sqrt{n} \leq t_{\frac{1+\gamma}{2}}^{(n-1)}\right) = \gamma$$

wie im Beispiel mit bekannter Varianz nach μ auflösen (mit S statt σ)

$$P\left(\bar{X} - \frac{t_{\frac{1+\gamma}{2}}^{(n-1)} \cdot S}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{t_{\frac{1+\gamma}{2}}^{(n-1)} \cdot S}{\sqrt{n}}\right) = \gamma$$

Damit ergibt sich:

Konfidenzintervall für μ bei unbekannter Varianz

$$\left[\bar{X} \pm \frac{t_{\frac{1+\gamma}{2}}^{(n-1)} \cdot S}{\sqrt{n}}\right]$$

Eigenschaften

- Es gelten analoge Aussagen zum Stichprobenumfang und Konfidenzniveau wie bei bekannter Varianz.
- Für jedes γ (und jedes ν) gilt

$$t_{\frac{1+\gamma}{2}} > z_{\frac{1+\gamma}{2}}$$

  also ist das t-Verteilungs-Konfidenzintervall (etwas) breiter.

  Hintergrund: Da σ^2 unbekannt ist, muss es geschätzt werden. Dies führt zu etwas größerer Ungenauigkeit.
- Je größer ν, umso kleiner ist der Unterschied. Für $n \geq 30$ rechnet man einfach auch bei der t-Verteilung mit $z_{\frac{1+\gamma}{2}}$.

Beispiel: Konfidenzintervall zum Konfidenzniveau γ

Eine Maschine füllt Gummibärchen in Tüten ab, die laut Aufdruck 250g Füllgewicht versprechen. Wir nehmen im folgenden an, dass das Füllgewicht normalverteilt ist. Bei 16 zufällig aus der Produktion herausgegriffenen Tüten wird ein mittleres Füllgewicht von 245g und eine Stichprobenstreuung (Standardabweichung) von 10g festgestellt.

a) Berechnen Sie ein Konfidenzintervall für das mittlere Füllgewicht zum Sicherheitsniveau von 95%.

Beispiel: Konfidenzintervall zum Konfidenzniveau γ

- Füllgewicht normalverteilt. ($\mu = 250g$ nicht benötigt)
- 16 Tüten gezogen $\Rightarrow n = 16$.
- Mittleres Füllgewicht in der Stichprobe: $\bar{x} = 245g$.
- Stichprobenstreuung: $s = 10g$.

Beispiel: Konfidenzintervall zum Konfidenzniveau γ

a) Konstruktion des Konfidenzintervalls:

- Da die Varianz σ^2 unbekannt ist, muss das Konfidenzintervall basierend auf der t-Verteilung konstruiert werden:

$$[\bar{X} \pm \frac{t_{\frac{1+\gamma}{2}}^{(n-1)} \cdot S}{\sqrt{n}}]$$

Aus dem Sicherheitsniveau $\gamma = 0.95$ errechnet sich $\frac{1+\gamma}{2} = 0.975$.
Quantil der t-Verteilung bei 0.975 und 15 Freiheitsgraden ($T = \frac{\bar{X}-\mu}{S}\sqrt{n}$ ist t-verteilt mit n-1 Freiheitsgeraden) liefert $t_{0.975}^{15} = 2.13$.

- Einsetzen liefert damit

$$[245 \pm 2.13 \cdot \frac{10}{4}] = [239.675; 250.325]$$

Approximative Konfidenzintervalle

Ist der Stichprobenumfang groß genug, so kann wegen des zentralen Grenzwertsatzes das Normalverteilungs-Konfidenzintervall auf den Erwartungswert beliebiger Merkmale (mit existierender Varianz) angewendet werden. Man erhält approximative Konfidenzintervalle, die meist auch der Berechnung mit Software zugrundeliegen.

Approximatives Konfidenzintervall für den Mittelwert (n groß)

$$\left[\bar{X} \pm z_{\frac{1+\gamma}{2}} \cdot \frac{S}{\sqrt{n}}\right]$$

$\frac{S}{\sqrt{n}}$ wird als Standardfehler (Standard error) bezeichnet.

Approximatives Konfidenzintervall für einen Anteil

Gesucht: Konfidenzintervall für den Anteilswert $p = P(X = 1)$ einer Bernoulli-Zufallsgröße X

- $X_1, \ldots, X_n$ i.i.d. Stichprobe
- n hinreichend groß (Faustregel $n > 30$)
- vorgegebenes Sicherheitsniveau γ („gamma")

Approximatives Konfidenzintervall für π

$$\hat{\pi} \pm z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

$\hat{\pi}$ = Anteil aus der Stichprobe

$z_{\frac{1+\gamma}{2}}$ ist das $\frac{1+\gamma}{2}$-Quantil der Standardnormalverteilung.

Beispiel: Wahlumfrage

- Gegeben:
  - $n = 500$
  - $\hat{\pi} = 46.5\%$
  - $\gamma = 95\%$ und damit $z_{\frac{1+\gamma}{2}} = 1.96$
- Konfidenzintervall:

$$\left[\hat{\pi} \pm z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right] = \left[0.465 \pm 1.96 \cdot \sqrt{\frac{0.465(1-0.465)}{500}}\right]$$

$$= [0.421; 0.508]$$

Inhaltliche Bemerkung (Beispiel: Wahlumfrage)

- Man beachte die relativ große Breite, trotz immerhin mittelgroßer Stichprobe
- Zum Sicherheitsniveau 95% ist keine eindeutige Aussage über die Mehrheitsverhältnisse möglich. Berücksichtigen, wenn man über Wahlumfrage urteilt
- In der Praxis werden bei Wahlumfragen Zusatzinformation verwendet (insbesondere auch frühere Wahlergebnisse). „Gebundene Hochrechnung“
- Zu der Unsicherheit durch die Stichprobenziehung kommen weitere Probleme wie falsche Antworten, Antwortverweigerung, Nicht–Erreichbarkeit von Personen. Dies kann zu Verzerrungen und deutlicher Unterschätzug des Fehlers führen

Bestimmung des Stichprobenumfangs für die Anteilsschätzung

- Genauigkeit ist inhaltlich vorzugeben
- Je genauer und sicherer, desto größer muss der Stichprobenumfang sein
- Genauigkeit: Halbe Länge g des Konfidenzintervalls
- Gib Konfidenzniveau (oft 95%) vor und bestimme n so, dass g kleiner ist als bestimmter Wert

Konkrete Umsetzung

γ: Konfidenzniveau
g: Genauigkeit

$$g \geq z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\pi(1-\pi)}{n}}$$

Auflösen nach n:

$$n \geq \frac{1}{g^2} z^2_{\frac{1+\gamma}{2}} \cdot \pi(1-\pi)$$

Beachte: $\pi(1-\pi) \leq 0.25$

Beispiel: Stichprobenplanung bei Anteilsschätzung

Gegeben:

- Konfidenzniveau: 0.95
- Genauigkeit: 10%

Bestimmung von n:

$$n \geq \frac{1}{g^2} z^2_{\frac{1+\gamma}{2}} \cdot \pi(1-\pi) = \frac{1}{0.1^2} 1.96^2 \cdot 0.25 = 96.04$$

Beachte: $\pi(1-\pi) \leq 0.25$

Also sollten ca. 100 Personen befragt werden.

Bei $g = 5\%$ ergibt sich $n = 385$
Bei $g = 1\%$ ergibt sich $n = 9604$

Konfidenzintervall für die Differenz von Mittelwerten (unabhängige Stichproben)

Unterschied der Mittelwerte zwischen zwei Gruppen $\mu_X - \mu_Y$

- Zwei voneinander stochastisch unabhängige Stichproben
  - Daten aus Gruppe 1: $X_1, \ldots, X_m$, X_i i.i.d.
  - Daten aus Gruppe 2: $Y_1, \ldots, Y_n$, Y_j i.i.d.
- Stichprobenumfänge hinreichend groß ($n \geq 30$, $m \geq 30$)
- Schätzung: $\bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{i=j}^{m} Y_j$

Approximatives KI für Differenz von Mittelwerten (unabhängigen Stichproben, n groß)

$$\left[(\bar{X} - \bar{Y}) - z_{\frac{1+\gamma}{2}} \cdot S_d; (\bar{X} - \bar{Y}) + z_{\frac{1+\gamma}{2}} \cdot S_d\right]$$

mit

- $S_d = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}$
- $z_{\frac{1+\gamma}{2}}$ ist das $\frac{1+\gamma}{2}$-Quantil der Standardnormalverteilung

Beispiel: Radiohördauer Ost-West

Westen:	$\bar{x} = 11.4$ Stunden und	$s_X = 8.4$	m $= 259$
Osten:	$\bar{y} = 9.5$ Stunden und	$s_Y = 8.4$	n$= 941$

$$\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}} \approx 0.6$$

Wir berechnen ein 99% - Konfidenzintervall:

$$k_u = \bar{x} - \bar{y} - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}} = 0.38$$

$$k_o = \bar{x} - \bar{y} + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}} = 3.42$$

Die Differenz liegt also zwischen 0.38 und 3.42 h/Woche
Werte für 95% - KI: 0.74 h; 3.1 h

Zusammenfassung

- Konfidenzintervalle sind zentrales Instrument statistischer Inferenz
- Unsicherheit der Aussagen direkt interpretierbar
- Interpretation des Sicherheitsniveaus problematisch
- (Fehl–)Interpretation als Wahrscheinlichkeit für den unbekannten Parameter in manchen Fällen vertretbar (Bayes-Inferenz)