

Der wahre Anteil der rot-grün Wähler 2009 war genau 33.7%. Wie groß ist die Wahrscheinlichkeit, in einer Zufallsstichprobe von 1000 Personen genau einen relativen Anteil von 33.7% von rot-grün Anhängern erhalten zu haben?

$$X_i = \begin{cases} 1, \text{rot/grün} \\ 0, \text{sonst} \end{cases}$$
$$P(X_i = 1) = \pi = 0.337$$
$$X = \sum_{i=1}^n X_i \sim B(n, \pi) \text{ mit } n = 1000$$

$$P(X = 337) = \binom{n}{x} \cdot \pi^x \cdot (1 - \pi)^{n-x}$$
$$= \binom{1000}{337} \cdot 0.337^{337} \cdot (1 - 0.337)^{663}$$
$$= 0.02668164$$

D.h., mit Wahrscheinlichkeit von etwa 97.3%, verfehlt der Schätzer den wahren Wert.

Konsequenzen

- Insbesondere Vorsicht bei der Interpretation „knapper Ergebnisse“ (z.B. Anteil 50.2%)
- Suche Schätzer mit möglichst kleiner Varianz, um „im Durchschnitt möglichst nahe dran zu sein“
- Es ist häufig auch gar nicht nötig, sich genau auf einen Wert festzulegen. Oft reicht die Angabe eines Intervalls, von dem man hofft, dass es den wahren Wert überdeckt: *Intervallschätzung*

Symmetrische Intervallschätzung

Basierend auf einer Schätzfunktion $T = g(X_1, \dots, X_n)$ 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: $I(T) = [0, 1]$ überdeckt sicher π , macht aber eine wertlose Aussage

- 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

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

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

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, \dots, X_n i.i.d. Stichprobe gemäß $X_i \sim N(\mu, \sigma^2)$, wobei σ^2 bekannt sei.

- Starte mit der Verteilung von \bar{X} :

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

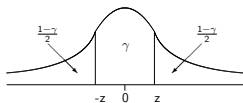
- Dann erfüllt

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

die obigen Bedingungen an eine Pivotgröße.

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

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



Bestimmung von z :

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

beziehungsweise

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



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: $P(\dots \leq \mu \leq \dots)$):

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



KI für Mittelwert (NV mit bekanntem σ)

Damit ergibt sich:

Konfidenzintervall für μ bei bekannter Varianz

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



- 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!

Neben dem Erwartungswert ist auch σ^2 unbekannt und muss entsprechend durch den UMVU-Schätzer

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

(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.

\rightarrow Wir führen deshalb ein neues Verteilungsmodell ein.

t-Verteilung

Definition

Gegeben sei eine i.i.d. Stichprobe X_1, \dots, X_n mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Dann heißt die Verteilung von

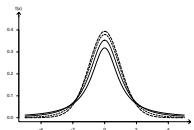
$$Z = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

t-Verteilung (oder Student-Verteilung) mit $\nu = n - 1$ Freiheitsgraden.

In Zeichen: $Z \sim t(\nu)$.

Dichte einer t-Verteilung

Die Dichte einer t-Verteilung ist der Dichte der Standardnormalverteilung sehr ähnlich: Sie ist auch symmetrisch um 0, besitzt aber etwas höhere Dichte für extreme Werte („schwerere Enden“).



Dichten von t-Verteilungen für $\nu = 1$ (—), $\nu = 2$ (···), $\nu = 5$ (- - -) und $\nu = 20$ (---) Freiheitsgrade.

Unsicherheit durch zusätzliche Schätzung von σ lässt Daten stärker schwanken.

- Je größer ν ist, umso ähnlicher sind sich die $t(\nu)$ -Verteilung und die Standardnormalverteilung.
 - Für $\nu \rightarrow \infty$ 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 $\frac{\bar{X} - \mu}{\sigma} \sqrt{n}$.

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 $z_{\frac{1-\gamma}{2}}$.

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

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

Bereite 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.

- Berechnen Sie ein Konfidenzintervall für das mittlere Füllgewicht zum Sicherheitsniveau von 95%.
- Wenn Ihnen zusätzlich bekannt würde, dass die Stichprobenstreuung gleich der tatsächlichen Streuung ist, wäre dann das unter a) zu berechnende Konfidenzintervall für das mittlere Füllgewicht breiter oder schmaler? Begründen Sie ihre Antwort ohne Rechnung.

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

a) Konstruktion des Konfidenzintervalls:

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

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

Aus dem Sicherheitsniveau $\gamma = 0.95$ errechnet sich $\frac{1+\gamma}{2} = 0.975$.

Nachschauen in t-Tabelle bei 0.975 und 15 Freiheitsgraden ($T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ ist t-verteilt mit $n-1$ Freiheitsgraden) liefert $t_{0.975} = 2.13$.

- Einsetzen liefert damit

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

b) Jetzt sei σ^2 bekannt.

- Dann kann man mit dem Normalverteilungs-Intervall rechnen:

$$[\bar{X} \pm z_{\frac{1-\gamma}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$$

Da jetzt σ bekannt, ist die Unsicherheit geringer und damit das Konfidenzintervall schmaler.

- In der Tat ist $z_{\frac{1-\gamma}{2}} < t_{\frac{1-\gamma}{2}}$.
- Rechnerisch ergibt sich mit $z_{\frac{1-\gamma}{2}} = 1.96$ das Konfidenzintervall

$$[240.100; 249.900]$$

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)

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

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

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

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

Approximatives Konfidenzintervall für π

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

R = Anteil aus der Stichprobe

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

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

$$\begin{aligned} \left[\bar{X} \pm z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right] &= \left[0.465 \pm 1.96 \cdot \sqrt{\frac{0.465(1-0.465)}{500}} \right] \\ &= [0.421; 0.508] \end{aligned}$$

- 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 sind aber Wahlumfragen etwas genauer, da man Zusatzinformation verwendet (insbesondere auch frühere Wahlergebnisse). „Gebundene Hochrechnung“

- 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

- γ : Konfidenzniveau
 g : Genauigkeit

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

Auflösen nach n :

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

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

Gegeben:

- Konfidenzniveau: 0.05
- Genauigkeit: 10%

Bestimmung von n :

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

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

Also sollten ca. 100 Personen befragt werden.

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

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

Weitere Konfidenzintervalle

- Differenz von Mittelwerten bei unabhängigen Stichproben
- Differenz von Anteilen bei unabhängigen Stichproben
- Differenz von Mittelwerten bei verbundenen Stichproben

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

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 $s_X = 8.4$

Osten: $\bar{y} = 9.5$ Stunden und $s_Y = 8.4$

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

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

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

Die Differenz liegt also zwischen 0.38 und 3.42 h/Woche

Konfidenzintervall für die Differenz von Anteilen (unabhängige Stichproben)

Vergleich von Anteilen über die Differenz $p_X - p_Y$

- Zwei voneinander stochastisch unabhängige Stichproben
 - Daten aus Gruppe 1: X_1, \dots, X_m, X_i i.i.d.
 - Daten aus Gruppe 2: Y_1, \dots, Y_n, Y_j i.i.d.
- Stichprobenumfänge hinreichend groß ($n \geq 30, m \geq 30$)
- Schätzung: $R_X - R_Y$ (R_X und R_Y : Anteil in den Gruppen 1 und 2)

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

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

mit

- $S_d = \sqrt{\frac{R_X \cdot (1-R_X)}{m} + \frac{R_Y \cdot (1-R_Y)}{n}}$
- $z_{\frac{1+\gamma}{2}}$ ist das $\frac{1+\gamma}{2}$ -Quantil der Standardnormalverteilung

Beispiel: Ist Fernsehen informativ?

	nein	ja	
alte BL: X	47	206	253
neue BL: Y	185	747	932
	232	953	1185

$$R_X - R_Y = \frac{206}{253} - \frac{747}{932} = 0.81 - 0.80$$

$$S_d = \sqrt{\frac{R_X \cdot (1-R_X)}{m} + \frac{R_Y \cdot (1-R_Y)}{n}} = 0.03$$

Konfidenzintervall: $[-0.04; 0.07]$

Verbundene Stichproben

- Gleiche Größe zweimal (davor - danach)
- Zwei Größen bei derselben Person
- „Matched Pair“

Hauptidee:

Verwende Differenzen $W_i = X_i - Y_i$

Konfidenzintervall für die Differenz von Mittelwerten (verbundene Stichproben)

Unterschied zwischen zwei verbundenen Gruppen $\mu_w = \mu_X - \mu_Y$

- Zwei verbundene Stichproben
- Daten: $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$,
- Stichprobenumfang hinreichend groß ($n \geq 30$)
- Schätzung: $\bar{W} = \frac{1}{n} \sum_{i=1}^n W_i = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i)$

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

$$\left[\bar{W} - z_{\frac{1+\gamma}{2}} \cdot \frac{S_w}{\sqrt{n}}; \bar{W} + z_{\frac{1+\gamma}{2}} \cdot \frac{S_w}{\sqrt{n}} \right]$$

mit:

- $S_w^2 = \frac{1}{n-1} \sum_{i=1}^n (W_i - \bar{W})^2$
- $z_{\frac{1+\gamma}{2}}$ ist das $\frac{1+\gamma}{2}$ -Quantil der Standardnormalverteilung