Statistik

Vorlesung 10 - Konfidenzintervalle

Prof. Dr. Sandra Eisenreich

Hochschule Landshut

Agenda

Konfidenzintervall

Bisher: Punktschätzung von Parametern ohne Gewissheit

Punktschätzer = ein Schätzwert für θ (z.B. die Trefferwahrscheinlichkeit) ohne Information wie gut die Schätzung ist

Beispiele:

- "Wir schätzen dass 10% der Wähler die Partei A gewählt haben"
- "Wir schätzen dass 1% der Schrauben defekt sind"

Ziel - erste Variante: Finde heraus wie sicher wir uns bei einer Schätzung sein können!

Problem: Bei der Schätzung einer stetigen Größe können wir einem einzelnen Punkt keine Wahrscheinlichkeit zuweisen. $\rightarrow$ aber Intervallen!

Motivation: Berechnen der Sicherheit eines geschätzten Intervalls

Gegeben: gewünschte Sicherheit unserer Prognose = Konfidenzniveau

Gesucht: Das Intervall, in dem mit dieser Sicherheit der richtige Wert liegt = Konfidenzintervall

Beispiele:

- "mit einer Wahrscheinlichkeit von 95% haben zwischen 9% und 11% der Wähler die Partei A gewählt"
- "mit 99% Sicherheit beträgt der Anteil der defekten Schrauben in der Kiste weniger als 1%".

Wie? Zwei (!) Schätzer: T_u und T_o, die für eine Stichprobe eine untere und obere Intervallsgrenze liefern.

Konfidenzintervall

Definition (Konfidenzintervall)

Es seien $(\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ ein statistisches Modell und $\alpha \in (0,1)$. Sei X eine gemäß P_θ verteilte Stichprobe. Es seien $T_u : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ und $T_o : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ zwei Stichprobenfunktionen mit

$$T_u(x) \leq T_o(x) \quad \text{für alle } x \in \mathcal{X}.$$

Wir sagen, dass $[T_u, T_o]$ ein Konfidenzintervall für θ zum Konfidenzniveau $1 - \alpha \in (0,1)$ ist, falls

$$P_\theta(T_u(X) \leq \theta \leq T_o(X)) \geq 1 - \alpha \quad \text{für alle } \theta \in \Theta.$$

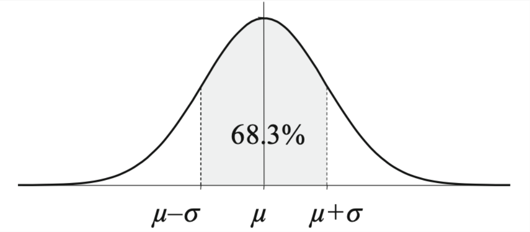
Wichtige Grundregeln zu Konfidenzintervallen

- in der Regel wird das Konfidenzniveau $\gamma = 1 - \alpha$ vorgegeben
- größeres γ/kleineres $\alpha \Rightarrow$ breiteres Intervall
- kleineres γ/größeres $\alpha \Rightarrow$ kleineres Intervall
- verschiedene Stichproben können verschiedene Intervalle ergeben

Konfidenzintervalle können für verschiedene Verteilungen bestimmt werden.
Hier: nur die Binomialverteilung.

Herleitung: Konfidenzintervall

Angenommen, eine Zufallsvariable X ist normalverteilt (binomial ist nahe dran). Dann kann man genau sagen, wie wahrscheinlich eine Stichprobe x von X in einem gewissen Intervall um den Mittelwert ist: zum Beispiel die Wahrscheinlichkeit, $< \sigma$ von μ weg zu sein:

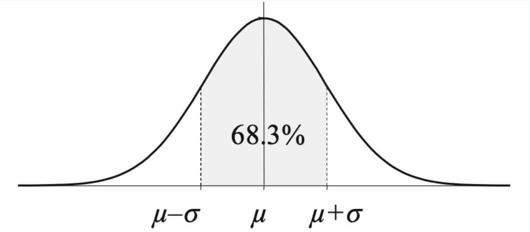


Herleitung: Konfidenzintervall

Gegeben: Stichprobe, gewünschte Konfidenz γ

Gesucht: In welchem Intervall liegt der Parameter höchstwahrscheinlich? (mit vorgegebener Konfidenz)

Wissen: Mit welcher Wahrscheinlichkeit P (abhängig vom Parameter) liegt eine Stichprobe wo?



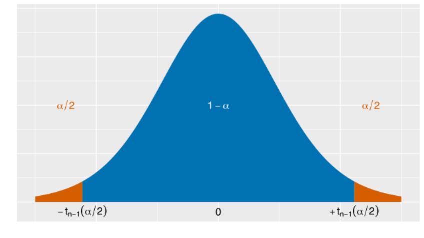
Idee: setze $P(x \in \text{Intervall}) = \gamma$ und löse nach dem Intervall auf.

Vorgehen: Konfidenzintervall für Binomialverteilung

Gegeben: Konfidenzniveau $\gamma = 1 - \alpha \in [0, 1]$, Stichprobe x, einer ZV $X \sim b_{n,p}$ mit

$$b_{np} \simeq N(np, np(1-p)).$$

- 1. Schritt: Beschreibe die Wahrscheinlichkeit, dass die Stichprobe x in einem Intervall mit Wahrscheinlichkeit γ liegt.
- 2. Schritt: Löse auf nach dem Intervall.



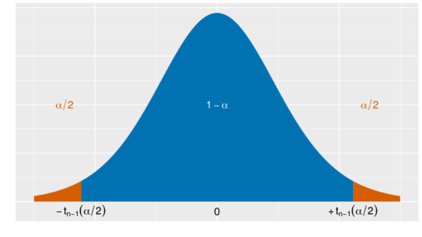
1. Schritt

Ziel: 1. Schritt: Beschreibe die Wahrscheinlichkeit, dass die Stichprobe x in einem Intervall mit Wahrscheinlichkeit γ liegt.

Eine Stichprobe x liegt im Intervall mit Wahrscheinlichkeit γ um den echten Mittelwert np genau dann wenn die Standardisierte

$$x^* = \frac{x - np}{\sqrt{np(1-p)}}$$

in der Standardnormalverteilung in dem Intervall mit Wahrscheinlichkeit γ um 0 ist



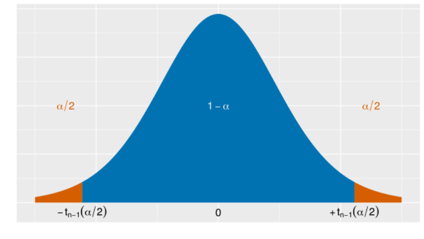
1. Schritt

Ziel: Beschreibe die Wahrscheinlichkeit, dass die Stichprobe x in einem Intervall mit Wahrscheinlichkeit γ liegt.

Wir erhalten also $\gamma = P(-c \leq X^* \leq c) = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1$ und berechnen daraus zunächst c:

$$2\Phi(c) - 1 = \gamma \Leftrightarrow \Phi(c) = \frac{\gamma + 1}{2}$$

$$\Rightarrow c = \Phi^{-1}(\frac{\gamma + 1}{2})$$



2. Schritt

Wir wissen bereits: Eine Stichprobe x liegt im Intervall mit Wahrscheinlichkeit γ um den echten Mittelwert np genau dann wenn $-c \leq x^* \leq c$, und $x^* = \frac{x-np}{\sqrt{np(1-p)}}$.

Ziel: Finde p, für die $|x^*| \leq c$ erfüllt ist. Das ist das Konfidenzintervall.

Rechnung:

$$|x^*| \leq c \Leftrightarrow \left|\frac{x-np}{\sqrt{np(1-p)}}\right| \leq c \Leftrightarrow \frac{(x-np)^2}{np(1-p)} \leq c^2 \Leftrightarrow (c^2+n)p^2 - (2x+c^2)p + \frac{x^2}{n} \leq 0$$

Die ist als Funktion von p eine Parabel, ≤ 0 zwischen den Nullstellen

$$p_{1,2} = \frac{1}{c^2+n}\left(X + \frac{c^2}{2} \pm c \cdot \sqrt{\frac{X(n-X)}{n} + \frac{c^2}{4}}\right)$$

Hier sind die Terme c^2, $\frac{c^2}{2}$ und $\frac{c^2}{4}$ sehr klein im Vergleich zum Rest, wir vernachlässigen sie.

Rechenregel

Ein Bernoulliexperiment mit unbekannter Wahrscheinlichkeit $p = P(A)$ werde n-mal durchgeführt, k-mal trete das Ereignis A ein. Es seien k und $n-k$ größer als 30. Zum Konfidenzniveau $\gamma = 1-\alpha$ gewinnt man das Konfidenzintervall $[p_u, p_o]$ für die unbekannte Wahrscheinlichkeit p von A durch

$$p_{u,o} = \frac{k}{n} \pm \frac{c}{n}\sqrt{\frac{k(n-k)}{n}}, \quad \text{mit } c = \Phi^{-1}\left(\frac{1+\gamma}{2}\right).$$

Die Konfidenzintervallbreite ist

$$B = 2 \cdot \frac{c}{n}\sqrt{\frac{k(n-k)}{n}}.$$

Beispiel: Wahlumfrage

Anzahl der Wähler der Partei A ist binomialverteilt. Umfrage bei 1000 Wählern, 550 wählen Partei A. Bestimme das Konfidenzintervall des Wähleranteils bei einem Konfidenzniveau von $\gamma = 1 - \alpha = 0.95$!

Achtung: Niemals γ und α verwechseln!

Ergebnis: Wahlumfrage

$$c = \Phi^{-1}\left(\frac{1.95}{2}\right) = \Phi^{-1}(0.975) = 1.96$$

$$p_{1,2} = \frac{1}{1000}\left(550 \pm 1.96\sqrt{\frac{550 \cdot 450}{1000}}\right)$$
$$\Rightarrow p_1 = 0.519, \quad p_2 = 0.581$$

$\Rightarrow$ Ergebnis: Mit 95% Wahrscheinlichkeit liegt das Resultat zwischen 51.9% und 58.1%.

Stichprobengröße

Motivation - Die Stichprobengröße

Je mehr Daten/Stichproben man hat, umso genauer (kleinere Intervallbreite, höhere Konfidenz) kann man Konfidenzintervalle bestimmen. Woher weiß ich, wie viele Stichproben ich brauche für gewünschte Genauigkeit?

$\rightarrow$ Frage: Wie groß muss n sein, damit bei vorgegebener Sicherheit (bzw. c) das Konfidenzintervall kleiner ist als eine gewisse Breite B?

Herleitung: Die Stichprobengröße

Wenn Konfidenzniveau und maximale Konfidenzintervallbreite vorgegeben sind, dann kann man die notwendige Stichprobengröße berechnen.

Breite B ist abhängig von k:

$$B(k) = \frac{2c}{n}\sqrt{\frac{k(n-k)}{n}}$$

Für welches k ist das maximal?

Ergebnis: Die Stichprobengröße

Bestimme Ableitung $B'(k) = 0$.

Damit ergibt sich: $k = \frac{n}{2}$ ist das Maximum.
Gegeben sei c, suche das n, für das eine vorgegebene Konfidenzintervallbreite erreicht wird.

$$B = \frac{2c}{n}\sqrt{\frac{\frac{n}{2}\cdot\frac{n}{2}}{n}} \quad \Rightarrow \quad B^2 = \frac{4c^2}{n^2}\frac{\frac{n^2}{4}}{n} = \frac{c^2}{n} \quad \Leftrightarrow \quad n = \frac{c^2}{B^2}$$

Rechenregel zur Stichprobengröße beim Bernoulliexperiment

Rechenregel

In einem Bernoulliexperiment sei das Konfidenzniveau γ vorgegeben. Das Konfidenzintervall um den unbekannten Parameter $p = P(A)$ hat maximal die Breite B, wenn für die Stichprobengröße gilt:

$$n \geq \frac{c^2}{B^2}, \quad \text{mit } c = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

Literatur

- Hartmann, Peter; Mathematik für Informatiker, Springer-Vieweg; 7. Auflage; 2019
- Henze, Norbert; Stochastik für Einsteiger; Springer; 10. Auflage; 2013