Konfidenzintervalle

- Grundlegendes Prinzip
- Erwartungswert
  - Bekannte Varianz
  - Unbekannte Varianz
- Anteilswert
- Differenzen von
  - Erwartungswert
  - Anteilswert

Beispiel für Konfidenzintervall

Im Prinzip haben wir bereits letztes mal ein Beispiel für ein Konfidenzintervall kennengelernt:

Sei $X \sim N(\mu,\sigma^2)$ und α eine vorgegebene Wahrscheinlichkeit. Dann ist $(\mu - x_\alpha , \mu + x_\alpha)$ ein sogenanntes α - Konfidenzintervall, wobei

$$x_\alpha = \sigma \cdot Q^{(N)}(1 - \tfrac{\alpha}{2})$$

Die mathematische Idee ist genau, dass

$$\mathbf{P}(\mu - x_\alpha \leq X \leq \mu + x_\alpha) = \alpha$$

Interpretation

In der Statistik wollen wir aufgrund von einer Stichprobe gewisse Kenngrößen wie Mittelwert oder Standardabweichung schätzen.

Wie wir diese Größen schätzen haben wir bereits in der deskriptiven Statistik kennengelernt. Es stellt sich nun die Frage, wie genau diese Schätzer sind.

Konfidenzintervalle sind dazu eine Möglichkeit. Sie liefern ein Intervall, in dem der geschätzte Parameter mit einer vorgegebenen Wahrscheinlichkeit (typischerweise 95%) liegt.

Arithmetisches Mittel, σ bekannt

Angenommen wir haben eine Stichprobe der Größe n und wir kennen die Standardabweichung σ der zugrunde liegenden Zufallsvariable.

Dann wissen wir: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

$\overline{X}$ soll unser Schätzer für μ sein.

Für ein 95% Konfidenzintervall brauchen wir zunächst $Q^{(N)}((1+0.95)/2) = Q^{(N)}(0.975) = 1.96$

Wir erhalten als 95% CI von μ:

$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

Beispiel 5-1

Stichprobe von 20 Brotwecken

Mittelwert 1015g, Varianz 2500g^2

Formel für das 95% Konfidenzintervall

$$
(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})
$$
$$
= (1015 - 22, 1015 + 22) = (993, 1037)
$$

Für Stichprobe mit 50 Brotleiben erhalten wir:

(1015-14, 1015+14) = (1001, 1029)

Die Länge des Intervalls schrumpft von 44g auf 28g.

Bestimmung der Stichprobengröße

Speziell bei der Planung von Studien steht man oft vor der Aufgabe die Stichprobengröße n festzulegen.

Eine Möglichkeit besteht darin, die maximale Länge L des gewünschten Konfidenzintervalls festzulegen und daraus n zu berechnen:

$$L \geq \mu + \frac{\sigma}{\sqrt{n}}\ Q^{(N)}() - [\mu - \frac{\sigma}{\sqrt{n}}\ Q^{(N)}()]$$

$$\rightarrow \quad n_L \geq [2\sigma\ Q^{(N)}(1 - \tfrac{\alpha}{2})/L]^2$$

In der Praxis schauen die Überlegungen zur Bestimmung der Stichprobengröße allerdings etwas anders aus (vgl. Kapitel Testen)

Beispiel 5-2

Mittlere Lebensdauer von Motoren,

Es wird angenommen $\sigma = 1000$

Gesucht: CI mit Breite $\pm$ 200

Lösung: Einfach in Formel Einsetzen

$$n_L \geq [2 \cdot 1000 \cdot 1.96/400]^2 = 96.04$$

Arithmetisches Mittel, σ unbekannt

Wir haben wieder eine Stichprobe der Größe n, in den allermeisten Fällen ist σ jedoch nicht bekannt.

Entsprechend den Überlegungen am Ende von Kapitel 4 müssen wir nun mit der T-Verteilung arbeiten.

Das α-Konfidenzintervall ist wieder von der Form $(\mu - x_\alpha , \mu + x_\alpha)$. Allerdings gilt nun:

$$x_\alpha = \frac{s}{\sqrt{n}} \cdot Q^{(t)}_{n-1}(1 - \tfrac{\alpha}{2})$$

Das heißt wir ersetzen die Quantile der Normalverteilung durch jene der T-Verteilung mit df = n-1 Freiheitsgraden.

Konfidenzintervalle für Anteilswerte

Binomialexperiment (Ziehen mit zurücklegen)

n groß → $\hat{p} \sim N(p, \frac{pq}{n})$

Approximatives α-Konfidenzintervall für Anteilswerte
Basierend auf Normalverteilung:

$$[\widehat{p} - Q^{(N)}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + Q^{(N)}(1-\tfrac{\alpha}{2})\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}]$$

Stichprobenumfang: $n_L \geq [Q^{(N)}(1-\frac{\alpha}{2})/L]^2$

Beachte: Abschätzung des Stichprobenumfangs ergibt sich hier daraus, dass man annimmt man weiß nichts über p!

Beispiel 5-5

Wieviel Prozent von Bankkunden suchen Wohnung?

Stichprobe n = 400, davon 88 auf Suche

Gesucht: 90%-Konfidentintervall für p

$$\hat{p} = 88/400 = 0.22 \qquad Q^{(N)}(0.95) = 1.645$$

$$[0.22 - 1.645\sqrt{\frac{0.22 \cdot 0.78}{400}}, 0.22 + 1.645\sqrt{\frac{0.22 \cdot 0.78}{400}}]$$

$$= [0.186, 0.254]$$

Konfidenzintervalle für Differenz zweier Mittelwerte

Häufige Fragestellung: Unterscheiden sich zwei Gruppen hinsichtlich eines (metrischen) Merkmals?

Antwort liefert der Vergleich der Mittelwerte des Merkmals pro Gruppe.

Zwei mögliche Szenarien:

1. Zwei unabhängige Stichproben jeweils von der ersten Gruppe und der zweiten Gruppe
2. Verbundene Stichproben: Zum Beispiel Vorher-Nachher Vergleich bei einer Stichprobe, oder Untersuchung von Paaren, etc.

Vorher-Nachher Vergleiche

Gepaarte Stichprobe der Größe n.

Bilde Differenzen der Werte zwischen den beiden Gruppen

(Vorher – Nachher) und berechne Konfidenzintervall dafür!

$(\mu - x_\alpha, \mu + x_\alpha)$ mit $x_\alpha = \frac{s}{\sqrt{n}} \cdot Q^{(t)}_{n-1}(1 - \frac{\alpha}{2})$

Hier ist μ der Mittelwert der Differenzen und s die geschätzte Standardabweichung der Differenzen

Bemerkung: Falls σ bekannt verwende Normalverteilung

Vorher-Nachher Vergleiche

Bsp. 5-7, Gewichtsabnahme bei Kurbesuch

Vor	85	78	92	103	94	89	84	82	109	102
Nh	78	75	90	93	93	83	85	79	98	96

Dif	7	3	2	10	1	6	-1	3	11	6

$n = 10$, $E(X) = 4.8$

$Q_9^{(t)}(0{,}975) = 2.262$

$Std(X) = 1.228$

$\rightarrow$ $CI = [2.0, 7.6]$

Unabhängige Stichproben

Die Berechnung der Konfidenzintervalls wird hier leider etwas komplizierter. Man unterscheidet wie gewohnt die Fälle mit bekannter Varianz und unbekannter Varianz.

Es stellt sich nun allerdings zusätzlich die Frage, ob sich in beiden Gruppen die Varianz unterscheidet oder nicht.

Wir betrachten folgende Fälle:

1. Bekannte Varianzen
2. Unbekannte Varianzen, in beiden Gruppen gleich
3. Unbekannte Varianzen, in beiden Gruppen verschieden

Unabhängige Stichproben, bekannte Varianzen σ_1^2, σ_2^2

Dieser Fall kommt in der Praxis kaum vor, wird aber aus didaktischen Gründen behandelt:

$$[\mu_D - Q^{(N)}(1-\tfrac{\alpha}{2})\sigma_D, \mu_D + Q^{(N)}(1-\tfrac{\alpha}{2})\sigma_D]$$

wobei $\mu_D = \overline{x}_1 - \overline{x}_2$ und $\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

σ_D ist die Standardabweichunge der Differenzen der Mittelwerte, deren Berechnung unmittelbar einsichtig ist

Unabhg. Stichproben, unbekannte Varianzen, Annahme: $\sigma_1^2 \neq \sigma_2^2$

α – Konfidenzintervall:

$$[\mu_D - Q_\nu^{(t)}(1-\tfrac{\alpha}{2})s_D, \mu_D + Q_\nu^{(t)}(1-\tfrac{\alpha}{2})s_D]$$

wobei $\mu_D = \overline{x}_1 - \overline{x}_2,$ $\quad s_D = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

und $\nu = \dfrac{s_D^4}{\frac{c_1}{n_1-1} + \frac{c_2}{n_2-1}}$ mit $c_i = \left(\dfrac{s_i^2}{n_i}\right)^2$

s_D ist nun einfach der aus den Daten berechnete Schätzer für σ_D. Freiheitsgrad ν muss gerundet werden!

Unabhg. Stichproben, unbekannte Varianzen, Annahme: $\sigma_1{}^2 = \sigma_2{}^2$

α – Konfidenzintervall:

$$[\mu_D - Q^{(t)}_{n_1+n_2-2}(1-\tfrac{\alpha}{2})s_P, \mu_D + Q^{(t)}_{n_1+n_2-2}(1-\tfrac{\alpha}{2})s_P]$$

wobei $\mu_D = \overline{x}_1 - \overline{x}_2$

und $$s_P = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\ \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

s_P ist die so genannte gepoolte Varianz

Bemerkung zur gepoolten Varianz

Unter der Annahme von $\sigma_1^2 = \sigma_1^2$ möchte man jene Gruppe mit mehr Beobachtungen stärker gewichten:

$$\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum(x_{1,i}-\overline{x_{1,.}})^2+\sum_j(x_{2,j}-\overline{x_{2,.}})^2}{n_1+n_2-2}$$

Die gepoolte Varianz ist als eine Art Mittelwert der Varianzen zwischen den beiden Gruppen zu interpretieren.

Wir haben df = n_1+n_2-2 Freiheitsgrade, weil zwei Mittelwerte aus den Daten geschätzt werden.

Konfidenzintervall für die Differenz zweier Anteilswerte

Wie für den Anteilswert selbst wird auch hier einfach mittels Normalverteilungs-Approximation das Konfidenzintervall bestimmt

$$[p_D - Q^{(N)}(1-\tfrac{\alpha}{2})s_D, p_D + Q^{(N)}(1-\tfrac{\alpha}{2})s_D]$$

mit $p_D = \widehat{p}_1 - \widehat{p}_2;\ s_D = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$

s_D ist der Schätzer für die Standardabweichung von $p_1 - p_2$

Verständnisfragen

- Was ist der Unterschied zwischen s und σ?
- Abhängigkeit der Intervallänge von
  1. Stichprobengröße
  2. Varianz der Daten
  3. Signifikanzniveau α
- Wie groß müsste ein Konfidenzintervall sein, um den Parameter μ von normalverteilten Zufalls-variablen sicher zu enthalten?