Wahrscheinlichkeitsrechnung und
Statistik für Biologen

7. Konfidenzintervalle

Martin Hutzenthaler & Dirk Metzler

1. Juni 2009

Inhaltsverzeichnis

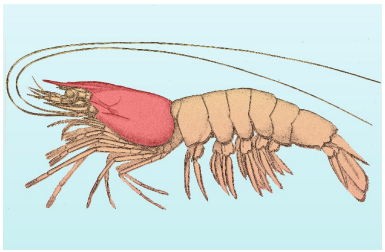
1 Konfidenzintervalle für Erwartungswerte

1.1 Beispiel: Carapaxlänge des Springkrebses

Beispiel: Springkrebs

Galathea squamifera

Carapaxlänge:



Wie groß ist die mittlere Carapaxlänge des weiblichen Springkrebses?

Alle weiblichen Springkrebse (also die Grundgesamtheit) zu vermessen, ist zu aufwändig.

Idee: Aus einer Stichprobe läßt sich die mittlere Carapaxlänge schätzen.

Wie genau ist diese Schätzung?

Ziel: Ein Intervall, in dem der Mittelwert der Carapaxlängen aller weiblichen Springkrebse mit hoher Wahrscheinlichkeit liegt.

Dieses Intervall nennen wir Konfidenzintervall oder Vertrauensbereich für den wahren Wert.

Galathea: Carapaxlänge in einer Stichprobe

Weibchen: $\overline{x} = 3.23$ mm $sd(x) = 0.9$ mm $n = 29$ $sem(x) = \frac{sd(x)}{\sqrt{n}} = \frac{0.9}{\sqrt{29}} = 0.17$ $(= sd(\overline{x}))$

Wir kennen bereits folgende Faustformeln:

- 2/3-Faustformel: Der wahre Mittelwert liegt im Intervall

$$\left[\overline{x} - sem(x), \overline{x} + sem(x)\right]$$

mit Wahrscheinlichkeit nahe bei 2/3

- 95%-Faustformel: Der wahre Mittelwert liegt im Intervall

$$\left[\overline{x} - 2 * sem(x), \overline{x} + 2 * sem(x)\right]$$

mit Wahrscheinlichkeit nahe bei 95%

Nun exakt: Sei $t_{5\%}$ `<- -qt(0.025,length(x)-1)`. Dann liegt der wahre Mittelwert mit Wahrscheinlichkeit 95% im Intervall

$$\left[\overline{x} - t_{5\%} * sem(x), \overline{x} + t_{5\%} * sem(x)\right]$$

Zur Begründung siehe nächster Abschnitt.

Setzt man die Zahlenwerte $\overline{x} = 3.23$, $t_{5\%} = 2.05$ und $\mathrm{sem}(x) = 0.17$ in

$$\left[\overline{x} - t_{5\%} * sem(x), \overline{x} + t_{5\%} * sem(x)\right]$$

ein, so erhält man das Konfidenzintervall

$$[2.88, 3.58]$$

für den wahren Mittelwert zum Irrtumsniveau 5%.

1.2 Theorie

Konfidenzintervall für den wahren Mittelwert

Ziel: Bestimme das Konfidenzintervall für den wahren Mittelwert zum Irrtumsniveau α

Das Konfidenzintervall für den wahren Mittelwert zum Irrtumsniveau α ist ein aus den Daten $x = (x_1, \ldots, x_n)$ geschätztes (zufälliges) Intervall

$$\left[\underline{I}(x), \overline{I}(x)\right]$$

mit folgender Eigenschaft: Ist der wahre Mittelwert gleich μ und ist $(x_1, \ldots, x_n)$ eine Stichprobe aus der Grundgesamtheit (mit Mittelwert μ), so gilt

$$\Pr\left(\mu \in \left[\underline{I}(x), \overline{I}(x)\right]\right) \geq 1 - \alpha$$

Selbstverständlich wollen wir das Konfidenzintervall möglichst klein wählen.

Konfidenzintervall für den wahren Mittelwert

Lösung: Wir wissen bereits (->Normalapproximation), dass die t-Statistik

$$t := \frac{\overline{x} - \mu}{\mathrm{sem}(x)}$$

annähernd Student-verteilt ist mit `length(x)-1` Freiheitsgraden (wenn `length(x)` groß genug ist).

Sei t_α `<- -qt(`$\alpha/2$`,length(x)-1)` das $\alpha/2$-Quantil der Student-Verteilung mit `length(x)-1` Freiheitsgraden. Dann ist

$$\left[\overline{x} - t_\alpha * sem(x), \overline{x} + t_\alpha * sem(x)\right]$$

das Konfidenzintervall zum Irrtumsniveau α.

Begründung:

$$\begin{aligned}
&\Pr\left(\mu \in \left[\overline{x} - t_\alpha * \mathrm{sem}(x), \overline{x} + t_\alpha * \mathrm{sem}(x)\right]\right) \\
&= \Pr\left(\mu - \overline{x} \in \left[-t_\alpha * \mathrm{sem}(x), t_\alpha * \mathrm{sem}(x)\right]\right) \\
&= \Pr\left(\frac{\mu - \overline{x}}{\mathrm{sem}(x)} \in \left[-t_\alpha, t_\alpha\right]\right) \\
&= \Pr\left(\left|\frac{\mu - \overline{x}}{\mathrm{sem}(x)}\right| \leq t_\alpha\right) \\
&= \Pr\left(|t| \leq t_\alpha\right) \\
&= 1 - \alpha
\end{aligned}$$

Beachte: t_α wird gerade so gewählt, dass die letzte Gleichung richtig ist.

Konfidenzintervall allgemein

Sei θ ein Parameter der zu Grunde liegenden Verteilung.

Das Konfidenzintervall für den Parameter θ zum Irrtumsniveau α ist ein aus den Daten $x = (x_1, \ldots, x_n)$ geschätztes (zufälliges) Intervall

$$\left[\underline{I}(x), \overline{I}(x)\right]$$

mit folgender Eigenschaft: Ist der wahre Parameter gleich θ und ist $(x_1, \ldots, x_n)$ eine Stichprobe aus der Grundgesamtheit (mit Parameter θ), so gilt

$$\Pr\left(\theta \in \left[\underline{I}(x), \overline{I}(x)\right]\right) \geq 1 - \alpha$$

Wie das Konfidenzintervall im Allgemeinen aussieht, ist unbekannt.

2 Konfidenzintervalle für Wahrscheinlichkeiten

2.1 Beispiel: Porzellankrebs

(c): public domain

Familie: *Porcellanidae*

In einem Fang vom 21.02.1992 in der Helgoländer Tiefe Rinne waren 23 Weibchen und 30 Männchen (*Pisidiae longicornis*), d.h. der Männchenanteil in der Stichprobe war $30/53 = 0{,}57$.

Was sagt uns dies über den Männchenanteil in der Population?

Was ist ein 95%-Konfidenzintervall für den Männchenanteil in der Population? (0,57±??)

2.2 Theorie

Wir beobachten X Männchen in einer Stichprobe der Größe n und möchten den (unbekannten) Männchenanteil p in der Gesamtpopulation schätzen.

Der offensichtliche Schätzer ist die relative Häufigkeit $\widehat{p} := \frac{X}{n}$ in der Stichprobe.

Frage: Wie verläßlich ist die Schätzung?

Gewünscht: Ein in Abhängigkeit von den Beobachtungen konstruiertes (und möglichst kurzes) Intervall $[\widehat{p}_u, \widehat{p}_o]$ mit der Eigenschaft

$$\Pr_p\left([\widehat{p}_u, \widehat{p}_o] \text{ überdeckt } p\right) \geq 1 - \alpha$$

für *jede Wahl* von p.

Lösungsweg:

Für gegebenes p ist X Binomial(n,p)-verteilt, $\mathrm{E}[X] = np$, $\mathrm{Var}[X] = np(1-p)$.

Wir wissen: Der Schätzer $\widehat{p}$ ist (in etwa) normalverteilt mit Erwartungswert p und Standardabweichung $\sqrt{\frac{p(1-p)}{n}}$. Dies können wir allerdings nicht verwenden, da p unbekannt ist.

Stattdessen nutzen wir wieder die Student-Verteilung, wobei hier als Schätzer für die Standardabweichung von $\widehat{p}$

$$\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}$$

verwendet wird. (Das Quadrat hiervon ist ein erwartungstreuer Schätzer der Varianz. In der Literatur wird auch der ML-Schätzer $\sqrt{\widehat{p}(1-\widehat{p})/n}$ verwendet.)

Lösung:

Sei $\widehat{p}$ die relative Häufigkeit in der Stichprobe der Länge n. Das Konfidenzintervall zum Irrtumsniveau α ist

$$\left[\widehat{p} - t_{\alpha,n-1} * \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}, \widehat{p} + t_{\alpha,n-1} * \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}\right]$$

wobei $t_{\alpha,n-1}$ `<- -qt(`α`/2,n-1)` das $\alpha/2$-Quantil der Student-Verteilung mit `n-1` Freiheitsgraden ist.

2.3 Beispiel: Porzellankrebs

Männchenanteil beim Porzellankrebs

Setzt man die Zahlenwerte $n = 53$, $\widehat{p} = 0.566$, $t_{5\%,52} = 2.007$ und $\sqrt{\widehat{p}(1-\widehat{p})/(n-1)} = 0.0687$ in

$$\left[\widehat{p} - t_{5\%,52} * \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}, \widehat{p} + t_{5\%,52} * \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}\right]$$

ein, so erhält man das Konfidenzintervall

$$[0.428, 0.704] = 0.566 \pm 0.138$$

für den wahren Männchenanteil zum Irrtumsniveau 5%.

2.4 Beispiel: Stockente

image (c) Andreas Trepte

Anas platyrhynchos

Stockente (engl. mallard)

Füchse jagen Stockenten. Durch ihre auffällige Färbung sind dabei Männchen leichter zu erspähen. Hat dies einen Einfluss auf das Geschlechterverhältnis bei amerikanischen Stockenten?

Daten: Stichprobe der Länge $n = 2200$. Relative Häufigkeit der Männchen war 0.564.

Daten aus:

Literatur

[Smi68] Johnson, Sargeant (1977) Impact of red fox predation on the sex ratio of prairie mallards *United States fish & wild life service*

Setzt man die Zahlenwerte $n = 2200$, $\widehat{p} = 0.564$, $t_{5\%,2199} = 1.961$ und $\sqrt{\widehat{p}(1-\widehat{p})/(n-1)} = 0.011$ in

$$\left[\widehat{p} - t_{5\%} * \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}, \widehat{p} + t_{5\%} * \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}\right]$$

ein, so erhält man das Konfidenzintervall

$$\left[0.543, 0.585\right] = 0.564 \pm 0.021$$

für den wahren Männchenanteil zum Irrtumsniveau 5%.

2.5 Anmerkungen

- Für die Gültigkeit der Approximation muss n genügend groß und p nicht zu nahe an 0 oder 1 sein. (Eine häufig zitierte „Faustregel“ ist „$np \geq 9$, $n(1-p) \geq 9$“.)
- Die Philosophie der Konfidenzintervalle entstammt der *frequentistischen* Interpretation der Statistik: Für jede Wahl des Parameters p würden wir bei häufiger Wiederholung des Experiments finden, dass in (ca.) $(1-\alpha)\cdot 100\%$ der Fälle das (zufällige) Konfidenzintervall den „wahren“ (festen) Parameter p überdeckt.