Hochschule Niederrhein
University of Applied Sciences

Elektrotechnik
und Informatik
Faculty of Electrical Engineering
and Computer Science

Konstruktionsmethoden für Konfidenzintervalle

Christoph Dalitz

Technischer Bericht Nr. 2017-01

IMPRESSUM

Technische Berichte des Fachbereichs Elektrotechnik und Informatik,
Hochschule Niederrhein

ISSN 2199-031X

HERAUSGEBER

Christoph Dalitz und Steffen Goebbels
Fachbereich Elektrotechnik und Informatik

ANSCHRIFT

Hochschule Niederrhein
Reinarzstr. 49
47805 Krefeld

http://www.hsnr.de/fb03/technische-berichte/

C. Dalitz: „Konstruktionsmethoden für Konfidenzintervalle.“ Technischer Bericht Nr. 2017-01, pp. 1-14, Hochschule Niederrhein, Fachbereich Elektrotechnik und Informatik, 2017

Technischer Bericht Nr. 2017-01, pp. 1-14, Hochschule Niederrhein, Fachbereich Elektrotechnik & Informatik (2017)

Konstruktionsmethoden für Konfidenzintervalle*

Christoph Dalitz
Institut für Mustererkennung
Hochschule Niederrhein
Reinarzstr. 49, 47805 Krefeld
christoph.dalitz@hsnr.de

Zusammenfassung

Statistik-Lehrbücher beschränken sich beim Thema "Konfidenzintervalle" üblicherweise auf das klassische "zwei sigma" Intervall für den statistischen Mittelwert und geben für andere Schätzwerte in der Regel keine Verfahren an, um Konfidenzintervalle zu erhalten. Dieser technische Bericht füllt diese Lücke, indem er zunächst verschiedene allgemeine Ansätze zur Konstruktion von Konfidenzintervallen beschreibt und diese dann auf relative Häufigkeiten, statistische Mittelwerte und beliebige Schätzer anwendet. Neben dem frequentistischen Ansatz werden auf dem Likelihood-Ratio und der Posterior Density basierende Ansätze erläutert. Für allgemeine Maximum-Likelihood Schätzer werden zwei Methoden (Hesse-Matrix, Jackknife) zum Schätzen der Varianz vorgestellt, und für beliebige Schätzer die Bootstrap-Methode. Die verschiedenen Konfidenzintervalle werden an typischen Beispielen quantitativ anhand von Coverage Probability und Intervallbreite mittels Monte-Carlo Simulationen verglichen. Für alle Methoden wird R-Code angegeben, mit dem die Konfidenzintervalle direkt berechnet werden können. Der Praktiker erhält Empfehlungen, welche Methode in welchen Fällen sinnvoll ist.

1 Überblick

Wenn man einen unbekannten Modellparameter aus Messdaten schätzt, so erhält man immer einen Wert, unabhängig davon, wie viele Messdaten in die Schätzung eingegangen sind. Bei einer größeren Datenbasis sollte der Schätzwert aber genauer sein. Ein *Konfidenzintervall* soll diese "Genauigkeit" des Schätzwerts quantifizieren, wobei es allerdings über die Definition von "Genauigkeit" unterschiedliche Auffassungen gibt. Deshalb gibt es verschiedene Ansätze zur Konstruktion von Konfidenzintervallen.

Der *frequentistische Ansatz* basiert auf der *Coverage Probability* und ist der verbreitetste Ansatz, der auch in Statistik-Einführungsbüchern gelehrt wird [1]. Er nimmt den unbekannten Parameter als bekannt an und wählt das Intervall um den Schätzwert so, dass der Parameter mit vorgegebener Wahrscheinlichkeit (meist 95%) hinein fällt. Der *evidenzbasierte Ansatz* basiert auf dem *Likelihood-Ratio* und wählt ein Intervall, in dem die Likelihood-Funktion über einem vorgegebenen Schwellwert (meist $1/8$) liegt [2]. Der *Bayessche Ansatz* betrachtet den gesuchten Parameter als Zufallsgröße, dessen Wahrscheinlichkeitsverteilung (die "Posterior Density") sich aus der Beobachtung ergibt, was zum *Highest Posterior Density* Intervall führt [3].

Sowohl für relative Häufigkeiten als auch für den statistischen Mittelwert kann man für alle drei Ansätze Formeln oder Algorithmen zur Berechnung eines Konfidenzintervalls angeben. Ein mögliches Kriterium zur Evaluation der verschiedenen Intervalle ist die Coverage Probability. Man könnte meinen, dass dieses Kriterium den frequentistischen Ansatz bevorzöge, aber auch bei diesem Ansatz kann die Coverage Probabilty je nach wahrem Wert des Parameters stark variieren. Bei Konfidenzintervallen, die nicht symmetrisch um den Schätzwert liegen, ist ein weiteres Gütekriterium die Breite des Intervalls, denn in diesem Fall können verschiedene Intervalle die gleiche Coverage Probability haben, von denen das kürzere Intervall zu bevorzugen ist.

Für andere Schätzer als die relative Häufigkeit oder den statistischen Mittelwert gibt es keine vorgefertigte Standardformel zur Berechnung eines Konfidenzintervalls. Ist dieser Schätzer jedoch ein Maximum-Likelihood Schätzer, so weiß man, dass er bei einer glatten Likelihoodfunktion asymptotisch normalverteilt ist [4]. Folglich kann man das Konfidenzintervall für den Mittelwert verwenden, sofern man die Varianz des Schätzers wiederum schätzen kann. Zum Schätzen der Varianz gibt es zwei allgemeine Metho-

*For an Englisch version of this article, see pp. 15-28 of this report.

den: die Diagonale der invertierten *Hesse-Matrix* der log-Likelihood Funktion oder die *Jackknife* Varianz.

Für nicht-glatte Likelihoodfunktionen oder für andere Schätzer ist universell anwendbar einzig das *Bootstrap*-Verfahren. Dabei werden neue Daten aus den Messdaten durch zufälliges Ziehen mit Zurücklegen generiert und aus deren Verteilung wird das Konfidenz-Intervall geschätzt. Im Prinzip ist die Bootstrap-Methode immer anwendbar, auch in den Fällen, in denen andere Methoden anwendbar sind, aber in den in diesem Bericht beschriebenen Experimenten war die Coverage Probability des klassischen Konfidenzintervalls durchgängig besser.

Dieser Bericht ist wie folgt aufgebaut: Zunächst werden in Abschnitt 2 die Grundbegriffe Schätzer, Coverage Probability, Likelihood Ratio und Posterior Density erläutert. Dann werden in den Abschnitten 3 und 4 die verschiedenen Ansätze auf die relative Häufigkeit und den statistischen Mittelwert angewandt und R-Code zur Berechnung der Konfidenzintervalle angegeben. Die Abschnitte 5 und 6 beschreiben Konfidenzintervalle für Maximum-Likelihood Schätzer und für beliebige Schätzer. Abschnitt 7 vergleicht die Coverage Probability der verschiedenen Konfidenzintervalle in Monte-Carlo Experimenten. Der letzte Abschnitt gibt Empfehlungen, welches Konfidenzintervall in welchem Fall sinnvoll ist.

2 Grundbegriffe

Die Wahrscheinlichkeitsverteilung einer Zufallsvariable X sei bis auf den Wert eines Parameters θ bekannt, d.h. die Form der Wahrscheinlichkeitsdichte $f_\theta(x)$ sei bekannt, nicht jedoch der Wert des Parameters θ. Im allgemeinen wird θ ein Vektor sein, also mehrere Parameterwerte repräsentieren. Wenn X z.B. normalverteilt ist, dann ist $\theta = (\mu, \sigma^2)$, beinhaltet also zwei Parameter. Eine Funktion zum Schätzen der unbekannten Parameter aus unabhängigen Messwerten $x_1, \ldots, x_n$ der Zufallsvariablen X heißt *Schätzer* und der damit berechnete Schätzwert wird mit $\hat{\theta}$ bezeichnet:

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \tag{1}$$

Einfache Beispiele für Schätzer sind die relative Häufigkeit als Schätzer für die Wahrscheinlichkeit oder der statistische Mittelwert als Schätzer für den Parameter μ der Normalverteilung.

2.1 Maximum-Likelihood (ML)

Für kompliziertere Fälle gibt es mit dem *Maximum-Likelihood Prinzip* eine generische Methode, um eine Schätzfunktion zu erhalten [4]. Dabei wird der Parameter θ so gewählt, dass die *Likelihood-Funktion* L oder[1] die *Log-Likelihood-Funktion* ℓ maximiert wird:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) \tag{2a}$$

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i) \tag{2b}$$

Salopp gesagt ist $L(\theta)$ ein Maß für die Wahrscheinlichkeit der Beobachtung $x_1, \ldots, x_n$ unter der Annahme, dass der Parameterwert θ zugrunde liegt. Wenn $\ell(\theta)$ differenzierbar ist und $\theta = (\theta_1, \ldots, \theta_t)$, dann ergeben sich aus dem Maximum-Likelihood Prinzip t Bestimmungsgleichungen für die t Parameter $\theta_1, \ldots, \theta_t$:

$$\frac{\partial}{\partial \theta_i} \ell(\theta) = 0 \quad \text{für } i = 1 \ldots, t \tag{3}$$

Maximum-Likelihood Schätzer haben eine Reihe attraktiver Eigenschaften, wie z.B. dass sie unter recht allgemeinen Bedingungen asymptotisch normalverteilt sind, was im Abschnitt 5 eine Rolle spielen wird. In vielen Fällen wird man die Gleichungen (3) nicht geschlossen lösen können, so dass man auf eine numerische Maximierung der Log-Likelihood-Funktion zurückgreifen muss. Falls auch das nicht möglich ist, kann man auf die *Methode der Momente* oder deren Generalisierung zurückgreifen [5].

2.2 Coverage Probability

Eine Schätzfunktion (1) liefert lediglich einen isolierten Wert und wird deshalb auch als *Punktschätzer* bezeichnet. Im Unterschied dazu gibt ein *Konfidenzintervall* $[\theta_l, \theta_u]$ einen Bereich an, in dem der Wert mit hoher Wahrscheinlichkeit liegt. Die Grenzen $\theta_{l,u}$ des Intervalls hängen von den Messwerten $x_1, \ldots, x_n$ ab und sind somit Zufallsgrößen. Der *frequentistische Ansatz* besteht in folgender Überlegung: Wenn θ der wahre Wert des Parameters ist, dann sollte er idealerweise mit einer vorgegebenen *Coverage Probability* von $(1 - \alpha)$ im Konfidenzintervall liegen:

$$P_{cov}(\theta) = P(\theta \in [\theta_l, \theta_u]) = 1 - \alpha \tag{4}$$

[1] Weil der Logarithmus eine monotone Funktion ist, haben $L(\theta)$ und $\log L(\theta)$ das Maximum an derselben Stelle.

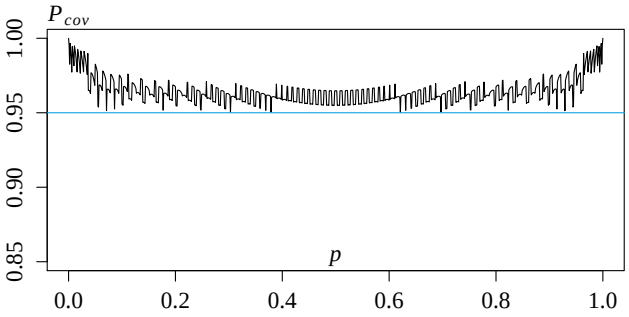


Abbildung 1: Coverage Probability P_{cov} des "exakten" Konfidenzintervalls für die relative Häufigkeit gemäß Glg. (5) als Funktion des wahren Parameters p für $n = 100$ und $\alpha = 0.05$.

Leider kann man Glg. (4) nicht zur Bestimmung von θ_l und θ_u benutzen, denn in diese Gleichung geht ja das unbekannte θ ein. Dieses Dilemma lässt sich lösen, wenn man das Problem als Hypothesentest uminterpretiert: unter der Hypothese $\theta \notin [\theta_l, \theta_u]$ soll die Wahrscheinlichkeit kleiner als α sein, dass der Schätzer stärker als der gemessene Wert für $\hat{\theta}$ von θ abweicht. In der Sprache des Hypothesentests ausgedrückt: wenn θ identisch mit einer der Intervallgrenzen wäre, dann läge alles jenseits von $\hat{\theta}$ im Ablehnungsbereich. Wenn man die Wahrscheinlichkeit α gleichmäßig auf große und kleine Abweichungen verteilt, erhält man somit die Definition des *frequentistischen Konfidenzintervalls*[2]:

$$P_{\theta=\theta_l}(\hat{\theta} \geq \theta_0) = \alpha/2 \quad \text{und} \tag{5a}$$

$$P_{\theta=\theta_u}(\hat{\theta} \leq \theta_0) = \alpha/2 \tag{5b}$$

wobei θ_0 der gemessene Wert für den Schätzer ist, und $P_{\theta=\theta_{l,u}}$ die Wahrscheinlichkeit unter der Annahme bezeichnet, dass der wahre Wert des Parameters die jeweilige Intervallgrenze wäre.

Das durch Auflösen von Glg. (5) nach θ_l und θ_u erhaltene Konfidenzintervall garantiert zwar unabhängig von θ eine Coverage Probability von mindestens $1-\alpha$, hat aber trotzdem zwei Haken: das Beispiel in Abb. 1 zeigt, dass selbst bei einem direkt mittels Glg. (5) berechneten "exakten" Konfidenzintervall die Coverage Probability für viele θ deutlich zu groß sein kann, so dass das Konfidenzintervall zu breit ist. Außerdem kennt man die Wahrscheinlichkeiten oft nur näherungsweise oder kann Glg. (5) nur asymptotisch lösen, so dass man nur ein approximatives Konfidenzintervall erhält, dessen $P_{cov}(\theta)$ auch kleiner als $1 - \alpha$ sein kann.

[2]Diese Definition liest sich bei DiCiccio & Efron [6] etwas anders: dort ist Glg. (5b) identisch, aber in Glg. (5a) verwenden sie "$>$" anstelle von "$\geq$". Für stetige Zufallsvariablen ist das egal, aber bei diskreten Zufallsvariablen würde dies die beiden Grenzen unterschiedlich behandeln.

2.3 Likelihood-Ratio

Ein anderer Ansatz, ein Konfidenzintervall zu erhalten, basiert auf der Likelihoodfunktion (2a). Der ML-Schätzer $\hat{\theta}$ wählt θ so, dass die Beobachtung maximale Wahrscheinlichkeit hat. Aber auch andere Werte in der Nähe von θ führen immer noch zu einer hohen Wahrscheinlichkeit der Beobachtung. Da ist es nahe liegend ein Intervall anzugeben, in dem das Verhältnis $L(\hat{\theta})/L(\theta)$ über einem Schwellwert liegt. Um dieses Intervall vom frequentistischen Konfidenzintervall zu unterscheiden, nennt man es das *Likelihood-Ratio Support Intervall* $[\theta_l, \theta_u]$:

$$\frac{L(\theta)}{L(\hat{\theta})} \geq \frac{1}{K} \quad \text{für alle } \theta \in [\theta_l, \theta_u] \tag{6}$$

wobei $\hat{\theta}$ der ML-Schätzer für θ ist. Meist wählt man $K = 8$, weil dieser Wert beim Konfidenzintervall für den statistischen Mittelwert ähnliche Werte wie das frequentistische Intervall für $\alpha = 0.05$ liefert (siehe Abschnitt 4.2).

2.4 Posterior Density

Ein dritter Ansatz für ein Konfidenzintervall versucht aus der Beobachtung $\hat{\theta}$ eine Wahrscheinlichkeitsverteilung für θ zu konstruieren. Der wahre Parameter θ wird hier also als eine Zufallsgröße betrachtet. Demzufolge ist $p_\theta(\hat{\theta})$ eine bedingte Wahrscheinlichkeitsdichte[3] $p(\hat{\theta}|\theta)$, die sich mit der Bayesschen Formel umschreiben lässt als

$$p(\theta|\hat{\theta}) = \frac{p(\hat{\theta}|\theta) \cdot p(\theta)}{\int_{\mathbb{R}} p(\hat{\theta}|\tau) \cdot p(\tau)\, d\tau} \tag{7}$$

Wenn man diese Dichte berechnen kann, dann ist das *Highest Posterior Density (HPD) Intervall* $[\theta_l, \theta_u]$ als der Bereich definiert, in dem die Wahrscheinlichkeit maximal ist und zugleich insgesamt den Wert $(1 - \alpha)$ hat. Formal führt das zu den gekoppelten Bestimmungsgleichungen (siehe Abb. 2)

$$1 - \alpha = \int_{\theta_l}^{\theta_u} p(\theta|\hat{\theta})\, d\theta \quad \text{und} \tag{8a}$$

[3]Weil θ und $\hat{\theta}$ stetige Größen sind, wird ihre Verteilung durch eine Dichte beschrieben, die hier mit dem Kleinbuchstaben p bezeichnet wird.

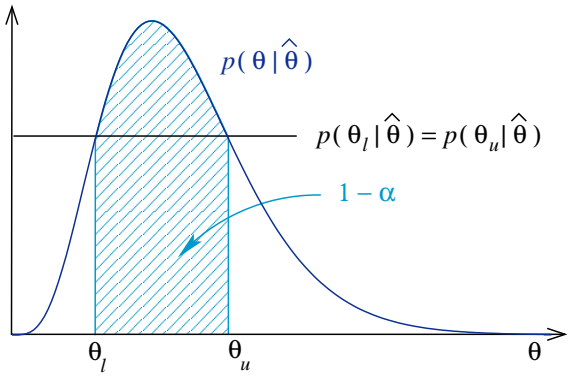


Abbildung 2: Bestimmung des Highest Posterior Density Intervalls $[\theta_l, \theta_u]$ gemäß Glg. (8).

$$p(\theta_l|\hat{\theta}) \quad = \quad p(\theta_u|\hat{\theta}) \tag{8b}$$

Abgesehen davon, dass dieses Gleichungssystem nur numerisch gelöst werden kann, hat das HPD Intervall einen fundamentalen Mangel: zur Bestimmung von $p(\theta|\hat{\theta})$ gemäß Glg. (7) muss man nämlich eine "a priori Verteilung" $p(\theta)$ für den unbekannten Parameter θ annehmen und diese Annahme ist willkürlich. Meist wählt man $p(\theta)$ konstant, so als gäbe es überhaupt kein Vorwissen über die ungefähre Lage des Parameters. Auch wenn diese Annahme in der Praxis niemals zutrifft, bedeutet das nicht notwendigerweise, dass das HPD-Intervall schlecht sein muss und vielleicht sogar eine brauchbare Coverage Probability aufweist.

3 Relative Häufigkeiten

Ein in der Praxis sehr häufig vorkommender Schätzer ist die relative Häufigkeit $\hat{p}$ als Schätzer der Wahrscheinlichkeit p. Die Wahrscheinlichkeitsverteilung der relativen Häufigkeit ist exakt durch die Binomialverteilung gegeben, d.h. wenn ein Ereignis die Wahrscheinlichkeit p hat, dann ist die Wahrscheinlichkeit, dass es bei n Versuchen k-mal auftritt

$$P_p(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{9}$$

bzw. für die relative Häufigkeit $\hat{p} = k/n$ gilt

$$P_p(\hat{p} = p_0) = \binom{n}{np_0} p^{np_0} (1-p)^{n(1-p_0)} \tag{10}$$

Glg. (10) bildet den Ausgangspunkt für alle Konfidenzintervalle der relativen Häufigkeit.

```
ci.binom <- function(n, k, alpha) {
  if (k == 0) {
    p1 <- 0.0
    p2 <- 1 - (alpha/2)**(1/n)
  }
  else if (k == n) {
    p1 <- (alpha/2)**(1/n)
    p2 <- 1.0
  }
  else {
    helper <- function(p, k, n, val) {
      return (pbinom(k, n, p) - val)
    }
    r <- uniroot(helper, k=(k-1),
                 n=n, val=1-alpha/2,
                 interval=c(0,1))
    p1 <- r$root
    r <- uniroot(helper, k=k,
                 n=n, val=alpha/2,
                 interval=c(0,1))
    p2 <- r$root
  }
  return (data.frame(p1=p1, p2=p2))
}
```

Listing 1: R Implementierung des exakten Clopper-Pearson Konfidenzintervalls für die relative Häufigkeit gemäß Glg. (11) & (12).

3.1 Frequentistisches Intervall für $\hat{p}$

Für die relative Häufigkeit ergibt Einsetzen der Verteilung (10) in Glg. (5) die Bestimmungsgleichungen für die Grenzen p_l und p_u:

$$1 - \text{pbinom}\,((k-1)/n, n, p_l) = \alpha/2 \tag{11a}$$

$$\text{und} \qquad \text{pbinom}\,(k/n, n, p_u) = \alpha/2 \tag{11b}$$

wobei $k/n = \hat{p}$ die gemessene relative Häufigkeit und *pbinom* die R-Funktion für die Verteilungsfunktion (CDF) der Binomialverteilung ist. Für die Spezialfälle $k = 0$ und $k = 0$ hat jeweils eine der Gleichungen (11) keine Lösung, weil p_l und p_u im Intervall $[0, 1]$ liegen müssen. In diesen Fällen ist $p_l = 0$ zu setzen ($k = 0$) bzw. $p_u = 1$ ($k = n$). Die andere Grenze lässt sich in diesem Fall analytisch lösen zu

$$k = 0 \quad \Rightarrow \quad [p_l, p_u] = [0, 1 - \sqrt[n]{\alpha/2}] \tag{12a}$$

$$k = n \quad \Rightarrow \quad [p_l, p_u] = [\sqrt[n]{\alpha/2}, 1] \tag{12b}$$

Für alle anderen Fälle muss Glg. (11) numerisch gelöst werden, z.B. mit der R-Funktion *uniroot*[4]. Der ent-

[4] Weil 1-*pbinom* die Incomplete Beta Function ist (siehe [7] Glg. 26.5.7), kann man die Gleichung auch über deren Inverse lösen, aber auch diese Inverse muss numerisch berechnet werden.

sprechende R-Code ist in Listing 1 zusammengefasst. Dieses Intervall ist das *Clopper-Pearson* Intervall [8], das auch die R-Funktion *binom.confint* im Paket *binom* mit der Option *method='exact'* berechnet.

Ein approximatives Konfidenzintervall erhält man, wenn man die Binomialverteilung nach dem zentralen Grenzwertsatz durch die Normalverteilung approximiert. Für große n ist $\hat{p}$ etwa normalverteilt zu $\mu = p$ und $\sigma^2 = p(1-p)/n$, so dass Glg. (5a) wird zu

$$\begin{aligned} & 1 - \text{pnorm}\left(\hat{p}, p_l, \sqrt{p_l(1-p_l)/n}\right) = \alpha/2 \\ \Leftrightarrow \quad & \text{pnorm}\left(\frac{\hat{p}-p_l}{\sqrt{p_l(1-p_l)/n}}, 0, 1\right) = 1-\alpha/2 \\ \Leftrightarrow \quad & \frac{\hat{p}-p_l}{\sqrt{p_l(1-p_l)/n}} = z_{1-\alpha/2} \end{aligned} \tag{13}$$

wobei *pnorm* die R-Funktion für die CDF der Normalverteilung ist und $z_{1-\alpha/2} = \text{qnorm}(1-\alpha/2)$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung. Wenn man die quadratische Glg. (13) und deren Entsprechung für p_u auflöst, erhält man das *Wilson Intervall*

$$\frac{1}{1+z^2/n}\left[\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}\right] \tag{14}$$

wobei zur kompakteren Notation $z = z_{1-\alpha/2}$ abgekürzt ist. Dieses Intervall haben Brown et al. in der vergleichenden Studie [9] aufgrund seiner Coverage Probability empfohlen. Im asymptotischen Fall großer n ergibt sich aus Glg. (14) das in einführenden Statistikbüchern gelehrte klassische *Wald Intervall*:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \tag{15}$$

3.2 Likelihood Ratio für $\hat{p}$

Wenn bei n Beobachtungen das interessierende Ereignis k-mal auftritt und $(n-k)$-mal nicht, dann ist die Likelihood-Funktion

$$L(p) = p^k(1-p)^{n-k} \tag{16}$$

Die relative Häufigkeit $\hat{p} = k/n$ ist der ML-Schätzer für p, so dass das Likelihood-Ratio Supportintervall der p-Bereich ist, für den gilt

$$\frac{L(p)}{L(\hat{p})} = \frac{p^k(1-p)^{n-k}}{\hat{p}^k(1-\hat{p})^{n-k}} \geq \frac{1}{K} \tag{17}$$

Der Verlauf der Funktion auf der linken Seite ist in Abb. 3 dargestellt. Glg. (17) muss numerisch gelöst werden, z.B. mit der R-Funktion *uniroot*. Eine mögliche Implementierung ist in Listing 2 angegeben.

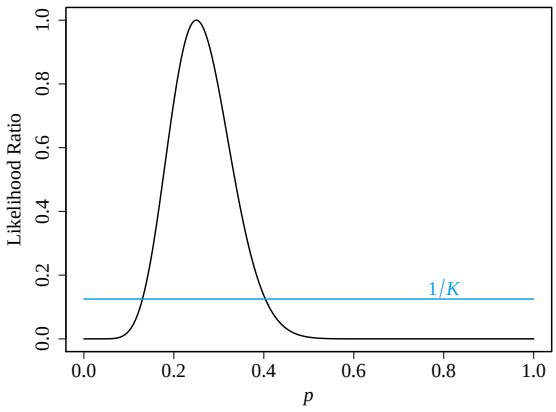


Abbildung 3: Likelihood-Ratio $L(p)/L(\hat{p})$ der Binomialverteilung für $n = 40$ und $k = 10$.

```
lr.binom <- function(n, k, K) {
  helper <- function(p, n, k, K) {
    return (p**k * (1-p)**(n-k) /
            ((k/n)**k * (1-k/n)**(n-k))
            - 1/K)
  }
  p1 <- rep(0,length(k))
  p2 <- p1
  if (k==0) {
    p1 <- 0
  } else {
    r <- uniroot(helper,n=n,k=k,K=K,
                 interval=c(0,k/n))
    p1 <- r$root
  }
  if (k==n) {
    p2 <- 1
  } else {
    r <- uniroot(helper,n=n,k=k,K=K,
                 interval=c(k/n,1))
    p2 <- r$root
  }
  return (data.frame(p1=p1, p2=p2))
}
```

Listing 2: R Implementierung des Likelihood-Ratio Supportintervalls für die relative Häufigkeit gemäß Glg. (17).

3.3 Highest Posterior Density für $\hat{p}$

Das HPD-Intervall kann in R mit der Funktion *hdi* aus dem R-Paket *HDInterval* berechnet werden. *hdi* benötigt als Übergabeparameter eine Funktion, die die Inverse von $\int_{-\infty}^{\theta} p(\tau|\hat{\theta})\, d\tau$ berechnet, so dass die Anwendbarkeit auf Fälle beschränkt ist, in denen man diese berechnen kann. Im Fall der Binomialverteilung geht das.

```
library(HDInterval)
ci <- hdi(qbeta, 1-alpha,
          shape1=(k+1),
          shape2=(n-k+1))
p1 <- ci[1]; p2 <- ci[2]
```

Listing 3: R-Code zur Berechnung des $(1-\alpha)$ HPD-Intervalls für die relative Häufigkeit.

Wenn man die Binomialverteilung (10) in Glg. (7) einsetzt und die "a priori" Wahrscheinlichkeit $p(\theta)$ konstant annimmt, ergibt sich

$$\begin{aligned} p(p|k) &= \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\int_0^1 \binom{n}{k} q^k (1-q)^{n-k}\, dq} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1} \\ &= \text{dbeta}(p,a,b) \end{aligned} \tag{18}$$

wobei $a = k+1$ und $b = n-k+1$ und *dbeta* die R-Funktion für die Wahrscheinlichkeitsdichte der Beta-Verteilung ist. Deren inverse Verteilungsfunktion ist in R als Funktion *qbeta* verfügbar, so dass das HPD-Intervall für die relative Häufigkeit $\hat{p}$ mit dem Code von Listing 3 berechnet werden kann.

4 Statistische Mittelwerte

Ein weiterer in der Praxis sehr häufig vorkommender Schätzer ist der statistische Mittelwert $\overline{x}$ als Schätzer für den Erwartungswert $\mu = E(X)$. Für den statistischen Mittelwert $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ kann man eine Größe konstruieren, die nur vom unbekannten μ abhängt und deren Verteilung bekannt ist, allerdings nur für den Fall, dass X normalverteilt ist. In diesem Fall ist

$$Z = \frac{\overline{x}-\mu}{\sqrt{s^2/n}} \quad \text{mit } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \tag{19}$$

t-verteilt mit $(n-1)$ Freiheitsgraden[5]. Wenn X nicht normalverteilt ist, weiß man aber durch den zentralen Grenzwertsatz, dass die Größe (19) immerhin näherungsweise standard-normalverteilt[6] ist [1]. Weil man in der Regel nicht weiß, ob X normalverteilt ist, können Konfidenzintervalle für den Mittelwert wahlweise auf der t-Verteilung oder der Normalverteilung basieren.

4.1 Frequentistisches Intervall für μ

Wenn μ_0 der gemessene Wert für $\overline{x}$ ist, dann wird die Bestimmungsglg. (5a) für μ_l bei Verwendung der t-Verteilung zu

$$\begin{aligned} & P_{\mu=\mu_l}(\overline{x} \geq \mu_0) = \alpha/2 \\ \Leftrightarrow\quad & P\left(Z \geq (\mu_0-\mu_l)/\sqrt{s^2/n}\right) = \alpha/2 \\ \Leftrightarrow\quad & 1 - \text{pt}\left((\mu_0-\mu_l)/\sqrt{s^2/n}, n-1\right) = \alpha/2 \\ \Leftrightarrow\quad & (\mu_0-\mu_l)/\sqrt{s^2/n} = \text{qt}(1-\alpha/2, n-1) \\ \Leftrightarrow\quad & \mu_l = \mu_0 - \text{qt}(1-\alpha/2, n-1)\cdot\sqrt{s^2/n} \end{aligned} \tag{20}$$

Dabei ist *pt* die Verteilungsfunktion der t-Verteilung, und *qt* deren Inverse. Ebenso lässt sich Glg. (5b) nach μ_u auflösen, wenn man die Symmetrieeigenschaft $\text{qt}(t) = -\text{qt}(1-t)$ berücksichtigt. Zusammenfassend ergibt sich das auf der t-Verteilung basierende Konfidenzintervall

$$\overline{x} \pm t_{1-\alpha/2}(n-1)\cdot\sqrt{s^2/n} \tag{21}$$

wobei $t_{1-\alpha/2}(n-1)$ das $(1-\alpha/2)$-Quantil der t-Verteilung ist, das sich in R berechnen lässt mit der Funktion *qt*.

Unter Zugrundelegung der Normalverteilung ergibt sich mit demselben Rechenweg das Konfidenzintervall

$$\overline{x} \pm z_{1-\alpha/2}\cdot\sqrt{s^2/n} \tag{22}$$

wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung ist, das sich in R berechnen lässt mit der Funktion *qnorm*.

Dass sich je nach zugrundegelegter Verteilung die zwei verschiedenen Konfidenzintervalle (21) oder (22) ergeben, ist kein Widerspruch. Zwar ist

$$t_{1-\alpha/2}(n-1) > z_{1-\alpha/2} \quad \text{für alle } n \tag{23}$$

und damit des Konfidenzintervall (21) immer etwas größer, aber asymptotisch für große n werden beide Konfidenzintervalle identisch wegen

$$\lim_{n\to\infty} t_{1-\alpha/2}(n-1) = z_{1-\alpha/2} \tag{24}$$

Für $\alpha = 0.05$ liegen beide Werte etwa bei zwei, so dass die obigen Konfidenzintervalle der Faustregel "zwei mal sigma" entsprechen mit $\sigma = \sqrt{s^2/n}$.

[5] Der esoterisch anmutende Begriff "Freiheitsgrade" bezeichnet lediglich den Parameter der t-Verteilung.

[6] Mit "Standard"-Normalverteilung bezeichnet man die Normalverteilung zu den Parametern $\mu = 0$ und $\sigma^2 = 1$.

4.2 Likelihood Ratio für μ

Bei Zugrundelegen der t-Verteilung ist die Bestimmungsgleichung (6) für das Likelihood-Ratio Supportintervall

$$\frac{L(\mu)}{L(\hat{\mu})} = \left(1 + \frac{n(\overline{x}-\mu)^2}{s^2(n-1)}\right)^{-n/2} \geq \frac{1}{K} \qquad (25)$$

Diese Gleichung lässt sich direkt nach μ auflösen, was als Supportintervall ergibt

$$\overline{x} \pm \sqrt{(K^{2/n}-1)s^2\frac{n-1}{n}} \qquad (26)$$

Bei Zugrundelegen der Normalverteilung ist die Bestimmungsgleichung

$$\frac{L(\mu)}{L(\hat{\mu})} = \exp\left(-\frac{n(\overline{x}-\mu)^2}{2s^2}\right) \geq \frac{1}{K} \qquad (27)$$

was sich ebenfalls elementar nach μ auflösen lässt und das Supportintervall ergibt:

$$\overline{x} \pm \sqrt{\frac{2s^2}{n}\ln K} \qquad (28)$$

Es scheint, als lieferten (26) und (28) völlig verschiedene Intervalle, aber dem ist nicht so: asymptotisch für große n sind beide Intervalle gleich wegen[7]

$$\ln x = \lim_{n\to\infty} n(x^{1/n}-1) \qquad (29)$$

Für sehr große n ist die rechte Seite von (26) jedoch durch Auslöschung der führenden Stelle ähnlicher Gleitkommazahlen bei Berechnung mit dem Computer ungenau, weshalb für große n Formel (28) auch im Falle der t-Verteilung zu bevorzugen ist.

Durch Vergleich des Supportintervalls (28) mit dem Konfidenzintervall (21) erkennt man auch den Grund für die Wahl $K = 8$. Es ist nämlich $\sqrt{2\ln 8} \approx 2.0393$, so dass das frequentistische Konfidenzintervall für $\alpha = 0.05$ und das Likelihood-Ratio Supportintervall für $K = 8$ etwa identisch sind. Für $K = 7$ ergibt sich sogar in guter Näherung $\sqrt{2\ln K} \approx z_{1-\alpha/2}$. Die Ergebnisse in Abschnitt 7 zeigen aber, dass das auf $z_{1-\alpha/2}$ basierende frequentistische Intervall typischerweise zu klein ist, so dass $K = 8$ eine sicherere Wahl ist.

[7] Dieser Grenzwert ergibt sich aus der Invertierung von [7] Glg. 4.2.21.

4.3 Highest Posterior Density für μ

Unter Zugrundelegung der Dichte der t-Verteilung und der Annahme der "a priori" Verteilung $p(\mu) = \textit{const.}$ wird Glg. (7):

$$p(\mu|\overline{x}) = \frac{\sqrt{n}\Gamma(\frac{n}{2})}{s\sqrt{\pi(n-1)}\Gamma(\frac{n-1}{2})}\left(1+\frac{(\overline{x}-\mu)^2 n}{s^2(n-1)}\right)^{-\frac{n}{2}}$$
$$= \sqrt{\frac{n}{s^2}}\cdot \mathrm{dt}\left(\frac{(\overline{x}-\mu)\sqrt{n}}{s}, n-1\right) \qquad (30)$$

wobei *dt* die R-Funktion für die Dichte der t-Verteilung ist. Unter Zugrundelegung der Normalverteilung ergibt sich (ebenfalls mit $p(\mu) = \textit{const.}$):

$$p(\mu|\overline{x}) = \sqrt{\frac{n}{2\pi s^2}}\cdot\exp\left(-\frac{(\overline{x}-\mu)^2 n}{2s^2}\right)$$
$$= \mathrm{dnorm}(\mu, \overline{x}, s^2/n) \qquad (31)$$

wobei *dnorm* die R-Funktion für die Dichte der Normalverteilung ist. In beiden Fällen ergeben sich dieselben Dichtefunktionen wie die im frequentistischen Eingehenden, die auch noch symmetrisch sind, so dass die Bestimmungsgleichung für das HPD-Intervall identisch zur Bestimmungsgleichung (20) für das frequentistische Intervall ist. Das HPD-Intervall für den statistischen Mittelwert ist also exakt identisch zum frequentistischen Konfidenzintervall (21) bzw. (21).

Diese Übereinstimmung ist kein Zufall, sondern liegt daran, dass μ ein "Lageparameter" ist, d.h. $p(\overline{x}|\mu) = f(\overline{x}-\mu)$. Wenn dieser funktionale Zusammenhang besteht, dann fallen frequentistisches Konfidenzintervall und HPD-Intervall immer zusammen [10].

5 Maximum-Likelihood Schätzer

Um ein Konfidenzintervall für andere Schätzer bestimmen zu können, muss man wissen, was die Wahrscheinlichkeitsverteilung der Schätzgröße $\hat{\theta}$ ist, was aber für andere als die in beiden vorhergehenden Abschnitten betrachteten Schätzer fast immer kompliziert bis unmöglich ist. Erfreulicherweise gibt es aber für eine große Klasse von Schätzern eine sehr allgemeine Aussage zu deren asymptotischer Verteilung: Maximum-Likelihood Schätzer sind bei "regulären" Log-Likelihoodfunktionen[8] $\ell(\theta)$ (siehe Glg. (2b)) für

[8] Die genauen Anforderungen sind, dass die Log-Likelihoodfunktion $\ell(\theta)$ dreimal stetig differenzierbar ist,

große n näherungsweise normalverteilt um den wahren Wert θ, d.h. die Wahrscheinlichkeitsdichte von $\hat{\theta}$ ist

$$p(\hat{\theta}) = \frac{\exp\left(-\frac{1}{2}\langle\hat{\theta}-\theta, \Sigma^{-1}(\hat{\theta}-\theta)\rangle\right)}{\sqrt{(2\pi)^t \det(\Sigma)}} \qquad (32)$$

wobei t die Anzahl der Parameter $\theta = (\theta_1, \ldots, \theta_t)$ ist, Σ die Kovarianzmatrix bezeichnet, und der Exponent "-1" für Matrixinversion steht.

Wenn man also die Kovarianzmatrix $(\sigma_{ij}) = \Sigma$ bestimmen kann, dann kann man über deren Diagonalelemente $\sigma_{ii} = \mathit{Var}(\theta_i)$ die auf der Normalverteilung basierenden Konfidenzintervalle aus Abschnitt 4 verwenden, also

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\sigma_{ii}} \qquad (33)$$

Alternativ würde es auch reichen, einen direkten Schätzer für die Varianz σ_{ii} der Parameter zu haben. Das führt zu zwei möglichen Ansätzen zum Schätzen der Varianz von Maximum-Likelihood Schätzern:

- Schätzen der Kovarianzmatrix durch Invertieren der Hesse-Matrix der Log-Likelihoodfunktion
- Jackknife-Schätzer der Varianz

Die erste Methode hat den Vorteil, geschlossene Formeln für die Varianz zu liefern, sofern man die Hesse-Matrix analytisch berechnen kann. Die zweite Methode hat den Vorteil, algorithmisch elementar zu sein und völlig ohne analytische oder numerische Berechnung von Ableitungen auszukommen.

Sollten die Voraussetzungen aus Fußnote 8 verletzt sein, kann man die Hesse-Matrix gar nicht berechnen, so dass die erste Methode ausscheidet. Zwar könnte man dann immer noch die Jackknife-Methode anwenden, aber es ist in diesem Fall weder garantiert, dass der Schätzer normalverteilt ist, noch dass die Jackknife-Varianz überhaupt ein guter Schätzer für die Varianz ist (siehe [11] für ein Gegenbeispiel). In diesem Fall sollte auf die in Abschnitt 6 beschriebene Bootstrap-Methode zurückgegriffen werden.

5.1 Hesse-Matrix

Unter den Voraussetzungen von Fußnote 8 kann man die Kovarianzmatrix in Glg. (32) schätzen durch [4]

$$(\sigma_{ij}) = \left(-\left.\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right|_{\theta=\hat{\theta}}\right)^{-1} \qquad (34)$$

dass die Erwartungswerte der ersten beiden Ableitungen existieren und dass die dritte Ableitung nach oben durch eine Funktion mit endlichem Erwartungswert beschränkt ist [4].

```
lnL <- function(theta1, theta2, ...) {
  # Definition der negativen (!)
  # Log-Likelihoodfunktion
  ...
}

# Startwerte der Optimierung
theta0 <- c(start1, start2, ...)

# Optimierung
p <- optim(theta0, lnL, hessian=TRUE)
if (p$convergence == 0) {
  theta <- p$par
  covmat <- solve(p$hessian)
  sigma <- sqrt(diag(covmat))
}
```

Listing 4: R Implementierung zur Bestimmung eines ML-Schätzers für $\theta = (\theta_1, \ldots, \theta_t)$ zusammen mit der Varianz der geschätzten Komponenten. Die Log-Likelihoodfunktion muss negativ definiert werden, weil *optim* minimiert statt maximiert.

wobei $\ell(\theta)$ die Log-Likelihood Funktion aus Glg. (2b) ist und der Exponent "-1" für Matrixinversion steht.

In vielen Fällen kann man weder die Bestimmungsgleichungen (3) für den ML-Schätzer $\hat{\theta}$ geschlossen lösen, noch die Hesse-Matrix und deren Inverse (34) analytisch berechnen. Das bedeutet aber nicht, dass diese Methode damit ausscheidet, denn beides kann ja trotzdem numerisch näherungsweise berechnet werden. Die R-Funktion *optim* bietet sogar mit *hessian=TRUE* die Option, die Hesse-Matrix gleich mit berechnen zu lassen. Eine Beispielimplementierung ist in Listing 4 angegeben.

5.2 Jackknife

Die Idee des Jackknife-Verfahrens besteht darin, den Schätzer $\hat{\theta}(x_1, \ldots, x_n)$ mehrmals zu berechnen, aber mit jeweils einem der Werte x_i weggelassen, und aus der Verteilung dieser Schätzwerte auf die Varianz von $\hat{\theta}$ zu schließen. Wenn $\theta_{(i)}$ der ohne den i-ten Inputwert bestimmte Schätzer ist, dann ist der Jackknife-Schätzer der Varianz von $\hat{\theta}$:

$$\sigma_{JK}(\hat{\theta}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\theta_{(i)} - \theta_{(.)})^2} \qquad (35)$$

$$\text{mit} \quad \theta_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\theta_{(i)}$$

```
theta.hat <- function(x) {
  # Implementierung des Schätzers
  ...
}
theta.jk <- rep(0, n)
for (i in 1:n) {
  theta.jk[i] <- theta.hat(x[-i])
}
theta.dot <- mean(theta.jk)
sigma.jk <- sqrt((n-1) *
        mean((theta.jk-theta.dot)^2))
```

Listing 5: Berechnung der Jackknife-Varianz eines Schätzers $\hat{\theta}(x_1, \ldots, x_n)$ in R.

Wenn θ ein Vektor mit mehreren Komponenten ist, dann kann man auch die komplette Kovarianzmatrix Σ mit dem Jackknife-Verfahren schätzen, aber weil für die Konfidenzintervalle (33) nur deren Diagonalelemente σ_{ii} benötigt werden, reicht es, Glg. (35) auf jede Komponente von θ anzuwenden. Für asymptotisch normalverteilte ML-Schätzer ist σ_{JK} ein asymptotisch erwartungstreuer und konsistenter Schätzer für deren Varianz [12]. Eine Implementierung von Formel (35) ist in Listing 5 angegeben.

6 Bootstrap

Ebenso wie beim Jackknife-Verfahren werden beim Bootstrap-Verfahren neue Datensätze aus den Originaldaten $x_1, \ldots, x_n$ generiert, aber nicht deterministisch durch zyklisches Weglassen, sondern per Zufallsauswahl. Diese Auswahl kann entweder erfolgen durch zufälliges n-maliges Ziehen mit Zurücklegen (*nicht-parametrischer Bootstrap*) oder durch n-malige Erzeugung von Zufallszahlen, die gemäß der mit dem ML-Schätzer geschätzten Dichte verteilt sind (*parametrischer Bootstrap*). Beim nicht-parametrischen Bootstrap gehen also die Messwerte alle in die Bootstrap-Simulation ein, beim parametrischen Bootstrap dagegen nur summarisch in Form des aus ihnen berechneten Schätzers $\hat{\theta}$.

Wenn man dieses Ziehen neuer Datensätze R-mal wiederholt, erhält man eine Monte-Carlo Simulation der Verteilung des Schätzers $\hat{\theta}$, aus der Konfidenzintervalle geschätzt werden können[9,10]. Dazu gibt es verschiedene Berechnungswege, die im Überblick zusammen mit ihrer asymptotischen Coverage Probability in [15] zusammengestellt sind und deren Hintergründe in [6] erläutert sind. Die wichtigsten sind:

Percentile & Basic. Die ursprünglich von Efron vorgeschlagene Methode (*Percentile Bootstrap*) nimmt einfach die Perzentile der simulierten Verteilung $\hat{\theta}_1, \ldots, \hat{\theta}_n$ von $\hat{\theta}$. Das *Basic Bootstrap Intervall* spiegelt dieses Intervall an $\hat{\theta}$. Venables & Ripley empfehlen den Basic Bootstrap gegenüber dem Percentile Bootstrap [16], die Untersuchungen in Abschnitt 7 legen aber genau die umgekehrte Empfehlung nahe.

Bias corrected accelerated (BC_a). Bei dieser Methode wird versucht, Transformationsparameter zu schätzen, die die Verteilung symmetrisch machen. Dies ist die von Efron empfohlene Methode.

Für das BC_a Intervall kann man zeigen, dass die Coverage Probability für große n mit $o(n^{-1})$ gegen den nominellen Wert $1 - \alpha$ konvergiert [6], was eine raschere Konvergenz ist als beim klassischen $z_{1-\alpha/2}\sigma$ Intervall, das nur eine Konvergenzrate von $o(n^{-1/2})$ hat. DiCiccio & Efron schlossen daraus, dass die Bootstrap-Methode grundsätzlich vorzuziehen sei (Kommentar zu [6], S. 228):

> "If the standard intervals were invented today, they might not be publishable."

Dies ist aber etwas irreführend, weil für große n die verschiedenen Konfidenzintervallen sowieso ähnlich sind, so dass die Konvergenzrate für große n eher von theoretischem als von praktischem Interesse ist. Praxisrelevanter ist das Verhalten bei kleinen n und dazu bemerkte der Erfinder der Bootstrap-Methode, Bradley Efron, als Reaktion auf eine Studie, in der das Bootstrap-Intervall schlecht abschnitt [17]:

> "Bootstrap methods are intended to supplement rather than replace parametric analysis, particularly when parametric methods can't be used because of modeling uncertainties or theoretical intractability."

[9] Man kann daraus auch die Varianz schätzen [13], aber das damit berechnete Konfidenzintervall würde wieder eine Normalverteilung von $\hat{\theta}$ voraussetzen.

[10] Man könnte auf die Idee kommen, statt dessen auch die Verteilung der n Jackknife "delete one" Schätzer $\theta_{(i)}$ zu nehmen, aber die ist selbst für reguläre ML-Schätzer nicht asymptotisch normal, also nicht repräsentativ für die Verteilung von $\hat{\theta}$ [14].

```
# Schätzfunktion
# über indices wählt boot() Werte aus
schaetzer <- function(x, indices) {
  x.auswahl <- x[indices]
  ... # berechne Schätzer aus x.auswahl
  return(theta.hat)
}

#bootstrap Konfidenzintervalle
boot.out <- boot(data=x,
        statistic=schaetzer, R=1000)
ci <- boot.ci(boot.out,
        conf=0.95, type="all")

# percentile interval:
theta1 <- ci$perc[4]
theta2 <- ci$perc[5]

# basic interval:
theta1 <- ci$basic[4]
theta2 <- ci$basic[5]

# BCa interval:
theta1 <- ci$bca[4]
theta2 <- ci$bca[5]
```

Listing 6: Berechnung von Bootstrap Konfidenzintervallen mit der R-Library *boot*.

In der R-Library *boot* kann die Funktion *boot.ci* verschiedene Bootstrap Konfidenzintervalle berechnen, darunter auch die drei oben angegebenen. Als minimalen Wert R für Bootstrap-Konfidenzintervalle geben Efron & Tibshirani $R = 1000$ an [18]. Die Benutzung zeigt Listing 6.

Neben der Frage, welches der vielen Bootstrap-Konfidenzintervalle man denn nun im Einzelfall nehmen soll, hat die Bootstrap-Methode noch einen weiteren Nachteil: weil sie auf einer Monte-Carlo Simulation beruht, sind die Ergebnisse nicht deterministisch reproduzierbar, d.h. wenn verschiedene Personen dieselben Daten auswerten, erhalten sie immer leicht unterschiedliche Ergebnisse.

7 Performance in Beispielen

Für eine vergleichende Evaluation der verschiedenen Konfidenzintervalle werden in diesem Abschnitt alle drei behandelten Fälle an Beispielen untersucht. Neben dem Verhalten der Coverage Probability P_{cov} ist auch die relative Größe der Konfidenzintervalle von Interesse.

Im Fall der relativen Häufigkeit gibt es bei festem n nur endlich viele mögliche Werte, so dass sich $P_{cov}(p)$ exakt ausrechnen lässt. Für den Mittelwert wird als Beispiel eine unsymmetrische Verteilung der Dichte $f(x) = 3x^2$ genommen, um mittels Monte-Carlo Simulationen zu prüfen, inwiefern die Bootstrap-Methode in solch einem Fall Verbesserungen bringt. Als Beispiel für einen ML-Schätzer wird der Parameter λ der Exponentialverteilung genommen, weil in diesem Fall sich auch die Inverse der Hesse-Matrix analytisch geschlossen berechnen lässt, so dass in einer Monte-Carlo Simulation alle Verfahren verglichen werden können. Beim Bootstrap wird nur die nicht-parametrische Methode angewandt, weil die parametrische Methode nicht universell einsetzbar ist, sondern immer speziell auf ein konkretes Problem zugeschnitten werden muss, was in der Regel nicht trivial ist[11].

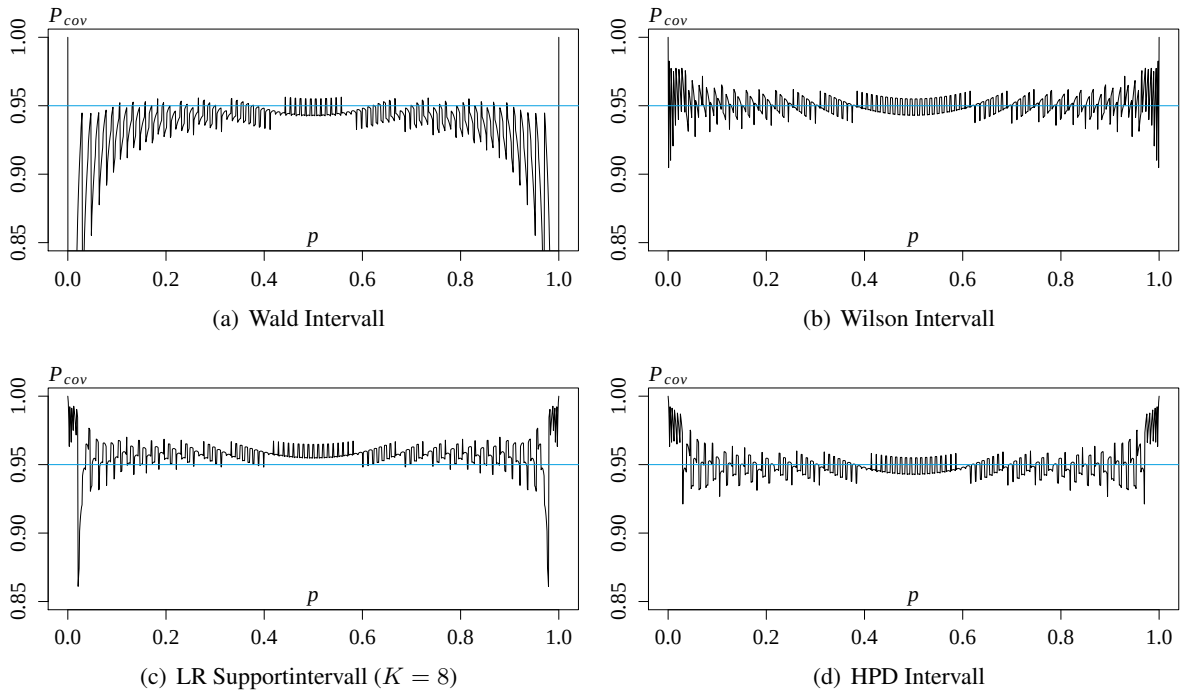
7.1 P_{cov} für relative Häufigkeit

Die Coverage Probability verschiedener Konfidenzintervalle für die relative Häufigkeit wurde bereits von Brown et al. untersucht, die aufgrund dessen das Wilson-Intervall empfahlen [9]. Sie haben allerdings nicht das LR-Supportintervall und das HPD-Intervall untersucht, weshalb die entsprechenden Verläufe für $P_{cov}(p)$ in Abb. 4 zusammenfassend dargestellt sind. Der Verlauf für das "exakte" (Clopper-Pearson) Intervall kann Abb. 1 entnommen werden. Die Kurven sind wie folgt berechnet:

- für jedes $0 \leq k \leq n$ wurde das Konfidenzintervall bestimmt
- für jeden gesampelten Wert $p \in [0, 1]$ wurden die Wahrscheinlichkeiten der Werte k aufaddiert, für die p im Konfidenzintervall liegt

Wie bereits Brown et al. bemerkten, hat das in Statistik-Lehrbüchern gelehrte klassische Wald-Intervall fast durchgängig eine zu niedrige Coverage Probability, die für kleine und große p sogar gegen Null geht, während das Wilson-Intervall etwa um den nominellen Wert schwankt, auch wenn an den Rändern größere Abweichungen nach unten auftreten. Interessanterweise hat das HPD-Intervall sogar eine noch

[11] Neben Kenntnis der Wahrscheinlichkeitstheorie des vorliegenden Problems erfordert es auch vertiefte numerische Kenntnisse über die Erzeugung von Zufallszahlen zu einer vorgegebenen Verteilung (Transformationsmethode, Rejection-Methode [19]).



(a) Wald Intervall (b) Wilson Intervall

(c) LR Supportintervall ($K = 8$) (d) HPD Intervall

Abbildung 4: Coverage Probability $P_{cov}(p)$ der Konfidenzintervalle für die relative Häufigkeit als Funktion des wahren Parameterwerts p für $n = 100$ und $1 - \alpha = 0.95$.

bessere Coverage Probability als das Wilson-Intervall, weil der Verlauf sehr ähnlich ist, aber an den Rändern keine zu niedrigen Werte auftreten. Das LR Supportintervall hat für $K = 8$ einen ähnlichen Verlauf wie das exakte Clopper-Pearson Intervall, hat aber Stellen, an denen P_{cov} deutlich unter den nominellen Wert fällt.

Ein anderes Bewertungskriterium ist die Länge der Intervalle, die bei vergleichbarer Coverage Probability möglichst klein sein sollte. Deren Maximalwert für verschiedene n ist in Abb. 5 dargestellt. Das exakte Intervall ist am größten, was der Preis dafür ist, dass es $P_{cov}(p) \geq 1 - \alpha$ garantiert und den nominellen Wert oft auch deutlich übertrifft. Kurioserweise ist aber die maximale Länge des Wald-Intervalls größer als die des Wilson- oder HPD-Intervalls. Dieser scheinbare Widerspruch erklärt sich, wenn man die Intervalllängen für verschiedene $\hat{p}$ bei festem n vergleicht (Abb. 6).

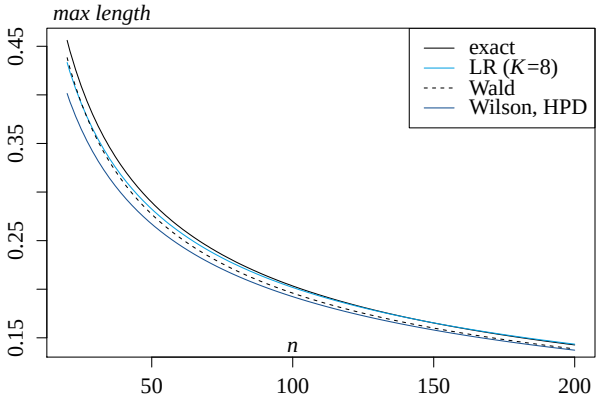


Abbildung 5: Maximale Länge der Konfidenzintervalle für die relative Häufigkeit als Funktion von n für $1 - \alpha = 0.95$. Die Werte für das HPD und Wilson-Intervall sind sehr ähnlich.

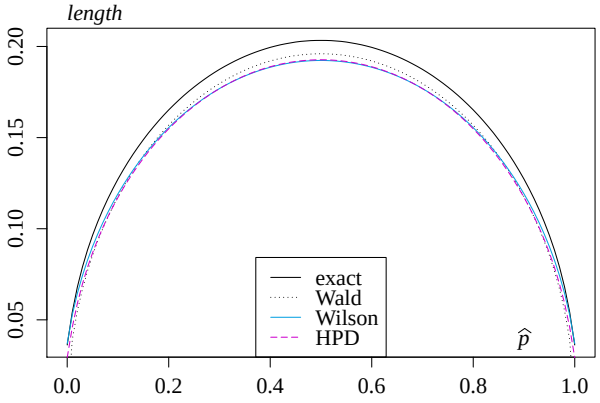


Abbildung 6: Länge der Konfidenzintervalle für die relative Häufigkeit als Funktion von $\hat{p}$ für $1 - \alpha = 0.95$ und $n = 100$.

Hier ist das klassische Wald-Intervall unnötig groß für $\hat{p} \approx 0.5$, aber zu klein für $\hat{p} \approx 0$ oder $\hat{p} \approx 1$.

Interessant ist, dass das HPD-Intervall für $\hat{p} \approx 0$ oder $\hat{p} \approx 1$ kürzer ist als das Wilson-Intervall, obwohl es eine höhere Coverage Probability in diesen Bereichen hat. Bezüglich der Kriterien Coverage Probability und Länge hat das HPD-Intervall also die besten Eigenschaften. Allerdings ist es nur numerisch berechenbar (Listing 3). Wenn eine geschlossene Formel benötigt wird, dann kann alternativ das Wilson-Intervall genommen werden (Glg. (14)), sofern $\hat{p}$ nicht zu nahe an Null oder Eins liegt.

7.2 P_{cov} für Mittelwert

Der Vergleich der klassischen Intervalle für den statistischen Mittelwert mit den Bootstrap-Intervallen erfolgt auf einer Zufallsvariablen, die verteilt ist zur Dichte

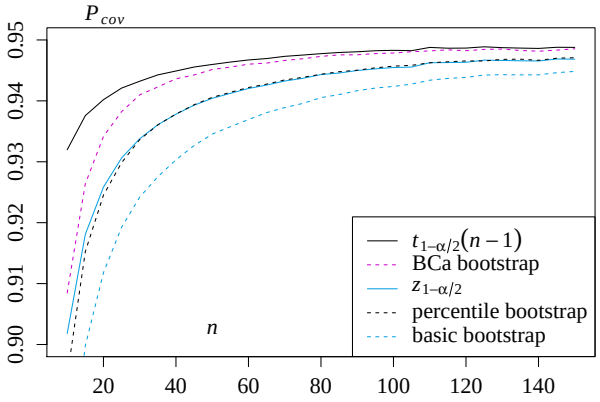
$$f(x) = \begin{cases} 3x^2 & \text{für } 0 \le x \le 1 \\ 0 & \text{sonst} \end{cases} \tag{36}$$

Der Erwartungswert dieser Verteilung ist $3/4$ und gemäß dieser Dichte verteilte Zufallszahlen lassen sich mit der Transformationsmethode [19] erzeugen durch

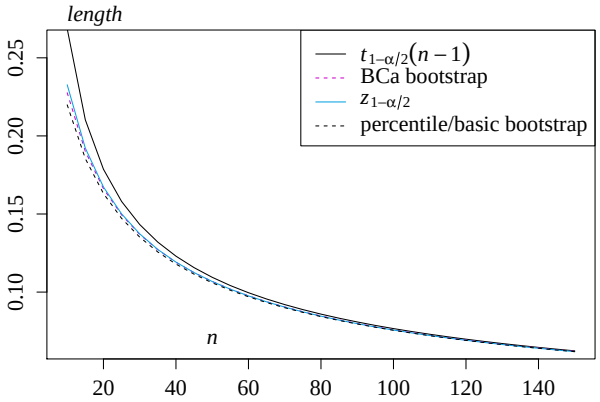
```
runif(N, min=0, max=1)**(1/3)
```

Für die Anzahl der simulierten Mittelwertmessungen habe ich $N = 10^6$ gewählt, so dass die Coverage Probability für $\alpha = 0.05$ mit einer Genauigkeit von ± 0.0004 geschätzt werden kann.

Der Verlauf von P_{cov} und Länge der verschiedenen Konfidenzintervalle als Funktion der Anzahl n der Messwerte ist in Abb. 7 dargestellt. Interessanterweise hat das klassische auf der t-Verteilung basierende Konfidenzintervall durchgängig die höchste Coverage Probability, obwohl die Verteilung von $\overline{x}$ für kleine n unsymmetrisch ist. Die Schwächen der Bootstrap Methode bei kleinen n werden also nicht durch die Berücksichtigung unsymmetrischer Verteilungen kompensiert. Das beste Bootstrap-Intervall ist in diesem Fall das BC_a-Intervall, das eine dem klassischen $z_{1-\alpha/2}$-Intervall vergleichbare Länge, aber ein höheres P_{cov} hat. Die Empfehlung von Venables & Ripley für das Basic gegenüber dem Percentile Bootstrap-Intervall kann nicht bestätigt werden, sondern, im Gegenteil, das Basic Intervall hat in diesem Beispiel ein deutlich zu niedriges P_{cov}, während das Percentile Intervall dem klassischen $z_{1-\alpha/2}$-Intervall vergleichbar ist.



(a) Coverage Probability



(b) Mittlere Länge

Abbildung 7: Coverage Probability und mittlere Länge der verschiedenen Konfidenzintervalle für den statistischen Mittelwert von n gemäß (36) verteilten Zufallsgrößen.

7.3 P_{cov} für ML-Schätzer

Um an einem Beispiel zu untersuchen, wie die verschiedenen Konfidenzintervalle sich bei Maximum-Likelihood Schätzern verhalten, sei die Exponentialverteilung betrachtet. Diese hat den Parameter λ und die Dichte

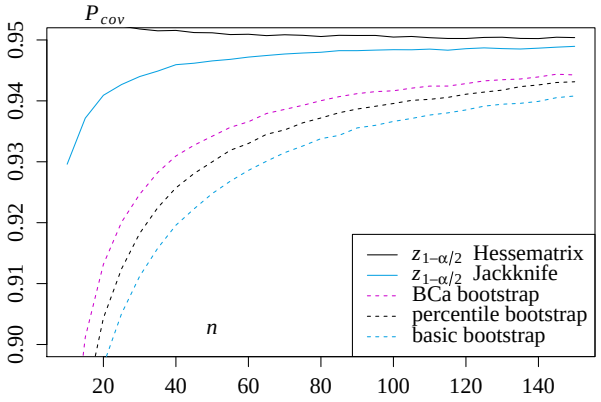
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für } x \ge 0 \\ 0 & \text{sonst} \end{cases} \tag{37}$$

Die Log-Likelihoodfunktion ist folglich

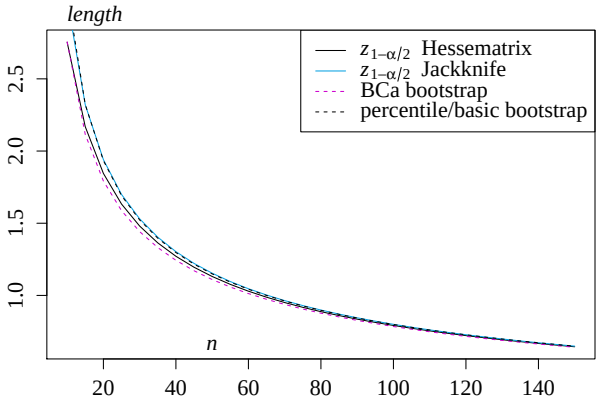
$$\ell(\lambda) = n \log(\lambda) + \lambda \sum_{i=1}^{n} x_i \tag{38}$$

Der ML-Schätzer für λ ergibt sich aus der Bestimmungsgleichung (3) zu

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}} \tag{39}$$



(a) Coverage Probability



(b) Mittlere Länge

Abbildung 8: Coverage Probability und mittlere Länge der verschiedenen Konfidenzintervalle für den ML-Schätzer des Parameters λ der Exponentialverteilung.

Weil die Verteilung nur einen Parameter enthält, ist die Hesse-Matrix lediglich ein 1×1 Matrix (also ein Skalar), die geschlossen berechnet werden kann zu

$$H(\lambda) = \left(\frac{\partial^2}{\partial\lambda^2}\ell\right) = \left(-\frac{n}{\lambda^2}\right) \tag{40}$$

Damit ergibt sich die aus der Hesse-Matrix geschätzte Streuung von $\hat{\lambda}$ zu

$$\hat{\sigma}_{HM} = \sqrt{\Big(-H(\hat{\lambda})\Big)^{-1}} = \frac{\hat{\lambda}}{\sqrt{n}} \tag{41}$$

Wieder wurden $N = 10^6$ mal n exponentialverteilte Zufallszahlen generiert mit $\lambda = 2$, um die Verteilung von $\hat{\lambda}$ zu simulieren und P_{cov} und Länge der verschiedenen Konfidenzintervalle zu vergleichen. Die Ergebnisse sind in Abb. 8 dargestellt. Die Coverage Probability des klassischen Intervalls mit $\hat{\sigma}_{HM}$ ist am besten, gefolgt vom klassischen Intervall mit $\hat{\sigma}_{JK}$. Von den Bootstrap-Intervalle ist wieder das BC_a-Intervall am besten, und wieder ist der Percentile Bootstrap dem Basic Bootstrap vorzuziehen. Die Empfehlung von Venables & Ripley zugunsten des Basic Bootstraps gegenüber dem Percentile Bootstrap muss also verworfen werden. Insgesamt zeigen aber alle Bootstrap-Intervalle eine deutlich zu niedrige Coverage Probability.

Erstaunlich ist, dass das auf der Jackknife-Varianz basierende Konfidenzintervall zwar breiter ist, aber eine geringere Coverage Probability hat als das auf der Hesse-Matrix basierende Intervall. Eine genauere Untersuchung der simulierten Verteilung von $\hat{\lambda}$ zeigt, dass das daran liegt, dass in diesem Fall der ML-Schätzer einen Bias hat und im Mittel zu groß ist[12]. Weil $\hat{\sigma}_{HM}$ gemäß Glg. (41) proportional zu $\hat{\lambda}$ ist, ist das Konfidenzintervall größer, wenn der Schätzer zu groß ist, was in diesem Fall den Bias des Schätzers kompensiert. So war die Korrelation von $|\hat{\lambda} - \lambda|$ mit $\hat{\sigma}_{HM}$ in der Monte-Carlo Simulation für $n = 20$ etwa 0.60, aber mit $\hat{\sigma}_{JK}$ nur etwa 0.40. Das erklärt, warum P_{cov} für das kleinere Intervall trotzdem größer sein kann.

8 Fazit

Für den Praktiker ergeben sich aus der vergleichenden Evaluation der Konfidenzintervalle folgende Empfehlungen:

1) Für die relative Häufigkeit sollte das HPD-Intervall (Listing 3) oder das Wilson-Intervall (Glg. (14)) genommen werden. Das Wilson-Intervall hat den Vorteil einer geschlossenen Formel, hat aber eine kleinere Coverage Probability als das HPD-Intervall für p-Werte nahe an Eins oder Null.

2) Für den statistischen Mittelwert sollte das klassische Konfidenzintervall basierend auf der t-Verteilung genommen werden, also Glg. (21).

3) Für ML-Schätzer mit einer glatten Log-Likelihoodfunktion sollte das Konfidenzintervall $z_{1-\alpha/2} \cdot \hat{\sigma}$ genommen werden, wobei $\hat{\sigma}$ über die Hesse-Matrix oder mittels Jackknife geschätzt werden kann (Listing 5).

4) In anderen Fällen sollte das BC_a Bootstrap-Intervall genommen werden.

[12] ML-Schätzer sind ja nur asymptotisch erwartungstreu, können also durchaus für kleine n verzerrt sein.

Die Ergebnisse dieses technischen Berichts bestätigen damit die oben zitierte Bemerkung von Efron [17]:

> "Bootstrap methods are intended to supplement rather than replace parametric analysis, particularly when parametric methods can't be used because of modeling uncertainties or theoretical intractability."

Literatur

[1] L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz, *Statistik*. Berlin: Springer, 5 ed., 2004.

[2] J. D. Blume, "Likelihood methods for measuring statistical evidence," *Statistics in medicine*, vol. 21, no. 17, pp. 2563–2599, 2002.

[3] N. Turkkan and T. Pham-Gia, "Computation of the highest posterior density interval in Bayesian analysis," *Journal of statistical computation and simulation*, vol. 44, no. 3-4, pp. 243–250, 1993.

[4] W. H. Greene, *Econometric Analysis*. New Jersey: Prentice Hall, 4 ed., 2000.

[5] P. Zsohar, "Short introduction to the generalized method of moments," *Hungarian Statistical Review*, vol. 16, pp. 150–170, 2010.

[6] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statistical science*, pp. 189–228, 1996.

[7] M. Abramowitz, I. Stegun, M. Danos, and J. Rafelski, *Pocketbook of mathematical functions*. Frankfurt: Harri Deutsch, 1984.

[8] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[9] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statistical science*, vol. 16, no. 2, pp. 101–117, 2001.

[10] D. Karlen, "Credibility of confidence intervals," in *Advanced Statistical Techniques in Particle Physics, Proceedings*, (Durham), pp. 53–57, 2002.

[11] R. G. Miller, "A trustworthy jackknife," *The Annals of Mathematical Statistics*, vol. 35, no. 4, pp. 1594–1605, 1964.

[12] J. A. Reeds, "Jackknifing maximum likelihood estimates," *The Annals of Statistics*, vol. 6, no. 4, pp. 727–739, 1978.

[13] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

[14] C.-F. J. Wu, "Jackknife, bootstrap and other resampling methods in regression analysis," *the Annals of Statistics*, vol. 14, no. 4, pp. 1261–1295, 1986.

[15] J. Carpenter and J. Bithell, "Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians," *Statistics in medicine*, vol. 19, no. 9, pp. 1141–1164, 2000.

[16] W. Venables and B. Ripley, *Modern Applied Statistics with S*. New York: Springer, 4 ed., 2002.

[17] B. Efron, "Bootstrap confidence intervals: Good or bad?," Technical Report 116, Stanford University, Division of Biostatistics, March 1987.

[18] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical science*, vol. 1, no. 1, pp. 54–75, 1986.

[19] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical recipes in Pascal: the art of scientific computing*. Cambridge University Press, 1989.