Wahrscheinlichkeitsrechnung und
Statistik für Biologen

7. Frequentistische und Bayessche Intervallschätzer

Dirk Metzler

24. März 2026

Inhaltsverzeichnis

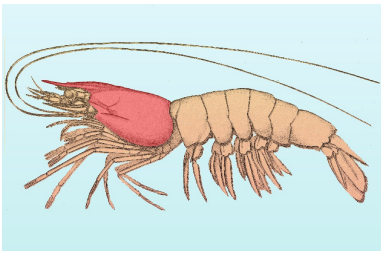
1 Konfidenzintervalle für Erwartungswerte

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 Faustregeln:

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

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

mit Wahrscheinlichkeit nahe bei 2/3

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

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

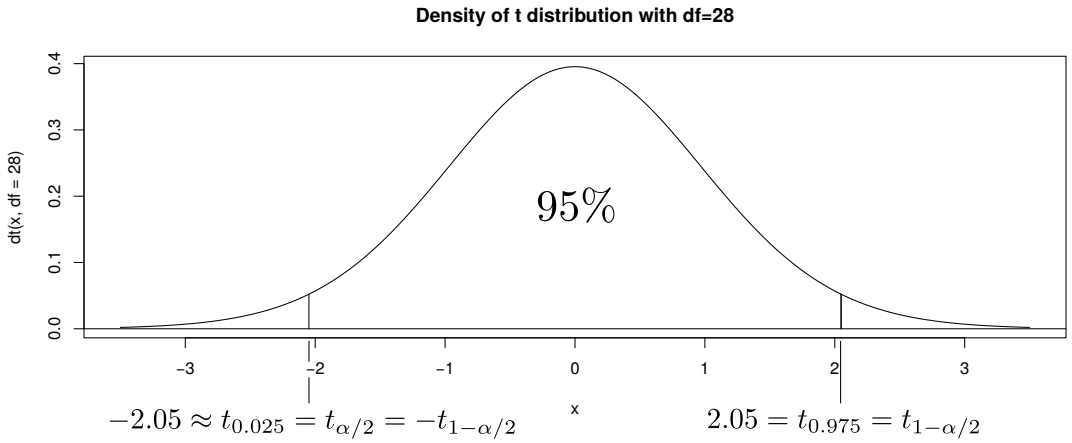
mit Wahrscheinlichkeit nahe bei 95%.

Nun exakt: Sei $t_{97.5\%}$ `<- qt(0.975,length(x)-1)` das 97.5%-Quantil von Student's t-Verteilung mit $n-1$ Freiheitsgraden.

Dann liegt der wahre Mittelwert mit Wahrscheinlichkeit 95% im Intervall

$$\left[\overline{x} - t_{97.5\%} \cdot sem(x),\ \overline{x} + t_{97.5\%} \cdot sem(x)\right]$$

(Beachte: $-t_{97.5\%} = t_{2.5\%}$).



Setzt man die Zahlenwerte $\overline{x} = 3.23$, $t_{97.5\%} = 2.05$ (bei $n-1 = 28$) und $\mathrm{sem}(x) = 0.17$ in

$$\left[\overline{x} - t_{97.5\%} \cdot sem(x),\ \overline{x} + t_{97.5\%} \cdot sem(x)\right]$$

ein, so erhält man das Konfidenzintervall

$$\left[2.88,\ 3.58\right]$$

für den wahren Mittelwert zum Irrtumsniveau 5%.

Das Konfidenzintervall zum Irrtumsniveau 5% nennt man üblicherweise

95%-Konfidenzintervall.

1.2 Erklärung, wieso das Intervall so passt

Konfidenzintervall für den wahren Mittelwert

Ziel: Bestimme das Konfidenzintervall für den wahren Mittelwert zum Irrtumsniveau α, also das $(1-\alpha)$-Konfidenzintervall.

Das Konfidenzintervall für den wahren Mittelwert zum Irrtumsniveau α ist ein aus den Daten $X = (X_1, \ldots, X_n)$ geschätztes (zufälliges) Intervall

$$\left[a(X),\ b(X)\right]$$

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

$$\Pr_\mu\left(\mu \in \left[a(X),\ b(X)\right]\right) \geq 1 - \alpha.$$

Selbstverständlich wollen wir das Konfidenzintervall möglichst klein wählen.

Konfidenzintervall für den wahren Mittelwert

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

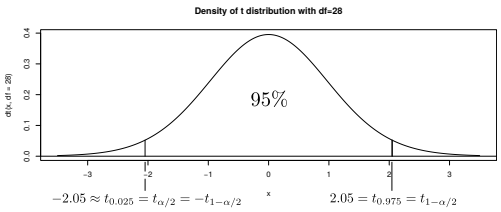
$$t := \frac{\overline{x} - \mu}{\mathrm{sem}(x)}$$

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

Sei $t_{1-\alpha/2}$ `<- qt(`$1-\alpha/2$`,length(x)-1)` das $1-\frac{\alpha}{2}$-Quantils (meistens $1-\frac{\alpha}{2} = 0.975$) der Student-Verteilung mit `length(x)-1` Freiheitsgraden. Dann ist

$$\left[\overline{x} - t_{1-\alpha/2} \cdot \mathrm{sem}(x),\ \overline{x} + t_{1-\alpha/2} \cdot \mathrm{sem}(x)\right]$$

das Konfidenzintervall zum Irrtumsniveau α.



Begründung:

$$\begin{aligned}
&\Pr_\mu\left(\mu \in \left[\overline{x} - t_{1-\alpha/2} \cdot \mathrm{sem}(x),\ \overline{x} + t_{1-\alpha/2} \cdot \mathrm{sem}(x)\right]\right) \\
&= \Pr_\mu\left(\overline{x} - t_{1-\alpha/2} \cdot \mathrm{sem}(x) \le \mu \le \overline{x} + t_{1-\alpha/2} \cdot \mathrm{sem}(x)\right) \\
&= \Pr_\mu\left(-t_{1-\alpha/2} \cdot \mathrm{sem}(x) \le \mu - \overline{x} \le t_{1-\alpha/2} \cdot \mathrm{sem}(x)\right) \\
&= \Pr_\mu\left(-t_{1-\alpha/2} \le \frac{\mu - \overline{x}}{\mathrm{sem}(x)} \le t_{1-\alpha/2}\right) \\
&= \Pr_\mu\left(t_{\alpha/2} \le -t \le t_{1-\alpha/2}\right) \qquad (t \text{ ist die t-Statistik, also t-verteilt mit } n-1 \text{ Freiheitsgraden}) \\
&= 1-\alpha
\end{aligned}$$

Beachte: $t_{\alpha/2}$ wird gerade so gewählt, dass die letzte Gleichung richtig ist.

1.3 Dualität von Tests und Konfidenzintervallen

Die wechselseitige Beziehung zwischen Test und Konfidenzintervall untersuchen wir am Beispiel des folgenden Datensatzes:

```
> X
 [1] 4.111007 5.023229 5.489230 4.456054 4.343212
 [5] 5.431928 3.944405 3.471677 4.337888 5.412292
> n <- length(X)
> m <- mean(X)
> sem <- sd(X)/sqrt(n)
> t <- -qt(0.025,n-1)
> konf <- c(m-t*sem,m+t*sem)
> konf
[1] 4.100824 5.103360
```

$[4.100824, 5.103360]$

```
> t.test(X,mu=4)

One Sample t-test

data:  X
t = 2.7172, df = 9, p-value = 0.02372
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 4.100824 5.103360
```

```
sample estimates:
mean of x
 4.602092
```

Beachte: R gibt beim t-Test auch das Konfidenzintervall an!
$[4.100824, 5.103360]$

```
> t.test(X,mu=4.1)

One Sample t-test

data:  X
t = 2.2659, df = 9, p-value = 0.0497
alternative hypothesis: true mean is not equal to 4.1
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Beachte: R gibt beim t-Test auch das Konfidenzintervall an!
$[4.100824, 5.103360]$

```
> t.test(X,mu=4.1009)

One Sample t-test

data:  X
t = 2.2618, df = 9, p-value = 0.05003
alternative hypothesis: true mean is not equal to 4.1009
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Beachte: R gibt beim t-Test auch das Konfidenzintervall an!
$[4.100824, 5.103360]$

```
> t.test(X,mu=5.1)

One Sample t-test

data:  X
t = -2.247, df = 9, p-value = 0.05125
alternative hypothesis: true mean is not equal to 5.1
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Beachte: R gibt beim t-Test auch das Konfidenzintervall an!
$[4.100824, 5.103360]$

```
> t.test(X,mu=5.1034)
```

```
One Sample t-test

data:  X
t = -2.2623, df = 9, p-value = 0.04999
alternative hypothesis: true mean is not equal to 5.1034
95 percent confidence interval:
 4.100824 5.103360
sample estimates:
mean of x
 4.602092
```

Beachte: R gibt beim t-Test auch das Konfidenzintervall an!

Dualität Tests $\leftrightarrow$ Konfidenzintervalle

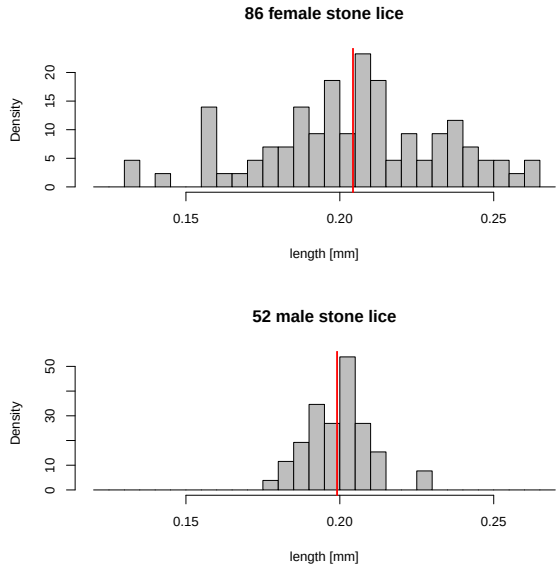
Ist $[a, b]$ ein $(1-\alpha)$-Konfidenzintervall für einen Parameter θ, so erhält man einen Test mit Signifikanzniveau α, wenn man die Nullhypothese $\theta = x$ genau dann verwirft, wenn $x \notin [a, b]$.[0.5cm]

Ist umgekehrt T_x ein Test mit Nullhypothese $\theta = x$ und Signifikanzniveau α, so bilden alle Werte x, für die die Nullhypothese $\theta = x$ *nicht* verworfen wird, ein $(1-\alpha)$-Konfidenzintervall für θ.

Konfidenzintervalle sind auch und gerade dann hilfreich, wenn ein Test *keine* Signifikanz anzeigt.

Beispiel: Gibt es bei Steinläusen geschlechtsspezifische Unterschiede in der Körperlänge?

Datenlage: die Längen von 86 weiblichen (F) und 52 männlichen (M) Steinläusen.




```
> t.test(F,M)

Welch Two Sample t-test

data:  F and M
t = 0.7173, df = 122.625, p-value = 0.4746
alternative hypothesis: true difference in means is
                                    not equal to 0
95 percent confidence interval:
```

```
 -0.004477856  0.009567353
sample estimates:
mean of x mean of y
0.2018155 0.1992707
```

Wie berichten wir über das Ergebis des Tests?

- Es gibt keinen Unterschied zwischen männlichen und weiblichen Steinläusen. ~~Es gibt keinen Unterschied zwischen männlichen und weiblichen Steinläusen.~~
- Männliche und weibliche Steinläuse sind im Mittel gleich lang. ~~Männliche und weibliche Steinläuse sind im Mittel gleich lang.~~
- Die Daten zeigen keine signifikanten Unterschiede zwischen den mittleren Längen männlicher und weiblicher Steinläuse. Die Daten zeigen keine signifikanten Unterschiede zwischen den mittleren Längen männlicher und weiblicher Steinläuse.✓
- Ein 95%-Konfidenzbereich für die Differenz zwischen der mittleren Länge der Weibchen und der Männchen ist [-0.0045,0.0096]. Ein 95%-Konfidenzbereich für die Differenz zwischen der mittleren Länge der Weibchen und der Männchen ist [-0.0045,0.0096].✓

2 Konfidenzintervalle für Wahrscheinlichkeiten

2.1 Beispiel: Porzellankrebs

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 Idee des Wald-Konfidenzintervalls

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

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

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 $\sqrt{p(1-p)/n}$.

Lösung:

Sei $\widehat{p}$ die relative Häufigkeit in der Stichprobe der Länge n. Das 95%-Konfidenzintervall ist

$$\left[\ \widehat{p} - 1.96 \cdot \sqrt{\widehat{p}(1-\widehat{p})/n},\ \widehat{p} + 1.96 \cdot \sqrt{\widehat{p}(1-\widehat{p})/n}\ \right]$$

2.3 Beispiel: Porzellankrebs

Männchenanteil beim Porzellankrebs

Setzt man die Zahlenwerte $n = 53$, $\widehat{p} = 0.566$, und $\sqrt{\widehat{p}(1-\widehat{p})/n} = 0.0681$ in

$$\left[\ \widehat{p} - 1.96 \cdot \sqrt{\widehat{p}(1-\widehat{p})/n},\ \widehat{p} + 1.96 \cdot \sqrt{\widehat{p}(1-\widehat{p})/n}\ \right]$$

ein, so erhält man das Konfidenzintervall

$$\big[0.433, 0.699\big] = 0.566 \pm 0.133$$

für den wahren Männchenanteil zum Irrtumsniveau 5%.

2.4 Beispiel: Stockente

image (c) Andreas Trepte, http://de.wikipedia.org/w/index.php?title=Datei:Mallard_male_female.jpg

Stockente (*Anas platyrhynchos*, 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 $n = 2200$, $\widehat{p} = 0.564$, und $\sqrt{\widehat{p}(1-\widehat{p})/n} = 0.011$ in

$$\left[\ \widehat{p} - 1.96 \cdot \sqrt{\widehat{p}(1-\widehat{p})/n},\ \widehat{p} + 1.96 \cdot \sqrt{\widehat{p}(1-\widehat{p})/n}\ \right]$$

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 Bessere Konfidenzintervalle

Das Konfidenzintervall

$$\left[\widehat{p} - 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\ ,\ \widehat{p} + 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\right]$$

nennt man auch Wald-Konfidenzintervall.

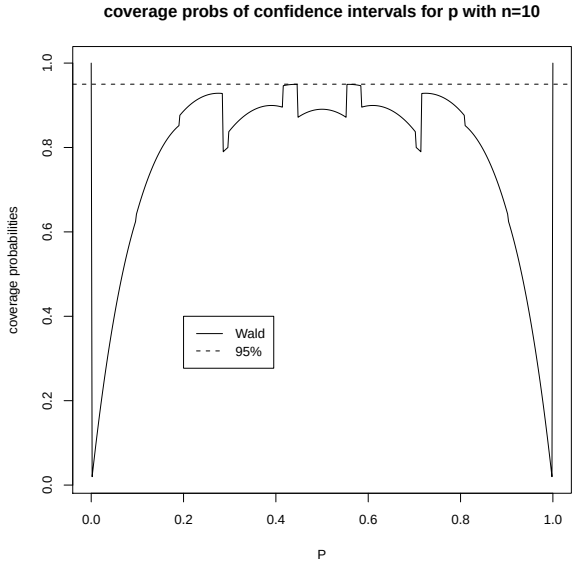
Es sollte gelten: Das Konfidenzintervall überdeckt (d.h. enthält) den wahren Parameterwert mit einer Wahrscheinlichkeit von mindestens 95%.

Diese *Überdeckungswahrscheinlichkeit* kann man berechnen, und das tun wir nun für $n = 10$ mit Werten für p zwischen 0 und 1.

Genauer: Wir zeichnen die Funktion

$$p \mapsto \Pr\left(p \in \left[\widehat{p} - 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\ ,\ \widehat{p} + 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/n}\right]\right)$$

wobei $\widehat{p} = X/n$ und X binomialverteilt ist mit Versuchslänge n und Erfolgswahrscheinlichkeit p .



Wie wir sehen, sacken die Überdeckungswahrscheinlichkeiten ab, wenn das wahre p nahe 0 oder nahe 1 ist.

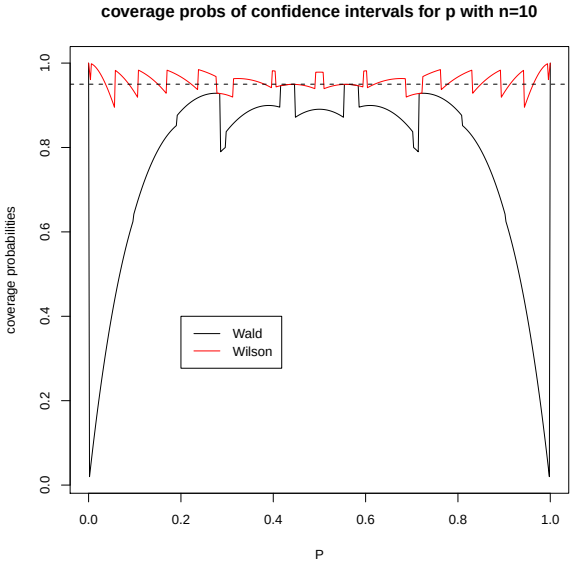
Grund: Angenommen, $p = 0.1$. Dann ist $K = 0$ relativ wahrscheinlich. Wir schätzen dann $\widehat{p} = K/n = 0/n = 0$ und $\widehat{p} \cdot (1 - \widehat{p})/n = 0$. Somit wird das Wald-Konfidenzintervall in etwa $[0, 0]$ sein und folglich das wahre $p = 0.1$ nicht enthalten.

Es gibt noch mehrere weitere Möglichkeiten, Konfidenzintervalle für das p einer binomialverteilten Zufallsvariablen vorzuschlagen. Einige sind in dem R-Befehl `binconf` aus dem Paket `Hmisc` implementiert.

Ein Beispiel ist die Methode von Wilson, die wir hier nicht im Detail ergründen aber mit dem Wald-Konfidenzintervall vergleichen wollen. (Sie wird standard-mäßig vom R-Befehl `binconf` verwendet).
Zur Erinnerung: Konfidenzintervalle sind zufällig, da sie von den Daten abhängen.

Eine ideale Methode zum Erzeugen von 95%-Konfidenzintervallen sollte mit Wahrscheinlichkeit 95% ein Intervall ausgeben, das den wahren Parameterwert überdeckt (d.h. enthält).

Diese *Überdeckungswahrscheinlichkeit* kann man berechnen, und das tun wir nun für die zwei genannten Methoden für $n = 10$ für alle p zwischen 0 und 1.



Wie wir sehen, sacken die Überdeckungswahrscheinlichkeiten für das einfache Wald-Konfidenzintervall ab, wenn das wahre p Nahe 0 oder nahe 1 ist.

Grund: Angenommen, $p = 0.1$. Dann ist $K = 0$ relativ wahrscheinlich. Wir schätzen dann $\widehat{p} = K/n = 0/n = 0$ und $\widehat{p} \cdot (1 - \widehat{p})/n \approx 0$. Somit wird das Wald-Konfidenzintervall in etwa $[0, 0]$ sein und folglich das wahre $p = 0.1$ nicht enthalten.

Ein einfacher Trick, das Problem zu lösen, besteht darin, das Konfidenzintervall so zu berechnen, als wäre nicht K sondern $K + 1$ beobachtet worden (um $\widehat{p} = 0$ im Fall $K = 0$ zu vermeiden) und als wäre die Gesamtzahl nicht n sondern $n + 2$ (um $\widehat{p} = 1$ im Fall $K = n$ zu vermeiden).

Der "k+1, n+2"-Trick

Siehe S. 121 in

Literatur

[KW08] Götz Kersting, Anton Wakolbinger (2008) *Elementare Stochastik*, Birkhäuser, Basel.

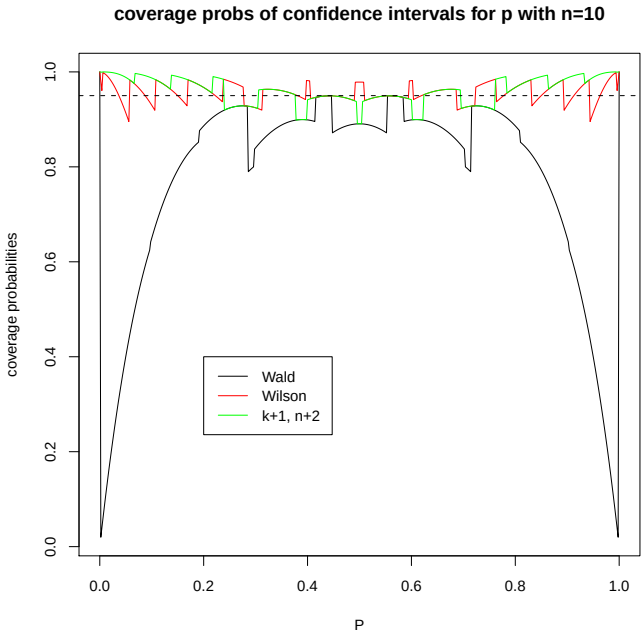
Sind k Erfolge in n Versuchen beobachtet worden, so schätze die Erfolgswahrscheinlichkeit durch

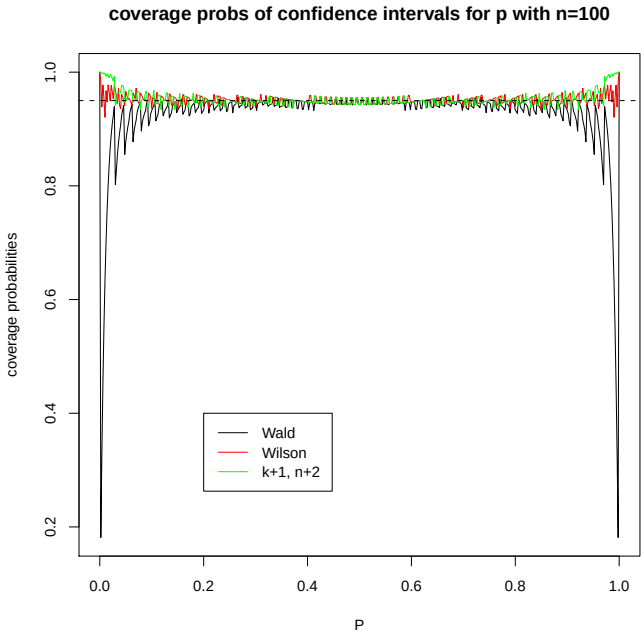
$$\widetilde{p} = (k + 1)/(n + 2)$$

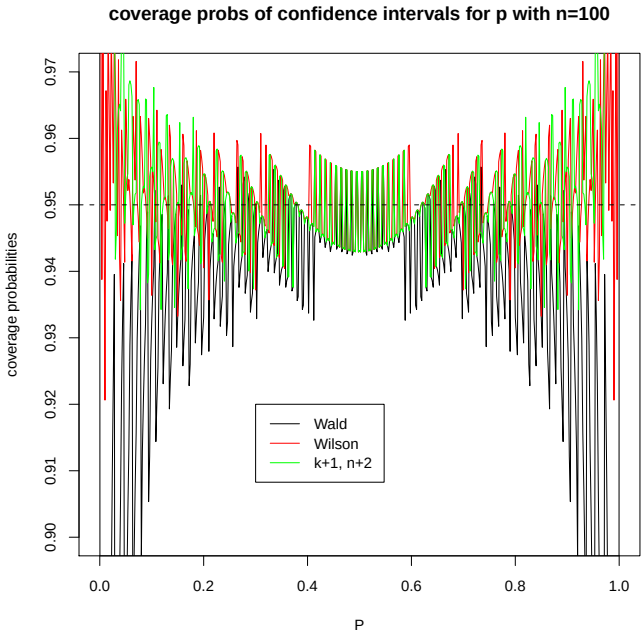
dieses $\widetilde{p}$ verwenden wir dann im einfachen Wald-Konfidenzintervall

$$\left[\widetilde{p} - 1.96 \cdot \sqrt{\widetilde{p} \cdot (1 - \widetilde{p})/n}\ ,\ \widetilde{p} + 1.96 \cdot \sqrt{\widetilde{p} \cdot (1 - \widetilde{p})/n}\right]$$

Das funktioniert erstaunlich gut, und zwar nicht nur für p in der Nähe von 0 oder 1.

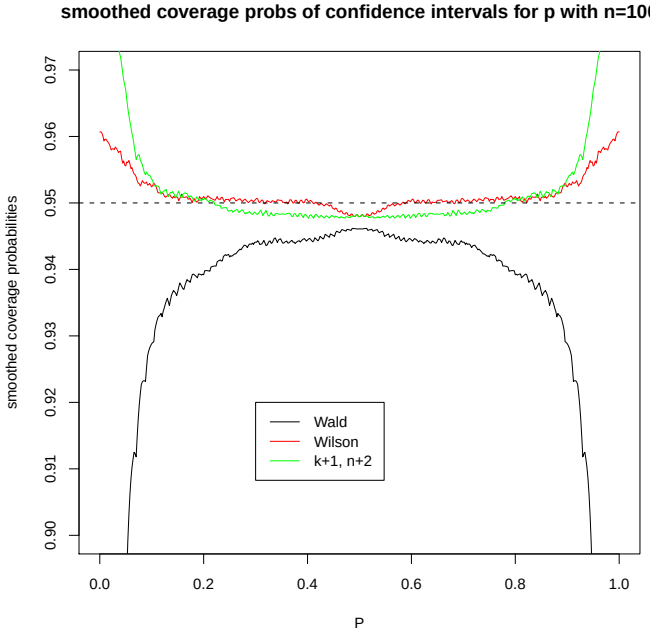

coverage probs of confidence intervals for p with n=10
coverage probabilities
0.0
0.2
0.4
0.6
0.8
1.0
P
Wald
Wilson
k+1, n+2


coverage probs of confidence intervals for p with n=100
coverage probabilities
0.2
0.4
0.6
0.8
1.0
P
Wald
Wilson
k+1, n+2



Die Überdeckungswahrscheinlichkeit hängt offensichtlich stark vom genauen Wert von p ab und liegt bei allen drei Methoden für bestimmte p auch mal unter 95%. Dafür kann sie gleich daneben für ein leicht verändertes p über 95% liegen.[1cm]

Um ein deutlicheres Bild zu bekommen, glätten wir die Kurven, indem wir jeweils über ein kleines Intervall von Werten für p mitteln.



Wir sehen also, dass die Wilson-Methode und die "k+1, n+2"-Wald-Methode sowohl bei $n = 10$ als auch bei $n = 100$ deutlich zuverlässigere Konfidenzintervalle liefern als die einfache Wald-Methode. Das gilt insbesondere

für p, die nahe bei 0 oder nahe bei 1 liegen.

Wir werden bei der Bayesschen Statistik noch einmal auf den "k+1, n+2"-Trick zurückkommen.

2.6 Grundsätzliches zur frequentistischen Statistik

- Parameter sind unbekannt aber nicht zufällig.
- Daten hängen von den Parametern und vom Zufall ab (gemäß Modellannahmen).
- frequentistischer Wahrscheinlichkeitsbegriff: Wenn ein Ereignis eine Wahrscheinlichkeit p hat, dann bedeutet das, dass es auf lange Sicht im Anteil p aller Fälle eintritt.
- Wenn ich meine Tests mit Signifikanzniveau α durchführe, verwerfe ich die Nullhypothese zu Unrecht nur in einem Anteil α der Fälle. (auf lange Sicht)
- Wenn ich 95%-Konfidenzintervalle angebe, enthalten 95% meiner Konfidenzintervalle den tatsächlichen Parameterwert. (auf lange Sicht)

2.7 Maximum-Likelihood-Schätzer

- Auch wenn es allgemein sinnvoll ist, Konfidenzintervalle für Parameterschätzungen anzugeben, möchte man manchmal einen einzelnen Schätzwert für einen Parameter angeben, und die frequentistische Statistik hat auch hierfür eine bevorzugte Methode, die *Maximum-Likelihood*-Schätzung (kurz ML).
- Es ist sinnlos, nach dem "wahrscheinlichsten" Parameterwert zu fragen, denn Parameter sind (aus Sicht der frequentistischen Statistik) nicht zufällig und haben daher auch keine Wahrscheinlichkeit.
- Statt dessen sucht man den Parameterwert, der die Daten am wahrscheinlichsten macht. Die *Likelihood* eines Werts x für einen Parameter θ ist die Wahrscheinlichkeit der beobachteten Daten D, falls $\theta = x$ gilt:

$$L_D(x) := \Pr_{\theta=x}(D)$$

- Die *Likelihood* eines Werts x für einen Parameter θ ist die Wahrscheinlichkeit der beobachteten Daten D, falls $\theta = x$ gilt:

$$L_D(x) := \Pr_{\theta=x}(D)$$

- Der *Maximum-Likelihood-Schätzer* (ML-Schätzer) ist der Parameterwert $\widehat{\theta}$, für den die Funktion L_D maximal wird:

$$\widehat{\theta} = \arg\max_x L_D(x)$$

  also dasjenige x, für das $L_D(x)$ maximal wird

Beispiel: Auf einem mtDNA-Abschnitt der Länge 100 werden zwischen Mensch und Schimpanse 7 Unterschiede festgestellt. Wie hoch ist die Wahrscheinlichkeit p, auch an der 101. Position einen Unterschied zu sehen?

Naheliegender Schätzer 7/100

ML-Schätzer: Modelliere die Anzahl K der beobachteten Mutationen als binomialverteilt mit $n = 100$ und unbekanntem p. Dann gilt

$$L(p) = \Pr_p(K = 7) = \binom{100}{7} p^7 \cdot (1-p)^{93}$$

und

$$\begin{aligned} \widehat{p} &= \arg\max_p \binom{100}{7} p^7 \cdot (1-p)^{93} = \arg\max_p p^7 \cdot (1-p)^{93} \\ &= \arg\max_p \log\left(p^7 \cdot (1-p)^{93}\right) \end{aligned}$$

Gesucht ist also die Maximalstelle von

$$f(p) := \log\left(p^7 \cdot (1-p)^{93}\right) \;=\; 7 \cdot \log(p) + 93 \cdot \log(1-p).$$

Wir finden Sie wie üblich durch Nullsetzen der Ableitung:

$$0 = f'(p) = 7 \cdot \frac{1}{p} + 93\frac{1}{1-p} \cdot (-1)$$

(dabei hilft es, zu wissen dass $\log'(x) = 1/x$.) Löst man die Gleichung nach p so erhält man:

$$\widehat{p} = 7/100$$

Wir haben also eine theoretische Begründung für den naheliegenden Schätzer 7/100 gefunden.

Der ML-Schätzer ist in vielen Fällen *konsistent*, d.h. wenn genügend viele Daten vorliegen und die Modellannahmen erfüllt sind, wird er den tatsächlichen Parameterwert finden.

Wenn eher wenig Daten vorhanden sind, ist manchmal ein anderer Schätzer zu bevorzugen.

Beispiel: ist $X_1, \ldots, X_n$ eine Stichprobe aus einer Normalverteilung, so ist $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ der ML-Schätzer für die Varianz σ^2. Meistens wird aber der Bias-korrigierte Schätzer $\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ bevorzugt.

Was Sie u.a. erklären können sollten

- Was sollte ein Konfidenzintervall leisten?
- Was ist dabei zufällig und was nicht?
- Studentisiertes Konfidenzintervall
- Dualität Konfidenzintervall $\leftrightarrow$ Test
- Konfidenzintervalle für Wahrscheinlichkeiten/Anteile
- Wald-Konfidenzintervall, seine Probleme mit den Überdeckungswahrscheinlichkeiten, Alternativen
- Wieso Konfidenintervalle besser sind als nur Signifikanzen
- Wieso Konfidenintervalle besonders nützlich sind bei Nichtsignifikanz
- Prinzip der ML-Schätzung

3 Bedingte Wahrscheinlichkeiten und die Bayes-Formel

3.1 Beispiel: Medizinischer Test

Daten zur Brustkrebs-Mammographie:

- 0.8% der 50-jährigen Frauen haben Brustkrebs.
- Das Mammogramm erkennt Brustkrebs bei 90% der Erkrankten.
- Das Mammogramm gibt bei 7% der Gesunden Fehlalarm.

Bei einer Vorsorgeuntersuchung zeigt das Mammogramm Brustkrebs an. Wie hoch ist die Wahrscheinlichkeit, dass die Patientin tatsächlich Krebs hat?

24 erfahrene Ärzte sollten diese Frage beantworten[1].

- 8 Ärzte gaben an: 90%
- 8 Ärzte gaben an: 50 bis 80%
- 8 Ärzte gaben an: 10% oder weniger.

Hier geht es um eine *bedingte Wahrscheinlichkeit*: Wie groß ist die *bedingte* Wahrscheinlichkeit, Krebs zu haben, *gegeben*, dass das Mammogramm dies anzeigt?[2cm]

Bedingte Wahrscheinlichkeiten können wir mit der Bayes-Formel berechnen.

[1] Gigerenzer, G. & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *BMJ*, **327**, 741-744

A, B Ereignisse

Bedingte Wahrscheinlichkeit von A, gegeben B (sofern $\Pr(B) > 0$):

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

($A \cap B$:= A und B treten beide ein)
"gegeben B" bedeutet: wenn man schon weiß, dass B eintritt oder eingetreten ist

Thomas Bayes,
1702–1761

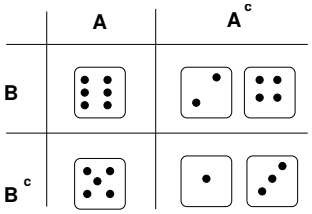
Satz von der totalen Wahrscheinlichkeit (mit B^c:={B tritt nicht ein}):

$$\Pr(A) = \Pr(B)\Pr(A|B) + \Pr(B^c)\Pr(A|B^c)$$

Bayes-Formel:

$$\Pr(B|A) = \frac{\Pr(B)\Pr(A|B)}{\Pr(A)}$$

Beispiel: Sei $W \in \{1,2,3,4,5,6\}$ das Ergebnis eines Würfelwurfs. Wie wahrscheinlich ist $W \geq 5$, wenn W eine gerade Zahl ist? [0.5cm]



$$\begin{aligned} A &:= \{W \geq 5\} \\ B &:= \{W \text{ ist gerade }\} \\ A \cap B &= \{W \text{ ist gerade und } \geq 5\} \end{aligned}$$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/6}{3/6} = \frac{1}{3}$$

$$\Pr(B|A) = \frac{\Pr(B) \cdot \Pr(A|B)}{\Pr(A)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{1/3} = \frac{1}{2}$$

Nun zurück zur Mammographie. Definiere Ereignisse:

A**:** Das Mammogramm zeigt Krebs an.

B**:** Die Patientin hat Krebs.

Die nicht bedingte Wahrscheinlichkeit $\Pr(B)$ heißt auch *a-priori*-Wahrscheinlichkeit für B, d.h. die Wahrscheinlichkeit, die man B zuordnet, *bevor* man die Daten A gesehen hat. In unserem Fall also 0.008, die Wahrscheinlichkeit, dass eine Vorsorgepatientin Brustkrebs hat.[0.5cm] Die bedingte Wahrscheinlichkeit $\Pr(B|A)$ heißt auch *a-posteriori*-Wahrscheinlichkeit für B. Das ist die Wahrscheinlichkeit, die man B zuweist, *nachdem* man die Daten A gesehen hat.

Die bedingte Wahrscheinlichkeit, dass die Patientin Krebs hat, gegeben, dass das Mammogramm dies anzeigt, ist:

$$\begin{aligned} \Pr(B|A) &= \frac{\Pr(B) \cdot \Pr(A|B)}{\Pr(A)} \\ &= \frac{\Pr(B) \cdot \Pr(A|B)}{\Pr(B) \cdot \Pr(A|B) + \Pr(B^C) \cdot \Pr(A|B^C)} \\ &= \frac{0.008 \cdot 0.9}{0.008 \cdot 0.9 + 0.992 \cdot 0.07} \approx 0.0939 \end{aligned}$$

Bedingt darauf, dass das Mammogramm Krebs anzeigt, beträgt die Wahrscheinlichkeit, dass die Patientin Krebs hat, also lediglich 9.4%. Das richtige Ergebnis "ungefähr 10%" hatten übrigens nur 4 der 24 Ärzte genannt. Zwei davon haben eine unzutreffende Begründung genannt und waren wohl nur zufällig auf das richtige Ergebnis gekommen.

3.2 Das Ziegenproblem

Das Ziegenproblem

- In der amerikanischen TV-Show *Let's Make a Deal* kann der Kandidat am Ende einen Sportwagen gewinnen, der sich hinter einer von drei Türen verbirgt.
- Hinter den anderen beiden Türen stehen Ziegen.
- Der Kandidat entscheidet sich zunächst für eine der drei Türen, z.B. Tür 1.
- Der Showmaster öffnet dann eine der beiden anderen Türen, und zwar eine, hinter der eine Ziege steht, z.B. Tür 2.
- Der Kandidat kann nun bei Tür 1 bleiben oder sich umentscheiden und Tür 3 wählen.
- Sollte er sich umentscheiden?

A : Der Showmaster öffnet Tür 2.

B : Das Auto ist hinter Tür 3.

C : Das Auto ist hinter Tür 1.

D : Das Auto ist hinter Tür 2.

$\Pr(B) = 1/3 = \Pr(C) = \Pr(D)$ $\Pr(A|B) = 1$, $\Pr(A|C) = 1/2$, $\Pr(A|D) = 0$.

$$\begin{aligned}\Pr(B|A) &= \frac{\Pr(B)\cdot\Pr(A|B)}{\Pr(B)\cdot\Pr(A|B)+\Pr(C)\cdot\Pr(A|C)+\Pr(D)\cdot\Pr(A|D)}\\ &= \frac{\frac{1}{3}\cdot 1}{\frac{1}{3}\cdot 1+\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot 0}\\ &= 2/3\end{aligned}$$

Es lohnt sich also, zu Tür 3 zu wechseln.

Durch das Öffnen einer Tür hat man Information hinzu gewonnen, denn: Der Showmaster öffnet eine Ziegen-Tür, niemals die Auto-Tür.

Mit Ws 2/3 wählt man zu Beginn eine Ziegen-Tür. Nachdem die zweite Ziegen-Tür geöffnet wurde, wechselt man automatisch zur Auto-Tür.

Mit Ws 1/3 wählt man zu Beginn die Auto-Tür. Nachdem eine Ziegen-Tür geöffnet wurde, wechselt man automatisch zu einer Ziegen-Tür.

4 Bayessche Statistik

Grundannahmen der Bayesschen Statistik

- Parameter werden auch als zufällig betrachtet
- Die *a-priori-Wahrscheinlichkeitsverteilung* eines Parameters gibt an, für wie wahrscheinlich man die möglichen Parameterwerte hält, **bevor** man die Daten gesehen hat.
- Mit der Bayes-Formel erhält man die *a-posteriori-Verteilung*, also die bedingte Wahrscheinlichkeitsverteilung der Parameterwerte θ gegeben die Daten D.

$$\Pr(\theta_0|D) = \frac{\Pr(D|\theta_0)\cdot\Pr(\theta_0)}{\Pr(D)} = \frac{\Pr(D|\theta_0)\cdot\Pr(\theta_0)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

Das Ganze geht nur, wenn die a-priori-Wahrscheinlichkeiten $\Pr(\theta)$ definiert sind. $\Pr(D|\theta_0)$ ist gerade die Likelihood $L_D(\theta)$ aus der frequentistischen Statistik.

- Falls die a-priori-Verteilung eine kontinuierliche Verteilung mit Dichte $f(\theta)$ ist, hat die a-posteri-Verteilung die Dichte
$$g_D(\theta_0) = \frac{\Pr(D|\theta_0) \cdot f(\theta_0)}{\int \Pr(D|\theta) \cdot f(\theta) d\theta},$$
also jedenfalls proportional zu $\Pr(D|\theta_0) \cdot f(\theta_0)$ und skaliert zu Fläche 1 unter der Kurve.
- Wenn man a-posteriori-Verteilungen für Parameter berechnen oder simulieren kann, kann man sich ein Bild davon machen, welche Parameterwerte angesichts der Daten in Frage kommen.
- Statt des ML-Schätzers verwendet man zur Parameterschätzung den Erwartungswert der a-posteriori-Verteilung oder den Wert mit der höchsten a-posteriori-Wahrscheinlichkeit(sdichte) [MAP=maximum a-posteriori].
- Analog zu den Konfidenzintervallen der frequentistischen Statistik gibt es in der Bayesschen Statistik die **Kredibilitätsbereiche**. Ein 95%-Kredibilitätsbereich ist ein Parameterbereich, in dem gemäß der a-posteriori-Verteilung der wahre Parameter mit 95%-iger Wahrscheinlichkeit liegt.

Bei Schätzungen von Anteilen oder Wahrscheinlichkeiten, also des Parameters p der Binomverteilung, werden oft Beta-verteilte a-priori-Verteilungen verwendet.

Wie wir sehen werden, lassen sich dann die a-posterior-Verteilungen leicht berechnen.

Die Dichte $f(x)$ der Beta(a,b)-Verteilung ist proportional zu
$$x^{a-1} \cdot (1-x)^{b-1},$$
also
$$f(x) = \frac{x^{a-1} \cdot (1-x)^{b-1}}{B(a,b)},$$
für $x \in [0,1]$, wobei die Beta-Funktion $B(a,b)$ im Nenner für Gesamtfläche 1 sorgt.

Spezialfall: Die Beta(1,1)-Verteilung hat Dichte $f(x) = 1$, ist also die uniforme Verteilung auf $[0,1]$.

Beispiel: $n = 20$ **Versuche,** $K = 3$ **Erfolge,** $p =?$
K ist binomialverteilt mit $n = 20$. Wir beobachten $K = 3$. Der ML-Schätzer ist also $\widehat{p} = 3/20$.

Wie sieht die a-posteriori-Verteilung für p aus?

Die ist nur definiert, wenn wir zunächst eine a-priori-Verteilung für p definieren. Wir gehen mal von der uniformen Verteilung auf $[0,1]$ aus, also Dichte $f(p) = 1$ ("alles gleich wahrscheinlich").

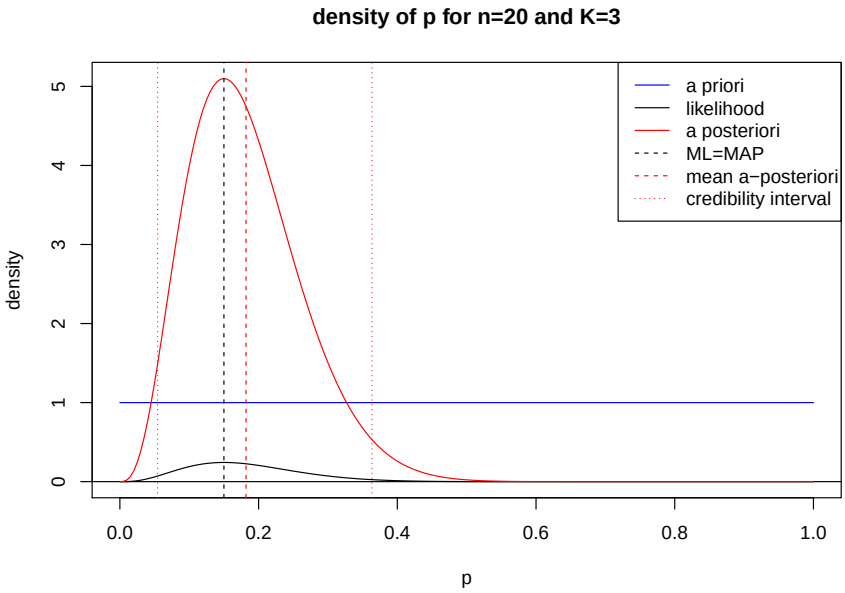
Die Likelihood-Funktion ist
$$L(p) \;=\; \binom{n}{K} p^K \cdot (1-p)^{n-K} \;=\; \binom{20}{3} p^3 \cdot (1-p)^{17}$$

Als a-posteriori-Verteilung ergibt sich dann die Beta($1+K$,$1+n-K$)-Verteilung mit Dichte $b(p)$ proportional zu $p^3 \cdot (1-p)^{17} \cdot f(p) = p^3 \cdot (1-p)^{17}$, und damit
$$b(p) \;=\; \frac{p^K \cdot (1-p)^{n-K}}{B(1+K, 1+n-K)} \;=\; \frac{p^3 \cdot (1-p)^{17}}{B(4,18)}.$$

Siehe auch S. 106 in

Literatur

[KW08] G. Kersting, A. Wakolbinger (2008) *Elementare Stochastik*, Birkhäuser, Basel.



- Der ML-Schätzer und der MAP-Schätzer stimmen in diesem Beispiel wegen der uniformen a-priori-Verteilung überein.
- Der Erwartungswert der a-posteriori-Verteilung beta$(1+K,1+n-K)$ ist

$$\mathbb{E}(p|K) = \frac{K+1}{n+2}.$$

  Diesen Schätzer kennen wir bereits vom "$k+1$, $n+2$"-Trick als $\widetilde{p}$. Wir erhalten hier also eine Bayessche Interpretation/Begründung für diesen Schätzer!

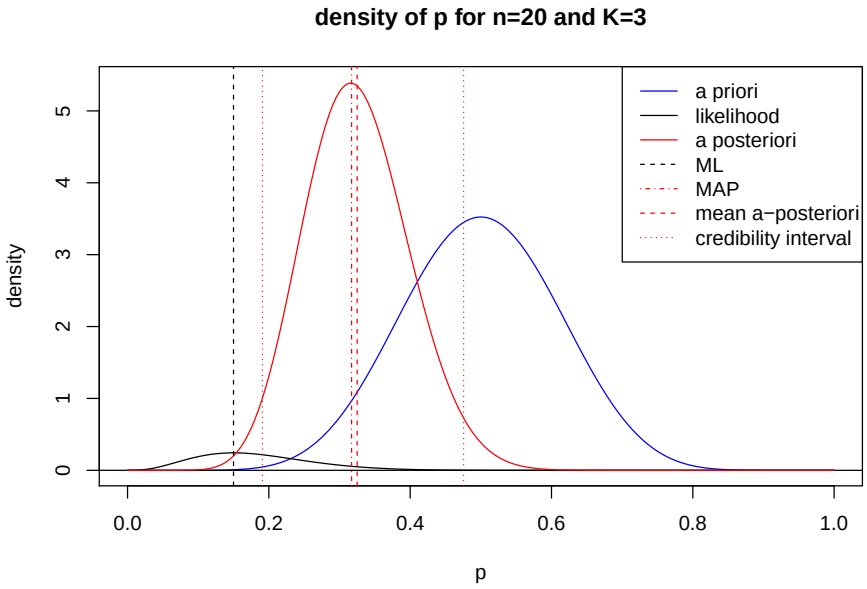
Jetzt mit anderem Prior

Beispiel: $n = 20$ **Versuche,** $K = 3$ **"Erfolge",** $p =?$

K ist binomialverteilt mit $n = 20$ (z.B. Anzahl beobachtete Stockenten). Wir beobachten $K = 3$ (z.B. Anzahl weiblich).

Wenn wir aufgrund von Vorwissen oder allgemeinen Plausibilitätsüberlegungen davon ausgehen, dass a priori z.B. Werte von p um 0.5 wahrscheinlicher sind als Werte nahe bei 0 oder 1, können wir für p z.B. einen beta(10, 10)-verteilten Prior verwenden, also mit Dichte $f(p)$ proportional zu $p^9 \cdot (1-p)^9$.

Damit hat die a-posteriori Verteilung eine Dichte proportional zu $p^3 \cdot (1-p)^{17} \cdot p^9 \cdot (1-p)^9 = p^{12} \cdot (1-p)^{26}$, ist also die beta(13, 27)-Verteilung.



	Wald-Konfidenzintervall:	[0, 0.306]
	"$k+1$, $n+2$"- Wald-Konfint.:	[0.013, 0.351]
Intervallschätzer: $(3/20 = 0.15)$	Wilson-Konfidenzintervall:	[0.052, 0.360]
	Kredibilitätsbereich mit uniformem Prior:	[0.054, 0.363]
	Kredibilitätsbereich mit beta(10, 10) Prior:	[0.191, 0.476]

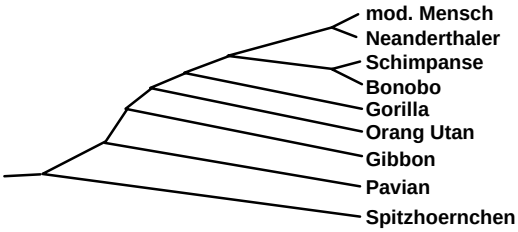
Allgemein gilt übrigens eine nützliche Eigenschaft der Familie der beta-Verteilungen: Verwendet man für den Parameter p einer Binomialverteilung mit gegebenem n und beobachtetem Wert k als Prior eine beta(a,b)-Verteilung, so ist der Posterior beta($a+k$,$b+n-k$)-verteilt. (Und die beta(1,1)-Verteilung ist übrigens die uniforme Verteilung auf [0,1]).

Frequentistische vs. Bayessche Statistik

- Lange Zeit stritten Frequentisten und Bayesianer darüber, welche Sicht auf die Statistik die "richtige" sei.
- Hauptkritikpunkt an den Bayesschen Methoden: Die Wahl einer a-priori-Verteilung ist subjektiv.
- Heute verwenden die meisten, die professionell Daten analysieren, sowohl frequentistische und Bayessche Methoden je nach Bedarf.
- Die Wahl der a-priori-Verteilung ist aber in der Tat ein heikler Punkt; eine uniforme Verteilung zu wählen, ist nicht immer eine Lösung.

Beispiel: Stammbaumschätzung

```
Bonobo      ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAAGCAAATTTAAGTGCCACCCAAGTATTGGCTCA...
Schimpanse  ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAAGCAAATTTAAGTACCACCTAAGTACTGGCTCA...
Gibbon      TATTCTCATGTGGAAGCCATTTTGGGTACAACCCCAGTACTAACCCACTTCTCCACAACTCTATGTACTT...
Gorilla     ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAGACAAATTTGGGTACCACCCAAGTATTGGCTAA...
mod. Mensch ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCA...
Neanderth   CCAAGTATTGACTCACCCATCAACAACCGCCATGTATTTCGTACATTACTGCCAGCCACCATGAATATTG...
Pavian      TATTTTATGTTGTACAAGCCCCACAGTACAACCTTAGCACTAGCTAACTTTTAATGCCACTATGTAATTC...
Oran Utan   TTCTTTCATGGGGGACCAGATTTGGGTGCCACCCCAGTACTGACCCATTTCTAACGGCCTATGTATTTCG...
Spitzhrn    CGTGCATTAATGCTTTACCACATTAATATATGGTACAGTACATAACTGTATATAAGTACATAGTACATTT...
```



- Parameterwerte müssen nicht immer Zahlen sein.
- In der Phylogenieschätzung ist der zu schätzende Baum der Parameter.
- ML-Programme wie PHYLIP/dnaml suchen den ML-Baum, also den Baum, für den die Sequenzdaten am wahrscheinlichsten sind.
- Bayessche Programme wie MrBayes oder BEAST erzeugen zunächst viele Bäume gemäß der a-posteriori-Verteilung (gegeben die Sequenzdaten) und fassen dann zusammen, welche Aussagen (z.B. "Mensch, Schimpanse und Bonobo bilden eine monophyletische Gruppe") für welchen Anteil der Bäume gelten.
- Mehr dazu erfahren Sie im LMU-EES-Master-Studiengang.

Was Sie u.a. erklären können sollten

- Bedingte Wahrscheinlichkeiten
- Satz von der totalen Wahrscheinlichkeit
- Bayes-Formel und wie man sie anwendet
- Unterschiede zwischen frequentistischer und Bayesscher Statistik
- a-priori- und a-posteriori-Verteilungen
- Kredibilitätsbereich, auch im Vergleich zum Konfidenzintervall

Bitte beachten Sie auch die Liste aus Seite 15.