

Wahrscheinlichkeitsrechnung und  
Statistik für Biologen

## 7. Frequentistische und Bayessche Intervallschätzer

Noémie Becker & Dirk Metzler

29. Mai 2013

### Inhaltsverzeichnis

#### 1 Konfidenzintervalle für Erwartungswerte

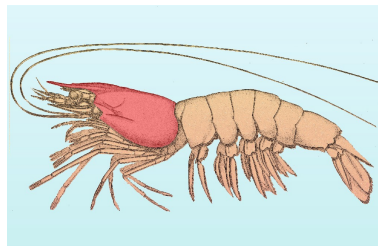
##### 1.1 Beispiel: Carapaxlänge des Springkrebses

Beispiel: Springkrebs



*Galathea squamifera*

Carapaxlänge:



Wie groß ist die mittlere Carapaxlänge des weiblichen Springkrebses?

Alle weiblichen Springkrebsse (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:  $\bar{x} = 3.23$  mm  $sd(x) = 0.9$  mm  $n = 29$   $sem(x) = \frac{sd(x)}{\sqrt{n}} = \frac{0.9}{\sqrt{29}} = 0.17$  ( $= sd(\bar{x})$ )

Wir kennen bereits folgende Faustformeln:

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

$$[\bar{x} - sem(x), \bar{x} + sem(x)]$$

mit Wahrscheinlichkeit nahe bei 2/3

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

$$[\bar{x} - 2 * sem(x), \bar{x} + 2 * sem(x)]$$

mit Wahrscheinlichkeit nahe bei 95%

Nun exakt: Sei  $t_{2.5\%} <- -qt(0.025, length(x)-1)$ . Dann liegt der wahre Mittelwert mit Wahrscheinlichkeit 95% im Intervall

$$[\bar{x} - t_{2.5\%} * sem(x), \bar{x} + t_{2.5\%} * sem(x)]$$

Zur Begründung siehe nächster Abschnitt.

Setzt man die Zahlenwerte  $\bar{x} = 3.23$ ,  $t_{2.5\%} = 2.05$  und  $sem(x) = 0.17$  in

$$[\bar{x} - t_{2.5\%} * sem(x), \bar{x} + t_{2.5\%} * sem(x)]$$

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  $\alpha$

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

$$[\underline{I}(x), \bar{I}(x)]$$

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

$$\Pr_{\mu} (\mu \in [\underline{I}(x), \bar{I}(x)]) \geq 1 - \alpha$$

Selbstverständlich wollen wir das Konfidenzintervall möglichst klein wählen.

## Konfidenzintervall für den wahren Mittelwert

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

$$t := \frac{\bar{x} - \mu}{\text{sem}(x)}$$

annähernd Student-verteilt ist mit  $\text{length}(x)-1$  Freiheitsgraden (wenn  $\text{length}(x)$  groß genug ist).

Sei  $t_{\alpha/2} \leftarrow -\text{qt}(\alpha/2, \text{length}(x)-1)$  das  $\alpha/2$ -Quantil der Student-Verteilung mit  $\text{length}(x)-1$  Freiheitsgraden. Dann ist

$$[\bar{x} - t_{\alpha/2} * \text{sem}(x), \bar{x} + t_{\alpha/2} * \text{sem}(x)]$$

das Konfidenzintervall zum Irrtumsniveau  $\alpha$ .

Begründung für Interessierte:

$$\begin{aligned} \Pr_{\mu}(\mu \in [\bar{x} - t_{\alpha/2} * \text{sem}(x), \bar{x} + t_{\alpha/2} * \text{sem}(x)]) &= \Pr_{\mu}(\mu - \bar{x} \in [-t_{\alpha/2} * \text{sem}(x), t_{\alpha/2} * \text{sem}(x)]) \\ &= \Pr_{\mu}\left(\frac{\mu - \bar{x}}{\text{sem}(x)} \in [-t_{\alpha/2}, t_{\alpha/2}]\right) \\ &= \Pr_{\mu}\left(\left|\frac{\mu - \bar{x}}{\text{sem}(x)}\right| \leq t_{\alpha/2}\right) \\ &= \Pr_{\mu}(|t| \leq t_{\alpha/2}) \\ &= 1 - \alpha \end{aligned}$$

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

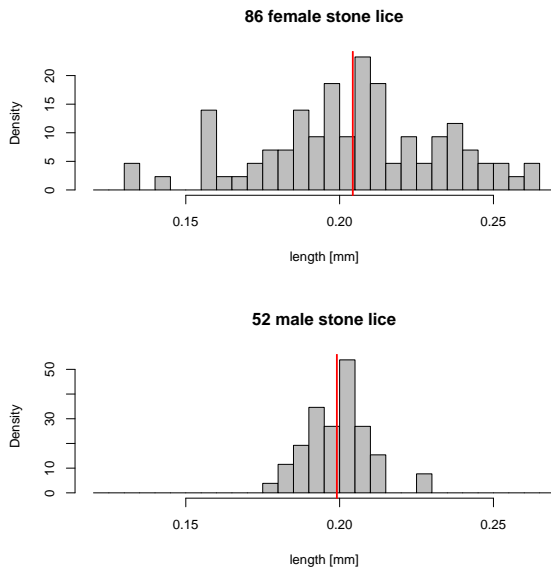
Ist  $[a, b]$  ein  $(1 - \alpha)$ -Konfidenzintervall für einen Parameter  $\theta$ , so erhält man einen Test mit Signifikanzniveau  $\alpha$ , wenn man die Nullhypothese  $\theta = x$  genau dann verwirft, wenn  $x \notin [a, b]$ . [0.5cm]

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

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 Ergebnis 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. ✓
- Der Konfidenzbereich für die Differenz zwischen der mittleren Länge der Weibchen und der Männchen ist  $[-0.0045, 0.0096]$ . Der 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? ( $0,57 \pm ??$ )

### 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  $\hat{p} := \frac{X}{n}$  in der Stichprobe.

Frage: Wie verlässlich ist die Schätzung?

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

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

für jede Wahl von  $p$ .

Lösungsweg:

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

Wir wissen: Der Schätzer  $\hat{p}$  ist (in etwa) normalverteilt mit Erwartungswert  $p$  und Standardabweichung  $\sqrt{p(1 - p)/n}$ .

Lösung:

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

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

### 2.3 Beispiel: Porzellankrebs

#### Männchenanteil beim Porzellankrebs

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

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

ein, so erhält man das Konfidenzintervall

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

für den wahren Männchenanteil zum Irrtumsniveau 5%.

## 2.4 Beispiel: Stockente



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

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

ein, so erhält man das Konfidenzintervall

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

für den wahren Männchenanteil zum Irrtumsniveau 5%.

## 2.5 Bessere Konfidenzintervalle

Das Konfidenzintervall

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

nennt man auch [Wald-Konfidenzintervall](#).

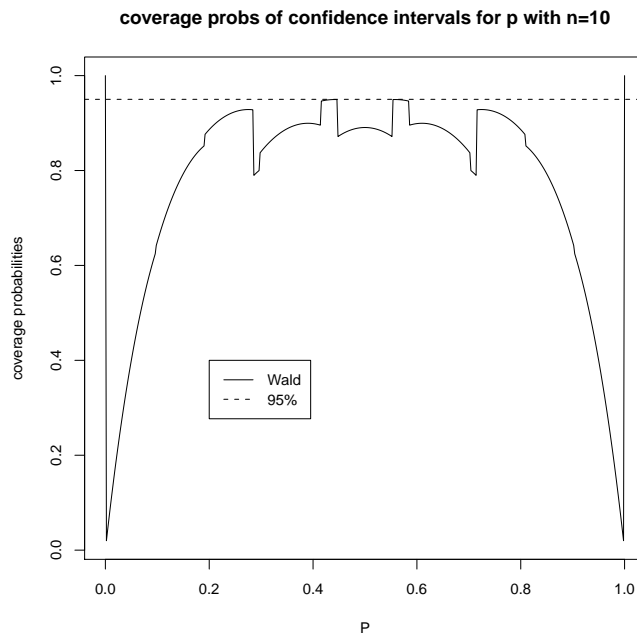
Es sollte gelten: Das Konfidenzintervall überdeckt (d.h. enthält) den wahren Parameterwert mit einer Wahrscheinlichkeit von mindestens 95%.

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

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

wobei  $\hat{p} = X/n$  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  $\hat{p} = K/n = 0/n = 0$  und  $\hat{p} \cdot (1 - \hat{p})/n = 0$ . 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).

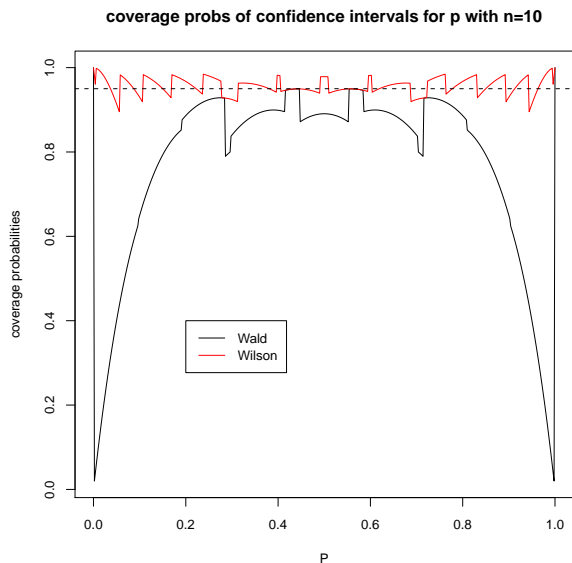
Zur Erinnerung: Konfidenzintervalle sind zufällig, da sie von den Daten abhängen.

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  $\hat{p} = K/n = 0/n = 0$  und  $\hat{p} \cdot (1 - \hat{p})/n \approx 0$ . 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  $\hat{p} = 0$  im Fall  $K = 0$  zu vermeiden) und als wäre die Gesamtzahl nicht  $n$  sondern  $n + 2$  (um  $\hat{p} = 1$  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.

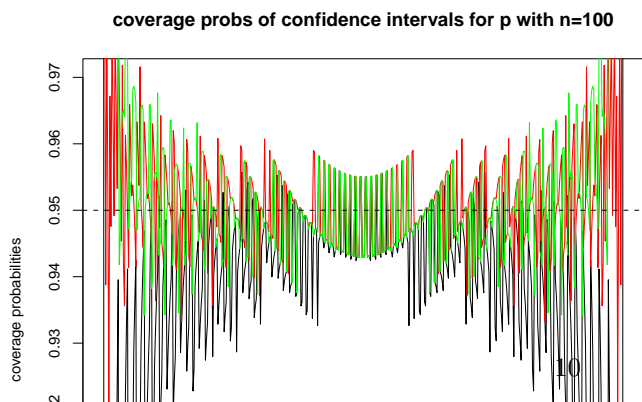
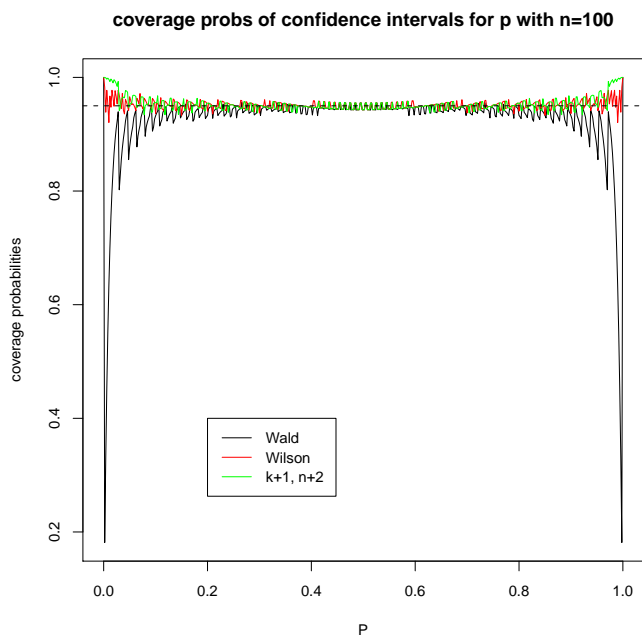
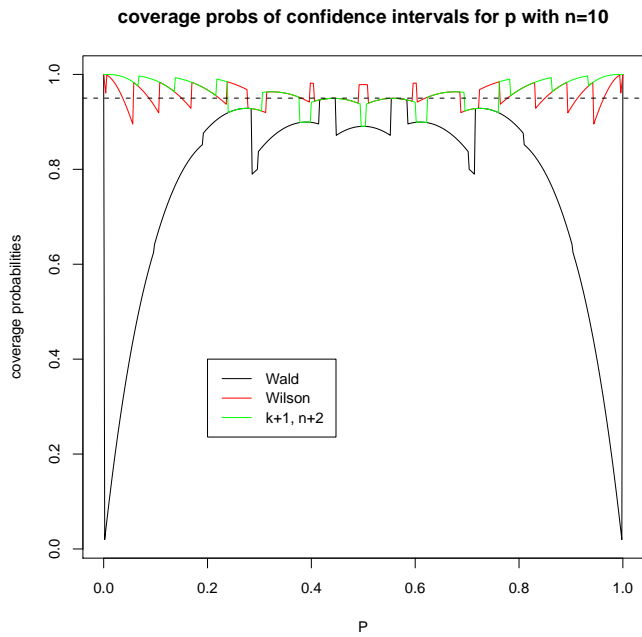
Sind  $k$  Erfolge in  $n$  Versuchen beobachtet worden, so schätze die Erfolgswahrscheinlichkeit durch

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

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

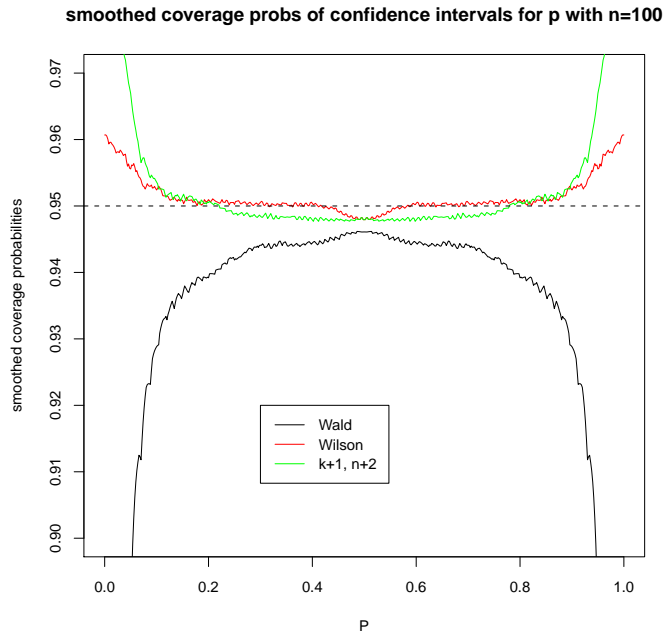
$$\left[ \tilde{p} - 1.96 \cdot \sqrt{\tilde{p} \cdot (1 - \tilde{p})/n}, \tilde{p} + 1.96 \cdot \sqrt{\tilde{p} \cdot (1 - \tilde{p})/n} \right]$$

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



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  $\alpha$  durchführe, verwerfe ich die Nullhypothese zu Unrecht nur in einem Anteil  $\alpha$  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  $\theta$  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  $\theta$  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  $\hat{\theta}$ , für den die Funktion  $L_D$  maximal wird:

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

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  $\tilde{p} = 7/100$

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

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

Gesucht ist also die Maximalstelle von

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

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:

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

Wir haben also eine theoretische Begründung für den naheliegenden Schätzer  $\tilde{p}$  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  $X_1, \dots, X_n$  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  $\sigma^2$ . Meistens wird aber der Bias-korrigierte Schätzer  $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  bevorzugt.

## 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<sup>1</sup>.

<sup>1</sup>Hoffrage, U. & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, **73**, 538-540

- 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.

$A, B$  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

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

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



Thomas Bayes,  
1702–1761

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?

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

•

[0.5cm]

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

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

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 28 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  $\theta$  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. In der Regel hat man es mit kontinuierlichen Parameterräumen zu tun. Dann sind die a-priori- und a-posteriori-Wahrscheinlichkeiten durch Dichten und die Summe durch ein Integral zu ersetzen.

- 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 Kreditabilitätsbereiche. Ein 95%-Kreditabilitätsbereich ist ein Parameterbereich, in dem gemäß der a-posteriori-Verteilung der wahre Parameter mit 95% liegt.

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

$K$  ist binomialverteilt mit  $n = 20$ . Wir beobachten  $K = 3$ . Der ML-Schätzer ist also  $\hat{p} = 3/20$ .

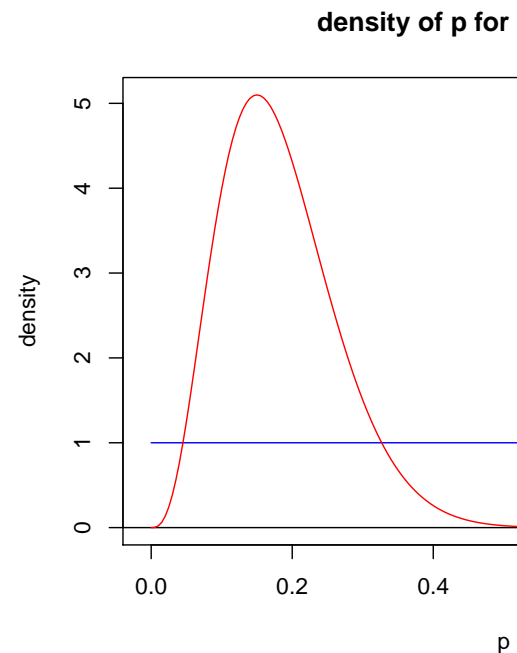
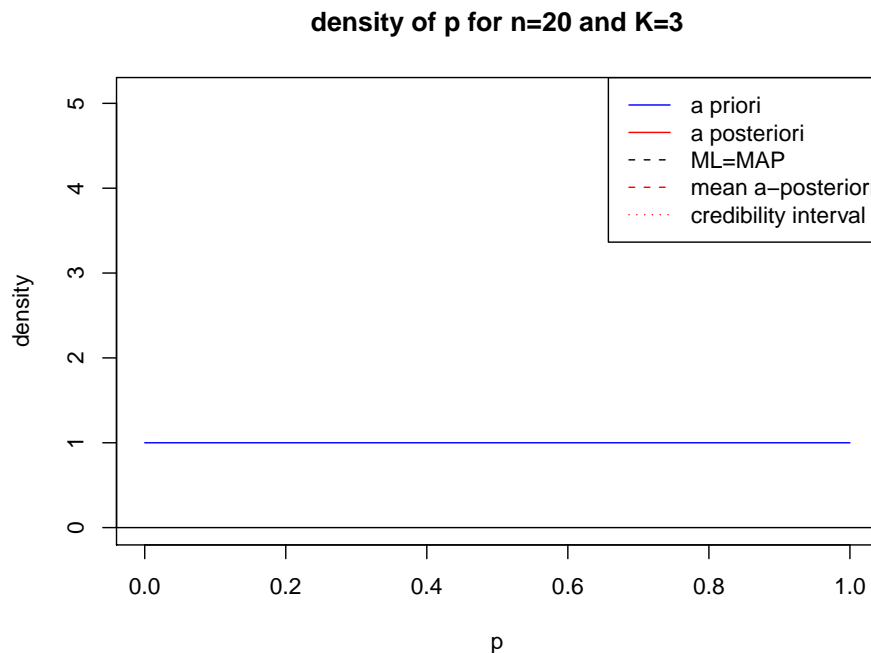
Wie sieht die a-posteriori-Verteilung für  $p$  aus?

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 (“alles gleich wahrscheinlich”).

Als a-posteriori-Verteilung ergibt sich dann die  $Beta(1 + K, 1 + n - K)$ -Verteilung, siehe S. 106 in

## Literatur

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



- Der ML-Schätzer und der MAP-Schätzer stimmen 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  $\tilde{p}$ . Wir erhalten hier also eine Bayessche Interpretation/Begründung für diesen Schätzer!

- Intervallschätzer: ( $3/20 = 0.15$ )
 

|                                    |                |
|------------------------------------|----------------|
| Wald-Konfidenzintervall:           | [0, 0.306]     |
| “ $k + 1, n + 2$ ”- Wald-Konfint.: | [0.013, 0.351] |
| Wilson-Konfidenzintervall:         | [0.052, 0.360] |
| Kreditabilitätsbereich:            | [0.054, 0.363] |

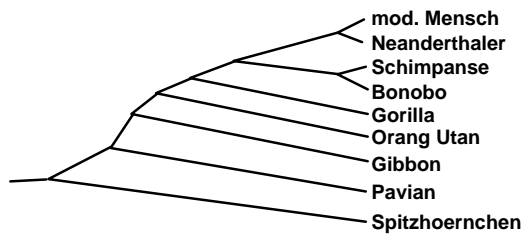
## Frequentisten vs. Bayesianer

- Lange Zeit stritten Frequentisten und Bayesianer darüber, welche Sicht auf die Statistik die “richtige” sei.
- Hauptkritikpunkt an den Bayesschen Methoden: Die Wahl einer a-posteriori-Verteilung ist subjektiv.
- Heute verwenden die meisten Statistiker 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      ATTCTAATTTAAACTATICTCTGTTCTTTTCATGGGGAAGCAAATTTAAGTGCCACCCAAGTATTGGCTCA...
Schimpanse  ATTCTAATTTAAACTATICTCTGTTCTTTTCATGGGGAAGCAAATTTAAGTACCACCTAAGTACTGGCTCA...
Gibbon      TATTCTCATGTGGAAGCCATTTTGGGTACAACCCCACTACTAACCCTTCTCCACAACCTCTATGTACTT...
Gorilla     ATTCTAATTTAAACTATICTCTGTTCTTTTCATGGGGAAGCAAATTTGGGTACCACCCAAGTATTGGCTAA...
mod. Mensch ATTCTAATTTAAACTATICTCTGTTCTTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCA...
Neanderth   CCAAGTATTGACTCACCATCAACAACCCGCGATGATTTTCGTACATTACTGCCAGCCACCATGAATATTG...
Pavian      TATTTTATGTTTACAAGCCCCACAGTACAACCTTAGCACTAGCTAACTTTAATGCCACTATGTAATTC...
Oran Utan   TTCTTTCATGGGGACAGATTTGGGTGCCACCCAGTACTGACCCATTTCTAACGGCCTATGTATTTCG...
Spitzhörn  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 [EESImu-Master’s-Studiengang](#).