Bayes in der Bucheckerschale

Ein fast minimal kurzer Einführungstext in die Grundkonzepte für bayesianische statistische Modellierung

Version: 1.0

Holger Sennhenn-Reulen
Waldinventur, Biometrie und Informatik
Waldwachstum
Nordwestdeutsche Forstliche Versuchsanstalt

Göttingen, 23. Juni 2020

Inhaltsverzeichnis

> Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations.
>
> (Gelman et al., 2014, Seite 1)

> […] science appears to work in accordance with Bayesian principles. At each stage in the development of a scientific study new information is used to adjust old information. […], this is how Bayesian modeling works. A posterior distribution created from the mixing of the model likelihood (derived from the model data) and a prior distribution (outside information we use to adjust the observed data) may itself be used as a prior for yet another enhanced model. New information is continually being used in models over time to advance yet newer models. This is the nature of scientific discovery. Yet, even if we think of a model in isolation from later models, scientists always bring their own perspectives into the creation of a model on the basis of previous studies or from their own experience in dealing with the study data. Models are not built independently of the context, so bringing in outside prior information to the study data is not unusual or overly subjective. Frequentist statisticians choose the data and predictors used to study some variable – most of the time based on their own backgrounds and external studies. Bayesians just make the process more explicit.
>
> Hilbe et al. (2017, Seite xiii)

Statistik und Erkenntnis durch Beobachtung

Unser Vertrauen auf *Statistik* baut auf auf der *'wissenschaftlichen Methode'* auf, welche im eigentlichen Sinn keine tatsächliche Methode ist, da es keine geordneten keine eindeutig festgelegten Verfahren gibt, sondern eine grob formulierte Kombination aus Empirismus – aus dem griechischen *empirikos - Erfahrung*, bedeutet 'basierend auf Beobachtung' – und Theorie darstellt, die mehrere verschachtelte und übergreifende Stufen der Argumentation mittels jeweiliger Methoden durchläuft:

Beobachtung, bei der die/der WissenschaftlerIn beobachtet, Informationen sammelt und die für das Problem relevanten Fakten untersucht und eingrenzt.
1) Bäume derselben Baumart sind bei gleichem Alter verschieden hoch und haben verschiedene Stammvolumen. 2) Bestände bestehen bei gleichem Alter aus unterschiedlich vielen Bäumen mit unterschiedlichen Höhen und Stammvolumen. 3) Diese Einzelgrößen sowie deren Verhältnisse zueinander sind in einem für die Praxis relevanten Anteil durch die Behandlung des Bestandes veränderbar. 4) Bestände können als Reinbestände einzelner Baumarten oder Mischbestände unterschiedlicher Baumarten geplant, begründet und behandelt werden. 5) Unterschiedliche Baumarten haben unterschiedliche Bedürfnisse bezüglich der für das Wachstum notwendigen Ressourcen wie Licht, Wärme, Wasser, Nährstoffe, Bodenphysik, …. 6) Diese Unterschiede haben unter Anderem ihren Ursprung in unterschiedlichem Wachstum und unterschiedlicher Dimensionen der Pflanzenbestandteile, wie etwa das Volumen des durch die Wurzeln erschlossenen Bodenraums. 7) Der Klimawandel führt zu einer veränderten Umgebung aktueller und zukünftiger Bestände. 8) In Trockenjahren bestimmt die der Pflanzen zur Verfügung stehende Wassermenge die Quantität und Qualität des Baumwachstums.

Theorie, in der die/der WissenschaftlerIn fundierte Vermutungen oder Erklärungen für beobachtete Befunde und Fakten vorbringt.
Pflanzenverfügbares Wasser kann in trockeneren Vegetationsphasen aufgrund einer wasserempfindlicheren Humusschicht und eines umfassenderen Wurzelsystems dann in gemischten Beständen besser gespeichert und genutzt werden wenn flach- und tiefwurzelnde Baumarten miteinander geeignet kombiniert werden.

Vorhersage, bei der die auf der Theorie basierenden vorausschauenden Ableitungen auf überprüfbare Weise dargestellt werden.
In einem einzelbaumweisen, homogen begründeten Mischbestand mit Buche und Eiche kann die zu erwartende Produktivität – erhoben als Stammholzvolumen in m^3/ha – eines Bestandes in einem Bestandealter von 100 Jahren um einen Faktor von mindestens 1.1 gegenüber vergleichbaren Reinbeständen aus Eiche oder Buche gesteigert werden.

Überprüfung, bei der durch Messung Daten gesammelt werden, um die Vorhersagen zu überprüfen.
Basierend auf der vergleichenden Messung auf 50 forstlichen Versuchsflächen ergab sich mit einer Wahrscheinlichkeit von 0.73[1] eine Steigerung des Stammholzvolumen – in m^3/ha – um mindestens den Faktor 1.1.

[1] Dies ist ein fiktives Beispiel und darum ist dieser Wert hier auch frei erfunden!

Die *Messung* ist hier eine Methode die in den Forstwissenschaften dazu genutzt wird, Daten von Bäumen, Waldbeständen oder Baum(-Umgebungs-)komponenten zu erzeugen. Empirische Forschung ist die der Beobachtung/Messung übergeordnete erkenntnistheoretische Methode die auf dem Ziel basiert – mittels der Messung zur Erzeugung von Daten –, auf einen (teilweise) unbekannten Prozess zuzugreifen um sich darauf basierend neue, reproduzierbare Erkenntnisse zu erarbeiten. Solch ein nicht direkt untersuchbarer Prozess wird auch *Messprozess* oder *Daten-generierender Mechanismus* genannt und durch die Anwendung statistischer Verfahren nutzen wir eine Vorgehensweise die sehr ähnlich ist zur *umgekehrten Entwicklung* eines Produktionsverfahrens.

> *Reverse Engineering* (englisch, bedeutet: umgekehrt entwickeln, rekonstruieren, Kürzel: RE; auch Nachkonstruktion) bezeichnet den Vorgang, aus einem bestehenden fertigen System oder einem meistens industriell gefertigten Produkt durch Untersuchung der Strukturen, Zustände und Verhaltensweisen die Konstruktionselemente zu extrahieren. Aus dem fertigen Objekt wird somit wieder ein Plan erstellt.
>
> Wikipedia: `https://de.wikipedia.org/wiki/Reverse_Engineering`

Als empirische ForscherIn haben wir das Ziel etwas über das Verfahren – einen unbekannten Daten-generierenden und in seinen Details nicht direkt beobachtbaren/messbaren Prozess – zu lernen, dadurch dass wir eine Menge von Endprodukten – die Realisation des Beobachtungs-/Messprozesses die wir *Daten* nennen – betrachten. Wird die *Vorhersage* dabei mit hoher Sicherheit als Bestandteil oder kein Bestandteil dieses Prozesses identifiziert, so ist dies Evidenz für oder gegen die aufgestellte *Theorie*.

Auffrischung: Wahrscheinlichkeit und Verteilungen

Weil wir ein paar Begriffe rund um Verteilungen und Wahrscheinlichkeit im Folgenden immer wieder gebrauchen werden, hier nur noch mal ganz kurz und grob ein paar Worte:

- In der Statistik ist eine *Verteilung* eine Menge von Wertepaaren:

$$\{(a_1, p\left(A = a_1\right)); (a_2, p\left(A = a_2\right)); (a_3, p\left(A = a_3\right)); \ldots\}$$

  mit *Ergebniswerten* $a_1, a_2, a_3, \ldots$ einer *Zufallsvariable* A mit ihrer jeweils dazugehörigen *Wahrscheinlichkeit* $p\left(A = a_1\right), p\left(A = a_2\right), p\left(A = a_3\right), \ldots$.

- Weil diese Wahrscheinlichkeiten als Funktionswert einer Funktion p – von engl. *probability* – angegeben werden spricht man von p hier auch als *Wahrscheinlichkeitsfunktion*.

- Die Wahrscheinlichkeit für irgendeinen beliebigen Ergebniswert a_i ist eine positive Zahl $p\left(A = a_i\right) \geq 0$ und über alle Ergebniswerte hinweg 'summieren' sich diese Wahrscheinlichkeiten zum Wert 1 auf.

- Ist die Menge an Ergebniswerten *abzählbar* oder *abzählbar unendlich* groß – wie etwa bei den natürlichen Zahlen $0, 1, 2, 3, \ldots$ –, so spricht man von den Wahrscheinlichkeiten als *Wahrscheinlichkeitsmassen* und das aufsummieren zum Wert 1 kann tatsächlich mit gewöhnlicher

Addition erfolgen.
Ein Beispiel: A kann die Werte a_1, a_2, a_3 annehmen: $A \in \{a_1, a_2, a_3\}$. Dann ist:

$$p\left(A = a_1\right) + p\left(A = a_2\right) + p\left(A = a_3\right) = \sum_{i=1}^{3} p\left(A = a_i\right) = 1$$

- Ist die Menge an Ergebniswerten *überabzählbar* groß – wie etwa bei den reellen Zahlen in einem beliebigen Intervall – so entspricht der Wert der Verteilung für jeden dieser Werte keiner Wahrscheinlichkeitmasse mehr, sondern ist der Wert einer *Dichte von Wahrscheinlichkeit*. Der Wert des *Integrals* dieser Dichte in irgendeinem Teilintervall ist dann aber wieder eine Wahrscheinlichkeitsmasse, und die Summe der Integralwerte der Dichte für alle möglichen, sich wechselseitig nicht überlappenden Teilintervalle, summiert sich dann wieder zum Wert 1 auf.
  Ein Beispiel: A kann die Werte im Intervall von 0 bis 1 annehmen: $A \in \{a; a \in [0, 1]\}$. Wir unterteilen dieses Intervall in drei gleich große Teilintervalle. Dann ist:

$$\int_{0}^{1/3} p\left(A = a\right) \partial a + \int_{1/3}^{2/3} p\left(A = a\right) \partial a + \int_{2/3}^{1} p\left(A = a\right) \partial a = 1$$

Vorwissen

In jeder statistischen Analyse werden wir neben den Messungen gewöhnlich noch durch eine zusätzliche Informationsquelle begleitet in der (vages) Wissen darüber gespeichert ist, was wir über das Herstellungsverfahren – um im obigen *Reverse Engineering* Bild zu bleiben – schon wissen bevor wir das erste Endprodukt gesehen haben.

In der wissenschaftlichen Anwendung ist dieses Vorwissen:

- **Inhaltlicher Natur** Beispielsweise die Kenntnis über den Wertebereich einer unbekannten Größe, oder Intervalle die unterschiedliche Ausmaße von Unsicherheit ausdrücken darüber wie diese Größe in verschiedenen verwandten Studien bisher variierte.
- **'Technischer' Natur** Wir wissen dass ein nützliches Modell auf verlässliche Art und Weise zu Ergebnissen kommt, dass es – um in klassischer (frequentistischer) Sprache zu bleiben – 'konvergiert'.
- **Eine Kombination dieser beiden Arten** Wir wissen dass Prozesse in der Natur auf einer Skala von *verlässlich* bis *chaotisch* eher im Bereich des Verlässlichen stattfinden. Beispielsweise geht eine Vergröss erung um 10% des Volumens des mit den Wurzeln erschlossenen Bodenraums eines Baumes nicht mit einer Steigerung des Stammdurchmessers von mehreren 10 000% einher. Solch eine Einschätzung gilt in jedem forstwissenschaftlichen Prozess für jeden Einfluss einer Größe die auf einer vernünftig Skala beruht[2].

[2] Natürlich sollte sich eine Steigerung des Stammdurchmessers von mehereren 10 000% ergeben bedingt darauf dass sich das Volumen des Wurzelraums ebenfalls um den Faktor 10 000 vergrößert. Diese Aussage hängt also immer davon ab dass die Skala einer natürlich vorkommenden Variation entspricht.

Ein gewichtiger Vorteil der probabilistischen bzw. bayesianischen Herangehensweise an statistische Modellierung ist dass wir hier technisch in der Lage sind über das Konzept der *Priori* mittels Wahrscheinlichkeitsverteilungen dieses Vorwissen in der Modellierung mit einzubeziehen[3]:

> Prior distributions provide the opportunity to formally include previous knowledge in a statistical analysis. If available, it is usually a good idea to use such previous knowledge. In almost every study, there exists some information, at least about the range of reasonable parameter values. If this information is used to construct the prior, the parameter estimates are kept within the range of reasonable values. In addition, such informative priors help making MCMC algorithms stable. *Weakly informative prior distributions* are priors that are constructed based on the range of reasonable parameter values, but without compiling information from existing studies. These priors contain some information, but obviously much less than what would be possible if the relevant information from the literature would be gathered (Gelman et al., 2014).
>
> Korner-Nievergelt et al. (2015, S. 266)

Hier die Stelle aus aus Gelman et al. (2014, S. 51) auf die Korner-Nievergelt et al. hier referenzierten:

> [...] the weakly informative prior distribution, which contains some information – enough to 'regularize' the posterior distribution, that is, to keep it roughly within reasonable bounds – but without attempting to fully capture one's scientific knowledge about the underlying parameter.
>
> Gelman et al. (2014, S. 51)

Weiterhin:

> Rather than trying to model complete ignorance, we prefer in most problems to use weakly informative prior distributions that include a small amount of real-world information, enough to ensure that the posterior distribution makes sense.
>
> Gelman et al. (2014, S. 55)

Mit dem Konzept der *posterior distribution* wird hier diejenige Größe aufgeführt mittels der wir unser Vorwissen mit der neu hinzugewonnenen Evidenz aus den Messungen kombiniert haben.

Beispiel (Erwarteter BHD im nördlichen Kalifornien) Dieses Beispiel wird in Stauffer (2007, Seite 56) geschildert. Das Ziel ist hier die Schätzung des erwarteten Brusthöhendurchmessers (BHD) eines 100-jährigen, gleichaltrigen Gemeindewaldes einer Kleinstadt im nördlichen Kalifornien (USA), welcher mit Küstenmammutbäumen und Douglasien bestockt ist. Die BHD-Werte variieren hier in der Regel zwischen 20 und 40 Inches ("[...] diameters generally ranging between 20 and 40 in."), und vorläufige Schätzer des mittleren BHD ergaben ungefähr den Wert 30 Inches ("[...,] with

[3] Ich will hier nicht verschweigen dass es natürlich auch in der klassischen Statistik eine Vielzahl an Konzepten und Methoden gibt um dies zu tun, wie etwa durch *penalisierte Splines*, *random effects*, *Regularisierung*, *Variablenselektion*, ... All diese Konzepte beruhen auf Vorinformation: Ein nichtlinearer Effekt sollte nicht zu wild verlaufen, die Menge an Koeffizienten sollte eine Tendenz zum Zentrum bezüglich der Population haben, Koeffizienten müssen eventuell stabilisiert werden damit die Ergebnisse in der Anwendung sinnvoll sind, der Kern des wahren Prozesses hängt womöglich nur 'von einer Hand voll' Effekten ab, ... All diese Konzepte haben jedoch auch eine Entsprechung in der bayesianischen Herangehensweise und eine Vielzahl davon lässt sich geschlossen mittels des Priori-Konzepts formulieren.

preliminary estimates of a mean around 30 in."). Weiterhin wird der Stadtförster hinzugezogen, welcher aussagt, dass kein BHD größer als 60 Inches ist. Bei keiner Waldinventur dieser Stadt wurde in den Vorjahren ein BHD größer 60 Inches gemessen, und es gibt keine sehr alten Bäume. Das Vorwissen für den erwarteten BHD wird hier nun in der Form einer Wahrscheinlichkeitsdichtefunktion kodiert, es wird eine stetige Gleichverteilung über dem Intervall von 0 bis 60 Inches angenommen. Diese Wahl ist komisch, da die zentrale Tendenz im Bereich zwischen 20 und 40 Inches 'vergessen' wird. Weiterhin sind Gleichverteilungen als Kodierung des Vorwissens parktisch nicht gut geeignet, da keine noch starke Evidenz aus den Daten uns jemals von einem Wert außerhalb des gleichverteilten Intervalls überzeugen könnte[4].

Ein eigenes Beispiel In einem bisher ununtersuchten 100 Jahre alten, nordwestdeutschen Mischbestand unbekannter Baumarten soll die Oberhöhe[5] des Bestandes geschätzt werden. Um diese Aufgabe tätigen zu können müssen Baumhöhen gemessen werden. Jedoch haben wir schon vor der ersten Messung (viele) Informationen über die Baumhöhen des unbekannten Bestandes, dadurch dass wir wissen:

1. dass eine Baumhöhe generell eine Größe ist die durch eine reellwertige und positive Zahl beschrieben wird,
2. dass die Höhen der Bäume in Nordwestdeutschland wohl ziemlich sicher[6] kleiner sind als die Höhe des aktuell bekannten höchsten Baumes der Welt.

Kenne ich darüber hinaus (3.) noch einen Referenzbestand in der Nähe meines unbekannten Bestandes und weiß (4.) weiterhin z.B. dass dieser etwas jünger oder etwas älter ist, dann kann ich durch mit dem Nebenwissen (5.) dass die Höhe eines Einzelbaumes als monoton steigende Funktion über dem Alter beschrieben werden kann, mit einiger Sicherheit – aber auch inhaltlich relevanter Unsicherheit! – die Baumhöhen meines Bestandes weiter eingrenzen. Kenne ich dann auch noch Unterschiede bezüglich des Standorts (6.), dann ....

Die Art, Menge und Präzision der Vorinformationen variiert also mit der Person die diese formuliert:

> A probability of an event or of the truth of a statement is a number between 0 and 1 that quantifies a particular person's subjective opinion as to how likely that event is to occur (or to have already occurred) or how likely the statement is to be true.
>
> Cowles (2013, Seite 4: Definition *„subjective interpretation of probability"*)

Ich, der Autor dieses Textes, mache eine Google-Bildersuche für `nw fva höhenkurve` und wähle das vierte Bild[7]. Ich erhalte die Beschreibung eines Naturwaldes in Niedersachsen. Darin steht:

> Die Altersstruktur des Naturwaldes ist recht einheitlich. Rund 150-jährige Stieleichen und eine vollflächig vorhandene Unter-und Zwischenschicht aus ca. 100-jährigen Hainbuchen bestimmen das Waldbild (Abb.3). Einzelne ebenfalls rund 100-jährige Buchen

[4] Siehe hierzu auch *Cromwell's Rule* in Lindley (2006, S. 90).
[5] Die zu erwartende Baumhöhe der 100 höchsten Bäume pro Hektar.
[6] Hier drücke ich Unsicherheit aus!
[7] Dieses verweist auf den folgenden Link: `https://www.nw-fva.de/NwInfo/pdf?nwf_nr=57`.

> sind zusammen mit Eschen eingemischt.
>
> Meyer et al. (2006)

Diese Beschreibung ist aus dem Jahr 2006 und die Abbildung 6 ist beschriftet mit *„Höhenkurven der wichtigsten Baumarten der Kernfläche 1985“*. Weiterhin ist in diesem Dokument zu lesen:

> Die Böden sind mehrschichtig aufgebaut und bestehen aus Feinsanden und verlehmten Geschiebesanden. Das für die Baumwurzeln leicht erreichbare Grundwasser und eine eutrophe Nährstoffversorgung schaffen günstige Wachstumsbedingungen.
>
> Meyer et al. (2006)

In Abbildung 6 (Meyer et al., 2006) sind Höhenkurven (x-Achse: Brusthöhendurchmesser, y-Achse: Baumhöhe) für die Baumarten Buche, Hainbuche, Eiche und Esche dargestellt. Die vier maximalen Höhen liegen dabei im Intervall zwischen[8] 25m und 33m. Ich gehe davon aus, dass die abgebildeten Kurven den jeweiligen geschätzten bedingten Erwartungswert aus einem nicht-linearen Regressionsmodell mit Normalverteilungsannahme (*non-linear least squares*) an die bedingte Höhe darstellen. Nun weiß ich ja nicht wie gut die Wuchsbedingungen in meinem unbekannten Bestand sind, wohl nicht viel besser, aber es ist auch unrealistisch anzunehmen dass die Bäume dort viel(!) weniger als halb so hoch wachsen können, da in Mitteleuropa wohl nur dort Waldbau betrieben wird wo dieser auch halbwegs Sinn macht!?! Ich treffe daher eine – meine! – aktuelle subjektive Einschätzung als:

> Baumhöhen im Alter 100 Jahre auf der mir unbekannten Fläche sind positive Zahlen mit Einheit *Meter* die mit einer Wahrscheinlichkeit von 0.95 zwischen ungefähr 10m und 50m liegen, Werte höher 33m – das war der maximale bedingte Erwartungswert im Naturwald Brand auf gutem Standort – sollten mit einer Wahrscheinlichkeit von ungefähr $1/4$ auftreten, richtig extrem hohe Werte wie etwa 100m sollten nur mit einer verschwinden kleinen Wahrscheinlichkeit von deutlich kleiner als 0.01 auftreten.

Eine solche Formulierung ist subjektiv, jemand Anderes wäre anders vorgegangen. Jedoch ist kein Vorwissen komplett richtig, und kein Vorwissen ist komplett falsch – solange Grundsätze guter wissenschaftlicher Praxis, wie der Ausschluss bewusster Falschaussage, gewahrt bleiben. Meine Formulierung beruht auf den Annahmen die ich oben beschrieben habe und auf der Vorgehensweise mit der ich diese und die weiteren Informationen miteinander verwoben habe. Sie ist mit hoher Sicherheit nicht ideal Gelman et al. (2017), aber die Vorgehensweise ist transparent und reproduzierbar beschrieben, steht einer offenen Diskussion zur Verfügung, kann und sollte demzufolge korrigiert werden sobald neue Informationen hinzukommen die auf mögliche Verbesserungen hinweisen, und erfüllt damit die Grundlagen des wissenschaftlichen Arbeitens.

Modell

Jedes statistische Verfahren basiert darauf dass wir ein Modell für den unbekannten, Daten-generierenden Mechanismus in mathematischer Notation formulieren, z.B.:

$$\log\left(Y\right) \sim \mathrm{N}\left(\mu, \sigma^2\right), \tag{1}$$

[8] Von mir visuell abgelesen und auf ganze Meter gerundet.

in Worten: *eine Beobachtung einer mit der Logarithmusfunktion transformierten Zufallsvariable Y ist eine Realisierung eines normalverteilten Zufallsprozesses mit Erwartungswertparameter μ und Varianzparameter σ^2.*

> […] modeling, that is, the ability to build up a probabilistic interpretation of an observed phenomenon […]. […] This means picking a parameterized probability distribution, denoted by f_θ, and extracting information about (shortened in "estimating") the unknown parameter θ of this probability distribution in order to provide a convincing interpretation of the reasons that led to the phenomenon at the basis of the dataset (and/or to be able to draw predictions about upcoming phenomena of the same nature). […] a model is an interpretation of a real phenomenon that fits its characteristics up to some degree of approximation rather than an explanation that would require the model to be "true". In short, there is no such thing as a "true model", even though some models are more appropriate than others!
>
> Marin and Robert (2013)

Ich nehme weiterhin an dass für eine log-Normalverteilung als Modell an den Daten-generierenden Mechanismus diese Aussage erfüllt sein soll, also $\log(\text{Baumhöhe [m]}) \sim \text{Normal}\,(\mu, \sigma^2)$.

Vorwissen als Modell Wird das Vorwissen durch Wahrscheinlichkeits(dichte)funktionen 'kodiert', so spricht man hier von erzeugendem (engl. *generative*) Vorwissen in dem Sinn das man aus dem kodierten Vorwissen heraus bereits Daten erzeugen bzw. simulieren kann die noch keine Evidenz aus den wirklichen Daten enthält. Mit solch einer Kodierung hat man ein Werkzeug in der Hand welches Auskunft darüber gibt wie realistisch oder unrealistisch die Kodierung selbst war:

> […] we can visualize simulations from the prior marginal distribution of the data to assess the consistency of the chosen priors with domain knowledge.

Wie finden wir nun aber nun nützliche Verteilungen und Parameter für solch ein erzeugendes Priori-Modell? Korner-Nievergelt et al. beschreiben mit einem Beispiel einen prinzipiellen Vorgang:

> For example, if we need a prior for a probability […], and we choose a beta distribution, what values should we use for the parameters a and b? If the prior knowledge can be expressed as an estimate and an uncertainty measure, say 0.05 $\pm$ 0.02 (estimate $\pm$ standard error), then the estimate and the square of the standard error (1/4 variance of the parameter) can be inserted in the formulas for the mean and variance of the beta distribution (see box), and the equations can be solved for a and b. This is done by the function `shapeparameter` from the package `carcass`. In other cases, the "trial and erroräpproach is also valuable. The functions `qgamma` , `qbinom`, and `qnorm` are helpful for this purpose.
>
> Korner-Nievergelt et al. (2015, S.)

Ich folge hier auch – mit der Hilfe geeigneter Grafiken – eher der 'Trial and Error'1-Strategie und finde, nach etwas Herumprobieren in R, heraus dass für Baumhöhen auf der Skala $\log(\text{Meter})$ für eine Normalverteilung mit Erwartungswertparameter $\mu = 3.2$ und Standardabweichung $\sigma = 0.43$ meine obere Aussage zu meinem Vorwissen ganz gut erfüllt ist. Weil ich aber – durch die Natur der Sache

– unsicher bin in dieser Aussage, formuliere ich für beide Parameter ebenfalls subjektive Wahrscheinlichkeitsverteilungen:

$$\mu \sim \text{Normal}\left(3.2, 0.1^2\right) \tag{2}$$

und

$$\log(\sigma) \sim \text{Normal}\left(\log(0.43), 0.1^2\right) \tag{3}$$

Mein auf diese Art und Weise formuliertes Vorwissen erzeugt dann durch folgenden R-Code die in Abbildung 1 auf Seite 11 dargestellte prädiktiven Priori-Wahrscheinlichkeitsverteilungen:

```
R <- 20
set.seed(123)
m <- rnorm(n = R, mean = 3.2, sd = 0.1)
s <- rnorm(n = R, mean = log(0.43), sd = 0.1)
par(mfrow = c(1, 1))
q99 <- exp(qnorm(p = 0.999, mean = m, sd = exp(s)))
plot(c(1, R), c(0, max(q99)), las = 1, bty = "n", xaxt = "n",
     type = "n", xlab = "Wiederholung", ylab = "Baumhöhe [m]")
axis(1, at = c(1, 10, 20))
abline(h = c(10, 33, 55, 90),lty = 2, col = rgb(0.4, 0.7, 0.4), lwd = 2)
for (i in 1:20) {
  lines(rep(i, 2), exp(qnorm(p = c(0.025, 0.975), mean = m[i], sd = exp(s[i]))),
        col = 1)
  points(i, q99[i], pch = 16, col = 1)
  lines(i + 0.2*c(-1, 1), rep(exp(qnorm(p = 0.75, mean = m[i], sd = exp(s[i]))), 2))
}
```

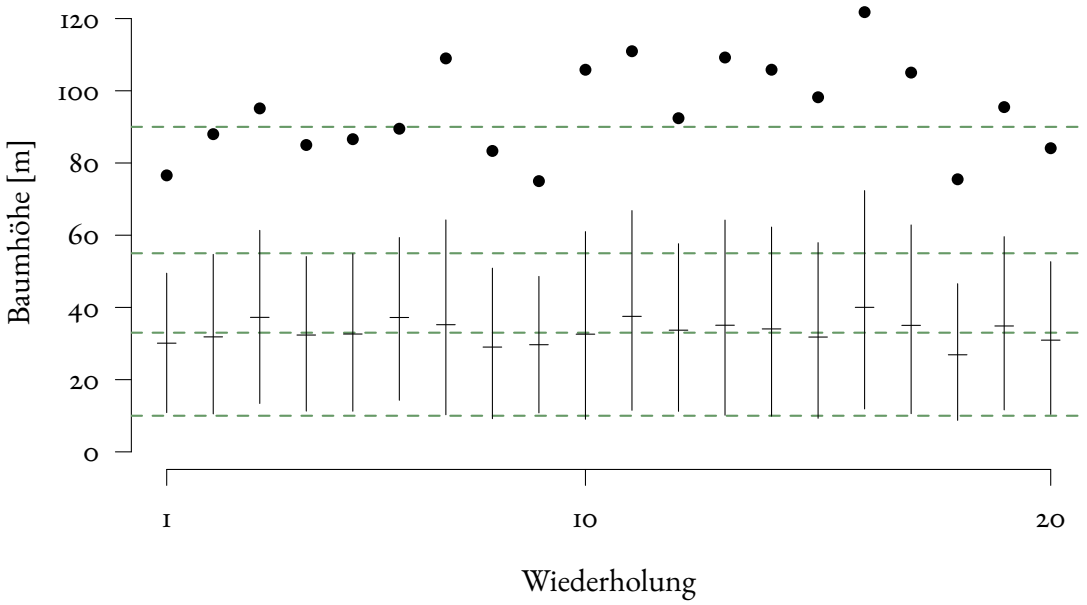

Abbildung 1: 20 Ziehungen aus den in den Gleichungen 2 (Seite 10) und 3 (Seite 10) formulierten Modellen für unser Vorwissen für die Parameter μ und σ führen zu den hier abgebildeten prädiktiven Priori-Verteilungen die durch ihre Quantile (99%-Quantil: gefüllter Punkt, 75% Quantil: kurzer horizontaler Strich; 2.5% bis 97.5% Quantil: vertikale Linie) mit dem verbalisierten Vorwissen abgeglichen werden können.

Aktualisierung des Vorwissens durch Messungen

Indem wir eine Menge y von Messungen unserer Zufallsvariablen Y erzeugen[9] erhalten wir eine Grundlage dafür, unser Vorwissen anhand der Evidenz durch diese neuen Messungen zu aktualisieren.

> Mit den Messungen selbst, einem *Modell für den Messprozess*, welches den Mechanismus zur Erzeugung dieser Messungen beschreibt, und einem *Modell für das Vorwissen*, welches das Vorwissen über die Parameter des Modells für den Messprozess beschreibt, haben wir alle Bestandteile vorliegen um in der Anwendung ein *bayesianisches Modell* zu definieren.

Für jeden Parameter folgt aus diesem bayesianisches Modell eine Aktualisierung in Form einer Punktschätzung sowie einer Quantifizierung der Unsicherheit dieser Punktschätzung, idealerweise in der Form einer kompletten Wahrscheinlichkeitsverteilung, genannt *Posteriori-Verteilung*.

Bayesianische Ansätze, die zu eben solch einer Wahrscheinlichkeitsverteilung für jeden Parameter des Modells führen, basieren, wie gerade beschrieben, auf der Verwendung eines vollständigen Wahrscheinlichkeitsmodells – das von einem oder mehreren unbekannten Parametern θ abhängt –, welches nicht nur unsere Unsicherheit bezüglich der auftretenden Werte einer Zielvariablen Y beschreibt (*'Beobachtungsmodell'*), sondern auch unsere *a priori* Formulierung unserer Vorinformationen (*'Vorwissenmodell'*) – kurz nur *Priori* – über die Parameter θ selbst. Im oberen Beispiel zur Schätzung der Baumhöhenverteilung eines 100-jährigen unbekannten Bestandes ist θ ein Vektor mit den Elementen μ und σ, also $\theta = (\mu, \sigma)^\top$. Das Ziel besteht nun darin, die Formulierung unserer Information über die Parameter θ – z. B. können das auch die Koeffizienten eines Regressionsmodells sein – unter Verwendung unseres Modells $\mathcal{M}$ und unserer Daten zu aktualisieren. Dies mündet in der *a posteriori* Formulierung der Information – kurz nur *Posteriori* – bezüglich θ als Kombination des Vorwissens mit der Evidenz aus dem Modell für die unbekannte Größe, ausgewertet an den gemessenen Werten. Sowohl die Priori als auch die Posteriori werden dabei als Wahrscheinlichkeitsverteilungen formuliert.

Die Beziehung zwischen der Priori bezüglich θ, also vor dem Erhalt der Daten, und unserer Posteriori bezüglich θ, also nachdem die Priori durch die Beobachtungen aktualisiert wurde, wird durch den *Satz von Bayes* beschrieben:

$$p(\theta \mid y) = \frac{p(\theta) \cdot p(y \mid \theta)}{p(y)} \tag{4}$$

der besagt, dass die posteriori Wahrscheinlichkeitsverteilung $p(\theta \mid y)$ des Parameters θ gegeben die Beobachtungen y dem Produkt einer Wahrscheinlichkeitsfunktion $p(y \mid \theta)$ (*'Likelihood'*) und der priori Wahrscheinlichkeitsverteilung $p(\theta)$, geteilt durch die Randverteilung der Daten $p(y)$, entspricht. Der Nenner, die Randverteilung der Daten $p(y)$, ist eine *Normalisierungskonstante* mit dem Zweck dass die linke Seite von Gleichung 4 einer notwendigen Eigenschaft für Wahrscheinlichkeits(dichte)funktionen genügt, und zwar sich zum Wert 1 integrieren zu lassen. $p(y)$ ergibt sich daraus, dass man die Wahrscheinlichkeitsfunktion $p(y \mid \theta)$ durch Integration über θ 'gewichtet aufsum-

[9] Diese Messungen von Y werden üblicherweise statt kurz y etwas präziser mit $y_1, \ldots, y_n$ bezeichnet, wobei n der *Stichprobenumfang* ist. Wenn unser statistisches Modell aus der Gruppe der *Regressionsmodelle* kommt, dann haben wir neben der Zielvariablen Y auch Einflussgrößen mit der Bezeichnung $X_1, \ldots, X_p$ mit Messungen $x_{1,1}, \ldots, x_{p,n}$.

miert', mit Gewichten definiert durch die priori Wahrscheinlichkeit $p(\theta)$ für den jeweiligen Wert von θ.

> The normalizing constant can be tricky. It is supposed to be the probability of seeing the data under any hypothesis at all, but in the most general case it is hard to nail down what that means.
>
> Downey (2013, S.)

Der Nenner ist also so etwas wie das 'Potenzial' des Zählers, hängt durch die Integration über θ nicht mehr von θ ab[10] bzw. ist konstant bezüglich θ und transportiert daher keine Informationen mehr über θ in Bezug darauf welche Werte von θ mehr oder weniger plausibel sind. Um unser Wissen über θ basierend auf den Daten y zu aktualisieren, müssen wir uns daher nur auf den Zähler $p(\theta) \cdot p(y \mid \theta)$ konzentrieren. Es ist aus diesem Grund üblich, die Gleichung 4 in der folgenden Äquivalenzform anzugeben:

$$p(\theta \mid y) \propto p(\theta) \cdot p(y \mid \theta), \tag{5}$$

wobei $\propto$ eine Proportionalität beschreibt. In allen Fällen hängt die Information – sowohl Priori als auch Posteriori – vom explizit gewählten Modell $\mathcal{M}$ ab. Daher:

$$p(\theta \mid y, \mathcal{M}) \propto p(\theta \mid \mathcal{M}) \cdot p(y \mid \theta, \mathcal{M}). \tag{6}$$

Für Regressionsmodelle könnten wir in unserer Aussage zum Satz von Bayes noch expliziter sein, indem wir die Beobachtungen x der Einflussgrößen X hier mit einbeziehen:

$$p(\theta \mid y, x, \mathcal{M}) \propto p(y \mid \theta, x)\, p(\theta \mid x), \tag{7}$$

was aber eigentlich nicht notwendig ist, siehe folgendes Beispiel: In einem ersten Modell beziehen wir die Werte x eine Einflussgröße X unverändert, in einem zweiten Modell um einen Wert c zentriert ein, also $x - c$. In der Regel wird sich damit unser Vorwissen über den Modellparameter verändern der festlegt, wo der zu erwartende Wert von y an der Stelle $x = 0$ liegt[11]. Wir haben also durch verschiedene Varianten von x verschiedene Modelle $\mathcal{M}$ und die Varianten von x sind damit ein Bestandteil der jeweiligen Modelle. Daher reicht die Formulierung in Gleichung 6, es kann aber in der ein oder anderen Situation sinnvoll sein x explizit in der Notation zu erwähnen.

[10] Man sagt diese sind hier nun *heraus integriert worden*, was in der umgangssprachlichen Bedeutung dieser Worte wohl gar keinen Sinn ergibt!?

[11] Dieser Parameter wird in der Regressionsmodellierung *Intercept* genannt. Eine solche lineare Verschiebung des Intercept wie hier beschrieben kann durch eine äquivalente Verschiebung in der Formulierung des Priori-Erwartungswertes – man addiert hier c mal den Priori-Erwartungswert des Steigungskoeffizienten von x zum bisherigen Priori-Erwartungswert – so ausgeglichen werden dass sich das Modell als Gesamtes 'auf dem Papier' nicht ändert, jedoch sollte man hier nicht vergessen dass es durchaus zu einer Änderung des Resultats führen kann da wir in der Regel auf Rechenalgorithmen angewiesen sind die in verschiedenen Bereichen reelle Zahlen auf unterschiedliche Weise konstruieren und verarbeiten. Es kann also praktisch durchaus hier zu leicht unterschiedlichen Modellen kommen – ich erinner mich noch gut an die Anpassung von gemischten Modellen für binomialverteile Zielvariablen mit Logit-Linkfunktion mittels `lme4::glmer` die konvergierten oder nicht konvergierten, je nach dem ob die Einflussgrößen sinnvoll skaliert waren.

Likelihood, Priori, Posteriori

Die oben eingeführten Komponenten des Satz von Bayes sind die grundlegenden mathematischen Objekte, mit denen wir vertraut sein sollten, wenn wir bayesianische Inferenz durchführen:

- Die (Wahrscheinlichkeits-)Funktion[12] $p\left(y \mid \theta, \mathcal{M}\right)$ ist eine Größe, die für ein festgelegtes Modell $\mathcal{M}$ die gemeinsame Wahrscheinlichkeit der Daten y berechnet, bedingt auf jeweils festgelegte, aber alle möglichen Werte des Parameters θ. Sie wird auch als *Modell für den Messprozess* oder *Likelihoodfunktion* bezeichnet. Für jede potentielle Messung y_i repräsentiert $p\left(y_i \mid \theta, \mathcal{M}\right)$ den Daten-erzeugenden Prozess von dem es als plausibel angenommen wird, dass er diese Messung erzeugt haben könnte.

  Wenn wir in $p\left(y \mid \theta, \mathcal{M}\right)$ den/die Wert(e) des/der Parameter θ festhalten, würde das aus praktischer Sicht bedeuten, dass eine jede weitere zusätzliche – hypothetische – Messung nicht mehr die Vorhersage jeder weiteren, darüber hinaus hinzukommenden potenziellen Messung verbessern könnte – die Parameter sind fest, und damit ist auch der Daten-generierende Mechanismus unveränderlich und komplett definiert. In diesem Fall spricht man davon dass die Messungen in Bezug auf die Parameter θ bedingt unabhängig sind. Die Likelihoodfunktion könnte dann interpretiert werden als eine Wahrscheinlichkeit, die aus dem Produkt der Wahrscheinlichkeitsbeiträge $p\left(y_i \mid \theta, \mathcal{M}\right)$ jeder einzelnen Messung y_i folgt.

  Halten wir die Daten fest und variieren hingegen die Werte der Parameter, so integriert sich die Likelihoodfunktion in Bezug auf die Parameter jedoch nicht mehr zum Wert 1 auf, und ist daher keine Wahrscheinlichkeits(dichte)funktion.

- Die Priori $p\left(\theta \mid \mathcal{M}\right)$ ist eine Wahrscheinlichkeitsverteilung, die unser Vorwissen oder auch unsere Unsicherheit über die Parameter θ des Modells für den Messprozess – vor Einbeziehung der Messungen y – beschreibt. In der Regel verfügen wir in jeder praktischen Anwendung über mindestens ein Mindestmaß an Vorinformationen. Wenn beispielsweise in einem Regressionsmodell alle Variablen – Zielgröße und Einflussgrößen – zueinander in einem angemessenen Verhältnis skaliert sind, wissen wir im Allgemeinen bereits vor der Realisierung einer ersten Messung, dass wir keine extremen Schätzungen der Regressionskoeffizienten zu erwarten haben.

  Zur Formulierung der Priori können bestehende Forschungsergebnisse eine wertvolle Quelle sein, die dann zur Entwicklung einer sogenannten informativen – und damit auch potentiell selbst Daten-erzeugenden (*generativen*) – Priori verwendet werden können.

- Die Posteriori $p\left(\theta \mid y, x\right)$ ist das Ziel: eine gemeinsame Wahrscheinlichkeitsverteilung aller Parameter θ, die unser aktualisiertes Vorwissen über die Parameter widerspiegelt, dadurch, dass wir y gemessen und das Modell $\mathcal{M}$ angenommen haben. Die Posteriori kann als Kompromiss zwischen der Likelihoodfunktion und der Priori betrachtet werden und beschreibt die relative Plausibilität aller vom Modell abhängigen Parameterwerte. Eine auf saubere und sinnvolle Weise angepasste Posteriori ermöglicht uns vielfältige, intuitive und flexible statistische Inferenzaussagen.

[12] Der Begriff 'Wahrscheinlichkeits-' steht hier in Klammern da $p\left(y \mid \theta, \mathcal{M}\right)$ nur in einer von zwei möglichen Interpretation die Bedingungen an Wahrscheinlichkeiten erfüllt. Die Begründung folgt gleich in diesem Abschnitt.

Bayesianismus und Frequentismus

Ein grundlegender Unterschied zwischen dem bayesianischen und dem frequentistischen Zugang zur statistische Inferenz besteht darin, dass angenommen wird, dass unbekannte Größen einen festen wahren Wert haben, oder ob wir die Unsicherheit über diese unbekannten Größen mittels der Verwendung der Wahrscheinlichkeitstheorie beschreiben dürfen.

Der Frequentismus behandelt Parameter als unbekannte Punkte, der Bayesianismus als Zufallsvariablen. Ein einfaches Beispiel: Es gibt eine Analogie zwischen beobachteten Häufigkeiten und Wahrscheinlichkeit. Werfe ich wiederholt einen sechsseitigen Würfel, so kann ich die Ergebnisse daraus in einer Häufigkeitstabelle zusammenfassen. Diese Häufigkeiten geteilt durch die Anzahl der Würfe ergibt den relativen, beobachteten Anteil eines jeden Ergebnisses. Diese relativen Anteile sind so etwas wie die bestmöglichste Schätzung der Wahrscheinlichkeiten für das Ergebnis des nächsten Wurfes. Nach jedem weiteren Wurf werden sich diese Anteile ändern und wenn ich immer und immer wieder meinen Würfel werfe, so wird das doch nie aufhören: sogar nach einer Trilliarde, Siebenhundert und Drei und Zwanzig Würfen werden meine relativen Anteile nach dem nächsten Wurf wieder anders sein. Es herrscht also *praktische Unsicherheit* bis ans Ende unserer Tage.

Der frequentistische Ansatz befasst sich mit der Wahrscheinlichkeit der beobachteten Daten bedingt auf festgezurrte Parameter – d.h. θ ist keine Zufallsvariable und hat demzufolge keine Wahrscheinlichkeitsverteilung. Die frequentistische Inferenz bezieht sich in ihrer Aussage immer auf fiktive, unendlich häufig und unabhängig durchgeführte Wiederholungen einer real durchgeführten Studie. Sie basiert dabei auf der Stichprobentheorie als Grundlage: die vorliegende Stichprobe ist nur ein Element aus einem unbekannten, unendlich großen, und auch variabel definierbaren Stichprobenmenge. So basiert beispielsweise die Definition des p-Wertes – eines der in der Praxis meistgenutzten Konzepte zur Kommunikation frequentistischer Inferenzaussagen – bezüglich des Tests einer Nullhypothese auf einer fiktiven und unendlich häufig durchgeführten Wiederholung der vorliegenden Studie. Für all diese Wiederholungen wird im Weiteren angenommen, dass diese aus einer formbaren Wirklichkeit stammen in welcher man festlegen konnte, dass dort die Nullhypothese immer wahr ist. Der p-Wert gibt dann die relative Häufigkeit (engl. *frequency*) an, wie viele dieser fiktiven Wiederholungen zu einem Ergebnis führten welches – im Vergleich zur einzigen wirklich durchgeführten Studie – gleich oder deutlicher in Richtung der Alternativhypothese zeigen (Held and Bové, 2014, S. 70).

Im Gegensatz dazu basiert die bayesianische Inferenz auf der Kombination der Formulierung des Vorwissens mit einzig den tatsächlich realisierten Messungen: bayesianische Aussagen beziehen sich auf die Wahrscheinlichkeitsverteilung der Parameter als Zufallsvariablen im Angesicht dieser festen, beobachteten Daten.

Eine der wichtigsten Implikationen dieser Unterscheidung ist, dass der frequentistische Ansatz zwar Punktschätzungen – und manchmal Standardfehler oder Konfidenzintervalle – auf der Grundlage asymptotischer Eigenschaften von Schätzern liefern kann, jedoch keine Wahrscheinlichkeitsaussagen über Parameter zulässt. Mittels frequentistischer Methoden können wir Fragen wie *'Wie hoch ist die Wahrscheinlichkeit, dass der Parameter θ in einem bestimmten Intervall liegt?'* nicht beantworten. So lässt beispielsweise auch ein Konfidenzintervall solch eine Interpretation nicht zu[13], wir können zu

[13] 'A X% confidence interval for a parameter θ is an interval (L, U) generated by an algorithm that in repeated sampling has an X% probability of containing the true value of θ.' (Neyman, 1937)

keinem Punkt innerhalb oder auch außerhalb des Intervalls eine Aussage darüber tätigen ob dieser Punkt eine höhere oder niedrigere Plausibiltät hat in der Nähe des unbekannten wahren Parameters zu liegen (Morey et al., 2016, Fallacy 3).

Unter Verwendung der Wahrscheinlichkeitsmethode ist dagegen die Beschreibung der Unsicherheit bezüglich eines Parameters θ für die bayesianische Inferenz von grundlegender Bedeutung. Anstelle von Punktschätzungen erhalten wir für jeden Parameter θ posteriori Wahrscheinlichkeitsverteilung über alle möglichen Parameterwerte die vom Modell und den beobachteten Daten abhängig ist. Durch Möglichkeit mittels moderner Verfahren und Algorithmen zufällige Ziehungen aus den posteriori-Verteilungen zu generieren können wir leicht Aussagen über interessierende Parameter (und Funktionen von Parametern) basierend auf Wahrscheinlichkeiten treffen, einschließlich der Wahrscheinlichkeit darüber, ob ein Parameterwert in einem bestimmten Intervall liegt.

Anwendung: Höhenverteilung von Fichten ähnlichen Durchmessers

Dieses Beispiel basiert auf klassischer, analytischer bayesianischer Inferenz wie diese betrieben wurde als Computer noch lange nicht so leistungsfähig waren wie diese heute sind. Die Herleitung der Posteriori wird in dieser analytischen Form schon in einfachen Anwendungen mathematisch ziemlich komplex. Heute müssen wir unsere praktischen Arbeit dank sehr leistungsfähiger Algorithmen und leistungsfähigen Rechnern nicht mehr auf solchen Herleitungen basieren, die dann schon für mittelkomplexe Situationen gar nicht mehr mathematisch gelöst werden können. Aber für die Intuition hinter den Begriffen Likelihood, Priori und Posteriori ergeben sich in dieser analytischen Form doch auch sehr wertvolle Inhalte.

In diesem angewandten Beispiel interessieren wir uns für die erwartete Baumhöhe von Fichten eines 116-jährigen Bestandes im Solling. Für einer 0.16ha großen Probefläche dieses Bestandes liegen uns dabei die Brusthöhendurchmesser vor (von Gadow, 2003, S. 116). Wir interessieren uns im Weiteren nur für das Teilkollektiv welches die Zielstärke von 45cm bereits erreicht oder überschritten hat.

```
## Von Gadow, Waldstruktur und Waldwachstum, Tabelle 4-1 (S. 116)
## http://www.iww.forst.uni-goettingen.de/doc/kgadow/lit/kvgwwbuch2003.pdf
bhd <- c(41, 41, 38, 53, 44, 42, 50, 43, 40, 44, 40, 33, 39, 32, 49, 47,
         38, 40, 37, 34, 47, 37, 41, 38, 38, 43, 40, 42, 34, 39, 41, 44,
         41, 45, 43, 36, 36, 46, 46, 34, 50)
## Mit erreichter Zielstärke von BHD >= 45cm:
(bhd <- bhd[bhd >= 45])
# [1] 53 50 49 47 47 45 46 46 50
(n <- length(bhd))
# [1] 9
```

Für dieses Teilkollektiv können wir hier näherungsweise annehmen, dass die Variabilität der Baumhöhen nicht durch die Veränderung der erwarteten Baumhöhe, bedingt auf den BHD, bestimmt wird (von Gadow, 2003, Abb. 4-8.) Die Baumhöhenmessungen liegen uns, entgegen der BHD-Messungen, jedoch aus von Gadow (2003) nicht vor. Um hier dieses Anschauungsbeispiel dennoch fortsetzen zu können 'erfinden' wir uns künstlich einen Daten-generierenden Prozess.

Simulation Wir wollen mit der Normalverteilung Daten erzeugen Aus Albert (2000, Tab. 3) erhalten wir Informationen darüber, wie die Höhenvariabilität mit der empirischen Variabiltät der BHD-Messungen in Bezug steht. Dies basiert auf der Standardabweichung σ einer Normalverteilung mit der erwarteten Baumhöhe bedingt auf den BHD als Erwartungswert. Wir können damit für die Standardabweichung der unbekannten Höhen Y unserer 9 Fichten eine Standardabweichung der Baumhöhen ableiten:

```
sd(bhd)
# [1] 2.571208
```

```
## Albert, 2000, Tab. 3
(sd_h <- 0.144 + 0.555 * sd(bhd))
# [1] 1.57102
```

Wir lesen weiterhin visuell aus von Gadow (2003, Abb. 4-8.) eine zu erwartende Baumhöhe von 34.5m in der Mitte unseres BHD-Intervalls $\geq$ 45cm ab. Mit diesen beiden Werten, $\mu = 34.5$ und $\sigma = 1.571 =$ `sd_h`, haben wir alle Teile unseres Simulations-Prozesses vollständig festgelegt und können uns künstlich neun Baumhöhenmessungen erzeugen (gerundet auf die erste Nachkommastelle):

```
set.seed(123456789)
y <- rnorm(n = length(bhd), mean = 34.5, sd = sd_h)
(y <- round(y, 1))
# [1] 35.3 35.1 36.7 33.4 33.5 32.0 34.7 34.3 32.1
```

Nehmen wir nun an diese Simulation wäre nie passiert, wir löschen also die beiden eingesetzten Parameterwerte $\mu = 34.5$ und $\sigma = 1.571 =$ `sd_h` aus unserem Gedächtnis. Wir vergessen ebenso dass eine Normalverteilung dieser Simulation zugrunde lag.

Statistisches Modell Wir nehmen als Grundlage für den unbekannten Daten-generierenden Prozess an, dass die Baumhöhenmessungen dieses Kollektivs einer Normalverteilung entstammen – das ist der erste Teil unseres statistischen Modells $\mathcal{M}$. Die Dichte dieser Verteilung verkörpert damit unser Modell an den Messprozess. Für eine beliebige Messung i erhalten wir:

$$p\left(y_i \mid \mu, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - \mu\right)^2}{2\sigma^2}\right)$$

Diese Formulierung nennt man auch *Likelihoodbeitrag* von Messung i. Für das gesamte Kollektiv von 9 Bäumen erhalten wir – nach ausmultiplizieren von $\left(y_i - \mu\right)^2$ und umstellen der Summe – die folgende Likelihoodfunktion:

$$p\left(\{y_1, \ldots, y_9\} \mid \mu, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\left(\sum_{i=1}^{9} y_i^2\right) - 2\left(\sum_{i=1}^{9} y_i\right) + 9 \cdot \mu^2\right)\right)$$

Für gegebene Messungen $\{y_1, \ldots, y_9\}$ können wir diese Likelihoodfunktion auch als eine Funktion der Parameter $\left(\mu, \sigma\right)^\top$ betrachten (Achtung, wie bereits auf Seite 14 erwähnt ist diese Funktion keine Wahrscheinlichkeitsdichte). Solch eine Funktion – zur besseren Darstellung und numerischen Stabilität[14] werden die Ergebnisse auf log-Skala angegeben, wir berechnen also die log-Likelihood – lässt sich in R direkt programmieren (siehe Anhang A auf Seite 25) und wie in Abbildung 2 auf Seite 19 mit Hilfe einer Konturlinien-Darstellung veranschaulichen.

[14] *Ein Computer ist eine Maschine!* Das heißt Zahlen sind nicht wie in der Mathematik einfach nur beliebig exakt einfach nur da, sondern nur konstruierbar. Dabei gibt es Bereiche die 'leichter fallen', und Bereiche wo eine Zahl nur 'näherungsweise' vorliegt. Z.B. ergibt `log(exp(-1000))` statt dem mathematisch korrekten Wert 1000 in R den Wert `-Inf` (negativ unendlich), `log(exp(-100))` ergibt jedoch den richtigen Wert von `-100`. Mal ausprobieren: `log(exp(-c(730:750)))`.

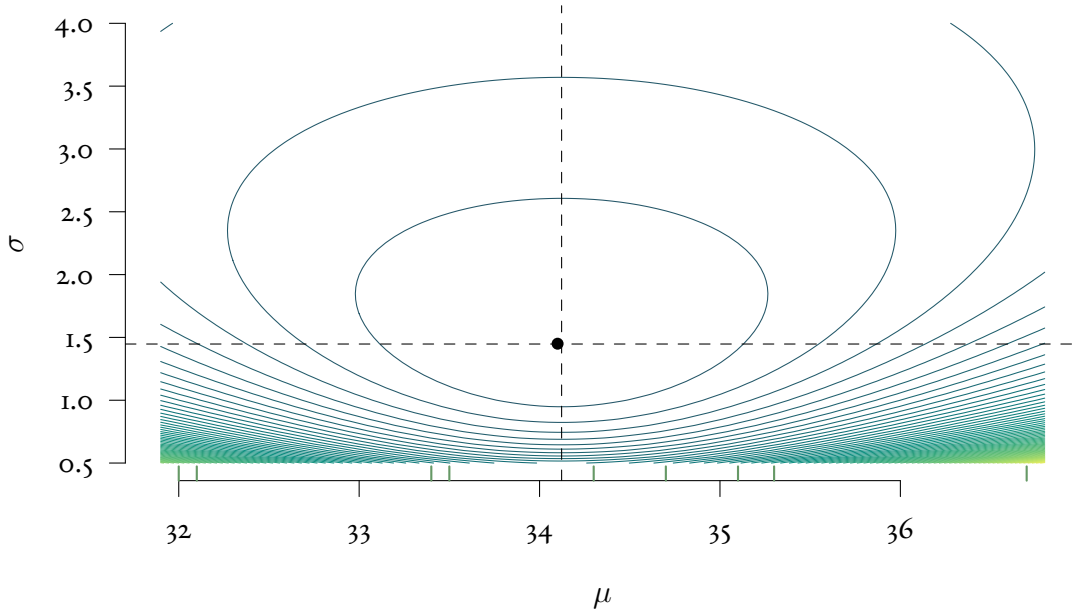

Abbildung 2: Visualisierung der log-Likelihood Oberfläche für μ (x-Achse) und σ (y-Achse). Die Konturlinien stellen den sich verändernden Wert der log-Likelihoodfunktion für sich ändernde Werte von μ und σ dar. Der grüne Punkt gibt den Ort auf dem Gitter – auf dem die log-Likelihood numerisch berechnet wurde – wieder, auf dem diese ihr Maximum annimmt. Die gestrichelten Linien sind auf der jeweiligen Achse der theoretische Wert des Maximum-Likelihood-Schätzers: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ für μ und $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ für σ. Die kurzen grünen Striche auf der x-Achse sind an den Werten der realisierten Baumhöhenmessungen.

In einer echten Anwendung komplett unrealistisch nehmen wir hier nun an, basierend auf der Information aus Albert (2000, Tab. 3), dass uns der Wert der Baumhöhenvariabiltät bekannt ist[15], $\sigma = 1.571$, und dass die Baumhöhenmessungen damit einer Normalverteilung mit nur noch unbekanntem Erwartungswert μ entstammen.

Unser Ziel ist die Posteriori für μ:

$$p\left(\mu \mid \{y_1, \ldots, y_9\}, \mathcal{M}\right),$$

die in Form einer Wahrscheinlichkeitsverteilung all die Informationen über μ, bedingt auf die Messungen und das Modell $\mathcal{M}$, bereit hält. In $\mathcal{M}$ ging bisher ein dass wir von einer Normalverteilung als den Daten-generierenden Mechanismus ausgehen und weiterhin annehmen dass uns die Varianz dieser Messungen bekannt ist. In Anhang B, beginnend auf Seite 26, wird diese Posteriori mathematisch hergeleitet (darin wird mit Abbildung 4 auf Seite 27 auch eine Visualisierung der Likelihoodfuntion für μ gegeben). Wir nehmen in dieser Herleitung an dass unser Vorwissen – das können wir als weiteren Teil von $\mathcal{M}$ verbuchen – über μ mit einer Normalverteilung *Priori-Erwartungswert* μ_0 und *Priori-Varianz* σ_0^2 'modelliert' werden kann:

$$\mu \sim \text{Normal}\left(\mu_0, \sigma_0^2\right).$$

Interpretation der Posteriori Mit unserer Herleitung haben wir herausgefunden dass, nachdem wir sowohl als Daten-generierenden Prozess jeder Messung, als auch für die Priori von μ Normalverteilungen angenommen haben, die Posteriori ebenfalls einer Normalverteilung entspricht:

$$\mu \mid \{y_1, \ldots, y_9\}, \mathcal{M} \sim \text{Normal}\left(\tilde{\mu}, \tilde{\sigma}^2\right),$$

mit Erwartungswert

$$\tilde{\mu} = \tilde{\sigma}^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{9\bar{y}}{\sigma^2}\right)$$

und Varianz

$$\tilde{\sigma}^2 = \left(\frac{1}{\sigma_0^2} + \frac{9}{\sigma^2}\right)^{-1},$$

Diese Posteriori-Varianz können wir direkt interpretieren: Mit einer *diffusen Priori*, also wenn σ_0^2 wesentlich größer als σ^2 ist, $\sigma_0^2 >> \sigma^2$, erhalten wir:

$$\tilde{\sigma}^2 \approx \left(0 + \frac{9}{\sigma^2}\right)^{-1} = \frac{\sigma^2}{9}.$$

bzw. für die Posteriori-Standardabweichung

$$\tilde{\sigma} \approx \frac{\sigma}{\sqrt{9}},$$

[15] Es ist hier nicht meine Intention dass diese Information hier als ein Beispiel für verfügbares, starken Vorwissen angesehen wird, sondern als ein Mittel zum Zweck hier ein didaktisch 'einfaches' Beispiel mit nur einem unbekannten Parameter zu konstruieren.

was direkt dem Standardfehler eines Stichprobenmittelwertes entspricht. Ist hingegen σ_0^2 gar nicht mal so klein, so bedeutet dies dass wir eine höhere *Präzision* unseres Vorwissens haben, was sich dann auch in einer höheren Posteriori-Präzision abbildet:

$$\frac{1}{\tilde{\sigma}^2} = \frac{1}{\sigma_0^2} + \frac{9}{\sigma^2} > \frac{9}{\sigma^2}.$$

Für die Interpretation des Posteriori-Erwartungswertes hilft eine Umformung:

$$\begin{aligned}
\tilde{\mu} &= \tilde{\sigma}^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{9\bar{y}}{\sigma^2} \right) \\
&= \frac{\frac{\mu_0}{\sigma_0^2} + \frac{9\bar{y}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{9}{\sigma^2}} \\
&= \mu_0 \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2} + \frac{9}{\sigma^2}} + \bar{y} \frac{\frac{9}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{9}{\sigma^2}}
\end{aligned}$$

Da $\sigma^2 > 0$ und $\sigma_0^2 > 0$, sind auch $\frac{1}{\sigma_0^2} > 0$, $\frac{9}{\sigma^2} > 0$ und folglich auch $\frac{1}{\sigma_0^2} + \frac{9}{\sigma^2} > 0$. Wir haben damit zwei *Gewichte* $w_0 = \frac{1}{\sigma_0^2} / \left(\frac{1}{\sigma_0^2} + \frac{9}{\sigma^2} \right)$ und $w_y = \frac{9}{\sigma^2} / \left(\frac{1}{\sigma_0^2} + \frac{9}{\sigma^2} \right)$, und damit:

$$\tilde{\mu} = w_0 \mu_0 + w_y \bar{y},$$

wobei $w_0 + w_y = 1$.

Daraus folgt: Mit steigender Priori-Präzision $\frac{1}{\sigma_0^2}$ – das heißt je sicherer wir uns a priori sind dass Werte direkt um μ_0 plausibler sind als Werte weit entfernt von μ_0 – steigt auch das Gewicht w_0 des Priori-Erwartungswertes in der Konstruktion des Posteriori-Erwartungswertes. Mit sinkender Priori-Präzision, oder eben steigender Priori-Varianz σ_0^2, sinkt das Gewicht für μ_0, und der Posteriori-Erwartungswert wird dominiert von den Daten, bzw. genauer gesagt von Lage $\bar{y}$ der Stichprobe $\{y_1, \ldots, y_9\}$.

Werte für μ_0 und σ_0^2 Aus dem Merkblatt Fichte geht hervor dass man bei gleichaltrigen Fichtenreinbeständen durch Jungdurchforstungsmaßnahmen ein Verhältnis von Höhe zu Durchmesser (h/d-Wert) von 80 nicht übersteigen möchte. Ebenfalls im Merkblatt Fichte steht im Abschnitt *Allgemeines zur Zielstärkennutzung und Verjüngung*:

> Sobald mehr als etwa 20 Fichten je ha ihre Zielstärke erreicht haben [...], soll mit der Zielstärkennutzung begonnen werden. Dies kann in sehr wüchsigen Beständen bereits ab dem Alter 60 der Fall sein

Jetzt ist unser Bestand im Solling 116 Jahre alt von Gadow (2003), und nur 9 von 41 Bäumen haben einen BHD$>$ 45cm. Vielleicht ist diese Versuchsfläche eine Nullfläche, weswegen durch fehlende

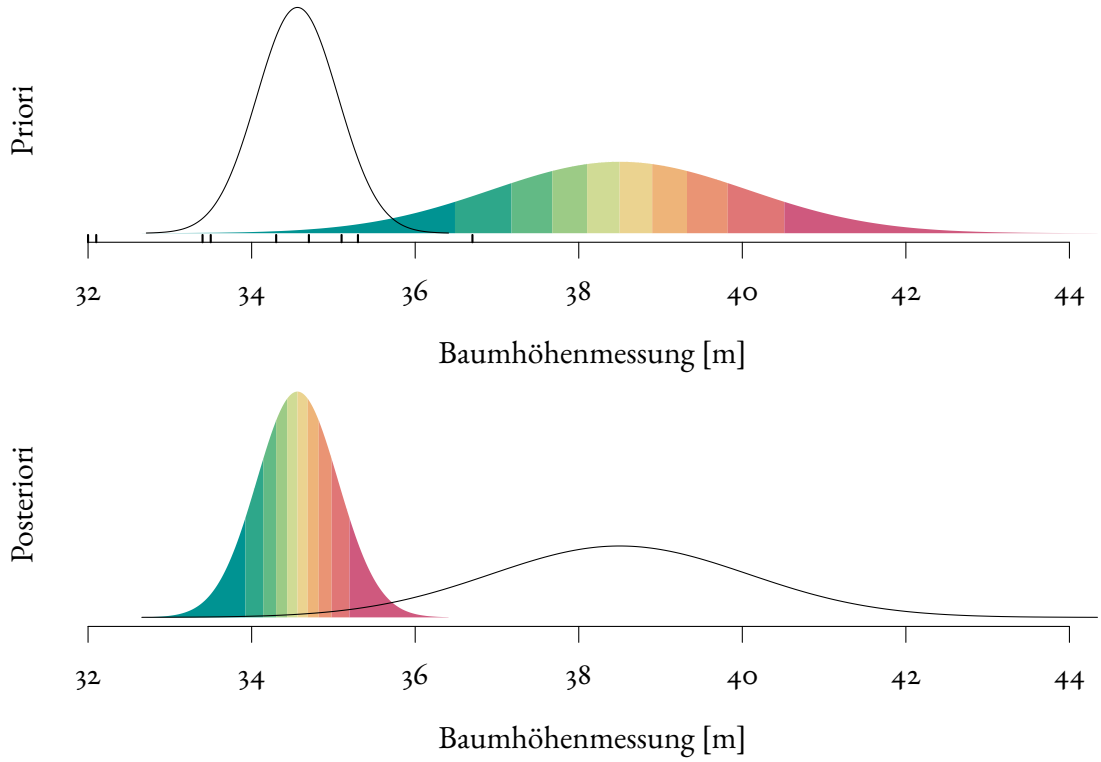

Abbildung 3: Visualisierung der Priori und Posteriori.

Durchforstungsmaßnahmen die Bäume viel in ihr Höhenwachstum und weniger in den Durchmesser investierten, wodurch der h/d-Wert nie abgesenkt wurde? Der mittlere BHD unserer neun Fichten über Zielstaärke war 48.1cm. Bei einem h/d-Wert $= 80$ wäre damit $\mu_0 = 80 \cdot 48.1/100 \approx 38.5$m. Für σ_0 nehm ich den gleichen Wert wie für die Populationsvariabilität $\sigma =$ `sd_h` $= 1.571$ schon feststeht, also $\sigma_0 = 1.571$. Wir haben damit die Priori festgelegt:

$$\mu \sim \text{Normal}\left(38.5, 1.571^2\right).$$

Abbildung 3 stellt Klassen mit je 10% Priori-Wahrscheinlichkeit für μ durch bunte Polygone dar.

Für die Posteriori folgt:

$$\tilde{\sigma}^2 = \left(\frac{1}{1.571^2} + \frac{9}{1.571^2}\right)^{-1} = \frac{1.571^2}{10} \approx 0.247 \approx 0.497^2,$$

und (das arithmetische Mittel der Baumhöhenmessungen ist $\bar{y} \approx 34.1$):

$$\tilde{\mu} = \frac{1.571^2}{10}\left(\frac{38.5}{1.571^2} + \frac{9 \cdot 34.1}{1.571^2}\right) = \frac{38.5 + 9 \cdot 34.1}{10} \approx 34.5.$$

also:

$$\mu \mid \{y_1, \ldots, y_9\}, \mathcal{M} \sim \text{Normal}\left(34.5, 0.497^2\right).$$

Mit der Wahl der Priori-Unsicherheit gleich der bekannten 'Populations-Variabilität', $\sigma_0^2 = \sigma^2$, ergibt sich hier in diesem speziellen Beispiel auch eine spezielle Erkenntnis über den 'Wert der Priori': Die Information die mit der Priori transportiert werden entspricht exakt der Information einer weiteren, hier also einer zehnten Höhenmessung, mit dem Wert $\mu_0 = 38.5$m – diese Analogie jedoch in einem Modell mit dann konstanter Priori, also komplett nicht-informativer Priori, wenn diese hier in dieser Interpretation selbst die Rolle einer Beobachtung übernimmt.

Literatur

Albert, M. (2000). Complementing missing tree heights using an algorithm for single tree growth models. *Jahrestagung des DVFFA Sektion Ertragskunde*.

Cowles, M. K. (2013). *Applied Bayesian Statistics: With R and OpenBUGS Examples*. Springer-Verlag New York.

Downey, A. B. (2013). *Think Bayes*. O'Reilly Media.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition.

Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10).

Held, L. and Bové, D. S. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Springer-Verlag Berlin Heidelberg.

Hilbe, J. M., de Souza, R. S., and Ishida, E. E. O. (2017). *Bayesian Models for Astrophysical Data: Using R, JAGS, Python, and Stan*. Cambridge University Press.

Korner-Nievergelt, F., Roth, T., von Felten, S., Guélat, J., Almasi, B., and Korner-Nievergelt, P. (2015). *Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and STAN*. Academic Press.

Lindley, D. V. (2006). *Understanding Uncertainty*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Marin, J.-M. and Robert, C. (2013). *Bayesian Essentials with R*. Springer-Verlag New York.

Meyer, P., Wevell von Krüger, A., Steffens, R., and Unkrig, W. (2006). Naturwald Brand. *Naturwaldreservate im Kurzportrait, Homepage der NW-FVA (www.nw-fva.de)*, 1(4).

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23.

Neyman, J. (1937). Outline of a theory of statistical estimation basedon the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A*, 236.

Stauffer, H. B. (2007). *Contemporary Bayesian and Frequentist Statistical Research Methods for Natural Resource Scientists*. Wiley.

von Gadow, K. (2003). *Waldstruktur und Wachstum*. Universitätsverlag Göttingen.

A. R-Code zur Berechnung und Visualisierung der log-Likelihood

A.1. Gemeinsame log-Likelihood für μ und σ

```
## log-Likelihood:
l_mu_sigma <- function(mu, sigma) {
  y <- c(35.3, 35.1, 36.7, 33.4, 33.5, 32, 34.7, 34.3, 32.1)
  n <- length(y)
  sum_y <- sum(y)
  sum_y2 <- sum(y^2)
  teil1 <- 1/(sqrt(2 * pi) * sigma)
  teil2 <- -1 * (sum_y2 - 2 * mu * sum_y + n * mu^2)/(2 * sigma^2)
  return(n * log(teil1) + teil2)
  ## sum(dnorm(x = y, mean = mu, sd = sigma, log = T))
}
## Gitter fuer mu- und sigma-Werte:
mu_seq <- seq(min(y) - 0.1, max(y) + 0.1, by = 0.05)
sigma_seq <- seq(0.5, 4, by = 0.05)
mu_sigma_grid <- expand.grid(mu_seq, sigma_seq)
## Ausrechnen aller log-Likelihoodwerte an den Gitterpunkten:
l <- l_mu_sigma(mu = mu_sigma_grid[, 1], sigma = mu_sigma_grid[, 2])
Z <- matrix(nrow = length(mu_seq),
            ncol = length(sigma_seq),
            data = l)
## Konturplot:
contour(x = mu_seq, y = sigma_seq, z = Z,
        levels = seq(min(l), max(l), length = 51)[-51],
        las = 1,
        xlab = '$\\mu$', ylab = '$\\sigma$', bty = "n", drawlabels = F)
points(mu_seq[which(Z == max(Z), arr.ind = T)[, 1]],
       sigma_seq[which(Z == max(Z), arr.ind = T)[, 2]],
       pch = 16, col = rgb(0.4, 0.6, 0.4))
## Theoretische ML-Schätzer:
abline(v = mean(y), lty = 2)
abline(h = sqrt(sum((y - mean(y))^2)/n), lty = 2)
## Beobachtete Daten:
rug(y, col = rgb(0.4, 0.6, 0.4), lwd = 2)
```

A.2. Log-Likelihood für μ bei bekannten σ

```
## log-Likelihood:
```

```
l_mu <- function(mu) {
  y <- c(35.3, 35.1, 36.7, 33.4, 33.5, 32, 34.7, 34.3, 32.1)
  n <- length(y)
  sum_y <- sum(y)
  sum_y2 <- sum(y^2)
  return(-1 * (sum_y2 - 2 * mu * sum_y + n * mu^2))
}
## Gitter fuer mu- und sigma-Werte:
mu_seq <- seq(min(y) - 0.1, max(y) + 0.1, by = 0.05)
## Ausrechnen aller log-Likelihoodwerte an dern Gitterpunkten:
l <- l_mu(mu = mu_seq)
## Konturplot im Bereich hoher Likelihood:
plot(x = mu_seq,
     y = l,
     type = "l", yaxt = "n",
     ylab = '$p\\left(\\left\\{y_1,\\ldots,y_9\\right\\}\\mid\\mu\\right)$',
     xlab = '$\\mu$', bty = "n")
## Maximum der log-Likelihoodfunktion:
points(mu_seq[which.max(l)], max(l),
       pch = 16, col = rgb(0.4, 0.6, 0.4))
## Theoretischer ML-Schätzer:
abline(v = mean(y), lty = 2)
## Beobachtete Daten:
rug(y, col = rgb(0.4, 0.6, 0.4), lwd = 2)
```

B. Herleitung Posteriori der erwarteten Höhenmessungen von Fichten im Solling die die BHD-Zielstärke erreicht haben

Mit einem Blick auf unsere bisherige Formulierung des Likelihood-Beitrags $p\left(y_i \mid \mu, \sigma\right)$ von Messung i können wir feststellen, dass $p\left(y_i \mid \mu, \sigma\right)$ ein Produkt zweier Teilfunktionen ist:

$$p\left(y_i \mid \mu, \sigma\right) = f_1\left(\sigma\right) \cdot f_2\left(y_i, \mu, \sigma\right)$$

mit

$$f_1\left(\sigma\right) = \frac{1}{\sqrt{2\pi}\sigma}$$

und

$$f_2\left(y_i, \mu, \sigma\right) = \exp\left(-\frac{\left(y_i - \mu\right)^2}{2\sigma^2}\right).$$

Ist nun durch den bekannten Parameter σ der Likelihood-Beitrag nur noch eine Funktion mit Parameter μ, $p\left(y_i \mid \mu\right)$, so ist $f_1\left(\sigma\right)$ nun eine Konstante c, und auch $f_2\left(y_i, \mu, \sigma\right)$ kann nicht mehr

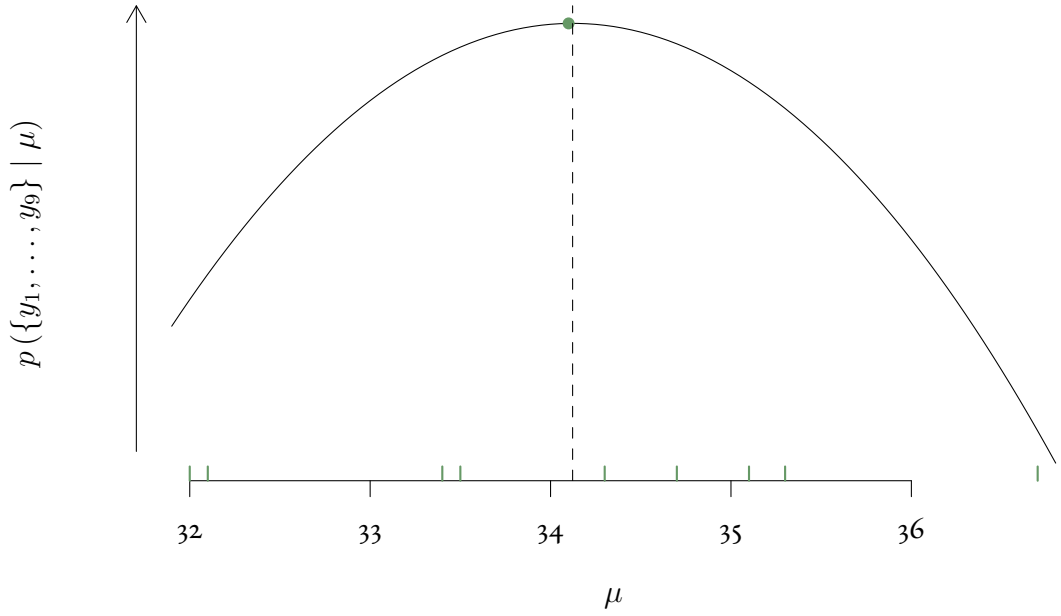

Abbildung 4: Visualisierung des log-Likelihoodfunktionsverlaufs für μ. Der grüne Punkt gibt den Ort auf der Sequenz – auf der die log-Likelihood numerisch berechnet wurde – wieder, auf dem diese ihr Maximum annimmt. Die gestrichelte Linie ist der theoretische Wert des Maximum-Likelihood-Schätzers $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ für μ. Die kurzen grünen Striche auf der x-Achse sind an den Werten der realisierten Baumhöhenmessungen.

bezüglich σ variieren, wir können also die Benennung von σ hier weglassen[16], $f_2(y_i, \mu)$. Wir erhalten demnach:

$$p(y_i \mid \mu) = c \cdot f_2(y_i, \mu),$$

und können folgende Äquivalenzformulierung nutzen:

$$p(y_i \mid \mu) \propto f_2(y_i, \mu),$$

da $f_1(\sigma) = c$ keine Information über μ enthält.

Für die gesamte Likelihoodfunktion aller neun Höhenmessungen erhalten wir (dargestellt in Abbildung 4 auf Seite 27):

$$p(\{y_1, \ldots, y_9\} \mid \mu) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\left(\sum_{i=1}^{9} y_i^2\right) - 2\left(\sum_{i=1}^{9} y_i\right)\mu + 9 \cdot \mu^2\right)\right)$$

[16] π wurde ja in $f_1(\sigma)$ auch nie als Parameter aufgeführt, da π ja hier einfach nur den Wert der 'Kreiszahl' gleich $3.1415\ldots$ beiträgt.

Nehmen wir weiter an, dass die erwartete Baumhöhe μ a priori ebenfalls Normalverteilt ist mit den *Hyperparametern* μ_0 (*Priori-Erwartungswert*) und σ_0^2 (*Priori-Varianz*), welche wir beide erst mal 'offen' lassen, also noch keine Werte festlegen. Unsere Priori $p\left(\mu \mid \mu_0, \sigma_0\right)$ für den einzig unbekannten Parameter $\theta = \mu$ ist damit:

$$\mu \sim \text{Normal}\left(\mu_0, \sigma_0^2\right).$$

Wir können $p\left(\mu \mid \mu_0, \sigma_0\right)$ durch die Unabhängigkeit des ersten Terms, $\frac{1}{\sqrt{2\pi}\sigma_0}$, von μ auch wieder in Äquivalenzform angeben:

$$p\left(\mu \mid \mu_0, \sigma_0\right) \propto \exp\left(-\frac{\left(\mu - \mu_0\right)^2}{2\sigma_0^2}\right)$$

Für die Posteriori folgt damit aus der Äquivalenzform des Satz von Bayes:

$$\begin{aligned} p\left(\mu \mid \{y_1, \ldots, y_9\}\right) &\propto p\left(\mu \mid \mu_0, \sigma_0^2\right) p\left(\{y_1, \ldots, y_9\} \mid \mu\right) \\ &\propto \exp\left(-\frac{\left(\mu - \mu_0\right)^2}{2\sigma_0^2} - \frac{\left(\sum_{i=1}^{9} y_i^2\right) - 2\left(\sum_{i=1}^{9} y_i\right)\mu + 9 \cdot \mu^2}{2\sigma^2}\right) \end{aligned}$$

In der letzten Zeile konnten haben wir wieder alle Terme durch die Äquivalenzformen weggelassen die nicht von μ abhängen. Wir können weiterhin a) $\left(\sum_{i=1}^{9} y_i^2\right) / \left(2\sigma^2\right)$ weglassen, b) mit der Definition des arithmetischen Mittels als $\bar{y} = \frac{1}{n}\sum_{i=1}^{9} y_i$ können wir durch $9\bar{y} = \sum_{i=1}^{9} y_i$ zu einer etwas kompakteren Notation finden, und c) den Term $\left(\mu - \mu_0\right)^2$ ausmultiplizieren und dann hier auch $\mu_0^2 / \left(2\sigma_0^2\right)$ wegen der Unabhängigkeit von μ weglassen. Wir erhalten:

$$\begin{aligned} p\left(\mu \mid \{y_1, \ldots, y_9\}\right) &\propto \exp\left(-\frac{\mu^2 - 2\mu\mu_0}{2\sigma_0^2} - \frac{-2 \cdot 9 \cdot \bar{y}\mu + 9\mu^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\mu^2}{2\sigma_0^2} + \frac{2\mu\mu_0}{2\sigma_0^2} + \frac{2 \cdot 9 \cdot \bar{y}\mu}{2\sigma^2} - \frac{9\mu^2}{2\sigma^2}\right) \end{aligned}$$

Wir konzentrieren uns auf den Ausdruck innerhalb der Exponentialfunktion und ziehen alle Terme zusammen die von μ oder μ^2 abhängen:

$$-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{9}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{9 \cdot \bar{y}}{\sigma^2}\right).$$

Wir definieren und zwei Parameter $\tilde{\sigma}^2$ und $\tilde{\mu}$ als:

$$\frac{1}{\tilde{\sigma}^2} := \frac{1}{\sigma_0^2} + \frac{9}{\sigma^2}$$

und basierend auf $\tilde{\sigma}^2$:

$$\tilde{\mu} := \tilde{\sigma}^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{9 \cdot \bar{y}}{\sigma^2}\right).$$

Wir erhalten damit:

$$\begin{aligned} -\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{9}{\sigma^2}\right)+\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{9\cdot\bar{y}}{\sigma^2}\right) &= -\frac{\mu^2}{2\tilde{\sigma}^2}+\frac{\mu\tilde{\mu}}{\tilde{\sigma}^2} \\ &= -\frac{\mu^2}{2\tilde{\sigma}^2}+\frac{2\mu\tilde{\mu}}{2\tilde{\sigma}^2} \end{aligned}$$

In der letzten Zeile haben wir den *Eins-Trick*, $1=\frac{2}{2}$, genutzt. Als nächstes nutzen wir den in mathematischen Herleitungen häufig benutzten *Null-Trick*, $0=\tilde{\mu}^2-\tilde{\mu}^2$, und erhalten damit:

$$\begin{aligned} -\frac{\mu^2}{2\tilde{\sigma}^2}+\frac{2\mu\tilde{\mu}}{2\tilde{\sigma}^2} &= -\frac{\mu^2-2\mu\tilde{\mu}+\tilde{\mu}^2-\tilde{\mu}^2}{2\tilde{\sigma}^2} \\ &= -\frac{\left(\mu-\tilde{\mu}\right)^2}{2\tilde{\sigma}^2}-\frac{\tilde{\mu}^2}{2\tilde{\sigma}^2} \end{aligned}$$

Der hier behandelten Ausdruck war ursprünglich ausgewertet in der Exponentialfunktion, d.h.:

$$\exp\left(-\frac{\left(\mu-\tilde{\mu}\right)^2}{2\tilde{\sigma}^2}-\frac{\tilde{\mu}^2}{2\tilde{\sigma}^2}\right)=\exp\left(-\frac{\left(\mu-\tilde{\mu}\right)^2}{2\tilde{\sigma}^2}\right)\exp\left(-\frac{\tilde{\mu}^2}{2\tilde{\sigma}^2}\right)$$

Stellen wir uns alle Terme die wir bisher – aufgrund der Nutzung von Äquivalenzformen – weggelassen haben als eine gemeinsame Funktion $f_{c_1}\left(\{y_1,\ldots,y_9\},\sigma,\mu_0,\sigma_0\right)$ vor. Wir können diese Funktion nun hier noch einmal um den zweiten Faktor der rechten Seite unserer letzten Gleichung erweitern und erhalten als unsere Posteriori:

$$p\left(\mu\mid\{y_1,\ldots,y_9\},\mathcal{M}\right)=\exp\left(-\frac{\left(\mu-\tilde{\mu}\right)^2}{2\tilde{\sigma}^2}\right)f_{c_2}\left(\{y_1,\ldots,y_9\},\sigma,\mu_0,\sigma_0\right),$$

wobei spätestens hier ein guter Zeitpunkt ist um mit $\mathcal{M}$ noch einmal auf all das zu verweisen was man so gesetzt hatte: Likelihood basiert auf Normalverteilung, Varianz bekannt, Priori für μ basiert auf Normalverteilung, Hyperparameter μ_0 und σ_0, ….

Wir nutzen noch einmal den Eins-Trick mit $1=\frac{1}{\sqrt{2\pi}\tilde{\sigma}}\sqrt{2\pi}\tilde{\sigma}$ und erhalten:

$$p\left(\mu\mid\{y_1,\ldots,y_9\},\mathcal{M}\right)=\frac{1}{\sqrt{2\pi}\tilde{\sigma}}\exp\left(-\frac{\left(\mu-\tilde{\mu}\right)^2}{2\tilde{\sigma}^2}\right)f_{c_3}\left(\{y_1,\ldots,y_9\},\sigma,\mu_0,\sigma_0\right),$$

Wir sind am Ziel: Wir konnten hiermit herleiten, dass die Posteriori proportional ist zu einer Normalverteilung:

$$\mu\mid\{y_1,\ldots,y_9\},\mathcal{M}\sim\text{Normal}\left(\tilde{\mu},\tilde{\sigma}^2\right),$$

mit Erwartungswert $\tilde{\mu}$ und Varianz $\tilde{\sigma}^2$. In der weiteren Nutzung können wir dann die Nebeninformation 'proportional zu' auch weglassen, wir sagen 'die Posteriori ist eine Normalverteilung mit …'.

C. R-Code plot_prior_posterior

```
plot_prior_posterior <- function(mu_0, sd_0, sd_y, y) {
  farben <- c("#009392", "#2EA78A", "#62BA85", "#9CCB86", "#D0DB95", "#EBD390",
              "#EEB479", "#EA9474", "#E07676", "#CF597E")
  print(sd_post <- sqrt(1/((1/sd_0^2) + length(y)/sd_y^2)))
  print(mu_post <- ((mu_0/sd_0^2) + (sum(y)/sd_y^2))*sd_post^2)
  par(mfrow = c(2, 1), mar = c(4, 1, 0, 0) + 0.2)
  x1 <- seq(qnorm(p = 0.0001, mean = mu_0, sd = sd_0),
            qnorm(p = 0.9999, mean = mu_0, sd = sd_0), length.out = 200)
  x2 <- seq(qnorm(p = 0.0001, mean = mu_post, sd = sd_post),
            qnorm(p = 0.9999, mean = mu_post, sd = sd_post), length.out = 200)
  x_lim <- range(c(x1, x2, y)) + c(-0.1, 0.1)
  y1 <- dnorm(x = x1, mean = mu_0, sd = sd_0)
  y2 <- dnorm(x = x2, mean = mu_post, sd = sd_post)
  plot(x1, y1, type = "n", yaxt = "n", ylim = c(0, max(c(y1, y2))),
       xlim = x_lim,
       ylab = "", xlab = "Baumhöhenmessung [m]", bty = "n")
  mtext(2, text = "Priori", line = 0)
  rug(y, lwd = 2)
  q <- qnorm(p = c(0.0001, seq(0.1, 0.9, by = 0.1), 0.9999),
             mean = mu_0, sd = sd_0)
  for (i in 2:length(q)) {
    x <- seq(q[i - 1], q[i], length.out = 50)
    y <- dnorm(x = x, mean = mu_0, sd = sd_0)
    polygon(c(x, rev(x)), c(y, 0*y), col = farben[i - 1], border = NA)
  }
  lines(x2, y2, lty = 1)
  plot(x1, y1, type = "n", yaxt = "n", ylim = c(0, max(c(y1, y2))),
       xlim = x_lim,
       ylab = "", xlab = "Baumhöhenmessung [m]", bty = "n")
  mtext(2, text = "Posteriori", line = 0)
  rug(y, lwd = 2)
  q <- qnorm(p = c(0.0001, seq(0.1, 0.9, by = 0.1), 0.9999),
             mean = mu_post, sd = sd_post)
  for (i in 2:length(q)) {
    x <- seq(q[i - 1], q[i], length.out = 50)
    y <- dnorm(x = x, mean = mu_post, sd = sd_post)
    polygon(c(x, rev(x)), c(y, 0*y), col = farben[i - 1], border = NA)
  }
  lines(x1, y1, lty = 1)
  abline(v = mean(y), lty = 2)
}
```