HTWD

Hochschule für Technik und Wirtschaft

Dokumentation

Projektseminar

Optimierung und Unsicherheitsquantifizierung mit Bayesianischer Statistik und MCMC-Methoden

(Prof. Schwarzenberger)

Clemens Näther, s85426
Jakub Kliemann, s85515

Inhaltsverzeichnis

1 Einleitung

Die Statistik ist ein wesentlicher Bestandteil der Wissenschaft. Sie ermöglicht es, aus Daten Informationen zu gewinnen und aus diesen Schlussfolgerungen zu ziehen. Es exsistieren dabei aber zwei verschiedene Ansätze, die klassische und die bayesianische Statistik. Beide bieten unterschiedliche Perspektiven und Werkzeuge zur Analyse und Interpretation von Unsicherheiten, und die Entscheidung für eine Methode kann erhebliche Auswirkungen auf die Schlussfolgerungen einer Studie haben.

In dieser Arbeit liegt der Fokus auf der bayesianischen Statistik, einem Ansatz, der die Wahrscheinlichkeit als subjektives Maß für die Unsicherheit interpretiert und es ermöglicht, Vorwissen systematisch in den Analyseprozess einzubringen. Der Vergleich mit der klassischen Statistik, die auf der relativen Häufigkeit basiert, zeigt, wie unterschiedlich die Ansätze sowohl in der Theorie als auch in der Praxis sind.

Ein zentraler Schwerpunkt der Arbeit liegt darin die theoretischen Grundlagen der bayesianischen Statistik zu erläutern und den Prinzipien der klassischen Statistik gegenüberzustellen. Dabei soll insbesondere auf die Markov Chain Monte Carlo (MCMC) Methoden eingegangen werden, die es ermöglichen, komplexe Modelle zu schätzen und zu simulieren. Inbesondere wird der Metropolis-Hastings-Algorithmus, einer der bekanntesten Algorithmen zur Erstellung von Markovketten vorgestellt.

Anschließend wird im zweiten Teil der Arbeit die praktische Anwendung der bayesianischen Statistik und der MCMC-Methoden anhand von Beispielen und Datensätzen in Python demonstriert. Dabei wird gezeigt, wie bayesianische Modelle implementiert und auf verschiedene Datensätze angewendet werden können. Die Ergebnisse werden interpretiert und mit klassischen Methoden verglichen.

2 Theoretischer Teil

2.1 Grundlagen der bayesianischen Statistik und das Bayes'sche Theorem

2.1.1 Einführung in die bayesianische Statistik

Die bayesianische Statistik ist ein Zweig der Statistik. Sie unterscheidet sich im wesentlichen in der Interpretation der Wahrscheinlichkeit von der klassischen Statistik. Die klassische Statistik definiert die Wahrscheinlichkeit als die **relative Häufigkeit** in einem Zufallsexperiment [5, S. 2]. In der bayesianischen Statistik hingegen wird die Wahrscheinlichkeit als Grad des Glaubens respektiv als **Plausibilität** eines Ereignisses oder einer Aussage interpretiert [2, S. 1].

Kern der bayesianischen Statistik ist es Wissen über ein Ereignis zu verfeinern, sobald neue Informationen vorliegen. Dazu nutzt man hauptsächlich das **Bayes'sche Theorem**, welches erlaubt das Vorwissen (Prior) mit neuen Daten (Likelihood) zu kombinieren und daraus eine aktualisierte Wahrscheinlichkeit (Posterior) zu berechnen.

Mit Hilfe des Bayes'schen Theorems kann man unbekannte Parameter schätzen, ein Konfidenzintervall für diese Parameter angeben und Hypothesen prüfen. Die klassische Statistik benötigt hingegen dafür Testgrößen, weshalb die bayesianische Statistik als flexibler und intuitiver gilt. [2, S. 1].

Problem der bayesianischen Statistik ist jedoch, dass die Berechnung der Posterioriverteilung analytisch oft nicht möglich ist. Da es nun aber gute numerische Methoden wie die **Markov Chain Monte Carlo (MCMC)** Methoden gibt, findet die bayesianische Statistik immer mehr Anwendungen. So zum Beispiel in der Medizin oder für künstliche Intelligenzen. [5, S. 1].

2.1.2 Das Bayes'sche Theorem und seine Bestandteile

Das Bayes'sche Theorem ist ein fundamentales Konzept der bayesianischen Statistik. Es beschreibt, wie man vorhandenes Vorwissen durch neue Daten aktualisiert.

Die **Prioriverteilung** beschreibt die anfänglichen Annahmen oder das Vorwissen über einen Parameter oder ein Ereignis, bevor neue Daten berücksichtigt werden. Dabei "enthält die Priorverteilung eines Parameters θ, ausgedrückt durch $f(\theta)$, was man vor Auswertung der Stichprobe über θ weiß." [5, S. 90].
Als Priori-Wahrscheinlichkeit wird somit die Wahrscheinlichkeit $P(A)$ bezeichnet.

Die **Posterioriverteilung** beschreibt das Wissen über einen Parameter oder ein Ereignis, nachdem alle vorhandenen Daten berücksichtigt wurden. Durch die neuen Daten, meist einer Stichprobe, wird die anfängliche Annahme, die durch die Prioriverteilung ausgedrückt wird, aktualisiert. Dies führt zu einer neuen Verteilung die widerspiegelt, wie wahrscheinlich verschiedene Werte des Parameters auf Grundlage sowohl des Vorwissens als auch der neuen Informationen sind. [5, S. 109]
Die Posteriori-Wahrscheinlichkeit wird somit als $P(A|B)$ bezeichnet.

Die **Likelihood-Funktion** enthält die Informationen, die die Daten über den Parameter oder das Ereignis liefern. Dabei beschreibt die Likelihood die Informationen aus den neuen Daten, die zur Aktualisierung der Prioriverteilung beitragen. [5, S. 88] Die Likelihood-Wahrscheinlichkeit wird somit als $P(B|A)$ bezeichnet.

Die Wahrscheinlichkeit $P(B)$ wird als Normierungskonstante bezeichnet. Sie sorgt dafür, dass die Posterioriverteilung korrekt normiert ist, das heißt, dass die Summe der Wahrscheinlichkeiten aller möglichen Werte des Parameters 1 ergibt. [5, S. 109]

Das Bayes'sche Theorem lässt sich somit wie folgt darstellen:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \tag{1}$$

Das Bayes'sche Theorem lässt sich auch rekursiv anwenden [2, S. 17].
Gegeben sei das Ereignis A sowie die Teilergebnisse $B_1, B_2, ..., B_n$. Dann ergibt sich die Wahrscheinlichkeit $P(A|B_1)$ zu:

$$P(A|B_1) = \frac{P(A) \cdot P(B_1|A)}{P(B_1)} \tag{2}$$

Nun wird die Information B_2 hinzugefügt. Die Wahrscheinlichkeit $P(A|B_1, B_2)$ ergibt sich bei Unabhängigkeit von den Teilereignissen $B_1, B_2, ..., B_n$ zu:

$$P(A|B_1, B_2) = \frac{P(A) \cdot P(B_1|A) \cdot P(B_2|A)}{P(B_1) \cdot P(B_2)} \tag{3}$$

Weiterhin lässt sich diese Formel umstellen, wodurch deutlich wird, dass beim Hinzufügen von neuen Informationen die Posterioriverteilung aktualisiert wird:

$$P(A|B_1, B_2) = \frac{P(A) \cdot P(B_1|A) \cdot P(B_2|A)}{P(B_1) \cdot P(B_2)} = P(A|B_1) \cdot \frac{P(B_2|A)}{P(B_2)} \tag{4}$$

Dies lässt sich allgemein formulieren für:

$$P(A|B_1, B_2, ..., B_n) = P(A|B_1, B_2, ..., B_{n-1}) \cdot \frac{P(B_n|A)}{P(B_n)} \tag{5}$$

Die Wahl der Prioriverteilung ist ein wichtiger Aspekt der bayesianischen Statistik. Sie wird immer so gewählt, dass die Entropie maximal ist. Die Entropie ist ein Maß für die Unsicherheit, was bedeutet, dass nur Informationen enthalten sind, die vor der Beobachtung bekannt sind. [2, S. 57]. Unter folgenden Bedingungen ist die Prioriverteilung optimal [2, S. 59]:

- Zufallsvariablen, die in $[a, b]$ definiert sind, sind **gleichverteilt**
- Zufallsvariablen mit gegebenen Mittelwert und Varianz sind **normalverteilt**
- Zufallsvariablen mit gegebenen Mittelwert sind **exponentialverteilt**
- Zufallsvariablen mit gegebenen Mittelwert und Varianz im Intervall [0,∞] besitzen eine **abgeschnittene Normalverteilung**

Wenn keine Informationen über den Parameter vorliegen, wird eine **uninformative** Prioriverteilung gewählt. Es handelt sich dabei um eine uneigentliche Verteilung. [2, S. 57].

2.1.3 Beispiele und praktische Anwendungen

Beispiel 1: m gleichgeformte Kugeln, unter denen sich k rote Kugeln und $m-k$ schwarze Kugeln befinden. Eine Kugel wird zufällig gezogen. Die Wahrscheinlichkeit, dass die gezogene Kugel rot ist, beträgt

$$P(A) = \frac{k}{m} = p \tag{6}$$

Der Versuch wird erweitert, sodass n-mal eine Kugel mit Zurücklegen gezogen wird. Die Wahrscheinlichkeit, dass x-mal eine rote Kugel bie n-maligem Ziehen gezogen wird, beträgt

$$P(x|n,p) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \tag{7}$$

Sei nun p unbekannt. Dieses p ist nun zu schätzen. Die Binomialverteilung wird nun als Likelihood-Funktion verwendet:

$$P(n,x|p) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \tag{8}$$

wobei $0 \leq p \leq 1$. Als Prioridichte wird die Gleichverteilung verwendet, da es keine Informationen über p gibt.

$$P(p) = \begin{cases} 1, & \text{für } 0 \leq p \leq 1 \\ 0, & \text{sonst} \end{cases} \tag{9}$$

Die Posterioridichte ergibt sich somit zu:

$$P(p|n,x) = \frac{\binom{n}{x} p^x \cdot (1-p)^{n-x}}{P(n,x)} \tag{10}$$

Vergleicht man dies mit der Dichtefunktion der Beta-Verteilung, so erkennt man dass die Posterioridichte einer Beta-Verteilung entspricht.

$$P(p|n,x) = \frac{(n+1)!}{x! \cdot (n-x)!} \cdot p^x \cdot (1-p)^{n-x} = \frac{\Gamma(n+1)}{\Gamma(x+1) \cdot \Gamma(n-x+1)} \cdot p^x \cdot (1-p)^{n-x} \tag{11}$$

Somit suche nach Maximum der Posterioridichte, um den Schätzer für p zu finden.

$$\frac{d}{dp} P(p|n,x) = x p^{x-1} \cdot (1-p)^{n-x} - (n-x) p^x \cdot (1-p)^{n-x-1} = 0 \tag{12}$$

$$\Rightarrow x p^{x-1} \cdot (1-p)^{n-x} = (n-x) p^x \cdot (1-p)^{n-x-1} \tag{13}$$

Vereinfacht ergibt sich:

$$\Rightarrow x(1-p) = (n-x)p \tag{14}$$

$$\Rightarrow \frac{x}{p} - \frac{n-x}{1-p} = 0 \tag{15}$$

$$\Rightarrow p = \frac{x}{n} \tag{16}$$

Der Schätzer für p ist somit der relative Anteil der roten Kugeln an der Gesamtanzahl der Kugeln.

Beispiel 2: Beispiel 1 wird erweitert. Es wird nun eine zweite Stichprobe gezogen. Stichprobe 1: $n_1 = 10$, $x_1 = 4$, Stichprobe 2: $n_2 = 20$, $x_2 = 6$. Daten sind unabhängig voneinander. Die Daten (Posterioriverteilung) der 1. Stichprobe dienen nun als Prioridichte für die 2. Stichprobe. Man erhält somit:

$$P(p|n_1, x_1, n_2, x_2) = \frac{P(p|n_1, x_1) \cdot P(p|n_2, x_2)}{P(n_1, x_1, n_2, x_2)} \tag{17}$$

Dabei ist die Prioridichte identisch zur Posterioridichte der 1. Stichprobe, siehe Gleichung (10). Die Posterioridichte der 2. Stichprobe ergibt sich somit zu:

$$P(p|n_1, x_1, n_2, x_2) = \frac{\binom{n_1+n_2}{x_1+x_2} \cdot p^{x_1} \cdot (1-p)^{n_1-x_1} \cdot p^{x_2} \cdot (1-p)^{n_2-x_2}}{P(n_1, x_1, n_2, x_2)} \tag{18}$$

Oder mithilfe der Beta-Verteilung:

$$P(p|n_1, x_1, n_2, x_2) = \frac{\Gamma(n_1 + n_2 + 1)}{\Gamma(x_1 + x_2 + 1) \cdot \Gamma(n_1 + n_2 - x_1 - x_2 + 1)} \cdot p^{x_1+x_2} \cdot (1-p)^{n_1+n_2-x_1-x_2} \tag{19}$$

Die Daten der 1. und 2. Stichprobe könnnen somit kombiniert werden. Für die Daten $n_1 = 10$, $x_1 = 4$, $n_2 = 20$, $x_2 = 6$:

$$P(p|10, 4, 20, 6) = 931395465p^{10} \cdot (1-p)^{20} \tag{20}$$

2.1.4 Punktschätzer, Konfidenzintervalle und Hypothesenprüfung in der bayesianischen Statistik

Punktschätzer

Im folgenden wird die Schätzung eines Parameters mithilfe der Bayes-Strategie erläutert. Die möglichen Schätzwerte der Parameter x werden als $\hat{x}$ bezeichnet. Die wahren Parameter werden als x bezeichnet.
Es wird eine Kostenfunktion $L(\hat{x}, x)$ definiert, die die Kosten für die Schätzung $\hat{x}$ des wahren Parameters x angibt. Dies bedeutet, dass die Kostenfunktion die Differenz zwischen dem wahren Parameter x und der Schätzung $\hat{x}$ angibt. Dabei gibt es verschiedene Kostenfunktionen, die verwendet werden können. [2, S. 65]

Die **quadratische Kostenfunktion** ist definiert als: $L(x - \hat{x}) = (x - \hat{x})\Sigma^{-1}(x - \hat{x})$. Diese gibt den quadratischen Abstand zwischen dem wahren Parameter x und der Schätzung $\hat{x}$ an. Die zu erwartenden Kosten werden berechnet mit dem Erwartungswert der Kostenfunktion. Diese Schätzung führt zu dem Erwartungswert von x, das heißt $\hat{x} = E(x)$. [2, S. 65–66]

Die **Kostenfunktion der absoluten Fehler** ist definiert als: $L(x, \hat{x}) = |x - \hat{x}|$. Diese gibt den absoluten Abstand zwischen x und $\hat{x}$ an. Die Schätzung mit dem absoluten Fehler ergibt den Median der Verteilung, das heißt $F(\hat{x}_{med}) = 0.5$ [2, S. 67–68]

Die **Null-Eins-Kostenfunktion** bedeutet, dass es entweder Kosten oder keine Kosten gibt. Diese ist definiert durch: $L(x-\hat{x}) = \begin{cases} 0 & \text{für } |x-\hat{x}| < b \\ a & \text{für } |x-\hat{x}| \geq b \end{cases}$, wobei a und b als Konstaten angenommen werden. Wenn der Fall $b \to 0$ betrachtet wird, ergibt sich als Schätzer das Argument des Maximums der Posterioriverteilung, das heißt $\hat{x}_M = \arg\max\ p(x|y)$. [2, S. 68–69]

Es ist so ersichtlich, dass es verschiedene Punktschätzer in der bayesianischen Statistik gibt:

- **Erwartungswert** $\hat{x} = E(x)$
- **Median** $\hat{x} = x_{0.5}$
- **Maximum der Posterioriverteilung** $\hat{x} = \arg\max\ p(x|y)$

Die Wahl zwischen den Schätzern hängt so von der Problemstellung ab.

Konfidenzintervalle

Um Unsicherheiten bei Schätzungen zu quantifizieren werden oftmals **Konfidenzintervall** verwendet. Konfidenzintervalle geben einen Bereich an, in dem ein unbekannter Parameter mit einer bestimmten Wahrscheinlichkeit liegt. In der Bayes Statistik kann eine Bereichschätzung unmittelbar aus der Posterioriverteilung abgeleitet werden. [2, S. 71]

Wenn $P(x|y)$ die Posterioridichte für den Parameter x ist, ist das Konfidenzintervall mit Konfidenzniveau 1-α definiert als:

$$P(x \in X_u|y) = \int_{X_u} P(x|y)dx = 1 - \alpha \tag{21}$$

Wobei für alle $x_1 \in X_u, x_2 \notin X_u$ gilt $P(x_1|y) \geq P(x_2|y)$; Xu ist dabei ein Unterraum auf dem Parameterraum [2, S. 71]. Meist wird 0.1, 0.05 oder 0.01 für α gewählt. Höhere Konfidenzniveaus führen zu größeren Intervallen, da mehr Unsicherheit abgedeckt wird, während niedrigere Konfidenzniveaus kleinere Intervalle liefern, die jedoch ein höheres Risiko haben, den wahren Parameter nicht einzuschließen.

Falls die Posterioriverteilung $P(x|y)$ beispielsweise der Normalverteilung $N(\mu, \sigma^2)$ entspricht, ist das Konfidenzintervall symmetrisch um den Mittelwert und die Grenzen lassen sich einfach durch die Quantile der Standardnormalverteilung z über $\mu - z_{1-\alpha/2} \cdot \sigma$ und $\mu + z_{1-\alpha/2} \cdot \sigma$ bestimmen.

Da die Konfidenzintervalle in der bayesianischen Statistik auf der Posterioriverteilung basieren, unterscheiden sich die Ergebnisse von denen der klassischen Statistik, da diese auf einer anderen Grundlage definiert sind.

Hypothesenprüfung

Hypothesenprüfung ist eine Methode, mit der man Annahmen über unbekannte Parameter überprüft. Dabei wird entschieden, ob solche Annahmen akzeptiert oder abgelehnt werden sollen. [2, S. 74]. Man unterscheidet zwischen der Nullhypothese H_0 und der Alternativhypothese H_1. Die Nullhypothese ist die Annahme, die überprüft wird, während die Alternativhypothese die Annahme ist, die akzeptiert wird, wenn die Nullhypothese abgelehnt wird.

In der bayesianischen Statistik werden Hypothesen durch Unterräume des Parameterraums definiert. Seien X_1 und X_2 disjunkte Unterräume des Parameterraums, so lauten die Hypothesen $H_0 : x \in X_1$ und $H_1 : x \in X_2$. Die Wahrscheinlichkeiten für diese Hypothesen ergeben sich direkt aus der Posterior-Dichte. Die Wahrscheinlichkeit, dass die Nullhypothese wahr ist, lautet:

$$P(H_0|y) = \int_{X_1} P(x|y)dx \tag{22}$$

Analog gilt dies für $P(H_1|y)$ [2, S. 74]. Ist jetzt $P(H_0|y) > P(H_1|y)$ rsp. $\frac{P(H_0|y)}{P(H_1|y)} > 1$, so wird die Nullhypothese akzeptiert, andernfalls die Alternativhypothese. [2, S. 77].

Bei einer Punkthypothese $H_0 : x = x_0$ ergibt sich $P(H_0|y) = 0$, da die Posterior-Dichte kontinuierlich ist und ein einzelner Punkt daher keine Wahrscheinlichkeit besitzt. In solchen Fällen greift man beispielsweise auf Konfidenzintervalle zurück: Liegt der Wert x_0 innerhalb des Konfidenzintervalls, wird die Nullhypothese akzeptiert, andernfalls abgelehnt [2, S. 84].

Ein wesentlicher Unterschied zur klassischen Statistik besteht darin, dass in der traditionellen Hypothesenprüfung die Nullhypothese solange als gültig angesehen wird, bis genügend Beweise dagegen vorliegen (Prinzip des Zweifelns). In der bayesiani-schen Statistik hingegen werden die Wahrscheinlichkeiten beider Hypothesen direkt geschätzt, wodurch beide Hypothesen von Anfang an gleichberechtigt betrachtet werden. [2, S. 83].

2.2 Markov Chain Monte Carlo (MCMC) Methoden

2.2.1 Einführung in MCMC-Methoden

Bei einer direkten Simulation wird vorrausgesetzt, dass die Verteilung der Zufallsvariablen bekannt ist. Dies ist jedoch in der Praxis nicht immer gegeben. Die Berechnung der Posterioriverteilung ist analytisch oft nicht möglich, vor allem bei komplexen Modellen oder hohen Dimensionen.

Die Markov Chain Monte Carlo (MCMC) Methoden sind eine Klasse von Algorithmen, die es ermöglichen, eine Stichprobe aus einer Verteilung zu ziehen, ohne die Verteilung zu kennen. [4, S. 179]
Diese Methoden verwenden zwei Konzepte: Markov-Ketten und Monte Carlo-Methoden. Eine Markov-Kette ist eine Folge von Zufallsvariablen, die die Markov-Eigenschaft erfüllen. Die Markov-Eigenschaft sagt aus, dass die nächste Zufallsvariable nicht von den vorherigen Zufallsvariablen, sondern nur von der letzten Zufallsvariable abhängt. Das bedeutet, dass die Wahrscheinlichkeit, im nächsten Zustand $X_n + 1$ zu landen, nur von X_n abhängt. Die Übergangswahrscheinlichkeit zwischen den Zuständen kann in einer Übergangsmatrix dargestellt werden. [4, 188f.]
Die Monte Carlo-Methoden sind eine Gruppe von Algorithmen, die es ermöglichen, Zufallsvariablen zu schätzen, indem Zufallszahlen generiert werden. Sie erzeugen zufällige Stichproben, um eine Näherung der Verteilung zu erhalten. [4, 14f.]

Die MCMC-Methoden nutzen die Monte Carlo-Methoden, um eine Markov-Kette zu simulieren. Diese Technik ist besonders nützlich, um eine Posterioriverteilung zu schätzen, wenn direkte Berechnungen nicht möglich sind. [4, S. 179]
Im folgenden wird der Metropolis-Hastings-Algorithmus erläutert, eine der bekanntesten MCMC-Methoden.

2.2.2 Der Metropolis-Hastings-Algorithmus

Der Metropolis-Hastings-Algorithmus erstellt eine Markov-Kette, die eine Posterioriverteilung $f(x)$ approximiert. $f(x)$ besteht aus einer bekannten Funktion $p(x)$, die sich aus der Likelihood-Funktion und der Prioriverteilung zusammensetzt und aus einer unbekannten Normierungskonstante. Für die Berechnung wird eine Hilfsfunktion q(y|x) benötigt, welche die Übergangswahrscheinlichkeit von einem Zustand x_t zum nächsten Zustand x_{t+1} angibt und ähnlich wie die Posterioriverteilung ist. Oftmals wird dafür die Normalverteilung gewählt sodass $q(y|x) = N(x, \sigma^2)$, wobei σ beeinflusst dabei die Schrittweite, die die Markov-Kette macht. [3, 226f.]

Der Algorithmus besteht aus folgenden Schritten:

1. Wähle einen Startwert x_0
2. Wähle einen neuen Wert y aus der Hilfsfunktion $q(y|x_t)$ im Fall einer Normalverteilung $N(x_t, \sigma^2)$ wird ein zufälliger Wert aus der Normalverteilung gezogen
3. Berechne die Akzeptanzwahrscheinlichkeit α:

$$\alpha = min\{1, \frac{p(y) \cdot q(x_t|y)}{p(x_t) \cdot q(y|x_t)}\} \tag{23}$$

   Diese gibt an, wie wahrscheinlich der neue Wert x_{t+1} akzeptiert wird.
4. Ziehe eine Zufallszahl u aus einer Gleichverteilung $U(0, 1)$
5. Wähle den nächsten Wert x_{t+1}:

$$x_{t+1} = \begin{cases} y, & \text{wenn u} < \alpha \\ x_t, & \text{sonst} \end{cases} \tag{24}$$

6. Wiederhole die Schritte 2-5 bis die Stichprobe die gewünschte Verteilung genau genug approximiert

Der Metropolis-Hastings-Algorithmus ist ein Verallgemeinerung des Metropolis-Algorithmus, der nur für symmetrische Hilfsfunktionen $q(y|x)$ funktioniert. Bei einer symmetrischen Hilfsfunktion ist $q(y|x_t) = q(x_t|y)$ und deswegen kürzt sich die Akzeptanzwahrscheinlichkeit zu $\alpha = min\{1, \frac{p(y)}{p(x_t)}\}$. [3, 226f.]

Der erste Teil der Markov-Kette wird als **Burn-in-Phase** bezeichnet, welche verworfen wird, weil sich die Kette erst an die Zielverteilung anpassen muss. Der restliche Teil wird als **Sampling-Phase** bezeichnet, welche dann die Verteilung simuliert. [3, 226f.]

3 Praktischer Teil

3.1 Implementierung des Metropolis-Hastings-Algorithmus in Python

Der Metropolis-Hasting-Algorithmus erstellt mit Hilfe der Posteriorifunktion eine Markov-Kette. Dafür wird die Normalverteilung als Hilfsfunktion verwendet, da sie ähnlich der Posteriorifunktion ist, welche in unserem Beispiel verwendet wird. Als μ wird der aktuelle Wert verwendet und als σ^2 wird 0.1 gewählt. Das σ beeinflusst die Schrittweite, ist es hoch werden mit höherer Wahrscheinlichkeit vom aktuellen Wert weiter entfernte Werte gewählt. Da in unserem Beispiel eine Wahrscheinlichkeit geschätzt wird muss beachtet werden, dass neue Werte immer zwischen 0 und 1 liegen. Der folgende Python Code zeigt die Implementierung des Metropolis-Hastings-Algorithmus.

```
1  def metropolis_hastings(x: int, n: int, n_samples: int, start: float
       = 0.5, proposal_width: float = 0.1) -> list:
2      samples = [] # Liste zur Speicherung der Stichproben (Markov-
         Kette)
3      current_p = start # Startwert fuer die Markov-Kette
4
5      for _ in range(n_samples):
6
7          # Vorschlag fuer einen neuen Wert basierend auf einer
             Normalverteilung um den aktuellen Wert
8          proposed_p = np.random.normal(current_p, proposal_width)
9
10         # Ablehnung des Vorschlags, falls er ausserhalb des
             Wertebereichs [0,1] liegt
11         if proposed_p < 0 or proposed_p > 1:
12             continue
13
14         # Berechnung der Akzeptanzwahrscheinlichkeit
15         current_posterior = posterior(current_p, x, n)
16         proposed_posterior = posterior(proposed_p, x, n)
17         acceptance_ratio = proposed_posterior / current_posterior
18
19         # Akzeptiere den neuen Vorschlag mit der berechneten
             Wahrscheinlichkeit
20         if np.random.rand() < acceptance_ratio:
21             current_p = proposed_p # Aktualisiere den aktuellen Wert
22
23         samples.append(current_p) # Speichere den aktuellen Wert in
             der Stichprobenliste
24
25     return samples[1000:] # Entferne die ersten 1000 Werte (Burn-in
         -Phase)
```

3.2 Anwendung der bayesianischen Statistik am Beispiel

3.2.1 Überblick und Aufbereitung des Datensatzes

Für den praktischen Teil der Arbeit nutzen wir einen CSV-Datensatz mit detatailierten Informationen zu Weltmeisterschaft-Fußballspielen. Dieser Datensatz wurde von der Plattform Kaggle heruntergeladen und ist unter folgendem Link zu finden: `https://www.kaggle.com/datasets/piterfm/fifa-football-world-cup`[1].

Der Datensatz enthält alle WM-Spiele zwischen 1930 und 2022 mit Informationen über die Manschaften, Ergebnisse, Austragungsorte, Spielereignisse (Auswechslungen etc.), Schiedrichterdaten und so weiter. Im ersten Teil wollen wir aber nur die Siegesquote der deutschen Nationalmannschaft analysieren. Dafür haben wir in `../src/world_cup_wins_germany.py` berechnet, wie viele Spiele die deutsche Nationalmannschaft gespielt und wie viele davon gewonnen hat. Die Anzahl gewonnener Spiele beträgt dabei 72 von insgesamt 112 Spielen.

Im späteren Verlauf unserer Untersuchungen wollen wir auch die Siegesquoten gegen europäische Teams mit denen gegen nicht-europäische Teams vergleichen. Dafür haben wir auch die Anzahl der gewonnenen Spiele gegen europäische und nicht-europäische Teams berechnet. Die Anzahl der gewonnenen Spiele gegen europäische Teams beträgt dabei 36 von insgesamt 65 Spielen und die Anzahl der gewonnenen Spiele gegen nicht-europäische Teams beträgt 36 von insgesamt 47 Spielen.

3.2.2 Anwendung der bayesianischen Statistik

Zuerst wollen wir die Siegesquote der deutschen Nationalmannschaft mit Hilfe des Metropolis-Hastings-Algorithmus schätzen. Als Vorwissen gehen wir dabei von einer um $\mu = 0.5$ normalverteilten Siegesquote mit einer Varianz von $\sigma^2 = 0.1$ aus, was somit unsere Prioiriverteilung ist. In die Likelihood Funktion fließt jetzt der gegebene Datensatz ein. Die Likelihood-Funktion ist somit eine Binomialverteilung mit den Parametern $x = 72$ und $n = 112$ ohne den Normierungsfaktor $\binom{n}{x}$. Die Posteriorifunktion ergibt sich aus der Multiplikation der Likelihood-Funktion und der Prioriverteilung. Dadurch ergeben sich folgende Python Funktionen, welche die Wahrscheinlichkeit für einen gegebenen Wert liefern.

```
1 def likelihood(p: float, x: int, n: int) -> float:
2     return (p**x) * ((1 - p)**(n - x))
3 def prior(p: float) -> float:
4     return stats.norm.pdf(p, 0.5, 0.1)
5 def posterior(p: float, x: int, n: int) -> float:
6     return likelihood(p, x, n) * prior(p)
```

Ist n die Gesamtanzahl an Spielen und x die Anzahl an Siegen ist die Prioriverteilung $P(p)$ somit mathematisch definiert als:

$$P(p) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right) \tag{25}$$

Und die Likelihood-Funktion $P(n, x|p)$ als:

$$P(n, x|p) = p^x(1-p)^{n-x} \tag{26}$$

Um nun den Normierungsfaktor P(n,x) der Posterioriverteilung zu berechnen zu können müsste man folgendes Integral berechnen:

$$P(n, x) = \int_0^1 (p^x(1-p)^{n-x}) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right) dp \tag{27}$$

Da dieses Integral analytisch schwer zu lösen ist, können wir unseren Metropolis-Hastings-Algorithmus aus 3.1 verwenden, um eine Markovkette zu erstellen, die die Posterioriverteilung approximiert. Starten wir nun den Algorithums mit den Werten n = 112, x = 72 so erhalten wir folgende Posterioriverteilung:

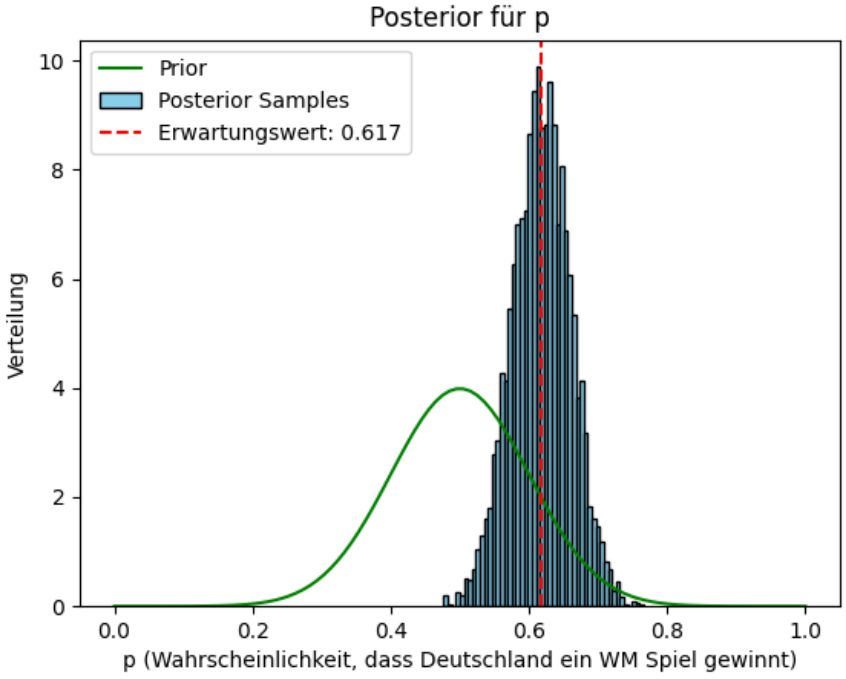
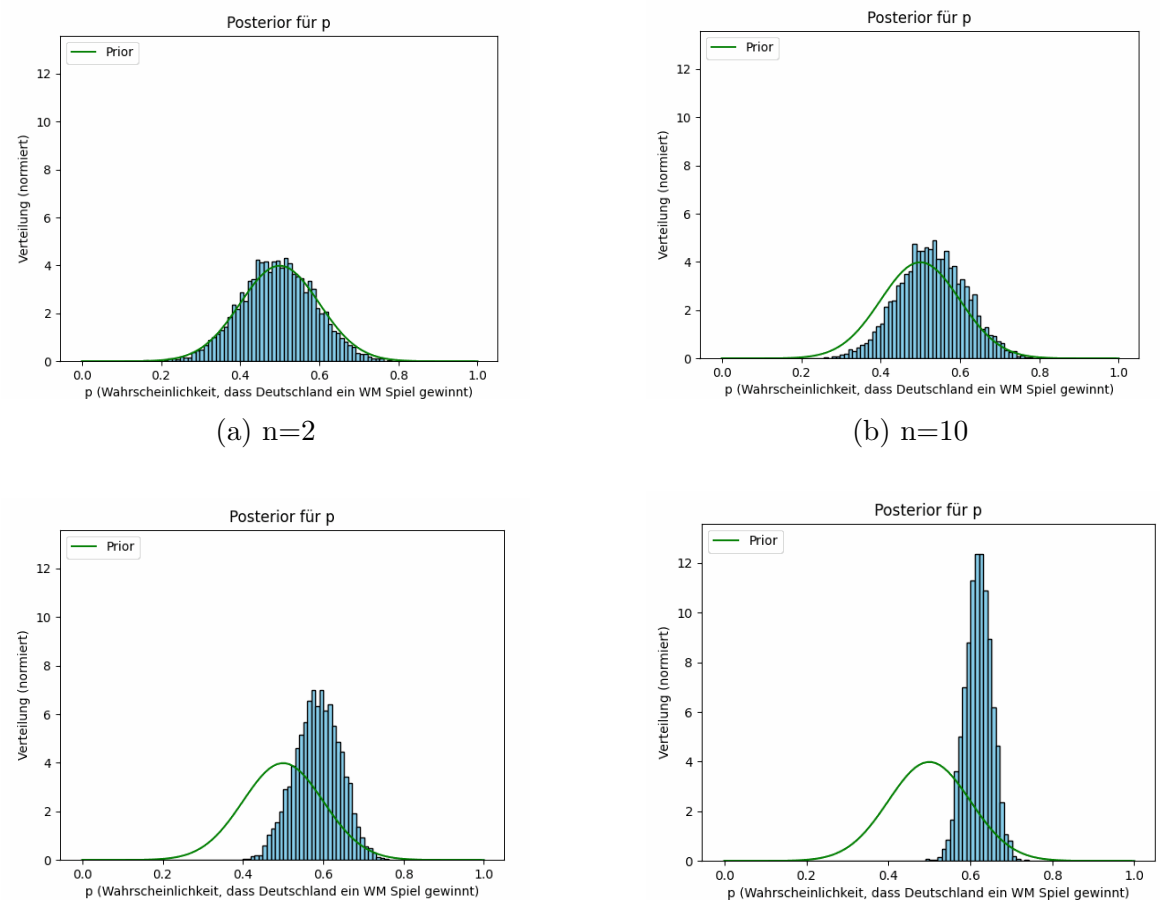


Abbildung 1: Posterioriverteilung der Siegesquote der deutschen Nationalmannschaft

In der Grafik sieht man, die zuvor angenommene Prioriverteilung und die Posterioriverteilung welche eine geringere Varianz besitzt und um einiges weiter rechts liegt als die Prioriverteilung. Es ist deutlich zu erkennen, dass die Siegesquote deutlich höher als die zuvor angenommenen 50% liegt. Der Erwartungswert der Posterioriverteilung beträgt ungefähr 0.615.

Wie stark die Posterioriverteilung nach rechts verschoben wird und wie stark die Varianz abnimmt, hängt von der Größe der Stichprobe ab. Je größer die Stichprobe, desto stärker wird die Posterioriverteilung verschoben und desto geringer wird die Varianz. Diesen Prozess haben wir durch `../src/showcases/binomStatisticsAnimated.py` simuliert und im Ergebniss ist die erwartete Verschiebung deutlich zu erkennen:

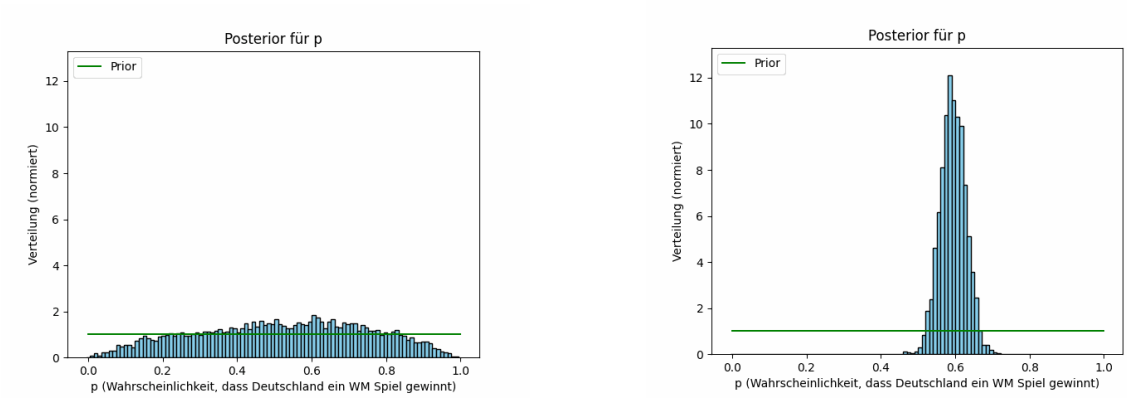


(a) n=2 (b) n=10

(c) n=50 (d) n=200

Abbildung 2: Posterioriverteilung für verschiedene Stichprobengrößen

Der Einfluss der Prioriverteilung auf die Posterioriverteilung nimmt mit steigender Stichprobengröße somit immer weiter ab. Das kann man auch gut in folgender Grafik erkennen, in welcher die Gleichverteilung auf [0,1] als Prioriverteilung verwendet wird. Es ist zu erkennen, dass die Posterioriverteilung sich immer weiter der oben abgebildeten Posterioriverteilung annähert also normalverteilt wird, obwohl die Prioriverteilung eine Gleichverteilung ist.



(a) n=2 (b) n=200

Abbildung 3: Posterioriverteilung für verschiedene Stichprobengrößen mit Gleichverteilung als Prioriverteilung

3.2.3 Statistische Auswertung der Ergebnisse

Im weiteren Wollen wir nun noch die Konfidenzintervalle und den Punktschätzer der Markovkette berechnen und diese mit den Ergebnissen der klassischen Statistik vergleichen. Wir gehen dabei wieder von der Stichprobengröße n=112 aus. Für das Konfidenzintervall in der klassischen Statistik nutzen wir das exakte Konfidenzintervall für eine unbekannte Wahrscheinlichkeit p. Dieses ist definiert als:

$$I_e = \left[\frac{S_n F_1}{n - S_n + 1 + S_n F_1}, \frac{(S_n + 1) F_2}{n - S_n + (S_n + 1) F_2} \right] \tag{28}$$

wobei F_1 das Quantil der F-Verteilung mit $\alpha/2$ und den Freiheitsgraden $2S_n$ und $2(n - S_n + 1)$ ist und F_2 das Quantil der F-Verteilung mit 1-$\alpha/2$ und den Freiheitsgraden $2(S_n + 1)$ und $2(n - S_n)$ ist. Sn ist dabei die Anzahl der gewonnenen Spiele. Für die Berechnung haben wir das folgende Python-Programm geschrieben:

```
1  # Berechnung der Bayes'schen Schaetzung und des 95%-
       Konfidenzintervalls
2  mean_bayes = np.mean(samples)
3  konf_bayes_low, konf_bayes_high = np.percentile(samples, [2.5,
       97.5])
4
5  # Berechnung der klassischen Schaetzung und des
       Konfidenzintervalls auf Basis der F-Verteilung
6  mean_classical = x / n
7  sn = x
8  f1 = stats.f.ppf(0.025, 2*sn, 2*(n-sn+1))
9  f2 = stats.f.ppf(0.975, 2*(sn+1), 2*(n-sn))
10 konf_classical_low = (sn*f1)/(n-sn+1+sn*f1)
11 konf_classical_high = ((sn+1)*f2)/(n-sn+(sn+1)*f2)
```

Dieses liefert den Punktschätzer 0.617 und das Konfidenzintervall [0.532, 0.7] mit der bayesianischen Statistik und 0.643 und [0.547,731] für die klassische Statistik.

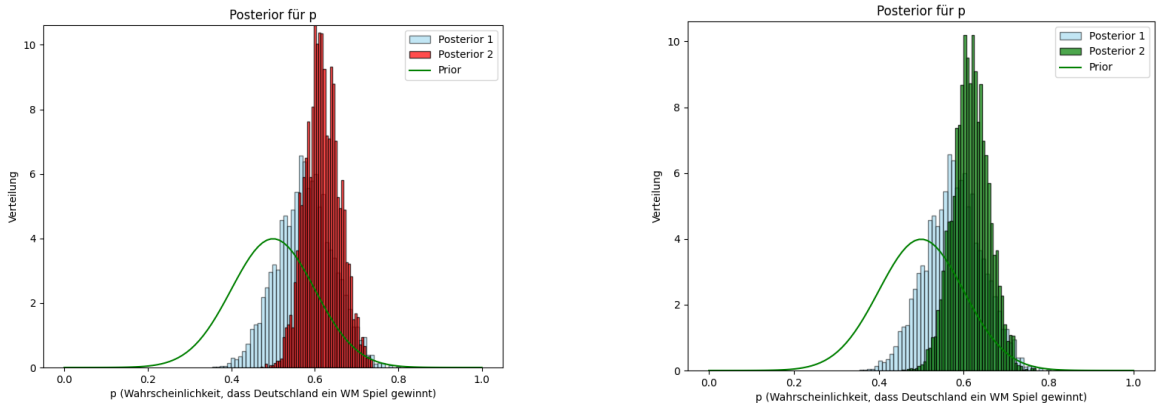
Es fällt auf, dass in der klassischen Statistik die geschätzen Werte für p höher sind als mit der Bayes'schen Statistik. Das liegt ganz einfach daran, dass die Verteilung in der bayesianischen Statistik durch die Prioriverteilung noch leicht nach in Richtung 0.5 gezogen wird. Man kann somit gut sehen, dass sich die Werte je nach Berechnung unterscheiden können und die Ergebnisse mit klassischer und bayesianischer Statistik nicht die gleichen sein müssen. Der große Vorteil der bayesianischen Statistik ist aber, dass man viel mehr Informationen über die Schätzung von p in Form der Posterioriverteilung erhält.

3.2.4 Rekursives Einbeziehen von neuen Daten

Weiter wollen wir überprüfen, wie sehr sich die Posterioriverteilungen unterscheiden, wenn man immer wieder neue Daten erhält. Die erste Möglichkeit wäre, dass man seine Posterioriverteilung als Prioiriverteilung verwendet und so eine neue Posterioriverteilung mit den neuen Daten als Likelihood berechnen kann. Eine weitere Möglichkeit ist es nochmal eine neue Posterioriverteilung zu berechnen mit der schon davor verwendeten Prioiriverteilung und der aktualisierten Stichprobe als Likelihood. Dafür haben wir das Python-Programm `../src/showcases/binomStatisticsRecursiv.py` erstellt, welche die beiden Varianten darstellt. Folgende Funktion simuliert die Prioriverteilung mit der Stichprobe der vorherigen Posterioriverteilung:

```
1 def prioriSamples(p: float, samples: list) -> float:
2     delta = 1e-3
3     samples = np.asarray(samples)
4     close_values = np.abs(samples - p) < delta
5     return np.mean(close_values)
```

Da es sich um eine Stichprobe handelt wird die Schätzung pro Iteration immer ungenauer, da man nicht die genaue Verteilung kennt. Ansonsten sieht man aber nicht so große Unterschiede zwischen den beiden Varianten, wie in der folgenden Grafik zu sehen ist:



(a) Posteriori als Priori (b) Zusammengefasste Stichprobe

Abbildung 4: Vergleich der 2 Varianten

3.3 Berechnungen für den Vergleich der Siegesrate gegen europäische und nicht-europäische Teams

Nachfolgend wollen wir die Siegesrate der deutschen Nationalmannschaft gegen europäische Teams mit der gegen nicht-europäische Teams vergleichen. Insbesondere wollen wir die Frage beantworten, ob die Siegesrate gegen europäische Teams signifikant niedriger ist als gegen nicht-europäische Teams. Für die nachfolgenden Berechnungen definieren wir somit:

- π_E als die Siegesrate der deutschen Nationalmannschaft gegen europäische Teams
- $\pi_{\neg E}$ als die Siegesrate der deutschen Nationalmannschaft gegen nicht-europäische Teams
- π_G als die Siegesrate der deutschen Nationalmannschaft insgesamt
- $H_0 : \pi_E \geq \pi_{\neg E}$ gegen $H_1 : \pi_E < \pi_{\neg E}$

Dabei gehen wir die Fragestellung mit klassischer sowie mit bayesianischer Statistik an.

Daten:

- Gegen europäische Teams wurden 65 Spiele gespielt, davon 36 gewonnen.
- Gegen nicht-europäische Teamsn wurden 47 Spiele gespielt, davon 36 gewonnen.
- Insgesamt wurden 112 Spiele gespielt, davon 72 gewonnen.

3.3.1 Berechnung mit klassischer Statistik

Punktschätzer:

$$\hat{\pi}_E = \frac{36}{65} = 0.5538$$

$$\hat{\pi}_{\neg E} = \frac{36}{47} = 0.7659$$

$$\hat{\pi}_G = \frac{72}{112} = 0.6429$$

Konfidenzintervall: Für $\alpha = 0.05$ beidseitige Konfidenzintervalle:

$$I_E = \left[0.5538 - 1.96 \cdot \sqrt{\frac{0.5538 \cdot (1 - 0.5538)}{65}},\ 0.5538 + 1.96 \cdot \sqrt{\frac{0.5538 \cdot (1 - 0.5538)}{65}}\right]$$

$$I_E = [0.4330, 0.6746]$$

$$I_{\neg E} = \left[0.7659 - 1.96 \cdot \sqrt{\frac{0.7659 \cdot (1 - 0.7659)}{47}},\ 0.7659 + 1.96 \cdot \sqrt{\frac{0.7659 \cdot (1 - 0.7659)}{47}}\right]$$

$$I_{\neg E} = [0.6448, 0.8870]$$

$$I_G = \left[0.6429 - 1.96 \cdot \sqrt{\frac{0.6429 \cdot (1 - 0.6429)}{112}},\ 0.6429 + 1.96 \cdot \sqrt{\frac{0.6429 \cdot (1 - 0.6429)}{112}}\right]$$

$$I_G = [0.5542, 0.7316]$$

Hypothesentest:
Es wird ein einseitiger z-Test der Differenz der Anteile durchgeführt.

Wir berechnen den Standardfehler für den Unterschied der Anteile:

$$SE = \sqrt{\hat{\pi}_G(1 - \hat{\pi}_G)\left(\frac{1}{65} + \frac{1}{47}\right)} = \sqrt{0.6429 \cdot (1 - 0.6429)\left(\frac{1}{65} + \frac{1}{47}\right)} \approx 0.0917$$

Berechnung des z-Werts:

$$z = \frac{\hat{\pi}_E - \hat{\pi}_{\neg E}}{SE} = \frac{0.5538 - 0.7659}{0.0917} \approx -2.31$$

Für ein Signifikanzniveau von $\alpha = 0.05$ und einen einseitigen Test ist der kritische z-Wert:

$$z_{\text{kritisch}} = -1.645$$

Da der berechnete z-Wert von -2.31 kleiner als der kritische z-Wert von -1.645 ist, können wir die Nullhypothese $H_0 : \pi_E \geq \pi_{\neg E}$ ablehnen.
Interpretation: Die Siegrate der deutschen Nationalmannschaft gegen europäische Teams ist signifikant niedriger als gegen nicht-europäische Teams.

3.3.2 Berechnung mit bayesianischer Statistik

Vorwissen: Für die bayesianische Analyse nehmen wir als Prior-Verteilungen Beta-Verteilungen für die Siegquote der deutschen Nationalmannschaft gegen europäische und nicht-europäische Teams an. Die Parameter dieser Verteilungen basieren auf den beobachteten Daten.

- Für europäische Teams:

  $$\alpha_E = 36 + 1 = 37, \quad \beta_E = 29 + 1 = 30$$

  Somit ist die Prior-Verteilung: Beta$(37, 30)$.

- Für nicht-europäische Teams:

  $$\alpha_{\neg E} = 36 + 1 = 37, \quad \beta_{\neg E} = 11 + 1 = 12$$

Somit ist die Prior-Verteilung: Beta$(37, 12)$.

Likelihood-Verteilung:
Für die gesamte Siegquote (Likelihood) wird ebenfalls eine Beta-Verteilung verwendet, basierend auf den gesamten Spielen der deutschen Nationalmannschaft:

$$\alpha_{\text{total}} = 72 + 1 = 73, \quad \beta_{\text{total}} = 40 + 1 = 41$$

Die Likelihood-Verteilung ist daher: Beta$(73, 41)$.

Posterior-Verteilung:
Die Posterior-Verteilung kann als Produkt von Prior und Likelihood berechnet werden. Sie wird durch Normierung auf den gesamten Bereich der Siegquote p bestimmt.

Die Verteilungen für die Prior-, Likelihood- und Posterior-Verteilungen werden wie folgt dargestellt:

- Prior für europäische Teams: Beta$(37, 30)$
- Prior für nicht-europäische Teams: Beta$(37, 12)$
- Likelihood für alle Spiele: Beta$(73, 41)$

Die Posterior-Verteilungen für europäische und nicht-europäische Teams werden durch das Produkt der Prior- und Likelihood-Verteilungen berechnet.

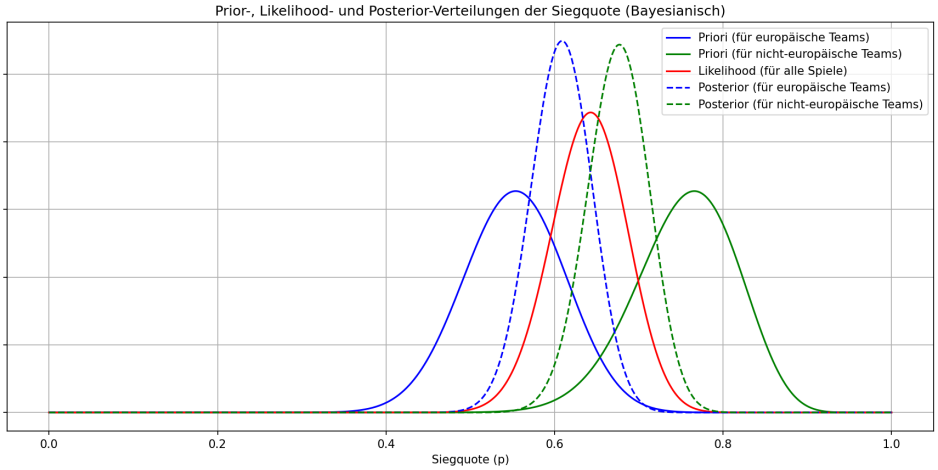


Abbildung 5: Prior-, Likelihood- und Posterior-Verteilungen der Siegquote der deutschen Nationalmannschaft

Punktschätzer:
Der Erwartungswert der Posterior-Verteilung ist der Mittelwert der Beta-Verteilung, der mit folgender Formel berechnet wird:

$$E = \frac{\alpha_{\text{posterior}}}{\alpha_{\text{posterior}} + \beta_{\text{posterior}}}$$

Für europäische Teams ist der Erwartungswert:

$$E_E = \frac{37 + 73}{37 + 73 + 30 + 41} = \frac{110}{181} \approx 0.6077$$

Für nicht-europäische Teams ist der Erwartungswert:

$$E_{\neg E} = \frac{37 + 73}{37 + 36 + 12 + 41} = \frac{110}{163} \approx 0.6748$$

Konfidenzintervall:
Das 95%-Konfidenzintervall für die Posterior-Verteilung wird durch die Quantilfunktion der Beta-Verteilung berechnet. Wir verwenden das 2.5%- und 97.5%-Quantil. Dies wurde mithilfe des Programms `world_cup_bayes.py` durchgeführt.
Für europäische Teams:

$$I_E = [0.5358, 0.6775]$$

Für nicht-europäische Teams:

$$I_{\neg E} = [0.6012, 0.7444]$$

Hypothesentest:
Um die Hypothese H_0 zu testen, dass die Siegquote gegen europäische Teams niedriger ist als gegen nicht-europäische Teams, wurden Monte-Carlo-Simulationen verwendet:

- Aus den Posterior-Verteilungen wurden 100.000 Zufallswerte (Samples) gezogen.
- Die Wahrscheinlichkeit $P(H_1) = P(\pi_E < \pi_{\neg E})$ wurde als der Anteil der Simulationen berechnet, bei denen $\pi_E < \pi_{\neg E}$.
- Die Wahrscheinlichkeit $P(H_0) = 1 - P(H_1)$ wurde ebenfalls berechnet.
- Je nachdem, welche Wahrscheinlichkeit größer ist, wurde entschieden, ob H_0 (Nullhypothese) oder H_1 (Alternativhypothese) akzeptiert wird.

Ergebnisse:
$P(H_0) = 0.0972$ und $P(H_1) = 0.9028$
Da $P(H_1) > P(H_0)$, wird die Nullhypothese $H_0 : \pi_E \geq \pi_{\neg E}$ abgelehnt.
Interpretation: Die Siegquote der deutschen Nationalmannschaft gegen europäische Teams ist signifikant niedriger als gegen nicht-europäische Teams.

4 Fazit

4.1 Vergleich der klassischen und bayesianischen Statistik

	Klassische Statistik	**Bayesianische Statistik**
Bevorzugte Verwendung	Human- und Sozialwissenschaften, Wirtschaftswissenschaften, Biologie	Technik und Künstliche Intelligenz
Schätzen von Parametern, Testen von Hypothesen	nur Stichprobe wird betrachtet	Stichprobe und Vorwissen wird betrachtet
Ergebnis beim Schätzen	Punkt- oder Intervallschätzer, z. B. Konfidenzintervall	Wahrscheinlichkeitsverteilung für den Parameter (Posterior)
Ergebnis beim Testen	p-Wert; Entscheidung basierend auf Signifikanzniveau, wird angenommen oder abgelehnt	Posterior-Wahrscheinlichkeit für Hypothesen

Tabelle 1: Vergleich der klassischen und bayesianischen Statistik

Die klassische Statistik und die bayesianische Statistik sind zwei unterschiedliche Ansätze zur Analyse von Daten und zur Entscheidungsfindung. Während die klassische Statistik vor allem in den Human- und Sozialwissenschaften sowie in der Wirtschaftswissenschaft und Biologie weit verbreitet ist, wird die bayesianische Statistik häufig in der Technik und im Bereich der Künstlichen Intelligenz eingesetzt.[5, S. 2]

Ein wesentlicher Unterschied zwischen den beiden Methoden liegt in der Art und Weise, wie Parameter geschätzt und Hypothesen getestet werden. Die klassische Statistik betrachtet dabei ausschließlich die vorliegende Stichprobe, während die bayesianische Statistik sowohl die Stichprobe als auch vorheriges Wissen in die Analyse einbezieht.

Auch die Ergebnisse unterscheiden sich: In der klassischen Statistik erfolgt die Schätzung von Parametern durch Punkt- oder Intervallschätzer, beispielsweise in Form eines Konfidenzintervalls. Die bayesianische Statistik hingegen liefert eine Wahrscheinlichkeitsverteilung für den gesuchten Parameter, die sogenannte Posterior-Verteilung.

Beim Testen von Hypothesen arbeitet die klassische Statistik mit dem p-Wert und einer Entscheidungsregel basierend auf einem festgelegten Signifikanzniveau. Eine Hypothese wird entweder ganz angenommen oder abgelehnt. [5, S. 2] Die bayesianische Statistik hingegen bewertet Hypothesen anhand der Posterior-Wahrscheinlichkeit und verwendet den Bayes-Faktor zur Entscheidungsfindung, wodurch ein flexiblerer und probabilistischerer Ansatz ermöglicht wird.

Diese Unterschiede wurden in der vorliegenden Arbeit besonders anhand der Berechnung der Fußball-WM im Abschnitt 3.3 deutlich.

Statistikmethode	Vorteile	Nachteile
Klassische Statistik	• Einfachheit • Gut für große Datensätze • keine subjektiven Annahmen	• abhängig von großen Stichproben • unflexibel gegenüber neuen Daten oder komplexen Datenstrukturen
Bayesianische Statistik	• flexibel und anpassbar an komplexe Modelle. • ermöglicht die Einbeziehung von Vorwissen • liefert direkte Wahrscheinlichkeiten für Parameter und Modelle	• Subjektivität bei der Wahl der Priors. • rechenintensiv, besonders bei großen Datensätzen.

Tabelle 2: Vergleich der Vor- und Nachteile der klassischen und bayesianischen Statistik.

Die klassische (frequentistische) Statistik ist einfach anzuwenden und weit verbreitet, da sie auf standardisierten Testverfahren basiert. Im Gegensatz dazu ist die bayesianische Statistik komplexer und erfordert einen höheren Rechenaufwand, insbesondere bei großen Datensätzen. Dies wurde deutlich in der vorliegenden Arbeit, da aus dem gewählten Datensatz die Siegesquoten ermittelt werden mussten.

Ein Vorteil der bayesianischen Statistik ist ihre Flexibilität und die Möglichkeit, Vorwissen in Form von Priors einzubeziehen. Hier konnten wir z.B. das Vorwissen einmal in Form einer Normalverteilung und einmal in Form einer Gleichverteilung einbeziehen. Dagegen ist die klassische Statistik unflexibel gegenüber neuen Daten und Modellen, da sie stark von großen Stichproben abhängt.

Die klassische Statistik benötigt keine subjektiven Annahmen über Wahrscheinlichkeiten, wodurch sie als objektiver gilt. Demgegenüber steht die bayesianische Statistik, bei der die Wahl der Priors subjektiv sein kann und dadurch die Ergebnisse beeinflusst. So kann allerdings mit der Wahl von viel Vorwissen die Schätzung verbessert werden. [5, S. 127]

Während die klassische Statistik gut für große Datensätze geeignet ist, können ihre Ergebnisse bei komplexen Datenstrukturen schwer interpretierbar sein. Besonders bei sehr wenigen vorliegenden Daten ist die klassische Statistik kaum aussagekräftig. Die bayesianische Statistik hingegen liefert direkte Wahrscheinlichkeiten für Parameter und

Modelle, was eine intuitivere Interpretation ermöglicht.

Zusammenfassend bietet die klassische Statistik eine einfache und weithin akzeptierte Methode zur Datenanalyse, ist aber in ihrer Anpassungsfähigkeit begrenzt. Die bayesianische Statistik ist flexibler und anpassungsfähiger, erfordert jedoch mehr Rechenleistung und die sorgfältige Wahl von Priors. Die Entscheidung für eine Methode sollte daher je nach Anwendungsfall getroffen werden.

5 Benutzerdokumentation

5.1 Voraussetzungen

Für die Ausführung der Python-Programme wurde Version 3.12 verwendet. Zusätzlich sind die folgenden Bibliotheken erforderlich:

- numpy
- matplotlib
- scipy
- sys
- csv

5.2 Starten der Programme

Zum Ausführen der Programme muss man sich im richtigen Verzeichnis befinden. Die Programme `binomStatistics.py`, `worldCupWinsGermany.py` und `worldCupBayes.py` befinden sich im Verzeichnis `src`. Alle übrigen Programme müssen aus dem Verzeichnis `src/showcases` ausgeführt werden. Die Programme werden über die Konsole mit den folgenden Befehlen gestartet.

binomStatistics.py

Dieses Programm ist unserer Hauptprogramm, welches mit Eingabe einer Stichprobe die Posterioriverteilung berechnet und weitere Auswertungen ausgibt. Als Priorverteilung wird die Normalverteilung mit $\mu = 0.5$ und $\sigma^2 = 0.1$ verwendet. Die Ausführung erfolgt mit folgendem Befehl:

```
1 python binomStatistics.py <Anzahl der Versuche> <Anzahl Erfolge>
```

worldCupWinsGermany.py

Dieses Programm analysiert die bereitgestellten CSV-Daten und berechnet die Anzahl der von der deutschen Nationalmannschaft gewonnenen Spiele. Es wird mit folgendem Befehl gestartet:

```
1 python worldCupWinsGermany.py
```

worldCupBayes.py

Dieses Programm führt die in Abschnitt 3.3 beschriebenen Berechnungen durch. Es wird mit folgendem Befehl gestartet:

```
python worldCupBayes.py
```

binomStatisticsUniform.py/binomStatisticsAnimated.py

Diese beiden Programme demonstrieren den Einfluss der Stichprobengröße auf die Posteriorverteilung:

- `binomStatisticsUniform.py` verwendet eine Gleichverteilung als Priorverteilung.
- `binomStatisticsAnimated.py` nutzt eine Normalverteilung als Priorverteilung.

Die Programme werden mit folgendem Befehl gestartet:

```
python binomStatisticsUniform.py <Ratio der Erfolge zu
   Gesamtversuchen>
```

binomStatisticsRecursiv.py

Dieses Programm demonstriert das rekursive Einbeziehen neuer Daten in die Analyse. Die Posteriorverteilung der ersten Stichprobe wird als Priorverteilung für die Berechnung der Posteriorverteilung der nächsten Stichprobe genutzt. Anschließend wird diese resultierende Posteriorverteilung mit einer Posteriorverteilung verglichen, die beide Stichproben zusammenfasst. Die Ausführung erfolgt durch den Befehl:

```
python binomStatisticsRecursiv.py
```

6 Literaturverzeichnis

Literatur

[1] Piter FM. *FIFA Football World Cup Dataset*. Zugriff am: 14. Dezember 2024. 2023. URL: https://www.kaggle.com/datasets/piterfm/fifa-football-world-cup.

[2] Karl-Rudolf Koch. *Einführung in die Bayes-Statistik*. Berlin [u.a.]: Springer, 2000. ISBN: 3540666702. URL: https://katalog.slub-dresden.de/id/0-306244284.

[3] Dirk P. Kroese, Thomas Taimre und Zdravko I. Botev. *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley, 2011. ISBN: 9780470177938. URL: https://www.wiley.com/en-us/Handbook+of+Monte+Carlo+Methods-p-9780470177938.

[4] Thomas Müller-Gronbach, Erich Novak und Klaus Ritter. *Monte Carlo-Algorithmen*. Berlin: Springer, 2012. ISBN: 9783540891406. URL: https://katalog.slub-dresden.de/id/0-618339728.

[5] Wolfgang Tschirk. *Statistik: Klassisch oder Bayes zwei Wege im Vergleich*. Berlin , , © 2014. ISBN: 3642543847. URL: https://katalog.slub-dresden.de/id/0-160866449X.