Analysemethoden für nicht-experimentelle Forschungsdesigns

Mirjam Laager

2026-08-24

Inhaltsverzeichnis

Einleitung **2**
Experimentelle und nicht-experimentelle Designs . . . 2
Beispiele nicht-experimenteller Daten . . . 3
Schulleistung im Jugendalter . . . 3
Familienstand und depressive Symptome . . . 4
Zufällige und systematische Fehler . . . 4
Die fehlenden Flugzeuge . . . 4
Systematische Fehler im Studienverlauf . . . 5
Attrition Bias . . . 5
Randomisierung und Modellierung . . . 6
Ungleichgewicht in Kovariablen bei Studienbeginn . . . 6
Ereignisse im Studienverlauf . . . 6

Bayesianische Statistik **8**
Geschichtlicher Hintergrund . . . 8
Bedingte Wahrscheinlichkeit . . . 9
Der Satz von Bayes . . . 10
Ein Anwendungbeispiel . . . 10
Vom Satz von Bayes zur bayesianischen Statistik . . . 11
Abgrenzung von frequentistischer Statistik . . . 11
Priorverteilung, Likelihoodfunktion und Posteriorverteilung . . . 12
Die Priorverteilung . . . 13
Die Likelihoodfunktion . . . 14
Die Posteriorverteilung . . . 16

Literatur **17**

> **Hinweis**
>
> Dieses Skript entwickelt sich mit der Vorlesung weiter und wird bis zur nächsten Durchführung im FS 27 laufend ergänzt.

Einleitung

Viele klassische statistische Verfahren gehen auf die Analyse randomisiert kontrollierter Studien zurück. So wurde zum Beispiel die Varianzanalyse (ANOVA) zu Beginn des 20. Jahrhunderts im Zusammenhang mit landwirtschaftlichen Feldexperimenten in Rothamsted in England entwickelt (Parolini 2015). Die zu analysierenden Daten stammten also aus Experimenten, in denen beispielsweise der Einfluss von verschiedenen Düngemitteln auf das Wachstum von Hafer untersucht wurde (Fisher 1926).

Bei solchen Daten gibt es häufig grosse Effekte mit wenig Streuung. Zudem sind die Bedingungen der Datenerhebung weitgehend bekannt. Die unabhängige Variable wird in der Regel gezielt manipuliert und der Einfluss der Störvariablen wird durch Randomisierung kontrolliert. Unter diesen Voraussetzungen lassen sich die statistischen Modelle eng auf das Design abstimmen. Nicht-experimentelle Daten entstehen demgegenüber unter deutlich weniger kontrollierten Bedingungen. Die unabhängige Variable wird nicht manipuliert, sondern lediglich beobachtet; deshalb spricht man auch von Beobachtungsstudien. Wo eine Randomisierung nicht möglich ist, fehlt das wichtigste Instrument zur Kontrolle unbekannter Störvariablen, weshalb es viel schwieriger ist, einen belastbaren Kausalschluss zu ziehen. Und das Wissen darüber, wie genau die Daten zusandegekommen sind, ist häufig unvollständig. Es gibt jedoch sehr viele spannende Forschungsfragen, die nicht im Rahmen von randomisiert kontrollierten Studien untersucht werden können. Um auch solche Fragestellungen evidenzbasiert bearbeiten zu können, ist es sinnvoll Analysemethoden für nicht-experimentelle Designs zu kennen.

Experimentelle und nicht-experimentelle Designs

Experimentelle und nicht-experimentelle Designs unterscheiden sich grundlegend in Bezug auf die unabhängige Variable. In experimentellen Studien wird sie zugewiesen; dadurch werden Manipulation und Kontrolle möglich. Durch die Kontrolle der bekannten und unbekannten Störvariablen kann eine hohe interne Validität erreicht werden - eine Grundbedingung für einen eindeutigen Kausalschluss. In nicht-experimentellen Designs wird die unabhängige Variable dagegen beobachtet. Kontrolle ist nur eingeschränkt möglich, und wichtige Unterschiede zwischen den Vergleichsgruppen können bereits vorliegen, bevor überhaupt Daten erhoben werden.

Nicht-experimentelle Designs lassen sich danach klassifizieren, wie sie mit Zeit umgehen. In Kohortenstudien wird zunächst die unabhängige Variable erhoben und die abhängige Variable später beobachtet. In Fall-Kontroll-Studien beginnt die Untersuchung mit der abhängigen Variable und rekonstruiert die unabhängige Variable rückblickend. In Querschnittstudien werden die unabhängige und die abhängige Variable gleichzeitig erfasst. Natürlich kann man nicht-experimentelle Designs auch nach anderen Aspekten klassifizieren. Die zeitliche

Abfolge ist aber insofern besonders relevant, als Zeitlichkeit eines der neun Hill Kriterien für Kausalzusammenhänge ist. Zeitlichkeit als Kriterium für Kausalschlüsse bedeutet, dass eine vermutete Ursache der Wirkung vorausgehen muss.

Abgesehen von der Randomisierung haben experimentelle Designs noch weitere Vorteile gegenüber nicht-experimentellen Designs: Da Datenerhebung und Analyse eng aufeinander abgestimmt sind, können Effekte und die damit verbundene Unsicherheit präziser geschätzt werden. Zudem liegen in der Regel genügend Informationen darüber vor, wie genau die Daten erhoben worden sind, sodass eine Studie unter gleichen oder ähnlichen Bedingungen wiederholt werden kann. Das hohe Mass an Kontrolle über die Bedingungen kann aber auch ein Nachteil sein, da die Verallgemeinerbarkeit teilweise eingeschränkt ist, etwa wenn die Ein- und Ausschlusskriterien sehr streng gewählt werden müssen. Hier zeigt sich ein Vorteil nicht-experimenteller Desings: ihre Nähe zu realen Bedingungen. Zudem gibt es Fragestellungen, die aus praktischen oder ethischen Gründen nicht im Rahmen einer experimentellen Studie untersucht werden können. Bei nicht-experimentellen Designs ist es hingegen möglich, bereits vorhandene Daten zu nutzen (zum Beispiel aus Registern). Dass die Analyse nicht zwingend durch das Design festgelegt ist, kann sowohl ein Vor- als auch ein Nachteil sein. Wenn diese Flexibilität durch sorgfältige Sensitivitätsanalysen und eine explizite Berücksichtigung der Modellunsicherheit ergänzt wird, wird daraus eine echte Stärke nicht-experimenteller Designs.

Beispiele nicht-experimenteller Daten

Die folgenden Beispiele zeigen, wie spannend die Fragestellungen sein können, die im Rahmen nicht-experimenteller Designs untersucht werden. Sie lassen sich grob dem Feld der psychologischen Forschung zuordnen, wobei nicht alle Variblen im engeren Sinne "psychologisch" sind. Mehr zur Unterscheidung von "psychologischen" und "nicht psychologischen" Variablen, die im Abschnitt zur kausalen Inferenz relevant werden wird, finden Sie hier.

Schulleistung im Jugendalter

In einer norwegischen Studie wurde der Zusammenhang zwischen schulischer Leistung in der frühen Lebensphase und vorzeitiger Sterblichkeit untersucht (Reme u. a. 2025). Dazu wurden Daten von insgesamt fast einer Million junger Menschen aus 17 Geburtenkohorten (Jahrgänge 1986 bis 2002) analysiert. Die unabhängige Variable war der Notendurchschnitt im Alter von 16 Jahren, die abhängige Variable war vorzeitges Versterben im Alter von 16 bis 30 Jahren. Das tiefste Quartil der Notendurchschnitte wies ein klar erhöhtes Risiko auf. Die Forschungsfrage dieser Studie kann nur in einem nicht-experimentellen Design beantwortet werden, weil die unabhängige Variable nicht manipuliert werden kann und weil um ein so seltenes Ereignis zu untersuchen eine sehr grosse Datenmenge nötig ist. Eine grosse Datenmenge lag in diesem Fall in Form eines Registers bereits vor, die Daten mussten also nicht spezifisch für diese Studie erhoben werden. Dass für die Analyse bereits vorhandene Daten genutzt werden konnten, hat allerdings auch einen Nachteil: Insgesamt sind in dem Register ca. 1.5 Millionen junge Erwachsene abgebildet, von denen aus verschiedenen Gründen (z.B. Wegzug ins Ausland), ca 500'000 nicht analysiert werden konnten. Unterscheiden sich diese ev. systematisch von den Personen, die im Sample verblieben sind? Bei solchen Fragen ist das zum Teil lückenhafte Wissen um den datengenerierenden Prozess eine Einschränkung. Die Autoren der Studie halten

unter anderem deshalb auch fest: “The main limitation of our study is that our results do not allow for causal interpretation.”

Familienstand und depressive Symptome

In einer länderübergreifenden Studie wurde der Zusammenhang zwischen Familienstand und depressiven Symptomen untersucht (Zhai u. a. 2024). Analysiert wurden Daten von ca. 100’000 Personen aus sieben Ländern. Die unabhängige Variable war der Familienstand, die abhängige Variable depressive Symptome. Bei dieser Studie handelt es sich um eine Querschnittstudie, d.h. die abhängige und die unabhängige Variable wurden gleichzeitig erfasst. Über alle Länder hinweg zeigte sich, dass unverheiratete Personen ein höheres Risiko für depressive Symptome hatten als verheiratete. Auch diese Fragestellung kann nur in einem nicht-experimentellen Design untersucht werden, da der Familienstand nicht randomisiert werden kann. Und auch in diesem Fall war die ursprüngliche Studienpopulation mit rund 150’000 Personen deutlich grösser als das letztlich analysierte Sample. Ein Teil der Personen wurde unter anderem deshalb nicht berücksichtigt, weil Angaben zu depressiven Symptomen fehlten. Es ist naheligend, dass sich gerade diese Personen systematisch von denjenigen Personen unterscheiden, die schlussendlich analysiert wurden. Unter anderem deshalb kann nicht ausgeschlossen werden, dass es noch unbekannte Störvariablen gibt, eine Limitation, die die Autoren so formulieren: “because this is an observational study, residual confounding […] cannot be ruled out.”

Zufällige und systematische Fehler

In einer Studie können zufällige und systematische Fehler auftreten. Zufällige Fehler enstehen durch Schwankungen der Messwerte um den wahren Wert. Es ist zu erwarten, dass Messwerte jeweils nicht exakt dem wahren Wert entsprechen, sondern um diesen herum streuen. Zufällige Fehler bilden die Unsicherheit in der Schätzung ab, die mit zunehmender Stichprobengrösse in der Regel kleiner wird. Demgegenüber stehen, und das ist bei nicht-experimentellen Designs oft ein Stolperstein, systematische Fehler. Systematische Fehler (biases) führen zu einer Verzerrung der Schätzung. Im Gegensatz zu zufälligen Fehlern streuen die Messwerte dabei nicht um den wahren Wert, sondern weichen systematisch davon ab. Ein systematischer Fehler lässt sich nicht durch eine Vergrösserung der Stichprobe reduzieren.

Die fehlenden Flugzeuge

Eine bekannte Geschichte zum Thema Bias dreht sich um Einschusslöcher in Flugzeugen. Die folgende Kurzfassung lehnt sich an die Darstellung bei Ellenberg (Ellenberg 2015) an.

Im zweiten Weltkrieg gab es in Manhattan eine “statistical research group”, welche die Armee unterstützte indem sie Daten analysierte. Eines Tages wurde einem Statistiker namens Abraham Wald ein Datensatz vorgelegt. Jemand hatte die Einschusslöcher in Flugzeugen, welche von Einsätzen zurückkamen, gezählt und in einer Tabelle zusammengefasst.

Teil des Flugzeugs	Einschusslöcher pro Quadratfuss
Motor	1.11
Rumpf	1.73
Kraftstoffsystem	1.55
Übrige	1.8

Es stellte sich die Frage, an welchen Stellen es sich lohnt, ein Flugzeug zu panzern. Das kann man nicht überall tun, sonst wird das Flugzeug zu schwer. Die naheliegende Überlegung wäre natürlich, das Flugzeug dort zu panzern, wo viele Einschüsslöcher sind, zum Beispiel am Rumpf. Walds Antwort war aber eine andere:

> "The armor, said Wald, doesn't go where the bullet holes are. It goes where the bullet holes aren't: on the engines."

An diesem Beispiel zeigt sich, wie wichtig das Wissen um den datengenerierenden Prozess für die Analyse ist. Es gibt hier zwei Geschichten zur Entstehung dieser Tabelle. Die erste Geschichte: aus irgend einem Grund schiessen Leute, die auf Flugzeuge schiessen, überdurchschnittlich oft auf den Rumpf. Dann lohnt es sich natürlich, den Rumpf zu panzern. Die zweite Geschichte: Die Daten, die hier vorliegen, sind keine zufällige Stichprobe. Da fehlen Flugzeuge. Und zwar überdurchschnittlich oft die, die am Motor getroffen wurden - diese schaffen es nicht mehr zurück. Die Einschusslöcher im Datensatz zeigen also gerade nicht die verwundbarsten Stellen, sondern die Bereiche, in denen Treffer häufig überlebt wurden. Das ist ein Beispiel für einen systematischen Fehler, ein sogenannter Survivor (oder Survivorship) Bias.

Systematische Fehler im Studienverlauf

Systematische Fehler können im gesamten Verlauf einer Studie auftreten: Bei der Literatursuche, bei der Auswahl der Teilnehmenden, beim Erfassen der Outcomes, bei der Datenanalyse und der Interpretation der Resultate und schlussendlich auch beim Publizieren der Resultate. Im Catalogue of Bias finden Sie eine grosse Sammlung verschiedener Biases mit jeweils einer kurzen Erklärung und Beispielen.

Attrition Bias

Ein häufiger Bias in Längsschnittstudien ist der sogenannte Attrition Bias. Gerade wenn Personen über einen langen Zeitraum beobachtet werden, kann es vorkommen, dass einzelne Teilnehmende die Studie vorzeitig verlassen (drop outs). Das ist nicht ungewöhnlich und wird bei der Studienplanung berücksichtigt. Ein Attrition Bias liegt vor, wenn sich Dropouts systematsich von den Personen unterscheiden, die bis zum Ende dabei bleiben. Ein Beispiel für Attrition Bias findet sich in einer Studie zu psychischer Belastung im Zusammenhang mit Spitalaufenthalten (Damen u. a. 2015). In dieser Studie füllten 1132 Patienten vor einer Operation einen Fragebogen zu Angstsymptomen und depressiven Beschwerden (Hospital Anxiety and Depression Scale) aus. Nach einem Jahr wurden sie erneut befragt. 71% der Teilnehmenden füllten den Follow-up Fragebogen aus. Die 29% Dropouts unterschieden sich hinsichtlich der zu Studienbeginn erfassten Variablen von den Personen mit vollständigem

Follow-up. So waren sie zum Beispiel im Durschnitt jünger und rauchten häufiger. Zudem waren die psychischen Beschwerden bei Studienbeginn unter den Dropouts stärker ausgeprägt. Würde man nun die Prävalenz nur auf Grundlage der Personen mit vollständigem Follow-up berechnen, fiele sie wahrscheinlich zu niedrig aus.

Randomisierung und Modellierung

Die in dieser Vorlesung behandelten statistischen Verfahren können auch für die Analyse experimenteller Daten nützlich sein. Das wollen wir zum Abschluss noch anhand von zwei Beispielen besprechen.

Ungleichgewicht in Kovariablen bei Studienbeginn

In einer randomisiert kontrollierten Studie wurde die Wirksamkeit von kognitiver Verhaltenstherapie bei der Behandlung von anhaltender Trauerstörung (prolonged grief disorder) untersucht und zwar konkret die Frage, ob die Therapie als Einzel- oder als Gruppentherapie gestaltet werden sollte (Komischke u. a. 2026). Zu Studienbeginn wurden verschiedene Variablen erfasst, zum Beispiel Einkommen, Ausbildung, Beziehung zur verstorbenen Person sowie die Zeit seit dem Verlust. Obwohl die Teilnehmenden nach dem Zufallsprinzip in die beiden Behandlungsgruppen eingeteilt worden waren, ergab es sich, dass in der Gruppe mit Einzeltherapie der Verlust im Schnitt 17.1 Monate zurück lag, während es in der Gruppe mit Gruppentherapie im Schnitt 41.5 Monate waren (siehe Tabelle 1 in (Komischke u. a. 2026)). Es liegt hier also ein Ungleichgewicht in den Kovariablen bei Studienbeginn (baseline imbalance) vor. Ein solches Ungleichgewicht stellt die Gültigkeit der Randomisierung zwar nicht grundsätzlich in Frage, da Randomisierung nicht auf perfekt ausbalancierte Gruppen abzielt, sondern auf eine unverzerrte Schätzung des Gruppenunterschieds. Trotzdem könnte man die Zeit seit dem Verlust in der Analyse berücksichtigen, da sie sehr wahrscheinlich einen wesentlichen Einfluss auf die Genesung hat. Die Frage ob, und wenn ja für welche Kovariablen adjustiert werden soll, ist also auch bei randomisiert kontrollierten Studien relevant. Einen Artikel, der in dieser Frage klar Position bezieht, finden Sie hier.

Ereignisse im Studienverlauf

In einer randomisiert kontrollierten Studie wurde die Wirkung verschiedener Antipsychotika untersucht (Tunis u. a. 2006). Zu Studienbeginn wurden 664 Personen in drei Gruppen randomisiert. Diese Personen wurden dann ein Jahr lang beobachtet, mit Follow-up Messungen nach 2 Wochen, 2 Monaten, 5 Monaten, 8 Monaten und 12 Monaten. Im Lauf dieser Zeit ist viel passiert: Manche Personen haben die zugeteilte Behandlung nie angetreten, andere haben die Therapie gewechselt, wieder andere haben die Studie frühzeitig verlassen. Beim letzten Follow-up waren nur noch 293 (44%) der Teilnehmenden in dem Arm, in den sie ursprünglich randomisiert worden waren (siehe Figure 1 in (Tunis u. a. 2006)). Solche Ereignisse im Studienverlauf (post randomisation events) sind nichts Aussergewönliches, sondern schlicht eine Folge davon, dass man Menschen über längere Zeit unter realen Bedingungen beobachtet. Es stellt sich jedoch die Frage, ob und wenn ja wie solche Ereignisse in der Analyse berücksichtigt werden sollen.

In einer Reanalyse der Studie wurde genau diese Frage untersucht (Hernán u. a. 2013). Grob gesagt wurden drei verschiedene Vorgehensweisen verglichen: Man kann so viele Teilnehmende wie möglich analysieren, das heisst, man schliesst alle Personen in die Analyse ein, unabhängig davon ob sie die zugeteilte Behandlung erhalten haben oder nicht (intention to treat analysis). Oder man kann sich auf diejenigen Teilnehmenden beschränken, die die Behandlung genau eingehalten haben (per protocol analysis). Die dritte in der Reanalyse beschriebene Methode versucht noch etwas anderes: Anhand der Baseline Variablen wird geschätzt, wie wahrscheinlich es ist, dass jemand die Behandlung abschliesst, und die Beobachtungen werden dann entsprechend gewichtet (inverse probability of censoring weights). Der Artikel zeigt auf, dass man je nach Vorgehen im Mittel unterschiedliche Ergebnisse bezüglich der Wirksamkeit der verschiedenen Behandlungsmethoden erhält. Im Abschnitt zur kausaler Inferenz werden wir uns ausführlicher mit den Vor- und Nachteilen dieser drei Vorgehensweisen befassen. Für den Moment halten wir fest:

> "In summary, the similarities between follow-up studies with and without baseline randomization are becoming increasingly apparent as more randomized trials study the effects of sustained interventions over long periods in real world settings. What started as a randomized trial may effectively become an observational study that requires analyses that complement, but go beyond, intention-to-treat analyses." (Hernán u. a. 2013)

Take-home messages

- Nicht-experimentelle Studien unterschieden sich von experimentellen Studien dadurch, dass die unabhängige Variable nicht zugewiesen sondern lediglich beobachtet wird.
- In experimentellen Studien sind Design und Analyse aufeinander abgestimmt. In nicht-experimentellen Studien ist der datengenerierende Prozess nicht vollständig kontrollierbar und oft auch nicht vollständig bekannt. Deshalb sind andere Analysemethoden nötig.
- Es gibt viele spannende Forschungsfragen, die nicht im Rahmen von experimentellen Studien untersucht werden können. Um auch solche Fragestellungen evidenzbasiert bearbeiten zu können, ist es sinnvoll Analysemethoden für nicht-experimentelle Designs zu kennen.
- Auch experimentelle Studien können nicht alles randomisieren und kontrollieren. Deshalb sind die in dieser Vorlesung behandelten Analysemethoden unter Umständen auch für experimentelle Studien nützlich.

Bayesianische Statistik

Geschichtlicher Hintergrund

Was wir heute als bayesiansische Statistik bezeichnen, hat sich über die letzten 250 Jahre entwickelt. Wir gehen zunächst ganz an den Anfang zurück: Zum ersten Mal formuliert wurde der Satz von Bayes in einem 1764 erschienenen Artikel mit dem unscheinbaren Titel "An Essay towards solving a Problem in the Doctrine of Chances" (der ursprüngliche Titel war deutlich gewagter: "A Method of Calculating the Exact Probability of All Conclusions Founded on Induction" ((Stigler 2016), S. 73). Der Artikel wurde ursprünglich von Thomas Bayes verfasst und nach seinem Tod von Richard Price überarbeitet, mit einem Begleitbrief versehen und an die Royal Society eingereicht. Ausgangspunkt für die Überlegungen von Thomas Bayes war ziemlich sicher die Auseinandersetzung mit dem schottischen Philosphen David Hume, insbesondere mit dessen 1748 erschienenen Essay "Of Miracles". In diesem Essay geht es unter anderem um die Frage, wie glaubwürdig ein Augenzeuge sein muss, damit man eine sehr unwahrscheinliche Aussage für wahr hält. Hume argumentiert: Erst wenn es unwahrscheinlicher ist, dass die Aussage falsch ist, als dass das behauptete Ereignis tatsächlich stattgefunden hat, sollte man der Aussage glauben. Thomas Bayes war nun, nach allem was man weiss, mathematisch interessiert. Zwar durfte er nicht in Oxford oder Cambridge studieren, er war aber Mitglied der Royal Society und hat mathematische Texte publiziert (McGrayne (2011), S. 4). Vermutlich hat er den Text von David Hume zum Anlass genommen, um zu überlegen: Könnte man das nicht irgendwie berechnen? Das heisst, statt einfach zu sagen, das eine muss glaubwürdiger sein als das andere, könnte man nicht konkret ausrechnen, wie viel Evidenz man bräuchte um eine Ausgangsüberzeugung zu ändern? Bayes selbst hat in seinem Artikel keine philosophischen Überlegungen festgehalten, die Einordnung in den Kontext von Hume stammt von Price. Dennoch ist der Bezug zu Hume aufschlussreich, weil er zeigt, dass das Umfeld, in dem die bayesianische Statistik entstanden ist, ein ganz anderes war als der Ausgangspunkt von anderen frühen wahrscheinlichkeitstheoretischen Überlegungen, die im Zusammenhang mit Glücksspiel und Wahlurnen entwickelt wurden. Das erklärt ein Stück weit die bis heute bestehende Kontroverse zwischen frequentistischer und bayesiansicher Statistik.

Der Artikel enthält noch keine voll funktionsfähige Methode und Stichworte wie Priorverteilung oder Likelihoodfunktion sucht man im Text vergebens. Dennoch lassen sich bereits drei zentrale Elemente der bayesianischen Statistik erkennen: Erstens wird das Problem bedingten Wahrscheinlichkeit in umgekehrter Reihenfolge formuliert: "*Given* the number of times in which an unknown event has happend or failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named." (Bayes 1763). Zweitens wird anhand eines Beispiels die Aktualisierung von Ausgangsüberzeugungen aufgrund neuer Daten eingeführt. Und drittens nimmt Price im Begleitbrief Bezug auf die aktuellen wahrscheinlichkeitstheoretischen Überlegungen der Zeit, erwähnt ausdrücklich Abraham de Moivres "The Doctrine of Chances" und umschreibt Bayes bedingte Wahrscheinlichkeit mit der Formulierung: "the peculiar definition he has given of the word chance". Damit deutet sich bereits im ersten Artikel zur bayesanischen Statistik ein neuer Wahrscheinlichkeitsbegriff an.

Die weitere Entwicklung erfolgte insbesondere durch Pierre-Simon Laplace, der den Satz von Bayes entweder unabhänig von Bayes Artikel nochmals entdeckte oder ihn aus diesem

übernahm. Er baute die Theorie zu einer Methode aus, die er in verschiedenen Bereichen anwandte, vor allem in der Astronomie (dort standen im 18. Jahrhundert besonders hochwertige Daten zur Verfügung). Erste Verwendungen von bayesianischer Statistik in der psychologischen Forschung gab es bereits in den 1960er Jahren (zum Beispiel (Edwards u. a. 1963)) und ab ca. 1990 wurden bayesianische Methoden zunehmend verwendet (Schoot u. a. 2017). Im Januar 2026 veröffentlichte die US-amerikanische Food and Drug Administration (FDA) einen Entwurf zu Guidelines für die Anwendung bayesianischer Methoden bei der Prüfung von Arzneimitteln (Research 2026). Aufgrund der regulatorischen Bedeutung der FDA kann diese Veröffentlichung als Hinweis auf eine wachsende Akzeptanz bayesianischer Methoden interpretiert werden (Gelman u. a. 2026).

Bedingte Wahrscheinlichkeit

Ein Zufallsversuch ist ein Vorgang, bei dem das Ergebnis nicht im Voraus feststeht. Ein möglicher Ausgang eines Zufallsversuchs wird Ergebnis genannt. Alle Ergebnisse zusammen bilden den Ergebnisraum und Teilmengen des Ergebnisraums werden als Ereignisse bezeichnet.

Die Unsicherheit in Bezug auf den Ausgang eines Zufallsversuchs wird auch aleatorische Unsicherheit genannt. Damit wird ausgedrückt, dass wir nicht wissen, was das Ergebnis ist und dass tatsächlich noch mehrere Ergebnisse möglich sind. Diese aleatorische Unsicherheit steht im Gegensatz zur sogenannten epistemischen Unsicherheit, bei der der Zufallsversuch bereits stattgefunden hat, das Ergebnis aber nicht bekannt ist. Alltagssprachlich kann mit dem Begriff "wahrscheinlich" beides ausgesagt werden: "Wahrscheinlich regnet es morgen" (aleatorische Unsicherheit) und "Wahrscheinlich wurde mein Fahrrad gestohlen" (epistemische Unsicherheit). Mehr zur Unterscheidung zwischen "chance" und "ignorance" können Sie hier nachlesen.

Stellen Sie sich nun vor, Sie werfen blind, ohne zu zielen, einen Stift auf ein Quadrat. Das ganze Quadrat ist unser Ereignisraum Ω. Um die Wahrscheinlichkeit zu berechnen, dass der Stift in einer Teilmenge A von Ω landet, können wir nun einfach die Fläche von A durch die Fläche von Ω dividieren. Je grösser A im Verhältnis zu Ω, desto wahrscheinlicher ist es, dass der Stift in A landet. Dasselbe gilt für B. Nun könnte man eine etwas merkwürdige Frage stellen: Wie gross ist die Wahrscheinlichkeit, dass der Stift in A landet, wenn ich schon weiss, dass er in B gelandet ist? Um das zu berechnen, dividieren wir die Schnittmenge von A und B (orange Fläche im Bild rechts) durch die Fläche von B. Die Wahrscheinlichkeit von A gegeben B ist also die Wahrscheinlichkeit von A und B dividiert durch die Wahrscheinlichkeit von B:

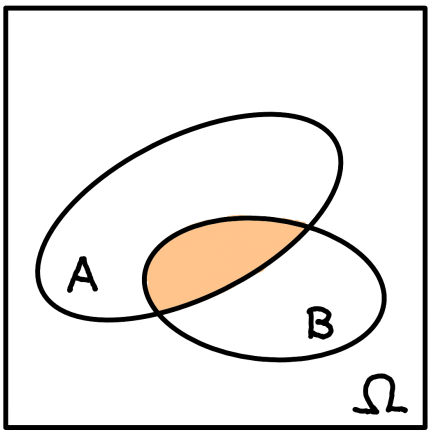


Abbildung 1: Bedingte Wahrscheinlichkeit

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

Auf den ersten Blick ist das eine simple Berechnung. Aber wenn man genauer über die Analogie mit dem Stift nachdenkt, könnte man fragen: Wenn ich bereits weiss, dass der Stift in B gelandet ist, müsste ich dann nicht auch wissen, ob er in A liegt oder nicht? Die Analogie passt hier nicht mehr ganz. Der Grund dafür ist, dass bedingte Wahrscheinlichkeiten eine merkwürdige Mischung aus aleatorischer und epistemischer Unsicherheit ausdrücken. Wir setzen voraus, dass B eingetreten ist und zugleich ist A noch offen. Dass sich unter diesen Voraussetzungen überhaupt eine Wahrscheinlichkeit definieren lässt, ist nicht selbstverständlich.

Der Satz von Bayes

Der Satz von Bayes ergibt sich direkt aus der Definition bedingter Wahrscheinlichkeiten. Der Ausdruck

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

kann durch Multiplizieren beider Seiten mit $P(A)$ umgeformt werden zu

$$P(A \cap B) = P(B|A) \cdot P(A).$$

Damit erhält man

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

In Worten ausgedrückt heisst das: Die Wahrscheinlichkeit eines Ereignisses A unter der Bedingung, dass das Ereignis B eingetreten ist, ist proportional zur Wahrscheinlichkeit des Ereignisses B unter der Bedingung, dass Ereignis A eingetreten ist, multipliziert mit der Wahrscheinlichkeit des Ereignisses A.

Ein Anwendungbeispiel

Eine einfache Berechnung, die den Satz von Bayes verwendet, wäre zum Beispiel folgende: Nehmen wir an, es findet ein Velorennen mit 1000 Fahrern statt, wovon 20 gedopt haben. Bei 19 von disen 20 Fahrern fällt der Dopingtest positiv aus (siehe Bild rechts). Unter den 980 nicht gedopten Fahrern, gibt es 49 positive Tests.

Die Wahrscheinlichkeit, dass ein Dopingtest postiv ausfällt, gegeben jemand hat gedopt, beträgt 95%, denn

$$P(B|A) = \frac{19}{20} = 0.95$$

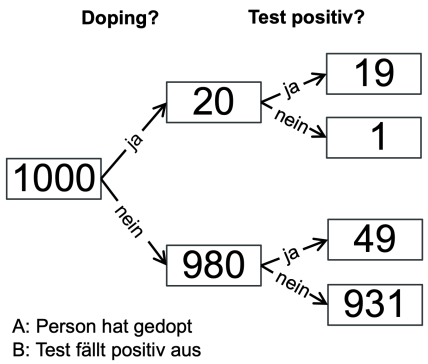


Abbildung 2: Anwendungsbeispiel für den Satz von Bayes

Viel interessanter als die Testsensitivität ist in der Praxis aber natürlich die umgekehrte Frage: Wie wahrscheinlich ist es, dass jemand gedpot hat, gegeben der Test ist positiv ausgefallen? Das kann man nun mit Hilfe des Satzes von Bayes berechnen:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.95 \cdot \frac{20}{1000}}{\frac{19+49}{1000}} \simeq 0.28$$

Die Wahrscheinlichkeit, dass jemand gedopt hat, wenn der Test positiv ausgefallen ist, beträgt also ungefähr 28%.

Vom Satz von Bayes zur bayesianischen Statistik

Bis hierhin haben wir nun einfach mit bedingten Wahrscheinlichkeiten gerechnet. Der eigentliche Schritt zur bayesiansischen Statistik besteht nun darin, die allgemeinen Ereignisse A und B durch Bestandteile eines statistischen Modells zu ersetzen, nämlich Daten D und Hypothesen H. Der Ausdruck $P(D|H)$ ("P von D gegeben H") drückt dann aus, wie wahrscheinlich die Daten unter einer bestimmten Hypothese sind und wird als Likelihoodfunktion bezeichnet. Der Ausdruck $P(H|D)$ ("P von H gegeben D") drückt aus, wie wahrscheinlich eine Hypothese ist, wenn bestimmte beobachtete Daten vorliegen und wird als Posteriorverteilung bezeichnet. Mit Hilfe des Satzes von Bayes kann nun ausgehend von der Likelihoodfunktion die Posteriorverteilung berechnet werden:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Allerdings benötigt man für diese Berechnung zwei weitere Bestandteile, nämlich $P(H)$, die sogenannte Priorverteilung, und $P(D)$. Mit $P(D)$ bezeichnen wir für den Moment einfach mal die "Wahrscheinlichkeit der Daten", eine genauere Klärung dazu folgt später. Relevant für den Moment ist, dass sich $P(D)$ nicht ändert, weshalb man diesen Bestandteil häufig weglässt und den Satz verkürzt wie folgt formuliert:

$$P(H|D) \propto P(D|H) \cdot P(H)$$

In Worten ausgedrückt heisst das: Die Posteriorverteilung ist proportional zum Produkt aus Likelihood und Prior.

Abgrenzung von frequentistischer Statistik

Bevor wir uns den drei zentralen Elementen (Posterior, Likelihood und Prior) im Detail zuwenden, wollen wir kurz der Frage nachgehen, inwiefern sich bayesianische Statistik von frequentistischer Statistik unterscheidet. Dafür ist ein Abschnitt aus Deobrah Mayos Buch "Statistical inference as severe testing. How to get beyond the statistics wars" sehr nützlich (der Titel lässt erahnen, dass diese Frage nicht immer nur neutral und nüchtern diskutiert worden ist). Mayo unterscheidet in Bezug auf Inferenz zwischen einer „performance philosophy"

und einer „probabilistic philosophy“ (Mayo (2018), S 13). Die “performance philosophy” sieht die Aufgabe statistischer Verfahren in erster Linie darin, die relative Häufigkeit fehlerhafter Schlussfolgerungen auf lange Sicht zu kontrollieren. Das bedeutet, über viele Studien hinweg soll sichergestellt werden, dass – unabhängig davon, welche Hypothese wahr ist – die Wahrscheinlichkeit einer fälschlichen Ablehnung von H_0 gering ist. Die “probabilistic philosophy” hingegen betrachtet statistische Verfahren als eine Möglichkeit, Hypothesen einen Grad an Überzeugung zuzuweisen. Das bedeutet, aufgrund von gegebenen Daten soll eine vergleichende Einschätzung gegeben werden, ob H_0 glaubwürdiger ist als H_1, oder dass H_0 angesichts der Daten weniger glaubwürdig ist als zuvor.

Grundsätzlich kann man frequentistische Verfahren eher der performance philosophy und bayesianische Verfahren eher der probabilistic philosophy zuordnen. Der Grund dafür ist genau die Umkehrwahrscheinlichkeit aus dem Satz von Bayes. Dort berechnen wir die Wahrscheinlichkeit einer Hypothese gegeben die Daten, während in einem frequentistischen Verfahren lediglich ausgesagt wird, wie wahrscheinlich es ist, die Daten erhalten zu haben, gegeben H_0 ist wahr. Der Satz von Bayes liefert also eine “Regel zur rationalen Revision von Wahrscheinlichkeiten im Lichte neuer Evidenz” (Sedlmeier und Renkewitz (2018), S 688) während frequentistische Verfahren darauf abzielen, langfristig, über viele Studien hinweg, möglichst wenig fehlerhafte Entscheidungen zu treffen.

Obwohl die Unterscheidung zwischen diesen beiden Philosophien eine wesentliche Abgrenzung zwischen bayesianischer und frequentistischer Statistik sichtbar macht, ist damit noch nicht der ganze Unterschied erfasst. Denn es gibt noch einen zweiten wichtigen Punkt: Auch innerhalb der frequentistischen Statistik bestehen seit Längerem Bestrebungen, sich vom Nullhypothesensignifikanztest zu lösen und stattdessen stärker auf die Schätzung von Effekten und die Quantifizierung von Unsicherheit zu fokussieren. Kruschke (Kruschke und Liddell (2018)) beschreibt in diesem Zusammenhang zwei Schwerpunktverlagerungen in der modernen Datenanalyse. Die erste besteht in einer Bewegung von der frequentistischen hin zur bayesianischen Statistik. Die zweite liegt in einer Verlagerung weg von Hypothesentests und hin zu einem stärkeren Fokus auf die Schätzung von Effektstärken sowie auf Masse der Unsicherheit dieser Schätzungen. Treffen beide Entwicklungen zusammen, ergibt sich eine bayesianische Quantifizierung von Unsicherheit in Form einer Posteriorverteilung über plausible Parameterwerte.

Gerade für nicht-experimentelle Daten ist diese Verlagerung weg vom Testen und hin zur Schätzung besonders wichtig, weil hier weniger die Entscheidung über eine Hypothese als die möglichst genaue Beschreibung plausibler Zusammenhänge unter Unsicherheit im Vordergrund steht. Bayesianische Statistik ist dafür besonders gut geeignet.

Priorverteilung, Likelihoodfunktion und Posteriorverteilung

Das Ziel einer bayesianischen Analyse ist es, eine Posteriorverteilung über einen Populationsparameter zu bestimmen. Wie das konkret funktioniert, lässt sich anhand eines einfachen Beispiels nachrechnen.

Angenommen man möchte zwei Lernmethoden vergleichen, Lernen mit Karteikärtchen (A) und Lernen durch wiederholtes Lesen (B). Dazu verwenden mehrere Personen jeweils beide Lernstrategien und notieren am Ende ob sie mit Methode A mehr Wörter lernen konnten

als mit Methode B. Für jede Person gibt es zwei mögliche Ergebnisse: 1, wenn sie mit A mehr Wörter lernen konnten und 0 wenn sie mit A nicht mehr Wörter lernen konnten. Der Populationsparameter θ, der uns hier interessiert, ist der Anteil Personen mit dem Ergebis 1. Beträgt θ zum Beispiel 0.6 bedeutet das, dass 60% der Personen mit Methode A mehr Wörter lernen konnten als mit Methode B. Ein Wert von $\theta = 0.5$ würde dann bedeuten, dass die beiden Methoden gleich gut funktionieren, Werte von unter 0.5 würden eher für Methode B und Werte von über 0.5 für Methode A sprechen.

Damit wir das Zusammenspiel von Priorverteilung, Likelihoodfunktion und Posteriorverteilung mit möglichst einfachen Berechnungen nachvollziehen können, nehmen wir eine Vereinfachung vor, die man in der Praxis so nicht machen würde. Wir nehmen für den Moment einmal an, dass θ diskret ist, d.h. wir betrachten nur Werte zwischen 0 und 1 im Abstand von jeweils 0.1. Später erweitern wir das auf kontinuierliche Populationsparameter.

Die Priorverteilung

Der Populationsparameter θ kann in unserem Beispiel also die folgenden Werte annehmen:

$$\theta_1 = 0, \theta_2 = 0.1, \theta_3 = 0.2, \ ..., \ \theta_{11} = 1$$

Die Priorverteilung ordnet nun jedem dieser Werte eine Wahrscheinlichkeit zu. Zum Beispiel könnten wir eine Priorverteilung p wählen mit

θ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$p(\theta)$	0.01	0.015	0.025	0.1	0.2	0.3	0.2	0.1	0.025	0.015

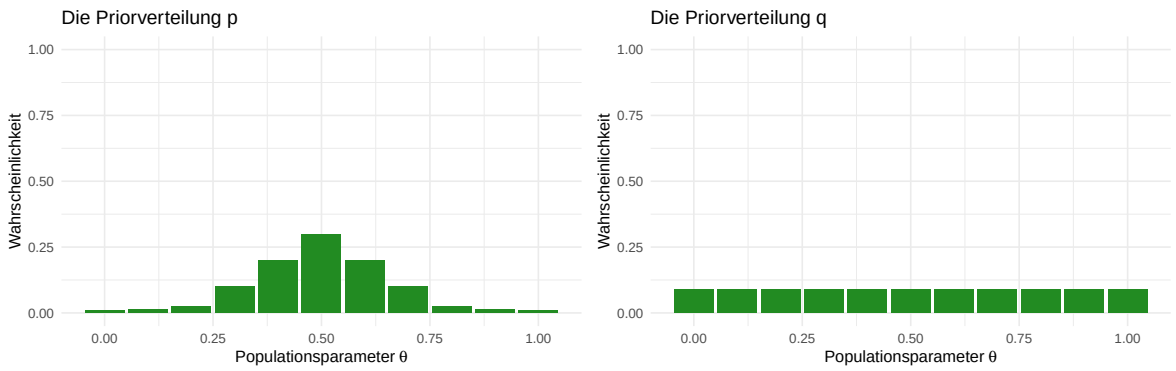
Diese Priorverteilung ordnet dem Wert 50% die höchste Wahrscheinlichkeit zu. Damit drücken wir aus, dass wir aufgrund unserer Vorannahmen relativ stark davon ausgehen, dass die beiden Methoden gleich gut sind. Zudem ist die Verteilung symmetrisch; a priori gehen wir also nicht davon aus, dass eine bestimmte Methode besser ist als die andere. Extremen Werten wie $\theta = 0$ oder $\theta = 1$ wird wenig Gewicht zugemessen.

Eine andere mögliche Priorverteilung, q wäre zum Beispiel folgende:

$$q(\theta_1) = q(\theta_2) = ... = q(\theta_{11}) = 1/11$$

Diese Priorverteilung ordnet allen möglichen Werten für den Populationsparameter genau die gleiche Ausgangswahrscheinlichkeit zu. Die Verteilung q ist ein sogenannter uniformer Prior.

Die Priorverteilung ist eine Wahrscheinlichkeitsverteilung, das heisst, die Wahrscheinlichkeiten, welche den einzelnen Werten des Populationsparameters zugeordnet werden, müssen zusammen 1 ergeben (im diskreten Fall werden sie summiert, im kontinuierlichen Fall integriert). Wenn man einem Wert des Populationsparameters in der Priorverteilung die Wahrscheinlichkeit 0 zuordnet, wird die Posteriorverteilung für diesen Wert immer 0 ergeben. Bei der Wahl der Priorverteilung sollte man daher nur solche Werte ausschliessen, die man aus inhaltlichen



Gründen tatsächlich für unmöglich hält. In unserem Beispiel sind das Werte kleiner als 0 und grösser als 1.

Wie eine Priorverteilung gewählt und interpretiert wird, hängt vom jeweiligen Anwendungsfall ab. In manchen Analysen ist die Priorverteilung tatsächlich Ausdruck einer inhaltlich begründeten Ausgangsüberzeugung. Häufig basiert sie jedoch auch auf den Ergebnissen früherer Studien. Man kann auch mehrere verschiedene Priorverteilungen wählen und die Ergebnisse miteinander vergleichen. Denn mit zunehmender Datenmenge sollte der Einfluss der Priorverteilung abnehmen und wenn eine Analyse unter verschiedenen Priorverteilungen zu ähnlichen Ergebnissen führt, stärkt das die Schlussfolgerung. Schliesslich kann die Wahl einer Priorverteilung auch algorithmisch motiviert sein: In diesem Fall dient die Priorverteilung nicht in erster Linie dazu, Vorwissen abzubilden, sondern dazu, sehr unplausiblen Parameterwerten nur geringe Wahrscheinlichkeit zuzuweisen und dadurch die Konvergenz des MCMC-Algorithmus zu beschleunigen.

Die Likelihoodfunktion

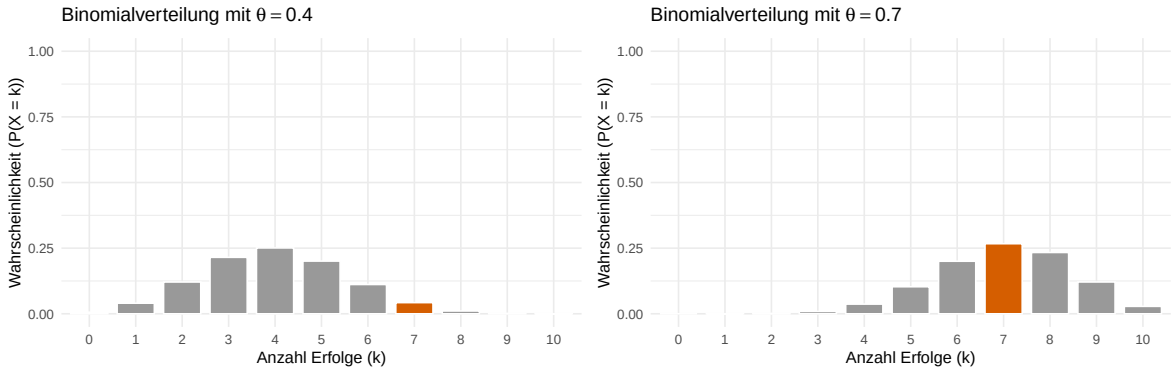
Nehmen wir nun an, wir hätten die beiden Lernmethoden bei 10 Personen verglichen und folgende Daten erhalten:

Person	1	2	3	4	5	6	7	8	9	10
Mehr Wörter mit Kärtchen?	1	0	1	1	1	1	0	1	0	1

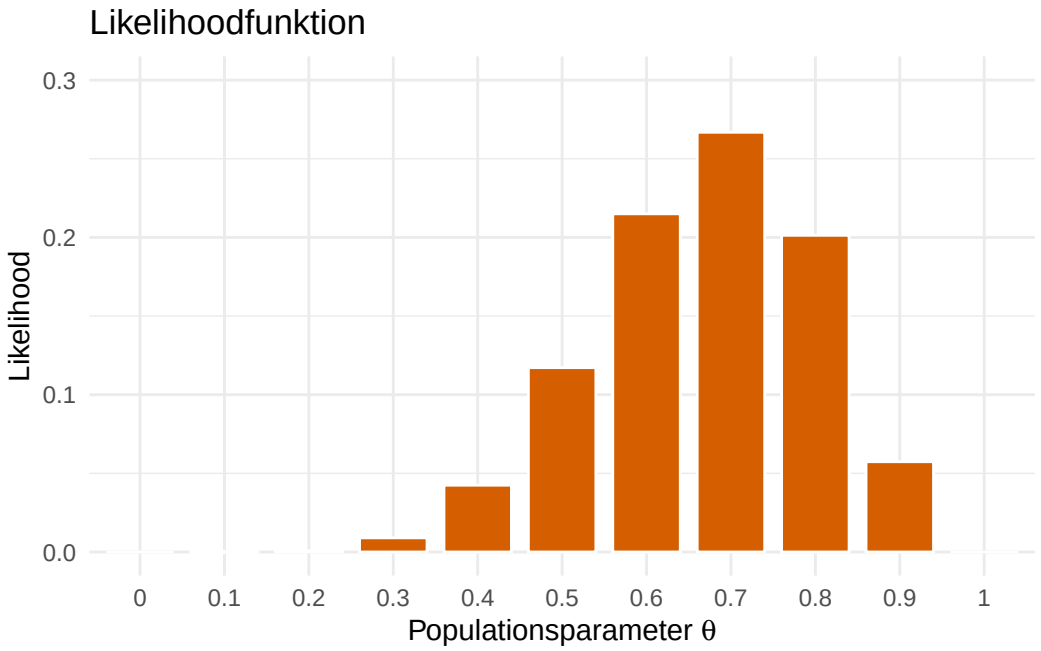
Die Likelihoodfunktion bildet die Wahrscheinlichkeit ab, diese Daten zu beobachten, wenn der Populationsparameter einen bestimmten Wert annimmt. In unserem Beispiel berechnen wir also für θ_1, ..., θ_{11} jeweils wie wahrscheinlich es ist, dass insgesamt 7 von 10 Personen mit Methode A mehr Wörter lernen konnten als mit Methode B. Das können wir das mit einer Binomialverteilung modellieren:

$$P(X = k) = \binom{10}{k} \cdot \theta^k \cdot (1 - \theta)^{10-k}$$

Diese Verteilung ordnet für einen gegebenen Wert des Populationsparameters θ jeder möglichen Anzahl beobachteter Erfolge k eine Wahrscheinlichkeit zu. Für $\theta = 0.4$ zum Beispiel beträgt die Wahrscheinlichkeit, bei 10 Versuchen 7 Erfolge zu beobachten, rund 5%, für $\theta = 0.7$ sind es ca. 27% (orange Balken).



Stellen wir diese Wahrscheinlichkeiten nun in Abhängigkeit des Populationsparameters θ dar, so erhalten wir die Likelihoodfunktion. Wir beobachten, dass die Likelihood für θ_1 und θ_{11} jeweils Null beträgt. Das macht Sinn, denn wenn die Erfolgswahrscheinlichkeit gleich Null ist, ist es unmöglich, 7 Erfolge zu beobachten. Ebenso, wenn die Erfolgswahrscheinlichkeit gleich 1 ist, denn dann müsste man 10 von 10 Erfolge beobachten, nicht nur 7. Wir beobachten auch, dass der Wert 0.7 für den Populationsparameter die grösste Likelihood aufweist. Warum das so ist, können Sie hier nachlesen.



Mit der Likelihoodfunktion können wir also für verschiedene Werte des Populationsparameters berechnen, wie wahrscheinlich die beobachteten Daten sind. In unserem Fall haben wir dazu eine Binomialverteilung verwendet. Das ist eine Annahme und man hätte hier auch eine andere Verteilung wählen oder eine fast beliebig komplizierte Funktion erstellen können. Diese

Flexibilität macht likelihood basierte Ansätze sehr nützlich. Das bedeutet aber auch, dass man mit der Wahl einer bestimmten Likelihoodfunktion eine Entscheidung trifft, die die Analyse stark beeinflusst. In dieser Hinsicht unterscheidet sich bayesiansiche Statistik nicht von klassischen Verfahren:

> "A great deal of ink has been spilled focusing on how Bayesian and non-Bayesian data analyses differ. Focusing on differences is useful, but sometimes it distracts us from fundamental similarites. Notably, the most influential assumptions in both Bayesian and many non-Bayesian models are the distributions assigned to data, the likelihood functions. The likelihoods influence inference for every piece of data, and as sample size increases, the likelihood matters more and more. This helps to explain why Bayesian and non-Bayesian inferences are often so similar. If we had to explain Bayesian inference using only one aspect of it, we should describe likelihood, not priors." (McElreath (2018), S. 34)

Die Posteriorverteilung

Nachdem wir eine Priorverteilung festgelegt und die Likelihoodfunktion bestimmt haben, können wir mithilfe des Satzes von Bayes die Posteriorverteilung berechnen. Dazu multiplizieren wir im Wesentlichen für jeden möglichen Wert des Populationsparameters den Wert der Priorverteilung mit dem entsprechenden Wert der Likelihoodfunktion. Für $\theta = 0.4$ ergibt sich die Posteriorverteilung beispielsweise aus dem Produkt von 0.2 und 0.04.

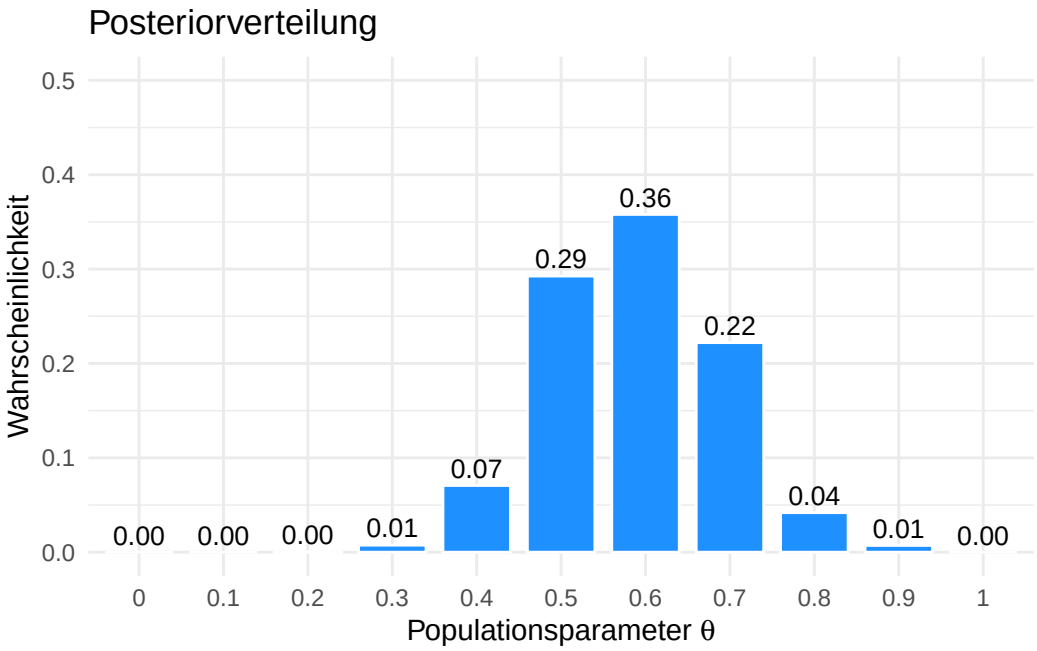
An dieser Stelle berücksichtigen wir nun jedoch noch den oben bereits erwähnten Nenner $P(D)$. Dieser Term beschreibt die "Wahrscheinlichkeit der Daten", also die Gesamtwahrscheinlichkeit dafür, dass die beobachteten Daten auftreten. Das bedeutet, dass wir über alle Werte des Populationsparameters hinweg berechnen, wie wahrscheinlich die Daten jeweils sind und diese Werte aufsummieren. Wir berechnen also folgendes:

$$P(D) = \sum_{k=1}^{11} P(D|\theta_k) \cdot P(\theta_k).$$

Dieser Nenner hängt nicht vom jeweils betrachteten Wert des Populationsparameter ab (weil alle Werte des Populationsparameters einfliessen). In Bezug auf die unten abgebildete Posteriorverteilung ist er deshalb lediglich eine Normierungskonstante. Er skaliert alle Balken gleichermassen nach oben oder nach unten und ändert nichts daran, welcher Balken der höchste ist.

Aus der Posteriorverteilung können wir nun zum Beispiel folgendes ablesen:

- Der Populationsparameter mit der höchsten Posteriorwahrscheinlichkeit ist der Wert 0.6. Die Wahrscheinlichkeit für diesen Wert beträgt ca. 36%.
- Der Populationsparameter liegt mit einer Wahrscheinlichkeit von ca. 90% zwischen 50% und 80%.
- Die Wahrscheinlichkeit, dass der Populationsparameter kleiner als 0.5 ist, beträgt ca. 8%.



Interessant ist auch der Vergleich der Posteriorverteilung mit der Likelihoodfunktion. Die Likelihoodfunktion nahm für den Wert $\theta = 0.7$ den höchsten Wert an. Bei der Posteriorverteilung ist der wahrscheinlichste Populationsparameter mit 0.6 etwas tiefer. Das liegt an der Priorverteilung, die in unserem Fall den höchsten Wert bei 0.5 annimmt. Die Posteriorverteilung ist also eine Art Kompromiss zwischen dem Vorwissen und den Daten eingegangen. In unserem Beispiel liegt der höchste Wert der Posteriorverteilung sogar gerade in der Mitte zwischen dem Prior und der Likelihoodfunktion. Wie sich das verändert, wenn mehr und mehr Daten erhoben werden, können Sie in Übung 1 nachrechnen.

Literatur

Bayes, Thomas. 1763. „LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S“. *Philosophical Transactions*, Nr. 53 (Dezember): 370–418. https://doi.org/10.1098/rstl.1763.0053.

Damen, Nikki L, Henneke Versteeg, Patrick W Serruys, u. a. 2015. „Cardiac patients who completed a longitudinal psychosocial study had a different clinical and psychosocial baseline profile than patients who dropped out prematurely“. *European Journal of Preventive Cardiology* 22 (2): 196–99. https://doi.org/10.1177/2047487313506548.

Edwards, Ward, Harold Lindman, und Leonard J. Savage. 1963. „Bayesian statistical inference for psychological research“. *Psychological Review* (US) 70 (3): 193–242. https://doi.org/10.1037/h0044139.

Ellenberg, Jordan. 2015. *How not to be wrong: the power of mathematical thinking*. Penguin Books.

Fisher, Ronald Aylmer. 1926. „The arrangement of field experiments“. *Journal of the Ministry of Agriculture* 33: 503–15. https://doi.org/10.23637/rothamsted.8v61q.

Gelman, Andrew, Erik van Zwet, und Witold Więcek. 2026. „FDA Draft Guidance for the Use of Bayesian Methods in Clinical Trials“. *JAMA* 335 (19): 1662–63. https://doi.org/10.1001/jama.2026.4178.

Hernán, Miguel A., Sonia Hernández-Díaz, und James M. Robins. 2013. „Randomized Trials Analyzed as Observational Studies“. *Annals of Internal Medicine* 159 (8): 560–62. https://doi.org/10.7326/0003-4819-159-8-201310150-00709.

Komischke, Katrine, Paul A. Boelen, Fiona Maccallum, und Maja O'Connor. 2026. „Group Vs Individual Grief-Focused Cognitive Behavioral Therapy for Older Adults: A Randomized Clinical Trial“. *JAMA Psychiatry* 83 (3): 229–37. https://doi.org/10.1001/jamapsychiatry.2025.4106.

Kruschke, John K., und Torrin M. Liddell. 2018. „Bayesian data analysis for newcomers“. *Psychonomic Bulletin & Review* 25 (1): 155–77. https://doi.org/10.3758/s13423-017-1272-1.

Mayo, Deborah G. 2018. *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

McElreath, Richard. 2018. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman; Hall/CRC.

McGrayne, Sharon Bertsch. 2011. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of C*. Yale University Press.

Parolini, Giuditta. 2015. „The Emergence of Modern Statistics in Agricultural Science: Analysis of Variance, Experimental Design and the Reshaping of Research at Rothamsted Experimental Station, 1919–1933“. *Journal of the History of Biology* 48 (2): 301–35. https://doi.org/10.1007/s10739-014-9394-z.

Reme, Bjørn-Atle, Ole Røgeberg, und Fartein Ask Torvik. 2025. „School performance and the social gradient in young adult death in Norway“. *Nature Human Behaviour* 9 (1): 84–89. https://doi.org/10.1038/s41562-024-02053-w.

Research, Center for Drug Evaluation and. 2026. *Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products*. FDA.

Schoot, Rens van de, Sonja D. Winter, Oisín Ryan, Mariëlle Zondervan-Zwijnenburg, und Sarah Depaoli. 2017. „A systematic review of Bayesian articles in psychology: The last 25 years“. *Psychological Methods* 22 (2): 217–39. https://doi.org/10.1037/met0000100.

Sedlmeier, P., und F. Renkewitz. 2018. *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler; Für Psychologen und Sozialwissenschaftler*. Pearson, Hallbergmoos.

Stigler, Stephen M. 2016. *The seven pillars of statistical wisdom*. Harvard University Press.

Tunis, Sandra L., Douglas E. Faries, Allen W. Nyhuis, Bruce J. Kinon, Haya Ascher-Svanum, und Ralph Aquila. 2006. „Cost-Effectiveness of Olanzapine as First-Line Treatment for Schizophrenia: Results from a Randomized, Open-Label, 1-Year Trial“. *Value in Health* 9 (2): 77–89. https://doi.org/10.1111/j.1524-4733.2006.00083.x.

Zhai, Xiaobing, Henry H. Y. Tong, Chi Kin Lam, u. a. 2024. „Association and causal mediation between marital status and depression in seven countries“. *Nature Human Behaviour* 8 (12): 2392–405. https://doi.org/10.1038/s41562-024-02033-0.