Unterlagen zum Weiterbildungskurs "Bayes-Statistik und Simulation" vom 6. Juni 2012

Hansruedi Künsch
Seminar für Statistik
ETH Zürich

Version vom 29. Mai 2012

1 Bedingte Wahrscheinlichkeiten und Bayes-Formel

1.1 Grundlagen

Ich nehme an, dass die mathematische Beschreibung eines Zufallsexperiments bekannt ist: Wir haben einen endlichen Stichprobenraum Ω, der aus den möglichen Ergebnissen des Experiments besteht, und zu jedem Ergebnis $\omega \in \Omega$ ist dessen Wahrscheinlichkeit $p(\omega)$ gegeben, wobei

$$p(\omega) \geq 0, \quad \sum_{\omega \in \Omega} p(\omega) = 1.$$

Ein Ereignis A ist dann eine Teilmenge von Ω und hat die Wahrscheinlichkeit

$$P(A) = \sum_{\omega \in A} p(\omega)$$

(d.h. $p(\omega) = P(\{\omega\})$). Die Wahrscheinlichkeit quantifiziert mit Werten zwischen 0 und 1, wie sicher es ist, dass das Ereignis A eintritt. Die frequentistische Auffassung der Wahrscheinlichkeit besagt, dass sich die relative Häufigkeit von A bei vielen unabhängigen Wiederholungen des Experiments unter gleichen Bedingungen bei $P(A)$ stabilisiert. In diesem Kurs sollen Sie auch eine andere, subjektive Interpretation von $P(A)$ als Mass für den Glauben an das Eintreten von A kennenlernen.

Das Gegenereignis von A bezeichne ich hier mit A^c (für Komplement).

1.2 Von unbedingten zu bedingten Wahrscheinlichkeiten

Es gibt oft Fälle, wo wir zwar nicht das genaue Ergebnis des Versuchs erfahren, aber wenigstens, dass B eingetreten ist. Dann modifizieren wir die Wahrscheinlichkeiten gemäss der folgenden Definition:
Die **bedingte Wahrscheinlichkeit von A gegeben B** ist

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

Eine "statische" Begründung dieser Definition geht wie folgt: $P(.|B)$ ist eine Wahrscheinlichkeit auf Ω, d.h. für $A \subseteq B$ kann man $P(A|B)$ wieder mit Hilfe der Wahrscheinlichkeiten der einzelnen Ergebnisse berechnen:

$$P(A|B) = \sum_{\omega \in A} p(\omega|B),$$

wobei

$$p(\omega|B) \geq 0, \quad p(\omega|B) = 0 \text{ für } \omega \notin B.$$

Die Information, dass B eingetreten ist, soll die relative Unsicherheit für zwei mögliche Ergebnisse in B nicht verändern:

$$\frac{p(\omega'|B)}{p(\omega|B)} = \frac{p(\omega')}{p(\omega)} \quad (\omega \in B, \omega' \in B).$$

Summiert man beide Seiten dieser Gleichung über $\omega' \in B$ so folgt wegen $\sum_{\omega \in B} p(\omega|B) = 1$:

$$p(\omega|B) = \frac{p(\omega)}{P(B)} \quad (\omega \in B).$$

Mit den bedingten Wahrscheinlichkeiten lassen wir also die Wahrscheinlichkeiten "im Wesentlichen" unverändert bei einer Einschränkung des Stichprobenraums.

Eine "dynamische" Begründung betrachtet einen binären Baum, bei dem auf der ersten Stufe zwischen B und B^c entschieden wird und auf der zweiten Stufe zwischen A und A^c. Dazu braucht das Experiment nicht in 2 Stufen durchgeführt zu werden, wir können uns einfach vorstellen, dass wir zunächst die Information bekommen, ob B oder B^c eingetreten ist, und nachher, ob A oder A^c eingetreten ist. Bei n Wiederholungen des Experiments kommt man ungefähr $nP(B)$ mal beim Knoten B vorbei und endet ungefähr $nP(A \cap B)$ mal beim Knoten $A \cap B$. Damit geht man aber mit der (ungefähren) relativen Häufigkeit $P(A \cap B)/P(B)$ vom Knoten B zum Knoten $A \cap B$.

1.3 Von bedingten zu unbedingten Wahrscheinlichkeiten

Viel häufiger möchte man aber den Stichprobenraum vergrössern, d.h. komplexere Experimente betrachten, anstatt ihn einzuschränken. In dieser Situation werden bedingte Wahrscheinlichkeiten benutzt, um unbedingte Wahrscheinlichkeiten auf dem vergrösserten Stichprobenraum festzulegen.

Wir kombinieren also zwei Experimente mit Stichprobenräumen $\Omega_1 = \{1, 2, \ldots, r\}$ und $\Omega_2 = \{1, 2, \ldots, s\}$ zu einem neuen Experiment, in dem die beiden Teilexperimente simultan oder hintereinander durchgeführt werden. Die "statische" Sichtweise entspricht der simultanen Durchführung. Sie nimmt als neuen Stichprobenraum das Produkt $\Omega = \Omega_1 \times \Omega_2$ und ordnet dessen Elemente als $r \times s$ Rechteck an. Die Wahrscheinlichkeiten $p(i, j)$ bilden dann eine Matrix mit nichtnegativen Elementen, die sich zu 1 summieren. Die Zeilen-, bzw. Spaltensummen dieser Matrix geben dann die Wahrscheinlichkeiten für die einzelnen Experimente an (die sogenannten Randverteilungen):

$$P(\omega_1 = i) = \sum_{j=1}^{s} p(i, j), \quad P(\omega_2 = j) = \sum_{i=1}^{r} p(i, j). \tag{2}$$

Man sieht leicht, dass die Angabe der Zeilen- und Spaltensummen die Elemente der Matrix nicht eindeutig bestimmen. Die bedingten Wahrscheinlichkeiten gegeben das Ergebnis im

ersten, bzw. zweiten Experiment erhält man, indem man die Zeilen-, bzw. Spaltensummen zu 1 normiert:

$$P(\omega_2 = j|\omega_1 = i) = \frac{p(i,j)}{\sum_{k=1}^{s} p(i,k)}, \quad P(\omega_1 = i|\omega_2 = j) = \frac{p(i,j)}{\sum_{k=1}^{r} p(k,j)}. \tag{3}$$

Löst man diese Gleichungen nach $p(i,j)$ auf, so findet man

$$p(i,j) = P(\omega_1 = i|\omega_2 = j)P(\omega_2 = j) = P(\omega_2 = j|\omega_1 = i)P(\omega_1 = i).$$

Die Wahrscheinlichkeiten im kombinierten Experiment sind also bestimmt durch die Wahrscheinlichkeiten des einen Experiments und die bedingten Wahrscheinlichkeiten des anderen.

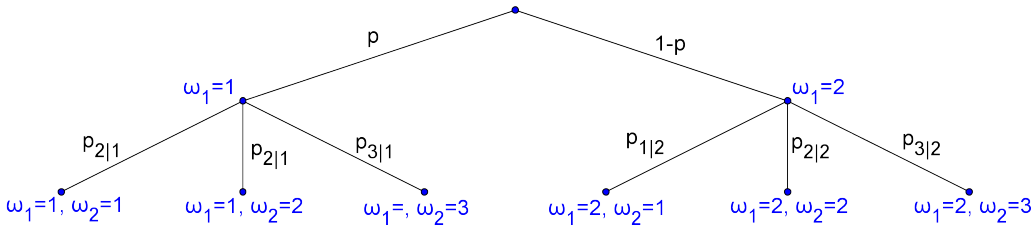

Abbildung 1: Darstellung eines kombinierten Experiments als Baum. Auf der ersten Stufe hat man 2 Möglichkeiten mit Wahrscheinlichkeiten p, bzw. $1-p$. Auf der zweiten Stufe hat man 3 Möglichkeiten, die Wahrscheinlichkeiten auf der zweiten Stufe hängen ab vom Ausgang der ersten Stufe.

Die "dynamische" Sichtweise entspricht der sukzessiven Durchführung. Sie nimmt als neuen Stichprobenraum die Endknoten eines zweistufigen Baums, bei dem sich auf der ersten Stufe entscheidet, welches Ergebnis aus Ω_1 eintritt und auf der zweiten Stufe, welches Ergebnis aus Ω_2 eintritt. Wahrscheinlichkeiten werden dann offensichtlich so festgelegt, dass man auf der ersten Stufe Wahrscheinlichkeiten für die r möglichen Äste festlegt, und dann – auf der zweiten Stufe – die bedingten Wahrscheinlichkeiten für die s möglichen Äste gegeben die Wahl auf der ersten Stufe, vergleiche Abbildung 1. Die Wahrscheinlichkeiten für einen Endknoten ergeben sich dann durch Multiplikation entlang des zugehörigen Pfades (Pfadregel)

$$p(i,j) = P(\omega_1 = i)P(\omega_2 = j|\omega_1 = i), \tag{4}$$

was wir auch schon in der "statischen" Sichtweise gesehen hatten.

Die beiden Sichtweisen sind natürlich äquivalent: Stellt man den Baum dreidimensional dar und dreht man die Äste auf der zweiten Stufe um 90 Grad, so bilden die Endknoten gerade ein Rechteck (siehe Abb. 3 in Gallin (2003) [1]). Der Vorteil der dynamischen Sichtweise ist, dass man den Baum wie ein "verallgemeinertes Galton-Brett" benutzen kann: Betrachtet man n Wiederholungen des zusammengesetzten Experiments so werden n Kugeln mit Hilfe des Baumes verteilt, beginnend an der Wurzel: An jedem Knoten werden die Kugeln auf die möglichen Äste im Verhältnis der Wahrscheinlichkeiten an diesem Knoten verteilt.

Betrachten wir als Beispiel folgende Situation eines medizinischen Tests. In einer bestimmten Altersgruppe habe einer von 200 Männern eine bestimmte Krankheit. Für das Ereignis

[1] Prädikatives und funktionales Denken in der Wahrscheinlichkeitsrechnung, http://www.gallin.ch/GallinZDMWahrsch.pdf

$B =$ "Krankheit" gilt also "a priori" $P(B) = 0.005$. Man kann die Krankheit nachweisen mit Hilfe eines Tests, bevor Symptome auftreten. Allerdings ist kein Test völlig fehlerfrei. Nehmen wir zum Beispiel an, dass für das Ereignis $A =$ "Test positiv" gilt:

$$P(A|B) = 0.90, \quad P(A^c|B^c) = 0.96.$$

Wenn wir nun ein Gedankenexperiment machen, bei dem sich 10'000 Männer mit diesem Test testen lassen, dann haben wir nach der ersten Stufe (ungefähr) 50 Männer bei Knoten B und 9950 beim Knoten B^c. Nach der zweiten Stufe haben wir also (ungefähr) folgende Verteilung: $45 = 50 \cdot 0.9$ bei $A \cap B$, $5 = 50 \cdot 0.1$ bei $A^c \cap B$, $398 = 9950 \cdot 0.04$ bei $A \cap B^c$ und $9552 = 9950 \cdot 0.96$ bei $A^c \cap B^c$.

1.4 Die Bayes-Formel

In der statischen Sichtweise werden die beiden Teilexperimente symmetrisch behandelt: Damit kann man direkt den Zusammenhang zwischen den beiden bedingten Verteilungen $P(\omega_1 = i|\omega_2 = j)$ und $P(\omega_2 = j|\omega_1 = i)$ herstellen:

$$P(\omega_1 = i|\omega_2 = j) = \frac{p(i,j)}{\sum_{k=1}^{s} p(k,j)} = \frac{P(\omega_2 = j|\omega_1 = i)P(\omega_1 = i)}{\sum_{k=1}^{s} P(\omega_2 = j|\omega_1 = k)P(\omega_1 = k)}. \tag{5}$$

Dies wird als **Formel von Bayes** bezeichnet.

Die dynamische Sichtweise impliziert, dass Teilexperiment 1 zuerst ausgeführt wird. Es kann aber trotzdem sein, dass das Ergebnis vom zweiten Teilexperiment zuerst bekannt wird, etwa im Beispiel mit dem medizinischen Test oben. Um $P(\omega_1 = i|\omega_2 = j)$ zu berechnen, muss man den Baum "umdrehen": Die Wahrscheinlichkeiten auf der ersten Stufe im "umgedrehten Baum" erhält man durch Addition. Die Wahrscheinlichkeiten für die Endknoten bleiben unverändert, und damit erhält man die Wahrscheinlichkeiten für die Äste der zweiten Stufe im "umgedrehten" Baum durch Division

$$P(\omega_1 = i|\omega_2 = j) = \frac{p(i,j)}{P(\omega_2 = j)} = \frac{p(i,j)}{\sum_{k=1}^{r} p(k,j)}.$$

Dies ist wieder die Formel von Bayes.

Im Beispiel mit den medizinischen Tests erhalten wir also

$$P(B|A) = \frac{0.005 \cdot 0.90}{0.005 \cdot 0.90 + 0.995 \cdot 0.04} = \frac{45}{45 + 398} = 0.1016.$$

Dies ist viel kleiner, als man naiverweise denken würde: Obwohl der Test kleine Fehlerwahrscheinlichkeiten hat, kann man aus dem Vorliegen eines positiven Testresultats noch nicht auf das Vorhandensein der Krankheit schliessen. Der Grund ist leicht einzusehen: Da die meisten der getesteten Männer gesund sind, erhält man trotz der kleinen Fehlerwahrscheinlichkeit eine grosse Zahl von 398 "falsch Positiven" gegenüber 45 "korrekt Positiven".

Bei einer häufigen Krankheit sähe das ganz anders aus. Das wird durchsichtiger, wenn wir das sogenannte Wettverhältnis (odds ratio) mit und ohne Test anschauen:

$$\frac{P(B|A)}{P(B^c|A)} = \frac{P(B \cap A)}{P(B^c \cap A)} = \frac{P(A|B)}{P(A|B^c)} \cdot \frac{P(B)}{P(B^c)} = \frac{0.90}{0.04} \cdot \frac{P(B)}{P(B^c)} = 22.5\frac{P(B)}{P(B^c)} \tag{6}$$

Der Test multipliziert also stets das Wettverhältnis mit dem Faktor 22.5. Er ist informativ, aber er kann ein extremes a priori Wettverhältnis nicht umkippen.

Dies ist ein sehr drastisches Beispiel, dass $P(B|A) \neq P(A|B)$. Noch extremer wird es, wenn $A \subset B$, weil dann $P(B|A) = 1$ während $P(A|B)$ beliebig klein sein kann (man nehme zum Beispiel die beiden Ereignisse "Ein zufällig ausgewählter Mensch ist männlich" und "Ein zufällig ausgewählter Mensch ist der Pabst"). Im Allgemeinen ist aber nicht einfach $P(B|A) \neq P(A|B)$, sondern man kann ohne zusätzliche Information aus $P(B|A)$ nicht $P(A|B)$ berechnen kann. Die zusätzliche Information, die man braucht, ist die "a priori" Wahrscheinlichlichkeit $P(A)$.

Obwohl diese Überlegungen eigentlich einleuchten, geschehen häufig Fehler auf Grund einer Verwechslung von $P(B|A)$ und $P(A|B)$, Im eher makabren Beispiel "Bei einem Drittel aller Verkehrsunfälle war der Lenker alkoholisiert, daher sollte man endlich die nüchternen Autofahrer von der Strasse verbannen, weil sie zwei Drittel aller Unfälle verursachen." ist das unmittelbar einsichtig. Beim sogenannten Trugschluss des Staatsanwalts ("prosecutors fallcy") wird $P(U|E)$ verwechselt mit $P(E|U)$ wobei E die vorliegende Evidenz bezeichnet und U die Unschuld des Angeklagten. Nach der Formel (6) ist

$$\frac{P(U|E)}{P(U^c|E)} = \frac{P(E|U)}{P(E|U^c)} \cdot \frac{P(U)}{P(U^c)},$$

und ein kleiner Wert des ersten Verhältnisses rechts impliziert kein kleines Verhältnis links.

1.5 Situationen mit 3 Stufen

Mit mehr als zwei Stufen geht alles im Prinzip analog, aber angesichts der Vielzahl von Stufen, auf die man bedingen kann, kann man leicht den Überblick verlieren. Wir beschränken uns im Moment auf drei binäre Stufen: C oder C^c auf der ersten, B oder B^c auf der zweiten und A oder A^c auf der dritten. Das kombinierte Experiment mit allen drei Stufen hat also 8 mögliche Ergebnisse, eine Wahrscheinlichkeitsverteilung darauf also 7 freie Parameter, da die Summe 1 ergeben muss. In der "dynamischen" Sichtweise sind die freien Parameter die Wahrscheinlichlickeiten, an einem der 7 Knoten jeweils den rechten Ast zu wählen.

Für das Folgende wollen wir $P(A|B)$ berechnen und sehen, wie es zusammenhängt mit $P(A|B \cap C)$ und $P(A|B \cap C^c)$. Das heisst, wie verändern sich die Unsicherheiten, wenn wir die Information nicht haben, ob C oder C^c eingetreten ist ? Intuitiv würde man vermuten, dass $P(A|B)$ ein gewichtetes Mittel von $P(A|B \cap C)$ und $P(A|B \cap C^c)$ ist. Mit Hilfe der Definition kann man leicht nachrechnen, dass

$$P(A|B) = P(A|B \cap C)P(C|B) + P(A|B \cap C^c)P(C^c|B). \tag{7}$$

Diese Vermutung ist also korrekt. Die Gewichte müssen allerdings mit Hilfe der Bayes'schen Formel berechnet werden: Wenn auf der ersten Stufe zwischen C und C^c entschieden wird, sind üblicherweise $P(C)$, $P(B|C)$ und $P(B|C^c)$ gegeben.

Als kleinen Exkurs wollen wir noch das sogenannte Simpson-Paradox besprechen: Die Beziehung (7) mit B^c statt B lautet

$$P(A|B^c) = P(A|B^c \cap C)P(C|B^c) + P(A|B^c \cap C^c)P(C^c|B^c).$$

Wir haben jetzt also ein gewichtetes Mittel von $P(A|B^c \cap C)$ und $P(A|B^c \cap C^c)$, aber mit anderen Gewichten. Es ist daher nicht schwierig Fälle zu konstruieren, wo $P(A|B) > P(A|B^c)$, obwohl $P(A|B \cap C) < P(A|B^c \cap C)$ *und* $P(A|B \cap C^c) < P(A|B^c \cap C^c)$ (ein Beispiel folgt in den Übungen). In Worten: Wenn wir nicht wissen, ob C eingetreten ist,

dann erhöht das Eintreten von B die Chancen von A, aber zusammen mit der Information über C reduziert das Eintreten von B die Chancen von A in beiden Fällen. Das zeigt, dass bedingte Wahrscheinlichkeiten mehr mit den vorliegenden Informationen zu tun hat als mit Ursache und Wirkung. Es zeigt auch eine prinzipielle Schwierigkeit von epidemiologischen oder soziologischen Studien, weil es schwierig ist, latente Faktoren, welche die Richtung eines Effekts umkehren können, auszuschliessen.

Eine andere Erklärung des Simpson-Paradoxes geht aus von

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B \cap C) + P(A \cap B \cap C^c)}{P(B \cap C) + P(B \cap C^c)},$$

während

$$P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)}, \quad P(A|B \cap C^c) = \frac{P(A \cap B \cap C^c)}{P(B \cap C^c)}.$$

Das heisst, $P(A|B)$ wird berechnet, indem man separat die Nenner und die Zähler von $P(A|B \cap C)$ und $P(A|B \cap C^c)$ addiert. Bei dieser Art des "Bruchrechnens" (welche auch auftritt bei der Bestimmung der mittleren Steigung von Strassenstücken), kann es nun vorkommen, dass sich bei der "Summe" die Richtung der Ungleichung bei den "Summanden" umkehrt.

2 Bayes'sche Statistik im Fall der Binomialverteilung

An Stelle von medizinischen Tests können wir auch statistische Tests betrachten: "Gesundheit" (B^c) wird dann zur Nullhypothese, "Krankheit" (B) zur Alternative und ein positiver Test (A) entspricht der Verwerfung der Nullhypothese. Der Fehler erster Art ist das Verwerfen der Nullhypothese, wenn diese richtig ist: Die Wahrscheinlichkeit eines Fehlers erster Art, das sogenannte Niveau des Tests, ist also in der Notation von oben gleich $P(A|B^c)$. Der Fehler zweiter Art ist das Akzeptieren der Nullhypothese, wenn diese falsch ist. Die Wahrscheinlichkeit eines Fehlers 2. Art ist also gleich $P(A^c|B)$, und die Macht des Tests (Eins minus die Wahrscheinlichkeit eines Fehlers 2. Art) ist gleich $P(A|B)$. Wenn man das Beispiel mit den medizinischen Tests verstanden hat, dann sollte es auch klar sein, dass das Niveau nicht gleich der Wahrscheinlichkeit ist, dass die Nullhypothese stimmt, wenn diese verworfen wird: Das wäre ja $P(B^c|A)$, und bei bedingten Wahrscheinlichkeiten spielt es eine Rolle, was links und was rechts vom senkrechten Strich steht !

Die Wahrscheinlichkeit, dass die Nullhypothese stimmt, gegeben dass sie verworfen wird, ist erst definiert, wenn man ausserdem noch eine a priori Wahrscheinlichkeit von B festlegt, d.h. man muss wissen, wie wahrscheinlich es a priori ist, dass die Nullhypothese falsch ist. An diesem Punkt tauchen prinzipielle Einwände auf: Viele Statistiker sind nicht bereit, die Nullhypothese als zufällig zu betrachten, denn es ist nicht klar, wie ein Zufallsexperiment aussieht, das bestimmt, ob eine Nullhypothese richtig oder falsch ist. Die klassische Theorie von Neyman und Pearson (1933) betrachtet die Korrektheit der Nullhypothese einfach als unbekannt, aber nicht als zufällig. Sie erlaubt es damit nicht, von der Wahrscheinlichkeit, dass die Nullhypothese stimmt, zu sprechen.

Umgangssprachlich benutzen wir aber Wahrscheinlichkeiten oft für Dinge, die unbekannt sind, aber nicht wiederholbar im Sinne eines Zufallsexperiments wie der Münzwurf, z.B. die Wahrscheinlichkeit, dass ich eine Prüfung bestehe. In der Fachsprache nennt man das "epistemische" im Unterschied zu "aleatorischer" Unsicherheit. Epistemische Wahrscheinlichkeiten sind meist subjektiver Natur, da sie von den individuellen Einschätzungen

abhängen. Um sie zu quantifizieren, kann man kein Experiment mehrmals wiederholen, aber man kann das Wettverhalten einer Person untersuchen. Man bietet die folgenden Alternativen an: Entweder p Franken sofort ausbezahlt, oder 100 Franken nur dann, wenn B eintritt. Derjenige Wert p, bei dem die beiden Angebote als gleichwertig taxiert werden, ist die subjektive Wahrscheinlichkeit von B (in Prozent).

Neben der Neyman-Pearson Theorie gibt es auch eine zweite Schule der Statistik, die sogenannte Bayes-Statistik, die auf Thomas Bayes (1701-1761) zurückgeht. In dieser Theorie betrachtet man nicht nur die Beobachtungen, sondern auch unbekannte Parameter (und damit auch Nullhypothesen) als zufällig und beschreibt die Unsicherheit über diese Parameter mit Hilfe von Verteilungen. Dann wird untersucht, wie sich diese Verteilungen auf Grund der beobachteten Daten verändern. Im Folgenden werden wir diesen Ansatz am Beispiel der Binomialverteilung etwas genauer diskutieren. Um das ganze etwas anschaulicher zu machen, wählen wir dazu eine Einkleidung, in der der Erfolgsparameter zufällig ist im aleatorischen Sinn.

2.1 Ziehung aus einer zufällig gewählten Urne

Wir betrachten folgendes Zufallsexperiment: Wir haben m verschiedene Urnen, jede mit $m+1$ Kugeln. In der k-ten Urne gibt es k rote und $m+1-k$ weisse Kugeln. Wir wählen eine Urne zufällig, können deren Zusammensetzung jedoch nicht direkt ersehen. Hingegen können wir n mal mit Zurücklegen aus der Urne eine Kugel ziehen und deren Farbe erkennen. Wie genau können wir damit die Zusammensetzung der gewählten Urne bestimmen ?

Offensichtlich ist das ein mehrstufiges Zufallsexperiment: Die erste Stufe ist die Wahl der Urne: B_k bezeichne das Ereignis, dass die Urne mit k roten Kugeln gewählt wird. Weil die Urne zufällig gewählt wird, gilt $P(B_k) = \frac{1}{m}$. Die nächsten n Stufen sind die Ziehungen: Wenn A_j das Ereignis "j-te Ziehung ergibt eine rote Kugel" bezeichnet, dann ist $P(A_j|B_k) = \frac{k}{m+1} =: p_k$. Weil wir mit Zurücklegen ziehen, sind die $A_1, \ldots, A_n$ unabhängig gegeben B_k und die bedingten Wahrscheinlichkeiten der Ereignisse $C_i =$ "Genau i rote Kugeln bei den n Ziehungen" sind durch die Binomialverteilung gegeben:

$$P(C_i|B_k) = \binom{n}{i} p_k^i \, (1-p_k)^{n-i}.$$

Wir haben in diesem Experiment daher zwei Zufallsvariablen: $X =$ Anzahl gezogener roter Kugeln und $Y =$ Anzahl roter Kugeln in der Urne. Y ist gleichverteilt auf $\{1, 2, \ldots, m\}$, und bedingt auf $Y = k$ ist X binomial$(n, \frac{k}{m+1})$-verteilt. Für grosses n sollte wegen des Gesetzes der grossen Zahlen $\frac{X}{n} \approx \frac{Y}{m+1}$ gelten.

2.1.1 Die a posteriori Verteilung von $\frac{Y}{m+1}$

Mit der Bayes-Formel erhalten wir sofort

$$P(Y = k|X = i) = P(B_k|C_i) = \frac{P(C_i|B_k)P(B_k)}{\sum_{h=1}^m P(C_i|B_h)} = \frac{p_k^i \, (1-p_k)^{n-i}}{\sum_{h=1}^m p_h^i \, (1-p_h)^{n-i}}.$$

Für gegebenes i und n lassen sich diese a posteriori Wahrscheinlichkeiten leicht numerisch berechnen, siehe z.B. die Abbildung 2, welche die Resultate für $m = 19$, $n = 10$ und $i = 0, 1, \ldots, 5$ zeigt. Der Nenner in der Bayes-Formel ist gerade $P(C_i 0 = P(X = i)$.

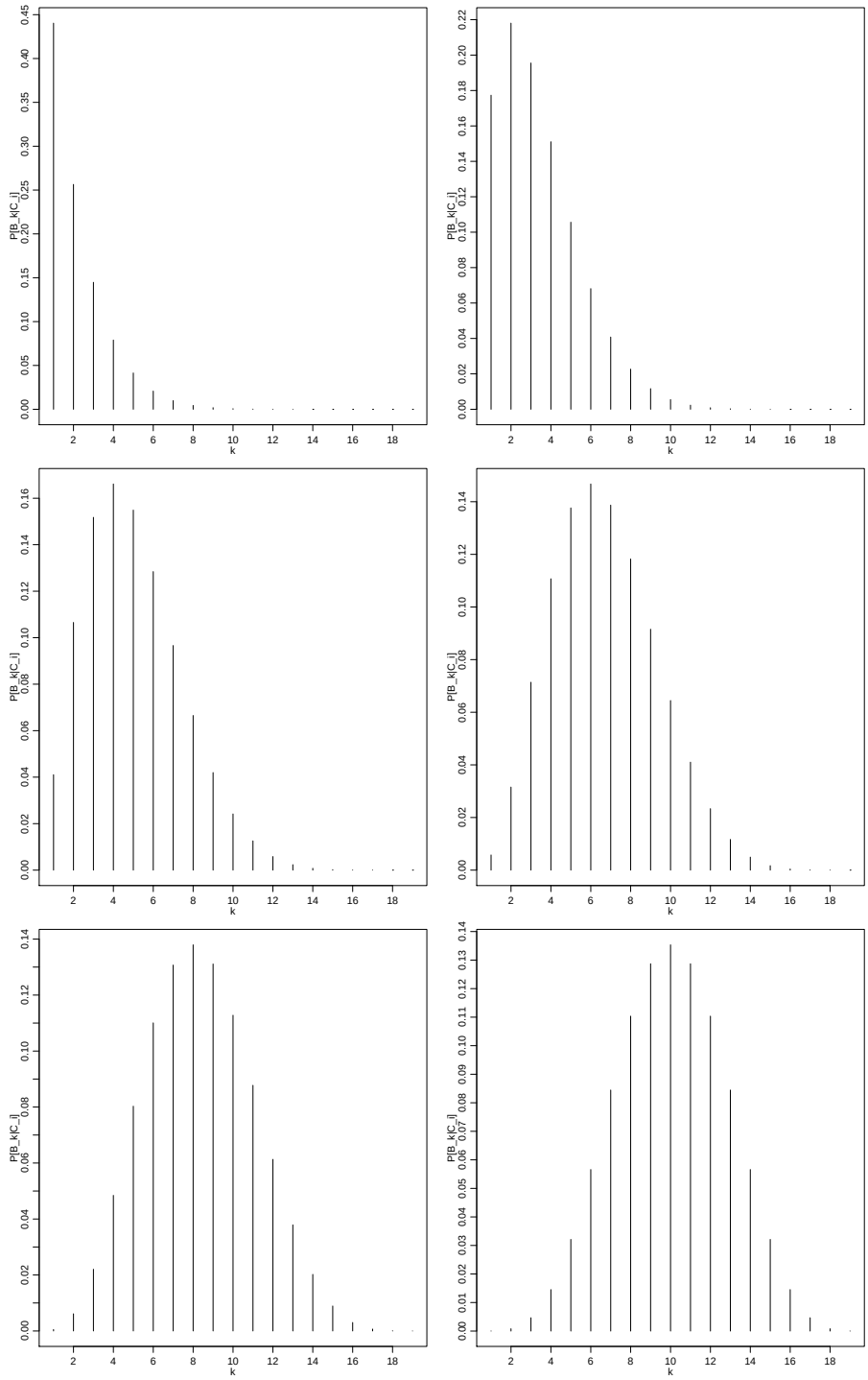

Abbildung 2: A posteriori Wahrscheinlichkeiten $P(B_k|C_i)$ für $m = 19$, $n = 10$ und $i = 0, 1, \ldots, 5$. Für $i = 6, \ldots, 10$ erhält man die entsprechenden Figuren aus Symmetrie.

Aus Symmetriegründen ist $P(X = i] = P(X = n - i)$, aber die Verteilung von X ist breiter als die Binomial$(n, E(\frac{Y}{m+1}) = \frac{1}{2})$-Verteilung, weil ja noch Unsicherheit über die Zusammensetzung der Urne besteht. Abbildung 3 zeigt die beiden Verteilungen für $n = 10$ und $m = 19$: X ist fast uniform verteilt.

Einen besseren Überblick, wie die Wahrscheinlichkeiten $P(B_k|C_i)$ aussehen, erhält man,

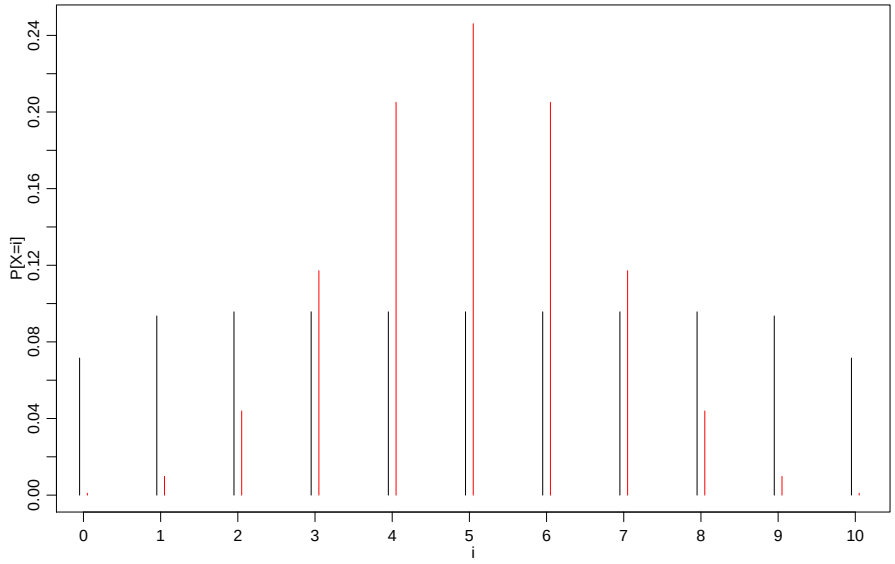

Abbildung 3: Randverteilung $P(X = i)$ der Anzahl roter Kugeln für $m = 19$, und $n = 10$ (schwarz), verglichen mit der Binomial$(10, \frac{1}{2})$-Verteilung (rot). Die erste Verteilung ist wesentlich breiter, was das Nichtwissen über die Zusammensetzung der Urne widerspiegelt.

wenn man beachtet, dass der Nenner in der obigen Formel nur eine Normierung darstellt, die bewirkt, dass $\sum_{k=1}^{m} P(B_k|C_i) = 1$. Diese Normierung kann man ignorieren und sich im Wesentlichen auf den Verlauf der Funktion

$$p \in (0,1) \mapsto p^i \, (1-p)^{n-i}$$

konzentrieren. Wenn i nicht zu nahe bei 0 oder n ist, ist diese Funktion glockenförmig mit dem Maximum an der Stelle $p = \frac{i}{n}$, und den Wendepunkten an den Stellen $p = \frac{i}{n} \pm \frac{1}{\sqrt{n-1}} \sqrt{\frac{i}{n}(1-\frac{i}{n})}$.

Wenn m nicht allzu klein ist, kann man auch den a posteriori Erwartungswert der Zufallsvariable $\frac{Y}{m+1}$, d.h. des relativen Anteils roter Kugeln in der gewählten Urne, approximieren

$$\begin{aligned} E\left(\frac{Y}{m+1}|C_i\right) &= \sum_{k=1}^{m} \frac{k}{m+1} P(B_k|C_i) = \frac{\sum_{k=1}^{m} p_k p_k^i \, (1-p_k)^{n-i}}{\sum_{k=1}^{m} p_k^i \, (1-p_k)^{n-i}} \\ &\approx \frac{\int_0^1 p^{i+1}(1-p)^{n-i} dp}{\int_0^1 p^i (1-p)^{n-i} dp} = \frac{i+1}{n+2}. \end{aligned} \tag{8}$$

Im zweitletzten letzten Schritt wurde benutzt, dass es sich bei Zähler und Nenner um eine Riemann-Summe handelt, und der letzte Schritt beruht darauf, dass

$$\int_0^1 p^{r-1}(1-p)^{s-1} dp = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}.$$

was man mit Integration von $e^{-x}x^{r-1}e^{-y}y^{s-1}$ über $\mathbb{R}_+^2$ zeigen kann. Damit liegt der a posteriori Erwartungswert von $\frac{Y}{m+1}$ sehr nahe bei der relativen Häufigkeit $\frac{X}{n}$ in der Stichprobe (der Maximum-Likelihood-Schätzung).

Ein ähnliches Argument erlaubt es auch die a posteriori Varianz von $\frac{Y}{m+1}$ approximativ zu berechnen:

$$E\left(\left(\frac{Y}{m+1}\right)^2 |B_k\right) = \sum_{k=1}^{m} p_k^2 P(B_k|C_i) \approx \frac{\int_0^1 p^{i+2}(1-p)^{n-i} dp}{\int_0^1 p^i (1-p)^{n-i} dp} = \frac{(i+1)(i+2)}{(n+2)(n+3)}.$$

Auf Grund der Formel $\mathrm{Var}(Y) = E(Y^2) - E(Y)^2$ ist die a posteriori Varianz daher approximativ gleich

$$\mathrm{Var}\left(\frac{Y}{m+1}\right) = \frac{i+1}{n+2}\left(\frac{i+2}{n+3} - \frac{i+1}{n+2}\right) = \frac{(i+1)(n+1-i)}{(n+2)^2(n+3)} \approx \frac{1}{n}\frac{i}{n}(1-\frac{i}{n}).$$

Die rechte Seite ist nichts anderes als die geschätzte Varianz der relativen Häufigkeit $\frac{X}{n}$ bei Ziehungen aus einer festen Urne. Man kann sogar zeigen, dass für grosses n die a-posteriori Verteilung von $\frac{Y}{m+1}$ gegen eine Normalverteilung mit Erwartungswert $\frac{X}{n}$ und Varianz $\frac{1}{n}\frac{X}{n}(1-\frac{X}{n})$ konvergiert, was man bereits auf Grund von Abbildung 2 vermuten konnte. Das heisst, trotz der unterschiedlichen Interpretationen sind die Schlussfolgerungen in der frequentistischen und der Bayes'schen Statistik sehr ähnlich.

2.1.2 Vorhersagen

Die Ergebnisse von 2 Ziehungen A_1 und A_2 sind bedingt auf B_k unabhängig. Kennt man B_k, d.h. die Zusammensetzung der Urne nicht, dann werden A_1 und A_2 jedoch abhängig. Es gilt

$$P(A_1) = \sum_{k=1}^{m} P(A_1|B_k)P(B_k) = \sum_{k=1}^{m} \frac{k}{m(m+1)} = \frac{1}{2},$$

während analog zu Formel (7) gilt:

$$P(A_2|A_1) = \sum_{k=1}^{m} P(A_2|A_1 \cap B_k)P(B_k|A_1) = \sum_{k=1}^{m} P(A_2|B_k)P(B_k|A_1) = \sum_{k=1}^{m} p_k \frac{p_k \frac{1}{m}}{\frac{1}{2}} = \frac{2m+1}{3(m+1)} \approx \frac{2}{3}$$

(beim zweiten Gleichheitszeichen haben wir benutzt, dass A_2 und A_1 unabhängig sind, wenn die Zusammensetzung der Urne bekannt ist). Das scheint auf den ersten Blick paradox, denn die beiden Ziehungen sind immer noch kausal unabhängig. Die erste Ziehung gibt uns jedoch Information über die Zusammensetzung der Urne, und das verändert die Wahrscheinlichkeit, das zweite Mal wieder eine rote Kugel zu ziehen.

Dies kann man für Prognosen benutzen: Angenommen, wir kennen das Ergebnis von n Ziehungen, dann ist die Wahrscheinlichkeit, bei der $(n+1)$-ten Ziehung eine rote Kugel zu ziehen, gleich

$$P(A_{n+1}|C_i) = \sum_{k=1}^{m} P(A_{n+1}|B_k)P(B_k|C_i) = \sum_{k=1}^{m} P(B_k|C_i) \approx \frac{i+1}{n+2}$$

(wir haben Gleichung (8) benutzt). Man nimmt also fast die beobachtete relative Häufigkeit als Vorhersage, fügt jedoch noch zwei hypothetische Ziehungen hinzu, eine mit einer weissen und eine mit einer roten Kugel. Laplace hat mit diesem Argument die Wahrscheinlichkeit berechnet, dass die Sonne am nächsten Tag aufgeht !

2.1.3 Bayes'sche Tests

Machen wir zum Schluss ein Beispiel eines Tests: Wir setzen $m = 19$ (d.h. wir ziehen nur Werte $0.05, 0.1, \ldots, 0.95$ für den Erfolgsparameter in Betracht) und $n = 20$, und wir testen die Nullhypothese "Die gewählte Urne enthält höchstens 10 rote Kugeln". Wenn wir in den 20 Ziehungen 15 mal eine rote Kugel erhalten haben, dann berechnen wir

$$P\left(\cup_{k=1}^{10} B_k|C_{15}\right) = \sum_{k=1}^{10} P(B_k|C_{15}) = \frac{\sum_{k=1}^{10} k^{15}(1-\frac{k}{20})^5}{\sum_{k=1}^{20} k^{15}(1-\frac{k}{20})^5} = 0.0223.$$

A posteriori ist also die Nullhypothese ziemlich unwahrscheinlich. Wenn die Erfolgswahrscheinlichkeit p_k nicht auf Grund eines Zufallsexperiments zustande kommt, würde man den P-Wert berechnen. Dieser gibt an, wie wahrscheinlich es unter der Nullhypothese ist, *mindestens* 15 rote Kugeln zu ziehen, also

$$\sup_{p\leq 1/2} \sum_{j=15}^{20} \binom{20}{j} p^j (1-p)^{20-j} = \sum_{j=15}^{20} \binom{20}{j} 2^{-20} = 0.0207.$$

Beide Grössen messen die Evidenz für die Nullhypothese, allerdings auf eine nicht vergleichbare Art. Insbesondere ist der P-Wert *keine* Wahrscheinlichkeit, sondern eine Realisierung einer Zufallsvariable ! Es ist daher erstaunlich, dass die konkreten Zahlenwerte ähnlich sind.

2.2 Allgemeine a priori Verteilung

Bisher hatten wir angenommen, dass alle Urnen gleich wahrscheinlich sind. Man kann jedoch auch Modelle betrachten, bei denen wir schon eine gewisse Information haben über die verschiedenen Urnen in der Form einer a priori Verteilung

$$\alpha_k = P(B_k), \text{ mit } \sum_{k=1}^{m} \alpha_k = 1.$$

Dann gilt

$$P(B_k|C_i) = \frac{P(C_i|B_k)P(B_k)}{\sum_{h=1}^m P(C_i|B_h)} = \frac{\alpha_k p_k^i \,(1-p_k)^{n-i}}{\sum_{h=1}^m \alpha_h p_h^i \,(1-p_h)^{n-i}}.$$

Approximative Berechnungen sind dann besonders einfach, wenn wir

$$\alpha_k = \frac{p_k^r (1-p_k)^s}{\sum_{h=1} p_h^r (1-p_h)^s}$$

mit $r, s \in \mathbb{N}$ wählen. Diese a priori Verteilung ist dann äquivalent zur Ziehung von r zusätzlichen roten und s zusätzlichen weissen Kugeln, sonst ändert sich bei den Rechnungen nichts.

2.3 Bayesformel im kontinuierlichen Fall

Die Betrachtung mit den m Urnen ist etwas künstlich und nicht wirklich nötig. Wenn wir die subjektive Interpretation der Wahrscheinlichkeit akzeptieren, dann kann man den Erfolgsparameter der Binomialverteilung als Zufallsvariable auffassen, ohne ein Experiment anzugeben, in dem der Paramter mit einem Zufallsexperiment bestimmt wird. Es ist dann auch nicht nötig, die möglichen Werte von p auf $p_k = \frac{2k-1}{2m}$ einzuschränken, sondern wir können p als stetige Zufallsvariable mit einer a priori Dichte $\alpha : [0,1] \to \mathbb{R}_+$ wählen. Wenn bedingt auf p die Zufallsvariable X eine Binomial(n,p)-Verteilung hat, dann ergibt sich bedingt auf $X = j$ für p die a posteriori Dichte

$$\alpha(p|X=i) = \frac{\alpha(p)\, p^i \,(1-p)^{n-i}}{\int_0^1 \alpha(u)\, u^i \,(1-u)^{n-i} du}.$$

Wählen wir als a priori Dichte die sogenannte Beta(r,s)-Dichte

$$\alpha(p) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\; p^{r-1}(1-p)^{s-1},$$

dann ist die a posteriori Dichte eine Beta($r+j$,$s+n-j$)-Dichte, und die Bayes-Formel kann ohne Rechnung angewendet werden. Insbesondere erspart man sich die Approximation von Riemann-Summen durch Integrale, die wir oben stets benutzt haben. Dies ist jedoch ein seltener Glücksfall: In komplexeren Anwendungen der Bayes-Statistik sind Berechnungen meist nicht in geschlossener Form, sondern nur mit Hilfe von Simulation möglich. Dies ergibt die Verbindung zum zweiten Thema dieses Kurses. Am Schluss werde ich einige angewandten Probleme vorstellen, wo die a posteriori Verteilung mit Hilfe von Simulationen untersucht wird.

Wie wir schon am Beispiel des medizinischen Tests gesehen hatten, hat die Wahl der a priori Verteilung einen grossen Einfluss auf das Resultat. Dies gilt ganz allgemein und ist ein Schwachpunkt der Bayes-Statistik. Ein mögliche Lösung besteht darin, möglichst uninformative a priori Verteilungen zu verwenden. Intuitiv würde man die uniforme Verteilung als nicht informativ bezeichnen, aber leider hängt das von der gewählten Parametrisierung ab.