

Wahrscheinlichkeit und Information - wie Henne und Ei ?

K. Schönhammer, U Göttingen

Einleitung

Ein Fazit der Bemühungen unserer Arbeitsgruppe zur Natur der Information ist wohl, dass jeder den Begriff “Information” intuitiv versteht, er aber beim Versuch ihn dingfest zu machen, eher schwimmt. Oft bewirkt eine Information direkt eine Handlung. Wegen der häufigen Verspätungen bei der Bahn, rufe ich meine Frau aus dem Zug an, wenn er in Kassel abgefahren ist. Fünf Minuten später fährt sie los, um mich abzuholen. Wie kurz kann das Gespräch sein, um die Aktion in Gang zu setzen? Reicht ein “jetzt” oder gar nur ein bit, falls man das mit dem Handy könnte? Das hängt natürlich von getroffenen Vereinbarungen ab. Shannon hat ein quantitatives Maß für den Informationsgehalt einer Nachricht vorgeschlagen. Es ist eine Funktion der *Wahrscheinlichkeit* der Nachricht.

Auch der Begriff der Wahrscheinlichkeit scheint intuitiv verständlich. Der Versuch ihn zu definieren, führt aber leicht zu Zirkelschlüssen. In meinen Vorträgen zum Entropiekonzept in der Physik habe ich dargelegt, dass der Wahrscheinlichkeitsbegriff in der Statistischen Mechanik eine zentrale Rolle spielt. Die von Neumannsche Entropiedefinition verwendet Wahrscheinlichkeiten. Daher ist es wichtig, sich zu verdeutlichen, wie man Wahrscheinlichkeiten angeben, bzw. interpretieren kann. Im Folgenden wird dargelegt, dass es zwei sehr unterschiedliche Interpretationen gibt. Das Wechselspiel zwischen Wahrscheinlichkeit und Information kehrt sich beim Übergang von der einen Interpretation zur anderen gerade um. Daher der Titel meines Vortrags.

Der Hauptteil des Vortrags ist eine hoffentlich einfache Einführung in die Wahrscheinlichkeitstheorie unter dem Gesichtspunkt, das “Henne-Ei-Problem” zu verdeutlichen. Es werden mehrere Beispiele ausführlich behandelt. Diese Einführung ist für nicht-Experten gedacht und erfordert nur geringe mathematische Vorkenntnisse.

Wahrscheinlichkeitstheorie

Den Begriff “Wahrscheinlichkeit” findet man nicht nur in der Mathematik oder den Natur- oder anderen Wissenschaften, sondern auch in der Alltagssprache. Bereits hier zeigt sich, dass der Begriff mit unterschiedlichen Bedeutungen belegt ist. Hier zwei Beispiele:

- Wie groß ist die Wahrscheinlichkeit, dass ich bei einmaligem Würfeln eine “6” erhalte (und damit das Spiel beginnen darf) ?

- Die Wahrscheinlichkeit, dass es morgen regnen wird, ist achtzig Prozent.

Auch ohne eine präzise Definition des Begriffs Wahrscheinlichkeit, die sich als problematisch herausstellt, weiß man in beiden Fällen, was gemeint ist. Das erste Beispiel steht in Beziehung zur Entstehung der Wahrscheinlichkeitstheorie, als praktisches Hilfsmittel zur Strategie in *Glücksspielen*. Wichtige Beiträge dazu aus dem siebzehnten Jahrhundert stammen von Blaise Pascal, Jakob Bernoulli und Pierre Simon Laplace [1]. Bereits bei diesen Anfängen stellte es sich als einfacher heraus, mit Wahrscheinlichkeiten zu hantieren, als sie zu definieren. Diese Regeln wurden 1933 von **A. Kolmogorov** axiomatisch formuliert [2]. Dazu benötigt man zwei Typen von Angaben:

a) die Menge Ω der möglichen Ereignisse E_i

b) die Wahrscheinlichkeiten $P(E_i)$ für das Auftreten der Ereignisse E_i

zu a) Die Menge kann **diskret** sein, wie bei der Augenzahl beim einmaligen Würfeln: $\Omega = \{E_1, E_2, E_3, E_4, E_5, E_6\} \equiv \{1, 2, 3, 4, 5, 6\}$, d.h. E_i ist das Ereignis mit der Augenzahl i .

Neben den "elementaren Ereignissen" betrachtet man häufig **zusammengesetzte** Ereignisse. Beim einmaligen Würfeln z.B. das (zusammengesetzte) Ereignis A , dass eine gerade Augenzahl gewürfelt wird, d.h. $A = \{2, 4, 6\}$.

Bei zweimaligem Würfeln besteht Ω aus den 36 Paaren (ij) , z.B. (61), wobei es auf die Reihenfolge von i und j ankommt, d.h. (61) und (16) sind verschiedene Ereignisse.

Die Geschwindigkeit eines Teilchens oder die Länge eines Stabes sind Beispiele, in denen die Ereignisse **kontinuierlichen** Charakter haben. Da dieser Fall mathematisch anspruchsvoller ist, werden wir uns hier auf den **diskreten** Fall beschränken.

zu b): Die Wahrscheinlichkeit $P(A)$ für ein zusammengesetztes Ereignis A ist die **Summe** aus den Wahrscheinlichkeiten $P(E_i)$, aus denen sich das Ereignis zusammensetzt. Für das obige Beispiel also $P(A) = P(E_2) + P(E_4) + P(E_6)$. Allgemeiner lauten die Axiome

- Für jede Teilmenge $A \in \Omega$ gilt: $P(A) \geq 0$
- Für alle elementfremden Teilmengen A und B gilt für die Vereinigungsmenge $A \cup B$:

$$P(A \cup B) = P(A) + P(B)$$

- Für die gesamte Menge Ω gilt: $P(\Omega) = 1$

Wo die “Wahrscheinlichkeiten” herkommen, dazu wird nichts gesagt. Dazu betrachten wir unser erstes Beispiel mit dem Würfeln einer sechs. Von einem nicht-präparierten Würfel erwartet man aus **Symmetriegründen**, dass alle Augenzahlen **gleichwahrscheinlich** sind. Da die Summe aller sechs Wahrscheinlichkeiten nach den Axiomen gleich eins ist, setzt man also

$$P(E_i) \equiv P(i) = \frac{1}{6}$$

für alle i von 1 bis 6.

Wenn man das “Experiment” *einmal Würfeln* sehr häufig wiederholt, so merkt man, dass im Mittel jede Augenzahl von 1 bis 6 **gleichhäufig** vorkommt. Hat man N -mal gewürfelt, so gilt für die Zahl N_6 der Ereignisse mit einer sechs ungefähr $N_6/N \approx 1/6$, was $P(6) = 1/6$ entspricht.

Wenn man sich die Wahrscheinlichkeiten $P(E_i)$ *nicht* aus einem Symmetrieargument verschaffen kann, so kann man die bei N -facher Wiederholung des entsprechenden “Experiments” gefundenen relativen Häufigkeiten N_i/N als die gesuchten Wahrscheinlichkeiten betrachten

$$P(E_i) = \frac{N_i}{N} .$$

Für unser Würfelbeispiel kommen wir so, wenn N nur groß genug ist, zu denselben Wahrscheinlichkeiten, wie mit dem Symmetrieargument.

Hier also die erste Interpretation von Wahrscheinlichkeiten:

I1: Wahrscheinlichkeiten als relative Häufigkeiten von Ereignissen

I1 ist die Interpretation von Wahrscheinlichkeiten, die z.B. Studenten der Physik in Vorlesungen wie “Statistische Mechanik” oder “Quantenmechanik” üblicherweise präsentiert bekommen. In der statistischen Mechanik wird ein “Ensemble” von identisch präparierten (makroskopischen) Systemen betrachtet und man berechnet physikalische Größen als Mittel über die Werte an den einzelnen Systemen. Das Ensemble existiert natürlich nur im Kopf des Theoretikers. Es wäre viel zu teuer all die Kopien anzuschaffen. Man wiederholt lieber das Experiment mit der *einen* Anordnung mehrfach.

Das mit Hilfe der Bornschen Wahrscheinlichkeitsinterpretation des Absolutquadrats der Wellenfunktion vorhergesagte Interferenzmuster bei der Streuung an einem Doppelspalt mit verdünnten Strahlen erscheint erst allmählich.

Nun zum ersten Zusammenhang mit dem **Informationsbegriff**:

Bei der Entwicklung des Morse-Alphabets war es eine wichtige Einsicht, auf die **relative Häufigkeit des Auftretens der Buchstaben** in einem üblichen (englischen) Text zu achten, und den *häufig* auftretenden Buchstaben *kurze* Symbole zuzuordnen. So wird der Buchstabe e durch *einen* Punkt, t durch *einen* Strich, die seltenen Buchstaben x,y und z aber durch *vier* Symbole dargestellt. Dadurch konnte man die Übertragungslänge wesentlich verkürzen.

Numeriert man die Buchstaben von a bis z mit $i = 1, 2, \dots, 26$ durch und nimmt als Ereignis E_{27} das Leerzeichen hinzu, so entspricht E_5 dem Buchstaben e, der mit der größten relativen Häufigkeit (als Wahrscheinlichkeit interpretiert!) $p_5 \approx 0.091$ auftritt, während E_{26} , das dem Buchstaben “z” entspricht mit $p_{26} \approx 0.007$ am wenigsten häufig auftritt [4].

Claude **Shannon** hat 1948 eine quantitative Analyse des **Informationsgehalts** einer Nachricht vorgelegt [3,4,5]. Er arbeitete hauptsächlich mit einem **binären** Alphabet mit den “Buchstaben” 0 und 1. Ist die Nachricht ein “string” der Länge N

$$011\dots101 ,$$

so gibt es 2^N verschiedene strings dieser Länge, d.h. die Ereignismenge besteht aus 2^N Elementen. Kommen alle strings mit **gleicher relativer Häufigkeit** vor, so gilt für alle strings $p(E_i) = 1/2^N$. Ein offensichtliches Maß für den Informationsgehalt der Nachricht ist die Zahl N der übertragenen 0 – 1-Entscheidungen (“**bits**”), d.h. die **Länge** des strings. Es gilt

$$h(E_i) := \log_2\left(\frac{1}{p(E_i)}\right) = \log_2 2^N = N \log_2 2 = N ,$$

wobei $\log_2$ der Logarithmus zur Basis 2 ist [6]. Also liefert die Funktion h das als “offensichtlich” angesehene Ergebnis. Shannons wichtige Einsicht war, dass der erste Teil der obigen Formel auch für den Fall nicht gleichhäufiger strings *das* nützliche **Maß für den Informationsgehalt** darstellt

$$h(E_i) := \log_2\left(\frac{1}{p(E_i)}\right) .$$

Der Mittelwert über das gesamte Ensemble wird **Informationsentropie** H genannt

$$H := \sum_i p(E_i) \log_2\left(\frac{1}{p(E_i)}\right) .$$

Auf den Zusammenhang mit der Entropiedefinition in der statistischen Mechanik (s. meine ersten beiden Vorträge) soll hier aus Zeitgründen nicht eingegangen werden.

Für das übliche Alphabet von a bis Leerzeichen gilt für die Buchstaben e und z

$$h(e) = \log_2\left(\frac{1}{0.091}\right) \approx 3.5 , \quad h(z) = \log_2\left(\frac{1}{0.007}\right) \approx 10.4 ,$$

und die Informationsentropie, d.h. der **mittlere Informationsgehalt** eines Buchstabens ist $H \approx 4.1$.

Bei Shannon stand das *praktische* Problem im Vordergrund, eine “Nachricht” so genau wie möglich von einem Ort an einen anderen zu transferieren. Ob sie tatsächlich eine **Bedeutung** hat oder nicht, spielte für ihn keine Rolle (“These semantic aspects of communication are irrelevant for the engineering problem” [3].)

Shannons Definitionen, die auf der Interpretation von Wahrscheinlichkeiten als relativen Häufigkeiten beruhen, besagen zum “Henne und Ei”-Problem

Wahrscheinlichkeit $\rightarrow$ Information(smaß).

Nun zu unserem **zweiten Beispiel** der Verwendung von Wahrscheinlichkeiten, der Frage, ob es morgen wohl regnen wird. Offensichtlich macht die Interpretation des Wahrscheinlichkeitsbegriffes als relative Häufigkeiten hier keinen Sinn. Man kann die heutige Großwetterlage nicht (mehrfach!) wiederholen. Hier ist die Angabe einer Wahrscheinlichkeit als **Maß für die Überzeugung (Erwartung)**, englisch “**degree of belief**”, anzusehen, dass es morgen regnen wird. Es klingt vielleicht überraschend, dass man mit solch einem vagen Begriff eine quantitative Theorie formulieren kann. Cox hat 1948 gezeigt, dass jeder Versuch, einer *konsistenten* **quantitativen Beschreibung von rationalem Schließen** bei unvollständiger Information notwendigerweise zu Regeln führt, die mit den Kolmogorov-Axiomen identisch sind [7]. Die Interpretation

I2: Wahrscheinlichkeiten als “degree of belief”

wird **Bayesscher Standpunkt** genannt, nach Thomas Bayes (1702-1761), dessen Arbeit dazu posthum veröffentlicht wurde. Wie an mehreren Beispielen im Folgenden gezeigt wird, gilt hier

Information $\rightarrow$ Wahrscheinlichkeit(s-update).

Zentrale Bedeutung hat dabei der Begriff der **bedingten Wahrscheinlichkeiten** und das **Bayessche Theorem**. Da wir mit der Häufigkeitsinterpretation begonnen haben, wird dieser Begriff im Folgenden auf diese Art eingeführt [8]. Die so motivierte Definition ist aber auch ohne diese Interpretation möglich.

Bedingte Wahrscheinlichkeiten:

Wir betrachten das Beispiel einer (bezüglich des folgenden Tests “zufälligen”) Ansammlung aus N Individuen, in der sich N_F Frauen befinden. Es wird getestet, wieviele der Anwesenden farbenblind sind. Die Gesamtzahl der Farbenblinden stellt sich als N_{fb} heraus, davon sind $N_{F,fb}$ Frauen. Das entspricht mit der Häufigkeitsinterpretation folgenden Wahrscheinlichkeiten

$$P(F) = \frac{N_F}{N} , \quad P(fb) = \frac{N_{fb}}{N} , \quad P(F, fb) = \frac{N_{F,fb}}{N} .$$

Jetzt kann man die “Verbundwahrscheinlichkeit” $P(F, fb)$ “Frau und farbenblind” elementar durch “Erweitern” umformen

$$P(F, fb) = \frac{N_{F,fb}}{N} = \frac{N_{F,fb}}{N_F} \frac{N_F}{N} = \frac{N_{F,fb}}{N_F} P(F) =: P(fb|F)P(F) .$$

Dabei haben wir die **bedingte Wahrscheinlichkeit** $P(fb|F) = N_{F,fb}/N_F$, “wenn Frau, dann farbenblind” (im Symbol von links nach rechts gelesen) eingeführt. Diese Größe wird mit dem Wert in medizinischen Statistiken ungefähr übereinstimmen. Alternativ kann man *formal* auch die bedingte Wahrscheinlichkeit $P(F|fb)$, “wenn farbenblind, dann Frau” einführen

$$P(F, fb) = \frac{N_{F,fb}}{N} = \frac{N_{F,fb}}{N_{fb}} \frac{N_{fb}}{N} = \frac{N_{F,fb}}{N_{fb}} P(fb) =: P(F|fb)P(fb) .$$

Diese bedingte Wahrscheinlichkeit $P(F|fb)$ kann man **nicht** irgendwelchen Statistiken entnehmen, da sie offensichtlich von der Zusammensetzung der Ansammlung abhängt. Besteht sie z.B. nur aus Frauen, so gilt $P(F|fb) = 1$. Allgemein ergibt sich $P(F|fb)$ durch Gleichsetzung der beiden Ausdrücke für $P(F, fb)$ und Division durch $P(fb)$ als

$$\boxed{P(F|fb) = \frac{P(fb|F)P(F)}{P(fb)} .}$$

Das ist **Bayes’ Theorem** (für unseren Fall). Für andere Beispiele (s.u.) muss man nur andere Bezeichnungen verwenden.

Für die elementaren Umformungen, die nötig waren, klingt der Ausdruck “Theorem” sehr hochgestochen. Bei der Herleitung kann man wie bereits erwähnt, auf die Frequenzinterpretation verzichten und in beiden Darstellungen für $P(F, fb)$ jeweils die drei mittleren Identitäten weglassen.

Wie später diskutiert wird, ist es die **Interpretation** obiger Formel, die zu teilweise heftigen Auseinandersetzungen geführt hat [9].

Beispiele

1) Urnen:

Ein Standardbeispiel zur Veranschaulichung der Wahrscheinlichkeitsrechnung sind Urnen, die verschiedene Kugeln enthalten. An diesem Beispiel wollen wir beide Interpretationen, die “direkte” I1 und die “inverse” I2 veranschaulichen.

Die hier verwendeten “Urnen” sollen so funktionieren, wie die im Fernsehen zu sehende *Lottomaschine*. Sie besitzt einen “Umrührmechanismus”, der dafür sorgt, dass die anschließend herausfallende Kugel als *völlig zufällig* gewählt betrachtet werden kann. Unsere Urnen sind aber, im Gegensatz zur Lottomaschine, in einen **undurchsichtigen Kasten** eingebaut. Nach Knopfdruck wird erst umgerührt und dann **eine** Kugel ausgespuckt. Wir nehmen an, **zwei** solcher Urnen vorliegen zu haben, mit Aufkleber A und B, sonst aber in ihrem Äußeren völlig identisch.

In **Urne A** befinden sich **47 weiße (w) und 2 rote (r) Kugeln**. Dann ist die Wahrscheinlichkeit nach Knopfdruck eine weiße oder eine rote Kugeln zu erhalten, leicht durch “Abzählen” zu finden

$$p_A^{(0)}(w) = \frac{47}{49}, \quad p_A^{(0)}(r) = \frac{2}{49}.$$

Dabei steht der Index A für die Bezeichnung der Urne und das hochgestellte (0) für die Tatsache, dass noch keine Kugel entnommen wurde. Diese Wahrscheinlichkeiten können auch als die *bedingten Wahrscheinlichkeiten*, wenn Urne A, dann w oder r aufgefasst werden

$$p^{(0)}(w|A) = p_A^{(0)}(w) = \frac{47}{49}, \quad p^{(0)}(r|A) = p_A^{(0)}(r) = \frac{2}{49}.$$

Sei nun die (erste) entnommene Kugel eine **rote** Kugel. In der Urne befinden sich dann noch 48 Kugeln, 47 weiße und eine rote. Nun werde eine zweite Kugel entnommen. Die neuen Wahrscheinlichkeiten sind

$$p^{(r)}(w|A) = p_A^{(r)}(w) = \frac{47}{48}, \quad p^{(r)}(r|A) = p_A^{(r)}(r) = \frac{1}{48}.$$

Dabei deutet das hochgestellte (r) an, dass beim ersten Knopfdruck eine rote Kugel erschienen ist. Jetzt kann man z.B. nach der Wahrscheinlichkeit fragen, dass **beide** gezogenen Kugeln rot sind, d.h. die Kugelfolge $\{r, r\}$ vorliegt. Dazu muss man einfach die beiden oben angegebenen Wahrscheinlichkeiten **multiplizieren**

$$p_A^{(0)}(\{r, r\}) = p_A^{(r)}(r)p_A^{(0)}(r) = \frac{1}{48} \cdot \frac{2}{49} \approx 0.00085.$$

Die Wahrscheinlichkeit, dass eine dritte gezogene Kugel auch rot ist, ist offensichtlich gleich Null.

In **Urne B** befinden sich vor der ersten Ziehung **49 rote Kugeln**, d.h. **alle** Kugeln sind **rot**. Für die entsprechenden Wahrscheinlichkeiten gilt also

$$p^{(0)}(w|B) = p_B^{(0)}(w) = 0, \quad p^{(0)}(r|B) = p_B^{(0)}(r) = 1,$$

sowie

$$p^{(r)}(w|B) = p_B^{(r)}(w) = 0, \quad p^{(r)}(r|B) = p_B^{(r)}(r) = 1.$$

Hier ist also alles deterministisch. Bei zweifachem Ziehen aus Urne B wird man z.B. immer zwei rote Kugeln finden

$$p^{(0)}(\{rr\}|B) = 1$$

usw.

Nun werden die Aufkleber A und B an den Urnen entfernt und in unserer Abwesenheit “zufällig” nebeneinander auf Position 1 und 2 auf einen Tisch gestellt. Dann kann man z.B. fragen, mit welcher Wahrscheinlichkeit man bei einmaligem Knopfdruck an der beliebig gewählten Urne, z.B. der an Position 1, eine rote Kugel findet. Da für die Wahrscheinlichkeit, Urne A oder B gewählt zu haben,

$$p_1^{(0)}(A) = p_1^{(0)}(B) = \frac{1}{2}$$

gilt, folgt für die Wahrscheinlichkeit, eine rote Kugel zu finden

$$p_1^{(0)}(r) = p^{(0)}(r|A)p_1^{(0)}(A) + p^{(0)}(r|B)p_1^{(0)}(B) = \frac{2}{49} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{51}{98}.$$

Völlig analog erhält man die Wahrscheinlichkeit, dass zweifaches Ziehen aus der Urne an Platz 1 zwei rote Kugeln liefert als

$$p_1^{(0)}(\{r, r\}) = p^{(0)}(\{r, r\}|A)p_1^{(0)}(A) + p^{(0)}(\{r, r\}|B)p_1^{(0)}(B) = \left(\frac{1}{48} \cdot \frac{2}{49} + 1 \right) \cdot \frac{1}{2}.$$

Soweit einfaches “**direktes**” Hantieren mit Wahrscheinlichkeiten.

Nun wollen wir ein “**inverses**” Problem betrachten und zur Lösung **Bayes’ Theorem** verwenden:

Die beiden noch voll besetzten Urnen stehen ohne die Aufkleber auf dem Tisch. Sie werden aufgefordert, Urne B möglichst schnell zu identifizieren. Was sollte man tun?

Man überlegt sich zuerst die sogenannten *priori*-Wahrscheinlichkeiten (“**priors**” [9]), wobei hier nur eine Zahlenangabe nötig ist

$$p_1^{(pr)}(B) = p_2^{(pr)}(A) = 1 - p_2^{(pr)}(B) = 1 - p_1^{(pr)}(A) .$$

Wenn man gar kein Vorwissen hat, wird man die beiden Anordnungen der Urnen als gleichwahrscheinlich annehmen, d.h. obige Wahrscheinlichkeiten gleich $1/2$ setzen, wie beim letzten “direkten” Beispiel.

Falls Sie einfach eine der beiden Urnen greifen, wird die Wahrscheinlichkeit, dass sie wirklich Urne B gefunden haben, also nur $1/2$ sein. Mit derselben Wahrscheinlichkeit haben Sie Urne A gegriffen. Daher wird man anfangen, den Knopf an einer der Urnen zu drücken.

Einmal drücken:

Erscheint eine weiße Kugel, so ist klar, dass man Urne A gewählt hat. Erscheint dagegen eine rote Kugel, so muss man bei Wahrscheinlichkeiten bleiben.

Bayes’ Theorem für die Wahrscheinlichkeit $p_1^{(0)}(B|r)$, dass die Urne auf Position 1 die Urne B ist, wenn dort eine rote Kugel erscheint, lautet (die benötigte Wahrscheinlichkeit $p_1^{(0)}(r)$ haben wir bereits berechnet)

$$p_1^{(0)}(B|r) = \frac{p^{(0)}(r|B)p_1^{(0)}(B)}{p_1^{(0)}(r)} = \frac{1 \cdot \frac{1}{2}}{\frac{51}{98}} = \frac{49}{51} \approx 0.96 \equiv p_1^{(post)}(B).$$

und $p_1^{(0)}(A|r) = 2/51 \equiv p_1^{(post)}(A)$. Dabei sind $p_1^{(post)}(B)$ und $p_1^{(post)}(A)$ die sogenannten **posteriori**-Wahrscheinlichkeiten, dass die Urne auf Position 1 die Urne B oder A ist, wenn dort eine rote Kugel gefunden wurde. Die Wahrscheinlichkeit, dass die Urne auf Position 1 die Urne B ist, ist also durch das Finden der roten Kugel von $1/2$ auf $49/51 \approx 0.96$ **angewachsen**. Wem das noch nicht genügt, der wird

Zweimal drücken:

Man kann nun entweder wieder bei der Urne auf Position 1 drücken oder aber die andere wählen. Erscheint eine weiße Kugel, so ist alles klar. Ist die zweite Kugel dagegen wieder rot, so muss man nochmals “Bayes verwenden”. Dazu verwendet man die posteriori-Wahrscheinlichkeiten vom ersten Drücken als *neue priors* (pr') für die zweite Runde

$$p_1^{post} \rightarrow p_1^{pr'} .$$

Für den Fall, dass **wieder bei der Urne auf Position 1 gedrückt wird**, erhält man

$$p_1^{(r)}(B|r) = \frac{p^{(r)}(r|B)p_1^{(pr')}(B)}{p_1^{(r)}(r)} = \frac{1 \cdot p_1^{(pr')}(B)}{p_1^{(r)}(r)} = \frac{49/51}{p_1^{(r)}(r)} .$$

Also benötigen wir noch

$$p_1^{(r)}(r) = p^{(r)}(r|A)p_1^{(pr')}(A) + p^{(r)}(r|B)p_1^{(pr')}(B) = \frac{1}{48} \cdot \frac{2}{51} + \frac{49}{51} = \frac{49}{51} \cdot \left(1 + \frac{1}{48} \cdot \frac{2}{49}\right).$$

Eingesetzt ergibt sich

$$p_1^{(r)}(B|r) = \frac{1}{1 + \frac{1}{48} \cdot \frac{2}{49}} \approx 0.99915.$$

Man kann also “fast sicher” sein, dass die Urne auf Platz 1 tatsächlich Urne B mit den ausschließlich roten Kugeln ist.

Als Test machen wir das “updating” der Wahrscheinlichkeiten nicht in zwei Schritten, sondern in **einem Schritt** und verwenden Bayes Theorem in der Form

$$p_1^{(0)}(B|\{r, r\}) = \frac{p^{(0)}(\{r, r\}|B)p_1^{(0)}(B)}{p_1^{(0)}(\{r, r\})} = \frac{1 \cdot \frac{1}{2}}{\left(\frac{1}{48} \cdot \frac{2}{49} + 1\right) \cdot \frac{1}{2}} = \frac{1}{1 + \frac{1}{48} \cdot \frac{2}{49}}$$

Also ergibt sich, wie erwartet, dasselbe Ergebnis.

Es ist interessant, sich auch den Fall anschauen, bei dem die zweite Kugel aus der Urne auf Platz 2 entnommen wird.

Urnenwechsel:

Nachdem aus der Urne auf Platz 1 eine rote Kugel gezogen wurde, ist die Urne auf Platz 2 mit hoher Wahrscheinlichkeit $p_2^{(post)}(A) = p_1^{(post)}(B) = 49/51$ die Urne A. Da die Urne auf Platz 2 noch voll ist, gilt für die Wahrscheinlichkeit dort eine rote Kugel zu erhalten

$$p_2(r) = p^{(0)}(r|A)p_2^{(post)}(A) + p^{(0)}(r|B)p_2^{(post)}(B) = \frac{2}{49} \cdot \frac{49}{51} + 1 \cdot \frac{2}{51} = \frac{4}{51}.$$

Also findet man mit der komplementären Wahrscheinlichkeit $47/51 \approx 0.92$ eine weiße Kugel und damit ist in diesem Fall klar, dass die Urne auf Platz 1 die Urne B ist.

Für den Fall, dass die **Kugel aus der Urne auf Platz 2 auch rot ist**, ist man “so klug als wie zuvor”, d.h. vor dem ersten Ziehen. Die *Zusatzinformation* hat die Situation relativ zur ersten Ziehung also verschlechtert. Das folgt natürlich auch wieder “mit Bayes” (“2 = B” steht für “Urne 2 ist Urne B”)

$$p(2 = B|r) = \frac{p(r|2 = B)p_2^{(pr')}(B)}{p_2(r)} = \frac{1 \cdot 2/51}{4/51} = \frac{1}{2} \equiv p_2^{post'}(B).$$

Zum Glück tritt dies nur in ca. acht Prozent der Zweitziehungen auf.

Dreimal drücken:

Hat man zweimal auf Platz 1 gedrückt und zwei rote Kugeln erhalten, so liefert die dritte Kugel aus dieser Urne die endgültige Antwort. Ist sie rot, so haben wir es mit Urne B zu tun, ist sie weiß mit Urne A.

Gab es rote Kugeln aus **beiden** Urnen, so ist man beim Erscheinen einer weiteren roten Kugel immer noch nicht fertig...

Dieses Beispiel zeigt klar, wie die jeweils **zusätzliche Information** über die gezogene Kugel die Wahrscheinlichkeiten modifiziert:

$$p^{(pr)} + \text{Information} \rightarrow p^{(post)} ,$$

wobei der Vorgang mehrfach durchlaufen werden kann. Dabei werden die $p^{(post)}$ des m -ten Durchlaufs die $p^{(pr)}$ des $m + 1$ -ten Durchlaufs. Wahrscheinlichkeiten als “degree of belief” haben also **keine absolute Bedeutung**, sondern hängen vom Informationsstand ab.

Ähnlich in der Durchführung, aber näher an der Praxis ist das zweite Beispiel.

2) Medizinische Tests

Dieses Beispiel macht die Nützlichkeit des Bayesschen Theorems *für jedermann* deutlich. Mit den überraschenden Ergebnissen sollten sowohl Ärzt(inn)e(n) als auch Patient(inn)en vertraut sein.

Eine Patientin, von der die Ärztin nicht weiß, ob sie eine spezielle Krankheit hat (“k” für krank) oder nicht (“g” für gesund), lässt einen *Test* über sich ergehen. Solche Tests sind nie fehlerfrei. Bei kranken Patientinnen wird der Test mit einer geringen Wahrscheinlichkeit negativ (n) sein, d.h. die Erkrankung wird übersehen. Die entsprechende bedingte Wahrscheinlichkeit sei

$$P(n|k) = \epsilon , \quad P(p|k) = 1 - \epsilon ,$$

wobei ϵ bei einem guten Test klein ist, z.B. $\epsilon = 0.02$ (zwei Prozent der Erkrankungen werden übersehen). Das Symbol “p” steht hier für positiven Test und **nicht** für Wahrscheinlichkeit. Ebenso passieren “falsch positive” Fälle

$$P(p|g) = \epsilon' , \quad P(n|g) = 1 - \epsilon' ,$$

wobei ϵ' bei einem guten Test ebenfalls klein ist, aber durchaus von ϵ verschieden sein kann. z.B. $\epsilon' = 0.01$.

Die Patientin habe nun ein **positives Testergebnis**. Mit welcher Wahrscheinlichkeit ist sie dann wirklich erkrankt?

Aus dem Urnenbeispiel sollte klar sein, dass man zur Verwendung der Bayesschen Regel einen “**prior**” benötigt. Hier sollte man zwei Fälle unterscheiden:

- a) die “Reihenuntersuchung”
- b) die “gezielte” Untersuchung,

bei der Resultate von einem *Vortest*, Symptome, die auf das Vorliegen der Krankheit schließen lassen könnten, oder die familiäre Vorgeschichte Grund zur Besorgnis geben.

Die Bayessche Regel für die Wahrscheinlichkeit, bei einem positiven Test erkrankt zu sein, lautet

$$P^{(post)}(k) = P(k|p) = \frac{P(p|k)P^{(pr)}(k)}{P(p)} = \frac{(1 - \epsilon)P^{(pr)}(k)}{(1 - \epsilon)P^{(pr)}(k) + \epsilon'(1 - P^{(pr)}(k))} .$$

Für das Folgende spielt es eine zentrale Rolle, welcher der beiden Summanden im Nenner der Größere ist. Wenn der (hintere) *falsch positive* Beitrag überwiegt, muss sich unsere Patientin bei positivem Befund noch nicht allzu große Sorgen machen. Wir betrachten nun die beiden Fälle a) und b):

- a) Falls keine Zusatzinformation vorliegt, kann man als $P^{(pr)}(k)$ die Häufigkeitsrate der Krankheit für die entsprechende Altersgruppe verwenden, z.B. $P^{(pr)}(k) = 0.005$. Dann erhält man für die angegebenen Werte von ϵ und ϵ'

$$P^{(post)}(k) = \frac{0.98 \cdot 0.005}{0.98 \cdot 0.005 + 0.01 \cdot 0.995} = \frac{49}{49 + 95} \approx 0.35 .$$

Obwohl das Verhältnis $P^{(post)}(k)/P^{(pr)}(k)$ sehr groß ist, stellt sich die Wahrscheinlichkeit wirklich erkrankt zu sein, als “nur” 35 Prozent heraus, also noch kein Grund zu verzweifeln, da es sich eher um einen **falsch positiven** Befund handelt.

- b) Um die Rechnung zu vereinfachen, nehmen wir als prior auf Grund der Vorinformation $P^{(pr)}(k) = 1/2$ an. Dann gilt $P(p) = (1 - \epsilon + \epsilon')/2$ und im Spezialfall $\epsilon = \epsilon'$ folgt $P(p) = 1/2$. Die Wahrscheinlichkeit bei einem positiven Test wirklich krank zu sein, ist dann direkt durch die bedingte Wahrscheinlichkeit gegeben

$$P^{(post)}(k) = P(k|p) = 1 - \epsilon = 0.98 .$$

Jetzt muss man sich wirklich Sorgen machen. Ist der Test dagegen **negativ**, so ist die Wahrscheinlichkeit tatsächlich **gesund** zu sein, ebenfalls durch $1 - \epsilon = 0.98$ gegeben und Freude ist angesagt.

3) Das Ziegenproblem [10]:

Sie nehmen an einer Spielshow im Fernsehen teil, bei der Sie eine von drei verschlossenen Türen auswählen sollen. Hinter einer Tür wartet der Preis, ein Auto, hinter den beiden anderen stehen Ziegen. Sie zeigen auf eine Tür, sagen wir Nummer **eins**. Sie bleibt vorerst geschlossen. Der Moderator weiß, hinter welcher Tür sich das Auto befindet; mit den Worten “Ich zeige Ihnen mal was” öffnet er eine andere Tür, zum Beispiel Nummer **drei**, und eine meckernde Ziege schaut ins Publikum. Er fragt: **“Bleiben Sie bei Nummer eins, oder wählen Sie Nummer zwei?”**

Dieses Problem, das in verschiedenen Zirkeln erstaunlich kontrovers diskutiert wurde, lässt sich *ohne Berechnung von bedingten Wahrscheinlichkeiten* mit einem einfachen Argument leicht lösen:

Steht das Auto hinter der Tür eins (Wahrscheinlichkeit $1/3$), so verliert man beim Wechsel (oberstes Bild). Steht das Auto aber hinter einer der beiden anderen Türen (Wahrscheinlichkeit $2/3$), dann ist der Wechsel erfolgreich (die beiden unteren Bilder). Die Gewinnchance verdoppelt sich also beim Wechsel.



Nun wollen wir darlegen, warum die Verwendung des Bayesschen Theorems hier nicht sinnvoll ist.

Der Ereignisraum besteht aus den Paaren (A_i, M_j) , wobei i und j von 1 bis 3 läuft. Dabei bedeutet A_i “das Auto steht hinter der Tür i ”, und M_j “der Moderator hat die Tür j ” geöffnet. Bei der Anwendung der Bayesschen Regel verwendet man die bedingten Wahrscheinlichkeiten $p_1(M_j|A_i)$, wobei der Index 1 andeutet, dass Tür 1 gewählt wurde. Die Ansage des Moderators legt aber nicht alle dieser Größen fest

$$\begin{aligned} p_1(M_1|A_1) &= 0, & p_1(M_2|A_1) &= ?, & p_1(M_3|A_1) &= ? \\ p_1(M_1|A_2) &= 0, & p_1(M_2|A_2) &= 0, & p_1(M_3|A_2) &= 1 \\ p_1(M_1|A_3) &= 0, & p_1(M_2|A_3) &= 1, & p_1(M_3|A_3) &= 0. \end{aligned}$$

Falls das Auto hinter der Tür 1 steht, könnte die Regel des Moderators sein, eine der beiden anderen Türen “zufällig”, d.h. mit Wahrscheinlichkeit $1/2$ zu öffnen. Er könnte aber die Anweisung haben, die Tür $i + 1$ zu öffnen, also in unserem Fall $i = 1$ die Tür 2 (bei $i = 3$ muss er modulo 3 rechnen, also Tür 1 öffnen). Oder die Anweisung für den Moderator könnte lauten, die Tür $i + 1$ (hier Tür 2) mit Wahrscheinlichkeit w_0 und Tür $i + 2$ (hier Tür 3) mit Wahrscheinlichkeit $1 - w_0$ zu öffnen.

Außer im Spezialfall $w_0 = 1/2$, (den G. von Randow in seinem Buch zum Problem [10] verwendet,) liefern die bedingten Wahrscheinlichkeiten $p_1(A_j|M_i)$ gar nicht direkt die gesuchte Gewinnwahrscheinlichkeit. Man kann die $p_1(A_j|M_i)$ gar nicht angeben, wenn man w_0 nicht kennt.

Es sind die **Verbundwahrscheinlichkeiten** $p_1(M_i, A_j)$ die direkt zur Lösung führen. Diese kann man leicht angeben. Da $p(A_j) = 1/3$ für alle j gilt, folgt mit $p_1(M_i, A_j) = p_1(M_i|A_j)p(A_j)$

$$\begin{aligned} p_1(M_1, A_1) &= 0, & p_1(M_2, A_1) &= \frac{w_0}{3}, & p_1(M_3, A_1) &= \frac{1 - w_0}{3} \\ p_1(M_1, A_2) &= 0, & p_1(M_2, A_2) &= 0, & p_1(M_3, A_2) &= \frac{1}{3} \\ p_1(M_1, A_3) &= 0, & p_1(M_2, A_3) &= \frac{1}{3}, & p_1(M_3, A_3) &= 0. \end{aligned}$$

Die gesamte Wahrscheinlichkeit, dass das Auto hinter der Tür 1 steht, ergibt sich als Summe der ersten Zeile zu $1/3$ (wobei w_0 herausgefallen ist), halb so groß, wie beim Türwechsel. Betrachtet man direkt die komplementären Wahrscheinlichkeiten, d.h. die Zeilen zwei und drei, so hätte man auch in den Verbundwahrscheinlichkeiten in der ersten Zeile die Fragezeichen stehen lassen können und trotzdem das richtige Ergebnis - doppelte Chance bei Wechsel - erhalten.

Zum Streit der Schulen

Im letzten Jahrhundert gab es einen heftigen Streit zwischen “Frequentisten” (F) und “Bayesianern” (B), wobei erstere das stärkere Lager bildeten. Exzellente Lehrbücher zu den beiden Standpunkten stammen von Feller (F) [10] und Jaynes (B) [8]. Unter meinen Kollegen sind die Fs in der Mehrheit, und wie angedeutet, ist das auch der “übliche” Standpunkt in den Vorlesungen zur theoretischen Physik. In der Frage der Interpretation “verrauschter” experimenteller Daten gibt es dagegen glühende Befürworter der Bayesschen Methode unter den Experimentalphysikern [11,12,13] und auch so bekannte Theoretiker wie der Nobelpreisträger P.W. Anderson wundern sich, dass es nicht mehr Bayesianer unter den Physikern gibt [14]. Zum Testen von Hypothesen stellen sie die *richtige Frage*, “wie wahrscheinlich ein Hypothese beim Vorliegen der Daten ist”, und nicht die Frage des anderen Lagers “angenommen, die Hypothese stimmt, was ist die Wahrscheinlichkeit der Daten”.

Ob sich eines der Lager langfristig durchsetzen wird, ist schwer zu sagen. Was ist die Wahrscheinlichkeit dafür....?

Literaturangaben:

- [1] zur Entwicklung des Wahrscheinlichkeitsbegriffs:
L. Daston, *Classical Probability in The Enlightenment*, Princeton University Press, 1988
- [2] A. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, 1933
- [3] C.E. Shannon, *A mathematical theory of communication*, Bell Sys. Tech. J. **27** 379-423, 623,-656 (1948)
- [4] ein modernes Lehrbuch dazu ist z.B.:
D. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003
- [5] ein allgemeinverständliche Diskussion dazu findet man z.B. in:
H.C. von Baeyer, *Information - The New Language of Science*, Harvard University Press, 2004
- [6] zu meiner Schulzeit spielte die Funktion $\log_{10}(x)$, der Logarithmus zur Basis 10, eine wichtige Rolle und man besaß dazu eine *Logarithmentafel*. Für beliebige

Basis $b > 0$ z.B. $b = 2$ gilt für die mit dem Argument x monoton anwachsende Logarithmusfunktion

$$\log_b(xy) = \log_b(x) + \log_b(y) , \quad \log_b(b) = 1 , \quad x \geq 1 : \log_b(x) \geq 0 .$$

[7] R.T. Cox, *Probability, Frequency and Reasonable Expectation*, Am. J. Physics, **1**, (1946)

[8] W. Feller, *An Introduction to Probability Theory and its Application*, Vol. I,II, John Wiley and Sons, 1957

[9] E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, 2003

[10] G. von Randow, *Das Ziegenproblem - Denken in Wahrscheinlichkeiten*, roro science, 1992

[11] V. Dose, *Bayesian inference in physics: case studies*, Rep. Prog. Phys. **66**, 1421-1461, (2003)

[12] R.Cousins, *Why isn't every physicist a Bayesian ?*, Am. J. Phys. **63**, 398-410 (1994)

[13] P. Coles, *From Cosmos to Chaos - The Science of Unpredictability*, Oxford University Press, 2006

[14] P.W. Anderson, *The Reverend Thomas Bayes, Needles in Haystacks, and the Fifth Force*, Physics Today, January 1992, 9-10