Skript zum Vorkurs Statistik

Katharina Glomb
Martina Michalikova

09.09.09

Inhaltsverzeichnis

1 Einleitung

Die Biostatistik-Vorlesung im 1.Semester der Monobachelor Biologie und Biophysik stellt viele Studienanfänger vor eine große Herausforderung. Oft fehlt so früh im Studium der Bezug zu den Vorlesungsinhalten und da viele, wenn nicht sogar die meisten Studierenden auch vorher noch nie Stochastik oder Statistik hatten, stellen sich schnell Verwirrung und Überforderung ein.

Seit einigen Jahren organisiert die Fachschaft Biologie zusammen mit den Tutoren der *Mathematik für Biologen*-Vorlesung vom Institut für Theoretische Biologie den Mathematik-Vorkurs. In dessen Rahmen findet auch eine eintägige Einführung in die Biometrie statt, mit dem Ziel, dass Studienanfänger der dazugehörigen Vorlesung besser folgen können und mehr aus ihr mitnehmen. Denn eins steht fest: Für Versuchsplanung und -durchführung sowie die Auswertung von Experimenten sind Kenntnisse in Statistik genau so wichtig wie für das Verständnis wissenschaftlicher Artikel.

Dieses Skript enthält kurz gefasst einige wichtige Begriffe der Biometrie sowie Übungsaufgaben mit Lösungen. Anregungen, Kritik oder Lob können gesendet werden an katharina.glomb@googlemail.com oder direkt an die Fachschaft Biologie (fsbio-berlin@gmx.de).

2 Literaturempfehlungen

Uneingeschränkt empfehlenswert sind folgende drei Bücher:

- Köhler, Schachtel, Voleske: Biostatistik, erschienen im Springer Verlag, 2.Auflage 1996
- Timischl: Biostatistik-Eine Einführung für Biologen, erschienen im Springer Verlag, 1990
- Riede: Mathematik für Biologen, erschienen im Vieweg Verlag, 1993

3 Zufallsversuche

Ein **Zufallsexperiment** hat folgende Eigenschaften:

- Alle möglichen Ereignisse sind vorher bekannt und bilden die *Ergebnismenge* (auch *Stichprobenraum* oder *Ergebnisraum* genannt und mit Ω (Omega) bezeichnet). Zum Beispiel ist beim wiederholten Würfeln bekannt, dass Kombinationen der Zahlen 1 bis 6 auftreten werden.
- Das Experiment kann unter gleichen Bedingungen beliebig oft durchgeführt werden.
- Der Ausgang des Experiments ist dabei für die Einzelfälle nicht bekannt, auch wenn sich Gesetzmäßigkeiten erkennen lassen, wenn man es sehr oft durchführt.

Einen bekannten Sonderfall stellen dabei so genannte *Laplace-Experimente* dar: Hierbei sind alle möglichen Ergebnisse gleich wahrscheinlich (z.B. fairer Würfel, Münzwurf).

Diese Bedingungen können meist nur im Labor hergestellt werden, dennoch ist das Ziel eines jeden Experiments, Rückschlüsse auf die tatsächlichen Vorgänge in der Natur zu ziehen. Insofern ist jedes Experiment ein Zufallsexperiment, weil Messungen immer in nicht vorhersehbarer Weise schwanken. Dann braucht es geeignete Werkzeuge, um Messergebnisse und wahren Wert in Beziehung zu setzen. Dabei muss man seinen Blick auf das Wesentliche richten und geeignete Größen wählen, um seine Fragestellung zu beantworten. Dazu folgendes Beispiel:

Fragestellung: Wie hängen Blattfläche/Sonneneinstrahlung und Energieumsatz bei Pflanzenart xy zusammen?

Problem: Wir haben kein "Energiemessgerät", mit dem wir die eintreffende Sonnenenergie direkt quantifizieren können.

Lösung: Wahl eines geeigneten Merkmals, das den Energieumsatz "repräsentiert", z.B. die Menge der gebildeten Stärke pro Zeiteinheit (Stärke lässt sich relativ einfach nachweisen) pro Blattfläche.

3.1 Merkmal und Merkmalsausprägung

An diesem Beispiel lassen sich auch zwei Begriffe erklären, die bei Zufallsexperimenten von zentraler Bedeutung sind und die man sich fest einprägen sollte: Zum einen das Merkmal, in diesem Fall könnte man es z.B. mit "Glucosemenge pro Zeiteinheit" bezeichnen, und zum anderen die konkrete Merkmalsausprägung, also die Mengen (z.B. in g), die man bei der Durchführung des Experimentes misst. Dabei bezeichnet man das Merkmal immer mit einem Großbuchstaben, oft "X", wodurch ein anderes Wort dafür, das oft verwendet wird, plausibel wird: *Zufallsvariable*. die Ausprägungen oder *Realisationen* werden mit dem dazugehörigen Kleinbuchstaben bezeichnet, mit einem Index für den 1., 2., usw. auftretenden Wert, also in diesem Fall "x_i". Dazu ein Beispiel:

Merkmal	Zufallsvariable	Realisationen (Bsp.)
Motiv auf einer Münze	X	x_1=Kopf, x_2=Zahl
Augenzahl beim Würfeln	Y	y_1=1, y_2=2
Hämatokritwert von Patienten	Z	z_1=43%, z_2=45%

Tabelle 1: Beispiele für Zufallsexperimente.

3.2 Von der Urliste zum Histogramm

Unter einer Urliste versteht man die Aufzeichnung, die man direkt während des Experimentes macht, also die Rohdaten (s. Tabelle 2). Dabei bezeichnet man die Anzahl der Messungen in der Regel mit **n**. Die Urliste ist natürlich nicht sehr übersichtlich oder aussagekräftig und daher überführt man sie in etwas in der Form, wie man es in Tabelle 3 sehen kann.

Realisation	Anzahl
1	IIII
2	II
3	II
4	IIIII
5	II

i	1	2	3	4	5	6	7	8	9
x_i	4	5	3	2	4	1	3	3	1

Tabelle 2: Urlisten für zwei fiktive Experimente mit n=15 (links) bzw. n=9 (rechts).

x_i	H_i	h_i
1	4	0,27
2	2	0,13
3	2	0,13
4	5	0,33
5	2	0,13
$\sum$		0 ,99

Tabelle 3: Liste mit absoluten und relativen Häufigkeiten für die oben links dargestellte Urliste (n=15). Die Abweichung der Summe der relativen Häufigkeiten von 1 kommt durch Rundungsfehler zustande.

Dabei steht das H_i ganz simpel für Häufigkeit, und zwar für *absolute Häufigkeit*, im Gegensatz zu h_i für relative Häufigkeit, also die Häufigkeit eines Ereignisses i im Verhältnis zur Anzahl aller Ereignisse n. Zu beachten ist auch, dass die Daten nun in geordneter Form vorliegen, man spricht von einer *geordneten Liste*.

Häufigkeiten werden oft in einem *Histogramm* dargestellt. Zu diesem Zweck ist es meist sinnvoll, die nach der Größe geordneten Daten in Klassen einzuteilen. In den bisher dargestellten Fällen wäre das nicht nötig, weil die Ergebnismenge ohnehin klein ist. Wenn man aber beispielsweise die Körpergrößen von 100 Studenten auf einen Zehntel-Zentimeter genau messen würde, würden nur wenige Messwerte mehrfach auftreten. Man würde sie dann in Gruppen (Klassen) einteilen, z.B. “160,1-165,0 cm”, “165,1-170,0 cm”, usw. Zudem handelt es sich bei der Körpergröße um eine *kontinuierliche Zufallsvariable*, was sich dadurch bemerkbar macht, dass die Messwerte nicht alle den gleichen Abstand zueinander haben.

Es ist zwar nicht zwingend erforderlich, dass die Klassen gleich breit sind, dass also jede Klasse einen gleich großen Messbereich abdeckt, ist jedoch in der Regel für die Übersichtlichkeit und weitere Interpretation der Daten förderlich.

Histogramme zeichnen sich dadurch aus, dass:

1. auf der y-Achse *immer* die absolute oder relative Häufigkeit aufgetragen ist.

2. die Anzahl der Messwerte, die in eine Klasse aufgenommen wurden, durch die *Fläche*, nicht etwa durch die Höhe der Säulen repräsentiert wird. Wenn alle Säulen gleich breit sind, ist das zwar wieder das selbe, trotzdem sollte man diese Eigenschaft von Histogrammen im Hinterkopf behalten.

Im Folgenden ist ein Beispiel für ein Histogramm gezeigt.

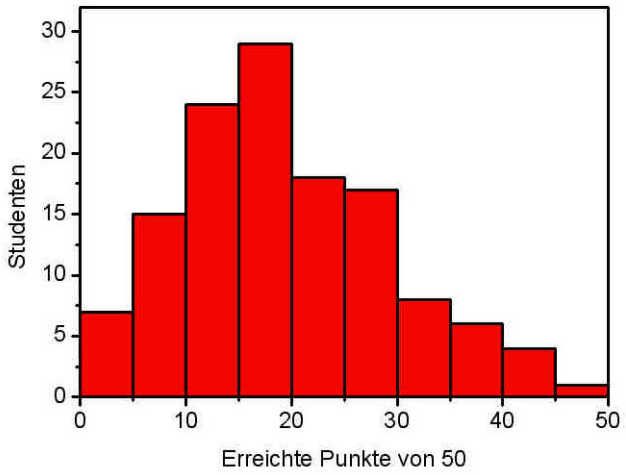


Abbildung 1: Histogramm mit aufgetragener absoluter Häufigkeit. Die Daten wurden in zehn Klassen eingeteilt. (Quelle: www.bb-sbl.de/ tutorial/ kennzahlen/ kennzahlengrafiken.html)

3.3 Wahrscheinlichkeitsrechnung

Bisher wurden Daten, die in Experimenten gemessen wurden, aufgenommen und dargestellt. Ein Versuch wird jedoch normalerweise nicht zum Selbstzweck durchgeführt, sondern ist mit einer Fragestellung verbunden. Man möchte herausfinden, welchen Gesetzen die Natur gehorcht und Regeln ableiten, die einem bei der Voraussage von Phänomenen helfen.

3.3.1 Das Urnenmodell

Als Modell und Gedankenstütze für diese Art von Experimenten hat sich das *Urnenmodell* als hilfreich erwiesen. Es ist überaus beliebt und weit verbreitet, daher sollte man es sich mit seinen verschiedenen Behältern mit vielen bunten Kugeln darin gut einprägen.

Für den Anfang stellen wir uns eine Urne vor, die 4 Kugeln enthält, alle mit unterschiedlichen Farben oder sonstwie unterscheidbar (z.B. durchnummeriert, Abbildung 2).

Wir ziehen nun hintereinander sämtliche Kugeln und fragen uns, in welcher Reihenfolge dies geschehen kann. Die Formel dafür kann man sich so überlegen:

Beim ersten Mal Ziehen sind 4 Kugeln in der Urne, also gibt es 4 Möglichkeiten, beim zweiten Mal sind nur noch 3 Kugeln da, beim dritten Mal 2, und so weiter, sodass wir zum Schluss auf diese einfache Formel kommen:

$$P = n! = 4! = 24$$

Das P steht für **Permutation**, was so viel heißt wie "das Vermischen".

Abbildung 2: Urne mit 4 unterscheidbaren Kugeln.

Ein einfaches Beispiel, auf das man diese Formel anwenden kann, ist das Szenario, wenn vier Leute sich an einen Tisch mit vier Stühlen setzen wollen - wie viele Möglichkeiten gibt es dann?

Bisher war jede Kugel nur einmal vorhanden und daher jede Anordnungsmöglichkeit gleich wahrscheinlich. Wir betrachten also im Grunde Laplace-Experimente.

Nun stellen wir uns vor, einige dieser Kugeln kommen doppelt vor, haben also die selbe Nummer und sind somit nicht unterscheidbar. In einem einfachen Fall könnte man sich vorstellen, dass es lediglich zwei Klassen gibt, die durch Kugeln in zwei verschiedenen Farben dargestellt werden (siehe Abbildung 3).

Wir ziehen wieder hintereinander alle Kugeln ohne zurückzulegen. Zuerst einmal wird uns klar, dass es weit weniger Anordnungen geben muss als wenn alle Kugeln unterscheidbar wären, denn wenn zwei Objekte derselben Klasse den Platz tauschen, zählt das als nur eine Anordnungsmöglichkeit. Die Frage ist nun, wie viele von diesen gleich erscheinenden Anordnungen es gibt. Anders formuliert: Wie viele verschiedene Anordnungen gibt es *innerhalb* jeder Klasse, die wir dann nicht unterscheiden können? Diese Frage können wir aber wieder mit der selben Formel beantworten, die wir schon kennen.

Es gibt also $8! = 40320$ Anordnungsmöglichkeiten. Betrachten wir die schwarzen Kugeln, können diese auf $3! = 6$, die roten auf $5! = 120$ Arten angeordnet sein. Die Formel sieht am Ende so aus:

$$P_{l_i} = \frac{n!}{l_1! \cdot l_2! \dots l_k!} = \frac{8!}{3! \cdot 5! \cdot 1!} = 56$$

Ein Beispiel ist die Frage, auf wie viele Arten man die Buchstaben im Wort “HONOLULU” anordnen kann, sodass ein neues “Wort” (Anagramm) entsteht. Man könnte sich dies mit einem Urnenmodell verdeutlichen, in dem wieder acht Kugeln sind. Es gibt 5 Klassen von Kugeln: je 1x “H” und “N” und je 2x “O”, “L” und “U”. Wären alle

Abbildung 3: Eine Urne, die Kugeln zwei verschiedener Farben enthält.

Buchstaben/Kugeln unterscheidbar, z.B. indem man ein U_1 und ein U_2 hätte, könnte man wieder unsere 40320 Anagramme bilden. Da aber $HONOLU_1LU_2$ genauso aussieht wie $HONOLU_2LU_1$ gibt es nur

$$P_{l_i} = \frac{8!}{2! \cdot 2! \cdot 2!} = \frac{40320}{8} = 5040$$

Möglichkeiten.

3.3.2 Baumdiagramme

Nun wissen wir zwar, *wie viele* Möglichkeiten es jeweils gibt, aber nicht, *welche*. Ab einer gewissen Größe von n ist es auch gar nicht mehr möglich, sich Anordnungsmöglichkeiten und Wahrscheinlichkeiten explizit klarzumachen und nötig ist es meist ohnehin nicht. 56 Möglichkeiten fallen bereits in eine Größenordnung, wo das weniger sinnvoll ist. Dennoch ist es in manchen Fällen nützlich.

Ein gängiges Hilfsmittel, um Anordnungsmöglichkeiten darzustellen, sind *Baumdiagramme*. Mit ihrer Hilfe kann man auch ausrechnen, wie wahrscheinlich ein bestimmter Ausgang des Experiments ist. Dann ist n meist auch sehr klein, sagen wir z.B., man zieht aus der Urne aus Abbildung 2 nur zwei Kugeln. Jedes Mal, wenn man zieht, legt man die Kugel wieder zurück, sodass sich die Wahrscheinlichkeiten nicht verändern. Bei $n = 4$ Kugeln und $k = 2$ Ziehungen gibt es genau

$$V_m = n^k$$

verschiedene Ergebnisse (siehe Abbildung 4). Das V steht für "Variation", da in diesem Fall die Reihenfolge der Kugeln wichtig ist, d.h. rot-grün ist ein anderes Ereignis als grün-rot. Manchmal liest man daher auch, dass *hintereinander* gezogen wurde und nicht gleichzeitig wie z.B. beim Lotto (später mehr dazu).

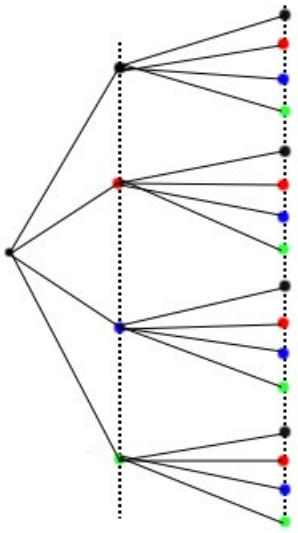


Abbildung 4: Baumdiagramm für zweimaliges Ziehen mit Zurücklegen aus einer Urne mit vier unterschiedlichen Kugeln.

Noch ein Beispiel: Das nacheinander Ziehen von 3 Kugeln aus der Urne in Abbildung 3. Ein passendes Baumdiagramm würde aussehen wie in Abbildung 5.

Will man die Wahrscheinlichkeit P für ein bestimmtes Ereignis ausrechnen, muss man den Ästen des Baumes folgen und die jeweiligen Wahrscheinlichkeiten miteinander *multiplizieren*. Man sieht, dass nicht jedes Ereignis gleich wahrscheinlich ist.

$$P_{sss} = \frac{3 \cdot 2 \cdot 1}{8 \cdot 7 \cdot 6} = \frac{6}{336} = \frac{1}{112}$$

$$P_{rrr} = \frac{5 \cdot 4 \cdot 3}{8 \cdot 7 \cdot 6} = \frac{60}{336} = \frac{5}{56}$$

Will man hingegen Wahrscheinlichkeiten für Ereignisse der Art "mindestens eine rote Kugel" ausrechnen, muss man die Wahrscheinlichkeiten für alle Ereigniss, auf die das Kriterium zutrifft, *addieren*. Mitunter ist es besser, das Gegenereignis zu berechnen und dann von 1 abzuziehen, denn **die Summe der Wahrscheinlichkeiten aller Ereignisse muss 1 sein**. Diese Regel sollte man sich übrigens gut merken!

$$P_{r \geq 1} = 1 - P_{sss} = 1 - \frac{1}{112} = \frac{111}{112}$$

3.4 Mengenlehre

Es gibt zur Darstellung von Ereignissen und ihren Wahrscheinlichkeiten eine oft verwendete und bequeme Schreibweise. Ganz oben haben wir schon den Ereignisraum kennen gelernt, der mit Ω bezeichnet wird. Ereignisse, wie sie in der oben stehenden Übung

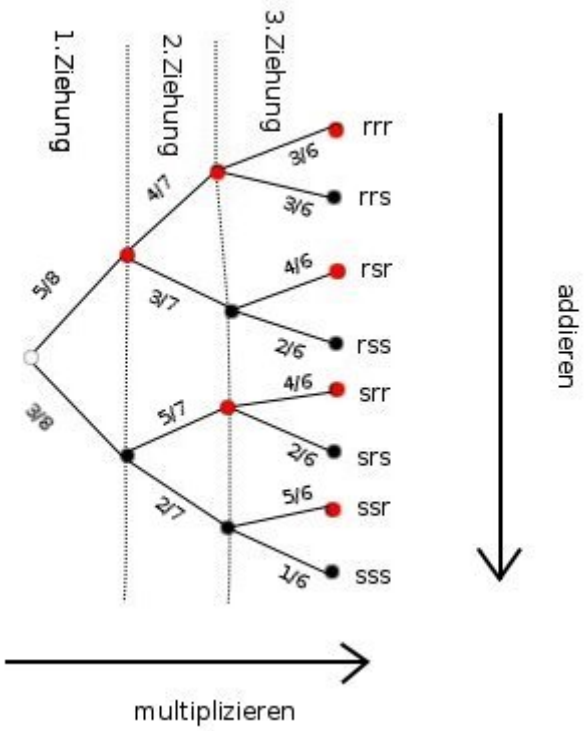


Abbildung 5: Baumdiagramm, wenn aus einer Urne mit 5 roten und 3 schwarzen Kugeln dreimal gezogen wird. Wenn es nur wenige Möglichkeiten gibt, kann man sich leicht überlegen, wie wahrscheinlich jedes einzelne Ereignis ist.

formuliert werden, sind Teil dieses Ereignisraumes. Man bezeichnet sie im Allgemeinen mit Großbuchstaben und gibt ihre Elemente in geschweiften Klammern an. Regeln aus der Mengenlehre kann man sich gut vorstellen wie in Abbildung 6 dargestellt.

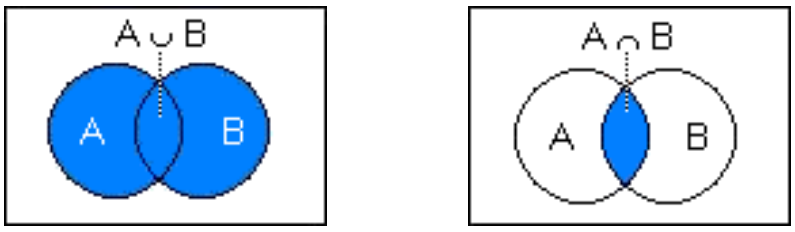


Abbildung 6: Vereinigungs- und Schnittmenge

Kommen wir noch einmal auf das Baumdiagramm in Abbildung 5 zurück und versuchen, die Abbildung darauf anzuwenden.

- A =höchstens zwei rote Kugeln=$\{rrs, rsr, rss, srr, srs, ssr, sss\}$
- B =genau eine schwarze Kugel=$\{rrs, rsr, srr\}$

Aus diesen **Mengen** lassen sich neue Mengen bilden:

- $D =$ "A oder B" $= A \cup B = \{rrs, rsr, rss, srr, srs, ssr, sss, srr\}$
- $E =$ "A und B" $= A \cap B = \{rrs, rsr, srr\}$

...oder sie lassen sich selbst aus anderen Mengen erstellen:

- $F =$genau eine rote Kugel $= \{rss, srs, ssr\}$
- $G =$keine rote Kugel $= \{sss\}$
- $H =$genau zwei rote Kugeln $= \{rrs, rsr, srr\}$
- $A = F \cap G \cap H$

Bei D handelt es sich um die *Vereinigungsmenge* von A und B und bei A um eine Vereinigungsmenge aus F, G und H (d.h. auch, dass D eine Vereinigungsmenge aus F, G, H und B ist), bei E um die *Schnittmenge* aus A und B. Man könnte auch *unmögliche* und *sichere Ereignisse* formulieren. Es ist z.B. nicht möglich, gleichzeitig genau drei rote und mindestens zwei schwarze Kugeln zu ziehen. Dies wird durch das Zeichen $\varnothing$ (leere Menge) dargestellt. Genau so gibt es auch sichere Ereignisse, wenn nämlich eine solche Menge mit Ω identisch ist.

Außerdem gibt es noch das *Gegenereignis*, dargestellt durch einen Strich über dem Buchstaben, der die Menge bezeichnet, also beispielsweise $\bar{A}$. Ist A "mindestens eine schwarze Kugel", ist $\bar{A}$ "keine schwarze Kugel", also $\bar{A} = rrr$.

Da wir nun das Werkzeug der *Mengenschreibweise* besitzen, können wir die drei Axiome für das Rechnen mit Wahrscheinlichkeiten formulieren:

1. Jedem Ereignis A aus dem Ereignisraum Ω wird eine Wahrscheinlichkeit zugeordnet, die zwischen 0 und 1 liegt.
2. Dabei hat das unmögliche Ereignis $\varnothing$ die Wahrscheinlichkeit 0 ($P_\varnothing = 0$), das sichere Ereignis die Wahrscheinlichkeit 1 ($P_\Omega = 1$).
3. Für zwei sich ausschließende Ereignisse A und B gilt: $P_{A \cup B} = P(A) + P(B)$ - und daher gilt auch: $P(A) + P(\bar{A}) = 1$.

3.4.1 Bayes-Formel

Bleiben wir noch einen Moment bei den Urnen, um eine andere Perspektive kennen zu lernen. Stellen wir uns jetzt statt nur einer Urne gleich zwei davon vor, in beiden sind wieder rote und schwarze Kugeln:

Bisher haben wir immer die Perspektive eingenommen, dass wir ziehen und uns fragen, wie wahrscheinlich es ist, eine bestimmte Farbe aus einer gegebenen Urne zu ziehen. Jetzt gehen wir von der anderen Seite an das Problem heran und stellen uns vor, dass wir bereits eine rote Kugel in der Hand halten - mit welcher Wahrscheinlichkeit stammt sie aus Urne A oder aus Urne B?

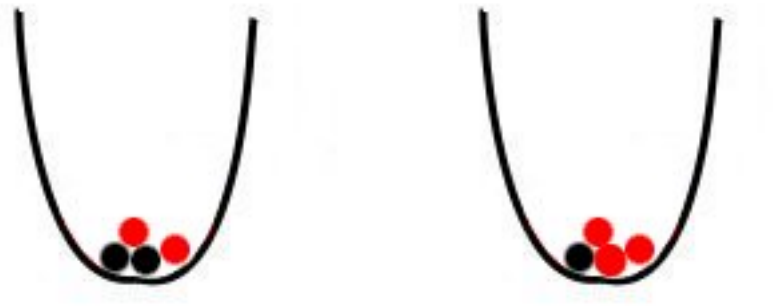


Abbildung 7: Zwei Urnen A und B, aus denen gezogen wurde.

Man nennte solche Wahrscheinlichkeiten *bedingte Wahrscheinlichkeiten*, weil man die Frage so formulieren könnte: "Unter der Voraussetzung, dass ich eine rote Kugel bereits gezogen habe, mit welcher Wahrscheinlichkeit stammt sie aus Urne A?", oder allgemeiner: "Unter der Bedingung, dass Ereignis A bereits eingetreten ist, mit welcher Wahrscheinlichkeit tritt nun Ereignis B ein?". Auch hier kann man wieder ein Baumdiagramm zu Hilfe nehmen.

Es ist schon beim Betrachten des Diagramms klar, dass die Wahrscheinlichkeit, dass eine Kugel einer bestimmten Farbe aus einer bestimmten Urne gezogen wurde, *nicht* nur von der Anzahl der Urnen abhängt. Wenn man jedoch die *Bedingung* einer bestimmten Farbe außen vor lässt, so ist das Ziehen aus jeder Urne zunächst einmal gleich wahrscheinlich. Dies bezeichnet man als *a priori-Wahrscheinlichkeit*. In diesem Fall können wir uns schon überlegen, wie wahrscheinlich es ist, dass unsere rote Kugel aus Urne A stammt: Da in Urne A zwei der insgesamt fünf roten Kugeln liegen, ist die bedingte Wahrscheinlichkeit:

$$P(Urne\,A|rot) = \frac{2}{5}$$

Da die Dinge jedoch in den seltensten Fällen so einfach und übersichtlich sind wie in diesem Beispiel, brauchen wir eine allgemeine Formel, und die sieht so aus:

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

Wenden wir die Formel auf unser Beispiel an und machen uns klar, was sie bedeutet. Zunächst lässt sich feststellen:

- A=Kugel wurde aus Urne A gezogen
- B=Kugel ist rot

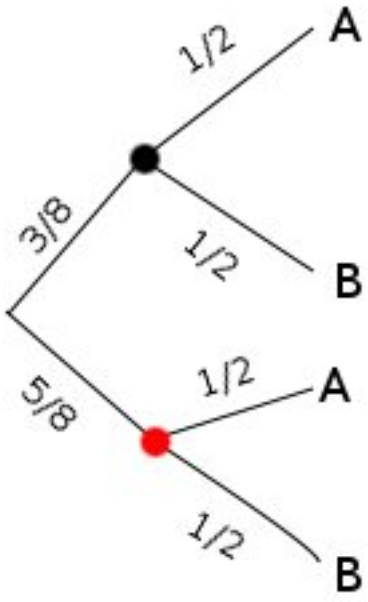


Abbildung 8: Die Wahrscheinlichkeit, eine bestimmte Farbe zu ziehen, ist abhängig von der Gesamtzahl der Kugeln in beiden Urnen. Da es zwei Urnen gibt, ist die *a priori*-Wahrscheinlichkeit jeweils 0,5.

Nun brauchen wir also die Wahrscheinlichkeit, dass eine rote Kugel gezogen wird unter der Voraussetzung, dass aus Urne A gezogen wird ($P(B|A)$). Da in Urne A insgesamt 4 Kugeln sind, von denen zwei rot sind, ist diese Wahrscheinlichkeit 0,5.

Die mit $P(A)$ bezeichnete Größe ist die bereits erwähnte a priori - Wahrscheinlichkeit von Urne A, also ebenfalls 0,5.

$P(B)$ nun meint die Wahrscheinlichkeit, eine rote Kugel zu ziehen insgesamt, also egal, ob aus Urne A oder Urne B. Da es insgesamt acht Kugeln gibt, von denen fünf rot sind, ist diese Wahrscheinlichkeit also $\frac{5}{8}$.

Setzen wir all das zusammen, erhalten wir:

$$P(Urne\,A|rot) = \frac{0,5 \cdot 0,5}{\frac{5}{8}} = \frac{2}{5}$$

Das entspricht dem Ergebnis, das wir schon vorher hatten.

Übung Ein medizinischer Test auf eine Krankheit, die bei 0,02% der Bevölkerung auftritt, besitzt eine Sensitivität von 98% (d.h. 98 von 100 Kranken werden als solche erkannt). Zudem schlägt er bei einem von hundert Getesteten an, obwohl die Krankheit gar nicht vorliegt (falschpositives Ergebnis). Wenn man nun ein positives Testergebnis hat, wie wahrscheinlich ist es, dass man tatsächlich krank ist? Überlegen Sie vorher, was Sie vermuten würden! Sie können auch hier ein Baumdiagramm zu Hilfe nehmen.

3.4.2 Statistische Unabhängigkeit

An dem Beispiel mit der Bayes-Formel gibt es einen wichtigen Unterschied zu den anderen, davor behandelten Fällen. Wenn eine Münze geworfen wird, ist es für den Ausgang des zehnten Wurfes völlig unerheblich, wie oft ich davor Kopf oder Zahl geworfen habe, die Wahrscheinlichkeit für jedes Ergebnis bleibt immer 0,5. Bei bedingten Wahrscheinlichkeiten ist das anders. Ob aus Urne A oder B gezogen wurde, hat sehr wohl einen Einfluss darauf, mit welcher Wahrscheinlichkeit die Kugel rot oder schwarz ist. Diese beiden Ereignisse sind nicht *statistisch unabhängig*. Mathematisch formuliert bedeutet statistische Unabhängigkeit:

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

oder

$$P(A|B) = P(A)$$

Beides bedeutet das selbe, nämlich dass die Wahrscheinlichkeit für das Ereignis A nicht vom Ereignis B abhängt und man daher die Wahrscheinlichkeit für das gemeinsame Auftreten der Ereignisse A und B einfach als Produkt der beiden Wahrscheinlichkeiten berechnen kann (entlang der Äste eines Baumdiagramms).

Beispiel Beim zweimaligen Würfeln gibt es 36 mögliche Ausgänge, wenn man die Reihenfolge beachtet (6^2). Um die Wahrscheinlichkeit für Ereignis $A = \{6, 6\}$ zu berechnen, reicht es, $P(A) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$ auszurechnen. Die Wahrscheinlichkeit, dass, wenn man eine 6 gewürfelt hat (B: beim ersten Wurf eine 6), noch einmal eine sechs gewürfelt wird (A), also $P(A|B)$ ist unverändert $\frac{1}{6} = P(A)$.

3.5 Übungsaufgaben

1. In einer Schachtel befinden sich in einer Reihe insgesamt 4 schlechte und 4 gute Äpfel. Wie viele mögliche Anordnungen gibt es, wenn jeweils die Schlechten und die Guten nicht unterscheidbar sind?

2. In einer Urne liegen 10 Kugel mit den Nummern 1 bis 10. Es werden mit einem Griff 6 Kugeln gezogen (Reihenfolge zählt daher nicht). Wie viele Möglichkeiten gibt es?

3. In einer Urne liegen 10 Kugeln mit den Nummern 1 bis 10. Man zieht eine Kugel zufällig, notiert ihre Nummer und legt sie dann wieder zurück. Wie viele verschiedene Zahlenfolgen erhält man, wenn man 6-mal zieht?

4. Bei einem Fahrradschloss können auf drei Ringen jeweils die Ziffern 1 bis 6 eingestellt werden.

   a) Wie viele verschiedene Möglichkeiten hat man, eine Zahlenkombination einzustellen?

b) Ein Dieb weiß, dass der Fahrradbesitzer eine Vorliebe für gerade Zahlen hat. Er möchte alle Zahlenkombinationen probieren, die an der ersten und an der letzten Stelle eine gerade Ziffer haben. Wie viele derartige Kombinationen gibt es?

5. 6 Biologiebücher, 4 Chemiebücher und 3 Physikbücher sollen in einem Regal so angeordnet werden, dass die Stoffgebiete zusammen bleiben. Auf wie viele Arten können diese angeordnet werden, wenn
   a) alle Bücher verschieden sind,
   b) die Biologie- und Chemiebücher gleich, die Physikbücher aber verschieden sind?

6. An einem Fußballturnier nehmen 8 Mannschaften teil. Wie viele Spiele müssen ausgetragen werden, wenn jede Mannschaft gegen jede spielt (ohne Rückspiel!)?

7. Von den Studenten einer Universität nutzen 35 % den Bus und 25 % die Straßenbahn für die Fahrt in die Uni. 15 % der Studenten kombinieren die beiden Verkerhrsmittel. Mit welcher Wahrscheinlichkeit nutzt ein zufällig ausgewählter Student
   a) mindestens eines der beiden Verkehrsmittel,
   b) weder Bus noch Straßenbahn,
   c) nur Bus,
   d) genau eines der beiden Verkehrsmittel?

8. In einer Urne befinden sich 2 rote, 3 blaue und 4 grüne Kugeln. Aus der Urne werden gleichzeitig 3 zufällige Kugeln entnommen. Mit welcher Wahrscheinlichkeit sind die entnommenen Kugeln
   a) alle von unterschiedlicher Farbe,
   b) alle grün,
   c) 2 grün und 1 blau?

9. An einem Volleyballturnier nehmen 12 Mannschaften teil, die zufällig in zwei gleich große Gruppen aufgeteilt werden. Wie groß ist die Wahrscheinlichkeit, dass zwei Mannschaften A und B
   a) zu der gleichen Gruppe gehören,
   b) zu unterschiedlichen Gruppen gehören?

10. In Urne A befinden sich 3 rote und 7 blaue Kugeln, in Urne B 5 rote und 5 blaue Kugeln. Es wird zufällig eine der Urnen A oder B ausgewählt und aus dieser werden zufällig 2 Kugeln entnommen. Wie groß ist die Wahrscheinlichkeit dafür, dass die beiden Kugeln blau sind?

4 Die wichtigsten Größen der empirischen Statistik

4.1 Grundgesamtheit und Stichprobe

In den meisten Situationen ist es unmöglich, die gesamte Population zu vermessen, so dass man sich mit einer Stichprobe zufrieden geben muss. Die Population wird häufig auch als *Grundgesamtheit* bezeichnet und Ziel des Experimentes sollte es sein, eine möglichst repräsentative Stichprobe zu verwenden. Dieses Problem mündet in einem intelligenten Versuchsdesign.

Aber selbst bei der vermutlich repräsentativsten Stichprobe muss davon ausgegangen werden, dass die statistischen Größen dieser Stichprobe nicht denen der Population entsprechen. Daher gibt es auch in der Regel unterschiedliche Symbole für auf den ersten Blick gleiche - und gleich bezeichnete - Größen.

Diese werden zwar nicht immer verwendet (z.B. wird die Wahrscheinlichkeit fast immer mit p oder P bezeichnet, obwohl es genau genommen einen Unterschied zwischen dem wahren und dem empirischen Wert $\hat{p}$ gibt), man sollte sich jedoch immer klar machen, was man eigentlich gerade ausrechnet.

4.2 Lagemaße

Um einen Überblick zu erhalten, in welcher Größenordnung sich die Messwerte bewegen, gibt es die *Lagemaße*.

4.2.1 Mittelwert

Aus den oben genannten Gründen ist der Mittelwert der Stichprobe nur eine Schätzung des Mittelwertes der Grundgesamtheit, welcher auch Erwartungswert genannt wird. Um dies deutlich zu machen, gibt es zwei verschiedene Symbole: Für den Erwartungswert wird μ geschrieben, die Symbolik $\bar{x}$ für den Mittelwert der Stichprobe verwendet.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Das Zeichen $\sum$ ist ein großes griechisches Sigma und bedeutet nichts anderes als Summe. Die Gleichung bedeutet, dass man alle Messwerte addieren und anschließend durch die Anzahl der Messwerte teilen muss. Daher kann der Mittelwert niemals größer als der größte Messwert und auch nicht kleiner als der kleinste Messwert sein. Man sollte allgemein darauf achten, Messwerte nur so genau anzugeben, wie man sie auch tatsächlich mit den zur Verfügung stehenden Geräten messen kann. Demzufolge darf man auch beim Mittelwert nicht mehr Kommastellen angeben als bei den Messwerten selbst. Dadurch entsteht eine vermeintliche Genauigkeit, die der Realität nicht entspricht und nicht exakt ist.

Neben dem hier angegebenen *arithmetischen Mittelwert* gibt es auch noch das *geometrische Mittel*. Dabei werden alle Werte multipliziert und dann die n-te Wurzel gezogen. Zu diesem Zweck verwendet man das Produktzeichen, was ein großes Pi ist:

$$\bar{x}_{geom} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Wann man welche Größe verwendet, kann nicht pauschal gesagt werden und hängt von der Fragestellung ab.

4.2.2 Median

Oben haben wir schon geordnete Listen kennen gelernt. Der Mittelwert hat den Nachteil, dass so genannte Ausreißer, also Messwerte, die ungewöhnlich groß oder klein sind, vergleichsweise stark ins Gewicht fallen. Eine Aussage darüber, in welchem Bereich sich die meisten Messwerte bewegen, kann auch der *Median*, symbolisiert durch $\tilde{x}$, geben. 50% aller Messwerte sind größer, 50% kleiner als er. Hat man eine ungerade Anzahl i an Messwerten, nimmt man in der geordneten Liste einfach den Wert, der in der Mitte steht. Hat man eine gerade Anzahl, nimmt man den Mittelwert aus den beiden mittleren Werten.

Anschaulich machen kann man sich die Bedeutung des Medians mit Hilfe der *kumulierten Wahrscheinlichkeit* (Tabelle 4). Man kann diese graphisch darstellen, indem man aus einer geordneten Liste für alle x_i die Häufigkeiten bis zu dieser Größe addiert. Daraus lässt sich dann ein Summenpolygon erstellen (siehe Abbildung 9).

Intervall	Anzahl H	Summe	Summe relativ
<160 cm	12	12	0,15
160-169 cm	15	27	0,33
170-179 cm	25	52	0,63
180-189 cm	21	73	0,88
>190 cm	10	83	1,00

Tabelle 4: Beispieldaten: Messung der Körpergrößen von 83 StudentInnen.

Neben dem Median kann man noch weitere Größen auf ähnliche Weise definieren, z.B. gibt es einen Wert, unter dem 25%, über dem 75% aller Daten liegen. Dieser Wert wird "unteres Quartil" genannt, weil die Daten auf diese Weise in vier gleich große Bereiche geteilt werden. Insofern ist auch der Median ein Quartil, nämlich das mittlere. Auf gleiche Weise kann man die Daten auch in 10 (Dezile), 20 oder 100 (Prozentile) Bereiche teilen, wobei es dann immer n-1 Quantilen gibt.

4.2.3 Modus

Beim Modus, auch Modalwert genannt, handelt es sich lediglich um den am häufigsten auftretenden Messwert. Messreihen bzw. Wahrscheinlichkeitsverteilungen können beliebig viele Modi besitzen.

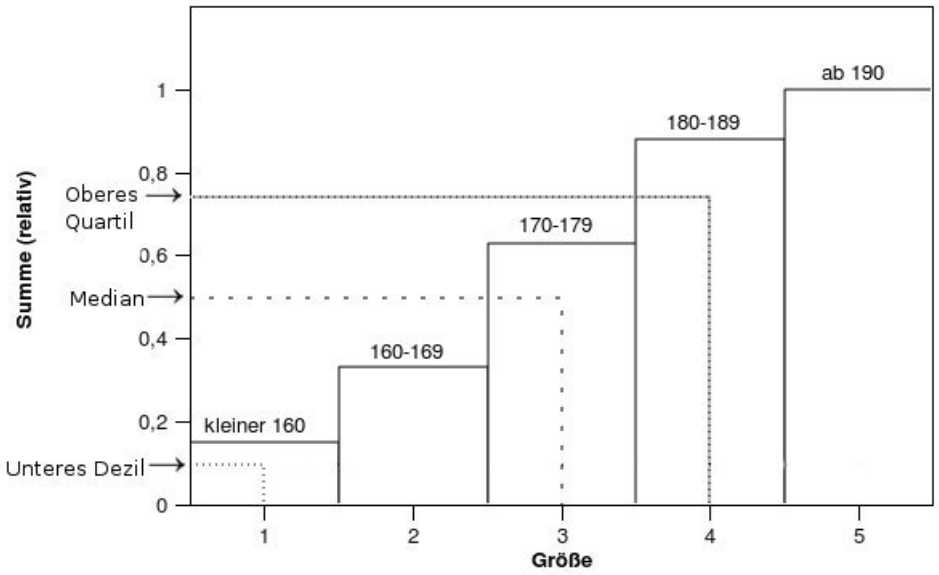


Abbildung 9: Beispiel für ein Summenpolygon mit eingezeichneten Quantilen.

4.3 Streuungsmaße

Hierbei geht es eher um die Frage, um wieviel die Messwerte im Allgemeinen von den errechneten Lagemaßen abweichen.

4.3.1 Varianz und Standardabweichung

Auch diese Werte sind nur Schätzwerte für die tatsächlichen Größen der Grundgesamtheit, die mit σ^2 (Standardabweichung) bzw. σ (Varianz) bezeichnet werden. Die Varianz einer Stichprobe ist gegeben durch:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Die Standardabweichung berechnet sich als

$$s = \sqrt{(s^2)}$$

Zu beachten ist, dass es beim Ziehen einer Wurzel immer zwei Ergebnisse gibt, da $x^2 = (-x)^2$. Daher schreibt man das Ergebnis immer in der Form $\pm x$.

Manchmal wird die Formal auch anders angegeben. Dann steht dort statt n $n-1$. Das hängt damit zusammen, dass man die Varianz normalerweise unterschätzt, je kleiner n ist, desto mehr. Es wird jedoch in der Regel beides als korrekt bewertet [1].

[1] siehe Anhang für mehr Erklärungen

4.3.2 Spannweite

Ein sehr viel einfacheres Lagemaß ist die Spannweite, die sich einfach als Differenz aus dem größten und dem kleinsten Messwert berechnet.

4.4 Übungsaufgaben

1. In einem Krankenhaus wurde das Gewicht von 10 Neugeborenen gemessen und ergab folgende Werte:

i	1	2	3	4	5	6	7	8	9	10
$x_i[g]$	2500	2900	2100	1850	2800	2700	2200	3150	2250	2550

   Berechnen Sie

   a) das mittlere Gewicht der Neugeborenen!
   b) die Standardabweichung!

2. In einem Würfelversuch wurde 20 Mal gewürfelt und folgende Augen notiert:

$$x_i = \{5, 6, 2, 4, 5, 2, 3, 6, 5, 4, 2, 3, 1, 5, 1, 2, 5, 4, 1, 6\}$$

   Stellen Sie die

   a) absoluten Häufigkeiten
   b) relativen Häufigkeiten

   in einem Histogramm dar!

3. Der Mittelwert von fünf Noten eines Schülers beträgt 3,2. Wie viele Einser müsste er noch bekommen, damit sein Mittelwert besser als 2,5 wird?

4. Die durchschnittliche Größe der Spieler einer Basketballmannschaft betrug 183 cm. Nachdem ein neuer Spieler mit der Größe von 199 cm in die Mannschaft aufgenommen wurde, vergrößerte sich die durchschnittliche Größe um 2 cm. Wie viele Basketballspieler gehören jetzt zu der Mannschaft?

5. Der Mittelwert von 7 verschiedenen natürlichen Zahlen beträgt 9, ihr Median beträgt 10. Welcher ist der maximale Wert, den die größte Zahl annehmen kann?

5 Wahrscheinlichkeitsverteilungen

Wenn man in einem Zufallsversuch eine komplette Datenreihe aufgenommen hat und in einem Histogramm darstellt, erhält man oft charakteristische Muster in der Verteilung der Häufigkeiten. Es gibt *Wahrscheinlichkeitsverteilungen*, die sehr häufig auftreten und auf eine gemeinsame Struktur der Versuche, auf die sie zutreffen, hinweisen. Im Folgenden werden die wichtigsten Verteilungen, die auch in der Vorlesung eine Rolle spielen werden, kurz vorgestellt und die Art von Experimenten, für die sie verwendet werden, beschrieben.

5.1 Diskrete Wahrscheinlichkeitsverteilungen

"Diskret" bedeutet in diesem Sinne, dass unterscheidbare, voneinander abgegrenzte Messwerte auftreten, zwischen denen nichts exisitiert, wie z.B. bei Farben von Kugeln, Kopf oder Zahl beim Münzwurf oder den Augen beim Würfeln. Daraus folgt, dass eine diskrete Zufallsvariable nur endlich viele bzw. abzählbar unendlich viele Werte annehmen kann. Anschaulich gesprochen, können die möglichen Werte durchnummeriert werden (auch wenn daraus nicht unbedingt eine Reihenfolge resultiert; Farben könnte man in kein zahlenmäßiges Verhältnis zueinander setzen).

5.1.1 Binomialverteilung

Die einfachste Verteilung ist die, in der nur zwei Werte möglich sind, z.B. rot und schwarz, 1 und 0, oder, allgemein gesprochen, Erfolg oder Misserfolg. Die Ergebnisse eines solchen Versuches sind *dichotom*, d.h. es gibt nur zwei von ihnen und sie schließen sich gegenseitig aus. Dieser Versuch wird nun mehrfach hintereinander durchgeführt, wobei sich die Wahrscheinlichkeiten für Erfolg (p) und Misserfolg ($1 - p = q$) von Versuch zu Versuch nicht verändern und die Einzelversuche voneinander unabhängig sind - wie beim Ziehen aus einer Urne mit Zurücklegen.

Wenn man sich z.B. vorstellt, dass man sechs mal würfelt und es als Erfolg gilt, wenn man eine 6 würfelt, kann man sich fragen, wie wahrscheinlich es ist, zwei mal einen Erfolg zu haben, vier mal einen Misserfolg. p ist jetzt $\frac{1}{6}$, q dagegen $\frac{5}{6}$. Bei sechs Würfen rechnet man also:

$$\left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^4 = \frac{625}{46656}$$

Nun muss man aber noch bedenken, dass es mehrere Möglichkeiten gibt, zwei 6en zu werfen, z.B. ganz am Anfang oder ganz am Ende. Dies kann man sich wieder als Urnenexperiment vorstellen, wo man aus n k Kugeln zieht ohne die jeweils gezogene Kugel zurück zu legen, wobei die Reihenfolge keine Rolle spielt ("Kombination", im Gegensatz zur Variation). Man kann sich auch vorstellen, dass man alle k Kugeln gleichzeit herauszieht. Zu diesem Zweck gibt es den *Binomialkoeffizienten* [2]:

$$\binom{6}{2} = 15$$

So ergibt sich die Formel und für unser Beispiel rechnen wir

$$P(k = 2) = 15 \cdot \frac{1}{6}^2 \cdot \frac{5}{6}^4 = 0,2$$

Wenn man alle P für jedes k von 0 bis n ausrechnet und addiert, ergibt sich 1. Trägt man diese Werte in einem Histogramm auf, erhält man ein charakteristisches Bild (siehe Abbildung 10).

An den unterschiedlichen Gestalten der Verteilungen kann man Folgendes erkennen:

[2] Mehr dazu im Anhang

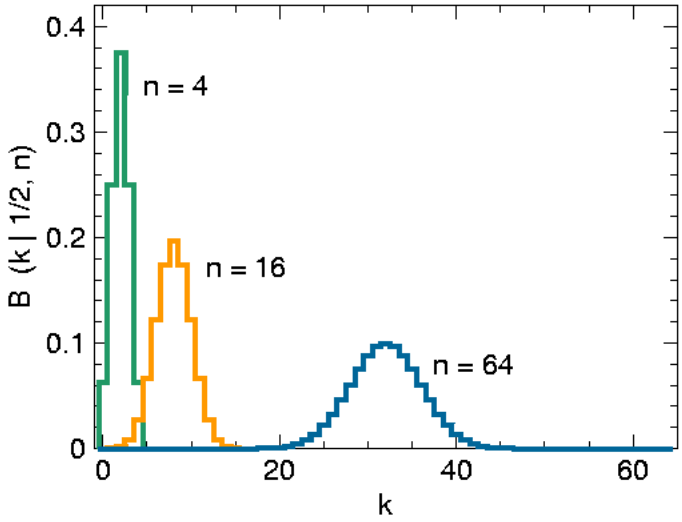


Abbildung 10: Mehrere Binomialverteilungen für unterschiedliche Versuchsanzahlen (n) mit der Erfolgswahrscheinlichkeit 0,5 (wie beim Münzwurf). Die Notation an der y-Achse liest sich so: B(k=Anzahl der Erfolge, auf x aufgetragen|p=1/2,Versuchsanzahl n). Quelle: http://de.wikipedia.org/wiki/Binomialverteilung

- Je mehr Versuche gemacht werden (n wird größer), desto weiter rückt die Verteilung nach rechts, weil durchschnittlich auch mehr Erfolge eintreten.
- Die Verteilung wird flacher, was bedeutet, dass die Werte mehr streuen (die Varianz wächst). Das Experiment hat mehr mögliche Ausgänge.

Eine weitere Eigenschaft der Binomialverteilung ist, dass sie für $p < 0,5$ linksschief ist, weil die Wahrscheinlichkeit für wenige Erfolge größer ist als für viele. Umgekehrtes gilt für $p > 0,5$ (siehe Abbildung 11). Für große n kann man beobachten, dass die Verteilung symmetrischer wird.

5.2 Kontinuierliche Wahrscheinlichkeitsverteilungen

Anders als bei diskreten Wahrscheinlichkeiten, kann man hier nicht sagen, wie viele mögliche Ergebnisse ein Experiment hat und man kann diese auch nicht durchnummerieren - es gibt unendlich viele mögliche Ergebnisse. Aus diesem Grund haben Einzelergebnisse auch immer die Wahrscheinlichkeit 0 und man kann nur die Wahrscheinlichkeit ausrechnen, dass ein Ergebnis in einem bestimmten Intervall liegt. Die Darstellung einer solchen Verteilung wird auch *Dichtefunktion* genannt. Wird bei Darstellungen von Dichtefunktionen auf der y-Achse das Intervall von 0 bis 1 aufgetragen, spricht man von einer *normierten Dichtefunktion*: Die Fläche unter einer normierten Dichtefunktion ist immer 1.

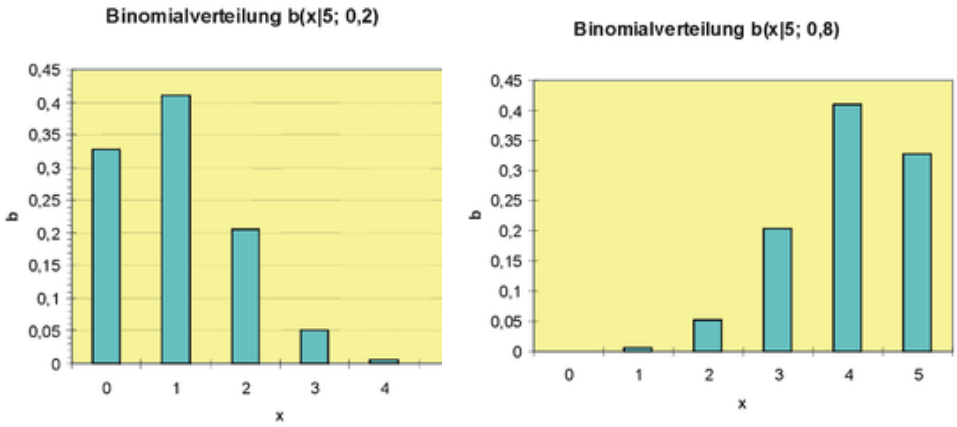


Abbildung 11: Zwei unsymmetrische Binomialverteilungen für sehr kleine n. Quelle: de.wikibooks.org

5.2.1 Gleichverteilung

Die einfachste unter den kontinuierlichen Verteilungen ist die Gleichverteilung, d.h. egal, welches Intervall man betrachtet, das Ereignis tritt immer mit gleicher Wahrscheinlichkeit darin auf (natürlich nur, solange die Intervalle gleich groß sind).

5.2.2 Normalverteilung

Dies ist die wahrscheinlich berühmteste Dichtefunktion überhaupt und auch unter dem namen Glockenkurve oder Gauß-Glocke bekannt. Grundsätzlich hat eine normalverteilte Größe folgende Eigenschaften:

- Die meisten Werte treten um den Mittelwert auf.
- Je größer die Abweichung vom Mittelwert ist, desto weniger Werte treten auf. (Kleine Abweichungen sind wahrscheinlicher als große.)
- Die Dichtefunktion ist symmetrisch zu der Achse, die durch den Mittelwert parallel zur y-Achse gelegt wird.

Durch diese Eigenschaften eignet sich die Normalverteilung für die meisten Messvorgänge, weil hier *zufällige Fehler* auftreten. Diese Fehler entstehen durch den Einfluss sehr vieler, voneinander unabhängiger Einflüsse, wodurch kleine Fehler viel wahrscheinlicher sind als große und man erwarten darf, in einem kleinen Intervall um den wahren Wert herum die meisten Messergebnisse zu erhalten. Zudem weist die Normalverteilung die sehr nützliche und interessante Eigenschaft auf, dass in dem Intervall $[\mu - \sigma, \mu + \sigma]$ genau 68,27% aller Werte liegen[3].

[3]Höchstens eine Abweichung von 2σ haben 95,45%, von 3σ 99,73% aller Daten.

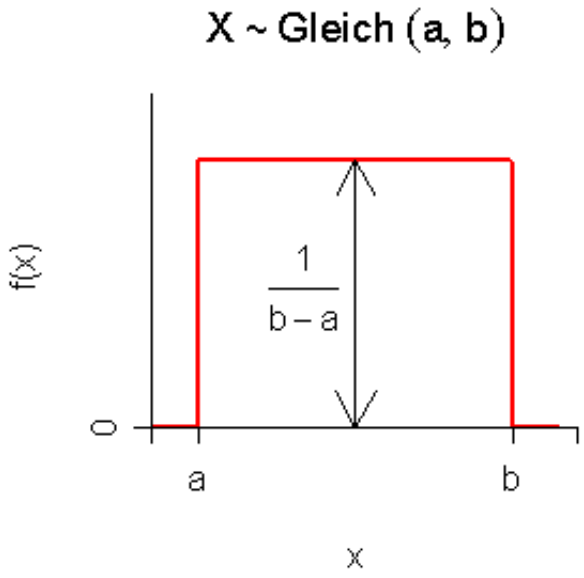


Abbildung 12: Schematische Darstellung einer stetigen Gleichverteilung. Die Fläche, die von der roten Linie eingeschlossen ist, muss 1 betragen, daher kann man die Höhe des Rechtecks auf in der Skizze bezeichnete Weise errechnen. Quelle: www.wikipedia.org

5.3 Aufgaben

1. Wie groß ist die Wahrscheinlichkeit dafür, dass beim gleichzeitigen Werfen von 6 Würfeln
   a) die Würfel unterschiedliche Augenzahlen zeigen,
   b) jeder Würfel die Augenzahl 6 zeigt,
   c) die Augenzahl 6 genau vier mal auftritt,
   d) jeder Würfel eine gerade Augenzahl zeigt,
   e) die gleiche Augenzahl auf genau drei Würfeln auftritt?

2. Beim Messen einer physikalischen Größe wird der erlaubte Messfehler mit einer Wahrscheinlichkeit von 0,4 überschritten. Es werden drei unabhängige Messungen durchgeführt. Wie groß ist die Wahrscheinlichkeit dafür, dass in zwei Messungen der erlaubte Messfehler überschritten wird?

6 Ausblick auf die Vorlesung

In der Vorlesung wird noch näher darauf eingegangen werden, wie man Daten darstellen kann und welche Eigenschaften sie haben können, z.B. unterschiedliche Skalentypen (Daten, die nach Farben geordnet sind vs. Daten, die tatsächlich numerisch sind und nach der Größe zu ordnen sind) sowie weitere Lage- und Streuungsmaße.

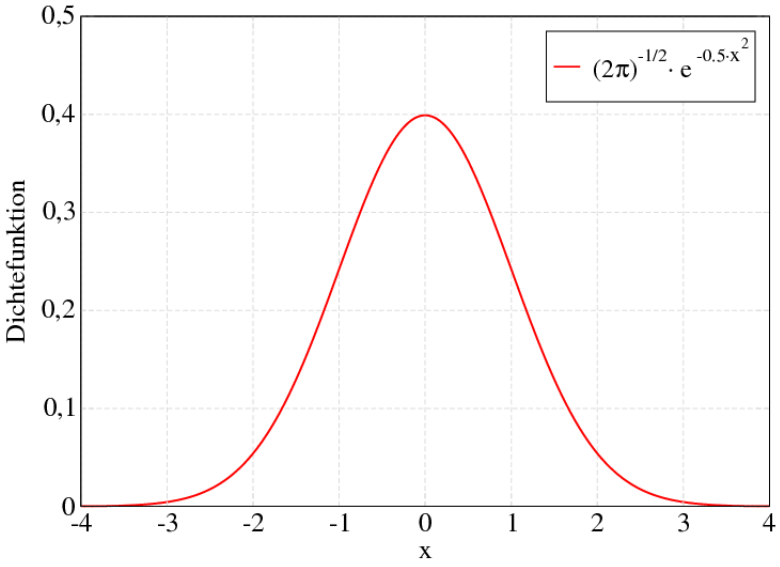


Abbildung 13: Normalverteilung mit dem Mittelwert 0 und der Standardabweichung 1.

Die nur kurz vorgestellten Verteilungen werden sehr viel tiefer behandelt; weitere kommen hinzu, die dann vor allem für die schließende Statistik von Bedeutung sind. Diese schließende Statistik ist das, was man letztendlich für die Auswertung von Daten am dringendsten braucht, wenn es nämlich darum geht, Hypothesen zu stützen oder zu verwerfen:

Ist der gemessene Unterschied tatsächlich so groß, wie er aussieht?

Lässt sich ein Zusammenhang zwischen zwei Größen bestätigen?

Welchen Gesetzmäßigkeiten folgen meine Daten, und kann ich daraus ein Modell ableiten, um diesen Vorgang zu beschreiben?

Die Biostatistik-Vorlesung kann in dieser Hinsicht nur einen Grundstein legen, im Laufe des Grundstudiums wird es aber für jede/n Gelegenheit geben, das Gelernte anzuwenden und zu erproben. Es ist daher nur zu empfehlen, sich von Anfang an mit den vorgestellten Konzepten vertraut zu machen und auch statistische Computerprogramme zu nutzen, um Berechnungen, so einfach sie auch erscheinen mögen, durchzuführen, um sich an größere und unübersichtlichere Datenmengen heranzutasten.

Viel Erfolg!

7 Lösungen zu den Aufgaben

7.1 Übungsaufgaben zum Kapitel Zufallsversuche

Bayes-Formel

A: Krankheitsstatus

B: Testergebnis
bekannt:

- $P(A^+) = 0,0002$
- $P(B^+|A^+) = 0,98$ (das ist die Sensitivität)
- $P(B^+|A^-) = 0,01$ (das sind die Falschpositiven)

gesucht:

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

Zwei der drei benötigten Größen stehen schon da und müssen nur eingesetzt werden ($P(A^+) = 0,0002$ und $P(B^+|A^+)$). Der dritte Wert ($P(B^+)$) errechnet sich, indem man alle Möglichkeiten, ein positives Testergebnis zu erhalten, in Betracht zieht und die Wahrscheinlichkeiten addiert:

$$P(B^+) = P(A^- \cap B^+) + P(A^+ \cap B^+) = 0,9998 \cdot 0,01 + 0,0002 \cdot 0,98 \approx 0,01$$

Nun können wir einsetzen:

$$P(A^+|B^+) = \frac{0,98 \cdot 0,0002}{0,01} \approx 0,02$$

Da die Krankheit also so selten ist, ist man im Falle eines positiven Tests mit nur 2%iger Wahrscheinlichkeit tatsächlich erkrankt!

1. $$n = 8,\ l_1(gut) = l_2(schlecht) = 4$$

   $$P_{l_i} = \frac{8!}{4! \cdot 4!} = 70$$

2. Reihenfolge egal: Kombination, ohne Zurücklegen; $n = 10$, $k = 6$

   $$K_o = \binom{n}{k} = \binom{10}{6} = 210$$

3. Reihenfolge wichtig: Variation, mit Zurücklegen; $n = 10$, $k = 6$

   $$V_m = n^k = 10^6$$

4. ist wie 3 mal Ziehen mit Zurücklegen (Zahlen können sich wiederholen), Reihenfolge wichtig (Variation)

a) $n = 6$, $k = 3$

$$V_m = n^k = 6^3 = 216$$

b) 1. und 3. Ring: 3 Möglichkeiten (2,4,6), 2. Ring: 6 Möglichkeiten

$$3 \cdot 6 \cdot 3 = 54$$

5. Permuatationen

a) Anordnungsmöglichkeiten innerhalb der Stoffgebiete: $6! = 720$, $4! = 24$, $3! = 6$, Anordnungsmöglichkeiten der Stoffgebiete untereinander: $3! = 6$

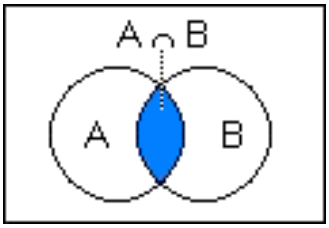
$$P = 6! \cdot 4! \cdot 3! \cdot 3! = 622080$$

b) $l_1 = 6$, $l_2 = 4$

$$P_{l_i} = \frac{622080}{6! \cdot 4!} = 36$$

6. Jede Mannschaft ($n = 8$) muss gegen 7 ($n - 1$) andere Mannschaften spielen. Um nicht jedes Spiel doppelt zu zählen, teilt man durch 2:

$$\frac{n \cdot (n-1)}{2} = \frac{56}{2} = 28$$

7. 

A: benutzen den Bus (35%)

B: benutzen die Straßenbahn (25%)

$A \cap B$: benutzen beides (15%)

a) "Flächeninhalt" von $A \cup B = A + B - A \cap B = 35\% + 25\% - 15\% = 45\%$
b) $100\% - A \cap B = 100\% - 45\% = 55\%$
c) $A - A \cap B = 35\% - 15\% = 20\%$
d) $A \cup B - A \cap B = 45\% - 15\% = 30\%$

8. a) Da man aus 9 Kugeln 3 zieht, erhält man die Gesamtzahl der Möglichkeiten über $\binom{9}{3} = 84$. Nun braucht man noch die Anzahl der Möglichkeiten, drei unterschiedliche Kugeln zu ziehen. Bei der Kombination rgb gibt es 2 Möglichkeiten für rot, 3 für blau und 4 für grün, also $2 \cdot 3 \cdot 4 = 24$. Die

Wahrscheinlichkeit, drei unterschiedliche Kugeln zu ziehen, erhält man nun über

$$p(1xr, 1xg, 1xb) = \frac{24}{84} = \frac{2}{7}$$

b) Möglichkeiten, 3 von 4 grünen Kugeln zu ziehen: $\binom{4}{3} = 4$

$$p(ggg) = \frac{4}{84} = \frac{1}{21}$$

c) Möglichkeiten, von 4 grünen 2 zu ziehen: $\binom{4}{2} = 6$

$$p(2xg, 1xb) = \frac{6 \cdot 3}{84} = \frac{3}{14}$$

9. a) Man kann sich das vorstellen, als ob aus einer Urne gezogen wird, in der 6 rote und 6 schwarze Kugeln sind. Zieht man z.B. beim ersten mal rot, würde das heißen, Mannschaft A wurde Gruppe 1 zugeordnet. Nun sind in der Urne nur noch 5 von 11 Kugeln rot, also beträgt die Wahrscheinlichkeit, noch einmal rot zu ziehen und somit Mannschaft B ebenfalls Gruppe 1 zuzuordnen, **5/11**.

   b) Dies ist das Gegenereignis, also **p=6/11**.

10. Die Kugeln können also aus Urne A oder aus Urne B stammen, man muss also die Wahrscheinlichkeiten für beide Möglichkeiten addieren.

$$p(bb|A) = 0,5 \cdot \frac{7}{10} \cdot \frac{6}{9} = \frac{21}{90}$$

$$p(bb|B) = 0,5 \cdot \frac{5}{10} \cdot \frac{4}{9} = \frac{10}{90}$$

$$p(bb) = \frac{21 + 10}{90} = \frac{31}{90}$$

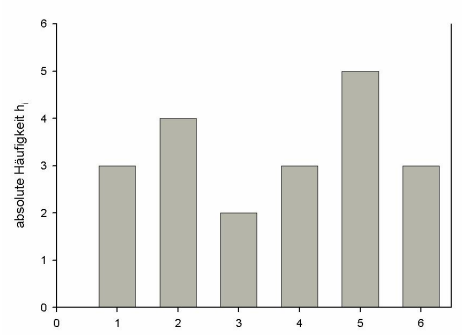
7.2 Übungsaufgaben zum Kapitel empirische Statistik

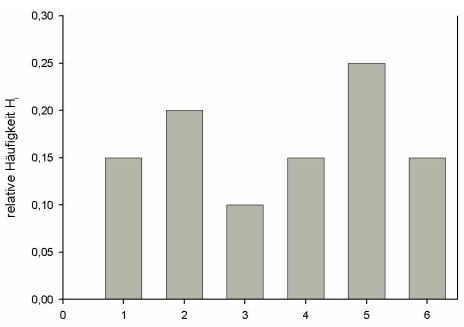
1. a)

$$\bar{x} = \frac{(2500 + 2900 + 2100 + 1850 + 2800 + 2700 + 2200 + 3150 + 2250 + 2550)g}{10} = 2500g$$

b)

$$s = \sqrt{\frac{(650^2 + 400^2 + 300^2 + 250^2 + 50^2 + 200^2 + 300^2 + 400^2 + 650^2)g^2}{10}} \approx \pm 401g$$

2. 



3. Bekannt (x ist die Summe aus allen 5 Noten des Schülers):

$$\frac{\sum x}{5} = 3,2 \rightarrow x = 16$$

Gesucht (b ist die Anzahl der noch benötigten Einser):

$$\frac{16+b}{5+b} \leq 2,5 \rightarrow b \geq 2,5$$

Der Schüler braucht also noch drei Einser, dann hat er einen Durchschnitt unter 2,5.

4. Bevor der neue Spieler kommt, sieht es so aus (x ist die Anzahl der Spieler):

$$\frac{183cm \cdot x}{x} = 183cm$$

Der neue Spieler verändert die Daten so:

$$\frac{183cm \cdot x + 199cm}{x+1} = 185cm \rightarrow x = 7$$

Das Team besteht also jetzt aus 8 Spielern.

5.

$$\frac{\sum x_i}{7} = 9 \rightarrow \sum x_i = 63$$

50% aller Werte müssen kleiner als der Median sein, die kleinste natürliche Zahl ist die 1. 50% aller Werte müssen größer als der Median sein, die kleinsten möglichen Zahlen sind also 11 und 12.

i	1	2	3	4	5	6	7
x_i	x	12	11	10	3	2	1

Die Tabelle zeigt die kleinest mögliche Summe der Summanden außer dem größten Summanden (x). Setzt man dies in die Gleichung ein, erhält man:

$$39 + x = 63 \rightarrow x = 24$$

Die größte Zahl kann also maximal 24 betragen.

7.3 Übungsaufgaben zum Kapitel Wahrscheinlichkeitsverteilungen

1. a) $p(1,2,3,4,5,6) = (1/6)^6$
   Anordnungsmöglichkeiten für $\{1,2,3,4,5,6\} : 6! = 720\, p(6\, unterschiedliche) = (1/6)^6 \cdot 720 = 5/324 \approx 0,015$
   b) $p(6 \cdot 6) = (1/6)^6 \approx 0,000021$
   c) $n = 6, k = 4, p = (1/6), q = (5/6)$
   $p(k = 4) = \binom{6}{4} \cdot (1/6)^4 \cdot (5/6)^2 = 375/6^6 \approx 0,008$
   d) $p = q = (1/2), n = k = 6$
   $p(k = 6) = \binom{6}{6} \cdot (1/2)^3 \cdot (1/2)^3 = 1/64 \approx 0,016$
   e) Zunächst wie drei Sechser:
   $p(k = 3) = \binom{6}{3} \cdot (1/6)^3 \cdot (5/6)^3$
   Es gibt aber sechs Augenzahlen, also das Ganze mit 6 multiplizieren.
   $= 625/6^5 \approx 0,322$

2. $p(k = 2) = \binom{3}{2} \cdot 0,4^2 \cdot 0,6 = 0,288$

8 Anhang

8.1 Binomialkoeffizient

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

Sprich "n über k".

Diese Zahl tritt als Koeffizient bei der Lösung der Aufgabe $(x+y)^n$ auf, also z.B. auch in den binomischen Formeln, deren Verallgemeinerung der binomische Lehrsatz ist.

$$(x+y)^{n=2} = \binom{n}{k=0} x^n + \binom{n}{k=1} x^{n-k} \cdot y^k + \binom{n}{k=n} y^k = x^2 + 2xy + y^2$$

n ist also der Grad des Binoms und k wird von 0 bis n durchgezählt. Während die Potenz von x von n bis 0 verringert wird, wird diejenige von y umgekehrt gesteigert.

Anschaulich sagt der Binomialkoeffizient aus, wie viele Möglichkeiten es gibt, k aus n gezogene Objekte anzuordnen, wenn die Reihenfolge nicht wichtig ist, wie z.B. beim Lotto:

$$\binom{49}{6} = 13.983.816$$

Um den Binomialkoeffizienten auszurechnen, kann das Pascal'sche Dreieck verwendet werden. Man fängt an der Spitze mit 1 an und addiert immer zwei nebeneinanderstehende Zahlen, deren Summe man in den Zwischenraum darunter schreibt:

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1

Daran kann man auch sehr schön die Symmetrie des Binomialkoeffizienten sehen sowie die Regel, dass $\binom{n}{0}$ immer 1 ist und $\binom{n}{1} = \binom{n}{n-1}$ immer n ist.

8.2 Mehr zur Varianz

Berechnet man die Varianz so, wie im Skript angegeben, besitzt dieser Schätzer für die Varianz der Grundgesamtheit einen systematischen Fehler, da man die Varianz immer unterschätzt. Dies liegt daran, dass man für das Berechnen des Mittelwertes der Stichprobe einen sogenannten *Freiheitsgrad* verliert. Folgendes Beispiel soll dies verdeutlichen:

Es sei der Mittelwert einer Stichprobe mit 5 Messwerten $\bar{x} = 2$ gegeben. Wieviele der 5 Messwerte lassen sich dadurch frei wählen? Die Antwort lautet: *Nur vier, denn der letzte Messwert muss so gewählt werden, dass der Mittelwert der Stichprobe stimmt.* Unser Beispiel hat also einen Freiheitsgrad durch den Mittelwert verloren.

Das Konzept der Freiheitsgrade wird einem noch öfter begegnen, gehört aber eher zum "advanced stuff" und ist nur etwas für Experten, die sich vertieft für Statistik interessieren.