Skriptum Elemente der Stochastik für Lehramtskandidaten WS 15/16

T. von der Twer

Inhaltsverzeichnis

Kapitel 1. Elementare Wahrscheinlichkeitstheorie und Statistik 1
1. Einleitung: Eine kleine historische Vorbemerkung, und Motivierung 1
2. Vorbereitung: Grundlegendes über das Rechnen mit Mengen 2
3. Die Grundbegriffe der Wahrscheinlichkeitstheorie 4
4. Der Begriff der Unabhängigkeit und das Rechnen mit bedingten Wahrscheinlichkeiten 8
5. Wahrscheinlichkeitsbäume und Pfadregel 11
6. Ergänzung: Der Begriff der Unabhängigkeit eines Systems von n Ereignissen $A_1, ...A_n$ 12

Kapitel 2. Zufallsvariablen und ihre Verteilungen: Einführung unter deskriptivem Aspekt 15
1. Begriff der Zufallsvariablen 15
2. Begriff der Verteilung einer Zufallsvariablen und der Verteilungsfunktion 16
3. Deskriptive Statistik 18

Kapitel 3. Mathematisches zu den wichtigsten Verteilungen und ein wenig Inferenzstatistik 27
1. Verknüpfung von Zufallsvariablen, Unabhängigkeit und Verteilungsparameter 27
2. Die wichtigen Verteilungsparameter Erwartungswert und Varianz (bzw. Streuung) 31
3. Die wichtigsten diskreten Verteilungen 35
4. Verteilungen mit Dichte 42
5. Einige Anwendungen der Normalverteilung 47
6. Anwendung der $t-$ Verteilung auf die Mittelwertschätzung 51
7. Die χ^2- Verteilungen 54
8. Das Schema des statistischen Hypothesentestens, mit Beispielen 58

Kapitel 4. Zusammenfassung der wichtigsten Begriffe und Resultate **sowie Formeln** 63

KAPITEL 1

Elementare Wahrscheinlichkeitstheorie und Statistik

1. Einleitung: Eine kleine historische Vorbemerkung, und Motivierung

Viele Aussagen kann man nicht mit Sicherheit behaupten, zum Beispiel kann man nicht mit Sicherheit sagen, dass es morgen regnet / nicht regnet, dass man nächste Woche im Lotto keine sechs Richtige haben wird. Die vernünftige Reaktion darauf ist nun nicht, dass man gar nichts mehr sagt (weil man nichts Gewisses sagen kann), auch nicht, dass man eine Aussage als gewiss ausgibt, die es nicht ist. Sondern man wird seine Behauptung durch 'wahrscheinlich', 'vermutlich' abschwächen. Das taten die Leute auch schon in der Antike, und so entstand die Bezeichnung 'Stochastik' aus dem Griechischen '$\sigma\tau\grave{o}\chi os$', was 'Vermutung' heißt. Stochastik ist also in der bloßen Wortbedeutung die Lehre vom Vermuten. Nun wurden oben jedoch durchaus heterogene Beispiele aufgeführt: Was man so im Alltag 'wahrscheinlich', 'unwahrscheinlich' nennt, hat nichts mit mathematischer Wahrscheinlichkeitstheorie zu tun - man hat den Begriff dann nur in völlig subjektiver Bedeutung. Raffiniert könnte man sogar einen Zahlenwert zuordnen, indem man Leute fragt, wie viel sie auf das Eintreten oder Nichteintreten des betreffenden Ereignisses setzen würden und die Quotienten mittelt.

Wahrscheinlichkeitstheorie und Statistik sind durchaus neuzeitliche Schöpfungen. Wahrscheinlichkeitstheorie begann im 17. Jahrhundert mit Pascal und Fermat, Jakob Bernoulli, war zum Teil durch objektive Fragen zu den Chancen bei gewissen Glücksspielen motiviert, gewann aber durchaus mathematische Substanz, gipfelnd im Bernoullischen Gesetz der großen Zahlen. Wahrscheinlichkeitstheorie wurde weiterentwickelt von Gauß, Chebyshev und anderen im 19. Jahrhundert (Gaußsche Charakterisierung der Normalverteilung, deren Bild nun 'Gaußsche Glocke' heißt, Chebyshevs Ungleichung - Beides werden wir genauer kennenlernen), dann förmlich eine Explosion im 20. Jahrhundert, die fortdauert: Axiomatische Grundlegung der abstrakten allgemeinen Wahrscheinlichkeitstheorie 1933 durch Kolmogorov, Theorie der stochastischen (oder Zufalls-) Prozesse, begründet von Markov und Wiener, starke Erweiterung der Theorie durch viele bedeutende Mathematiker, wobei die Wahrscheinlichkeitstheorie immer stärker mit Analysis, Funktionalanalysis, Maßtheorie usw. vernetzt wurde. Es entstand noch ein eigenes 'Kind' der Wahrscheinlichkeitstheorie, die Infomationstheorie, von Claude Shannon begründet (nach dem Ende des II. Weltkrieges, etwa 1946). Ein besonders aktives Gebiet ist das der stochastischen Differentialgleichungen, das bereits mit Wiener begann, aber bis heute immer mehr an Dynamik gewann. Zum Ende des 19. Jahrhunderts und mit der Quantenmechanik entstand auch ein fruchtbares Wechselspiel zwischen Wahrscheinlichkeitstheorie und Theoretischer Physik ('Statistische Mechanik', die immer noch Forschungsgegenstand ist sowie Quantenphysik). Wahrscheinlichkeitstheorie ist also (mit anderen Theorien) im Zentrum eines riesigen theoretischen Komplexes, den wohl niemand mehr ganz überblicken kann.

Was ist Statistik? Wir kennen den Begriff 'Statistiken' aus dem Alltag: Es werden Häufigkeiten (oder relative Häufigkeiten, oder prozentuale) gezählt und diese Daten werden graphisch dargestellt oder auch weiter für Schlüsse (nach Wahrscheinlichkeit) verwendet. Das gibt es aber nicht nur vom Statistischen Bundesamt oder von Wirtschaftsverbänden, sondern auch als 'Sozialstatistik', 'Wetterstatistik' usw. Ingenieure, Physiker, Chemiker, Biologen, Agronomen, Meteorologen, Archäologen ... wenden Statistik an, wenn es um die Verarbeitung von Messdaten geht. Man sieht sofort, dass eine riesige Industrie von Statistik entstanden ist. Was hat nun Statistik mit Wahrscheinlichkeitstheorie zu tun? Man macht (in der moderneren Entwicklung Folgendes: Es wird ein Modell hinter die Daten gebaut und dann mit den Schlüssen weitergearbeitet, deren theoretische Grundlage die Wahrscheinlichkeitstheorie ist. Statistik besteht also aus Datensammlung ('sampling'), Darstellung der Daten (insbesondere graphische) und Anwendungen jeweils fachspezifischer Modellbildung sowie der Wahrscheinlichkeitstheorie. Statistik befasst sich auch eigenständig (jeweils fachspezifisch) mit den Problemen, die sich bei der Datenerhebung

ergeben. (Gute Daten zu produzieren macht viel Arbeit und erfordert großes Geschick, und überall sind Datensammler am Werk, denen Beides fehlt - anschließend wird die Sache durch ein paar Computerprogramme gejagt, die eigentlich gut sind, aber aus miserablen Daten nur elaborierteren Quatsch machen.) Umgekehrt können gute Leute ihres Faches mit bloßem Geschick beim Anschauen der Daten zu sehr feinen und weitreichenden Einsichten gelangen, ohne viel 'Inferenzstatistik' ('schließende Statistik') zu benötigen. Das geht aber nicht so billig, wie schlechte Daten zu sammeln und durch den Computer zu jagen. Wir gehen zurück zum Anfang, das (verständige) Sammeln von Daten - auch das ist neuzeitlich, und nennen zwei ganz verschiedene Protagonisten, die eigenständig damit begannen: Vauban war im 17. Jahrhundert berühmter Festungsbaumeister, Mathematiker und Vieles mehr. Bei seinen Reisen durch das ganze Frankreich bemerkte er das wirtschaftlich und soziale Elend der Bevölkerung überall, wollte daran etwas ändern und begann, Sozialstatistiken anzufertigen. Cotta (an der Wende des 18. Jahrhunderts zum 19.) war der Begründer der Forstwissenschaft - seine Methoden verbreiteten sich rasch auf der ganzen Erde. Es gelang ihm, den miserablen Bestand der Wälder, die rücksichtslos nur ausgebeutet worden waren und vielfach verschwunden waren oder völlig darniederlagen, 'aufzuforsten'. Seine Grundlage dafür war sorgfältige statistische Erhebung und Rechnung, woraus er schießen konnte, wie viel nur entnommen werden durfte, was und wie viel neu anzupflanzen und zu pflegen war - und das eben über rieseige Landstriche!

Abschließend nennen wir aus dem heterogenen Feld, das man 'Statistik' nennt, typische Beispiele, abschreckende (ganze Beispielgruppen) und sehr gute und schöne:

Erste Beispielguppe (da sind es jeweils nur Vertreter riesiger Klassen): Es werden Statistiken angefertigt zu unglaublich unsinnigen Themen, wie z.B. ob Menschen mit 'fiesen' Initialen wie P.I.G tendenziell schlechte, solche mit 'guten' Initialen oder Namen wie 'König', 'S.I.R.' aber bessere Lebensläufe haben. Es werden auch 'streng wissenschaftlich' unglaublich banale Dinge 'bewiesen', die jeder Mensch mit etwas Alltagswissen bereits weiß, zum Beispiel, dass Menschen, die viel lachen, gesünder sind und glücklicher als andere (wer hätte das gedacht?). Schließlich werden statistische Daten entweder schon in unnützer Weise erhoben oder so unzureichend aufbereitet und mitgeteilt, dass man höchstens noch einen Unfug daraus machen kann, wenn es zum Beispiel heißt, dieses Jahr habe es 'mehr' Verkehrstote gegeben als im letzten, oder wenn nicht mehr mitgeteilt wird als der gesamte Bundesdurchschnitt bei Einkommen etc.

Eine Bemerkung zur ersten Beispielgruppe - ein besonders abschreckendes Beispiel, das aber eben doch noch etwas Gutes zeigt und einen guten Ausgang nimmt: Im 19. Jahrhundet war die 'Phrenologie', wörtlich 'Hirnforschung', groß in Mode - und die Vorstellungen dazu waren ein kindischer Quatsch, dazu widerlich ideologisch bestimmt - man bedenke aber, dass man von so vielem nun gar nichts wissen konnte, von dem man heute noch kaum eine Ahnung hat (!) : Man dachte tatsächlich, die Qualität eines Gehirns an seinem Gewicht 'messen' zu können und wähnte sich auf gutem Wege, weil ja so liebe Vorstellungen wie der 'physiologische Schwachsinn des Weibes' (so ein Titel!) sehr gut dazu passten. Nun stelle man sich vor: Die Leute nahmen das ernst mit diesen Messungen und entfalteten riesige Anstrengungen, in geeigneter Weise 'normiert' das Hirngewicht zu messen, das war gar nicht leicht, weil nach dem Tod sich so schnell davon verlor. Also wurde mit peinlicher Sorgfalt alles genau bedacht wie Temperatur, Räume, Zeitpunkt nach dem Tod usw. So weit klingt das noch schlimmer als die heute abschreckendsten Beispiele. Der gute Ausgang der Sache aber: Es wurde damit immer klarer, dass sich die ganze Vorstellung von der Sache nicht halten ließ: Man fand objektiv heraus, dass das Hirngewicht mit der Intelligenz nichts zu tun hat, das es vielmehr deutlich korreliert ist mit dem Körperbau. Und mehr noch: Man hielt sich daran und gab die für selbstverständlich gehaltenen Vorstellungen auf - also: 'Das Weib' ist ebenso wenig schwachsinnig wie der zartgebaute männliche südfranzösische Professor! Das würde man sich auch heute manchmal wünschen: Eine mehr empirisch-pragmatisch als ideologisch ausgerichtete Pädagogik z.B.!

Zweite Beispielgruppe: Grundvorstellungen der Mendelschen Genetik konnten gestützt werden durch Voraussagen mit wahrscheinlichkeitstheretischen Mitteln, mit entsprechend weitreichenden Anwendungen ('Hardy-Weinberg-Gesetz', wir werden das ein wenig behandeln).

2. Vorbereitung: Grundlegendes über das Rechnen mit Mengen

Das Folgende wird ziemlich bekannt sein und sollte daher wohl nur 'überflogen' werden. Allerdings sollte man sich die de Morganschen Gesetze sehr bewusst machen, das kann sonst leicht eine Quelle

unangenehmer Fehler sein. Außerdem führen wir eine Notation für die Anzahl einer endlichen Menge ein und ein praktisches Rechengesetz dafür.

Vorab wollen wir uns versichern, dass wir die Aussgenveknüpfung 'Wenn α, so β' (dabei sind α, β Aussagen) gut verstehen, formaler geschrieben: $\alpha \implies \beta$. Der Wahrheitswert dieser Verknüpfung ist definiert als der von $\neg\alpha \vee \beta$ (in Worten: '(nicht α) oder β'). Dabei ist 'oder' als nicht ausschließendes zu verstehen, also ist $\alpha \vee \beta$ wahr genau dann, wenn *mindestens* eine der beiden Aussagen α, β wahr ist. Es ist auch günstig, die Aussage $\neg(\alpha \implies \beta)$ intuitiv richtig zu verstehen - sie bedeutet nach dem Vorangehenden: 'α ist wahr und β ist falsch', formal geschrieben: $\alpha \wedge \neg\beta$. Naheliegend formalisiert man die Aussagenverknüpfung $(\alpha \implies \beta) \wedge (\beta \implies \alpha)$ als $\alpha \iff \beta$, der Sinn ist klar: α und β haben (stets) denselben Wahrheitswert, in Worten: 'α genau dann, wenn β'. Diese Dinge sind grundlegend für das Verständnis der mathematischen Sprache in Definitionen und Sätzen überhaupt, und man versteht mit ihnen auch problemlos die folgenden Ausführungen über die einfache Boolesche Mengenalgebra.

Zwischen zwei Mengen A, B kann die folgende Teilmengenbeziehung bestehen (oder auch nicht):

$$A \subset B :\iff \text{ für alle } x : (\text{wenn } x \in A, \text{ dann } x \in B).$$

(Das ist eine Definition, zu bemerken am Doppelpunkt.) Insbesondere ist definiert:

$$A = B :\iff A \subset B \text{ und } B \subset A.$$

Man verknüpft Mengen A, B zu neuen Mengen, insbesondere zu

$$\begin{aligned} A \cup B &: = \{\, x |\, x \in A \text{ oder } x \in B\} \text{ (Vereinigung)} \\ A \cap B &: = \{\, x |\, x \in A \text{ und } x \in B\} \text{ (Durchschnitt)} \end{aligned}$$

Wir werden jeweils eine Grundmenge Ω haben, und alle dann zu betrachtenden Mengen werden Teilmengen von Ω sein. Dann ist definiert:

$$\mathcal{P}(\Omega) := \{\, A |\, A \subset \Omega\} \text{ (Potenzmenge von } \Omega).$$

In diesem Rahmen definieren wir dann die Operation der Komplementbildung:

$$\overline{A} := \{\, x \in \Omega |\, x \notin A\} \text{ (Komplement in } \Omega)$$

Allgemeiner bilden wir auch Mengendifferenzen:

$$\begin{aligned} A \setminus B &: = \{\, x \in A |\, x \notin B\} \ (A \text{ minus } B), \text{ also} \\ A \setminus B &= A \cap \overline{B}. \end{aligned}$$

Also ist $\overline{A} = \Omega \setminus A$, $\Omega \setminus \Omega = \emptyset$ (die leere Menge, welche überhaupt kein Element enthält). Folgende Booleschen Rechengesetze sind wichtig und entsprechen der simplen Aussagenlogik, also dem logischen Umgang mit 'und', 'oder', 'nicht' (Assoziativität, Allmenge, leere Menge und Komplement, Absorption und speziell Idempotenz, Distributivgesetze, de Morgansche Gesetze der Reihe nach):

$$\begin{aligned} A \cup \overline{A} &= \Omega, \ A \cap \overline{A} = \emptyset \\ A \cup (B \cup C) &= (A \cup B) \cup C, \ A \cap (B \cap C) = (A \cap B) \cap C \\ \text{insbesondere } A \cap A &= A, \ A \cup A = A. \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C), \ A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \\ \overline{A \cup B} &= \overline{A} \cap \overline{B}, \ \overline{A \cap B} = \overline{A} \cup \overline{B} \end{aligned}$$

zu beachten ist: Zu jeder Formel gibt es auch die duale, die dadurch entsteht, dass man $\cup, \cap$ austauscht sowie $\Omega, \emptyset$. Eine algebraische Struktur mit zwei Operationen und Konstanten dieser Art heißt eine **Boolesche Algebra**.Man leitet z.B. ab:

$$A \setminus (B \cup C) = A \cap \overline{B \cup C} = A \cap \left(\overline{B} \cap \overline{C}\right).$$

Schließlich bezeichnen wir für endliche Mengen mit $|A|$ die Anzahl der Elemente von A. Das ist dann eine Zahl aus $\mathbb{N}_0$. Es ist sicher klar, dass $|A \cup B| = |A| + |B|$, **wenn** $A \cap B = \emptyset$. Für die Wahrscheinleichkeitsrechnung ziemlich wichtig ist folgende Rechenregel für die Menge

$$A^B := \{\, f |\, f : B \to A\},$$

und für endliche Mengen A, B hat man:

$$\left|A^B\right| = |A|^{|B|}.$$

Beispiel: Ein Wort der Länge 5 (also mit 5 Buchstaben) aus einem Alphabet mit 26 Buchstaben kann man auf 26^5 Weisen bilden. Denn jedes derartige Wort ist aufzufassen als Abbildung $\{1,2,3,4,5\} \to \{n \in \mathbb{N} | \, 1 \leq n \leq 26\}$.

Der Zusammenhang zwischen der Booleschen Algebra auf einer Potenzmenge von Ω mit $\cap, \cup$, und der Aussagenlogik mit $\wedge, \vee, \neg$ (und, oder, nicht) ist sehr einfach:

$$A \cup B \text{ entspricht der Aussage } x \in A \vee x \in B, \text{ deren Form } \alpha \vee \beta \text{ ist, usw.}$$

die Aussage über Mengen

$$A = B$$

hat die aussagenlogische Form

$$\alpha \iff \beta \text{ (nämlich } x \in A \iff x \in B).$$

Also kann man z.B. das de Morgan-Gesetz

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

zeigen, indem man die aussagenlogische Allgemeingültigkeit beweist für die aussagenenlogische Verknüpfung

$$\neg(\alpha \vee \beta) \iff \neg\alpha \wedge \neg\beta,$$

das heißt: Was immer die Wahrheitswerte der Aussagenvariablen α, β sind: Stets haben $\neg(\alpha \vee \beta)$ und $\neg\alpha \vee \neg\beta$ denselben Wahrheitswert. Wir zeigen das mit folgender 'Wahrheitstafel' (wobei 'W' steht für 'wahr', 'F' für 'falsch'):

α	W	W	F	F
β	W	F	W	F
$\alpha \vee \beta$	W	W	W	F
$\neg(\alpha \vee \beta)$	F	F	F	W
$\neg\alpha \wedge \neg\beta$	F	F	F	W

3. Die Grundbegriffe der Wahrscheinlichkeitstheorie

3.1. Kolmogorovs Axiomensystem für Wahrscheinlichleitsräume. Wir erklären diese Grundbegriffe an einem sehr einfachen und einem etwas komplizierteren Beispiel: Das einfache: Mit einem gewöhnlichen Würfel wird einmal gewürfelt. Das kompliziertere: Mit einem gewöhnlichen Würfel wird drei mal gewürfelt.

DEFINITION 1. *Ein Ergebnisraum (oder die 'Ergebnismenge') ist eine nicht leere Menge Ω. Inhaltliche Deutung: Die Elemente von Ω sind die möglichen Ausgänge eines Zufallsexperiments.*

Beispiele: Beim einmaligen Würfeln hat man $\Omega = \{1,2,3,4,5,6\}$. Denn es kommt eine der Zahlen $1,2,3,4,5,6$ heraus. Beim dreimaligen Würfeln hat man

$$\Omega = \{(a_1,a_2,a_3) | \, a_i \in \{1,2,3,4,5,6\}\}.$$

Denn es kommt eine Folge der Länge drei der möglichen Augenzahlen von 1 bis 6 heraus.

Bemerkung: In der einfachsten Wahrscheinlichkeitsrechnung ist Ω endlich, wichtig sind dann auch noch die Fälle, dass Ω abzählbar unendlich ist (also mit den natürlichen Zahlen durchzuzählen), dann aber noch der Fall, dass Ω der Menge aller reellen Zahlen entspricht.

Der entscheidende Schachzug bei der mathematischen Fassung des Begriffs der Wahrscheinlichkeit (das Ganze gelang erst Kolmogorov 1933 (!)) ist nun folgender: Man fragt nur nach Wahrscheinlichkeiten von Ereignissen im Rahmen eines Zufallsexperiments, zu dem ein fester Ergebnisraum Ω gehört.

Wir geben nun ein paar Beispiele, welche die folgenden Definitionen motivieren werden: Beim einmaligen Würfeln fragt man etwa, wie wahrscheinlich es ist, dass eine Augenzahl unter 3 herauskommt. Oder bei dreimaligem Würfeln nach der Wahrscheinlichkeit dafür, dass höchstens eine Sechs dabei ist. Stets ist es ein Ereignis, das bei Durchführung des betreffenden Zufallsexperiments geschehen kann, nach dessen Wahrscheinlichkeit man fragt, und stets kann man ein solches Ereignis so fassen im Rahmen des betreffenden Ω:

(1) 'das herauskommende $\omega \in \Omega$ hat die Eigenschaft E'.

Im ersten Beispiel: '$\omega < 3$', im zweiten Fall '$\omega = (a_1, a_2, a_3)$, und es gilt nicht, dass es $i \neq j$ gibt, $1 \leq i, j \leq 3$, so dass $a_i = a_j = 6$'. Nun kann man jede Eigenschaft im Rahmen einer Grundmenge Ω identifizieren mit einer Teilmenge von Ω, man bildet also zum Ereignis (1):

$$(2)\ A_E := \{\omega \in \Omega | \, \omega \text{ hat die Eigenschaft } E\} \subset \Omega.$$

Auf diese Weise werden die im Rahmen von Ω formulierbaren Ereignisse einfach Teilmengen von Ω. Wir wollen nun die Menge aller Ereignisse in den Griff bekommen. Jedenfalls ist das eine Teilmenge von Ω. Im endlichen Fall nimmt man einfach die gesamte Potenzmenge von Ω, also die Menge aller Teilmengen von Ω. Insbesondere für $\Omega = \mathbb{R}$ macht das Schwierigkeiten, dann verlangt man nur, was für alle praktischen Bedürfnisse ausreicht: Man nimmt eine sogenannte $\sigma-$ Algebra $\mathcal{S}$ von Teilmengen von Ω, und die soll alle Intervalle enthalten und abgeschlossen sein unter abzählbaren Vereinigungen und Durchschnitten sowie Komplementbildung. $\emptyset$ und $\mathbb{R} = \Omega$ selbst gehören dann insbesondere dazu. Nun kommt ein Drittes dazu: jedem Element aus $\mathcal{S}$ wird eine Wahrscheinlichkeit eindeutig zugeordnet. Das sollte eine Zahl im Bereich $[0, 1]$ sein. Man hat also eine Wahrscheinlichkeitsfunktion $P : \mathcal{S} \to [0, 1]$. Aber es sollte diese Funktion P nicht beliebig sein. Denn denken wir an eine bekannte Erfahrungstatsache: Wenn man ein Zufallsexperiment oft wiederholt und betrachtet die relative Häufigkeit, mit der ein Ereignis $A \in \mathcal{S}$ dabei vorgekommen ist, so stabilisiert sich auf Dauer gerade die Wahrscheinlichkeit als zu beobachtende relative Häufigkeit (diese ist die Anzahl der Versuche, bei denen sich A ereignet hat, geteilt durch die Anzahl aller Versuche). Die beobachtete relative Häufigkeit wird fast niemals genau die Wahrscheinlichkeit sein, aber stets sehr nahe daran bei großer Versuchsreihe. Darum sollte die Funktion P die wesentlichen Eigenschaften besitzen, welche der Begriff 'relative Häufigkeit' besitzt, nämlich: $P(\Omega) = 1$, und $P(A \cup B) = P(A) + P(B)$, falls $A \cap B = \emptyset$.

Wir fassen nunmehr das Ganze zum Kolmogorovschen Begriffssystem zusammen:

DEFINITION 2. *Ein Wahrscheinlichkeitsraum ist ein Tripel* $(\Omega, \mathcal{S}, P)$, *wobei* Ω *ein Ergebnisraum,* $\mathcal{S}$ *eine* $\sigma-$ *Algebra von Mengen in* $\mathcal{P}(\Omega)$ *(= Potenzmenge von* Ω*) ist und* P *eine Wahrscheinlichkeitsfunktion auf* $\mathcal{S}$, *welche folgende Eigenschaften hat:*

$$P: \begin{array}{ccc} \mathcal{S} & \to & [0,1] \\ A & \mapsto & P(A) \end{array}, \text{ und}$$

$$(i)\ P(\Omega) = 1$$

$$(ii)\ P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k), \text{ falls } A_m \cap A_k = \emptyset \text{ für } k \neq m,\ k, m \in \mathbb{N}.$$

Speziell im Fall endlicher Menge Ω *kann man stets* $\mathcal{S} = \mathcal{P}(\Omega)$ *wählen, und anstelle des Axioms* (ii) *genügt es, nur zu fordern, dass*

$$(ii)'\ P(A \cup B) = P(A) + P(B), \text{ falls } A \cap B = \emptyset.$$

Bemerkung: Klar folgt $(ii)'$ aus (ii). Wenn aber Ω endlich ist, dann gilt das auch umgekehrt: Es gibt dann nur endlich viele Teilmengen von Ω, also nur endliche viele Ereignisse, und aus $(ii)'$ kann man mit Induktion die Aussage auch für disjunkte Vereinigungen von endlich vielen Mengen folgern.

Noch eine Bemerkung zu $(ii)'$: Die Existenz der unendlichen Summe ist völlig unproblematisch, da sie beschränkt und monoton wachsend ist.

Wir kommen zu ein paar Folgerungen aus dem Axiomensystem, die große Bedeutung für das praktische Wahrscheinlichkeitsrechnen haben:

FOLGERUNG 1. *Für jede Wahrscheinlichkeitsfunktion* P *und für alle Ereignisse* A, B *in* $\mathcal{S}$ *gilt:*

$$\begin{aligned} (1)\ P\left(\overline{A}\right) &= 1 - P(A), \\ (2)\ P(A \cup B) &= P(A) + P(B) - P(A \cap B), \text{ (ohne Bedingung an } A, B), \\ (3)\ A \subset B &\implies P(A) \leq P(B). \end{aligned}$$

Begründungen: $1 = P(\Omega) = P\left(A \cup \overline{A}\right) = P(A) + P\left(\overline{A}\right)$ mit den Axiomen $(i), (ii)$ und der Tatsache $A \cap \overline{A} = \emptyset$. Also (1).

Zu (2) : $A \cup B = A \cup \left(B \cap \overline{A}\right)$, und diese Vereinigung ist disjunkt. Also mit

$$\begin{aligned} B &= \left(B \cap \overline{A}\right) \cup P\left(B \cap A\right) \text{ (diskunkte Vereinigung):} \\ P\left(B\right) &= P\left(B \cap \overline{A}\right) + P\left(B \cap A\right), \end{aligned}$$

insgesamt

$$\begin{aligned} P(A \cup B) &= P(A) + P\left(B \cap \overline{A}\right) \\ &= P\left(A\right) + P\left(B\right) - P\left(A \cap B\right). \blacksquare \end{aligned}$$

(3) überlassen wir als einfache Übungsaufgabe.

Eine Warnung: Aus (1) folgt sofort $P\left(\emptyset\right) = 0$. Aber man darf nicht aus $P\left(A\right) = \emptyset$ darauf schließen, dass $A = \emptyset$. Hier ein Beispiel: Würfelt man unendlich oft, dann ist das Ereignis: 'Eine Folge konstant Sechs wird gewürfelt' nicht leer, es enthält eine Folge, aber seine Wahrscheinlichkeit ist Null. Oder: Man schießt zufällig auf das Intervall $[0, 1]$. Dann ist das Ereignis $[2, 2]$ nicht leer, aber seine Wahrscheinlichkeit ist Null. Ebenso kann ein Ereignis $A \subsetneqq \Omega$ existieren mit $P\left(A\right) = 1$. Also: Wenn ein Ereignis logisch unmöglich ist, so hat es Wahrscheinlichkeit Null, aber das ist nicht umkehrbar. Ebenso folgt aus logischer Sicherheit die Wahrscheinlichkeit 1, aber das ist wieder nicht umkehrbar.

Wir kommen nun zur elementaren Wahrscheinlichkeitsrechnung durch 'Abzählen', die man immer in Laplace-Räumen hat:

DEFINITION 3. $\left(\Omega, \mathcal{P}\left(\Omega\right), P\right)$ *heißt Laplace-Raum, wenn* Ω *endlich ist und alle Elementarereignisse*

$$\{\omega\},\ \omega \in \Omega,$$

dieselbe Wahrscheinlichkeit haben. Folgerungen sind sofort:

$$\begin{aligned} \textit{Mit } |\Omega| &= n \textit{ gilt dann} \\ P\left(\{\omega\}\right) &= \frac{1}{n}, \textit{ und für alle } A \subset \Omega \textit{ gilt:} \\ (*)\ P\left(A\right) &= \frac{|A|}{|\Omega|} = \textit{ relative Häufigkeit von } A \textit{ in } \Omega. \end{aligned}$$

Die folgende Aussage gilt stets bei endlichem Ω*, auch wenn die Elementarereignisse nicht dieselben Wahrscheinlichkeiten besitzen:*

$$(+)\ P\left(A\right) = \sum_{\omega \in A} P\left(\{\omega\}\right).$$

Bemerkungen: Die Formel $(*)$ beschreibt man traditionell verbal auch so: 'Die Wahrscheinlichkeit ist die Anzahl der günstigen Fälle geteilt durch die Anzahl der möglichen Fälle'. Dabei heißen nur einfach genau die zu A gehörigen Fälle 'günstig', auch wenn sie fatal sein sollten. Die Formel $(+)$ gilt auch für abzählbar unendliches Ω, wenn $\mathcal{S} = \mathcal{P}\left(\Omega\right)$, so dass alle $\{\omega\} \in \mathcal{S}$ für $\omega \in \Omega$. Axiom (ii) liefert dann wieder $(+)$.

Warnung: Man beachte jedoch, dass diese Formel *nur* dann gilt, wenn alle Elementarereignisse dieselbe Wahrscheinlichkeit haben. Man findet ein Gegenbeispiel sofort, wenn man bei zweimaligem Würfeln die Menge $\Omega = \{2, 3, ...12\}$ der Augensummen bildet. Diese sind nicht gleich wahrscheinlich, z.B. $P\left(\{2\}\right) \neq P\left(\{7\}\right)$.Bei der Anwendung der Laplace-Formel käme jedoch beide Male 1/11 heraus.

Begründung der Formeln: Zunächst ist $(+)$ richtig, weil mit $A = \{\omega_{i_1}, ..., \omega_{i_k}\}$ (die aufgezählten Elemente seien paarweise verschieden) gilt:

$$P\left(A\right) = P\left(\bigcup_{i=1}^{k} \{\omega_{i_k}\}\right) =^{(ii)} \sum_{i=1}^{k} P\left(\{\omega_{i_k}\}\right) = \sum_{\omega \in A} P\left(\{\omega\}\right).$$

(Man beachte: Die letzte Schreibweise ist viel eleganter, weil sie ohne Aufzählung auskommt. Wir werden solche Formulierungen stets vorziehen.) Für die leere Menge A gilt die Formel ebenfalls, die Summe rechts wird dann leer, bekommt also den Wert Null. Aus $(+)$ folgt mit der Laplaceeigenschaft dann sofort $(*)$, weil dann gilt:

$$\sum_{\omega \in A} P\left(\{\omega\}\right) = \frac{1}{n} \cdot |A|. \blacksquare$$

3.2. Beispiele für Zufallsexperimente und zugehörige Wahrscheinlichkeitsräume: 1.) Zufallsexperiment: Drei mal wird gewürfelt mit einem gewöhnlichen symmetrischen Würfel.

Dann definiert man zweckmäßig

$$\Omega := \{1,2,3,4,5,6\}^3 \left[= \{(k,m,n) | \, k,m,n \in \mathbb{N} \text{ und } 1 \leq k,m,n \leq 6\}\right].$$

Ferner setzt man $\mathcal{S} := \mathcal{P}(\Omega)$ und $P(A) := \dfrac{|A|}{|\Omega|}$. Damit hat man einen Laplace-Raum.

Bemerkung: Oft gibt es praktisch recht verschieden erscheinende Zufallsexperimente, die auf denselben Wahrscheinlichkeitsraum führen: So kann man das dreimalige Würfeln auch ersetzen durch:

A) Aus einer Urne mit sechs unterscheidbaren Kugeln wird drei mal eine gezogen, jedoch **'mit Zurücklegen'**, d.h. nach einem Zug wird die Kugel wieder in die Urne gelegt und werden die Kugeln gemischt.

B) Man würfelt nicht drei mal, sondern auf einmal mit drei unterscheidbaren Würfeln (die man etwa gefärbt hat, dann setzt man willkürlich eine Reihenfolge der Farben)

Wir betrachten (bei dreimaligem Würfeln) das Ereignis A : 'Es werden genau zwei Sechsen gewürfelt'. Dann kann man die Zahl der 'günstigen Fälle', d.h. die Zahl $|A|$, so ausrechnen:

$$|A| = 3 \cdot 5,$$

weil man 3 Möglichkeiten hat, die Positionen für die Sechsen zu wählen, unabhängig noch 5 Möglichkeiten für den verbleibenden Wurf, der keine Sechs ist. Außerdem ist in diesem Beispiel

$$|\Omega| = 6^3,$$

also

$$P(A) = \frac{15}{6^3} = \frac{5}{72}.$$

Im selben Rahmen betrachten wir das Ereignis B : 'Die Augensumme aller drei Würfe ist ≤ 4'. Dann zählen wir wieder ab, wie viele Elementarereignisse, also konkrete Folgen von drei Augenzahlen, dazu gehören, es sind:

$$(1,1,1),$$
$$(1,1,2),(1,2,1),(2,1,1),$$

also

$$P(B) = \frac{4}{6^3} = \frac{1}{54}.$$

2.) Wir geben ein Beispiel mit abzählbar unendlicher Menge Ω, hier $\Omega = \mathbb{N}$: **Zufallsexperiment: Man wirft so lange eine Münze, bis zum ersten Mal 'Kopf' kommt. Das Ergebnis ist dann die Anzahl der Würfe, die man benötigt hat.** (Es ist auch logisch möglich, dass man niemals 'Kopf' bekommt und damit keine endliche Anzahl zuordnen könnte. Diesen Ausgang lassen wir weg, er hätte ohnedies Wahrscheinlichkeit Null.) Damit haben wir $\Omega = \mathbb{N}$. Ferner wählen wir $\mathcal{S} := \mathcal{P}(\Omega)$ .Schließlich bekommen wir leicht (überlegen Sie das!)

$$P(\{n\}) = \frac{1}{2^n}, \text{ für alle } n \in \mathbb{N}.$$

Aus der Analysis sollten Sie wissen ('geometrische Reihe'), dass tatsächlich damit

$$P(\Omega) = P(\mathbb{N}) = \sum_{k=1}^{\infty} \frac{1}{2^n} = 1 \text{ wird,}$$

und mit Formel (+) haben wir

$$P(A) = \sum_{n \in A} \frac{1}{2^n} \text{ für alle } A \subset \mathbb{N}.$$

3.) Wir geben ein Beispiel mit überabzählbarer Menge Ω, hier $\Omega = [0,1]$, die Menge aller reellen Zahlen in diesem Intervall. **Zufallsexperiment: Ein idealer Pfeil mit punktförmiger Spitze wird auf das Einheitsintervall zufällig geworfen.** Für ein Intervall $I \subset [0,1]$ der Länge d hat man offenbar $P(I) = d$. Es ist nun ein berühmter Satz, dass man hier *nicht* $\mathcal{S} = \mathcal{P}(\Omega)$ wählen kann (was aber mit tiefer liegenden Resultaten der axiomatischen Mengenlehre zusammenhängt), *wenn* man das

Auswahlaxiom für die Mengenlehre haben will (und damit für die gesamte Mathematik). Daher wählt man $\mathcal{S} :=$ die kleinste $\sigma-$ Algebra, welche von den Intervallen erzeugt wird (also Boolesche Algebra, die unter **abzählbaren** Vereinigungen und Durchschnitten abgeschlossen ist und die Intervalle enthält). Das klingt kompliziert, dennoch ist der Wahrscheinlichkeitsraum $([0,1],\mathcal{S},P)$ damit sehr einfach: $P(A)$ wird einfach gemäß Axiom (ii) gebildet, 'hochgezogen' von den Intervallen. Nun kommen zwei wesentliche Vereinfachungen hinzu: Man wird sich in der Praxis kaum für komlizierteste Mengen aus $\mathcal{S}$ interessieren, sondern vielleicht nur für die Intervalle oder gerade noch für Mengen wie $[0,0.1]\cup[0.9,1]$. Dann sieht man, wie einfach hier die Wahrscheinlichkeitsrechnung ist. Kommt hinzu, dass man die Wahrscheinlichkeiten solcher Mengen hier sehr gut veranschaulichen kann durch Flächeninhalte: Man zeichnet einen Kasten der Höhe 1 über dem Intervall $[0,1]$ (oder auch den Graphen der Funktion $f(x)=1$), dann sind die Flächeninhalte über Intervallen (und Mengenverknüpfungen von Intervallen) einfach die Integrale der Funktion f in den entsprechenden Grenzen.

Bemerkung: Dies Beispiel der Gleichverteilung ist das einfachste in einer Reihe von sehr wichtigen theoretischen 'Verteilungen', die durch 'Dichtefunktionen' gegeben sind. Die Normalverteilung ist das wichtigste dieser Beispiele, wir werden sie und ein paar weitere kennenlernen - es gibt viele nützliche Anwendungen davon.

4. Der Begriff der Unabhängigkeit und das Rechnen mit bedingten Wahrscheinlichkeiten

Ein wichtiger Schachzug bei praktischer Berechnung von Wahrscheinlichkeiten besteht darin, beim Abzählen in verschiedene disjunkte Teilmengen zu unterteilen. Der Grund dafür, dass man die Wahrscheinlichkeiten dieser Teilmengen besser berechnen kann, liegt nicht nur darin, dass die Teilmengen kleiner sind und nicht so viele Fälle jeweils aufgeführt werden müssen. Sondern man betrachtet ein Ereignis A als eingetreten und zählt nur noch die Fälle von B, die schon in A liegen. Das gibt dann die Anzahl der günstigen Fälle von $B\cap A$. Dies wollen wir nun nicht mehr nur über Anzahlen, sondern allgemein mit Wahrscheinlichkeiten ausführen. Denn es hat weitere wichtige Aspekte: Die Wahrscheinlichkeit dafür, dass jemand die Piratenpartei wählt, ist sicher in der Wählergruppe der Leute unter 25 Jahren größer als in der Wählerschaft über 25. Man sagt zu einem solchen Sachverhalt dann: Die Ereignisse: 'Wähler über 25' und 'Wähler der Piratenpartei' sind nicht unabhängig. Einerseits ist ein solcher Sachverhalt vielfach als solcher interessant (wenn die Wahrscheinlichkeiten in den verschiedenen Töpfen **deutlich** verschieden sind), andererseits hängen viele sehr wichtige Dinge der Wahrscheinlichkeitstheorie und Statistik daran. Erwähnen wir hier erst nur den Umstand, dass wir bisher keine wirklich 'vernünftige' Formel für $P(A\cap B)$ haben - die Formel $P(A\cap B)=P(A)+P(B)-P(A\cup B)$ ist für praktische Zwecke wenig brauchbar. Bereits aus elementarem Alltagsverständnis wird man gern $P(A\cap B)=P(A)\cdot P(B)$ rechnen wollen, aber das gilt keineswegs allgemein. Wir werden nunmehr genau erklären, wann das gilt, und für den andern Fall eine brauchbare Alternative bereitstellen. Wir definieren nunmehr den äußerst wichtigen Grundbegriff der Unabhängigkeit zweier Ereignisse auf der Grundlage des Begriffs der bedingten Wahrscheinlichkeit:

DEFINITION 4. *Für zwei Ereignisse A,B mit $P(B)\neq 0$ definiert man*

$$(4)\quad P(A|B) := \frac{P(A\cap B)}{P(B)} \quad \textit{('Wahrscheinlichkeit von } A \textit{ bedingt durch } B\textit{')}.$$

Die Ereignisse A,B heißen unabhängig, wenn $P(A\cap B)=P(A)\cdot P(B)$, sonst abhängig.

Zum Verständnis: Die bedingte Wahrscheinlichkeit versteht man anschaulich gut, wenn man denkt an die Wahrscheinlichkeit von A, wenn man aus dem Topf B zieht. Denn im Topf B bedeutet A dasselbe wie $A\cap B$ (der Zähler), und nun ist durch $P(B)$ zu teilen, weil nunmehr B die Wahrscheinlichkeit 1 bekommen soll. Wie oben eingeführt, ist es natürlich, Ereignisse A,B unabhängig zu nennen, wenn

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

Das ist aber nur brauchbar, wenn $P(B)\neq 0$. Andererseits ergibt die Definition von $P(A|B)$ sofort, dass

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

Daher ist die eingeführte Definition des Begriffs der Unabhängigkeit gerade nur die Verallgemeinerung auf den Fall, dass $P(A)$ oder $P(B)$ auch Null sein kann.

Die Definitionen ergeben sofort zwei brauchbare Formeln zu Berechnung von $P(A\cap B)$:

FOLGERUNG 2. *Man hat*

$$(5)\quad P(A\cap B) = P(A)\cdot P(B)\,,\ \textbf{\textit{wenn}}\ A,B\ \textit{unabhängig sind.}$$
$$(6)\quad P(A\cap B) = P(A|B)\cdot P(B)\ \textit{(wenn nur } P(B)\neq 0\textit{).}$$

Bemerkung: Nur auf den ersten Blick erscheinen diese Formeln wenig praktikabel. Aber man bedenke: Die erste Formel kann man oft dann anwenden, wenn aus der Situation klar ist, dass zwei Ereignisse 'nichts miteinander zu tun' haben und daher sofort ohne jede Rechnung als unabhängig erkannt werden können. Z.B. haben die Augenzahlen der einzelnen Würfe bei fortgesetztem Würfeln nichts miteinander zu tun, und so sind Ereignisse wie: 'Sechs im ersten Wurf', 'Augenzahl unter 3 beim vierten Wurf' sicher unabhängig voneinander. Andererseits hat man vielfach schon intuitiv eine Abhängigkeit zu erwarten und rechnet dann nicht mit der ersten Formel, sondern mit der zweiten. Hier erscheint zunächst der Ausdruck $P(A|B)$ als Problem, aber das ist er im Allgemeinen nicht, vielmehr kann man solche bedingten Wahrscheinlichkeiten oft leichter direkt ermitteln als die unbedingte $P(A)$. Ein Beispiel: Man würfelt zwei mal. Wie wahrscheinlich ist eine Augensumme < 6 (Ereignis A)? Das ist schwieriger zu beantworten als die Frage, wie wahrscheinlich A ist bedingt dadurch, dass der erste Wurf eine 1 $(2,3,4,5,6)$ ergab. Klar ist

$$\begin{aligned} P(A|\text{ erster Wurf ist 1}) &= \frac{4}{6}=\frac{2}{3},\\ P(A|\text{ erster Wurf ist 2}) &= \frac{3}{6}=\frac{1}{2},\\ P(A|\text{ erster Wurf ist 3}) &= \frac{2}{6}=\frac{1}{3},\\ P(A|\text{ erster Wurf ist 4}) &= \frac{1}{6}. \end{aligned}$$

Offenbar kommt hinzu:

$$P(A|\text{ erster Wurf ist }\geq 5)=0.$$

Damit hat man über die zweite Formel z.B.:

$$P(A\cap\text{ 'erster Wurf ist 1'})=\frac{2}{3}\cdot\frac{1}{6}=\frac{1}{9}.$$

Nun kann man auch $P(A)$ über folgende Formel ausrechnen:

FOLGERUNG 3 (Formel von der totalen Wahrscheinlichkeit). *Wenn die Mengen $B_1,...,B_m$ eine Klasseneinteilung von Ω bilden, also*

$$\bigcup_{i=1}^{m} B_i=\Omega\ \textit{und}\ B_i\cap B_j=\emptyset\ \textit{für}\ i\neq j,\ 1\leq i,j\leq m,$$

dann hat man

$$(7)\quad P(A)=\sum_{k=1}^{m} P(A|B_k)\,P(B_k)\,.$$

Begründung: Man hat

$$P(A)=P\left(\bigcup_{k=1}^{n}(A\cap B_k)\right)=^{Ax.\ (ii)}\sum_{k=1}^{m}P(A\cap B_k)=^{(6)}\sum_{k=1}^{m}P(A|B_k)\,P(B_k)\,.\blacksquare$$

Beispiel: Im angeführten Beispiel für das Ereignis A : 'Augensumme < 6' ergibt sich

$$\begin{aligned} P(A) &= \frac{2}{3}\cdot\frac{1}{6}+\frac{1}{2}\cdot\frac{1}{6}+\frac{1}{3}\cdot\frac{1}{6}+\frac{1}{6}\cdot\frac{1}{6}\\ &= \frac{10}{36}=\frac{5}{18}. \end{aligned}$$

Wir haben noch eine weitere typisch interessante Fragestellung: Man hat gewisse Wahrscheinlichkeiten dafür, dass ein (nicht zu teurer) Material-Test etwa anzeigen soll, ob ein Werkstück gefährliche Risse hat oder nicht. Aber das wird nicht mit voller Sicherheit funktionieren, sondern stets werden defekte Stücke als 'sicher' durchgehen und ordentliche Stücke als 'Ausschuss' deklariert werden. Nun kann man mit viel Erfahrungswerten (etwa auch durch Materialprüfung bis zur Zerstörung oder andere sehr aufwendige

Verfahren folgende Daten (zumindest in guter Näherung) ermitteln - wir bezeichnen das Ereignis 'das zufällig gewählte Stück ist sicher' mit S, entsprechend das Ereignis 'das zufällig gewählte Stück wird vom Test als 'sicher' diagnostziert' mit 'S_T' ('sicher laut Test'). Zugleich führen wir Beispiele für diese Daten an, wie sie nicht untypisch vorkommen:

$$(+)\ \ P\left(S_T|\overline{S}\right) = 0.01,\ P\left(\overline{S_T}|S\right) = 0.02.$$

Man beachte: Ein vorsichtiger Test wird vielleicht eher ein ungefährliches Stück noch als gefährlich einstufen als ein gefährliches Stück als unbedenklich. Daher müssen diese Wahrscheinlichkeiten nicht gleich sein. Beide sind hier aber gering. Wir haben dann zwangsläufig:

$$(*)\ \ P\left(\overline{S_T}|\overline{S}\right) = 0.99 \text{ und } P\left(S_T|S\right) = 0.98.$$

Begründung von $(*)$ aus $(+)$**:**

$$P\left(\overline{S_T}|\overline{S}\right) + P\left(S_T|\overline{S}\right) = \frac{P\left(\overline{S_T} \cap \overline{S}\right)}{P\left(\overline{S}\right)} + \frac{P\left(S_T \cap \overline{S}\right)}{P\left(\overline{S}\right)} = \frac{1}{P\left(\overline{S}\right)}\left(P\left(\overline{S}\right)\right) = 1.$$

(analog für die zweite Aussage). Aber was uns praktisch interessiert, ist die Frage, mit welcher Wahrscheinlichkeit ein vom Test als sicher diagnostiziertes Teil nun wirklich sicher ist - und dasselbe mit 'gefährlich laut Test' und 'tatsächlich gefährlich'. Z.B. die erste Frage ist die nach $P\left(S|S_T\right)$, das dürfen wir nicht verwechseln mit $P\left(S_T|S\right) = 0.98$ (häufiger Fehler). Dazu haben wir folgende wichtige Bayessche Formel, welche einen Schluss auf die bedingte Wahrscheinlichkeit mit umgedrehter Bedingungsrichtung erlaubt - was man für Daten dazu braucht, werden wir sehen:

SATZ 1. *Es gilt für beliebige Ereignisse A, B mit $P\left(A\right), P\left(B\right) \neq 0$:*

$$(8)\quad P\left(B|A\right) = \frac{P\left(A|B\right)P\left(B\right)}{P\left(A\right)}.$$

Dabei ist oft praktisch $P\left(A\right)$ gemäß der Formel von der totalen Wahrscheinlichkeit auszurechnen als

$$P\left(A\right) = P\left(A|B\right)P\left(B\right) + P\left(A|\overline{B}\right)P\left(\overline{B}\right),$$

so dass entsteht:

$$(8')\quad P\left(B|A\right) = \frac{P\left(A|B\right)P\left(B\right)}{P\left(A|B\right)P\left(B\right) + P\left(A|\overline{B}\right)P\left(\overline{B}\right)}.$$

Begründung: (8) ist offenbar gleichwertig zu $P\left(B|A\right)P\left(B\right) = P\left(A|B\right)P\left(B\right)$, aber das bedeutet wieder dasselbe wir $P\left(B \cap A\right) = P\left(A \cap B\right)$, was klar ist.■

Diese simple Formel ermöglicht es nun gerade, eine Bedingungsrichtung *korrekt* umzudrehen, wie es unser Problem verlangte. Wir führen das aus, im Beispiel des Materialprüfungstests haben wir also

$$\begin{aligned} P\left(S|S_T\right) &= \frac{P\left(S_T|S\right)P\left(S\right)}{P\left(S_T|S\right)P\left(S\right) + P\left(S_T|\overline{S}\right)P\left(\overline{S}\right)} \\ &= \frac{0.98 \cdot P\left(S\right)}{0.98 \cdot P\left(S\right) + 0.01 \cdot P\left(\overline{S}\right)}. \end{aligned}$$

Wir haben eingefüllt, was wir wissen. Nun benötigen wir eine entscheidende Größe, um die gesuchte bedingte Wahrscheinlichkeit $P\left(S|S_T\right)$ auszurechnen, nämlich $P\left(S\right)$. Nehmen wir eine ordentliche Produktion an mit $P\left(S\right) = 0.99$. Dann erhält man:

$$P\left(S|S_T\right) = \frac{0.98 \cdot 0.99}{0.98 \cdot 0.99 + 0.01 \cdot 0.01} \approx 0.9999.$$

Die Wahrscheinlichkeit, dass das als unbedenklich eingestufte Stück auch wirklich unbedenklich ist, liegt dann glücklicherweise noch viel höher als $P\left(S_T|S\right) = 0.98$. Aber betrachten wir, was bei einer sehr schlampigen Produktion herauskäme mit $P\left(S\right) = 0.01$. Dann hat man tatsächlich

$$P\left(S|S_T\right) = \frac{0.98 \cdot 0.01}{0.98 \cdot 0.01 + 0.01 \cdot 0.99} < \frac{1}{2},$$

es ist dann gar kein Verlass mehr auf ein Testresultat 'sicher laut Test'. Die wesentliche Feststellung ist also, dass die Verlässlichkeit auf dem Anteil der gefährlichen Stücke beruht. Man kann $P\left(S|S_T\right)$ beliebig weiter nach unten drücken, wenn man $P\left(S\right)$ beliebig absenkt.

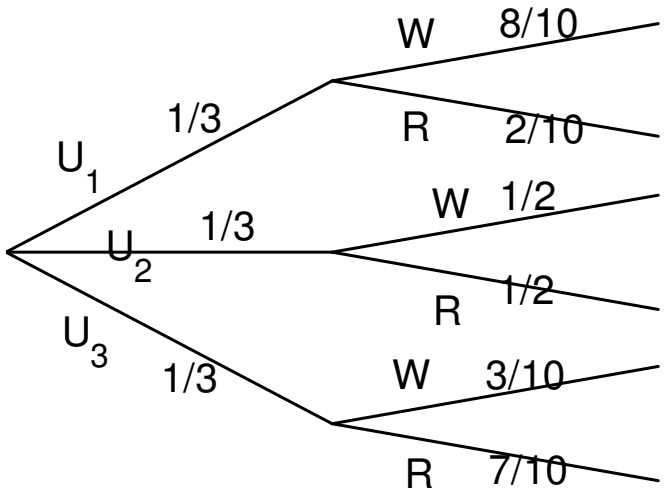
Ein gehobener Zweck von Statistik besteht nun darin, anhand von Messdaten Parameter zu schätzen bzw. aus Modellklassen ein Modell auszuwählen. Dafür verwendet man durchaus gerade die sogenannte Bayessche Methode, die auf (unter Umständen recht komplizierten) Ausführungen der Bayesschen Formel basiert. Die Sache fußt darauf, dass man die Wahrscheinlichkeit eines Modells bedingt durch die vorhandene Information berechnet, indem man das Wissen darüber benutzt, wie wahrscheinlich diese Information zu bekommen wäre, wenn das Modell zuträfe.

5. Wahrscheinlichkeitsbäume und Pfadregel

Es handelt sich nur darum, die Formel

$$(*)\ \ P(A \cap B) = P(A|B)\, P(B)\ \ (P(B) \neq 0)$$

in größerem Stil anzuwenden und dabei insbesondere fertig zu werden mit Wahrscheinlichkeitsberechnungen für mehrstufige Zufallsexperimente, die aus einer Folge von Zufallsexperimenten bestehen. Passend zur Fragestellung und auch zur Bequemlichkeit der Rechnung wendet man die Formel immer wieder an. Dazu bildet man geeignet jeweils eine Klasseneinteilung von Ausgängen zu jedem Experiment der Folge, aber möglicherweise abhängig davon, welches Ereignis der Klassifikation sich beim vorigen Experiment ereignet hat. Wir illustrieren das ein einem einfachen Beispiel: Man hat drei Urnen mit jeweils 10 Kugeln, die rot oder weiß sein können. Die erste Urne hat 2 rote, die zweite 5 rote, die dritte 7 rote Kugeln. Das Experiment bestehe nun darin, dass man zuerst eine Urne zufällig wählt. Dann zieht man aus der gewählten Urne zufällig eine Kugel und legt sie beiseite, in der dritten Stufe wird aus derselben Urne noch eine Kugel gezogen. Was ist die Wahrscheinlichkeit für 'eine rote und eine weiße Kugel wurde gezogen' (die Reihenfolge soll gleichgültig sein)? Auf der ersten Stufe haben wir die Klassifikation U_i, $i = 1, 2, 3$ (für die beschriebenen Urnen). Nun ist die Wahrscheinlichkeit für 'zuerst wird eine rote Kugel gezogen' abhängig von U_i. Auf der zweiten Stufe und der dritten haben wir die Klassifikation 'rot'-'weiß' wir nennen diese Ereignisse R_1, W_1 auf der zweiten Stufe und R_2, W_2 auf der dritten Stufe. Daher ergibt sich folgende Verzweigung und damit ein Baum, von dem hier nur die ersten beiden Stufen gezeichnet sind:



Ein Pfad in diesem Baum führt vom Ausgangspunkt zu einer der letzten Spitzen rechts. Die Wahrscheinlichkeit für das Durchlaufen eines Pfades ist also die Wahrscheinlichkeit für einem Durchschnitt über Klassen der verschiedenen Stufen. Diese berechnet man gemäß

$$\begin{aligned} P(A_1 \cap ... \cap A_{n+1}) &= P(A_{n+1}|A_1 \cap ... \cap A_n) \cdot P(A_1 \cap ... \cap A_n) \\ &= P(A_{n+1}|A_1 \cap ... \cap A_n) \cdot P(A_n|A_1 \cap ... \cap A_{n-1}) \cdot P(A_1 \cap ... \cap A_{n-1}) \\ ... &= P(A_{n+1}|A_1 \cap ... \cap A_n)\, P(A_n|A_1 \cap ... \cap A_{n-1}) \cdot ... \cdot P(A_2|A_1) \cdot P(A_1) \end{aligned}$$

Nun schreibt man an die einzelnen Strecken des Pfades jeweils diese bedingten Wahrscheinlichkeiten an (und die eine nicht bedingte) und multipliziert über die am Pfad stehenden Zahlen, um die Wahrscheinlichkeit des Pfades zu bekommen. Je zwei verschiedene Pfade schließen einander aus. Um die Wahrscheinlichkeit eines Ereignisses zu bilden, das global für die Pfade formuliert ist (wie im Beispiel 'es wird eine rote und eine weiße Kugel gezogen'), hat man also die Wahrscheinlichkeiten der zugehörigen Pfade zu addieren. Im Beispiel hat man die zugehörigen Pfade

$$\begin{aligned} U_i, R, W, \, i &= 1, 2, 3, \\ U_i, W, R, \, i &= 1, 2, 3, \end{aligned}$$

Man hat

$$\begin{aligned} P\left(U_1 \cap R_1 \cap W_2\right) &= \frac{1}{3} \cdot \frac{2}{10} \cdot \frac{8}{9} = \frac{8}{135} \\ P\left(U_1 \cap W_1 \cap R_2\right) &= \frac{1}{3} \cdot \frac{8}{10} \cdot \frac{2}{9} = \frac{8}{135}, \end{aligned}$$

dasselbe in beiden Fällen (!), dann weiter:

$$\begin{aligned} P\left(U_2 \cap R_1 \cap W_2\right) &= P\left(U_2 \cap W_1 \cap R_2\right) = \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{5}{9} = \frac{5}{54}, \\ P\left(U_3 \cap R_1 \cap W_2\right) &= P\left(U_3 \cap W_1 \cap R_2\right) = \frac{1}{3} \cdot \frac{7}{10} \cdot \frac{3}{9} = \frac{7}{90}, \end{aligned}$$

summiert über alle diese Pfade kommt also:

$$\begin{aligned} P\left(\text{'eine weiße und eine rote Kugel'}\right) &= \frac{16}{135} + \frac{5}{27} + \frac{7}{45} \\ &= \frac{62}{135} \approx 0.46. \end{aligned}$$

Bemerkungen: Man sollte einfacher schreiben (U_i, R, W) statt $U_i \cap R_1 \cap W_2$ usw. Dann braucht man keine Indizes für fortgesetzte Verzweigungen nach 'weiß'-'rot'. Es sollte klar sein, dass man auf diese Weise auch fortgesetztes Ziehen aus einer Urne ohne Zurücklegen behandeln kann, oder auch die Vorgänge beim Würfeln, das etwa mit vier Würfeln, von denen beliebig viele auf einmal geworfen werden können, strategisch mit größtmöglicher Wahrscheinlichkeit auf einen Viererpasch hinauslaufen soll.

6. Ergänzung: Der Begriff der Unabhängigkeit eines Systems von n Ereignissen $A_1, ...A_n$

Vorab das Wichtigste: $P\left(A_1 \cap ... \cap \ A_n\right) = P\left(A_1\right) \cdot ... \cdot P\left(A_n\right)$ reicht nicht als Bedingung! Sie sollten Analoges von der Vektorrechnung kennen: Ein System von n Vektoren ist noch lange nicht linear unabhängig, wenn die Vektoren paarweise linear unabhängig sind, d.h. keiner ist ein Vielfaches des anderen. Beispiel: Die Vektoren $(1,1), (1,2), (1,3)$ sind paarweise linear unabhängig, keiner parallel zum anderen, aber das System der drei Vektoren ist linear abhängig, weil einer (sogar jeder) sich als Linearkombination der anderen schreiben lässt. Nun sollte man sinngemäß auch bei der wahrscheinlichkeitstheoretischen Unabhängigkeit von Ereignissen verlangen, dass bei einem unabhängigen Mengensystem insbesondere alle Teilsysteme der Länge ≥ 2 unabhängig sind. Dafür genügt die Bedingung $P\left(A_1 \cap ... \cap \ A_n\right) = P\left(A_1\right) \cdot ... \cdot P\left(A_n\right)$ nicht.

Wir konstruieren ein Beispiel beim Experiment 'Drei mal Würfeln'. Dazu legen wir A_1 und A_2 so an, dass sie nicht unabhängig sind, z.B. so: A_1: 'Der erste Wurf bringt eine Eins'. Dann A_2: 'Die Augensumme der ersten beiden Würfe ist höchstens Drei'. Offenbar $P\left(A_1\right) = 1/6$, $P\left(A_2\right) = 1/18$, aber $P\left(A_2|A_1\right) \cdot P\left(A_1\right) = 1/3 \cdot 1/6 \neq P\left(A_1\right) \cdot P\left(A_2\right)$.Nun suchen wir A_3 derart, dass $P\left(A_1 \cap A_2 \cap \ A_3\right) = P\left(A_1\right) \cdot P\left(A_2\right) \cdot P\left(A_3\right)$. Dafür könnten wir ganz banal nehmen: $A_3 = \emptyset$. Das *System* A_1, A_2, A_3 sollte nun sicher nicht als unabhängig gelten, weil A_1, A_2 bereits abhängig sind. Man kann auch weniger triviale Beispiele bilden, vgl. eine entsprechende Übung.

Damit lautet die adäquate Formulierung für die Unabhängigkeit von n Ereignissen $A_1, ..., A_n$ für beliebige Zahl $n \geq 3$:

DEFINITION 5. *Die Ereignisse $A_1, ..., A_n$ $(n \geq 3)$ sind unabhängig genau dann, wenn*

$A_{i_1}, ..., A_{i_{n-1}}$ stets unabhängig sind für $i_1, ..., i_{n-1}$ paarweise verschieden, und

$P\left(A_1 \cap ... \cap \ A_n\right) \ = \ P\left(A_1\right) \cdot ... \cdot P\left(A_n\right)$.

Beispiel: Bei dreimaligem Würfeln sind die Ereignisse A: 'Der erste Wurf zeigt eine gerade Zahl', B: 'Der zweite Wurf zeigt eine Zahl unter 5', C: 'Der dritte Wurf ergibt eine Sechs' sichtlich unabhängig.

Bemerkung: Es können auch drei Ereignisse A_1, A_2, A_3 paarweise unabhängig sein, also stets A_i, A_j ein unabhängiges Paar bilden für $i \neq j$, ohne dass A_1, A_2, A_3 ein unabhängiges Tripel bilden. Dazu kommt auch ein Beispiel in den Übungen. (Dies Phänomen kennen Sie auch aus der Vektorrechnung mit der linearen Unabhängigkeit.)

KAPITEL 2

Zufallsvariablen und ihre Verteilungen: Einführung unter deskriptivem Aspekt

1. Begriff der Zufallsvariablen

Diese Worte bezeichnen dasselbe, manchmal sagt man auch nur 'Variable', wenn schon klar ist, dass man im Zusammenhang der Statistik bzw. Wahrscheinlichkeitstheorie spricht. Wenn wir an den Gebrauch von 'Variable' bei Funktionen anknüpfen, so haben wir es hier mit abhängiger Variablen zu tun. Genauer also: Eine Zufallsvariable ist ein Abbildung, und diese ist jeweils völlig präzise definiert. Was ist dann das 'Zufällige' daran? Nun, man befindet sich im Rahmen eines Wahrscheinlichkeitsraums $(\Omega, \mathcal{S}, P)$ und hat nun eine Abbildung

$$X : \Omega \to \mathbb{R}.$$

Man denkt nun daran, das zugehörige Zufallsexperiment werde durchgeführt. Dann kommt ein zufälliges Ergebnis $\omega \in \Omega$ heraus. An diesem wird nun der Wert $X(\omega)$ beobachtet. Dieser ist nun zufällig, in dem Maße, wie zufällig ω herauskam. Nun darf insbesondere im Fall $\Omega = \mathbb{R}$ die Abbildung X keine *ganz* beliebige sein, man möchte bei Zufallsvariablen fragen können, wofür man sich typisch interessiert: 'Mit welcher Wahrscheinlichkeit kommt ein $X-$ Wert unter 150 heraus?', usw. Man verlangt also, dass folgende Ereignisse jedenfalls in $\mathcal{S}$ liegen:

$$\begin{aligned} X \quad &\leq \quad \alpha := \{\omega \in \Omega | \, X(\omega) \leq \alpha\}, \text{ analog} \\ \alpha \quad &\leq \quad X, \ \alpha \leq X \leq \beta, \text{ überall acuh mit } < \text{ statt } \leq . \end{aligned}$$

Achtung: Ein Ausdruck wie '$X < \alpha$' ist eigentlich Unsinn. Aber hier ist genau erklärt, als welche Menge in der Ereignisalgebra er zu nehmen ist. Umgekehrt lenkt der exakte Ausdruck für die Menge durch Vieles, das immer gleich bleibt, vom Wesentlichen ab - nämlich von den Grenzen. Außerdem haben wir bereits am Anfang gesehen, dass Ereignisse zunächst ganz natürlich durch Aussagen informell beschrieben werden. Das macht man gern in praktischer Anwendung. Die kurzen Ausdrücke wie '$X < \alpha$' knüpfen da an, verkürzen nur symbolisch die Aussagen, machen sie aber direkt verständlich, genauer formuliert wäre '$X < \alpha$' in Worten: 'Das im Experiment herauskommende $\omega \in \Omega$ hat die Eigenschaft, dass $X(\omega) < \alpha$.' Oder ein wenig kürzer: 'Es kommt ein $X-$ Wert $< \alpha$ heraus'. Man sieht, dass sich die Beziehungen wie $<$ auf $X(\omega)$ beziehen und nicht auf X. Im sonstigen mathematischen Sprachgebrauch wäre '$X < \alpha$' gerade so zu verstehen, dass alle Elemente von Bild(X) diese Eigenschaft haben. Das ist nun hier gerade nicht gemeint.

Wir werden zunächst im Wesentlichen mit endlichem Ω und $\mathcal{S} = \mathcal{P}(\Omega)$ arbeiten, und dann entfällt die genannte Bedingung der $P-$ Messbarkeit der Mengen wie $X < \alpha$ usw. In diesem einfachen Fall ist *jede Abbildung* $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Insbesondere sind dann auch stets die Ereignisse $X = \alpha$, für eine Zahl $\alpha \in \mathbb{R}$, von Interesse. Wir fassen zusammen:

DEFINITION 6. *Es sei ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{S}, P)$ gegeben. Dann heißt eine Abbildung*

$$X : \Omega \to \mathbb{R}$$

eine Zufallsvariable (Zufallsgröße, zufällige Größe), wenn alle Ereignisse der Form $X \in I$ mit einem Intervall I in $\mathcal{S}$ liegen, also $P(X \in I)$ existiert. Im Falle eines endlichen Ω und $\mathcal{S} = \mathcal{P}(\Omega)$ ist jede Abbildung $X : \Omega \to \mathbb{R}$ eine Zufallsvariable.

Bemerkung: Im praktischen Gebrauch gibt man gern bequem eine verbale Beschreibung, die nur die Ermittlung des Wertes bei gegebenem Ausgang aus Ω kennzeichnet, also keine formale mathematische Definition. Wir werden das in den folgenden Beispielen in typischer Weise ausführen.

Beispiele:

1.) Für $A \in \mathcal{S}$ ist die Größe

$$X_A(\omega) = \begin{cases} 1, & \text{wenn } \omega \in A, \\ 0, & \text{wenn } \omega \notin A \end{cases}$$

eine Zufallsvariable.

Aus diesen baut man weitere wichtige zusammen, z.B.

2.) Es wird n mal gewürfelt. Dann ist die Zufallsvariable 'X = Anzahl der Sechsen' (bequeme Beschreibung!) so formal zu definieren:

$$X((a_1, ..., a_n)) := |\{k \in \mathbb{N} | \, 1 \leq k \leq n \text{ und } a_k = 6\}|, \text{ für } a_1, ..., a_n \in \{1, 2, 3, 4, 5, 6\}.$$

Das Ereignis $X < 4$ bedeutet dann: 'Es werden bei den n Würfen höchstens 3 Sechsen gewürfelt'.

3.) Entsprechend ist die Zufallsvariable 'X = Augensumme bei n Würfen mit einem Würfel' formaler zu schreiben:

$$X((a_1, ..., a_n)) := \sum_{k=1}^{n} a_k, \text{ für } a_1, ..., a_n \in \{1, 2, 3, 4, 5, 6\}.$$

Im Folgenden wird es wichtig, typische Bildungen wie hier im 2. und 3. Beispiel aufzufassen als zusammengesetzt aus elementaren Zufallsvariablen, z.B. für das 2. Beispiel:

$$X_k((a_1, ..., a_n)) := \begin{cases} 1, & \text{wenn } a_k = 6 \\ 0 & \text{sonst} \end{cases}, \text{ für } 1 \leq k \leq n.$$

Dann hat man für X im 2. Beispiel: $X = \sum_{k=1}^{n} X_k$. Analog wird man im 3. Beispiel X ebenso als Summe von Zufallsvariablen X_k auffassen, nur diesmal mit $X_k :=$ 'Augenzahl im k. Wurf', also $X_k((a_1, ..., a_n)) := a_k$.

4.) Im Alltagsleben werden ständig Zufallsvariablen wie 'durchschnittliches Nettoeinkommen des letzten Jahres', 'Anzahl der Schuljahre' usw. betrachtet. Dann hat man eine endliche Population Ω spezifiziert und denkt an das Zufallsexperiment: 'Jemand wird zufällig aus Ω gezogen'. Allerdings besitzt man dann oft auch nur die Daten einer Stichprobe und stellt diese graphisch und mit ein paar Kennwerten dar. Daher hat man viele Beispiele für das, was man 'deskriptive Statistik' nennt.

2. Begriff der Verteilung einer Zufallsvariablen und der Verteilungsfunktion

Unter der Verteilung einer empirischen Zufallsvariablen X versteht man zunächst noch etwas unscharf: Irgendwie wird mitgeteilt, wie viele Werte von X in welchem Bereich liegen, oder auch, welche Werte von X welche relativen Häufigkeiten haben. Man interessiert sich also nicht mehr dafür, welches Objekt $\omega \in \Omega$ welchen $X-$ Wert $X(\omega)$ bekommt, sondern nur noch dafür, welche Werte wie häufig sind. Man fertigt dann eine Tabelle der Häufigkeiten an - es können absolute Häufigkeiten, relative oder prozentuale sein. Wenn eine Variable nur wenige Werte annimmt, wird man zu jedem einzelnen $X-$ Wert dessen Häufigkeit angeben. Wenn eine Variable dagegen sehr viele (sehr dicht liegende) Werte annimmt, so wird man den Bereich der vorkommenden Werte in Intervalle einteilen und die Häufigkeiten nur für diese Intervalle zählen. (Solche Dinge werden mit den Mitteln der 'deskriptiven Statistik' beschrieben, vgl. den nächsten Abschnitt.)

Aber auch Zufallsvariablen $X : \Omega \to \mathbb{R}$ auf der Ergebnismenge eines allgemeinen Wahrscheinlichkeitsraums $(\Omega, \mathcal{S}, P)$ die mathematisch exakt definiert sind, haben eine Verteilung. Das ist dann sinngemäß eine Wahrscheinlichkeitsverteilung. Dabei hat man diskrete oder kontinuierliche Verteilungen (auch Mischformen sind denkbar). Im diskreten Fall hat man die Möglichkeit, jedem Wert von X seine Wahrscheinlichkeit zuzuordnen. Das kann man dann in einer Tabelle festhalten, oder auch mit einer Formel, also z.B.:

$X-$ Wert a	a_1	a_2	...	a_n
$P(X = a)$	$P(X = a_1)$	$P(X = a_2)$	...	$P(X = a_n)$

Konkretes Beispiel: X = Augensumme bei zweimaligem Würfeln, dann hat man

$X-$ Wert a	2	3	4	5	6	7	8	9	10	11	12
$P(X = a)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

.

Eine solche Tabelle wird man dann gern auch graphisch darstellen, vgl. den nächsten Abschnitt. In vielen systematisch wichtigen Beispielen hat man auch (das ist bei einer Vielzahl von $X-$ Werten oder gar bei abzählbar unendlich vielen) eine Formel, welche die Funktion $a \mapsto P(X = a)$ genau beschreibt durch

einen Rechenausdruck in a. So werden wir für $X =$ Anzahl der Sechsen bei n Würfel-Würfen bekommen: $P(X=k) = \binom{n}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{n-k}$, für alle $k = 0, 1, ..., n$.

Wie beschreibt man kontinuierliche Verteilungen oder auch vergröbernd solche mit sehr dicht liegenden Werten (auch empirische)? Wieder hat man den Gegensatz zwischen mathematisch definierten Verteilungen dieser Art oder empirischen. Für beide gibt es einmal (vergröbernd) die Möglichkeit des Histogramms: Man teilt die Werte im interessierenden Bereich in Intervalle ein, ordnet dann jedem Intervall die zugehörige Wahrscheinlichkeit bzw. empirische relative Häufigkeit zu. Das gibt dann eine Tabelle wie

Intervall I	$[a_0, a_1]$	$(a_1, a_2]$	...	$(a_{n-1}, a_n]$
$P(X \in I)$	p_1	p_2	...	p_3

.

Wieder gehört eine graphische Darstellung dazu (Histogramm, vgl. den nächsten Abschnitt).

Ein sehr nützliches Mittel, eine Verteilung zu beschreiben, ist die Verteilungsfunktion. Sie funktioniert in jedem Fall - ob eine Zufallsvariable nun wenige oder endlich viele Werte in großer Anzahl oder ein Kontinuum von Werten annimmt:

DEFINITION 7 (Verteilungsfunktion). *Die Verteilungsfunktion (manchmal auch 'kumulierte Verteilung' genannt) einer Zufallsvariablen X bezeichnet man gern mit F_X. Sie ist definiert:*

$$\begin{array}{rccc} F_X: & \mathbb{R} & \to & [0,1] \\ & a & \mapsto & P(X \le a) \end{array}.$$

Bemerkungen: Die zuvor genannten Darstellungen von Verteilungen nennt man auch 'nichtkumulativ'. Diese heißt verständlich 'kumulativ', weil stets alle Wahrscheinlichkeiten für die Werte bis a 'aufgesammelt' werden. Der Wert $P(X \le a)$ kann dabei als endliche Summe oder auch als Integral über eine 'Dichtefunktion' berechnet werden, oder man hat vergröbert nur für endlich viele Werte $a_1, ..., a_n$ den Wert $P(X \le a_i)$ ermittelt.

Beispiele:

1.) Für die Variable $X =$ Augensumme bei zwei Würfel-Würfen wird F_X durch folgende Tabelle vollständig beschrieben:

$X-$ Wert a	2	3	4	5	6	7	8	9	10	11	12
$F_X(a) = P(X \le a)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36

.

Dabei versteht man mit: Werte unter 2 haben Wahrscheinlichkeit Null, also $F_X(a) = 0$ für $a < 2$. $F_X(a) = F_X(k)$, wenn k eine der möglichen Augensummen ist und $k \le a < k+1$. $F_X(a) = 1$, wenn $a \ge 12$. Klar gilt für solche diskreten Verteilungen stets:

$$F_X(a) = \sum_{b \le a} P(X = b).$$

So entstand diese Tabelle aus der oben angeführten 'nicht kumulativen'.

2.) Für die Variable $X =$ Zufallszahl aus $[0,1]$ mit der Dichtefunktion

$$f(x) = \begin{cases} 1 \text{ für } x \in [0,1] \\ 0 \text{ sonst} \end{cases}$$

hat man

$$F_X(a) = \int_{-\infty}^{a} f(x) = \begin{cases} 0 \text{ für } a < 0 \\ \int_0^a 1 \cdot dx = a \text{ für } 0 \le a \le 1 \\ 1 \text{ für } a > 1. \end{cases}$$

Folgende Eigenschaften sind selbstverständlich für alle Verteilungsfunktionen:

SATZ 2. *Jede Verteilungsfunktion F hat nur Werte ≥ 0, und sie ist monoton steigend im nichtstrengen Sinn, also*

$$a \le b \implies F(a) \le F(b).$$

Ferner ist sie halbstetig nach unten, d.h. $\lim_{x \to a+} F(x) = F(a)$.

Bemerkungen:

1.) Manchmal wird die Verteilungsfunktion definiert durch $P(X < a)$. Dann hat man Halbstetigkeit nach oben, nicht mehr nach unten.

2.) $a < b \implies F(a) < F(b)$ gilt im Allgemeinen nicht. Diese Eigenschaft ist für das zweite Beispiel oben nur richtig für $a, b \in [0,1]$, für das erste ist sie radikal falsch.

Abschließend wollen wir zeigen, dass F_X für jede Verteilung brauchbar ist: Wenn X diskret ist mit Werten

$$a_i, i \in \mathbb{N}, \text{ so dass } a_i < a_j \text{ für } i < j,$$

dann hat man

$$\begin{aligned} F_X(b) &= 0, \text{ wenn } b \neq a_i \text{ für alle } i, \\ P(X = a_1) &= F_X(a_1), \\ P(X = a_{i+1}) &= F_X(a_{i+1}) - F_X(a_i). \end{aligned}$$

Die Verteilungsfunktion lässt also die nichtkumulierte Verteilung erschließen.

Wenn X mit Dichte f verteilt ist, f stückweise stetig, so erhält man f aus F_X mit $F_X' = f$, bis auf die in den interessierenden Beispielen nur endlich vielen Stellen, an denen F_X nicht differenzierbar ist. Dort kann man den Wert von f beliebig festlegen. Das praktisch Wichtigste: Man hat $F_X(b) - F_X(a) = P(a < X \leq b)$, was wiederum dasselbe ist wie $P(a \leq X \leq b)$ oder auch $P(a \leq X < b)$, da jede nur aus einem Punkt bestehende Menge die Wahrscheinlichkeit Null bekommt.

3. Deskriptive Statistik

Es geht hier einmal um graphische Illustration von empirischen Häufigkeitstabellen, aber auch um die Angabe von gewissen beschreibenden Zahlenwerten, welche Information geben, die immer wieder gefragt ist. Natürlich sind dieselben Elemente auch nützlich zur Beschreibung mathematisch definierter Verteilungen. Insbesondere bekommt Beides im Verein einen Sinn, wenn man auf die Idee kommt, eine empirische Verteilung durch eine mathematische näherungsweise zu beschreiben, also ein mathematisches Modell der ersten zu bilden.

3.1. Die Idee der statistischen Abstraktion. In allen erdenklichen 'Statistiken' des Bundes, der Länder, der Versicherungen usw. hat man mit großen endlichen Populationen Ω zu tun und einer Abbildung $X : \Omega \to \mathbb{R}$. Zum Beispiel ordnet man jeder Wohnung in NRW ihre Quadratmeterzahl zu, jedem Menschen in NRW sein Alter usw. Dabei sagt man gewöhnlich, wie genau man die Zahlangaben nimmt, z.B. in den erwähnten Beispielen: 'Wohnfläche in ganzen [m^2]', 'Alter in vollendeten ganzen Lebensjahren'. Die volle Information über X könnte man nur geben mit einer fürchterlich langen Tabelle der Form

ω_1	$X(\omega_1)$
ω_2	$X(\omega_2)$
$\vdots$	$\vdots$
ω_n	$X(\omega_n)$

.

Damit könnte man wenig anfangen, das ist völlig unhandlich.

Hier setzt die statistische Abstraktion ein: Man konstruiert eine Abbildung, die jedem Wert von X dessen relative Häufigkeit zuordnet. Man will also nicht mehr wissen, welche $X-$ Werte die einzelnen Populationsmitglieder haben, sondern nur noch, wie häufig diese Werte vorkommen. Das führt zu einer Tabelle der Form

a	$P(X = a)$
a_1	$P(X = a_1)$
$\vdots$	$\vdots$
a_s	$P(X = a_s)$

.

Wenn s überschaubar klein ist, dann ist diese Tabelle schon ziemlich übersichtlich. Aber in vielen Beispielen ist das nicht der Fall, so auch bei 'Wohnfläche', 'Alter', zusätzlich kann man daran denken, dass man im Prinzip die Werte auch feiner messen könnte. Dann ist auch diese Abstraktion noch nicht ausreichend,

und man teilt die Werte besser in (nicht zu viele und nicht zu wenige) Intervalle ein und bekommt damit etwas Überschaubares, in Tabellenform:

a	$P(X=a)$
$[a_0,a_1]$	$P(a_0 \leq X \leq a_1)$
$(a_1,a_2]$	$P(a_1 < X \leq a_2)$
$\vdots$	$\vdots$
$(a_{m-1},a_m]$	$P(a_{m-1} < X \leq a_m)$

Diese Tabelle ist dann endgültig kurz genug. Zur Konkretisierung ein Beispiel - in solchen Fällen wird gewöhnlich nicht die relative Häufigkeit angegeben, sondern die prozentuale:

Wohnfläche $[m^2]$	prozentuale Häufigkeit
unter 40	4.7
$40-60$	17.9
$60-80$	25.7
$\vdots$	$\vdots$
$160-200$	4
über 200	2.4

(Das sind die tatsächlichen Zahlen für NRW 2014.) Aber so sah die Tabelle nicht aus - hier begegnet man einem Genauigkeitswahn, gepaart mit völligem mathematischen Unverstand: Man denkt daran, dass die Klassen disjunkt sein sollten und schreibt dann in diesem Beispiel (und analog auch in anderen): '$40-59$'. '$60-79$' usw. Natürlich wäre das bei Messung in ganzen Zahlen in Ordnung, aber eben nicht mehr bei etwas feinerer Messung. Gibt es etwa keine Wohnflächen zwischen 59 und 60 [m^2]? Zudem treten blöde Zahlen auf, die das Bild nicht übersichtlicher machen. Da ist man besser weniger pedantisch und schreibt wie in der Tabelle oben. Allenfalls kann man in Worten noch dazu sagen, dass $40-60$ heißt: über 40 und höchstens 60. Bei hinreichend feinen Messungen ist es völlig egal, ob ein Objekt, dessen Wert genau auf einer Klassengrenze liegt, zur unteren oder zur oberen gezählt wird, das ändert dann gar nichts.

Besser wäre es, daran zu denken, dass man für nach oben offene Intervalle wie das letzte im Beispiel dessen Mittelwert nennen sollte, damit man näherungsweise das Populations-Mittel ermitteln kann. (Alternativ könnte man auch dies angeben. Dasselbe gilt natürlich auch für nach unten offene Intervalle - im Beispiel aber ist das wenigstens mit dem Wert Null begrenzt.

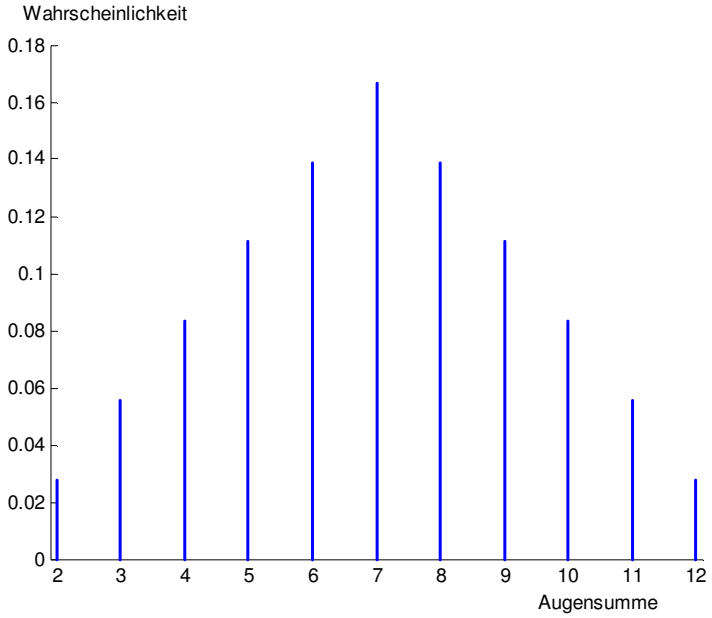
Andererseits begegnet man zumeist nur einer dürren Mittelwert-Angabe. Dann hat man keinerlei Information über den Verlauf der Verteilung, nicht einmal über deren grobe Form. Der Mittelwert allein ist fast immer unzureichend, wenigstens sollte man dann noch die Streuung angeben (s.u.). Im Falle einer einigermaßen 'normalverteilten' Variablen wären dann sogar die Häufigkeitsinformationen für die einzelnen Intervalle einigermaßen zu entnehmen, sonst aber nicht.

3.2. Graphische Darstellung von Verteilungen durch Stabdiagramm und Histogramm. Hat X nur wenige Werte, so empfiehlt es sich, eine Tabelle der Form (mit kleinem Wert von n, mit Computergraphik dürfen es schon hunderte sein)

$X-$ Wert a	a_1	a_2	...	a_n
$P(X=a)$	$P(X=a_1)$	$P(X=a_2)$	...	$P(X=a_n)$

mit einem Stabdiagramm darzustellen: An jeder Stelle a auf der Werte-Achse wird ein Stab senkrecht zu dieser Achse errichtet, dessen Länge proportional zu $P(X=a)$ ist. Das sieht dann im Beispiel der

genauen Wahrscheinlichkeitsverteilung für die Augensumme von zwei Würfel-Würfen so aus:



Bemerkung: Das Stabdiagramm ist nichts anderes als die gewöhnliche Darstellung von Funktionsgraphen, hier der Funktion $a \mapsto P(X = a)$. Lediglich malt man zur Verdeutlichung Stäbe statt der Punkte - alle anderen Werte liegen auf der Achse, sie sind Null. Man sollte sich um korrekte Beschriftung der Achsen bemühen (also Größenwerte mit Einheiten, 'relative Häufigkeit' (absolute, prozentuale) bei der vertikalen Achse, oder auch gegebenenfalls 'Wahrscheinlichkeit'.

Nehmen wir nun folgende empirische Verteilung der Variablen X mit folgende Tabelle (sogenannter 'gruppierter' Daten:

1.)

Intervall	$[1,2]$	$(2,3]$	$(3,4]$	$(4,5]$	$(5,6]$
relative Häufigkeit	0.1	0.2	0.4	0.2	0.1

Dieselbe Verteilung nur noch einmal vergröbert:

2.)

Intervall	$[1,2]$	$(2,3]$	$(3,4]$	$(4,6]$
relative Häufigkeit	0.1	0.2	0.4	0.3

Dann sind zwei Fälle zu unterscheiden:

1. Fall: Die Intervalle haben alle dieselbe Breite - so bei 1.). Dann kann man 'Histogramm'-Kästen über jedem Intervall errichten, deren Höhen proportional zu den Häufigkeiten der Tabelle sind - das bedeutet automatisch, dass auch die *Flächeninhalte* der Kästen proportional zu den Häufigkeiten der zugehörigen Intervalle sind. Da man Beides ablesen kann, ist es zweckmäßig, zwei vertikale Achsen mit den entsprechenden Beschriftungen anzubringen.

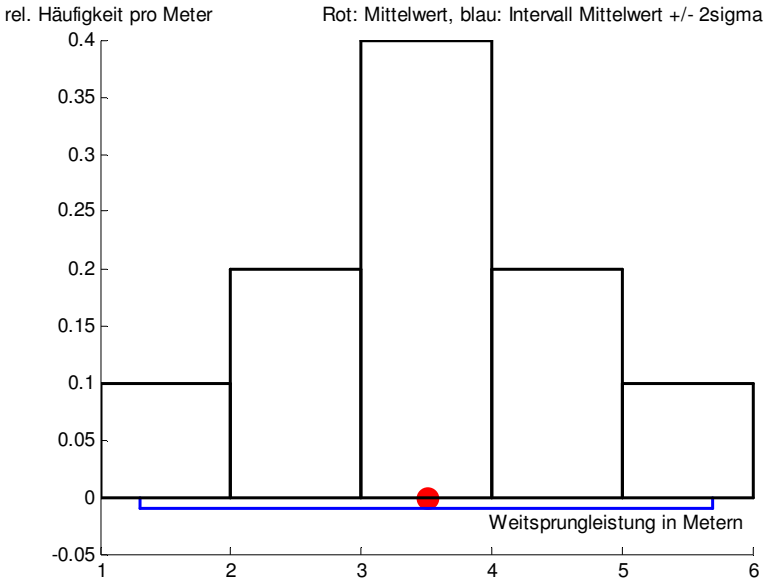
2. Fall (der oft vorkommt, so bei 2.)): Die Intervalle haben verschiedene Breiten. Dann *muss* man die Kastenhöhen ausrechenen mit

$$h_I = \frac{P(X \in I)}{\text{Breite von } I} = \text{ auf } I \text{ konstanter Dichtewert.}$$

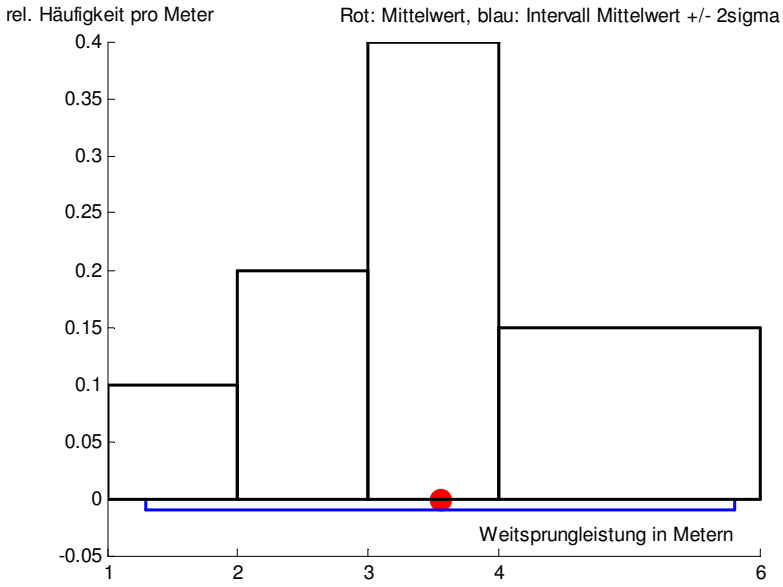
damit man korrekt auch intuitiv die Flächeninhalte als Wahrscheinlichkeiten deutet (!). Sonst kommt man auf eine üble Irreführung, und die Sache liest sich nicht mehr wie bei einer Dichtefunktion. Insbesondere muss die korrekte Beschriftung der vertikalen Achse sein: 'relative Häufigkeit pro Größeneinheit' bzw. 'Wahrscheinlichkeit pro Größeneinheit' - und nicht etwa 'relative Häufigkeit' oder 'Wahrscheinlichkeit'.

Hier sind die Beispiele 1.) und 2.) gezeichnet wie beschrieben:

Beispiel 1.):



Beispiel 2.):



Man beachte: Die Höhe des letzten Kastens ist nicht etwa 0.3, sondern 0.15, sonst erschiene er viel zu gewichtig.

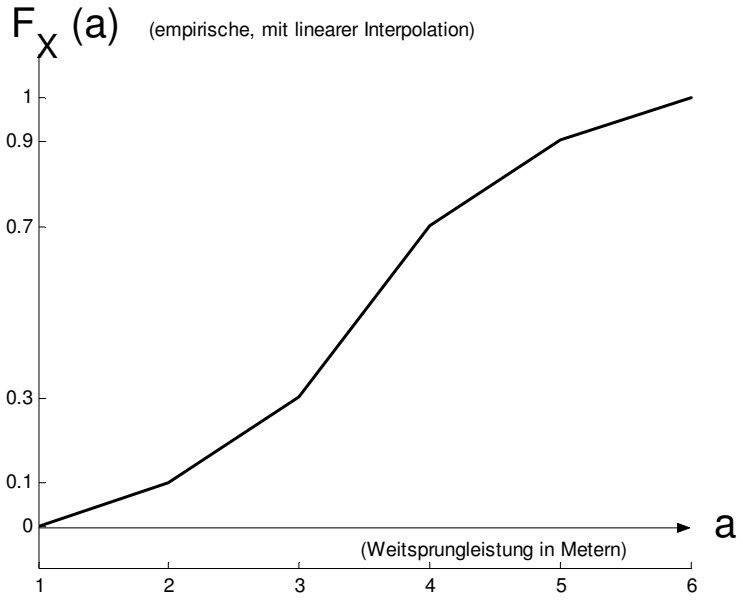
Empirische statistische Daten enthalten oft oben ein Intervall der Form '$> a$'. (Es könnte auch noch eines am linken Ende geben.) Das muss man dann aus der Zeichnung lassen und verbal hinzufügen. Allerdings ist es eine (verbreitete) Unsitte, dass für dies Intervall kein Mittelwert angegeben wird, mit der Folge, dass man auch den Mittelwert der in Frage stehenden Population oder Stichprobe gar nicht kennt.

3.3. Graphische Darstellung der Verteilungsfunktion. Diese Darstellung ist stets der Funktionsgraph von F_X. Wenn die nichtkumulierte Darstellung ein Stabdiagramm war, ist dieser Graph stückweise konstant, die jeweilige Höhe ist immer die der links davon stehenden Stäbe. Wenn die nichtkumulierte graphische Darstellung die eines Histogramms war, so malt man einen Polygonzug: Beim linken Ende des ersten Intervalls a_0 startet man mit dem Punkt $(a_0, 0)$. Diesen verbindet man mit dem Punkt $(a_1, P(a_0 \leq X \leq a_1))$, für das nächste Intervall verbindet man diesen Punkt mit $(a_2, P(a_0 \leq X \leq a_2))$, usw. (Noch einmal: In empirischen Beispielen steht 'relative Häufigkeit der Werte zwischen a_0 und a_2' statt $P(a_0 \leq X \leq a_2)$, usw. Bei a_n ist der Wert 1, von da ab auch weiter konstant 1, wenn man kein offenes letztes Intervall hat. Links von a_0 ist die Funktion als konstant Null zu verstehen. Einfach gesagt: Man verbindet die relevanten bekannten Graphenpunkte geradlinig miteinander, also 'linear interpolierend'. Es ergibt sich dann für jedes Intervall: Die Steigung des Polygonzuges auf diesem Intervall ist die Kastenhöhe des Histogramms = Dichtewert für den betreffenden Kasten

Beispiele:

1.) Für die Verteilungsfunktion zu einem Stabdiagramm vgl. das Beispiel im Unterabschnitt: 'Praktisches Anlegen einer Statistik'

2.) Zum Histogramm im ersten Beispiel oben (also zu stückweise konstanter Dichtefunktion):



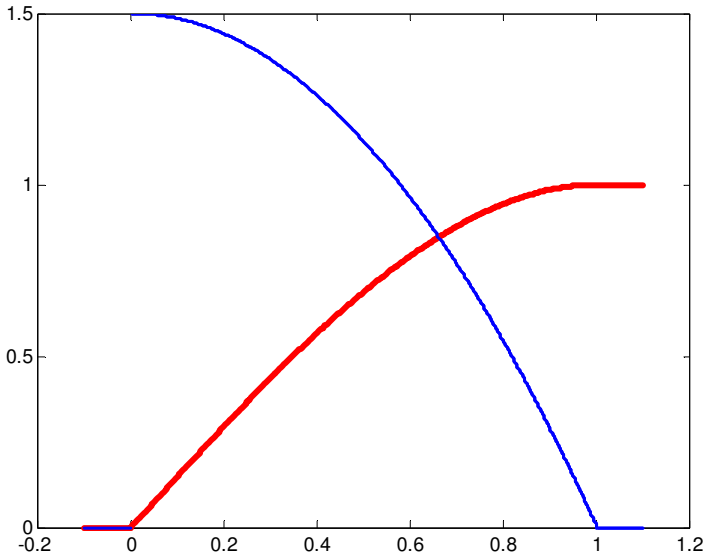
3.) Zur Dichtefunktion

$$f(x) = \begin{cases} \frac{3}{2}\left(1 - x^2\right) & \text{für } 0 \leq x \leq 1 \\ 0 \text{ sonst} \end{cases}$$

bekommt man

$$F_X(a) = \begin{cases} 0 \text{ für } a < 0, \\ \int_0^a \frac{3}{2}\left(1 - x^2\right) dx = \frac{3}{2}a - \frac{1}{2}a^3 \text{ für } 0 \leq x \leq 1, \\ 1 \text{ für } a > 1. \end{cases}$$

Hier ist ein Bild, das Dichte (blau) und Verteilungsfunktion (rot) zugleich zeigt:



Bemerkung: im Intervall $(-\infty, 0]$ sind sowohl Dichte als auch Verteilungsfunktion konstant Null.

3.4. Erwünschte beschreibende Zahlangaben: Verschiedene Arten von 'Mittelwert' und Standardabweichung. Die wichtigen mathematisch definierten Verteilungen sind durch sachlich passende Parameter beschrieben (einer oder mehrere, oft zwei). Natürlich gibt man diese stets an, wenn man mit einer solchen Verteilung arbeitet. Zu den wichtigsten dieser Verteilungen später mehr, z.B. gehören die Binomialverteilungen dazu und die Normalverteilungen.

Bei empirischen Verteilungen sind die (zur 'deskriptiven Statistik' gehörenden) Parameter stets: Arithmetisches Mittel und Streuung (in der Stichprobe als Gesamtheit betrachtet) oder eher die sogenannte 'Standardabweichung' (die davon zu unterscheiden ist). Zuweilen kann man urteilen, dass der Median angemessener ist als der arithmetische Mittelwert.

3.4.1. *Der arithmetische Mittelwert.* Dieser ersetzt alle beobachteten Zahlenwerte durch einen einzigen, beschreibt von der (empirischen) Verteilung den Punkt auf der Größenachse, an dem man das Verteilungsbild unterstützen muss, damit man das Bild im Gleichgewicht hält.

DEFINITION 8 (arithmetischer Mittelwert). *Hat man die Größenwerte $x_1, ..., x_n$ (Urliste, also mit möglichen Wiederholungen!) ermittelt in einer Stichprobe (des Umfangs n) von $X-$ Werten, so ist*

$$\overline{x}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Man nennt diesen Wert dann auch gern 'beobachtetes Stichprobenmittel'.

Bemerkungen:

1.) Meist nennt man n im Text und schreibt nur $\overline{x}$. Wenn die Variable Y heißt, so bezeichnet man entsprechend mit $\overline{y}$, usw.

2.) Wenn man nur gruppierte Daten hat und das zur Urliste passende $\overline{x}$ nicht benannt ist, kann man einen (bei hinreichend feiner Intervalleinteilung) als Näherungswert angeben:

$$\overline{x} \approx \sum_{k} m_k h_k,$$

wobei m_k der Mittelpunkt des k. Intervalls und h_k der zugehörigen Dichtewert ist.

3.) Für die Daten eines Stabdiagramms hat man mit den Werten a_i (ohne Wiederholung!) und zugehörigen relativen Häufigkeiten p_i:

$$\overline{x} = \sum_{i} a_i p_i.$$

(Das folgt aus der Definition von $\overline{x}$ (!)).

4.) $\overline{x}$ ist ein brauchbarer Schätzwert für $\mu(X)$, den Erwartungswert der Variablen X, der für eine empirische Variable X mit großer Population Ω einfach der arithmetische Mittelwert der Gesamtpopulation ist. Wie gut dieser Schätzwert ist, das hängt von n sowie von der später einzuführenden Streuung ab.

3.4.2. *Die Streuung innerhalb der Stichprobe und die Standardabweichung.* Diese bilden ein Maß dafür, wie breit die Verteilung ist. Man beachte: Es ist nicht sehr informativ, den größten und den kleinsten beobachteten Wert oder gar den kleinstmöglichen und größtmöglichen Wertzu nennen. Es könnte dabei immer noch sein, dass sich alle Werte in enger Umgebung von diesen beiden oder bei einem Wert dazwischen häufen. Der hier zu diskutierende Parameter soll gerade darüber nähere Auskunft geben.

DEFINITION 9. *Die Streuung der Variablen X innerhalb der Stichprobe von Werten x_i (Urliste!), $1 \leq i \leq n$, von X ist*

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} = \sqrt{\sum_i (a_i-\overline{x})^2 p_i}$$

(a_i Aufzählung der beobachteten Werte ohne Wiederholung, p_i die zugehörigen relativen Häufigkeiten.)

Bemerkung: Man nimmt also das arithmetische Mittel der quadratischen Abweichungen der einzelnen Werte Vom arithmetischen $X-$ Mittelwert, dann die Wurzel davon.

Nun möchte man aber häufig mit dem Streuungsparameter einer Stichprobe die Streuung in der Gesamtpopulation nähern, und dann nimmt man besser:

DEFINITION 10 (Standardabweichung). *Diese ist für eine Stichprobe $x_1, ..., x_n$ von $X-$ Werten ($n \geq 2$) definiert als*

$$s^{(n)}(X) := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

(Meist notiert man nur '$s(X)$'.

Diskussion: Für $n = 1$ kommt für die zuvor eingeführte 'Streuung innerhalb der Stichprobe' stets Null heraus, man hat also gar keinen Anhaltspunkt für die Streuung in der Gesamtpopulation. Außerdem wird auch bei $n > 1$ mit dem Vorfaktor $\frac{1}{n}$ die Streuung stets systematisch unterschätzt. Dagegen wird mit

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

das Quadrat der Streuung in der gesamten Population (auch 'Varianz') genannt, verzerrungsfrei geschätzt. Schließlich müsste es korrekt heißen: $s(X; x_1, ...x_n)$, da der Schätzwert von den Werten der konkreten Stichprobe abhängt und nicht nur von der Variablen X.

Manchmal wird der arithmetische Mittelwert von wenigen extremen Populationsmitgliedern dominiert oder gar von nur einem. Das ist vielfach unerwünscht. Z.B. könnte man durch eine Umfrage ermitteln lassen, wie hoch die Militärausgaben eines Landes sein sollten. Dann könnte ein einziger 'Militärkopf' durch Nennung einer unsinnig großen Zahl bestimmen, dass für andere Ausgaben nichts übrig bliebe. In solchen Fälle eignet sich besser der Median:

DEFINITION 11. *Für Verteilungen mit Dichte ist der Median der Verteilung m definiert durch $F_X(m) = \frac{1}{2}$. Im Fall einer diskret verteilten Variablen X definiert man den Median als eine Zahl (oder einen möglichst kleinen Bereich von Zahlen), für die gilt, dass der mittlere absolute Abstand aller Einzelwerte von dieser Zahl (diesen Zahlen) minimal wird.*

Bemerkung: Zuweilen findet man bei Statistik-Anwendern (völlig unsinnige!) Versuche, den Median doch in jedem Fall als exakten Zahlenwert auch bei diskreten Verteilungen zu definieren. Entscheidend ist der heilsame Effekt im Falle der angesprochenen Problematik: Wenn in einer Population von $10^4 + 1$ Menschen der eine 10^6 (...) Nettoeinkommen (pro ...) hat, die anderen 10^4 alle ein Nettoeinkommen von 10^3, dann ist der Median 10^3. Der eine verändert den Median der übrigen überhaupt nicht. Allgemein kann man einsehen, dass bei einer Aufzählung $x_1, x_2, ..x_n$ der einzelnen Werte (mit Wiederholungen) in aufsteigender Reihenfolge, also $x_i \leq x_{i+1}$, bei ungerader Bevölkerungszahl $n = 2m + 1$ der Median stets

den genauen Wert x_{m+1} besitzt, bei gerader Anzahl $n = 2m$ kann man nur sagen, der Median liege zwischen x_m und x_{m+1}.

Als Maß der Streuung in der gesamten Population wäre auch denkbar, den mittlern absoluten Abstand von μ statt der Wurzel des mittleren quadratischen Abstandes zu nehmen, zuweilen geschieht das auch. Oft macht das keinen wesentlichen Unterschied. Aber für theoretische Zwecke ist es besser, eine differenzierbare Funktion zu haben statt des nicht differenzierbaren Absolutbetrages.

3.5. Praktisches Anlegen einer 'Statistik'. Man legt eine Zufallsvariable X fest (oder mehrere), je nach Fragestellung. Dann bildet man eine Zufalls-Stichprobe (bei 'Totalerhebungen' nimmt man die gesamte Population). Das ergibt eine Urliste $x_1, x_2, ...x_n$ von Werten. (Bei einer weiteren Zufallsvariablen auf demselben Ω könnte das $y_1, ..., y_n$ sein, bei Vergleich mit Y auf einer anderen Population auch eine andere Zahl von Werten y_i.)

Dann bildet man die Folge der verschiedenen Werte $a_1, ..., a_k$ ohne Wiederholungen (wenn k relativ klein ist) und findet die absoluten und damit auch die relativen Häufigkeiten, mit welcher diese vorgekommen sind. Daraus kann man dann alle weiteren aus der Stichprobe resultierenden Verteilungsinformationen berechnen. Wenn k allzu groß ist, wird man in Intervalle gruppieren und die graphische Darstellung darauf beschränken, aber arithmetisches Mittel und Standardabweichung (letztere im Fall der Totalerhebung durch die tatsächliche Streuung zu ersetzen) genau angeben.

Graphische Darstellungen sind auch für theoretisch-mathematische Verteilungen wichtig, damit man sie sich besser vorstellen kann.

Ein Beispiel: Man hat 200 mal gewürfelt und findet die absoluten Häufigkeiten für die Augenzahl X - hier ist das mit dem Computer getan worden, es resultierten die absoluten Häufigkeiten:

$a =$ Wert von X	1	2	3	4	5	6
abs. Häufigkeit von a	34	37	25	35	34	35
relative Häufigkeit von a	34/200	37/200	25/200	35/200	34/200	35/200

Wir finden

$$\begin{aligned}\overline{x} &= 1 \cdot \frac{34}{200} + 2 \cdot \frac{37}{200} + 3 \cdot \frac{25}{200} + 4 \cdot \frac{35}{200} + 5 \cdot \frac{34}{200} + 6 \cdot \frac{35}{200} \\ &= \frac{1}{200}\left(1 \cdot 34 + 2 \cdot 37 + ...\right) \\ &= \frac{703}{200} \approx 3.515.\end{aligned}$$

Man sieht, dass der Erwartungswert 3.5 recht genau genähert ist. Wir finden weiter die Standardabweichung

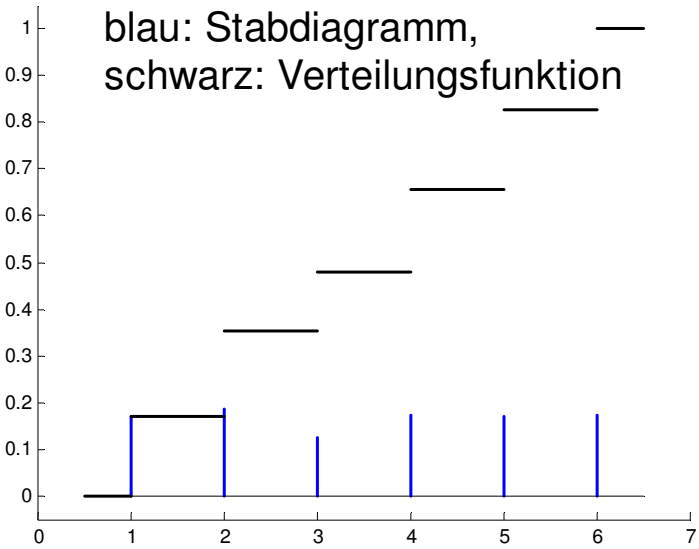
$$\begin{aligned}s\left(X\right) &= \sqrt{\frac{1}{199}\left(34 \cdot \left(1 - \frac{703}{200}\right)^2 + 37 \cdot \left(2 - \frac{703}{200}\right)^2 ... + 35 \cdot \left(6 - \frac{703}{200}\right)^2\right)} \\ &\approx 1.735.\end{aligned}$$

Die tatsächliche Streuung ist (unter Ausnutzung der Symmetrie):

$$\begin{aligned}\sigma\left(X\right) &= \sqrt{\frac{1}{3}\left(\left(1 - \frac{7}{2}\right)^2 + \left(2 - \frac{7}{2}\right)^2 + \left(3 - \frac{7}{2}\right)^2\right)} \\ &= \frac{1}{6}\sqrt{105} \approx 1.7078.\end{aligned}$$

$s\left(X\right)$ nähert also auch sehr gut $\sigma\left(X\right)$. Würfeln wir erneut 200 mal, so können die Verhältnisse auch anders sein, extremere Abweichungen von der Gleichverteilung können vorkommen, $\overline{x}$ und $s\left(X\right)$ werden anders ausfallen, auch weiter von den 'Sollwerten' abweichen. Wir zeigen die erwähnten Darstellungen der empirisch gefundenen Verteilung (die kumulierte ist nach links mit konstantem Wert 0 fortzusetzen,

nach rechts mit konstantem Wert 1):



Man erkennt kleine, aber deutliche Abweichungen von der Wahrscheinlichkeitsverteilung, bei welcher die Stäbe gleich lang und die Stufen gleich hoch wären.

KAPITEL 3

Mathematisches zu den wichtigsten Verteilungen und ein wenig Inferenzstatistik

1. Verknüpfung von Zufallsvariablen, Unabhängigkeit und Verteilungsparameter

Vorbemerkungen:

1.) Die Zusammensetzung von einfacheren Zufallsvariablen zu komplizierteren ein sehr wichtiger Vorgang, um eben etwas kompliziertere Probleme zu bewältigen. Einen völlig analogen Vorgang hatten wir bei der elementaren Wahrscheinlichkeitsrechnung: Dort wurden kompliziertere Ereignisse aus einfacheren zusammengesetzt und aus leicht zu berechnenden Wahrscheinlichkeiten solche zusammengesetzt, die ohne diese Reduktion kaum zu erhalten wären.

2. Man beachte auch, dass die gesamte elementare Wahrscheinlichkeitsrechnung beim Umgang mit Zufallsvariablen benutzt wird. Ein wichtiger Gesichtspunkt dieses Kapitels wie des folgenden ist dieser: Die technischen Mittel zur Bewältigung komplizierterer Fragen werden wesentlich erweitert, über das Zusammensetzen von Ereignissen mit Mengenoperationen hinaus. Man kann mit diesen Mitteln dann leicht bewerkstelligen, was ohne Theorie mit 'konkretem Gefummel' (trotzdem: dies behält seinen eigenen Wert!) nicht zu machen wäre.

3. Ein neuer Gesichtspunkt tritt auf: Viele systematisch bei Anwendungen wichtige Verteilungen werden komplett durch Parameter (ein bis zwei Zahlangaben sind das hier gewöhnlich nur, in Ausnahmen mehr) beschrieben. Die wichtigsten sind Erwartungswert und Streuung. Hat man diese allein, so kann man bei einem bekannten Verteilungstyp die gesamte Verteilung allein daraus gewinnen. Also wird es bedeutsam sein, diese Parameter für komplexere Zufallsvariablen aus denen für einfache berechnen zu können. Das wird in diesem Kapitel betrieben.

4. In diesem Kapitel wird das Allgemeine getan, das immer wieder benötigt wird, vor allem das Rechnen mit Erwartungswert und Streuung. Es kommen auch zwei wichtige Sätze. Aber es wird auch etwas Konkretes damit getan. Dann sollte klar sein, wie das nächste Kapitel über 'mathematische Verteilungstypen' anschließt: Damit bekommt man das Rüstzeug, die Kraft von 'Verteilungstyp kennen und rechnen mit Verteilungsparametern' auszunutzen.

1.1. Verknüpfungen von Zufallsvariablen. Wie man Funktionen $f, g : \mathbb{R} \to \mathbb{R}$ verknüpft durch Addition, Multiplikation mit einer Zahl, Multiplikation der Funktionen, usw., so verknüpft man auch Zufallsvariablen, welche auf demselben Ergebnisraum definiert sind, also

DEFINITION 12. *Seien $X, Y : \Omega \to \mathbb{R}$ Zufallsvariablen, dann sind es auch die folgenden:*

$$\begin{aligned}
(\alpha X)(\omega) &= \alpha(X(\omega)), \\
(X+Y)(\omega) &= X(\omega) + Y(\omega), \\
(XY)(\omega) &= X(\omega)\, Y(\omega), \\
\frac{X}{Y}(\omega) &= \frac{X(\omega)}{Y(\omega)}, \text{ wenn stets } Y(\omega) \neq 0.
\end{aligned}$$

Ferner kann man mit einer stetigen Funktion $f : Bild(X) \to \mathbb{R}$ auch bilden:

$$f \circ X(\omega) = f(X(\omega)).$$

Bemerkung: Besonders wichtig sind zunächst die ersten beiden Verknüpfungen, mit denen die Zufallsvariablen auf Ω einen Vektorraum bilden.

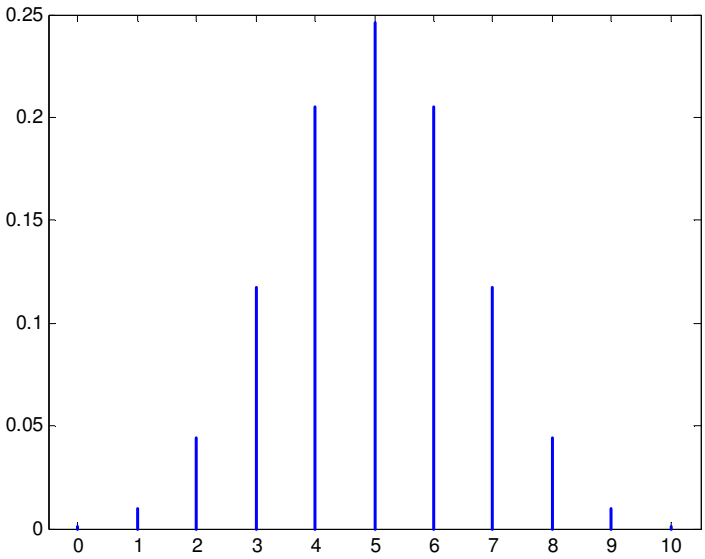
1.2. Die Verteilung von Zufallsvariablen. Diesen Abschnitt können Sie übergehen. Denn der Begriff der Verteilung tauchte zwangsläufig bereits im Kapitel der deskriptiven Statistik auf. Aber einmal lohnt sich eine Wiederholung dieser wichtigsten Grundbegriffe, zum anderen ist der Gesichtspunkt hier anders: Wir wollen auf die 'wahre' Verteilung hinaus, um von da aus nach Schätzungen zu sehen, im Kapitel der deskriptiven Statistik war es umgekehrt. Außerdem finden Sie hier ein paar mehr Bilder.

DEFINITION 13 (nichtkumulierte Verteilung bei diskreten Zufallsvariablen). *Sei Ω endlich oder höchstens abzählbar unendlich (dann heißt die Verteilung von X auf Ω diskret). Dann ist die (nichtkumulierte)* ***Verteilung von*** *X die folgende Abbildung:*

$$f_X(\alpha) := P(X = \alpha), \ \alpha \in \mathbb{R}.$$

Beispiel: Sei X = Augensumme bei zwei Würfen mit symmetrischem Würfel, dann ist $f_X(2) = \frac{1}{36}$, weiter ist $f_X(3.5) = f_X(50) = 0$ usw. (Viel mehr Beispiele im nächsten Abschnitt über die wichtigsten diskreten Wahrscheinlichkeitsverteilungen.)

Zum allgemeinen Verständnis: Die Verteilung von X ordnet also jeder reellen Zahl zu, mit welcher Wahrscheinlichkeit sie als $X-$ Wert vorkommt. Zur graphischen Darstellung: Man zeichnet sinnvoll Stäbe, deren Längen proportional zu den Werten sind, wie sonst Funktionsgraphen auch, nur, dass hier die Stäbe dazu da sind, dass man überhaupt etwas deutlich sieht (sonst hätte man nur einzelne Punkte) Das sieht dann z. B. so aus, hier im Beispiel X = Anzahl der 'Köpfe' bei 10 Münzwürfen:



Die Werte sind Wahrscheinlichkeiten, entsprechend ist die Ordinate zu beschriften. Es sollte klar sein, dass es keinen Sinn hat, die Stabenden miteinander zu verbinden. Abgesehen von den Stäben ist der Funktionswert der Wharscheinlichkeitsfunktion p_X überall Null.

Nun hat diese Beschreibung keinen Sinn bei kontinuierlichen Verteilungen: Für eine solche Verteilung gilt in der Regel, dass $P(X = \alpha) = 0$ für alle $\alpha \in \mathbb{R}$. Die so gebildete Funktion f_X wäre konstant Null und gäbe keinerlei Information! Dafür hat man nun eine Beschreibung der Verteilung, die *immer* funktioniert, bei jeder Zufallsvariablen (also auch bei den diskreten):

DEFINITION 14 (Verteilungsfunktion einer Zufallsvariablen). *Es sei X eine beliebige Zufallsvariable. Dann ist folgende Abbildung die* ***Verteilungsfunktion von*** *X (auch kumulierte Verteilung genannt):*

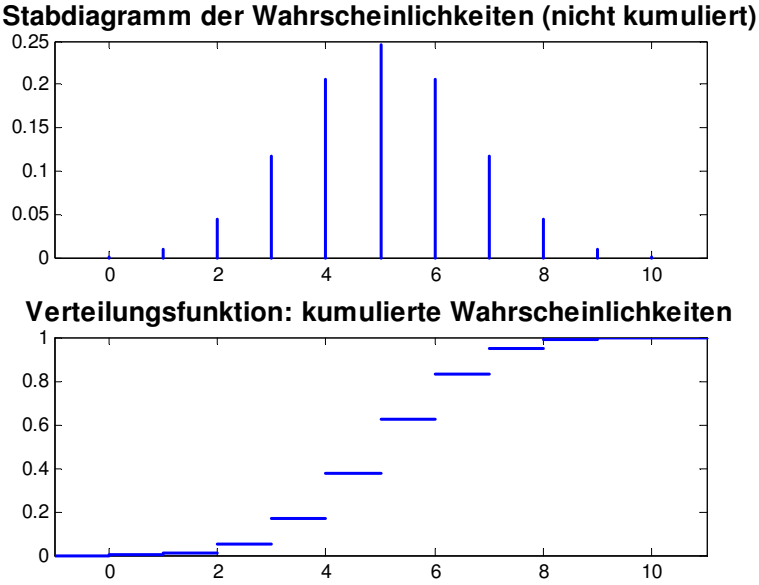
$$F_X(\alpha) := P(X \leq \alpha), \ \alpha \in \mathbb{R}.$$

Beispiele:

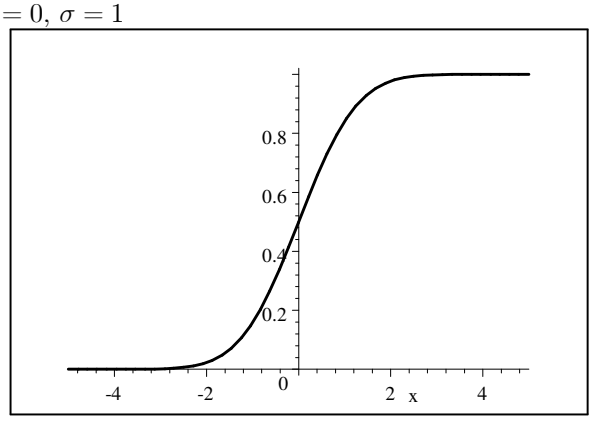
1.) Im Fall einer diskreten Verteilung mit Werten $x_1, ..., x_n$, welche mit den Wahrscheinlichkeiten p_i angenommen werden, $i = 1, ..., n$, ist die Verteilungsfunktion stückweise konstant, und man hat

$$F(x) = \sum_{\{\, i | x_i \leq x\}} p_i.$$

Im Beispiel der Variablen X = Anzahl der 'Köpfe' bei 10 Münzwürfen sieht das so aus, unteres Bild, hier noch einmal zum Vergleich das Stabdiagramm zu f_X:



2.) Im Fall einer Normalverteilung ist die Verteilungsfunktion differenzierbar und insbesondere stetig, sie sieht so aus, hier für $\mu = 0$, $\sigma = 1$



Wir fragen nach einem geeigneten nichtkumulierten Bild für eine stetige Verteilung, z.B. für eine Normalverteilung. Wir wir sahen, kann man f_X dann nicht brauchen, weil die Werte alle Null werden. Aber für kleine Intervalle der Form $[a, a + \varepsilon]$ hat man eine nicht verschwindende Wahrscheinlichkeit

$$P(a \leq X \leq a + \varepsilon),$$

und wenn man diese durch ε teilt, so bekommt man im Limes $\varepsilon \to 0$ gerade die Ableitung von F_X an der Stelle a. Diese ist dann als Wahrscheinlichkeitsdichte für X zu deuten, sie gibt an jeder Stelle a den Zuwachs der Wahrscheinlichkeit pro $X-$ Werte-Differenz. Wir definieren daher:

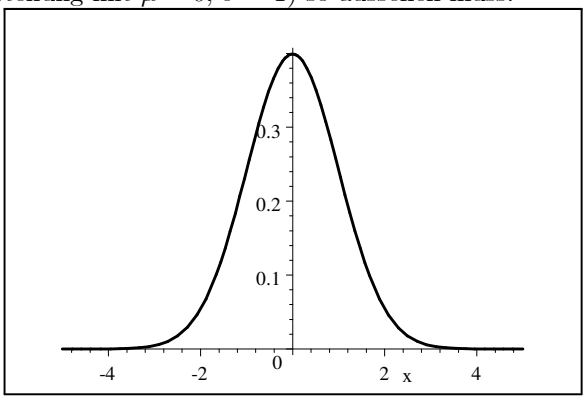
DEFINITION 15. *Eine Zufallsariable X mit stückweise differenzierbarer Verteilungsfunktion F_X hat die Wahrscheinlichkeitsdichte*

$$f_X(a) = F_X'(a).$$

(An den (höchstens abzählbar unendlich vielen) Stellen b, an denen F_X nicht ableitbar ist, kann der Wert $f_X(b)$ beliebig festgelegt werden, er ist irrelevant.) Umgekehrt sagt man, dass die Variable X mit der Dichte f verteilt sei, wenn für alle $a \in \mathbb{R}$ gilt:

$$P(X \leq a) = \int_{-\infty}^{a} f(x)\, dx.$$

Zum Verständnis: Wenn X nach dieser Definition mit der Dichte f verteilt ist, so folgt mit dem Hauptsatz der Differential- und Integralrechnung sofort, dass $F'_X(a) = \frac{d}{da}\int_{-\infty}^{a} f(x)\, dx = f(a)$. Die Wahrscheinlichkeitsdichte ist also stets die Ableitung der Verteilungsfunktion. Da F_X monoton steigt, ist diese Ableitung stets ≥ 0. Man kann nun auch anschaulich verstehen, dass die Normalverteilungsdichte (hier wieder zur Normalverteilung mit $\mu = 0$, $\sigma = 1$) so aussehen muss:



Ihre Werte sind keine Wahrscheinlichkeiten, sondern eben Wahrscheinlichkeitsdichten, welche angeben, wie stark die Verteilungsfunktion jeweils lokal ansteigt. Man kann also nicht etwa auf der Ordinate Wahrscheinlichkeiten ablesen. Aber man findet Wahrscheinlichkeiten wieder, mit der Flächendeutung des Integrals: Die Wahrscheinlichkeit $P(a \leq X \leq b)$ ist gerade der Flächeninhalt, der zwischen $x_1 = a$ und $x_2 = b$ vom Graphen der Dichtefunktion mit der $x-$ Achse eingeschlossen wird. Insbesondere ist $F_X(a)$ graphisch abzulesen als der Flächeninhalt der Fläche unter der Dichtekurve links von a. Man beachte, dass bei stetigen Verteilungen stets gilt: $P(a \leq X \leq b) = P(a < X < b)$, da auf den einzelnen Punkt nur die Wahrscheinlichkeit Null entfällt.

1.3. Der Begriff der Unabhängigkeit von Zufallsvariablen. Für die Berechnung von Wahrscheinlichkeiten für Ereignisse, die im Zusammenhang mit Zufallsvariablen formuliert werden, auch für die Berechnung von Erwartungswerten und Varianzen (vgl. den nächsten Abschnitt) spielt es häufig eine Rolle, ob zwei Variablen X, Y unabhängig sind oder nicht. Unabhängigkeit bedeutet: Kennt man den Wert der einen Variablen, so bleiben die Wahrscheinlichkeiten für die Werte der anderen unverändert. Oft ist es aus der Situation klar, dass zwei Variablen unabhängig sind. Beispiel: Man würfelt 10 mal, X = Augenzahl im dritten Wurf, Y = Augenzahl im fünften Wurf, diese Variablen sind sicher unabhängig. Ebenso versteht man sofort, dass die Augensumme aller 10 Würfe sicher nicht unabhängig ist vom Resultat des ersten Wurfs (oder auch des letzten). Man sollte hier aber stets bedenken, dass 'Abhängigkeit' im wahrscheinlichkeitstheoretischen Sinn deutlich weniger besagt als Abhängigkeit im Sinne einer Funktion. Wenn x bestimmt ist, dann ist $f(x)$ damit eindeutig bestimmt, das ist die stärkste Form der Abhängigkeit. Wir definieren nun den statistischen Begriff der Abhängigkeit:

DEFINITION 16. *Die Zufallsvariablen $X_1, ..., X_n$ sind genau dann unabhängig, wenn für alle $a_1, ..., a_n \in \mathbb{R}$ das System der Ereignisse*

$$X \leq a_1, ..., X_n \leq a_n \text{ unabhängig sind.}$$

Bemerkung: Es selbst im häufigsten Fall von nur zwei Zufallsvariablen weniger darum, diese Definition rechnerisch in Beispielen nachzuprüfen, als vielmehr darum, eine Unabhängigkeit per Konstruktion herbeizuführen und die Vereinfachungen zu nutzen, die sich daraus ergeben. Das wichtigste

Beispiel: Man führt das Zufallsexperiment, bei dem ein Wert von X beobachtet wird, n mal aus. Dann sei X_i der $X-$ Wert, der im i. Versuch beobachtet wird. Nach Konstruktion sind die X_i unabhängig,

sogar das ganze Tupel $(X_1, ..., X_n)$ ist unabhängig. Wir sprechen von den X_i als von n 'unabhängigen Kopien von X'. In diesem Rahmen hat man die äußerst wichtige 'Mittelungsgröße'

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i, \text{ wir schreiben manchmal verdeutlichend } \overline{X}^{(n)},$$

deren beobachteten Wert (bei den n Versuchen) man nutzt, um den Erwartungswert von X zu schätzen. Die Untersuchung der Wahrscheinlichkeit, mit welcher diese Schätzung eine bestimmte Genauigkeit besitzt, ist ein elementares Hauptanliegen der Statistik. Das werden wir behandeln.

2. Die wichtigen Verteilungsparameter Erwartungswert und Varianz (bzw. Streuung)

Vielfach interessiert man sich nicht für die volle Information der Verteilung einer Zufallsvariablen, sondern vor allem für den Erwartungswert und die Streuung (bzw. Varianz). Das sind Kennzahlen der Verteilung, mit folgender Bedeutung: Der Erwartungswert ist derjenige, auf den sich in langer Versuchsreihe das beobachtete empirische arithmetische Mittel einpegelt. Im Fall einer empirischen Verteilung hat man nur als Näherungswert dafür den arithmetischen Mittelwert aller beobachteten Werte. Die Varianz ist der Erwartungswert des quadratischen Abstandes der einzelnen Werte von X zum Erwartungswert von X. Man geht dann über zur Streuung, welche ein Maß für die Differenz der Größenwerte selbst zu ihrem Erwartungswert liefert. Sie ist definiert als Wurzel der Varianz. Wir definieren das nun und stellen die üblichen Symbole vor:

DEFINITION 17. *Der Erwartungswert einer Variablen X ist im Falle einer diskreten Verteilung*

$$E(X) = \mu(X) := \sum_a aP(X = a).$$

Ist X mit der Dichte f_X verteilt, so ist

$$E(X) = \mu(X) := \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Die Varianz von X ist so definiert:

$$Var(X) = \sigma^2(X) := E\left((X - E(X))^2\right).$$

Die Streuung von X ist definiert durch

$$\sigma(X) := \sqrt{\sigma^2(X)}.$$

Bemerkung: Man findet oft die Bezeichnung '$E(X)$', manchmal auch mit stilisiertem Buchstaben wie $\mathcal{E}$. Wir werden jedoch μ vorziehen, urteilen Sie selber, was Sie besser lesen können:

$$\sigma^2(X) = \mu\left((X - \mu(X))^2\right) \text{ oder } E\left((X - E(X))^2\right).$$

Wir werden für die Lesbarkeit manchmal auch schreiben: μ_X statt $\mu(X)$, dann hat man

$$\sigma^2(X) = \mu\left((X - \mu_X)^2\right).$$

Zum Verständnis: In der Definition des Erwartungswertes für den diskreten Fall hat man zumeist nur endliche viele Werte $a \in \mathbb{R}$, für welche $P(X = a) \neq 0$ ist, die Summe wird dann endlich und ist unproblematisch. Wenn aber unendlich viele Werte a_i, $i \in \mathbb{N}$, existieren mit $P(X = a_i) \neq 0$, dann könnte die unendliche Reihe nicht mehr konvergieren, und dann existiert $E(X)$ nicht. Dasselbe kann im kontinuierlichen Fall mit dem Integral $\int_{-\infty}^{\infty} x f_X(x)\, dx$ geschehen, dass es nicht existiert. Entsprechendes ist zur Varianz zu sagen. Aber die praktisch wichtigsten Verteilungen haben alle Erwartunswert und Varianz.

2.1. Berechnung von μ und σ für zusammengesetzte Zufallsvariablen. Es gelten dafür folgende leicht zu verifizierende Formeln:

$$\begin{array}{lrcl}
(i) & \mu(\alpha X+\beta Y) & = & \alpha\mu(X)+\beta\mu(Y) \text{ (stets)},\\
(ii) & \mu(X\cdot Y) & = & \mu(X)\mu(Y), \textit{ wenn } X,Y \text{ unabhängig sind},\\
(iii) & \sigma^2(X) & = & \mu\left(X^2\right)-\mu^2(X),\\
(iv) & \sigma(X+\alpha) & = & \sigma(X), \text{ entsprechend } \sigma^2(X+\alpha)=\sigma^2(X),\\
(v) & \sigma(\alpha X) & = & |\alpha|\,\sigma(X), \text{ entsprechend } \sigma^2(\alpha X)=\alpha^2\sigma^2(X),\\
(vi) & \sigma^2(X+Y) & = & \sigma^2(X)+\sigma^2(Y), \text{ wenn } X,Y \text{ unabhängig sind}^*, \text{ also}\\
(vii) & \sigma(X+Y) & = & \sqrt{\sigma^2(X)+\sigma^2(Y)}, \text{ wenn } X,Y \text{ unabhängig sind}^*.
\end{array}$$

*) : Hier genügt bereits die *lineare Unabhängigkeit*, die gleichbedeutend damit ist, dass

$$Cov(X,Y) := \mu((X-\mu(X))(Y-\mu(Y))) = 0 \text{ ('Kovarianz von } X,Y\text{')}.$$

Einige Begründungen:

Zu (i) zeigt man zunächst leicht: $\mu(\alpha X)=\alpha\mu(X)$ mit:

$$\mu(\alpha X)=\sum_b \alpha b P(X=b)=\alpha\sum_b bP(X=b)=\alpha\mu(X).$$

Wir benötigen noch die Aussage, dass $\mu(X+Y)=\mu(X)+\mu(Y)$. Beides zusammen ergibt (i). Tatsächlich ist die Additivität nur mit einer kleinen Raffinesse einsichtig, damit man nicht etwa die Unabhängigkeit vom X,Y voraussetzen muss (siehe (ii), dort wird sie benötigt, allerdings nur in einer schwachen Form): Wir zeigen zuerst mittels der Formel von der totalen Wahrscheinlichkeit:

$$(1)\ \ \mu(X)=\sum_b\sum_a aP(X=a\cap Y=b),$$

weil

$$\sum_b\sum_a aP(X=a\cap Y=b)=\sum_a a\sum_b P(X=a\cap Y=b)=^{*(\text{totale Wahrscheinlichkeit})}\sum_a a\sum_b P(X=a).$$

Nun folgt analog:

$$(2)\ \ \mu(Y)=\sum_b\sum_a bP(X=a\cap Y=b),$$

Mit (1) und (2) hat man dann:

$$\begin{array}{rcl}
\mu(X+Y) & = & ^{(\text{Definition})}\sum_{a,b}(a+b)\,P(X=a\cap Y=b)\\
& = & \sum_{a,b} aP(X=a\cap Y=b)+\sum_{a,b} bP(X=a\cap Y=b)\\
& = & ^{(1),(2)}\mu(X)+\mu(Y).
\end{array}$$

Bemerkung: Dieser Beweis ist formuliert für den Fall, dass alle Summen endlich sind - zu summieren ist über alle Wertepaare, welche die Variablen X,Y mit einer Wahrscheinlichkeit >0 annehmen. Aber man kann alles auch für unendliche Reihen lesen, also für den Fall abzählbar unendlich vieler Werte. Was noch besser ist: Man braucht nur die Summenzeichen durch Integrale zu ersetzen, da, db hinzuzufügen und $P(X=a\cap Y=b)$ durch eine gemeinsame Dichte für die Paare, also $f(a,b)$ zu ersetzen, um den Beweis für Zufallsvariablen mit Dichtefunktionen zu bekommen. (Integriert wird jeweils von $-\infty$ bis ∞, und natürlich muss es Bedingungen für die Existenz der Integrale geben.)

Zu (ii) für den diskreten Fall (im anderen geht das wie in (i) ganz ähnlich mit den Integralen anstelle der Summen):

$$\begin{array}{rcl}
\mu(XY) & = & \sum_{a,b} abP(X=a\cap Y=b)=^{X,Y \text{ unabh.}}\sum_{a,b} abP(X=a)\,P(Y=b)\\
& = & \left(\sum_a aP(X=a)\right)\left(\sum_b bP(Y=b)\right)
\end{array}$$

Wir begründen (*iii*):

$$\begin{aligned}\sigma^2(X) &= \mu\left((X-\mu(X))^2\right) = \mu\left(X^2 - 2\mu(X)X + \mu^2(X)\right)\\ &= \mu\left(X^2\right) - 2\mu^2(X) + \mu^2(X) = \mu\left(X^2\right) - \mu^2(X).\end{aligned}$$

Das funktioniert offensichtlich, wenn nur $\mu(X)$ und $\mu\left(X^2\right)$ existieren.

Zu (*vi*):

$$\begin{aligned}\sigma^2(X+Y) &= \mu\left((X-\mu(X)+Y-\mu(Y))^2\right)\\ &= \mu\left((X-\mu(X))^2\right) + \mu\left((Y-\mu(Y))^2\right) + 2\mu\left((X-\mu(X))(Y-\mu(Y))\right)\\ &= \sigma^2(X) + \sigma^2(Y) + 2Cov(X,Y).\end{aligned}$$

Also (*vii*) unter der Voraussetzung, dass $Cov(X,Y) = 0$.■

Beispiele zur Anwendung:

1.) Wenn X eine beliebige Zufallsvariable ist mit $\sigma(X) \neq 0$, dann hat man

$$\begin{aligned}\mu\left(\overline{X}\right) &= \mu(X),\\ \sigma\left(\overline{X}^{(n)}\right) &= \frac{\sigma(X)}{\sqrt{n}},\\ \mu\left(\frac{X-\mu(X)}{\sigma(X)}\right) &= 0,\\ \sigma\left(\frac{X-\mu(X)}{\sigma(X)}\right) &= 1.\end{aligned}$$

2.) Mittelwertsgrößen: Man wiederholt das zu X gehörige Zufallsexperiment n mal *unabhängig*, dann bildet man das arithmetische Mittel der beobachteten $X-$ Werte als Wert von $\overline{X}$. Diese Variable kann man mit der Definition: $X_i := X-$ Wert im i. Versuch so schreiben:

$$\overline{X} := \sum_{i=1}^{n} \frac{1}{n} X_i.$$

Nun hat man

$$\begin{aligned}\mu\left(\overline{X}\right) &= \mu\left(\sum_{i=1}^{n}\frac{1}{n}X_i\right) = \frac{1}{n}\cdot n\cdot\mu(X) = \mu(X),\\ \sigma^2\left(\overline{X}\right) &= \sigma^2\left(\sum_{i=1}^{n}\frac{1}{n}X_i\right) = \sum_{i=1}^{n}\frac{1}{n^2}\sigma^2(X) = \frac{1}{n}\sigma^2(x), \text{ also}\\ &\quad \sigma\left(\overline{X}\right) = \frac{\sigma(X)}{\sqrt{n}}.\end{aligned}$$

Letzteres ist das berühmte 'Wurzel $-n-$ Gesetz', das man sich wegen seiner großen Bedeutung merken sollte. Dass die Streuung des Mittelwertes nur mit $\sqrt{n}$ fällt, hat die Konsequenz, dass man für sehr genaue Schätzungen von unbekannten Erwartungswerten unangenehm große Stichproben benötigt.

2.2. Eine Faustregel für die Streuung, welche diese erläutert. Eine Faustregel heißt so, weil sie nichts Präzises sagt, aber eben doch etwas Substantielles. Hier lautet sie:

In aller Regel liegen über 90% der Populationswerte im Bereich $\mu(X) \pm 2\sigma(X)$, genauer formuliert :

Für beinahe alle Zufallsvariablen, denen man begegnet, gilt : $P(\mu(X) - 2\sigma(X) \leq X \leq \mu(X) + 2\sigma(X)) \geq 0.9$.

Im nächsten Abschnitt wird mit der Tschebyscheffschen Ungleichung diese Faustregel durch etwas ganz Genaues ersetzt. Aber dabei verliert man für viele praktische Zwecke zu viel, man kommt nur auf '≥ 0.75' statt '≥ 0.9'. Dabei wird man häufig die angesprochene Wahrscheinlichkeit über 0.95 finden.

Illustration: Nehmen Sie die Variablen: 'Wasserverbrauch von Haushalten', 'Stromverbrauch...', 'Wohnfläche...', 'Körperlänge...', 'Augensumme von n Würfel-Würfen', 'Lebensdauer von Geräten bestimmten Typs', 'Messfehler bei einer bestimmten physikalischen Messung', usw.,usw. Sie werden keine finden, für welche die Faustregel nicht gilt.

Die von $\sigma(X)$ beschriebene 'Breite' einer Verteilung begrenzt also die Wahrscheinlichkeit dafür, das ein zufällig beobachteter Wert von X weiter als $2\sigma(X)$ vom Erwartungswert $\mu(X)$ liegt. Darum ist $\sigma(X)$ auch viel besser als etwa einen kleinsten und größten möglichen Wert von X zu nennen, selbst wenn diese existieren. Diese Zahlangaben mögen zuweilen eine kleine zusätzliche Information geben, die manchmal sehr wichtig sein wird, z.B. der größtmögliche Ansturm auf eine medizinische Versorgungsstelle. Aber über die Verteilung von X und die meisten interessierenden Fragen dazu sagen sie fast gar nichts.

2.3. Die Tschebyscheffsche Ungleichung. Eine Vorbemerkung zur Transkription dieses russischen Namens: Die hier verwandte ist im deutschsprachigen Raum traditionell. Moderner wird man so etwas wie 'Chebyshev' finden, mit dem Vorteil, das man das im Englischen gut wiedererkennt.

SATZ 3. *Für jede beliebige Zufallsvariable X (für die $\mu(X)$ und $\sigma(X)$ existieren) und jede Zahl $\alpha > 0$ gilt*

$$P(|X - \mu(X)| \geq \alpha) \leq \frac{\sigma^2(X)}{\alpha^2}.$$

Beweis:

$$\sigma^2(X) = \sum_{|a-\mu(X)|\leq\alpha} (a - \mu(X))^2 P(X = a) + \sum_{|a-\mu(X)|\geq\alpha} (a - \mu(X))^2 P(X = a),$$

also

$$\alpha^2 P(|X - \mu(X)| \geq \alpha) \leq \sum_{|a-\mu(X)|\geq\alpha} (a - \mu(X))^2 P(X = a) \leq \sigma^2(X). \blacksquare$$

Zum Verständnis: Diese Ungleichung macht klar, dass die Varianz von X auch mit mathematischer Genauigkeit die Wahrscheinlichkeit dafür begrenzt, dass X einen Wert mindestens im Abstand $\alpha > 0$ vom Mittelwert annimmt. Sie gibt also eine gültige wichtige Deutung für die Varianz im Stile der Faustregel. Natürlich liefert die Ungleichung nur dann ein nichttriviales Resultat, wenn die rechte Seite einen Wert unter 1 annimmt.

Eine systematisch wichtige Anwendung: Die Wahrscheinlichkeit dafür, dass die Variable $\overline{X}$ einen Wert annimmt, der von $\mu(X) = \mu\left(\overline{X}\right)$ um wenigstens $\alpha > 0$ entfernt ist, geht nach Null mit $n \to \infty$. Denn $\sigma^2\left(\overline{X}\right) = \frac{1}{n}\sigma^2(X)$ geht nach Null für $n \to \infty$. Mehr noch: Man kann mit einer sehr groben Abschätzung für $\sigma^2(X)$ nach oben für festen Wert $\alpha > 0$ angeben, welches n man braucht, um mit einem beobachteten Wert von $\overline{X}$ den Mittelwert $\mu(X)$ mit Genauigkeit α zu schätzen, mit einer Wahrscheinlichkeit von p (die man natürlich groß wählt).

Beispiel: Man möchte eine unbekannte Wahrscheinlichkeit $P(A) = p_0$ schätzen durch die relative Häufigkeit, mit der man das Ereignis A in n Versuchen beobachtet. Das Resultat soll mit Wahrscheinlichkeit 0.99 eine Ungenauigkeit von nur 0.01 aufweisen. Dann hat man jedenfalls für die Variable X (Wert 1, wenn A beobachtet wird, 0 sonst):

$$\sigma^2(X) \leq \frac{1}{4}, \text{ also } \sigma^2\left(\overline{X}\right) = \frac{1}{4n}.$$

wir suchen also n, so dass

$$\frac{1}{4n \cdot 0.0001} = \frac{2500}{n} \leq 0.01.$$

Wir finden, dass $n = 250000$ genügt. Bei 250000 unabhängigen Ausführungen des Zufallsexperiments haben wir also eine Wahrscheinlichkeit von 0.99, dass die beobachtete relative Häufigkeit, mit der A herauskommt, von $P(A) = p_0$ um höchstens 0.01 entfernt ist.

Bemerkung: Später werden wir dies Resultat mittels Normalverteilung und dann noch $t-$ Verteilung noch wesentlich verbessern können. Der Vorteil der Tschebyscheffschen Ungleichung liegt aber darin, dass ihr keinerlei Voraussetzung über die Verteilung von X zugrunde liegt.

In dieser Linie liegt der folgende Satz, der immerhin erlaubt, das erwähnte 'empirische Gesetz der großen Zahlen' nachzuzeichnen:

SATZ 1 (das schwache Gesetz der großen Zahlen). *Wenn X eine beliebige Zufallsvariable ist und $\mu(X), \sigma(X)$ existieren, dann hat man für alle $\varepsilon > 0$:*

$$\lim_{n\to\infty} P\left(\left|\overline{X}^{(n)} - \mu(X)\right| \geq \varepsilon\right) = 0.$$

Gleichwertig ist (über Komplementbildung):

$$\lim_{n\to\infty} P\left(\left|\overline{X}^{(n)} - \mu(X)\right| < \varepsilon\right) = 1$$

Beweis: Mit der Tschebyscheffschen Ungleichung haben wir - ε bleibt im Verlauf fest gewählt:

$$0 \leq P\left(\left|\overline{X}^{(n)} - \mu(X)\right| \geq \varepsilon\right) \leq \frac{\sigma^2\left(\overline{X}^{(n)}\right)}{\varepsilon^2} = \frac{\sigma^2(X)}{\varepsilon^2 n} \underset{n\to\infty}{\to} 0.$$

Es folgt sofort, dass $\alpha_n(\varepsilon) := P\left(\left|\overline{X}^{(n)} - \mu(X)\right| \geq \varepsilon\right)$ eine Nullfolge ist, d.h. mit $n \to \infty$ gegen Null geht.■

Bemerkungen:

1. Das oft eigens für sich genannte 'Bernoullische Gesetz der großen Zahlen' ist nur der Spezialfall davon, dass X eine Bernoulli-Variable ist (also nur mit Werten $1, 0$). Es lohnt sich aber auch, diesen Fall konkret anzuschauen: Dann ist $\mu(X)$ einfach die Wahrscheinlichkeit eines Ereignisses A in einem beliebigen Wahrscheinlichkeitsraum. Man hat weiter $\mu(X) = P(A) =: p$ und $\sigma(X) = \sqrt{p(1-p)}$. Die Voraussetzungen des Satzes sind also erfüllt, und der Satz besagt dann in Worten: Wenn man den Stichprobenumfang n groß genug wählt, so hat man eine beliebig hohe Wahrscheinlichkeit (also beliebig nahe an 1) dafür, dass die gefundene empirische relative Häufigkeit so nahe bei dem tatsächlichen $P(A)$ liegt, wie man will. Die empirische relative Häufigkeit ist der zu findende Wert von $\overline{X}^{(n)}$. Wir haben hier die zweite Lesart verwendet und oben bereits konkrete Beispiele gegeben.

2. Man sollte auch den allgemeinen Fall in der zweiten Lesart verbal kompakt formulieren: Bei genügendem Stichprobenumfang ist das arithmetische Mittel der Werte einer Zufallsvariablen mit beliebig hoher Wahrscheinlichkeit beliebig nahe am Erwartungswert der Variablen. (Für eine empirische Variable in einer riesigen Population ist der Erwartungswert einfach der arithmetische Mittelwert in der Gesamtpopulation.)

3. Der Satz besagt also, dass man sich am eigenen Schopf aus dem Sumpf ziehen kann und ermittelt, wie groß der Abstand zum Gewünschten ist, obgleich man das Gewünschte niemals genau zu wissen bekommt. Allerdings muss man einen Preis bezahlen: Die Aussage ist nur eine mit hoher Wahrscheinlichkeit, keine absolut sichere. Außerdem macht die Forderung nach einer hohen Wahrscheinlichkeit auch Arbeit - sie verlangt große Stichprobenumfänge und insbesondere ordentliche Stichproben - nicht etwa solche, die 'zufällig' bei der Hand liegen. ('Bequem zur Hand' ist alles andere als wirklich zufällig!) Es sei noch einmal bemerkt, dass dies 'am eigenen Schopf aus dem Sumpf ziehen' auch einen geeigneten Umgang mit der dann natürlich auch unbekannten Streuung verlangt, aber nicht daran scheitert.

3. Die wichtigsten diskreten Verteilungen

3.1. Kombinatorische Vorbereitungen. Wir stellen hier einige kombinatorische Tatsachen zusammen, die für das Folgende nützlich sind:

Wenn wir n Objekte haben, aus denen wir beliebige Folgen der Länge k bilden (mit erlaubten Wiederholungen), so gibt es dafür offenbar $n \cdot n \cdot ... \cdot n = n^k$ Möglichkeiten. Wenn wir Folgen ohne Wiederholungen bilden, so gibt es dafür offenbar $n(n-1)...(n-k+1)$ Möglichkeiten. Diese Zahl können wir mit $n! = \begin{cases} 1 \text{ für } n = 0 \\ 1 \cdot ... \cdot n \text{ für } n \geq 1 \end{cases}$ dann auch so schreiben:

$$n(n-1)...(n-k+1) = \frac{n!}{(n-k)!}.$$

Definition 18. *$n!$ ('n Fakultät') ist definiert durch $0! := 1$, $(n+1)! = (n+1)\,n!$ Die Binomialkoeffizienten sind definiert durch*

$$\binom{n}{k} := \frac{n!}{k!\,(n-k)!}.$$

Satz 4. *$n!$ ist die Anzahl der Möglichkeiten, n Objekte anzuordnen. $\binom{n}{k}$ ist die Anzahl der Möglichkeiten, aus einer Menge von n Elementen eine $k-$ elementige Teilmenge auszuwählen, für $k = 0, 1, ..., n$.*

Begründung: Es ist offenbar die Zahl der Möglichkeiten, $n = 0$ Objekte anzuordnen, Eins. (Oder fangen Sie an mit $n = 1$, dann gibt es wieder genau eine Anordnungsmöglichkeit.) Nun kann man bei $n+1$ Objekten zunächst den Platz für das $n+1$. Objekt auf $n+1$ Weisen bestimmen, dann hat man noch unabhängig alle Möglichkeiten, die restlichen n Objekte anzuordnen. Wenn wir mit Induktionsvoraussetzung für letztere Anzahl $n!$ haben, dann also insgesamt $(n+1)\,n!$ Möglichkeiten. Damit ist die erste Aussage durch Induktion bewiesen.

Zur zweiten Aussage: Man hat mit den Überlegungen zum Eingang

$$\frac{n!}{(n-k)!} \text{ Möglichkeiten,}$$

k Objekte aus n Objekten in einer bestimmten Reihenfolge auszuwählen. und nun gibt es $k!$ mögliche Reihenfolgen, in denen dieselbe Teilmenge ausgewählt werden kann, daher ist die fragliche Anzahl gerade

$$\frac{n!}{k!\,(n-k)!} = \binom{n}{k}.\blacksquare$$

Satz 5.

Folgerung 4 (Verallgemeinerung auf Multinomialkoeffizienten).

$$\begin{aligned} \frac{n!}{k_1! \cdot \ldots \cdot k_r!} &= \textit{Anzahl der Möglichkeiten, } n \textit{ Objekte auf } r \textit{ Töpfe zu verteilen,} \\ &\quad \textit{so dass } k_s \textit{ Objekte jeweils in den } s. \textit{ Topf kommen, für} \\ k_1 + \ldots + k_r &= n. \end{aligned}$$

Zur Begründung: Für $r = 1$ stimmt die Aussage offensichtlich, für $r = 2$ ist sie oben schon bewiesen. Sie gelte für r Töpfe. Wir zeigen, dass sie dann auch für $r+1$ Töpfe gilt: Es mögen k Objekte in die ersten r Töpfe kommen, $n - k = k_{r+1}$ Objekte in den $r+1$. Topf. Dann gibt es dafür nach dem vorigen Satz

$$\frac{n!}{k!\,(n-k)!} \text{ Möglichkeiten.}$$

Unabhängig davon gibt es nun nach Induktionsvoraussetzung noch

$$\frac{k!}{k_1! \cdot \ldots k_r!} \text{ Möglichkeiten,}$$

die erwähnten k Objekte auf die ersten r Töpfe zu verteilen, mit $k_1 + \ldots + k_r = k$. Das macht zusammen

$$\frac{k!}{k_1! \cdot \ldots k_r!} \cdot \frac{n!}{k!\,(n-k)!} = \frac{n!}{k_1! \cdot \ldots k_r! k_{r+1}!} \text{ Möglichkeiten,}$$

die n Objekte auf $r+1$ Töpfe so zu verteilen, dass k_s Objekte im s. Topf landet, $s = 1, \ldots, r+1$.$\blacksquare$

Anwendungsbeispiele:

1.) Mit welcher Wahrscheinlichkeit haben bei 5 zufällig ausgewählten Menschen wenigstens zwei im selben Monat Geburtstag? Dass alle fünf in verschiedenen Monaten Geburtstag haben, dafür gibt es

$$12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \text{ Möglichkeiten,}$$

unter insgesamt 12^5 Möglichkeiten für die Geburtstagsmonate der Fünf, also ist die gefragte Wahrscheinlichkeit

$$1 - \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8}{12^5} = \frac{89}{144} \approx 0.62.$$

Vorausgesetzt wird allerdings, dass alle Geburtstagsmonate gleich wahrscheinlich sind. Diese Wahrscheinlichkeit ist erstaunlich hoch, man würde sie intuitiv wohl stets unterschätzen.

2.) Wie viele Möglichkeiten gibt es, 10 Kugeln auf drei Schachteln zu verteilen, so dass drei jeweils in die erste, drei in die zweite und 4 in die dritte Schachtel kommen? Diese Anzahl ist

$$\frac{10!}{3!3!4!} = 4200.$$

3.2. Binomialverteilung.

DEFINITION 19. *Eine Zufallsvariable heißt $(n,p)-$ binomialverteilt, mit $n \in \mathbb{N}$ und $0 \leq p \leq 1$, wenn sie die Gestalt hat: $X =$ Trefferzahl bei n unabhängigen Versuchen mit Einzeltrefferwahrscheinlichkeit p.*

Beispiel: Würfelt man 10 mal mit einem gewöhnlichen Würfel, so ist die Größe 'Anzahl der Sechsen' $\left(10, \frac{1}{6}\right)-$ binomialverteilt.

SATZ 6. *Ist X $(n,p)-$ binomialverteilt, so hat man:*

$$f_X(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ für } k = 0,1,...,n.$$

Ferner sind Erwartungswert und Streuung:

$$\begin{aligned} \mu(X) &= np, \\ \sigma(X) &= \sqrt{np(1-p)}. \end{aligned}$$

Bemerkung: Man muss dabei die Wahrscheinlichkeiten für Intervalle über längere Summen ausrechnen, dazu gehören auch die Werte der Verteilungsfunktion. Wir werden diese Werte später über Normalverteilung sehr gut nähern können.

Beispiele:

1.) Die Wahrscheinlichkeit für höchstens 3 Sechsen bei 20 Würfen mit einem Würfel ist:

$$\begin{aligned} P(X_{20} \leq 3) &= \sum_{k=0}^{3} \binom{20}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{20-k} \\ &= \left(\frac{5}{6}\right)^{20} + 20 \cdot \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^{19} + 190 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{18} + 1140 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{17} \\ &\approx 0.57. \end{aligned}$$

2.) Die Wahrscheinlichkeit für höchstens 30 Sechsen bei 200 Würfen ist

$$P(X_{200} \leq 30) = \sum_{k=0}^{30} \binom{200}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{200-k} \approx 0.3.$$

Die Wahrscheinlichkeit für höchstens 300 Sechsen bei 2000 Würfen ist

$$P(X_{2000} \leq 300) = \sum_{k=0}^{300} \binom{2000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{2000-k} \approx 0.0233.$$

Hier bekommt man die Resultate allenfalls noch mit einem Computeralgebraprogramm.

Denken Sie darüber nach, warum bei diesen analogen Problemen nicht stets dasselbe Resultat herauskommt.

Begründung des Satzes: Dass k 'Treffer' und $n-k$ 'nicht Treffer' kommen in einer bestimmten, aber beliebigen Reihenfolge, hat wegen der Multiplikation der Wahrscheinlichkeiten für $\cap$ die Wahrscheinlichkeit

$$p^k (1-p)^{n-k}.$$

Nun gibt es $\binom{n}{k}$ mögliche Anordnungen für die 'Treffer' und 'Nieten', und die Ereignisse 'genau k Treffer in Anordnung (1)', 'genau k Treffer in Anordnung (2)' schließen einander aus, daher die Formel für $P(X=k)$. Die Formeln für Erwartungswert und Streuung zeigt man zuerst für $n=1$. Wir bezeichnen mit X_1 eine $(1,p)-$ binomialverteilte Zufallsvariable, mit X_n eine $(n,p)-$ binomialverteilte Zufallsvariable. Dann hat man

$$\begin{aligned} \mu(X_1) &= 1 \cdot P(X_1 = 1) + 0 \cdot P(X_1 = 0) \\ &= p, \text{ also} \\ \mu(X_n) &= n\mu(X_1) = np. \end{aligned}$$

Ferner für die Varianz:

$$\begin{aligned}\sigma^2(X_1) &= \mu(X_1 \cdot X_1) - \mu^2(X_1) \\ &= P(X_1 = 1) - p^2 = p - p^2 \\ &= p(1-p).\end{aligned}$$

Daraus folgt sofort, dass (mit der Unabhängigkeit der Versuche):

$$\sigma^2(X_n) = n\sigma^2(X_1) = np(1-p). \blacksquare$$

Wir können die Wahrscheinlichkeitsformel noch ein wenig verallgemeinern: Hat man die Ereignisse $B_1, ..., B_r$, die eine Klasseneinteilung bilden, dabei $P(B_s) = p_s$ für $s = 1, ..., r$, dann ist die Wahrscheinlichkeit dafür, dass bei n Versuchen genau k_s mal B_s herauskommt, gleich

$$P(X_1 = k_1 \cap ... \cap X_r = k_r) = \frac{n!}{k_1! \cdot ... \cdot k_r!} p_1^{k_1} \cdot ... \cdot p_r^{k_r}.$$

Wir haben hier mit X_s die Variable bezeichnet: 'Anzahl der Vorkommen von B_s bei n Versuchen'. Man beachte, dass stets $k_1 + ... + k_r = n$ gelten muss, und dass ohnehin $p_1 + ... + p_r = 1$.

Beispiel: Mit welcher Wahrscheinlichkeit hat man bei 10 Würfen mit einem gewöhnlichen Würfel drei mal eine Zahl unter 3, 2 mal eine 3 und 5 mal eine Zahl über 3 (in beliebiger Reihenfolge (!))? Diese Wahrscheinlichkeit ist

$$\frac{10!}{3! \cdot 2! \cdot 5!} \left(\frac{1}{3}\right)^3 \left(\frac{1}{6}\right)^2 \left(\frac{1}{2}\right)^5 = \frac{35}{432} \approx 0.08.$$

3.3. Hypergeometrische Verteilung. Wir stellen uns vor, dass wir aus einer Urne mit N Kugeln n Kugeln *ohne Zurücklegen* ziehen. In der Urne seien K 'Trefferkugeln'. Wir interessieren uns für die Variable X = Trefferzahl. Dann hat man wieder

$$X = \sum_{k=0}^{n} X_k,$$

wobei X_k = Trefferzahl bei Ziehen der k. Kugel. Aber die Variablen X_k sind nicht mehr unabhängig (bei Ziehen mit Zurücklegen wäre die gesamte Trefferzahl $\left(n, \frac{K}{N}\right)$ − binomialverteilt). Hier liegt die (N, K, n) − hypergeometrische Verteilung vor. Die Wahrscheinlichkeiten für die möglichen Trefferzahlen sind leicht auszurechnen: Wir haben die Laplace-Situation, und Ω ist die Menge aller Teilmengen der N Kugeln, welche n Elemente enthalten. Also

$$|\Omega| = \binom{N}{n}.$$

Ebenso ist für das Ereignis $X = k$ die Anzahl der Fälle einfach

$$\binom{K}{k}\binom{N-K}{n-k},$$

weil man dabei k Kugeln aus K Trefferkugeln und unabhängig $n-k$ Kugeln aus den $N-K$ 'Nietenkugeln' ziehen muss. Also ist die Verteilung einfach:

$$P(X = k) = f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Beispiel: In einer Urne sind $N = 5$ Kugeln, davon sind $K = 3$ rote Kugeln, wir ziehen $n = 3$ Kugeln. Mit welcher Wahrscheinlichkeit ziehen wir dabei genau zwei rote? Wir haben

$$P(X = 2) = \frac{\binom{3}{2}\binom{2}{1}}{\binom{5}{3}} = \frac{6}{10} = \frac{3}{5}.$$

Zum Vergleich: Bei Ziehen mit Zurücklegen bekommt man mit der $\left(3, \frac{3}{5}\right)$ −Binomialverteilung:

$$P(X_{binv} = 2) = \binom{3}{2}\left(\frac{3}{5}\right)^2\left(\frac{2}{5}\right)^1 = 0.432.$$

Die letztere Wahrscheinlichkeit ist also deutlich geringer. Das ist auch so zu erwarten: Tatsächlich ist die Varianz der entsprechenden hypergeometrischen Verteilung geringer als die der Binomialverteilung.

Wir berechnen μ und σ^2 für eine (N, K, n) −hypergeometrisch verteilte Variable.

Zunächst stellen wir fest, dass mit den eingeführten Variablen X_k gilt:

$$(1)\quad \mu(X_k) = \frac{K}{N},\ \text{also unabhängig von } k.$$

Dafür haben wir natürlich mit bedingten Wahrscheinlichkeiten zu rechnen außer für $k = 1$, klar gilt die Aussage (1) für $k = 1$. Nun hat man aber weiter mit Induktion - wir setzen die Aussage für k voraus:

$$\begin{aligned}
\mu(X_{k+1}) &= P(X_{k+1} = 1|X_k = 1)\, P(X_k = 1) + P(X_{k+1} = 1|X_k = 0)\, P(X_k = 0) \\
&= \frac{K-1}{N-1}\frac{K}{N} + \frac{K}{N-1}\left(1 - \frac{K}{N}\right) \\
&= \frac{K}{N-1}\left(\frac{K-1+N-K}{N}\right) \\
&= \frac{K}{N}.
\end{aligned}$$

Es folgt aus (1) nun:

$$\mu(X) = \sum_{k=1}^{n} \mu(X_k) = n\frac{K}{N}.$$

Der Erwartungswert ist also wir bei der entsprechenden Binomialverteilung. Aber die Varianz ist kleiner, wir berechnen sie so:

$$\sigma^2(X) = \mu\left(X^2\right) - \mu^2(X).$$

Nun ist

$$\mu\left(X^2\right) = \mu\left(\left(X_1 + ... + X_n\right)^2\right) = \sum_{k=1}^{n} \mu\left(X_k^2\right) + \sum_{i \neq j} \mu(X_i X_j).$$

Dabei hat man mit $X_k = X_k^2$ (in diesem Falle!):

$$\sum_{k=1}^{n} \mu\left(X_k^2\right) = n\frac{K}{N}.$$

Ferner für $i \neq j$, $1 \leq i, j \leq n$:

$$\begin{aligned}
\mu(X_i X_j) &= 1 \cdot P(X_i = 1 \cap X_j = 1) \\
&= P(X_i = 1|X_j = 1) \cdot P(X_j = 1) \\
&= \frac{K-1}{N-1} \cdot \frac{K}{N}.
\end{aligned}$$

Die Paare (i, j) mit $i \neq j$ haben die Anzahl $n(n-1)$. Also

$$\begin{aligned}
\mu\left(X^2\right) &= n\frac{K}{N} + n(n-1)\frac{K-1}{N-1} \cdot \frac{K}{N} \\
&= n\frac{K}{N}\left(1 + (n-1)\frac{K-1}{N-1}\right).
\end{aligned}$$

Damit wird

$$\begin{aligned}
\sigma^2(X) &= n\frac{K}{N}\left(1 + (n-1)\frac{K-1}{N-1}\right) - n^2\frac{K^2}{N^2} \\
&= n\frac{K}{N}\left(1 + (n-1)\frac{K-1}{N-1} - n\frac{K}{N}\right) \\
&= n\frac{K}{N} \cdot \frac{N^2 - Nn - NK + nK}{(N-1)N} \\
&= n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}
\end{aligned}$$

Jetzt sieht man deutlich: Die Varianz der entsprechenden Binomialverteilung wäre

$$n\frac{K}{N}\frac{N-K}{N} = np(1-p).$$

Hier kommt also noch der Faktor $\frac{N-n}{N-1}$ hinzu, der kleiner als 1 ist, wenn $n > 1$. Wenn N viel größer ist als n, so ist der Unterschied sehr gering, auch kann man dann die hypergeometrische Verteilung überhaupt durch Binomialverteilung nähern. Aber mit mäßigem N und beträchtlichem n werden die Unterschiede sehr groß. Etwa für $N = 11$ und $n = 6$ wird die Varianz durch den Faktor halbiert. Entsprechend wird die Wahrscheinlichkeit größer, nahe beim Erwartungswert zu liegen. Das sah man oben im Wahrscheinlichkeitsbeispiel.

Wir fassen zusammen:

SATZ 7. *Für eine $(N, K, n)-$ hypergeometrisch verteilte Zufallsvariable X gilt:*

$$\begin{aligned} P(X = k) &= f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \\ \mu(X) &= n\frac{K}{N}, \\ \sigma^2(X) &= n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}. \end{aligned}$$

Eine Anwendung der hypergeometrischen Verteilung: Man hat etwa folgende absolute Häufigkeiten beobachtet, mit denen Geräte nach Reparatur defekt oder in Ordnung waren:

	defekt	funktionstüchtig
Werkstatt A	4	1
Werkstatt B	1	4

Wie wahrscheinlich wäre eine solche Abweichung (nach beiden Seiten, also auch zugunsten von A), wenn man annimmt, dass von den 12 Geräten genau sechs objektiv reparierbar waren und von den 10 Geräten zufällig je sechs zu beiden Werkstätten geschickt worden wären, so dass das Resultat nichts mit der Qualität der Werstätten zu tun hätte? Diese Wahrscheinlichkeit wäre:

$$2 \cdot \left(\frac{\binom{5}{1}\binom{5}{4}}{\binom{10}{5}} + \frac{\binom{5}{0}\binom{5}{5}}{\binom{10}{5}} \right) \approx 0.21.$$

Eine Diskrepanz, welche mindestens so groß ist wie die beobachtete, wäre also gar nicht unwahrscheinlich. Für solche Tafeln wird es natürlich lästig, hypergeometrisch zu rechnen, wenn die Beobachtungszahlen viel größer sind. Dann benutzt man die χ^2- Verteilung mit einem Freiheitsgrad (s.u.). Statt an Geräte kann man auch denken an 'geheilt' - 'nicht geheilt' und 'Medikament A', Medikament B'. Auch dann kann man fragen, wie wahrscheinlich eine solche Abweichung wäre, wenn man nur Leute, die ohne Medikamente überhaupt 'geheilt' oder 'nicht geheilt' würden, zufällig den Medikamenten zugeführt hätte - so etwas soll ja wohl vorkommen.

Bemerkungen:

1.) Es sollte wohl klar sein, dass eine beobachtete Tafel mit denselben Verhältnissen wie im Beispiel, aber viel größeren absoluten Häufigkeiten, etwa 40 und 10 statt 4 und 1, offensichtlich gegen die Hypothese der Unabhängigkeit spricht (auch rechnerisch), aber dafür benutzen wir dann lieber die Näherung durch die χ^2- Verteilung.

2.) Zuweilen wäre das Modell von zwei unabhängigen binomialverteilten Variablen, deren absolute Differenz man betrachtet, passender. Die dabei resultierenden Wahrscheinlichkeiten sind etwas anders.

3.3.1. *Verallgemeinerung der hypergeometrischen Verteilung.* Die Wahrscheinlichkeitsformel für die $(N, K, n)-$ hypergeometrische Verteilung hatten wir sehr einfach ermittelt über die Laplace-Formel. Nun kann man analog auch dieselbe Verallgemeinerung anbringen wie die von der Binomialverteilung zur Multinomialverteilung, nur ist es hier noch einfacher; wir betrachten also eine Urne mit N Kugeln, aus denen wir wieder n herausziehen, auf einmal bzw. 'ohne Zurücklegen'. Nun sind in der Urne $r > 2$ Sorten von Kugeln, jeweils mit den Anzahlen $K_1, ..., K_r$. (Offenbar $K_1 + ... + K_r = N$.) Die natürliche Frage: Mit welcher Wahrscheinlichkeit bekommen wir k_1 Kugeln der ersten Sorte und ... und k_r Kugeln der r.

Sorte? Nennen wir $X_i :=$ Anzahl der gezogenen Kugeln der Sorte i. Dann hat man offenbar:

$$\begin{aligned} \text{Für } 0 \quad \leq \quad & k_i \leq K_i,\ 1 \leq i \leq n, \text{ mit } k_1 + ...k_r = n \text{ gilt:} \\ & P\left(X_1 = k_1 \cap ... \cap X_r = k_r\right) \ = \ P\left(\bigcap_{i=1}^{r} P\left(X_i = k_i\right)\right) \\ = \quad & \frac{\binom{K_1}{k_1}...\binom{K_r}{k_r}}{\binom{N}{n}} = \frac{\prod_{i=1}^{r}\binom{K_i}{k_i}}{\binom{N}{n}}. \end{aligned}$$

Die einfache Begründung: wieder hat man $\binom{N}{n}$ gleich wahrscheinliche Möglichkeiten, n Kugeln aus alle n herauszuziehen. Nun ist aber die Zahl der 'günstigen Fälle' so zu sehen: Man hat $\binom{K_1}{k_1}$ Möglichkeiten, die k_1 verlangten Kugeln der ersten Sorte zu wählen, unabhängig davon $\binom{K_2}{k_2}$ Möglichkeiten, die k_2 Kugeln der zweiten Sorte zu wählen, usw., und diese Anzahlen multiplizieren sich nach unserem bekannten kombinatorischen Grundprinzip.

3.4. Poisson-Verteilung. Wir wollen die Poissonverteilung betont nicht als Näherung der Binomialverteilung für gewisse Parameterbereiche auffassen, weil sie ihre eigene Rolle spielt. Es handelt sich darum: Man hat eine Variable 'Trefferzahl', die sich nicht auf eine Zahl n diskreter Versuche bezieht, sondern 'Treffer' sollen zu jedem Zeitpunkt passieren können und vor allem unabhängig davon eintreten, ob zuvor viele oder wenige 'Treffer' passierten. Ein Beispiel: $X =$ Zahl kosmischer Teilchen, die pro Sekunde auf einer Fläche eintreffen. Oder auch die Zahl atomarer Zerfälle, die in einem radioaktiven Material pro Sekunde festzustellen sind. In solchen Fällen hat man einen festen Erwartungswert $\lambda > 0$ für die Zahl der Treffer pro Zeiteinheit (oder pro Raum-Zeit-Einheit). Dieser Erwartungswert und die Unabhängigkeit sind die definierenden Merkmale allein.

Bemerkung: Oft wird heute noch von der Poissonverteilung als von der Verteilung 'seltener Ereignisse' gesprochen. Das ist ein Relikt aus frühen bekannten Anwendungen ('Tödliche Unfälle mit Pferden in der preußischen Armee'), hat aber nichts mit der Sache zu tun. Kosmische Teilchen kommen als wahre Schauer, dabei wird λ riesig sein. Gerade heute lägen ganz andersartige Anwendungen nahe. Darum ist es stets ungünstig, mathematische Begriffe und Resultate auf zunächst liegende Anwendungen herunter zu projizieren. Das ergibt nicht nur eine Blockierung weitergehender Anwendungen, sondern es entstehen dabei auch Vorstellungen, die schief liegen und nicht zur Sache Gehöriges als definierend umdeuten. Dies Phänomen ist sehr häufig zu beobachten.

Wenn es nun zur Wahrscheinlichkeitsfunktion geht, dann werden wir allerdings Binomialverteilungen verwenden, doch in dem Sinne, dass wir die $\lambda-$ Poissonverteilung als Limes der Binomialverteilungen mit $\lambda = np$ fest und $n \to \infty$ (also entsprechend $p = \lambda/n \to 0$) bilden. Damit nähern wir uns der kontinuierlichen Zeit, die gemeint ist.

Wir haben (k wird festgehalten (!)):

$$\begin{aligned} P\left(X_{binv(n,\lambda/n)} = k\right) \quad &= \quad \binom{n}{k} p^k \left(1-p\right)^{n-k} \\ &= \quad \frac{n!}{k!\,(n-k)!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \quad \frac{n!}{(n-k)!n^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \quad \frac{n\,(n-1)\cdot ...\cdot (n-k+1)}{n^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k}. \end{aligned}$$

Nun bilden wir den Limes für $n \to \infty$. Der erste Faktor geht offenbar gegen 1, der zweite ist konstant, und der dritte geht gegen $e^{-\lambda}$. Wir kennen

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda},$$

und daraus folgt auch

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-k}=e^{-\lambda},$$

weil

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{k}=\left(\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)\right)^{k}=1^k=1.$$

Damit haben wir für eine $\lambda-Poisson-verteilte$ Zufallsvariable X :

$$P\left(X=k\right)=e^{-\lambda}\frac{\lambda^k}{k!}.$$

Wir berechnen noch Erwartungswert und Varianz - der Erwartungswert ist λ nach Konstruktion. Nun hat man

$$\mu\left(X^2\right)=\sum_{k=0}^{\infty}k^2e^{-\lambda}\frac{\lambda^k}{k!}=\lambda\left(\lambda+1\right).$$

Damit also

$$\sigma^2\left(X\right)=\lambda\left(\lambda+1\right)-\lambda^2=\lambda.$$

Die Varianz ist also wieder λ. Das sollte auch nicht überraschen, da die Varianzen der Binomialverteilungen, als deren Limes wir die Poissonverteilung bildeten, ebenfalls gegen $\lambda=np$ gehen - sie sind $np\left(1-p\right)=\lambda\left(1-\frac{\lambda}{n}\right)$, und das geht gegen λ.

SATZ 8. *Sei X $\lambda-Poisson-verteilt$ (stets $\lambda>0$). Dann gilt:*

$$\begin{aligned}P\left(X=k\right) &= e^{-\lambda}\frac{\lambda^k}{k!},\ \textit{für } k\in\mathbb{N}_0,\\ \mu\left(X\right) &= \lambda,\\ \sigma^2\left(X\right) &= \lambda,\ \textit{also } \sigma\left(X\right)=\sqrt{\lambda}.\end{aligned}$$

Beispiel: Erwartet man auf einer winzigen Fläche 5 kosmische Teilchen pro Sekunde, so ist die Wahrscheinlichkeit dafür, in der nächsten Sekunde genau 3 Teilchen zu beobachten:

$$P\left(X=3\right)=e^{-5}\frac{5^3}{3!}\approx 0.14.$$

4. Verteilungen mit Dichte

4.1. Normalverteilungen. Das sind die wichtigsten überhaupt. Das liegt daran, dass lange Summen unabhängiger Variablen beinahe immer näherungsweise normalverteilt sind. Dabei bedeutet 'lang' tatsächlich oft in der Praxis bereits, dass Längen unter 10 bereits ausreichen. Insbesondere ist die Verteilung einer Mittelwertsgröße $\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$ mit X_k Kopie von X stets näherungsweise normalverteilt, bereits mit geringem n. Das gilt natürlich dann auch für $\sum_{k=1}^{n}X_k$, also z.B. für eine binomialverteilte Variable. Welches n für eine gute Näherung ausreicht, das hängt von der Schiefe der Verteilung ab. Im Allgemeinen ist $n=10$ schon sehr gut. Aber beispielsweise für eine Binomialverteilung ist die Faustregel $np\left(1-p\right)\geq 10$. Das bedeutet: Wenn $p\left(1-p\right)$ klein ist, d.h. die Binomialverteilung sehr schief, braucht man größeres n.

Wir werden zunächst das technische Grundwissen bereitstellen, um dann typische Anwendungen auszuführen.

DEFINITION 20. *Eine Variable X heißt $(\mu,\sigma)-normalverteilt$, wenn sie verteilt ist mit der Dichte*

$$\varphi_{\mu,\sigma}\left(x\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Also lautet ihre Verteilungsfunktion

$$F_X\left(a\right)=\Phi_{\mu,\sigma}\left(a\right)=\int_{-\infty}^{a}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx.$$

Die Parameter μ und σ sind dabei gerade Erwartungswert und Streuung. Die Normalverteilung mit $\mu=0$ und $\sigma=1$ heißt Standard-Normalverteilung.

FOLGERUNG 5. *Man hat folgende praktisch wichtigen Formeln:*

$$(1)\quad \Phi_{\mu,\sigma}(a) = \Phi_{0,1}\left(\frac{a-\mu}{\sigma}\right).$$

Sei $a_{\mu,\sigma}(p)$ *für* $0 < p < 1$ *dadurch definiert, dass*

$$\Phi_{\mu,\sigma}(a_{\mu,\sigma}(p)) = p \text{ (also } a_{\mu,\sigma}(p) = \Phi^{-1}_{\mu,\sigma}(p) \text{).}$$

Dann hat man

$$(2)\quad \alpha_{\mu,\sigma}(p) = \mu + \sigma\cdot\Phi^{-1}_{0,1}(p),$$
$$(3)\quad \alpha_{0,1}(p) = \Phi^{-1}_{0,1}(p).$$

Also: Für eine $(\mu,\sigma)-$ *normalverteilte Zufallsvariable* X *gilt:*

$$(4)\quad P(X \le a) = \Phi_{0,1}\left(\frac{a-\mu}{\sigma}\right).$$

(5) *Das Vertrauensintervall zur Wahrscheinlichkeit* α *symmetrisch um* μ *für die* $(\mu,\sigma)-$ *Normalverteilung lautet :* $\mu \pm \sigma\cdot\Phi^{-1}_{0,1}\left(\alpha + \frac{1-\alpha}{2}\right)$

Bemerkung: Der Parameter σ bedeutet eine Streckung längs der $x-$ Achse mit Faktor σ und eine Stauchung längs der $y-$ Achse mit Faktor σ. Die Dichte sieht glockenförmig aus, der Gipfel liegt bei $x = \mu$, die Streuung ist der Abstand zwischen Mittelwert und den Wendepunkten. Die Verteilung ist symmetrisch um μ.

Die Folgerung ergibt sich direkt aus Streckung und Verschiebung sowie der Symmetrie. Als Beispiel zeigen wir (2) :

$$\Phi_{\mu,\sigma}\left(\mu+\sigma\cdot\Phi^{-1}_{0,1}(p)\right) = {}^{\text{mit (1)}}\Phi_{0,1}\left(\frac{\mu+\sigma\Phi^{-1}_{0,1}(p)-\mu}{\sigma}\right) = \Phi_{0,1}\left(\Phi^{-1}_{0,1}(p)\right) = p, \text{ also}$$
$$a_{\mu,\sigma}(p) = \Phi^{-1}_{\mu,\sigma}\left(\Phi_{\mu,\sigma}\left(\mu+\sigma\cdot\Phi^{-1}_{0,1}(p)\right)\right) = \mu+\sigma\cdot\Phi^{-1}_{0,1}(p).$$

Nun zu (5) : Zu zeigen ist, dass $\Phi_{\mu,\sigma}\left(\mu+\sigma\cdot\Phi^{-1}_{0,1}\left(\alpha+\frac{1-\alpha}{2}\right)\right) - \Phi_{\mu,\sigma}\left(\mu+\sigma\cdot\Phi^{-1}_{0,1}\left(\alpha+\frac{1-\alpha}{2}\right)\right) = \alpha$. Man hat mit (2):

$$\Phi_{\mu,\sigma}\left(\mu+\sigma\cdot\Phi^{-1}_{0,1}\left(\alpha+\frac{1-\alpha}{2}\right)\right) = \alpha+\frac{1-\alpha}{2},$$

nun mit der Symmetrie der $\mu,\sigma-$ Normalverteilung um μ:

$$\Phi_{\mu,\sigma}\left(\mu-\sigma\cdot\Phi^{-1}_{0,1}\left(\alpha+\frac{1-\alpha}{2}\right)\right) = 1-\left(\alpha+\frac{1-\alpha}{2}\right),$$

wir bilden die Differenz:

$$\alpha+\frac{1-\alpha}{2}-1+\left(\alpha+\frac{1-\alpha}{2}\right) = \alpha.$$

Praktische Bemerkung: Man kann das Integral nicht elementar ausrechnen. Darum benötigt man eine Tabelle für die Standard-Normalverteilung, dann kann man die anderen auch berechnen. Mit einem Computeralgebraprogramm ist das nicht nötig, man bekommt alles numerisch Benötigte sofort für beliebige μ,σ. Aber es bleibt nützlich, ein paar Grundwahrscheinlichkeiten und entsprechende Grenzen für die Standard-Normalverteilung zu kennen, damit kann man Vieles im Kopf überschlagen. Hier ist eine solche Tabelle:

a	-3.1	-2.58	-2.33	-1.96	-1	1	1.96	2.33	2.58	3.1
$\Phi_{0;1}(a)$	0.001	0.005	0.01	0.025	0.159	0.841	0.975	0.99	0.995	0.999

Wichtigste Anwendungen:

1.) Wenn X eine $(\mu,\sigma)-$ normalverteilte Zufallsvariable ist, dann hat man z.B.

$$P(X > \mu+\sigma) = 1-\Phi_{0;1}\left(\frac{\mu+\sigma-\mu}{\sigma}\right) = 1-\Phi_{0;1}(1) = \Phi_{0;1}(-1) = 0.159.$$

2.) Wenn X eine $(100, 10)-$ normalverteilte Zufallsvariable ist, dann lautet das zweiseitige 95%− Vertrauensintervall für X, symmetrisch um $\mu = 100$:

$$100 \pm 10 \cdot 1.96.$$

Das ist also grob das Intervall $[80; 120]$. Man hat also um Bereich $\mu \pm 2\sigma$ etwas mehr als 95% aller Fälle. Das ist eine gern benutzte Faustregel. Ebenso macht man sich klar, wie groß dann erst die Anteile in den Bereichen $\mu \pm 3\sigma$, $\mu \pm 5\sigma$ sind. Im letzteren Fall fehlt nur noch weniger als ein Millionstel an 1.

3.) Es sei $p = 0.4$ eine 'Treffer'-Wahrscheinlichkeit im Einzelversuch. Für die Variable 'relative Trefferhäufigkeit bei 1000 unabhängigen Versuchen' hat man dann in sehr guter Näherung das zweiseitige 99%− Vertrauensintervall:

$$0.4 \pm \frac{\sqrt{0.4 \cdot 0.6}}{\sqrt{1000}} \cdot 2.33,$$

also wird mit Wahrscheinlichkeit 0.99 die beobachtete relative Häufigkeit der Treffer in folgendem Bereich liegen:

$$[0.363\,9; 0.4361].$$

Es sei bemerkt, dass man dies exakt auszurechnen hätte mit der entsprechenden Binomialverteilung. Das wäre aber sehr viel mühsamer und unnötig genau.

Abschließend wollen wir noch den wichtigen Zentralen Grenzwertsatz formulieren, der sehr allgemein aussagt, dass lange Summen unabhängiger Variablen näherungsweise normalverteilt werden unter gewissen Bedingungen an die Streuungen. Klar muss man ausschließen, dass eine der Variablen Streuung > 0 hat und alle anderen Streuung Null, um einen Extremfall zu nennen. Das besorgt die Lyapunov-Bedingung im folgenden Satz.

Satz 2 (Zentraler Grenzwertsatz). *Es sei $(X_k)_k$ eine Folge von Zufallsvariablen mit $\mu(X_k) = \mu_k$ und $\sigma(X_k) = \sigma_k$ und ferner existiere ein $\delta > 0$, so dass $\mu\left((X_k - \mu_k)^{2+2\delta}\right)$ existiert und folgende Lyapunov-Bedingung erfüllt ist:*

$$(L) \quad \lim_{n\to\infty} \frac{\sum_{k=1}^n \mu\left(|X_k - \mu_k|^{2+2\delta}\right)}{\left(\sum_{k=1}^n \sigma_k^2\right)^{1+\delta}} = 0,$$

dann geht die Verteilungsfunktion der Variablen

$$U_n := \frac{\sum_{k=1}^n (X_k - \mu_k)}{\sqrt{\sum_{k=1}^n \sigma_k^2}}$$

gegen $\Phi_{0,1}$, also die Verteilungsfunktion der Standard-Normalverteilung. Das heißt:

$$\lim_{n\to\infty} P(U_n \le a) = \Phi_{0,1}(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a e^{-\frac{1}{2}x^2} dx.$$

Bemerkung: Unsere Standard-Anwendung dieses Satzes ist die auf Variablen der Form $\overline{X}$. Dann hat man $X_k = X$ und $\mu_k = \mu$, $\sigma_k = \sigma$, ferner $\mu\left(|X_k - \mu_k|^{2+2\delta}\right) = \mu_\delta$, alle unabhängig von k, also

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n \mu\left(|X_k - \mu_k|^{2+2\delta}\right)}{\left(\sum_{k=1}^n \sigma_k^2\right)^{1+\delta}} = \lim_{n\to\infty} \frac{n\mu_\delta}{(n\sigma^2)^{1+\delta}} = \lim_{n\to\infty} \left(\frac{1}{n^\delta} \cdot \frac{\mu_\delta}{\sigma^{2+2\delta}}\right) = 0.$$

Mit dem Satz folgt nun im Spezialfall, dass

$$U_n = \frac{\sum_{k=1}^n (X_k - \mu_k)}{\sqrt{\sum_{k=1}^n \sigma_k^2}} = \frac{\sum_{k=1}^n (X_k - \mu)}{\sigma\sqrt{n}} = \frac{\sum_{k=1}^n X_k - n\mu}{\sigma\sqrt{n}}$$

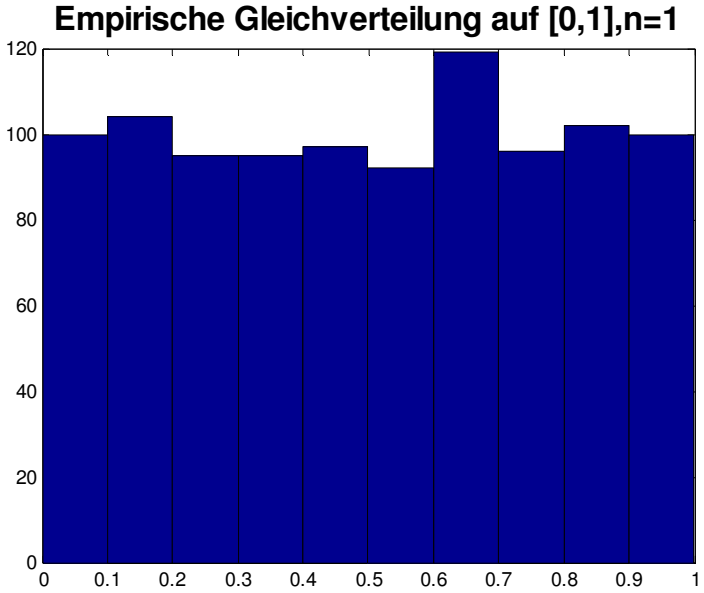
näherungsweise standard-normalverteilt ist. Nun haben wir

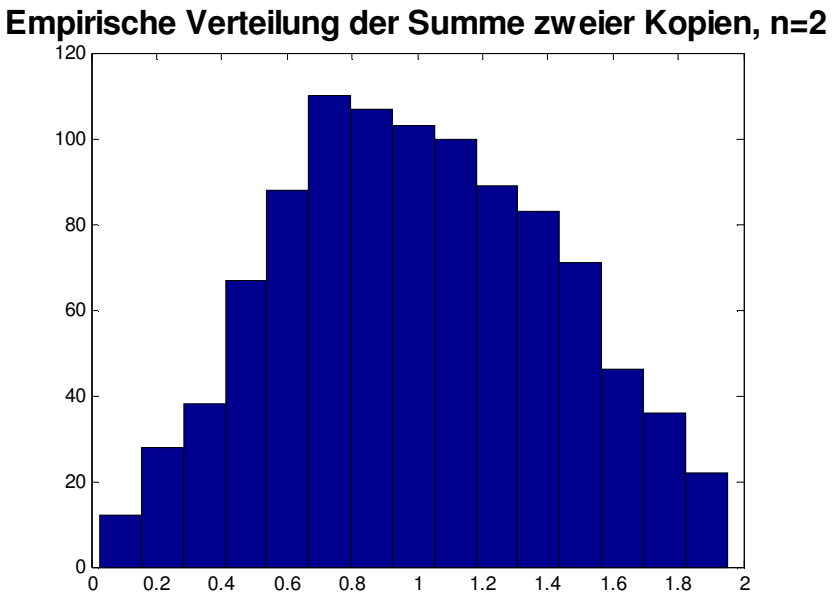
$$\frac{1}{n}\sum_{k=1}^n X_k = \frac{\sigma}{\sqrt{n}} U_n + \mu.$$

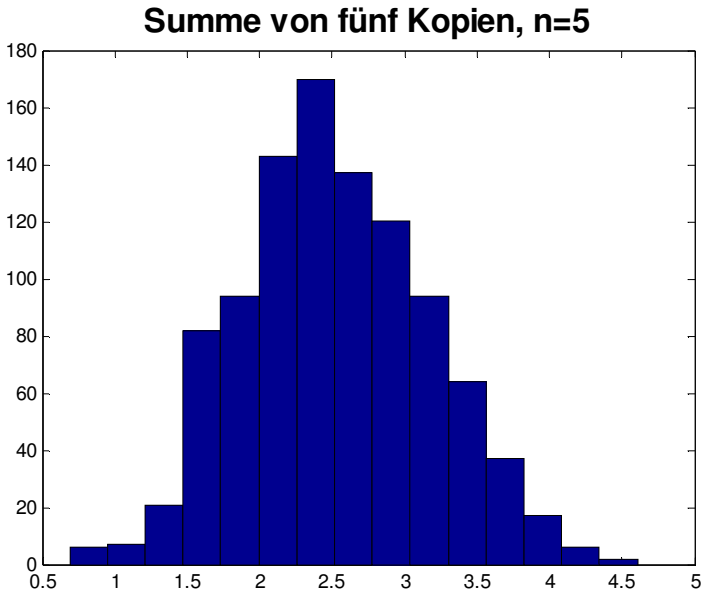
Da U_n näherungsweise $(0,1)-$*normalverteilt* ist, ist $\frac{1}{n}\sum_{k=1}^n X_k$ also näherungsweise $\left(\mu, \frac{\sigma}{\sqrt{n}}\right)-$*normalverteilt*.

Abschließend wollen wir noch illustrieren, dass auch die Verteilung von $\sum_{k=1}^n X_k$ mit unabhängigen Kopien X_k einer gleichverteilten Variablen auf $[0, 1]$ gegen eine Normalverteilung geht, und dies tun wir

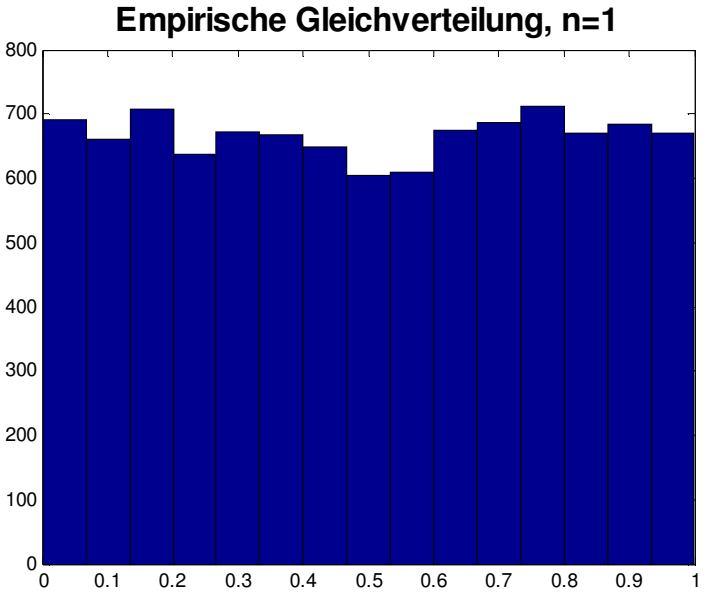
anhand empirischer Verteilungen, mit Stichprobenumfang 1000, um zugleich zu demonstrieren, dass sich die empirischen Verteilungen mit wachsendem Stichprobenumfang der exakten Verteilung nähern. Wir zeigen beide Umstände in Dichtefunktionsbildern - n ist jeweils der Index, für den $\sum_{k=1}^{n} X_k$ gebildet wird:

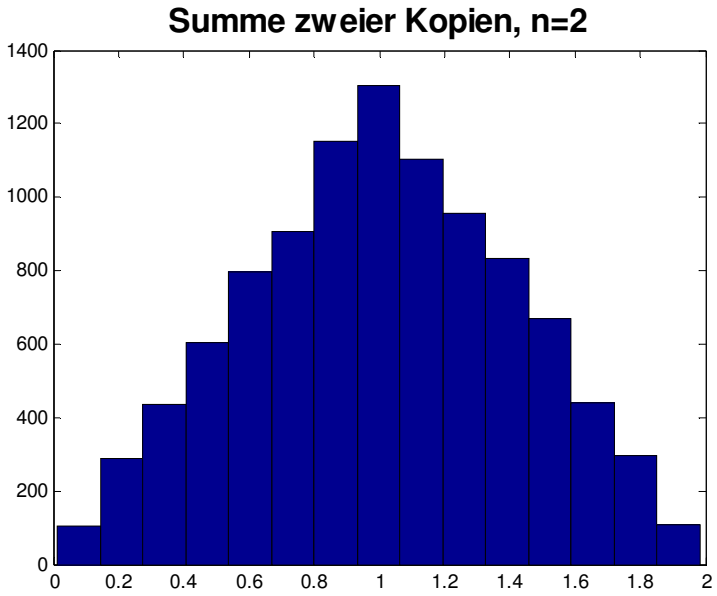


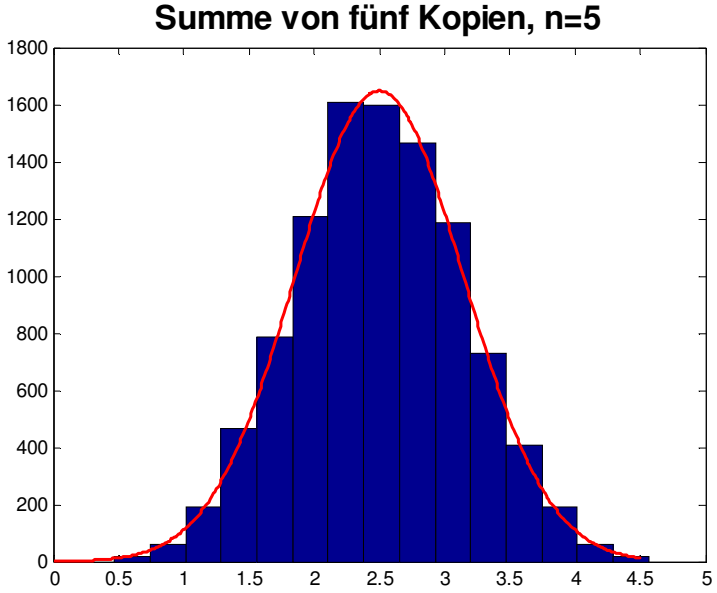




Bemerkung: Man sieht, wie aus der Gleichverteilung bereits mit der Summe von fünf Kopien die Normalverteilungs-Gestalt entsteht. Die Näherungen sind noch grob bei dem gewählten Stichprobenumfang 1000. Für $n = 2$ ergibt sich exakt eine Dreiecksverteilung. Wir zeigen dasselbe noch einmal mit Stichprobenumfang 10000 mit viel besserer Näherung an die mathematisch exakten Dichten:







Bemerkung: Im letzten Bild wurde mit passender Skalierung die zugehörige Normalverteilungs-Glocke mit eingezeichnet (rot). Dazu beachtet man $\mu = 5/2$ und $\sigma = \sqrt{5/12}$, da die Varianz der Gleichverteilung auf $[0, 1]$ lautet:

$$\sigma^2 = \int_0^1 x^2 dx - \frac{1}{4} = \frac{1}{12}.$$

5. Einige Anwendungen der Normalverteilung

5.1. Anwendung auf die Binomialverteilungen. Wenn wir wissen wollen, in welchem (möglichst kleinen) Bereich um 1000 die Anzahl der 'Treffer' bei 2000 Versuchen mit einer Wahrscheinlichkeit von mindestens 0.95 liegt, bei Trefferwahrscheinlichkeit $p = 1/2$, dann wäre es auch noch mit einem Computeralgebraprogramm sehr mühsam, diese Frage zu beantworten. Mit der Normalverteilung, welche die Parameter μ, σ mit der angesprochenen Binomialverteilung gemeinsam hat, ist das sehr leicht und fast im Kopf zu machen: Wir haben $\mu = 1000$, $\sigma = \sqrt{2000 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 10\sqrt{5}$ und nun bilden wir das

95%-Vertrauensintervall symmetrisch um $\mu = 1000$ für die $\left(1000, 10\sqrt{5}\right)-$ Normalverteilung, das ist

$$1000 \pm 1.96 \cdot 10 \cdot \sqrt{5},$$

die Grenzen sind also 956. 173 067 6 und 1043. 826 932 und wir haben das Intervall

$$[956; 1044],$$

da unsere binomialverteilte Variable nur ganzzahlige Werte zulässt. Das ist die *exakte Antwort* auf unsere Frage. Tatsächlich bestätigt man mit einem geeigneten Computeralgebraprogramm, dass für eine Zufallsvariable X, welche binomialverteilt ist mit $n = 2000$ und $p = \dfrac{1}{2}$ gilt:

$$\begin{aligned} P(956 \leq X \leq 1044) &= 0.9509 \text{ (gerundet) und} \\ P(957 \leq X \leq 1043) &= 0.9455 \text{ (gerundet).} \end{aligned}$$

Wir haben geringere Genauigkeit zu erwarten bei $p \neq 1/2$ und bei kleinerem n. Versuchen wir uns am Beispiel $n = 278$, $p = 0.1$. Dann bekommen wir für die Frage nach dem kleinsten Intervall um $\mu = 27.8$, das die Wahrscheinlichkeit 0.95 besitzt, mit der Normalverteilung:

$$27.8 \pm 1.96 \cdot \sqrt{278 \cdot 0.1 \cdot 0.9},$$

ganzzahlig also das Intervall

$$[17; 38]$$

Auch das sind die exakten Grenzen, welche für die Binomialverteilung gelten, unter der Voraussetzung, dass das Intervall so weit wie möglich symmetrisch zum Mittelwert liegt. Lassen wir diese Forderung beiseite, so können wir die untere Grenze durch 18 ersetzen. Auch das gibt die Näherung mit der Normalverteilung her, wenn man die sogenannte **Stetigkeitskorrektur** vornimmt:

BEMERKUNG 1 (Stetigkeitskorrektur). *Nähert man eine diskrete Verteilung einer Zufallsvariablen X, die nur ganzzahlige Werte annimmt, durch eine stetige Verteilung (deren Verteilungsfunktion sei F), z.B. eine Normalverteilung, dann werden die Resultate genauer, wenn man die Näherung mit folgender sogenannter Stetigkeitskorrektur für die Grenzen vornimmt:*

$$\begin{aligned} P(X \leq k) &\approx F\left(k + \frac{1}{2}\right), \\ P(X \geq k) &\approx F\left(k - \frac{1}{2}\right), \\ P(k_1 \leq X \leq k_2) &\approx F\left(k_2 + \frac{1}{2}\right) - F\left(k_1 - \frac{1}{2}\right). \end{aligned}$$

Für die Bildung von (nahezu um μ symmetrischen) Vertrauensintervallen zur Wahrscheinlichkeit α heißt das: Wenn $[a, b]$ das entsprechende Intervall zu F ist, also $F(b) - F(a) = \alpha$, dann setze für X das Intervall an: $[\alpha, \beta]$, mit

$$\begin{aligned} \alpha &= k_1, \text{ mit } k_1 := \text{ größte ganze Zahl mit } k_1 - \frac{1}{2} \leq a, \\ \beta &= k_2, \text{ mit } k_2 := \text{ kleinste ganze Zahl mit } k_2 + \frac{1}{2} \geq b. \end{aligned}$$

Erklärung: Man kann die exakte diskrete Verteilung mit einem Histogramm exakt wiedergeben, bei dem für jeden Wert k von X ein Histogrammkästchen gebildet wird mit den Grenzen $k - \frac{1}{2}, k + \frac{1}{2}$ und der Höhe $P(X = k)$. Dann nähert sich die Dichte $f = F'$ der Treppenfunktion. Nun ist die Wahrscheinlichkeit $P(k_1 \leq X \leq k_2)$ also näherungsweise der Flächeninhalt der vom Graphen von f mit der $x-$ Achse eingeschlossenen Fläche im Bereich der Histogrammkästen zu den Werten $k_1, k_1 + 1, ..., k_2$, und dazu gehört als untere Grenze $k_1 - \frac{1}{2}$, als obere $k_2 + \frac{1}{2}$. Ebenso erklären sich alle anderen Aussagen.

Beispiele zum Genauigkeitsvergleich mit und ohne Stetigkeitskorrektur:

1. Beispiel: Wir betrachten die Wahrscheinlichkeit für höchstens 6 mal 'Kopf' bei 10 Münzwürfen, also $P(X \leq 6)$ für eine binomialverteilte Variable X mit $n = 10$, $p = \frac{1}{2}$. Das gibt exakt

$$P(X \leq 6) = \sum_{k=0}^{6} \binom{10}{k} \left(\frac{1}{2}\right)^{10} = \frac{53}{64} = 0.828\,125.$$

Mit Näherung durch entsprechende Normalverteilung, $\mu = 5$, $\sigma = \frac{1}{2}\sqrt{10}$ bekommt man

bei Näherung ohne Stetigkeitskorrektur :

$$P(X \leq 6) \approx \Phi_{0,1}\left(\frac{6-5}{\frac{1}{2}\sqrt{10}}\right) = \Phi_{0,1}\left(\frac{2}{\sqrt{10}}\right) = 0.7365,$$

bei Näherung mit Stetigkeitskorrektur :

$$P(X \leq 6) \approx \Phi_{0,1}\left(\frac{6.5-5}{\frac{1}{2}\sqrt{10}}\right) = \Phi_{0,1}\left(\frac{3}{\sqrt{10}}\right) = 0.828\,609.$$

Man sieht sehr deutlich, wie viel besser das Resultat mit der Stetigkeitskorrektur wird. Natürlich werden die Unterschiede geringer bei größerem n, aber das ist kein vernünftiger Grund dafür, die Korrektur fallenzulassen, wenn n noch so groß ist (solche irrationalen 'Regeln' finden sich bei 'Praktikern', die sich nichts überlegen, sondern lieber irgendwelchen unsinnigen Traditionen folgen). Denn es kostet doch nichts, die Korrektur zu machen!

Bemerkung: Das Beispiel zeigt, dass die Näherung sogar noch in einem Bereich gute Resultate liefern kann, auch wenn man die Faustregel $np(a-p) \geq 5$ verletzt. Aber selbstverständlich würde man in solchen Beispielen exakt rechnen.

2. Beispiel:

Wir fragen nach der Wahrscheinlichkeit für mindestens 1700 Sechsen bei 10000 Würfel-Würfen. Nennen wir die Anzahl der Sechsen X, dann haben wir das exakte Ergebnis

$$P(X \geq 1700) = 0.1890.$$

Näherung durch Normalverteilung ergibt:

$$\textbf{ohne Stetigkeitskorrektur} \quad : \quad P(X \geq 1700) \approx 1 - \Phi_{0,1}\left(\frac{1699 - 10000/6}{\sqrt{10000 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) = 0.192\,81,$$

$$\textbf{mit Stetigkeitskorrektur}: \quad P(X \geq 1700) \approx 1 - \Phi_{0,1}\left(\frac{1699.5 - 10000/6}{\sqrt{10000 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) = 0.189\,16.$$

Man beachte: Mit der Korrektur wir der Näherungsfehler eine ganze Größenordnung kleiner als ohne Korrektur!

5.2. Anwendung der Normalverteilung auf die Schätzung von Mittelwerten, insbesondere Schätzung von Wahrscheinlichkeiten. Es geht hier um folgendes Problem: Wenn wir ein Experiment 10 mal durchführen, dabei das Ereignis A k mal auftritt, dann werden wir $P(A)$ durch die Zahl $k/10$ schätzen. Wenn wir das Experiment 100 mal durchführen und dabei A l mal auftritt, so wird $l/100$ als Schätzwert genommen. Wir werden erwarten, dass letzterer deutlich besser ist als ersterer. Doch das ist so offenbar falsch, wir könnten bei nur 10 Versuchen 'zufällig' auch ein viel besseres Resultat (also näher an $P(A)$ erzielt haben. Was können wir dazu **Genaueres** sagen, das nicht so falsch ist? Zunächst sollte man klären, wie eine vernünftige Aussage lauten würde. Sie kann nicht besagen, dass ein Schätzwert mit Sicherheit höchstens um $\varepsilon > 0$ vom tatsächlichen Wert $P(A)$ abweicht, sondern nur so:

$(*)$ 'Der Schätzwert $\dfrac{k}{n}$ für $P(A)$ liegt mit Wahrscheinlichkeit α höchstens um ε von $P(A)$ entfernt.'

Dabei kann man nun α (nahe bei 1 ist erwünscht) vorgeben und ε bestimmen. Man kann auch ε vorgeben und α bestimmen, beides bei vorgegebenem n. Für die Praxis ist es natürlich wichtig, α und ε vorzugeben, so dass die Aussage $(*)$ inhaltlich befriedigt, und zu fragen, wie groß man n wählen muss, um diese Sicherheit und Genauigkeit der Schätzung zu erreichen.

Beispiel: Wir geben $n = 100$ vor und fragen, wie groß ε wird bei $\alpha = 0.95$. Für die Variable $X =$ absolute Häufigkeit, mit welcher A bei 100 Versuchen eintritt, haben wir das zweiseitige Vertrauensintervall, bei Näherung durch Normalverteilung:

$$\mu \pm 2.58 \cdot \sqrt{100 \cdot p(1-p)}.$$

Dabei ist $\mu = 100 \cdot p$ und $p = P(A)$. Nun wissen wir nicht, wie groß p ist, also auch nicht, wie groß die Streuung $\sqrt{100 \cdot p \cdot (1-p)}$ ist. Aber wir wissen, dass diese Streuung maximal den Wert

$$\sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 5$$

annehmen kann. Daher liegt der Schätzwert für die mittlere absolute Häufigkeit $\mu = 100 \cdot p$ mit 95% Sicherheit im Bereich

$$\mu \pm 1.96 \cdot 5 \text{ oder } \mu \pm 9.8.$$

Wenn wir also $P(A)$ durch die beobachtete relative Häufigkeit schätzen, so haben wir für den Fehler ε der Schätzung, der mit 95% Sicherheit eingehalten wird:

$$\varepsilon \leq \frac{10}{100}.$$

Wenn wir nun bei 100 Versuchen 40 mal das Ereignis A sehen, so können wir sagen:

Die unbekannte Wahrscheinlichkeit $P(A)$ liegt mit mindestens 95% Sicherheit im Bereich $[0.3; 0.5]$.

Das ist keine sehr befriedigende Genauigkeit. Wir wollen daher fragen, welchen Stichprobenumfang n wir brauchen, um mit (wenigstens) 95% Sicherheit eine Genauigkeit von 2 Prozentpunkten (also $\varepsilon \leq 0.02$) zu erreichen. Dazu lösen wir die Gleichung - wir rechnen hier sofort mit der relativen Häufigkeitsgröße, deren Streuung entsprechend nur *höchstens* $\frac{\sqrt{1/4}}{\sqrt{n}}$ beträgt:

$$1.96 \cdot \frac{\sqrt{1/4}}{\sqrt{n}} = 0.02$$

und finden

$$n \geq 2401.$$

Das ist also der Stichprobenumfang, den wir für die Aussage mit der Genauigkeit 0.02 und der Sicherheit 0.95 benötigen. Steigern wir nun die geforderte Sicherheit auf 0.99 oder 99%, dann ist die Zahl 1.96 zu ersetzen durch 2.58, und es kommt (man rechne das nach):

$$n = 4161.$$

Ein zweites Beispiel: Man möchte bei einer Fabrikation von Schrauben der Soll-Länge 10 cm die Qualität so gut haben, dass die tatsächliche mittlere Länge der Schrauben um höchstens 0.1 mm von 10 cm abweicht. Kann man anhand einer Stichprobe diese Aussage mit 99% Sicherheit treffen? Das lässt sich offenbar nur ermitteln, wenn wir wie im vorigen Beispiel eine Obergrenze für die Streuung der Schraubenlänge L angeben können. Nehmen wir nun an, $\sigma(L) \leq 0.02$ mm. Dann hat man

$$\sigma\left(\overline{L}\right) = \frac{0.02}{\sqrt{n}}.$$

Wir nehmen nun an, dass in einer Stichprobe von 200 Schrauben gefunden wurde: $\overline{l} = 10.07$. ($\overline{l}$ als beobachteter Wert von $\overline{L}$). Dann haben wir $\mu(L)$ mit Sicherheit 0.99 im Intervall

$$10.05 \pm 2.58 \cdot \frac{0.02}{\sqrt{200}}, \text{ also im Bereich}$$
$$[10.066; 10.074].$$

Es kann also mit 99% Sicherheit gesagt werden, dass das Soll-Mittel um höchstens 0.074 mm verfehlt wird, somit im erwünschten Bereich. Wir können die Frage auch ein wenig modifizieren und fragen, mit welcher Sicherheit man noch sagen kann, dass der Mittelwertfehler unter 0.1 mm liegt. Dazu lösen wir

$$10.05 + x \cdot \frac{0.02}{\sqrt{200}} = 10.1$$

und finden $x > 35.35$. Das ist eine 'astronomisch große' Zahl für die $(0,1)-$ Normalverteilung. Wir sind praktisch sicher, dass das Soll-Mittel um höchstens 0.1 mm verfehlt wird. (Im Bereich $\mu \pm 11\sigma$ hat man mehr als $1 - 10^{-27}$ Wahrscheinlichkeit!)

Bemerkung: In beiden Beispielen benötigten wir eine Streuung, deren Wert unbekannt war, aber nach oben abgeschätzt werden konnte, einmal mit mathematischer Sicherheit, einmal (so angenommen)

aus Erfahrung begündet. Es liegt die Frage nahe, ob man die Information über die Streuung, welche aus der Stichprobe selbst zu entnehmen ist, für genauere Resultate benutzt werden könnte. Das ist so, und wir besprechen im nächsten Abschnitt, wie man das technisch durchführt.

6. Anwendung der $t-$ Verteilung auf die Mittelwertschätzung

Wir wollen für einen unbekannten Mittelwert $\mu(X)$ ein Vertrauensintervall zur Wahrscheinlichkeit w angeben. Dabei nutzen wir wiederum einen beobachteten Wert von $\overline{X}$. Dazu wollen wir bei unbekannter Streuung $\sigma(X)$ folgenden empirischen Schätzwert für σ verwenden:

$$s(X) := \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2}.$$

Dabei ist n der Stichprobenumfang, x_k der im k. Versuch beobachtete $X-$ Wert, $\overline{x}$ das beobachtete Stichprobenmittel.

Erläuterung: Es liegt zunächst nahe, σ^2 zu schätzen wie einen Mittelwert, es handelt sich doch um die mittlere quadratische Abweichung vom Mittelwert. Aber man stellt fest, dass der Erwartungswert von

$$\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2$$

gerade $\sigma^2(X)$ ist. Man kann sich auch klarmachen, dass $n = 1$ keinerlei Information über die Varianz geben kann. Somit würde man bei Nenner n statt $n-1$ die Streuung systematisch unterschätzen. (Die Schätzung hätte einen Bias, wie man sagt.)

Naiv würde man nun weiter so vorgehen, dass man $s(X)$ für die unbekannte Streuung $\sigma(X)$ in die Überlegungen des letzten Abschnitts einsetzt. (Übrigens wären die Resultate ab $n = 100$ etwa ziemlich gut.) Aber dabei ist zu überlegen, dass $s(X)$ eben nicht dasselbe ist wie $\sigma(X)$. Somit liegt es nahe, dass die Vertrauensgrenzen gegenüber der Normalverteilung ein wenig nach außen gerückt werden müssen, um die Möglichkeit zu kompensieren, dass $s(X)$ eben auch zu klein ausgefallen sein könnte. Die Verteilung, mit der man dabei genau genommen zu arbeiten hat, heißt (Student's) $t-$ Verteilung - der humorvolle Brite Gosset erfand sie, und er nannte sich selbst 'Student' -, und dazu hat man folgenden

SATZ 9 (Student's $t-$ Verteilung). *Wenn* $n \geq 2$, $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ *und* $S(X) = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2}$, *dann ist die Zufallsvariable*

$$T = \frac{X - \mu(X)}{S(X)}$$

$t-$ *verteilt mit* $n-1$ *Freiheitsgraden. Für die Anwendung interessiert vor allem:*

$$U = \frac{\overline{X} - \mu(X)}{S\left(\overline{X}\right)}, \text{ mit } S\left(\overline{X}\right) := \frac{S(X)}{\sqrt{n}},$$

ist $t-$ *verteilt mit* $n-1$ *Freiheitsgraden. Die zugehörige Dichtefunktion lautet*

$$f_{t,n-1}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \frac{1}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}.$$

Dabei ist Γ *die bekannte Gamma-Funktion, definiert durch*

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt, \text{ für } x \geq 0.$$

Sie hat die Eigenschaft, dass $\Gamma(n+1) = n!$ *für alle* $n \in \mathbb{N}_0$, *wie man über partielle Integration nachrechnet.*

Praktische Bemerkung: Man wird auch hier erwarten, dass man die Werte der Verteilungsfunktion einer Tabelle oder heute besser einem Computeralgebraprogramm entnimmt, da man das Integral

nicht elementar ausrechnen kann. Folgende Tabelle gibt Anhaltspunkte dafür, wie stark die $t-$ Vertrauensgrenzen gegen die der Standard-Normalverteilung verschoben sind - zum Vergleich sind diese mit hinzugegeben:

Wahrscheinlichkeit linksseitig von a		0.001	0.01	0.025	0.975	0.99	0.999
Grade ↓							
∞	Grenzzahl a für $\Phi_{0,1}$	-3.1	-2.33	-1.96	1.96	2.33	3.1
200	Grenzzahl für $F_{t,200}$	-3.131	-2.345	-1.972	1.972	2.345	3.131
100	Grenzzahl für $F_{t,100}$	-3.174	-2.364	-1.984	1.984	2.364	3.174
80	Grenzzahl für $F_{t,80}$	-3.195	-2.374	-1.990	1.990	2.374	3.195
60	Grenzzahl für $F_{t,60}$	-3.232	-2.390	-2.000	2.000	2.390	3.232
40	Grenzzahl für $F_{t,40}$	-3.307	-2.423	-2.021	2.021	2.423	3.307
20	Grenzzahl für $F_{t,20}$	-3.552	-2.528	-2.086	2.086	2.528	3.552
15	Grenzzahl für $F_{t,15}$	-3.733	-2.602	-2.131	2.131	2.602	3.733
10	Grenzzahl für $F_{t,10}$	-4.144	-2.764	-2.228	2.228	2.764	4.144
5	Grenzzahl für $F_{t,5}$	-5.893	-3.365	-2.571	2.571	3.365	5.893
2	Grenzzahl für $F_{t,2}$	-22.327	-6.965	-4.303	4.303	6.965	22.327

Bemerkung: Es ist zu beachten, dass bei Stichprobenumfang n die Zahl der Freiheitsgrade $n-1$ ist. Die Tabelle zeigt, dass die Unterschiede zur Normalverteilung bei 100 oder mehr Freiheitsgraden noch gering sind, bei 50 Freiheitsgraden schon merklich, unter 10 Freiheitsgraden drastisch. Mit einem Computeralgebraprogramm benötigt man keine Tabellen, sondern kann zu jeder Grenzzahl die Wahrscheinlichkeit (linksseitig) bekommen und umgekehrt zu jeder Wahrscheinlichkeit die zugehörige Grenzzahl, für alle Freiheitsgrade. Das sieht dann mit MAPLE etwa so aus, hier im Beispiel für 8 Freiheitsgrade:

$$\mathrm{TDist}\,(x;\,8) = 0.02$$

ergibt die Lösung: $x = -2.\,448\,984\,990$. Also für eine Variable X, welche $t-$ verteilt ist mit 8 Freiheitsgraden: $P\,(X \leq -2.\,448\,984\,990) = 0.02$.

In umgekehrter Richtung:

$$\mathrm{TDist}\,(-3\,;8) = 0.008536,$$

also $P\,(X \leq 3) = 0.008536$.

Anwendungsbeispiele:

1. Beispiel: Man hat bei 61 zufällig ausgewählten Werkstücken 6 Stücke mit Fabrikationsfehler entdeckt. Gesucht ist ein 95%− Vertrauensintervall für den Anteil der fehlerhaften Stücke in der Gesamtpopulation. Wir haben beobachtet - $\overline{X}$ ist die Zufallsvariable: Relative Häufigkeit (oder Anteil) der fehlerhaften Werkstücke unter 61 zufällig ausgewählten:

$$\begin{aligned} \overline{x} &= \frac{6}{61}, \\ s\left(\overline{X}\right) &= \sqrt{\frac{\frac{6}{61}\cdot\frac{55}{61}}{60}}. \end{aligned}$$

Wir wollen die zweite Aussage erklären: Man hat mit der Bernoulli-Variablen X, welche jedem Werkstück den Wert 1 für 'fehlerhaft' zuordnet, sonst den Wert Null, x_k ist der k. beobachtete Wert von X:

$$\begin{aligned} s^2\,(X) &= \frac{1}{60}\sum_{k=1}^{61}\left(x_k - \frac{6}{61}\right)^2 = \frac{61}{60}\left(\frac{1}{61}\sum_{k=1}^{61}\left(x_k - \frac{6}{61}\right)^2\right) \\ &= \frac{61}{60}\left(\frac{6}{61}\cdot\frac{55}{61}\right), \end{aligned}$$

daher

$$s\left(\overline{X}\right) = \frac{s\,(X)}{\sqrt{61}} = \frac{\sqrt{s^2\,(X)}}{\sqrt{61}} = \sqrt{\frac{\frac{6}{61}\cdot\frac{55}{61}}{60}}.$$

Bemerkung: Hier haben wir den Sonderfall, dass der beobachtete Mittelwert $\overline{x} = \dfrac{6}{61}$ unmittelbar erlaubt, $s\left(\overline{X}\right)$ zu bestimmen. Normalerweise ist die Streuung völlig unabhängig vom Mittelwert. (Vgl. das zweite Beispiel weiter unten.)

Also ist das gesuchte Vertrauensintervall:

$$\frac{6}{61} \pm \sqrt{\frac{\frac{6}{61} \cdot \frac{55}{61}}{60}} \cdot 2.000,$$

da wir aus der Tabelle oben für 60 Freiheitsgrade die Grenzzahlen $-2,2$ für die Wahrscheinlichkeiten $0.025, 0.975$ ablesen. Das Resultat: Mit 95% Sicherheit liegt der Anteil der fehlerhaften Werkstücke im Bereich

$$[0.02147; 0.1753].$$

2. Beispiel: Man hat bei 81 zufällig ausgewählten Schrauben einer Fabrikation die mittlere Länge 10.02 cm gefunden und den Streuungsschätzwert - l_k ist die Länge der k. Schraube der Stichprobe:

$$s(X) = \sqrt{\frac{1}{80} \sum_{k=1}^{81} (l_k - 10.02)^2} = 0.03.$$

Gesucht ist ein 99%$-$ Vertrauensintervall für die mittlere Länge der fabrizierten Schrauben. Das findet man sofort so:

$$10.02 \pm \frac{0.03}{\sqrt{81}} \cdot 3.195,$$

also liegt mit einer Sicherheit von 0.99 die mittlere Länge der fabrizierten Schrauben im Bereich

$$[10.00935; 10.03065].$$

Bemerkung: In diesem Beispiel sieht man, dass der empirische Mittelwert nichts über die Streuung aussagt und die Berechnung von $s(X)$ wie hier noch einmal mit Einsetzen in die Formel wiederholt gesondert auszuführen ist. Das war im ersten Beispiel mit den relativen Häufigkeiten anders.

6.1. Exponentialverteilungen. Zur inhaltlichen Deutung: Eine Variable T ist $\lambda-$ exponentialverteilt, wenn sie die Wartezeit bis zum ersten $\lambda - Poisson - Treffer$ ist. Wir zeigen nunmehr, dass gilt:

$$P(T \leq t) = \int_0^t \lambda e^{-\lambda s} ds = 1 - e^{-\lambda t}.$$

Die zugehörige Dichte ist damit:

$$f_\lambda(t) = \begin{cases} 0 \text{ für } t < 0, \\ \lambda e^{-\lambda t} \text{ für } t \geq 0 \end{cases}.$$

Man hat weiter:

$$\begin{aligned} \mu(T) &= \int_0^\infty t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}, \\ \sigma^2(T) &= \int_0^\infty t^2 \lambda e^{-\lambda t} dt - \mu^2(T) \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\ &= \frac{1}{\lambda^2}, \text{ also} \\ \sigma(T) &= \frac{1}{\lambda}. \end{aligned}$$

Klar bedeutet der Parameter λ, dass der Graph der Grundfunktion e^{-x} $(x > 0)$ mit Faktor λ längs der $x-$ Achse gestaucht und mit Faktor λ längs der $y-$ Achse gestreckt ist.

Bemerkung: Der Erwartungswert $1/\lambda$ versteht sich von selbst: Erwartet man λ Poisson-Treffer pro Zeiteinheit, so sollte intuitiv die mittlere Wartezeit bis zum ersten Treffer $1/\lambda$ (in der entsprechenden Zeiteinheit) sein.

Begründung der Verteilungsfunktion: Wir haben für die Verteilungsfunktion $F(t) := P(T \leq t)$ (wir setzen dabei $\Delta t > 0$):

$$F(t + \Delta t) - F(t) = P(t < T \leq t + \Delta t) = (1 - P(T \leq t)) \cdot \left(1 - e^{-\lambda \Delta t}\right).$$

Die erste Gleichung ergibt sich aus der Definition für F. Die zweite Gleichung kommt so: Das Ereignis $t \leq T \leq t + \Delta t$ bedeutet: Im Zeitintervall $[0, t]$ findet kein 'Treffer' statt, und im Zeitintervall $(t, t + \Delta t]$ findet mindestens ein Treffer statt (d.h. 'nicht kein Treffer'). Aus der Poissoneigenschaft folgt, dass diese Ereignisse unabhängig sind, für die Wahrscheinlichkeit der Konjunktion also zu multiplizieren ist. Das erste hat die Wahrscheinlichkeit $1 - P(T \leq t)$ nach Definition, das zweite hat aber die Wahrscheinlichkeit $1 - e^{-\lambda \Delta t}$, weil die erwartete Trefferzahl für das Zeitintervall der Länge Δt gerade $\lambda \Delta t$ ist, somit gemäß Poisson-Wahrscheinlichkeitsformel:

$$P(\text{'kein Treffer im Zeitintervall der Länge } \Delta t\text{'}) = e^{-\lambda \Delta t} \frac{\lambda^0}{0!} = e^{-\lambda \Delta t}.$$

Die Wahrscheinlichkeit für mindestens einen Treffer ist also $1 - e^{-\lambda \Delta t}$. Aus der Gleichung

$$(2) \quad F(t + \Delta t) - F(t) = (1 - P(T \leq t)) \cdot \left(1 - e^{-\lambda \Delta t}\right) = (1 - F(t)) \cdot \left(1 - e^{-\lambda \Delta t}\right).$$

folgt aber nach Division durch Δt und Grenzwertbildung $\Delta t \to 0$:

$$(2) \quad F'(t) = \lambda (1 - F(t)).$$

Denn

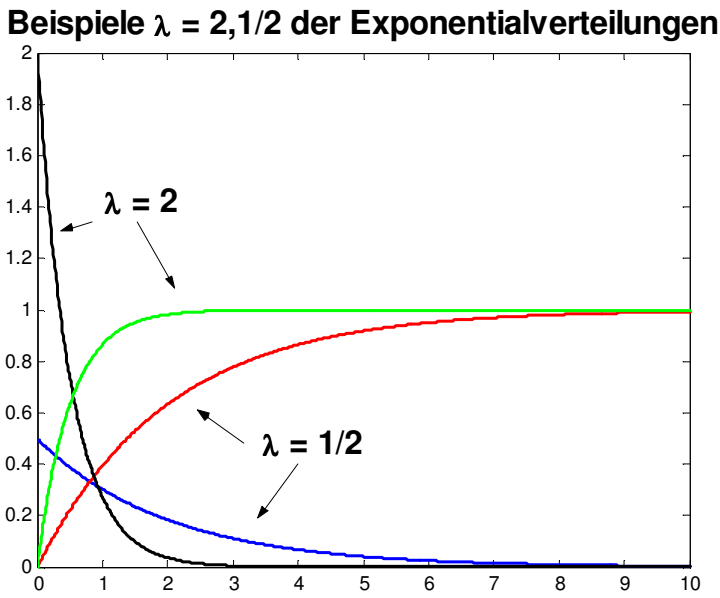
$$\begin{aligned} \lim_{\Delta t \to 0} \frac{1 - e^{-\lambda \Delta t}}{\Delta t} &= -f'(0) \text{ mit } f(t) = e^{-\lambda t}, \text{ aber} \\ f'(0) &= -\lambda. \end{aligned}$$

Nun lässt sich die DGL (2) mittels Separation ohne weiteres mit der Anfangsbedingung $F(0) = 0$ eindeutig lösen zu

$$F(t) = 1 - e^{-\lambda t} \text{ (für } t \geq 0, \text{ für } t < 0 \text{ ist der Wert Null).}$$

Ableiten ergibt die behauptete Dichtefunktion.

Hier ist ein Bild zu zwei Exponentialverteilungen, sowohl im Dichtebild (fallend) als mit der Verteilungsfunktion jeweils - die Werte sind Null für $t < 0$ - die Verteilungsfunktionen (ansteigend) nähern sich dem Wert 1 nur asymptotisch für $t \to \infty$.



7. Die χ^2- Verteilungen

Die technische Grundlage ist folgende

DEFINITION 21. *Eine Variable X ist genau dann χ^2- verteilt mit n Freiheitsgraden, wenn $X = \sum_{k=1}^{n} Y_k^2$ mit unabhängigen Variablen Y_k, die alle $(0, 1) -$ normalverteilt sind.*

Bemerkung: Die zugehörigen Dichtefunktionen lassen sich aus der Standard-Normalverteilungsdichte entwickeln über Faltung und dann numerisch integrieren. Wiederum hat jedes nennenswerte Computeralgebraprogramm alle zugehörigen Verteilungsfunktionen eingebaut.

Zum praktischen Auftreten dieser Variablen: Sehr oft läuft ein statistisches Problem darauf hinaus, zu prüfen, ob beobachtete Häufigkeiten mit einer theoretischen Annahme verträglich sind. Von dieser Art ist insbesondere das wichtige Problem: Hat eine Zufallsvariable X eine bestimmte Verteilung bzw. eine Verteilung bestimmten Typs? Nun liegt es nahe, die tatsächlich zu beobachtenden Häufigkeiten von Werten in den Intervallen $I_1, ..., I_s$ mit den laut theoretischer Verteilung zu erwartenden Häufigkeiten zu vergleichen. Man bildet dann einen empirischen Wert der Variablen

$$U = \sum_{k=1}^{s} \frac{(B_k - E_k)^2}{E_k}.$$

Dabei ist

$$E_k := nP_t\left(X \in I_k\right)$$

die laut theoretischer Verteilung zu erwartende absolute (!) Häufigkeit für das Intervall I_k, mit dem Stichprobenumfang n. $P_t\left(X \in I_k\right)$ ist die auf I_k entfallende Wahrscheinlichkeit laut theoretischer Verteilung. Ferner ist

$$B_k := \text{ Anzahl der Stichprobenwerte von } X \text{ in } I_k.$$

Man bildet die quadrierten Differenzen dieser absoluten Häufigkeiten und teilt durch die erwartete absolute Häufigkeit. Letzteres geschieht, um die Streuungen der Summanden auf 1 zu normieren. Damit ist die Variable U näherungsweise χ^2- verteilt. Was sind die Freiheitsgrade? Deren Zahl ist allenfalls $s-1$, weil die Häufigkeiten B_k und damit die $(B_k - E_k)^2$ nicht unabhängig sind, sondern nur $s-1$ sind unabhängig, der Wert des letzten Summanden steht dann fest, weil

$$\begin{aligned} B_s &= n - \sum_{k=1}^{s-1} B_k, \\ E_s &= n - \sum_{k=1}^{s-1} E_k. \end{aligned}$$

Wenn man einen Würfel auf Symmetrie testet, so ist $s = 6$ und $E_k = \frac{1}{6}n$ für alle $k = 1, ..., 6$. Dann hat man 5 Freiheitsgrade. Wenn man aber empirische Daten daraufhin testet, ob etwa eine Normalverteilungvorliegt, so nutzt man die Stichprobe auch, um μ und σ zu bestimmen. Analog braucht man zwei Parameter, um das Vorliegen einer Normalverteilung zu testen. (Für eine Binomialverteilung hätte man wiederum nur p zu bestimmen, einen Parameter, weil n bereits festliegt.) Nun hat man für jeden geschätzten Parameter jeweils einen Freiheitsgrad abzuziehen. In unserer Beispielfrage für eine Normalverteilung ergibt das also $s-3$ Freiheitsgrade.

In der folgenden Tabelle geben wir die Vertrauensgrenzen zu den Wahrscheinlichkeiten $0.95, 0.99, 0.999$ für verschiedene Freiheitsgrade:

$F_{\chi^2,n}(a)$	0.95	0.99	0.999	
a	3.8415	6.6349	10.8276	für $n = 1$
a	5.9915	9.2103	13.8155	$n = 2$
a	11.0705	15.0863	20.5150	$n = 5$
a	18.3070	23.2093	29.5883	$n = 10$
a	24.9958	30.5779	37.6973	$n = 15$
a	31.4104	37.5662	45.3147	$n = 20$
a	67.5048	76.1539	86.6608	$n = 50$
a	124.3421	135.8067	149.4493	$n = 100$
a	233.9943	249.4451	267.5405	$n = 200$

Beispiel: Findet man den Wert der $\chi^2 - verteilten$ Variablen U mit 15 Freiheitsgraden zu $u = 32$, dann besteht bereits gewisser Zweifel daran, dass die theoretische Verteilung auf die Daten passt. Denn in weniger als 1/100 der Fälle würde man (bei Zutreffen der theoretischen Verteilung) einen so hohen oder höheren Wert einer $\chi^2 - verteilten$ Variablen mit 15 Freiheitsgraden beobachten.

Das Beispiel der Vierfelder-Kreuztabellen: Die Daten sehen so aus - a, b, c, d sind die gefundenen absoluten Häufigkeiten, und es sind bereits die Randsummen eingetragen, die später benötigt werden:

	B	$\overline{B}$	
A	a	b	$m_1 = a + b$
$\overline{A}$	c	d	$m_2 = c + d$
	$n_1 = a + c$	$n_2 = b + d$	$N = a + b + c + d$

Typisch untersucht man dazu die Frage, ob A und B unabhängig sind, genauer testet man die Hypothese

$$H_0 : \text{ '}A \text{ und } B \text{ sind unabhängig'}.$$

Nun schätzt man die Wahrscheinlichkeiten $P(A)$ und $P(B)$ über die Randsummen, durch die empirischen relativen Häufigkeiten, also

$$\widehat{P(A)} = \frac{m_1}{N},\ \widehat{P(B)} = \frac{n_1}{N}.$$

Diese Schätzwerte implizieren solche auch für $P\left(\overline{A}\right)$ und $P\left(\overline{B}\right)$. Gemäß der Unabhängigkeitshypothese sind nun die erwarteten absoluten Häufigkeiten auszurechnen, wir stellen sie in folgender Tafel dar:

	B	$\overline{B}$
A	$\frac{m_1 n_1}{N}$	$\frac{m_1 n_2}{N}$
$\overline{A}$	$\frac{m_2 n_1}{N}$	$\frac{m_2 n_2}{N}$

Nach der zuvor gegebenen Regel haben wir dann zu bilden:

$$\chi^2 = \frac{\left(a - \frac{m_1 n_1}{N}\right)^2}{\frac{m_1 n_1}{N}} + \frac{\left(b - \frac{m_1 n_2}{N}\right)^2}{\frac{m_1 n_2}{N}} + \frac{\left(c - \frac{m_2 n_1}{N}\right)^2}{\frac{m_2 n_1}{N}} + \frac{\left(d - \frac{m_2 n_2}{N}\right)^2}{\frac{m_3 n_2}{N}}.$$

Dies ist ein beobachteter Wert einer näherungsweise χ^2- verteilten Variablen mit einem Freiheitsgrad; denn nach der Regel oben hat man in diesem Beispiel:

$$\begin{aligned} &\text{Zahl der Freiheitsgrade} \ = df \\ = \ &\text{Anzahl der 'Häuser' minus 1 minus Anzahl der geschätzten Parameter, hier} \\ = \ &4 - 1 - 2 = 1. \end{aligned}$$

In unserem Beispiel kann man χ^2 wesentlich bequemer berechnen, es gilt:

$$\chi^2 = \frac{(ad - bc)^2 N}{m_1 m_2 n_1 n_2}.$$

(Kleine Herausforderung an Ihre Fertigkeiten im Bruchrechnen: Rechnen Sie das nach.)

Tatsächlich wird die Näherung an die stetige χ^2- Verteilung deutlich besser, wenn man folgende Stetigkeitskorrektur nach Yates vornimmt:

$$\chi^2_{korr} = \begin{cases} \dfrac{\left(\left(a - \frac{1}{2}\right)\left(d - \frac{1}{2}\right) - \left(b + \frac{1}{2}\right)\left(c + \frac{1}{2}\right)\right)^2 N}{m_1 m_2 n_1 n_2}, \text{ wenn } ad > bc \text{ und dabei } \chi^2_{korr} < \chi^2 \ (\chi^2 \text{ wie oben}) \\ \dfrac{\left(\left(a + \frac{1}{2}\right)\left(d + \frac{1}{2}\right) - \left(b - \frac{1}{2}\right)\left(c - \frac{1}{2}\right)\right)^2 N}{m_1 m_2 n_1 n_2}, \text{ wenn } ad < bc \text{ und dabei } \chi^2_{korr} < \chi^2. \end{cases}$$

Bemerkungen:

1) Dass bei diesem Vorgang $\chi^2_{korr} > \chi^2$ wird, dafür kann man pathologische Beispiele bringen - allerdings würde kein vernünftiger Mensch in solchen Beispielen auf den Gedanken kommen, der empirische Befund könne gegen die Unabhängigkeitshypothese sprechen, so dass man erst gar keinen Test anstrengen würde.

2) Eine Faustregel: Die absoluten Häufigkeiten a, b, c, d sollten alle ≥ 5 sein, sonst sind die früher angesprochenen exakten Berechnungen mit einer hypergeometrischen Verteilung oder mit einer Differenz von zwei Binomialverteilungen leicht ausführbar und genauer.

Beispiel: Bei dem Befund

	B	$\overline{B}$
A	40	50
$\overline{A}$	40	25

bekäme man

$$\chi^2_{korr} = \frac{(40.5 \cdot 25.5 - 39.5 \cdot 49.5)^2 \cdot 155}{90 \cdot 65 \cdot 80 \cdot 75} \approx 3.758,$$

und wir bekommen

$$P\left(\chi^2_{df=1} \geq 3.758\right) = 2\Phi_{0,1}\left(-\sqrt{3.758}\right) \approx 0.053.$$

Man hat also Verwerfen auf $5\%-$ Niveau knapp verpasst. Aber man wird an der Unabhängigkeit zweifeln und ein signifikantes Resultat bei erhöhten Stichprobenumfängen erwarten. Außerdem sollte man stets daran denken, dass bei großen Stichproben auch die unbedeutendsten Unterschiede statistisch 'signifikant' werden, für die man sich kaum interessieren wird. Nur wenn $P(A|B)$ viel größer oder kleiner ist als $P\left(A|\overline{B}\right)$, dann wird das interessieren. Beispiel: Wenn betrunkene Autofahrer kaum schlechter Auto führen als nüchterne, könnte man getrost Alkohol am Steuer erlauben, so wie man auch nicht verlangt, ein Autofahrender dürfe kurz vor der Fahrt keinen Kuchen gegessen haben.

Bemerkungen:

1) Bei $n = 1$ erkennt man genau die Quadrate der (zweiseitigen) Vertrauensgrenzen zu den Wahrscheinlichkeiten der $(0, 1) -$ Normalverteilung, was klar sein sollte und oben im Beispiel illustriert wurde. Mit wachsender Zahl der Freiheitsgrade steigen die 'zulässigen' χ^2- Werte stark an.

2) Noch eine praktisch nicht unwichtige Bemerkung: Man sollte auch zu niedrigen χ^2- Werten misstrauen, dann kann man daran zweifeln, dass 'ehrliche Daten' vorliegen. (Man wird jemandem nicht glauben, der behauptet, er habe mit einem Würfel bei 120 Würfen genau je 20 Einsen, Zweien, ..., Sechsen gehabt, und damit den Würfel als symmetrisch erweisen will.)

8. Das Schema des statistischen Hypothesentestens, mit Beispielen

Dies Schema ist weit verbreitet, es gehört ein gewisser Jargon dazu, und es ist begleitet von einigen fulminanten Missverständnissen, die ebenfalls weit verbreitet sind. Daher darf eine Besprechung nicht fehlen im Rahmen auch der elementarsten Wahrscheinlichkeitsrechnung, die typische praktische Anwendungen berücksichtigen will. Der Rahmen, in den das Thema des Hypothesentestens gehört, heißt gewöhnlich 'Inferenzstatistik' oder 'schließende Statistik': Es geht darum, statistische Daten nicht einfach zu beschreiben ('deskriptive Statistik'), sondern nach Wahrscheinlichkeit (nicht mit logischer Sicherheit) auf zugrundeliegende Verhältnisse zu schließen.

8.1. Die Grundidee, und die verbreiteten Missverständnisse. Eine Hypothese H_0 soll mit statistischen Daten plus Wahrscheinlichkeitsrechnung getestet werden. Das funktioniert so:

Aus H_0 folgt eine Aussage über die Verteilungsparameter einer (oder mehrerer) Zufallsvariablen, oder H_0 ist selbst unmittelbar eine solche Aussage. Beobachtete Daten zeigen eine qualitative Abweichung von dem, was gemäß der genannten Aussage zu erwarten wäre. Man bildet das Ereignis A mit dem Inhalt: 'Eine mindestens so große Abweichung vom Erwarteten wie beobachtet tritt ein' und berechnet $P(A)$ unter Voraussetzung der Gültigkeit der Hypothese H_0 und natürlich der mathematischen Wahrscheinlichkeitsrechnung, manchmal auch nur unter zusätzlichen Modellannahmen, die natürlich einigermaßen begründet sein sollten. Wenn sich $P(A) < \alpha$ mit einer kleinen Zahl α erweist, wird man ungefähr so denken: 'Sollten wir gerade im Lotto gewonnen haben? Das ist nicht glaubhaft, also verwerfen wir die Hypothese.' Technisch sagt man: 'H_0 wird auf dem Niveau α verworfen.' Man findet häufig Formulierungen wie '$\alpha < 0.01$' oder '$\alpha = 0.01$' oder auch '$p = 0.001$', vielfach gibt man die Wahrscheinlichkeit an wie '$p = 3 \cdot 10^{-5}$' und überlässt die Beurteilung den Adressaten. Ob man sich dann bis auf weiteres praktisch so verhält, als sei H_0 falsch, hängt von der jeweiligen Sachproblematik und von der (auch) subjektiven Einschätzung der 'Kleinheit' von α in diesem Kontext ab.

Zum Jargon gehört auch, dass H_0 genannt wird: 'Nullhypothese'. Diese Bezeichnung ist ein ziemlich irreführender Unfug, jedenfalls nicht hilfreich. (Die rührt daher, dass man häufig nach einem 'Effekt' fragt und dann H_0 formuliert als: 'Es ist kein Effekt da, d.h. der Effekt hat das Maß Null.') Liest man also 'Nullhypothese', so denke man einfach daran, dass es die zu testende Hypothese ist. Hier schließt sich sogleich ein Missverständnis an, das stark verbreitet ist bei 'Anwendern': Man unterstellt eine Bedeutung der Hypothese im Sinne einer wissenschaftlich bereits gut unterstützten Vermutung, und dann nimmt man es als weitere Bestätigung der Hypothese, dass sie mit einem statistischen Test nicht verworfen wurde. Typisch findet man dann Formulierungen wie: 'Die Hypothese wird nicht verworfen, also angenommen (als wahr). Aber bei den alltäglich getesteten Hypothesen handelt es sich überhaupt nicht um bereits zu einem guten Teil bestätigte Vermutungen oder gar Aussagen, die von einer guten und bewährten Theorie begründet werden. Außerdem weiß man überhaupt nichts über den Wahrheitsgehalt einer Aussage, wenn man sie nicht widerlegen konnte. Entsprechend weiß man auch nicht, ob eine Aussage mit hoher Wahrscheinlichkeit wahr ist, wenn man sie mit einem Test nicht als 'wahrscheinlich falsch' erweisen konnte. Das gilt vor allem, wenn der Test nur eine schwache Datenlage (also etwa eine kleine Stichprobe) zur Verfügung hatte. Einen solchen Test kann auch eine sehr falsche Hypothese überstehen. Es kommt hinzu: Wenn H_0 auf dem Niveau α (etwa mit $\alpha = 1/1000$) verworfen wurde, dann ist α nur die Wahrscheinlichkeit dafür, dass man die Hypothese fälschlich verwirft, also auf dem Niveau α verwirft, obwohl sie wahr ist. α ist nicht etwa die Wahrscheinlichkeit dafür, dass die Hypothese wahr ist, so dass man mit der Sicherheit $1-\alpha$ sagen kann, H_0 sei falsch. Vielmehr kann man eine solche Aussage nur nach einer Bayesschen Analyse gewinnen, mit zusätzlciehn Annahmen. (Wir wissen, dass man $P(A|B)$ nicht mit $P(B|A)$ identifizieren darf.)

Das praktische Fazit: Wenn man eine Aussage H statistisch unterstützen will, so muss man die Hypothese $H_0 = \neg H$, also die Verneinung von H, in das Schema des statistischen Hypothesentestens hineinstecken. Eine Aussage der Form etwa: H :'$\mu(X) = \mu_0$' kann man nicht statistisch bestätigen, sondern nur mittels einer großen Stichprobe herausbekommen, dass sie mit einer gewissen großen Wahrscheinlichkeit näherungsweise richtig ist (unter Angabe des Fehlers, der mit dieser Wahrscheinlichkeit eingehalten wird). Aber eine Hypothese der Form H_0 :'$\mu(X) = \mu_0$' kann verworfen werden auf einem guten (d.h. kleinen) Niveau α.

8.2. Einseitige und zweiseitige Tests, und die zugehörigen Fragestellungen bzw. Hypothesen. Eine Hypothese der Form H_0 :'$Par(X) = q_0$' mit einem Verteilungsparameter $Par(X)$ ist zweiseitig in dem Sinne, dass Abweichungen nach beiden Seiten von q_0 zählen. Entsprechend besteht der empirische Verwerfungsbereich aus zwei Intervallen, die in gewissen Abständen links und rechts von q_0 liegen. Dagegen ist eine Hypothese der Form H_0 :'$Par(X) \leq q_0$' oder '$Par(X) \geq q_0$' einseitig - offensichtlich liegen die Verwerfungsbereiche nur jeweils oberhalb von q_0 bzw. unterhalb von q_0, wieder in gewissen Abständen von q_0. Man spricht von zweiseitigen bzw. einseitigen Tests.

Auch eine Unabhängigkeitshypothese wie H_0 :'A, B sind unabhängig' oder 'X, Y sind unabhängig' kann zweiseitig oder einseitig präzisiert werden, z.B. im ersten Fall zu H_0 :'$P(A|B) = P(A)$' oder zu '$P(A|B) \leq [\geq] P(A)$'. Dasselbe gilt für eine Hypothese zu einem Vergleich von Erwartungswerten wie H_0 :'$\mu(X) = [\leq, \geq] \mu(Y)$'.

8.2.1. *Die typisch auftretenden Missverständnisse und deren Vermeidung.* **1**) **Das erste Missverständnis** wurde bereits genannt: Eine Hypothese H_0 kann anhand des Datenmaterials nicht auf einem guten Niveau verworfen werden, und man denkt, damit sei H_0 bestätigt. Hier hilft es, wenn man sich klar macht, dass man kein Ergebnis hat, wenn man nicht verwerfen kann. Dann sollte man differenzieren: Wird das Verwerfen nur knapp verpasst bei bescheidenem Datenmaterial, so wird man mit reichhaltigerer Beobachtung hoffen, zum Verwerfen zu kommen. Oder aber man hofft, mit einer großen Stichprobe zeigen zu können, dass H_0 nicht in einem interessierenden Maße falsch ist. In diesem Fall ist es jedoch besser, das Hypothesentest-Schema zu verlassen und stattdessen für den fraglichen Parameter ein Vertrauensintervall mit hoher Wahrscheinlichkeit und goßer Genauigkeit (d.h. ein schmales Intervall) anzugeben (je nach Fragestellung kann dies Intervall wieder zweiseitig sein (zumeist in diesem Fall) oder einseitig (wenn praktisch nur eine Seite interessiert).

2) **Das zweite Missverständnis** bezieht sich auf die Relevanz der Aussage. Hinzu kommt der unglückliche Jargon 'Das Resultat ist signifikant' oder 'Die Daten weichen signifikant von der Hypothese ab'. Dabei muss man verstehen, dass 'signifikant' dasselbe heißt wie 'bedeutsam', dass aber mit diesem Jargon nur die **statistische Bedeutsamkeit** gemeint ist, also die Möglichkeit eines Schlusses nach Wahrscheinlichkeiten von den Daten auf die Hpothese (zu deren Verneinung genauer). Völlig ausgespart ist dagegen die Frage nach der **inhaltlichen Bedeutsamkeit** ('Signifikanz' in diesem Sinne also). Stellen Sie sich vor, dass jemand sagt: 'Maria kann mehr als 0 cm weit springen'. Alle würden lachen und wüssten, warum. Aber mit Bierernst werden andauernd solche Aussagen gemacht mit strenger statistischer 'Wissenschaftlichkeit'. Wie kann das passieren, und keiner lacht? Nun, vielen Menschen fehlt leider ein gewisses Bildungselement von quantitativem Denken. Dann ist eine Aussage wie 'Medikament a_1 ist besser als Medikament a_2' typisch, und man merkt nicht, dass alles darauf ankommt, **wie viel besser** Medikament a_1 als Medikament a_2 sei. Sie werden eine Fülle solcher Aussagen ohne jeden inhaltlichen Wert im alltäglichen Denken antreffen. Dazu gehören insbesondere 'besorgniserregende Tatsachen' wie 'Durch das Essen von ... wird das Krebsrisiko erhöht'. (Meist kommt dann etwas, das eine geringere Erhöhung bedeutet als ein einziger Flug oder der Aufenthalt in einem neuen Gebäude mit viel Granit.) Besonders häufig tritt dieser Unfug auf, wenn von abstrakteren Dingen wieKorrelationskoeffizienten die Rede ist: Die 'Anwender haben dann typisch keine Ahnung von der inhaltlichen Bedeutung eines solchen Koeffizienten und wissen insbesondere nicht, dass ein Korrelationskoeffizient $\rho(X, Y)$ zwischen den Variablen X und Y von $\rho(X, Y) = 0.1$ oder $\rho(X, Y) = -0.1$ so gut wie nichts bedeutet: Andauernd wird mit großem statistischem Aplomb eine Aussage der Form etabliert, dass $\rho(X, Y) \neq 0$ sei.

Wir wiederholen hier noch einmal ein drittes Missverständnis, das anderer Art ist: Wenn Vegetarier gesünder leben (durchaus auch quantitativ bedeutsam), dann liegt das nicht am Vegetarier-Sein, sondern an einer überhaupt gesünderen Lebensweise und daran, dass Leute mit einer guten sozialen Stellung und bewussten Gesundheitspflege eher Vegetarier sind als die anderen. Die Kenntnis von $Y - Modellen$ hilft, solche Missverständnisse zu vermeiden.

Wie vermeidet man das zweite Missverständnis? Zunächst erkennt man, dass an Aussagen wie 'besser als' oder 'Manager verdienen mehr als Arbeiter' nichts gelegen ist. Stattdessen formuliert man zweckmäßig bereits eine einseitige Hypothese wie H_0 :'$\mu(X) \leq \mu(Y) - a$' mit positiver Zahl $a > 0$. Die Verneinung (die mit statistischer Signifikanz belegt werden könnte durch Beobachtung) heißt dann '$\mu(X) > \mu(Y) + a$', und dann hat man nach Wahrscheinlichkeit das Resultat '$\mu(X)$ ist mindestens um a größer als $\mu(Y)$'. Zuvor überlegt man nach empirischer Kenntnis und praktischen Bedürfnissen, welcher Differenzbetrag a **inhaltlich bedeutsam** ist.

Der Witz bei der Sache ist der: Man muss statistisch nichts Raffinierteres tun, eine vernünftige Aussage zu stützen statt eine Aussage ohne jeden Wert: Es ist nur in die Formulierung der zu testenden Hypothese eine inhaltlich interessierende Differenz einzubauen nach dem gezeigten Muster. So kann man dann z.B. bestätigen, ein Korrelationskoeffizient sei mindestens 0.7 (und das wäre schon etwas).

8.2.2. *Anleitungen für Tests von typisch auftretenden einfachen Hypothesen.* **Vorbemerkung**: Es gibt eine große Fülle von Hypothesen (über Vergleich von zwei Varianzen, Korrelationskoefizienten, das Vorliegen einer bestimmten Verteilungsform usw., usw.) Dabei benötigt man ein gewisses (unter Umständen großes) Arsenal von Verteilungen (die technisch-mathematisch schwierig sind, aber mit Computerprogrammen leicht numerisch zu bedienen). Aber wenn man einmal die Sache grundsätzlich verstanden hat, ist es kein Problem, sich Weiteres mit einschlägiger Literatur zuänglich zu machen. (Vgl. Kreyszig, aber es gibt noch viel mehr.) Weiter sei bemerkt, dass die im vorigen Abschnitt besprochenen χ^2- Tests und zugehörige Fragestellungen ebenfalls Beispiele bilden, die man häufig antrifft. Dort hatten wir bereits darauf hingewiesen, dass eine Auskunft wie 'A, B sind abhängig' wieder eine sehr schwache Aussage ist und dass es viel mehr auf den Grad dieser Abhängigkeit ankommt, was man dann im Sinne der Bemerkungen wie zu '$\mu(X) \leq \mu(Y)$' besser anreichern würde zu einer quantitativen Aussage. Etwa: 'Die Wahrscheinlichkeit eines Unglücks unter Bedingung B ist (mit hoher Wahrscheinlichkeit) mindestens 0.4,während sie unter der Bedingung $\overline{A}$ bei 1/1000 liegt.' Dagegen hat man wieder, dass A und B nahezu unabhängig sind, wenn ein Test mit sehr hohem Stichprobenumfang bestanden wird.

Wir besprechen nur noch ein paar einfache Tests:

1) Einfache Hypothesen über Erwartungswerte:

H_0 :'$\mu(X) = \mu_0$': Hier sollte man besser für $\mu(X)$ ein gutes Vertrauensintervall angeben (schmal mit einer hohen Wahrscheinlichkeit) und zuvor ausrechnen, welchen Stichprobenumfang man dabei benötigt. Die verwandte Zufallsvariable ist $\overline{X}^{(n)}$.

H_0 :'$\mu(X) \leq \mu_0$', mit $a > 0$: Hat man $\overline{x}^{(n)} > \mu_0$ gefunden, so berechnet man (mittels Normalverteilung oder $t-$ Verteilung (bei geschätzter Streuung notwendig insbesondere für mäßige n)

$$P\left(\overline{X}^{(n)} \geq \overline{x}^{(n)}\right)$$

unter Voraussetzung von $^{*)}$ $\mu(X) = \mu_0$ (plus Streuungsschätzung oder Einsatz maximaler Streuung für X usw.). Fällt diese Wahrscheinlichkeit $< \alpha$ aus, kann man die Hypothese auf dem Niveau α verwerfen.

$^{*)}$ Der Grund dafür, dass man $\mu(X) = \mu_0$ setzen kann, obwohl die Hypothese das nicht festlegt, liegt darin: Wenn $\mu(X) < \mu_0$ wäre, so bekäme man für $P\left(\overline{X}^{(n)} \geq \overline{x}^{(n)}\right)$ einen noch kleineren Wert als im Falle $\mu(X) = \mu_0$. Man nimmt also den für H_0 günstigsten Fall an.

Bemerkung: Man bedenke, dass man bei einseitiger Hypothese vor der Durchführung des Tests die Richtung formuliert haben sollte, sonst hält man das Signifikanzniveau nicht ein. Sinngemäß ist abzuwandeln für '$\mu(X) \geq \mu_0$'. Ferner sollte die Verneinung von H_0 eine im Kontext inhaltlich interessierende Aussage sein, vgl. insbesonder Beispiel **2**), das nun folgt.

2 a) Hypothesen über einen Vergleich von $\mu(X)$ und $\mu(Y)$ bei Variablen X, Y **auf disjunkten Populationsteilen anhand von Stichproben der Umfänge** n_1 für X und n_2 für Y:

H_0 :'$\mu(X) = \mu(Y)$': Man formuliert gleichwertig um zu '$\mu\left(\overline{X}^{(n_1)} - \overline{Y}^{(n_2)}\right) = 0$' und geht dann am besten so vor wie unter **1**) im Fall der zweiseitigen Hypothese. Zu beachten ist, dass hier $\overline{X}^{(n_1)}$ und $\overline{Y}^{(n_2)}$ unabhängig näherungsweise normalverteilte Variablen sind, also ist $\overline{X}^{(n_1)} - \overline{Y}^{(n_2)}$ wieder näherungsweise normalverteilt mit Streuung $\sigma\left(\overline{X}^{(n_1)} - \overline{Y}^{(n_2)}\right) = \sqrt{\dfrac{\sigma^2(X)}{n_1} + \dfrac{\sigma^2(Y)}{n_2}}$. Mehr noch: Auch bei mäßigen Werten von n_1, n_2 macht man keinen großen Fehler, wenn man Streuungsschätzungen $s(X)$ und $s(Y)$ anhand der Stichproben einsetzt für $\sigma(X)$ und $\sigma(Y)$ und trotzdem einfach mit Normalverteilung rechnet (wenn n_1, n_2 kritisch klein werden, sollte man allerdings bei Flannery, Press 'Numerical Recipes' nachsehen und nicht den vielfach üblichen Unfug betreiben, dass zuerst mit falscher Logik (s.o.) aus dem Nichtverwerfen (selbstverständlich gerade bei den dann kleinen Stichprobenumfängen) von '$\sigma(X) = \sigma(Y)$' auf $\sigma(X) = \sigma(Y)$ schließen und dann den $t-$ Test unter Voraussetzung gleicher Streuungen machen. Merke: Wenn $\mu(X)$ wesentlich größer ist als $\mu(Y)$, so wird man tendenziell auch $\sigma(X)$ deutlich größer als $\sigma(Y)$ finden.

H_0 :'$\mu(X) \geq \mu(Y) + a$' mit $a > 0$, a 'interessant': Man formuliert um zu '$\mu\left(\overline{X}^{(n_1)} - \overline{Y}^{(n_2)}\right) \geq a$' und rechnet mit $\mu\left(\overline{X}^{(n_1)} - \overline{Y}^{(n_2)}\right) = a$ (vgl. die Bemerkung zu **1**), zweiseitiger Fall) und der Streuung wie oben $\left(\sqrt{\frac{\sigma^2(X)}{n_1} + \frac{\sigma^2(Y)}{n_2}}\right)$ die Wahrscheinlichkeit aus:

$$P\left(\overline{X}^{(n_1)} - \overline{Y}^{(n_2)} < \overline{x}^{(n_1)} - \overline{y}^{(n_2)}\right),$$

für den Fall natürlich nur, dass $\overline{x}^{(n_1)} - \overline{y}^{(n_2)} < a$ ausgefallen ist. (Sonst hat man keinen empirischen Befund, der gegen die Hypothese sprechen könnte, und es ist nichts mehr zu rechnen!) Ist diese Wahrscheinlichkeit $< \alpha$, so kann man H_0 auf dem Niveau α verwerfen. Für '$\mu(X) + a \leq \mu(Y)$', $a > 0$, braucht man offenbar nur X, Y oben zu vertauschen.

2 b) Hypothesen zu einem Vergleich zwischen $\mu(X)$ und $\mu(Y)$ bei abhängigen Variablen X, Y. Typische Beispiele sind von der Art: X ist eine Messung der Leistung von Individuen vor einem Training, Y die Leistung nach dem Training, oder auch: X ist die Leistung mit einer bestimmten Messmethode gemessen, Y dieselbe Leistung mit einer anderen Methode gemessen (es könnten auch subjektiv beeinflusste Bewertungen durch zwei verschiedene Individuen sein), usw.,usw.

In diesem Fall wird man stets Daten haben der Form $(X(\omega_i), Y(\omega_i)) = (x_i, y_i)$, $1 \leq i \leq n$. Nun kann man nicht wie in **2 a**) mit $\overline{X}^{(n)} - \overline{Y}^{(n)}$ rechnen, weil diese Variablen abhängig sind. Stattdessen benutzt man die Testvariable $\overline{X - Y}^{(n)}$. Für die Streuung ist dann auch $\sigma\left(\overline{X - Y}^{(n)}\right)$ zu bilden und zu schätzen durch $s\left(\overline{X - Y}^{(n)}\right)$. Anschließend ist $t-$ Verteilung zu benutzen. Sonst läuft alles analog zu **2 a**).

2 b (i) : Noch eine letzte Bemerkung: Man kann für eine Frage wie 'Ist $\mu(X)$ größer als $\mu(Y)$?' im Falle der zuletzt besprochenen Beispiele abhängiger Zufallsvariablen auch einen einfachen Vorzeichentest durchführen, der allerdings nicht alle Information der Daten benutzt, den sogenannten **Vorzeichentest**: Dann zählt man nur die gefundenen Fälle mit $x_i > y_i$ und $x_i \neq y_i$. Diese sei nun $k > n/2$, mit dem Stichprobenumfang n. Mit H_0 :'$\mu(X) \leq \mu(Y)$' und normalverteilten X, Y hätte man die Konsequenz für den für die Hypothese günstigsten Fall $\mu(X) = \mu(Y)$, dass $p = P(X > Y) = 1/2$. Dann rechnet man für die Variable $U =$ Anzahl der gefundenen Vorzeichen näherungsweise mit Binomialverteilung zu $n =$ Stichprobenumfang und $p = 1/2$ aus: $P(U \geq k)$ .Ist diese Wahrscheinlichkeit $< \alpha$, so kann man die Hypothese '$\mu(X) \leq \mu(Y)$' verwerfen auf dem Niveau α. Zusätzlich könnte man auch bereits eine interessante Differenz etablieren, indem man H_0' :'$P(X \geq Y) \leq 1/2 + a$' verwerfen kann. Aber man hätte eben nicht die gesamte Information über $\mu(X) - \mu(Y)$ ausgeschöpft, dafür einen simpel und schnell auch von Hand durchführbaren Test. Mit dem Computer kann man jedoch den Test wie oben in **2 b**) ebenso schnell ausführen, so dass man ein Vorzeichentest nur noch für das unmittelbare Überschlagen mit 'Rechnen von Hand' von Bedeutung ist. Andererseits hat man mit der schnellen Durchführbarkeit raffiniertester und kompliziertester Tests auf dem Computer großen möglichen Gewinn, aber neue Gefahren, dass inhaltlich 'der Wald vor lauter Bäumen' nicht mehr gesehen wird und die Interpretation zu kurz kommt oder auch falsch wird.

KAPITEL 4

Zusammenfassung der wichtigsten Begriffe und Resultate sowie Formeln

Grundbegriffe und Formeln der elementaren Wahrscheinlichkeitsrechnung
P ist eine Funktion, welche jedem Ereignis A seine Wahrscheinlichkeit zuordnet,
also $P(A)$ = Wahrscheinlichkeit dafür, dass A eintritt.
Axiomatisch wird gefordert, dass
(i) $0 \leq P(A) \leq 1$,
(ii) $P(\Omega) = 1$,
(iii) $P\left(\bigcup_{k=0}^{\infty} A_k\right) = \sum_{k=0}^{\infty} P(A_k)$, *falls* $A_i \cap A_j = \emptyset$ für $i \neq j$, $i, j \in \mathbb{N}$.
(Es folgt $P\left(\overline{A}\right) = 1 - P(A)$.)
Die Ereignisse A, B heißen **unabhängig** : $\Longleftrightarrow$ $P(A)\,P(B) = P(A \cap B)$
Bedingte Wahrscheinlichkeit: $P(A|B) := \frac{P(A \cap B)}{P(B)}$, für $P(B) \neq 0$.
Anschauliches Verständnis: $P(A|B)$ = Wahrscheinlichkeit für A im 'Topf' B (als neuem Ω).
Es folgt: Wenn $P(B) \neq 0$, dann: (A, B unabhängig $\Longleftrightarrow$ $P(A|B) = P(A)$).
Weiter folgen:
(i) **Formel von der totalen Wahrscheinlichkeit**: Wenn $\bigcup_{i=1}^{n} B_i = \Omega$ und $B_i \cap B_j = \emptyset$ für $i \neq j$,
dann $P(A) = \sum_{i=1}^{n} P(A|B_i)\,P(B_i)$
(ii) **Bayessche Formel**: $P(B|A) = \dfrac{P(A|B)\,P(B)}{P(A|B)\,P(B) + P\left(A|\overline{B}\right)P\left(\overline{B}\right)}$

Zufallsvariablen und ihre Verteilungen, allgemein
Definition: Eine Zufallsvariable X ist eine Abbildung $X : \Omega \to \mathbb{R}$, so dass von allen Ereignissen der Form $X \leq a$ usw. die Wahrscheinlichkeiten bildbar sind.
Definition: Die Verteilungsfunktion von X ist die Funktion $F_X(a) = P(X \leq a)$ für alle $a \in \mathbb{R}$.
Definition: X ist mit Dichte f verteilt : $\Longleftrightarrow$ für alle $a \in \mathbb{R}$: $\left(P(X \leq a) = \int_{-\infty}^{a} f(x)\,dx\right)$.
(Dann ist nach dem Hauptsatz $F_X' = f$.)
Für diskrete Verteilungen hat man einfach $F_X(a) = \sum_{b \leq a} P(X = b)$.
Die Verteilungsparameter Erwartungswert und Varianz bzw. Streuung
$\mu(X) := \sum_{a \text{ Wert von } X} a \cdot P(X = a)$ im diskreten Fall, also höchstens abzählbar unendlich viele Werte von a (im unendlichen Fall muss die Reihe konvergieren, was sie in den praktisch relevanten Fällen tut).
$\mu(X) := \int_{-\infty}^{\infty} x \cdot f(x)\,dx$ für X mit Dichte f verteilt, vorauszusetzen ist die Existenz des Integrals.
$\sigma^2(X) := \mu\left((X - \mu(X))^2\right)$ $(= \mu(X^2) - \mu^2(X))$, (wenn Summe bzw. Integral existieren).
$\sigma(X) = \sqrt{\sigma^2(X)}$
Es gelten die wichtigen Formeln (geichgültig, wie X, Y **verteilt sind** bis auf die Voraussetzung der Unabhängigkeit bzw. linearen Unabhängigkeit bei $\sigma(X+Y)$**:**
$\mu(\alpha X + \beta Y) = \alpha\mu(X) + \beta\mu(Y)$,
$\sigma(X+Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)}$, **wenn** X, Y unabhängig sind, d.h. alle Ereignispaare $(X \leq a, Y \leq b)$ unabhängig. (Es genügt aber schon $Cov(X,Y) = \mu((X - \mu(X))(Y - \mu(Y))) = 0$.)
$\sigma(\alpha X + \beta) = \sigma(\alpha X) = |\alpha|\,\sigma(X)$.

Die wichtigsten diskreten Verteilungen:
1. Binomialverteilungen:
X ist genau dann $(n,p)-$*binomialverteilt*, wenn
$X=$ Trefferzahl bei n unabhängigen Versuchen mit Wahrscheinlichkeit
p für einen Treffer im einzelnen Versuch.
Es gilt dann: $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ für $k=0,...,n$
(Also $F_X(a)=\sum\limits_{k\leq a}P(X=k)$.)
$\mu(X)=np$, $\sigma(X)=\sqrt{n\cdot p\cdot(1-p)}$

2. Hypergeometrische Verteilungen:
X ist genau dann $(N,K,n)-$*hypergeometrisch* verteilt, wenn
$X=$ Zahl der gezogenen 'Trefferkugeln', wenn aus einer Urne mit N Kugeln,
darunter K 'Trefferkugeln', auf einmal n Kugeln herausgezogen werden.
Es gilt dann: $P(X=k)=\dfrac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$ für $k=0,...,n$.
Ferner $\mu(X)=\frac{nK}{N}$, $\sigma^2(X)=n\frac{K}{N}\left(1-\frac{K}{N}\right)\frac{N-n}{N-1}$ (man beachte: 'σ^2' ist das).

3. Poisson-Verteilungen:
X ist $\lambda-Poisson-verteilt:\Longleftrightarrow$ pro raumzeitlicher Einheit werden
λ Treffer erwartet ($\lambda>0$), unabhängig von dem, was in anderen raumzeitlichen
Einheiten passiert oder passiert ist.
(Statt 'raumzeitlich' hat man oft nur 'zeitlich'.)
Es gilt dann: $\mu(X)=\lambda$, $\sigma(X)=\sqrt{\lambda}$.

4. Folgende Verteilung mit Dichte ist eng mit der $\lambda-Poisson-Verteilung$
verknüpft: Sei $T=$ Wartezeit bis zum ersten $\lambda-Poisson-Treffer$.
Dann gilt: T ist $\lambda-$ exponentialverteilt, d.h. T hat die Dichte
$f_\lambda(t)=\lambda e^{-\lambda t}$ für $t\geq 0$, $f_\lambda(t)=0$ für $t<0$.
Es folgen: $F_T(a)=P(T\leq a)=\int_0^a\lambda e^{-\lambda t}dt=1-e^{-\lambda t}$, für $a\geq 0$,
$\mu(T)=\frac{1}{\lambda}$, $\sigma(T)=\frac{1}{\lambda}$

Bedingte Verteilungen und entsprechend
bedingte Erwartungswerte sowie Streuungen:
Sei X eine mit Dichte f verteilte Zufallsvariable. Sei A ein Ereignis
(typisch $X\leq a_0$ oder $X\geq a_0$),
dann ist $f_{X|A}(x):=\frac{1}{P(A)}f(x)$.
Es folgt: $\mu(X|A)=\int_A x\cdot\frac{1}{P(A)}f(x)\,dx=\frac{1}{P(A)}\int_A x\cdot f(x)\,dx$,
d.h. für $A=(X\leq a_0)$ z.B.
$\mu(X|X\leq a_0)=\frac{1}{P(X\leq a_0)}\int_{-\infty}^{a_0}x\cdot f(x)\,dx$.
(Usw., entsprechend abzuwandeln für $A=X\geq a_0$ etc.)
Analog: $\sigma^2(X|X\leq a_0)=\frac{1}{P(X\leq a_0)}\int_{-\infty}^{a_0}(x-\mu(X|A))^2\cdot f(x)\,dx$
$=\frac{1}{P(X\leq a_0)}\left(\int_{-\infty}^{a_0}x^2\cdot f(x)\,dx-\mu^2(X|A)\right)$.

Analog zur Formel von der totalen Wahrscheinlichkeit
hat man für Erwartungswerte:
Sei $\bigcup\limits_{i=1}^{n}A_i=\Omega$ und $A_i\cap A_j=\emptyset$ für $i\neq j$, seinen die **bedingten Erwartungswerte**
$\mu(X|A_i)$ bekannt, $P(A_i)$ bekannt, dann $\mu(X)=\sum\limits_{i=1}^{n}\mu(X|A_i)\,P(A_i)$.

Die Normalverteilungen:
Definition: Sei $\mu \in \mathbb{R}, \sigma > 0$, dann ist die $(\mu,\, \sigma) - Normalverteilung$
die Verteilung mit der Dichte $\varphi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$,
die $(0,1) - Normalverteilung$ nennt man auch **Standard-Normalverteilung.**
Die Verteilungsfunktion der $(\mu,\sigma) - Normalverteilung$ ist dann also
$\Phi_{\mu,\sigma}(a) = \int_{-\infty}^{a} \varphi_{\mu,\sigma}(x)\, dx = \Phi_{0,1}\left(\frac{x-\mu}{\sigma}\right)$ (letzterer Wert aus Tabelle abzulesen!)
Ist also X normalverteilt mit μ, σ, dann hat man:
$\mu(X) = \mu$, $\sigma(X) = \sigma$, und $F_X(a) = P(X \leq a) = \Phi_{0,1}\left(\frac{a-\mu}{\sigma}\right)$.

Bildung eines zweiseitigen Vertrauensintervalls symmetrisch um μ
zur Wahrscheinlichkeit α (gewöhnlich nahe bei 1) **für eine Variable** X, die
$(\mu,\sigma) - normalverteilt$ ist:
$\mu \pm \Phi_{0,1}^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) \cdot \sigma$

Was für Variablen sind normalverteilt oder näherungsweise normalverteilt?
1. Summen unabhängiger normalverteilter Variablen sind (exakt) normalverteilt.
2. Variablen der Form $\sum_{i=1}^{n} X_i$ mit X_i unabhängige Wiederholung einer festen Variablen X
sind (schon bei mäßigem n**) näherungsweise normalverteilt.**
Anwendung von **2.** auf die Binomialverteilungen: Sei $np(1-p) \geq 5$, dann hat man
$P(X \leq a) \approx \Phi_{0,1}\left(\frac{a+0.5-np}{\sqrt{np(1-p)}}\right)$ (Stetigkeitskorrektur ist angebracht!)
3. Variablen der Form $\overline{X}^{(n)} := \frac{1}{n}\sum_{i=1}^{n} X_i$ mit X_i unabhängige Wiederholung
einer festen Variablen X sind näherungsweise normalverteilt, man rechnet entsprechend
mit $\mu\left(\overline{X}^{(n)}\right) = \mu(X)$ und $\sigma\left(\overline{X}^{(n)}\right) = \frac{\sigma(X)}{\sqrt{n}}$.
Achtung: X ist dabei in typischen Anwendungen keineswegs binomialverteilt,
sondern hat recht dicht liegende Werte, dann also keine sinnlose 'Stetigkeitskorrektur'!
(Selbst im Falle einer Binomialverteilung bei X wäre entsprechend anzupassen!)

a) Zweiseitiges Vertrauensintervall zur Wahrscheinlichkeit α für $\mu(X)$ anhand einer beobachteten Stichprobe von unabhängig genommenen $X-$ Werten x_i mit empirischem Mittel
$\overline{x}^n := \frac{1}{n}\sum_{i=1}^{n} x_i$ (beobachteter Wert von $\overline{X}^{(n)}$):
$\overline{x}^n \pm \Phi_{0,1}^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) \cdot \sigma\left(\overline{X}^{(n)}\right)$, dabei $\sigma\left(\overline{X}^{(n)}\right) = \frac{\sigma(X)}{\sqrt{n}}$.
bei unbekanntem $\sigma(X)$ kann man für große $n \geq 100$ einfach
$s_n(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}^{(n)}\right)^2}$ für $\sigma(X)$ einsetzen (mit geringer Ungenauigkeit),
genauer wird das mit $t - Verteilung$ so (notwendig bei recht kleinen n):
$\overline{x}^n \pm TV_{df=n-1}^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) \cdot \frac{s_n(X)}{\sqrt{n}}$,
dabei ist $TV_{df=n-1}^{-1}$ die Umkehrfunktion zur $t-$ Verteilungsfunktion mit $n-1$ Freiheitsgraden.

b) Zweiseitiges Vertrauensintervall zu α für eine Mittelwertsdifferenz $\mu(Y) - \mu(X)$ bei unabhängigen Variablen X und Y :
In einer Stichprobe für X hat man beobachtet: $\overline{x}^{n_1}$, $s_{n_1}(X)$, in einer Stichprobe für Y hat man beobachtet: $\overline{y}^{(n_2)}$, $s_{n_2}(Y)$. Dann lautet das gesuchte Intervall:
$\overline{y}^{(n_2)} - \overline{x}^{(n_1)} \pm \Phi_{0,1}^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) \cdot \sqrt{\frac{s_{n_1}^2(X)}{n_1} + \frac{s_{n_2}^2(Y)}{n_2}}$, mit großer Genauigkeit auch bei mäßigen n_1, n_2, sagen wir beide ≥ 30.
Typische Anwendung: Man hat zwei getrennte Teilpopulationen A, B, eine Variable U auf $A \cup B$,und $Y = U_{|B}$, $X = U_{|A}$.

c) Man unterscheide davon die Situation, dass man ein Vertrauensintervall für eine erzielte Verbesserung (ausgedrückt in einer Mittelwertdifferenz $\mu_{nachher} - \mu_{vorher}$) angeben will:
Dann sind $Y = U_{nachher}$ und $X = U_{vorher}$ im Allgemeinen nicht unabhängig, man nimmt eine Stichprobe n von Paaren $(u_{vorher,i}, u_{nachher,i})$, $1 \leq i \leq n$,
und beobachtet $\overline{d}^{(n)}$ als Wert von $\frac{1}{n}\sum_{i=1}^{n}(u_{nachher,i} - u_{vorher,i})$, bildet
dann das Vertrauensintervall $\overline{d}^{(n)} \pm \Phi_{0,1}^{-1}\left(\alpha + \frac{1-\alpha}{2}\right) \cdot \sigma\left(\overline{D}^{(n)}\right)$
(bzw. genauer mit $t - Verteilung$) völlig analog zu **a)**.