Stochastik

Notizen zu «Elementare Stochastik» von Kütting, H. und Sauer, M. (2011)

Torsten Linnemann

Pädagogische Hochschule Fachhochschule Nordwestschweiz

torsten.linnemann@fhnw.ch

11. August 2015

Fachhochschule Nordwestschweiz
Pädagogische Hochschule

Inhaltsverzeichnis

1 Statistik

Dieses Kapitel habe ich zu von meinem Kollegen Dr. Boris Girnat übernommen, dem ich an dieser Stelle ganz herzlich dafür danken möchte.

1.1 Grundbegriffe der beschreibenden Statistik

Die *beschreibende* oder *deskriptive* Statistik hat das Ziel, empirisch erhobene Daten durch *Tabellen*, *Kennzahlen* bzw. *Parameter* oder *grafische Darstellungen* so darzustellen oder aufzubereiten, dass relevante Eigenschaften eines oder mehrerer Datensätze leicht überblickt bzw. miteinander verglichen werden können. In diesem Skript wird den folgenden Fragen nachgegangen:

1) Welche Arten von Daten gibt es?
2) Auf welche Weise können Daten grafisch dargestellt werden und welche Vor- und Nachteile haben die jeweiligen Darstellungen?
3) Mit welchen Parametern oder Kennzahlen kann man wichtige Informationen eines Datensatzes auf eine oder wenige Zahlen reduzieren?
4) Wie kann man Datensätze miteinander vergleichen (z. B. ist dieselbe Klausur in einer Klasse besser ausgefallen als in einer anderen)?
5) Wie kann man Beziehungen zwischen verschiedenen Merkmalen eines Datensatzes untersuchen und beschreiben (z. B. hängt die Schuhgrösse einer Person von ihrer Körpergrösse ab und – wenn ja – wie lässt sich dieser Zusammenhang mathematisch darstellen)?

1.1.1 Grundgesamtheit, Stichproben, Merkmale

Bei statistischen Erhebungen werden Eigenschaften oder *Merkmale* $M_1, M_2, \ldots, M_k$ von Objekten einer Menge, der sogenannten *Grundgesamtheit* G, erhoben. Sind diese «Objekte» beispielsweise Menschen, so können statistische Merkmale etwa Alter, Geschlecht, Körpergrösse, Jahreseinkommen, Intelligenzquotient, Besuch einer bestimmten Schulform oder ähnliches sein.

Oft werden nicht alle Objekte der Grundgesamtheit betrachtet, sondern nur eine Auswahl Ω daraus. Eine solche Auswahl nennt man *Stichprobe*. Die Erhebung kann eine Befragung, Beobachtung oder Messung sein. Jedes Merkmal hat üblicherweise verschiedene *Merkmalsausprägungen*, d. h. es kann verschiedene Werte annehmen wie z. B. zwei Ausprägungen beim Geschlecht oder potentiell beliebig viele Ausprägungen beim Jahreseinkommen oder der Körpergrösse (was allerdings auch mit der jeweiligen Mess- oder Erhebungsgenauigkeit zu tun hat).

Das Merkmal X Geschlecht kann beispielsweise die Ausprägungen $X = w$ oder $X = m$ annehmen. (Das Merkmal X ist eine auf Ω definierte Funktion.

Beispiel 1.1

Zwei Würfel werden 50 Mal nacheinander geworfen. M_{42} ist dann das Ergebnis des zweiund-

vierzigsten Wurfes. Das Merkmal X, Augensumme, kann dabei zum Beispiel die Ausprägung $X = 3$ angenommen haben.

Beispiel 1.2

In der Tabelle **??** sind Daten einer Studentengruppe erhoben. Es handelt sich um eine Totalerhebung. Die Studenten haben alle an einem Seminar im ersten Semesters eines Mathikstudiums für Lehramtsanwärtern teilgenommen. Es wurden fünf Merkmale erhoben: das Geschlecht, die Ergebnisse in drei Klausuren A, B und S in Prozent von der Gesamtpunktzahl und schliesslich eine sogenannte Studienleistung, welche die Studenten durch eine längere Hausarbeit ableisten sollten. Für die Studienleistung wurden drei Noten vergeben, nämlich 0 für «nicht abgegeben bzw. ungenügend», 1 für «mit merklichen Fehlern» und 2 für «im wesentlichen korrekt oder besser».

Nr.	Geschl.	Studienlst.	A	B	S
1	w	0	71	98	67
2	w	1	44	72	82
3	m	0	62	66	99
4	m	0	48	76	78
5	w	2	57	80	82
6	m	1	52	64	78
7	w	2	28	25	39
8	w	2	48	82	95
9	w	1	27	78	98
10	w	2	75	100	92
11	m	2	81	95	62
12	w	1	51	53	78
13	m	2	52	80	45
14	w	2	92	100	98
15	w	1	52	74	83
16	m	0	64	79	100
17	w	2	45	73	98
18	w	2	62	88	92
19	m	0	58	82	76
20	w	2	88	100	71
21	w	1	71	86	92
22	m	2	67	95	54

1.1.2 Skalierung

An den Merkmalen bzw. Merkmalsausprägungen in Beispiel 1.2 wird deutlich, dass sie sich auf unterschiedlicher Weise mathematisch weiterverarbeiten lassen. Manche Merkmalsausprägungen sind Zahlenwerte, manche nicht.

Aber auch bei den Zahlenwerten gibt es Unterschiede: Die Klausurergebnisse sind in Prozent angegeben, bei denen sich sinnvoll nach Abständen und Vielfachen fragen lässt: Jemand mit 40 Prozent der Punkte hat nur halb so viel gelöst wie jemand mit 80 Prozent; 73 Prozent sind nur wenig besser als 71, aber zwischen 12 Prozent und 91 besteht ein grosser Unterschied.

Analoge Fragen lassen sich zur Studienleistung nicht stellen: Sie wird so vage bewertet, dass man nicht sagen kann, ein Student mit dem Ergebnis 2 sei doppelt so gut gewesen wie einer mit

1. Erst recht lässt sich die Frage nach einer Reihenfolge oder Rangordnung beim Merkmal «Geschlecht» nicht stellen. Daran würde sich auch nichts ändern, wenn man statt «w» und «m» die Zahlen 0 und 1 als Kodierung des Geschlechts benutzte.

Für die mathematische Weiterverarbeitung ist es entscheidend, ob sich die Ausprägungen eines Merkmals in eine Hierarchie bringen lassen und – wenn ja – ob man Abständen und Vielfachen zwischen Merkmalsausprägungen eine sinnvolle Interpretation zukommen lassen kann. Diese Unterschiede werden als verschiedene Arten der *Skalierung* oder des *Skalenniveaus* von Daten bzw. Merkmalen bezeichnet.

Definition 1.1

Es sei M das Merkmal einer Stichprobe S. Das Merkmal M heisst ...

1) *... nominal- oder kategorialskaliert, wenn die Merkmalsausprägungen bis auf Identität unterschieden werden.*
2) *... ordinalskaliert, wenn für die Merkmalsausprägungen eine Ordnungsrelation definiert ist.*
3) *... metrisch skaliert, wenn für die Merkmalsausprägungen eine Ordnungsrelation definiert und die Datenerhebung additiv ist.*

Nach dieser Definition ist das Merkmal
«Geschlecht» aus Beispiel 1.2 nominalskaliert,
die Studienleistung ordinalskaliert, und
die Klausurergebnisse sind metrisch skaliert.

Jede metrische Skalierung ist auch eine ordinale Skalierung und jede ordinale auch eine nominale.

Bei ordinalskalierten Merkmalen mit Zahlenwerten als Ausprägungen haben die Abstände der Zahlenwerte keine Bedeutung, d. h. lassen sich – wie oben dargestellt – nicht sinnvoll im Rahmen des Sachkontextes interpretieren. Bei metrischen Skalen haben die Abstände eine Bedeutung, und zwar bei einer Intervallskala nur die Abstände und bei einer Verhältnisskala zusätzlich auch die Quotienten der Skalenwerte, also die ihre Verhältnisse.

Von der Art der Skalierung hängt es ab, wie Daten mathematisch weiterverarbeitet werden können. z. B. ist es nur bei metrisch skalierten Daten sinnvoll, die Merkmalsausprägungen zu addieren oder Durchschnittswerte zu berechnen.

1.1.3 Häufigkeiten und Klassierungen

Definition 1.2

Die absolute Häufigkeit $H_n(x_i)$ einer Merkmalsausprägung x_i in einer Stichprobe vom Umfang n gibt an, wie oft diese Merkmalsausprägung angenommen wurde.

Für die relative Häufigkeit gilt $h_n = \frac{H_n}{n}$. Falls n unwichtig ist, kann n als Index weggelassen werden.

Beispiel 1.3

In der Stichprobe aus Beispiel 1.2 ist $H_{22}(w) = 14$ und $h_{22}(w) = \frac{14}{22} \approx 0.64$ für die beiden Merkmalsausprägungen des Merkmals «Geschlecht».

Für die Merkmale «Geschlecht» und «Studienleistung» aus Beispiels 1.2 mag es aufschlussreich sein, die absoluten oder relativen Häufigkeiten der Merkmalsausprägungen zu berechnen. Weniger sinnvoll erscheint das im Fall der Klausurergebnisse, da hier verhältnismässig viele Merkmalsausprägungen mit jeweils relativ geringen Häufigkeiten auftreten. Die Häufigkeiten würden daher kaum mehr aussagen als die Datenwerte in der Tabelle. In solchen Fällen ist es üblich «ähnliche» Merkmalsausprägungen zu einer *Klasse* zusammenzufassen. Was als «ähnlich» gilt, hängt vom Zweck der Erhebung und dem Entscheidungsspielraum des Statistikers ab.

Beispiel 1.4

Ein Dozent vergibt nach den Prozentwerten aus 1.2 Noten auf die Klausuren.
Bis 39 Prozent eine 0 für nicht bestanden. Ab 40 Prozent eine 4, ab 52 Prozent eine 4.5 und so weiter in Schritten von 12 Prozent.
Damit hat der Dozent eine Klassierung der Klausurergebnisse vorgenommen.

Definition 1.3

Es sei $\{x_1, x_2, \ldots, x_s\}$ die Menge der Merkmalsausprägungen eines Merkmals X. Ist K eine disjunkte und vollständige Zerlegung von $\{x_1, x_2, \ldots, x_s\}$ in Teilmengen $K_1, K_2, \ldots, K_t$, so nennt man K eine Klassierung von X.

Durch eine Klassierung des Merkmals X wird ein neues Merkmal X_K eingeführt, das jedem x_i die Klasse K_j zuordnet, in der sich x_i befindet. Auf diese Weise erhalten die Klassen $K_1, K_2, \ldots, K_t$ absolute Häufigkeiten (und damit indirekt auch relative Häufigkeiten), nämlich jeweils die Anzahl der Elemente aus Ω, die unter X Werte in der jeweiligen Klasse annehmen.

Beispiel 1.5

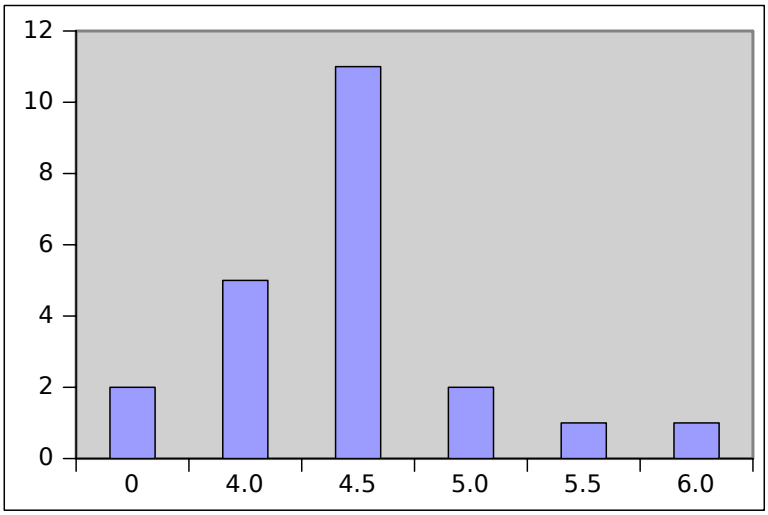
Nach dem oben angegebenen Notenschlüssel sind in der Tabelle **??** die Noten für die Klausur A ergänzt. Zur besseren Übersicht wurden die Daten nach den Prozentwerten von A sortiert.

Geschlecht	Prozent A	Note A
w	27	0
w	28	0
w	44	4.0
w	45	4.0
m	48	4.0
w	48	4.0
w	51	4.0
m	52	4.5
m	52	4.5
w	52	4.5
w	57	4.5
m	58	4.5
m	62	4.5
w	62	4.5
m	64	4.5
m	67	4.5
w	71	4.5
w	71	4.5
w	75	5.0
m	81	5.0
w	88	5.5
w	92	6.0

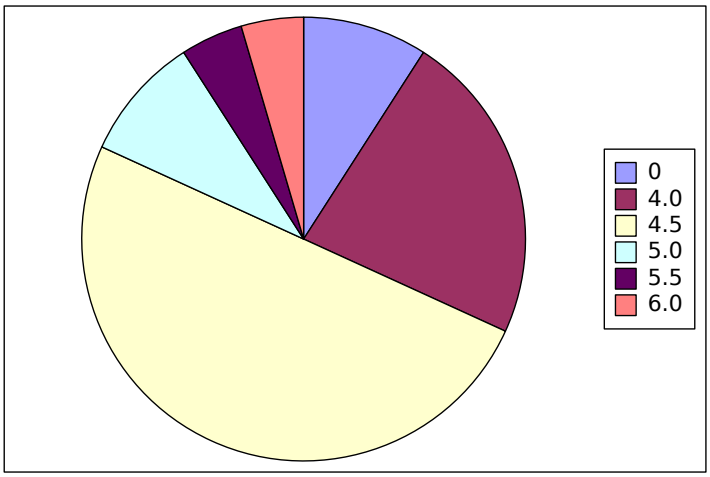
Für die sechs Klassen, die durch die Benotung entstanden sind, lassen sich nun absolute Häufigkeiten (und darauf aufbauend auch relative) angeben, und zwar $H(0) = 2$, $H(4.0) = 5$, $H(4.5) = 11$, $H(5.0) = 2$ und $H(5.5) = 1$ und $H(6.0) = 1$.

1.1.4 Grafische Aufbereitung von Häufigkeiten

Absolute und relative Häufigkeiten werden oft durch Diagramme veranschaulicht. Typische Darstellungsweisen sind das *Balken-* oder *Säulendiagramm* und das *Kreisdiagramm*. Im Balken- oder Säulendiagramm werden die absoluten oder relativen Häufigkeiten so dargestellt, dass die Höhe der aufrecht stehenden Säulen bzw. die Länge der waagerecht liegenden Balken der relativen oder absoluten Häufigkeit der jeweiligen Merkmalsausprägung entspricht. Die Abbildung 1.1.4 zeigt ein Säulendiagramm mit den absoluten Häufigkeiten der Noten aus der Klausur A.



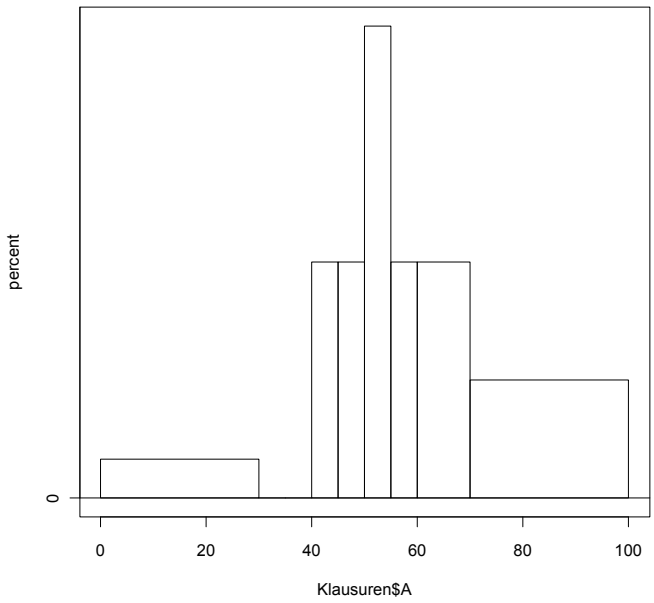
Das Kreisdiagramm eignet sich nur, um relative Häufigkeiten darzustellen. Aus einem Vollkreis wird der Merkmalsausprägung x_i ein Kreissegment zugeordnet, das den Winkel $360° \cdot h_n(x_i)$ einschliesst. Die Abbildung 1.1.4 zeigt ein Kreisdiagramm mit den relativen Häufigkeiten der Noten aus der Klausur A.



Kreis-, Säulen- und Balkendiagramme lassen sich bei jeder Art der Skalierung eines Merkmals benutzen, um Häufigkeiten darzustellen. Das sogenannte *Histogramm* eignet sich nur, wenn ein Merkmal metrisch skaliert ist. Es ähnelt dem Säulendiagramm, allerdings wird die (absolute oder relative) Häufigkeit nicht proportional zur Säulen*höhe*, sondern zur Säulen*fläche* abgetragen. Dies setzt voraus, dass die x-Achse metrisch skaliert ist.

Ein Histogramm mit äquidistanter Klasseneinteilung (d. h. mit Klassen derselben Länge auf der x-Achse) unterscheidet sich nicht wesentlich von einem Säulendiagramm. Daher werden Histogramme gern benutzt, um Klassen unterschiedlicher metrischer Länge darzustellen. So zeigt die Abbildung 1.1.4 ein Histogramm zu den Prozentwerten der Klausur A, bei dem von 30 % bis 60 % eine Klassenlänge von 5 % gewählt wurde und darunter bzw. darüber ein bzw. zwei Klassen grösserer Länge. Die feinere Einteilung zwischen 30 und 60 Prozent könnte dadurch motiviert sein, sich vom Übergangsbereich zwischen «bestanden» und «nicht bestanden» ein genaueres Bild zu machen. Die Klassenlängen wurden dann auf der x-Achse abgetragen und nach oben hin so zu einem Rechteck ergänzt, sodass Fläche des Rechtecks der Häufigkeit im jeweiligen Pro-

zentbereich entspricht.



Ein Histogramm kann irritieren, da es Menschen üblicherweise schwerer fällt, Flächenmasse miteinander zu vergleichen anstelle von Längenmassen wie beim Säulen- oder Balkendiagramm. So kann hier der Eindruck entstehen, dass zwischen 55 und 60 Prozent genau so viele Ergebnisse lägen wie zwischen 60 und 70 Prozent – was natürlich falsch ist. Ebenso können ungünstig gewählte Grenzen falsche Vermutungen implizieren. So ist die erste Klasse von 0 bis 30 gewählt und entsprechend gibt es ein Rechteck, das den gesamten Bereich von 0 bis 30 Prozent überspannt. Das erweckt den Eindruck, es gäbe tatsächlich Klausurergebnisse bis hinunter zu 0 Prozent. Die Untergrenze liegt jedoch bei 27 %.

1.2 Lageparameter

Als Lageparameter bezeichnet man Zahlen, die etwas Charakteristisches über ein Merkmal oder einen Datensatz aussagen sollen. Sie «fassen» einen Datensatz in einem einzigen Zahlenwert zusammen. Damit ist notwendigerweise eine Verringerung von Detailinformationen verbunden. Lageparameter befinden sich von daher immer in einem Spannungsverhältnis zwischen erwünschter «Informationskonzentration» und unerwünschter «Informationsvernichtung». An welcher Stelle man zwischen diesen Extremen sich befindet, ist von Fall zu Fall unterschiedlich und erfordert einen vorsichtigen und verständigen Umgang mit Lageparametern.

1.2.1 Der Modalwert

Definition 1.4

Der Modalwert ist die Merkmalsausprägung, die in einer Stichprobe am häufigsten vorkommt. Kommen mehrere Ausprägungen gleich häufig vor, so wird eine Liste von Modalwerten aufgestellt.

Beispiel 1.6

Eine Stichprobe ergibt die Merkmalsausprägungen $\{3, 3, 4, 4, 4, 5, 6, 6\}$. Der Modalwert ist 4. Bei den Merkmalsausprägungen $\{3, 3, 4, 4, 4, 5, 6, 6, 6\}$ gibt es eine Liste von Modalwerten: $\{3, 6\}$.

Beispiel 1.7

Im Beispiel der Benotung der Klausur A ist $x_{\text{mod}} = 4.5$.

Der Modalwert bezieht sich meist auf eine Klasseneinteilung.

1.2.2 Quantile, Quartile und Median

Quantile, Quartile und der Median sind Lageparameter, die nicht auf Häufigkeiten aufbauen, sondern direkt unter den Ausprägungen eines Merkmals «Übersicht schaffen» sollen. Der grundlegende Begriff ist der des *Quantils*: Wenn die Merkmalsausprägungen mindesten ordinal skalierte Zahlenwerte sind, so soll das p-Quantil zu einen vorgegebenen Pronzentwert $p \cdot 100\,\%$ angeben, welchen Zahlenwert die niedrigsten $p \cdot 100\,\%$ aller Merkmalsausprägungen nicht überschreiten.

Beispiel 1.8

Der Median bezeichnet die Merkmalsausprägung unter der die Hälfte der Stichprobenergebnisse liegt, Also das $0.5-$Quantil.
Der Dozent der Studentengruppe aus Beispiel 1.2 könnte daran interessiert sein herauszufinden, unter welchem Ergebnis die Hälfte seiner Studenten in der Klausur A geblieben sind. Er sortiert die Liste nach den Ergebnissen von A in der Tabelle **??** neu und betrachtet, bis wohin die Hälfte gekommen ist.

Nr.	Geschl.	Studienlst.	A	B	S
1	w	1	27	78	98
2	w	2	28	25	39
3	w	1	44	72	82
4	w	2	45	73	98
5	m	0	48	76	78
6	w	2	48	82	95
7	w	1	51	53	78
8	m	1	52	64	78
9	w	1	52	74	83
10	m	2	52	80	45
11	w	2	57	80	82
12	m	0	58	82	76
13	m	0	62	66	99
14	w	2	62	88	92
15	m	0	64	79	100
16	m	2	67	95	54
17	w	1	71	86	92
18	w	0	71	98	67
19	w	2	75	100	92
20	m	2	81	95	62
21	w	2	88	100	71
22	w	2	92	100	98

Es gibt 22 Teilnehmer. Die Hälfte sind 11. Der 11. hat 57 Prozent, der 12. Teilnehmer 58 Prozent. Als Median wird 57.5.festgelegt.
Sucht man nicht nach der Grenze für die Hälfte der Teilnehmer, sondern sondern etwa eine Grenze für «den überwiegenden Teil» – sagen wir also beispielsweise für 85 % der Teilnehmer –, so wird es noch schwieriger, die Grenze zu ziehen, denn einen Teilnehmer mit der Nummer $0.85 \cdot 22 = 18.7$ gibt es nicht. Um wenigsten 85 % einzuschliessen, entscheidet man sich dafür, auf den 19. Teilnehmers aufzurunden, also das Klausurergebnis 75 als Grenzen zu wählen.
Zur Definition brauchen wir die Gaussklammer, $[\ ]$, die jeder reellen Zahl die nächstkleinere ganze Zahl zuordnet

Definition 1.5

Ist $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ *eine Stichprobe,* X *ein auf* Ω *definiertes reellwertiges, mindestens ordinalskaliertes Merkmal mit den Merkmalsausprägungen* $x_i = X(\omega_i)$ *für* i *mit* $1 \leq i \leq n$*, sowie* $p \in \mathbb{R}$ *mit* $0 < p < 1$*, so ist*

$$x_p = \begin{cases} \dfrac{x_{p\cdot n} + x_{p\cdot n+1}}{2}, & \text{falls } p \cdot n \text{ ganzzahlig ist;} \\ x_{[n\cdot p]+1}, & \text{falls } p \cdot n \text{ nicht ganzzahlig ist.} \end{cases}$$

das p*-Quantil von* Ω*.*

$p_{0.00}$ *ist das Minimum,* $p_{1.00}$ *das Maximum der Merkmalsausprägungen.*

Hier wird dem Objekt ω_i der Stichprobe die Merkmalsausprägung x_i zugeordnet, d. h. anders als bisher ist die Anzahl der x_i und ω_i identisch und die Werte x_i sind nicht mehr notwendigerweise verschieden.

Die Indizes der Quantile werden hier als Dezimalzahlen mit zwei Nachkommastellen angegeben, um Verwechselungen mit den ausschliesslich natürlichzahligen Indizes zu vermeiden, mit denen die Merkmalsausprägungen $x_1, x_2, \ldots, x_n$ durchnummeriert werden.

Manche Computerprogramme unterscheiden nicht die beiden Fälle «$p \cdot n$ ist ganzzahlig» und «$p \cdot n$ ist nicht ganzzahlig», sondern verwenden immer die Definition für den ganzzahligen Fall, d. h. wenn Sie solche Programme zur Kontrolle einsetzen, könnten deren Werte ggf. von denen abweichen, die Sie von Hand berechnet haben.

Man beachte: Für die «Aufteilung» einer Stichprobe in Quantile muss von den einzelnen Elementen der Stichprobe $\omega_1, \omega_2, \ldots, \omega_n$ ausgegangen werden, und nicht von den Merkmalsausprägungen $x_1, x_2, \ldots, x_s$, die möglicherweise für verschiedene ω_i und ω_j identisch sind und in diesem Fall die gewünschten Prozentgrenzen verschieben würden. Die Zahl x_p gibt also einen Wert von X an, den *mindestens*, aber durch die Rundungsoperation ggf. auch (etwas) mehr als $p \cdot 100\,\%$ der Elemente der Stichprobe nicht überschreiten.

Beispiel 1.9

Man möchte ermitteln, welchen Wert 60 % der Studenten in der Klausur A nicht überschritten haben, d. h. das 0.6-Quantil soll ermittelt werden. Die Zahl $0.6 \cdot 22 = 13.2$ ist nicht ganzzahlig, also ist Fall 2 der Quantilsdefinition zu benutzen. Die Gauss-Klammer liefert $[0.6 \cdot 22] = [13.2] = 13$, also ist $x_{0.60} = x_{[0.6\cdot 22+1]} = x_{[14.2]} = x_{14} = 62$ gemäss Tabelle **??**.
Wie man an diesem diesem Beispiel sieht, ist das «mindestens» in der Formulierung «mindestens $p \cdot 100$ Prozent der Werte überschreiten x_p nicht» wichtig: In unserem Fall ist $x_{0.60} = x_{14} = 62$. Allerdings ist $\frac{14}{22} \approx 63.63$, also etwas mehr als $60\,\%$, was sich zwangsläufig durch die Gesamtzahl 22 der Werte ergibt. Der Prozentwert könnte sogar noch über $63.63\,\%$ liegen, wenn nämlich auch x_{15} oder noch weitere Merkmalsausprägungen zufälligerweise den Wert 62 hätten, was hier wegen $x_{15} = 64$ nicht der Fall ist.

Für einige ausgewählte und für die weitere Datenanalyse und -darstellung besonders wichtige Werte von p haben sich feststehenden Bezeichnungen der p-Quantile eingebürgert:

Definition 1.6

Das p-Quantil x_p heisst ...

1) ... Minimum oder 0. Quartil $x_{0.00}$ bzw. $x_{\min}$ für $p = 0.00$;
2) ... 1. Quartil $x_{0.25}$ für $p = 0.25$;
3) ... Median oder 2. Quartil $x_{0.50}$ oder x_{med} oder $\tilde{x}$ für $p = 0.50$;
4) ... 3. Quartil $x_{0.75}$ für $p = 0.75$;
5) ... Maximum oder 5. Quartil $x_{1.00}$ bzw. $x_{\max}$ für $p = 1.00$.

Die Berechnungsformeln der Quantile lassen sich im Falle des Medians folgendermassen vereinfachen:

Satz 1.1

Ist X ein auf Ω definiertes reellwertiges, mindestens ordinalskaliertes Merkmal, so ist

$$x_{0.50} = \begin{cases} \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right), & \text{falls } n \text{ gerade ist;} \\ x_{\frac{n+1}{2}}, & \text{falls } n \text{ ungerade ist.} \end{cases}$$

1.2.3 Der Boxplot

Sinn und Zweck der fünf Quartile ist es, den Wertebereich von X in vier hierarchisch angeordnete Teile zu unterteilen, die jeweils von ungefähr 25 % der Elemente der Stichprobe angenommen werden. Üblicherweise wird der Bereich vom Minimum bis zum 1. Quartil als Ort der unteren Ausreisser angesehen, der Bereich vom 1. bis zum 3. Quartil als «normales Mittelfeld» mit dem Median als eine Art Mittelwert darin und der Bereich vom 3. Quartil zum Maximum als Abschnitt der oberen Ausreisser. Je nach Sachsituationen können aber auch andere Einteilungen für die Unterscheidung in Ausreisser und normale Fälle sinnvoll sein.

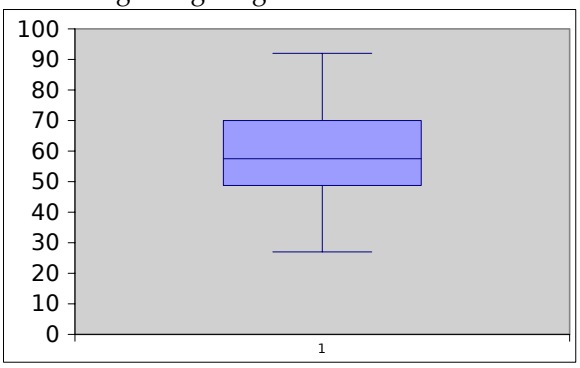
Eine typische Visualisierung der fünf Quartile ist der *Boxplot*, der eine Box, begrenzt durch das 1. und 3. Quartil mit dem Median als Trennlinie darstellt und vom 1. Quartil zum Minimum bzw. von 3. Quartil zum Maximum durch zwei Linien, den sogenannten *Wimpern*, fortgesetzt wird; d. h. die Box enthält das Mittelfeld und die Wimpern die Ausreisser nach unten bzw. oben. Gelegentlich wird auch der Bereich in den Wimpern zwischen $x_{0.00}$ und $x_{0.05}$ sowie zwischen $x_{0.95}$ und $x_{1.00}$ gestrichelt oder gepunktet dargestellt, um die «extremen» Ausreisser grafisch deutlich zu machen. Diese Darstellung wird hier jedoch nicht verwandt. Im deutschsprachigen Bereich wird der Boxplot meist stehend, im englischsprachigen Bereich eher liegend dargestellt (was anscheinend die Bezeichnung «Wimpern» bzw. «whiskers» für die dann seitlich angebrachten Linien motiviert hat).

Beispiel 1.10

Für die Klausurergebnisse A, die in Beispiel 1.8 bereits aufsteigend sortiert worden sind, ergeben sich die Quartilswerte so, wie sie in der nächsten Tabelle eingetragen sind.

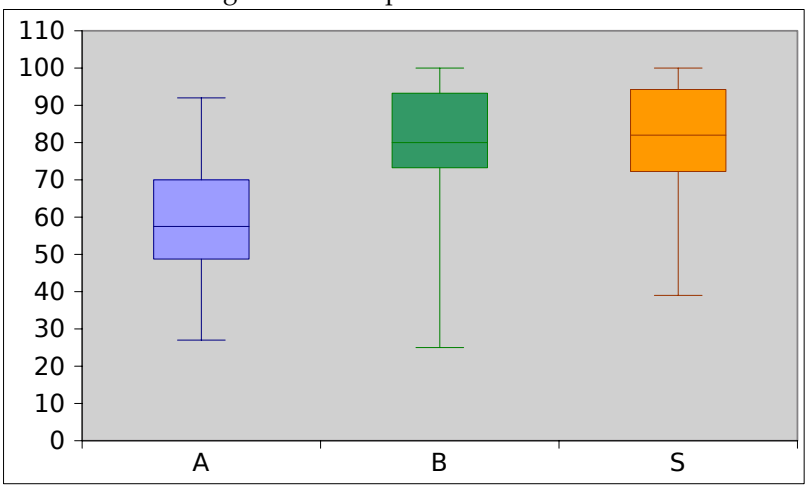
Quartilstyp	Wert
$x_{0.00}$	27
$x_{0.25}$	48
$x_{0.50}$	57.5
$x_{0.75}$	71
$x_{1.00}$	92

Mit den Werten aus der Tabelle lässt sich der Boxplot nach deutschsprachigem Muster wie in der Abbildung 1.10 gezeigt darstellen.



Beispiel 1.11

Da zu allen Teilnehmern die Ergebnisse dreier Klausuren erhoben worden sind, lassen sich die Quartile (oder andere Lageparameter) bzw. ihr visueller Eindruck zum Vergleich benutzen. So stellt die Abbildung 1.11 die Boxplots aller drei Klausuren nebeneinander.



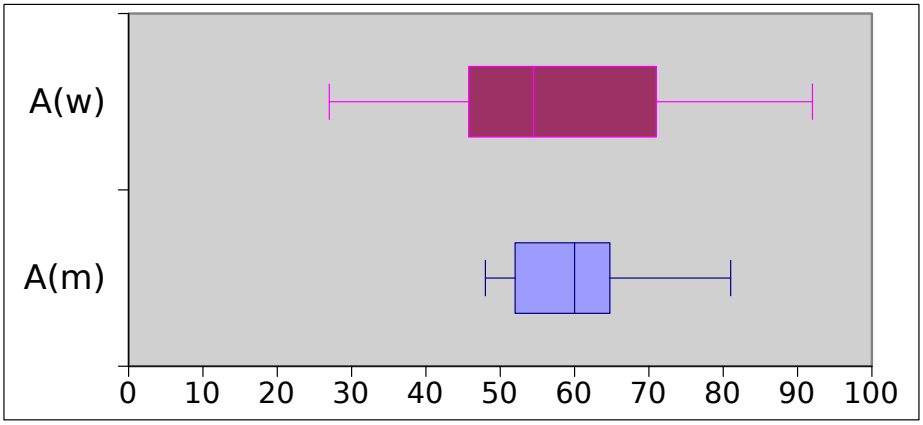
Auftrag 1.1

Überlegen Sie, ob man aus den drei Boxplots in Abbildung 1.11 Informationen zu einem Vergleich der drei Klausuren entnehmen kann. Falls ja, dann formulieren Sie möglichst prägnante Behauptungen und begründenden Sie diese anhand der Boxplots.

Beispiel 1.12

Zu den Klausuren wurde auch ein nominales oder kategoriales Merkmal erhoben, nämlich das Geschlecht der Teilnehmer. Über nominale Merkmale erhält man automatisch eine Einteilung der Stichprobe in Klassen. Wenn man Lageparameter für diese Klassen getrennt berechnet, hat man eine erste, noch sehr rudimentäre Möglichkeit, Zusammenhänge zwischen Merkmalen zu unter-

suchen. So könnte man fragen, ob das Geschlecht in einem Zusammenhang zum Klausurergebnis steht. Die Abbildung 1.12 zeigt die Boxplots der Prozentwerte für Klausur A getrennt nach Geschlecht (dieses Mal sind die Boxplots in angloamerikanischer Manier liegend gezeichnet).



Auftrag 1.2

Begründen Sie, ob und welche Behauptungen über Geschlechterunterschiede durch die Boxplots in der Abbildung 1.12 gestützt werden. Überlegen Sie, wie ein Histogramm mit äquidistanter Klasseneinteilung aussehen könnte, das zum Boxplot $A(w)$ bzw. $A(m)$ passt (ohne die Histogramme aus den Daten tatsächlich zu erstellen). Erstellen Sie zu einer der anderen beiden Klausuren geschlechtergetrennte Boxplots.

1.2.4 Das arithmetische Mittel

Definition 1.7

Ist $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ eine Stichprobe, X ein auf Ω definiertes metrisches Merkmal mit den Merkmalsausprägungen $x_i = X(\omega_i)$ für i mit $1 \leq i \leq n$, so ist

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

das arithmetische Mittel bzw. der (arithmetische) Mittelwert von X.

Hier werden Merkmalsausprägungen addiert, daher muss das Merkmal metrisch skaliert sein, damit sich die Summe der Merkmale sinnvoll interpretieren lässt.

Der Median und das arithmetische Mittel treten beide mit dem Anspruch auf, so etwas wie «die Mitte» eines Datensatzes anzugeben. Beim arithmetischen Mittel ist es der Durchschnittwert der Merkmalsausprägungen; beim Median der Wert, der den Datensatz (ungefähr) in zwei Hälften teilt. Im Fall der Prozentwerte von Klausur A liegen die beiden Werte nicht weit auseinander: So ist $\overline{x} = 58.86$ und $x_{0.50} = 57.5$, d. h. es tritt hier keine grosse Uneinigkeit darüber auf, was die «besser» Mitte des Datensatzes ist. Können diese Werte dennoch weit auseinander liegen? Haben Sie mathematisch gesehen unterschiedliche Eigenschaften?

Auftrag 1.3

In einem Klinikum arbeiten sechs Stationsärzte mit einem Jahressalär von jeweils $96'000$ CHF

und der prominente Chefarzt, der $840'000$ CHF pro Jahr verdient. Erläutern Sie anhand dieses Datensatzes, inwiefern ein «kreativer» Umgang mit statistischen Methoden, insbesondere eine Entscheidung für oder gegen den Median oder das arithmetische Mittel als Ausdruck für «die Mitte» eines Datensatzes, zu beliebigen Antworten auf die Frage «Gehören Ärzte zu den Spitzenverdienern?» führen kann.

1.3 Streuparameter

Lageparameter sollen über die *absolute* Lage von Daten Auskunft geben. Wie man beispielsweise schon in der Abbildung 1.12 erkennen konnte, ist nicht nur die absolute Lage der Daten von Interesse, sondern auch ihre *relative*, d. h. die Lage der Daten zueinander. So weichen die beiden Mediane in diesem Fall nicht allzu sehr voneinander ab; die übrigen Daten «streuen» aber sehr unterschiedlich um diese beiden Werte: Im Fall der weiblichen Teilnehmer ist die Box breiter und sind die Wimpern länger als bei den männlichen Teilnehmern. Bei diesen ist nicht nur die geringere Streuung für sich interessant, sondern auch, dass die Streuung nach unten deutlich geringer ist als die nach oben.

Die Verteilung der Merkmalsausprägungen um einen Mittelwert (in der Regel um den Median oder das arithmetische Mittel) ist der interessanteste, aber nicht der einzige Fall der relativen Lage von Daten zueinander, der statistisch untersucht wird. Mit einem *Streuparameter* möchte man zu einem Datensatz einen einzigen Zahlenwert angeben, der eine relevante Information über das Streuverhalten der Daten enthält.

1.3.1 Spannweite und Quartilsabstände

Bevor wir zur Streuung um einen Mittelwert kommen, werden Streuparameter vorgestellt, die «interessante Ausschnitt» aus dem Datensatz betrachten, ohne einen Bezug zu einem Zentrum herzustellen. Naheliegende «interessante Bereiche» sind dabei der gesamte Datensatz und das «Mittelfeld»:

Definition 1.8

Ist X ein wenigstens ordinalskaliertes Merkmal, so ist

$$SW = |x_{\max} - x_{\min}|$$

die Spannweite und

$$Q_{0.5} = |x_{0.75} - x_{0.25}|$$

der Quartilsabstand von X.

Analog könnte man auch Abstände für andere Quantile definieren, was aber in der Regel unterbleibt.

1.3.2 Mittlere absolute Abweichung

Betrachten wir nun Abweichungen der Daten von einem Mittelwert. Zwar sind Median und arithmetisches Mittel als Mittelwerte am interessantesten, doch betrachten wir zunächst allgemein den Fall, wie weit die Daten von einem beliebigen Zahlenwert c «im Durchschnitt» abweichen.

Definition 1.9

Es sei $c \in \mathbb{R}$ und $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ eine Stichprobe, X ein auf Ω definiertes metrisches Merkmal mit den Merkmalsausprägungen $x_i = X(\omega_i)$ für i mit $1 \leq i \leq n$. Dann ist

$$d_c = \frac{1}{n} \sum_{i=1}^{n} |x_i - c|$$

die mittlere absolute Abweichung von X bezüglich c.

Von der Realsituation her gesehen, sind mittlere absolute Abweichungen von Mittelwerten von besonderem Interesse. Der Median hat darüber hinaus noch eine mathematisch interessante Eigenschaft: Die mittlere absolute Abweichung zum Median ist minimal gegenüber allen anderen Werten von c:

Satz 1.2

Es sei X ein metrisch skaliertes Merkmal. Dann gilt $d_{x_{0.50}} \leq d_c$ für alle $c \in \mathbb{R}$.

Auftrag 1.4

Man könnte versuchen, statt der *absoluten* Abweichung der Merkmalsausprägungen die durchschnittlichen Abstände selbst als ein Mass der Abweichung zu benutzen. Beweisen Sie, dass diese Idee im Falle der Abweichung zum arithmetischen Mittel zu einer überraschenden Konsequenz führt, die diesen Ansatz vollkommen unbrauchbar macht. Es gilt nämlich

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}) = 0,$$

d. h. für jeden beliebigen Datensatz wäre diese «mittlere Abweichung» gleich Null und würde daher nichts Gehaltvolles über den Datensatz aussagen.

1.3.3 Varianz und Standardabweichung

Definition 1.10

Ist $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ eine Stichprobe, X ein auf Ω definiertes metrisches Merkmal mit den Merkmalsausprägungen $x_i = X(\omega_i)$ für i mit $1 \leq i \leq n$, so ist

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

die empirische Varianz und

$$\sigma_X = \sqrt{\sigma_X^2}$$

die Standardabweichung von X. Falls klar ist, um welchen Datensatz es sich handelt, kann der Index X auch wegfallen und einfach σ bzw. σ^2 geschrieben werden.

Die empirische Varianz ist nicht die mittlere Abweichung vom arithmetischen Mittel, sondern der Mittelwert der *quadrierten* mittleren Abweichungen vom arithmetischen Mittel. Sie ist von daher der Mittelwert von *Flächenmassen* und kann daher *nicht* als Abstand, also als ein Längenmass, interpretiert werden. Die Standardabweichung hingegen hat dieselbe masstheoretische Dimension wie die Merkmalsausprägungen und lässt sich daher als Mass für einen Abstand zum arithmetischen Mittel interpretieren.

Warum man die Wurzel aus dem Mittelwert von Flächenmassen betrachten sollte, um einen Abstand zum arithmetischen Mittel zu definieren, erschliesst sich aus dem Sachkontext nicht und kann erst dann begründet werden, wenn eine Verbindung zwischen Statistik und Wahrscheinlichkeitsrechnung hergestellt ist. Hier kann allenfalls darauf hingewiesen werden, dass diese «merkwürdig abwegige» Definition dieselbe Minimalitätseigenschaft bezüglich des arithmetischen Mittels hat wie die mittlere absolute Abweichung bezüglich des Medians:

Satz 1.3

Es sei X ein metrisch skaliertes Merkmal. Dann gilt

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \leq \frac{1}{n} \sum_{i=1}^{n} (x_i - c)^2$$

für alle $c \in \mathbb{R}$.

BEWEIS: Wie Sie im Auftrag 1.4 bewiesen haben, gilt $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. Damit folgt für ein beliebiges $c \in \mathbb{R}$:

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - c)^2 &= \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - c)^2 \\
&= \sum_{i=1}^{n}\left((x_i - \overline{x})^2 + 2(x_i - \overline{x})(\overline{x} - c) + (\overline{x} - c)^2\right) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2(\overline{x} - c)\underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{=0} + \sum_{i=1}^{n}(\overline{x} - c)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + \underbrace{n(\overline{x} - c)^2}_{=:C} \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + C
\end{aligned}
$$

Da n und $(\overline{x} - c)^2$ nicht negativ sind, trifft das auch auf C zu. Nur im Fall $c = \overline{x}$, ist $C = 0$. Andernfalls ist $C > 0$ und daher die Summe $\sum_{i=1}^{n}(x_i - c)^2$ grösser als $\sum_{i=1}^{n}(x_i - \overline{x})^2$.

1.4 Lineare Regression

Die bisherigen statistischen Methoden haben stets nur *eine* Merkmalsausprägung für sich ins Auge gefasst. Solche Analysen nennt man *eindimensional* oder *univariat*. Werden *Zusammenhänge* zwischen *mehreren* Merkmalsausprägungen untersucht, so spricht man von *mehrdimensionalen* oder *multivariaten* Analysemethoden. In der Abbildung 1.12 wurde ein erster Schritt in diese Richtung gegangen: Man hat ein kategoriales Merkmal zur Klassierung des Datensatzes benutzt, dann die univariaten Methoden für jede Klasse getrennt angewandt und die Ergebnisse anschliessend miteinander verglichen – sei es grafisch oder sei es anhand der arithmetischen Werte der Lage- und Streuparameter. Dieses Verfahren nennt sich *Clustern* und eignet sich nur, wenn man Abhängigkeiten eines Merkmals von einem *kategorialen* Merkmal untersuchen möchte.

In diesem Abschnitt werden Methoden vorgestellt, mit denen man Zusammenhänge zwischen *metrischen* Merkmalen analysieren kann. Wir beschränken uns dabei auf zwei Merkmale, d. h. auf den *zweidimensionalen* oder *bivariaten* Fall. Ausserdem behandeln wir hier nur die Möglichkeit eines *linearen Zusammenhangs*. Dieser Spezialfall ist einerseits als realitätsbezogenen Gründen wichtig, weil solche Zusammenhänge verhältnismässig oft auftreten; andererseits ist ein linearer Zusammenhang der Fall, für den gute und trotzdem verhältnismässig einfach zu handhabende mathematische Methoden zur Verfügung stehen, die sich teilweise auch auf nicht-lineare Fälle übertragen lassen.

1.4.1 Punktwolken

Abhängigkeiten zwischen Merkmalen lassen sich oft aus dem Sachkontext heraus vermuten: Das Gewicht einer Person hängt vermutlich (auch) von ihrer Grösse ab; wer mehr verdient, hat eher

eine grössere Wohnung als einer, der wenig verdient; je stärker man ein Gewebe radioaktiv bestrahlt, desto stärker ist der Schaden. In all diesen Fällen scheint ein Zusammenhang zu bestehen. Ob der aber tatsächlich besteht und – wenn ja – ob er dann auch linear ist, soll nun mit statistischen Methoden untersucht werden. Ein erster Schritt besteht darin, sich einen grafischen Überblick über Daten zu verschaffen, die zu zwei Merkmalen gehören.

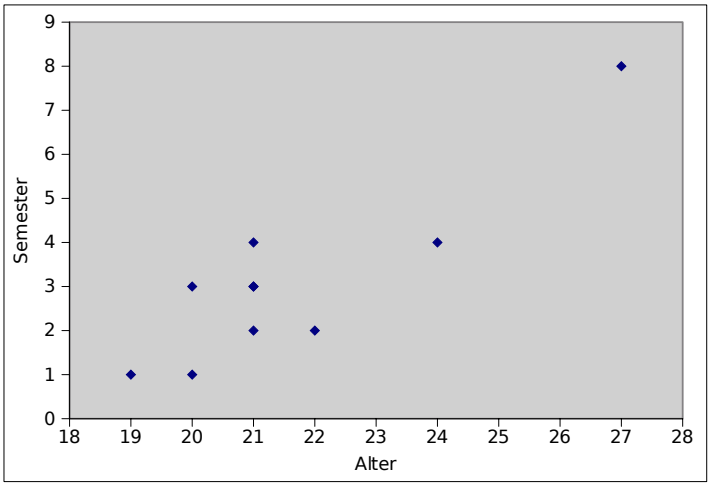
Beispiel 1.13

Wir beginnen mit einem kleinen Datensatz, in dem das Alter und die Semesterzahl einer Studentengruppe erhoben ist. Man vermutet: Je älter ein Student ist, desto höher ist sein Semester. Auch dass dieser Zusammenhang linear sein könnte, ist nicht unplausibel: Sieht man von Studienwechsel und Auslands-, Urlaubs- und Freisemestern ab, so erhöhen sich Alter und Semesterzahl proportional zueinander, sogar mit dem Proportionalitätsfaktor 1.

Alter	20	21	24	21	22	21	27	19	20	21
Semester	3	3	4	2	2	3	8	1	1	4

Wenn ein linearer Zusammenhang zwischen diesen beiden Merkmalen besteht, so könnte man die zehn Paare von Messwerten in ein Koordinatensystem einzeichnen, und die Datenpunkte müssten dann «ungefähr» auf einer Geraden liegen. Den Graphen, der aus den Datenpaaren besteht, nennt man *Punktwolke*.

Punktwolken kann man wie jede Menge reeller Zahlenpaare grafisch in einem Koordinatensystem darstellen. In der Abbildung ist die Punktwolke der beiden Merkmale aus der letzten Tabelle in einem kartesischen Koordinatensystem dargestellt.



1.4.2 Lineare Regression nach Augenmass

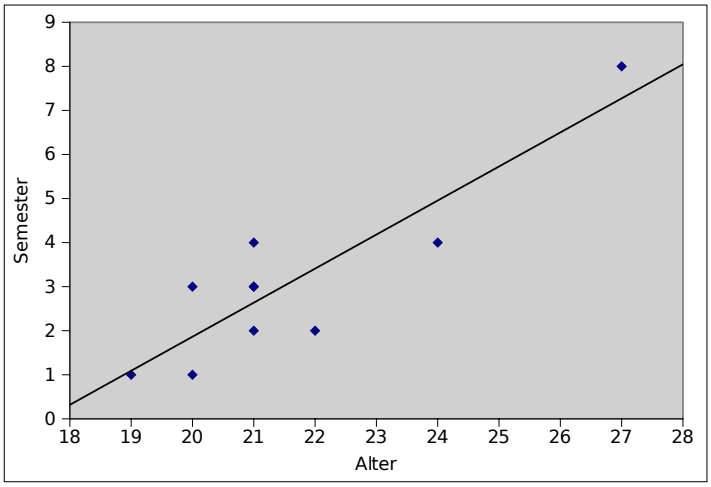
Die Punktwolke in der obigen Abbildung scheint linear anzusteigen. Eine Gerade, auf der alle Punkte liegen, wird man nicht finden – allein schon, weil dem Alter 23 drei verschiedene Semesterzahlen zugeordnet sind. Aufgabe der linearen Regression ist es, dennoch eine Gerade zu finden, welche die Punktwolke «möglichst gut» annähert, d. h. die «lineare Grundtendenz» der Punktwolke möglichst gut darzustellen, so dass die Abweichung von dieser «Tendenz» möglichst gering sind. Das Ziel ist also eine lineare Funktion y_{fit} zu finden, die zu jedem x-Wert x_i einen y-Wert $y_{\text{fit}}(x_i)$ liefert, sodass die Abweichung vom tatsächlich gemessenen y-Wert y_i für

alle Messwerte x_i möglichst gering ist. Diese Abweichung $r_i = y_i - y_{\text{fit}}(x_i)$ bezeichnet man als *Residuum*.

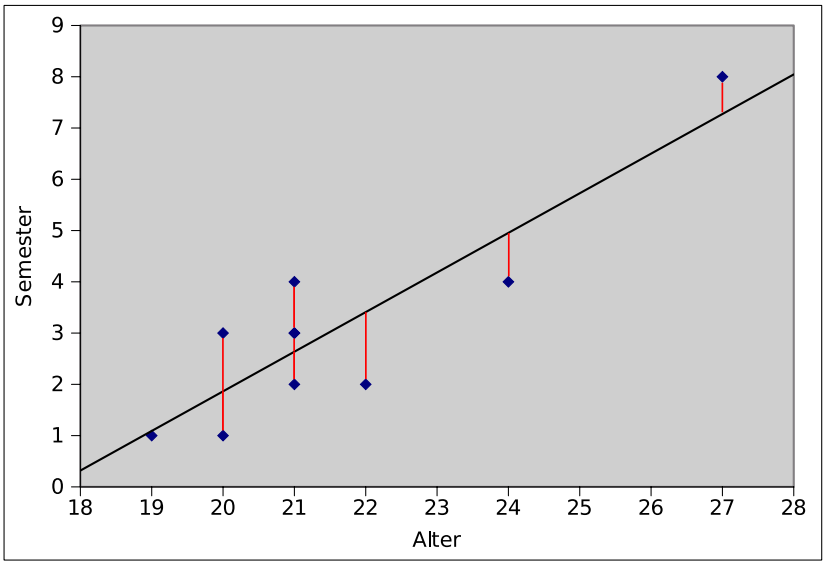
Definition 1.11

Sind $M_X = (x_1, x_2, \ldots, x_n)$ und $M_Y = (y_1, y_2, \ldots, y_n)$ die n-Tupel der Merkmalsausprägungen der Merkmale X und Y einer Stichprobe Ω und ist y_{fit} eine reelle Funktion, die X als Definitionsbereich einschliesst, so ist $r_i = y_i - y_{\text{fit}}(x_i)$ das Residuum von y_{fit} an der Stelle x_i.

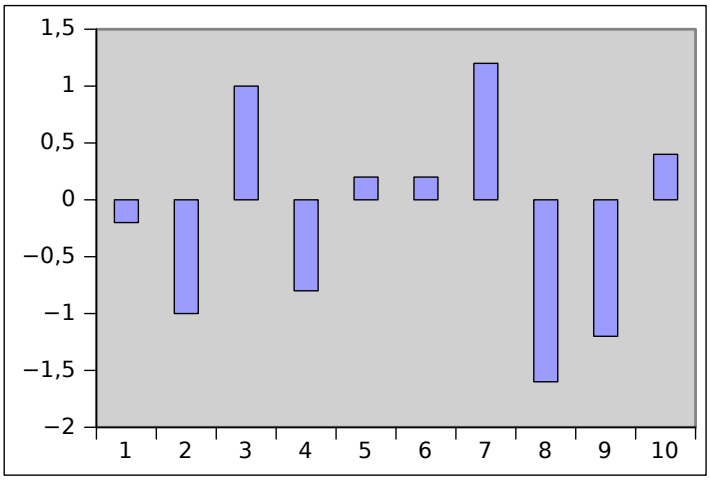
Eine einfache und für Schulzwecke oft ausreichende Methode, eine passende Regressionsfunktion y_{fit} zu finden – insbesondere wenn es sich (vermutlich) um einen linearen Zusammenhang handelt und die Regressionsfunktion eine Gerade ist – besteht darin, den Graphen von y_{fit} nach Augenmass möglichst «passend» in die Punktwolke einzuzeichnen und aus der grafischen Darstellung die Funktionsgleichung von y_{fit} abzulesen. In der Abbildung **??** ist per Augenmass eine Regressionsgerade eingezeichnet worden, so dass die Gerade möglichst dicht an den Punkten verläuft und möglichst gleichmässig Punkte ober- und unterhalb der Geraden liegen. Als Funktionsgleichung der Regressionsgeraden kann mit den üblichen Verfahren näherungsweise die Funktionsgleichung $y_{\text{fit}} = 0.8x - 14$ aus dem Schaubild ermitteln.



In der nächsten Grafik sind die Residuen $r_i = y_i - y_{\text{fit}}(x_i)$ rot eingezeichnet. Eine Möglichkeit, die Einpassung der Geraden per Augenmass zu verbessern, ist es, die Residuen als Säulendiagramm darzustellen und nach «systematischen Fehlern» zu suchen.

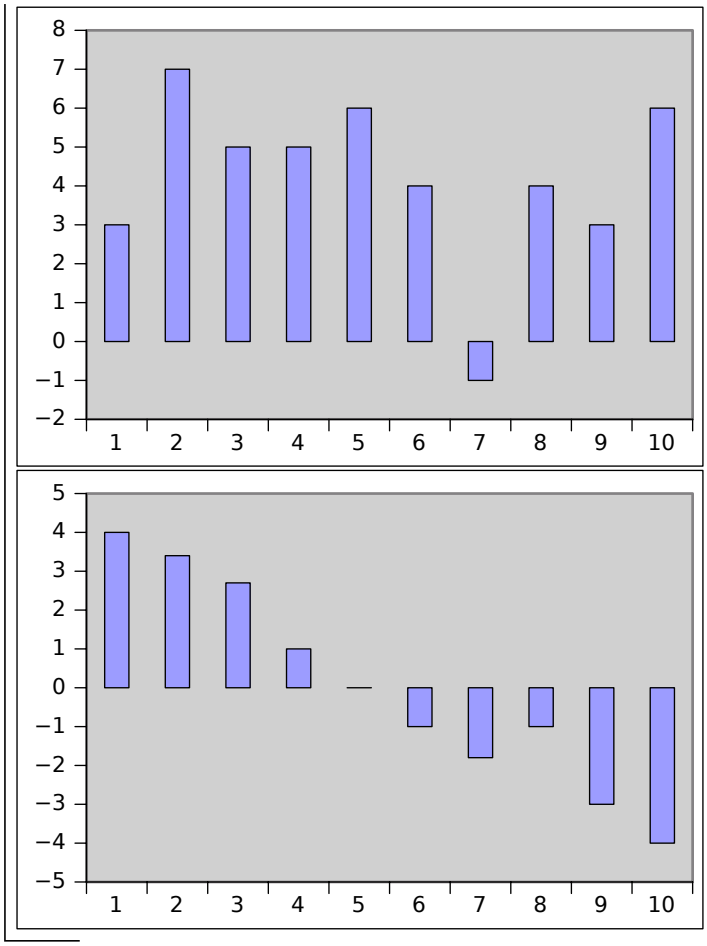


In der nächsten Abbildung sind die Residuen zur Regressionsgeraden aus der Grafik 1.4.2 aufgetragen. Man kann keinen systematischen Fehler erkennen: Die Residuen sind nicht allzu gross und verteilen sich einigermassen gleichmässig im positiven wie im negativen Bereich.



Auftrag 1.5

In den Abbildungen 1.5 und 1.5 sehen Sie Residuenplots, die auf einen systematischen Fehler beim Einzeichnen der Regressionsgeraden hindeuten. a) Beschreiben Sie, worin der Fehler besteht; b) erläutern Sie, wie sich dieser Fehler auf die Lage der Regressionsgeraden bezüglich der Punktwolke auswirkt, und c) geben Sie begründet an, wie man die Regressiongeraden verändern sollte, um den Fehler zu vermeiden, und wie sich diese Änderung in der Funktionsgleichung der Regressionsgeraden bemerkbar macht.



1.4.3 Lineare Regression mit der Methode der kleinsten Quadrate

Bei der Einpassung einer guten Regressionsgerade geht es darum, die Residuen möglichst klein zu halten. Das heisst nicht anderes, als die Summe $\sum_{i=1}^{n} |r_i|$ zu minimieren. Wenn man sich nicht auf das Augenmass verlassen kann oder will, sondern wenn man einer rechnerische Lösung sucht, die man insbesondere auch einem Computer anvertrauen kann, dann ist die Forderung, die Summe $\sum_{i=1}^{n} |r_i|$ zu minimieren, problematisch. Minimierungsprobleme lassen sich oft mit Methoden der Analysis lösen. Dass in der Summe Beträge auftreten, verhindert eine Anwendung des Ableitungskalküls. Auch andere brauchbare Lösungen ohne analytische Methoden hat man nicht gefunden. Aus diesem Grunde hat man sich dafür entschieden, statt der Summe $\sum_{i=1}^{n} |r_i|$ die Summe $\sum_{i=1}^{n} r_i^2$ zu minimieren, also nicht die Summe der absoluten Abstände der Datenpunkte zur Geraden, sondern die Summe der Quadrate der Abstände zur Geraden, die man auch schon für die Minimalitätseigenschaft des arithmetischen Mittels in Abschnitt 1.3.3 betrachtet hat. Die Minimierung dieser Summe ist über den Ableitungskalkül möglich und führt sogar für jeden Datensatz zu einer eindeutigen Lösung.

Satz 1.4

Ist $P = \{(x_i, y_i) \in \mathbb{R}^2 \mid 1 \leq i \leq n\}$ *eine Punktwolke, so ist*

$$y = a \cdot x + b$$

genau dann eine Regressionsgerade, welche die Summe

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - a \cdot x_i - b)^2$$

minimiert, wenn

$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sigma_X^2} = \frac{Cov(X,Y)}{\sigma_X^2}$$

und

$$b = \overline{y} - a\overline{x}$$

gilt.

Der Beweis benötigt Mittel aus der Differentialrechnung und findet sich beispielsweise in Kütting (2011).

Im Nenner des Terms, der die Steigung einer optimalen Regressionsgeraden nach der Methode der kleinsten Quadrate ausdrückt, tritt die empirische Varianz des Merkmals X auf. Im Zähler steht ein Ausdruck, der eine hohe Ähnlichkeit mit der Varianz hat, nur dass beide Merkmale X und Y darin «verarbeitet» werden. Diesen Ausdruck nennt man *Kovarianz* von X und Y.

Definition 1.12

Es seien $X = (x_1, x_2, \ldots, x_n)$ *und* $Y = (y_1, y_2, \ldots, y_n)$ *metrische Merkmale, dann heisst*

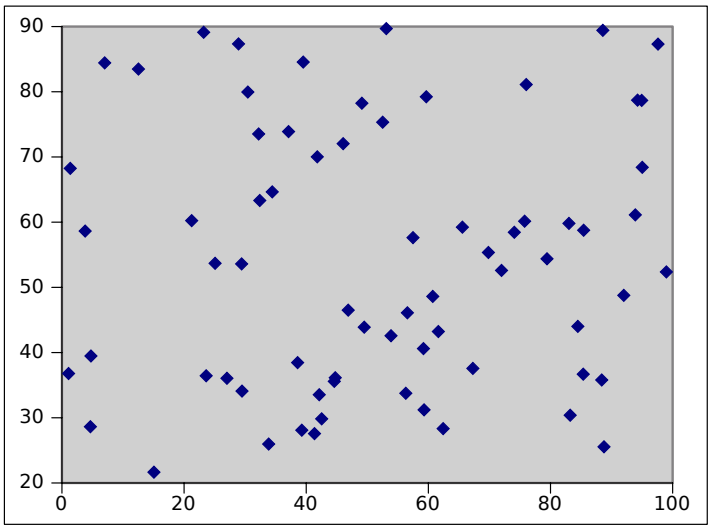
$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

die Kovarianz von X *und* Y*.*

1.4.4 Korrelationskoeffizienten

Mit der Methode der kleinsten Quadrate hat man ein Verfahren, das für jeden Datensatz mit zwei numerischen Merkmalen eine optimale Regressionsgerade angibt. So schön es ist, eine algorithmische Lösung für dieses Problem zu haben, so fraglich ist es, ob diese Lösung in jedem Fall brauchbar ist. In der Abbildung 1.4.4 ist eine Punktwolke eingezeichnet, in der kein linearer Zusammenhang (und auch nicht irgendein anderer) zwischen den x- und y-Werten erkennbar ist. Auch für diese Punktwolke würde die Methode der kleinsten Quadrate eine Regressionsgerade ermitteln. Die Summe der Residuenquadrate wäre zwar minimal, aber trotzdem noch so hoch,

dass die Abweichung von der Geraden im allgemeinen sehr gross ist. *Jede* Gerade wäre ungeeignet, die Punktwolke anzunähern, auch die Regressionsgerade nach der Methode der kleinsten Quadrate.



Erforderlich wäre also eine Entscheidung darüber, ob zwischen zwei Merkmalen überhaupt ein linearer Zusammenhang besteht. Hier wäre eine Ja-Nein-Entscheidung allerdings ungeeignet, da selbst «sehr schön lineare» Punktwolken wie jene in der Abbildung 1.4.1 nicht vollkommen auf einer Geraden liegen. Sinnvoller wäre ein *Mass*, das *Grade der Linearität* angibt. Ein solches Mass nennt man *Korrelationskoeffizienten*. Für die Interpretation dieses Masses wird folgende Konvention getroffen:

Definition 1.13

Es seien $X = (x_1, x_2, \ldots, x_n)$ und $Y = (y_1, y_2, \ldots, y_n)$ metrische Merkmale. Eine Funktion $r(X, Y)$ heisst Korrelationskoeffizient, wenn r die folgenden Eigenschaften besitzt:

1) $r(X, Y) = r(Y, X)$,
2) $r(X, X) = 1$,
3) $r(X, -X) = -1$.

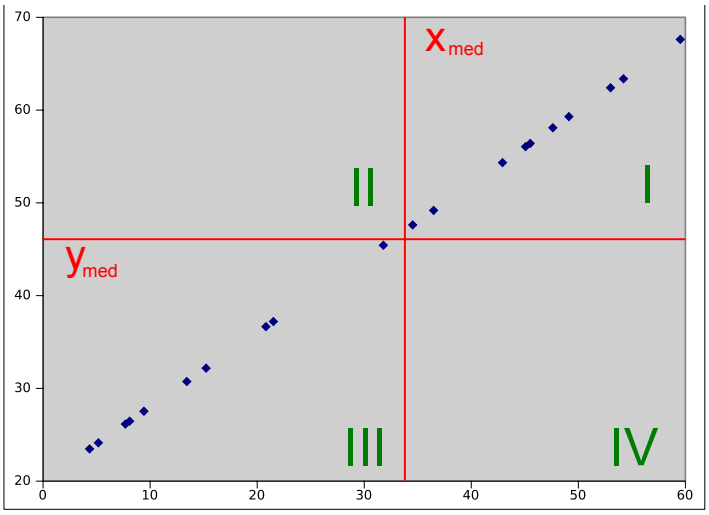
Man kann leicht zeigen, dass r nur Werte aus dem Intervall $[-1, 1]$ annimmt. Mit der zweiten und dritten Eigenschaft der Definition wird die Interpretation von r deutlich: X hängt von X klarerweise perfekt linear ab; und ebenso hängt auch X von $-X$ perfekt linear ab; die Regressionsgerade hätte allerdings eine negative statt einer positiven Steigung. Dies gilt allgemein: Je dichter r bei 1 oder -1 liegt, desto linearer ist der Zusammenhang; und je dichter r bei 0 liegt, desto geringer ist der lineare Zusammenhang. Im Falle $r = 0$ besteht überhaupt kein linearer Zusammenhang, so wie in den Grenzfällen 1 und -1 ein perfekter linearer Zusammenhang besteht. Zusätzlich gibt das Vorzeichen des Korrelationskoeffizienten das Vorzeichen einer optimalen Regressionsgerade an. Die Interpretation zwischen den Extremfällen $-1, 0$ und 1 unterliegt einer gewissen Willkür. Üblich, aber nicht zwingend ist es, die Grenzen folgendermassen zu ziehen:

Wertebereich von r	Interpretation
$-1 \leq r \leq -0.7$	starker negativer linearer Zusammenhang
$-0.7 < r < -0.3$	Grauzone
$-0.3 \leq r \leq 0.3$	kein linearer Zusammenhang
$0.3 < r < 0.7$	Grauzone
$0.7 \leq r \leq 1$	starker positiver linearer Zusammenhang

1.4.5 Der resistente Korrelationskoeffizient

Bisher ist es offen, ob sich überhaupt Funktionen definieren lassen, welche die Erfordernisse eines Korrelationskoeffizienten erfüllen. Der *resistente Korrelationskoeffizient* ist eine einfache Funktion, welche diese Aufgabe erfüllt und welche sich bei kleineren Datensätzen ohne allzu grossen Aufwand auch gut von Hand berechnen lässt. Aus diesen Gründen wird er gern an der Schule verwandt.

Die Grundüberlegung zum resistenten Korrelationskoeffizienten ist die folgenden: Nehmen wir an, die Datenpunkte lägen perfekt auf einer Geraden mit positiver Steigung. Betrachten wie nun die Werte des Merkmals X: Der Median x_{med} teilt die Werte in zwei Hälfte. Für die x-Werte, die unterhalb von x_{med} liegen, liegen auch die zugehörigen y-Werte unterhalb von y_{med}, da lineare Funktionen monoton steigen. Andererseits liegen die y-Werte, die zu x-Werten gehören, die oberhalb von x_{med} liegen, ebenfalls oberhalb von y_{med}; d. h. wenn man die beiden Medianwerte x_{med} und y_{med} als senkrechte bzw. waagerechte Gerade in die Punktwolke einzeichnet, dann liegen sämtliche Punkte der Wolke im I. und III. Quadranten, die durch das Mediankreuz gebildet werden. Die Abbildung 1.4.5 zeigt einen solchen Idealfall.



Die Idee des resistenten Korrelationkoeffizienten besteht darin zu messen, wie sehr eine Punktwolke diesem Idealfall nahe kommt, d. h. letztendlich zu zählen, wie viele der Punkte im I. und III. Quadranten liegen. Genauer:

Definition 1.14

Man zählt die Anzahl der Punkte, die im I. und III. Quadranten liegen. Diese Anzahl sei n^+. Falls ein Punkt nicht eindeutig in einem der Quadranten, sondern auf mindestens einer der beiden Mediangeraden liegt, dann wird er mit 0.5 zu n^+ hinzugezählt. Man berechnet den resistenten Korrelationskoeffizienten r_{rst} dann

durch

$$r_{rst} = \sin\left(\left(\frac{n^+}{n} - 0.5\right) \cdot 180^\circ\right)$$

1.4.6 Der Korrelationskoeffizient nach Pearson

Wie Sie eben selbst nachgewiesen haben, weist der resistente Korrelationskoeffizient unter Umständen auch dann einen linearen Zusammenhang aus, wenn gar keiner vorliegt. Das ist wenig erfreulich. Aus diesem Grunde werden lieber andere Korrelationskoeffizienten benutzt, die in der Regel aber einen höheren Rechenaufwand erfordern und weniger anschaulich zu begründen sind. Am gebräuchlichsten ist der Korrelationskoeffizient von Pearson.

Definition 1.15

Es seien $X = (x_1, x_2, \ldots, x_n)$ und $Y = (y_1, y_2, \ldots, y_n)$ metrische Merkmale. Dann ist

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

der Korrelationskoeffizient nach Pearson.

Warum das eine sinnvolle Definition für einen Korrelationskoeffizient ist, lässt sich am besten aus Sicht der analytischen Geometrie verstehen. Betrachten wir die beiden Vektoren

$$\vec{x} = \begin{pmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_n - \overline{x} \end{pmatrix} \text{ und } \vec{y} = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix}$$

Diese Vektoren beinhalten als Komponenten die Abweichungen der einzelnen Messwerte vom arithmetischen Mittel. Verwendet man das Skalarprodukt und die euklidische Norm, so sieht der Pearsonsche Korrelationskoeffizient in vektorieller Schreibweise folgendermassen aus:

$$r = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \cdot ||\vec{y}||}$$

Über diesen Term werden überlicherweise Winkel definiert. Der Winkel $\sphericalangle$ zwischen $\vec{x}$ und $\vec{y}$ ist

$$\sphericalangle(\vec{x}, \vec{y}) = \arccos\left(\frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| \cdot ||\vec{y}||}\right) = \arccos(r)$$

Nun wird deutlich: Wenn r nahe bei 1 liegt, dann schliessen $\vec{x}$ und $\vec{y}$ nahezu einen Winkel von $0°$ ein, d. h. die Änderungen von X und Y «laufen» nahezu in «dieselbe Richtung», und da der Winkel nicht von der Länge von $\vec{x}$ und $\vec{y}$ abhängt, sind die beiden Änderungen sogar proportional zueinander, d. h. ein linearer Zusammenhang liegt vor. Bei r nahe bei Null liegt der Winkel zwischen $vecx$ und $\vec{y}$ nahe bei $90°$, d. h. wenn X sich in die einer Richtung ändert, ändert Y sich in eine «ganz andere» Richtung als X, d. h. es liegt kein Zusammenhang zwischen X und Y vor, erst recht kein linearer. Im Fall, dass r nahe bei -1 liegt, ist der Winkel zwischen $\vec{x}$ und $\vec{y}$ ungefähr $180°$, d. h. wenn sich X verändert, verändert sich Y genau in die entgegengesetzte Richtung, und zwar proportional zu X. Also liegt dann ein negativer linearer Zusammenhang vor.

2 Endliche Wahrscheinlichkeitsräume

2.1 Grundlegung

2.1.1 Grundtatsachen der Wahrscheinlichkeitsrechnung

Definition 2.1 (*Zufallsexperiment*)

Ein Zufallsexperiment ist ein realer Vorgang (Experiment) unter exakt festgelegten Bedingungen mit den folgenden Eigenschaften

- *Alle möglichen Ergebnisse des Experiments sind vorab bekannt.*
- *Das Ergebnis eines einzelnen Experiments kann nicht vorhergesagt werden (Zufälligkeit).*
- *Das Experiment kann unter identischen Bedingungen beliebig oft wiederholt werden.*

Ergebnisse und Ereignisse

Die Unterscheidung zwischen Ergebnis und Ereignis scheint zunächst mühsam, später brauchen wir aber genau diese Unterscheidung, um uns richtig über die Probleme in der Wahrscheinlichkeitsrechnung unterhalten zu können.

Definition 2.2 (*Ergebnisraum*)

Eine Menge $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_n\}$ heisst Ergebnisraum eines Zufallsexperiments, wenn jedem Versuchsausgang höchstens ein Element ω_i aus Ω zugeordnet ist. Die ω_i heissen dann die Ergebnisse des Zufallsexperiments.

Definition 2.3 (*Ereignisraum*)

Jede Teilmenge des endlichen Ergebnisraums Ω heisst Ereignis. A tritt genau dann ein, wenn sich ein Versuchsergebnis ω einstellt, das in A enthalten ist. Die Menge aller Ereignisse heisst Ereignisraum, bezeichnet mit $\wp(\Omega)$.

1. *Die einelementigen Teilmengen von Ω, also die Teilmengen, die genau ein Ergebnis enthalten, werden als Elementarereignisse bezeichnet.*
2. *Die Teilmenge Ω von Ω, also die Ergebnismenge selbst, wird als sicheres Ereignis bezeichnet.*
3. *Die leere Menge $\emptyset$ heisst unmögliches Ereignis.*
4. *Die Ereignisse A und B heissen* unvereinbar *oder auch* disjunkt *genau dann, wenn*

   $A \cap B = \emptyset$.
5. *Mit $\bar{A}$ wird das Gegenereignis zu A bezeichnet. $\bar{A}$ umfasst ω ohne A. Also $\bar{A} \cup A = \Omega$ und $\bar{A} \cap A = \emptyset$.*

Beispiel 2.1

Beim Werfen eines 4-seitigen (tetraederförmigen) Würfels lautet der
Ergebnisraum=Ω= {1,2,3,4} und der
Ereignisraum=
$\{\{\}, \{1\}, \{2\}, \{3\}, \{4\}, \ \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\},$
$\{1,2,3\}, \{1,3,4\}, \{1,2,4\}, \{2,3,4\}, \ \{1,2,3,4\}\}$
Insbesondere gehören das leere Ereignis $\{\} = \emptyset$ und das sichere Ereignis $\{1,2,3,4\} = \Omega$ zum Ereignisraum.

Empirische Zufallsexperimente

Definition 2.4 (*Relative Häufigkeit*)

Tritt ein Ereignis A bei n Versuchen k-mal ein, so heisst $h_n(A) = \frac{k}{n}$ die relative Häufigkeit des Ereignisses A in dieser Versuchsfolge.

Wahrscheinlichkeit und relative Häufigkeit sind grundsätzlich verschiedene Begriffe: Wahrscheinlichkeiten dienen der Prognose. Sie geben Auskunft über Chancen in bevorstehenden Zufallsversuchen. Dagegen machen relative Häufigkeiten immer Aussagen über durchgeführte Zufallsversuche.

Satz 2.1 (*Eigenschaften der relativen Häufigkeiten*)

- $0 \leq h_n(A) \leq 1$
- $h_n(A) = \sum_{\omega \in A} h_n(\omega)$ Das Zeichen $\sum$ heisst Summenzeichen. Hier bedeutet es: es werden alle relativen Häufigkeiten $h_n(\omega)$ von Ergebnissen ω, die zum Ereignis A gehören zusammengezählt.
- $h_n(\emptyset) = 0$
- $h_n(\Omega) = 1$
- $h_n(A \cup B) = h_n(A) + h_n(B) - h_n(A \cap B)$
- $A \cap B = \emptyset \ \rightarrow \ h_n(A \cup B) = h_n(A) + h_n(B)$
- $h_n(\bar{A}) = 1 - h_n(A)$

Satz 2.2 (*Empirisches Gesetz der grossen Zahlen*)

Die relative Häufigkeit eines Ereignisses stabilisiert sich mit zunehmender Versuchszahl um einen festen Wert.

Theoretische Zufallsexperimente

Definition 2.5 (*Endlicher Wahrscheinlichkeitsraum – Kolmogorov-Axiome*)

Die Funktion $P : \wp(\Omega) \rightarrow \mathbb{R}$ *heisst Wahrscheinlichkeitsverteilung über dem Ergebnisraum* Ω*, wenn sie die folgenden Eigenschaften hat:*

1. **Nichtnegativität** $P(A) \geq 0$ *für alle* $A \in \wp(\Omega)$
2. **Normierung** $P(\Omega) = 1$
3. **Additivität** $P(A \cup B) = P(A) + P(B)$ *für alle* $A, B \in \wp(\Omega)$ *mit* $A \cap B = \emptyset$.

Der Funktionswert von $P(A)$ *heisst die Wahrscheinlichkeit des Ereignisses A. Das Tripel* $(\Omega, \wp(\Omega), P)$ *oder auch das Paar* (Ω, P) *heisst endlicher Wahrscheinlichkeitsraum.*

Folgerungen

Beweise finden sich beispielsweise in Kütting (2011).

Folgerung 2.1

Folgerungen aus dem Axiomensystem, Satz 2.5

1. **Gegenereignis** $P(\bar{A}) = 1 - P(A)$.
2. *Die Wahrscheinlichkeit $P(A)$ eines Ereignisses A nimmt nur Werte zwischen Null und eins an:*
   $0 \leq P(A) \leq 1$.
3. **Unmögliches Ereignis** $P(\emptyset) = 0$.
4. **Verallgemeinerte Fassung von Axiom 3** *Sind je zwei der Ereignisse $A_1, A_2, \ldots A_n$ unvereinbar, so gilt*

   $$P(A_1 \cup A_2 \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

5. **Allgemeine Additionsregel/ Zerlegungssatz** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6. **Formel von Sylvester für n=3**
   $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Äquivalentes Axiomensystem

Das folgende Axiomensystem ist zu Definition 2.5 äquivalent. Es betont die Rolle der Elementarereignisse - lässt sich aber nicht auf unendliche Wahrscheinlichkeitsräume verallgemeinern.

Definition 2.6

Es sei $\Omega = \{\omega_1, \omega_2 \ldots \omega_n\}$ *eine nichtleere, endliche Ergebnismenge, und seien* $\{\omega_i\}$ *für* $i = 1, \ldots n$ *die Elementarereignisse. Die Abbildung*

$$P : \wp(\Omega) \to \mathbb{R}$$

ist genau dann eine Wahrscheinlichkeitsverteilung, wenn gilt:

- *Für alle Elementarereignisse gilt* $P(\omega_i) \geq 0$.
- *Die Summe der Wahrscheinlichkeiten aller Elementarereignisse ist 1, das heisst*

$$\sum_{\omega \in \Omega} P(\omega) = 1$$

- *Die Wahrscheinlichkeit des unmöglichen Ereignisses ist 0, das heisst* $P(\emptyset) = 0$.
- *Die Wahrscheinlichkeit eines Ereignisses* A *ist die Summe der Wahrscheinlichkeiten seiner Elementarereignisse, das heisst*

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

2.1.2 Laplace-Experimente

Experimente, bei denen alle Ausgänge gleich wahrscheinlich sind, werden *Laplace–Experimente* genannt. Diese beschränken sich nicht auf Würfel. Es gibt auch Laplace–Münzen, Laplace–Roulette und so weiter.

Satz 2.3 (*Wahrscheinlichkeiten bei Laplace-Experimenten*)

Bezeichne mit A *ein Ereignis, dass mehrere Ausgänge des Laplace-Experimentes zusammenfasst. Diese Ausgänge werden günstig genannt. Dann gilt*

$$P(A) = \frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl der möglichen Fälle}}$$

Mit dieser Regel lassen sich nun viele Wahrscheinlichkeiten berechnen.

Beispiel 2.2

Eine Lotterie umfasst 400 Lose, darunter 10 Hauptgewinne, 90 Trostpreise und 300 Nieten. Berechnet werden die Wahrscheinlichkeiten für die folgenden Ereignisse:

Ereignis A: Das Los ist ein Hauptgewinn P(A)=10/400
Ereignis B: Das Los ist ein Trostpreis P(B)=90/400
Ereignis C: Das Los ist ein Gewinn P(C)=100/400

2.1.3 Pfadregeln

Oft bestehen Zufallsexperimente aus Teilexperimenten, oder ein Zufallsexperiment lässt sich aus Teilexperimenten zusammengesetzt denken. Diese Teilexperimente bezeichnet man auch als Stufen im Gesamtexperiment. Allgemein wird von mehrstufigen Zufallsexperimenten gesprochen.

Regel Die Ergebnisse eines n-stufigen Zufallsexperiments sind n-Tupel $(a_1|a_2|a_3|\ldots|a_n)$, kurz auch $a_1a_2a_3\ldots a_n$, wobei a_i irgendein Ergebnis des i-ten Teilexperiments ist. Ω ist dann die Menge aller dieser n-Tupel.

Beispiel 2.3

Zwei Münzen werden gleichzeitig geworfen. Dann lautet der Ergebnisraum $\Omega = kk, kz, zz$.[1] Es gilt:
Ereignisraum= $\{\{\}, \{kk\}, \{kz\}, \{zz\}, \{kk, zk\}, \{kk, zz\}, \{kz, zz\}, \{kk, kz, zz\}\}$

Mehrstufige Zufallsexperimente können mit Hilfe von Baumdiagrammen dargestellt werden. Ergebnisse eines mehrstufigen Versuchs werden durch Streckenzüge (Pfade) im Baumdiagramm dargestellt. Zu jedem Ergebnis gehört ein Pfad. Dargestellt ist dies zum Beispiel in Kütting (2011) auf den Seiten 45, 46 und 137.

Beispiel 2.4

Dreimaliges Werfen einer Münze

Beispiel 2.5

Gleichzeitiges Werfen einer Münze und eines Würfels

Satz 2.4 (*Multiplikationspfadregel*)

Die Wahrscheinlichkeit eines Elementarereignisses in einem mehrstufigen Zufallsexperiment ist gleich dem Produkt der Wahrscheinlichkeiten auf dem Pfad, der zu diesem Elementarereignis führt.

Satz 2.5 (*Additionspfadregel*)

Die Wahrscheinlichkeit eines Ereignisses ist gleich der Summe der Wahrscheinlichkeiten der Pfade, die zu diesem Ereignis führen.

2.1.4 Aufgaben

Aufgabe 2.1

Bei einem Zufallsversuch gilt $\Omega = \{a, b, c\}$. Es gilt $P(\{b\}) = 0.2$ und $P(\{c\}) = 0.6$.
a) Wie gross ist $P(\{a\})$?

b) Bestimmen Sie die Wahrscheinlichkeiten für alle Ereignisse.

Aufgabe 2.2

Bei einem Zufallsversuch gilt $\Omega = \{a_1, a_2, a_3, a_4\}$.
Welche der folgenden Wahrscheinlichkeiten sind gemeinsam möglich, welche sind nicht möglich? Begründen Sie Ihre Aussage.

a) $P(\{a_1, a_2\} = 0.5$ und $P(\{a_3, a_4\}) = 0.4$

b) $P(\{a_3, a_4\} = 0.6$ und $P(\{a_3\}) = 0.4$

c) $P(\{a_1, a_2\}) = 0.6$ und $P(\{a_2, a_3, a_4\} = 0.5$

d) $P(\{a_4\}) = 0.3$ und $P(\{a_2, a_3\}) = 0.2$ und $P(\{a_1, a_2\}) = 0.6$

Aufgabe 2.3

Betrachtet werden Würfel, die keine Laplace-Würfel sind. Vielmehr gelten die folgenden Wahrscheinlichkeiten.

	1	2	3	4	5	6
p	0.2	0.1	0.1	0.1	0.2	0.3

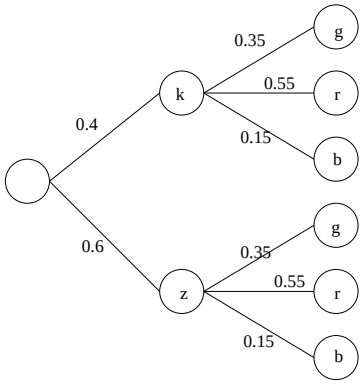
Es werden drei dieser Würfel gleichzeitig geworfen. Wie gross ist die Wahrscheinlichkeit, dass die Summe der Augenzahlen 8 ist?

Aufgabe 2.4

Die Kontrolle eines Produkts wird mehrfach durchgeführt. Die letzte Kontrolle ist dabei besonders gründlich. 65 Prozent der Fehler werden entdeckt und korrigiert. Bei den ersten beiden Kontrollen werden nur 50 beziehungsweise 30 Prozent der Fehler gefunden (und jeweils korrigiert). Mit welcher Wahrscheinlichkeit wird ein Fehler entdeckt?

Aufgabe 2.5

Handelt es sich bei dem nebenstehenden Baumdiagramm um die Beschreibung eines Zufallsexperiments? Warum nicht?



Aufgabe 2.6

Von einem Kartenspiel werden in dieser Aufgabe nur ein Bube, eine Dame und ein König verwendet. Es werden zwei Karten gezogen. Welche Elemente enthält der Ergebnisraum

a) wenn nach dem Ziehen der ersten Karte diese zurückgelegt wird,

b) wenn die beiden Karten nacheinander gezogen werden, die erste aber nicht zurückgelegt wird.

c) beide Karten gleichzeitig gezogen werden.

d) Nun werden zwei aus Bube, Dame, König bestehende Spiele verwendet. Aus jedem der beiden "Spiele" wird eine Karte gezogen; es wird nicht notiert, welche Karte aus welchem Spiel stammt. Beschreiben Sie auch hier den Ergebnisraum.

Aufgabe 2.7

Formulieren Sie das Gegenereignis in Worten *und* auch als Menge. Bestimmen Sie die Wahrscheinlichkeit von Ereignis und Gegenereignis.
Die Aufgabenteile beziehen sich auf die Aufgabenteile von Aufgabe 2.6.

a) Teil a, Ereignis *" Weder ein Bube noch ein König werden gezogen."*

b) Teil b, Ereignis *" An erster Stelle wird weder ein König noch ein Bube gezogen."*

c) Teil c, Ereignis *" Weder ein Bube noch ein König werden gezogen."*

Aufgabe 2.8

Nun werden den drei Karten Werte zugewiesen. Bube=0, Dame=1 und König=2. Berechnen Sie bei allen vier Teilaufgaben der vorletzten Aufgabe die Wahrscheinlichkeit, dass die Summe der Werte gleich 3 ist.

Aufgabe 2.9

a) Beschreiben Sie die Zufallsexperimente a) und b) aus Aufgabe 2.6 mit Hilfe von Baumdiagrammen (inklusive Wahrscheinlichkeiten an den Pfeilen).

b) Berechnen Sie die Wahrscheinlichkeiten für das Ereignis *" Dame und König"* der Teilaufgaben d) und c) von 2.6. Dabei *müssen* die Baumdiagramme aus dem ersten Teil von 2.9 verwendet werden. (Ja, das geht, obwohl es andere Teilaufgaben sind.)

Lösung 2.1: a) 0.2

Lösung 2.2: nein, ja, ja, ja

Lösung 2.3: 0.078

Lösung 2.4: 0.88 **Lösung 2.5:** nein, Wahrscheinlichkeiten in der letzten Stufe zu gross.

Lösung 2.6: a) { BB, BD, BK, DB, DD, DK, KB, KD, KK} b) { BD, BK, DB, DK, KC, KD}

c) { BD, BK, DK} d) { BB, BD, BK, DD, DK, KK}

Lösung 2.7: a)„Bube oder König wird gezogen"b) „An erster Stelle wird Bube oder König gezogen"c) „Es wird gezogen"

Lösung 2.8: $2/9$; $1/3$; $1/3$; $2/9$

Lösung 2.9: b) $1/3$; $2/9$

2.2 Kombinatorisches Zählen

2.2.1 Das Zählprinzip

Beispiel 2.6

Drei Parallelklassen haben 19, 22 bzw. 17 Schülerinnen und Schüler. Aus jeder Klasse wird ein Klassensprecherin oder ein Klassensprecher gewählt. Wieviele Möglichkeiten der Zusammensetzung der "Konferenz der Klassensprechenden" gibt es?
Antwort 19·22·17

Satz 2.6 (*Zählprinzip*)

Gegeben sind k Mengen. Die erste Menge hat n_1 Elemente, die zweite Menge hat n_2 Elemente, und so weiter bis zur letzten Menge, die n_k Elemente hat.
Aus jeder dieser Mengen wird ein Element ausgewählt. Dann erhalten wir die Gesamtzahl der möglichen Auswahlen, indem wir die Anzahlen der Elemente der Mengen miteinander multiplizieren.

$$\text{Anzahl der Möglichkeiten} = n_1 \cdot n_2 \ldots n_k$$

Diese Art, Anzahlen zu berechnen nennt sich das Zählprinzip. Es liegt der ganzen Kombinatorik zugrunde.

Beispiel 2.7

An einem Pferderennen nehmen 20 Pferde teil. Bei einem Wettabschluss sollen die ersten drei Plätze richtig angegeben werden. Wie viele Möglichkeiten gibt es für die Besetzung der ersten drei Plätze?
Für den ersten Platz haben wir 20 Möglichkeiten, für den zweiten dann noch 19, für den dritten noch 18. Insgesamt also 20·19·18=6840 Möglichkeiten.

Beispiel 2.8

Beim Würfeln mit 4 (verschiedenen) Würfeln gibt es je 6 Möglichkeiten, insgesamt also 6·6·6·6=1296

Definition 2.7 (*Potenzmenge*)

Die Menge aller Teilmengen einer Menge Ω heisst Potenzmenge $\wp(\Omega)$ von Ω, geschrieben $\wp(\Omega)$.
Ein Ereignis ist also ein Element von $\wp(\Omega)$.

Satz 2.7 (*Anzahl der Elemente*)

Wenn $\sharp\Omega = n$, dann gilt $\sharp\wp(\Omega) = 2^n$.

Beweis:

Folgendermassen lässt sich eine beliebige Teilmenge (ein beliebiges Ereignis) $A \in \wp(\Omega)$ erhalten:

Wir führen ein n-stufiges Zufallsexperiment durch, bei dem die Teilexperimente jeweils den Ausgang 0 oder 1 haben. Dabei bedeutet 0 im i-ten Teilexperiment, dass $\omega_i \notin A$ und analog bedeutet 1 im i-ten Experiment, dass $\omega_i \in A$.

Es entspricht also $\{0, 1, 1, 0\}$ Der Teilmenge $\{\omega_2, \omega_3\}$ einer vierelementigen Menge.

Wir müssen nun zählen, wie viele Versuchsausgänge unser n-elementiges Zufallsexperiment hat. Nach dem Zählprinzip, Satz 2.6, sind dies

$$2 \cdot 2 \cdot 2 \ldots 2 = 2^n$$

□

Spezialfall zum Zählprinzip: Es gibt n Mengen. Die erste Menge hat n Elemente, die zweite Menge eines weniger und so weiter bis die letzte Menge nur noch ein Element hat. Dann müssen zur Ermittlung der Gesamtzahl alle natürlichen Zahlen zwischen 1 und n miteinander multipliziert werden.
Das ist so wichtig, dass es in der Mathematik eine eigene Bezeichnung bekommen hat: $n!$ (sprich n Fakultät) bezeichnet das Produkt der ersten n natürlichen Zahlen).

Beispiel 2.9

Wie viele Möglichkeiten gibt es, 8 Türme auf einem Schachbrett so anzuordnen, dass sie sich nicht gegenseitig bedrohen?

2.2.2 Das Urnenmodell

Fast alle Probleme in der Kombinatorik lassen sich auf das sogenannte Urnenmodell zurückführen.
In einer Urne liegen n verschieden bezeichnete Kugeln. Davon werden k gezogen. Es sind verschiedene "Zugtechniken" möglich

- Nach jeder Ziehung wird die Kugel zurückgelegt, oder nicht zurückgelegt. (Im ersten Fall kann bei verschieden farbigen Kugeln das Ergebnis rot, rot, grün lauten, im zweiten Fall nicht)
- Es kommt auf die Reihenfolge an, oder auch nicht (im ersten Fall sind "rot, schwarz, grün" und "grün, rot, schwarz" verschiedene Ereignisse, im zweiten Fall nicht).

Die folgende Tabelle fasst die Möglichkeiten zusammen. Vorweg noch eine Definition.

Definition 2.8 (*Binomialkoeffizienten*)

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

Das Urnenmodell

von n Kugeln werden k gezogen	unter Beachtung der Reihenfolge (Permutation)	ohne Beachtung der Reihenfolge (Kombination)
mit Wiederholung	n^k	$\binom{n+k-1}{k}$
ohne Wiederholung	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Bei den folgenden Beispielen liegen in einer Urne jeweils 4 Kugeln, nummeriert von 1 bis 4. Es werden 3 gezogen.

Beispiel 2.10

Ziehen mit Zurücklegen, unter Beachtung der Reihenfolge: 4^3 Möglichkeiten.

Beispiel 2.11

Ziehen ohne Zurücklegen, unter Beachtung der Reihenfolge: In einer Urne liegen vier Kugeln, nummeriert von 1 bis 4. Es werden nacheinander 4 Kugeln entnommen. Mögliche Ergebnisse sind:
(1,2,3); (1,3,2); (2,1,3); (2,3,1); (3;1;2); (3;2;1)
(1,2,4); (1,4,2); (2,1,4); (2,4,1); (4;1;2); (4;2;1)
(1,4,3); (1,3,4); (4,1,3); (4,3,1); (3;1;4); (3;4;1)
(4,2,3); (4,3,2); (2,4,3); (2,3,4); (3;4;2); (3;2;4)

Beispiel 2.12

Ziehen ohne Zurücklegen, ohne Beachtung der Reihenfolge: In einer Urne liegen vier Kugeln, nummeriert von 1 bis 4. Es werden gleichzeitig 4 Kugeln entnommen. Mögliche Ergebnisse sind:
(123); (124); (134); (234)

Beispiel 2.13

Ziehen mit Zurücklegen, ohne Beachtung der Reihenfolge: Nach dem Ziehen wird zurückgelegt, es wird drei Mal gezogen. Einige mögliche Ergebnisse sind: (es fehlen diejenigen, bei denen eine vier vorkommt.)
(1,1,1); (1,1,2); (1,1,3); (1,2,2); (1,2,3); (1,3,3)
(2,2,2); (2,2,3); (2,3,3); (3,3,3)
Hier ist zu beachten, dass bei der Berechnung von Wahrscheinlichkeiten nicht alle Elementarereignisse gleich wahrscheinlich sind. Bei (1,1,1) muss drei Mal hintereinander die 1 gezogen werden, bei (1,1,2) gibt es drei Möglichkeiten, an welcher Stelle die 2 gezogen wird: dieses Elemenarereignis ist drei Mal so wahrscheinlich.

3 Diskrete Verteilungen

3.1 Zufallsgrössen

Die Definitionen in diesem Kapitel gelten für endliche Ergebnismengen und, mit kleinen Einschränkungen auch für abzählbare (diskrete) Ergebnismengen (die Ergebnisse lassen sich durchnummerieren; es können aber unendlich viele sein.)

Diese Einschränkung wird nicht immer konsequent erwähnt.

Eine Verallgemeinerung folgt in einem weiteren Kapitel.

3.1.1 Zufallsgrössen und Verteilungen

Definition 3.1

Eine diskrete Zufallsgrösse ist eine auf einem abzählbaren Ergebnisraum Ω definierte Funktion.

Definition 3.2

Für eine (diskrete) Zufallsgrösse (oder Zufallsvariable) X ist $X = k$ das Ereignis: es tritt ein Ergebnis auf, dessen Wert bei Anwendung der Zufallsgrösse X gleich k ist.

$$\{X = r\} = \{\omega \in \Omega \,|X(\omega) = r\}$$

Definition 3.3

Mit $P(X = k)$ wird die Wahrscheinlichkeit des Ereignisses $X = r$ bezeichnet:

$$P(X = k) := P(\{\omega \in \Omega \,|X(\omega) = k\}$$

Zu beachten ist die Abweichung der Notation von der Analysis.

Definition 3.4

Die Auflistung aller Wahrscheinlichkeiten $P(X = r)$ ist die Wahrscheinlichkeitsverteilung der Zufallsgrösse.

Beispiel 3.1

Es wird zwei Mal gewürfelt. $X =$„Augensumme“.

Wahrscheinlichkeits-verteilung	zugehörige Ergebnisse	Wahrscheinlichkeits-verteilung	zugehörige Ergebnisse
$P(X = 2) = 1/36$	11	$P(X = 3) = 1/18$	12;21
$P(X = 4) = 1/12$	13;22;31	$P(X = 5) = 1/9$	14;23;32;41
$P(X = 6) = 5/36$	15;24;33;42;51	$P(X = 7) = 1/6$	16;25;34;43;52;61
$P(X = 8) = 5/36$	26;35;44;53;62	$P(X = 9) = 1/9$	36;45;54;63
$P(X = 10) = 1/12$	46;55;64	$P(X = 11) = 1/18$	56;65
$P(X = 12) = 1/36$	66		

Beispiel 3.2

Es wird zwei Mal gewürfelt. $X =$„Maximum der Augenzahlen“.

Wahrscheinlich-keitsverteilung	zugehörige Ergebnisse	Wahrscheinlich-keitsverteilung	zugehörige Ergebnisse
$P(X = 1) = 1/36$	11	$P(X = 4) = 7/36$	14;24;34;44;41;42;43
$P(X = 2) = 1/12$	12;21;22	$P(X = 5) = 1/4$	15;25;35;45;55;54;53;52;51
$P(X = 3) = 5/36$	13;23;33;32;31	$P(X = 6) = 11/36$	16;26;36;46;56;66;65;64;63;62;61

Hinweise

1. $P(X = k)$ ist für alle $k \in \mathbb{R}$ definiert. Für alle $k \in \mathbb{R}$, die nicht zur Wertemenge $X(\Omega)$ gehören, gilt nach obiger Definition $P(X = k) = 0$, denn für alle $k \notin X(\Omega)$ gilt

   $$\{\omega \in \Omega \mid X(\omega) = k\} = \emptyset$$

2. Ist M eine Teilmenge von $\mathbb{R}$, so berechnet sich die Wahrscheinlichkeit, dass die Zufallsgrösse einen Wert in M annimmt zu

   $$P(X^{-1}(M) := \sum_{k \in M} P(X = k).$$

3.1.2 Kumulative Verteilungsfunktion

Definition 3.5 (*Verteilungsfunktion*)

Sei X eine diskrete Zufallsgrösse, dann heisst die Funktion

$$F : \mathbb{R} \to [0, 1] mit F(x) := P(X \leq x) := \sum_{x_i \leq x} P(X = x_i)$$

die (kumulative) Verteilungsfunktion.

Diese summiert also alle Wahrscheinlichkeiten für Werte der Zufallsgrösse kleiner gleich x auf. Das Wort kumulativ wird oft weggelassen. Es besteht dann die Gefahr der Verwechslung mit der Wahrscheinlichkeitsverteilung. Der Begriff ist aber sinnvoll, wie die folgenden Eigenschaften zeigen:

Satz 3.1

Sei F eine Verteilungsfunktion. Dann gilt:

1. $0 \leq F(x) \leq 1$ *für alle* $x \in \mathbb{R}$.
2. *Sind* a, b *beliebige reelle Zahlen mit* $a < b$*, dann gilt:*

$$P(a < X \leq b) = F(b) - F(a).$$

3. F *ist eine monoton steigende Funktion.*

Beachten Sie die Verwandtschaft der zweiten Bedingung mit einer Stammfunktion in der Integralrechnung.

3.1.3 Erwartungswert und Streuung

Definition 3.6

Der Erwartungswert μ einer Zufallsgrösse ist:

$$\mu = E(x) = \sum a_k \cdot P(X = a_k)$$

wobei die Summe alle Werte a_k der Zufallsgrösse umfasst.

Bemerkung: Bei Zufallsgrössen mit (abzählbar) unendlich vielen Werten muss die absolute Konvergenz gefordert werden. Das spielt aber bei unseren Beispielen keine Rolle.

Bemerkung: Natürlich gilt stets

$$\sum_{a_k \in Im(X)} P(X = a_k) = 1$$

Beispiel 3.3

In einer Prüfung gibt es die folgenden Noten

Note	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
Anzahl	0	0	0	2	3	3	5	6	3	2	1

Der Mittelwert beträgt $2.5 \cdot 0.08 + 3 \cdot 0.12 + 3.5 \cdot 0.12 + 4 \cdot 0.2 + 4.5 \cdot 0.24 + 5 \cdot 0.12 + 5.5 \cdot 0.08 + 6 \cdot 0.04 = 4.14$.
Vom Erwartungswert wird bei einer *theoretischen* Wahrscheinlichkeitsverteilung gesprochen.
Beispiel: Es wird in einer Klasse davon ausgegangen, dass eine beliebige Person in der Klasse mit einer Wahrscheinlichkeit von 0.08 eine 2.5 schreibt usw. Das ergäbe dann die obige Wahrscheinlichkeitsverteilung.
Warnung: Der Erwartungswert kann oft nicht als Wert der Wahrscheinlichkeitsverteilung auftreten: niemand schreibt eine 4.14, es wird aber ein Notenschnitt von 4.14 erwartet.

Definition 3.7

Varianz einer Zufallsgrösse:

$$V(X) = \sum (a_k - \mu)^2 \cdot P(X = a_k)$$

Definition 3.8

Standardabweichung und Streuung σ:

$$\sigma = \sqrt{V(X)}$$

Analoge Definitionen gelten für empirische Verteilungen.

Die beiden Grössen messen, wie gross die Abweichung vom Mittelwert ist. Grosse Abweichungen werden stark gewertet. Im obigen Beispiel ist die Standardabweichung (auch Streuung genannt):

$\sigma = (2.5 - 4.14)^2 \cdot 0.08 + (3 - 4.14)^2 \cdot 0.12 + (3.5 - 4.14)^2 \cdot 0.12 + (4 - 4.14)^2 \cdot 0.2 + (4.5 - 4.14)^2 \cdot 0.24 + (5 - 4.14)^2 \cdot 0.12 + (5.5 - 4.14)^2 \cdot 0.08 + (6 - 4.14)^2 \cdot 0.04 = 0.93.$

Beispiel 3.4

Eine Prüfung mit gleichem Mittelwert und viel kleinerer Streuung ($\sigma = 0.64$) wäre

Note	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
Anzahl	0	0	0	0	2	5	7	7	3	1	0

Satz 3.2 (*Rechenregeln für Erwartungswert und Varianz*)

Sei X eine Zufallsgrösse und seien a und b reelle Zahlen. Dann gilt:

1. $E(aX + b) = aE(x) + b$ *und*
2. $V(aX + b) = a^2V(x)$

Ein Beweis findet sich zum Beispiel in Kütting (2011). Wichtig ist insbesondere, dass sich die Varianz nicht ändert, wenn die Zufallsgrösse um b verschoben wird.

Beispiel 3.5

Sei X die Zufallsgrösse, die jeder Person die Punktzahl bei einer Prüfung zuordnet. Dann könnte $Y = \frac{1}{5} \cdot X + 1$ die Zufallsgrösse sein, die jeder Person die Note zuordnet.
Ist $E(X) = 18$ und $\sigma(X) = 4$, so ist $E(Y) = 4.6$ und $\sigma(Y) = 0.8$

Satz 3.3

Für die Summe zweier diskreter Zufallsgrössen X und Y auf derselben Ergebnismenge Ω gilt $E(X + Y) = E(X) + E(Y)$

Bemerkung: Ein analoger Satz für die Varianz gilt nur, wenn X und Y unabhängige Zufallsvariablen sind, wennn also für alle x und y gilt: $P(X = x \text{und} Y = y) = P(X = x) \cdot P(Y = y)$

Satz 3.4

Es gilt $V(x) = E((X - \mu)^2)$

Beweis: Hat X für $\omega \in \Omega$ den Wert a_k, so hat $(X-\mu)^2$ den Wert $(a_k-\mu)^2$. Es ist also $P((X-\mu)^2=(a_k-\mu)^2)=P(X=a_k)$. Damit gilt

$$E((X-\mu)^2)=\sum(a_k-\mu)^2\cdot P((X-\mu)^2=(a_k-\mu)^2)=\sum(a_k-\mu)^2\cdot P((X=a_k)$$

□

3.2 Binomialverteilungen

3.2.1 Bernoulli-Ketten

Definition 3.9

Ein Bernoulli-Experiment ist ein Zufallsexperiment, dessen Ergebnisbeimenge nur aus zwei Elementen besteht. Häufig wird gesagt, dass es nur Erfolg (mit Wahrscheinlichkeit p) und Misserfolg (mit Wahrscheinlichkeit $q=1-p$) gibt.

Definition 3.10 (*Bernoulli-Kette*)

Ein Zufallsexperiment, das aus n getrennten, unabhängigen, gleichartigen Bernoulli-Experimenten besteht, heisst Bernoulli-Kette der Länge n.

Beispiel 3.6

Sören wirft fünf Mal hintereinander einen Schneeball auf eine Viehtränke. Seine Trefferwahrscheinlichkeit beträgt 45 Prozent.
Zu berechnen ist nun zum Beispiel die Wahrscheinlichkeit, dass Sören 2 Mal trifft. Wir zeichen einen Baum:
Na ja, das ist sehr aufwändig, machen wir doch nicht...
Ein Pfad, bei dem zwei Erfolge und drei Misserfolge vorkommen, hat die Wahrscheinlichkeit $0.45^2\cdot 0.55^3$. Es gibt $\left(\begin{array}{c}5\\2\end{array}\right)$ solcher Pfade (aus 5 Stufen 2 auswählen). Es gilt also $P(X=2)=\left(\begin{array}{c}5\\2\end{array}\right)\cdot 0.45^2\cdot 0.55^3$

3.2.2 Sätze zur Binomialverteilung

Satz 3.5 (*Binomialverteilung*)

Für die Verteilung der Zufallsgrösse X = Anzahl der Erfolge (mit Bezeichnung $B_{n,p}$) ist

$$P(B_{n,p}=k)=P(X=k)=\left(\begin{array}{c}n\\k\end{array}\right)\cdot p^k\cdot q^{n-k}$$

Satz 3.6 (*Kumulierte Binomialverteilung*)

$$P(X \leq k) = \sum_{i=0}^{k} \left(\begin{array}{c} n \\ i \end{array} \right) \cdot p^i \cdot q^{n-i}$$

Tipp zur schnellen Berechnung:
http://www.mathematik.ch/anwendungenmath/wkeit/binomialvert.php

Beispiel 3.7

Wie oben wird auf eine Viehtränke geworfen. Sörens Onkel wettet: „Sören wird höchstens zwei Mal treffen". Sörens Vater wettet „Sören wird mindestens drei Mal treffen". Zu berechnen sind die Wahrscheinlichkeiten, dass Onkel bzw Vater gewinnen. (Erinnerung $p = 0.45$)

$$P(x \leq 2) = \sum_{i=0}^{2} \left(\begin{array}{c} 5 \\ i \end{array} \right) \cdot p^i \cdot q^{5-i} = 0.59$$

$$P(x \geq 3) = \sum_{i=3}^{5} \left(\begin{array}{c} 5 \\ i \end{array} \right) \cdot p^i \cdot q^{5-i} = 1 - P(x \leq 2) = 0.41$$

Satz 3.7

$$P(B_{n,p} \leq k) = P(B_{n,q} \geq n - k)$$

denn: k Erfolge = $n - k$ Misserfolge.

Satz 3.8 (*Erwartungswert μ einer Binomialverteilung*)

$$\mu = n \cdot p$$

Satz 3.9 (*Standardabweichung σ einer Binomialverteilung*)

$$\sigma = \sqrt{n \cdot p \cdot q}$$

Beweise finden sich zum Beispiel in Kütting (2011).

Satz 3.10

Binomialansatz bei Stichprobenentnahmen *Ist das Verhältnis Umfang der Gesamtheit zu Umfang der Stichprobe sehr gross, dann ist für die Wahrscheinlichkeitsverteilung näherungsweise ein Binomialansatz zulässig.*

Beispiel 3.8

Bei der Herstellung von Kerzen ist nicht zu vermeiden, dass einige Kernzen kleinere Mängel aufweisen. Auf Grund langjähriger Erfahrung weiss der Hersteller, dass etwa 5% der Kerzen zur zweiten Wahl gehören.
Zur Qualitätssicherung werden täglich 10 Kerzen entnommen und geprüft.
Eines Tages werden unter den 10 Kerzen 2 Kerzen zweiter Wahl entdeckt. Wie wahrscheinlich ist dieses Ereignis?

3.3 Hypergeometrische Verteilung – Die Zahl der Erfolge

Beispiel 3.9

Wie gross ist die Wahrscheinlichkeit auf 4 Richtige beim Lotto 6 aus 49?

Beispiel 3.10

Ein Händler prüft einen Karton mit 100 Kerzen, unter denen sich 30 Kerzen zweiter Wahl befinden. Er betrachtet dazu 4 beliebige Kerzen. Mit welcher Wahrscheinlichkeit findet er keine zu beanstandende Kerze.

Bezeichnet N die Gesamtmenge, n die Stichprobe, k die Zahl der Erfolge, so ergibt sich die Formel:

$$P(X=k) = \frac{\left(\begin{array}{c} K \\ k \end{array}\right) \cdot \left(\begin{array}{c} N-K \\ n-k \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}$$

3.4 Geometrische Verteilung – Warten auf den ersten Erfolg

Beispiel 3.11

Wie oft muss im Beispiel 3.8 eine Kerze gezogen werden, bis erstmals eine defekte Kerze erscheint?

$$P(X=n) = p \cdot (1-p)^{n-1}$$

Hier tritt erstmals ein unendlicher Wahrscheinlichkeitsraum auf.

Es gilt

$$\lim_{n\to\infty} \sum_{k=1}^{n} P(X=k) = 1$$

Der Beweis erfolgt mit einer geometrischen Reihe.

3.5 Testen mit der Binomialverteilung

3.5.1 Einführendes Beispiel

Beispiel 3.12

Die Schülerinnen und Schüler einer Schule werden jeweils nach der jährlichen Spezialwoche befragt, ob sie weiterhin eine solche wünschen. Um Papier und Zeit zu sparen, sollen nur noch 50 Schülerinnen und Schüler befragt werden.

a) Das Rektorat stellt die Hypothese auf, dass 60 oder mehr Prozent aller Schülerinnen und Schü- ler die Spezialwoche befürworten. Es verwirft die Hypothese, falls weniger als 25 der befragten Schülerinnen und Schüler die Spezialwoche befürworten. Ist das sinnvoll?

b) Bei welcher Anzahl Ja-Antworten lässt sich mit einem Risiko von weniger als 5 Prozent annehnmen, dass mehr als die Hälfte der Schülerinnen und Schüler der Kantonsschule weiterhin Spezialwochen wünschen?

Die Ideen aus dieser Problemstellung werden nun formalisiert.

3.5.2 Tests

Allgemeines Vorgehen bei einem Test:

- Formulierung einer Nullhypothese H_0 (und einer Alternativhypothese H_1) – *im Beispiel oben:* $p = 0.6$
- Wahl des geeigneten Tests, der geeigneten Verteilung *im Beispiel oben: Binomialverteilung*
- Wahl des kritischen Bereichs, in dem die Nullhypothese abgelehnt wird. *im Beispiel oben: erst bei Teil b durchgeführt, weniger als 24 Ja-Stimmen.*
  Zum Bereich gehört ein Signifikanzniveau α. Oft wird α gleich 5 Prozent *wie oben* oder gleich 1 Prozent gewählt.
- Bestimmung des Werts des Tests aus der Stichprobe. *wurde oben nicht durchgeführt.*
- Entscheidung
- Liegt der Wert nicht im kritischen Bereich, wo wird H_0 beibehalten, andernfalls wird H_0 verworfen.

Problem bei Signifikanztests ohne Alternativhypothese: wenn α genügend klein gewählt wird, kann jede Nullhypothese „gerettet« werden. Die Alternativhypothese ist aber oft, zum Beispiel oben, nicht sinnvoll bestimmbar. Sie muss recht willkürlich gewählt werden.

Beispiel 3.13

Eine Population soll auf eine bestimmte Eigenschaft untersucht werden. Die Nullhypothese ist, dass 60 Prozent dieser Population die Eigenschaft haben. Die Alternativhypothese ist, dass 50 Prozent diese Eigenschaft haben. Es werden 50 Individuen untersucht, getestet wird mit der Binomialverteilung, Signifikanzniveau ist $\alpha = 5$ Prozent.
Nach den Berechnungen aus dem vorigen Beispiel wird die Nullhypothese verworfen, wenn in der Stichprobe weniger als 24 Individuen die Eigenschaft haben.

Schön gezeigt wird dies zum Beispiel auf
http://mathematik.ch/anwendungenmath/wkeit/hypotest/hypotest.php

Zusammenfassung:

	H_0 trifft zu	H_0 trifft nicht zu
H_0 wird abgelehnt	Fehler 1. Art	richtige Entscheidung
H_0 wird beibehalten	richtige Entscheidung	Fehler 2.Art

Dies soll an einem Beispiel noch weiter verdeutlicht werden. Ein Rauchmelder soll so eingestellt werden, dass er bei normalen Kochvorgängen in einer Küche nicht anschlägt, wohl aber bei einem Brand.

Hier zeigt sich, dass die Wahl des Bereichs von H_0 sehr wichtig sein kann:

	H_0 trifft zu Köchin ist beim Anbraten	trifft nicht zu *Es brennt*
H_0 wird vom Rauchmelder abgelehnt Alarm	Fehler 1. Art Feuerwehr kommt umsonst	richtige Entscheidung Feuerwehr löscht Brand
H_0 wird beibehalten kein Alarm	richtige Entscheidung Alles in Ordnung	Fehler 2.Art *Restaurant brennt ab*

3.5.3 Links- und rechtsseitige Tests

In unserem Beispiel war die Alternativhypothese p_1 kleiner als p_0. Der Ablehnungsbereich befindet sich auf der linken Seite der Verteilung – Dies heisst linksseitiger Signifikanztest. Analog definiert ist der rechtsseitig Signifikanztest. Bei diesem wird die Nullhypothese abgelehnt. wenn zu viele Ereignisse eintreten.

Bei einem zweiseitigen Test wird die Nullhypothese abgelehnt, wenn die Abweichung nach oben oder unten zu gross ist.

Beispiel 3.14

Getestet wird, ob bei einem Würfel die 6 nicht zu oft und nicht zu selten fällt: 50 Versuche, H_0 lautet $p = \frac{1}{6}$, $\alpha = 0.05$.
Der Verwerfungsbereich ist, wenn weniger als 4 oder mehr als 14 Sechsen fallen[1] (der Annahmebereich ist breit, wie jeder bestätigen wird, der in einem Spiel mal auf die Sechser geachtet hat...)

4 Allgemeine Wahrscheinlichkeitsräume

4.1 Borelmengen

4.1.1 Abzählbar unendliche Wahrscheinlichkeitsräume

Beispiel 4.1

Wie gross ist die Wahrscheinlichkeit, dass bei einem idealen Würfel die Augenzahl 6 zum ersten Mal bei einer geraden Anzahl von Würfen fällt?

Definition 4.1

Abzählbar unendlicher Wahrscheinlichkeitsraum – Kolmogorov-Axiome *Sei Ω eine abzählbar unendliche Ergebnismenge. Die Funktion $P : \wp(\Omega) \to \mathbb{R}$ heisst Wahrscheinlichkeitsmass über dem Ergebnisraum Ω, wenn sie die folgenden Eigenschaften hat:*

1. **Nichtnegativität** $P(A) \geq 0$ *für alle* $A \in \wp(\Omega)$
2. **Normierung** $P(\Omega) = 1$
3. **Additivität** *Für ein System von Mengen* $A_i \in \wp(\Omega), (i \in \mathbb{N})$ *mit* $A_i \cap A_j) = \emptyset$ *für* $i \neq j$ *gilt*

$$P\left(\bigcup_{i=i}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

4.1.2 Überabzählbar-unendliche Wahrscheinlichkeitsräume

Beispiel 4.2

Eine S-Bahn fahre immer genau alle 15 Minuten ab. Wie gross ist die Wahrscheinlichkeit, dass die Wartezeit kleiner oder gleich x ist?
Die einleuchtende Antwort lautet:

$$P(]-\infty, x]) = \begin{cases} 0, \text{ falls } x < 0 \\ \frac{1}{15}x, \text{ falls } x \in [0, 15] \\ 1, \text{ falls } x > 15 \end{cases}$$

An dieser Stelle ergibt sich ein mathematisches Problem. Die Wahrscheinlichkeit, dass die Wahrscheinlichkeit ganz genau x beträgt, ist 0. Die Wahrscheinlichkeit für Elementarereignisse wäre Null. Gefordert ist also

$$[\text{N}] \qquad P(\{x\}) = 0$$

Dies ist nicht verträglich mit unserer bisherigen Definition für Wahrscheinlichkeitsräume. 1929 haben Banach und Kuratowski gezeigt, dass es kein Wahrscheinlichkeitsmass, das auf der ganzen Potenzmenge von $\mathbb{R}$ definiert ist, gibt, das die drei Kolmogorov-Bedingungen aus Definition 4.1 und [N] erfüllt.

Der Ausweg ist, nur Teilmengen von $\wp(\mathbb{R})$ für die Berechnung zuzulassen. Dieses System sind die sogenannten Borel-Mengen. Dieses Mengensystem wird nun konstruiert.

Sei $\mathcal{I}$ die Menge aller (nach links) halboffene Intervalle in $\mathbb{R}$, also $\mathcal{I} = \{]a, b] | a, b \in \mathbb{R}, a < b\}$.

Dann betrachten wir eine Teilmenge $\mathcal{A}$ der Potenzmenge $\wp(\mathbb{R})$, welche die folgenden Eigenschaften erfüllt.

[B0] Die Menge $\mathcal{I}$ ist Teilmenge von $\mathcal{A}$.

[B1] $\mathbb{R} \in \mathcal{A}$.

[B2] $A \in \mathcal{A} \Rightarrow \bar{A} \in \mathcal{A}$.

[B3] Sind für alle n in $\mathbb{N}$ die Mengen $A_n \in \mathcal{A} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$

Satz 4.1

Unter allen Mengensystemen, welche die Eigenschaften B0, B1, B,2 und B3 erfüllen, gibt es ein kleinstes Mengensystem, nämlich die Schnittmenge aller Mengensysteme, welche diese Eigenschaften erfüllen.

Für den Beweis wird auf [2], Kapitel VII.2.2 verwiesen.

Definition 4.2 (*Borelmengen*)

Das nach obigem Existenzsatz existierende Mengensystem heisst System der Borelmengen auf $\mathbb{R}$ und wird mit $\mathcal{B}(\mathcal{I})$ bezeichnet.
Das Paar $(\mathbb{R}, \mathcal{B}(\mathcal{I}))$ wird als Messraum der reellen Zahlen bezeichnet.

Das System der Borelmengen umfasst alle möglichen halboffenen, offenen und abgeschlossenen Intervalle, die Gesamtmenge $\mathbb{R}$ und auch jede einzelne reelle Zahl x. Alle uns interessierenden Mengen sind also Borel-Mengen. Auch hierzu sei auf [2] verwiesen.

Definition 4.3 (*Wahrscheinlichkeitsmass,Wahrscheinlichkeitsraum*)

Die Funktion $P : \mathcal{B}(\mathcal{I}) \to \mathbb{R}$ heisst Wahrscheinlichkeitsmass auf $\mathbb{R}$, wenn sie die folgenden Eigenschaften hat:

1. **Nichtnegativität** $P(A) \geq 0$ *für alle* $A \in \mathcal{B}(\mathcal{I})$
2. **Normierung** $P(\mathbb{R}) = 1$
3. **Additivität** *Für ein System von Mengen* $A_i \in \mathcal{B}(\mathcal{I}), (i \in \mathbb{N})$ *mit* $A_i \cap A_j) = \emptyset$ *für* $i \neq j$ *gilt*

$$P\left(\bigcup_{i=i}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Für ein Wahrscheinlichkeitsmass P auf $\mathbb{R}$ heisst $(\mathbb{R}, \mathcal{B}(\mathcal{I}), P)$ Wahrscheinlichkeitsraum zu $\mathbb{R}$.

4.1.3 Dichtefunktionen

Ein Rechteck mit Länge a und Breite b hat die Fläche ab. Wird nun die Breite $b = x$ variabel gewählt, so ist für jedes Intervall $[x - \epsilon, x + \epsilon]$ die Teilfläche des Rechtecks $2\epsilon \cdot b$. Ist $\epsilon = 0$, so ist auch die Fläche Null.

Handelt es sich nicht um ein Rechteck, sondern ist auch die Länge variabel und durch eine Funktion f gegeben so lässt sich die Fläche mit einem Integral berechnen:

$$\text{Fläche im Intervall } x_0, x_1 = \int_{x_0}^{x_1} f(t)\, dt$$

Das hat Bezug zum Beispiel 4.2.

Betrachtet wird die «Rechteckfunktion»

$$f : \mathbb{R} \to \mathbb{R}, t \mapsto f(t) = \begin{cases} 0, \text{ falls } t < 0 \\ \frac{1}{15}, \text{ falls } t \in [0, 15] \\ 0, \text{ falls } t > 15 \end{cases}$$

Mit dieser Funktion gilt

$$P(]-\infty, x]) = \int_{-\infty}^{x} f(t)\, dt$$

Definition 4.4

Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ heisst Dichtefunktion (oder kurz Dichte) falls gilt
[D1] f ist integrierbar,
[D2] $f(t) \geq 0$ für alle $t \in \mathbb{R}$
[D3] $\int_{-\infty}^{\infty} f(t)\, dt = 1$.

Aus Dichtefunktionen ergeben sich dann Wahrscheinlichkeiten, indem integriert wird. Zunächst folgt aber ein Beispiel.

Beispiel 4.3

Sei λ eine positive reelle Zahl. Die Funktion $f : \mathbb{R} \to \mathbb{R}$ wird wie folgt definiert:

$$f(t) = \begin{cases} 0, \text{ falls } t < 0 \\ \lambda \cdot e^{-\lambda t}, \text{ falls } t \geq 0 \end{cases}$$

Dies ist eine Dichtefunktion.

Zu jeder Dichtefunktion ist $F(x) = \int_{-\infty}^{x} f(t)\,dt$ eine Verteilungsfunktion im Sinne der folgenden Definition:

Definition 4.5

Ist P ein Wahrscheinlichkeitsmass auf $(\mathbb{R}, \mathcal{BI})$, so heisst die Funktion

$$F : \mathbb{R} \to \mathbb{R}, x \mapsto F(x) := P([-\infty, x])$$

Verteilungsfunktion bezüglich P.

Satz 4.2 (*Fundamentalsatz zu Verteilungsfunktionen*)

Sei $F : \mathbb{R} \to \mathbb{R}$ eine Funktion, welche die drei folgenden Eigenschaften besitzt:
[V1] F ist monoton wachsend
[V2] F ist rechtsseitig stetig.
[V3] $\lim_{x \to -\infty} F(x) = 0$ *und* $\lim_{x \to \infty} F(x) = 1$
Dann existiert ein Wahrscheinlichkeitsmass P auf $(\mathbb{R}, \mathcal{B}(\mathcal{I}))$, so dass F die Verteilungsfunktion zu P ist.

Der Satz geht vertieft in die Theorie zu den Borelmengen ein und kann hier nicht bewiesen werden.

Das typische Vorgehen zur Konstruktion von Wahrscheinlichkeitsmassen ist also:

Es wird eine Funktion f gefunden, die eine Dichtefunktion ist. Integration liefert dann die Verteilungsfunktion $F(x)$ und diese das Wahrscheinlichkeitsmass per Fundamentalsatz. Es gilt also

$$\int_{-\infty}^{x} f(t)\,dt = F(x) = P([-\infty, x])$$

4.2 Verteilungsfunktionen zu vorgegebenen Dichtefunktionen

Vorgehen:

1. Vorgabe einer Dichtefunktion f

2. Definition einer Stammfunktion

$$F(x) = \int_{\infty}^{x} f(t)\,dt$$

(Nachprüfung der Eigenschaften aus dem Fundamentalsatz)

3. Verwendung der Stammfunktion als Verteilungsfunktion

Beispiel 4.4

Rechteckverteilung

$$f(t) = \begin{cases} 0, \text{ falls } t < a \\ \dfrac{1}{b-a}, \text{ falls } a \leq t \leq b \\ 0, \text{ falls } t > b \end{cases}$$

Beispiel 4.5

Exponentialverteilung

$$f(t) = \begin{cases} 0, \text{ falls } t < 0 \\ \lambda \cdot e^{-\lambda t}, \text{ falls } t \geq 0 \end{cases}$$

4.3 Normalverteilung

Teile dieses Kapitels beruhen auf einem Skript von Christoph Drollinger, dem ich an dieser Stelle dafür danken möchte.

4.3.1 Einführung: Transformationen der Binomialverteilung

Ein idealer Würfel wird 5000-mal geworfen. Wie wahrscheinlich ist es, dass 500-mal die 6 auftritt? Die Lösung ist $\binom{500}{50} \cdot \left(\frac{5}{6}\right)^{4500} \cdot \left(\frac{1}{6}\right)^{500} = 6.2 \cdot 10^{-42}$ Für grosse n ist bei einer Binomialverteilung die Wahrscheinlichkeit $P(X = k)$ sehr klein. Der genaue Wert ist daher uninteressant. Wichtiger ist die Wahrscheinlichkeitsfunktion $F : x \mapsto P(X \leq x)$. Für $P(X \leq 500)$ müsste 501-mal die Formel der Binomialverteilung anwenden. Dies ist sehr rechenaufwändig.

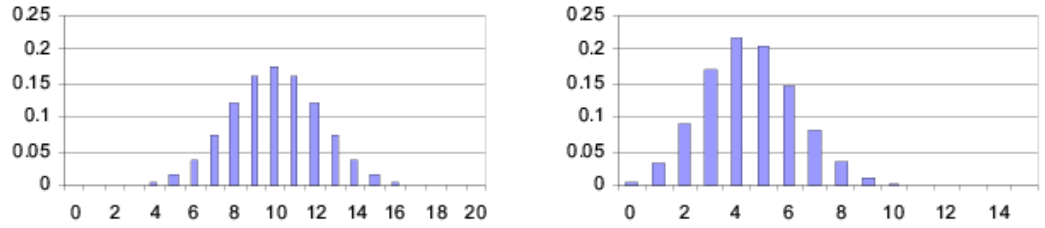
Andererseits sieht die graphische Darstellung der Wahrscheinlichkeitsverteilung fast wie eine stetige Kurve aus – tatsächlich existiert eine Dichtefunktion, mit der die Binomialverteilung angenähert werden kann – dazu mehr im nächsten Abschnitt.

In diesem Abschnitt soll einführend gezeigt werden, wie diese Dichtefunktion aussehen muss – und welche Umformungen gemacht werden müssen, um die Binomialverteilung damit annähern zu können. Es werden beispielhaft zwei Binomialverteilungen verglichen.

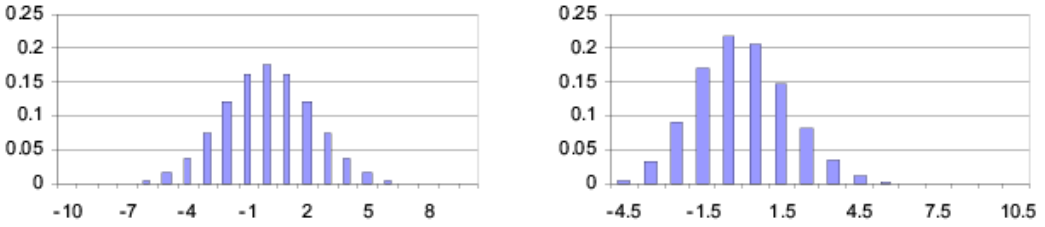
Sei X_1 eine binomialverteilte Zufallsvariable mit $n = 20$ und $p = 0.5$. Es gilt $E(X) = n \cdot p = 10$ und $\sigma = \sqrt{npq} = \sqrt{10 \cdot 0.5 \cdot 0.5} = \sqrt{5} \cong 2.24$

Sei X_2 eine binomialverteilte Zufallsvariable mit $n = 15$ und $p = 0.3$. Also $E(X) = 4.5$ und $\sigma = \sqrt{3.15} \cong 1.77$

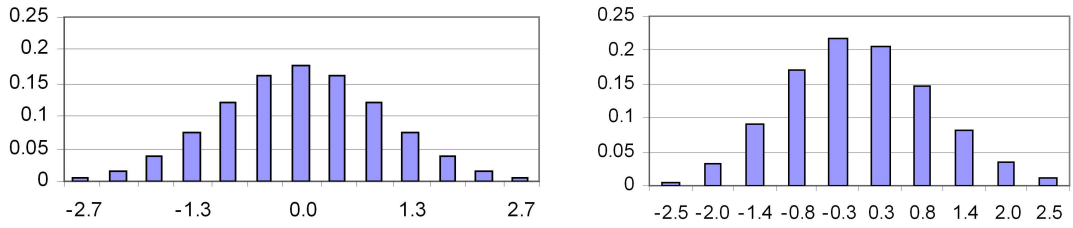
Erstellt wir je eine Wahrscheinlichkeitsverteilung, so ergeben sich folgende Histogramme:



Dabei sind auf der x-Achse die Werte für k und auf der y-Achse die zugehörigen Wahrscheinlichkeiten abgetragen. Eine Näherungslösung ergibt sich nun in mehreren Schritten. Zuerst verschieben wir die Histogramme um den Erwartungswert nach links. Für führen also die Transformation $x = k - \mu$ durch. Wir erhalten je eine Wahrscheinlichkeitsverteilung mit $\mu = 0$

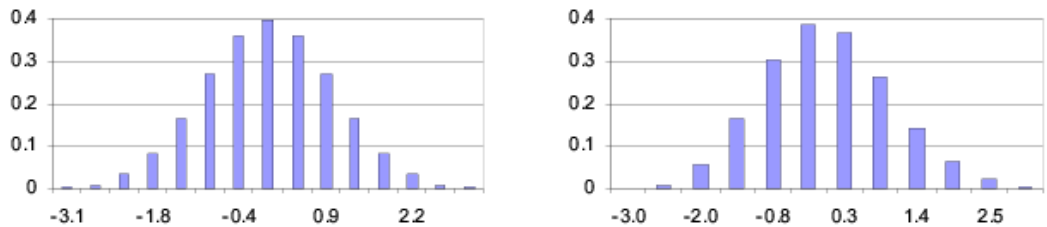


Der Erwartungswert der Wahrscheinlichkeitsverteilungen ist je 0. Die Varianzen sind aber noch unterschiedlich. Teilen wir die Werte für x je durch die Standardabweichung, so erhalten wir Wahrscheinlichkeitsverteilungen mit je $\sigma = 1$.

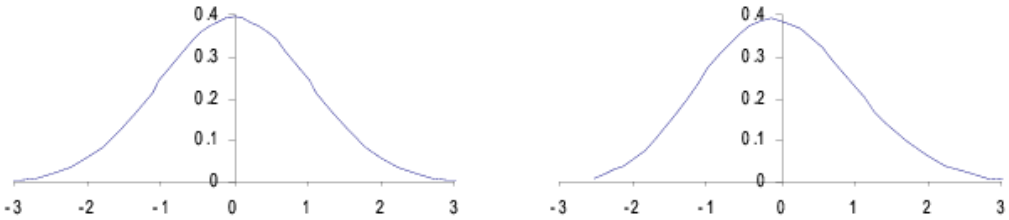


Da n für X_1 und X_2 unterschiedlich ist, sind die Säulen im linken und im rechten Histogramm nicht gleich hoch. Um dies anzupassen, werden die y-Werte mit der Standardabweichung σ multipliziert. Wir verwenden somit folgende Transformationen:

$$x = \frac{k - \mu}{\sigma} \text{ und } y = \sigma \cdot B_{n;p}(k)$$



Zeichnen wir nicht mehr Säulen, sondern verbinden wir die Punkte oben auf der Säule miteinander, so folgen zwei fast identische Diagramme:



Im nächsten Abschnitt wird eine Dichtefunktion betrachtet, die ein ähnliches Aussehen hat. Später wird, ohne Beweis, ein Approximationssatz der Binomialverteilung durch die entsprechende Verteilungsfunktion formuliert und die Konsequenzen analysiert.

4.3.2 Die Normalverteilung

Zunächst führen wir eine neue Schreibweise für die Exponentialfunktion e^x ein, die praktisch für komplizierte Exponenten ist: $e^x = \exp(x)$

Gausssche Glockenkurve und Normalverteilung

Satz 4.3

Seien μ und σ reelle Zahlen mit $\sigma > 0$. Die Funktion

$$f(t) := \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)$$

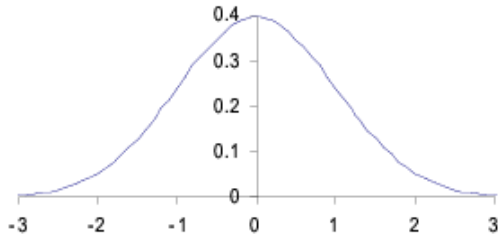
ist eine Dichtefunktion.

Die Funktion ist klar positiv, da die Exponentialfunktion immer positiv ist. Sie ist integrierbar, da sie beschränkt ist (Ein Maximum wir bei der Behandlung der geometrischen Eigenschaften berechnet).

Die zentrale Eigenschaft, dass das Integral von $-\infty$ bis ∞ gleich 1 ist, können wir an dieser Stelle nicht beweisen.

Die folgende Graphik zeigt den Graphen für $\sigma = 1$ und $\mu = 0$.

Der Graph dieser Funktion heisst Gausssche Glockenkurve.



Definition 4.6

Die Verteilungsfunktion

$$F(x) := \int_{-\infty}^{x} f(t)\,dt$$

zur Dichtefunktion $f(t) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)$ *aus Satz 4.3 heisst Normalverteilung mit den Parametern* μ *und* σ^2.

Wenn eine Funktion F *eine Normalverteilung mit den Parametern* μ *und* σ^2 *st, sagen wir kurz:* F *ist* $N(\mu, \sigma^2)$*-verteilt.*

Satz 4.4

Die Wahrscheinlichkeitsverteilung zur Verteilungsfunktion $F(x)$ *aus 4.6 hat den Erwartungswert* μ *und die Standardabweichung* σ.

Der Beweis ist eine aufwändige Übung in Differentialrechnung. Er wird hier nicht durchgeführt, ist aber zum Beispiel in Kütting (2011) auf Seite 278f nachzulesen.

Satz 4.5 (*Eigenschaften der Dichtefunktion zur Normalverteilung*)

Die Dichtefunktion $f(t) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)$ *hat die folgenden Eigenschaften:*

1. *Sie ist symmetrisch zur Achse* $t = \mu$.
2. *Sie nimmmt an der Stellle* $t = \mu$ *ihr Maximum an, dort gilt* $f(u) = \frac{1}{\sigma\sqrt{2\pi}}$.
3. *Sie hat in* $t_1 = \mu - \sigma$ *und* $t_2 = \mu + \sigma$ *Wendestellen*

Folgerungen:

Je grösser μ ist, um so mehr ist der Graph der Glockenkurve nach rechts verschoben.

Je grösser σ ist, um so kleiner ist das Maximum der Glockenkurve und um so schwächer fällt die Kurve nach beiden Seiten ab.

Berechnung von Wahrscheinlichkeiten mit der Standard-Normalverteilung

Wie bereits im einleitenden Abschnitt 4.3.1 angedeutet wurde, lassen sich viele Verteilungen auf die Normalverteilung zurückführen – und zwar auf die Standardnormalverteilung mit $\mu = 0$ und $\sigma = 1$.

In diesem Abschnitt wird die Rückführung aller Normalverteilungen auf die Standardnormalverteilung gezeigt.

Definition 4.7

Die Verteilungsfunktion Φ zu $\varphi(t) := \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-t^2}{2}\right)$ heisst Standard-Normalverteilung. Es gilt also

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\, dt.$$

Für dieses Integral gibt es keinen geschlossenen Ausdruck. Es ist aber so wichtig, dass ausführliche Tabellen vorliegen, zum Beispiel in Kütting (2011) oder auf
http://de.wikipedia.org/wiki/Tabelle_Standardnormalverteilung

Folgendermassen werden Wahrscheinlichkeiten für beliebige Normalverteilungen aus der Standard-Normalverteilung berechnet:

Satz 4.6 (*Transformation zur Standard-Normalverteilung*)

Ist F eine $N(\mu, \sigma^2)$-verteilte Funktion, so gilt:
$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.

Gleichbedeutend ist:

1. $P_{\mu,\sigma^2}(]-\infty, x]) = P_{0,1}(]-\infty, z)$ *mit* $z = \frac{x-\mu}{\sigma}$.
2. $$\int_{-\infty}^{x} \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) dt$$
$$= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-u^2}{2}\right) du \text{ mit } z = \frac{x-\mu}{\sigma}.$$
3. *Muss die Wahrscheinlichkeit von $-\infty$ bis x einer normalverteilten Zufallsgrösse berechnet werden, so kann stattdessen die Standard-Normalverteilung von $-\infty$ bis $\frac{x-\mu}{\sigma}$ verwendet werden.*

Beispiele zur Berechnung finden sich beispielsweise in Kütting (2011) auf den Seiten 266 und 269. Hier werden Beispiele innerhalb des Abschnitts 4.3.3 vertieft behandelt.

Sigma-Regeln

Die Normalverteilung lässt sich grob abschätzen durch die folgenden Regeln:

- Rund 68% der Beobachtungswerte liegen im Intervall von $\mu-\sigma$ bis $\mu+\sigma$ (Sigma-Umgebung).
- Die 2-Sigma-Umgebung enthält rund 95% der Beobachtungswerte (Intervall $[\mu-2\sigma, \mu+2\sigma]$).
- Die 3-Sigma-Umgebung enthält rund 99% der Beobachtungswerte.

4.3.3 Binomialverteilung und Normalverteilung

Satz 4.7 (*Näherungsformeln von De Moivre-Laplace*)

Sei Φ die Verteilungsfunktion aus Definition 4.7.

Für eine $B_{n;p}$-verteilte Zufallsvariable X mit dem Erwartungswert $E(X) = np$ und der Standardabweichung $\sigma = \sqrt{npq}$ gelten bei grossen Werten von n:

$$P(X \leq k) \cong \Phi(x) \text{ mit } x = \frac{k + 0.5 - \mu}{\sigma}$$

$$P(k_1 \leq X \leq k_2) \cong \Phi(x_2) - \Phi(x_1)$$

$$\text{mit } x_1 = \frac{k_1 - 0.5 - \mu}{\sigma} \text{ und } x_2 = \frac{k_2 + 0.5 - \mu}{\sigma}$$

Die Formeln liefern brauchbare Werte für $\sigma \geq 3$, d.h für $npq \geq 9$.

Ohne Beweis.

Bemerkung: Die 0.5 tritt auf, da der k-te Balken bei der Binomialverteilung von $k-0.5$ bis $k+0.5$ geht. Bei der Berechnung von x_1 und x_2 wird oft die Zahl 0.5 oder -0.5 weggelassen. Für grosse n spielt spielt dies keine Rolle.

Beispiel 4.6

Wir betrachten die Binomialverteilung $B_{64;0.5}$ und interessieren uns für die Wahrscheinlichkeit $P(29 \leq X \leq 36)$

http://www.mathematik.ch/anwendungenmath/wkeit/approx_bin_norm.php

Beispiel 4.7

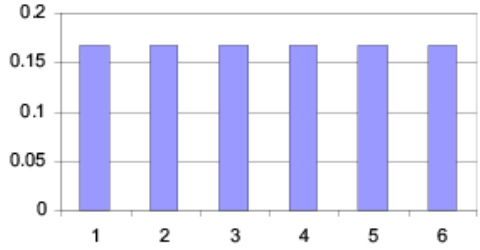
Ein Händler bietet Gurkensamen an, die erfahrungsgemäss zu 95% keimfähig sind. Mit welcher Wahrscheinlichkeit keimen von 500 ausgesäten Körnern höchstens 472?

4.3.4 Der zentrale Grenzwertsatz

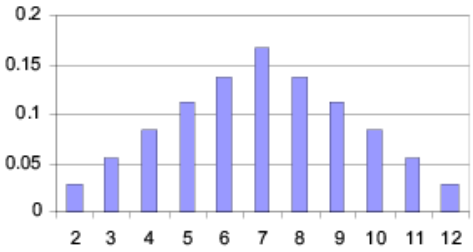
In diesem Abschnitt wird anschaulich angedeutet, dass die Normalverteilung immer anwendbar ist, wenn ein Experiment oft wiederholt wird.

Werfen wir einen Würfel. Die Zufallsvariable X sei die Augenzahl. Dann können wir die Wahrscheinlichkeitsverteilung von X einfach tabellieren:

k	1	2	3	4	5	6
$P(X = k)$	0.167	0.167	0.167	0.167	0.167	0.167

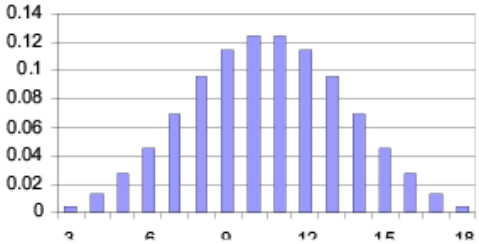


Werfen wir zwei Würfel. Die Zufallsvariable X_i sei die Augensumme beim i-ten Würfel. Es sei $X = X_1 + X_2$ die Augensumme der zwei Würfel. Die Augensumme 2 kann nur in einem Fall auftreten: Beide Würfel haben die Augensumme 1. Daher ist $P(X = 2) = 1/36$. Die Augensumme 3 kann in zwei Fällen auftreten: Der eine Würfel hat eine 1 und der andere Würfel eine 2. Daher ist $P(X = 3) = 2/36$.

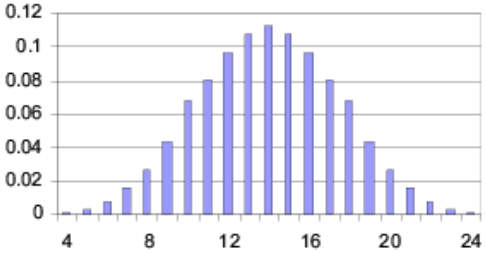


k	2	3	4	5	6	7	8	9	10	11	12
$P(X = k)$	0.028	0.056	0.083	0.111	0.139	0.167	0.139	0.111	0.083	0.056	0.028

Für drei Würfel folgt: Die Zufallsvariable X_i sei die Augensumme beim i-ten Würfel. Es sei $X = X_1 + X_2 + X_3$ die Augensumme der drei Würfel.



Für vier Würfel folgt:



Es scheint sich bei der Augensumme von Würfeln wieder eine Glockenkurve anzudeuten, wie sie uns bei Binomialverteilungen mit grossen n begegnet ist. Sie zeigt sich mit zunehmender Deutlichkeit, je mehr Würfel man verwendet. Obwohl die Zufallsvariablen X_i in diesem Fall keine Bernoulli-Variablen sind, scheint sich die Verteilung von $X = X_1 + X_2 + \ldots + X_n$ wieder näherungsweise mit Hilfe der Funktion Φ beschreiben zu lassen. Tatsächlich gilt der folgende Satz:

Satz 4.8 (*Zentraler Grenzwertsatz*)

Sind $X_1, X_2, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen, so gilt für die Zufallsvariable

$X = X_1 + X_2 + \ldots + X_n$ mit $E(X) = \mu$ und $V(X) = \sigma^2$ bei hinreichend grossen Werten von n:

$$P(X \leq x) \cong \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Die Näherung ist umso besser, je grösser die Anzahl n ist.

Mehr Hintergrund findet sich unter http://de.wikipedia.org/wiki/Zentraler_Grenzwertsatz

Literaturverzeichnis

[1] Jahnke, T. (Hrsg) et al (2005): Stochastik. Cornelsen: Berlin (Es gibt mehrere Bücher gleichen Titels: ISBN 978-3-464-57218-4)

[2] Kütting, H. und Sauer, M. (2011): Elementare Stochastik. Spektrum: Heidelberg.