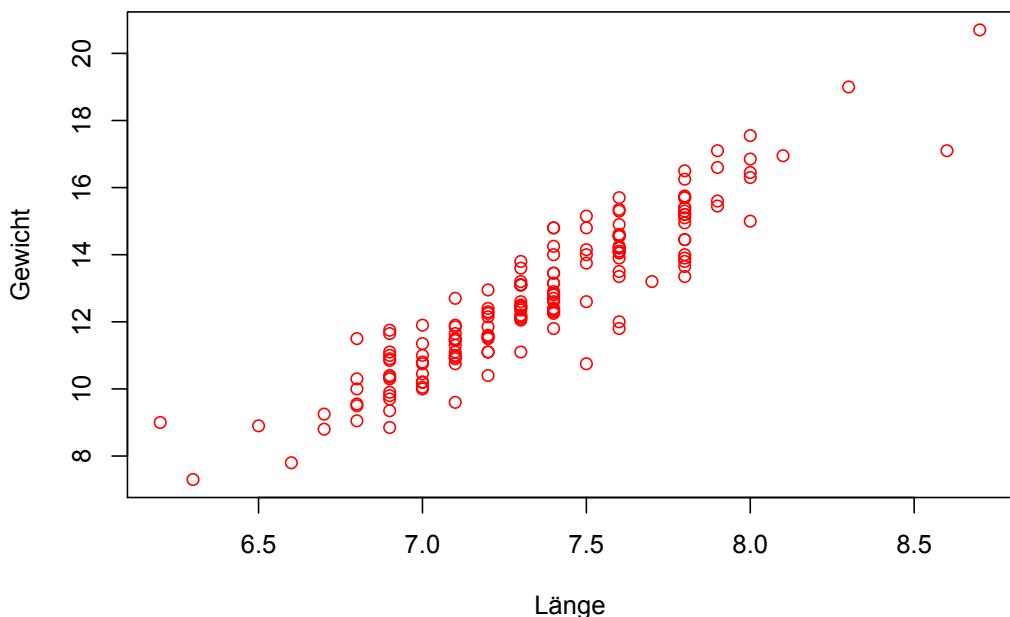


Deskriptive Statistik und Wahrscheinlichkeitsrechnung



Scatterplot der Länge und des Gewichts von Krabben

Diese Seite ist leer.

Inhaltsverzeichnis

1 Grundbegriffe	1
1.1 Zufallsexperimente	1
1.2 Stichprobenraum und Ereignisse	3
1.3 Wahrscheinlichkeit	5
1.4 Bedingte Wahrscheinlichkeit	7
1.5 Unabhängigkeit	8
1.6 Formel der totalen Wahrscheinlichkeit	10
1.7 Mehrstufige Zufallsexperimente	11
1.8 Formel von Bayes	13
1.9 Diagnostische Tests: Sensitivität und Spezifität	14
2 Zufallsvariablen	18
2.1 Begriff der Zufallsvariablen	18
2.2 Die Verteilung einer Zufallsvariable	18
2.3 Unabhängigkeit von Zufallsvariablen	20
2.4 Erwartungswert und Varianz einer diskreten Zufallsvariable	20
2.4.1 Erwartungswert	20
2.4.2 Varianz und Standardabweichung	23
2.5 Die Kovarianz*	25
2.6 Binomialverteilung	27
2.7 Stetige Zufallsvariablen	30
2.8 Normalverteilung	33
3 Deskriptive Statistik	36
3.1 Daten	36
3.2 Kategoriale Merkmale	36
3.3 Quantitative Merkmale	39
3.4 Formen der Histogramme	42
3.5 Multivariate Daten	43
4 Statistische Masszahlen	46
4.1 Lagemasse	46
4.1.1 Mittelwert	46
4.1.2 Median	47
4.1.3 Modus	49
4.1.4 Quantile	49
4.2 Masszahlen für die Streuung	50
4.2.1 Varianz und Standardabweichung	50
4.2.2 Variationskoeffizient	52
4.2.3 Interquartilsabstand und Spannweite	52
4.3 Boxplots	52
4.4 Schiefe	54
4.5 Kurtosis*	56
4.6 Korrelation	57

5 Normalverteilte Daten	59
5.1 Einleitung	59
5.2 QQ-Plot	61
5.3 Der zentrale Grenzwertsatz	65
5.4 Transformation von Daten	68
Stichwortverzeichnis	70

1 Grundbegriffe

1.1 Zufallsxperimente

In der Wahrscheinlichkeitsrechnung studiert man die Gesetzmässigkeiten von **Zufalls-experiment**. Dies sind Experimente deren Resultat nicht durch logische Gründe oder durch die Versuchsbedingungen determiniert sind. Die Experimente müssen unter den gleichen Bedingungen wiederholbar sein, und zwar so, dass das Ergebnis nicht notwendig stets das gleiche ist, sondern nur statistischen Regelmässigkeiten folgt.

Beispiel 1 Wir können eine Geburt in Bezug auf das Geschlecht des Neugeborenen als Zufallsexperiment ansehen.

Die nachfolgende Tabelle (Quelle: Bundesamt für Statistik) enthält die Anzahl Lebendgeburten in der Schweiz nach Geschlecht in den Jahren 2001-2018:

Jahr	Total	Knaben	Mädchen
2001	72'295	37'123	35'172
2002	72'372	37'318	35'054
2003	71'848	36'902	34'946
2004	73'082	37'340	35'742
2005	72'903	37'569	35'334
2006	73'371	37'766	35'605
2007	74'494	38'184	36'310
2008	76'691	39'549	37'142
2009	78'286	40'407	37'879
2010	80'290	41'111	39'179
2011	80'808	41'626	39'182
2012	82'164	42'435	39'729
2013	82'731	42'595	40'136
2014	85'287	43'850	41'437
2015	86'559	44'649	41'910
2016	87'883	44'932	42'951
2017	87'381	44'873	42'508
2018	87'851	45'013	42'838

Aufgrund dieser Tabelle können wir **empirische Wahrscheinlichkeiten** definieren. Die Wahrscheinlichkeit p_1 für eine Mädchengeburt im Jahre 2018 ist gegeben durch:

$$p_1 = \frac{42'838}{87'851} \doteq 48.8\%$$

Die Wahrscheinlichkeit p_2 für eine Knabengeburt in den Jahren 2001-2018 ist gegeben durch:

$$p_2 = \frac{733'242}{1'426'296} \doteq 51.4\%$$

In diesem Fall sind die Wahrscheinlichkeiten **relative Häufigkeiten**. ◇

Beispiel 2 Würfelspiel: die Flächen sind von 1 bis 6 nummeriert. Es ist nicht möglich, das Ergebnis eines Wurfs vorauszusagen; es handelt sich also um ein Zufallsexperiment.

Falls wir alle relevanten Faktoren kennen würden, wie etwa die Ausgangsposition der Hand, die Position und Bewegung der Hand in Bezug zum Tisch, die Elastizität des

Würfels und des Tisches etc., könnten wir nach den Gesetzen der Physik das Resultat theoretisch vorausberechnen. In der Praxis sind diese Vorgänge aber derart komplex, dass eine Berechnung völlig unmöglich ist.

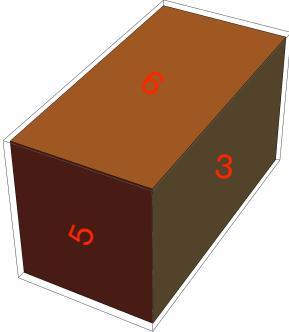
Mit *Mathematica* wurden zwei Serien zu 6'000 Würfen eines perfekten Würfels simuliert. mal der Wurf eines Würfels simuliert. Nachfolgend sind die relativen Häufigkeiten k/n für die ersten n Würfe abgebildet:



Da es sich um einen perfekten, homogenen Würfel handelt, wird man aus Symmetriegründen erwarten, dass die Wahrscheinlichkeit p für das Resultat „6“ gleich $\frac{1}{6} \doteq 0.1\bar{6}$ ist. Man wird erwarten, dass wenn wir n immer grösser und grösser machen, die relative Häufigkeit gegen $\frac{1}{6}$ strebt. Die Wahrscheinlichkeit p kann hier als Grenzwert einer Folge von relativen Häufigkeiten angesehen werden, wenn die Anzahl der Versuche gegen Unendlich strebt. \diamond

Beispiel 3 Wir betrachten einen aus Holz gefertigten Quader, dessen Seitenflächen ebenfalls von 1 bis 6 nummeriert sind [7]. Dieser Quader ist nicht ideal. Wir können die Wahrscheinlichkeiten der einzelnen Resultate nicht durch Symmetrieüberlegungen festlegen. Wir werfen den Quader 1'000 mal und erhalten die folgenden absoluten und relativen Häufigkeiten:

Zahl	absolute Häufigkeit	relative Häufigkeit	festgelegte Wahrscheinlichkeit
1	97	0.097	0.10
2	73	0.073	0.07
3	329	0.329	0.33
4	334	0.334	0.33
5	68	0.068	0.07
6	99	0.099	0.1
Summe	1'000	1	1



Wir erhalten die *festgelegten Wahrscheinlichkeiten* durch Rundung. Das entsprechen vermutlich nicht den exakten Wahrscheinlichkeiten, sind aber für die Praxis sicher akzeptable Näherungen. \diamond

Beispiel 4 Die Lungenkrankheit Tuberkulose kann indirekt mit einem Hauttest nachgewiesen werden. Das Resultat dieses Tests kann *positiv*, *negativ* oder *ungewiss* sein. Falls die Wahrscheinlichkeit für ein positives Resultat 10% wäre, dann würde dies bedeuten, dass wenn wir den Test n mal anwenden, wobei n gross sein soll, wir in etwa $n/10$ der Fälle ein positives Resultat erhalten würden. Wenn wir n grösser und grösser machen würden, dann würde sich diese relative Häufigkeit immer mehr dem Prozentsatz 10% annähern. \diamond

Die obigen 4 Beispiele sind Beispiele für **Zufallsexperimente**. Wenn das Geschlecht des Kindes nicht vor der Geburt bestimmt worden ist, dann weiss man nicht, ob das Neugeborene ein Knabe oder ein Mädchen sein wird. Wenn Sie mit einem perfekten Würfel würfeln, dann wissen Sie das Resultat nicht im Voraus. Dasselbe gilt für Resultat des quaderförmigen Würfels sowie für das Resultat des Hauttests. Zufallsexperimente können wiederholt werden: so gibt es viele Geburten (allerdings nicht mit der gleichen Mutter!), den Würfel und den Quader kann man so oft werfen wie man möchte und auch den Hauttest kann man theoretisch beliebig oft anwenden.

1.2 Stichprobenraum und Ereignisse

Die Menge aller möglichen Resultate eines Zufallsexperiments heisst **Stichprobenraum**. Die Stichprobenräume der vier Beispiele sind:

$$\Omega = \{\text{Knabe, Mädchen}\}, \quad \Omega = \{1, 2, 3, 4, 5, 6\} \quad \text{und} \quad \Omega = \{\text{positiv, negativ, ungewiss}\} .$$

Bevor wir das nächste Beispiel in Angriff nehmen können, müssen wir noch den Begriff des geordneten Paares sowie das kartesische Produkt zweier Mengen A und B einführen. Die Mengenlehre wird ausführlicher im Modul *Mathematik für Medizininformatik* behandelt.

Unter einem **geordneten Paar** versteht man eine Menge bestehend aus zwei Elementen, wobei die Reihenfolge eine Rolle spielt. Anstelle der geschweiften Klammern verwendet man runde Klammern: (x, y) . Das Element x ist die erste und das Element y die zweite Komponente. Zwei geordnete Paare sind genau dann gleich, wenn die entsprechenden Komponenten gleich sind:

$$(x, y) = (u, v) \iff x = u \wedge y = v$$

Seien A und B zwei Mengen. Unter dem **kartesischen Produkt** $A \times B$ versteht man die Menge aller geordneter Paare (x, y) , wobei $x \in A$ und $y \in B$ ist:

$$A \times B = \{(x, y) : x \in A \wedge y \in B\}$$

Beispiel 5 Gegeben seien die Mengen $A = \{1, 2, 3\}$ und $B = \{x, y\}$. Bestimmen Sie $A \times B$, $B \times A$ und $A \times A$.

Die Lösung wird im Unterricht erarbeitet. \diamond

Natürlich lässt sich der Begriff des geordneten Paares verallgemeinern:

- Ein geordnete Folge bestehend aus 3 Elementen nennt man ein Tripel. Zwei Tripel sind genau dann gleich, wenn die entsprechenden Komponenten gleich sind:

$$(a_1, a_2, a_3) = (b_1, b_2, b_3) \iff a_1 = b_1 \wedge a_2 = b_2 \wedge a_3 = b_3$$

Wenn A_1 , A_2 und A_3 drei Mengen sind, dann ist $A_1 \times A_2 \times A_3$ die Menge aller Tripel, wo die erste Komponenten in A_1 , die zweite Komponente in A_2 und die dritte Komponente in A_3 ist.

- Sei $n \in \mathbb{N}$. Eine geordnete Folge von n Elementen, bezeichnet man als n -Tupel. Zwei n -Tupel sind genau dann gleich, wenn die entsprechenden Komponenten gleich sind:

$$(a_1, a_2, \dots, a_n) = (b_1, b_2, \dots, b_n) \iff a_1 = b_1 \wedge a_2 = b_2 \wedge \dots \wedge a_n = b_n$$

Beispiel 6 Bestimmen Sie den Stichprobenraum der folgenden Zufallsexperimente:

- (a) Wir werfen eine Münze.
- (b) Wir werfen zwei mal hintereinander eine Münze.
- (c) Wir werfen zwei Würfel mit den Farben schwarz und rot.
- (d) Wir werfen eine Münze sowie einen roten und einen schwarzen Würfel.

Die Lösung wird im Unterricht erarbeitet. \diamond

Wir betrachten das Beispiel mit dem Würfel. Manchmal interessiert man sich nicht in erster Linie für die gewürfelte Augenzahl. Vielleicht möchte man nur wissen, ob die Augenzahl gerade ist. Man interessiert sich also in diesem Fall für die Teilmenge

$$E = \{2, 4, 6\}$$

des Stichprobenraums. Ganz allgemein versteht man unter einem **Ereignis** eine Teilmenge E des Stichprobenraums Ω . Das Ereignis *Die gewürfelte Augenzahl ist gleich 5* entspricht der Teilmenge $E = \{5\}$. Das Ereignis *Die gewürfelte Augenzahl ist ungerade* entspricht der Teilmenge $E = \{1, 3, 5\}$. Man sagt weiter, dass das Ereignis E realisiert wird, wenn das Resultat ω des Zufallsexperiments in E enthalten ist.

Wir betrachten einige spezielle Ereignisse:

- **Das sichere Ereignis:** $E = \Omega$
- **Das unmögliche Ereignis:** $E = \emptyset$
- **Gegenereignis von $E \subset \Omega$:** E^c

- **Ein Elementarereignis:** $E = \{\omega\}$, wobei $\omega \in \Omega$.

Da die Ereignisse *Teilmengen* von Ω sind, kann man die Operationen \cup und \cap der Mengenlehre auf sie anwenden.

Beispiel 7 Sei Ω ein Stichprobenraum eines Zufallsexperiments und E und F zwei Ereignisse. Interpretieren Sie die folgenden Ereignisse:

- $E \cup F$;
- $E \cap F$.

Die Lösung wird im Unterricht erarbeitet. \diamond

1.3 Wahrscheinlichkeit

Man kann sagen, dass eine **Wahrscheinlichkeit**

- eine relative Häufigkeit sein kann (Beispiel 1) oder
- der Grenzwert einer Folge von relativen Häufigkeiten, wenn wir die Anzahl der Versuche immer grösser und grösser machen (Beispiele 2, 3 und 4).

Unter einer **Wahrscheinlichkeit** versteht man in der Mathematik eine Funktion P , welche jedem Ereignis $E \subset \Omega$ eine Zahl $P(E)$ zuordnet. Dabei müssen die folgenden 3 plausiblen Eigenschaften erfüllt sein:

- (1) $0 \leq P(E) \leq 1$
- (2) $P(\Omega) = 1$
- (3) Falls E und F nicht gleichzeitig eintreten können ($E \cap F = \emptyset$), gilt:

$$P(E \cup F) = P(E) + P(F) \quad (1)$$

Zwei Ereignisse E und F eines Stichprobenraums Ω , die nicht gleichzeitig eintreten können, heissen **unvereinbar**. Die Formel (1) ist auch gültig, für mehr als 2 Ereignisse, sofern diese **paarweise** unvereinbar sind. Seien $A, B, C \subset \Omega$ drei paarweise unvereinbare Ereignisse: $A \cap B = A \cap C = B \cap C = \emptyset$. Es gilt dann

$$\boxed{P(A \cup B \cup C) = P(A) + P(B) + P(C)} .$$

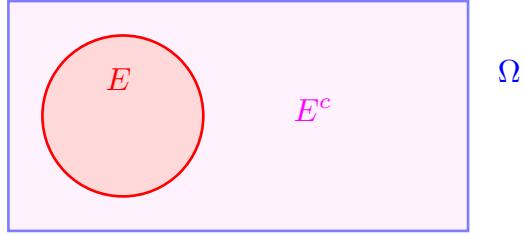
Da E und sein Gegenereignis E^c unvereinbar sind und $E \cup E^c = \Omega$ ist, folgt mit den Eigenschaft (2) und (3):

$$1 = P(\Omega) = P(E) + P(E^c) \implies P(E^c) = 1 - P(E)$$

Also:

$$\boxed{P(E^c) = 1 - P(E)} \quad (2)$$

Wir werden weiter unten ein Beispiel sehen, wo es einfacher ist zuerst $P(E^c)$ zu berechnen.



Wir betrachten jetzt einen wichtigen Spezialfall. Es sei $n \in \mathbb{N}$. Wir betrachten ein Zufallsexperiment mit dem Stichprobenraum

$$\Omega = \{x_1, x_2, \dots, x_n\}.$$

Wenn jedes Resultat die gleiche Wahrscheinlichkeit besitzt, das heisst, wenn gilt

$$P(\{x_1\}) = P(\{x_2\}) = \dots P(\{x_n\})$$

spricht man von einem **Laplace-Raum**. Aus der obigen Eigenschaft (2) ergibt sich leicht, dass dann

$$P(\{x_i\}) = \frac{1}{n}, \quad 1 \leq i \leq n$$

ist. Wenn das Ereignis E k Elemente umfasst, muss

$$P(E) = \frac{k}{n} = \frac{|E|}{|\Omega|}$$

sein, wobei $|E|$ die Anzahl Elemente von E ist. Man formuliert dieses Resultat oft so:

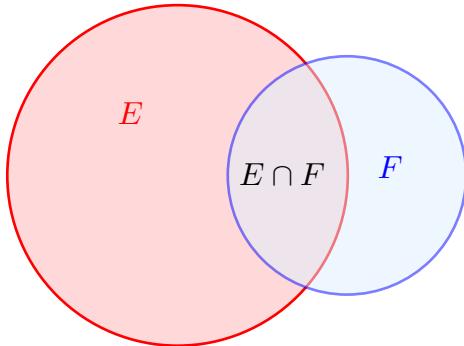
$P(E) = \frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl der möglichen Fälle}}$
--

Dies ist die **klassische Definition einer Wahrscheinlichkeit**. Beachten Sie bitte, dass diese Definition nur gültig ist, wenn alle Resultate die gleiche Wahrscheinlichkeit besitzen, es sich also bei (Ω, P) um einen Laplace-Raum handelt.

Beispiel 8 Wir machen einen Wurf mit einem perfekten Würfel. Bestimmen Sie die Wahrscheinlichkeit, dass die gewürfelte Augenzahl gerade ist.

Die Lösung wird im Unterricht erarbeitet. ◇

Wir wollen uns jetzt überlegen wie wir die Formel (1) auf zwei beliebige Ereignisse $E, F \subset \Omega$ verallgemeinern können. Die nachfolgende Zeichnung macht ersichtlich, dass die Wahrscheinlichkeit des Schnittbereichs subtrahiert werden muss, weil er sonst **doppelt** gezählt wird.



Wir erhalten so die Formel:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \quad (3)$$

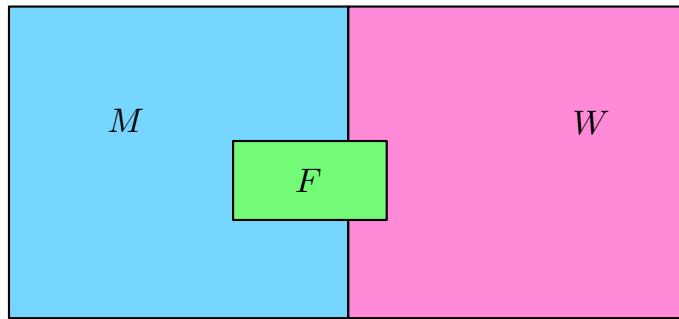
Beispiel 9 28% aller amerikanischen Männer rauchen Zigaretten, 7% rauchen Zigarre und 5% rauchen sowohl Zigarette als auch Zigarre. Wie hoch ist der Prozentsatz der Männer, die weder Zigarette noch Zigarre rauchen?

Die Lösung wird im Unterricht erarbeitet. \diamond

1.4 Bedingte Wahrscheinlichkeit

Wir untersuchen 10'000 Personen auf Farbenblindheit (rot/grün). Die Resultate dieser Untersuchung sind in der nachfolgenden Tabelle enthalten:

	männlich M	weiblich W	total
farbenblind F	423	65	488
normal N	4'848	4'664	9'512
total	5'271	4'729	10'000



Falls wir zufällig eine Person auswählen, dann ist diese:

- ein Mann mit der (empirischen) Wahrscheinlichkeit $\frac{5'271}{10'000}$;
- farbenblind mit der Wahrscheinlichkeit $\frac{488}{10'000}$.

Wir wählen zufällig eine Person aus und stellen sogleich fest, dass es sich um einen Mann handelt. Mit welcher Wahrscheinlichkeit ist der Mann farbenblind?

$|M| = 5'271$ ist die Zahl der untersuchten Männer. $|M \cap F| = 423$ ist die Zahl der Männer, die farbenblind sind. Falls wir die Information benutzen, dass die gewählte Person ein Mann ist, dann reduziert sich die Zahl der möglichen Fälle auf $|M| = 5'271$. Also gilt

$$P = \frac{|M \cap F|}{|M|} = \frac{423}{5'271} = 8.03\% .$$

Diese neue Wahrscheinlichkeit heisst **bedingte Wahrscheinlichkeit**. Sie wird mit

$$P(F|M)$$

bezeichnet. (gelesen: „bedingte Wahrscheinlichkeit von F , falls man weiss, dass M eingetroffen ist“)

Ein weiteres Beispiel einer bedingten Wahrscheinlichkeit ist

$$P(M|F) = \frac{|M \cap F|}{|F|} = \frac{423}{488} = 86.68\% .$$

Falls wir zufällig eine farbenblinde Person auswählen, dann ist diese mit einer Wahrscheinlichkeit von 86.68% ein Mann.

In den obigen Beispielen haben wir die absoluten Häufigkeiten $|M \cap F|$ und $|M|$ (und $|F|$) zur Berechnung der bedingten Wahrscheinlichkeit benutzt. Häufig kennt man jedoch nur die entsprechenden Wahrscheinlichkeiten

$$P(F \cap M) \text{ und } P(M) .$$

Deshalb suchen wir eine Formel für $P(F|M)$, wo diese Wahrscheinlichkeiten auftreten.

$$P(M) = \frac{|M|}{|\Omega|} \quad P(F \cap M) = \frac{|F \cap M|}{|\Omega|}$$

Wir erhalten so

$$P(F|M) = \frac{|F \cap M|}{|M|} = \frac{|F \cap M|/|\Omega|}{|M|/|\Omega|} = \frac{P(F \cap M)}{P(M)} \quad (4)$$

Wir definieren: falls $A, B \subset \Omega$ zwei Ereignisse sind mit $P(B) \neq 0$, dann ist die **bedingte Wahrscheinlichkeit, dass A eintrifft, wenn man weiss, dass B eingetroffen ist**, geschrieben $P(A|B)$, gegeben durch:

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}} \quad (5)$$

Aus (5) erhalten wir die Multiplikationsformel:

$$\boxed{P(A \cap B) = P(B) \cdot P(A|B)} \quad (6)$$

Um auf der rechten Seite eine alphabetische Reihenfolge zu erhalten, vertauschen wir die Ereignisse A und B und erhalten wegen $B \cap A = A \cap B$ so:

$$\boxed{P(A \cap B) = P(A) \cdot P(B|A)} \quad (7)$$

Diese Formel kann leicht auf 3 (oder mehr Ereignisse) verallgemeinert werden. Seien A, B und C drei Ereignisse des gleichen Stichprobenraums Ω . dann gilt:

$$\boxed{P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)} \quad (8)$$

Diese Formeln kommen bei mehrstufigen Zufallsexperimenten zur Anwendung. Wir werden später Beispiele dazu betrachten.

1.5 Unabhängigkeit

Wir betrachten ein Ereignis A als **unabhängig** von B , wenn gilt:

$$P(A|B) = P(A) \quad (9)$$

Die Wahrscheinlichkeit für das Eintreten von A ist unabhängig davon, ob B eingetreten ist oder nicht. Die Tatsache, dass B eingetreten ist oder nicht liefert also keinerlei Information über das Eintreten von A . Falls (9) nicht gilt, ist A **abhängig** von B .

Man kann leicht zeigen, dass falls (9) gilt, auch $P(B|A) = P(B)$ gilt, das heisst, B ist dann auch unabhängig von A . In der Wahrscheinlichkeitsrechnung ist also die Unabhängigkeit, und damit auch die Abhängigkeit, ein symmetrischer Begriff. Wenn A zeitlich nach B eintritt, und A **abhängig** von B ist, dann entsteht die paradoxe Situation, dass B auch abhängig von A ist. Es erscheint eigenartig hier von Abhängigkeit zu sprechen, da B vor A eintritt. Man muss einfach daran denken, dass in der Wahrscheinlichkeitsrechnung Abhängigkeit keine Kausalität beinhalten muss.

Wenn $P(A|B) = P(A)$ ist, können wir in Formel (6) $P(A|B)$ durch $P(A)$ ersetzen. Wir erhalten so die Formel:

$$P(A \cap B) = P(A) \cdot P(B) \quad (10)$$

Man benutzt in der Regel diese Formel für die Definition der Unabhängigkeit. Die Ereignisse A und B heissen **unabhängig**, wenn (10) gilt. Wenn (10) nicht gilt, heissen A und B **abhängig**.

Die Formel (10) kann auf zwei Arten benutzt werden:

- (i) Falls man $P(A)$, $P(B)$ und $P(A \cap B)$ kennt, so kann man nachprüfen, ob A und B unabhängig sind.
- (ii) Bei der **Modellierung** eines Problems, nimmt man manchmal die Unabhängigkeit der Ereignisse A und B als gegeben an. Man kann dann die Formel (10) benutzen.

Beispiel 10 Man wirft dreimal eine (ideale) Münze. Der Stichprobenraum ist gegeben durch

$$\Omega = \{KKK, KKZ, KZK, KZZ, ZKK, ZKZ, ZZK, ZZZ\} .$$

Aus Symmetriegründen ist es ein Laplace-Raum, das heisst, jedes Element besitzt die gleiche Wahrscheinlichkeit.

Man betrachtet die Ereignisse

$$\begin{aligned} A &= \{\text{der erste Wurf ergibt Kopf}\} , \\ B &= \{\text{der zweite Wurf ergibt Kopf}\} , \\ C &= \{\text{mindestens zwei aufeinanderfolgende Würfe ergeben Kopf}\} . \end{aligned}$$

Welche Paare der obigen Ereignisse sind unabhängig?

Die Lösung wird im Unterricht erarbeitet. ◇

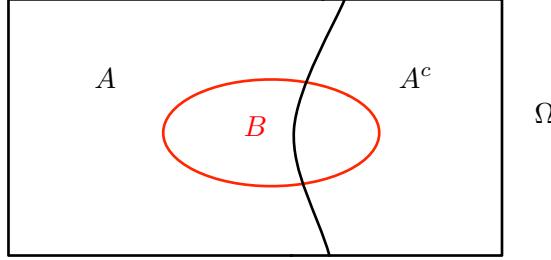
Beispiel 11 Es gibt zwei Tests, um eine bestimmte Krankheit zu entdecken. Der erste Test liefert im Fall einer Person, die tatsächlich an dieser Krankheit leidet, in 70% der Fälle das richtige Resultat und der zweite Test in 80% der Fälle. Ein Ärztin wendet beide Tests auf eine Person an, die an dieser Krankheit leidet. Wir nehmen an, dass die Tests unabhängig voneinander sind. Wie gross ist die Wahrscheinlichkeit, dass mindestens einer der beiden Tests positiv ist?

Die Lösung wird im Unterricht erarbeitet. ◇

1.6 Formel der totalen Wahrscheinlichkeit

Wir betrachten einen Stichprobenraum Ω und zwei Ereignisse A und B . Das Ereignis A und sein Gegenereignis A^c bilden dann eine Partition von Ω , das heißt, es gilt:

$$A \cup A^c = \Omega \quad \text{und} \quad A \cap A^c = \emptyset$$



Es gilt:

$$B = \Omega \cap B = (A \cup A^c) \cap B = (A \cap B) \cup (A^c \cap B)$$

Da $A \cap B$ und $A^c \cap B$ unvereinbar sind, folgt:

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

Mit der Multiplikationsformel (6) ergibt sich dann die **Formel der totalen Wahrscheinlichkeit**:

$$P(B) = P(A) \cdot P(B|A) + P(A^c) \cdot P(B|A^c) \quad (11)$$

Mit der Formel (11) lässt sich eigentlich das Problem lösen. Wenn ein konkretes Problem gelöst werden muss, ist es manchmal sicherer einen kleinen Umweg zu machen. Es gibt zwei Möglichkeiten wie $P(B)$ systematisch berechnet werden kann. Die erste Möglichkeit ist die **Vierfeldertafell**:

	B	B^c	
A	$ A \cap B $	$ A \cap B^c $	$ A $
A^c	$ A^c \cap B $	$ A^c \cap B^c $	$ A^c $
	$ B $	$ B^c $	$ \Omega $

Beachten Sie, dass $|M|$ die Anzahl Elemente einer Menge M bezeichnet. Wir haben also in der Vierfeldertafel die absoluten Häufigkeiten eingetragen. Als konkretes Beispiel kann das Problem mit der Farbeinblindheit auf Seite 7 dienen. Indem wir diese durch $|\Omega|$ dividieren, erhalten wir die relativen Häufigkeiten, welche wir als Wahrscheinlichkeiten interpretieren können:

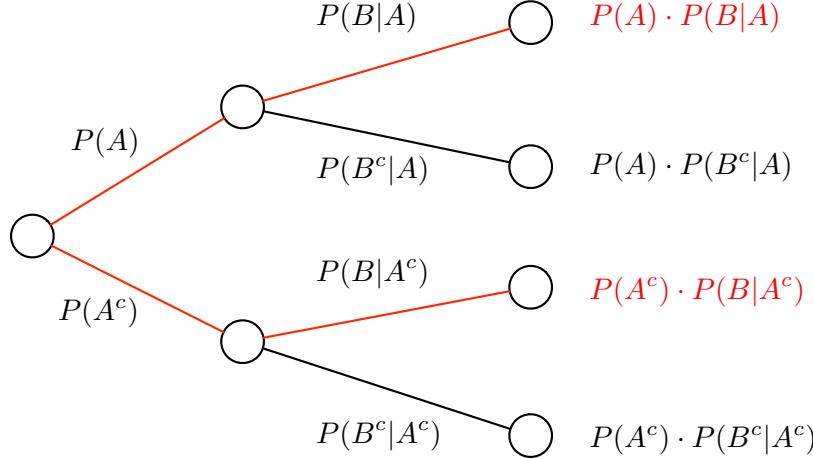
	B	B^c	
A	$P(A \cap B)$	$P(A \cap B^c)$	$P(A)$
A^c	$P(A^c \cap B)$	$P(A^c \cap B^c)$	$P(A^c)$
	$P(B)$	$P(B^c)$	$P(\Omega) = 1$

Es genügt jetzt die Wahrscheinlichkeiten der ersten Spalte zu addieren um $P(B)$ zu erhalten:

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

Die zweite Möglichkeit besteht aus einem **Baumdiagramm**. Damit wird das einstufige Zufallsexperiment *Wir wählen zufällig eine Person aus* als **zweistufiges Experiment** interpretiert.

- (1) Man trägt auf den Kanten die angegebenen Wahrscheinlichkeiten ab.
- (2) Durch Multiplikation erhält man dann die Wahrscheinlichkeiten bei den Blättern des Baums (Knoten ganz rechts).
- (3) Man addiert dann die roten Wahrscheinlichkeiten.



Beispiel 12 Angenommen von 100'000 Frauen mit negativer Mammographie wird in den nächsten 2 Jahren bei 20 Brustkrebs diagnostiziert, während bei den Frauen mit positiver Mammographie bei einer aus 10 Brustkrebs in den nächsten 2 Jahren diagnostiziert wird. Wir nehmen weiter an, dass in der Population aller Frauen 7% eine positive Mammographie haben. Wie gross ist die Wahrscheinlichkeit, dass bei einer aus der Population der Frauen zufällig ausgewählten Frau in den nächsten zwei Jahren Brustkrebs diagnostiziert wird? Verwenden Sie sowohl eine Vierfeldertafel wie auch ein Baumdiagramm.

Die Lösung wird im Unterricht erarbeitet. ◇

Bevor wir die Formel von Bayes herleiten, schieben wir einen Abschnitt über mehrstufige Zufallsexperimente ein:

1.7 Mehrstufige Zufallsexperimente

Wir betrachten zuerst ein ganz einfaches Beispiel. Wir werfen zwei Würfel und interessieren uns für die Wahrscheinlichkeit, dass beide die Augensumme 6 anzeigen. Wir können dieses Problem auf zwei Arten lösen:

- (1) Wir werfen beide Würfel miteinander. Auch wenn beide Würfel die gleiche Farbe habe, können wir sie im Geiste unterscheiden, denn es sind zwei verschiedene Körper. Wir geben dem einen Würfel die Nummer 1 und dem zweiten die Nummer 2. Der Stichprobenraum ist gleich

$$\Omega = \{(i, j) : 1 \leq i \leq 6 \wedge 1 \leq j \leq 6\} .$$

Dieser Raum enthält 35 Elemente. Aus Symmetriegründen handelt es sich um einen Laplace-Raum. Damit ist

$$P(\{(6, 6)\}) = \frac{1}{36} .$$

Auf der linken Seite haben wir viele Klammern. Allerdings ist das die richtige Schreibweise.

- (2) Wir nehmen den ersten Würfel und werfen ihn. Dann nehmen wir den zweiten Würfel und werfen ihn. Wenn beide Würfel komplett identisch sind, könnte man den gleichen Würfel auch zweimal verwenden. Es liegt jetzt ein *zweistufiges* Zufallsexperiment vor. Die Stichprobenräume der beiden Teilexperimente sind gleich

$$\Omega_1 = \Omega_2 = \{1, 2, 3, 4, 5, 6\}$$

Der Stichprobenraum beider Zufallsexperimente zusammen ist dann gleich

$$\Omega = \Omega_1 \times \Omega_2 = \{(i, j) : 1 \leq i \leq 6 \wedge 1 \leq j \leq 6\} .$$

Wir haben dann die beiden Ereignisse

$$A = \{\text{der erste Wurf gibt } 6\} = \{6\} \times \Omega_2 = \{(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

und

$$B = \{\text{der zweite Wurf gibt } 6\} = \Omega_1 \times \{6\} = \{(1, 6), (2, 6), (3, 6), (4, 6), (5, 6), (6, 6)\}$$

Wir verwenden dann die Formel

$$P(A \cap B) = P(A) \cdot P(B|A) = P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} .$$

Wir haben benutzt, dass A und B unabhängig sind.

Das nächste Beispiel [6] ist interessanter, da A und B abhängig sind:

Beispiel 13 Eine Urne enthält 3 schwarze und 1 rote Kugel. Man zieht zufällig eine Kugel, notiert sich deren Farbe und legt die Kugel zusammen mit einer *zusätzlichen Kugel der gleichen Farbe* in die Urne zurück. Man zieht dann wieder zufällig eine Kugel. Wie gross ist die Wahrscheinlichkeit, dass diese zweite gezogene Kugel rot ist?

Lösung: Der Stichprobenraum Ω besteht aus den folgenden geordneten Paaren:

$$\Omega = \{(s, s), (s, r), (r, s), (r, r)\}$$

Hier ist die erste Komponente gleich dem Resultat der ersten Ziehung und die zweite Komponente gleich dem Resultat der zweiten Ziehung. Der Buchstabe s steht für schwarze Kugel und der Buchstabe r für rote Kugel. Gesucht wird die Wahrscheinlichkeit des Ereignisses

$$B = \{(s, r), (r, r)\} .$$

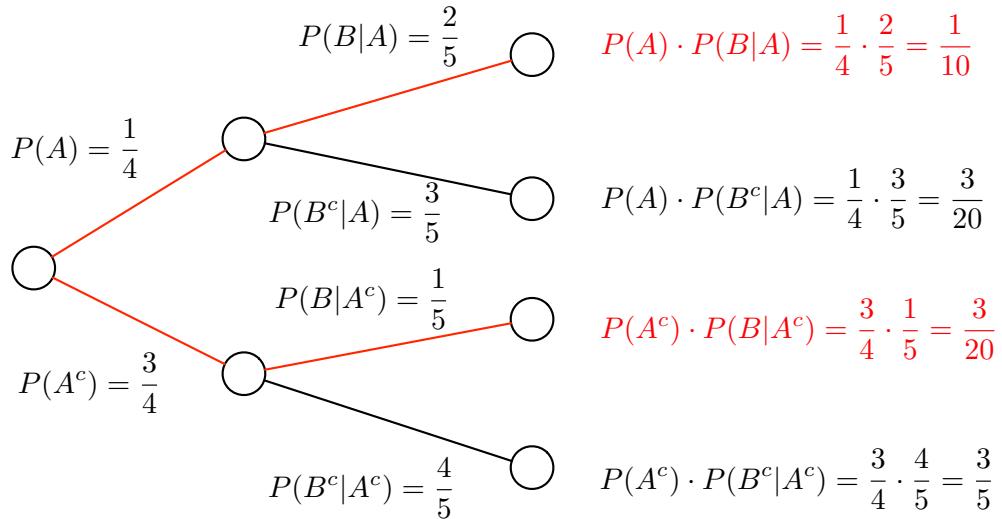
Wir definieren das Ereignis

$$A : \text{Die erste gezogene Kugel ist rot}$$

Natürlich ist dann

$$A^c : \text{Die erste gezogene Kugel ist schwarz} .$$

Wir können jetzt direkt den obigen Baum wiederverwenden, wobei wir zusätzlich die konkreten Wahrscheinlichkeiten notieren.



Indem wir die roten Wahrscheinlichkeiten addieren, erhalten wir $P(B)$:

$$P(B) = \frac{1}{10} + \frac{3}{20} = \frac{5}{20} = \frac{1}{4}$$

Für die Lösung des Problems benötigt man nicht den ganzen Baum. Es genügt $P(\{s, r\})$ und $P(\{r, r\})$ zu berechnen und dann die beiden Wahrscheinlichkeiten zu addieren. Die Berechnung der beiden Wahrscheinlichkeiten geschieht einfach über die Formel (7). Das entspricht den beiden roten Ästen im obigen Baum. \diamond

Im nächsten Beispiel ist es einfacher zuerst $P(E^c)$ zu berechnen.

Beispiel 14 Sie sind zusammen mit $n - 1$ Kolleginnen und Kollegen $n \geq 2$. Wir nehmen an, dass Niemand am 29. Februar Geburtstag hat.

- (a) Wie gross ist die Wahrscheinlichkeit, dass mindestens eine Kollegin bzw. ein Kollege am gleichen Tag (gleicher Monat und Tag) Geburtstag hat wie Sie?
- (b) Wie gross ist die Wahrscheinlichkeit, dass mindestens zwei Personen in der Gruppe am gleichen Tag (gleicher Monat und Tag) Geburtstag haben? Programmieren Sie die Formel und berechnen Sie die Wahrscheinlichkeit für

$n =$	5	10	15	20	22	23	24	30	40	50	70
$P(E) =$											

Die Lösung wird im Unterricht erarbeitet. \diamond

Das obige Problem (b) ist unter dem Namen **Geburtstagsparadoxon** bekannt. Es wird für gewisse Attacken in der Kryptographie verwendet.

1.8 Formel von Bayes

Wenn man $P(B|A)$ und $P(B|A^c)$ kennt, kann man $P(A|B)$ berechnen. Nach Definition der bedingten Wahrscheinlichkeit gilt:

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

Wir ersetzen den Zähler durch $P(A) \cdot P(B|A)$ und den Nenner durch den Ausdruck in der Formel der totalen Wahrscheinlichkeit (11):

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A^c) \cdot P(B|A^c)} \quad (12)$$

Auf der linken Seite steht die bedingte Wahrscheinlichkeit, dass A eintritt, wenn man weiss, dass B eingetreten ist. Auf der rechten Seite ist es gerade umgekehrt: hier stehen die bedingten Wahrscheinlichkeiten, dass B eintritt, wenn man weiss, dass A bzw. A^c eingetreten ist. Dies ist die **Formel von Bayes**¹. Im nächsten Abschnitt werden wir eine Anwendung dieser Formel sehen.

Wir geben noch die folgende interessante Interpretation der Formel von Bayes [6]:

1.9 Diagnostische Tests: Sensitivität und Spezifität

Eine wichtige Anwendung der Formel von Bayes ergibt sich bei einem diagnostischen Test. Dieser sollte idealerweise nur richtige Entscheidungen bringen, was natürlich in der Praxis nicht der Fall ist.

Wir betrachten einen hypothetischen diagnostischen Test, mit welchem eine bestimmte Krankheit erkannt werden kann. Wir wenden den Test auf 50 Personen an, die tatsächlich an dieser Krankheit leiden, und auf 50 Personen, die diese Krankheit nicht haben. Wir nehmen an, dass sich die folgende Vierfeldertafel ergibt:

		tatsächlich krank ja	tatsächlich gesund nein	Total
Test		pos.	47	10
		neg.	3	40
	Total	50	50	100

(13)

Wir werden anhand dieses Beispiels verschiedene Größen einführen:

Die **Sensitivität** eines diagnostischen Tests ist die Wahrscheinlichkeit, dass der Test positiv ist, wenn die Person an der Krankheit leidet:

$$\text{Sensitivität} = P(\text{Test}^+ | \text{Krankheit}) .$$

Die Sensitivität ist also die Rate der richtig-positiven Klassifikationen. In unserem Beispiel erhalten wir als Schätzwert

$$\frac{47}{50} \doteq 0.94$$

Die **Spezifität** eines diagnostischen Tests ist die Wahrscheinlichkeit, dass der Test negativ ist, wenn die Person **nicht** an der Krankheit leidet:

$$\text{Spezifität} = P(\text{Test}^- | \text{keine Krankheit}) .$$

Die Spezifität ist also die Rate der richtig-negativen Klassifikationen. In unserem Beispiel erhalten wir den folgenden Schätzwert:

$$\frac{40}{50} = 0.8$$

¹Thomas Bayes (London 1702 - Turnbridge Wells (Kent) 1761): englischer Theologe. Seine Beiträge zur Wahrscheinlichkeitsrechnung wurden erst posthum veröffentlicht.

Ideal wäre natürlich wenn die Sensitivität und die Spezifität eines diagnostischen Tests beide gleich 100% wären. Es existiert aber in der Praxis kein Test, der dieses Ziel erreicht. Die obigen Begriffe sind auch sinnvoll im Zusammenhang mit einem Symptom statt eines Tests, das typisch für eine Krankheit ist.

Sensitivität und Spezifität charakterisieren die Güte eines diagnostischen Tests. Sie geben die Wahrscheinlichkeit für ein positives oder negatives Testergebnis an, wenn man krank bzw. gesund ist. Eine Patientin oder ein Patient interessiert sich aber für die Wahrscheinlichkeit gesund oder krank zu sein, wenn der Test negativ bzw. positiv ist.

Unter einem **positiven Vorhersagewert (predictive value positive)** eines diagnostischen Tests, verstehen wir die bedingte Wahrscheinlichkeit, dass die Person die Krankheit hat, wenn der Test positiv ist:

$$PV^+ = P(\text{Krankheit} | \text{Test}^+)$$

Im obigen Beispiel (13) erhalten wir den folgenden Schätzwert:

$$\widehat{PV^+} = \frac{47}{57} \doteq 0.825$$

Der **negative Vorhersagewert (predictive value negative)** ist die Wahrscheinlichkeit, dass die Person die Krankheit nicht hat, wenn der Test negativ ist:

$$PV^- = P(\text{keine Krankheit} | \text{Test}^-)$$

In unserem Beispiel erhalten wir den folgenden Schätzwert:

$$\widehat{PV^-} = \frac{40}{43} \doteq 0.930$$

Falls die **Prävalenz $P(K)$** (=Krankheitshäufigkeit in der Referenzpopulation) bekannt ist, können mithilfe der Bayeschen Formel die Vorhersagewerte mit der Sensitivität und der Spezifität ausgedrückt werden:

$$PV^+ = \frac{\text{Sensitivität} \cdot P(K)}{\text{Sensitivität} \cdot P(K) + (1 - \text{Spezifität}) \cdot (1 - P(K))},$$

und

$$PV^- = \frac{\text{Spezifität} \cdot (1 - P(K))}{\text{Spezifität} \cdot (1 - P(K)) + (1 - \text{Sensitivität}) \cdot P(K)}.$$

Beispiel 15 Eine automatische Blutdruckmesseinrichtung klassifiziert 84% der Personen, die an Bluthochdruck leiden, richtig und 23% der Personen mit normalen falsch. Wir nehmen weiter an, dass 20% der erwachsenen Bevölkerung an Bluthochdruck leiden. Gesucht werden PV^+ und PV^- .

Die Lösung wird im Unterricht erarbeitet. ◇

Wir betrachten ein weiteres Beispiel zur Formel von Bayes, welches aus [11] entnommen wurde. Es handelt sich um den ELISA-Test (enzyme-linked immunosorbent assay) für den Nachweis von HIV-Antikörpern. Wir definieren die folgenden Ereignisse:

- A^+ : Im Blut sind Antikörper vorhanden
- A^- : Im Blut sind keine Antikörper vorhanden
- T^+ : Test auf Antikörper ist positiv
- T^- : Test auf Antikörper ist negativ

Für eine Anwendung dieses Tests im Jahre 1989 wurden die folgenden bedingten Wahrscheinlichkeiten angegeben:

$$\begin{aligned} P(T^+|A^+) &= 0.999 \quad (\text{Sensitivität des Tests}) \\ P(T^-|A^-) &= 0.995 \quad (\text{Spezifität des Tests}) \end{aligned}$$

Diese Wahrscheinlichkeiten sind sehr hoch und man hat den Eindruck, dass der Test sehr gut sein müsse.

Die Prävalenz wurde mit $P(A^+) = 0.001$ angegeben. Wir berechnen zuerst die Wahrscheinlichkeit, dass der Test für eine zufällig ausgewählten Person ein positives Resultat liefert:

$$\begin{aligned} P(T^+) &= P(T^+|A^+) \cdot P(A^+) + P(T^+|A^-) \cdot P(A^-) \\ &= 0.999 \cdot 0.001 + 0.005 \cdot 0.999 = 0.00599 \end{aligned}$$

Wir berechnen jetzt die Wahrscheinlichkeit, dass im Blut tatsächlich Antikörper vorhanden sind, wenn der Test positiv wird:

$$PV^+ = P(A^+|T^+) = \frac{P(T^+|A^+) \cdot P(A^+)}{P(T^+)} = \frac{0.999 \cdot 0.001}{0.00599} \doteq 0.167$$

Das heisst nur in 16.7% der Fälle in denen der Test angibt, dass Antikörper vorhanden sind, sich solche auch tatsächlich vorhanden! Wir berechnen jetzt noch die Wahrscheinlichkeit, dass keine Antikörper vorhanden sind, wenn der Test negativ ist:

$$PV^- = P(A^-|T^-) = \frac{P(T^-|A^-) \cdot P(A^-)}{P(T^-)} = \frac{0.995 \cdot (1 - 0.001)}{1 - 0.00599} \doteq 0.999995$$

Hier ist der Test sehr gut.

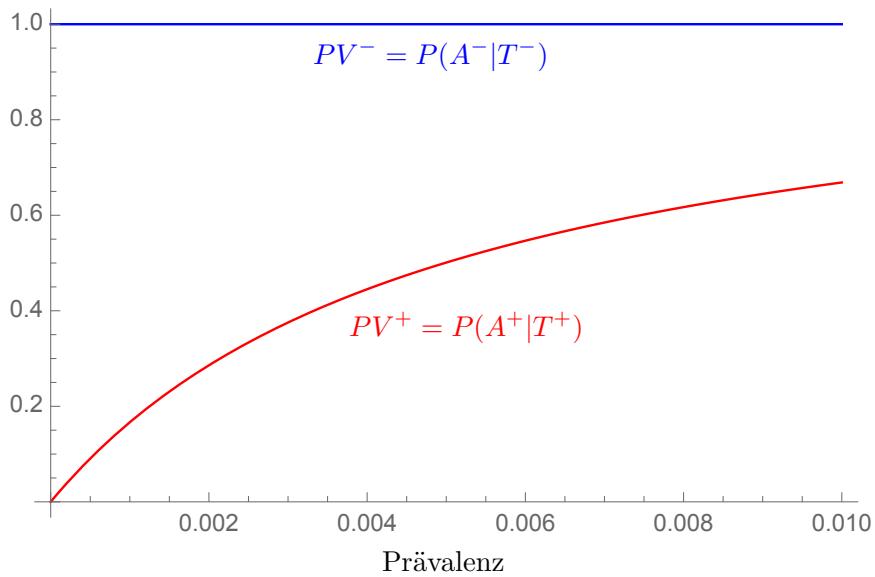
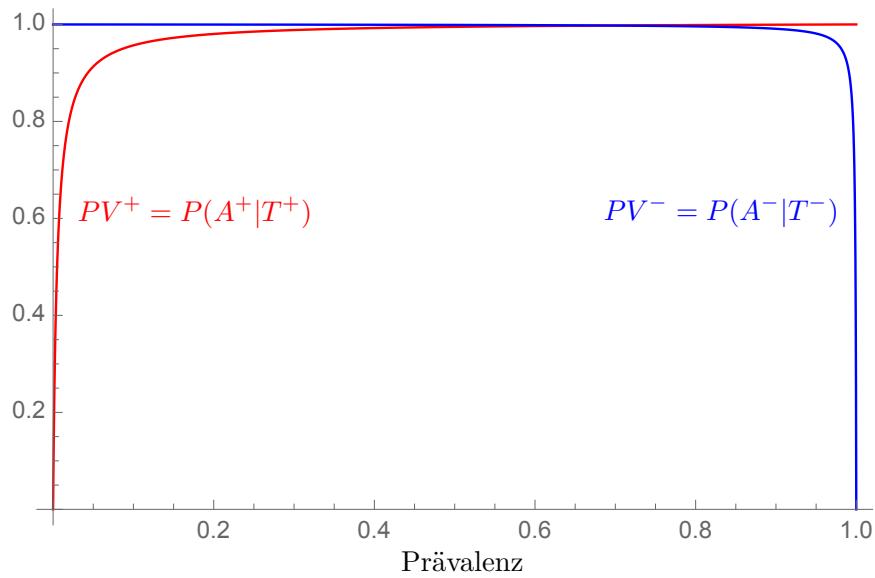
Der Test könnte verwendet werden, um Blutkonserven zu kontrollieren. Wenn der Test negativ ist, habe wir eine sehr hohe Sicherheit, dass tatsächlich keine Antikörper im Blut vorhanden sind. Allerdings werden in 83.3% der Fälle, in denen der Test positiv ist, Blutkonserven vernichtet, die keine Antikörper enthalten. Da die Wahrscheinlichkeit für T^+ aber weniger als 1% ist, ereignet sich dies sehr selten!

Wir wenden uns jetzt noch der Frage zu, ob die zu Beginn genannten Wahrscheinlichkeiten

$$P(T^+|A^+) = 0.999, \quad P(T^-|A^-) = 0.995 \quad \text{und} \quad P(A^+) = 0.001$$

realistisch sind. Die Sensitivität kann relativ genau bestimmt werden. Wir können Blutkonserven mit Antikörper versetzen und dann den Test darauf anwenden. Die Spezifität ist schwieriger zu bestimmen, denn wir benötigen Blutkonserven, von denen wir sicher sein müssen, dass sie keine Antikörper enthalten. Richtig schwierig wird es aber bei der Prävalenz. Bei dieser Wahrscheinlichkeit kann es sich nur um eine Schätzung handeln.

Die nachfolgenden Graphiken zeigen die beiden Wahrscheinlichkeiten $P(A^+|T^+)$ und $P(A^-|T^-)$ als Funktion der Prävalenz $P(A^+)$. Die beiden Graphen sind symmetrisch zur vertikalen Geraden durch den Punkt Prävalenz = 0.5. Wir sehen, dass $P(A^+|T^+)$ für sehr kleine Prävalenzen auch klein ist und empfindlich auf Änderungen reagiert.



Falls die Prävalenz sehr klein ist und nur ungenau bestimmt werden kann, wie das hier der Fall ist, wird die Wahrscheinlichkeit $P(A^+|T^+)$ ebenfalls klein und ungenau sein. Hier liegt ein Problem in der Anwendung der Formel von Bayes.

2 Zufallsvariablen

2.1 Begriff der Zufallsvariablen

Wir betrachten ein Zufallsexperiment mit Stichprobenraum Ω . Unter einer **Zufallsvariable** X versteht man eine reellwertige Funktion, die auf Ω definiert ist:

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) \end{aligned}$$

Die Zufallsvariable ordnet also jedem Element des Stichprobenraums eine reelle Zahl zu. Im Gegensatz zu Funktionen in der Mathematik, die häufig mit f, g, h, \dots bezeichnet werden, verwendet man für Zufallsvariablen Grossbuchstaben und zwar häufig X, Y oder Z .

Betrachten wir einige Beispiele:

- Wir würfeln mit einem Würfel. Falls die gewürfelte Zahl gerade ist, gewinnen wir so viele Franken, falls sie ungerade ist, verlieren wir das doppelte der Zahl. Wir bezeichnen den Gewinn oder Verlust mit X . X ist eine **diskrete** Zufallsvariable mit den Werten:

ω	1	2	3	4	5	6
$X(\omega)$	-2	2	-6	4	-10	6

- Wir betrachten die in der Schweiz wohnhaften Familien mit 4 Kindern. Dies ist unser Stichprobenraum Ω . Sei Y die Anzahl der Mädchen in einer zufällig aus Ω ausgewählten Familie. Y ist eine diskrete Zufallsvariable, welche die Werte 0, 1, 2, 3 oder 4 annimmt.
- Wir betrachten die Menge der Patienten einer Hausärztin. Sei Z der systolische Blutdruck eines zufällig ausgewählten Patienten. Z ist eine Zufallsvariable. Im Gegensatz zu den beiden vorhergehenden Beispielen kann Z - zumindest theoretisch - jeden Wert in einem bestimmten Intervall annehmen. Man spricht von einer **stetigen** Zufallsvariable.

Der Name Zufallsvariable ist historisch bedingt. Früher, und ab und zu auch heute noch, bezeichnete man bei einer Funktion $y = f(x)$ die Variable y als abhängige Variable, die man mit der Funktion identifizierte. Es ist auch so, dass man sich bei einer Zufallsvariable in erster Linie für die Werte interessiert, die diese annimmt. Beachten Sie auch, dass die Zufallsvariable selber nicht zufällig ist. Sobald wir beim Glücksspiel die gewürfelte Augenzahl kennen, kennen wir den Gewinn oder den Verlust. Das was zufällig ist, ist das Resultat des Zufallsexperiments, z.B. die gewürfelte Augenzahl oder der zufällig ausgewählte Patient.

Unter einer **diskreten** Zufallsvariablen versteht man eine Zufallsvariable, die nur **endlich viele** oder **abzählbar unendliche** viele Werte annimmt.

2.2 Die Verteilung einer Zufallsvariable

Wir betrachten im Folgenden nur diskrete Zufallsvariablen, welche endlich viele Werte annehmen. Sei n eine natürliche Zahl und

$$x_1, x_2, \dots, x_n$$

die Werte, welche die Zufallsvariable annimmt. Wir nehmen weiter an, dass der Wert x_i mit der Wahrscheinlichkeit p_i angenommen wird. Dies schreibt man manchmal so:

$$p_i := P(X = x_i)$$

Unter der **Verteilung von X** versteht man die Tabelle:

Werte von X	x_1	x_2	x_3	\cdots	x_n
$P(X = x_i)$	p_1	p_2	p_3	\cdots	p_n

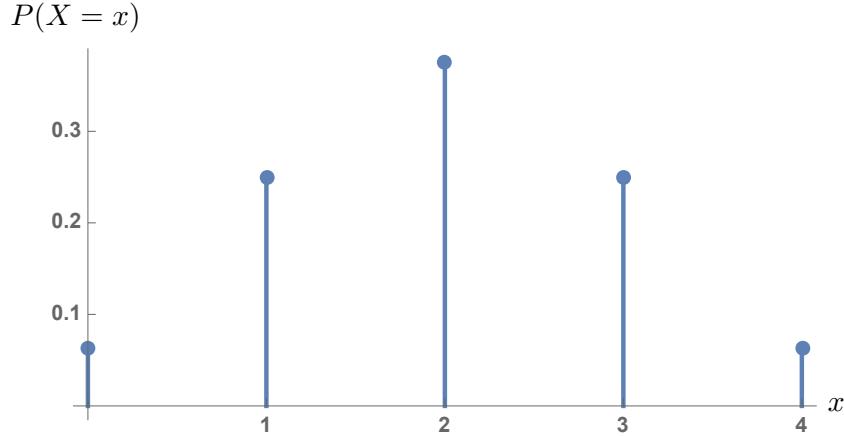
Mathematisch ist die Verteilung von X die Funktion

$$\begin{array}{ccc} \{x_1, x_2, \dots, x_n\} & \rightarrow & [0, 1] , \\ x & \rightarrow & P(X = x) , \end{array}$$

welche jedem Wert, den X annimmt, die Wahrscheinlichkeit zuordnet, mit welcher dieser angenommen wird. Wir können die Verteilung einer diskreten Zufallsvariable als Balkendiagramm darstellen. Betrachten wir das Beispiel mit den Familien mit 4 Kindern. Wenn wir annehmen - was gemäss Beispiel 1 nicht ganz richtig ist - dass die Wahrscheinlichkeit für eine Mädchengeburt gleich 0.5 ist, dann ergibt sich die folgende Verteilung für die Anzahl Mädchen X in solchen Familien (eine Begründung folgt später):

Werte von X	0	1	2	3	4
$P(X = x_i)$	6.25%	25%	37.5%	25%	6.25%

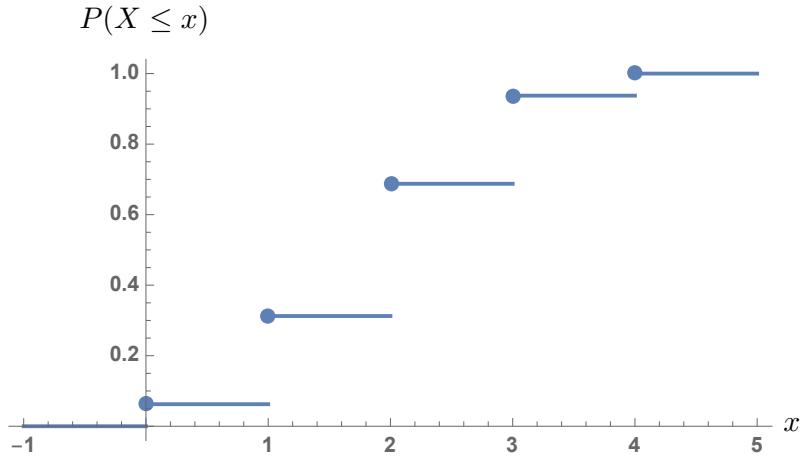
Beachten Sie, dass die Summe der Wahrscheinlichkeiten in der zweiten Zeile 1 ergeben muss, denn wir haben alle möglichen Werte von X berücksichtigt. Das Balkendiagramm der Verteilung ist unten abgebildet:



Gerade bei stetigen Zufallsvariablen, welche wir erst später behandeln, ist eine zweite Verteilungsfunktion sehr nützlich. Unter der **kumulativen Verteilungsfunktion** der Zufallsvariable X versteht man die Funktion

$$F(x) = P(X \leq x) , \quad x \in \mathbb{R} .$$

Beachten Sie, dass diese Funktion auf ganz \mathbb{R} definiert ist. Im Fall einer diskreten Zufallsvariablen X ergibt sich als Graph eine Treppenfunktion. Nachfolgend die kumulative Verteilungsfunktion für das Beispiel mit den Mädchengeburten:



Beispiel 16 Man wirft zwei (ideale) Würfel. Sei X die Zufallsvariable, die als Werte das Maximum der beiden gewürfelten Zahlen annimmt. Bestimmen Sie die Verteilung und die kumulative Verteilungsfunktion von X .

Die Aufgabe wird im Unterricht gelöst. \diamond

Beispiel 17 Man wirft dreimal eine manipulierte Münze, für welche $P(K) = 2/3$ und $P(Z) = 1/3$ ist. Sei Y die Zufallsvariable, welche die grösste Anzahl aufeinanderfolgender „Köpfe“ zählt. Bestimmen Sie die Verteilung und die kumulative Verteilungsfunktion von Y .

Die Aufgabe wird im Unterricht gelöst. \diamond

2.3 Unabhängigkeit von Zufallsvariablen

Wir kennen bereits die Unabhängigkeit von Ereignissen. Die **Unabhängigkeit von Zufallsvariablen** wird analog definiert.

Seien X und Y zwei Zufallsvariable, welche die Werte x_i ($1 \leq i \leq n$) bzw. y_j ($1 \leq j \leq m$) annehmen. Dann heissen X und Y **unabhängig**, wenn gilt:

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) \quad , \quad i = 1, 2, \dots, n ; j = 1, 2, \dots, m .$$

Falls wir noch eine dritte Zufallsvariable Z haben, welche die Werte z_k ($1 \leq k \leq K$) annimmt, dann sind X , Y und Z unabhängig, wenn für alle i , j und k gilt:

$$P(X = x_i, Y = y_j, Z = z_k) = P(X = x_i) \cdot P(Y = y_j) \cdot P(Z = z_k) \quad (14)$$

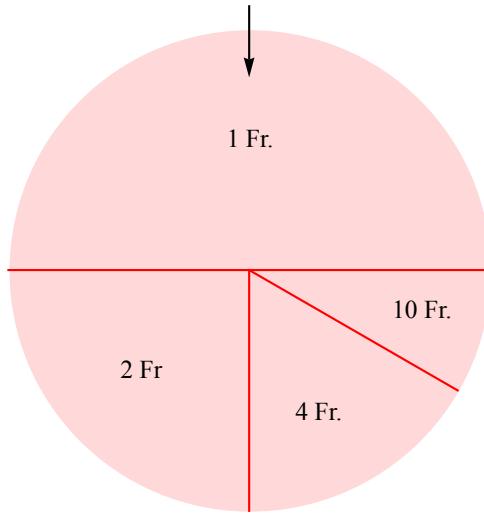
Analog definiert man die Unabhängigkeit für eine beliebige, endliche Anzahl von Zufallsvariablen.

2.4 Erwartungswert und Varianz einer diskreten Zufallsvariable

2.4.1 Erwartungswert

Wenn wir die Verteilung einer Zufallsvariable X kennen, ist vom Standpunkt der Wahrscheinlichkeitsrechnung alles über X bekannt. Wir werden in diesem Abschnitt zwei Kennzahlen einführen, welche X auf kompakte Art beschreiben. Die eine entspricht dem mittleren Wert, den X annimmt, und die andere ist ein Mass für die Streuung.

Wir betrachten ein Glücksrad, das die folgende Einteilung besitzt (Sektoren von 180° , 90° , 60° , 30°):



Die entsprechenden Gewinne sind 1, 2, 4 bzw. 10 Fr. Es liegt also eine Zufallsvariable vor, deren Verteilung durch die folgende Tabelle gegeben ist:

Wert x_i von X	1	2	4	10
$P(X = x_i)$	$1/2$	$1/4$	$1/6$	$1/12$

Wir stellen die folgende Frage: was ist der durchschnittliche Gewinn?

Falls wir 12-mal spielen, können wir den folgenden Gewinn erwarten:

$$\begin{array}{rcc}
 6 \text{ mal} & 1.- & : & 6.- \\
 3 \text{ mal} & 2.- & : & 6.- \\
 2 \text{ mal} & 4.- & : & 8.- \\
 1 \text{ mal} & 10.- & : & 10.- \\
 \hline
 \text{total} & & & 30.- \\
 \end{array}$$

Der durchschnittliche Gewinn ist also

$$30.- : 12 = 2.50$$

Es ist klar, dass wir **nicht genau** die obige Gewinnverteilung erwarten dürfen, wenn wir blass 12-mal spielen. Wenn wir jedoch n mal spielen, wo n eine sehr grosse Zahl ist, dann können wir auf Grund der intuitiven Interpretation der Wahrscheinlichkeit erwarten, dass der mittlere Gewinn sehr nahe bei diesem Wert liegt und im Grenzfall mit ihm identisch ist.

Dieser mittlere, hypothetische Gewinn heisst **Erwartungswert** der Zufallsvariablen X .

Beispiel 18 Wir würfeln mit einem perfekten Würfel. Sei X die erhaltene Augenzahl. Bestimmen Sie $E(X)$.

Die Lösung wird im Unterricht erarbeitet. ◇

Wir betrachten den allgemeinen Fall. Die Verteilung von X sei gegeben durch

Wert x_i von X	x_1	x_2	\dots	x_n
$P(X = x_i)$	p_1	p_2	\dots	p_n

Falls wir n mal spielen, so erwarten wir np_1 -mal das Resultat x_1 , np_2 -mal das Resultat x_2, \dots, np_n -mal das Resultat x_n . Der totale Gewinn ist also

$$np_1x_1 + np_2x_2 + \dots + np_nx_n .$$

Wir erhalten den durchschnittlichen Gewinn, indem wir durch n dividieren:

$$p_1x_1 + p_2x_2 + \dots + p_nx_n .$$

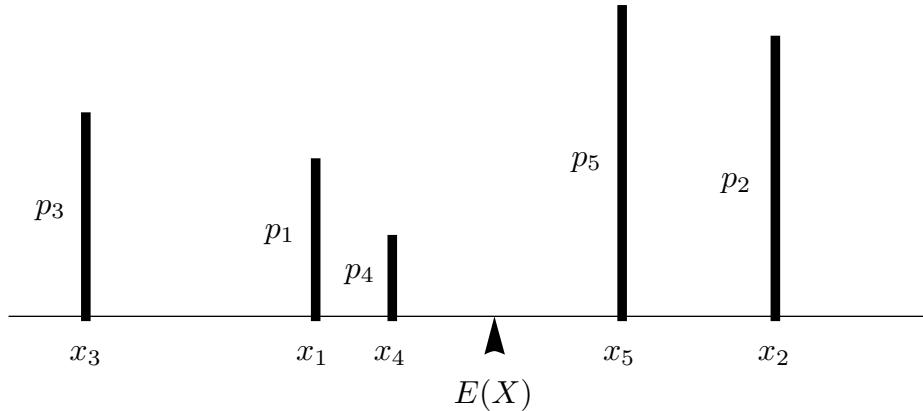
Wir bezeichnen diesen Wert als Erwartungswert, $E(X)$, von X :

$$E(X) = p_1x_1 + p_2x_2 + \dots + p_nx_n = \sum_{i=1}^n p_i x_i$$

Auf der rechten Seite haben wir die Summe mithilfe des **Summenzeichens** Σ kompakt geschrieben. Das Zeichen Σ ist der griechische Grossbuchstabe Sigma und steht als Abkürzung für das Wort Summe. Der Buchstabe i heisst **Laufindex** und nimmt ausgehen von 1 bis n in Schritten von 1 alle natürlichen Zahlen an. Die Beträge

$$p_1x_1, p_2x_2, \dots, p_nx_n$$

müssen aufaddiert werden.



Für jene, die sich dafür interessieren, geben wir noch eine **physikalische** Interpretation des Erwartungswertes: wir denken uns die Zahlenwerte x_i auf der horizontalen Zahlengerade abgetragen. An diesen Stellen platzieren wir Massen p_i die in der obigen Graphik durch vertikale Stäbe dargestellt sind. Wir nehmen an, dass die Zahlengerade gewichtslos sei. An welcher Stelle x müssen wir die Zahlengerade unterstützen, damit sie im Gleichgewicht bleibt? Die Summe der Drehmomente der Gewichtskräfte $p_i g$, wo g die Erdbeschleunigung bezeichnet, muss Null ergeben:

$$\sum_{i=1}^n g p_i (x - x_i) = 0 \implies x = \sum_{i=1}^n p_i x_i$$

Die Stelle $x = E(X)$ ist der Schwerpunkt.

Der Erwartungswert besitzt die folgende wichtige Eigenschaft. Wenn X und Y zwei Zufallsvariablen sind, dann ist der Erwartungswert der Summe gleich der Summe der Erwartungswerte:

$$E(X + Y) = E(X) + E(Y) \quad (15)$$

Beachten Sie, dass diese Formel auch gilt, wenn X und Y abhängig sind. Diese Formel kann man sofort auf n Zufallsvariablen X_i verallgemeinern:

$$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n) \quad (16)$$

Weiter gilt für eine beliebige reelle Zahlen a und b :

$$E(aX + b) = aE(X) + b \quad (17)$$

2.4.2 Varianz und Standardabweichung

Es fehlt uns noch ein Mass für die Streuung der Werte einer Zufallsvariable. Die drei Zufallsvariablen

$W = w_i$	0	$Y = y_i$	-1	+1	$Z = z_i$	-100	+100
$P(W = w_i)$	1	$P(Y = y_i)$	$\frac{1}{2}$	$\frac{1}{2}$	$P(Z = z_i)$	$\frac{1}{2}$	$\frac{1}{2}$

besitzen alle den gleichen Erwartungswert, nämlich 0. Es gibt aber viel grössere Unterschiede zwischen den Werten von Y als zwischen jenen der Zufallsvariable W (die konstant ist) und es gibt nochmals grössere Unterschiede zwischen den Werten von Z als zwischen jenen von Y .

Da eine Zufallsvariable X ihre Werte x_i , $i = 1, 2, \dots, n$ um den Erwartungswert $E(X)$ (=Mittelwert) herum annimmt, scheint es vernünftig Differenzen zwischen den Werten und dem Erwartungswert zu bilden. Damit sich positive und negative Differenzen nicht aufheben, bildet man das Quadrat dieser Differenzen. Zusätzlich werden diese Quadrate mit den Wahrscheinlichkeiten p_i multipliziert, mit denen x_i angenommen wird. Man erreicht damit, dass wenn die Wahrscheinlichkeit p_i sehr klein ist, der Beitrag von $(x_i - E(X))^2$ an der Gesamtsumme nicht zu gross wird. Man erhält so die folgende Grösse:

$$V(X) := \sum_{i=1}^n p_i(x_i - E(X))^2 = E((X - E(X))^2)$$

Diese Grösse heisst **Varianz** von X . Die Varianz hat als Masseinheit das Quadrat der Masseinheiten der Werte x_i . Wenn z.B. die x_i die Masseinheit cm besitzen, dann besitzt die Varianz die Masseinheit cm^2 .

Wir erhalten eine Grösse mit der gleichen Masseinheit, wenn wir die Wurzel aus der Varianz ziehen:

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

Diese Grösse heisst **Standardabweichung**. Für die Interpretation ist die Standardabweichung wichtiger als die Varianz. Man kann sich fragen wieso es die Varianz überhaupt braucht. Das hat mit wichtigen Eigenschaften zu tun, welche für die Varianz, aber nicht für die Standardabweichung gelten. Wir werden diese Eigenschaften weiter unten auflisten.

Auch für die Varianz gibt es eine *physikalische Interpretation*. Wir nehmen an, dass die Zahlengerade (siehe Graphik auf Seite 22)) um die vertikale Gerade durch den Schwerpunkt rotiere. Wir bezeichnen die konstante Winkelgeschwindigkeit mit ω (dieses ω hat nichts mit dem Stichprobenraum zu tun!). Die kinetische Energie eines Massenpunktes mit der Masse m und der Geschwindigkeit v ist bekanntlich $\frac{1}{2}mv^2$. In unserem Fall bewegt sich die Masse p_i mit der Geschwindigkeit $\omega(x_i - E(X))$. Die Energie des Systems ist also gegeben durch:

$$\frac{1}{2}\omega^2 \sum_{i=1}^n p_i(x_i - E(X))^2$$

Die Zahl

$$\sum_{i=1}^n p_i(x_i - E(X))^2$$

nennt man in diesem Zusammenhang das Trägheitsmoment.

Die Varianz einer Summe von zwei Zufallsvariablen ist nur gleich der Summe der Varianzen, wenn X und Y **unabhängig** sind:

$$V(X + Y) = V(X) + V(Y) , \quad X \text{ und } Y \text{ unabhängig} \quad (18)$$

Beachten Sie bitte, dass die analoge Formel für die Standardabweichung **nicht gilt**, denn im allgemeinen ist

$$\sqrt{a+b} \neq \sqrt{a} + \sqrt{b} .$$

Ein konkretes Beispiel möge genügen:

$$5 = \sqrt{25} = \sqrt{3^2 + 4^2} \neq \sqrt{3^2} + \sqrt{4^2} = 3 + 4 = 7$$

Die Formel (18) kann man sofort auf n **unabhängige** Zufallsvariablen X_i verallgemeinern:

$$V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n) \quad (19)$$

Weiter gilt für eine beliebige reelle Zahlen a und b :

$$V(aX + b) = a^2V(X) \quad (20)$$

Die nachfolgende Formel ist unter anderem nützlich, wenn die Varianz von Hand berechnet wird. Wir werden diese Formel selten anwenden.

Beispiel 19 Beweisen Sie die folgende Formel:

$$V(X) = E(X^2) - E(X)^2$$

Beweis: Es gilt:

$$\begin{aligned} V(X) &= E((X - E(X))^2) = E(X^2 - 2E(X)X + E(X)^2) \\ &= E(X^2) - 2E(X)^2 + E(X)^2 = E(X^2) - E(X)^2 \end{aligned}$$

◇

Beispiel 20 Zwei Personen spielen folgendes Glücksspiel: Der Spieler A leistet einen bestimmten Einsatz, würfelt und erhält vom Spieler B :

- 10 Rappen beim Würfeln einer 1 oder 2
- 20 Rappen beim Würfeln einer 3 oder 4
- 40 Rappen beim Würfeln einer 5
- 80 Rappen beim Würfeln einer 6.

Welche durchschnittliche Einnahme pro Spiel kann der Spieler A erwarten? Wie gross ist die Varianz der Einnahme?

Die Lösung wird im Unterricht erarbeitet. \diamond

Es stellt sich die Frage, ob es auch quantitative Aussagen über die Standardabweichung gibt. Dies ist in der Tat der Fall. Für eine beliebige Zufallsvariable mit Mittelwert μ und Standardabweichung σ gelten die folgenden Ungleichungen:

$$\begin{aligned} P(|X - \mu| < \sigma) &\geq 0 \\ P(|X - \mu| < 2\sigma) &\geq \frac{3}{4} = 0.75 \\ P(|X - \mu| < 3\sigma) &\geq \frac{8}{9} \doteq 0.889 \\ P(|X - \mu| < 4\sigma) &\geq \frac{15}{16} \doteq 0.938 \end{aligned}$$

Wenn wir beispielsweise die zweite Ungleichung betrachten, so besagt diese, dass die Wahrscheinlichkeit, dass die Zufallsvariable X einen Wert annimmt, der von μ um weniger als 2σ abweicht, grösser als 0.75 ist. Der Vorteil dieser Ungleichungen besteht darin, dass sie für beliebige Zufallsvariablen gelten. Allerdings sind sie oft sehr grob, wie wir später im Zusammenhang mit der Normalverteilung sehen werden.

Beispiel 21 Im Beispiel 20 haben wir gesehen, dass die Zufallsvariable X den Erwartungswert $\mu = 30$ und die Standardabweichung $\sigma = \sqrt{600} \doteq 24.49$ besitzt. Berechnen Sie $P(|X - \mu| < \sigma)$.

Die Lösung wird im Unterricht erarbeitet. \diamond

2.5 Die Kovarianz*

Wir wollen noch eine Formel für die Varianz der Summe zweier Zufallsvariablen herleiten, die auch gültig ist, wenn X und Y abhängig sind:

$$\begin{aligned} V(X + Y) &= E[((X + Y) - E(X + Y))^2] = E[((X - E(X)) + (Y - E(Y)))^2] \\ &= E[(X - E(X))^2 + 2(X - E(X))(Y - E(Y)) + (Y - E(Y))^2] \\ &= E[(X - E(X))^2] + E[(Y - E(Y))^2] + 2E[(X - E(X))(Y - E(Y))] \\ &= V(X) + V(Y) + 2E[(X - E(X))(Y - E(Y))] \end{aligned}$$

Man bezeichnet den Ausdruck

$$\boxed{\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]}$$

als **Kovarianz** von X und Y . Es gilt also:

$$\boxed{V(X + Y) = V(X) + V(Y) + 2 \text{Cov}(X, Y)}$$

Wir werden die Kovarianz in diesem Modul nicht benötigen. Allerdings werden wir später eine mit der Kovarianz verwandte Grösse studieren. Deshalb vertiefen wir den Begriff ein wenig.

Wir werden zuerst eine explizite Formel für die Kovarianz herleiten. Seien X und Y zwei Zufallsvariablen auf dem Stichprobenraum Ω mit n Elementen. X und Y nehmen dann ebenfalls n Werte an, die nicht alle verschieden sein müssen:

$$X : x_1, x_2, \dots, x_n \quad , \quad Y : y_1, y_2, \dots, y_n$$

Die Kovarianz von X und Y ist dann gegeben durch

$$\text{Cov}(X, Y) = \sum_{i=1}^n \sum_{j=1}^n p_{ij}(x_i - E(X))(y_j - E(Y)) ,$$

wobei

$$p_{ij} := P(X = x_i, Y = y_j) .$$

Es tritt jetzt also eine Doppelsumme auf. Betrachten wir den Fall $n = 3$. Wir müssen 9 Terme addieren, die wir in einem quadratischen Schema anordnen können:

$$\begin{array}{lll} p_{11}(x_1 - E(X))(y_1 - E(Y)) & p_{12}(x_1 - E(X))(y_2 - E(Y)) & p_{13}(x_1 - E(X))(y_3 - E(Y)) \\ p_{21}(x_2 - E(X))(y_1 - E(Y)) & p_{22}(x_2 - E(X))(y_2 - E(Y)) & p_{23}(x_2 - E(X))(y_3 - E(Y)) \\ p_{31}(x_3 - E(X))(y_1 - E(Y)) & p_{32}(x_3 - E(X))(y_2 - E(Y)) & p_{33}(x_3 - E(X))(y_3 - E(Y)) \end{array}$$

Diese 9 Terme können in irgendeiner Reihenfolge aufaddiert werden. Mit der Doppelsumme geschieht dies Zeilenweise.

Wir notieren noch einige Eigenschaften der Kovarianz:

- Wenn X und Y **unabhängig** sind, dann ist $\text{Cov}(X, Y) = 0$. Die Umkehrung ist aber nicht richtig, das heisst, die Kovarianz kann gleich 0 sein und X und Y können dennoch abhängig sein.
- Wenn eine Zunahme von X auch eine Zunahme von Y zur Folge hat, ist die Kovarianz positiv.
- Wenn eine Zunahme von X eine Abnahme von Y zur Folge hat, ist die Kovarianz negativ.
- Der Wert der Kovarianz ist schwierig zu interpretieren. Man erhält eine Grösse, die besser interpretierbar ist, indem man eine Normierung durchführt. Diese besteht darin, dass man durch die Standardabweichungen der x - und der y -Werte dividiert.

$$\rho = \frac{\text{Cov}(X, Y)}{s_X \cdot s_Y}$$

Man bezeichnet diese Grösse als den **Korrelationskoeffizienten**. Man kann zeigen, dass stets $-1 \leq \rho \leq 1$ gilt. Der Korrelationskoeffizient ist ein Mass für die Stärke des linearen Zusammenhangs zwischen X und Y . Wir werden später darauf zurückkommen.

2.6 Binomialverteilung

Es kommt durchaus vor, dass man mit beliebigen Verteilungen arbeitet. Viel häufiger ist aber die Situation, dass man eine bereits bekannte Verteilung verwendet. Wir werden in diesem Abschnitt die wichtigste diskrete Verteilung diskutieren. Es handelt sich um die **Binomialverteilung**.

Das einfachst mögliche Zufallsexperiment besteht aus nur zwei möglichen Ergebnissen. Ein solches Experiment ist auch unter dem Namen Bernoulli²-Experiment bekannt:

Kopf	Zahl
gesund	krank
gewürfelte Zahl ist 6	gewürfelte Zahl ist nicht 6
Test ist negativ	Test ist positiv
1	0

Wie das Beispiel mit dem Würfel zeigt, kann das Zufallsexperiment durchaus einen Stichprobenraum besitzen, der mehr als 2 Elemente umfasst. Weil wir uns aber für die beiden Ereignisse

$$\begin{aligned} A &: \text{die gewürfelte Zahl ist 6} \\ A^c &: \text{die gewürfelte Zahl ist verschieden von 6} \end{aligned}$$

interessieren, liegt ein Bernoulli-Experiment vor.

Wir werden oft die zwei Ergebnisse mit 1 oder 0 codieren. Wir bezeichnen die Wahrscheinlichkeit für das Resultat 1 mit p und jene für das Resultat 0 mit $q := 1 - p$. Beachten Sie, dass p nicht unbedingt gleich 0.5 sein muss!

Wir wiederholen unser Zufallsexperiment n mal. Wir setzen voraus, dass die einzelnen Wiederholungen voneinander **unabhängig** sind, d.h. das Ergebnis eines bestimmten Versuchs soll nicht von den Ergebnissen der vorangegangenen Versuche abhängen. Diese Voraussetzung ist sehr wichtig. Ob man sie in einem konkreten Beispiel als gegeben annehmen darf, muss derjenige entscheiden, der den Versuch durchführt und die Wahrscheinlichkeitsrechnung anwenden will.

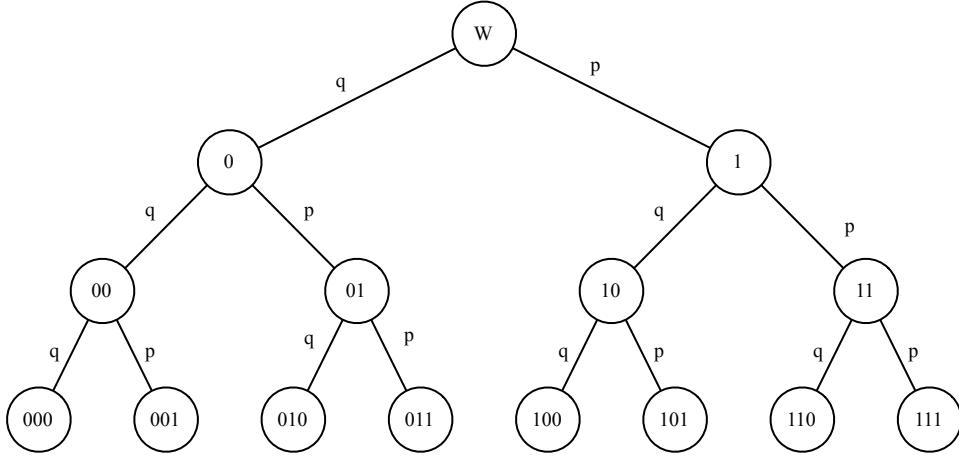
Unseren n Versuchen entspricht eine Folge von n Symbolen 0 und 1. Im Fall $n = 5$ könnten wir z.B.

$$00000, 01000, 01010 \text{ oder } 01100$$

erhalten. Im allgemeinen Fall mit n Versuchen besteht die Folge aus n Symbolen 0 oder 1. Da wir für jedes der n Symbole 2 Möglichkeiten haben, gibt es insgesamt 2^n solche Folgen.

Das nachfolgende Baumdiagramm zeigt für den Fall $n = 3$ alle möglichen 8 Folgen. Für jede Folge erhalten wir die Wahrscheinlichkeit, mit der diese realisiert wird, indem wir die Wahrscheinlichkeiten, mit der die Kanten durchlaufen werden, miteinander multiplizieren.

²Jakob Bernoulli (1654 Basel - 1705 Basel): Sein Werk über die Wahrscheinlichkeitsrechnung, die *Ars Conjectandi*, wurde erst 8 Jahre nach seinem Tod veröffentlicht. Mit dem *Gesetz der grossen Zahlen* machte Jakob Bernoulli einen ganzen wichtigen Beitrag zur Definition der Wahrscheinlichkeit.



Wir interessieren uns für die Anzahl möglichen Erfolge X . Offenbar nimmt X die Werte 0, 1, 2 und 3 an. Wir bestimmen jetzt die Verteilung von X . Der Wert 0 kommt nur durch eine einzige Folge mit der Wahrscheinlichkeit q^3 zustande, der Wert 1 kommt durch 3 Folgen, welche alle die Wahrscheinlichkeit pq^2 haben, zustande, der Wert 2 kommt ebenfalls durch 3 Folgen, welche die Wahrscheinlichkeit p^2q haben, zustande und der Wert 3 kommt nur durch eine einzige Folge mit der Wahrscheinlichkeit p^3 zustande. Wir haben also die folgende Verteilung:

$$\begin{aligned} P(X = 0) &= q^3 \\ P(X = 1) &= 3pq^2 \\ P(X = 2) &= 3p^2q \\ P(X = 3) &= p^3 \end{aligned}$$

Wir betrachten jetzt den allgemeinen Fall mit n Versuchen. Sei X wiederum die Anzahl der Erfolge. Wir interessieren uns für die Wahrscheinlichkeit, dass X gleich k ist mit $0 \leq k \leq n$. Jede Folge mit k Erfolgen hat die gleiche Wahrscheinlichkeit $p^k q^{n-k}$. Das Problem ist die Bestimmung der Anzahl dieser Folgen. Solche Fragestellungen werden in der Kombinatorik behandelt. Man kann zeigen, dass es

$$\binom{n}{k} := \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$$

solche Folgen gibt. Man bezeichnet diese Grösse als **Binomialkoeffizienten**. Sie wird „ n tief k “ ausgesprochen. Sie finden eine Herleitung dieser Formel im Anhang.

Die gesuchte Wahrscheinlichkeit ist also gegeben durch:

$$P(X = k) = \binom{n}{k} p^k q^{n-k} . \quad (21)$$

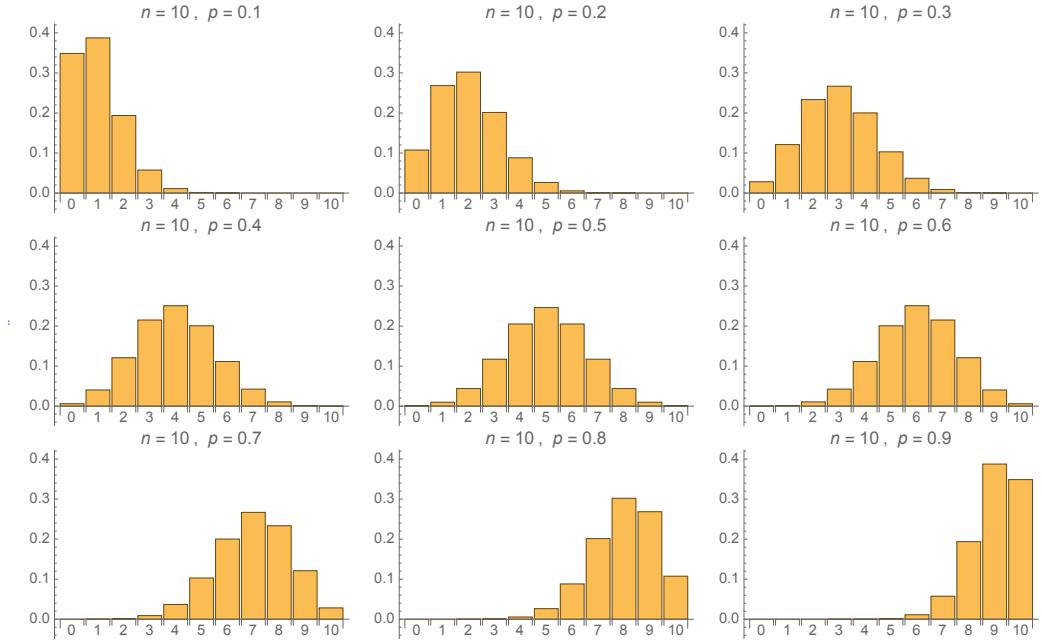
Man bezeichnet diese Verteilung als **Binomialverteilung**. Der Name erklärt sich aus der Tatsache, dass diese Wahrscheinlichkeiten gleich den Termen des Binoms $(p + q)^n$ sind, wenn wir dieses entwickeln:

$$(p + q)^n = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k}$$

Man bezeichnet diese Formel als **binomische Formel**. Sie finden ebenfalls im Anhang Erklärungen dazu.

Die Binomialverteilung hängt von n und p ab, die Parameter der Verteilung heissen. Für jedes $n \in \mathbf{N}$ und p ($0 \leq p \leq 1$) existiert eine eigene Binomialverteilung.

Falls X eine binomialverteilte Zufallsvariable mit den Parametern n und p ist, so drücken wir dies durch folgende Notation aus: $X \sim \mathcal{B}(n, p)$.



Beispiel 22 Eine homogene Münze wird 6 mal geworfen. Berechnen Sie die Wahrscheinlichkeit, dass

- (a) genau zweimal Kopf auftritt;
- (b) mindestens viermal Kopf auftritt;
- (c) nie Kopf auftritt.

Lösung wird im Unterricht erarbeitet. ◇

Wir bestimmen noch den Erwartungswert und die Varianz einer binomialverteilten Zufallsvariable X mit den Parametern n und p . Zu diesem Zweck definieren wir für $i = 1, 2, \dots, n$ die Zufallsvariable

$$X_i = \begin{cases} 1 & \text{falls das Resultat des } i\text{-ten Versuchs 1 ist,} \\ 0 & \text{falls das Resultat des } i\text{-ten Versuchs 0 ist.} \end{cases}$$

Nach Definition gilt:

$$E(X_i) = q \cdot 0 + p \cdot 1 = p$$

und

$$V(X_i) = q \cdot (0 - p)^2 + p(1 - p)^2 = qp^2 + pq^2 = pq(p + q) = pq$$

Wir haben benutzt, dass $p + q = 1$ ist. Da $X = X_1 + X_2 + \dots + X_n$ ist, folgt aus den Formeln (16) und (19):

$$E(X) = E(X_1) + E(X_2) + \dots + E(X_n) = np$$

und

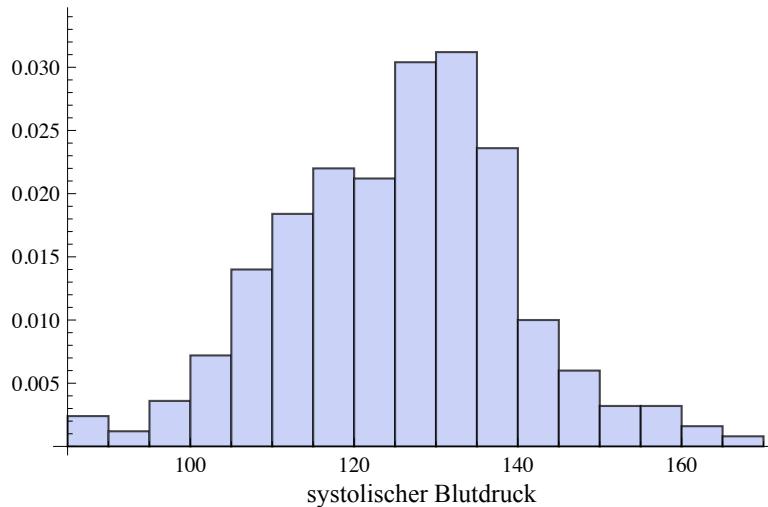
$$V(X) = V(X_1) + V(X_2) + \dots + V(X_n) = npq .$$

Weitere wichtige diskrete Verteilungen sind die Poisson- und die hypergeometrische Verteilung. Wir werden diese bei Bedarf später einführen.

2.7 Stetige Zufallsvariablen

In diesem Abschnitt soll erklärt werden wie sogenannte ***stetige Zufallsvariablen*** in der Wahrscheinlichkeitsrechnung beschrieben werden können. Als Beispiel betrachten wir den systolischen Blutdruck X einer zufällig ausgewählten Person in der Population der gesunden, 25-jährigen Männern. Wenn wir den Blutdruck beliebig genau messen könnten, dann nimmt X eine beliebige reelle Zahl in einem Intervall an. Der Stichprobenraum ist also nicht mehr diskret.

Wir wählen jetzt aus der oben genannten Population $n = 500$ Männer aus und zeichnen ein sogenanntes ***Histogramm***:

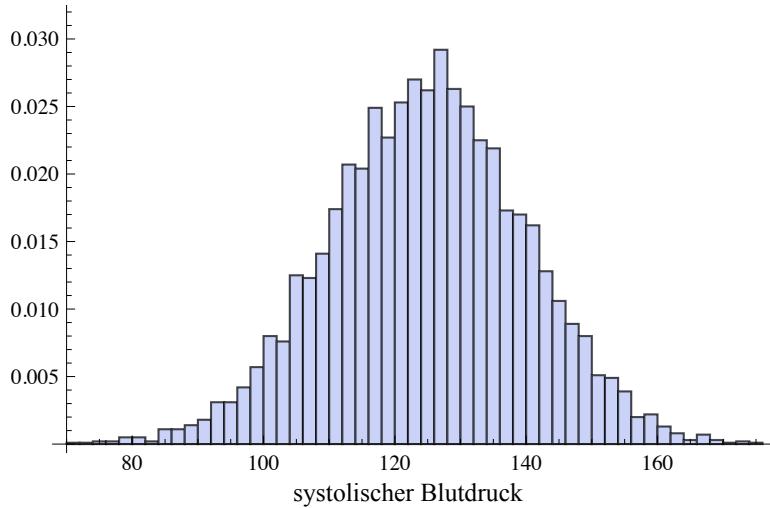


Auf der horizontalen Achse ist der Blutdruck abgetragen. Die Breite der Rechtecke ist gleich 5 mm Hg. Der ***Flächeninhalt*** des i -ten Rechtecks mit den Grenzen x_{i-1} und x_i entspricht dem prozentualen Anteil der Männer, deren Blutdruck X die Ungleichungen

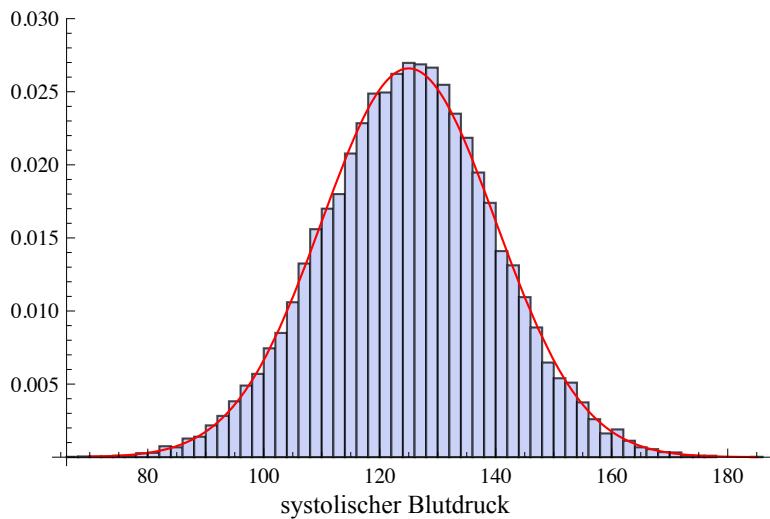
$$x_{i-1} \leq X < x_i$$

erfüllt. Der Flächeninhalt des Rechtecks kann also als empirische Wahrscheinlichkeit gedeutet werden, dass X einen Wert zwischen x_{i-1} und x_i annimmt.

Wenn wir die gleiche Untersuchung bei einer Stichprobe von $n = 5'000$ Männern durchführen, könnte das Histogramm folgendermassen aussehen:



Da die Stichprobe viel grösser ist, kann die Breite der Rechtecke viel kleiner gewählt werden. Das setzt natürlich voraus, dass wir den Blutdruck genauer messen können. Wenn wir schliesslich eine Stichprobe von $n = 20'000$ Männern wählen, könnte das Histogramm so aussehen:



Wenn wir unsere Stichprobe immer grösser wählen und wir den Blutdruck immer genauer bestimmen, dann wird sich das Histogramm immer mehr einer glatten Kurve annähern. Dies führt zum Konzept einer **Dichte**. Eine Funktion $f : \mathbf{R} \rightarrow \mathbf{R}$ heisst Dichte, wenn sie die beiden folgenden Eigenschaften besitzt:

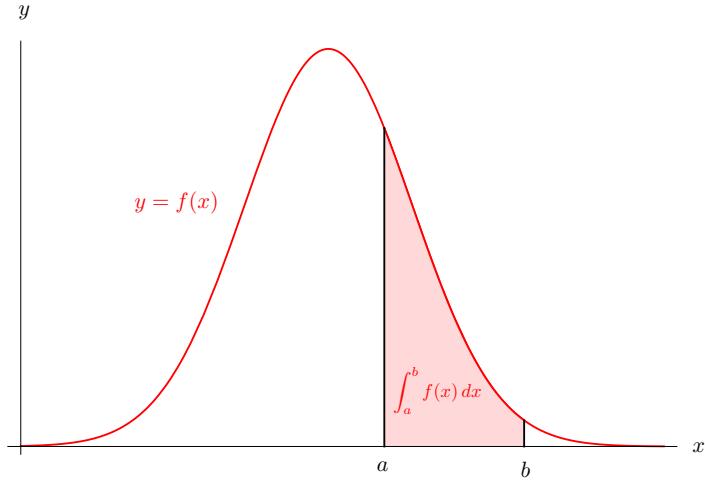
- (i) Die Funktion nimmt keine negativen Werte an: $f(x) \geq 0$ für alle $x \in \mathbf{R}$.
- (ii) Die Fläche zwischen x -Achse und Graph der Funktion ist gleich 1:

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Der Flächeninhalt kann als bestimmtes Integral ausgedrückt werden, gelesen *bestimmtes Integral der Funktion $f(x)$ von minus Unendlich bis plus Unendlich*. Sie werden im

Modul *Mathematik für Medizininformatik* lernen wie ein bestimmtes Integral für eine konkrete Funktion $f(x)$ berechnet wird.

Man bezeichnet Zufallsvariablen, welche eine Dichte besitzen als *stetige* Zufallsvariablen. Seien a und b , $a < b$, zwei gegebene reelle Zahlen. Die Wahrscheinlichkeit, dass eine stetige Zufallsvariable X mit Dichte $f(x)$ einen Wert annimmt, der zwischen a und b liegt, ist gleich dem **Inhalt** der roten Fläche:



Dieser Inhalt ist gleich dem bestimmten Integral der Funktion $f(x)$ von a bis b :

$$P(a < X < b) = \int_a^b f(x) dx \quad (22)$$

Da die beiden Strecken, welche die Fläche an der Stelle $x = a$ oder $x = b$ begrenzen, den Flächeninhalt Null besitzen, sind die folgenden Wahrscheinlichkeiten alle gleich gross:

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

Für eine *stetige* Zufallsvariable spielt es also keine Rolle, ob wir die Grenzen mit einbeziehen oder nicht. Aus dem gleichen Grund macht es bei einer stetigen Zufallsvariablen X mit Dichte f keinen Sinn, danach zu fragen, wie gross die Wahrscheinlichkeit ist, dass X exakt gleich einem Wert c ist. Diese Wahrscheinlichkeit ist nämlich Null, denn der Flächeninhalt der Strecke mit den Endpunkten $(c, 0)$ und $(c, f(c))$ ist gleich 0.

Analog zum Fall einer diskreten Zufallsvariablen ist die (*kumulative*) **Verteilungsfunktion** für stetige Zufallsvariablen definiert

$$F(u) := P(X < u) = \int_{-\infty}^u f(x) dx .$$

Wenn die kumulative Verteilungsfunktion bekannt ist, lässt sich das Integral (22) ganz leicht berechnen. Es ist:

$$P(a < X < b) = F(b) - F(a)$$

Völlig analog zu den diskreten Zufallsvariablen, definiert man den Erwartungswert und die Varianz einer stetigen Zufallsvariablen X mit der Dichte $f(x)$:

$$\begin{aligned} E(X) &:= \int_{-\infty}^{\infty} x f(x) dx \\ V(X) &:= \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx \end{aligned}$$

Lassen Sie sich bitte durch diese Integrale nicht einschüchtern: es handelt sich jeweils einfach um die Fläche zwischen dem Graphen der Funktion, die hinter dem Integralzeichen \int steht, und der x -Achse und zwar von $-\infty$ bis $+\infty$.

Beispiel 23 Bestimmen Sie die Dichte, die kumulative Verteilungsfunktion sowie den Erwartungswert einer Zufallsvariable X , welche im Intervall $[a, b]$ **gleichmässig verteilt** ist.

Lösung wird im Unterricht erarbeitet. \diamond

2.8 Normalverteilung

Wenn Sie die Histogramme mit den systolischen Blutdrücke betrachten, stellen Sie fest, dass die resultierende Dichte einer glockenförmigen Kurve gleicht. Dies ist die **Normalverteilung**. Ihre Dichte ist gegeben durch:

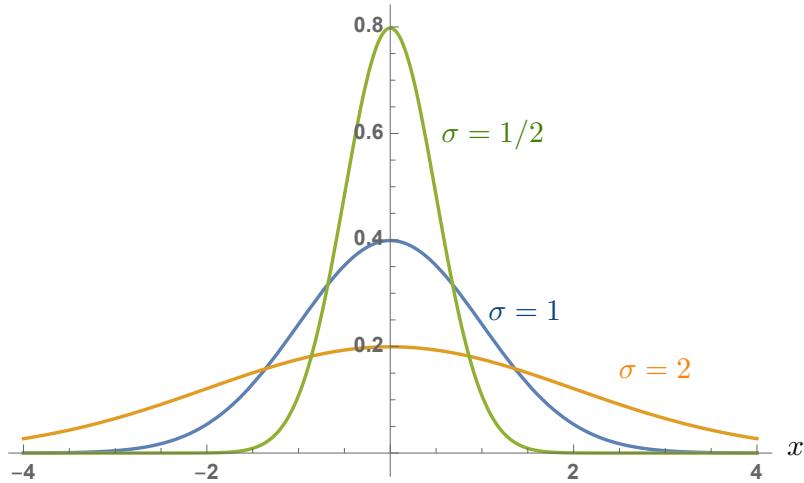
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Hier sind $\mu \in \mathbf{R}$ und $\sigma \in \mathbf{R}^+$ Parameter. Man kann zeigen, dass der Parameter μ gerade der Erwartungswert und der Parameter σ die Standardabweichung ist:

$$\mu = E(X) \quad \text{und} \quad \sigma^2 = V(X)$$

Wenn $\mu = 0$ und $\sigma = 1$ ist, spricht man von der **Standardnormalverteilung**. Wenn X eine normalverteilte Zufallsvariable mit Erwartungswert μ und Varianz σ^2 ist, dann schreiben wir $X \sim \mathcal{N}(\mu, \sigma^2)$. Beachten Sie, dass wir in dieser Schreibweise die Varianz und nicht die Standardabweichung verwenden.

Der Graph von $f(x)$ ist symmetrisch zur vertikalen Geraden $x = \mu$. Die nachfolgende Figur zeigt die Graphen von $f(x)$ für $\mu = 0$ und für $\sigma = 0.5$, $\sigma = 1$ und $\sigma = 2$:

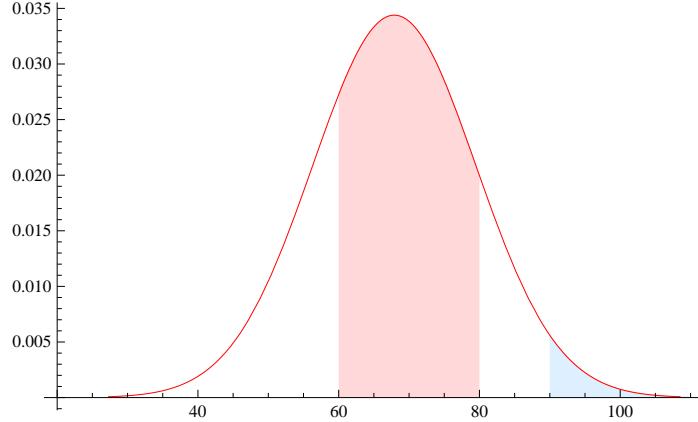


Je grösser σ ist, desto flacher wird die Kurve. Die Fläche zwischen x -Achse und Kurve ist für jede Wahl von μ und σ gleich 1.

Beispiel 24 In den USA wird der Mittelwert des diastolischen Blutdrucks bei 17-jährigen Knaben auf $\mu = 67.9$ mm Hg mit einer Standardabweichung von $\sigma = 11.6$ mm Hg geschätzt. Wir nehmen an, dass der Blutdruck in dieser Population normalverteilt ist.

- (a) Wie gross ist der Anteil der Knaben, deren diastolischer Blutdruck zwischen 60 und 80 mm Hg ist?
- (b) Wenn wir für Bluthochdruck das gleiche Kriterium wie bei Erwachsenen verwenden, nämlich ≥ 90 mm Hg, welcher Prozentsatz der Knaben leidet dann an Bluthochdruck?

Lösung: Gemäss Problemstellung ist $X \sim \mathcal{N}(67.9, 11.6^2)$. Wir müssen $P(60 < X < 80)$ bestimmen. Diese Wahrscheinlichkeit entspricht dem Flächeninhalt der roten Fläche.



Dieser Flächeninhalt ist mathematisch durch das bestimmte Integral

$$P(a < X < b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad (23)$$

Wir werden dieses Integral mit dem Taschenrechner berechnen:

$$P(60 < X < 80) = \frac{1}{11.6 \cdot \sqrt{2\pi}} \int_{60}^{80} e^{-\frac{(x-67.9)^2}{2 \cdot 11.6^2}} dx \doteq 0.604$$

Wir erwähnen nochmals, dass die Wahrscheinlichkeit gleich bleibt, wenn wir die Grenzen 60 und 80 einschliessen

$$P(60 \leq X \leq 80) = P(60 < X < 80).$$

In Teil (b) müssen wir den Flächeninhalt der blauen Fläche berechnen. Dieser ist gegeben durch das Integral:

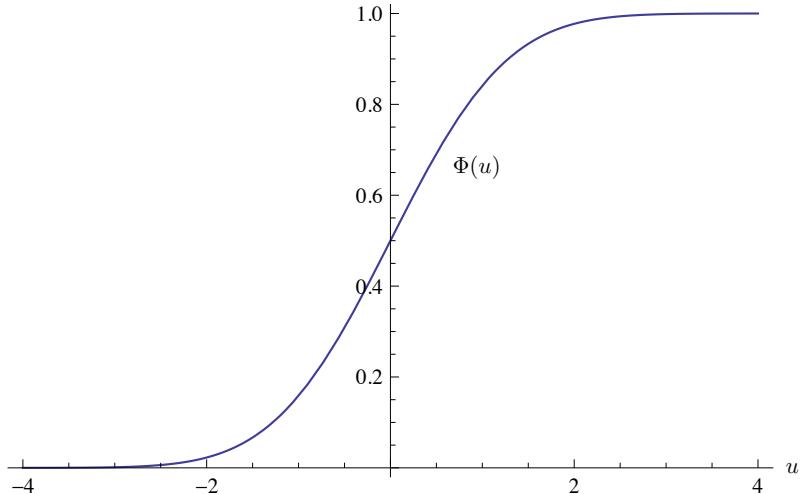
$$P(90 \leq X < \infty) = \frac{1}{11.6 \cdot \sqrt{2\pi}} \int_{90}^{\infty} e^{-\frac{(x-67.9)^2}{2 \cdot 11.6^2}} dx \doteq 0.0284$$

◇

Im Zusammenhang mit der Berechnung des Integrals (23) erwähnen wir die **kumulative Verteilungsfunktion** der Standardnormalverteilung. Diese ist definiert durch

$$\Phi(u) := P(X < u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{x^2}{2}} dx.$$

Sie gibt also die Wahrscheinlichkeit an, dass eine standardnormalverteilte Zufallsvariable X einen Wert annimmt, der kleiner als u ist.



Man kann mit Hilfe einer Variablentransformation zeigen, dass sich das Integral (23) folgendermassen ausdrücken lässt:

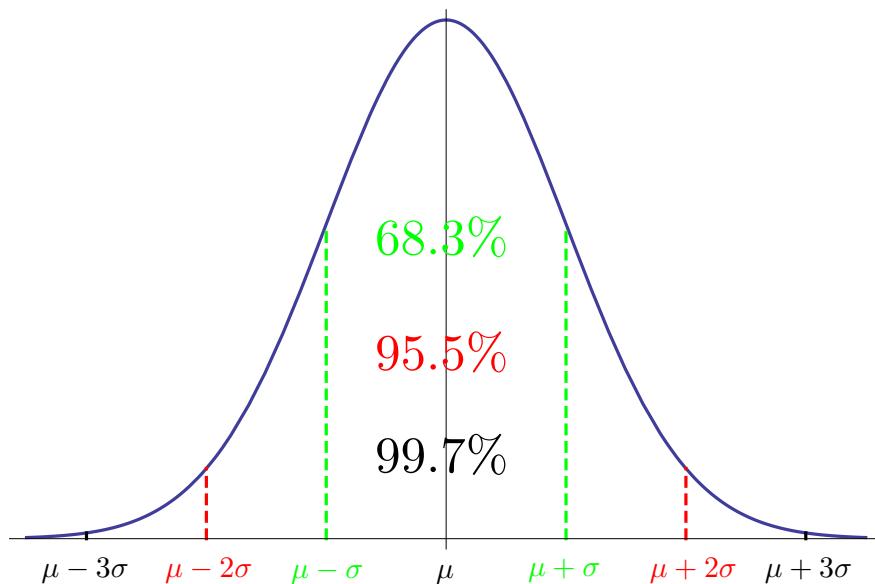
$$\frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Aus diesem Zusammenhang ergeben sich die folgenden wichtigen Wahrscheinlichkeiten:

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &\doteq 68.3\% \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &\doteq 95.5\% \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &\doteq 99.7\% \end{aligned}$$

In Worten: Die Wahrscheinlichkeit, dass eine normalverteilte Zufallsvariable X mit Mittelwert μ und Standardabweichung σ einen Wert annimmt, der vom Mittelwert μ um weniger als

- σ abweicht, ist gleich 68.3%;
- 2σ abweicht, ist gleich 95.5%;
- 3σ abweicht, ist gleich 99.7%.



3 Deskriptive Statistik

Wir verlassen jetzt die Wahrscheinlichkeitsrechnung und wenden uns der **deskriptiven Statistik** zu. In diesem Gebiet geht es um die graphische Darstellung von Daten sowie um deren kompakte Beschreibung durch Masszahlen.

3.1 Daten

Im Zentrum einer statistischen Auswertung stehen **Daten**. Wir betrachten einen realen Datensatz. In einer Studie wurden von 189 Müttern, die eben geboren haben, die folgenden Daten erfasst:

- Alter in Jahren (age)
- Gewicht in kg bei der letzten Menstruation (lwt)
- Ethnische Zugehörigkeit (ethn)
- Raucherverhalten (smoke)
- Gewicht des Neugeborenen in Gramm (bwt)

Nachfolgend finden Sie einen kleinen Auszug dieser Daten:

Nr.	age	lwt	ethn	smoke	bwt	
1	19	82.6	2	0	2523	
2	33	70.3	3	0	2551	
3	20	47.6	1	1	2557	
4	21	49.0	1	1	2594	
5	18	48.5	1	1	2600	

age, *lwt*, *ethn*, *smoke* und *bwt* sind Zufallsvariablen. Die Datei enthält Werte dieser Variablen. In der Statistik bezeichnet man diese Variablen oft als **Merkmale**.

Die verschiedenen Werte, welche die Merkmale annehmen können, werden **Ausprägungen** genannt. Bei den realisierten Werten sprechen wir auch von **Beobachtungen**.

Wir unterscheiden zwischen zwei Typen von Merkmalen:

- (1) **Kategoriale Merkmale**: die Ausprägungen dieser Merkmale sind Kategorien. Beispiele für kategoriale Merkmale sind die Blutgruppe mit den Kategorien A, B, AB und O oder das Geschlecht mit den Kategorien *männlich* und *weiblich*. Im obigen Datensatz sind *ethn* und *smoke* kategoriale Merkmale. Das Merkmal *ethn* hat die Ausprägungen 1 (white), 2 (black) und 3 (other). Das Merkmal *smoke* hat die Ausprägungen 0 (Nichtraucherin) und 1 (Raucherin). Beachten Sie, dass die Zahlen bloss Kodierungen sind. Es handelt sich um kategoriale Daten.
- (2) **Quantitative Merkmale**: die Ausprägungen sind Messwerte (z.B. Blutdruck) oder Anzahlen, also **Zahlen**. Im obigen Datensatz sind *age*, *lwt* und *bwt* quantitative Merkmale.

3.2 Kategoriale Merkmale

Bei den kategorialen Merkmalen unterscheiden wir zwischen den folgenden Untergruppen:

- **nominale Merkmale:** die Kategorien können nicht in eine sinnvolle Reihenfolge gebracht werden. Beispiele sind: Blutgruppe, Geschlecht, ethnische Zugehörigkeit.
- **ordinale Merkmale:** die Kategorien können angeordnet werden. Ein Beispiel hierfür ist die Schmerzintensität mit den Ausprägungen *leicht*, *mittel* und *stark*. Die Ausprägungen können auch durch Zahlen codiert sein. Es macht aber keinen Sinn Differenzen zu betrachten.

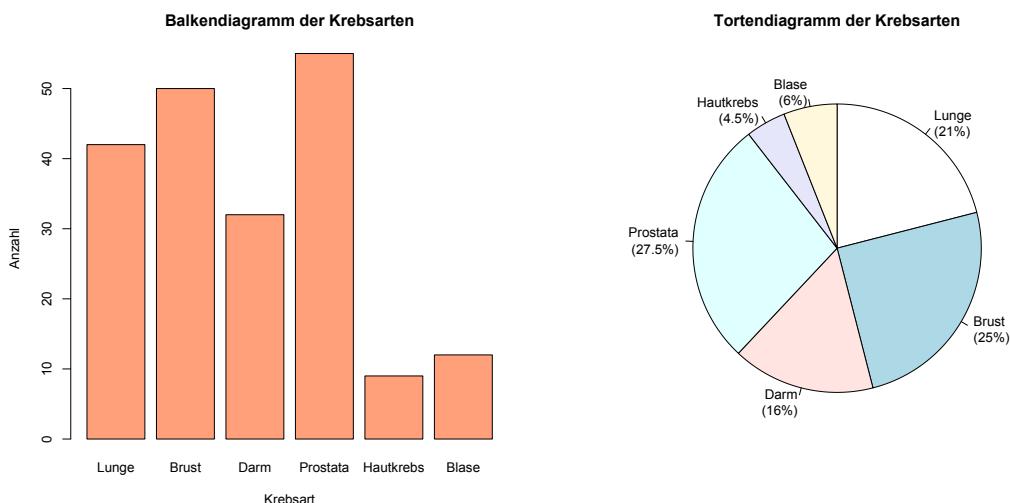
Wir wenden uns jetzt der graphischen Darstellung eines kategoriellen Merkmals zu.

Beispiel 25 In einer Klinik für Krebskranke wurde die Krebsart der letzten 200 Patientinnen und Patienten aufgezeichnet. Es ergab sich das folgende Bild:

Krebsart	Anzahl Fälle	Relative Häufigkeit
Lunge	42	21%
Brust	50	25%
Darm	32	16%
Prostata	55	27.5%
Hautkrebs	9	4.5%
Blase	12	6%
Total	200	100%

Die Daten sollen graphisch Dargestellt werden. Als erste Möglichkeit präsentieren wir ein **Balkendiagramm (bar chart)** (siehe unten links). Im Gegensatz zum Histogramm fügen wir zwischen den Balken Abstände ein, denn die Grösse, die auf der horizontalen Achse abgetragen wird ist kategoriall. Die horizontale Achse ist also keine Zahlengerade. Die Balken können auch horizontal dargestellt werden.

Als zweite Möglichkeit präsentieren wir ein **Tortendiagramm (pie chart)**. (unten rechts). Es ist hier schwieriger die Unterschiede zwischen den einzelnen Kategorien zu erkennen, dafür kann man gut abschätzen, wie gross der Anteil am Total ist.



◇

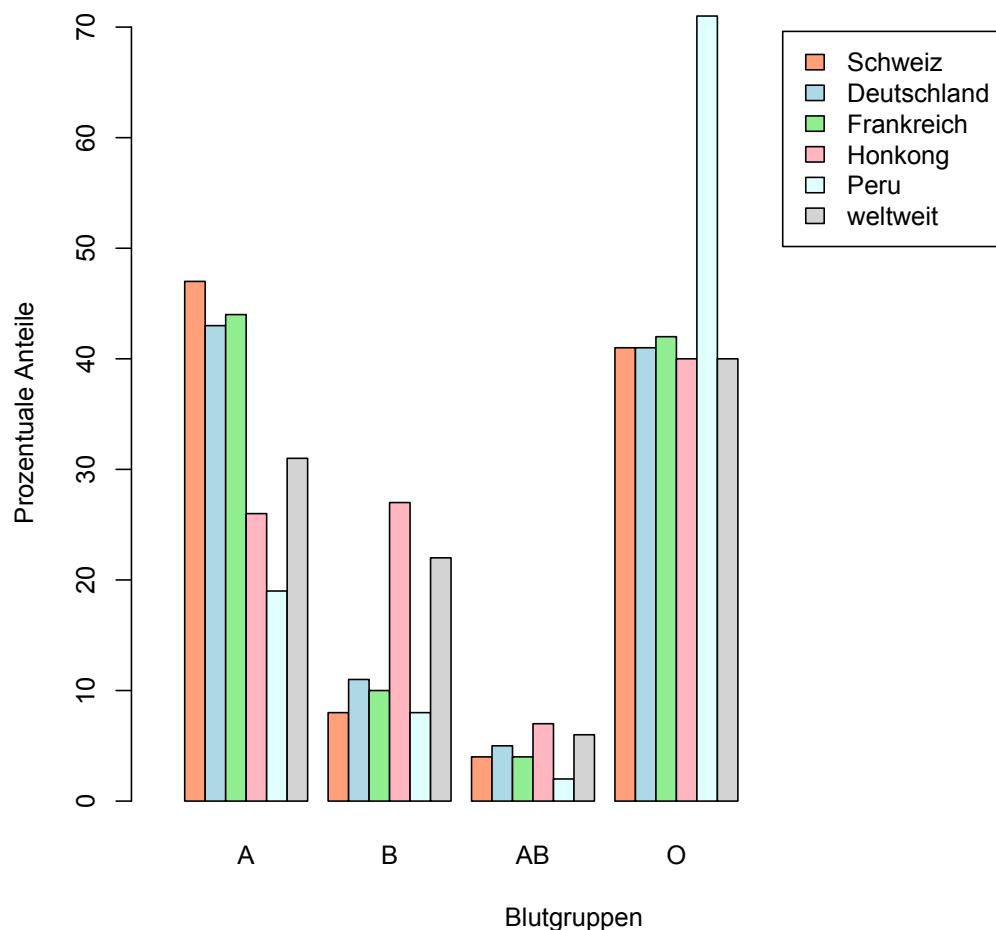
Beispiel 26 In einer frühen Untersuchung über Blutgruppen aus dem Jahre 1939 bestimmten Forscherinnen und Forscher in England die Blutgruppe von 3'696 Personen. Es ergaben sich die folgenden Resultate [14]:

Blutgruppe	Häufigkeit
A	1'634
B	327
AB	119
O	1'616
Total	3'696

Stellen Sie die Daten graphisch dar.

Die Lösung wird im Unterricht erarbeitet. ◇

Wir interessieren uns für die Verteilung der Blutgruppen in verschiedenen Ländern. Mit dem Merkmal *Land* haben wir eine zweite kategoriale Variable. Wir möchten aufzeigen, ob es Unterschiede in der Verteilung zwischen den einzelnen Ländern gibt. Für solche Vergleiche eignet sich ein *gruppiertes Balkendiagramm*:



Die Daten für die Graphik stammen aus Wikipedia. Wir sehen, dass teilweise grosse Unterschiede bestehen.

3.3 Quantitative Merkmale

Wir haben oben definiert, dass **quantitativen Merkmale** Messwerte oder Anzahlen sind. Wir unterscheiden zwischen den folgenden Untergruppen:

- **Diskrete Merkmale:** die Ausprägungen sind natürliche Zahlen. In unserem Datensatz ist *age* ein diskretes Merkmal.
- **Stetige Merkmale:** theoretisch sind alle Werte in einem bestimmten Intervall möglich. Z.B. Temperatur, Blutdruck und Körperlänge. Stetige Merkmale sind eine theoretische Konstruktion, die in der Praxis so nicht vorkommen. Da die Messgenauigkeit endlich ist, sind nämlich alle gemessenen Größen gerundet und damit eigentlich diskret. Die Behandlung als stetige Größe liefert eine mathematische Vereinfachung. Im Datensatz (24) sind *lwt* und *bwt* stetige Merkmale. Das Gewicht des Kindes *bwt* ist zwar immer eine natürliche Zahl, denn das Gewicht wird auf Gramm genau gemessen und wird in dieser Einheit angegeben. Trotzdem ist es sinnvoller *bwt* als stetige Variable zu betrachten.

Quantitative Merkmale können noch feiner unterteilt werden:

- **Intervallskaliertes Merkmal:** im Gegensatz zu ordinalskalierten Merkmalen können Differenzen sinnvoll interpretiert werden. Aus diesem Grund können bei solchen Variablen auch Mittelwerte berechnet werden. Intervallskalierte Merkmale brauchen aber keinen absoluten Nullpunkt zu besitzen. Ein typisches Beispiel ist die Temperaturangabe in °C. Wenn es gestern 10°C warm war und heute 20°C, dann ist es heute nicht doppelt so warm wie gestern!
- **Absolut-, verhältnis- oder rationalskalierte Merkmale:** Bei solchen Merkmalen existiert ein absoluter Nullpunkt. Aus diesem Grund können Verhältnisse sinnvoll interpretiert werden. Beispiele sind das Alter, das Gewicht oder die Länge. Eine bestimmte Person kann beispielsweise doppelt so schwer sein wie eine andere Person.

Die folgenden Daten stammen aus [8]. Es handelt sich um die Hämoglobinwerte in [g/100 ml] von 70 Patientinnen und Patienten:

10.2	13.7	10.4	14.9	11.5	12.0	11.0
13.3	12.9	12.1	9.4	13.2	10.8	11.7
10.6	10.5	13.7	11.8	14.1	10.3	13.6
12.1	12.9	11.4	12.7	10.6	11.4	11.9
9.3	13.5	14.6	11.2	11.7	10.9	10.4
12.0	12.9	11.1	8.8	10.2	11.6	12.5
13.4	12.1	10.9	11.3	14.7	10.8	13.3
11.9	11.4	12.5	13.0	11.6	13.1	9.7
11.2	15.1	10.7	12.9	13.4	12.3	11.0
14.6	11.1	13.5	10.9	13.1	11.8	12.2

Obwohl gewisse Werte mehrmals vorkommen, ist es nicht sinnvoll ihre Häufigkeit zu erfassen, denn wir würden in den meisten Fällen nur 1 erhalten. Gewisse Werte, wie etwa 8.9, 9.0, 9.1 oder 9.2 kommen überhaupt nicht vor. Das zugehörige Balkendiagramm wäre überhaupt nicht aussagekräftig. Wir müssen in diesem Fall **Klassen** bilden. Wir bezeichnen mit *X* den Hämoglobinwert und halten fest, wie viele Werte sich jeweils im

entsprechenden Intervall befinden:

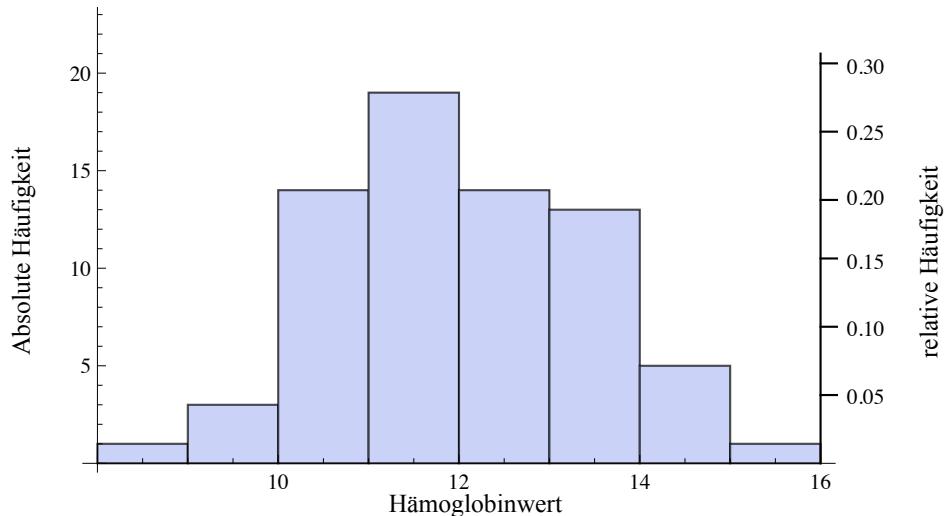
Klassen		absolute Häufigkeit	relative Häufigkeit
8 $\leq X < 9$:	1	0.014
9 $\leq X < 10$:	3	0.043
10 $\leq X < 11$:	14	0.200
11 $\leq X < 12$:	19	0.271
12 $\leq X < 13$:	14	0.200
13 $\leq X < 14$:	13	0.186
14 $\leq X < 15$:	5	0.071
15 $\leq X < 16$:	1	0.014

Wir haben hier als Klassenbreite den Wert 1 gewählt. Wenn man die Graphik mit einem Statistikprogramm zeichnen lässt, wird die die Anzahl der verwendeten Klassen und damit die Klassenbreite automatisch bestimmt. Das Statistikprogramm R verwendet für die Bestimmung der Anzahl Klassen die Formel von Sturges:

$$K = \lceil 1 + \log_2 n \rceil ,$$

wobei n die Anzahl der Werte und K die Anzahl der Klassen ist. Mit $\lceil x \rceil$ bezeichnet man die kleinste ganze Zahl, die grösser oder gleich x ist. Die Regel von Sturges wird von einer Binomialverteilung abgeleitet und liefert nur gute Resultate, wenn die Daten näherungsweise normalverteilt sind.

Die nachfolgende Graphik ist ein sogenanntes **Histogramm**. Über den Klassen auf der horizontalen Achse werden die Häufigkeiten abgetragen. Die linke Achse enthält die Skala für absoluten Häufigkeiten, die Achse auf der rechten Seite jene für die relativen Häufigkeiten.

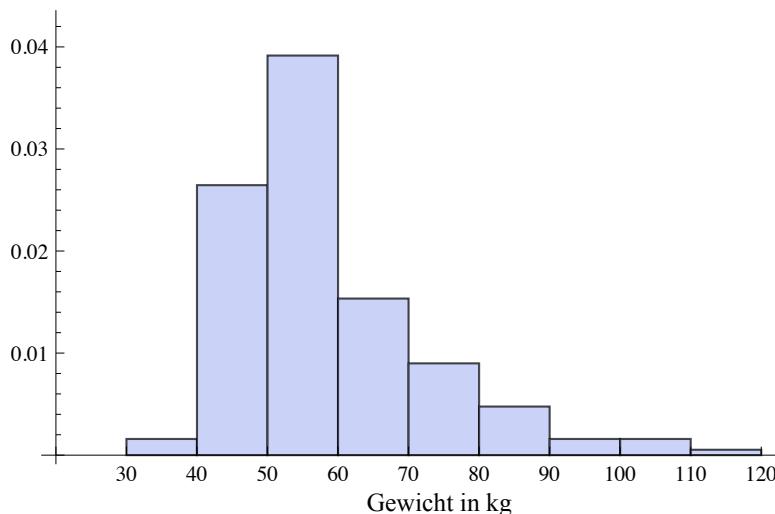


Als zweites Beispiel betrachten wir das Gewicht *lwt* (in kg) der Mütter im Datensatz

(24). Wenn wir als Klassenbreite 10 verwenden, erhalten wir:

Klasse	absolute Häufigkeit	relative Häufigkeit
$30 \leq lwt < 40$	3	1.6%
$40 \leq lwt < 50$	50	26.5%
$50 \leq lwt < 60$	74	39.2%
$60 \leq lwt < 70$	29	15.3%
$70 \leq lwt < 80$	17	9.0%
$80 \leq lwt < 90$	9	4.8%
$90 \leq lwt < 100$	3	1.6%
$100 \leq lwt < 110$	3	1.6%
$110 \leq lwt < 120$	1	0.5%
Total	189	100%

Im nachfolgenden Histogramm wird die Höhe jeden Rechtecks so gewählt, dass sein Flächeninhalt der relativen Häufigkeit der Daten, die in das entsprechende Intervall fallen, entspricht. Man erhält so ein **Dichte-Histogramm**:



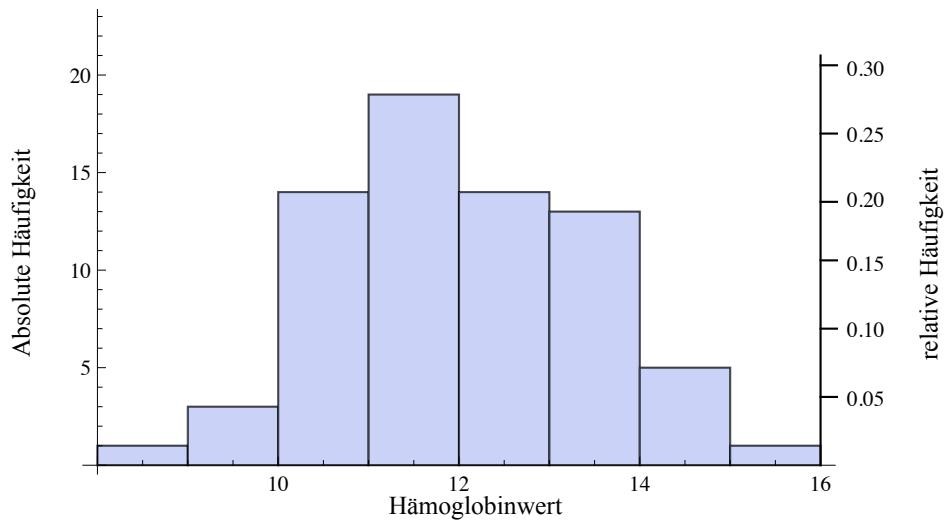
Beispiel 27 Kreatin-Kinase (CK) ist ein Enzym, das man in Muskelzellen und im Gehirn findet. Um die natürliche Variation der CK-Konzentration zu studieren, wurde 36 männlichen Freiwilligen Blut entnommen. Es ergaben sich die folgenden Konzentrationen (Einheit U/l):

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

Gesucht wird das Histogramm, wobei eine Klassenbreite von 20 verwendet werden soll.

Die Lösung wird im Unterricht erarbeitet. \diamond

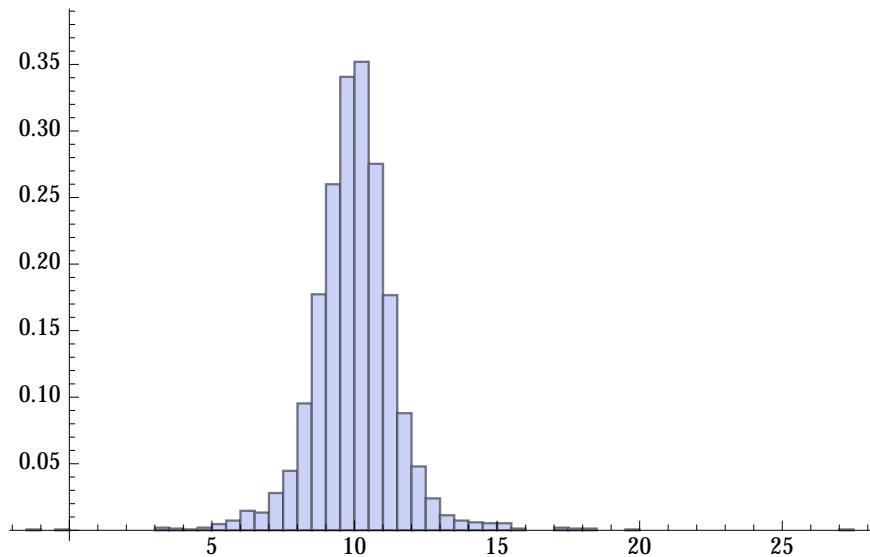
Bei einem Dichte-Histogramm ist die Fläche jedes Rechtecks gleich der Wahrscheinlichkeit, dass die Zufallsvariable einen Wert in der entsprechenden Klasse annimmt. Betrachten wir nochmals das Beispiel mit den Hämoglobinwerten:



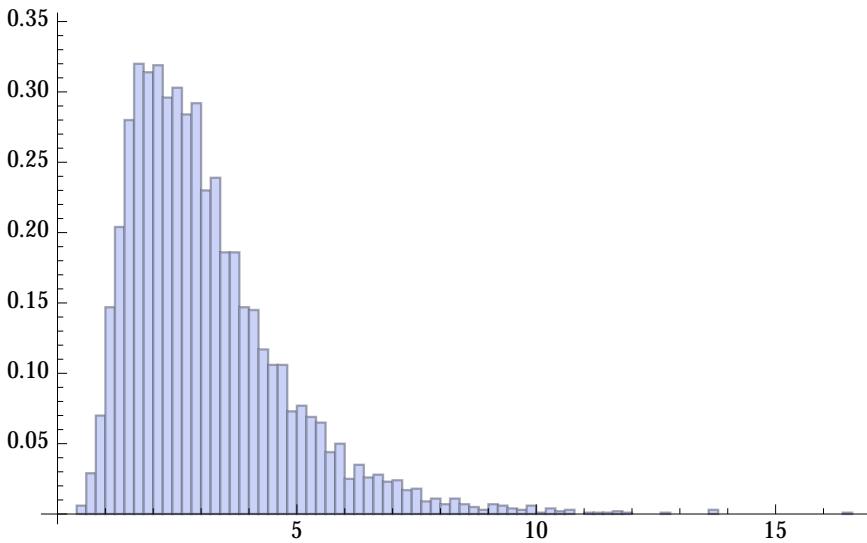
Da die Klassenbreite gleich 1 ist, erhalten wir bei Verwendung der Skala mit den relativen Häufigkeiten gerade das Dichte-Histogramm:

3.4 Formen der Histogramme

Ein Histogramm kann *symmetrisch* sein

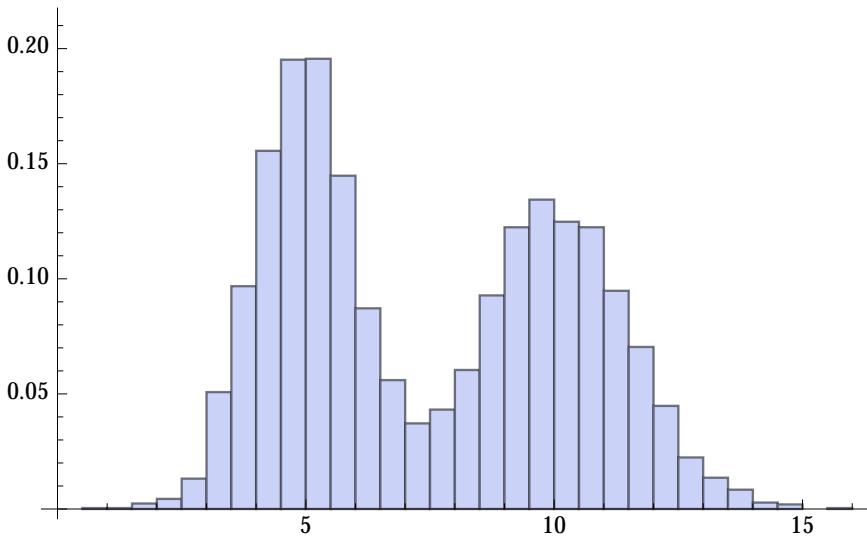


Eine exakte Symmetrie wird sich natürlich selten einstellen. Man spricht bereits von einem *symmetrische* Histogramm, wenn es näherungsweise symmetrisch ist. Wenn das Histogramm nicht symmetrisch ist, dann ist es entweder *linksschief* (left-skewed) oder *rechtsschief* (right-skewed). Das nachfolgende Histogramm ist rechtsschief. Es geht rechts weniger schnell gegen Null als links.



Wir erhalten ein symmetrisches Histogramm für die Körpergrösse. Diese ist sogar normalverteilt. Histogramme von Laborwerten wie Hämoglobin oder Cholesterin sind rechtsschief. Linksschiefe Histogramme sind in der Biostatistik selten (siehe [5]). Ein Beispiel für ein linksschiefes Histogramm ist die Anzahl der Schwangerschaftswochen bis zur Geburt. Einige Kinder kommen sehr früh zur Welt, der Hauptanteil liegt aber zwischen Woche 37 und 41.

Weiter kann ein Histogramm nur einen Gipfel besitzen (unimodal) oder 2 und mehr Gipfel besitzen. Bei 2 Gipfeln spricht man von einer bimodalen Verteilung.



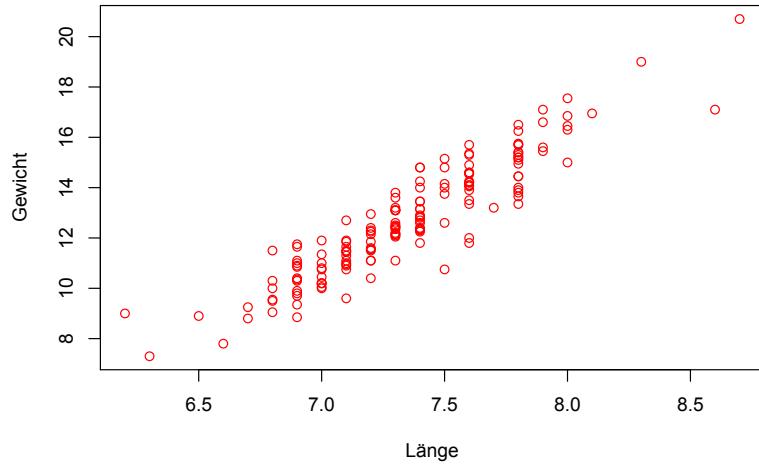
Die gleichen Begriffe werden auch bei Dichten verwendet.

3.5 Multivariate Daten

Häufig werden gleichzeitig zwei oder mehr Variablen gemessen. So werden beispielsweise bei einem Patienten gleichzeitig Gewicht, Grösse und Alter bestimmt. Solche Daten werden als ***multivariate Daten*** bezeichnet.

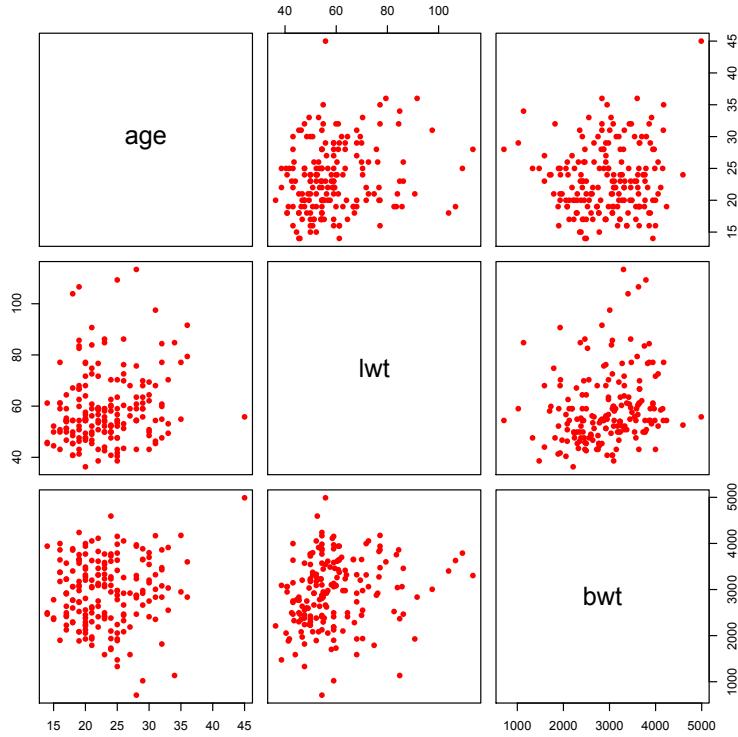
Von 162 Krabben werden Länge und Gewicht bestimmt. Wir können für jede Krabbe die Länge als x - und das Gewicht als y -Koordinate interpretieren und zu einem Punkt

(x, y) zusammenfassen, den wir in einem Koordinatensystem darstellen. Wir erhalten so ein **Streudiagramm** oder **Scatterplot**.

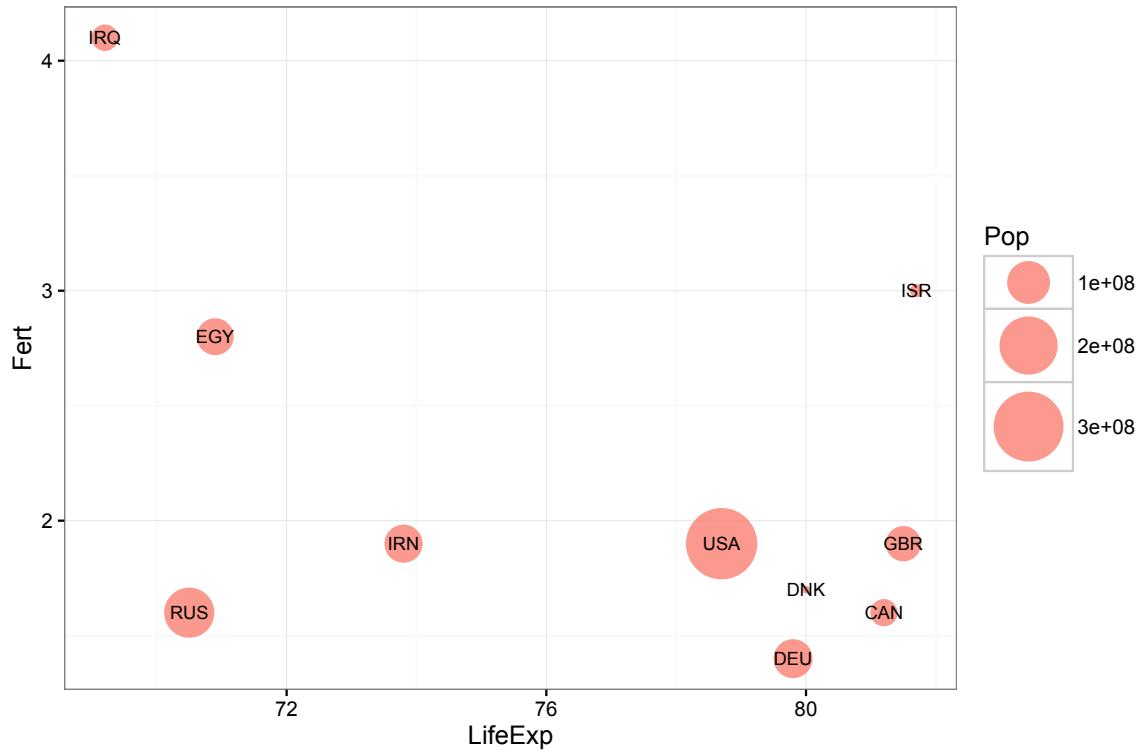


Wir erkennen, dass zwischen der Länge und dem Gewicht ein linearer Zusammenhang besteht. Wir kommen im Kapitel über die lineare Regression darauf zurück.

Betrachten wir wieder den Datensatz (24). Wir interessieren uns für die Merkmale age (Alter), lwt, Gewicht der Mutter in kg) und bwt (Gewicht des Neugeborenen in g). Mit der Funktion `pairs` werden in R alle Paarungen der Merkmale als Scatterplot gezeichnet:



Bubble Plots erlauben die Darstellung von 3 Merkmalen. Im nachfolgenden Plot wird für einige Länder auf der horizontalen Achse die Lebenserwartung in Jahren und auf der vertikalen Achse die Fertilitätsrate (= Anzahl Kinder pro Frau) angegeben. Die Grösse der Bubbles gibt die Bevölkerungsgrösse wieder.



4 Statistische Masszahlen

Dank den statistischen Masszahlen können wir Daten auf kompakte Art und Weise beschreiben.

4.1 Lagemasse

4.1.1 Mittelwert

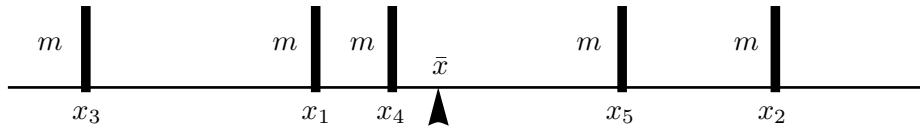
Wir betrachten n Beobachtungen x_i ($1 \leq i \leq n$). Diese n Beobachtungen bilden eine **Stichprobe**. Die natürliche Zahl n heisst **Umfang** der Stichprobe. Unter dem **Stichprobenmittel (sample mean)** oder kurz **Mittelwert** \bar{x} versteht man das arithmetische Mittel der Stichprobenwerte:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (25)$$

Im Fall der Hämoglobinwerte auf Seite 39 war $n = 70$. Wir erhalten für das Stichprobenmittel:

$$\bar{x} = 11.9$$

Wir geben noch eine nützliche physikalische Interpretation des Mittelwerts. Auf der als masselos angesehenen Zahlengerade sind an den Stellen x_i Klötze mit der Masse m kg aufgestellt. Auf diesen Stellen wirkt dann auf die Zahlengerade eine Gewichtskraft von $m \cdot g$ Newton, wobei $g = 9.8 \text{ m/s}^2$ die Erdbeschleunigung ist.



Wir stellen uns die Frage, wo wir die Zahlengerade unterstützen müssen, damit sie im Gleichgewicht bleibt. Wir bezeichnen die Koordinate des gesuchten Punktes mit x . Aus der Physik ist bekannt, dass die Summe der Drehmomente gleich Null sein muss:

$$\sum_{i=1}^n mg(x - x_i) = 0 \implies \sum_{i=1}^n (x - x_i) = 0 \implies nx - \sum_{i=1}^n x_i = 0 \implies x = \frac{1}{n} \sum_{i=1}^n x_i$$

Offenbar muss die Zahlengerade an der Stelle \bar{x} unterstützt werden.

Der Mittelwert (25) wird für uns die wichtigste Masszahl zur Beschreibung der mittleren Lage sein. Er hat jedoch einen Nachteil: er ist empfindlich gegenüber extremen Werten, sogenannten **Ausreissern (outliers)**. Angenommen in einem Dorf leben 67 steuerpflichtige Einwohner mit einem durchschnittlichen Jahreseinkommen von 88'000 Fr. Nun zieht ein CEO einer Grossunternehmung ins Dorf mit einem jährlichen Einkommen von 20 Mio. Fr. Das durchschnittliche Einkommen der steuerpflichtigen Personen beträgt neu:

$$\bar{x} = \frac{67 \cdot 88'000 + 20'000'000}{68} \doteq 380'824 \text{ Fr.}$$

Dass das obige Beispiel nicht frei erfunden ist, zeigt ein Artikel aus der NZZ vom 16. April 2011. Gemäss diesem Artikel ist die Gemeinde Vaux-sur-Morges die reichste Gemeinde der Schweiz. Die Gemeinde hat 170 Einwohner und pro Kopf werden Fr. 338'779

an Steuern bezahlt. 90% der Steuereinnahmen stammen aber von einer einzigen Person, dem Roche-Erben André Hoffmann. Ohne André Hoffmann würden die durchschnittlichen Steuereinnahmen pro Kopf bei Fr. 34'078 liegen.

4.1.2 Median

Eine zweite Masszahl für die mittlere Lage ist der **Median** \tilde{x} . Wir denken uns die n beobachteten Werte **aufsteigend der Grösse nach geordnet**:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)}$$

Beachten Sie, dass die Indizes in runden Klammern gesetzt sind. Damit will man ausdrücken, dass die Werte in aufsteigender Grösse angeordnet sind. Wir müssen jetzt zwei Fälle unterscheiden. Sei n ungerade, z.B. $n = 7$. Dann ist $n+1 = 8$ gerade und $\frac{n+1}{2} = 4$ ist der mittlere Wert in der sortierten Liste:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)} \leq x_{(5)} \leq x_{(6)} \leq x_{(7)}$$

Wir setzen

$$\tilde{x} = x_{((n+1)/2)} .$$

Falls n gerade ist, z.B. $n = 8$,

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)} \leq x_{(5)} \leq x_{(6)} \leq x_{(7)} \leq x_{(8)} ,$$

definiert man den Median als arithmetisches Mittel von $x_{(n/2)} = x_{(4)}$ und $x_{(n/2+1)} = x_{(5)}$:

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Dies ergibt die folgende allgemeine Definition:

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & \text{falls } n \text{ ungerade;} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{falls } n \text{ gerade.} \end{cases}$$

Der Median ist unempfindlich gegen einzelne extreme Werte, denn nur der mittlere Wert oder die beiden mittleren Werte werden für seine Berechnung verwendet.

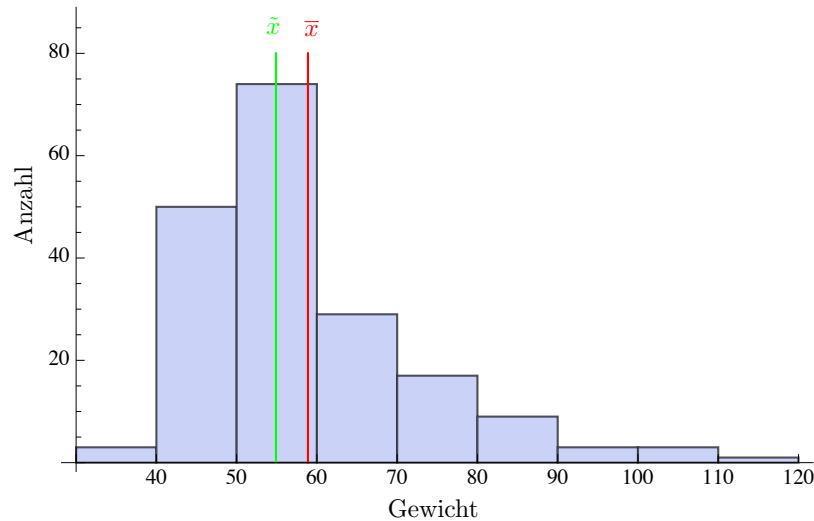
MITTELWERT ODER MEDIAN?

Falls das Histogramm der Werte der Stichprobe symmetrisch ist, stimmen Median und Mittelwert überein. Bei der Interpretation mit der Zahlengerade sind dann die Klötze symmetrisch bezüglich des Mittelwerts verteilt.

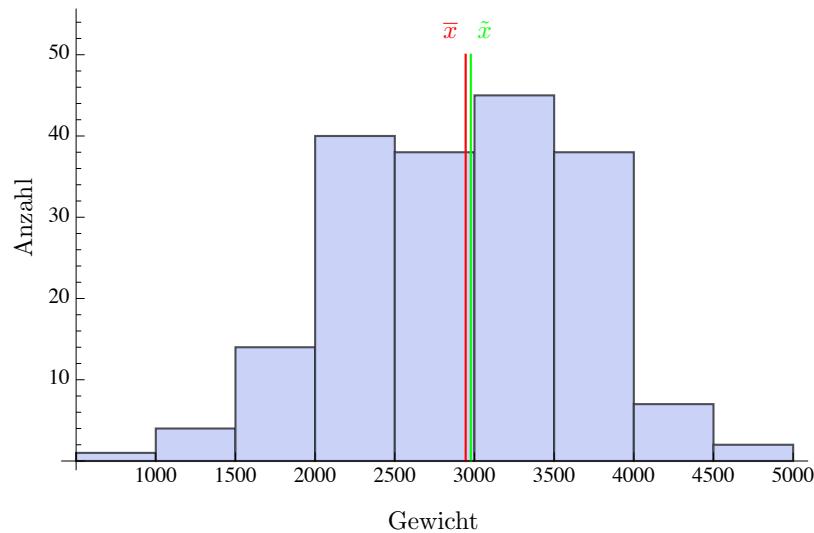
Wenn das Histogramm asymmetrisch ist, dann kann man nicht generell sagen, die eine Kennzahl sei besser als die andere. Es kommt darauf an, was man damit machen will. Betrachten wir als Beispiel die Verteilung der Einkommen in einer Stadt. Da das Einkommen durch Null nach unten beschränkt ist, es aber andererseits sehr hohe Einkommen geben kann, ist die Verteilung meistens rechtsschief. Das mittlere Einkommen kann für die Steuerbehörde interessant sein, da sie durch Multiplikation mit der Anzahl Einwohner das totale Einkommen der Stadt und damit die totalen Steuereinnahmen abschätzen kann. Für die einzelne Steuerzahlerin bzw. den einzelnen Steuerzahler ist dieses mittlere Einkommen aber meistens zu hoch, denn der Mittelwert ist empfindlich gegenüber Ausreisern. Für diese Person ist es interessanter zu wissen, ob ihr Einkommen in der oberen oder unteren Hälfte liegt.

Ein weitere Aspekt ist theoretischer Natur. Der Mittelwert hat interessante mathematische Eigenschaften, welche den Aufbau einer mathematischen Theorie ermöglichen. Das ist für den Median weniger der Fall.

Die nachfolgende Graphik zeigt das Histogramm des Gewichts der 189 Mütter aus dem Datensatz (24). Da das Histogramm rechtsschief ist, ist der Mittelwert $\bar{x} = 58.9$ grösser als der Median $\tilde{x} = 54.9$:



Das Histogramm des Gewichts (in Gramm) der Neugeborenen ist näherungsweise symmetrisch. Mittelwert $\bar{x} = 2'945$ und Median $\tilde{x} = 2'977$ sind fast gleich:



Beispiel 28 In einer Studie über den Einfluss verschiedener Risiken auf die Sterblichkeit von Mäusen wurde eine Gruppe von Labormäusen mit einer Strahlendosis von 300 rad bestrahlt. Dann teilte man die Mäuse in 2 Gruppen ein: die erste Gruppe wurde in einer keimfreien Umgebung gehalten und die zweite Gruppe in normaler Laborumgebung. Man beobachtete dann die Überlebenszeit in Tagen. Bei den beiden folgenden Diagrammen handelt es sich um *Stengel-Blatt-Diagramme*. Die Überlebenszeiten sind Zahlen im Hunderter-Bereich. Man spaltet die erste Ziffer (Hunderterziffer) ab und notiert sie links vom vertikalen Strich. Sie bilden den Stengel. Rechts davon notiert man

die dazugehörigen Zehner- und Einerstellen. Diese bilden die Blätter. Man spart so einerseits Schreibarbeit und erhält andererseits bereits ein Histogramm.

Keimfreie Umgebung										Normale Umgebung							
1	58	92	93	94	95					1	59	89	91	98			
2	02	12	15	29	30	37	40	44	47	2	35	45	50	56	61	65	66
3	01	01	21	37						3	43	56	83				
4	15	34	44	85	96					4	03	14	28	32			
5	29	37															
6	24																
7	07																
8	00																

Bestimmen Sie den Mittelwert und den Median der beiden Stichproben.

Die Lösung wird im Unterricht erarbeitet. \diamond

4.1.3 Modus

Bei nominalen Merkmalen (Farbe, ethnische Zugehörigkeit, etc.) kann man keinen Mittelwert bilden. Man kann dann die Frage nach der Merkmalsausprägung mit der grössten Häufigkeit stellen. Der **Modalwert** ist die Ausprägung mit der grössten Häufigkeit. Betrachten wir als Beispiel die ethnische Zugehörigkeit beim Datensatz (24):

white: 96, black: 26, other: 67

Der Modalwert ist hier *white*.

Der Modalwert ist auch für quantitative Merkmale definiert.

- Diskrete Merkmale: häufigster Wert
- Stetige Merkmale mit Klasseneinteilung: Mitte der häufigsten Klasse

4.1.4 Quantile

Wir betrachten eine Verallgemeinerung des Medians. Seien

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}$$

n Beobachtungen und sei p eine Zahl zwischen 0 und 1. Das p -Quantil, q_p , teilt die n Beobachtungen in 2 Gruppen:

- $p \cdot 100\%$ der Beobachtungen sind kleiner als das Quantil;
- $(1 - p) \cdot 100\%$ der Beobachtungen sind grösser als das Quantil.

Die genaue Formel lautet folgendermassen:

$$q_p = \begin{cases} \frac{1}{2}(x_{(n \cdot p)} + x_{(n \cdot p + 1)}) & \text{falls } n \cdot p \text{ ganzzahlig ist,} \\ x_{(\lceil n \cdot p \rceil)} & \text{falls } n \cdot p \text{ nicht ganzzahlig ist.} \end{cases}$$

Hier ist $\lceil n \cdot p \rceil$ die kleinste ganze Zahl grösser oder gleich $n \cdot p$.

Beachten Sie, dass es verschiedene Definitionen für die Quantile gibt. Aus diesem Grund können verschiedene Statistik-Programme für die gleiche Stichprobe auch unterschiedliche Resultate liefern.

Beispiel 29 Wir betrachten die $n = 20$ der Grösse nach geordneten Beobachtungen

$$1, 1, 2, 6, 6, 7, 13, 14, 15, 21, 24, 29, 34, 34, 36, 38, 42, 46, 48, 49$$

- Berechnung von $q_{1/4}$: Es ist $n \cdot p = 20 \cdot \frac{1}{4} = 5$, also ganzzahlig. Damit gilt:

$$q_{1/4} = \frac{1}{2}(x_{(5)} + x_{(6)}) = \frac{1}{2}(6 + 7) = 6.5$$

- Berechnung von $q_{1/3}$: Es ist $n \cdot p = 20 \cdot \frac{1}{3} = 6.\bar{6}$. Die nächstgrössere ganze Zahl ist 7. Also:

$$q_{1/3} = x_{(7)} = 13$$

◇

Beachten Sie die folgenden Spezialfälle:

- Das 0.5-Quantil, $q_{0.5}$, ist gleich dem Median.
- Das 0.25- und das 0.75-Quantil, $q_{0.25}$ und $q_{0.75}$, heissen **unteres** bzw. **oberes Quartil**. Das untere Quartil wird auch als **1. Quartil** und das obere Quartil als **3. Quartil** bezeichnet. Der Bereich zwischen dem unteren und dem oberen Quartil umfasst also 50% der Beobachtungen.

4.2 Masszahlen für die Streuung

Wir interessieren uns jetzt für Masszahlen, mit denen man die Spannweite oder die Variabilität einer Stichprobe beschreiben kann.

4.2.1 Varianz und Standardabweichung

Wir betrachten wiederum eine Stichprobe mit n Werten x_1, x_2, \dots, x_n . Man könnte die Variabilität der Stichprobe messen, indem man die Beträge der Abweichungen der Werte zum Mittelwert aufaddiert und dann durch n teilt:

$$\frac{1}{n} [|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|]$$

Allerdings hat dieses Mass keine guten mathematischen Eigenschaften. Es ist viel besser, wenn wir die Quadrate der Abweichungen aufaddieren und dann durch $n-1$ teilen. Man erhält so die **empirische Varianz** s^2 der Stichprobe:

$$s^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2] = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Die Gründe, wieso wir durch $n-1$ und nicht durch n teilen, sind theoretischer Natur und werden später erklärt werden. Für grosses n ist der Unterschied sowieso unbedeutend.

Wenn die Werte der Stichproben Längen sind, die in cm gemessen werden, dann hat die Varianz die Masseinheit cm^2 . Um zu einer Grösse zu kommen, welche die gleiche Masseinheit wie die Werte der Stichprobe besitzt, ziehen wir die Wurzel aus der Varianz. Wir erhalten so die **empirische Standardabweichung** der Stichprobe:

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

Beispiel 30 Wir betrachten die Stichprobe

$$\{11.4, 8.3, 9.4, 19.3, 1.0, 16.3, 16.4\}.$$

Der Mittelwert ist gegeben durch

$$\bar{x} = \frac{11.4 + 8.3 + 9.4 + 19.3 + 1.0 + 16.3 + 16.4}{7} \doteq 11.7$$

und die Varianz

$$\begin{aligned}s^2 &= \frac{1}{6} [(11.4 - 11.7)^2 + (8.3 - 11.7)^2 + (9.4 - 11.7)^2 + (19.3 - 11.7)^2 \\ &\quad + (1.0 - 11.7)^2 + (16.3 - 11.7)^2 + (16.4 - 11.7)^2] \\ &\doteq 38.7\end{aligned}$$

Für die Standardabweichung erhalten wir schliesslich $s = \sqrt{38.7} \doteq 6.2$. \diamond

Es gibt die folgenden quantitativen Aussagen über die empirische Standardabweichung:

Für **beliebig verteilte** Beobachtungen x_i ($1 \leq i \leq n$) gilt:

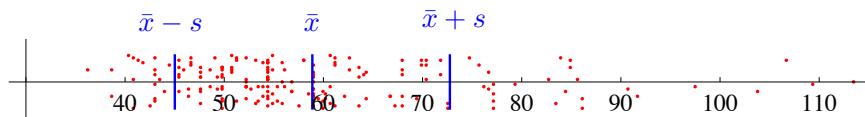
- (1) Intervall $[\bar{x} - s, \bar{x} + s]$: keine Angabe möglich
- (2) Intervall $[\bar{x} - 2s, \bar{x} + 2s]$: enthält **mindestens** 75% der Beobachtungen
- (3) Intervall $[\bar{x} - 3s, \bar{x} + 3s]$: enthält **mindestens** 88.8% der Beobachtungen

Falls die Beobachtungen x_i ($1 \leq i \leq n$) **normalverteilt** sind, gilt:

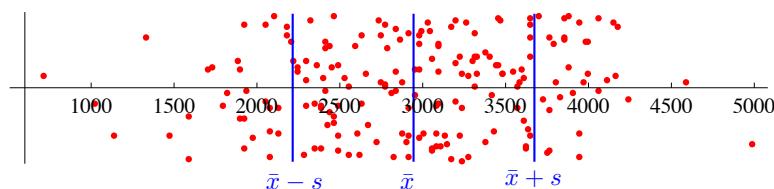
- (1) Intervall $[\bar{x} - s, \bar{x} + s]$: enthält **ungefähr** 68% der Beobachtungen
- (2) Intervall $[\bar{x} - 2s, \bar{x} + 2s]$: enthält **ungefähr** 95% der Beobachtungen
- (3) Intervall $[\bar{x} - 3s, \bar{x} + 3s]$: enthält **ungefähr** 99.5% der Beobachtungen

Betrachten wir wiederum den Datensatz (24):

- Die empirische Standardabweichung des Gewichts der Frauen beträgt $s = 13.9$ kg . Das Intervall $[\bar{x} - s, \bar{x} + s]$ enthält 86.2% der Beobachtungen.



- Die empirische Standardabweichung des Gewichts der Neugeborenen beträgt $s = 729$ g . Das Intervall $[\bar{x} - s, \bar{x} + s]$ enthält 83.6% der Beobachtungen.



4.2.2 Variationskoeffizient

Um die Variabilität von Variablen, die auf verschiedenen Skalen gemessen worden sind, vergleichen zu können, benötigen wir eine Kennzahl, die unabhängig von der Maßeinheit ist. Wir erhalten eine solche Masszahl, indem wir die Standardabweichung durch den Mittelwert dividieren. Man bezeichnet diese Grösse als **Variationskoeffizienten (coefficient of variation)**:

$$cv = \frac{s}{\bar{x}}$$

Der Variationskoeffizient ist eine **dimensionslose** Grösse. Er ist unabhängig gegenüber Skalenänderungen: Da s und \bar{x} die gleiche Maßeinheit besitzen, spielt es keine Rolle, ob wir zum Beispiel bei Längen mm, cm oder m verwenden. Der Variationskoeffizient wird nur für positive, verhältnisskalierte Daten, wie etwa Konzentrationen oder Anzahlen, verwendet.

Beispiel 31 Eine Psychologin und ein Psychologe entwickeln unabhängig voneinander zwei Fragebogen zur Messung des Merkmals *soziale Kompetenz*. Beide Fragebogen ergeben einen verhältnisskalierten Messwert und werden an der gleichen Stichprobe evaluiert. Es ergeben sich die folgenden Mittelwerte und Standardabweichungen:

$$\begin{aligned} \text{Fragebogen 1: } & \bar{x}_1 = 50, \quad s_1 = 15 \\ \text{Fragebogen 2: } & \bar{x}_2 = 8, \quad s_2 = 4 \end{aligned}$$

Bei welchem Fragebogen ist die Variabilität grösser?

Die Lösung wird im Unterricht erarbeitet. ◇

4.2.3 Interquartilsabstand und Spannweite

Wir betrachten jetzt ein Streumass, welches auf den Quartilen basiert.

- Das Intervall $[q_{0.25}, q_{0.75}]$ enthält 50% der Beobachtungen.
- Man bezeichnet die Differenz $q_{0.75} - q_{0.25}$ als **Quartils-Differenz** oder als **Interquartilsabstand** (inter-quartile range, IQR).

Zum Schluss erwähnen wir noch die **Spannweite**, die gleich der Differenz aus der grössten und der kleinsten Beobachtung ist:

$$\text{Spannweite} = \max_{1 \leq i \leq n} x_i - \min_{1 \leq i \leq n} x_i$$

Dieser Wert reagiert sehr empfindlich auf Extremwerte.

4.3 Boxplots

Minimum, 1. Quartil, Median, 3. Quartil und Maximum werden oft als 5-Punkt-Zusammenfassung für ein Merkmal angegeben. Darin enthalten sind mit dem Median ein Lagemass und implizit auch Spannweite und Interquartilsabstand, also zwei Streuungsmaße. Aus diesen 5 Grössen ist auch der **Boxplot** aufgebaut. Die Grenzen der Box sind durch die Quartile bestimmt. An der Stelle des Medians ist die Box unterteilt. Von der Box führen Linien (Whiskers) zum Minimum und Maximum. Der Boxplot wurde im Jahre 1977 von John Tukey³ publiziert.

³John Wilder Tukey (1915 Bedford - 2000 Princeton): amerikanischer Statistiker. Tukey hat zusammen mit J. W. Cooley auch den Algorithmus für die schnelle Fourier-Transformation entwickelt.

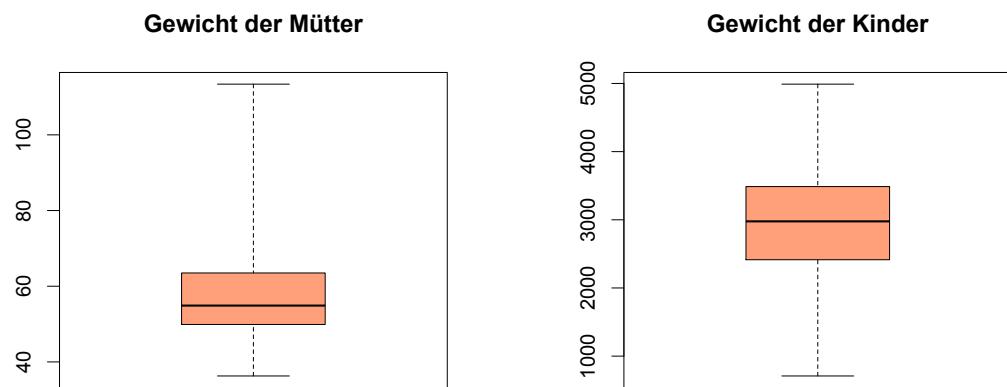
Im Fall des Datensatzes (24) lauten die 5 Werte für das Merkmal lwt (Gewicht der Mutter) folgendermassen:

Min.	1. Quartil	Median	3. Quartil	Max.
36.3	49.9	54.9	63.5	113.4

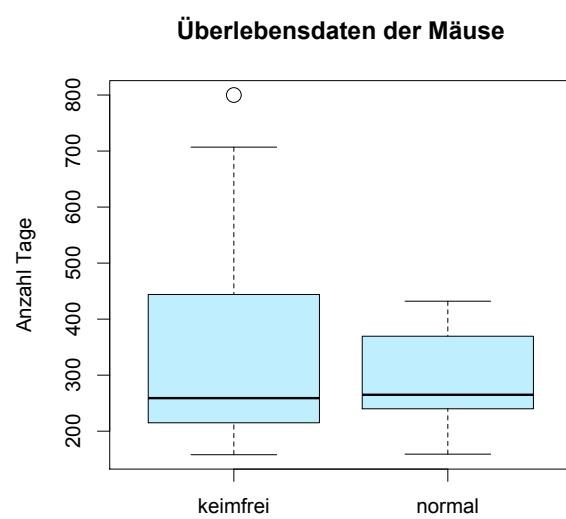
Der Boxplot ist nachfolgend links abgebildet. Für das Gewicht der Neugeborenen erhalten wir die folgenden 5 Zahlen:

Min.	1. Quartil	Median	3. Quartil	Max.
709	2'414	2'977	3'487	4'990

Der Boxplot ist unten rechts gegeben abgebildet:



Die nachfolgende Graphik zeigt die Boxplots der Überlebensdaten der beiden Gruppen von Mäusen aus dem Beispiel 28:



Standardmäßig werden im Statistikprogramm R im Boxplot **Ausreisser** als einzelne Punkte dargestellt, die nicht mehr durch die gestrichelte Linie (Whisker) verbunden werden. In den ersten beiden Boxplots haben wir dies mit der Option `outline=F` unterbunden. Ein Wert wird als Ausreisser betrachtet, wenn er nicht mehr im Intervall

$$[q_{0.25} - 1.5 \cdot \text{IQR}, q_{0.75} + 1.5 \cdot \text{IQR}]$$

liegt. Der obere Whisker wird bis zum grössten Wert, der kleiner oder gleich

$$q_{0.75} + 1.5 \cdot \text{IQR}$$

ist, gezeichnet und der untere Whisker bis zum kleinsten Wert, der grösser oder gleich

$$q_{0.25} - 1.5 \cdot \text{IQR}$$

ist.

In unserem Fall ist für die Gruppe *keimfrei* $\text{IQR} = 229$, das untere Quartil $q_{0.25} = 215.0$ und das obere Quartil $q_{0.75} = 444.0$. Wir erhalten so das Intervall $[-128.5, 787.5]$. Der beobachtete Wert 800 wird als Ausreisser betrachtet.

Die obige Regel wird von normalverteilten Daten motiviert. Falls die Daten normalverteilt sind, dann ist die Wahrscheinlichkeit für eine Abweichung vom oberen Quartil um mehr als $1.5 \cdot \text{IQR}$ gleich 0.00349. Analog nach unten. Auf etwa 300 Werte gäbe es dann im Mittel einen Ausreisser nach oben sowie nach unten.

Ausreisser können die folgenden Ursachen haben (siehe [4]):

- Fehler bei der Datenaufnahme durch ein defektes Messgerät;
- Codierfehler; unklare Maßeinheiten (z.B. Körpergrösse in Metern statt Zentimetern);
- Schreib- oder Tippfehler;
- Die Werte sind korrekt. Wenn in diesem Fall mehrere Ausreisser vorkommen, so deutet dies darauf hin, dass die Daten **nicht normalverteilt** sind.

In unserem Beispiel wird vermutlich der letzte Fall zutreffen. Ausreisser können statistische Auswertungen stark beeinflussen und unter Umständen verfälschen. Im Beispiel mit den Überlebensdaten der Mäuse wird der Mittelwert durch die Beobachtung 800 von 327.8 auf 344.1 angehoben. Wie muss man mit Ausreisern verfahren? Falls ein eindeutiger Datenfehler vorliegt, kann man versuchen den Fehler zu korrigieren. Ist eine Korrektur nicht möglich, weil man den tatsächlichen Wert nicht kennt, muss man die Beobachtung weglassen oder durch `NA` (not available) ersetzen. Schwieriger ist es, wenn der Wert theoretisch möglich ist. Man kann dann versuchen die Auswertung mit und ohne Ausreisser durchzuführen. Es gibt auch statistische Methoden, die wenig empfindlich gegenüber Ausreisern sind. Solche Methoden werden als **robust** bezeichnet. Wir haben oben bereits erwähnt, dass der Median ein robustes Lagemaß ist. Der Quartilsabstand (IQR) ist ein robustes Streuungsmaß.

4.4 Schiefe

Die **Schiefe** (E: skewness) ist ein Mass für die Abweichung von einer symmetrischen Verteilung. Im Fall einer Stichprobe $x_1, x_2, x_3, \dots, x_n$ ist die empirische Schiefe gegeben durch:

$$v = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3, \quad (26)$$

wobei \bar{x} der Mittelwert und s die empirische Standardabweichung der Stichprobe ist.
Mithilfe der **standardisierten Werte** (Mittelwert=0 und Standardabweichung=1)

$$z_i := \frac{x_i - \bar{x}}{s}$$

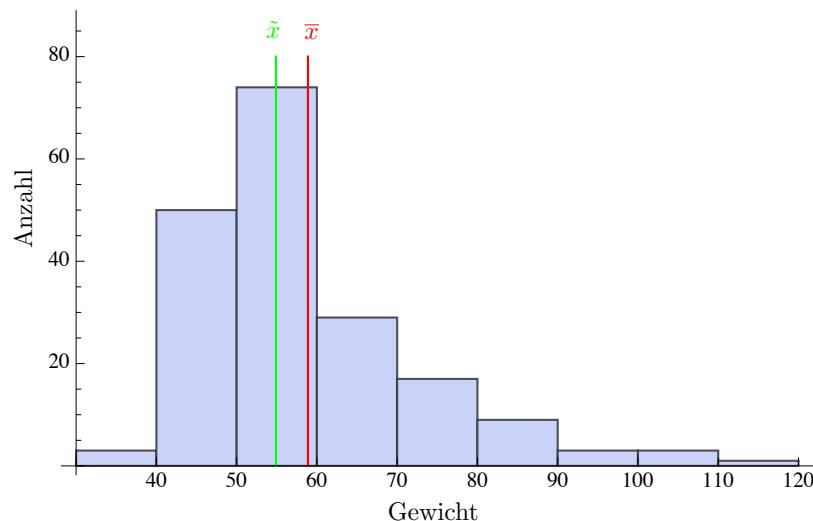
können wir die Schiefe auch so ausdrücken:

$$v = \frac{1}{n} \sum_{i=1}^n z_i^3 .$$

Beachten Sie, dass v **dimensionslos** ist. Sie ist invariant gegenüber Skalenänderungen.
Bei einer rechtsschiefen Verteilung ist $v > 0$ und bei einer linksschiefen Verteilung $v < 0$.
Bei einer symmetrischen Verteilung ist $v = 0$.

Es gibt noch kompliziertere Definitionen für die Schiefe als (26).

Die Verteilung des Gewichts der Mütter ist rechtsschief. Wir erhalten einen Schiefekoeffizienten von 1.4.



Als **Daumenregel** notieren wir:

- falls der Schiefekoeffizient kleiner als -1 oder grösser als 1 ist, ist die Verteilung stark schief;
- falls der Schiefekoeffizient zwischen -1 und -0.5 oder zwischen 0.5 und 1 ist, ist die Verteilung moderat schief;
- falls der Schiefekoeffizient zwischen -0.5 und $+0.5$ ist, ist die Verteilung näherungsweise symmetrisch.

Wir können dieses Mass auch verwenden, um zu überprüfen, ob Daten normalverteilt sind oder nicht. Wenn die Verteilung schief ist, können die Daten **nicht normalverteilt** sein. Umgekehrt können wir aber bei einer symmetrischen Verteilung nicht automatisch folgern, dass die Daten normalverteilt sind. Dazu muss die Höhe der Balken genügend schnell gegen 0 gehen.

4.5 Kurtosis*

Die **Kurtosis** ist durch den Ausdruck

$$w = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - 3 = \frac{1}{n} \sum_{i=1}^n z_i^4 - 3 , \quad (27)$$

definiert, wobei \bar{x} wiederum der Mittelwert, s die empirische Standardabweichung der Stichprobe und z_i die standardisierten Werte sind. Die Kurtosis ist dimensionslos und damit invariant gegenüber Skalenänderungen. Wenn grosse Abweichungen von \bar{x} auftreten, dann wird aufgrund der 4. Potenz w gross. Beachten Sie, dass in einigen Büchern die oben definierte Grösse als **Exzess** bezeichnet wird und die Kurtosis ohne die Subtraktion von 3 definiert wird.

Wie für die Schiefe gibt es auch für die Kurtosis noch kompliziertere Definitionen als (27). Dies kann zu kleinen Abweichungen führen.

Die Interpretation der Kurtosis ist die folgende:

- Wenn die Daten normalverteilt sind, dann gilt:

$$w \approx 0$$

Man bezeichnet diesen Fall als **mesokurtisch**.

- Wenn

$$w < 0$$

ist, dann sind die Schwänze der Verteilung kürzer und dünner als bei der Normalverteilung. Man bezeichnet diesen Fall als **platykurtisch**.

- Wenn

$$w > 0$$

ist, dann sind die Schwänze der Verteilung länger und dicker als bei der Normalverteilung. Man bezeichnet diesen Fall als **leptokurtisch**.

Leider hält sich in Lehrbüchern und selbst Forschungsartikeln hartnäckig die Ansicht, dass die Kurtosis etwas mit der Steilheit oder Flachheit des Gipfels zu tun hat. Diese falsche Ansicht wird in [17] widerlegt.

Für uns ist wichtig, dass $w > 0$ auf **Ausreisser** hindeutet.

Nachfolgend 3 Zufallsstichproben, die alle 3 auf Verteilungen mit Mittelwert 0 und Standardabweichung 1 basieren.

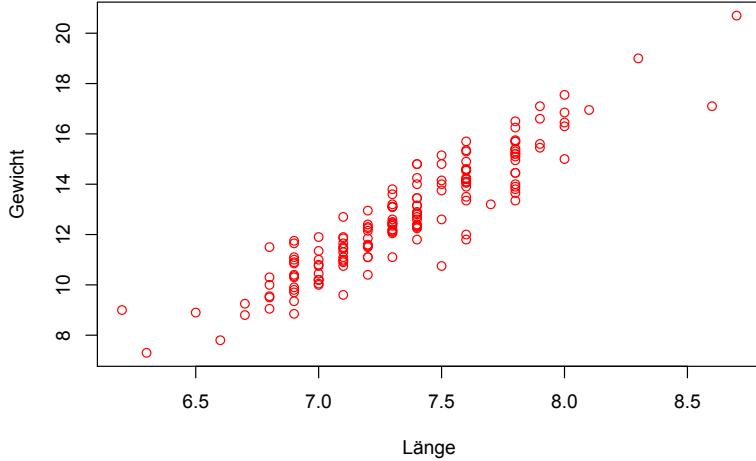
Distribution	Parameter	Kurtosis Category
gleichverteilt	$w = -1.20745$	platykurtisch
normalverteilt	$w = -0.0539041$	mesokurtisch
logistisch verteilt	$w = 1.11336$	leptokurtisch

Wenn wir von der Gleichverteilung zur Normalverteilung übergehen, stellen wir fest, dass ein Teil der Masse im Randbereich der Gleichverteilung ins Zentrum und in die Schwänze transferiert worden ist. Die mittleren und extremen Werte werden so wahrscheinlicher. Die Kurtosis nimmt zu. Beim Übergang von der Normalverteilung zur logistischen Verteilung geht dieser Trend weiter.

56

4.6 Korrelation

Wir kehren zurück zum Beispiel mit den Krabben. Nachfolgend zeigen wir nochmals den Scatterplot:



Eine längere Krabbe wird in der Regel auch schwerer sein. Der Scatterplot lässt vermuten, dass ein **linearer Zusammenhang** zwischen dem Gewicht y und der Länge x besteht:

$$y = ax + b .$$

Die Stärke dieses **linearen** Zusammenhangs kann mit dem **Korrelationskoeffizienten** gemessen werden. Dieser ist definiert durch:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} ,$$

wobei s_x und s_y die Standardabweichungen der x - bzw. y -Werte sind. Mit Hilfe des Skalarprodukts lässt sich zeigen, dass

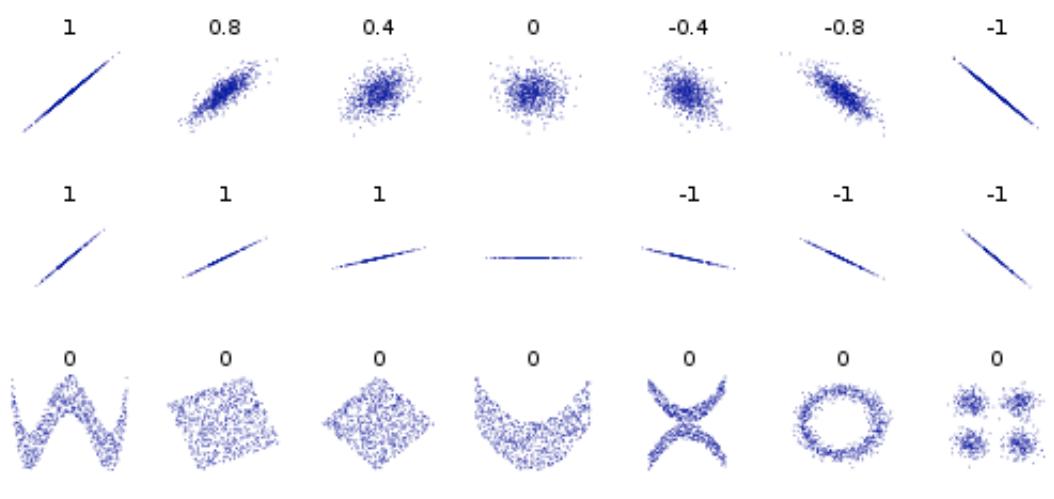
$$-1 \leq r \leq 1$$

ist. Es gilt:

- Wenn kein linearer Zusammenhang zwischen X und Y besteht, dann ist $r \approx 0$.
- Wenn $r \approx 1$, dann besteht ein positiver linearer Zusammenhang zwischen X und Y . Wenn X zunimmt, dann nimmt Y auch zu.
- Wenn $r \approx -1$, dann besteht ein negativer linearer Zusammenhang zwischen X und Y . Wenn X zunimmt, dann nimmt Y ab.

Für die Daten mit den Krabben erhalten wir $r = 0.92263$. Es liegt also ein starker positiver linearer Zusammenhang zwischen Länge und Gewicht vor.

Bevor man den Korrelationskoeffizienten interpretiert, sollt man sich mit einem Scatterplot der Daten davon überzeugen, dass tatsächlich ein linearer Zusammenhang zwischen X und Y besteht. Die nachfolgende Graphik aus Wikipedia zeigt die Korrelation für einige Punktwolken.



Beachten Sie, dass im Fall, wo die Punkte auf einer Geraden liegen, die Korrelation gleich 1 ist, wenn die Steigung positiv und gleich -1 ist, wenn die Steigung negativ ist. Wenn die Gerade horizontal ist, dann ist die Korrelation nicht definiert, da $s_y = 0$ ist.

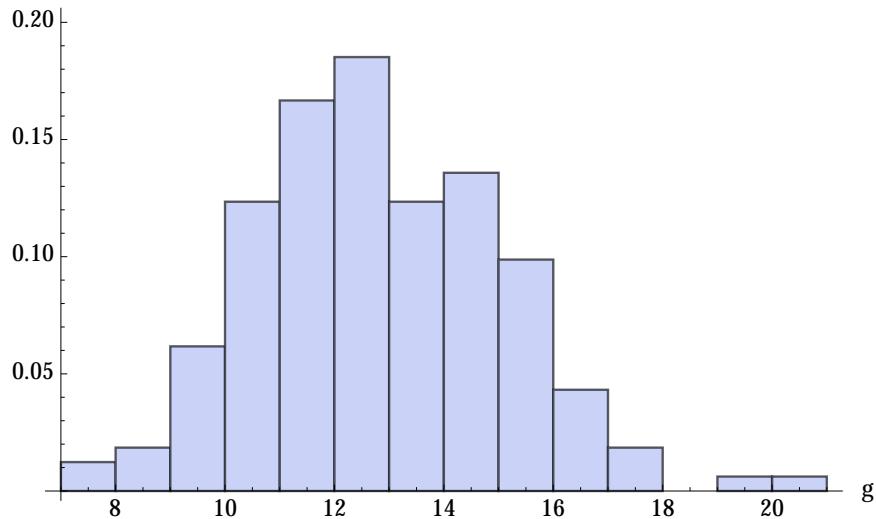
Wenn die Korrelation gleich 0 ist, bedeutet dies nur, dass es **keinen linearen** Zusammenhang gibt. Einen nichtlinearen Zusammenhang könnte es geben.

Wir werden im Kapitel über die lineare Regression auf die Korrelation zurückkommen.

5 Normalverteilte Daten

5.1 Einleitung

Wir haben bereits den Datensatz aus [3] mit den Messwerten zu den Krabben betrachtet. Nachfolgend das Histogramm der Gewichtsdaten in Gramm von 162 Krabben:

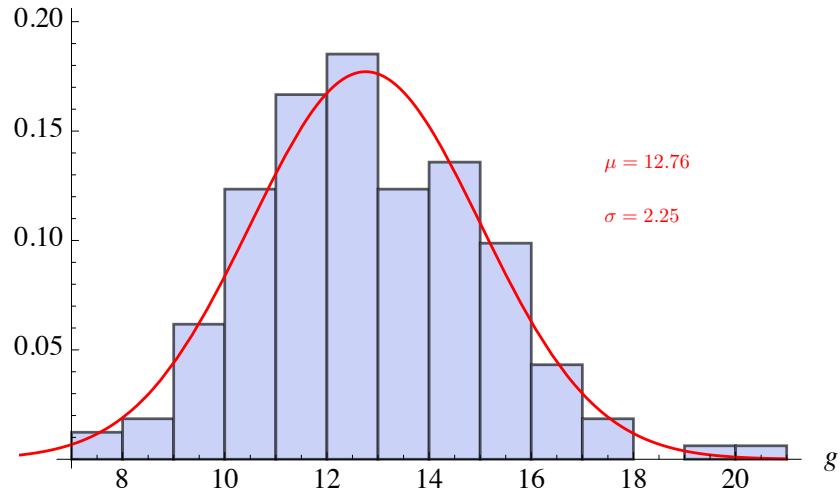


Die Höhe über jeder Klasse wird so gewählt, dass die Rechtecksfläche der empirischen Wahrscheinlichkeit entspricht, dass ein zufällig ausgewählter Wert in die entsprechende Klasse fällt. Wir können beispielsweise aus dem Histogramm herauslesen, dass die Wahrscheinlichkeit, dass eine Krabbe ein Gewicht zwischen 9 und 10 Gramm besitzt etwa gleich 6% ist. Wir können das Histogramm als Verteilung interpretieren. Für eine andere Stichprobe werden wir natürlich *nicht exakt* das gleiche Histogramm erhalten. Aus diesem Grund spricht man präziser von einer **empirischen Verteilungsfunktion**.

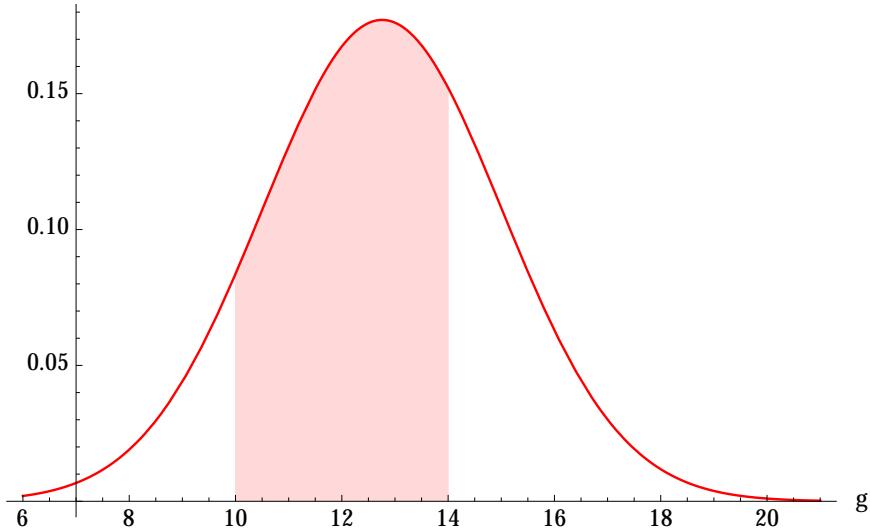
Viele, aber längst nicht alle, quantitativen Daten besitzen ein vergleichbares Histogramm: ein Gipfel in der Mitte, der auf beide Seiten mehr oder weniger symmetrisch abklingt. Dieses Abklingen muss darüberhinaus genügend schnell erfolgen. Solche Daten können oft mit der **Normalverteilung** modelliert werden. Wir erinnern daran, dass die **Dichte** der Normalverteilung gegeben ist durch:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

In der nachfolgenden Graphik haben wir den Graphen der Normalverteilung in das Histogramm gezeichnet. Für den Parameter μ haben wir das Stichprobenmittel und für σ die empirische Standardabweichung verwendet. Wir stellen fest, dass die Approximation zufriedenstellend ist, aber nicht mehr. Das liegt daran, dass die Klassenbreite mit der Länge 1 relativ gross ist. Hätte man mehr Krabben zur Verfügung, so könnte man die Klassen kleiner wählen und es würde sich eine bessere Approximation ergeben.



Wir nehmen jetzt an, dass das Gewicht der Krabben exakt normalverteilt ist mit Mittelwert $\mu = 12.76$ g und Standardabweichung $\sigma = 2.25$ g. Wenn wir zum Beispiel die Wahrscheinlichkeit berechnen möchten dass das Gewicht X einer Krabbe zwischen $a = 10$ g und $b = 14$ g liegt, $P(10 \leq X \leq 14)$, dann müssen wir den Inhalt der rot abgebildeten Fläche berechnen:



Dieser Flächeninhalt ist durch das bestimmte Integral

$$\frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad (28)$$

gegeben. Wir haben bereits erwähnt, dass wir dieses nicht mehr elementar mit Hilfe einer Stammfunktion berechnen können. Mit einem guten Taschenrechner, mit *Mathematica* oder der Statistiksoftware R erhalten wir:

$$\frac{1}{2.25 \cdot \sqrt{2\pi}} \int_{10}^{14} e^{-\frac{(x-12.76)^2}{2 \cdot 2.25^2}} dx \doteq 0.5989$$

5.2 QQ-Plot

Wir wenden uns in diesem Abschnitt der Frage zu, wie man prüfen kann, ob Daten normalverteilt sind. Das nach wie vor beste Verfahren besteht in einer visuellen Überprüfung anhand eines sogenannten Quantile-Quantile-Plot, kurz **QQ-Plot**. Alle Statistik-Softwarepakete besitzen eine solche Funktion und Sie müssen diese nur anwenden können.

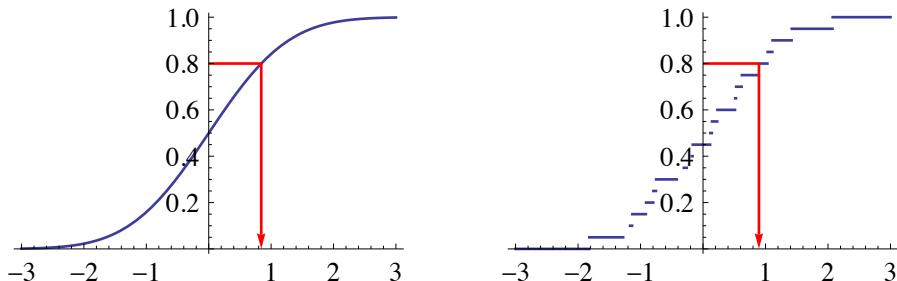
Wir betrachten die kumulative Verteilungsfunktion $\Phi(x)$ der Standardnormalverteilung. Diese Funktion ist streng monoton steigend und besitzt eine Umkehrfunktion. Für $0 < p < 1$ bezeichnet man den Wert dieser Umkehrfunktion als p -Quantil, geschrieben q_p :

$$q_p = \Phi^{-1}(p) .$$

Wir betrachten jetzt eine Stichprobe mit 20 standardnormalverteilten Zufallszahlen, die wir der Grösse nach geordnet haben:

$$\begin{array}{cccccc} -1.82175 & -1.27641 & -1.13323 & -0.913251 & -0.795239 \\ -0.751543 & -0.417967 & -0.259316 & -0.165893 & 0.114769 \\ 0.142405 & 0.228509 & 0.513383 & 0.528883 & 0.619885 \\ 0.895036 & 1.03179 & 1.11101 & 1.41913 & 2.07339 \end{array} \quad (29)$$

Die nachfolgende Graphik zeigt die **empirische kumulative Verteilungsfunktion** dieser Daten. Die Funktion ist gleich Null für $x < -1.82175$ und springt dann an dieser Stelle auf den Wert $\frac{1}{20} = 0.05$. Sie bleibt dann auf diesem Wert bis zur Stelle $x = -1.27641$ und springt dann auf den Wert $\frac{2}{20} = 0.1$ usw.

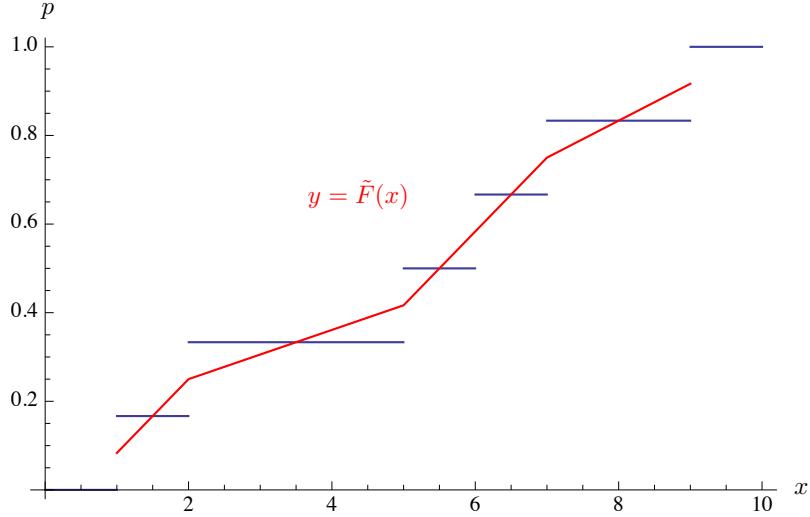


Natürlich ist diese Funktion nicht injektiv. Wir können aber leicht zu einer injektiven Ersatzfunktion kommen, indem wir an den Sprungstellen das arithmetische Mittel der beiden y -Werte bilden und diese Punkte durch Strecken verbinden. Es ergibt sich so eine geglättete Version der empirischen Verteilungsfunktion.

Wir zeigen die beiden Funktionen nochmals für die einfachere Stichprobe mit $n = 6$ Beobachtungen:

$$x_{(1)} = 1, x_{(2)} = 2, x_{(3)} = 5, x_{(4)} = 6, x_{(5)} = 7, x_{(6)} = 9 .$$

Die empirische kumulative Verteilungsfunktion $\hat{F}(x)$ ist blau gezeichnet. Die geglättete Verteilungsfunktion $\tilde{F}(x)$ ist rot gezeichnet:



Für diese geglättete Verteilungsfunktion $\tilde{F}(x)$ gilt dann:

$$\tilde{F}(x_{(i)}) = \frac{i - 0.5}{n} \quad \text{und} \quad \tilde{F}^{-1}\left(\frac{i - 0.5}{n}\right) = x_{(i)} .$$

Der Wert $x_{(i)}$ (der i -te Wert der sortierten Stichprobe) ist also gerade das $(i - 0.5)/n$ -Quantil.

Zu den Werten

$$p_i := \frac{i - 0.5}{n} \quad , \quad i = 1, 2, \dots, n ,$$

berechnet man die Quantile $q_{p_i} = \Phi^{-1}(p_i)$ der kumulativen Verteilungsfunktion der Standardnormalverteilung $\Phi(x)$. Wir fassen dann das empirische Quantil $q_{p_i}^* = x_{(i)}$ und das zugehörige theoretische Quantil q_{p_i} zu einem Punkt zusammen und erhalten so die folgende Menge von Punkten, die wir in einem Koordinatensystem darstellen können:

$$\{(q_{p_i}, x_{(i)}) : i = 1, 2, \dots, n\} .$$

Falls die Daten gut standardnormalverteilt sind, dann sollten die empirischen Quantile näherungsweise gleich den theoretischen Quantilen sein. Die Punkte sollten also näherungsweise auf der Geraden $y = x$ liegen.

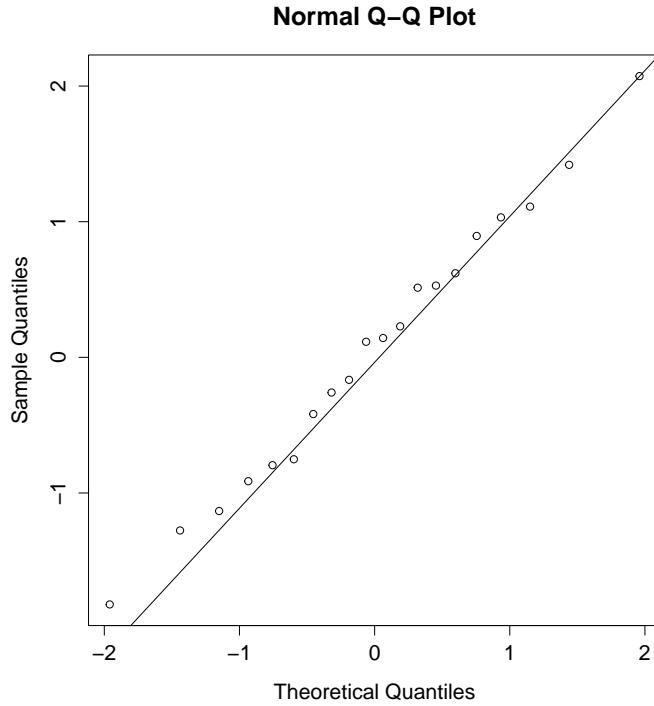
Kehren wir zu den 20 sortierten Zufallszahlen zurück. Diese Zahlen stellen die empirischen Quantile

$$q_{0.025}^*, q_{0.075}^*, q_{0.125}^*, q_{0.175}^*, \dots, q_{0.925}^*, q_{0.975}^*$$

dar. Wir bestimmen dann die zugehörigen theoretischen Quantile

$$q_p \quad , \quad p = 0.025, 0.075, \dots, 0.975 ,$$

der Standardnormalverteilung und fassen die entsprechenden Quantile zu einem Punkt (q_p, q_p^*) zusammen, den wir in einem Koordinatensystem darstellen. Die nachfolgende Graphik zeigt das Resultat:



Man bezeichnet diese Graphik als QQ-Plot. Falls die empirischen Quantile gleich den theoretischen Quantilen wären, dann müssten alle Punkte auf der Geraden $y = x$ liegen. Aufgrund von Zufallsschwankungen ist das natürlich nicht der Fall. Wir müssen einfach prüfen, ob die Punkte in etwa auf einer Geraden liegen. Beim oben vorliegenden Bild würde man die Daten als normalverteilt annehmen.

Wenn die Daten irgendwie normalverteilt sind, mit unbekannten μ und σ , dann funktioniert das obige Verfahren auch. Die Punkte müssen nach wie vor auf einer Geraden liegen. Diese Gerade hat die Steigung σ und den y -Achsenabschnitt μ .

Begründung für die interessierte Leserin und den interessierten Leser: Sei $F(u)$ die kumulative Verteilungsfunktion der Normalverteilung mit Mittelwert μ und Standardabweichung σ :

$$F(u) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^u e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx .$$

Sei $0 < p < 1$. Wir suchen u , so dass

$$F(u) = p .$$

Dieses u ist das p -Quantil für die Normalverteilung mit Mittelwert μ und Standardabweichung σ . Wir machen die Variablentransformation

$$t = \frac{x - \mu}{\sigma}, \quad dt = \frac{1}{\sigma} dx .$$

Wir erhalten so

$$F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{u-\mu}{\sigma}} e^{-\frac{t^2}{2}} dt = \Phi\left(\frac{u-\mu}{\sigma}\right) ,$$

wo $\Phi(x)$ die kumulative Verteilungsfunktion der Standardnormalverteilung ist. Wenn wir die Gleichung

$$F(u) = \Phi\left(\frac{u-\mu}{\sigma}\right) = p$$

nach u lösen, erhalten wir:

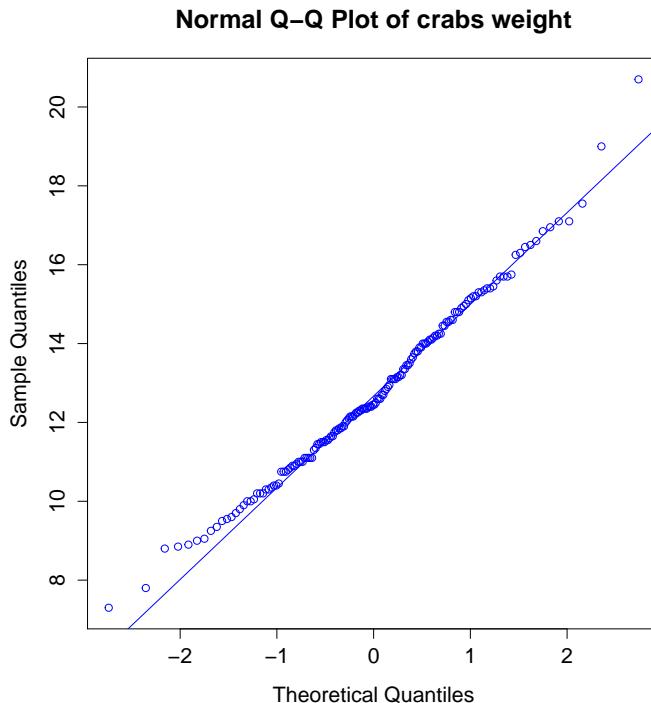
$$\frac{u - \mu}{\sigma} = \Phi^{-1}(p) \implies u = \sigma \cdot \Phi^{-1}(p) + \mu$$

Hier ist aber $\Phi^{-1}(p)$ das entsprechende Quantil der Standardnormalverteilung. Die Punkte $(\Phi^{-1}(p), u)$ liegen also auf einer Geraden mit Steigung σ und y -Achsenabschnitt μ .

Wir merken uns:

Wenn die Punkte im QQ-Plot nicht näherungsweise auf einer Geraden liegen, dann sind die Daten nicht normalverteilt.

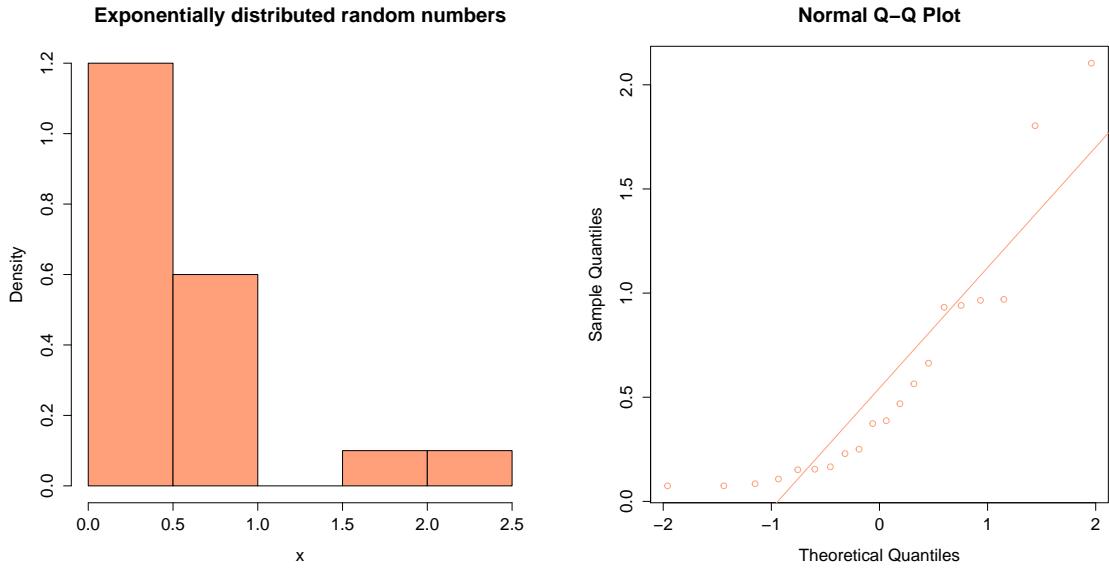
Die nachfolgende Graphik zeigt den QQ-Plot der Gewichte der Krabben. Auch hier würde man die Daten als normalverteilt annehmen, obwohl am Anfang und am Ende die Punkte von der eingezeichneten Geraden abweichen. Es zeigt sich bereits im Histogramm, dass die Verteilung weniger schnell gegen Null geht als die Normalverteilung. Sie ist *langschwänziger*.



Zum Abschluss betrachten wir noch einen Datensatz, der nicht normalverteilt ist. Die 20 nachfolgenden Zufallszahlen sind exponentialverteilt mit dem Mittelwert 0.5:

0.0745053	0.0746619	0.0847834	0.107161	0.151886
0.154398	0.166114	0.229101	0.250508	0.373426
0.387284	0.468921	0.564208	0.663514	0.931911
0.940774	0.96517	0.969738	1.80358	2.10385

Bereits das Histogramm zeigt, dass keine Normalverteilung vorliegt. Der QQ-Plot auf der rechten Seite bestätigt diesen Eindruck.



5.3 Der zentrale Grenzwertsatz

Seien X_i ($i = 1, 2, 3, \dots, n$) n Zufallsvariablen. Beispielsweise kann X_i das Gewicht der i -ten Krabbe in unserer Stichprobe von 162 Krabben sein oder X_i kann die Länge der i -ten Person sein aus einer Stichprobe von Personen mit gleichem Alter und gleichem Geschlecht. Wir machen die folgende Annahme:

Annahme: Alle X_i besitzen die gleiche Verteilung mit gleichem Mittelwert μ und gleicher Standardabweichung σ und sind unabhängig voneinander. Man schreibt dies kurz so: die X_i sind *i.i.d.* (*independant identically distributed*). Über die Art der Verteilung machen wir keine Aussage, insbesondere müssen die X_i nicht normalverteilt sein.

Wir interessieren uns für das Stichprobenmittel

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Wegen (16) und (17) gilt

$$E(\bar{X}) = \frac{1}{n} (\mu + \mu + \dots + \mu) = \mu$$

und wegen (19) und (20) gilt

$$V(\bar{X}) = \frac{1}{n^2} (\sigma^2 + \sigma^2 + \dots + \sigma^2) = \frac{\sigma^2}{n} .$$

Für die Standardabweichung gilt also:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Das Stichprobenmittel streut also weniger stark als die einzelnen X_i . Diese plausible Eigenschaft ist sehr wichtig.

Die nachfolgende Aussage sagt etwas über die *Verteilung* von \bar{X} .

Aussage: Das Stichprobenmittel

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

ist **approximativ normalverteilt** mit Mittelwert μ und Varianz σ^2/n , wenn n hinreichend gross ist:

$$\bar{X} \stackrel{\text{approx}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad (30)$$

Indem wir \bar{X} standardisieren kann die Aussage (30) auch so geschrieben werden:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \stackrel{\text{approx}}{\sim} \mathcal{N}(0, 1) \quad (31)$$

Die mathematisch exakte Aussage verwendet den Begriff des Grenzwerts. Die Verteilung ist exakt gleich der Normalverteilung, wenn n gegen Unendlich strebt. Die oben formulierte Aussage ist aber wichtiger für uns, denn wir haben es immer mit Stichproben zu tun, die eine endliche Grösse n besitzen. Was heisst n müsse hinreichend gross sein? Es gibt die Daumenregel, dass

$$n > 10 \cdot |\text{Kurtosis}|$$

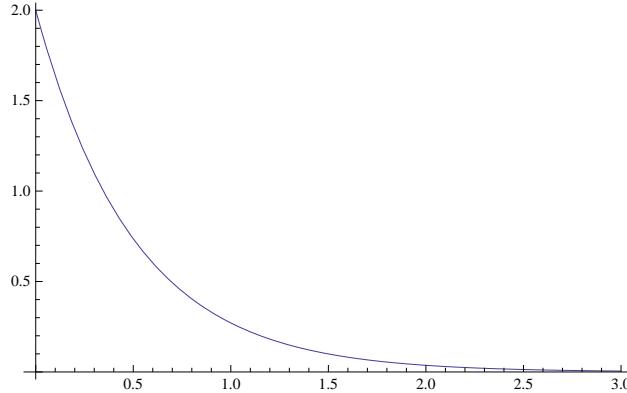
sein müsse. Ohne diese Daumenregel lässt sich sagen, dass in den allermeisten Fällen die Approximation für $n \geq 30$ oder besser $n \geq 50$ genügend gut ist.

Wir erwähnen noch, dass \bar{X} **exakt** normalverteilt ist, wenn die X_i normalverteilt sind.

Beispiel 32 Die Dichte der Exponentialverteilung mit Parameter $\lambda > 0$ ist gegeben durch:

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

Ihr Mittelwert und ihre Standardabweichung ist gleich $1/\lambda$. Die Exponentialverteilung wird oft dazu benutzt die Lebensdauer von Objekten (z.B. Autobatterien) oder Wartezeiten zu modellieren. Die nachfolgende Graphik zeigt die Dichte der Exponentialverteilung für $\lambda = 2$.

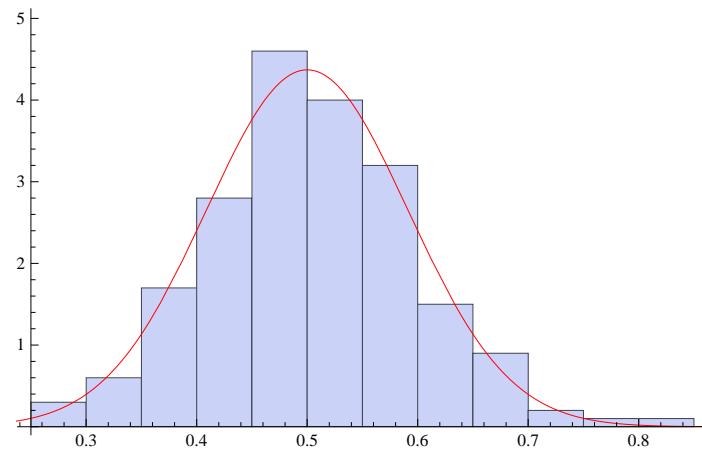


Wir stellen fest, dass sich die Verteilung stark von der Normalverteilung unterscheidet, insbesondere ist sie nicht symmetrisch bezüglich der Geraden $x = 0.5$.

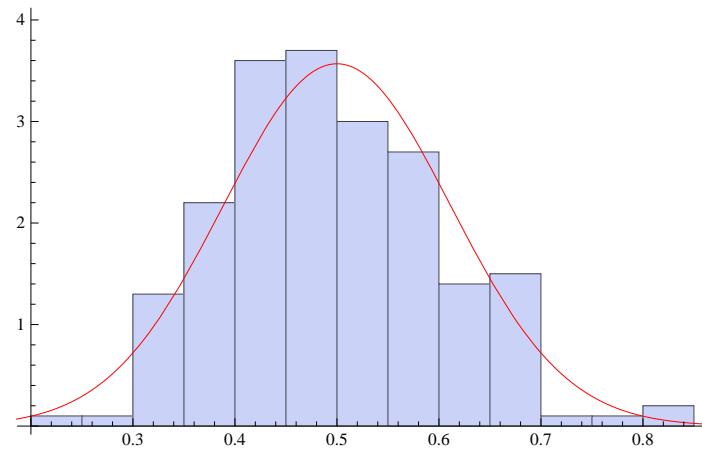
Wir erzeugen jetzt mit einer Software 30 exponentialverteilte Zufallszahlen mit Mittelwert $1/\lambda = 0.5$. Nachfolgend die sortierte Liste dieser Zahlen:

0.00521219	0.0113855	0.0229525	0.040869	0.101144	0.107409
0.112364	0.117128	0.14362	0.194052	0.200271	0.221093
0.248641	0.264576	0.352157	0.40014	0.405607	0.45959
0.473239	0.507268	0.540414	0.629203	0.77925	0.821205
1.00318	1.04255	1.09196	1.70623	1.98752	2.32097

Wir berechnen den Mittelwert und erhalten $\bar{x} = 0.543707$. Wir wiederholen dies 200 mal und erhalten so 200 Mittelwerte. Die nachfolgende Graphik zeigt das Histogramm dieser Mittelwerte zusammen mit der Dichte der Normalverteilung für $\mu = 0.5$ und $\sigma = 0.5/\sqrt{30} \doteq 0.09129$:



Hätten wir Stichproben mit nur 20 Zufallszahlen verwendet, dann wäre die Approximation noch nicht so gut:



◇

Beispiel 33 Wiederholen Sie das Experiment mit normalverteilten Zufallszahlen.

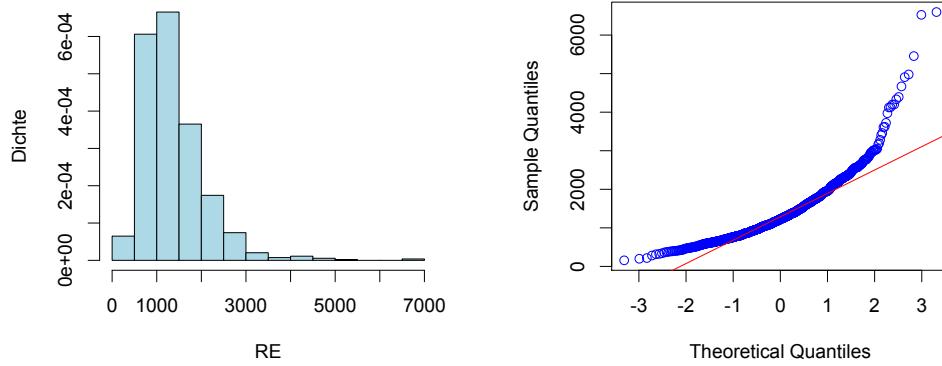
Das Problem wird im Unterricht gelöst.

◇

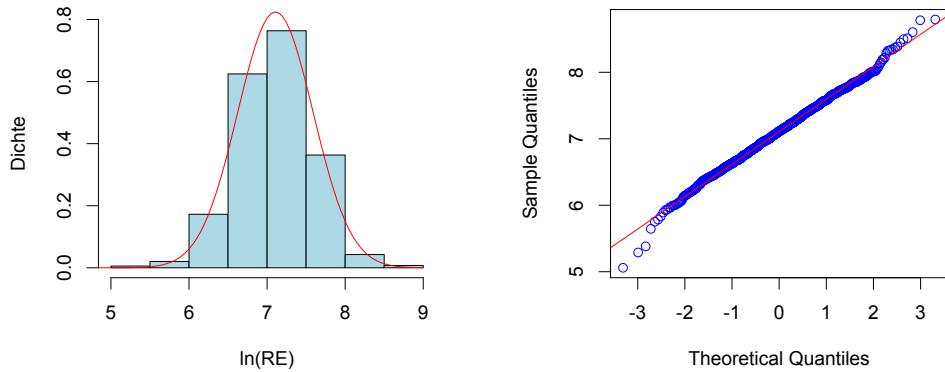
Mit dem zentralen Grenzwertsatz können wir auch erklären wieso viele Daten normalverteilt sind (siehe [16]). Ist die Variabilität, speziell ein Messfehler, eine Summe von vielen kleinen, unabhängigen Effekten, so ist sie näherungsweise normalverteilt. Allerdings zeigt es sich, dass die Annahme, dass sich die Effekte **addieren**, selten realistisch ist. Wir kommen im nächsten Abschnitt darauf zurück.

5.4 Transformation von Daten

Falls die Daten nicht normalverteilt sind, kann manchmal eine geeignete Transformation helfen. Wir verwenden einen Datensatz aus [3], welcher aus einer Studie aus dem Jahre 1985 resultierte, in der die Nahrungsaufnahme der dänischen Bevölkerung studiert wurde. Wir interessieren uns für die Aufnahme an Vitamin A (gemessen in Retinoläquivalenten, RE) bei den Männern. Die Studie umfasst 1'079 Männer.



Sowohl das Histogramm wie der QQ-Plot zeigen deutlich, dass die Daten **nicht** normalverteilt sind. Das Histogramm zeigt, dass die Verteilung nicht symmetrisch, sondern **rechtsschief** ist. Bei Zufallsvariablen, die nur positive Werte annehmen können, ist dies zu erwarten: es können grosse positive, aber keine negativen Werte angenommen werden. In diesem Fall hilft oft die **Logarithmus-Transformation**. Wir wenden auf die Daten den natürlichen Logarithmus \ln an. Das Histogramm und der QQ-Plot der transformierten Daten zeigen, dass diese approximativ normalverteilt sind. Die eingezeichnete rote Kurve ist die Dichte der Normalverteilung mit $\mu = 7.48$ (Mittelwert der transformierten Daten) und $\sigma = 0.44$ (Standardabweichung der transformierten Daten).



Beispiel 34 Es soll mit den oben erwähnten Daten die Wahrscheinlichkeit berechnet werden, dass ein zufällig ausgewählter Mann eine Aufnahme an Vitamin A hat, die zwischen 3'000 und 4'000 liegt.

Lösung: Es ist wichtig, dass $y = \ln x$ eine streng **monoton wachsende** Funktion ist, das heisst es gilt:

$$x_1 < x_2 \implies \ln x_1 < \ln x_2$$

Wir transformieren die oben angegebenen Grenzen:

$$a := \ln 3'000 \doteq 8.006 \quad \text{und} \quad b := \ln 4'000 \doteq 8.294$$

Wir berechnen dann mit $\mu = 7.48$ und $\sigma = 0.44$:

$$\frac{1}{\sigma \cdot \sqrt{2\pi}} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \doteq 8.5\%$$

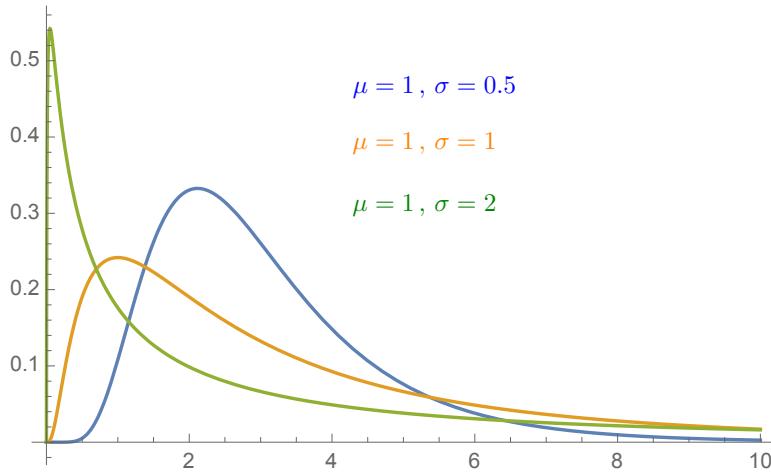
◇

Wir erwähnen noch das Folgende: eine Zufallsvariable Y heisst **lognormal-verteilt**, wenn $\ln Y$ normalverteilt ist. Die obigen Daten sind also näherungsweise lognormalverteilt.

Wir haben bei der Normalverteilung die Hypothese der additiven Elementarfehler erwähnt (siehe Seite 68). Viele zufällige Effekte mit kleinen Variationskoeffizienten, die **multiplikativ** wirken, erzeugen eine Lognormal-Verteilung. Da in den Naturgesetzen Additionen viel seltener vorkommen als Multiplikationen und Divisionen passt für gemessene Daten die Lognormal-Verteilung oft viel besser als die Normalverteilung (siehe [16]). Im Artikel [9] werden diese Überlegungen vertieft.

Nachfolgend drei Dichten einer lognormal-verteilten Zufallsvariablen Y mit

$$\ln Y \sim \mathcal{N}(\mu, \sigma^2).$$



Es gilt:

$$\text{Mittelwert: } e^{\mu + \frac{1}{2}\sigma^2}$$

$$\text{Varianz: } e^{\sigma^2} (e^{\sigma^2} - 1) e^{2\mu}$$

$$\text{Median: } e^\mu$$

Stichwortverzeichnis

- Ausprägungen, 36
- Ausreisser, 46, 54
- Balkendiagramm, 37
- Baumdiagramm, 10
- Beobachtungen, 36
- Binomialkoeffizient, 28
- Binomialverteilung, 27
- binomische Formel, 28
- Boxplot, 52
- Bubble Plot, 45
- Dichte, 31
- Dichte-Histogramm, 41
- Ereignis, 4
 - Gegenereignis, 4
 - sichere, 4
 - unmöglich, 4
- Ereignisse
 - abhängige, 9
 - unabhängige, 9
 - unvereinbare, 5
- Erwartungswert, 21
- Exzess, 56
- Formel
 - der totalen Wahrscheinlichkeit, 10
 - von Bayes, 14
- Geburtstagsparadoxon, 13
- geordnetes Paar, 3
- Grenzwertsatz
 - zentraler, 65
- Histogramm, 40
 - linksschiefes, 42
 - rechtsschiefes, 42
 - symmetrisch, 42
- i.i.d, 65
- Interquartilabstand, 52
- IQR, 52
- kartesisches Produkt , 4
- Korrelationskoeffizient, 26, 57
- Kovarianz, 26
- Kurtosis, 56
- Laplace-Raum, 6, 9
- Laufindex, 22
- leptokurtisch, 56
- Median, 47
- Merkmal, 36
 - diskretes, 39
 - intervallskaliertes, 39
 - kategorielles, 36
 - nominales, 37
 - ordinales, 37
 - quantitatives, 36, 39
 - stetiges, 39
 - verhältnisskaliertes, 39
- mesokurtisch, 56
- Mittelwert, 46
- Modus, 49
- Normalverteilung, 33
- platykurtisch, 56
- Prävalenz, 15
- QQ-Plot, 61
- Quantil, 49
- Quartil
 - oberes, 50
 - unteres, 50
- Schiefe, 54
- Sensitivität, 14
- Spannweite, 52
- Spezifität, 14
- Standardabweichung, 23
 - empirische, 50
- Standardnormalverteilung, 33
- Statistik
 - deskriptive, 36
- Stichprobenmittel, 46
- Stichprobenraum, 3
- Streudiagramm, 44
- Summenzeichen, 22
- Tortendiagramm, 37
- Transformation
 - Logarithmus-, 68
- Varianz, 23
 - empirische, 50
- Variationskoeffizient, 52
- Verteilung, 19
 - lognormal, 69

Verteilungsfunktion, 32	Wert
empirische, 61	standardisierter, 55
Vierfeldertafel, 10	
Vorhersagewert	Zufallsexperiment, 1, 3
negativer, 15	Zufallsvariable, 18
positiver, 15	diskrete, 18
Wahrscheinlichkeit, 5	stetige, 18, 30
bedingte, 7	unabhängige, 20

Literatur

- [1] Jürgen Bortz, *Statistik für Human- und Sozialwissenschaftler*, 6. Auflage, Springer, 2005
- [2] Wayne W. Daniel, *Biostatistics - Basic Concepts and Methodology for the Health Sciences*, 9th Edition (International Student Version), Wiley, 2010
- [3] Claus Thorn Ekstrøm, Helle Sørensen, *Statistical Data Analysis for the Life Sciences*, CRC Press, 2011
- [4] Reinhold Hatzinger, Kurt Hornik, Herbert Nagel, *R - Einführung durch angewandte Statistik*, Pearson, 2011
- [5] Leonhard Held, Kaspar Rufibach, Burkhardt Seifert, *Medizinische Statistik - Konzepte, Methoden, Anwendungen*, Pearson 2013
- [6] Norbert Henze, *Stochastik für Einsteiger*, 12. Auflage, Springer Spektrum, 2018
- [7] Manfred Baum et al., *LS Stochastik*, 1. Auflage, Ernst-Klett-Verlag, 2003
- [8] Jürg Hüsler, Heinz Zimmermann, *Statistische Prinzipien für medizinische Projekte*, 5. Auflage, Huber, 2010
- [9] Eckhard Limpert, Werner A. Stahel, Markus Abbt, *Log-normal Distributions across the Sciences: Keys and Clues*, BioScience 51: 341-352
- [10] Göran Kauermann, Helmut Küchenhoff, *Stichproben*, Springer, 2011
- [11] Ulrich Kockelkorn, *Statistik für Anwender*, Springer Spektrum, 2012
- [12] Bernard Rosner, *Fundamentals of Biostatistics*, 7th Edition (International Edition), Wadsworth Cengage Learning, 2011
- [13] Sheldon M. Ross, *Statistik für Ingenieure und Naturwissenschaftler*, 3. Auflage, Spektrum, 2006
- [14] Myra L. Samuels, Jeffrey A. Witmer, Andrew Schaffner, *Statistics for the life sciences*, 4th edition, Pearson, 2012
- [15] Babak Shahbaba, *Biostatistics with R - An Introduction to Statistics Through Biological Data*, Springer, 2012
- [16] Werner A. Stahel, *Statistische Datenanalyse - Eine Einführung für Naturwissenschaftler*, 5. Auflage, Vieweg+Teubner, 2007
- [17] Peter H. Westfall, *Kurtosis as Peakedness*, *The American Statistician*, 68(3), 2014