Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Einführung in die Statistik

Grundlagen der Mathematik II

Lineare Algebra und Statistik

FS 2014

Dr. Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

CH-8401 Winterthur

1 Was ist Statistik?

Statistische Datenanalyse ist dazu da, um unter der Präsenz von Unsicherheit und Variation korrekte Fakten zu gewinnen und intelligente Aussagen abzuleiten, die nicht von den Launen des Zufalls beeinflusst sind. Es geht darum, Messwerte und Beobachtungen in systematische Effekte und zufällige Variation zu separieren.

Messdaten, Ausgänge von Experimenten und Resultate aus Umfragen weisen immer Unsicherheit und Variation auf. Darum ist die Statistik zum unentbehrlichen Hilfsmittel in Wissenschaft, Technik und Alltag geworden. In der Forschung, und sei dies nur für Studentenarbeiten oder das Lesen von Fachliteratur, kommt man ohne Statistik-Kenntnisse nirgendwo hin. Selbst in der Alltagspresse finden immer häufiger Resultate von mit statistischen Methoden ausgewerteten Studien Erwähnung. Selbst zu deren Interpretation sind gewisse Grundkenntnisse in der Statistik unverzichtbar.

1.1 Beispiel 1: Wirksamkeit von Schlafmitteln

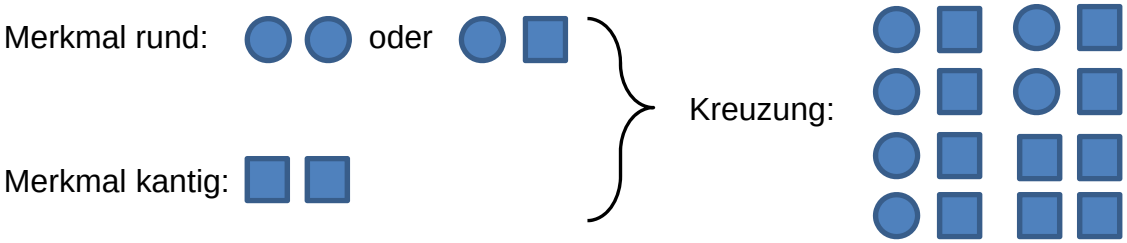
Untersucht werden soll die Wirksamkeit von zwei Schlafmitteln A und B. Dazu wurde bei 10 Probanden die durchschnittliche Schlafverlängerung von A vs. B in Stunden gemessen. Die beobachteten Werte sind:

+1.2 +2.4 +1.3 +1.3 +0.0 +1.0 +1.8 +0.8 +4.6 +1.4

Es scheint offensichtlich, dass man mit Mittel A einen längeren Schlaf geniesst. Eine erste und wichtige Aufgabe ist es, die durchschnittliche Verlängerung des Schlafes zu bestimmen, was man z.B. mit dem *arithmetischen Mittel* tun kann. Wir müssen uns aber bewusst sein, dass mit anderen Probanden andere Messwerte erzielt würden. Daher wäre es nützlich, die durchschnittliche Schlafverlängerung mit einer Genauigkeitsangabe zu versehen. Dies kann man mit dem *Vertrauensintervall* tun. Eine weitere spannende Frage ist es, ob A signifikant besser als B ist, und mit welcher Sicherheit wir dies sagen können. Die Antwort darauf gibt ein *statistischer Test*, bzw. dessen *p-Wert*.

1.2 Beispiel 2: Mendels Vererbungsgesetze

Gregor Mendel publizierte im Jahr 1866 eine Studie über den Vererbungsvorgang bei Merkmalen, deren Ausprägung nur von einem einzelnen Gen bestimmt wird. Als Beispiel studierte er 2 Erbsensorten mit runden bzw. kantigen Samen. Dabei werden runde Samen dominant vererbt. Daher sollte das folgende Vererbungsgesetz gelten:

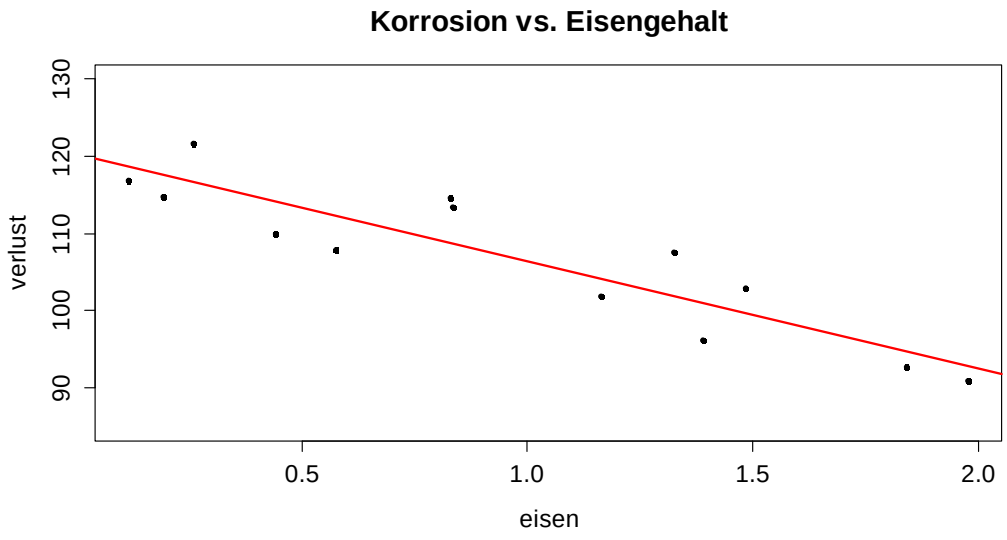


Wenn man runde und kantige Erbsen kreuzt und Mendels Modell gilt, dann sollte man bei den Nachkommen ein Verhältnis von 3:1 zugunsten der runden Samen beobachten.

Mendel führte Experimente durch, um seine Theorie zu verifizieren. Bei einer Auszählung erhielt er 5‘474 runde, und 1‘850 kantige Samen. Dies entspricht einem Verhältnis von 2.96:1. Es stellt sich nun natürlich die Frage, ob die Pflanzen dem Vererbungsmodell folgen oder nicht. Oder etwas mehr im Statistik-Jargon formuliert: Wie gross ist die zufällige Schwankungsbreite für das Verhältnis, wenn man insgesamt 7‘324 Samen auszählt? Man kann dies wiederum mit einem *Vertrauensintervall* beantworten, bzw. einen *statistischen Test* ausführen, welcher das postulierte Verhältnis von 3:1 testet. Im vorliegenden Fall liegt übrigens kein Widerspruch gegen das Vererbungsmodell vor.

1.3 Beispiel 3: Korrosion in Abhängigkeit vom Eisengehalt

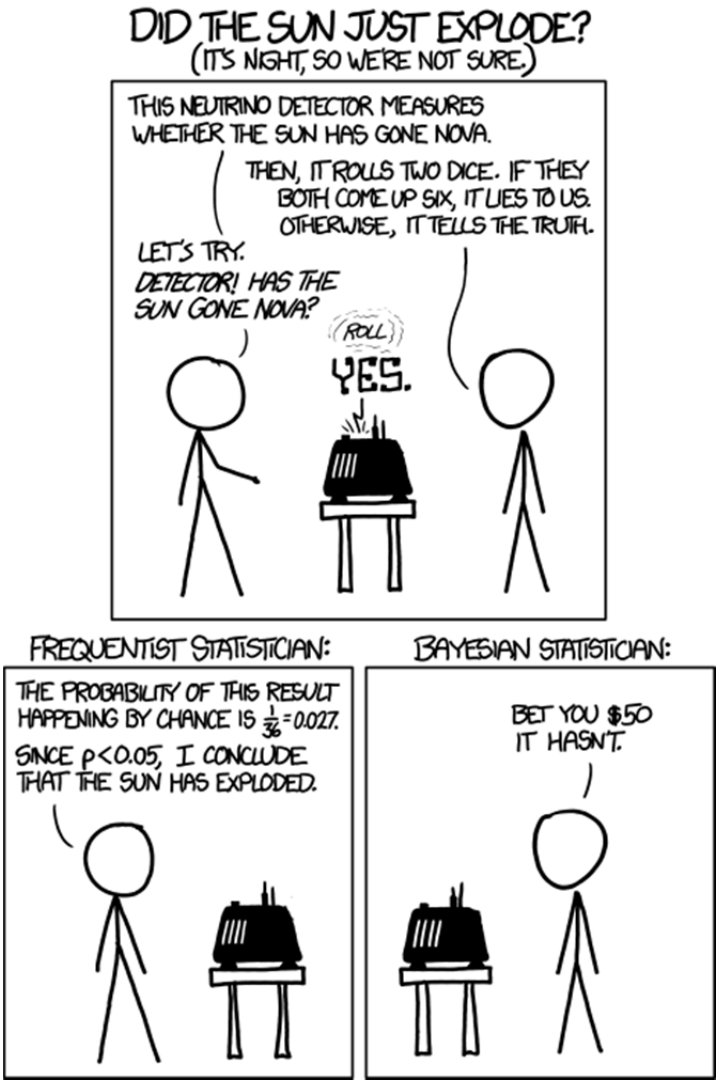
In einem Experiment soll der Zusammenhang zwischen der Korrosion einer Kupfer-Nickel-Legierung in Abhängigkeit von deren Eisengehalt studiert werden. Dazu wurden 13 Räder hergestellt und während 60 Tagen gedreht. Gemessen wurde der korrosionsbedingte Gewichtsverlust in Milligramm.



Die Daten können mit einem Streudiagramm dargestellt werden. Wir beobachten, dass der Gewichtsverlust mit zunehmendem Eisengehalt abnimmt. Einige Fragestellungen, die in der Praxis interessieren sind z.B., um wie viel der Verlust kleiner ist, wenn der Eisengehalt um eine Einheit zunimmt. Ebenso interessant ist es, ob diese Abnahme gesichert ist, oder bloss auf die zufällige Streuung der Datenpunkte, d.h. einen Zufallseffekt zurückzuführen ist. Weiter wäre es auch spannend zu wissen, wie und mit welcher Genauigkeit man den Gewichtsverlust in Abhängigkeit vom Eisengehalt vorhersagen kann. Die Antworten auf all diese Fragen liefert die *Regression*.

1.4 Übersicht über das Skript

Im Rahmen der folgenden Kapitel werden wir alle Methoden kennenlernen, welche zur Behandlung der 3 vorgestellten Anwendungsprobleme notwendig sind. Zuerst müssen wir aber eine kurze Einführung in die Wahrscheinlichkeitsrechnung machen. Sie liefert den Unterbau, damit wir uns überhaupt der statistischen Datenanalyse widmen können.



2 Konzepte der Wahrscheinlichkeitsrechnung

2.1 Zufall und Wahrscheinlichkeit

Ein *Zufallsexperiment* ist ein Versuch, bzw. eine Situation, wo das Ergebnis nicht deterministisch vorbestimmt ist. Um zu entscheiden, ob eine Situation ein Zufallsexperiment ist, stellt man sich am besten die Frage, ob bei einer Wiederholung exakt dasselbe Resultat erneut auftreten würde. Ist die Antwort auf diese Frage „nein", so ist die Definition für ein Zufallsexperiment erfüllt. In der Praxis, bzw. im täglichen Alltag gibt es sehr viele solche Zufallsexperimente:

- Ein nahe liegendes Beispiel, auf welches wir im Rahmen dieses Skript einige Male zurückgreifen werden, ist der Münzen- oder der Würfelwurf.
- Auch bei der Anzahl Studenten, welche in der Vorlesung Lineare Algebra und Statistik anwesend sind, stellt ein Zufallsexperiment dar. Man kennt sie nicht im Voraus, und sie ist immer wieder verschieden.
- Ebenso ist die Regenmenge innerhalb von 24h bei der Messstation von Meteoschweiz am Züriberg ein Zufallsexperiment. Sie ist nicht exakt vorbestimmt, und am nächsten Tag möglicherweise wieder anders.

Immer dann, wenn das Ergebnis eines Versuchs oder einer Beobachtung nicht mit Sicherheit vorausgesagt werden kann, so behilft man sich mit einer Angabe der Wahrscheinlichkeit. Beim Münz- und Würfelwurf ist es noch offensichtlich, wie man die Wahrscheinlichkeit auf Kopf/Zahl bzw. die Augenzahlen aufteilt. Auch bei der Studentenzahl teilen wir jedem Wert eine gewisse Wahrscheinlichkeit zu. Bei der Regenmenge macht man dies ebenso, allerdings ist es bei dieser stetigen Grösse weniger offensichtlich, wie dies geht – wir werden darauf zurückkommen.

Wir führen nun einige weitere Begriffe im Zusammenhang mit Zufall und Wahrscheinlichkeit ein. Wir beginnen mit dem Ereignis- oder Wahrscheinlichkeitsraum Ω. Er enthält alle möglichen Ausgänge eines Zufallsexperiments. In unseren Beispielen ist er wie folgt:

Münzenwurf:	$\Omega = \{kopf, zahl\}$
Würfelwurf:	$\Omega = \{1,2,3,4,5,6\}$
Studentenzahl:	$\Omega = \{0,1,2,...\} = \mathbb{N}$
Regenmenge:	$\Omega = [0,+\infty) = \mathbb{R}_+$

Die einzelnen Elemente von Ω nennt man Elementarereignisse. Unter gewissen mathematisch motivierten Voraussetzungen, die bei den von uns betrachteten Beispielen stets erfüllt sind, kann man Teilmengen A von Ω definieren. Eine solche Teilmenge A nennt man Ereignis. Als Beispiel nehmen wir das Ereignis „alle geraden Zahlen" beim Würfelwurf. Dann ist $A = \{2,4,6\} \subset \Omega$. Bei der Regenmenge könnte z.B. das Ereignis, dass zwischen 10mm und 20mm Regen fällt, von Interesse sein. Dann wäre $A' = [10,20]$. Als nächstes kümmern wir uns um den Begriff der Wahrscheinlichkeit.

Def: Die Wahrscheinlichkeit $P(\cdot)$ ist eine Funktion, die jedem Ereignis, d.h. jeder Teilmenge von Ω eine Zahl zwischen 0 und 1 zuordnet. Diese beschreibt die idealisierte relative Häufigkeit, mit welcher das Ereignis beim Durchführen des Zufallsexperiments eintritt.

Es seien nun A, B Ereignisse zum Wahrscheinlichkeitsraum Ω mit zugehöriger Wahrscheinlichkeitsfunktion $P(\cdot)$ und A^c sei das Komplement, d.h. das Gegenereignis zu A. Wir fordern axiomatisch die Aussagen i)-iii), woraus dann iv)-vii) folgen.

i) $P(\Omega) = 1$
ii) Für $A \subseteq \Omega$ gilt $0 \le P(A) \le 1$
iii) $P(A \cup B) = P(A) + P(B)$, falls A und B disjunkt sind, d.h. $A \cap B = \varnothing$.
iv) $P(A^c) = 1 - P(A)$
v) $P(\varnothing) = 0$
vi) Wenn $A \subset B$, so gilt $P(A) < P(B)$
vii) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Daraus entnehmen wir, dass Wahrscheinlichkeiten stets zwischen 0 (unmöglich eintretendes Ereignis) und 1 (sicher eintretendes Ereignis) liegen. Oft gibt man die Wahrscheinlichkeiten auch in Prozenten an, dann liegen sie zwischen 0% und 100%. Für den mathematischen Umgang mit Zufallsexperimenten und deren unsicherem Ausgang erweist sich das Konzept von Zufallsvariable und Wahrscheinlichkeitsverteilung als äusserst nützlich.

Def: Eine *Zufallsvariable* beschreibt den quantitativen Ausgang eines Zufallsexperiments. Sie dient quasi als „Platzhalter" und „liefert" einen Zahlenwert, der beim Durchführen des Experiments abgeleitet werden kann. Man kann sich eine Maschine vorstellen, welche im Innern das Zufallsexperiment durchführt und eine Zahl liefert. Es gibt *diskrete* und *stetige* Zufallsvariablen. Diskret ist eine Zufallsvariable dann, wenn sie nur endlich viele oder abzählbar unendlich viele Werte annehmen kann. Dies ist typischerweise bei Zählgrössen der Fall. Falls die Zufallsvariable jeden reellen Wert in einem Intervall annehmen kann ist sie stetig. Mathematisch ist eine Zufallsvariable eine Funktion $X : \Omega \to \mathbb{N}$ oder $X : \Omega \to \mathbb{R}$.

Die folgenden Beispiele illustrieren das Konzept der Zufallsvariable:

U =	"Augenzahl bei 1x Würfelwurf"	$\in \{1,2,3,4,5,6\}$,	diskret
V =	"Augensumme bei 2x Würfelwurf"	$\in \{2,3,...,12\}$,	diskret
W =	"Anzahl Kopf bei 1x Münzenwurf"	$\in \{0,1\}$,	diskret
X =	"Anzahl Kopf bei 10x Münzenwurf"	$\in \{0,1,2,...,10\}$,	diskret
Y =	"Regenmenge in Zürich in 24h"	$\in [0,+\infty)$,	stetig

Die Wahrscheinlichkeitsverteilung einer Zufallsvariable gibt nun an, welche Werte die Zufallsvariable mit welcher Wahrscheinlichkeit annimmt. Die konzeptuelle Beschreibung ist deutlich einfacher im Falle von diskreten Zufallsvariablen, weshalb wir damit beginnen.

2.2 Diskrete Wahrscheinlichkeitsverteilungen

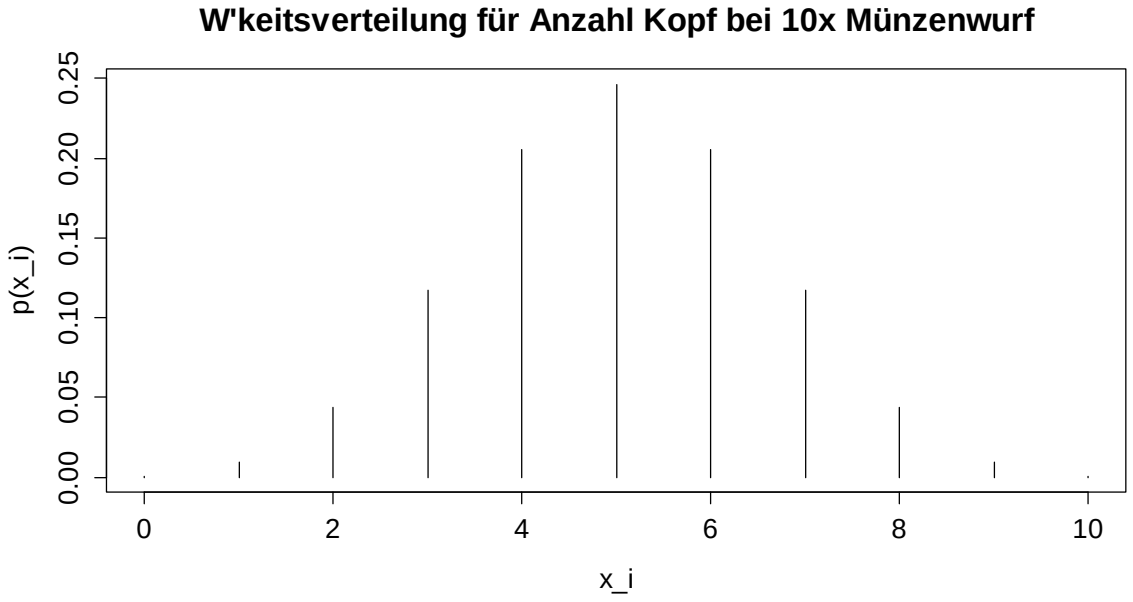
Es sei X eine beliebige diskrete Zufallsvariable. Wir bezeichnen die Werte, die X annehmen kann mit $x_1, x_2, x_3, \ldots$. Die zugehörigen Wahrscheinlichkeiten notieren wir mit $p(x_1), p(x_2), p(x_3), \ldots$. Es ist also:

$$p(x_i) = P(X = x_i) .$$

Weil $P(\Omega) = 1$ sein muss, gilt auch $\sum_i p(x_i) = 1$. Man kann die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariable mit einer Tabelle darstellen:

X	x_1	x_2	x_3	…	x_k
$p(\cdot)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	…	$p(x_k)$

Wie immer sind Tabellen aber nur mässig übersichtlich und es ist besser, eine grafische Darstellung zu wählen. Mit einem Stabdiagramm kann eine diskrete Wahrscheinlichkeitsverteilung viel besser wahrgenommen werden. Als Beispiel zeigen wir die Verteilung der Zufallsvariable Y' ="Anzahl Kopf bei 10x Münzenwurf".

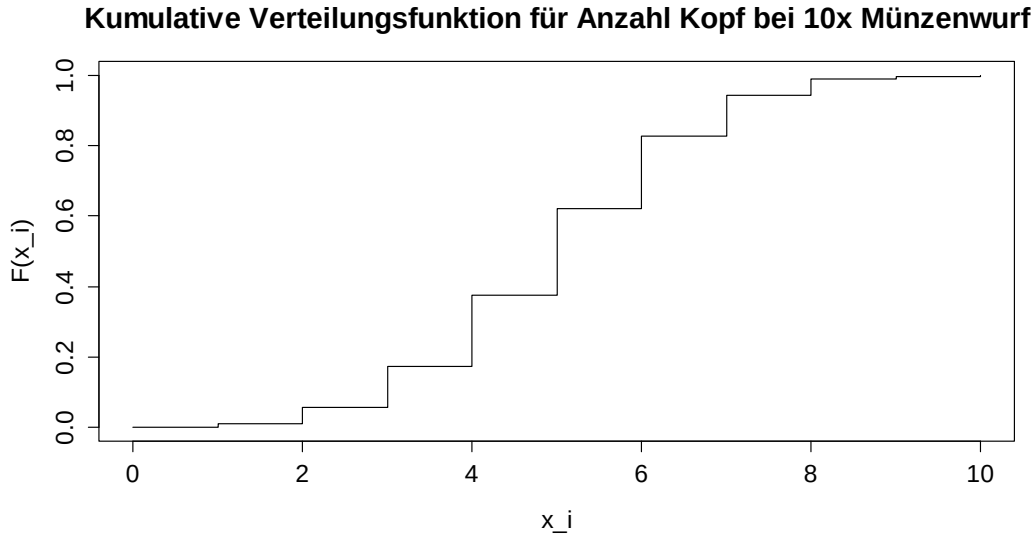


Auf der x-Achse wird der Wertebereich aufgetragen (0-10x Kopf), auf der y-Achse mit den Stäben die entsprechenden Wahrscheinlichkeiten. Wir werden in Kürze darauf zurückkommen, wie man diese bestimmt. Zu jeder (diskreten) Wahrscheinlichkeitsverteilung gibt es auch eine kumulative Verteilungsfunktion. Die mathematische Definition lautet:

$$F(x_i) = P(X \le x_i) .$$

Die kumulative Verteilungsfunktion gibt also die Wahrscheinlichkeit an, mit welcher die Zufallsvariable X einen Wert annimmt, der kleiner oder gleich einem vorgegebenen Wert x_i ist. Für $F(x_i)$ gilt $F(-\infty) = 0$, bzw. $F(+\infty) = 1\, F(+\infty) = 1$.

Zudem ist eine kumulative Verteilungsfunktion immer monoton wachsend. Im Falle von diskreten Zufallsvariablen handelt es sich um eine Treppenfunktion:

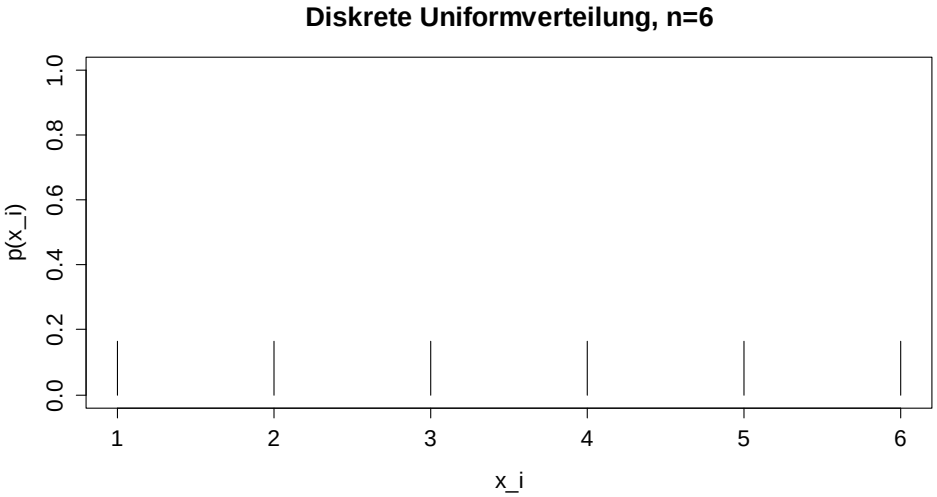


Als Beispiel sehen wir im Plot die kumulative Verteilungsfunktion der Anzahl Kopf bei 10x Münzenwurf. Wir können daraus z.B. Ablesen, dass die Wahrscheinlichkeit für 6x oder weniger Kopf bereits über 80% beträgt.

Grundsätzlich hat jede (diskrete) Zufallsvariable ihre eigene, individuelle Verteilung. In gewissen Situationen ist diese einfacher zu bestimmen, in anderen wird das schon schwieriger. Wie wir aber in der Folge sehen werden, tauchen einige prototypische Verteilungsfamilien bzw. Wahrscheinlichkeitsmodelle immer wieder auf. Wir führen diese nun der Reihe nach ein.

Diskrete Uniformverteilung

Bei der diskreten Uniformverteilung hat *jedes Elementarereignis exakt dieselbe Wahrscheinlichkeit*. Sie ist gleich $1/n$, wobei n die Anzahl Elemente von Ω ist. Als Beispiel können wir die Augenzahl bei 1x Würfelwurf heranziehen. Jede Seite liegt mit derselben Wahrscheinlichkeit oben, somit ist die Verteilung



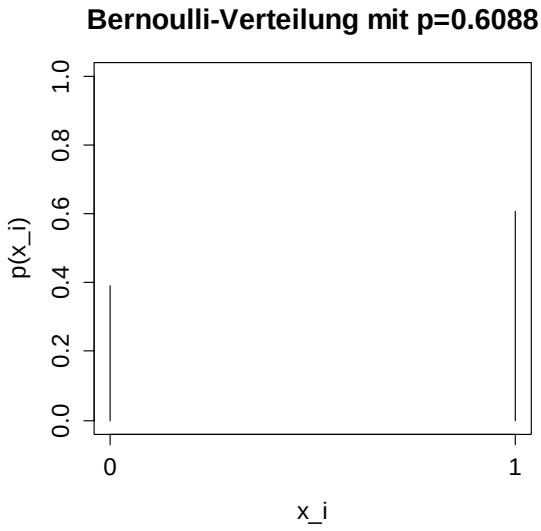
Bernoulli-Verteilung

Die Bernoulli-Verteilung kommt immer dann zur Anwendung, wenn ein Zufallsexperiment nur zwei mögliche Ergebnisse hat, nämlich „Misserfolg" und „Erfolg", bzw. „Nein" und „Ja", etc. Die entsprechende Zufallsvariable X codiert man mit 0 und 1. Ein wichtiger Begriff im Zusammenhang mit der Bernoulli-Verteilung ist *Erfolgswahrscheinlichkeit* $p = P(X = 1)$. Sie ist der Parameter, welcher die Verteilung exakt spezifiziert. Wir schreiben in diesem Fall $X \sim Bernoulli(p)$.

Als Beispiel betrachten wir die Zufallsvariable X ="Ein Kandidat besteht die Fahrprüfung". Wir codieren das Ergebnis mit 0 für „nein" und 1 für „ja". Den Parameter p können wir aus vergangenen Daten schätzen. Im Kanton Zürich bestanden im Jahr 2011 z.B. 15'100 von 24'801 Kandidaten die Prüfung. Somit schätzen wir:

$$\hat{p} = \frac{15'100}{24'801} = 0.6088 = 60.88\%$$

Wir können auch diese Verteilungsfunktion grafisch darstellen. Weil es nur 2 verschiedene Werte gibt, sind nur 2 Stäbe vorhanden:



Binomial-Verteilung

Die Binomial-Verteilung tritt dann auf, wenn n unabhängige Bernoulli-Experimente mit jeweils konstanter Erfolgswahrscheinlichkeit p nacheinander ausgeführt werden, und wir uns für die gesamte Anzahl Erfolge interessieren. Als Beispiel betrachten wir einen Fahrlehrer, der aktuell 7 Kandidaten zur Prüfung angemeldet hat. Er fragt sich, wie viele davon wohl bestehen werden, und ihm so als Kunden verloren gehen. Die interessierende Zufallsvariable ist also X ="Anzahl Fahrschüler von 7, welche die Prüfung bestehen". Das Bestehen jedes einzelnen Prüflings stellt ein Bernoulli-Experiment mit $p = 0.6088$ dar. Aber natürlich können alle Fahrschüler bestehen, oder auch alle durchfallen. Der Wertebereich von X ist $\{0,1,2,3,4,5,6,7\}$.

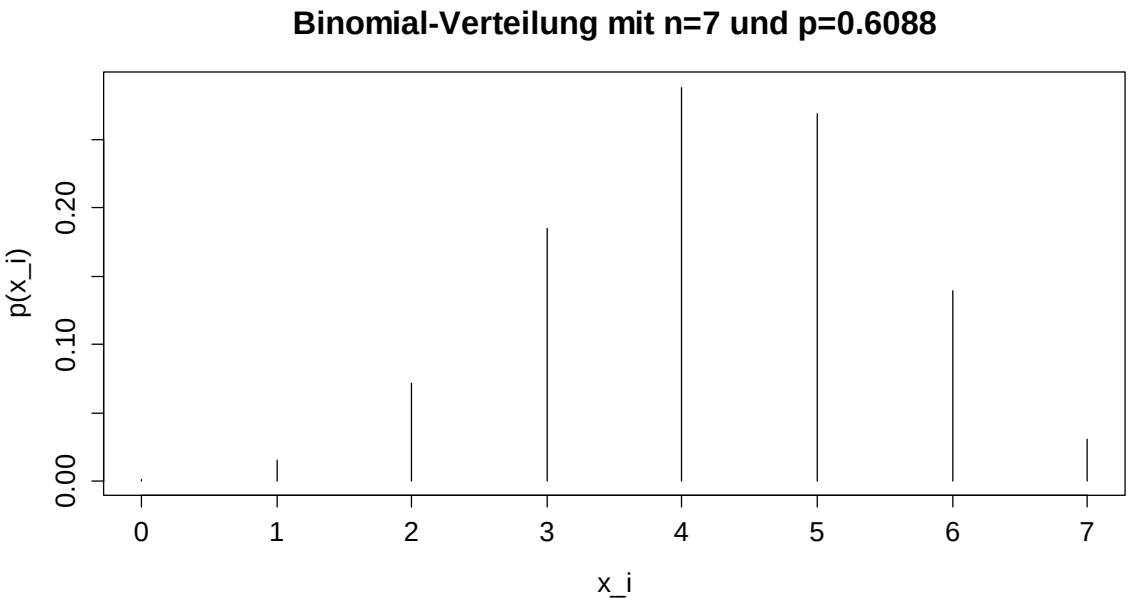
Es sei nun n die Anzahl Versuche (im Bsp. $n=7$), p sei die Erfolgswahrscheinlichkeit pro Versuch (im Bsp. $p=0.6088$) und k sei die Anzahl Erfolge, die uns interessiert. Dann gilt:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \text{ für } k=0,1,2,...,n$$

Hierbei charakterisiert $p^k(1-p)^{n-k}$ die Wahrscheinlichkeit, in n Versuchen eine bestimmte Konfiguration mit k Erfolgen und $n-k$ Misserfolgen zu haben. Der Term „n tief k" gibt schliesslich an, wie viele unterscheidbare Konfigurationen mit k Erfolgen in n Versuchen es gibt. Letzteres kann man als Auswahlproblem sehen: es gilt, die k erfolgreichen Fahrschüler aus den total n zur Verfügung stehenden zu bestimmen. Wir berechnen als Beispiel $P(X=5)$:

$$P(X=5)=\binom{7}{5}(0.6088)^5(1-0.6088)^2=0.2688.$$

Am interessanten ist die grafische Darstellung der Verteilung. Sie ist wie folgt:



Wenn eine Zufallsvariable binomialverteilt ist, so schreiben wir kurz $X \sim Bin(n, p)$.

Poisson-Verteilung

Ein weiteres, wichtiges, universell einsetzbares und häufig gebrauchtes diskretes Modell ist die Poisson-Verteilung. Wenn die einzelnen Ereignisse unabhängig voneinander mit einer konstanten Rate λ passieren, so hat die Zufallsvariable

$$X = "Anzahl \;\; Ereignisse"$$

eine Poisson-Verteilung mit Parameter λ. Man interessiert sich also für die Anzahl Vorkommnisse in einer bestimmten Zeitspanne, in einem festgelegten Gebiet/Einheit, oder ähnlichem. Wir schreiben $X \sim Pois(\lambda)$.

Typische Beispiele sind Defekte in Geräten, an Fahr- und Flugzeugen, Unfälle auf der Strasse, in der Luft oder einer Fabrik, das Eintreffen von Klienten an einem Schalter, et cetera. Die Wahrscheinlichkeit für k Ereignisse ist gegeben durch:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \text{ für } k = 0,1,2,...$$

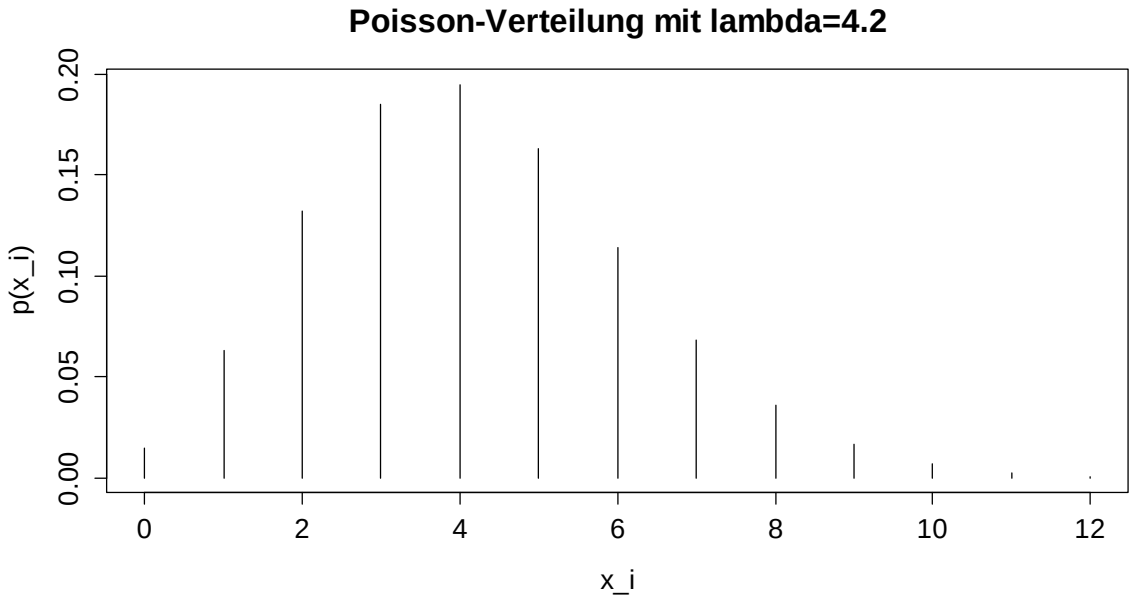
Auf die Herleitung der Formel verzichten wir an dieser Stelle. Wir weisen aber nochmals auf die Unterschiede zur Binomialverteilung hin:

- wir haben es nicht mehr mit einer bekannten Zahl von n Einzelversuchen zu tun, sondern die Grösse der Population ist unbekannt.
- der Wertebereich der Zufallsvariablen, d.h. die Anzahl Vorkommnisse, die wir beobachten können, ist nicht mehr auf $0,1,...,n$ beschränkt, sondern umfasst theoretisch alle natürlichen Zahlen und ist unendlich abzählbar.
- wir kennen nicht mehr wie bei der Binomialverteilung die Erfolgswahrscheinlichkeit p für den Einzelversuch, sondern nur noch die Rate λ, mit welcher das Ereignis auftritt.

Ein typisches Beispiel zur Poisson-Verteilung ist die Anzahl tödlicher Segelflugunfälle in der Schweiz. In den vergangenen 20 Jahren haben sich insgesamt 84 tödliche Segelflugunfälle ereignet. Pro Jahr ergibt dies im Schnitt 4.2 tödliche Unfälle. Wir setzen dies gleich dem Parameter λ. Es gilt also:

$$X = "Anzahl\ \ tödliche\ \ Segelflugunfälle\ \ pro\ \ Jahr\ \ in\ \ der\ \ CH" \sim Pois(4.2)$$

Die zugehörige Wahrscheinlichkeitsverteilung ist unten abgebildet. Es ist wichtig zu wissen, dass auch $P(X = 13)$, $P(X = 14)$, … Werte grösser als null annehmen. Allerdings werden diese Wahrscheinlichkeiten mit steigendem k sehr rasch extrem klein.



2.3 Stetige Wahrscheinlichkeitsverteilungen

Gegeben sei eine stetige Zufallsvariable, also eine Grösse, die jeden beliebigen Wert eines Intervalls annehmen kann, bzw. zumindest annehmen könnte. Diese treten in der Regel überall dort auf, wo gemessen wird (*Länge*, *Gewicht*, *Temperatur*, *Zeit*, …). Wir werden auf drei Beispiele Bezug nehmen:

$X =$ "Wartezeit, bis der nächste Kunde an den Bahnschalter kommt"
$Y =$ "Hämatokrit-Wert eines Menschen (wichtig für Dopingproben)"
$Z =$ "Regenmenge an einem Tag, wo es Niederschlag gibt"

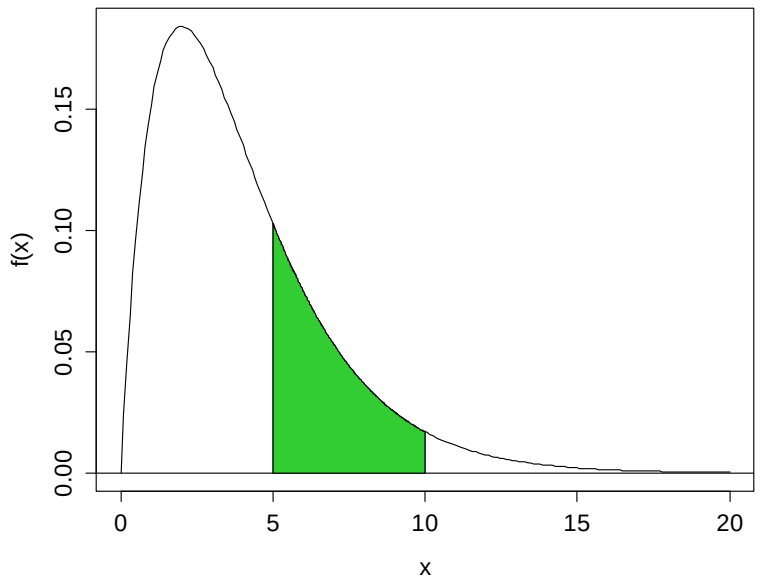
Wir wollen nun die Wahrscheinlichkeitsverteilungen dieser stetigen Zufallsgrössen charakterisieren. Das bisherige Stabdiagramm eignet sich aber nicht mehr, denn wir bräuchten unendlich viele, sehr dicht liegende Stäbe. Die sogenannte Dichtefunktion liefert genau das, was wir brauchen. Die Definition ist wie folgt:

Def: Die Verteilung einer stetigen Zufallsvariable X ist charakterisiert durch die *Dichtefunktion* $f(x)$, welche den folgenden beiden Bedingungen genügt:

- $f(x) \geq 0$ für alle $x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

Eine Dichtefunktion muss also überall positiv sein, und das Integral über die ganze Kurve ist eins. Man sei sich aber bewusst, dass die Werte der Dichtefunktion $f(x)$ keine Wahrscheinlichkeiten sind, insbesondere ist auch $f(x) > 1$ erlaubt.

Beispiel einer Dichtefunktion



Wenn man Wahrscheinlichkeiten berechnen will, so sind diese als Integrale über die Dichtefunktion definiert:

$$P(a \leq X < b) = \int_a^b f(x)\,dx \in [0,1]$$

Im Beispiel auf der vorangehenden Seite wäre also z.B. $P(5 \leq X < 10)$ von Interesse. Der zugehörige Wert ist die grün markierte Fläche, welche durch Integration über die Dichtekurve bestimmt werden muss. Wir verzichten an dieser Stelle auf die Rechnung, und halten einfach fest, dass das Resultat 0.247 beträgt. Sehr wichtig ist die Feststellung, dass das Ereignis $P(X = a)$ keine Wahrscheinlichkeit trägt, denn es gilt ja:

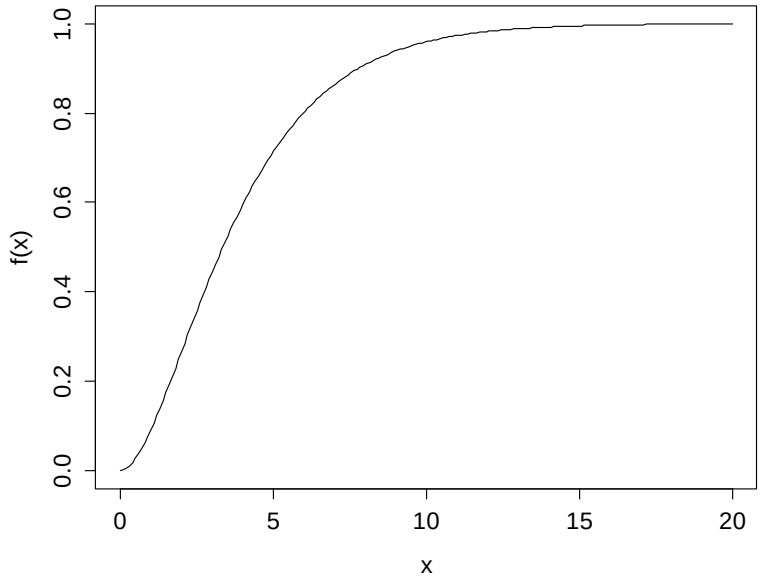
$$P(X = a) = \int_a^a f(x)\,dx = 0\,.$$

Aus diesem Sachverhalt leitet sich auch ab, dass wir bei stetigen Zufallsvariablen (*im strikten Gegensatz zu den diskreten!!!*) keine Unterscheidung zwischen "$\leq$" und "$<$" machen müssen, d.h. es gilt $P(a \leq X \leq b) = P(a < X < b)$. Zu jeder Dichtefunktion kann man auch die *kumulative Verteilungsfunktion* $F(x) = P(X \leq x)$ definieren. Diese erhält man durch Integration vom linken Rand bis zur Stelle x.

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(z)\,dz \in [0,1]$$

Die kumulative Verteilungsfunktion einer stetigen Zufallsvariable ist monoton wachsend, zudem gelten $F(-\infty) = 0$ und $F(+\infty) = 1$. Sie eignet sich vor allem zum Ablesen von Wahrscheinlichkeiten und Quantilen einer Verteilung. Für unsere beispielhafte Dichtefunktion von oben ist $F(x)$ wie folgt:

Beispiel einer kumulativen Verteilungsfunktion



Natürlich ist die Dichtefunktion $f(x)$ die Ableitung der Verteilungsfunktion $F(x)$. Umgekehrt ist $F(x)$ die Stammfunktion von $f(x)$. Grundsätzlich hat jede stetige Zufallsvariable ihre eigene, typische Verteilung. Wir stellen im Folgenden einige prototypische Verteilungsfamilien vor, die universell einsetzbar sind.

Exponentialverteilung

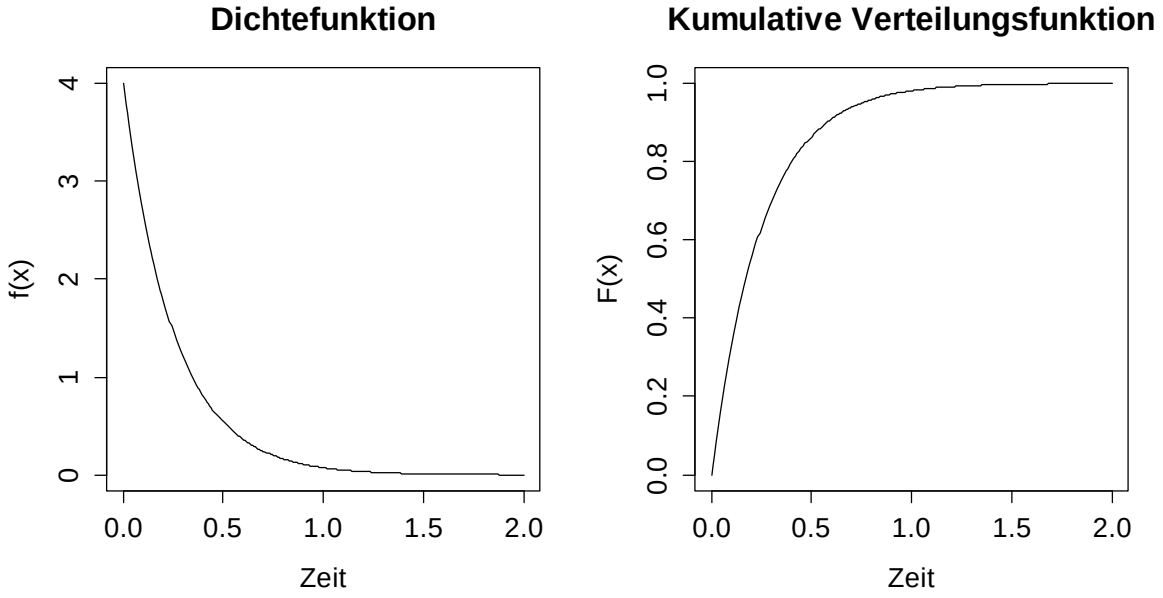
Ein wichtiges, stetiges W'keitsmodell ist durch die *Exponentialverteilung* gegeben. Sie eignet sich für alle Zufallsvariablen X, welche eine *Wartezeit* oder eine *Lebensdauer* beschreiben. Definiert ist sie durch die Dichtefunktion

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

bzw. ihre kumulative Verteilungsfunktion

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

Dabei steht der Parameter λ für die erwartete Anzahl an Ereignissen pro Zeiteinheit, oder umgekehrt ist $1/\lambda$ die mittlere Lebensdauer bzw. Wartezeit. Dichtefunktion und kumulative Verteilungsfunktion sehen bei $\lambda = 4$ wie folgt aus:



Die Einheit auf der x-Achse ist in beiden Fällen die Warte- oder Lebenszeit. Auf der y-Achse haben wir Dichte (links) bzw. Wahrscheinlichkeit (rechts). Offensichtlich ist die Exponentialverteilung stark rechtsschief, d.h. kurze Warte- oder Lebenszeiten sind häufig. Lange Warte- oder Lebenszeiten kommen zwar auch vor, sind aber selten. Wichtig ist es auch zu wissen, dass rechts ausserhalb des Darstellungsbereichs (d.h. für $x > 2$) weder die Dichte gleich 0, noch die Wahrscheinlichkeit gleich 1 ist. Weil es aber nur wenige Wartezeiten grösser als 2 Minuten gibt, macht es trotzdem Sinn, den Darstellungsbereich einzuschränken.

Doch warum wählt man gerade die Exponentialverteilung als Modell für Warte- und Lebenszeiten? Der Grund liegt in der mathematischen Einfachheit und Klarheit, bzw. der Verwandtschaft zum sogenannten *Poisson-Prozess*. Es lässt sich nämlich zeigen, dass folgendes gilt:

> Beschreibt die Zufallsvariable X die Wartezeit zwischen zwei Ereignissen und folgt einer $\exp(\lambda)$-Verteilung, so gilt für die Zufallsvariable $Y =$ "Anzahl Ereignisse in t Zeiteinheiten" eine Poissonverteilung mit Parameter λt, d.h. aus $X \sim \exp(\lambda)$ folgt $Y \sim Pois(\lambda t)$.

Dies impliziert, dass die sogenannte *Ausfallrate* (engl. *Hazard Rate*), also die Wahrscheinlichkeit, dass im nächsten Moment etwas passiert konstant und unabhängig davon ist, wie lange man schon gewartet hat. Wegen dieser Eigenschaft sagt man auch, dass die Exponentialverteilung „gedächtnislos" sei, denn es spielt ja für die weitere Betrachtung keine Rolle, wie lange das System schon in Betrieb ist. Natürlich gibt es auch komplexere, schwierigere und für manche Situationen realistischere Modelle wie z.B. die *Weibull-Verteilung*, welche eine nicht-konstante Ausfallrate haben (z.B. hoch zu Beginn, tief nach einer gewissen Zeit, dann wieder ansteigend). Sie werden hier jedoch nicht besprochen.

Als Beispiel stellen wir uns einen Schalter vor, an welchem im langfristigen Schnitt 4 Kunden pro Minute eintreffen, d.h. alle 15 Sekunden bzw. jede Viertelminute einer. Wir betrachten die Zufallsvariable $X =$"Wartezeit zwischen dem Eintreffen zweier Kunden". Wir können dafür eine Exponentialverteilung ansetzen, deren Parameter λ schätzen wir aus dem beobachteten Mittelwert, d.h.:

$X \sim \exp(\lambda)$, wobei λ auf $\hat{\lambda} = 4$ gesetzt wird.

Der Vorteil am Modell der Exponentialverteilung liegt darin, dass wir nur mittlere Wartezeit (bzw. die mittlere Anzahl Ereignisse pro Zeiteinheit) zu kennen brauchen, um die Wahrscheinlichkeit einer beliebigen Wartezeit angeben zu können. Andererseits erhalten wir via den Link zum Poisson-Prozess auch sogleich die Verteilung für die Anzahl eintreffender Kunden in einer beliebigen Zeit t als Zufallsvariable $Y \sim Pois(\lambda t)$, d.h. konkret $Y \sim Pois(4t)$. Wir wollen nun ein konkretes Beispiel durchrechnen und die Wahrscheinlichkeit berechnen, dass die Wartezeit zwischen 15 und 30 Sekunden beträgt. Diese erhalten wir durch Integration über die Dichtefunktion:

$$P(0.25 \le X < 0.50) = \int_{0.25}^{0.50} f(z)\, dz = \left[-e^{-4z} \right]_{0.25}^{0.50} = -e^{-2} + e^{-1} = 0.233.$$

Falls wir die kumulative Verteilungsfunktion auch kennen (was ja hier der Fall ist), so lässt sich die entsprechende Wahrscheinlichkeit auch damit, und daher ohne Integration berechnen. Es gilt nämlich:

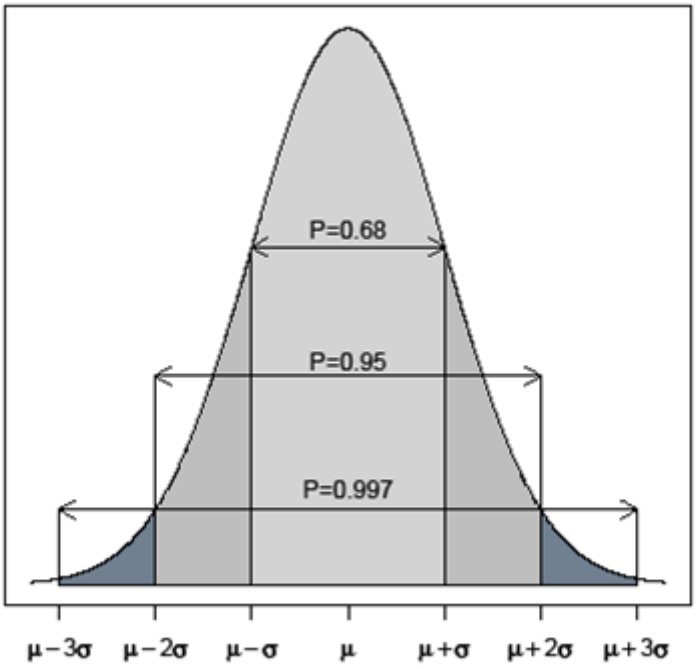
$$P(0.25 \le X < 0.50) = F(0.50) - F(0.25) = (1 - e^{-2}) - (1 - e^{-1}) = -e^{-2} + e^{-1} = 0.233$$

Dies gilt für stetige Zufallsvariablen universell, d.h. $P(a \le X < b) = F(b) - F(a)$.

Normalverteilung

Die Normalverteilung ist die wichtigste stetige Verteilung. Sie wurde als Modell für Messwerte bzw. Messfehler entwickelt und passt dort meist sehr gut. Doch das ist nicht der einzige Grund: wie wir später sehen werden, hat die Summe von stochastisch unabhängigen Zufallsvariablen mit beliebiger, identischer Verteilung approximativ eine Normalverteilung. Sie kann daher stets zum Einsatz kommen, falls sich ein Resultat aus einer additiven Überlagerung von vielen „kleinen Einflüssen" ergibt. Die Normalverteilung mit ihrer Glockenform ist also so etwas wie eine „Naturkonstante" in der Welt der Zufallsvariablen. Die Dichtefunktion einer Zufallsvariable X mit Normalverteilung $N(\mu, \sigma^2)$ ist gegeben durch:

$$f(x) := \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\left\{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}\right\} \text{ für } x\in(-\infty,+\infty).$$



Die Normalverteilung hat offensichtlich zwei Parameter. Der Lageparameter $\mu\in(-\infty,+\infty)$ ist für Verschiebungen der Glockenkurve auf der x-Achse zuständig, ist Zentrum und Symmetriepunkt der Verteilung, und ist gleichzeitig auch der Erwartungswert $E[X]$. Der Skalenparameter σ^2 bestimmt die „Breite" der Verteilung, d.h. streckt oder staucht die Glockenkurve und ist gleichzeitig die Varianz $Var(X)$. Für die Interpretation ist die sogenannte Standardabweichung σ zugänglicher. Bei $\mu\pm\sigma$ befinden sich die Wendepunkte der Dichtefunktion und das Intervall $[\mu-\sigma,\mu+\sigma]$ enthält rund 68% der Wahrscheinlichkeit (d.h. der gesamten Fläche unter der Kurve). Im Intervall $[\mu-2\sigma,\mu+2\sigma]$ sind es rund 95%, und innerhalb von $\mu\pm3\sigma$ liegen bereits ca. 99.7% der Wahrscheinlichkeit.

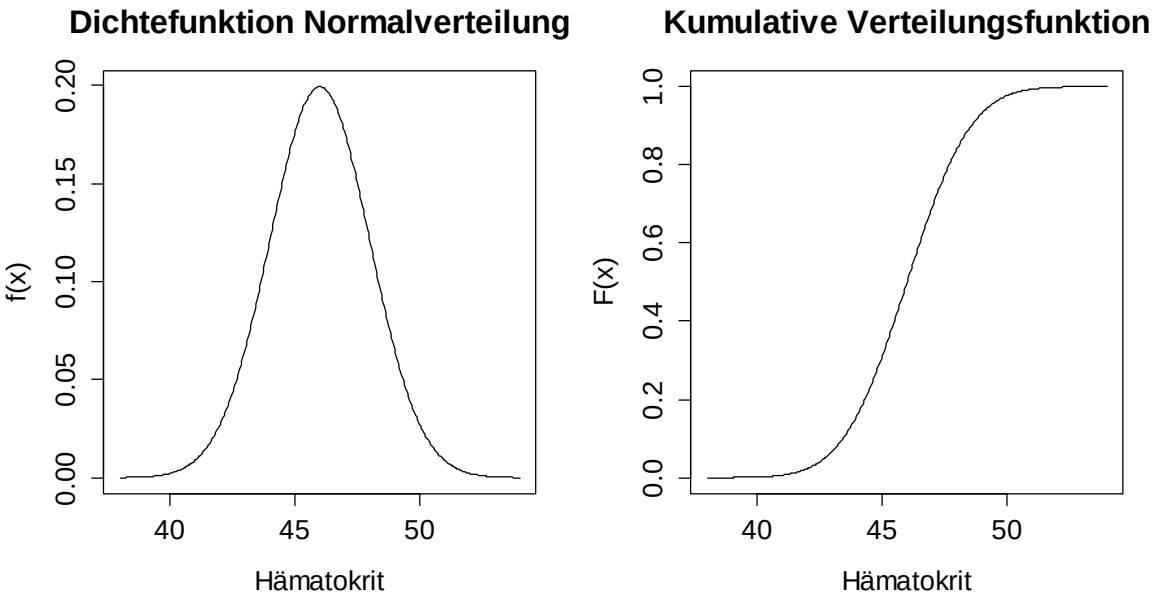
Um diese Wahrscheinlichkeiten zu berechnen, muss man bekanntlich die Dichtefunktion integrieren. Dies ist im Fall der Normalverteilung aber schwierig, gibt es doch keine Stammfunktion

$$F(x) = P(X\le x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\left(-\frac{1}{2}\cdot\frac{(z-\mu)^2}{\sigma^2}\right)dz$$

welche man in geschlossener Form schreiben könnte. Die Integrale existieren aber, und man kann/muss zur Berechnung auf numerische Methoden zurückgreifen. Während früher häufig mit Tabellen gearbeitet wurde, verwendet man heute in der Regel Computerprogramme.

Als Beispiel betrachten wir den Hämatokrit. Er bezeichnet den Anteil der *Erythrozyten* am Volumen des Blutes. Diese stellen rund 99% des

Gesamtvolumens der Blutzellen dar, somit entspricht der Hämatokrit ungefähr dem Anteil des Zellvolumens im Blut und gibt Aufschluss über den Wasserhaushalt einer Person. Bei Menschen liegen die normalen Hämatokrit-Werte zwischen 42% und 50%. Wir können also den Hämatokrit mit einer Normalverteilung beschreiben, wo $\mu = 46$ und $\sigma = 2$. Dichte und kumulative Verteilungsfunktion sind dann wie folgt:



Im Nordischen Skisport wurde durch Doping-Kontrollbehörden ein Hämatokrit-Grenzwert von 51.5% festgelegt. Überschreitet ein Athlet diese Schwelle, so ist er nicht mehr zu den Wettkämpfen zugelassen. Wir können nun aufgrund der obigen Verteilung die Wahrscheinlichkeit $P(X > 51.5) = 0.3\%$ berechnen, mit welcher ein Athlet (ohne Manipulation) die Schwelle überschreitet. In der Praxis, d.h. wenn eine Stichprobe von Beobachtungen vorliegt, schätzt man die beiden Parameter μ und σ^2 oftmals aus den Daten durch Mittelwert und Stichproben-Varianz, d.h.:

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} x_i \text{ und } \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 ,$$

Lognormal-Verteilung

Die Lognormal-Verteilung ist ein Modell für Zufallsgrössen, welche nur strikt positive Werte annehmen können. Beispiele in der Natur und im Alltag sind zahlreich: so die in der Einleitung erwähnte Zufallsgrösse

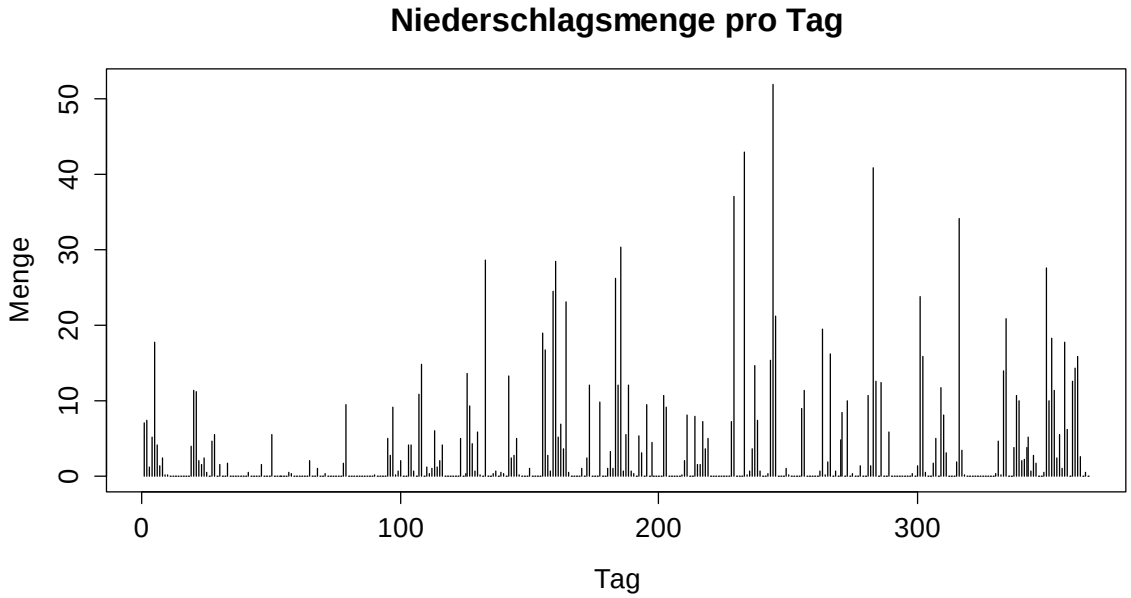
$Z =$ "Regenmenge an einem Tag, wo es Niederschlag gibt",

oder auch das Einkommen einer erwerbstätigen Person, die Nettokosten einer Person in der Krankenversicherung, et cetera. In all diesen Beispielen wird die Zufallsvariable meistens einen eher kleinen Wert annehmen, in seltenen Fällen stellt sich auch einmal ein grosser Wert ein. Die Verteilung ist also rechtsschief. Etwas abstrakter kann man sagen, dass sie sich überall dort eignet, wo Unterschiede besser durch Verhältnisse als durch Differenzen ausgedrückt

werden, bzw. sich das Resultat aus einer multiplikativen Überlagerung von vielen kleinen Einflüssen ergibt. Zur Definition der Lognormal-Verteilung gilt:

Es gilt $Z \sim \log N(\mu, \sigma^2)$, genau dann wenn $Z' = \log(Z) \sim N(\mu, \sigma^2)$.

Wir sprechen also von einer Lognormal-Verteilung, wenn die mit dem Logarithmus transformierten Daten einer Normalverteilung folgen. Dies ist auch der einfachste Weg, um die Parameter μ und σ^2 zu schätzen. Als Beispiel haben wir hier die tägliche Regenmenge (Einheit mm) in Winterthur im Jahr 2012:

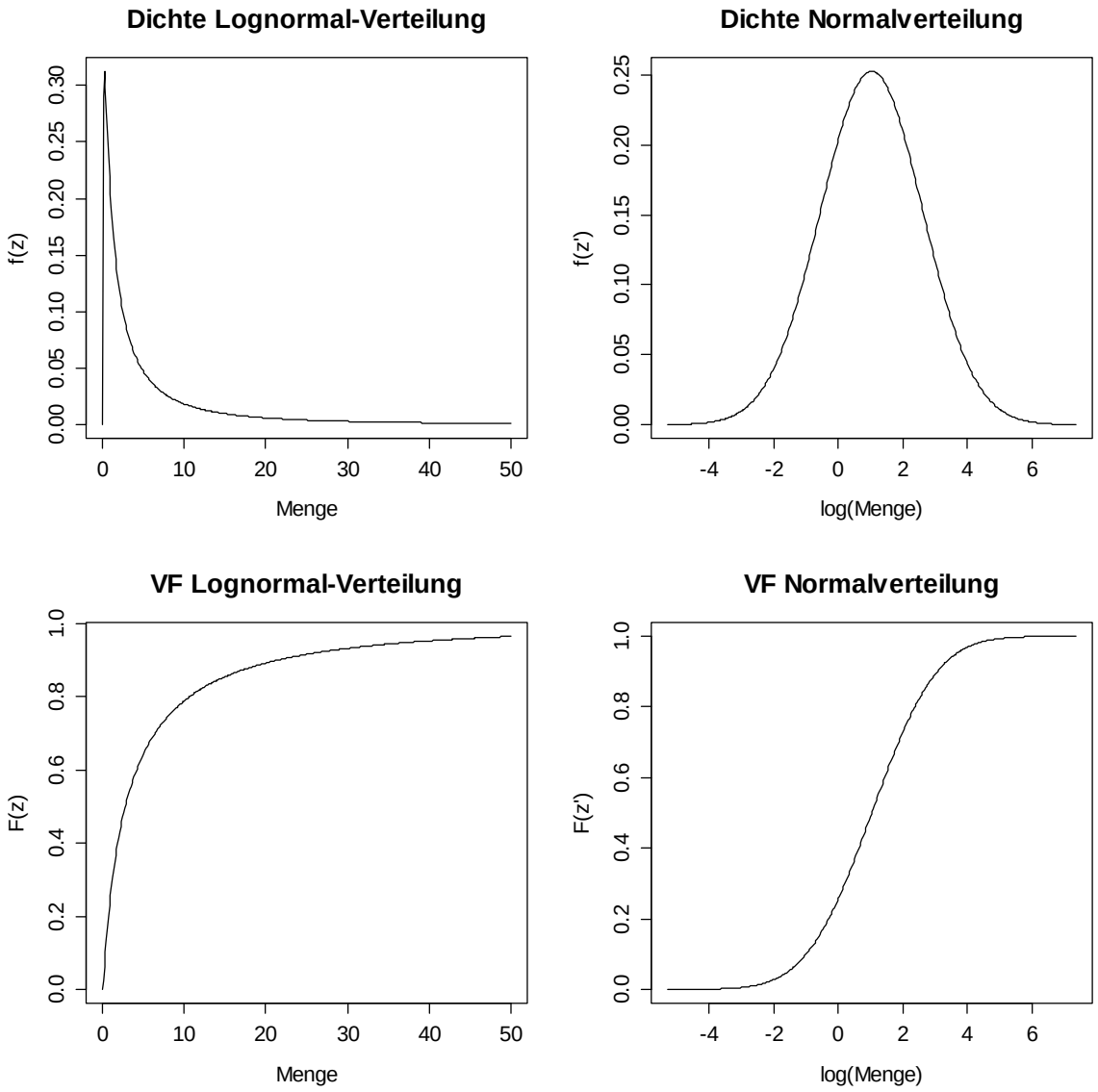


Wir sehen, dass es einerseits viele trockene Tage gibt (welche wir in Hinsicht auf die Zufallsvariable Z nicht betrachten), dann etliche mit kleinen Mengen und einige wenige mit grossen bis sehr grossen. Wir schätzen die Parameter μ und σ^2 durch Mittelwert und Stichproben-Varianz der logarithmierten Daten. Als Werte ergeben sich $\hat{\mu} = 1.033$ und $\hat{\sigma}^2 = 2.495$.

Mit diesen beiden Parametern können wir nun sowohl die Dichtefunktion der Normalverteilung für die logarithmierten Daten, wie auch jene der Lognormal-Verteilung für die Originaldaten aufzeichnen, siehe nächste Seite. Ebenso sind dort auch die kumulative Verteilungsfunktion der Lognormal-Verteilung sowie der zugehörigen Normalverteilung für die logarithmierten Daten aufgetragen. Wenn wir uns nun fragen, wie gross die Wahrscheinlichkeit ist, dass an einem Regentag weniger als 10mm Niederschlag fallen, so berechnen wir am einfachsten:

$$P(Z' \leq \log(10)) = 78.92\%$$

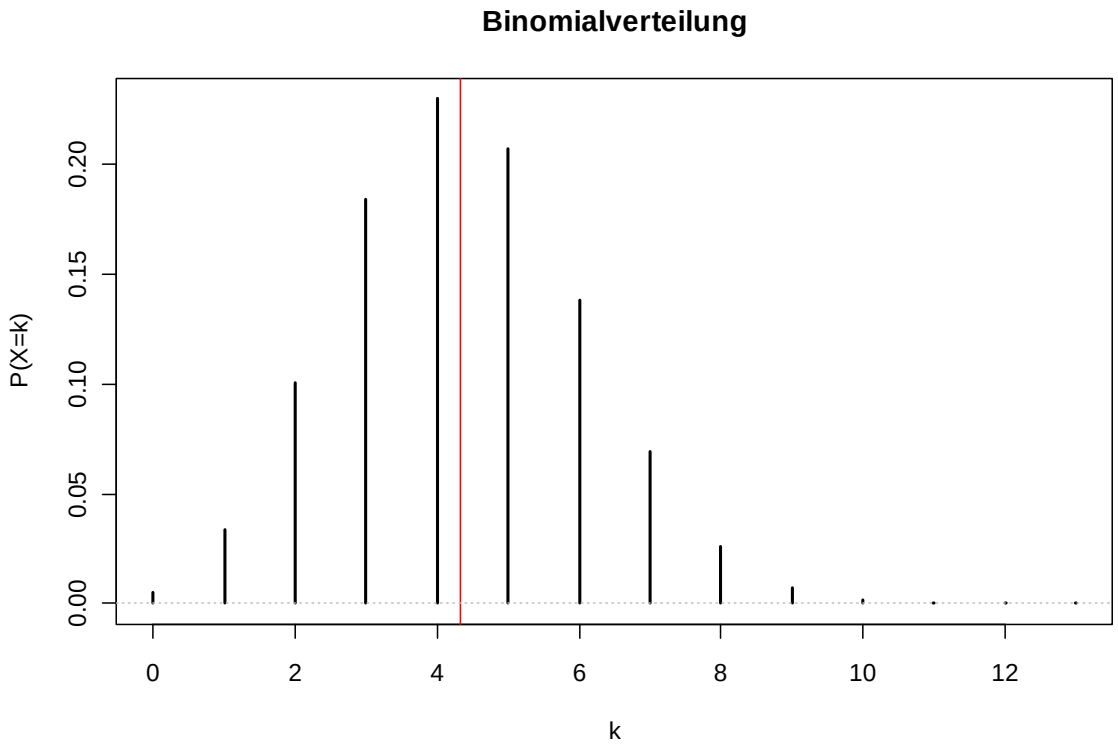
Wir beschliessen das Kapitel mit der Erkenntnis, dass die Logarithmus-Transformation beim Umgang mit Daten ein wichtiges Hilfsmittel ist.



2.4 Erwartungswert und Varianz

Falls möglich, ist es stets sinnvoll, die diskrete Wahrscheinlichkeitsfunktion (das Stabdiagramm) bzw. die Dichtefunktion komplett und vollständig anzugeben. Manchmal ist dies aber schwierig, und man begnügt sich mit der Angabe der Kenngrössen Erwartungswert und Varianz. Auch für das tiefere Verständnis und einige theoretische Betrachtungen sind die beiden Begriffe unerlässlich. Zu verstehen sind sie aber leicht: genau so wie Mittelwert und Stichproben-Varianz die Lage und Streuung einer Datenreihe charakterisieren, sind Erwartungswert und (theoretische) Varianz Kenngrössen einer Verteilung.

Der Erwartungswert $E[X]$ einer Zufallsvariable ist das Resultat, was man „im Schnitt", bei unendlich vielen Realisierungen, erhält. Er ist ein Lagemass, bildlich gesprochen handelt es sich um die „Position" der Verteilung auf der x-Achse. Oder geometrisch: es ist der Schwerpunkt (in x-Richtung) der Dichte- bzw. Wahrscheinlichkeitsfunktion. Somit ist auch klar, worin der Schlüssel zur Berechnung des Erwartungswerts liegt. Die Grafik zeigt eine $Bin(13,1/3)$-Verteilung, für welche der Erwartungswert $13/3$ in roter Farbe markiert ist.



Es sei X eine diskrete Zufallsvariable mit diskreter Wahrscheinlichkeitsfunktion $p(\cdot)$. Der Erwartungswert $E[X]$ ist definiert als:

$$E(X) = \sum_i x_i \cdot P(X = i) = \sum_i x_i p(x_i) = \sum_i x_i p_i \,,$$

falls die Summe existiert. Wir schreiben oft auch $\mu = E[X]$ und stellen fest, dass zur Berechnung die Höhe bzw. das Gewicht jedes Stabes mit seiner Position auf der x-Achse multipliziert wird, genauso wie wir dies in der Physik zur Berechnung eines Schwerpunkts machen würden.

Für eine stetige Zufallsvariable Z wird der Erwartungswert konzeptuell auf genau dieselbe Art und Weise berechnet. Jedoch gibt es dort unendlich viele „Stäbe" und die Summe ist eine infinitesimale, d.h. ein Integral. Konkret:

$$E(Z) = \int_{-\infty}^{+\infty} z \cdot f(z)\, dz \,,$$

falls das Integral existiert. Für viele wichtige Verteilungen wurde der Erwartungswert bereits berechnet und ist als Funktion der Parameter bekannt. Er kann aus der unten stehenden Tabelle entnommen werden.

Die Varianz $Var(X)$ ist ein Streuungsmass einer Zufallsvariable, charakterisiert also, wie breit die angenommenen Werte streuen. Es sei wiederum X eine diskrete Zufallsvariable mit Erwartungswert $E[X]$. Die Formel zur Berechnung der Varianz lautet:

$$Var(X) = \sum_i (x_i - \mu)^2 \cdot p(x_i) \,,$$

falls die Summe existiert, d.h. die „Stäbe" der Wahrscheinlichkeitsfunktion werden hier nun mit der quadrierten Differenz zwischen Erwartungswert und x-Wert

multipliziert, und aufsummiert. Für eine stetige Zufallsvariable Z mit Erwartungswert $E[Z]$ gilt:

$$Var(Z) = \int_{-\infty}^{+\infty} (z-\mu)^2 f(z)\, dz,$$

falls das Integral existiert. Die Standardabweichung von X bzw. Z ist die Wurzel aus der Varianz. Oft wird die Varianz mit σ^2 und die Standardabweichung mit σ bezeichnet. Für die üblichen parametrischen Verteilungsfamilien ist die Varianz als Funktion der Parameter bekannt. Wir stellen sie mit der folgenden Tabelle dar:

Verteilung		E[X] bzw. E[Z]	Var(X) bzw. Var(Z)
Bernoulli	$X \sim Bernoulli(p)$	p	$p(1-p)$
Binomial	$X \sim Bin(n, p)$	np	$np(1-p)$
Poisson	$X \sim Pois(\lambda)$	λ	λ
Exponential	$Z \sim Exp(\lambda)$	$1/\lambda$	$1/\lambda^2$
Normal	$Z \sim N(\mu, \sigma^2)$	μ	σ^2
Lognormal	$Z \sim \log N(\mu, \sigma^2)$	$e^{(\mu+0.5\sigma^2)}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$

Es ist nicht selten der Fall, dass wir an der Summe Z von zwei beliebigen Zufallsvariablen X und Y interessiert sind. Selbst wenn wir die Verteilung von X und Y kennen, so ist es nur in wenigen Ausnahmefällen einfach möglich, die Verteilung von Z zu spezifizieren. Immerhin lassen sich aber die Kenngrössen Erwartungswert und Varianz meist einfach bestimmen. Dies gilt ebenso für skalare Vielfache von Zufallsvariablen. Es seien also a, b skalare Grössen, und $Z = X + Y$ Zufallsvariablen. Für den Erwartungswert gelten stets die folgenden Rechenregeln:

$$E[aX + b] = aE[X] + b$$

$$E[Z] = E[X] + E[Y]$$

Für die Varianz ist die Sache komplizierter. Für skalare Lineartransformationen erhalten, bzw. die Summe von Zufallsvariablen erhalten wir:

$$Var(aX + b) = a^2 Var(X)$$

$$Var(Z) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$$

Hierbei ist $Cov(X, Y)$ die Kovarianz der beiden Zufallsvariablen. Die Varianz addiert sich beim Summieren von zwei Zufallsvariablen also nur dann, falls dieser Term den Wert 0 annimmt. Dies ist insbesondere dann der Fall, falls X und Y stochastisch unabhängig sind. Im nächsten Kapitel brauchen wir diese Rechenregeln wieder. Wir kümmern uns dort um die Summe von Zufallsvariablen und werden sehen, dass die dort erzielten Rechenregeln von fundamentaler Bedeutung für die Anwendung der statistischen Methoden in der Praxis sind.

2.5 Zentraler Grenzwertsatz

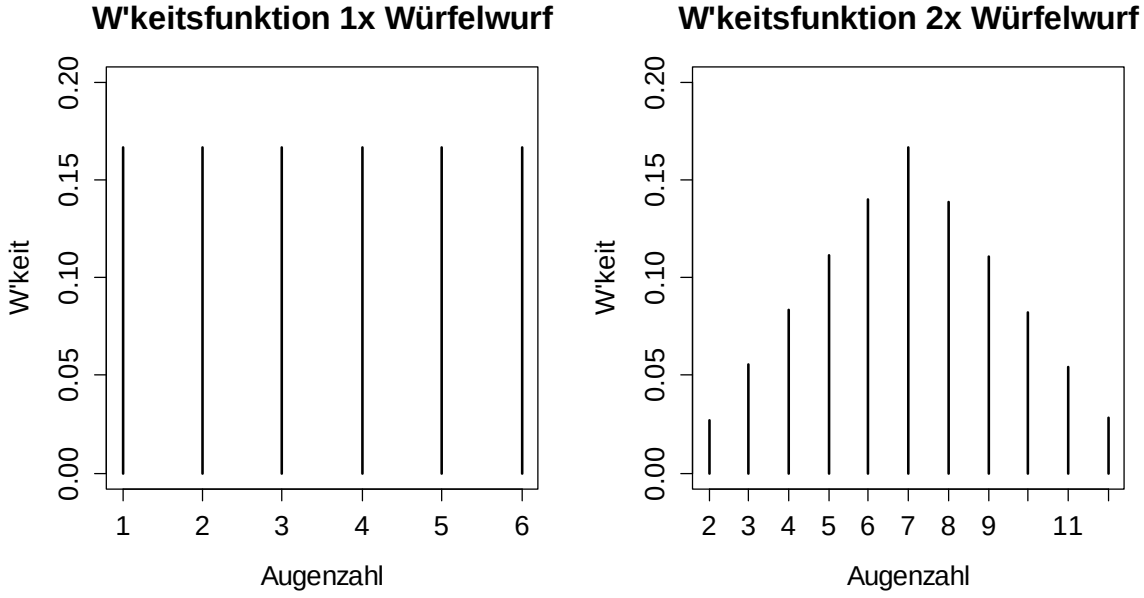
Wir kümmern uns hier zuerst um die Frage, welcher Verteilung die Summe von Zufallsvariablen folgt. Um die Sache möglichst einfach zu halten, beginnen wir mit $S = X_1 + X_2$, wobei X_1, X_2 zwei stochastisch unabhängige Zufallsvariablen mit identischer Verteilung seien. Gemäss den obigen Rechenregeln können wir problemlos ein Aussage zu Lage und Streuung von S machen, nämlich:

$$E[S] = E[X_1] + E[X_2] \text{ und } Var(S) = Var(X_1) + Var(X_2).$$

Über die exakte Form der Verteilung, bzw. deren diskrete W'keits- bzw. stetige Dichtefunktion können wir aber im Allgemeinen keine einfache Aussage treffen, sondern müssen diese (oft schwierig) durch die Faltungsformel bestimmen. Nur für einige wenige Spezialfälle bleibt nämlich die Summe S innerhalb derselben Verteilungsfamilie. Die uns bekannten Beispiele sind:

Falls $X_1, X_2 \sim Bin(n, p)$: $S \sim Bin(2n, p)$
Falls $X_1, X_2 \sim Pois(\lambda)$: $S \sim Pois(2\lambda)$
Falls $Z_1, Z_2 \sim N(\mu, \sigma^2)$: $S \sim N(2\mu, 2\sigma^2)$

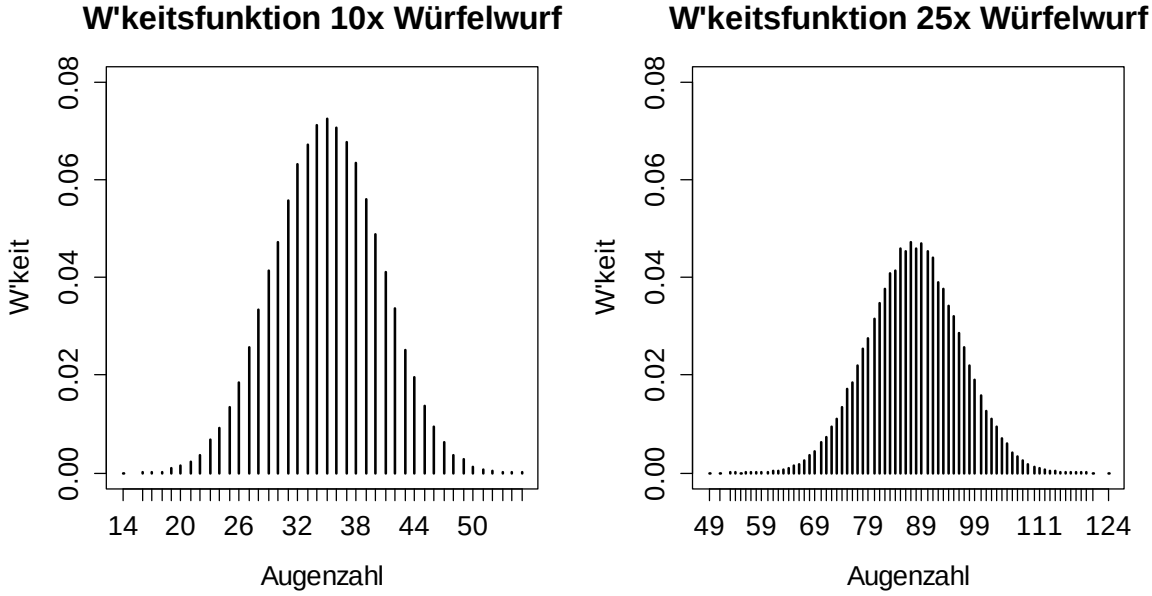
Bei allen anderen uns bekannten Verteilungen ist hingegen S nicht mehr innerhalb derselben Familie. Wir betrachten das Beispiel des einfachen und doppelten Würfelwurfs. Während wir beim einmaligen Werfen eine diskrete Uniformverteilung haben, ergibt sich beim doppelten Wurf eine sogenannte Dreiecksverteilung, also etwas grundlegend anderes. Dies kommt dadurch zustande, dass es für die Augenzahl 2 nur gerade die Kombination $(1,1)$ gibt, für die Augenzahl 4 sind hingegen $(3,1)$, $(2,2)$ und $(1,3)$ günstig.



Eine interessante Frage ist nun, was passiert wenn wir $n \gg 2$ Zufallsvariablen mit identischer Verteilung addieren. Darüber lässt sich eine allgemeine Konvergenzaussage treffen, die als Zentraler Grenzwertsatz bekannt ist. Er ist für die Statistik fundamental.

Satz: Seien $X_1,...,X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $E[X_i]=\mu<\infty$ und $0<Var(X_i)=\sigma^2<\infty$. Dann konvergiert die Verteilung der Summe S_n gegen eine Normalverteilung mit $E[S_n]=n\mu$ und $Var(S_n)=n\sigma^2$, d.h. $S_n \sim N(n\mu, n\sigma^2)$ für $n\to\infty$.

In der Praxis erreicht man, je nachdem wie stark sich die Verteilung der X_i von der Glockenform unterscheidet, meist schon bei $n\approx 10-50$ eine gute Näherung zur Normalverteilung. In besonders ungünstigen Fällen, d.h. Verteilungen mit extremen Ausreissern oder starker Schiefe, kann auch ein grösseres n nötig sein. Wir illustrieren den Zentralen Grenzwertsatz mit der Verteilung der Summe aus 10 und 25 Würfelwürfen. *Hinweis:* diese Verteilungen wurden nicht mehr theoretisch hergeleitet, sondern mittels einer Simulation aus 100'000 Würfen geschätzt.



Wie wir sehen können, ist die Glockenform beinahe perfekt. Allerdings handelt es sich immer noch um diskrete Verteilungen. Dennoch können Wahrscheinlichkeiten der Art $P(S_n \le s)$ dennoch gut durch die entsprechende (stetige) Normalverteilung abgeschätzt werden. Im Grenzübergang $n\to\infty$ wird der Unterschied zwischen stetiger und diskreter Verteilung hingegen irrelevant.

Von grosser Wichtigkeit sind auch die Folgerungen aus dem Zentralen Grenzwertsatz für den Mittelwert-Schätzer. Es seien wiederum $X_1,\ldots,X_n$ unabhängige und identisch verteilte Zufallsvariablen mit $E[X_i]=\mu<\infty$ und $0<Var(X_i)=\sigma^2<\infty$, also z.B. eine repräsentative Stichprobe aus derselben Grundgesamtheit. Dann ist auch der Mittelwert approximativ normalverteilt, d.h.:

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i \sim N\left(\mu,\frac{\sigma^2}{n}\right) \text{ für } n\to\infty$$

Eminent wichtig ist neben der approximativen Normalverteilung auch die Tatsache, dass die Varianz $Var(\bar{X}_n)\to 0$ für $n\to\infty$. Mit immer grösseren Stichproben kann der Erwartungswert μ also immer genauer durch den Mittelwert geschätzt werden. Dies ist salopp der Grund, warum „Statistik funktioniert".

3 Statistisches Testen

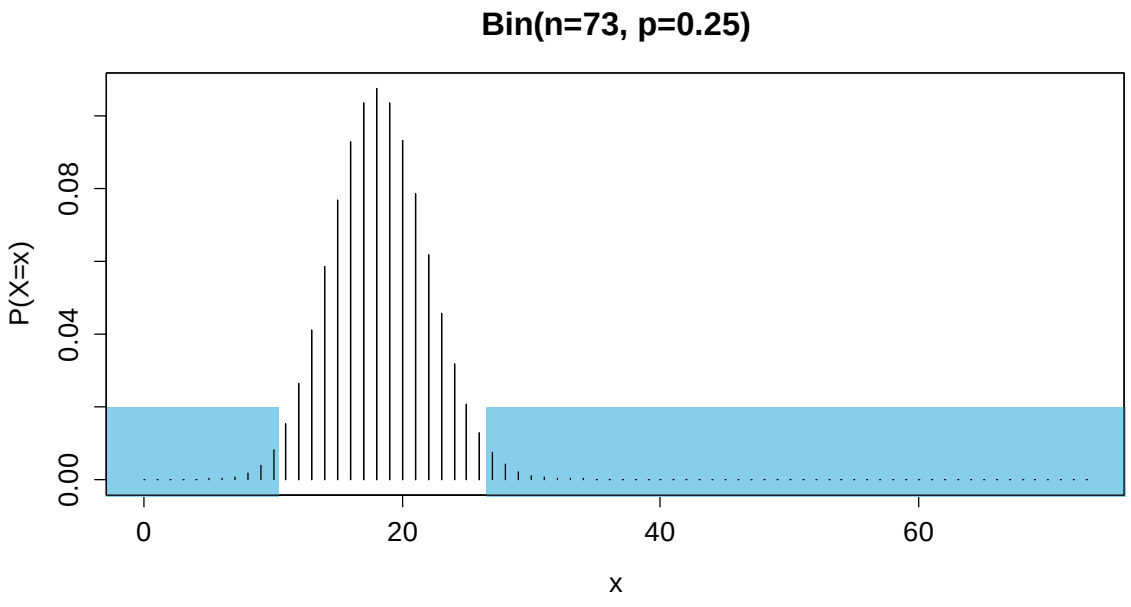
Mit einem *statistischen Test* wird beurteilt, ob der Unterschied zwischen einem Schätzwert auf einer Stichprobe und einem gegebenen Sollwert zufällig oder systematisch ist. Wie wir später sehen werden, lässt sich das Vorgehen auch Verallgemeinern, so dass auch zwei Stichproben, bzw. die daraus gewonnen Schätzwerte miteinander verglichen werden können. Für besseres Verständnis sollten wir uns vergegenwärtigen, dass zwei Messwerte (z.B. die Mittelwerte von zwei Stichproben derselben Grundgesamtheit) nie exakt übereinstimmen. Deshalb brauchen wir ein Verfahren, das uns solche Unterschiede einstufen lässt.

3.1 Binomialtest für einen Anteil p_0

Hier geht es darum zu prüfen, ob die auf einer Stichprobe vom Umfang n beobachtete Anzahl Erfolge X im Rahmen einer $Bin(n, p_0)$-Verteilung mit vorgegebener Wahrscheinlichkeit p_0 liegt. Wir erklären das Vorgehen am Beispiel.

Binomialtest für das Mendel-Experiment

Falls Mendels Vererbungsgesetze korrekt sind, so treten Erbsen mit runden Samen gegenüber solchen mit kantigen dreifach gehäuft auf. Durch eine Feldstudie soll dies verifiziert werden: untersucht wurden dabei 73 Samen, wovon 55 rund und 19 kantig waren. Dies entspricht einem Verhältnis von $2.89:1$. Nun sind dies nicht exakt $3:1$, was aber bei 73 Samen auch gar nicht möglich ist. Zudem ist beim Ausgangsmaterial für die Kreuzung ja jeweils nicht bekannt, ob die verwendeten Pflanzen homozygot oder heterozygot sind. Wir betrachten nun die Zufallsvariable $X =$"Anzahl kantige Samen". Falls Mendel's Theorie richtig ist, dann hätte diese auf unserer Stichprobe eine $Bin(73, 0.25)$-Verteilung. Dies nennen wir die sogenannte Nullhypothese $H_0 : p = p_0 = 0.25$. Wir analysieren nun die Verteilung von X unter dieser Annahme:



Die Verteilung ist um den Erwartungswert $np_0 = 18.25$ zentriert. Treten beim Feldversuch also ungefähr 18 kantige Samen auf, so besteht sicher kein Widerspruch zur Nullhypothese, d.h. Mendels Vererbungsgesetz. Anders sieht dies aus, falls sehr wenige oder sehr viele kantige Samen beobachtet werden. Um in dieser Situation zu einer quantitativen Aussage zu kommen, definieren wir uns den 95%-Annahmebereich für die Nullhypothese H_0 unter der zweiseitigen Alternative $H_A: p \neq p_0 = 0.25$. Er enthält einen zentralen Bereich um den Erwartungswert, der mindestens 95% der Wahrscheinlichkeit enthält, und an den Rändern je höchstens 2.5% aussen vor lässt. Weil wir hier:

$P(X \leq 10) = 1.41\%$ und $P(X \leq 11) = 2.93\%$, bzw.
$P(X \geq 26) = 2.82\%$ und $P(X \geq 27) = 1.54\%$

haben, enthält der Annahmebereich die Ereignisse $\{11,12,...,25,26\}$. Sein Komplement ist der in der obigen Skizze blau eingefärbte Verwerfungsbereich, d.h. die unter der Nullhypothese extremen Ereignisse. Er besteht aus $\{0,1,...,10\} \cup \{27,28,...,73\}$. Weil der beobachtete Wert von $x = 19$ kantigen Samen im Annahmebereich liegt, wird die Nullhypothese H_0 beibehalten. Wir konnten also am Experiment keinen Widerspruch zu Mendels Vererbungsgesetz finden. Achtung, dies ist kein Beweis, dass das Verhältnis von 3:1 richtig ist. Wir können nur folgern, dass die Beobachtung mit unserer Nullhypothese verträglich ist (d.h. „die Nullhypothese ist plausibel“).

p-Wert

Die Frage, ob die Daten mit der Nullhypothese kompatibel seien, kann nach dem Test also nur grob mit „ja“ oder „nein“ beantwortet werden. Ein feineres Mass für die *Verträglichkeit zwischen Daten und Nullhypothese* ist der *p-Wert*. Er berechnet sich formell als die Wahrscheinlichkeit, gemäss dem Modell der Nullhypothese einen Wert zu erhalten, welcher im Betrag gleich weit oder weiter vom Erwartungswert np_0 weg liegt wie die tatsächlich gemachte Beobachtung x.

$$p - Wert = P[|X - np_0| \geq |x - np_0|]$$

Per Definition liegt der p-Wert im Intervall $[0,1]$, wobei die Nullhypothese zu verwerfen ist, falls der p-Wert kleiner als 0.05 ist. In vielen wissenschaftlichen Publikationen hat es sich eingebürgert, sämtliche Resultate mit p-Werten zu untermauern. In unserem Beispiel mit den Erbsen wo $x = 19$ ist, erhalten wir konkret:

$$p - Wert = P[|X - 18.25| \geq 0.75] = P[X \leq 17] + P[X \geq 19] = 0.8925$$

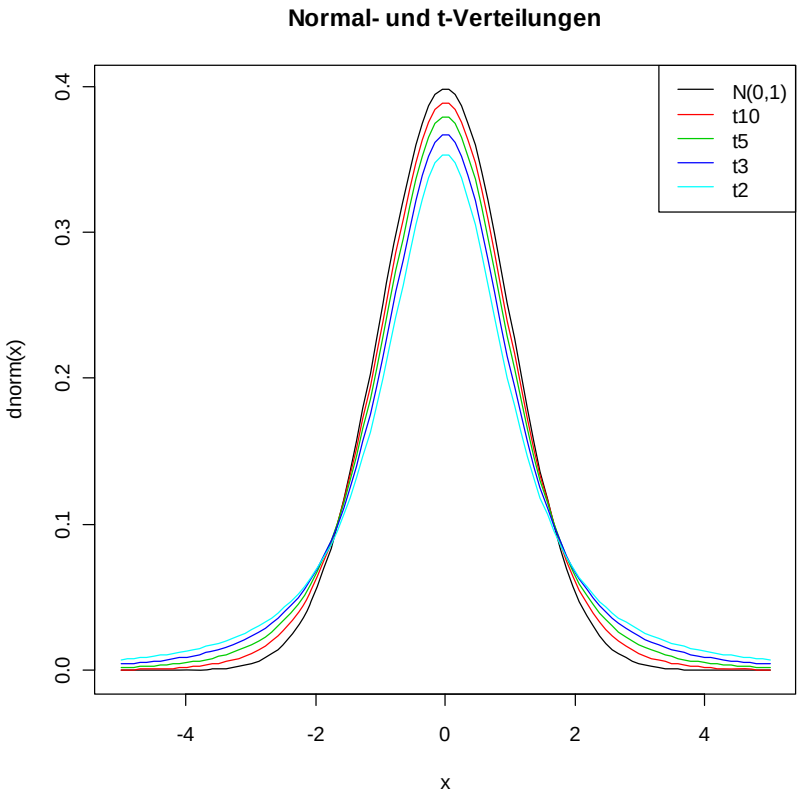
Der p-Wert ergibt sich also zu 0.8925, wir sind damit weit davon entfernt, die Nullhypothese zu verwerfen. Während die obige Formel zwar korrekt und auch hilfreich ist, so ist die Berechnung in der Praxis damit oft aufwendig. In vielen Statistik-Paketen ist denn der Binomialtest auch implementiert. Meistens genügt es, die Anzahl Versuche n, den Beobachtungswert x sowie die vermutete Erfolgswahrscheinlichkeit p_0 anzugeben, und man erhält als Output direkt den p-Wert.

3.2 t-Test für einen Erwartungswert μ_0

Der t-Test dient dazu den Mittelwert einer Stichprobe mit einem Sollwert μ_0 zu vergleichen. Eine typische Anwendung ist diese: ein schmerzstillendes Medikament, das in Spitälern angewandt wird, soll im Schnitt spätestens nach 120 Sekunden Erleichterung bringen. Die tatsächliche Zeit bis zur Wirkung schwankt jedoch von Patient zu Patient und hängt von seinen individuellen Gegebenheiten ab. Bei einer praktischen Erprobung eines neuen Medikaments an 30 Patienten zeigte ergab sich ein Mittelwert von 100.65 Sekunden, mit einer empirischen Standardabweichung von 40.41 Sekunden. Sind wir damit signifikant besser als der Gold-Standard von 120 Sekunden? Wir prüfen die Nullhypothese $H_0 : \mu = \mu_0 = 120$ gegen die Alternative $H_A : \mu \neq 120$ mit der folgenden Teststatistik:

$$T = \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\hat{\sigma}_X} \sim t_{n-1}$$

Sie misst den Unterschied zwischen Mittelwert und Sollwert auf standardisierte Art und ist intuitiv leicht nachvollziehbar. Der beobachtete Unterschied $\bar{X} - \mu_0$ ist umso bedeutender, je grösser der Stichprobenumfang n und je kleiner die Streuung der Messwerte $\hat{\sigma}_X$ ist. Unter der Nullhypothese folgt die Testgrösse T einer sogenannten Student- oder t-Verteilung mit $n-1$ Freiheitsgraden. Diese Verteilung ähnelt der Standard-Normalverteilung, ist aber etwas breiter. In der Grafik unten sind t-Verteilungen mit 2, 3, 5 und 10 Freiheitsgraden, sowie die $N(0,1)$-Verteilung aufgezeichnet. Man könnte daraus vermuten, dass mit wachsendem n eine Annäherung stattfindet. Das ist tatsächlich der Fall, für $n \to \infty$ konvergiert die t-Verteilung gegen die $N(0,1)$-Verteilung.



Der Annahmebereich für den t-Test bestimmt sich aus dem 2.5% und dem 97.5%-Quantil der t-Verteilung mit entsprechender Anzahl Freiheitsgrade. Er hängt von der Stichprobengrösse ab, erstreckt sich aber ganz grob über das Intervall $[-2,+2]$. Für die exakte Bestimmung der Grenzen sind Computerprogramme oder Tabellen nötig. In unserem Beispiel ist die t_{29}-Verteilung massgebend, und der Annahmebereich ist $[-2.045,+2.045]$. Alles was sich ausserhalb davon befindet, ist der Bereich der extremen Ereignisse, wo die Nullhypothese verworfen wird. Wir berechnen nun also den Beobachtungswert für die Teststatistik:

$$t = \sqrt{30} \cdot \frac{100.65 - 120}{40.41} = -2.62$$

Dieser Wert liegt im Verwerfungsbereich. Wir haben also einen Widerspruch zur Nullhypothese gefunden, sie wird verworfen. Somit unterschreitet unsere Stichprobe den Sollwert von $\mu_0 = 120$ signifikant. Doch wie bedeutend ist dieses Resultat? Auch hier können wir wieder eine präzisere Aussage machen, wenn wir den p-Wert als Mass für die Verträglichkeit zwischen Daten und Nullhypothese berechnen. Die allgemeine Formel lautet:

$$p - Wert = 2 \cdot P[T \leq -|t|]$$

Um Wahrscheinlichkeiten für Zufallsvariablen mit einer t-Verteilung zu berechnen, muss man ebenfalls auf Computerprogramme oder Tabellen zurückgreifen. In unserem Beispiel ergibt sich $2 \cdot P[T \leq -2.62] = 0.014$ als p-Wert.

3.3 Vertrauensintervalle

Ein Vertrauensintervall enthält alle Parameterwerte, die mit den Beobachtungen vereinbar sind. Man kann es auch als Genauigkeitsangabe für einen Schätzwert interpretieren. Es wird bestimmt, indem man alle Nullhypothesen H_0 bestimmt, welche der zugehörige Test nicht verwirft. Wir führen dies am Beispiel der beiden besprochenen Tests für Mittelwert und Anteil explizit aus.

Vertrauensintervall für den Erwartungswert μ

Wir erhalten ein 95%-Vertrauensintervall für den Erwartungswert μ, indem wir alle Nullhypothesen $H_0 : \mu = \mu_0$ bestimmen, welche der t-Test auf demselben Niveau nicht verwirft. Eine (schlechte) Möglichkeit wäre es, dies durch Ausprobieren zu tun, besser und bequemer ist die mathematische Herleitung. Wir beginnen mit dem 95%-Annahmebereich für $H_0 : \mu = \mu_0$. Dieser ist:

$$qt_{0.025;n-1} \leq T \leq qt_{0.975;n-1}$$

Wir ersetzen die Teststatistik T mit der entsprechenden Berechnungsformel:

$$qt_{0.025;n-1} \leq \frac{\bar{X} - \mu_0}{\hat{\sigma}_X / \sqrt{n}} \leq qt_{0.975;n-1}$$

In anderen Worten bedeutet dies, dass die Nullhypothese genau dann beibehalten wird, falls die obigen Ungleichungen für einen realisierten Mittelwert $\overline{x}$ erfüllt sind. Wir betrachten nun $\overline{x}$ als gegeben/fest und formen die Ungleichungen um:

$$qt_{0.025;n-1} \leq \frac{\overline{x} - \mu_0}{\hat{\sigma}_X / \sqrt{n}} \leq qt_{0.975;n-1}$$

$$qt_{0.025;n-1} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}} \leq \overline{x} - \mu_0 \leq qt_{0.975;n-1} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}$$

$$\overline{x} + qt_{0.025;n-1} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}} \leq \mu_0 \leq \overline{x} + qt_{0.975;n-1} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}$$

Somit haben wir eine Ungleichung für μ_0 gefunden, die wir benützen können, um die nicht verworfenen Nullhypothesen zu bestimmen. Weil die t-Verteilung symmetrisch ist, gilt zudem:

$$-qt_{0.025;n-1} = qt_{0.975;n-1}$$

Somit lautet die Formel für das Intervall der nicht verworfenen Nullhypothesen, und damit für das 95%-Vertrauensintervall für den Erwartungswert μ:

$$\overline{x} \pm qt_{0.975;n-1} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}.$$

Mit dieser Formel ist auch klar, dass das Vertrauensintervall als Genauigkeitsangabe der Schätzung interpretiert werden kann. Ebenso ist sofort ersichtlich, dass das Vertrauensintervall für grössere Stichproben kürzer ist, bzw. dass es bei grosser Streuung der Einzelwerte länger wird. Am Beispiel des Schmerzmittels ergibt sich folgendes Resultat:

$$100.65 \pm 2.045 \cdot \frac{40.41}{\sqrt{30}} = [93.27, 108.03]$$

Es ist der Bereich von plausiblen Werten für μ, d.h. der erwarteten Dauer, bis das Schmerzmittel eine Wirkung zeigt. Der Wert von 120 Sekunden gehört nicht zu diesem Intervall und wird somit signifikant unterschritten. Das ist keine Neuigkeit, dies hatten wir bereits beim t-Test erkannt.

Vertrauensintervall für einen Anteil p

Auch hier enthält das 95%-Vertrauensintervall alle Nullhypothesen $H_0: p = p_0$, welche der Binomialtest nicht verwirft. Da es sich um eine diskrete Verteilung handelt, ist die Angabe einer exakten, expliziten Formel umständlich. Man behilft sich daher mit der Näherungsformel, welche aus dem zentralen Grenzwertsatz abgeleitet wird. Sie ist genügend genau, falls $n\hat{p}(1-\hat{p}) \geq 10$:

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Im Beispiel mit Mendel's Erbsenexperiment ergibt sich folgendes Zahlenesultat: $\hat{p} = 19/73 = 0.26$, und weil $73 \cdot 0.26 \cdot 0.74 = 14.05$ ist die Näherungsformel gültig, woraus sich ein 95%-Vertrauenintervall von $[0.16, 0.36]$ ergibt. Der Wert von 0.25 ist darin enthalten und damit ein plausibler Wert. Wie wir bereits erkannt hatten, wird diese Nullhypothese denn auch nicht verworfen. Ein weiteres, wichtiges Anwendungsfeld für dieses Vertrauensintervall sind Meinungsumfragen, z.B. vor einer Abstimmung.

3.4 Vergleich von 2 Stichproben

Die bisher besprochenen Tests gegen einen Sollwert sind in manchen Situationen nützlich. In der Praxis tritt aber oft auch der Fall auf, dass zwei Stichproben bzw. ihre Kennwerte (*Mittelwert*, bzw. *Anteil Erfolge*) gegeneinander verglichen werden müssen. Somit weisen beide Kenngrössen eine Unsicherheit auf, und dies muss beim Testen entsprechend berücksichtigt werden.

Beispiel: Weisse Weihnachten

Ein Beispiel hierzu ist die Frage, ob es früher häufiger weisse Weihnachten gab. Die Archive von Meteoschweiz besagen, dass von 1961-1990 an 39 von total 90 Weihnachtstagen, d.h. 24./25./26. Dezember, mindestens 1cm Schnee lag. In der späteren Periode von 1991-2010 war dies an 16 von 60 Tagen der Fall. Nun ist der Anteil $\hat{p}_{1961–1990} = 39/90 = 0.433$ höher als $\hat{p}_{1991–2010} = 16/60 = 0.267$. Doch ist dieser Unterschied signifikant?

Beispiel: Hipparions

Hipparions sind eine Pferdeart, welche vor knapp 1 Mio. Jahren ausgestorben sind. Untersucht wurden 77 Backenzähne von zwei verschiedenen Unterarten, welche in Malawi (Afrika) gefunden wurden. Davon liessen sich 39 Stück dem *Hipparion Africanum* zuordnen, welches vor etwa 4 Mio. Jahren lebte. Die anderen 38 Stück kamen vom *Hipparion Libycum*, das vor rund 2.5 Mio. Jahren existierte. Vor rund 2.8 Mio. Jahren kühlte sich das Klima weltweit ab, und wechselte in Ostafrika von warm-feucht zu kühl-trocken. Man vermutet, dass die Hipparion deshalb von Laub- zu Grasfressern wurden. Durch diese Evolution veränderten sich auch die Zähne, was anhand der mesiodistalen Länge (in mm) der Fundstücke untersucht werden soll. Es ergaben sich Messwerte von $\hat{\mu}_A = 25.9$ mit einer Streuung von $\hat{\sigma}_A = 2.2$, respektive $\hat{\mu}_L = 28.4$ mit $\hat{\sigma}_L = 4.3$. Die beiden Mittelwerte sind offensichtlich nicht identisch, doch unterscheiden sie sich signifikant?

Ad-Hoc-Vergleich

Ein einfaches und schnell durchzuführendes Ad-Hoc-Verfahren für alle obigen Situationen besteht darin, jeweils für beide Stichproben das Vertrauensintervall zu erzeugen. Falls diese sich *nicht überlappen*, so wird die Nullhypothese $p_1 = p_2$ bzw. $\mu_1 = \mu_2$ verworfen, d.h. der Unterschied zwischen den beiden Stichproben ist statistisch signifikant. Der Nachteil bei diesem Vorgehen ist jedoch, dass man auf diese Weise erstens keinen p-Wert erhält und das Verfahren auch nur ungefähr

ist, in dem Sinne dass ein genauerer Test die Nullhypothesen möglicherweise verwirft. Wir rechnen nun die Resultate für unsere Beispiele nach.

Für die weissen Weihnachten ergeben sich Vertrauensintervalle von $[0.33, 0.54]$ für die Periode von 1961-1990, und von $[0.15, 0.38]$ für den zweiten Abschnitt. Die beiden Intervalle überlappen sich, somit kann die Nullhypothese $p_{1961–1990} = p_{1991–2010}$ nicht verworfen werden, d.h. es ist möglich (aber nicht sicher bzw. bewiesen), dass der Anteil von weissen Weihnachtstagen identisch ist. Bei den Hipparion ergeben sich Vertrauensintervalle von $[25.19, 26.61]$ und $[26.99, 29.81]$. Diese überlappen sich nicht, damit wird die Nullhypothese verworfen und wir haben statistisch nachgewiesen, dass sich die mesiodistale Länge der Zähne der beiden Arten unterscheidet.

2-Stichproben-Tests

Noch genauere Aussagen und die Angabe von p-Werten sind möglich, wenn man entsprechende 2-Stichproben-Tests benützt. Die Nullhypothese $H_0 : p_1 = p_2$ kann mit dem sogenannten Proportionalitäts-Test oder 2-Stichproben-Binomial-Test geprüft werden, falls $n\hat{p}_1(1-\hat{p}_1) \geq 5$ und $n\hat{p}_2(1-\hat{p}_2) \geq 5$ erfüllt sind. Man benützt dann die Teststatistik:

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \chi_1^2$$

Man bildet also eine standardisierte Differenz der beiden geschätzten Anteile $\hat{p}_1$ und $\hat{p}_2$. Sie folgt einer Chi-Quadrat-Verteilung mit einem Freiheitsgrad. Für die Bestimmung von Annahme- und Verwerfungsbereich sowie die Angabe von p-Werten sind Computerprogramme nötig. In unserem Beispiel ergibt sich für die Teststatistik ein Wert von 3.62, was einem p-Wert von 0.057 entspricht, die Nullhypothese wird also (knapp) beibehalten.

Für die Hipparion, bzw. generell für den Vergleich von zwei Mittelwerten ist der 2-Stichproben-t-Test einzusetzen. Wir evaluieren damit die Nullhypothese $H_0 : \mu_1 = \mu_2$ unter der Annahme, dass die beiden Varianzen identisch sind, d.h. $\sigma = \sigma_1 = \sigma_2$. Die Teststatistik lautet:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}, \text{ mit } S_p = \sqrt{\frac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}}$$

Sie ist auf dem Unterschied zwischen den beiden Mittelwerten aufgebaut, berücksichtigt werden daneben die Grösse der beiden Stichproben sowie die Streuung der Einzelwerte, wofür ein gewichtetes Mittel der beiden auf den Stichproben geschätzten Varianzen gebildet wird. Unter der Nullhypothese hat T eine t-Verteilung mit $n_1 + n_2 - 2$ Freiheitsgraden. Als Zahlenresultat ergibt sich bei den Hipparion ein Wert der Teststatistik von 3.2, was einen p-Wert von 0.002 ergibt. Die Nullhypothese wird also (deutlich) verworfen.

4 Regression

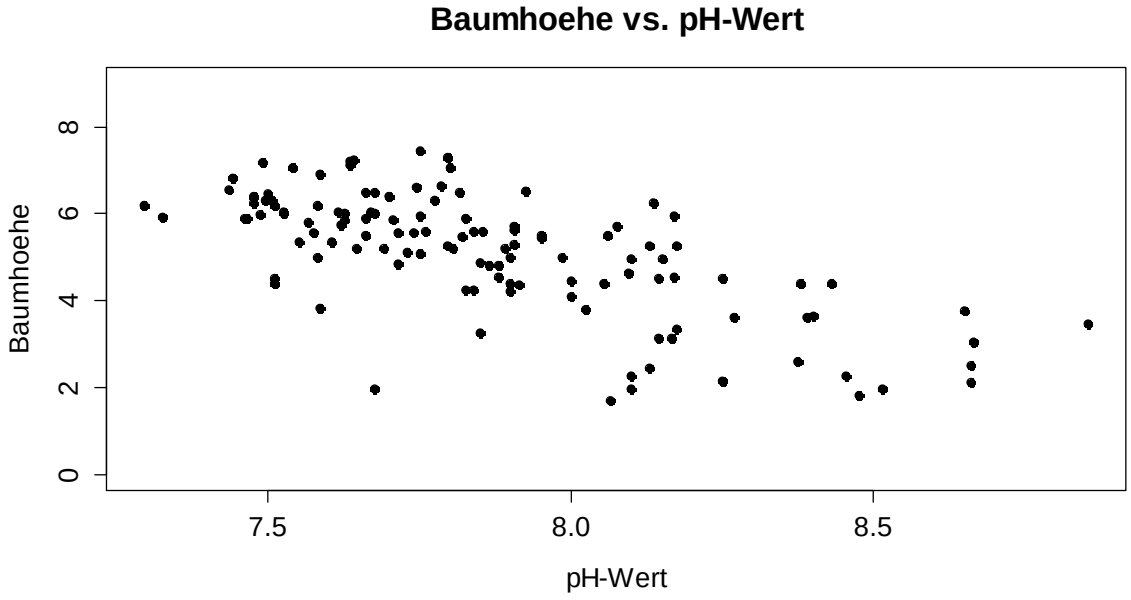
4.1 Einführung, Beispiel, Fragestellung

In Wissenschaft und Alltag fragen wir uns immer wieder, wie eine Grösse von Interesse von anderen Einflussfaktoren abhängt. Diese grundlegende Frage wird durch die Regression quantitativ untersucht. So erstaunt es auch nicht, dass es sich um eine der am meisten angewandten Techniken der Statistik handelt.

Beispiel: In Indien behindern basische Böden Pflanzen beim Wachstum. Es werden daher Baumarten gesucht, die eine hohe Toleranz gegen solche Umweltbedingungen zeigen. In einem Versuch wurden auf einem Feld mit grossen lokalen Schwankungen des pH-Werts 120 Bäume einer bestimmten Art gepflanzt. Nach 3 Jahren wurde deren Höhe y_i gemessen, ebenso war der pH-Wert x_i des Bodens an derselben Stelle bekannt. Weiter waren Informationen zur Sodium Absorption Rate vorhanden, welche einen anderen Aspekt der Basizität erfasst.

Baum	Höhe	pH	SAR
1	5.91	7.325	0.0969
2	5.20	7.690	0.4393
3	4.40	7.900	1.0000
4	4.50	8.145	1.3160
5	6.05	7.615	0.0607
6	6.00	7.525	0.2041

Wir untersuchen zuerst den Zusammenhang zwischen Höhe und pH-Wert. Die Variable SAR bringen wir erst später ins Spiel. Wir betrachten den Scatterplot:



Mit zunehmendem pH-Wert nimmt die Baumhöhe tendenziell ab. Der Zusammenhang scheint linear. Es bietet sich also an, eine Gerade zur Beschreibung zu verwenden. Wir interpretieren die x-Variable *pH* als *Prädiktor* und die y-Variable *Höhe* als *Zielvariable*. Der Zusammenhang wird durch eine lineare Funktion beschrieben, von welcher es zufällige Abweichungen gibt. Das Hauptziel der Untersuchung wird darin bestehen, den Zusammenhang mit einer Formel zu beschreiben. Weiter soll auch entschieden werden, ob der pH-Wert einen signifikanten Einfluss auf die Baumhöhe hat. Ebenfalls von Interesse ist die erwartete Baumhöhe für einen gegebenen pH-Wert, inklusive eines Vertrauensintervalls.

4.2 Modell

Der Zusammenhang zwischen der Zielvariablen y und dem Prädiktor x und wird folgendermassen beschrieben:

$y_i = \beta_0 + \beta_1 x_i + E_i$, für alle $i = 1, ..., n$.

Dabei haben die einzelnen Grössen die folgende Bedeutung:

y_i ist die *Zielvariable* der i-ten Beobachtung. Im vorliegenden Beispiel handelt es sich um die Höhe des i-ten Baums. Dies ist eine Zufallsgrösse.

x_i ist die *Prädiktorwert* der i-ten Beobachtung. Im vorliegenden Beispiel handelt es sich um den pH-Wert des i-ten Baums. Die erklärende Variable wird als feste, nicht zufällige Grösse betrachtet.

β_0, β_1 sind unbekannte Parameter, die sogenannten *Regressionskoeffizienten.* Diese sollen mit Hilfe der vorliegenden Beobachtungen geschätzt werden. Dabei ist β_0 der so genannte *Achsenabschnitt* (engl. *Intercept*) und β_1 die *Steigung* (engl. *Slope*). Letztere gibt an, um wie viel sich der Wert der Zielvariablen erhöht, wenn der x-Wert um eine Einheit zunimmt.

E_i ist der *Restterm* oder *Fehler*. Es handelt sich um eine Zufallsgrösse, d.h. die Abweichung zwischen dem beobachteten Wert y_i und dem angepassten Wert auf der Gerade wird als zufällig interpretiert.

Modellvoraussetzungen: Wir setzen voraus, dass der Erwartungswert $E[E_i] = 0$ ist, d.h. der Zusammenhang zwischen Zielgrösse und Prädiktor wird durch eine Gerade beschrieben, von welcher es keine systematische Abweichung gibt. Weiter muss die Varianz $Var(E_i) = \sigma_E^2$ konstant sein, und die Fehler dürfen keine Korrelation aufweisen, d.h. $Cov(E_i, E_j) = 0$ für $i \neq j$.

Wir sprechen hier von einfacher linearer Regression, weil nur eine einzige erklärende Variable im Modell enthalten ist. Achtung: das Modell heisst nicht linear, weil eine Gerade angepasst wird, sondern weil die Modellgleichung linear in den Parametern β_0, β_1 ist. So ist zum Beispiel auch $y_i = \beta_0 + \beta_1 x_i^2 + E_i$ ein einfaches lineares Modell, während es sich bei $y_i = \beta_0 + \beta_1 x_i^{\beta_2} + E_i$ um ein nichtlineares Regressionsproblem handelt.

4.3 KQ-Verfahren zur Anpassung der Gerade

Hier stellen wir uns die Frage, welche Gerade am besten zu den n Punktepaaren (x_i, y_i) passt. Für jeden Punkt betrachten wir die vertikale Abweichung vom zugehörigen Punkt $\hat{y}_i$ auf der Geraden, das sogenannte *Residuum:*

$$r_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$$

Die Gerade soll nun so gelegt werden, dass die Summe der quadrierten Residuen so klein wie möglich ist. Man spricht auch von der *Methode der kleinsten Quadrate* (engl. *Least Squares*), bzw. abgekürzt vom *KQ-Verfahren* (engl. *OLS*). Formuliert man dieses Paradigma mathematisch aus, so lautet es:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = \min!$$

Das Minimierungsproblem kann entweder durch Nullsetzen der partiellen Ableitungen, oder auch durch geometrische Projektionsüberlegungen gelöst werden. Beachten sie die Analogie zur Ausgleichsrechnung in der Linearen Algebra. Ohne die Details aufzuzeigen, sei hier bloss erwähnt, dass wir hier ebenfalls die Normalengleichungen erhalten:

$$(X^T X)\beta = X^T y\,.$$

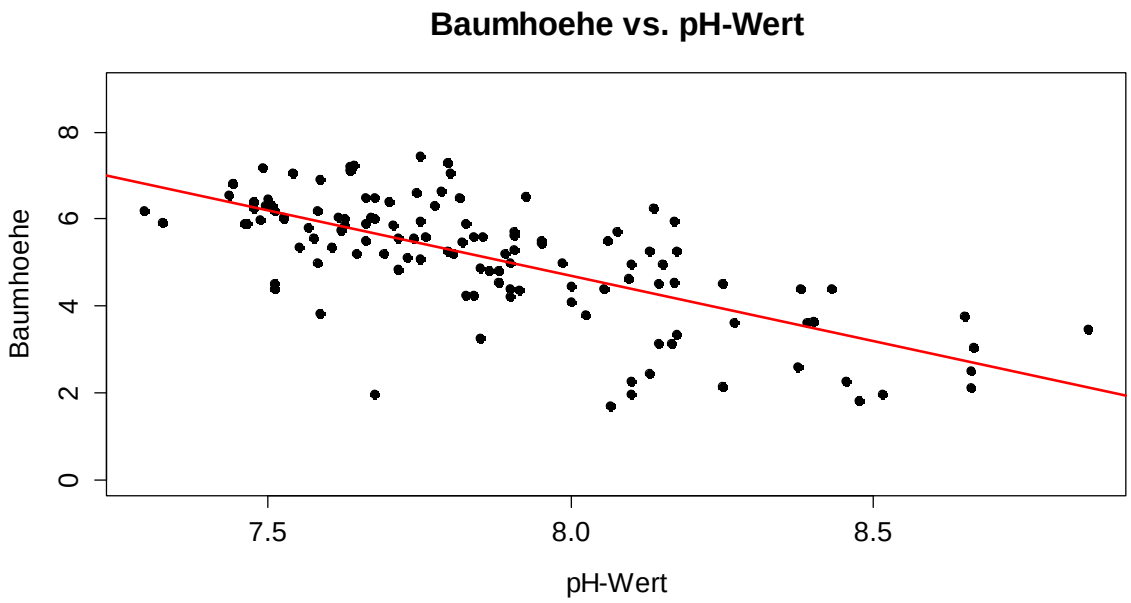
Die Normalengleichungen können explizit gelöst werden. Aus den Lösungen erhalten wir die Schätzungen für die beiden Regressionsparameter, welche von folgender Art sind:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \text{ und } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\,.$$

Durch die geschätzten Regressionskoeffizienten ist die Regressionsgerade bestimmt. Sie ist gleich:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \text{ für alle } i = 1, ..., n\,.$$

und wir können sie im Scatterplot einzeichnen, siehe nächste Seite. Hierbei ist $\hat{y}_i$ der vom Modell geschätzte Wert der Zielgrösse. Man spricht auch vom *gefitteten Wert* (engl. *fitted* bzw. *predicted value*). Beachten sie, dass die Residuen r_i für alle Beobachtungen der Differenz zwischen gefittetem und beobachtetem Wert entsprechen. Natürlich ist es ein Stück weit willkürlich, gerade die Quadratsumme der Residuen zu minimieren. So könnte z.B. auch die Summe der absoluten Abweichungen, d.h. die L_1-Norm minimiert werden. Diese Norm, sowie einige weitere Minimierungskriterien, finden in der Praxis durchaus manchmal Anwendung. Sie zeichnen sich typischerweise durch grössere Robustheit gegenüber Ausreissern aus. Andererseits zeichnet sich das KQ-Verfahren durch seine Einfachheit aus, da die Lösung explizit als Funktion der Datenpaare (x_i, y_i) angegeben werden kann. Ebenso lassen sich Optimalitätsaussagen beweisen, insbesondere bei Annahme von normalverteilten Fehler E_i.



4.4 Eigenschaften der Schätzungen

Wie erwähnt gibt es gute Gründe, das KQ-Verfahren einzusetzen. Diese wollen wir hier etwas beleuchten. Das *Gauss-Markov-Theorem* besagt, dass unter den Modellvoraussetzungen von oben die Schätzungen $\hat{\beta}_0, \hat{\beta}_1$ erwartungstreu sind (d.h. $E[\hat{\beta}_0] = \beta_0$ und $E[\hat{\beta}_1] = \beta_1$). Weiter haben sie unter allen erwartungstreuen, linearen Schätzern minimale Varianz, es sind also die genauesten Schätzungen, die man erhalten kann. Es gilt:

$$Var(\hat{\beta}_0) = \sigma_E^2 \cdot \left(\frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right), \text{ und } Var(\hat{\beta}_1) = \frac{\sigma_E^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Dieses Resultat zeigt uns auch, wie wir genauere Schätzungen, d.h. eine „exakter bestimmte Regressionsgerade“ erhalten können.

- Wir können die Anzahl Beobachtungen n erhöhen.
- Wir können auf eine gute Streuung der x-Werte achten
- Wir können durch geeignete erklärende Variablen σ_ε^2 klein halten
- Für $\hat{\beta}_0$ hilft es, wenn der Mittelwert $\overline{x}$ nahe bei null liegt.

4.5 Schätzung von σ_E^2

Neben den Regressionskoeffizienten ist auch noch die *Varianz der zufälligen Fehler* zu schätzen, die wir für alle möglichen Tests und Vertrauensintervalle benötigen. Sie basiert auf der *Residuenquadratsumme* (RSS, engl. *Residual Sum of Squares*) und lautet:

$$\hat{\sigma}_E^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} r_i^2$$

4.6 Tests und Vertrauensintervalle

Bis zu diesem Punkt haben wir nur Annahmen über Erwartungswert, Varianz und Korrelation der zufälligen Fehler gemacht, jedoch keine bestimmte Verteilung vorausgesetzt. Das bedeutet konkret, dass die obigen Resultate unabhängig von der Verteilung gelten. Für die folgenden Abschnitte über Tests, Vertrauens- und Prognoseintervalle müssen wir nun aber voraussetzen, dass:

$E_i \sim N(0, \sigma_E^2)$, zudem müssen die Fehler stochastisch unabhängig sein.

Diese Annahme muss mit einem Normal Plot überprüft werden, falls man sich auf die im Folgenden vorgestellten Tests und Intervalle verlassen will. Ist die Voraussetzung verletzt, so kann es zu groben Fehlschlüssen kommen. Wir widmen uns als erstes der Frage der Genauigkeit der Steigung, bzw. möchten ein Intervall mit plausiblen Werten für β_1 angeben. Dies kann man mit einem 95%-Vertrauensintervall tun. Es ist wie folgt:

$$\hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \hat{\sigma}_{\hat{\beta}_1} = \hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \sqrt{\hat{\sigma}_E^2 \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

In unserem Beispiel ergibt sich die Punktschätzung $\hat{\beta}_1 = -3.003$, mit einem 95%-VI von $(-3.566, -2.440)$. Um zu entscheiden, ob die erklärende Variable x einen signifikanten Einfluss auf die Zielgrösse y hat, testet man die Nullhypothese $H_0 : \beta_1 = 0$ gegen die Alternative $H_A : \beta_1 \neq 0$. Als Testgrösse verwenden wir

$$T = \frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{\sqrt{Var(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}_E^2 \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2}} \sim t_{n-2} .$$

Die Testgrösse hat eine t-Verteilung mit $n-2$ Freiheitsgraden. Diese wird zur Bestimmung von Verwerfungs- und Annahmebereich, bzw. für die Bestimmung des p-Werts verwendet. Verwirft man die Nullhypothese, so wird der Zusammenhang zwischen erklärender Variable und Zielgrösse als statistisch gesichert betrachtet, d.h. die Steigung ist signifikant von null verschieden.

Will man prüfen, ob sich der Achsenabschnitt $\hat{\beta}_0$ signifikant von null unterscheidet, so geht man auf analoge Weise vor. Dieser Test ist aber weniger wichtig, da der Achsenabschnitt eigentlich immer im Modell verbleiben soll, auch wenn der y-Wert an der Stelle $x = 0$ eigentlich null sein sollte. Dies gilt insbesondere, wenn die Daten weit rechts und links von $x = 0$ liegen, und auch wenn der Achsenabschnitt als nicht signifikant ausgegeben wird.

4.7 Output von Statistikpaketen

Wird die Regressionsanalyse mit einem Statistikpaket ausgeführt, so werden im Output nicht nur die Punktschätzungen für β_0, β_1 angegeben (Spalte „*Estimate*“), sondern in der Regel auch deren Standardabweichungen (Spalte „*Std. Error*“), der Wert der Testgrösse T (Spalte „*t value*“), sowie der p-Wert zu den jeweiligen Nullhypothesen (Spalte „*Pr(>|t|)*“).

```
Coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -716015.56  279804.37  -2.559   0.0121 *
ATM             103.10      12.43   8.293 9.45e-13 ***
---
Residual standard error: 196800 on 91 degrees of freedom
Multiple R-squared: 0.4304,   Adjusted R-squared: 0.4242
F-statistic: 68.77 on 1 and 91 DF,  p-value: 9.452e-13
```

Weiter ist auch die Punktschätzung für σ_E angegeben („*Residual standard error*“), mit der zugehörigen Anzahl Freiheitsgrade $n-2$ („*degrees of freedom*“), aus welcher sich direkt die Anzahl Beobachtungen ablesen lässt, mit welcher die Regression gerechnet wurde. Zusätzlich wird auch noch die Grösse *Multiple R-squared* angegeben. Sie berechnet sich als

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0,1]$$

und gibt den Anteil der durch die Regressionsgerade erklärten Streuung an der gesamten Streuung der Daten in y-Richtung an.

4.8 Prognose

Die geschätzten Parameter, bzw. die angepasste Regressionsgerade können nun benützt werden, um für ein beliebiges x^* den Wert der Zielvariablen vorherzusagen. Wir verwenden dazu einfach:

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

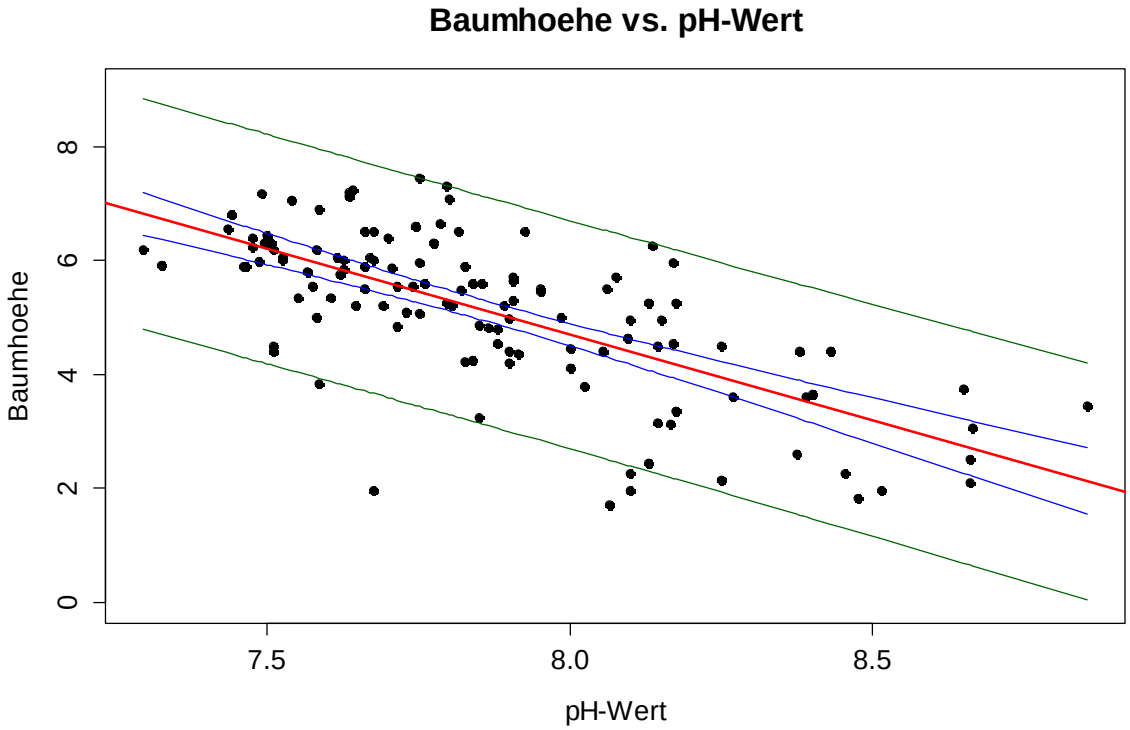
Hierbei ist zu berücksichtigen, dass normalerweise nur Vorhersagen innerhalb des Bereichs der für die Anpassung der Gerade verwendeten x-Werte verlässlich sind. Man spricht in diesem Fall von *Interpolation*. Hingegen sind *Extrapolationen*, über den Bereich der verwendeten x-Werte hinaus, generell mit Vorsicht zu geniessen.

Beispiel: Für eine pH-Wert von 8.0 erwarten wir eine Baumhöhe von $28.7227 + (-3.0034 \cdot 8.0) = 4.4955$ Metern. Hingegen ist es nicht sinnvoll, mit Hilfe der Regressionsgerade die Baumhöhe für einen pH-Wert von 5.0 anzugeben. Und erst recht ist es hier offensichtlich, dass eine Vorhersage für einen x-Wert von 0 komplett unsinnig ist.

Wir können nun ein 95%-Vertrauensintervall für den Fitted Value $\hat{y}^*$ angeben.

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Wir können dieses Intervall für beliebige x^* angeben, und in der folgenden Grafik in blauer Farbe als *Vertrauensbereich* für die angepasste Gerade einzeichnen. Dieser ist, wie aufgrund der Gleichung leicht sichtbar, in der Mitte schmaler als an den Rändern.



Der Vertrauensbereich für die Regressionsgerade zeigt auf, welche Geraden aufgrund der Daten als plausibel zu erachten sind. Wie wir schon im Intervall für den Parameter β_1 erkannt haben, könnte die Steigung der Geraden auch etwas grösser oder kleiner sein. Hingegen ist mit diesem Intervall für die Fitted Values noch nicht klar, wo eine neue Beobachtung zu liegen kommen wird. Die einzelnen Beobachtungen streuen ja noch zusätzlich um den erwarteten Wert herum. Das 95%-Prognoseintervall für y^* ist gegeben durch:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \,.$$

Wiederum können wir das Prognoseintervall für alle x^* bestimmen und als *Prognosebereich* mit dunkelgrüner Farbe im Scatterplot einzeichnen. Dieses Intervall ist natürlich stets breiter als das blaue Vertrauensintervall für die Gerade.

4.9 Prüfen der Voraussetzungen: Residuenanalyse

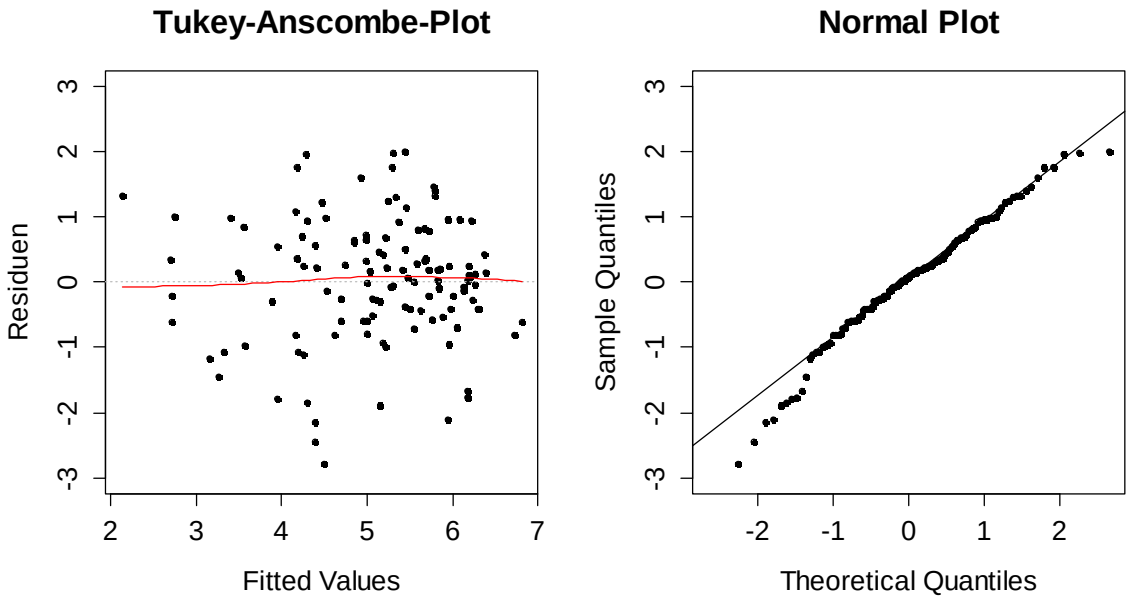
Nach der Anpassung einer Regressionsgeraden soll überprüft werden, ob die Modellvoraussetzungen und die Normalverteilungsannahme erfüllt, und die Resultate somit gültig sind. Zu prüfen ist:

- Der Zusammenhang zwischen x und y zeigt keine systematische Abweichung von einer Geraden, d.h. der Erwartungswert der Fehler E_i ist auf dem ganzen Wertebereich null.

- Die Fehler E_i haben über den ganzen Wertebereich eine konstante Streuung σ_E^2 und sind zudem (zeitlich) unkorreliert.

- Falls Tests oder Vertrauens-/Prognosebereiche berechnet werden, so müssen die Fehler E_i zusätzlich auch normalverteilt sein.

Die Überprüfung findet mit grafischen Methoden statt. Dabei werden typischerweise die Residuen r_i gegen verschiedene andere Grössen dargestellt. Hier werden die wichtigsten Plots vorgestellt.



Tukey-Anscombe-Plot

Mit dieser Darstellung können in erster Linie systematische Fehler entdeckt werden, und auch nichtkonstante Varianz ist manchmal auffällig. Man trägt auf der x-Achse die Fitted Values, und auf der y-Achse die Residuen auf. Im Idealfall befinden sich alle Residuen in einem horizontalen Band mit konstanter Breite, und streuen zufällig, die rote Linie des Glätters liegt auf der x-Achse. Dies ist im vorliegenden Beispiel mit hinreichender Genauigkeit erfüllt. Falls der Glätter systematisch von der Nulllinie abweicht, so ist das angepasste Regressionsmodell falsch. Die damit erzeugten Fits, Vorhersagen, VIs und Tests sind nicht korrekt. Abhilfe können Variablentransformationen oder zusätzliche, ins Modell aufgenommene Variablen schaffen, was hier aber nicht weiter besprochen wird.

Normalplot

Die Annahme der Normalverteilung wird mit dem sogenannten *Normal Plot* überprüft. Dabei werden die nach Grösse geordneten Residuen gegen die entsprechenden Quantile der Normalverteilung geplottet. Wenn die Fehler E_i

normalverteilt sind, so gilt dies auch für die Residuen. Die Punkte im Normal Plot sollten also entlang einer Gerade liegen. Im vorliegenden Beispiel erkennt man zwar an den Rändern einige Abweichungen, doch liegen diesen noch im Rahmen des Tolerierbaren, so dass wir unseren Resultaten vertrauen können.

Würde der Normalplot eine rechtsschiefe Verteilung zeigen, so kann eine Logarithmustransformation der Zielgrösse und möglicherweise auch des Prädiktors Abhilfe schaffen. Zeigt der Normalplot einzelne Beobachtungen mit betragsmässig grossen Residuen, so soll abgeklärt werden, ob es sich hierbei um grobe (Abschreib-)fehler etc. handelt. Falls ja, so können diese Beobachtungen entfernt oder korrigiert, und die Analyse wiederholt werden. Für andere, systematische Abweichungen von der Normalverteilungsannahme sind Methoden nötig, die über den Rahmen einer Einführungsvorlesung hinausgehen. Konsultieren sie falls nötig einen Spezialisten!

4.10 Ausblick: Multiple Regression

Das Modell der einfachen linearen Regression lässt sich auf den Fall verallgemeinern, wo die Zielvariable y nicht nur von einer, sondern von mehreren erklärenden Variablen $x_{i1}, x_{i2}, ..., x_{ip}$ beeinflusst wird. Die Regressionsgleichung lautet dann:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + E_i \text{ für alle } i = 1, ..., n.$$

Über die zufälligen Fehler werden hier wieder genau dieselben Annahmen wie in der einfachen Regression getroffen. Absolut zentral ist die Erkenntnis, dass das Modell der multiplen Regression mehr Information liefert als das Ausführen einer Sequenz von einfachen Regressionen gegen alle Prädiktoren einzeln. Die multiple Regression kann nämlich das Zusammenspielen der verschiedenen Einflussfaktoren berücksichtigen.

Beispiel: Bei den Baumhöhen war nicht nur der pH-Wert des Bodens verfügbar, sondern auch die *Sodium Absorption Ratio*, welche einen etwas anderen Aspekt der Basizität erfasst. Erklärt man die Baumhöhe mit beiden Variablen gleichzeitig, so ist eine genauere Beschreibung der Baumhöhe möglich.

Die geschätzten Parameter $\hat{\beta}_1, ..., \hat{\beta}_p$ sowie $\hat{\sigma}_E^2$ erhält man wieder durch Anwendung der Methode der Kleinsten Quadrate, welche auch hier zu den Normalengleichungen führen. Bei der Verwendung von Software-Paketen zur Regressionsrechnung ist die Erweiterung zu multipler Regression meist „straightforward“. Im Output werden weiterhin Schätzwerte, Standardfehler, Testgrössen und p-Werte für die Nullhypothesen $H_0 : \beta_j = 0$ für alle $j = 0, ..., p$ geliefert. In unserem Beispiel ist das Resultat wie folgt:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.9466     2.7445   9.818  < 2e-16 ***
ph           -2.7558     0.3603  -7.649  5.6e-12 ***
sar          -0.2519     0.2255  -1.117    0.266
```

```
---
Residual standard error: 1.007 on 120 degrees of freedom
Multiple R-squared:  0.485, Adjusted R-squared:  0.4764
F-statistic: 56.51 on 2 and 120 DF,  p-value: < 2.2e-16
```

Es zeigt sich aber, dass der Einfluss der Variable `sar` auf die Zielgrösse nicht signifikant ist. D.h., es ist (im Gegensatz zum pH-Wert) nicht gesichert, dass die Sodium Absorption Rate einen Einfluss auf die Baumhöhe hat. Ebenso ist auch das Bestimmtheitsmass durch das Hinzufügen der Variable nur minimal angestiegen. Somit ist die neu hinzugefügte Grösse sicher kein wesentlicher Prädiktor für das Wachstum der Bäume.

4.11 Ausblick: Verallgemeinerte lineare Modelle

Für die bisher besprochenen Regressionsmethoden gingen wir von der Annahme aus, dass die Zielvariable y eine stetige Zufallsgrösse ist. Dies ist nicht bei allen Regressionsproblemen der Fall. Es kann z.B. auch interessant sein, das Eintreten eines Ereignisses, d.h. eine 0/1-Variable in Abhängigkeit von erklärenden Grössen zu untersuchen.

Hierzu passt weder das Modell der einfachen, noch der multiplen linearen Regression, sondern es ist eine logistische Regression gefragt, die ihrerseits den verallgemeinerten linearen Modellen zugeordnet ist. Falls sie vor einem solchen Problem stehen, bilden sie sich weiter, oder konsultieren sie einen Spezialisten.