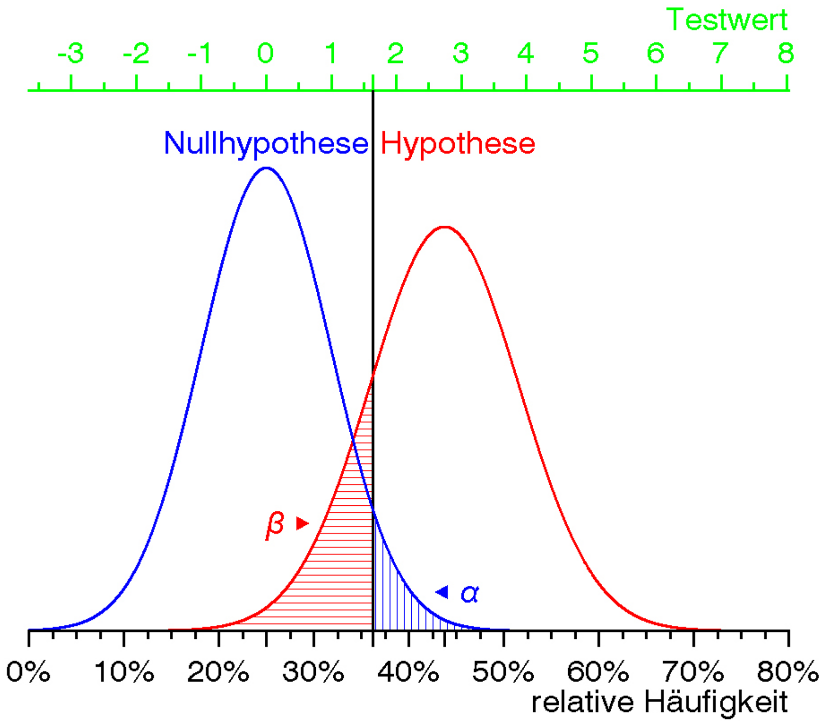
Skript zur

Statistik in den Naturwissenschaften

Gerhard Osius



Mathematik Arbeitspapiere Nr. 59
Teil B: Hochschuldidaktisches Material

Fachbereich Mathematik/Informatik
Universität Bremen

April 2009

Anderson's law:[1]
There is no problem, no matter how complex, which, upon careful analysis, does not become more complex.

Vorwort

Dieses Skript ist im Laufe mehrerer Jahre entstanden und liegt jetzt in einer relativ vollständigen Form vor. Es ist primär als Begleit- und Nachschlagewerk zur Lehrveranstaltung „Statistik in Naturwissenschaft und Informatik“ konzipiert. Das Ziel dieser Veranstaltung ist es, eine Einführung in statistische Methoden und deren Grundlagen zu geben, die dann weiterführende eigene Literatur-Studien ermöglicht. Deshalb sollen hier nicht möglichst viele, sondern exemplarisch einige grundlegende Verfahren ausführlich erläutert und angewandt werden. Auf strenge Herleitungen der Verfahren wird bewußt verzichtet, aber auf präzise Formulierungen wird Wert gelegt und für (mathematisch) Interessierte werden auch Begründungen erwähnt.

Die effiziente Anwendung statistischer Verfahren auf konkrete Daten erfordert heutzutage einen Computer mit entsprechender Software. Ein wissenschaftlich-technischer Taschenrechner (in Kombination mit statistischen Tabellen) bietet insbesondere bei umfangreicheren Daten zu wenig Komfort und (wenn überhaupt) auch nur eingeschränkte Möglichkeiten für graphische Darstellungen. Für die hier behandelten Verfahren ist allerdings keine spezielle Statistik-Software erforderlich, sondern lediglich ein Tabellen-Kalkulationsprogramm. Da sich hierunter das Programm *Microsoft Excel* als Standard etabliert hat, werden die Tipps für die statistische Auswertung hier nur für dieses Programm genauer ausgeführt (Grundkenntnisse hierzu werden im Rahmen der Übungen zur Veranstaltung vermittelt). Das Program *OpenOffice Calc* ist für die hiesigen Erfordernisse ebenfalls ausreichend und ist weitgehend ähnlich bedienbar.

Das Skript gliedert sich in drei große Teile sowie einen Anhang mit statistischen Tabellen, einem umfangreicheren Datensatz (StatLab-Auswahl 1985) und einer Zusammenstellung der Tipps für Microsoft Excel.

Der erste Teil ist der beschreibenden Statistik gewidmet. Hier werden erst die elementaren Techniken zur Datenbeschreibung behandelt: Tabellen, Diagramme und Maßzahlen (Mittelwert, Median, Standardabweichung). Danach wird auf den Zusammenhang zweier Mermale eingegangen, und der lineare Zusammenhang ausführlicher erläutert. Hierbei wird (wegen ihrer Bedeutung) auch auf die *gewichtete* lineare Regression eingegangen. Anschließend werden zunächst solche nicht-linearen Zusammenhänge behandelt, die sich unter Verwendung von Transformationen auf lineare Zusammenhänge zurückführen lassen. Als Erweiterung des linearen Zusammenhangs wird noch der quadratische Zusammenhang betrachtet.

[1] vom Science-Fiction-Autor Poul Anderson, zitiert aus Hodges et al (1975), StatLab, Box 8.

Im zweiten Teil werden die wichtigsten stochastischen Grundbegriffe erläutert, die der deduktiven Statistik zugrunde liegen. Zuerst werden Wahrscheinlichkeiten und Zufallsvariablen eingeführt, wobei von (geringen) Vorkenntnissen aus der Schule ausgegangen wird. Als konkrete Verteilungen werden die Gleichverteilung, die Normalverteilung, die Binomialverteilung und die Poissonverteilung ausführlicher behandelt (die erst im Teil 3 benötigten Chiquadrat-, t- und F-Verteilungen werden dort vorgestellt). Danach wird die Idee des Schätzens von Parametern wird am Beispiel von Erwartungswert und Standardabweichung erläutert.

Der dritte Teil ist der deduktiven (schließenden) Statistik gewidmet und behandelt Schätzungen mit Konfidenzgrenzen und statistische Tests sowie damit zusammenhängende statistische Analysen. Zuerst wird am Beispiel einer einzelnen Wahrscheinlichkeit (für ein interessierendes Ziel-Ereignis) die Idee von Konfidenzgrenzen und Tests ausführlich erläutert. Anschließend werden Vergleiche zweier Wahrscheinlichkeiten und danach Tests für ein und zwei Erwartungswerte behandelt. Hierbei werden überwiegend *asymptotische* Konfidenzgrenzen und Tests verwendet, die bei nicht zu kleinem Stichprobenumfang stets anwendbar sind – im Gegensatz zu den teilweise auch behandelten *exakten* Verfahren, die spezielle Verteilungsannahmen (wie z.B. Normalverteilung) voraussetzen. Danach wird die lineare (ungewichtete und gewichtete) Regressionsanalyse aus der Sicht der schließenden Statistik erneut behandelt. Weiter werden Anpassungstests zum Vergleich von beobachteten mit erwarteten Häufigkeiten erläutert und als Anwendung hiervon werden zweidimensionale Kontingenztafeln analysiert. Bei allen behandelten Tests wird großer Wert auf die Analyse der Testschärfe und des Fehlerrisikos 2. Art inklusive der darauf aufbauenden Versuchsplanung (Bestimmung des erforderlichen Stichprobenumfangs) gelegt.

Die statistischen Verfahren werden immer in mehreren Schritten vorgestellt, die in der Vorlesung teilweise parallel dargestellt werden, aber im Skript typischerweise in der Reihenfolge behandelt werden:

- einführendes Beispiel zur Motivation,
- allgemeines Ziel und Durchführung (mit illustrienden Graphiken),
- Zusammenfassung der Rechenschritte in einer Box,
- Analyse des einführenden Beispiels (und ggf. weiterer Beispiele).

Beim (dringend empfohlenen) Nachvollziehen der Rechen-Beispiele ist zu beachten, daß die Zwischen- und Endergebnisse hier sinnvoll gerundet angegeben sind, aber in darauf aufbauenden Berechnungen trotzdem mit den nicht-gerundeten Werten (d.h. mit voller Genauigkeit des Computers) weitergerechnet wird. Deshalb können sich beim eigenen Überprüfen der hiesigen Rechnungen geringfügige Unterschiede ergeben, allerdings sollten die Abweichungen erst in der letzten oder (bei genaueren Angaben) in den letzten beiden hier angegeben Dezimalstellen auftreten.

Der Kurs hat den Umfang von zwei Semesterwochenstunden und wird durch Übungen gleichen Umfangs begleitet Da die Veranstaltung primär für Studierende im Fach Biologie vorgesehen ist, sind auch die Anwendungsbeispiele schwerpunktmäßig aus diesem Bereich. Bei den Beispielen im Skript (und den Übungen) handelt es sich weitgehend um konkrete und realistische Anwendungen (teilweise in stark vereinfachter Darstellung). Die jeweils wechselnden Übungsaufgaben sind nicht ins Skript aufgenommen, aber die aktuelle Serie findet man im Internet (vgl. unten).

Die vorliegende 4. Auflage wurde vollständig neu (im Satzsystem Latex) von Heidi Eckl-Reichelt geschrieben. Joachim Schalthöfer hat das endgültige Layout erstellt und insbesondere die bisherigen Graphiken konvertiert bzw. neu angefertigt. Beiden möchte ich an dieser Stelle herzlich danken. Inhaltlich unterscheidet sich diese Auflage von der 3. Auflage (Juli 2004) nur durch geringgfügige Umformulierungen und Korrekturen. Ein wesentlicher Vorzug der Neuauflage ist, daß man in der digitalen PDF-Version jetzt nach Stichwörtern suchen kann (was vorher nicht möglich war).

Erfahrungsgemäß finden sich trotz Korrekturlesens vorwiegend in den neueren Passagen (aber nicht nur dort) einige Druckfehler. Bevor man daher am eigenen Verständnis zweifelt, sollte man auch einen Fehler im Skript in Erwägung ziehen. Für Hinweise auf Druckfehler oder andere Kommentare bin ich dankbar (am besten per e-mail: siehe unten).

Bremen, im April 2009 *Gerhard Osius*

PS Die aktuelle Fassung des Skriptes (sowie Übungsaufgaben und weiteres Material zur Veranstaltung) findet man unter:

http://www.math.uni-bremen.de/ osius/download/lehre/Statistik_NW/.

Die Zugangsdaten für diese Internet-Seite sowie das Kennwort für das Skript können Studierende per e-Mail erfragen:

osius@math.uni-bremen.de .

Inhaltsverzeichnis

Anhang

1 Beschreibende Statistik

Die Aufgabe der beschreibenden Statistik besteht darin, eine Menge von Daten (Datensatz) in möglichst übersichtlicher Form darzustellen. Bei größeren Datenmengen steht hierbei eine Reduktion (Zusammenfassung) der Daten im Hinblick auf wesentliche Aspekte im Vordergrund.

Für die beschreibende Darstellung von Daten verwendet man

- *Tabellen* (z.B. Häufigkeitstabellen)
- *Graphiken* (z.B. Häufigkeitsdiagramme, Punktwolken)
- *Maßzahlen* (z.B. Mittelwert, Standardabweichung, Korrelationskoeffizient).

1.1 Daten und Merkmale

Die darzustellenden Daten werden typischerweise im Rahmen einer *Studie* (z.B. über die Entwicklung von Neugeborenen), eines *Experiments* (z.B. über das Wachstum einer Pflanze) oder durch *Beobachtungen* (z.B. Monitoring der Ozon-Konzentration in der Luft) erhoben. Für jedes *Untersuchungobjekt* (z.B. Individuum einer Studie, Pflanze in einem Experiment oder Beobachtungszeitpunkt) werden dabei ein oder mehrere *Merkmale* erfasst, wobei anzustreben ist, daß alle interessierenden Merkmale vollständig, d.h. an *jedem* Untersuchungsobjekt, erhoben werden.

Als konkretes Beispiel eines umfangreichen Datensatzes dienen die *StatLab-Daten*. Die Daten wurden im Rahmen einer medizinischen Studie von Neugeborenen („Kaiser Foundation Health Plan") in Oakland (Kalifornien) unter Leitung von J. Yerushalmy in den Jahren 1961-1972 erhoben. Für jedes Kind (und für seine Eltern) wurden zahlreiche Daten zu zwei Zeitpunkten registriert: bei der Geburt (in den Jahren 1961-1963) und ca. 10 Jahre danach (zur Kontrolle). Die vollständigen Daten von je 1296 Mädchen und Jungen sind in dem Lehrbuch von Hodges, Krech und Crutchfield (1975) "*StatLab: an empirical introduction to statistics*" enthalten. Wir betrachten hier eine (zufällig erzeugte) Auswahl von je 50 Mädchen und Jungen aus der gesamten Studie und dabei auch nur einige besonders interessierende Merkmale (vgl. Anhang **C**).

Arten von Merkmalen (Variablen)

Die interessierenden Merkmale (auch *Variablen* genannt) können unterschiedlicher Natur sein und man unterscheidet (auch im Hinblick auf die anzuwendenden Methoden) verschiedene Arten.

Ein Merkmal ist *qualitativ*, wenn es verschiedene Ausprägungen besitzt, die sich *nicht* inhaltlich durch *Zahlen* repräsentieren lassen. Beispiele hierfür sind die folgenden Merkmale (mit ihren Ausprägungen): der Familienstand (ledig, verheiratet, verwitwet) und die AB0-Blutgruppe ($0, A, B$ und AB).

Ein qualitatives Merkmal mit nur *zwei* Ausprägungen heißt auch *alternativ* und ist typischerweise durch das Vorhandensein oder Fehlen eines Charakteristikums gegeben (etwa: Raucher - Nichtraucher oder weiblich - männlich).

Ein *ordinales* (qualitatives) Merkmal liegt vor, wenn sich seine Ausprägungen *sinnvoll anordnen* lassen, z.B. die Bewertung eines Zustands durch die Kategorien: schlecht - mittel - gut.

Obwohl sich die Ausprägungen qualitativer Merkmale nicht inhaltlich durch Zahlen darstellen lassen, werden sie trotzdem oft *formal* durch eine Zahl (typischerweise eine laufende Nummer) *kodiert* bzw. *klassifiziert*, z.B. bei den vier Blutgruppen durch $1 = Null, 2 = A, 3 = B, 4 = AB$. Eine solche Kodierung dient ausschließlich der Identifikation der Ausprägung durch Zahlen und stellt keinen *inhaltlichen* Zusammenhang zwischen Ausprägung und Zahlen-Kode dar. Insbesondere ist es unsinnig Mittelwerte solcher Zahlen-Kodes zu bilden, z.B. ist eine *mittlere Blutgruppe* von 2,25 für eine Population eine sinnlose Angabe.

Im Gegensatz zu den qualitativen Merkmalen sind *quantitative* Merkmale solche, deren Ausprägungen auf natürliche Weise durch *Zahlen* beschrieben oder gemessen werden. Man unterscheidet hier noch zwischen *diskreten* und *stetigen* (kontinuierlichen) Merkmalen. Ein *diskretes* Merkmal nimmt nur isolierte Werte an, typischerweise ganze Zahlen wie z.B. die Anzahl von Mutanten einer Zellkultur oder die konsumierten Zigaretten pro Tag. Demgegenüber kann ein *stetiges* Merkmal kontinuierliche Werte aus einem Zahlenintervall annehmen (z.B. das Gewicht in kg oder die Größe in cm), die in der Praxis allerdings typischerweise *gerundet* werden (z.B auf eine ganze Zahl oder eine Nachkommastelle).

Rundung:
Unter Berücksichtigung der Meßgenauigkeit (oder aus Gründen der Übersicht) werden Meßwerte oft *gerundet* angegeben (*Zur Erinnerung*: Die Ziffern 0 bis 4 werden *abgerundet* und 5 bis 9 *aufgerundet*). Eine gerundete Angabe ist immer so zu interpretieren, daß der genaue Wert in einem entsprechenden *Intervall* liegt, z.B.

$$\text{Gewicht} \;=\; 75\ [kg] \qquad \text{bedeutet} \qquad 74{,}5 \leq Gewicht < 75{,}5.$$

Hierbei ist das Symbol "$\leq$" die mathematische Abkürzung für "*kleiner oder gleich*" und "$<$" steht für "*echt kleiner*". (Als Eselbrücke: die "kleine" Spitze zeigt zum kleineren Wert und die "große" Öffnung zum größeren Wert).

Manchmal wird auch nur *abgerundet*, z.B. beim Alter in *vollendeten Lebensjahren*:

$$\text{Alter} \;=\; 30 \text{ Jahre} \qquad \text{bedeutet} \qquad 30 \leq \text{ Alter} < 31.$$

Durch Rundung werden stetige Merkmale in diskrete überführt, wie z.B. das Alter in vollendeten Jahren. Daher ist der Unterschied zwischen stetigen und diskreten Merkmalen in der Praxis eher fließend, aber dennoch von theoretischer Bedeutung.

1.2 Häufigkeiten und Diagramme

Unser Ausgangspunkt ist ein Datensatz über mehrere Untersuchungsobjekte, deren Anzahl n auch als Stichprobenumfang bezeichnet wird. Tritt ein Merkmal im Datensatz in insgesamt K *verschiedenen* Ausprägungen $A_1, \ldots, A_K$ auf, so interessiert man sich in erster Linie für die *Häufigkeiten* dieser Ausprägungen A_k.

Die absolute *Häufigkeit* n_k von A_k ist die Anzahl aller Untersuchungsobjekte mit dieser Ausprägung. Die Summe aller absoluten Häufigkeiten ist die Gesamtzahl n aller Untersuchungsobjekte:

$$n = n_1 + n_2 + \ldots + n_K = \sum_{k=1}^{K} n_k = \sum_k n_k \qquad (\textit{Stichprobenumfang})$$

Die relative *Häufigkeit (Frequenz)* von A_k ist der Anteil

$$f_k = \frac{n_k}{n}.$$

Er wird meist in Prozent angegeben, d.h. als $100 \cdot f_k\%$. Es gilt immer

$$0 \leq f_k \leq 1 \qquad \text{bzw.} \qquad 0 \leq 100 \cdot f_k \leq 100,$$

und die Summe der relativen Häufigkeiten ergibt immer 1 bzw. 100% – was in der Praxis zur Kontrolle überprüft werden sollte:

$$\sum_{k=1}^{K} f_k = 1 \qquad \text{bzw.} \qquad \sum_{k=1}^{K} 100 \cdot f_k = 100.$$

Tabellierung:
Die absoluten und relativen Häufigkeiten werden in einer *Häufigkeitstabelle* für das Merkmal zusammengefasst. Eine solche Tabelle (vgl. *Tabelle 1*) enthält typischerweise je eine Spalte für die laufende Nr. k, die Ausprägung A_k und die Häufigkeiten n_k und f_k. Zur Kontrolle sollten auch die Summen über alle n_k bzw. f_k berechnet werden.

Graphische Darstellungen:
Zur Veranschaulichung von Häufigkeiten werden diese durch Flächen dargestellt, deren Flächeninhalte den Häufigkeiten entsprechen. Typische Darstellungen sind (vgl. Abb. 1):

- *Säulen-Diagramm (Histogramm)*
  Die Häufigkeiten entsprechen den Flächen von Säulen und (bei *gleicher* Säulenbreite) somit auch der *Höhe* der Säulen.

- *Kreis-Diagramm (Torten-Diagramm)*
  Aufteilung eines Kreises in Segmente, wobei der Winkel w_k (und damit auch die zugehörige Fläche) des A_k repräsentierenden Kreissegments der relativen Häufigkeit f_k entspricht, d.h.

$$w_k = f_k \cdot 360°.$$

Im Säulendiagramm lassen sich die einzelnen Häufigkeiten direkt (durch die Säulenhöhe) untereinander vergleichen. Demgegenüber vermittelt das Kreis-Diagramm eine übersichtliche (den Häufigkeiten entsprechende) Aufteilung der Kreisfläche in der auch „runde“ Anteile wie z.B. $\frac{1}{4}$, $\frac{1}{2}$ und $\frac{3}{4}$ leicht erkennbar sind.

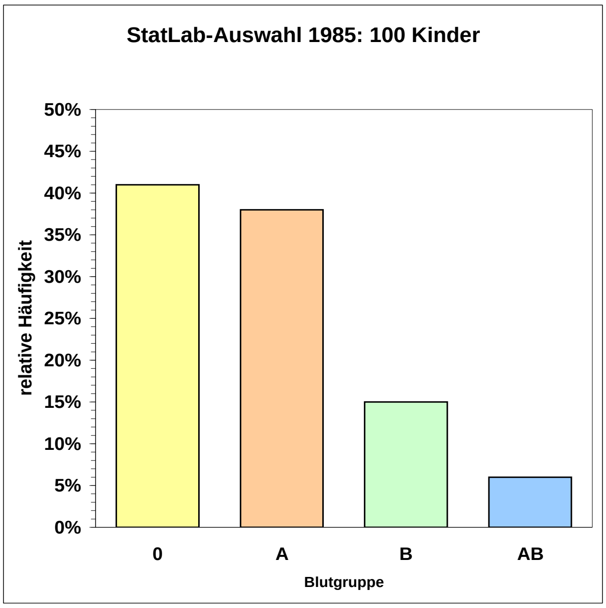
Beispiel: AB0-Blutgruppe in der StatLab-Auswahl 1985

Tabelle 1: Häufigkeitstabelle für die $n = 100$ Kinder

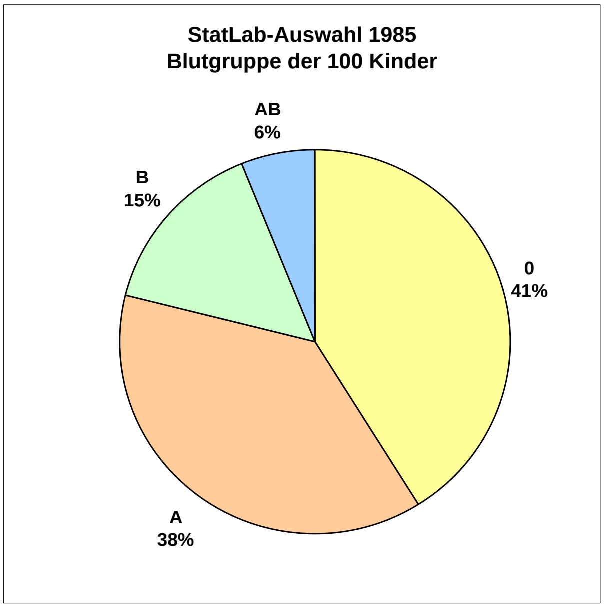
Nr k	Merkmal/Klasse A_k	Häufigkeit absolut n_k	Häufigkeit relativ f_k
1	0	41	41,0%
2	A	38	38,0%
3	B	15	15,0%
4	AB	6	6,0%
$\sum$		100	100,0%

Abb. 1: Darstellung der Blutgruppenverteilung der $n = 100$ Kinder

Säulen-Diagramm (Histogramm)



Kreis-Diagramm



Klassifizierung eines Merkmals

Liegen sehr viele verschiedene Ausprägungen eines Merkmals vor (typischerweise bei stetigen Merkmalen), so liefert die Häufigkeitstabelle keine sinnvolle Zusammenfassung mehr, und die Diagramme werden unübersichtlich. Hier ist meist eine Reduktion sinnvoll, indem man das ursprüngliche Merkmal neu klassifiziert, z.B. durch starke Rundung bei einem stetigen Merkmal. Die so entstandenen, neuen Klassen $k = 1, \ldots, K$ entsprechen dann den Ausprägungen $A_1, \ldots A_K$ des *klassififzierten* Merkmals. Die Anzahl K der Klassen soll dabei nicht zu groß sein, als grobe Faustregel kann gelten:

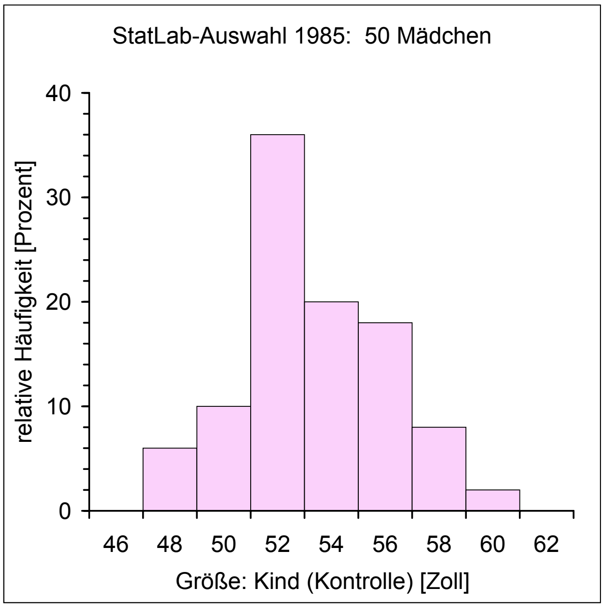
$$\text{(Anzahl } K \text{ der Klassen)} \quad \leq \quad \frac{1}{5} \text{ (Anzahl } n \text{ der Untersuchungsobjekte)}$$

Beispiel: Größe [Zoll] der 50 Mädchen zum Kontrollzeitpunkt in der StatLab-Auswahl 1985

Tabelle 2: Häufigkeitstabelle der klassifizierten Größe

Nr	Merkmal			Klassen-Mitte	Häufigkeit absolut	Häufigkeit relativ
	von (incl.)		bis (excl.)			
k	[Zoll]		[Zoll]	[Zoll]	n_k	f_k
1	47	-	49	48	3	$6,0\%$
2	49	-	51	50	5	$10,0\%$
3	51	-	53	52	18	$36,0\%$
4	53	-	55	54	10	$20,0\%$
5	55	-	57	56	9	$18,0\%$
6	57	-	59	58	4	$8,0\%$
7	59	-	61	60	1	$2,0\%$
$\sum$		-			50	100%

Abb. 2: Säulen-Diagramm (Histogramm) der Größen-Verteilung der $n = 50$ Mädchen



Tipps für ***Microsoft Excel***

Diagramme: Die oben genannten (und zahlreiche weitere) *Grafiken* lassen sich unter Verwendung des *Diagramm-Assistenten* von *Excel* einfach erstellen.

Klassifizierung: Eine *Häufikeitstabelle* für ein *qualitatives* Merkmal (z.B. die Blutgruppe) erzeugt man aus den Originaldaten mit dem Befehl ZÄHLENWENN. Und für ein *quantitatives* Merkmal (z.B. die Körpergröße) läßt sich eine Häufigkeitstabelle mit dem Befehl HÄUFIGKEIT in einer *Matrix-Formel* erstellen (zur *Eingabe* von *Matrix-Formeln* konsultiere man die *Excel-Hilfe*).

1.3 Maßzahlen: Median, Mittelwert, Standardabweichung

Der Ausgangspunkt für die folgenden Betrachtungen ist ein quantitatives Merkmal x in einem Datensatz mit n Untersuchungsobjekten. Unser Ziel ist es, die beobachteten x-Werte $x_1, \ldots, x_n$ (*Stichprobe vom Umfang* n) durch aussagekräftige *Maßzahlen* zusammenzufassen, d.h. wir wollen die Daten auf wesentliche Aspekte reduzieren. Hierfür geben wir *Lage-* und *Streuungsmaße* an, die folgende Fragen beantworten:

Wo *liegen* die Werte (im Mittel)?
Wie stark *streuen* die Werte?

Die Maßzahlen werden zunächst inhaltlich erläutert (**1.3.1-2**), dann wird auf ihre Berechnung eingegangen (**1.3.3**) und abschließend wird ein Anwendungsbeispiel durchgerechnet (**1.3.4**).

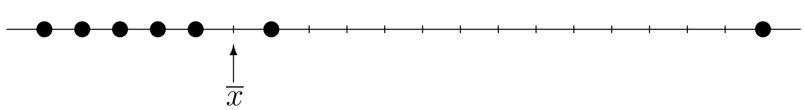
1.3.1 Lagemaße: Mittelwert und Median

Zur Charakterisierung der *Lage* der Werte $x_1, \ldots, x_n$ sucht man einen *mittleren Wert.* Die Präzisierung eines solchen mittleren Wertes kann auf verschiedene Weise erfolgen, von denen wir nur die beiden wichtigsten besprechen und vergleichen: das *arithmetische Mittel (Mittelwert)* und den *Median.*

Der *Mittelwert* $\overline{x}$ der Werte $x_1, \ldots, x_n$ ist definiert als ihr arithmetisches Mittel,

$$\boxed{\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textit{(Mittelwert, arithmetisches Mittel)}\ .}$$

Stellt man sich die Werte als gleichschwere Kugeln (Massen) auf der Zahlengeraden vor, so entspricht der Mittelwert dem *Schwerpunkt* der x-Werte bzw. Kugeln,



d.h. die Summe aller Abweichungen vom Mittelwert ist Null:

$$\sum_i (x_i - \overline{x}) = 0.$$

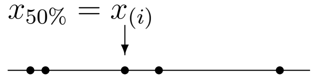
Der *Median* $x_{50\%}$ (zentraler Wert) teilt die Werte $x_1, \ldots, x_n$ derart, daß *gleichviel* Werte *unterhalb* wie *oberhalb* des Medians liegen. Zur Bestimmung des Medians werden die Werte zunächst *aufsteigend* nach ihrer Größe *sortiert* und man erhält die sogenannte *Rangliste* der Werte:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} \qquad \textit{(Rangliste der x-Werte)}\ .$$

Insbesondere ist also $x_{(1)}$ der *kleinste* und $x_{(n)}$ der *größte* Wert. Im Gegensatz zu $x_{(1)}$ und $x_{(n)}$ sind x_1 der erste und x_n der letzte Wert in der *gegebenen* Reihenfolge der x-Werte in der Stichprobe.

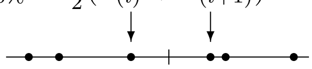
Der Median $x_{50\%}$ zerlegt die Rangliste derart, daß *gleichviele* Werte unterhalb wie oberhalb liegen, d.h.

- für *ungerades* n ist: $x_{50\%} = x_{(i)}$ *mit* $i = \frac{1}{2}(n+1)$



$(n = 5)$

- für *gerades* n ist: $x_{50\%} = \frac{1}{2}(x_{(i)} + x_{(i+1)})$ *mit* $i = \frac{1}{2}n$

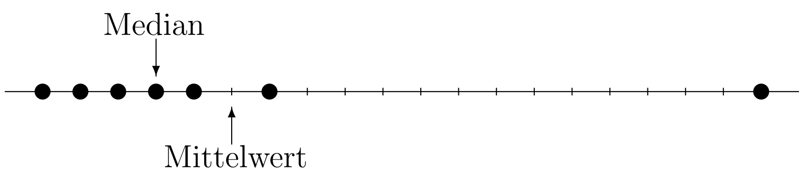


$(n = 6)$

Median vs. Mittelwert

Der Mittelwert hängt von allen x-Werten ab, und wird von extremen *Ausreißern* (nach oben oder unten) stark beeinflußt. Der Median ist dagegen *robust* gegen solche Ausreißer, weil er nur von den *zentralen* x-Werten abhängt.

Beispiel 1:



Verschiebt man den größten Wert weiter nach rechts, wird nur der Mittelwert größer, aber der Median ändert sich nicht.

Beispiel 2:
Bei einem Prozeß in den USA machen drei Gutachter folgende Vorschläge über die Höhe des Schadensersatzanspruchs:

Gutachter	Betrag x[\$]
1	3 000
2	30 000
3	3 000 000

Es ist hier von entscheidender Bedeutung, ob das Gericht den Mittelwert $\overline{x} = 1\,011\,000$ oder den Median $x_{50\%} = 30\,000$ der Gutachtervorschläge wählt.

Beispiel 3:
Bei der Angabe eines “mittleren Preises” für ein Produkt ist es sinnvoll, den *Median* der Einzelpreise verschiedener Anbieter zu verwenden, wie es z.B. in der Zeitschrift *test* der *Stiftung Warentest* üblich ist. Hierdurch werden Verzerrungen sowohl durch „Dumpingpreise“ (in Ballungsgebieten mit viel Konkurrenz) als auch durch „Wucherpreise“ (in ländlichen Gebieten ohne Konkurrenz) vermieden.

1.3.2 Streuungsmaße: Standardabweichung, Varianz und Spannweite

Der umgangssprachliche Begriff der Streuung von Zahlen (oder Punkten) wird in der Statistik durch unterschiedliche Streuungsmaße präzisiert. Als Maßzahl für die Streuung der Werte $x_1, \ldots, x_n$ wird vorwiegend die *Standardabweichung* bzw. deren Quadrat, die *Varianz*, und zusätzlich auch die *Spannweite* verwendet.

Standardabweichung und Varianz

Ausgangspunkt für die Berechnung der Standardabweichung sind die *Abweichungen* $(x_i - \overline{x})$ der x-Werte von ihrem Mittelwert. Da man auf das *Vorzeichen* dieser Abweichung keinen Wert legt (d.h. ob x_i *unter-* oder *oberhalb* von $\overline{x}$ liegt), wird die *quadratische Abweichung* $(x_i - \overline{x})^2$ verwendet. Man könnte auch den *Absolutbetrag* $\mid x_i - \overline{x} \mid$ nehmen, aber das hat sich nicht durchgesetzt, weil die Berechnungen mit dem *Quadrat* einfacher sind.

Die *(empirische) Varianz* der x-Werte ist nun die *mittlere quadratische Abweichung vom Mittelwert*:

$$\boxed{s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad (\textit{empirische Varianz})}$$

Man beachte, daß im Nenner die um 1 *reduzierte* Anzahl $n - 1$ aller x-Werte und nicht der Stichprobenumfang n steht. Eine genaue Begründung hierfür wird erst später (im Abschnitt **2.6.5**) gegeben. Grob gesprochen liegt es daran, daß man zur Bestimmung von s_x^2 bereits *einen* aus den x-Werten *berechneten* Parameter (den Mittelwert $\overline{x}$) verwendet.

Da die Varianz s_x^2 als *Dimension* das Quadrat der Dimension der x-Werte hat (wenn x eine Länge ist, so ist die Varianz eine Fläche), verwendet man ihre Wurzel als Streuungsmaß, die sogenannte *(empirische) Standardabweichung* s_x, die von gleicher Dimension wie die x-Werte ist:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (\textit{empirische Standardabweichung})$$

Je *größer* s_x^2 bzw. s_x ist, desto *stärker* streuen die x-Werte. Für $s_x = 0$ sind alle x-Werte *gleich* (keine Streuung) – und umgekehrt. Im Bereich $\overline{x} \pm 2s_x$ liegen immer mindestens 75% aller x-Werte, oft jedoch bis zu 95% (z.B. bei der Normalverteilung, vgl. **2.4**).

Warnung: Im Gegensatz zur obigen Definition werden gelegentlich auch die entsprechenden Streuungsmaße verwendet, bei denen man durch n statt durch $n-1$ dividiert (was wir hier *nicht* wollen). Bei Rechnerprogrammen und Taschenrechnern sind meist beide Versionen (mit $n-1$ und mit n) implementiert und man sollte sich vergewissern, daß man die obige Version verwendet.

Spannweite

Ein völlig anderes und leicht zu berechnendes Streuungsmaß ist die Spannweite, definiert als Differenz vom größten Wert $x_{(n)}$ (Maximum) zum kleinsten Wert $x_{(1)}$ (Minimum) in der Rangliste der x-Werte (vgl. **1.3.1**).

$$\textit{Spannweite} \;=\; (\textit{größter Wert}) - (\textit{kleinster Wert}) \;=\; \textit{Maximum} - \textit{Minimum}.$$

Für die Berechnung wird nicht die gesamte Rangliste, sondern nur das Maximum $x_{(n)}$ und das Minimum $x_{(1)}$ der x-Werte benötigt. Diese extremen Werte sind meist ohnehin von Interesse, z.B. für die Skalierung von graphischen Darstellungen.

1.3.3 Berechnung von Mittelwert und Standardabweichung

Die Varianz s_x^2 der x-Werte $x_1, \ldots, x_n$ kann auch berechnet werden ohne hierzu die einzelnen Abweichungsquadrate $(x_i - \overline{x})^2$ vom Mittelwert $\overline{x}$ bestimmen zu müssen. Durch algebraische Umformungen läßt sich die Varianz durch die Summe $\sum x$ der x-Werte und die Summe $\sum x^2$ der *quadrierten* x-Werte darstellen. Die zugehörigen Schritte zur Berechnung von Mittelwert, Varianz und Standardabweichung sind in der *Rechen-Box 1* zusammengefaßt.

Rechen-Box 1 für Daten einer Variablen x	
Stichprobenumfang	n
x-Werte:	$x_1,\ x_2,\ \ldots,\ x_n$
Summen:	
x-Werte	$\sum x\ =\ x_1 + x_2 + \ldots + x_n$
x-Quadrate	$\sum x^2\ =\ x_1^2 + x_2^2 + \ldots + x_n^2$
Mittelwert:	$\bar{x} = \frac{1}{n}(\sum x)$
Summe der quadratischen Abweichungen:	
	$Sxx = \sum_i (x_i - \bar{x})^2 = (\sum x^2) - \frac{1}{n}(\sum x)^2 = (\sum x^2) - n\bar{x}^2$
(empirische) Varianz:	$s_x^2 = \dfrac{1}{n-1} Sxx$
(empirische) Standardabweichung:	$s_x = \sqrt{\dfrac{1}{n-1} Sxx}$

Skalenverschiebung:
Bei (heute nicht mehr zeitgemäßen) Berechnungen per Hand oder mit einem Taschenrechner kann man zur Vermeidung großer Zahlen (beim Quadrieren der x-Werte) die x-Werte um einen Wert c *verschieben* und mit den verschobenen Werten $u = x - c$ rechnen. Die *Mittelwerte* sind dann entsprechend verschoben,

$$\overline{u} = \bar{x} - c \qquad \text{bzw.} \qquad \bar{x} = \bar{u} + c,$$

und die *Standardabweichungen* der $x-$ und u-Werte stimmen sogar überein:

$$s_x = s_u \qquad \text{und} \qquad Sxx = Suu.$$

Lineare Transformation:
Transformiert man die x-Werte (z.B. beim Übergang auf eine andere Maß-Einheit) linear zu

$$u = a + bx,$$

so transformiert sich der Mittelwert entsprechend, d.h.

$$\overline{u} = a + b\overline{x}$$

und die Standardabweichung der u-Werte ergibt sich zu:

$$s_u = |b| s_x \qquad \text{und} \qquad Suu = b^2 Sxx$$

wobei $|b|$ den Absolutbetrag von b bezeichnet.

Beispiele: Umrechnung der Länge von Zoll bzw. Inch (x) in cm (u):

$$u = 2{,}54x \ .$$

Umrechnung der Temperatur von °Celsius (x) in °Fahrenheit (u)

$$u = 32 + 1{,}8x \ .$$

Umrechnung eines Umsatzes (x) auf den Rückgang (u) des Umsatzes im Bezug auf den zur Kostendeckung erforderlichen Umsatz a:

$$u = a - x \ .$$

Regeln zur Genauigkeit

Generell sollen alle Berechnungen mit *maximaler* Genauigkeit erfolgen, d.h. weder die Eingangsdaten (x-Werte) noch die *Zwischenergebnisse* $\sum x, \sum x^2$ und Sxx dürfen gerundet werden. Lediglich die *Endergebnisse* sind sinnvoll zu runden.

Konkret ist die Summe $\sum x$ mit *derselben* Anzahl von Nachkommastellen anzugeben wie die x-Werte, aber die Summe $\sum x^2$ und Sxx mit der *doppelten Anzahl* von Nachkommastellen. Wenn man mit einem Taschenrechner arbeitet, sollte man die Zwischenergebnisse mit voller Genauigkeit im Rechner *speichern* und mit diesen (genauen) Werten weiterrechnen.

Als Endergebniss werden *Mittelwert* und *Standardabweichung* meist mit ein bis zwei Nachkommastellen *mehr* als die x-Werte angegeben, bei extrem großen Umfängen sogar mit *zusätzlichen* Nachkommastellen.

1.3.4 Anwendung: StatLab-Auswahl 1985

Beispiel: Körpergröße der Mädchen zum Kontrollzeitpunkt

Zur Bestimmung des Medians und der Spannweite sind die Körpergrößen [Zoll] der $n = 50$ Mädchen zunächst aufsteigend zu sortieren und die entsprechende Rangliste ist in Spalte 2 der *Tabelle 1* (auszugsweise) wiedergegeben. Die Tabelle enthält weiter die quadrierten Größen x^2 sowie die zur Berechnung von Mittelwert und Standardabweichung erforderlichen Summen $\sum x$ und $\sum x^2$.

Tabelle 1: Rangliste.

Rang	x	x^2	Nr.
1	47,4	2246,76	44
2	48,7	2371,69	9
3	48,9	2391,21	8
:	:	:	:
:	:	:	:
:	:	:	:
25	52,9	2798,41	22
26	52,9	2798,41	49
:	:	:	:
:	:	:	:
:	:	:	:
48	57,5	3306,25	45
49	57,9	3352,41	32
50	60,0	3600,00	50
$\sum$	2661,3	142002,59	

Aus der Tabelle 1 ergeben sich mit den Formeln aus der *Rechen-Box 1* die folgenden Maßzahlen:

Spannweite:	$x_{(50)} - x_{(1)} = 12{,}6;$	$n = 50$
Median:	$x_{50\%} = 52{,}9$	
Mittelwert:	$\bar{x} = 53{,}23;$	
Standardabweichung:	$s_x = 2{,}68;$	$Sxx = 352{,}2362.$

Man beachte, daß der Median und der Mittelwert sich hier zwar nur geringfügig unterscheiden, aber nicht *exakt* übereinstimmen. Zur Veranschaulichung sind die Meßwerte in *Abbildung 1* zusammen mit dem Median, dem Mittelwert und den Bereichen $\bar{x} \pm 1s_x$ sowie $\bar{x} \pm 2s_x$ dargestellt.

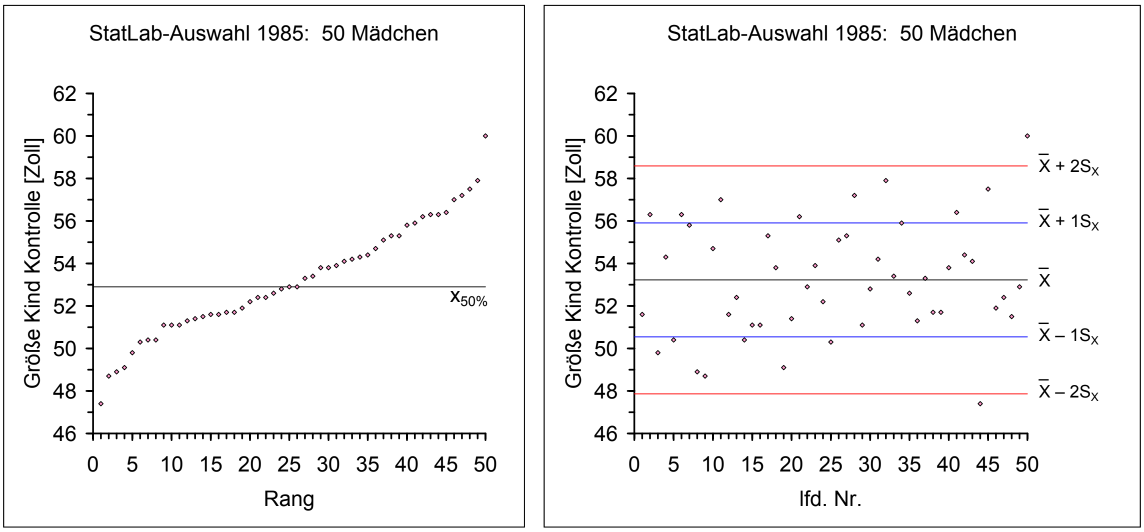


Abb. 1. *Links:* Die *geordneten* x-Werte mit ihrem Median $x_{50\%}$.

Rechts: Die *ungeordneten* x-Werte mit ihrem Mittelwert $\bar{x}$ und den durch die Standardabweichung s_x gegebenen Bereichen $\bar{x} \pm 1s_x$ sowie $\bar{x} \pm 2s_x$. Man erkennt deutlich, daß die Standardabweichung ein Maß für die *Streuung* ist: der größte Teil der x-Werte "streut" im Bereich $\bar{x} \pm 1s_x$ und *außerhalb* des Bereichs $\bar{x} \pm 2s_x$ liegen nur sehr wenig Werte.

Tipps für ***Microsoft Excel***

Funktionen: Die folgenden *Statistik-Funktionen* stehen in *Excel* direkt zur Verfügung (mit Erläuterung in Klammern, sofern erforderlich):

MITTELWERT		MEDIAN	
STABW	(= *Standardabweichung*)	VARIANZ	
MIN	(= *Minimum*)	MAX	(= *Maximum*).

Rangliste: Man erhält sie mit dem Befehl *Sortieren* im *Excel*-Menü *Daten.*

Diagramme: Eine Darstellung (wie *Abb. 1*) erstellt man mit dem *Diagramm-Assistenten.* Die waagerechten Hilfslinien in *Abb. 1* werden dabei *nicht* dargestellt, lassen sich aber als „Geraden“ einzeichnen (vgl. auch **1.5.7**).

Formate: Zur Übersicht sollte man Zahlen in *Excel formatieren*, indem man im Menü *Format - Zellen* für jedes Feld (oder ganze Bereiche) eine sinnvolle Anzahl von *Nachkommastellen* einstellt, wie z.B. in den *Regeln zur Genauigkeit* in **1.3.1**. Die *Formatierung* hat *keinen* Einfluß auf die interne *Rechengenauigkeit* von *Excel.*