JUMBO

Java-unterstützte Münsteraner Biometrie-Oberfläche

Institut für Medizinische Informatik
und Biomathematik

Münster
2009

I. Deskriptive Statistik I

1.1 Lernziele zur Deskriptiven Statistik I

- Merkmalstypen: qualitativ nominal und ordinal, quantitativ stetig und diskret.
- Identifikationsgröße, Einflussgröße, Zielgröße.
- Grundgesamtheit, Stichprobe, Beobachtungseinheit.
- Untersuchungstypen: retrospektive Erhebung, prospektive Erhebung, Experiment.
- Tabellarische und grafische Darstellung der Daten eines qualitativen Merkmals
- Tabellarische und grafische Darstellung der Daten eines quantitativen Merkmals
- Statistische Maßzahlen
- Lagemaße
- Streuungsmaße
- Empirische Verteilungsfunktion

1.2 Studienbeschreibung

Die statistischen Methoden sollen anhand eines realen Beispiels eingeführt und erläutert werden. Hierzu dient eine multizentrische, klinische Studie zur Therapie der akuten myeloischen Leukämie beim Erwachsenen, die im Folgenden kurz erläutert wird [1]. Diese Studie wurde von der deutschen AML-Cooperative Group (AMLCG) durchgeführt.

Die akute myeloische Leukämie (AML) ist eine maligne Erkrankung des Teils des blutbildenden Systems, der für die Bildung von Granulozyten und Monozyten verantwortlich ist. Sie führt zu einer z. T. massiven Vermehrung unreifer Vorstufen (Blasten) der Myelopoese im Knochenmark und im peripheren Blut.

Die Behandlung der AML erfolgt durch intensive Chemotherapie. Diese wird in zwei Stufen durchgeführt. Die erste Stufe bildet die so genannte Induktionstherapie, deren Ziel das Erreichen einer kompletten Remission (CR für engl.: complete remission) ist. Erreichen der CR bedeutet, dass alle Krankheitssymptome beseitigt werden und sich das Blutbild wieder normalisiert. Hierfür gibt es feste Kriterien, die u. a. vorsehen, dass der Anteil der Blasten < 5 % liegen muss. Die Behandlung besteht aus mehrtägigen Chemotherapiekursen. Sie werden häufig durch die eingesetzten Substanzen gekennzeichnet, in der hier betrachteten Studie z. B. durch **TAD** für **T**hioguanin, **A**RA-C (Cytarabin) und **D**aunorubicin, bzw. **HAM** für **H**igh dose **A**RA-C und **M**itoxantron. Darreichungsform und Dosis der Substanzen sind in den entsprechenden Studienprotokollen genau festgelegt. Wird eine CR erreicht, dies ist heute je nach Patientenauswahl bei etwa 60 - 70 % aller Patienten der Fall, kommt es zur zweiten Stufe, der so genannten Postremissions- oder Erhaltungstherapie, deren Ziel es ist, die Remission möglichst lange zu erhalten.

In dieser Erhaltungsphase kommen als weitere Therapieverfahren auch allogene oder autologe Stammzelltransplantationen in Betracht.

Bei der Mehrzahl der Patienten muss mit einem Rezidiv der AML gerechnet werden.

Langzeitremissionen von mehr als 5 Jahren Dauer werden bei etwa 25% der Patienten erreicht. Diese dürfen als geheilt angesehen werden.

Die hier als Beispiel betrachteten multizentrischen, randomisierten, klinischen Studien der deutschen AML-Cooperative Group (AMLCG) wurden in Münster, München, und Braunschweig koordiniert. Die Studien wurden im Institut für Medizinische Informatik und Biomathematik (IMIB) in Münster biometrisch ausgewertet. Ziel der Studien war der randomisierte Vergleich zweier Erhaltungstherapien (Studie AMLCG-92) bzw. der randomisierte Vergleich zweier Induktions- und zweier Erhaltungstherapien (AMLCG-99). Gegenstand der Publikation [1] war es, die Altersabhängigkeit der Therapieergebnisse aufzuzeigen.

[1] Thomas Büchner, Wolfgang E. Berdel, Claudia Haferlach et al.: Age-Related Risk Profile and Chemotherapy Dose Response in Acute Myeloid Leukemia: A Study by the German Acute Myeloid Leukemia Cooperative Group
Journal of Clinical Oncology 27, 61-69, 2009.

1.3 Grundlegende Begriffe

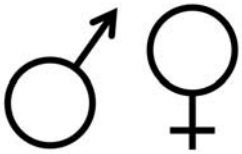
1.3.1 Merkmalstypen

In wissenschaftlichen Untersuchungen werden Objekte oder Individuen, die sogenannten **Beobachtungseinheiten**, im Hinblick auf vorgegebene Fragestellungen untersucht. Diese Beobachtungseinheiten sind Träger von **Merkmalen**, die bei den verschiedenen Beobachtungseinheiten in unterschiedlichen **Ausprägungen** vorliegen. Sie können z. B. unterschiedliches Gewicht, unterschiedliches Alter oder unterschiedliche Farbe haben. Man unterscheidet qualitative und quantitative Merkmale.

Ein Merkmal heißt **qualitativ**, wenn seine verschiedenen Ausprägungen begrifflich voneinander unterschiedene Kategorien ohne zahlenmäßige Ordnung sind, die sich gegenseitig ausschließen (**disjunkt**) und alle denkbaren Fälle abdecken.

Beispiel 1.1

- *Geschlecht mit den beiden Ausprägungen "männlich" und "weiblich"*



- *Blutgruppe mit den Ausprägungen "A", "B", "AB" und "0"*

Blutgruppe	Genotyp	Häufigkeiten in Deutschland	Antigene der Erythrozyten	Antikörper im Serum
A	AA oder A0	44%	A	Anti-B (=β)
B	BB oder B0	12%	B	Anti-A (=α)
AB	AB	6%	A und B	keine
0	0	38%	weder A noch B	α und β

- *Schweregrad einer Erkrankung mit den Ausprägungen leicht, mittel und schwer*

Wenn wie im Beispiel "*Schweregrad einer Erkrankung*" eine **natürliche Ordnung** zwischen den verschiedenen Ausprägungen gegeben ist, nennt man das Merkmal qualitativ **ordinal**. Wenn das wie im Beispiel "*Geschlecht*" und "*Blutgruppe*" nicht gegeben ist, nennt man das Merkmal qualitativ **nominal**.

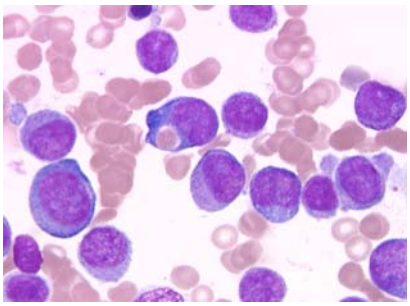
Ein Merkmal heißt **quantitativ**, wenn seine unterschiedlichen Ausprägungen unterschiedliche **Vielfache einer gegebenen Maßeinheit** sind.

Beispiel 1.2

- *Körpergröße in cm*

- *Gewicht in kg*

- *Anzahl AML-Zellen*



- *Anzahl der Geschwister*

Wenn wie im Beispiel "*Körpergröße*" und "*Gewicht*" zwischen je zwei Ausprägungen auch jeder dazwischenliegende Wert als Ausprägung des Merkmals denkbar ist, nennt man das Merkmal **quantitativ stetig**. Wenn das wie im Beispiel "*Anzahl der Geschwister*" und "*Anzahl AML-Zellen*" nicht gegeben ist, nennt man das Merkmal **quantitativ diskret**.

Vor Beginn einer Untersuchung muss für jedes betrachtete Merkmal die Liste der möglichen Ausprägungen festgelegt werden. Nur anhand dieser Liste lässt sich der Typ des Merkmals erkennen, und nur wenn man den Merkmalstyp kennt, lässt sich bei gegebener Fragestellung die richtige statistische Auswertungsmethode bestimmen.

Die Liste der Ausprägungen eines Merkmals muss so beschaffen sein, dass sich je zwei verschiedene Ausprägungen gegenseitig ausschließen, und es muss auch jede denkbare Ausprägung genannt sein. Man sagt auch, die Liste der Ausprägungen muss **disjunkt** und **vollständig** sein.

Nur auf diese Weise ist sichergestellt, dass bei jeder Beobachtungseinheit der Untersuchung genau eine Ausprägung des betrachteten Merkmals vorliegt. Die im Laufe einer Untersuchung an den Beobachtungseinheiten festgestellten Ausprägungen eines Merkmals sind die **Daten**. Diese sind Gegenstand der statistischen Auswertung.

1.3.2 Identifikationsgröße, Zielgröße, Einflussgröße

Die Merkmale, die in einer Untersuchung betrachtet werden, haben im **Versuchsplan** dieser Untersuchung unterschiedliche Aufgaben. Sie können auftreten als

- **Identifikationsgröße**,
- **Zielgröße**,
- **Einflussgröße**.

Identifikationsgrößen braucht man zur Identifikation der Beobachtungseinheiten. Dies wird z. B. bei Fehlerkontrollen erforderlich. Wenn z. B. bei Plausibilitätsbetrachtungen unmögliche Werte festgestellt werden, kann man nur mit Hilfe einer Identifikationsgröße feststellen, zu welcher Beobachtungseinheit dieser Wert gehört. Außerdem braucht man Identifikationsgrößen, wenn in einer Untersuchung Daten aus unterschiedlichen Quellen zusammengeführt werden müssen.

Beispiel 1.3
In einer klinischen Untersuchung sollen anamnestische Daten, Labordaten und Röntgenbefunde ausgewertet und in Beziehung gesetzt werden. Dies ist nur möglich, wenn es eine Identifikationsgröße gibt, die in allen drei Datensätzen vorliegt - z. B. eine Krankenblattnummer oder Name, Geburts- und Aufnahmedatum des Patienten.

Patient

Nachname _____

Vorname _____

Geb. Datum ☐☐.☐☐.☐☐☐☐ **Geschl.** ☐**M** / ☐ **W**

Pat.-Nr. ☐☐—☐☐☐☐☐☐☐

Nur über diese Identifikationsgröße bzw. Identifikationsgrößen ist es möglich, die Daten, die zum gleichen Patienten gehören, zusammenzuführen. Dieses Beispiel zeigt auch, dass zur Identifikation möglicherweise mehrere Merkmale erforderlich sein können.

Zielgröße nennt man das Merkmal, das der eigentliche Gegenstand der Untersuchung ist. Eine Untersuchung kann mehrere Zielgrößen haben.

Beispiel 1.4
*In klinischen Studien zur Behandlung der **akuten myeloischen Leukämie (AML)** sind mögliche **Zielgrößen***

- *die Überlebenszeit*
- *die Zeit bis zum Auftreten eines Rezidivs*

- *die Lebensqualität der behandelten Patienten.*

Einflussgrößen nennt man die Merkmale, die einen Einfluss auf die Zielgröße haben. Meist ist es Gegenstand der Untersuchung, einen solchen Einfluss nachzuweisen, ihn genauer zu beschreiben oder auch nachzuweisen, dass kein Einfluss besteht.

Beispiel 1.5
*In klinischen Studien zur Behandlung der AML sind mögliche **Einflussgrößen***

- *das Alter*
- *die Therapie*
- *der Zelltyp*
- *das Geschlecht.*

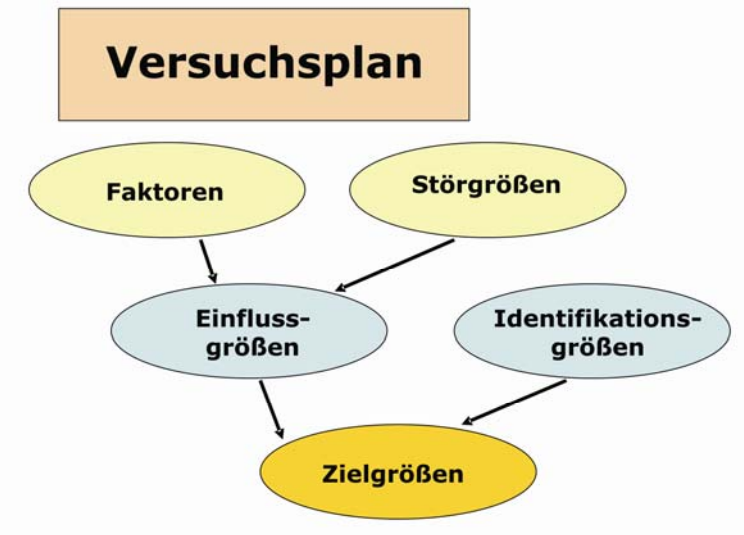
Natürlich kann man in einer Untersuchung nicht alle Einflussgrößen erfassen. Man nennt die erfassten Einflussgrößen **Faktoren**, die nicht erfassten **Störgrößen**.

Beispiel 1.6
*In klinischen Studien zur Behandlung der AML sind mögliche **Störgrößen***

- *Stresssituationen*
- *psychische Situationen.*

Die **Faktoren** unterscheidet man weiter in **zuteilbare** und **nicht zuteilbare** Faktoren. Die zuteilbaren Faktoren werden den Beobachtungseinheiten zugeteilt, die nicht zuteilbaren sind fest mit den Beobachtungseinheiten verbunden.

Beispiel 1.7
In klinischen Studien zur Behandlung der AML ist z.B. die **Therapie** ein **zuteilbarer** Faktor und das **Geschlecht** ein **nicht zuteilbarer** Faktor.



1.3.3 Grundgesamtheit, Stichprobe, Beobachtungseinheit

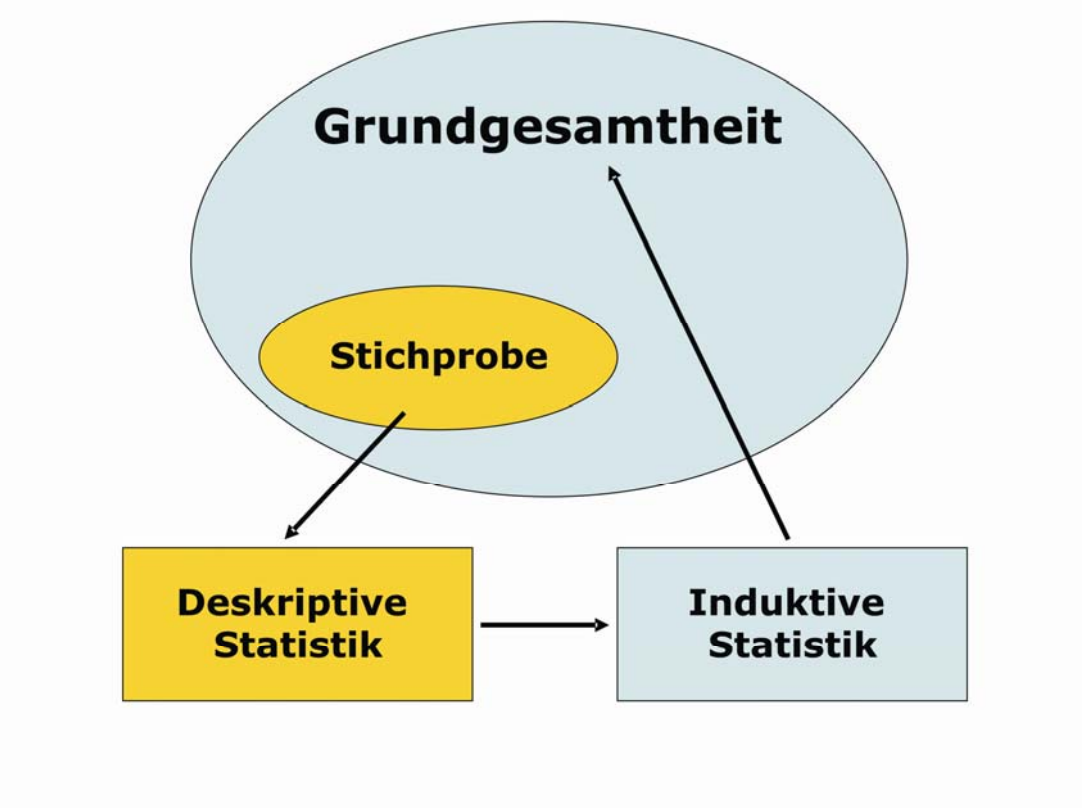
Bei einer statistischen Untersuchung schließt man aus den Daten einer **Stichprobe** zurück auf die zugehörige **Grundgesamtheit**. Die Grundgesamtheit ist die Menge der Objekte oder Individuen, über die man etwas aussagen möchte.

Beispiel 1.8
Eine mögliche Grundgesamtheit sind z.B. alle an AML erkrankten Patienten.

Die Stichprobe ist die Teilmenge der Grundgesamtheit, die man tatsächlich untersucht hat.

Beispiel 1.9
Eine mögliche Stichprobe sind z.B. die in einer klinischen Studie behandelten Patienten mit AML.

Die einzelnen Elemente der Stichprobe sind die in einer Untersuchung tatsächlich beteiligten Beobachtungseinheiten. Da die Stichprobe meist nur ein verschwindend kleiner Teil der Grundgesamtheit ist, ist der Rückschluss von der Stichprobe auf die Grundgesamtheit nicht unproblematisch. Damit er überhaupt sinnvoll ist, muss die Ziehung der Stichprobe so organisiert sein, dass sich in ihr die Verhältnisse der Grundgesamtheit widerspiegeln. Störgrößen bewirken oft eine systematische, nicht zufällige **Verzerrung (Bias)** von Stichproben.



1.3.4 Erhebung, Experiment

Bei statistischen Untersuchungen unterscheidet man die Typen "**retrospektive Erhebung**", "**prospektive Erhebung**" und "**Experiment**".

In einer **retrospektiven** Erhebung wird eine Fragestellung anhand von Daten bearbeitet, die in der Vergangenheit nicht im Hinblick auf diese Fragestellung erhoben wurden. Musterbeispiel für eine retrospektive Erhebung ist die Untersuchung einer Fragestellung anhand von routinemäßig erhobenen Daten.

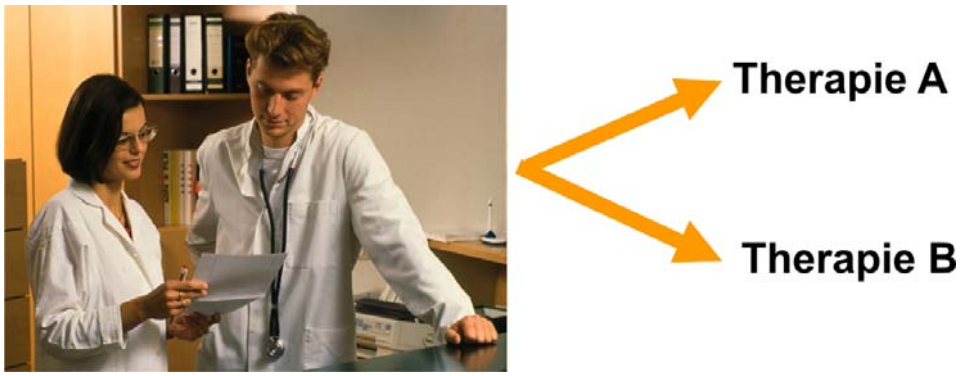
Beispiel 1.10
In einer chirurgischen Klinik werden alle Krankenblätter von Patientinnen, die in den Jahren 1987 bis 1997 an einem Mammakarzinom operiert wurden, herausgesucht. Es wird festgestellt, welche Operationsmethode angewendet wurde, wie lange die einzelne Patientin überlebte und welche Todesursache vorlag.

Bei einer **prospektiven** Erhebung werden die Daten erst nach Vorliegen der Fragestellung an einer zufälligen Stichprobe aus einer definierten Grundgesamtheit im Hinblick auf die Fragestellung neu erhoben.

Beispiel 1.11
In einer prospektiven Erhebung sollen die Nebenwirkungen eines Ovulationshemmers untersucht werden. Die Frauen einer Großstadt werden gebeten, sich für diese Studie zur Verfügung zu stellen. Aus der Menge der Frauen, die sich gemeldet haben (Grundgesamtheit), wird eine zufällige Stichprobe gezogen. In einem festgelegten Zeitraum werden Art und Zeitpunkt der aufgetretenen Nebenwirkungen registriert.

Das **Experiment** erfüllt alle Voraussetzungen der **prospektiven** Erhebung. Zusätzlich wird mindestens eine zuteilbare Einflussgröße den Beobachtungseinheiten vom Versuchsleiter unter statistischen Gesichtspunkten frei zugeteilt.

Beispiel 1.12
In einer klinischen Studie zur Behandlung der AML soll die remissionserhaltende Wirkung zweier Therapien A und B verglichen werden. Zielgröße ist die rezidivfreie Überlebenszeit nach Erreichen der ersten Remission. Sobald sich ein an AML erkrankter Patient bereit erklärt hat, an der Studie teilzunehmen, wird ihm eine der beiden Faktorstufen (Therapie A bzw. B) zufällig zugeteilt.



1.4 Univariate Verfahren

Aufgabe der **deskriptiven Statistik** ist es, die in den Daten der **Stichprobe** enthaltene Information übersichtlich und unverfälscht in **Tabellen, Grafiken und statistischen Maßzahlen** zusammenzufassen. Wie das zu geschehen hat, hängt entscheidend vom Typ des betrachteten **Merkmals** ab.

1.4.1 Tabellarische und grafische Darstellung bei qualitativen Merkmalen

Im Wesentlichen wird die Auswertung **qualitativer** Merkmale hier auf die Darstellung der **absoluten** bzw. der **relativen Häufigkeiten** beschränkt.

Sei A ein qualitatives Merkmal mit den Ausprägungen $A_1,...,A_k$. In einer Stichprobe vom Umfang n habe man die Ausprägung A mit der absoluten Häufigkeit n_i beobachtet ($i=1,2,...,k$). Bei n_0 Beobachtungseinheiten aus der Stichprobe fehle die Angabe zum Merkmal A.

Dann ist offenbar

$$n_1 + n_2 + \ldots + n_k = n - n_0$$

Unter den **relativen Häufigkeiten**, genauer den **adjustierten** relativen Häufigkeiten, versteht man

$$h_i = \frac{n_i}{n - n_0} (i = 1,2,\ldots,k)$$

Der Zusatz "adjustiert" soll betonen, dass man bei der Berechnung nur diejenigen Beobachtungseinheiten berücksichtigt, bei denen die Angaben zum Merkmal tatsächlich vorliegen. Meist werden die relativen Häufigkeiten in **Prozent** angegeben:

$$h_i = h_i \cdot 100\%$$

also z. B. 30 % statt 0.3.

Eine Tabelle der adjustierten relativen Häufigkeiten sollte stets auch den Stichprobenumfang und die Anzahl bzw. den Anteil fehlender Werte enthalten.

Beispiel 1.13
Bei einer Stichprobe von Patienten, die unter Krampfadern im Unterschenkelbereich litten, wurde eine Salbe zur Linderung der Beschwerden angewandt. Eine halbe Stunde nach Auftragen der Salbe wurden die Patienten befragt, ob eine Besserung eingetreten sei. Es ergab sich folgende Urliste:

Tabelle 1.1: Urliste für das Merkmal "Besserung nach Salbenbehandlung"

Patient	Besserung	Patient	Besserung
1	gering	13	gering
2	deutlich	14	gering
3	gering	15	keine
4	deutlich	16	keine Angabe
5	gering	17	gering
6	keine	18	deutlich
7	deutlich	19	deutlich
8	deutlich	20	gering
9	keine Angabe	21	keine Angabe
10	gering	22	gering
11	keine	23	gering
12	keine Angabe	24	deutlich

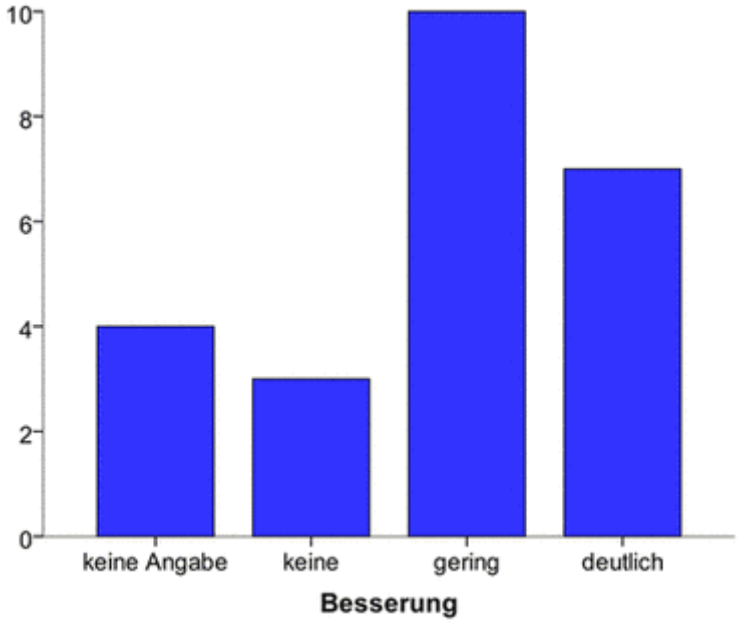
Aus dieser Urliste ergibt sich die folgende Tabelle der absoluten und der relativen Häufigkeiten.

Tabelle 1.2: Häufigkeiten für das Merkmal "Besserung nach Salbenbehandlung"

Besserung	absolute Häufigkeit	relative Häufigkeit	adjustierte relative Häufigkeit
keine	3	12.5 %	15 %
gering	10	41.7 %	50 %
deutlich	7	29.2 %	35 %
keine Angabe	4	16.4 %	
Gesamt:	**24**	**100 %**	**100 %**

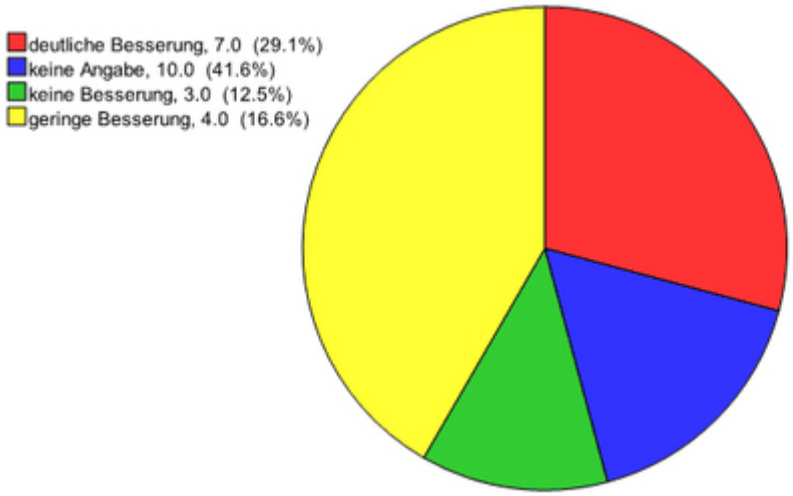
Für die grafische Darstellung gibt es viele Varianten. Am einfachsten ist das Blockdiagramm. Die Merkmalsausprägungen werden an einer Achse in beliebiger (nominales Merkmal) bzw. in der natürlichen Reihenfolge (ordinales Merkmal) angetragen. Darüber wird ein Block gezeichnet, dessen Höhe der absoluten bzw. der relativen Häufigkeit der Ausprägung entspricht. Die Breite der Blöcke ist beliebig, sie soll aber für alle Blöcke gleich sein.

Abbildung 1.1: Blockdiagramm für das Merkmal "*Besserung nach Salbenbehandlung*"



Bei einem Kreisdiagramm entspricht der absoluten bzw. der relativen Häufigkeit der Ausprägung der **zentrale Winkel** des zugeordneten Kreissegments.

Abbildung 1.2: Kreisdiagramm für das Merkmal "*Besserung nach Salbenbehandlung*"



1.4.2 Tabellarische und grafische Darstellung bei quantitativen Merkmalen

Die Darstellung für ein **quantitativ diskretes** Merkmal folgt im Wesentlichen der bei den qualitativen Merkmalen. Zusätzlich werden die **absoluten** und die **relativen** Häufigkeitssummen betrachtet, die definiert sind als

$N_i = \sum_{j=1}^{i} n_j (i = 1,2,\ldots,k)$ absolute Häufigkeitssumme,

$H_i = \sum_{j=1}^{i} h_j (i = 1,2,\ldots,k)$ relative Häufigkeitssumme.

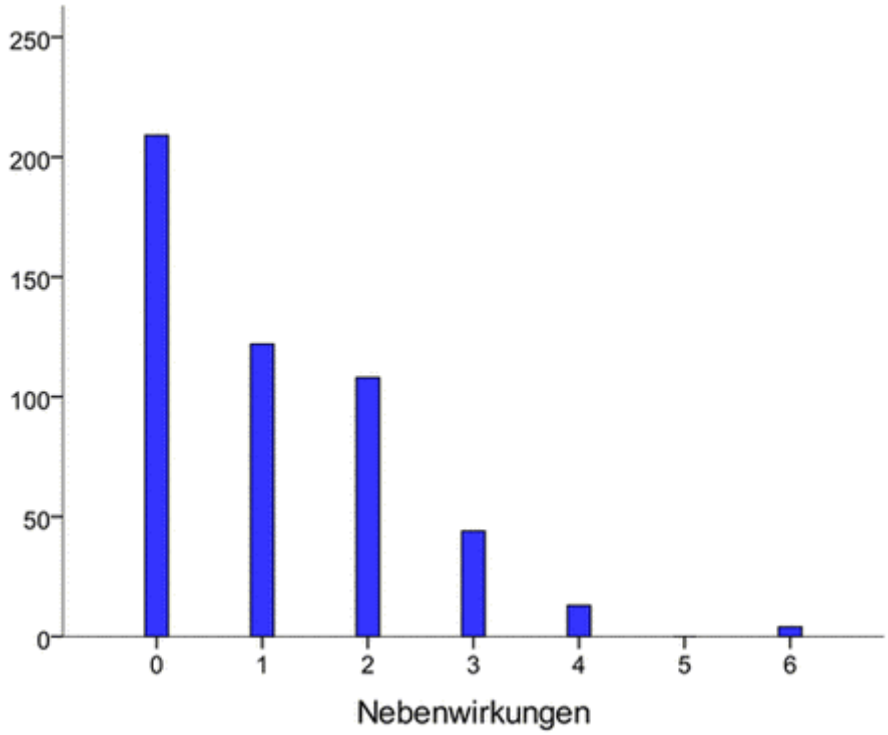
Diese Definition benutzt, dass die Ausprägungen eines quantitativ diskreten Merkmals stets in natürlicher Weise von den kleinen zu den großen Werten geordnet sind. N_i bzw. H_i geben dann Antwort auf die Frage, wie groß die Anzahl bzw. der Anteil der Beobachtungseinheiten mit Ausprägungen kleiner oder gleich der i-ten ist (i=1,2,...,k). Die geeignete grafische Darstellung für die Häufigkeiten bei einem diskreten Merkmal ist das im Wesentlichen dem **Blockdiagramm** entsprechende **Stabdiagramm**.

Beispiel 1.14
In einer Therapiestudie wurden die Häufigkeiten für das diskrete Merkmal "Anzahl der gemeldeten Nebenwirkungen" ermittelt. Tabelle 1.3 enthält das Ergebnis. In Abbildung 1.3 sind die Häufigkeiten als Stabdiagramm dargestellt.

Tabelle 1.3: Häufigkeiten für das quantitativ diskrete Merkmal "Anzahl der Nebenwirkungen"

Anzahl der Nebenwirkungen	absolute Häufigkeit	relative Häufigkeit	absolute Häufigkeitssumme	relative Häufigkeitssumme
0	209	41.8 %	209	41.8 %
1	122	24.4 %	331	66.2 %
2	108	21.6 %	439	87.8 %
3	44	8.8 %	483	96.6 %
4	13	2.6 %	496	99.2 %
5	0	0.0 %	496	99.2 %
6	4	0.8 %	500	100.0 %

Abbildung 1.3: Stabdiagramm für das Merkmal "*Anzahl gemeldeter Nebenwirkungen*"



Wenn man die Häufigkeitsverteilung eines quantitativ **stetigen** Merkmals tabellarisch oder grafisch darstellen will, muss man das Merkmal **klassieren**. Das bedeutet, man zerlegt den gesamten Wertebereich des Merkmals in k Klassen. Die Klassengrenzen bezeichnet man mit $a_0 < a_1 < ... < a_k$.

$a_i - a_{i-1}$ (i=1,2,...,k) ist die Breite der i-ten Klasse, die man normalerweise für alle Klassen gleich groß wählt. Wenn der Wertebereich des Merkmals allerdings nach links bzw. rechts unbegrenzt ist, führt man eine linke bzw. eine rechte Restklasse ein, die nach links bzw. rechts unbegrenzt ist. Die Anzahl k der Klassen sollte nicht zu groß und nicht zu klein sein. Als Faustregel für die Wahl von k gilt

$$k = \begin{cases} \sqrt{n} & n \le 1000 \\ 10 \cdot \lg n & n > 1000 \end{cases}.$$

Die zugehörige grafische Darstellung ist das **Histogramm**. Hier werden die absoluten oder die relativen Häufigkeiten als **Höhe** eines Rechtecks über der gesamten Klasse dargestellt.

Beispiel 1.15
In Tabelle 1.4 liegen die bereits klassierten Altersangaben von 25 Patienten einer klinischen Studie vor.

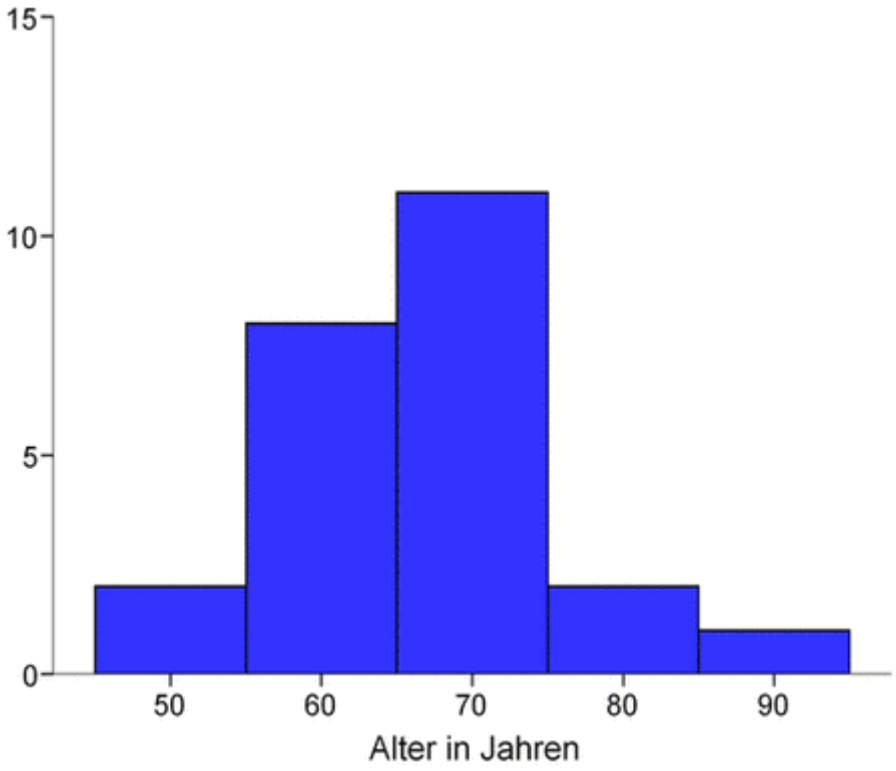
Tabelle 1.4: Häufigkeitsverteilung des klassierten stetigen Merkmals "Alter in Jahren"

Klasse	Alter in Jahren	Klassenmitte	Häufigkeiten absolut	Häufigkeiten relativ
1	(45,55]	50	2	2/24=0.08
2	(55,65]	60	8	8/24=0.33
3	(65,75]	70	11	11/24=0.46
4	(75,85]	80	2	2/24=0.08
5	(85,95]	90	1	1/24=0.04
Summe	-----	------	24	1

Die runde Klammer besagt, dass die entsprechende Klassengrenze selbst nicht zur Klasse gehört, die eckige Klammer zeigt an, dass die entsprechende Klassengrenze dazugehört.

Das zugehörige Histogramm ist in Abbildung 1.4 zu sehen.

Abbildung 1.4: Histogramm für das Merkmal "*Alter in Jahren*"



1.4.3 Statistische Maßzahlen

Liegen Daten x_i zu einem **quantitativen** Merkmal vor, lässt sich die darin enthaltene Information übersichtlich in so genannten statistischen **Maßzahlen** zusammenfassen. Man unterscheidet Lagemaße und Streuungsmaße.

Lagemaße charakterisieren den Durchschnittswert von Daten.

Die bekanntesten Lagemaße sind der arithmetische Mittelwert $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

und der empirische Median $\tilde{x}$.

Zur Berechnung des empirischen Medians müssen die Daten der Größe nach geordnet werden, d. h., man geht von der Urliste der Daten $x_1, x_2, \ldots, x_n$ zur Rangliste $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ über, indem man die Daten der Größe nach ordnet; (*i*) heißt **Rangzahl**. Die Rangzahl gibt den Platz auf der Rangliste an. (*1*) ist die Rangzahl des kleinsten Wertes, (*n*) ist die Rangzahl des größten Wertes. Der empirische Median ist der Wert "in der Mitte" der Rangliste, d.h.

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \qquad \text{falls } n \text{ ungerade,}$$

und

$$\tilde{x} = x_{\left(\frac{n}{2}\right)} \qquad \text{falls } n \text{ gerade.}$$

Oft verwendet man für ein gerades *n* auch die Formel

$$\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right).$$

Beim Vergleich von Mittelwert und empirischem Median stellt man fest, dass man zur Berechnung des Mittelwertes alle Daten x_1, x_2, ..., x_n vollständig kennen muss, während zur Berechnung des empirischen Medians grob gesprochen die erste Hälfte der Rangliste der Daten ausreicht.

Hat man z. B. eine Stichprobe vom Umfang n=3 und kennt x_1=2, x_2=4 und von x_3 weiß man nur, dass es größer ist als x_2=4, dann kann man den Mittelwert nicht angeben, aber für den empirischen Median gilt $\tilde{x} = 4$,

gleichgültig wie groß x_3 ausfällt.

Aus dieser Beobachtung folgt, dass der empirische Median **robust** ist gegenüber **Ausreißern**. Dieser Sachverhalt wird bei der Auswertung von **Überlebenszeiten** noch eine Rolle spielen.

Man kann den empirischen Median auch mit Hilfe der **empirischen Verteilungsfunktion** F_n

definieren.

F_n gibt für jedes x auf der Zahlengeraden an, wie groß der Anteil der Daten ist, die kleiner oder gleich x sind. Für $x_{(1)}$, den kleinsten Wert, gilt

$$F_n\left(x_{(1)}\right) = \frac{1}{n}$$

- aber nur, falls alle Daten voneinander verschieden sind - für $x_{(n)}$, den größten Wert, gilt

$$F_n\left(x_{(n)}\right) = \frac{n}{n} = 1$$

Damit ist der empirische Median $\tilde{x}$ der kleinste Wert, für den gilt

$$F_n(\tilde{x}) \geq 0.5\,.$$

Entsprechend definiert man mit Hilfe der empirischen Verteilungsfunktion als weitere Lagemaße die so genannten empirischen **Quantile** $x_p\,(0 < p < 1)$: x_p ist der kleinste Wert, für den gilt: $F_n(x_p) \geq p$. Insbesondere werden $x_{0.25}$ und $x_{0.75}$ betrachtet. $x_{0.25}$ heißt **1. Quartil**, $x_{0.75}$ **3. Quartil**. In dieser Terminologie ist der empirische Median das 2. Quartil

$$\tilde{x} = x_{0.5}$$

In Analogie zur Berechnung des Medians werden die Quartile oft folgendermaßen berechnet. Die gesamte Rangliste wird in zwei Hälften geteilt. Das erste Quartil ist der Wert "in der Mitte" der ersten Hälfte der Rangliste, d. h. die Hälfte der halbierten Messreihe ist kleiner bzw. größer als das erste Quartil. Analog wird das 3. Quartil berechnet. Das oben beim Median erwähnte Verfahren, bei geraden Anzahlen den Mittelwert der Rangwerte (n/2) und (n/2+1) zu verwenden, wird oft auch bei der Quartilsberechnung benutzt.

Streuungsmaße sind Maßzahlen für die Abweichung der Messwerte vom Durchschnittswert. Die bekanntesten Streuungsmaße sind die empirische **Varianz** s^2

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

und die empirische **Standardabweichung**

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

Wegen des Quadrierens lässt sich die empirische Varianz als Zahlwert anschaulich kaum interpretierten, während sich die empirische Standardabweichung grob als mittlere Abweichung der Daten von ihrem Mittelwert deuten lässt.

Als weitere Streuungsmaße betrachtet man die empirische **Spannweite** R (engl.: range):

$$R = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}$$

und den empirischen **Interquartilsabstand**

$$q = x_{0.75} - x_{0.25}$$

Die empirische Spannweite ist offenbar extrem ausreißerempfindlich, der empirische Interquartils-abstand ist ein stabileres Streuungsmaß.

Mit dem Adjektiv "**empirisch**" bei den Maßzahlen, soll betont werden, dass sich diese Maßzahlen tatsächlich aus der Stichprobe berechnen lassen. Später sollen sie den analogen Maßzahlen der Grundgesamtheit gegenübergestellt werden, die sich zumeist nicht berechnen lassen. Vielmehr werden die empirischen Maßzahlen der Stichprobe als Schätzwerte für die theoretischen Maßzahlen der Grundgesamtheit dienen. Wenn aus dem Zusammenhang ersichtlich ist, ob die Grundgesamtheit oder die Stichprobe gemeint ist, soll in Zukunft das Adjektiv empirisch entfallen.

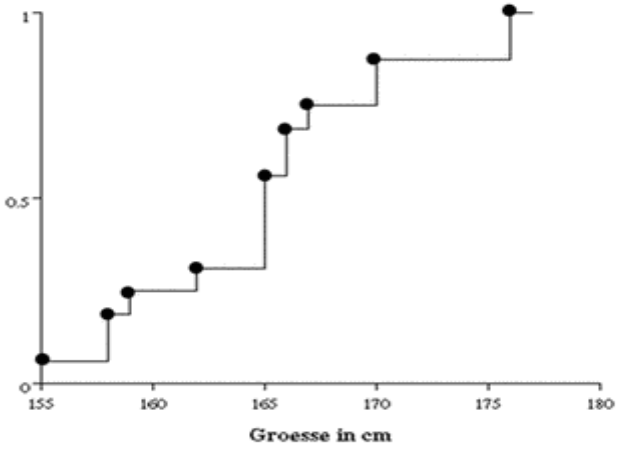
Beispiel 1.16
In Tabelle 1.5 finden Sie für 16 weibliche Patienten einer klinischen Studie die Angaben zur Körpergröße in cm. Die Daten liegen in Form einer Rangliste vor, d. h. sie sind bereits aufsteigend sortiert.

Tabelle 1.5: Körpergröße von 16 Patienten

Lfd. Nr. *i*	**Größe** x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	**empirische Verteilungsfunktion** $F_{16}(x_i)$
1	155	-10.1875	103.785	1/16 = 0.0625
2	158	-7.1875	51.660	
3	158	-7.1875	51.660	3/16 = 0.1875
4	159	-6.1875	38.285	4/16 = 0.2500
5	162	-3.1875	10.160	5/16 = 0.3125
6	165	-0.1875	0.035	
7	165	-0.1875	0.035	
8	165	-0.1875	0.035	
9	165	-0.1875	0.035	9/16 = 0.5625
10	166	0.8125	0.660	
11	166	0.8125	0.660	11/16 = 0.6875
12	167	1.8125	3.285	12/16 = 0.7500
13	170	4.8125	23.160	
14	170	4.8125	22.160	14/16 = 0.8750
15	176	10.8125	116.910	
16	176	10.8125	116.910	16/16 = 1
Summe	2643	0.0000	540.437	

Aus den Werten der Tabelle 1.5 erhält man die folgende grafische Darstellung der empirischen Verteilungsfunktion.

Abbildung 1.5: Empirische Verteilungsfunktion für das Merkmal "*Größe*"



Aus den Werten der Tabelle 1.5 erhält man die folgenden **Lagemaße** und **Streuungsmaße**.

Lagemaße

empirisches Minimum	$x_{\min} = 155$
empirisches 0.25-Quantil (1. Quartil)	$x_{0.25} = 159$
alternativ	$0.5 \cdot (x_{(4)} + x_{(5)}) = 160.5$
empirischer Median (2. Quartil)	$\tilde{x} = 165$
alternativ	$x_{0.5} = 165$
empirisches 0.75-Quantil (3. Quartil)	$x_{0.75} = 167$
alternativ	$0.5 \cdot (x_{(12)} + x_{(13)}) = 168.5$
empirisches Maximum	$x_{\max} = 176$
Mittelwert	$\bar{x} = 165.1875$

Streuungsmaße

empirische Spannweite (Range)	$R = x_{\max} - x_{\min} = 21$
empirischer Interquartilsabstand	$q = x_{0.75} - x_{0.25} = 8$
empirische Varianz	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 36.0291$
empirische Standardabweichung	$s = \sqrt{s^2} = 6.0024$.

1.4.4 Boxplot

Der **Boxplot** ist eine grafische Darstellung, mit der man sich einen guten Überblick über die Verteilung der Daten einer Stichprobe verschaffen kann.

In einem Koordinatensystem, an dessen *y*-Achse eine Skala für das betrachtete Merkmal angetragen ist, wird der **Interquartilsabstand** q_i als Kasten (engl.: box) eingezeichnet. Vom oberen Ende des Kastens wird eine Strecke bis zum maximalen Wert gezeichnet, die aber nicht länger als das 1.5-fache des Interquartilsabstandes gezogen wird. Falls es Werte gibt, die mehr als $1{,}5 \cdot q$ vom oberen Ende entfernt sind, so werden diese einzeln als Punkte eingetragen.

Entsprechend verfährt man am unteren Ende des Kastens mit dem **minimalen** Wert. Zusätzlich wird die Position des empirischen **Medians** und manchmal auch die des **Mittelwertes** markiert. In dieser Definition umfasst der Kasten gerade die mittleren 50% der Daten. Es gibt auch andere Varianten des Boxplots. Wenn man statistische Software benutzt, muss man prüfen, ob diese oder eine andere Definition benutzt wird.

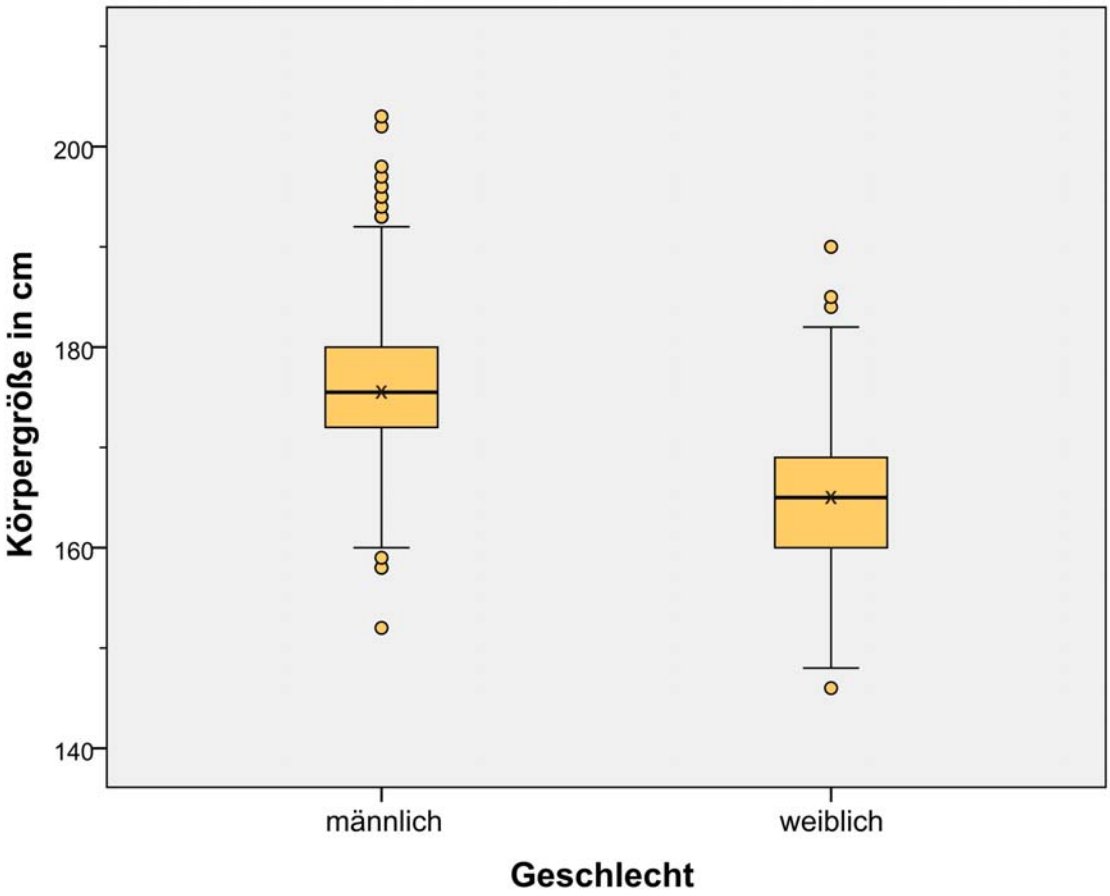
Die Darstellung als Boxplot ist sehr nützlich, insbesondere dann, wenn man Untergruppen der Stichprobe vergleichen will.

Beispiel 1.17
Tabelle 1.6 enthält die statistischen Angaben für alle 140 Patienten (59 männliche, 81 weibliche) einer Studie.

Größe nach Maßzahl	Geschlecht: Männlich	Geschlecht: Weiblich
N	57	76
fehlend	2	5
$x_{\min}$	160.00	148.00
$x_{0.25}$	171.00	160.00
$\tilde{x}$	176.00	165.00
$x_{0.75}$	180.00	170.00
$x_{\max}$	190.00	180.00
$\bar{x}$	176.12	164.87
Spannweite R	30.00	32.00
empirischer Interquartilsabstand *q*	9.00	10.00
empirische Varianz s^2	49.72	44.76
empirische Standardabweichung s	7.05	6.69

Aus den Angaben in Tabelle 1.6 erhält man in Abbildung 1.6 einen nach Geschlecht getrennten Boxplot für die Körpergröße. Hierbei wird die Lage des arithmetischen Mittels durch einen Punkt dargestellt.

Abbildung 1.6: Boxplot für das Merkmal "*Körpergröße*"



Die Quantile werden in der Medizin häufig angewandt z.B. zur Festlegung von **Normbereichen**. Abbildung 1.7 zeigt ein Diagramm, in das bei den Säuglingsvorsorgeuntersuchungen die Körpergröße in Abhängigkeit vom Alter eingetragen wird. Dargestellt sind das 0.03-Quantil, der Median und das 0.97-Quantil der Körpergröße in Abhängigkeit vom Alter des Kindes. Im Bereich zwischen den Quantilen befinden sich etwa 94% der Grundgesamtheit. Kinder, deren Körpergröße dauerhaft nicht in diesen Bereich fällt, gelten als auffällig groß bzw. auffällig klein. Bei ihnen wird untersucht, ob eine Entwicklungsstörung vorliegt.

Abbildung 1.7: Somatogramm