

2. EINFÜHRUNG IN R UND BESCHREIBENDE STATISTIK

2.1. EINFÜHRUNG IN R

Kommt noch nach dem Semester

2.2. BESCHREIBENDE STATISTIK

2.2.1. MERKMALE UND SKALEN

(2.2.1.1) Überblick

Bei Untersuchungen in den Natur-, Geistes- und Sozialwissenschaften geht es allgemein formuliert stets darum, gewisse *Merkmale* der zu untersuchenden Objekte zu beschreiben. In diesem Kapitel werden diese Merkmale genauer klassifiziert; so unterscheidet man etwa *qualitative* und *quantitative* Merkmale, letztere werden weiter in *stetige* und *diskrete* Merkmale unterteilt. (2.2.3)

Für die rechnerische Behandlung eignen sich selbstverständlich die durch Zahlen beschreibbaren Merkmale. Diesen Zahlen kann aber ein mehr oder weniger grosser Informationsgehalt innewohnen, was einen Einfluss darauf hat, welche Rechenoperationen mit diesen Zahlen sinnvoll sind und welche nicht. Diese Unterschiede drücken sich in den so genannten *Skalen* aus: (2.2.4)

- Nominalskala,
- Ordinalskala,
- Intervallskala,
- Verhältnisskala.

Das Ziel dieses Kapitels besteht vor allem darin, Sie darauf aufmerksam zu machen, dass beim Umgang mit Daten auf deren Natur Rücksicht zu nehmen ist. (2.2.5)

(2.2.1.2) Allgemeines zur beschreibenden Statistik

Eine der Aufgaben der *beschreibenden Statistik* ist es, Ergebnisse von Beobachtungen und Versuchen auf übersichtliche Weise darzustellen. Diesem Thema wird Teil 2.2.2 gewidmet sein. Eine weitere Aufgabe besteht aber darin, die gewonnenen Daten auf

prägnante Art und Weise durch einige wenige zusammenfassende Zahlen zu charakterisieren. Dazu dienen die so genannten statistischen Masszahlen, auf die in Kapitel 2.2.3 eingegangen wird, und von denen Sie als wichtiges Beispiel jedenfalls den Durchschnitt kennen. Im vorliegenden Kapitel 2.2.1 werden zunächst einige grundlegende Tatsachen im Zusammenhang mit der zahlenmässigen Auswertung von Versuchen und Beobachtungen besprochen.

Neben der beschreibenden gibt es noch die *beurteilende Statistik*, die es erlaubt, das Beobachtungsmaterial auszuwerten und daraus weitergehende Schlüsse zu ziehen. So kann man etwa versuchen, aus einer Umfrage in einem beschränkten Personenkreis das Ergebnis einer Abstimmung vorherzusagen. Die beurteilende Statistik beruht auf der Wahrscheinlichkeitsrechnung und wird ab Kapitel 8 zur Sprache kommen.

(2.2.1.3) Untersuchungsobjekte und Merkmale

Das Ziel der folgenden Betrachtungen ist es, einige wichtige Begriffe zu klären und zu benennen. Bei naturwissenschaftlichen Untersuchungen (und auch bei solchen auf anderen Gebieten) geht es sehr oft darum, einem *Untersuchungsobjekt* ein bestimmtes *Merkmal* (oder mehrere Merkmale gleichzeitig) zuzuordnen. Die nachstehenden Beispiele sollen diese etwas allgemeine Terminologie erläutern.

Untersuchungsobjekt	Merkmal
Ortschaft	Einwohnerzahl
Mensch	Körpergrösse
Mensch	Alter
Mensch	Augenfarbe
Mensch	Geschlecht
Mensch	Blutgruppe (0, A, B oder AB)
Flüssigkeit	Siedepunkt
Batterie	Spannung
Blatt einer Pflanze	Form (lineal, lanzettlich etc.)

(2.2.1.4) Qualitative und quantitative Merkmale

Wir unterscheiden:

Quantitative Merkmale	Qualitative Merkmale
Können durch Messen oder Zählen erfasst werden.	Können nicht durch Messen oder Zählen erfasst werden.
Beispiele: Einwohnerzahl Körpergrösse Alter Siedepunkt Spannung	Beispiele: Augenfarbe Geschlecht Blutgruppe Blattform

Der Unterschied zwischen quantitativen und qualitativen Merkmalen kann von der Art der Beobachtung abhängen. Ein an sich quantitatives Merkmal kann auch qualitativ beschrieben werden. Schliesslich spricht man ja etwa von grossen oder kleinen Äpfeln, obwohl man die Angabe auch quantitativ (durch Gewicht oder Umfang) machen könnte.

Für die mathematische Behandlung kommen natürlich in erster Linie die quantitativen Merkmale in Frage. Diese unterteilen wir in (2.2.1.5) weiter.

(2.2.1.5) Diskrete und stetige Merkmale

Ein quantitatives Merkmal heisst *stetig*, wenn es von seiner Natur her jeden Wert, also im Prinzip jede reelle Zahl (zumindest innerhalb bestimmter Grenzen), annehmen kann. Insbesondere sind wenigstens theoretisch unendlich viele Messwerte möglich. Stetige Merkmale werden in der Regel durch *Messen* bestimmt. Beispiele dafür sind etwa Körpergrösse, Alter, Siedepunkt oder Spannung.

Wir haben dabei insofern idealisiert, als wir unbeschränkte Messgenauigkeit vorausgesetzt haben, die in der Praxis ja nie erreicht werden kann (daher der Einschub “im Prinzip”). Für die Anwendung mathematischer Verfahren ist diese Idealisierung meist sehr zweckmässig.

Ein quantitatives Merkmal heisst *diskret*, wenn es nur endlich viele (oder höchstens “abzählbar unendlich viele”) Werte annehmen kann.

Der Begriff “abzählbar unendlich” wird für uns vor allem in der Wahrscheinlichkeitsrechnung von Bedeutung sein und zwar im Zusammenhang mit den so genannten diskreten Zufallsgrössen. Er wird aber schon jetzt erwähnt, obwohl er hier eher von theoretischem Interesse ist. Eine unendliche Menge heisst *abzählbar*, wenn man ihre Elemente in eine Folge (a_0, a_1, a_2, \dots) anordnen kann. So sind etwa die natürlichen

Zahlen \mathbb{N} , aber auch die ganzen Zahlen \mathbb{Z} abzählbar. Eine mögliche Anordnung der ganzen Zahlen in eine Folge ist gegeben durch

$$0, 1, -1, 2, -2, 3, -3, \dots$$

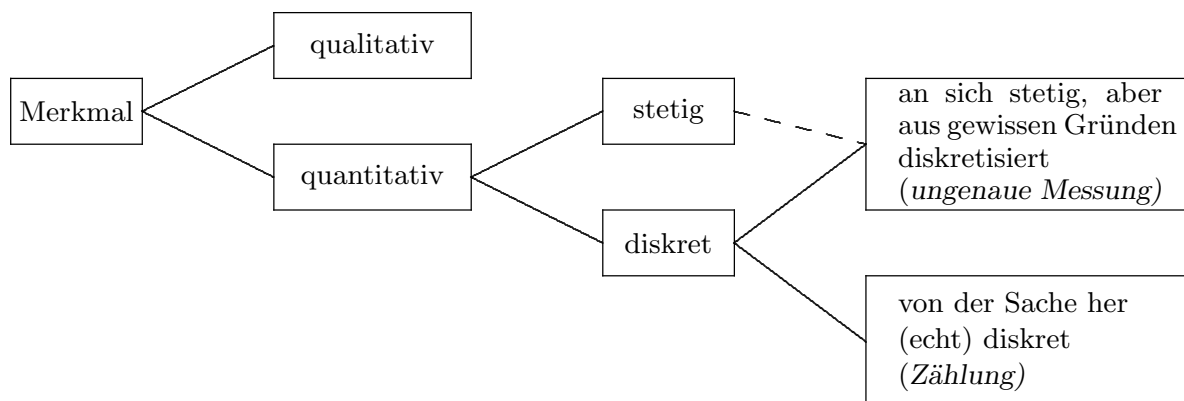
Man kann ferner beweisen, dass die Menge \mathbb{R} der reellen Zahlen nicht abzählbar (“überabzählbar”) ist.

Die wichtigsten Beispiele von diskreten Merkmalen sind jene, die durch *Zählen* ermittelt werden, wie etwa die Einwohnerzahl einer Ortschaft, die Zahl der Blütenblätter einer Blume etc. Das Resultat der Zählung ist offensichtlich eine natürliche Zahl. Solche Zahlen werden oft *Häufigkeiten* genannt.

Daneben treten aber diskrete Merkmale auch im Zusammenhang mit stetigen Merkmalen auf. Die Werte eines stetigen Merkmals werden nämlich meist in Klassen zusammengefasst (Lebensalter in Jahren, Körpergrösse in cm, etc.). Dies ist allein schon aus Gründen der praktisch beschränkten Messgenauigkeit notwendig. Ein im Prinzip stetiges Merkmal wird auf diese Weise *diskretisiert*, und die Messung wird im Grunde durch eine Zählung ersetzt.

Misst man etwa beim Hundertmeterlauf die Zeit in Hundertstelsekunden, so läuft die Ablesung der Stoppuhr im Prinzip auf eine Zählung von Hundertstelsekunden heraus. Die möglichen Resultate werden dann bequemlichkeitshalber in Sekunden ausgedrückt (9.98, 10.04 etc.). Sie sind also keine natürlichen Zahlen, aber trotzdem diskrete Messwerte. (Würde man alles in Hundertstelsekunden ausdrücken, so erhielte man ja natürliche Zahlen!)

Die Überlegungen von (2.2.1.4) und (2.2.1.5) lassen sich im folgenden Schema zusammenfassen:



(2.2.1.6) Beschreibung von qualitativen Merkmalen durch Zahlen

Auch qualitative Merkmale werden oft unter Zuhilfenahme von Zahlen beschrieben; sie werden aber dadurch nicht etwa zu quantitativen Merkmalen. Diese Zuordnung von Zahlen kann auf zwei Stufen geschehen:

a) Nominale Merkmale

Die Zahlen dienen bloss zur Codifizierung oder zur einfacheren Verarbeitung der Daten:

- Bei der computerisierten Auswertung einer Umfrage ist es praktisch, das Merkmal “Geschlecht” durch 0 (= männlich) bzw. 1 (= weiblich) auszudrücken.
- Jeder erwachsene Einwohner und jede erwachsene Einwohnerin der Schweiz ist durch die AHV-Nummer eindeutig gekennzeichnet.
- Bei einer Untersuchung über Berufe mag es angezeigt sein, die einzelnen Berufe zu nummerieren: Architektin = 1, Bäcker = 2, Chemiker = 3, Drogistin = 4, ...

Da die Zahlen hier eigentlich nur andere Namen für die untersuchten Ausprägungen eines Merkmals sind, spricht man auch von *nominalen Merkmalen*. Charakteristisch ist, dass die Zahlen hier durch andere Symbole ersetzt werden können, ohne dass sich der Informationsgehalt ändert.

Es ist deshalb offensichtlich sinnlos, mit diesen Zahlen irgendwelche Rechnungen durchführen zu wollen. Man könnte zwar z.B. die “durchschnittliche AHV-Nummer” aller Zürcherinnen und Zürcher berechnen, das Ergebnis dieser Mühe wäre aber von der Sache her völlig absurd.

b) Ordinale Merkmale

Die Zahlen dienen dazu, eine Reihenfolge (Rangordnung) festzulegen:

- In einer gewöhnlichen Rangliste bedeutet die Aussage “ $1 < 2$ ”, dass der Teilnehmer mit Rang 1 besser ist als jener mit Rang 2.
- Eine Ärztin unterscheidet bei einer gewissen Krankheit vier Stufen: 0 = Krankheit nicht vorhanden, 1 = leicht, 2 = mittel, 3 = schwer. Auch hier hat die Aussage “ $1 < 3$ ” eine konkrete Bedeutung.
- In der Mineralogie kennt man die Härteskala für Mineralien (benannt nach MOHS). Ein Mineral mit Härte 7 (z.B. Quarz) ritzt dabei eines mit Härte 6 (z.B. Feldspat). Die Relation $>$ bedeutet also “härter”.

In derartigen Fällen, wo also die Ordnungsrelation $<$ der Zahlen eine konkrete Bedeutung in Bezug auf das untersuchte Merkmal hat, spricht man von einem *ordinalen Merkmal*. Bezeichnend ist, dass man bei solchen Merkmalen die Zahlen durch irgendwelche Zeichen ersetzen könnte, sofern zwischen diesen eine klar ersichtliche Reihenfolge besteht, z.B. durch das Alphabet A, B, C, ... oder durch “Jass-Striche” |, ||, |||, ...

Auch hier ist die Durchführung von Rechenoperationen mit diesen Zahlen nicht sinnvoll. So ist zwar etwa im Beispiel der Krankheit $1 - 0 = 3 - 2$; diese arithmetische Tatsache hat aber keine praktische Bedeutung, denn man kann nicht sagen, der Unterschied zwischen nicht vorhandener Krankheit und einem leichten Fall sei derselbe wie zwischen einem mittleren und einem schweren Fall. Auch die folgende Aussage ist

unsinnig: Ein gesunder und ein mittelschwer erkrankter Patient sind im Durchschnitt leicht krank.

(2.2.1.7) Skalen

Wenn man irgendwelchen quantitativen oder qualitativen Merkmalen Zahlen zuordnet, so sagt man, man habe eine *Skala* eingeführt. Man ist dann natürlich versucht, mit diesen Zahlen auch zu rechnen. Wie wir aber eben gesehen haben, ist nicht jede an sich mögliche Rechnung auch wirklich sinnvoll. Es kommt darauf an, was durch die Zahlen beschrieben wird.

Um Klarheit darüber zu erhalten, welche Rechenoperationen bei einem bestimmten Merkmal sinnvoll sind, unterscheidet man vier verschiedenen Skalen, wobei a) die “schwächste”, d) die “stärkste” ist:

- a) Nominalskala,
- b) Ordinalskala,
- c) Intervallskala,
- d) Verhältnisskala.

Je höher das “Niveau” der Skala ist, desto mehr Rechenoperationen sind erlaubt. Wir besprechen nun diese vier Skalen im Einzelnen.

a) Nominalskala

Hier handelt es sich einfach um die Darstellung eines nominalen Merkmals (2.2.1.6.a) durch Zahlen (vgl. die dortigen Beispiele). Diese Zahlen haben nur die Bedeutung von Namen, mehr darf nicht hineingelesen werden. Jede Rechenoperation ist, wie bereits erwähnt, sinnlos. Auch die Grössenbeziehung der Zahlen hat keine Bedeutung. Im Beispiel der Berufsbezeichnung etwa soll die Zuordnung “Architektin = 1”, “Bäcker = 2” keine Wertung oder Rangordnung ausdrücken.

b) Ordinalskala

Hier drücken die Zahlen eine Rangordnung aus, sie charakterisieren ein ordinales Merkmal (2.2.1.6.b), vgl. die dortigen Beispiele. Die Beziehung $<$ beschreibt eine gewisse wertende Eigenschaft, wie z.B. “grösser”, “schneller”, “kleiner”, “schlechter”, “intensiver” usw.

Zahlenmässig gleiche Unterschiede auf der Ordinalskala müssen aber nicht gleichen Unterschieden der Merkmale entsprechen, denn die Unterschiede der Merkmale brauchen gar nicht vergleichbar zu sein: Man kann z.B. nicht sagen, der Härteunterschied zwischen Gips (Härte 2) und Talk (Härte 1) sei grösser oder kleiner als jener zwischen Diamant (10) und Korund (9). Bei einer Ordinalskala haben die Abstände zwischen zwei Werten auf der Skala somit im Allgemeinen keine Bedeutung.

c) Intervallskala

Wenn eine Ordinalskala vorliegt, bei der auch die Abstände zwischen zwei Messwerten (anders gesagt die Intervalle) eine Bedeutung haben, dann spricht man von einer Intervallskala. Intervallskalen sind nur für quantitative Merkmale möglich, bedingen also einen Mess- oder Zählprozess.

Beispiele

- Temperatur in °C:

Hier liegt zunächst sicher eine Ordinalskala (für ein stetiges Merkmal!) vor: 0° ist "wärmer" als -5° , 10.5° ist wärmer als 10.4° etc. Wir haben aber sogar eine Intervallskala, denn Intervalle gleicher Länge, etwa zwischen 0° und 1° bzw. zwischen 11° und 12° , haben dieselbe physikalische Bedeutung. Beachten Sie aber, dass der Nullpunkt der Skala willkürlich (als Schmelzpunkt von Eis) festgelegt wurde. In der Fahrenheit-Skala z.B. liegt der Nullpunkt anderswo (0°C entspricht 32°F , 100°C entspricht 212°F).

- Höhe von Bergen:

Wir haben wiederum eine Ordinalskala; die Ordnungsrelation $<$ entspricht dem Begriff "niedriger". Aber auch die Differenz von Zahlen hat eine klar festgelegte Bedeutung: Der Höhenunterschied zwischen Uetliberg (871) und Pfannenstil (853) ist derselbe wie zwischen Monte Brè (930) und San Salvatore (912). Der Nullpunkt aber ist willkürlich festgelegt. Man spricht zwar von der Höhe über Meer, die Angaben beziehen sich jedoch auf einen Bezugspunkt bei Genf (Repère Pierre du Niton). Die Höhe des letztern ist übrigens früher einmal um etwa 3 m nach unten korrigiert worden, was die Relativität der Sache aufzeigt.

In allen diesen Beispielen ist der Nullpunkt der Skala willkürlich festgelegt worden. Das Fehlen eines natürlichen "absoluten Nullpunkts" bewirkt, dass das Bilden von Verhältnissen im Allgemeinen nicht sinnvoll ist:

- Zwei Flugzeuge mögen in Höhen von 3000 bzw. 5000 Meter über Meer fliegen. Wenn ich auf einem Berg von der Höhe 1000 m.ü.M. bin, dann kann ich zwar sagen, das eine Flugzeug sei doppelt so hoch wie das andere. Würde ich auf einem Berg von 2000 m.ü.M. stehen, so wäre diese Aussage falsch. Die Höhendifferenz (2000 m) aber hat einen vom Standort des Beobachters unabhängigen Sinn.
- Ebenso wenig ist die Verwendung von Prozentzahlen möglich. Die Aussage "Der Uetliberg ist um 18 m höher als der Pfannenstil" ist absolut sinnvoll, nicht aber die Aussage "Der Uetliberg ist um 2.1% höher als der Pfannenstil", da dies vom gewählten Ausgangspunkt der Höhenmessung abhängt (dessen Höhenangabe nicht absolut ist und, wie oben erwähnt, tatsächlich einmal geändert wurde).

Derartige Probleme treten nun aber nicht auf, wenn ein natürlicher Nullpunkt vorhanden ist, nämlich bei einer so genannten Verhältnisskala.

d) Verhältnisskala

Wenn eine Intervallskala vorliegt, bei der zusätzlich der *Nullpunkt* der Skala *in natürlicher Weise* gegeben ist, dann hat auch die Bildung von Verhältnissen (und von Prozentzahlen) einen Sinn. Man spricht von einer Verhältnisskala.

Beispiele

- Gewicht: Das Gewicht 0 ist ein natürlicher Nullpunkt. Es ist sinnvoll, zu sagen, ein Mensch von 140 kg wiege doppelt so viel wie einer von 70 kg.
- Entsprechendes gilt für Länge, Flächeninhalt, Volumen, elektrische Stromstärke etc.
- Auch die Temperatur in K (Kelvin) ist eine Verhältnisskala, da der absolute Nullpunkt (-273.15°C) ein naturgegebener Nullpunkt der Skala ist.
- Durch Zählungen ermittelte *Häufigkeiten* entsprechen normalerweise einer Verhältnisskala. (Die Häufigkeit 0 ist ein natürlicher Nullpunkt.)

e) Zusammenfassung

Mit Zahlen, die Merkmalen von Untersuchungsobjekten entsprechen, darf nicht unbesehen gerechnet werden. Man muss sich jeweils überlegen, zu welcher der folgenden vier Skalen diese Zahlen gehören.

Nominalskala	Keine Rangordnung. Rechenoperationen sinnlos.
Ordinalskala	Rangordnung festgelegt. Rechenoperationen sinnlos.
Intervallskala	Rangordnung festgelegt, Abstände zwischen Zahlen haben Bedeutung, Nullpunkt willkürlich. Bildung von Verhältnissen (und Prozenten) nicht sinnvoll.
Verhältnisskala	Rangordnung festgelegt, Abstände zwischen Zahlen haben Bedeutung, Nullpunkt absolut. Bildung von Verhältnissen (und Prozenten) sinnvoll.

Für die mathematische Behandlung interessieren hauptsächlich die Intervall- und Verhältnisskalen, gelegentlich auch die Ordinalskalen.

In Einzelfällen kann die Zuordnung eines Merkmals zu einer bestimmten Skala diskutabel sein.

- a) Zu welcher Skala gehören Schulnoten? Sicher bilden sie eine Ordinalskala. Ob sie auch eine Intervallskala bilden, hängt von der Situation ab. So wird z.B. eine Stilnote in einem Aufsatz zu einer Ordinalskala gehören, die Orthographienote dagegen zu einer Intervallskala, sofern sie sich direkt auf die Anzahl der gemachten Fehler bezieht.

- b) Wie sind Startnummern bei einem Wettkampf einzureihen? Auf jeden Fall handelt es sich um nominale Merkmale, da die Startnummern die Teilnehmerinnen bezeichnen. Bei vielen Wettkämpfen, z.B. bei einer Ski-Abfahrt, geben sie auch die Startreihenfolge an. In diesem Fall liegt eine Ordinalskala vor.

Die Zuordnung zu einer bestimmten Skala hängt also auch davon ab, in welchem Licht man die Merkmale betrachtet.

2.2.2. DARSTELLUNG VON VERSUCHSERGEBNISSEN

(2.2.2.1) Überblick

Bei vielen Untersuchungen fallen Daten in grosser Zahl an. Um diese in übersichtlicher Form darstellen zu können, verwendet man verschiedene Verfahren, von denen einige in diesem Kapitel vorgestellt werden. Dies wird anhand je eines Beispiels für ein diskretes und für ein stetiges Merkmal durchgeführt.

Die beiden wichtigsten Begriffe sind die *Häufigkeitsverteilung* (dargestellt mit Stab- bzw. Balkendiagrammen) und die *Summenhäufigkeitsverteilung*.

(2.2.2.2), (2.2.2.3)

(2.2.2.6), (2.2.2.7)

(2.2.2.2) Häufigkeitsverteilung bei diskreten Merkmalen

Wir beginnen mit einem Beispiel:

Beispiel 2.2.2.2.A

Eine Gruppe von 40 Studierenden erzielte an einer Prüfung die folgenden Noten. Da diese einer alphabetischen Liste entstammen, sind sie noch ungeordnet.

4	$4\frac{1}{2}$	5	3	$5\frac{1}{2}$	$4\frac{1}{2}$	$3\frac{1}{2}$	$5\frac{1}{2}$	3	4
$4\frac{1}{2}$	1	5	4	$5\frac{1}{2}$	$3\frac{1}{2}$	6	$4\frac{1}{2}$	5	$4\frac{1}{2}$
6	4	3	$4\frac{1}{2}$	5	3	6	4	$2\frac{1}{2}$	$4\frac{1}{2}$
$4\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$	$5\frac{1}{2}$	4	$4\frac{1}{2}$	3	5	3	4

☒

Diese Angaben nennt man die *Urliste* oder die *Rohdaten*. Die Anzahl der untersuchten Objekte bezeichnen wir mit n (hier ist $n = 40$). Um sich etwas mehr Übersicht zu verschaffen, verfertigt man zweckmässigerweise eine so genannte *Strichliste*:

1		1
$1\frac{1}{2}$		0
2		0
$2\frac{1}{2}$		1
3		6
$3\frac{1}{2}$		4
4		7
$4\frac{1}{2}$		9
5		5
$5\frac{1}{2}$		4
6		3

Die erhaltenen Anzahlen nennt man die *absoluten Häufigkeiten* des betrachteten Merkmalswerts. Wir formulieren den Sachverhalt noch mit allgemeinen Grössen:

Die möglichen Werte des Merkmals seien

$$w_1, w_2, \dots, w_k .$$

(Im Beispiel ist $w_1 = 1, w_2 = 1\frac{1}{2}, \dots, w_{11} = 6$, somit $k = 11$.) Die absolute Häufigkeit des Merkmalswerts w_i bezeichnen wir mit H_i (hier ist also $H_1 = 1, \dots, H_{11} = 3$). Natürlich ist dann die Summe aller absoluten Häufigkeiten gleich n : $\sum_{i=1}^k H_i = n$.

Neben den absoluten Häufigkeiten H_i betrachtet man auch die *relativen Häufigkeiten* h_i , da diese oft übersichtlicher sind. Eine solche relative Häufigkeit kommt in zwei Varianten vor: Als Zahl zwischen 0 und 1 oder als Prozentzahl zwischen 0% und 100%. Für die Theorie ist die erste Variante zweckmässiger, in der Praxis verwendet man gerne die zweite. Die relative Häufigkeit h_i des Merkmalswerts w_i ist gegeben durch

$$h_i = \frac{H_i}{n} = \frac{\text{Absolute Häufigkeit}}{\text{Anzahl der Untersuchungsobjekte}}$$

bzw.

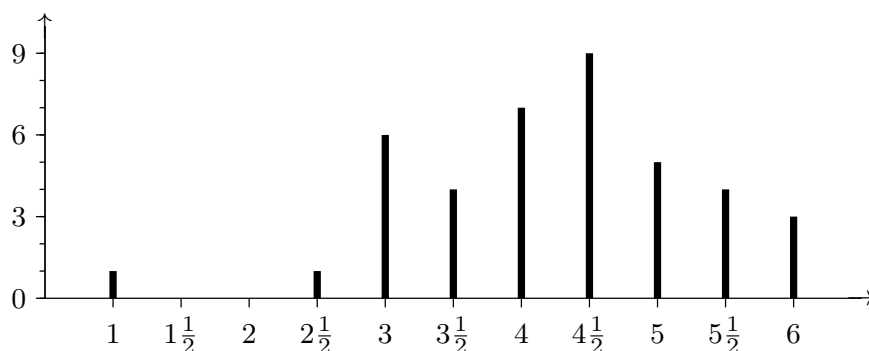
$$h_i = \frac{H_i}{n} \cdot 100\% .$$

Aus der obigen Strichliste erhalten wir so die folgende Tabelle:

w_i	H_i	h_i	h_i (%)
1	1	0.025	2.5%
$1\frac{1}{2}$	0	0	0%
2	0	0	0%
$2\frac{1}{2}$	1	0.025	2.5%
3	6	0.150	15%
$3\frac{1}{2}$	4	0.100	10%
4	7	0.175	17.5%
$4\frac{1}{2}$	9	0.225	22.5%
5	5	0.125	12.5%
$5\frac{1}{2}$	4	0.100	10%
6	3	0.075	7.5%

Es ist klar, dass die Summe der relativen Häufigkeiten h_i 1 bzw. 100% ergeben muss.

Besonders anschaulich sind graphische Darstellungen wie etwa das unten stehende *Stabdiagramm*:



Man kann die Ordinate sowohl mit den absoluten als auch mit den relativen Häufigkeiten beschriften. Die Höhe der eingetragenen Strecken (Stäbe) ist proportional zu den betreffenden Häufigkeiten. Anstelle der Strecken könnte man auch schmale Balken einzeichnen; diese sollten sich aber nicht gegenseitig berühren, vgl. (2.2.2.3).

Unter dem im Titel des Abschnitts gebrauchten zusammenfassenden Begriff "*Häufigkeitsverteilung*" versteht man einfach die Menge der Werte w_i mit den zugehörigen Häufigkeiten. Man kann sie wahlweise mit einer Tabelle oder einem Stabdiagramm beschreiben.

(2.2.2.3) Häufigkeitsverteilung bei stetigen Merkmalen

Wir beginnen wieder mit einem Beispiel.

Beispiel 2.2.2.3.A

Im Rahmen einer Untersuchung wurden 50 zweiwöchige Küken gewogen, wobei die Gewichte auf Gramm gerundet wurden. Man erhielt folgende Urliste (Angaben in Gramm):

100	87	101	107	102	105	91	104	103	102
99	96	104	93	105	107	103	106	96	111
104	92	107	101	109	90	106	97	103	112
101	103	108	105	105	110	97	109	112	103
108	98	104	106	97	119	99	115	100	106

Wie schon im Beispiel 2.2.2.2.A sind die Daten ungeordnet. ☒

Wiederum stellen wir eine Strichliste auf (siehe folgende Seite), wobei wir gleichzeitig die absoluten und die relativen Häufigkeiten (in %) eintragen.

Zum Beispiel (2.2.2.2) der Schulnoten (eines diskreten Merkmals) besteht der folgende wesentliche Unterschied: Das Gewicht ist ein stetiges Merkmal, welches (innerhalb

Messwert		absolute Häufigkeit	relative Häufigkeit (in %)
86		0	0
87		1	2
88		0	0
89		0	0
90		1	2
91		1	2
92		1	2
93		1	2
94		0	0
95		0	0
96		2	4
97		3	6
98		1	2
99		2	4
100		2	4
101		3	6
102		2	4
103		5	10
104		4	8
105		4	8
106		4	8
107		3	6
108		2	4
109		2	4
110		1	2
111		1	2
112		2	4
113		0	0
114		0	0
115		1	2
116		0	0
117		0	0
118		0	0
119		1	2
120		0	0
Total		50	100

gewisser Grenzen) jeden Wert annehmen kann. Bei den auf Gramm genau angegebenen Gewichten handelt es sich um gerundete (“diskretisierte”, vgl. (2.2.1.5)) Werte. Die Angabe “91” bedeutet, dass das betreffende Küken zwischen 90.5 und 91.5 Gramm wiegt.

Um eine eindeutige Festlegung zu erhalten, runden wir an den Intervallgrenzen jeweils ab: Ein Küken, das 91.5 g wiegt, zählt somit zur Gewichtsklasse 91. Dies ist bloss von theoretischem Interesse, wiegt doch ein Küken kaum genau 91.5 g.

Die Angaben in der Urliste und in der Strichliste beziehen sich also auf eine so genannte *Klasseneinteilung*:

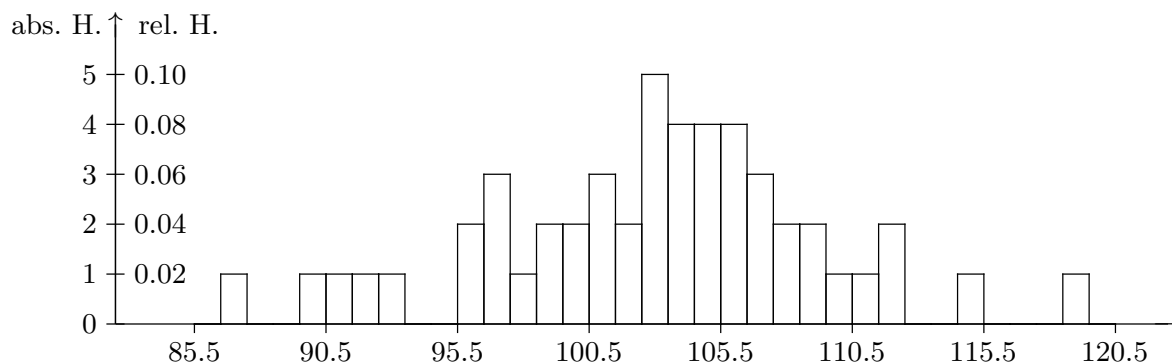
Gewicht x : $85.5 < x \leq 86.5$: Klasse 86,

Gewicht x : $86.5 < x \leq 87.5$: Klasse 87, etc.

Weiter gebraucht man den Begriff der *Klassenbreite* (hier 1 bzw. 1 g); die Zahlen

86, 87 etc. heissen die *Klassenmitten*.

Wir stellen nun auch diese Daten graphisch dar. Um klarzumachen, dass es sich hier um ein stetiges Merkmal handelt, zeichnen wir im Gegensatz zu (2.2.2.2) aneinanderliegende Balken. Die Höhe dieser Balken entspricht zunächst der relativen (oder absoluten) Häufigkeit, vgl. aber (2.2.2.4). Man erhält so ein *Histogramm* oder *Blockdiagramm*.



(2.2.2.4) Änderung der Klasseneinteilung

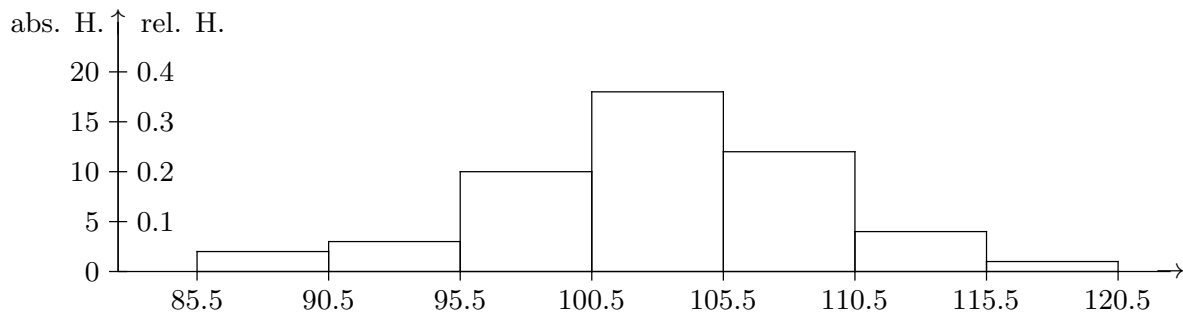
Das obige Histogramm ist noch nicht sehr übersichtlich. Das liegt daran, dass die absoluten Häufigkeiten der einzelnen Klassen zu klein sind, so dass sich Lücken und weitere Unregelmässigkeiten ergeben. Dies könnte man durch Erhöhung der Zahl der Untersuchungsobjekte ändern. Muss man aber mit der gegebenen Anzahl auskommen, so kann man versuchen, die Klassenbreite zu vergrössern.

Wir wählen* (mehr oder weniger willkürlich) die Klassenbreite 5 (5 Gramm) und beginnen wie vorhin mit 85.5 als unterster Klassengrenze. Wir erhalten so folgende Tabelle:

Klasse	Klassenmitte	abs. H.	rel. H.
$85.5 < x \leq 90.5$	88	2	0.04
$90.5 < x \leq 95.5$	93	3	0.06
$95.5 < x \leq 100.5$	98	10	0.20
$100.5 < x \leq 105.5$	103	18	0.36
$105.5 < x \leq 110.5$	108	12	0.24
$110.5 < x \leq 115.5$	113	4	0.08
$115.5 < x \leq 120.5$	118	1	0.02

* Die Anzahl der Klassen kann man an sich frei wählen. Zu viele Klassen ergeben aber ein unregelmässiges Bild, zu wenige unterdrücken zuviel Information.

Das zugehörige Histogramm lässt die wesentlichen Züge besser hervortreten als jenes in (2.2.2.3):

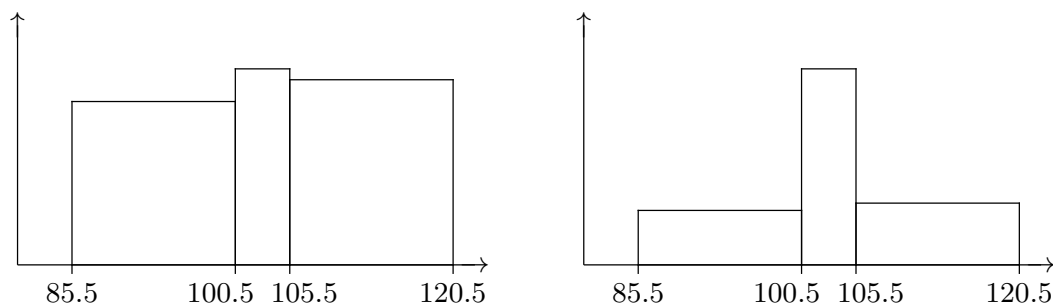


Beim Vergleich der beiden Histogramme fällt auf, dass im zweiten der Massstab der Ordinate verkürzt wurde, und zwar um den Faktor 5. Der Grund dafür ist der, dass die neuen Klassen fünfmal breiter sind als die alten. Die Massstabsänderung bewirkt nun, dass in den beiden Histogrammen der Flächeninhalt pro Untersuchungsobjekt (ein einzelnes Küken) derselbe ist. Anders ausgedrückt: Nicht die Höhe der Balken, sondern ihr Inhalt soll ein Mass für die Häufigkeit sein. Insbesondere ist dann der totale Flächeninhalt unter allen Balken zusammen in beiden Fällen derselbe (und zwar $= 1$, bzw. $= 100\%$, wenn wir relative Häufigkeiten betrachten). Dies ermöglicht einen guten Vergleich der beiden Histogramme.

Die Regel, dass der Flächeninhalt und nicht die Höhe der Balken massgebend ist, muss zwingend angewendet werden, wenn man, was gelegentlich nötig ist, mit verschiedenen Klassenbreiten im selben Histogramm arbeitet. Als Beispiel dazu betrachten wir die drei folgenden Klassen:

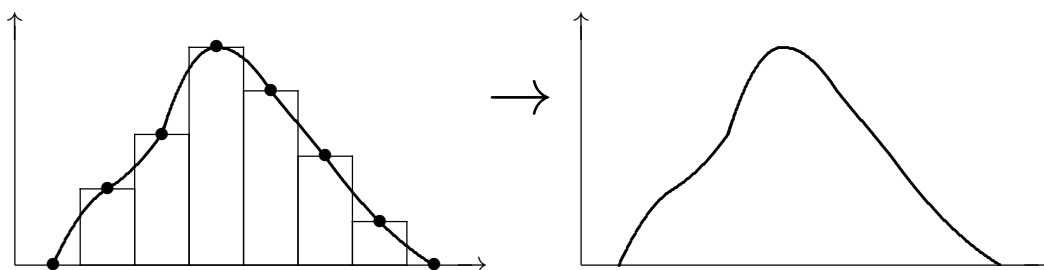
Klasse	Klassenbreite	abs. H.	rel. H.
$85.5 < x \leq 100.5$	15	15	0.30
$100.5 < x \leq 105.5$	5	18	0.36
$105.5 < x \leq 120.5$	15	17	0.34

Wir stellen diese Daten auf zwei Arten dar: In der Figur links ist (fälschlicherweise) die Höhe, in jener rechts (korrekterweise) der Flächeninhalt proportional zu den relativen Häufigkeiten. Man stellt sofort fest, dass die linke Darstellung rein optisch den falschen Eindruck erweckt, die mittlere Klasse sei viel kleiner als die beiden andern, obwohl sie in Tat und Wahrheit die grösste ist.

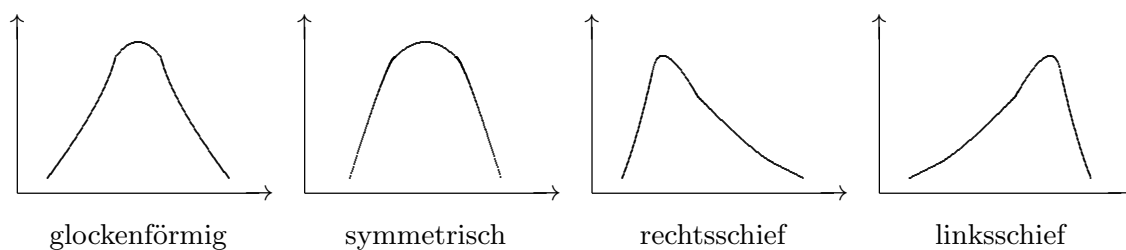


(2.2.2.5) Eine Idealisierung

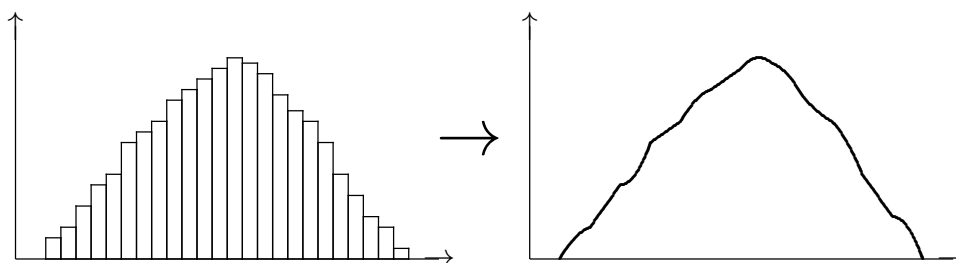
Gelegentlich verbindet man die Mitten der oberen Rechtecksseiten durch eine glatte Kurve und betrachtet diese dann für sich allein:



Dies tut man zum Beispiel dann, wenn man grob die Form der Häufigkeitsverteilung skizzieren will. Man unterscheidet etwa:



Bei diesem Prozess handelt es sich natürlich um eine Idealisierung. Hat man aber sehr viele und sehr genaue Messdaten, so kann man eine derart feine Klasseneinteilung wählen, dass diese “glatte” Idealisierung von der “Treppenkurve” des Histogramms kaum mehr abweicht.



Der eben beschriebene Übergang von der “Treppenkurve” des Histogramms zu einer “glatte” Kurve wird in der Wahrscheinlichkeitsrechnung wieder aufgenommen werden. Da er gewissermassen auf einer Art Grenzübergang beruht (immer mehr, dafür

immer feinere Klassen), kann er nur in Gedanken durchgeführt werden (daher der Ausdruck “Idealisierung”). Aus der relativen Häufigkeit wird bei diesem Übergang die Wahrscheinlichkeit, und die “glatte” Kurve bestimmt dann eine so genannte Wahrscheinlichkeitsverteilung. Näheres dazu finden Sie im Kapitel 4.

(2.2.2.6) Summenhäufigkeitsverteilung bei diskreten Merkmalen

Eine weitere Art der Darstellung von Versuchsergebnissen ist die Summenhäufigkeitsverteilung. Wir betrachten zur Erläuterung das Beispiel 2.2.2.2.A der Noten.

Wenn man fragt, wieviele Studierende eine ungenügende Note, d.h., zumindest in der Schweiz, eine Note $\leq 3\frac{1}{2}$ erzielt haben, dann ist die Antwort nicht direkt aus der Tabelle oder dem Stabdiagramm von (2.2.2.2) abzulesen. Vielmehr muss man die Häufigkeit aller Noten $\leq 3\frac{1}{2}$ addieren, und man erhält

$$1 + 0 + 0 + 1 + 6 + 4 = 12 .$$

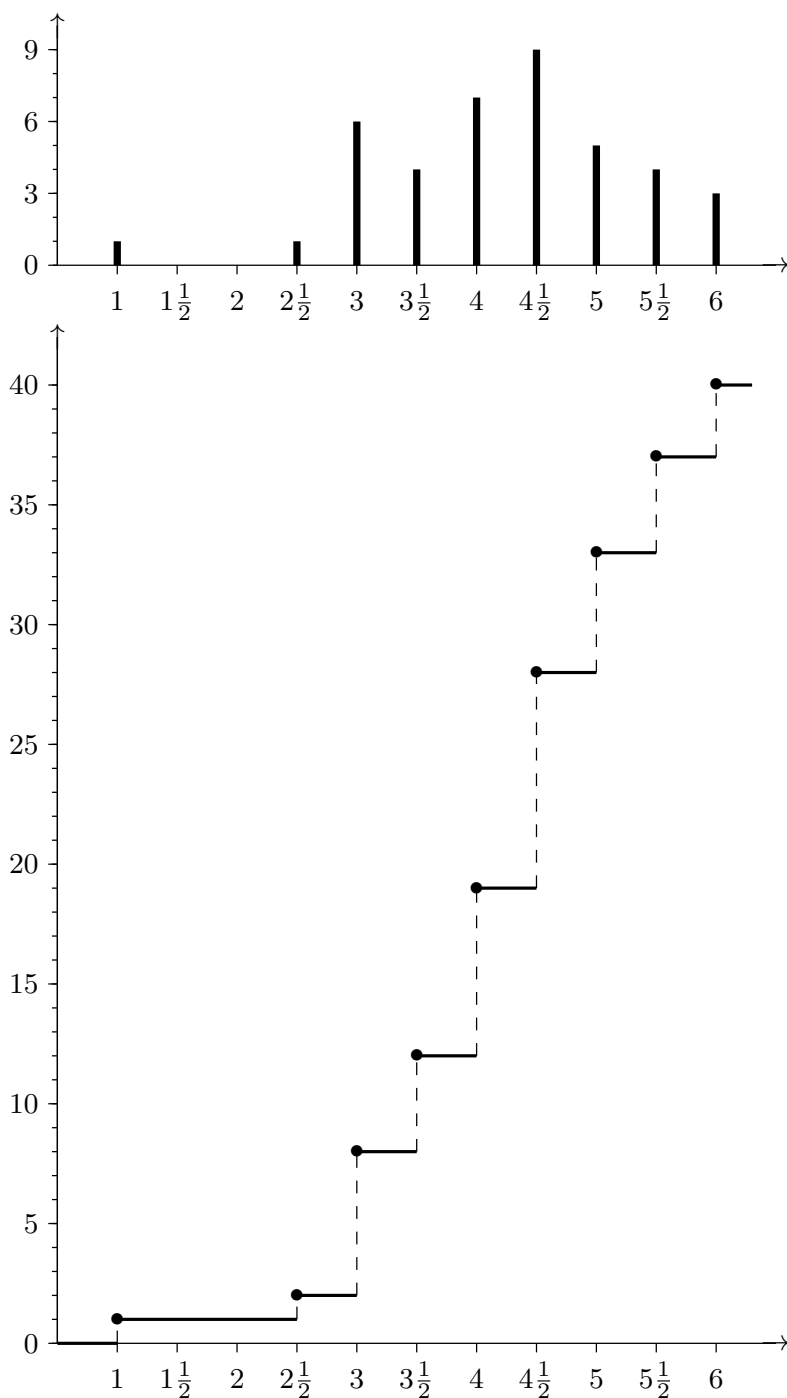
Die so erhaltene Zahl heisst eine *Summenhäufigkeit* (oder *kumulative Häufigkeit*). Diese Summenhäufigkeit gibt also an, wieviele Messwerte kleiner oder gleich* einer gewissen Zahl sind. Selbstverständlich kommt die Summenhäufigkeit sowohl in einer absoluten als auch in einer relativen Variante vor. In der folgenden Tabelle sind die Werte zu unserem Beispiel eingetragen (H. bedeutet Häufigkeit, SH. Summenhäufigkeit).

Note	abs. H.	rel. H.	abs. SH.	rel. SH.
1	1	0.025	1	0.025
$1\frac{1}{2}$	0	0	1	0.025
2	0	0	1	0.025
$2\frac{1}{2}$	1	0.025	2	0.050
3	6	0.150	8	0.200
$3\frac{1}{2}$	4	0.100	12	0.300
4	7	0.175	19	0.475
$4\frac{1}{2}$	9	0.225	28	0.700
5	5	0.125	33	0.825
$5\frac{1}{2}$	4	0.100	37	0.925
6	3	0.075	40	1.000

Natürlich ist die letzte Summenhäufigkeit gerade gleich der Anzahl n der untersuchten Objekte, bzw. gleich der relativen Häufigkeit 1.

* Es sei erwähnt, dass in manchen Büchern die Summenhäufigkeit mit $<$ statt mit \leq definiert wird.

Diese Summenhäufigkeit ist unten graphisch aufgetragen. Ein Vergleich mit dem nochmals aufgezeichneten Stabdiagramm zeigt, dass die “Sprünge” bei der Summenhäufigkeitsverteilung gerade die Höhe der entsprechenden Stäbe haben. Die gestrichelt eingetragenen vertikalen Linien kann man wahlweise eintragen oder weglassen. Zeichnet man sie ein, hat man eine schöne “Treppe”, lässt man sie weg, erhält man den Graphen einer Funktion.



Bemerkungen

- a) Um einen direkten Vergleich mit dem Stabdiagramm zu ermöglichen, musste in der obigen Figur der Massstab auf der Ordinate recht gross gewählt werden. Mit einem verkürzten Massstab bei der Summenhäufigkeitsverteilung würde die Darstellung übersichtlicher.
- b) An den Sprungstellen ist jeweils der obere Wert zu nehmen, was durch die ausgefüllten Punkte angedeutet wird. Dies kommt daher, dass wir zur Definition der Summenhäufigkeitsverteilung die Beziehung \leq und nicht die Beziehung $<$ verwendet haben.
- c) Die waagrechten Teilstücke in der Darstellung der Summenhäufigkeitsverteilung sind mit gutem Grund ausgezogen worden. Man kann nämlich die Summenhäufigkeiten auch an andern Stellen als den effektiv vorkommenden Noten $(1, 1\frac{1}{2}, 2, \dots)$ betrachten, indem man etwa die Frage stellt: Wieviele Personen haben eine Note ≤ 3.25 erzielt? Da diese Note in unserer Skala nicht vorkommt, gehört zu 3.25 dieselbe Summenhäufigkeit wie zu 3. Mit andern Worten: Der Graph hat an der Stelle 3.25 (und ebenso für alle Zahlen x mit $3 \leq x < 3.5$) dieselbe Höhe wie an der Stelle 3. Erst bei 3.5 "passiert wieder etwas".
- d) Die waagrechten Stücke in der Figur können als Graph einer Funktion aufgefasst werden, der so genannten (*empirischen*) *Verteilungsfunktion* $\tilde{F}(x)$. Der Zusatz "empirisch" soll den Begriff gegenüber der in der Wahrscheinlichkeitsrechnung vorkommenden "Verteilungsfunktion" (vgl. (4.1.7)) abgrenzen.

(2.2.2.7) Summenhäufigkeitsverteilung bei stetigen Merkmalen

Ähnlich wie in (2.2.2.6) bilden wir zu den Häufigkeiten von Beispiel 2.2.2.3.A die zugehörigen Summenhäufigkeiten. Es genügt, einen Ausschnitt aus der entsprechenden Tabelle zu geben:

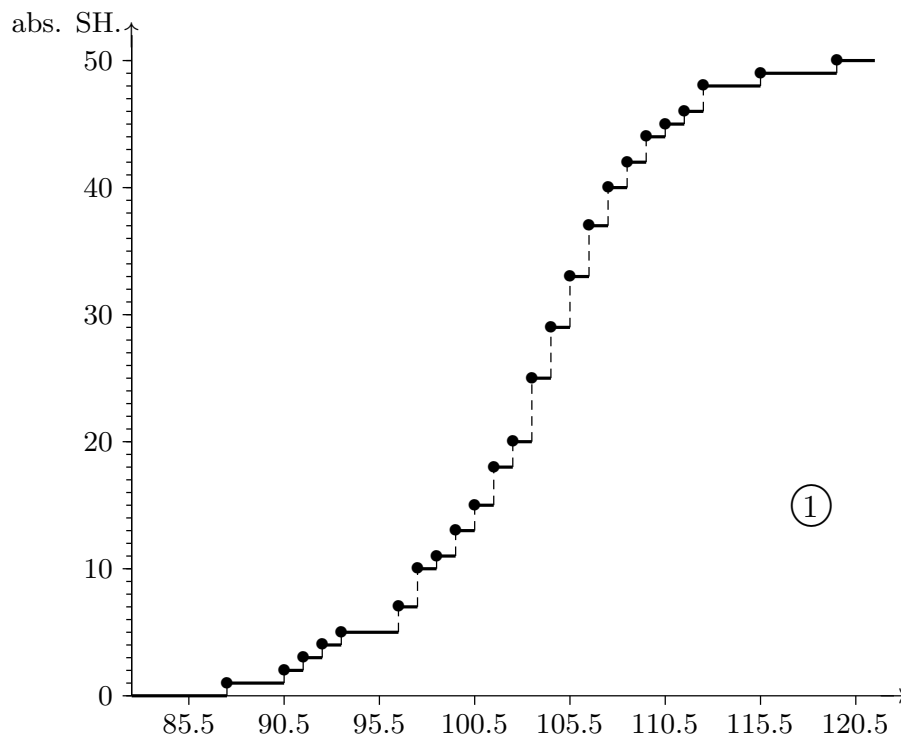
Gewicht	abs. H.	abs. SH.
86	0	0
87	1	1
88	0	1
89	0	1
90	1	2
	...	
101	3	18
102	2	20
103	5	25
104	4	29
105	4	33
	...	

Nun stellen wir diese Summenhäufigkeit graphisch dar.

Dabei tritt ein kleines Problem auf, das mit der Stetigkeit des Merkmals zu tun hat. Zur Klasse "87" beispielsweise gehören alle Hühner, deren Gewicht im (halboffenen) Intervall $(86.5, 87.5]$ liegt. Da also ein Küken mit einem Gewicht von 87.5 Gramm gerade

noch in dieser Klasse liegt, lassen wir die Treppenkurve des Graphen an der Stelle 87.5 springen. Andere Lehrmeinungen besagen, dass man den Sprung in die Klassenmitte, also an die Stelle 87, legen soll. Beide Verfahren bringen aber kleine Ungenauigkeiten.

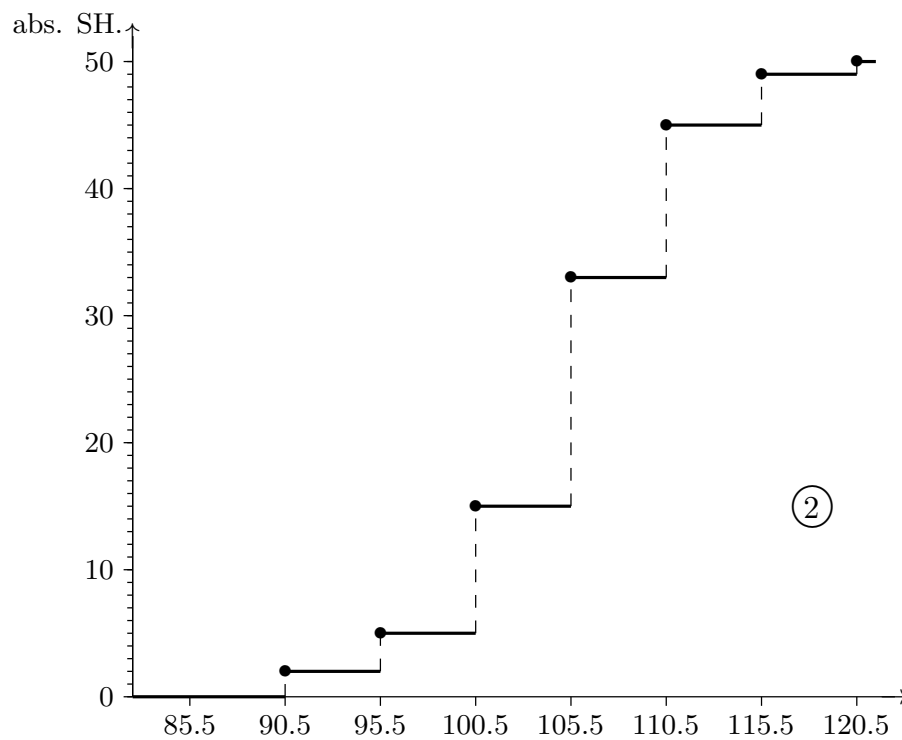
Wir erhalten das folgende Bild:



Das ganze Prozedere kann selbstverständlich mit anderen Klasseneinteilungen wiederholt werden. Mit der Einteilung von (2.2.2.4) finden wir die folgende Tabelle:

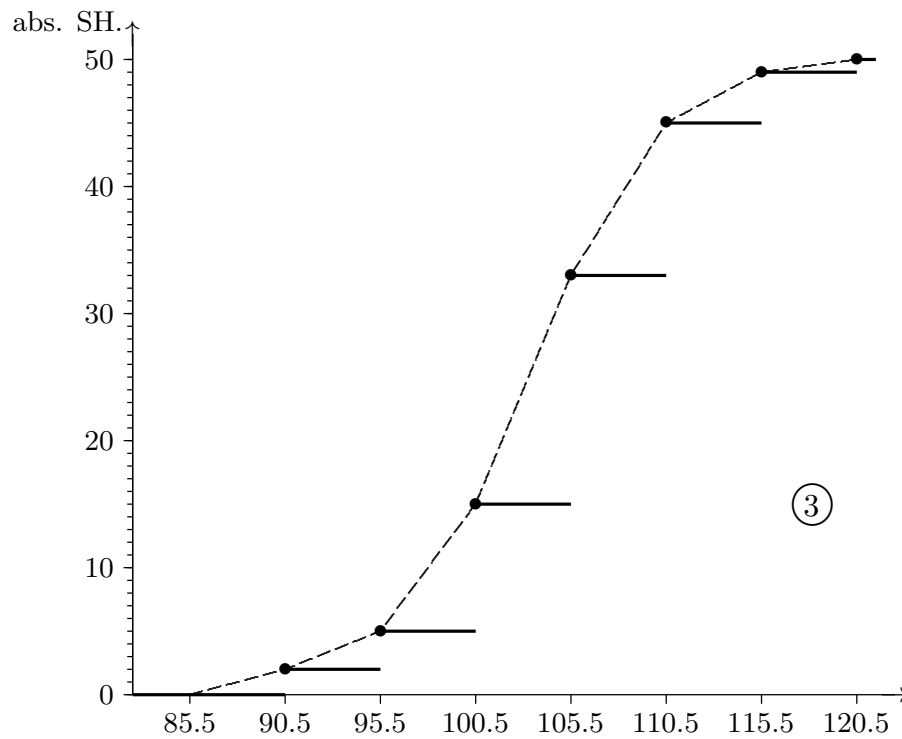
Klasse	abs. H.	abs. SH.
$85.5 < x \leq 90.5$	2	2
$90.5 < x \leq 95.5$	3	5
$95.5 < x \leq 100.5$	10	15
$100.5 < x \leq 105.5$	18	33
$105.5 < x \leq 110.5$	12	45
$110.5 < x \leq 115.5$	4	49
$115.5 < x \leq 120.5$	1	50

Auf diese Weise erhalten wir natürlich einen Graphen mit “grösseren Treppenstufen”. Die Sprungstellen liegen bei 90.5, 95.5 usw. Beachten Sie, dass die Figuren ① und ② denselben Ordinatenmassstab aufweisen. Dies ist hier — im Gegensatz zu (2.2.2.4) — zweckmässig.

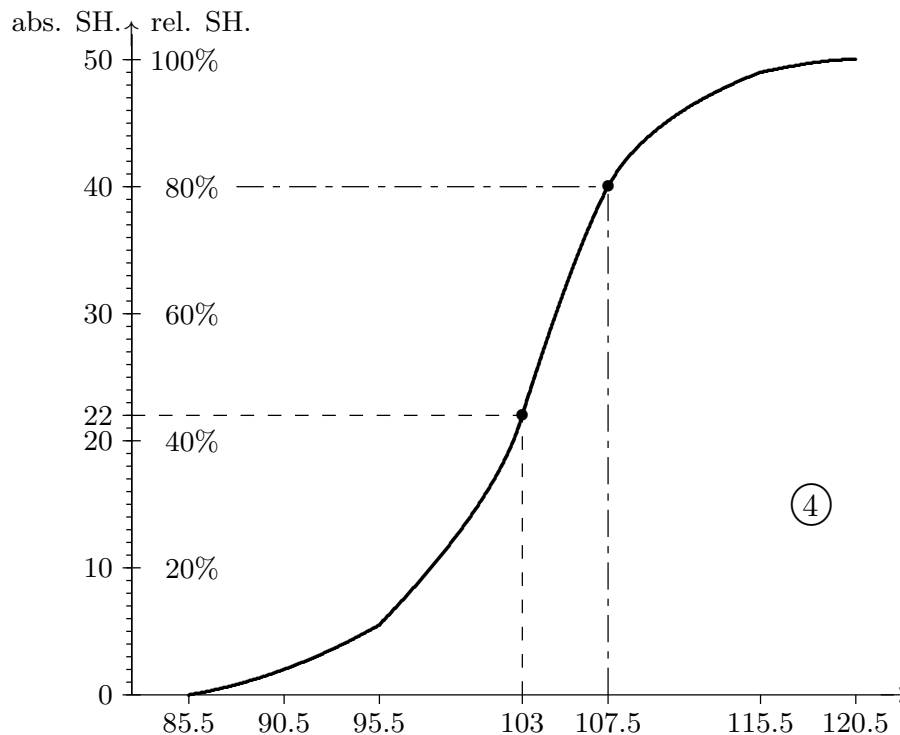


(2.2.2.8) Idealisierung der Summenhäufigkeitsverteilung

Verbinden wir in der Figur ② von (2.2.2.7) die Eckpunkte der “Treppe” durch Geradenstücke, so erhalten wir einen Streckenzug (gestrichelte Linie in Figur ③).



Wird die Klasseneinteilung verfeinert, d.h., unterwerfen wir Figur ① demselben Prozess, so nähert sich der Streckenzug einer glatten Kurve. Verfeinern wir in Gedanken die Klasseneinteilung noch mehr, so erhalten wir als Idealisierung eine glatte Kurve, die etwa so aussieht:



Eine solche Kurve kann man dazu benützen, gewisse Daten auf einfache Weise abzulesen. Wir illustrieren das Vorgehen an Figur ④.

- a) Wieviele Küken wiegen ≤ 103 Gramm?

Wir gehen von 103 aus entlang der gestrichelten Geraden nach oben bis zur Kurve und dann nach links. Wir finden so die Antwort: 22.

Aufgrund der ursprünglichen Daten wissen wir allerdings nur, dass 20 Küken unter 102.5 g und 25 Küken unter 103.5 g wiegen. Die erhaltene Zahl 22 ist das Resultat einer Interpolation aufgrund der (idealisierten) Kurve und kein exakter Wert.

Noch ein zweiter Einwand: Es hätte ja als Schnittpunkt mit der Ordinate auch z.B. die Zahl 22.5 herauskommen können, was als Häufigkeit sinnlos ist. Immerhin ist man bei relativen Häufigkeiten (Prozentzahlen) eher gewohnt, auch nicht ganze Zahlen zu akzeptieren.

- b) Für welches Gewicht x gilt, dass 80% der Tiere höchstens x Gramm wiegen?

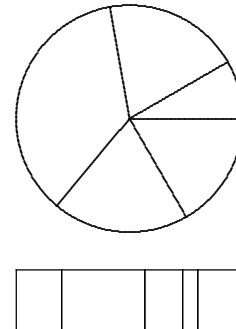
Hier gehen wir von der Zahl 80% auf der Ordinate entlang der strichpunktieren Geraden bis zur Kurve und dann senkrecht nach unten. Als Antwort finden wir $x = 107.5$. Natürlich gilt dieselbe Kritik wie unter a).

(2.2.2.9) Schlussbemerkungen

a) Andere Darstellungsarten

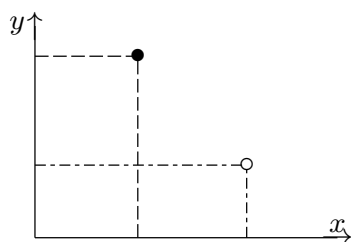
Aus dem täglichen Leben sind Ihnen Darstellungsarten wie “Kreisdiagramme” oder “Balkendiagramme” sicher vertraut. Wir werden diese hier nicht weiter betrachten. Immerhin sollten Sie an diese verschiedenen Möglichkeiten denken, wenn es darum geht, einen Sachverhalt mit einfachen Mitteln anschaulich zu beschreiben.

In derartigen Darstellungen ist der Flächeninhalt des Kreissektors bzw. des Teilrechtecks ein Mass für die Häufigkeit.

b) Untersuchung von mehreren Merkmalen

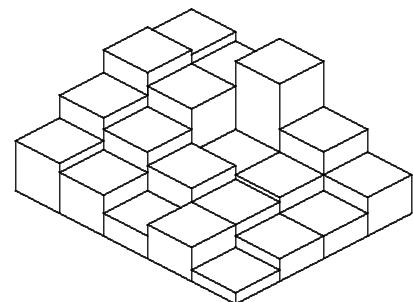
Im ganzen Kapitel wurde davon ausgegangen, dass an jedem Untersuchungsobjekt nur ein einziges Merkmal beobachtet wurde. Das ist natürlich nicht immer der Fall. So kann an einem einzelnen Menschen etwa das Gewicht, die Körperlänge, der Brustumfang, die Schuhgrösse, der systolische Blutdruck usw. usw. gemessen werden. Die übersichtliche Darstellung der gewonnenen Daten ist natürlich umso schwieriger, je mehr Grössen gleichzeitig erfasst werden sollen.

Im Fall von zwei Grössen (wie etwa Gewicht und Körperlänge) kann man in einem kartesischen Koordinatensystem das Gewicht auf der x -Achse, die Körpergrösse auf der y -Achse abtragen. Zu jeder untersuchten Person gehört dann ein Punkt in der x - y -Ebene:



- : Niedriges Gewicht, grosse Körperlänge.
- : Hohes Gewicht, kleine Körperlänge.

Das “eindimensionale” Histogramm von (2.2.2.3) wird im Fall von zwei zu untersuchenden Merkmalen durch ein “zweidimensionales” Histogramm gemäss nebenstehender Figur ersetzt.



2.2.3. STATISTISCHE MASSZAHLEN

(2.2.3.1) Überblick

Statistische Masszahlen dienen dazu, die Verteilung von beobachteten Daten auf prägnante Weise zusammenzufassen. Wir betrachten hier nur zwei Arten von derartigen Grössen, nämlich *Lagemasse* und *Streuungsmasse* und zwar im Einzelnen (2.2.3.2)

- Lagemasse:
 - Durchschnitt, (2.2.3.3)
 - Median, (2.2.3.4)
 - Modus. (2.2.3.5)
- Streuungsmasse:
 - Variationsbreite, (2.2.3.6)
 - Interdezilbereich, (2.2.3.6)
 - Varianz, (2.2.3.7)
 - Standardabweichung. (2.2.3.7)

(2.2.3.2) Allgemeine Betrachtungen

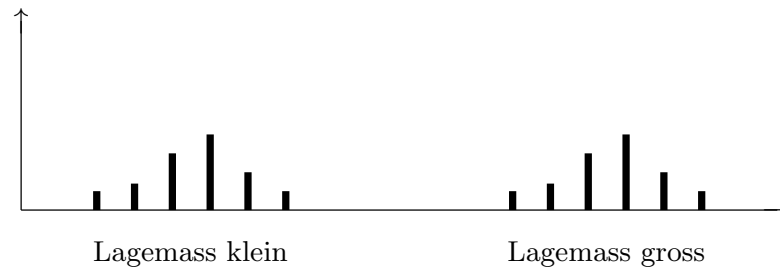
Wenn wir eine Anzahl von Beobachtungsdaten vor uns haben, wie etwa in den Beispielen von Teil 2.2.2, dann steckt die volle Information in der Urliste ((2.2.2.2.A) bzw. (2.2.2.3.A)), die sämtliche Daten enthält. Meist sind diese Angaben aber zu zahlreich, so dass man gezwungen ist, eine übersichtlichere Darstellung in kondensierter Form zu suchen. Eine Möglichkeit besteht darin, das Stabdiagramm bzw. das Histogramm der Häufigkeitsverteilung ((2.2.2.2) bzw. (2.2.2.3)) oder die Summenhäufigkeitsverteilung (2.2.2.6) zu zeichnen.

Oft verkleinert man dabei bewusst den Informationsgehalt, wenn man dafür die charakteristischen Züge besser darstellen kann. Dies haben wir etwa bei der Gruppierung nach Klassen gesehen.

Häufig möchte man die Verteilung der Beobachtungsdaten nicht graphisch, sondern mit möglichst wenigen, aussagekräftigen Zahlen darstellen (natürlich unter Verlust der einzelnen Detailangaben). Dazu dienen die *statistischen Masszahlen*, von denen uns zwei Gruppen interessieren werden:

a) Lagemasse

Gesucht ist eine Zahl, welche die Lage der Beobachtungsdaten auf der Abszisse (also der horizontalen Achse) beschreibt:



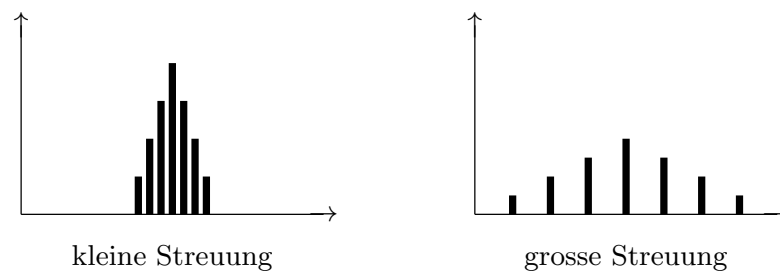
Das bekannteste Lagemass ist der Durchschnitt (das arithmetische Mittel). Daneben gibt es aber auch noch andere Lagemasse. Diese sind insofern von Bedeutung, als das arithmetische Mittel gar nicht immer sinnvoll angewandt werden kann. Dies hängt nämlich davon ab, welche Skala vorliegt (eine entsprechende Bemerkung wurde schon am Schluss von (2.2.1.6) gemacht).

Wir werden die folgenden Lagemasse besprechen:

- 1) Durchschnitt (oder arithmetisches Mittel) in (2.2.3.3),
- 2) Median (oder Zentralwert) in (2.2.3.4),
- 3) Modus (oder Dichtemittel) in (2.2.3.5).

b) Streuungsmaße

Gesucht ist eine Zahl, die angibt, wie die Daten um das Lagemass herum "streuen", d.h., ob sie nahe beieinander oder weit gestreut liegen. Anschaulich:



Auch hier gibt es verschiedene derartige Masse; die folgenden vier werden weiter unten behandelt:

- 1) Variationsbreite in (2.2.3.6),
- 2) Interdezilbereich in (2.2.3.6),
- 3) Varianz in (2.2.3.7),
- 4) Standardabweichung in (2.2.3.7).

Dabei unterscheiden sich die beiden letzten Grössen insofern nicht wesentlich, als die Standardabweichung die Wurzel aus der Varianz ist.

(2.2.3.3) Das arithmetische Mittel

a) Die Definition

Wir gehen davon aus, dass n beobachtete (gemessene oder gezählte) Werte bekannt sind. Diese bezeichnen wir mit

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

Dabei kann derselbe Wert ohne weiteres mehrfach vorkommen. Im Beispiel 2.2.2.3.A etwa ist $n = 50$, ferner ist

$$x_1 = 100, x_2 = 87, x_3 = 101, \dots, x_{49} = 100, x_{50} = 106 .$$

Das *arithmetische Mittel* (oft auch *Durchschnitt* genannt) wird mit \bar{x} bezeichnet und ist definiert durch die Formel

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

b) Beispiele

1. Die Durchschnittsnote der 40 Werte aus (2.2.2.2) ist

$$\frac{1}{40}(4 + 4.5 + 5 + \dots + 3 + 4) = \frac{169}{40} = 4.225 .$$

2. Das Durchschnittsgewicht der 50 Küken aus (2.2.2.3) beträgt

$$\frac{1}{50}(100 + 87 + 101 + \dots + 100 + 106) = \frac{5148}{50} = 102.96 .$$

Mit vielen Taschenrechnern kann man den Durchschnitt direkt ausrechnen. Die Daten werden oft mit der Taste $\boxed{\Sigma+}$ eingegeben (falsch eingegebene Werte werden mit $\boxed{\Sigma-}$ gelöscht); der Durchschnitt wird mit $\boxed{\bar{x}}$ abgerufen.

c) Anwendungsbereich des arithmetischen Mittels

Das arithmetische Mittel ist

- sinnvoll für Intervall- und Verhältnisskalen (2.2.1.7), sofern nicht die Verteilung sehr schief (2.2.2.5) ist. In diesem Fall pflegt man den Median zu verwenden (für ein Beispiel siehe (2.2.3.4)),
- nicht sinnvoll für Nominal- und Ordinalskalen (vgl. den Schluss von (2.2.1.6)).

d) Bildung des Durchschnitts bei zu Klassen gruppierten Daten

Liegt eine Klasseneinteilung wie etwa in (2.2.2.4) vor, so kennt man den genauen Messwert für ein einzelnes Objekt nicht und nimmt deshalb (mangels präziserer Information) an, er sei gleich der Klassenmitte der Klasse, zu welcher das Objekt gehört. Zur Berechnung des Durchschnitts zählt man dann diese Zahl so oft, wie es Objekte in der Klasse hat. Wir betrachten die Zahlen aus dem Beispiel von (2.2.2.4), wo eine Klassenbreite von 5 Gramm vorliegt:

Klasse	Klassenmitte	abs. Häufigkeit.
$85.5 < x \leq 90.5$	88	2
$90.5 < x \leq 95.5$	93	3
$95.5 < x \leq 100.5$	98	10
$100.5 < x \leq 105.5$	103	18
$105.5 < x \leq 110.5$	108	12
$110.5 < x \leq 115.5$	113	4
$115.5 < x \leq 120.5$	118	1

Wir finden

$$\bar{x} = \frac{1}{50}(2 \cdot 88 + 3 \cdot 93 + 10 \cdot 98 + 18 \cdot 103 + 12 \cdot 108 + 4 \cdot 113 + 1 \cdot 118) = 103.1 .$$

Es ist nicht verwunderlich, dass das Ergebnis etwas anders als im Beispiel 2. von b) ausgefallen ist. (Genau betrachtet betraf auch dieses Beispiel bereits eine Klasseneinteilung mit der Klassenbreite 1 Gramm.)

Die allgemeine Formel lautet

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k H_i x_i = \sum_{i=1}^k \frac{H_i}{n} x_i = \sum_{i=1}^k h_i x_i .$$

Dabei wurden die folgenden Bezeichnungen verwendet: k = Anzahl der Klassen, H_i = Anzahl Objekte in der i -ten Klasse (absolute Häufigkeit), x_i = Mitte der i -ten Klasse, $h_i = H_i/n$ = relative Häufigkeit. Selbstverständlich ist $n = \sum_{i=1}^k H_i$.

Diese Formel wurde hier auf stetige Daten angewandt. Sie kann aber auch bei diskreten Merkmalen benutzt werden, wenn bekannt ist, dass der Wert x_i gerade H_i -mal vorkommt. Im Beispiel 2.2.2.2.A tritt etwa die Note 3 genau sechsmal auf. Wir finden so für \bar{x} den Wert

$$\bar{x} = \frac{1}{40}(1 \cdot 1 + 1 \cdot 2.5 + 6 \cdot 3 + 4 \cdot 3.5 + 7 \cdot 4 + 9 \cdot 4.5 + 5 \cdot 5 + 4 \cdot 5.5 + 3 \cdot 6) = 4.225 ,$$

also dieselbe Zahl wie in b) oben.

(2.2.3.4) Der Median

Bei diesem zweiten Lagemass behandeln wir zunächst den Fall, wo die Beobachtungs- oder Messdaten in Form einer Urliste einzeln bekannt sind. Nachher untersuchen wir die Situation, wo die Daten bereits zu Klassen gruppiert sind und wo also nur die absoluten Häufigkeiten pro Klasse bekannt sind.

a) Die Daten sind einzeln bekannt

Wir ordnen die Daten der Grösse nach. Beachten Sie, dass dies bereits im Fall einer Ordinalskala (und erst recht natürlich für Intervall- und Verhältnisskalen) möglich ist. Im Beispiel (2.2.2.3) der Küken erhalten wir die folgende Anordnung der Gewichte:

87, 90, 91, 92, 93, 96, 96, 97, 97, 97, 98, 99, 99, . . . , 111, 112, 112, 115, 119 .

Der *Median* \tilde{x} (Synonym: *Zentralwert*) dieser Daten ist derjenige Wert, der in der Mitte der obigen Folge steht.

Anders formuliert: Der Median teilt die geordnete Folge der Daten in zwei gleich grosse Hälften. Im folgenden Beispiel mit fünf Daten

12, 14, 14, 16, 20,

ist $\tilde{x} = 14$.

Eine kleine Schwierigkeit ergibt sich, wenn die Anzahl n der Daten gerade ist. Hier gibt es zwei "mittlere" Werte und wir definieren den Median einfach als Durchschnitt dieser beiden Zahlen*. Somit ist der Median von

12, 14, 14, 16, 20, 25

gleich $\frac{1}{2}(14 + 16) = 15$.

Im oben erwähnten Beispiel der Küken ist $n = 50$, also gerade. Wir nehmen also den Durchschnitt des 25. und des 26. Werts in der aufsteigenden Folge der Gewichte. Der Strichliste aus (2.2.2.3) entnimmt man sofort, dass diese Werte = 103 bzw. = 104 sind, und es folgt $\tilde{x} = 103.5$.

b) Anwendungsbereich des Medians

Der Median ist

- sinnvoll für Ordinal-, Intervall- und Verhältnisskalen,
- nicht sinnvoll für Nominalskalen.

Die Begründung dafür wurde bereits eingangs des Abschnitts gegeben: Da von der Möglichkeit der Anordnung Gebrauch gemacht wird, muss (mindestens) eine Ordinalskala vorliegen.

* Bei blossen Ordinalskalen ist die Durchschnittsbildung streng genommen nicht erlaubt, doch sehen wir hier darüber hinweg. In der Literatur findet man auch andere Möglichkeiten, dieses (kleine) Problem zu meistern.

Bei Intervall- und Verhältnisskalen ist zwar der Durchschnitt das häufigste Lagemass. Manchmal ist aber auch hier der Median vorzuziehen. Er ist nämlich gegenüber so genannten Ausreissern viel weniger empfindlich als der Durchschnitt. (Unter einem *Ausreisser* versteht man einen extrem hohen oder niedrigen Wert in einer Reihe sich sonst wenig unterscheidenden Werte.) Betrachten wir etwa die Daten

$$12, 14, 14, 16, 20 \quad \text{bzw.} \quad 12, 14, 14, 16, 2000,$$

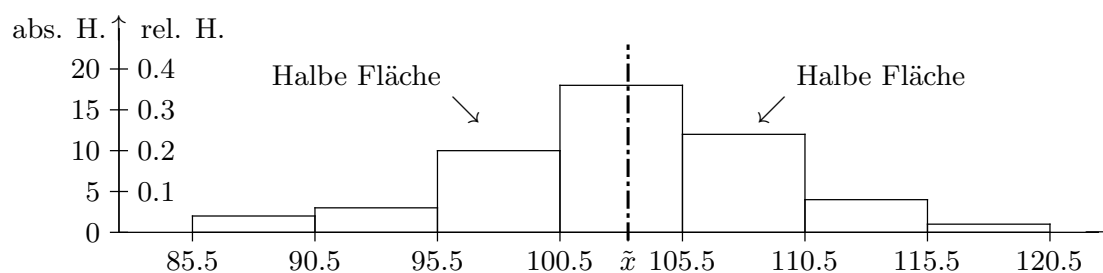
so ist der Median in beiden Fällen derselbe, die Durchschnitte unterscheiden sich aber sehr stark.

Zur weiteren Illustration betrachten wir die (fiktiven) Einkommensverhältnisse von 101 Personen: 30 Personen verdienen je 3000.–, 50 Personen je 4000.–, 20 Personen je 5000.– und eine Person verdient 100000.–. Der Median ist der 51. Wert in der geordneten Folge, es ist somit $\tilde{x} = 4000$. (Dabei haben wir in Gedanken die Einkommen der Grösse nach geordnet: Zuerst steht dreissigmal die Zahl 3000, dann kommt fünfzigmal die Zahl 4000, usw.) Das arithmetische Mittel $\bar{x} = 13762.38$ liefert hier offensichtlich einen ganz falschen Eindruck.

c) Die Daten sind zu Klassen gruppiert

Als Beispiel betrachten wir die Daten von (2.2.2.4). Diesen Angaben können wir nun, da sie in Klassen zusammengefasst sind, nur noch entnehmen, dass sowohl der 25. als auch der 26. Wert in der Klasse “101 bis 105” (genauer “100.5 bis 105.5”) liegt. Die einfachste Lösung besteht darin, diese Klasse als *Medianklasse* zu bezeichnen und es dabei bewenden zu lassen.

Man kann auch etwas raffinierter vorgehen. Dazu betrachten wir das Histogramm der zugehörigen Häufigkeitsverteilung:



Da der Flächeninhalt der einzelnen Balken ein Mass für die zugehörigen Häufigkeiten ist, kann man den Median als jene Zahl \tilde{x} bezeichnen, welche die Eigenschaft hat, dass die durch sie gelegte Parallele zur Ordinate den Flächeninhalt des Histogramms halbiert.

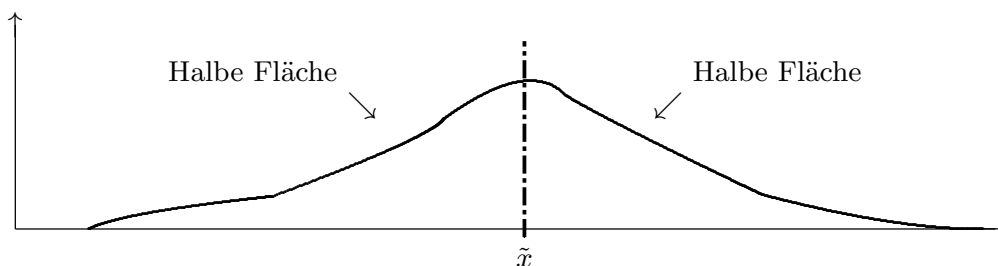
In unserm Beispiel sieht der Prozess zahlenmässig so aus: Die ersten drei Klassen enthalten zusammen 15 Küken, die letzten drei enthalten 17 Küken. Wir müssen also die mittlere Klasse (mit 18 Küken) “gerecht” auf die linken drei bzw. die rechten drei

Klassen verteilen, nämlich 10 Küken nach links und 8 nach rechts. Wir teilen also die Klasse “100.5 bis 105.5” mit der Breite 5 im Verhältnis 10 : 8. Um den Median \tilde{x} zu finden, addieren wir $\frac{10}{18}$ der Klassenbreite 5 zum linken Randpunkt 100.5 und erhalten

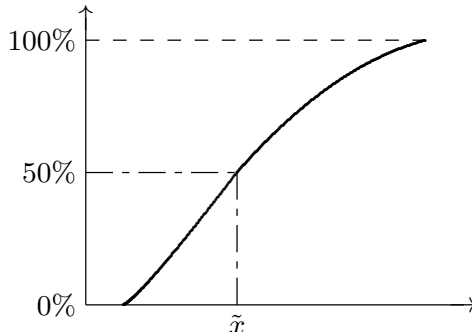
$$\tilde{x} = 100.5 + \frac{10}{18} \cdot 5 = 103.28 .$$

Wir verzichten darauf, die Methode in eine allgemeine Formel zu kleiden.

Das eben beschriebene Vorgehen lässt sich in natürlicher Weise idealisieren. Ersetzen wir nämlich das Histogramm im Sinne von (2.2.2.5) durch eine glatte Kurve, so können wir auch hier \tilde{x} durch die Bedingung definieren, dass die Senkrechte durch \tilde{x} die Fläche unter der Kurve halbiert.



Der Median \tilde{x} lässt sich auch aus der Kurve der (idealisierten) Summenhäufigkeit ablesen, indem man von der 50%-Marke nach rechts bis zur Kurve und dann senkrecht nach unten geht.

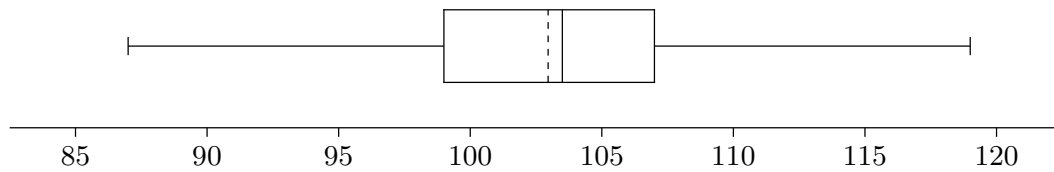


In (2.2.2.8) ist eine analoge Überlegung (allerdings für 80% statt für 50%) bereits durchgeführt worden.

d) Ergänzung

Ohne detaillierte Erläuterungen sei hier noch auf einige weitere Masszahlen hingewiesen, die mit dem Median verwandt sind. Dieser gibt die Mitte der geordneten Messreihe an, analog markiert das *untere bzw. obere Quartil* die beiden “Viertel-Positionen”. Somit sind 25% der Werte kleiner als das untere Quartil. Statt 25% kann man natürlich auch einen andern Prozentsatz nehmen: Wenn p eine Zahl mit $0 < p < 1$ ist, dann ist das p -Quantil dadurch definiert, dass 100 p % der Werte kleiner sind. Das 0.75-Quantil beispielsweise ist also dasselbe wie das obere Quartil.

Im Fall der Küken (Beispiel 2.2.2.3.A) ist das untere Quartil der 13. Wert von unten, also 99, das obere Quartil der 13. Wert von oben, nämlich 107. Diese Zahlen lassen sich, zusammen mit andern wichtigen Daten, in einem so genannten *Boxplot* darstellen.



Der “Kasten” reicht vom untern bis zum obern Quartil, also über 50% der Werte, die beiden Strecken beginnen beim kleinsten, bzw. enden beim grössten Wert. Jede deckt 25% der Werte ab. Die durchgezogene Linie im Kasten gibt den Median, die gestrichelte das arithmetische Mittel an.

(2.2.3.5) Der Modus

Als *Modus* (Synonyme: *Modalwert*, *Dichtemittel*) bezeichnet man jeden Wert, der am häufigsten auftritt. In der Wertereihe

10, 10, 11, 11, 11, 11, 12, 12, 13, 13, 13, 14

kommt der Wert 11 am häufigsten vor und ist deshalb der Modus.

Die Reihe

10, 10, 11, 11, 11, 11, 12, 12, 13, 13, 13, 13, 14

hat zwei Modi, nämlich 11 und 13.

Im Beispiel 2.2.2.3.A der Küken ist der Modus = 103.

Der Modus ist sicher für Ordinal-, Intervall- und Verhältnisskalen sinnvoll, man kann ihn aber sogar auf Nominalskalen anwenden, wie das folgende Beispiel zeigt: Im Zürcher Kantonsrat gilt in der Legislaturperiode 2023–2027 die folgende Sitzverteilung (Zahlen vom Wahlabend vor Parteiwechseln):

AL 5, EDU 3, EVP 7, FDP 29, Grüne 19, Grünliberale 24, Die Mitte 11, SP 36
SVP 46.

Der Modus liegt hier bei der SVP.

(2.2.3.6) Die Variationsbreite und der Interdezilbereich

Wir kommen nun zu den bereits am Schluss von (2.2.3.2) vorgestellten Streuungsmassen. Das einfachste Streuungsmass ist die *Variationsbreite* (Synonym: *Spannweite*). Sie ist als Differenz zwischen dem grössten und dem kleinsten Wert definiert. Im Fall der Küken aus (2.2.2.3) ist sie gleich $119 - 87 = 32$. Diese Variationsbreite ist äusserst einfach zu berechnen. Sie hat den Nachteil, dass die zwischen den Extremen liegenden Werte überhaupt nicht beachtet werden; ferner ist sie sehr empfindlich gegenüber Ausreissern.

Diese Abhängigkeit von Ausreissern kann man dadurch vermindern, dass man bei der Berechnung der Variationsbreite die kleinsten und die grössten Werte nicht

berücksichtigt. Lässt man auf beiden Seiten der nach der Grösse geordneten Folge der Werte jeweils 10% der Gesamtzahl der Beobachtungen weg und berechnet die Differenz zwischen dem jetzt gültigen grössten bzw. kleinsten Wert, so erhält man den so genannten *Interdezilbereich*. Im Beispiel der Küken (mit $n = 50$) sind also jeweils 5 Beobachtungen wegzulassen. Der kleinste übrig bleibende Wert ist dann 96, der grösste 110, so dass der Interdezilbereich gleich $110 - 96 = 14$ ist. Natürlich gibt es auch die Möglichkeit, statt 10% einen anderen Prozentsatz zu verwenden.

Bei diesen Verfahren werden Differenzen gebildet (und auch die Grössenordnung wird benützt), so dass eine Intervall- oder eine Verhältnisskala vorhanden sein muss.

(2.2.3.7) Die Varianz und die Standardabweichung

a) Einleitende Betrachtungen

Die wichtigsten Streuungsmasse sind die Standardabweichung und ihr Quadrat, die Varianz. Da diese Masse auf dem arithmetischen Mittel basieren, sind sie nur für Intervall- und Verhältnisskalen sinnvoll.

Wir stellen zuerst einige allgemeine Betrachtungen an. Wenn n Beobachtungen mit den Werten x_1, \dots, x_n vorliegen, dann können wir gemäss (2.2.3.3) das arithmetische Mittel

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

bilden. Die Abweichungen der einzelnen Daten vom Durchschnitt sind gegeben durch die Differenzen

$$x_i - \bar{x}, \quad i = 1, \dots, n.$$

Man könnte nun versucht sein, die Summe dieser Abweichungen als Mass für die Streuung zu verwenden. Dies ist aber nicht sinnvoll, weil diese Summe immer $= 0$ ist, denn $(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = x_1 + x_2 + \dots + x_n - n\bar{x} = 0$ nach Definition von \bar{x} . Die Abweichungen sind eben teils positiv, teils negativ und heben sich auf.

Dies ändert sich, wenn man die Absolutbeträge $|x_i - \bar{x}|$ betrachtet. In der Tat ist

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

ein Streuungsmass, die *mittlere absolute Abweichung*, das aber nur selten gebraucht wird, u.a. deshalb, weil der Umgang mit Absolutbeträgen etwas mühsam ist.

Man schafft die Vorzeichen der Differenzen lieber auf andere Weise weg*, indem man quadriert und $(x_i - \bar{x})^2$ betrachtet. Für die Summe dieser Quadrate der Abweichungen

* Bei der Methode der kleinsten Quadrate (23.7) im ersten Band wird analog vorgegangen.

ist die Bezeichnung SS_{xx} üblich (SS steht für Sum of Squares; die Indizes xx haben nichts mit partiellen Ableitungen (23.4) aus dem ersten Band zu tun), es ist also

$$SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 .$$

Zur Berechnung von SS_{xx} sind folgende Umformungen nützlich*. Es ist

$$\begin{aligned} SS_{xx} &= (x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \\ &= (x_1^2 - 2x_1\bar{x} + \bar{x}^2) + \dots + (x_n^2 - 2x_n\bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 . \end{aligned}$$

Nun ist aber $\sum_{i=1}^n x_i = n\bar{x}$, und durch Einsetzen finden wir

$$SS_{xx} = \sum_{i=1}^n x_i^2 - 2\bar{x} \cdot n\bar{x} + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 ,$$

oder, wenn statt $n\bar{x}$ wieder $\sum_{i=1}^n x_i$ eingesetzt wird,

$$SS_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 .$$

Diese Formeln sind bequem, weil man nicht n Differenzen $x_i - \bar{x}$ berechnen muss, vgl. Beispiel 1. weiter unten.

b) Definition

Nun kommen wir endlich zur Definition der *Varianz*, welche mit s^2 bezeichnet wird:

$$s^2 = \frac{1}{n-1} SS_{xx} .$$

Diese Definition ist nur für $n > 1$ sinnvoll. Die Wahl von $n-1$ (und nicht n) als Nenner können wir später in (8.2.6) begründen. Für den Moment ist es am besten, die obige Formel einfach als Definition zu akzeptieren.

Da für SS_{xx} verschiedene gleichwertige Ausdrücke zur Verfügung stehen, hat man für die Varianz die folgenden drei Formeln zur Auswahl:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 , \\ s^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) , \\ s^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) . \end{aligned}$$

* Eine ähnliche Rechnung finden Sie am Schluss von (23.7) im ersten Band.

Die Bezeichnung s^2 wurde deshalb gewählt, weil man häufig nicht die Varianz, sondern ihre Quadratwurzel s betrachtet, welche *Standardabweichung* genannt wird. In Formeln ausgedrückt ist

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

Es kann auch eine der oben angegebenen Varianten verwendet werden. Beachten Sie, dass gemäss der üblichen Konvention, (26.3.c.4) im ersten Band, das Wurzelzeichen stets die positive Wurzel bedeutet.

Einer der Gründe dafür, dass man die Standardabweichung statt der Varianz gebraucht, ist der folgende: Die Standardabweichung hat dieselbe Dimension wie die gegebenen Daten, die Varianz aber nicht. Wenn beispielsweise die x_i in cm gemessen werden, dann hat s^2 die Dimension cm^2 , aber s hat wieder die Dimension cm.

c) Beispiele

1. Als einfache konkrete Anwendung einer der Formeln für s^2 betrachten wir die fünf Werte

$$x_1 = 5, x_2 = 6, x_3 = 6, x_4 = 8, x_5 = 10.$$

Nun stellen wir folgende Tabelle auf:

x_i	x_i^2
5	25
6	36
6	36
8	64
10	100
35	261

Wir lesen die Summen ab: $\sum_{i=1}^5 x_i = 35$, $\sum_{i=1}^5 x_i^2 = 261$ und finden mit $n = 5$ und unter Verwendung der letzten Formel für s^2 :

$$s^2 = \frac{1}{4} \left(\sum_{i=1}^5 x_i^2 - \frac{1}{5} \left(\sum_{i=1}^5 x_i \right)^2 \right) = \frac{1}{4} (261 - \frac{1}{5} \cdot 35^2) = \frac{1}{4} (261 - 245) = 4.$$

Es folgt, dass die Varianz $s^2 = 4$ und die Standardabweichung $s = 2$ ist. ☒

Der Vorteil der eben verwendeten Formel für s^2 ist der folgende: Wenn weitere Werte x_i hinzukommen oder wenn einer dieser Werte abgeändert werden muss, dann sind auch die Summen $\sum_{i=1}^n x_i$ und $\sum_{i=1}^n x_i^2$ schnell angepasst. Würde man dagegen mit der ursprünglichen Form von SS_{xx} , also mit $SS_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ arbeiten, so müssten neben \bar{x} auch alle n Differenzquadrate $(x_i - \bar{x})^2$ neu berechnet werden.

2. Führt man dieselbe Rechnung für das Beispiel 2.2.2.3.A der Küken durch, so erhält man $s = 6.3662$. \boxtimes

d) Bildung der Varianz bei zu Klassen gruppierten Daten

Wie in (2.2.3.3.d) gehen wir davon aus, dass die zu untersuchenden Werte in k Klassen aufgeteilt sind. Mit H_i bezeichnen wir die Anzahl der Messwerte in der i -ten Klasse, mit x_i die Klassenmitte dieser Klasse ($i = 1, \dots, k$). In Ermangelung genauerer Information ordnen wir jedem Objekt, das in der i -ten Klasse liegt, den Wert x_i zu, der somit H_i -mal angenommen wird. Dies führt zu folgenden Formeln:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k H_i (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^k H_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^k H_i x_i \right)^2 \right).$$

Dabei ist n die Gesamtzahl der Untersuchungsobjekte ($n = \sum_{i=1}^k H_i$); der Durchschnitt \bar{x} wird gemäss (2.2.3.3.d) bestimmt. Der Übergang von der ersten zur zweiten Formel geht ganz ähnlich wie die entsprechende, in (2.2.3.7.b) oben vorgenommene Umformung im Fall von SS_{xx} .

Als Beispiel verwenden wir die gruppierten Daten von (2.2.3.3) und stellen die folgende Tabelle auf:

x_i	H_i	$H_i x_i$	$H_i x_i^2$
88	2	176	15488
93	3	279	25947
98	10	980	96040
103	18	1854	190962
108	12	1296	139968
113	4	452	51076
118	1	118	13924
Summe	50	5155	533405

Setzen wir (mit $n = \sum_{i=1}^7 H_i = 50$) die Summen

$$\sum_{i=1}^7 H_i x_i = 5155 \quad \text{und} \quad \sum_{i=1}^7 H_i x_i^2 = 533405$$

in die obige Formel ein, so finden wir

$$s^2 = \frac{1}{49} (533405 - \frac{1}{50} \cdot 5155^2) = 39.2755.$$

Für die Standardabweichung s ergibt sich $s = 6.267$. Natürlich differiert dieser Wert etwas von dem in Beispiel 2. von (2.2.3.7.c) ermittelten.

Das obige Beispiel betraf ein stetiges Merkmal. Die Formel kann aber auch auf diskrete Merkmale angewandt werden, wenn wie am Schluss von (2.2.3.3.d) die Werte x_i mit ihren Häufigkeiten H_i gegeben sind.

e) Weitere Bemerkungen

- Viele Taschenrechner haben eingebaute Routinen, welche den Durchschnitt, die Standardabweichung und die Varianz berechnen (vgl. (2.2.3.3.b)). Dabei muss man allerdings kontrollieren, ob die Definition dieser Grössen mit der hier gegebenen übereinstimmt; der Unterschied liegt im Nenner, der bei uns gleich $n - 1$, an manchen Orten aber gleich n ist. Bei vielen Rechnern sind die entsprechenden Tasten mit $\boxed{\sigma_{n-1}}$ bzw. mit $\boxed{\sigma_n}$ oder ähnlichen Symbolen bezeichnet.
- In diesem Zusammenhang sei noch vorausblickend erwähnt, dass wir in der Wahrscheinlichkeitsrechnung ebenfalls eine "Varianz" definieren werden, die dort mit σ^2 bezeichnet wird (vgl. Kapitel 5). Ferner schreibt man μ für das Analogon zu \bar{x} . Unter gewissen Annahmen gilt dann die Formel

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 ,$$

wo im Nenner n und nicht $n - 1$ steht. Es gilt also, sich hier nicht verwirren zu lassen. Falls nötig, unterscheiden wir die *statistische* (oder *empirische*) Varianz bzw. Standardabweichung einerseits und die *wahrscheinlichkeitstheoretische* Varianz bzw. Standardabweichung andererseits.

- Eine weitere oft gebrauchte Grösse ist der *Standardfehler* $s_{\bar{x}}$. Die genauere Beschreibung verschieben wir auf die beurteilende Statistik in Kapitel 8 und erwähnen hier nur die Definition:

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

(2.2.3.8) Zur Bedeutung der Standardabweichung

Aus dem täglichen Leben haben wir eine gute Vorstellung davon, was unter dem Durchschnitt \bar{x} einer Anzahl von Daten zu verstehen ist. Eine derartige Vorstellung fehlt aber zumindest am Anfang für die Standardabweichung. Von der Definition her ist zwar klar, dass eine Verteilung mit einer grossen Standardabweichung breiter gestreut ist als eine solche mit einer kleinen. Diese Aussage ist aber eher qualitativ, und man möchte auch gerne eine zahlenmässige Interpretation haben. Dazu kann die folgende Faustregel verhelfen:

Wenn die Verteilung glockenförmig und symmetrisch ist, dann liegen

- im Bereich von $\bar{x} - s$ bis $\bar{x} + s$ ungefähr 68% aller Werte (also etwa 2/3),
- im Bereich von $\bar{x} - 2s$ bis $\bar{x} + 2s$ ungefähr 95% aller Werte.

Nehmen wir als Beispiel einmal mehr unsere Küken. Weiter oben wurden $\bar{x} = 102.96$ und $s = 6.36$ berechnet. Wir setzen gerundet

$$\bar{x} - s = 97, \quad \bar{x} + s = 109, \quad \bar{x} - 2s = 90, \quad \bar{x} + 2s = 116 .$$

Im Intervall $[97, 109]$ müssten nach der Faustregel 68% von 50, also 34 Messwerte liegen; Abzählen liefert in unserm Beispiel 37 Messwerte. Entsprechend erwarten wir im Intervall $[90, 116]$ 48 Werte, und in der Tat liegen genau so viele darin.

Die obigen Prozentangaben sind bis auf praktische Rundungen exakt, wenn die Verteilung eine Normalverteilung ist (vgl. (5.10.4) - (5.10.7)). Diese Voraussetzung trifft in vielen praktischen Fällen wenigstens näherungsweise zu (siehe auch in (7.2) "Zentraler Grenzwertsatz").

(2.∞) Aufgaben

2-1 Geben Sie bei den folgenden Merkmalen an, ob sie qualitativ, quantitativ und stetig bzw. quantitativ und diskret sind.

- a) Länge des Fusses eines Menschen, b) Schuhnummer dieser Person, c) Farbe des Schuhs, d) Preis des Schuhs, e) Gewicht des Schuhs, f) Form des Absatzes.

2-2 Zu welcher Skala gehören die folgenden Merkmale?

- a) Rückennummer eines Fussballspielers, b) Rang des Teams in der Meisterschaft, c) Effektive Dauer des Fussballspiels, d) Anzahl Zuschauer, e) Höhe des Fussballplatzes gemäss Landkarte, f) Länge des Fussballfeldes.

2-3 In der folgenden Geschichte kommen fettgedruckte Zahlen vor. Legen Sie in jedem Fall fest, zu welcher Skala (Nominal-, Ordinal-, Intervall- oder sogar Verhältnisskala) diese Zahl gehört. In unklaren Fällen ist Diskussion erwünscht.

Frau X., wohnhaft an der Rechnerstrasse **7** in **3141** Piwil, bestieg am **27.** Juli auf Gleis **14** einen Wagen **1.** Klasse des Zugs Nr. **708**, fahrplanmässige Abfahrtszeit **703**, der mit **4** Minuten Verspätung abfuhr. Bald danach erstand sie sich für Fr. **5.-** (inkl. Trinkgeld) an der Minibar eine Zwischenverpflegung (der Kaffee war ihr mit seinen **75°** zunächst zu heiss) und las im Heft Nr. **13** ihrer Lieblingszeitschrift. Die Zeit verging wie im Fluge, und ehe sie sich's versah, hatte der Zug die **129** Kilometer zwischen Zürich und Bern zurückgelegt.

2-4 Zwanzig Studentinnen und Studenten absolvierten eine Prüfung, in welcher maximal 12 Punkte erzielt werden konnten und erreichten dabei die folgenden Resultate:

6, 8, 3, 10, 1, 12, 3, 12, 7, 5, 10, 9, 6, 6, 11, 7, 6, 10, 9, 7.

- a) Zeichnen Sie das Stabdiagramm.
b) Stellen Sie die Summenhäufigkeitsverteilung graphisch dar.

2-5 Bei 24 Personen ergaben sich folgende, auf 0.5 cm genau gemessene, Körpergrössen:

182.0	176.0	191.0	173.5	183.0	178.0	179.0	174.5
184.5	174.0	180.5	174.5	178.5	169.0	178.5	168.5
179.0	177.0	172.5	171.0	176.5	167.0	175.0	175.0

- a) Zeichnen Sie das zugehörige Histogramm für eine Klassenbreite von 2 cm, beginnend mit der Klasse $(166.25, 168.25]$.
b) Stellen Sie die Summenhäufigkeitsverteilung bezüglich der in a) angegebenen Klasseneinteilung graphisch dar.

- c) Zeichnen Sie das Histogramm für eine Klassenbreite von 4 cm, beginnend mit der Klasse (164,25,168,25]. Ändern Sie den Massstab der Ordinate gegenüber a) gemäss (2.2.2.4).
 2–6 Bestimmen Sie das arithmetische Mittel und den Median der folgenden Messdaten:

2.5, 1.8, 3.0, 2.2, 2.9, 1.9, 2.3, 2.6.

- 2–7 Berechnen Sie zu den Daten der Aufgabe 2–4 a) Durchschnitt, b) Median, c) Modus, d) Interdezilbereich, e) Varianz, f) Standardabweichung.
 2–8 Gegeben sind die ganzzahligen Daten 20, 30, 22, 25, 23, 25, U , wobei U leider unleserlich ist. Welche Werte kommen für den Median dieser Daten in Frage?
 2–9 In einer Ortschaft wurde die Kinderzahl von Familien untersucht. Es fanden sich 120 Familien ohne Kinder, 200 mit einem Kind, 220 mit zwei, 150 mit drei, 40 mit vier, 6 mit fünf und eine Familie mit acht Kindern.
 a) Zeichnen Sie das zugehörige Stabdiagramm.
 b) Berechnen Sie Durchschnitt, Median und Varianz.
 2–10 Bei einer Wägung von 36 Hühnereiern wurden die folgenden Gewichtsklassen (Gewichte in Gramm) gebildet:

[58,61], (61,64], (64,67], (67,70], (70,73], (73,76], (76,79].

Auf diese Klassen entfielen der Reihe nach 4, 4, 8, 10, 7, 1, 2 Eier.

- a) Zeichnen Sie das Histogramm.
 b) Berechnen Sie Durchschnitt, Median und Varianz.
 2–11 Eine idealisierte Summenhäufigkeitsverteilung ist im Intervall $[0,2]$ durch die Funktion $F(x) = x - \frac{1}{4}x^2$ gegeben. Skizzieren Sie den Graphen und berechnen Sie den Median dieser Verteilung.
 2–12 Ein idealisiertes Histogramm eines stetigen Merkmals sei durch die Funktion

$$f : [0, 1] \rightarrow \mathbb{R}, f(x) = 4(x - x^3)$$

gegeben. Zeichnen Sie den Graphen und berechnen Sie den Median dieser Verteilung.

- 2–13 Die Zahlen x_1, x_2, \dots, x_n seien gegeben. Gesucht ist die Zahl x , für welche der Ausdruck $\sum_{i=1}^n (x_i - x)^2$ minimal wird.
 2–14 Zeichnen Sie den Graphen der Funktion $g(x) = \sum_{i=1}^n |x_i - x|$, und bestimmen Sie ihr Minimum
 a) für $x_1 = 1, x_2 = 2, x_3 = 4$ und b) für $x_1 = 1, x_2 = 2, x_3 = 4, x_4 = 6$. Suchen Sie einen Zusammenhang mit dem Median.