Grundlagen der Statistik

Dr. Sabine Lauer

Wintersemester 2005/2006

Inhaltsverzeichnis

1 Einführung **3**

1.1 Was ist Statistik? 3
1.2 Die statistische Untersuchung 5
1.3 Statistische Grundbegriffe 6
1.4 Merkmalstypen 7

2 Eindimensionale Häufigkeitsverteilungen **9**

2.1 Einzelhäufigkeiten 9
2.2 Summenhäufigkeiten 10
2.3 Empirische Verteilungsfunktion 12
2.4 Graphische Darstellung von eindimensionalen Häufigkeitsverteilungen 13

3 Kennzahlen eindimensionaler Häufigkeitsverteilungen **24**

3.1 Lageparameter 24
3.2 Streuungsparameter 29
3.3 Konzentrationsmessung 33

4 Zweidimensionale Häufigkeitsverteilungen **38**

4.1 Streuungsdiagramm und gemeinsame Verteilung 38
4.2 Randverteilungen 40
4.3 Bedingte Verteilung und statistische Unabhängigkeit 41
4.4 Statistischer Zusammenhang von mindestens ordinal skalierten Merkmalen . . 43
4.5 Statistischer Zusammenhang zwischen nominal skalierten Merkmalen 55

5 Wahrscheinlichkeitsrechnung und schließende Statistik **58**

5.1 Einführung in die Wahrscheinlichkeitsrechnung 58
5.2 Theoretische Wahrscheinlichkeitsverteilungen 63
5.3 Erwartungswert und Varianz von Zufallsvariablen 66
5.4 Schließende Statistik 72

6 Literaturverzeichnis **78**

1 Einführung

1.1 Was ist Statistik?

1.1.1 Einleitung

Statistik ist die Herrscherin über die Daten und ist zu einem Grundansatz geworden, die Wirklichkeit zu erfassen. In der Vielfalt und Variation der uns umgegebenden Erscheinungen hilft sie Gesetzmäßigkeiten zu erkennen und das Wesentliche herauszuarbeiten.

Schon beim Lesen einer normalen Tageszeitung werden wir mit einer Unmenge von Statistiken überhäuft, wie z.B.

- Im Jahr 2003 vernaschten die Deutschen 8.4 Liter (ca. 100 Kugeln) Eis pro Kopf. 52.4 Prozent davon wurden im Handel und nicht in der Eisdiele verkauft.
- 21 Prozent der Bundesbürger wünschen sich die Mauer zurück.
- 25 Prozent mehr Frauen als Männer biegen falsch in Einbahnstraßen ein.
- Ehemänner leben im Schnitt drei Jahre länger als unverheiratete Männer und verdienen im Durchschnitt jährlich 4500 Euro mehr.

Leider sind viele publizierte Statistiken falsch, sei es bewußt manipuliert oder nur unpassend ausgesucht. Deswegen ist ein grundlegendes Verständnis von Statistik so wichtig, um die Behauptungen solcher Veröffentlichungen kritisch hinterfragen zu können.

1.1.2 Definition der Statistik

Die Statistik hat zwei große Wege entwickelt, mit denen sie versucht aus Daten Rückschlüsse auf die Realität zu ziehen:

1. die beschreibende (deskriptive) Statistik
2. die schließende Statistik.

Die beschreibenden Statistik kümmert sich nur um die ihr vorliegenden Daten und macht darüberhinaus keine Aussagen (Beispiele: Ausfüllen eines Polizei-Unfallberichtes, Auswertung der Noten einer Klassenarbeit). Dazu steht ihr ein ganzes Repertoire an unterschiedlichen Methoden zur Verfügung, wie z.B.

- Darstellen in Tabellen
- Visualisieren in Diagrammen
- Verdichten zu Indexzahlen
- Aufschlüsseln nach Einflußfaktoren
- Analysieren von Zusammenhängen
- ...

Die schließende Statistik dagegen macht über die vorliegenden Daten hinausgehende Aussagen, indem sie von ihnen auf die Allgemeinheit schließt. Dies funktioniert nach dem Motto "Man muß nicht den ganzen Kuchen essen, um zu wissen, wie gut er schmeckt". Die Aussagen der schließenden Statistik sind zwangsläufig mit einer gewissen Ungenauigkeit behaftet, die allerdings mathematisch quantifizierbar ist. Wenn die untersuchte Stichprobe genügend groß und außerdem repräsentativ für das zu untersuchende Phänomen ist, dann sind die Aussagen im allgemeinen sehr genau (Beispiel: Wahlumfragen, bei denen anhand der Antworten der Befragten, auf das Stimmverhalten der gesamten Wahlbevölkerung geschlossen wird).

1.1.3 Daten und Wirklichkeit

Ein Datum ist alles, was konkrete Information über einen bestimmten Aspekt der Wirklichkeit liefert. Daten können dabei die unterschiedlichsten Formen annehmen, wie z.B. Worte, Symbole, Zahlen, Diagramme, Töne, Zeichen oder Bilder. Daten sind aber nicht die Wirklichkeit - sie geben nur bestimmte ausgewählte Aspekte der Wirklichkeit wider, und zwar solche, die als Daten erfaßbar sind ("Gesundheit" an sich kann z.B. statistisch nicht erfaßt werden, wohl aber Gesundheitsindikatoren wie z.B. Temperatur, Blutdruck oder Hämoglobinwert).

Außerdem sagen Daten nichts über ihren eigenen Wahrheitsgehalt aus, ob man Daten für "wahr" oder "falsch" hält, ist eine persönliche Entscheidung.

Daten ohne Angabe von Quellen sowie Daten mit politischer, wirtschaftlicher oder wissenschaftlicher Brisanz sind immer mit Vorsicht zu verwenden, ebenso wie Daten mit zu stimmigem Datenbild oder übergroßer Genauigkeit. Den Vertrauensvorschuß, den wir krummen vor glatten Zahlen geben, machen sich viele Publikationen zu nutze, indem sie mit Hife äußerst präziser Angaben versuchen, Korrektheit an Stellen vorzuspiegeln, an denen Aussagen in dieser Genauigkeit eigentlich nicht möglich sind. Zwei Beispiele hierzu:

1. Methusalem in der Bibel lebte nicht einfach ungewöhnlich lang, sondern wurde exakt 969 Jahre alt.
2. Laut Bild-Zeitung arbeitet die typisch deutsche Ehefrau pro Tag insgesamt eine Stunde, 50 Minuten und 13 Sekunden nur für ihren Mann (darunter 4 Minuten Hemden bügeln, 2 Minuten 30 Sekunden Bett machen, 1 Minute Barthaare aus dem Ausguß fischen und 15 Sekunden Klobrille schließen).

1.2 Die statistische Untersuchung

Eine statistische Untersuchung gliedert sich in fünf Phasen:

1. Planung
2. Erhebung
3. Aufbereitung
4. Analyse
5. Interpretation

Jede dieser Phasen wollen wir im folgenden einzeln etwas genauer betrachten, und jeweils am Beispiel einer klinischen Studie verdeutlichen.

1.2.1 Planung

Der Planungsphase einer statistischen Untersuchung wird zugerechnet:

- die genaue Ausarbeitung der Fragestellung
- die sachliche, räumliche und zeitliche Abgrenzung der Untersuchung
- die Klärung organisatorischer Fragen
- die exakte Auswahl der zu erhebenden Daten bzw. der sich daraus ergebenden statistischen Größen
- die Festlegung der zu verwendenden statistischen Verfahren

Bei einer klinischen Studie werden diese Punkte vor Beginn der Studie in einem sogenannten Studienprotokoll schriftlich fixiert: Was sind die primären und sekundären Ziele der Studie? In wie vielen Prüfzentren und welchen Ländern soll die Studie durchgeführt werden? Wie viele Patienten müssen in die Studie eingeschlossen werden, um die primäre Fragestellung beantworten zu können? Welche Daten sollen zu welchem Zeitpunkt erhoben werden, um den Krankheitsverlauf und etwaige Nebenwirkungen möglichst gut beschreiben zu können? Die statistischen Analysemethoden werden allerdings im Protokoll in der Regel nur grob skizziert, die detaillierte Festlegung erfolgt erst später im sogenannten statistischen Analyseplan.

Bei jeder statistischen Untersuchung bedarf insbesondere der genaue Wortlaut der Fragen besonderer Aufmerksamkeit. Bei vielen Umfragen sind die Antworten durch suggestiv gestellte Fragen geradezu vorprogrammiert: So lehnten z.B. nach einer Umfrage der IG Metall 95% aller bundesdeutschen Arbeiter das Arbeiten am Samstag ab, aber in einer zeitgleich von Marburger Marplan-Institut durchgeführten Umfrage waren 72% aller Beschäftigten grundsätzlich auch zum Arbeiten am Wochenende bereit.

1.2.2 Erhebung

In diesem Schritt wird das statische Datenmaterial gesammelt. Bei klinischen Studien erfolgt dies in der Regel über sogenannte Case report forms, in die der Prüfarzt alle relevanten Daten eines jeden Patienten einträgt. Teilweise werden die Daten auch schon vollelektronisch, z.B. mit Hilfe von PDAs (electronic patient diary) erfaßt.

1.2.3 Aufbereitung

Zur Aufbereitung des Datenmaterials gehört die Zusammenfassung der Daten in Tabellen, Schaubildern oder auch ganzen Datenbanken. Außerdem zählt dazu das sogenannte "Daten-Cleaning", bei dem Unstimmigkeiten zwischen den erhobenen Daten bereinigt werden. Bei klinischen Studien werden die Daten der Case report forms via zweifacher, unabhängiger Dateneingabe elektronisch erfaßt. Umstimmigkeiten werden dann programmgesteuert aufgedeckt und durch Rückfragen beim Prüfarzt geklärt.

1.2.4 Analyse

In der Analysephase wird das aufbereitete Datenmaterial mit Hilfe geeigneter statistischer Verfahren untersucht. Häufig sind dazu zunächst aus den vorhandenen Daten neue Variablen herzuleiten, aus denen dann Maßzahlen der beschreibenden oder auch schließenden Statistik errechnet werden. Heutzutage geschieht dies in der Regel unter Verwendung moderner Statistikprogramme wie R, SAS oder SPSS.

1.2.5 Interpretation

Anhand der Ergebnisse wird versucht, die ursprüngliche Fragestellung zu beantworten: sei es entweder nur rein deskriptiv die untersuchten Daten genauer zu beschreiben, oder aber zusätzlich mit Hilfe der Methoden der schließenden Statistik Rückschlüsse auf die Allgemeinheit zu ziehen. Bei klinischen Studien werden die Ergebnisse mit ihrer Interpretation im sogenannten Studienbericht veröffentlicht, der eine tragende Rolle bei der Zulassung neuer Medikamente durch die Behörden spielt.

1.3 Statistische Grundbegriffe

Im folgenden wollen wir einige grundlegende Begriffe einführen, die in der Statistik immer wieder auftauchen.

Merkmalsträger (statistische Einheit): Träger der interessierenden statistischen Information (z.B. den Wahlberechtigten im Politbarometer, oder den Patienten in einer klinischen Studie für ein neues Medikament gegen Herzrhythmusstörungen)

Grundgesamtheit (statistische Masse): Gesamtheit aller für die statistische Untersuchung relevanten Merkmalsträger (z.B. alle Wahlberechtigten im Politbarometer oder alle Patienten mit Herzrhythmusstörungen)

Stichprobe: Eine mit einer bestimmten Methode erzeugte Teilmenge der Grundgesamtheit. In der Statistik wird häufig mit Stichproben gearbeitet, da es vielfach extrem schwierig bis völlig unmöglich ist, die Grundgesamtheit zu untersuchen. Um zu gültigen Aussagen über die Grundgesamtheit zu kommen, muß die Stichprobe *repräsentativ* sein, d. h. sie muß in ihrer Zusammensetzung der Population möglichst stark ähneln.

Merkmal: Eigenschaft der Merkmalsträger, die bei der statistischen Untersuchung von Interesse ist (z.B. politische Partei oder Herzfrequenz)

Merkmalswert (Merkmalsausprägung): Wert, der bei der Datenerhebung festgestellt werden kann (z.B. CDU/FDP/Grüne/SPD oder 100 Herzschläge pro Minute)

Datenreihe (Messreihe): Menge aller Ergebnisse einer statistischen Untersuchung. Eine besondere Datenreihe ist die sogenannte *Urliste*, die die Daten in der Reihenfolge enthält, in der sie angefallen sind.

1.4 Merkmalstypen

Wir unterscheiden folgende Merkmalstypen:

Nominale Merkmale: Den Merkmalswerten können lediglich Namen zugeordnet werden, z.B. männlich/weiblich, manisch/depressiv/paranoisch. Nominale Merkmale können zwar auch mittels Zahlen codiert werden (z.B. durch die Festlegung 1 = männlich, 2 = weiblich), diese Zahlenwerte dienen dann aber lediglich der Unterscheidung; eine Mittelwertbildung macht z.B. keinen Sinn. Ein für die Praxis brauchbares Nominalsystem muß erschöpfend und eindeutig sein, d.h. es müssen alle theoretisch möglichen Ausprägungen angegeben werden und diese müssen sich außerdem gegenseitig ausschließen, damit keine Mehrfachzuordnungen erfolgen können (z.B. beim Familienstand: ledig/verheiratet/verwitwet/geschieden).

Ordinalmerkmale: Den Merkmalswerten können Namen zugeordnet werden, die in eine logische Reihenfolge gebracht werden können, z.B. Schulnoten (1-2-3-4-5-6) oder eine Gewichtsklassifizierung (dünn-normal-dick). Werden die Merkmalswerte von Ordinalmerkmalen durch Zahlen codiert, so sollte dies so geschehen, daß die logische Reihenfolge erhalten bleibt.

Intervallskalierte Merkmale: Während Ordinalmerkmale lediglich die Information liefern, daß Merkmal A besser als Merkmal B ist, verraten intervallskalierte Merkmale auch, wie groß dieser Unterschied ist. Bei intervallskalierten Merkmalen sind die Abstände zwischen zwei aufeinanderfolgenden Stufen gleich groß, so daß man Aussagen über das Ausmaß

des Unterschiedes zwischen zwei Daten machen kann. Typische Intervalldaten sind z.B. Temperaturangaben in Celsius.

Verhältnisskalierte Merkmale: Verhältnisskalierte Merkmale erhält man, wenn die Meßskala auf einen natürlichen Nullpunkt bezogen ist. Sie erlauben es nicht nur den Unterschied, sondern auch das zahlenmäßige Verhältnis zwischen Merkmalsausprägungen zu erfassen, z.B. bei Körpergröße und Monatseinkommen: A ist doppelt so groß wie B, C verdient dreimal soviel wie D.

Nominelle und ordinale Merkmale werden auch als *kategorielle* Merkmale bezeichnet.

Werden Merkmale nicht nach der Beziehung ihrer Ausprägungen untereinander, sondern nach der Anzahl ihrer möglichen Ausprägungen unterschieden, so differenziert man zwischen *stetigen* und *diskreten* Merkmalen. Merkmale sind diskret, wenn sie nur ganz bestimmte Werte annehmen können, sozusagen in Stufen auftreten. Stetige Merkmale können dagegen zwischen zwei Stufen jeden beliebigen Zwischenwerte annehmen, wenigstens prinzipiell.

Aus praktischen Gründen werden in der Statistik häufig stetige Merkmals wie diskrete behandelt und umgekehrt:

- Angabe in Fragebögen wie z.B. “Alter in Jahren” oder “Körpergröße in cm” machen aus einem stetigen Merkmal ein diskretes (*Diskretisierung*)
- Diskrete Merkmale, die sehr viele Werte annehmen können, werden oft wie stetige Merkmale behandelt (z.B. Preise in Euro mit 2 Nachkommastellen). Solche Merkmale heißen auch *quasistetig*.

2 Eindimensionale Häufigkeitsverteilungen

2.1 Einzelhäufigkeiten

Bei einer statistischen Erhebung werde an n Merkmalsträgern ein Merkmal X beobachtet. Die resultierenden Merkmalswerte seien Zahlen, die in der Urliste wie folgt vorliegen:

$$x_1, x_2, ..., x_n.$$

Diese Darstellung enthält zwar die gesamte Information für den Statistiker, aber in der Regel ist sie selbst für kleine n unübersichtlich.

Beispiel 2.1.1: Bei einer Untersuchung wurde der Umfang des Wortschatzes von $n = 25$ Kleinkindern im Alter von ca. 13 Monaten ermittelt. Die Anzahl der verschiedenen Wörter ist in der folgenden Urliste aufgeführt:

$$0,2,0,0,4,2,4,2,1,3,6,2,2,0,0,0,1,3,2,1,4,1,0,3,0.$$

Somit bezeichnen $x_1 = 0$, $x_2 = 2$, $x_3 = 0$, ..., $x_{25} = 0$. In der Urliste gibt es $k = 6$ verschiedene Ausprägungen der Merkmalswerte, nämlich $y_1, y_2, ..., y_k$:

$$y_1 = 0, y_2 = 1, y_3 = 2, y_4 = 3, y_5 = 4, y_6 = 6.$$

Wir wollen im folgenden immer annehmen, daß die verschiedenen Ausprägungen $y_1, y_2, ..., y_k$ so angeordnet seien, daß

$$y_1 < y_2 < ... < y_k.$$

Die Ausprägung $y_1 = 0$ kommt mit einer absoluten Häufigkeit von $h_1 = 8$ bzw. mit einer relativen Häufigkeit von $f_1 = \frac{8}{25}$ in der Urliste vor. Allgemein definieren wir für $i = 1, 2, ..., k$ wie folgt :

$$\begin{aligned} \text{Absolute Häufigkeit } h_i &:= \text{Anzahl der Merkmalswerte mit Ausprägung } y_i \text{ in der Urliste} \\ \text{Relative Häufigkeit } f_i &:= \frac{h_i}{n} = \frac{\text{Absolute Häufigkeit}}{\text{Gesamtanzahl der Merkmalsträger}} \end{aligned}$$

Damit gilt $0 \leq h_i \leq 1$, für alle $i = 1, 2, ..., k$ und:

$$\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{h_i}{n} = \frac{1}{n} \sum_{i=1}^{k} h_i = \frac{1}{n} n = 1$$

Bemerkung: Werden in Publikationen nur relative Häufigkeiten ohne die zugehörigen absoluten Häufigkeiten aufgeführt, so ist in der Regel Vorsicht angebracht, da sich kleine Zahlen trefflich

hinter imponierenden relativen Häufigkeiten verstecken lassen (man denke nur an die berühmten x Prozent der Testpersonen in Werbebotschaften...).

Nimmt das untersuchte Merkmal sehr viele verschiedene Werte an (im Extremfall können sogar alle Merkmalswerte paarweise verschieden sein), so macht die Bildung von relativen und absoluten Häufigkeiten nach der oben beschriebenen Methode wenig Sinn. Hier bietet es sich an, die verschiedenen Merkmalsausprägungen zunächst in k Klassen zusammenzufassen, um erst dann die absoluten bzw. relativen Häufigkeiten pro Klasse (*Klassenhäufigkeiten*) zu ermitteln. Analog zur oben eingeführten Notation bei nicht-klassifizierten Merkmalen bezeichnen wir die absolute bzw. relative Häufigkeit der i-ten Klasse K_i mit h_i bzw. f_i. Die Klassen benötigen hierbei eindeutige Grenzen und als Faustregel sollte die Anzahl der Klassen, je nach Merkmalstyp, zwischen 4 und 20 liegen. Aus Gründen der Übersichtlichkeit sollten die meisten Klassen auch möglichst eine identische Breite haben. Zur Verdeutlichung betrachten wir ein Beispiel:

Beispiel 2.1.2: Adrenalin ist ein Streßhormon, das vom sympathischen Nervensystem und von der Nebenniere ausgeschüttet wird und den Körper auf Angriff präpariert. In einer Studie soll untersucht werden, ob Rauchen den Adrenalinspiegel verändert. Dazu wurde an einer Fachhochschule ein Anschlag an das schwarze Brett gemacht, auf den sich 19 Nichtraucher- und 29 Raucher-Studenten meldeten. Unser Beispieldatensatz umfaßt die Adrenalinwerte der 19 Nichtraucher-Studenten:

Adrenalinwerte in μg /24 Std. (Urin)				
7	8	39	33	24
19	16	10	12	18
15	33	14	28	37
25	41	58	19	

Eine mögliche Klasseneinteilung zeigt die folgende Tabelle:

Klasse K_i	Adrenalingehalt	Klassenmitte x_i^*	Klassenbreite ΔK_i	Abs. Klassenhäufigkeit h_i	Rel. Klassenhäufigkeit f_i
K_1	0 bis unter 10	5	10	2	0.105
K_2	10 bis unter 20	15	10	8	0.421
K_3	20 bis unter 30	25	10	3	0.158
K_4	30 bis unter 40	35	10	4	0.211
K_5	40 bis unter 50	45	10	1	0.053
K_6	50 bis unter 60	55	10	1	0.053

2.2 Summenhäufigkeiten

Häufig ist auch die Anzahl aller Merkmalsausprägungen unterhalb eines gewissen Wertes von Interesse, z.B. die Anzahl der Kleinkinder mit einem Wortschatz von bis zu 2 Wörtern in Beispiel 2.1.1. Dies führt zur Definition von absoluten bzw. relativen Summenhäufigkeiten (auch *kumulative* Häufigkeiten genannt):

$$\begin{aligned}
\text{Absolute Summenhäufigkeit } H_j &:= \sum_{i=1}^{j} h_i \\
&= \text{Anzahl der Merkmalswerte mit Ausprägung } y_i \leq y_j \\
\text{Relative Summenhäufigkeit } F_j &:= \sum_{i=1}^{j} f_i \\
&= \frac{\text{absolute Summenhäufigkeit}}{\text{Gesamtanzahl der Merkmalsträger}}
\end{aligned}$$

So gilt in Beispiel 2.1.1:

$$\begin{aligned}
H_3 &= h_1 + h_2 + h_3 = 8 + 4 + 6 = 18 \\
\text{und} \quad F_3 &= \frac{H_3}{25} = \frac{18}{25} = 72\%
\end{aligned}$$

Demnach verfügen 18 von den 25 Kleinkindern bzw. 72% über ein Vokabular von höchstens 2 Wörtern.

Alle absoluten und relativen Einzel- sowie Summenhäufigkeiten von Beispiel 2.1.1 sind übersichtlich in der folgenden Tabelle zusammengefaßt:

Index i	Anzahl der Merkmalswerte mit Ausprägung y_i	Absolute Häufigkeit h_i	Relative Häufigkeit f_i	Abs. Summenhäufigkeit H_i	Rel. Summenhäufigkeit F_i
1	0	8	0.32	8	0.32
2	1	4	0.16	12	0.48
3	2	6	0.24	18	0.72
4	3	3	0.12	21	0.84
5	4	3	0.12	24	0.96
4	6	1	0.04	25	1

Diese Häufigkeitstabelle beschreibt die *(eindimensionale) Häufigkeitsverteilung* des quantitativ-diskreten Merkmals "Anzahl der Wörter im Wortschatz pro Kleinkind". Allgemein bezeichnet eine (eindimensionale) Häufigkeitsverteilung die Zuordnung von absoluten oder relativen Einzel- und Summenhäufigkeiten zu den verschiedenen Ausprägungen $y_1, y_2, \ldots, y_k$ aller Merkmalswerte $x_1, x_2, \ldots, x_n$.

Summenhäufigkeiten von klassifizierten Merkmalen werden ganz analog gebildet. Für Beispiel 2.1.2 errechnen sich die Summenhäufigkeiten wie folgt:

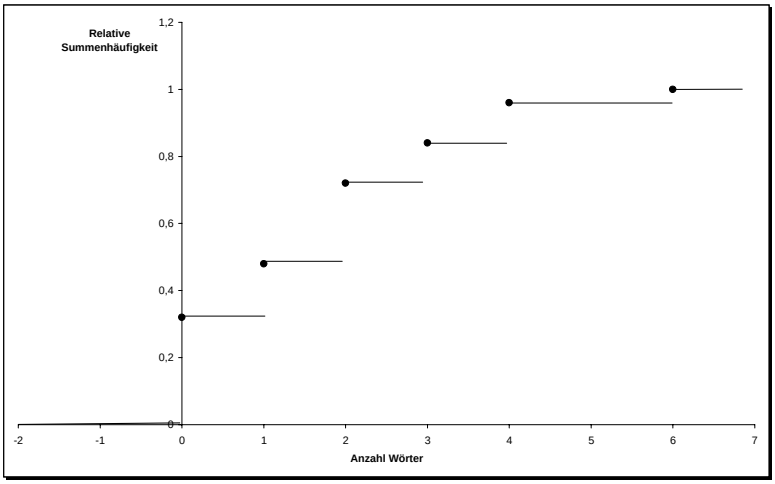
Klasse K_i	Adrenalingehalt	Abs. Summenhäufigkeit H_i	Rel. Summenhäufigkeit F_i
K_1	0 bis unter 10	2	0.105
K_2	10 bis unter 20	10	0.526
K_3	20 bis unter 30	13	0.684
K_4	30 bis unter 40	17	0.895
K_5	40 bis unter 50	18	0.947
K_6	50 bis unter 60	19	1

2.3 Empirische Verteilungsfunktion

Einen genauen Überblick über die Summenhäufigkeiten einer Datenreihe vom Umfang n liefert die *empirische Verteilungsfunktion* bzw. *Summenhäufigkeitsfunktion* F, die wie folgt definiert ist:

$$F : \mathbb{R} \mapsto \mathbb{R}, \quad x \mapsto F(x) := \frac{\text{Anzahl der Merkmalsträger mit Merkmalsausprägung} \leq x}{n}$$

Für Beispiel 2.1.1 stellt sich die empirische Verteilungsfunktion F als Treppenfunktion wie folgt dar:



Offensichtlich gilt für alle $x, y \in R$:

$$0 \leq F(x) \leq 1 \quad \text{und} \quad x \leq y \;\rightarrow\; F(x) \leq F(y),$$

d.h. die empirische Verteilungsfunktion nimmt Werte zwischen 0 und 1 an und ist monoton wachsend.

Wie sich die empirische Verteilungsfunktion bei klassifzierten Daten errechnet, werden wir in Abschnitt 2.4.2 vorstellen.

2.4 Graphische Darstellung von eindimensionalen Häufigkeitsverteilungen

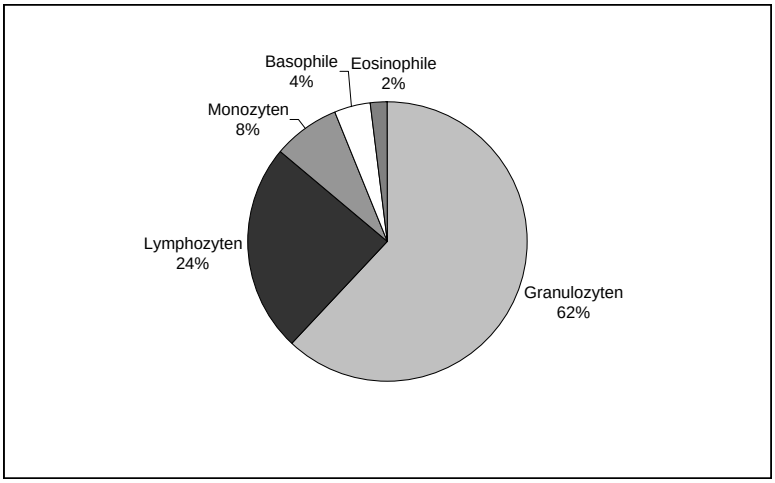
Graphische Symbole werden auf der ganzen Welt zur schnelleren Orientierung und Kommunikation benutzt, man denke nur an die Verkehrszeichen. In kürzester Form geben sie auf einen Blick einen komplizierten Sachverhalt wieder. Deswegen erstaunt es nicht, daß angesichts der überquellenden Datenfülle die visuelle Darstellung auch in der heutigen Statistik einen festen Platz erworben hat. Sie erlaubt eine schnelle Orientierung über das Wesentliche, indem sie die Datengestalt auf einen Blick wiedergibt.

2.4.1 Diagramme für einfache Zahlenwerte

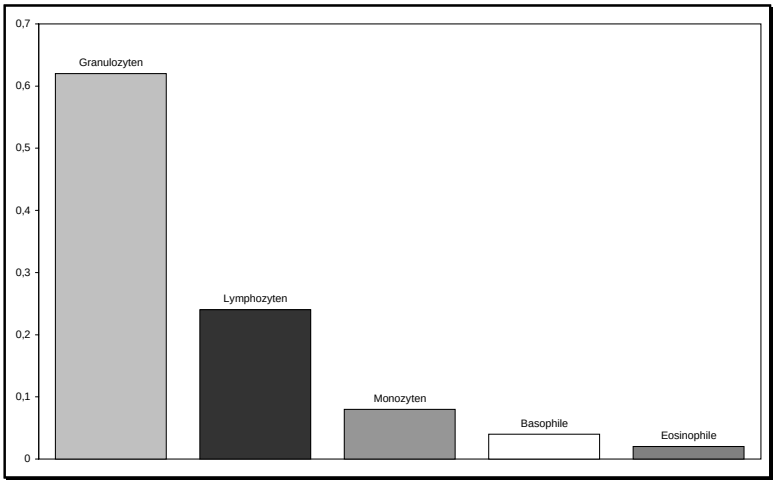
Im folgenden werden anhand eines einfachen Datensatzes aus dem klinischen Alltag (Blutbild) verschiedene Möglichkeiten der graphischen Darstellung erläutert. Die Rohdaten mit dem prozentualen Anteil der verschiedenen Blutzellen am weißen Blutbild seien wie folgt vorgegeben: Granulozyten 62%, Lymphozyten 24%, Monozyten 8%, Basophile 4%, Eosinophile 2%.

Diese Zahlenwerte werden nun nacheinander durch verschiedene Diagrammtypen dargestellt:

Kreisdiagramm: Die Zahlenwerte werden durch Flächenabschnitte eines Kreises repräsentiert. Das Kreisdiagramm eignet sich besonders zur Darstellung des Anteils von Teilkomponenten an einem Ganzen.

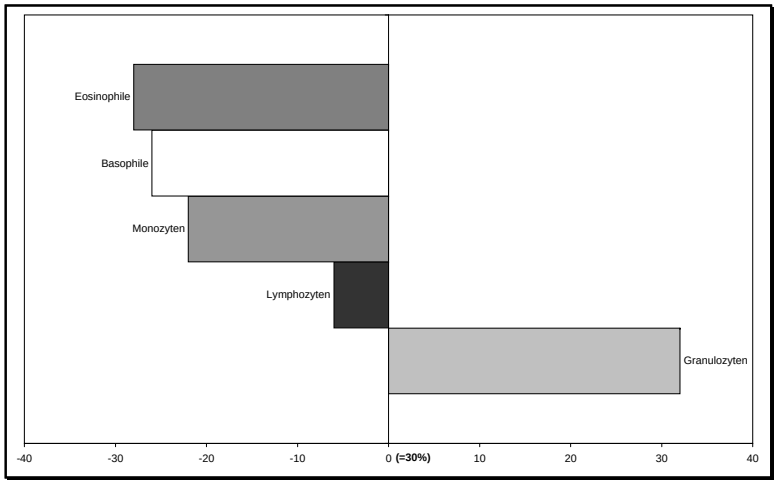


Stabdiagramm: Die Zahlenwerte werden durch die Länge von Stäben repräsentiert, die Breite der Stäbe ist ohne Bedeutung. Das Stabdiagramm wird gerne für den anschaulichen Vergleich von Einzelgrößen verwendet.

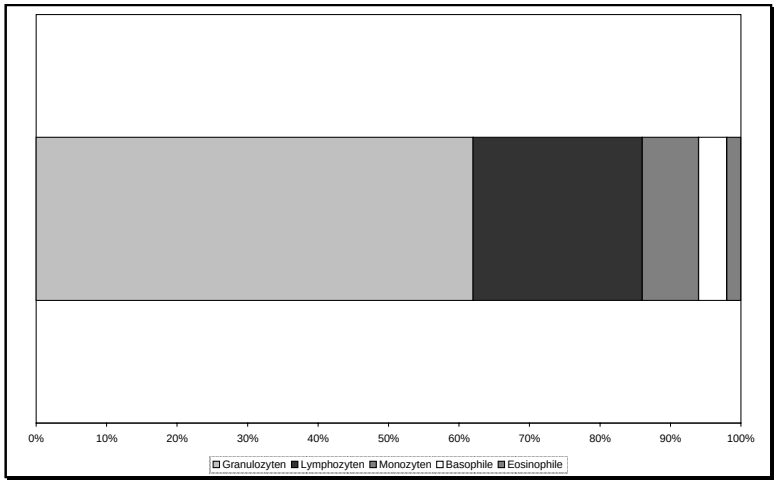


2-Richtungs-Stabdiagramm: Wieder werden die Zahlenwerte durch die Länge von Stäben repräsentiert, die Skala ist allerdings zweigeteilt und geht von einem Bezugswert aus in

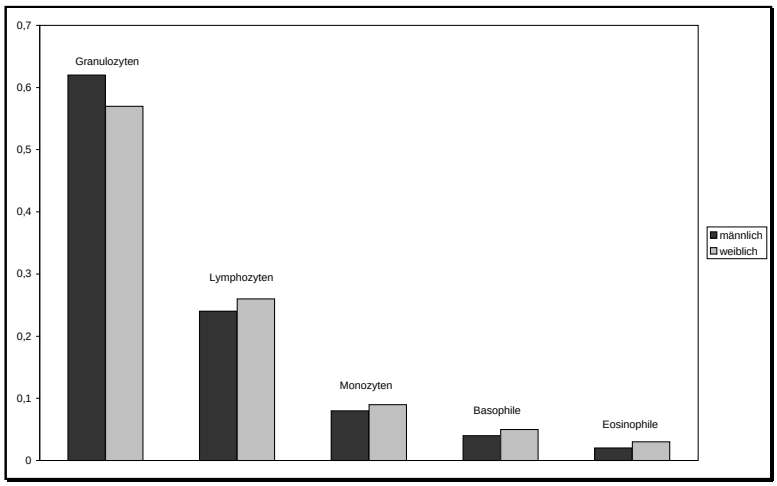
zwei Richtungen. Diese Darstellung eignet sich gut zum Vergleich zum Zahlenwerten mit einem Fixwert.



Komponenten-Stabdiagramm: Die Zahlenwerte werden durch Längsabschnitte eines Stabes dargestellt, die Breite des Stabes ist wieder ohne Bedeutung. Die Darstellung ist besonders günstig, wenn mehrere Komponentenstäbe miteinander verglichen werden sollen.



Gruppen-Stabdiagramm: Das Gruppen-Stabdiagramm ist eine Ausweitung des einfachen Stabdiagramm-Prinzips von einer auf mehrere Datengruppen; mehrere Datensätze werden einander gegenübergestellt.



2.4.2 Diagramme für Zahlenreihen

Handelt es sich nicht um Einzelwerte, die dargestellt werden sollen, sondern um ganze Zahlenreihen, bei denen ein- und dasselbe Merkmal an verschiedenen Objekten gemessen wurde, so werden die Einfachdiagramme schnell unpraktikabel: Wurde z.B. bei 1000 Schulkindern der Blutdruck gemessen, so ist es sehr mühsam im Stabdiagramm 1000 Stäbe nebeneinander zu malen. Hier existieren besondere Diagrammtypen, die übersichtlich zeigen, wie sich die Einzelwerte einer Datenreihe über den gemessenen Bereich verteilen. Wir verdeutlichen die verschiedenen Diagrammtypen anhand von Beispiel 2.1.2.

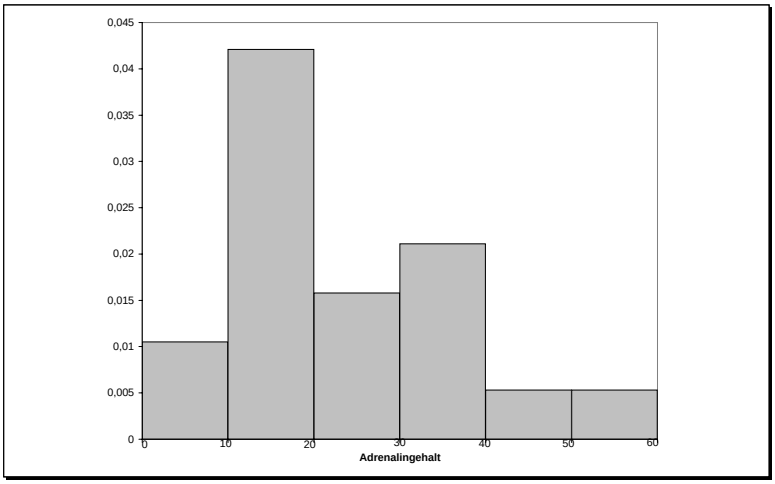
Stamm-Blatt-Diagramm: Bei dieser Darstellung werden die 10-er Stellen als "Stamm" eines Datenbaumes vorne hingeschrieben. Rechts vom Längsstrich folgen die Einerstellen als "Blätter", der Größe nach geordnet. Sämtliche Daten sind noch in Rohform vorhanden, zusätzlich kann man jedoch aus der Länge der einzelnen Zeilen gestalthaft die Verteilung der Daten entnehmen.

0	7 8
1	0 2 4 5 6 8 9 9
2	4 5 8
3	3 3 7 9
4	1
5	8

Histogramm: Das Histogramm ist wohl die bekannteste graphische Darstellung von Zahlenreihen. Es beruht auf der Zusammenfassung der verschiedenen Merkmalsausprägungen in k Klassen K_i. Das zugehörige Histogramm besteht dann aus k Rechtecken der Breite ΔK_i und der Höhe d_i, wobei

$$d_i = \frac{\text{relative Häufigkeit der i-ten Klasse}}{\text{Breite der i-ten Klasse}} = \frac{f_i}{\Delta K_i}.$$

Für Beispiel 2.1.2 stellt sich das Histogramm also wie folgt dar:



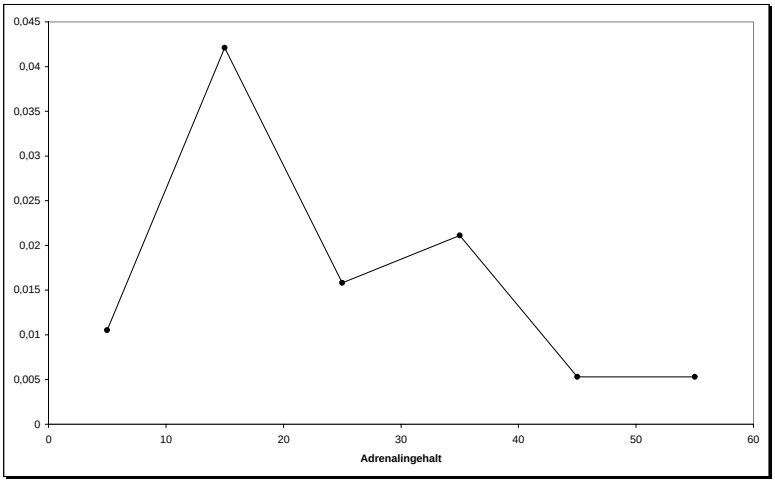
Demnach gilt:

1. Die Fläche jedes Rechtecks ist gleich der relativen Klassenhäufigkeit.

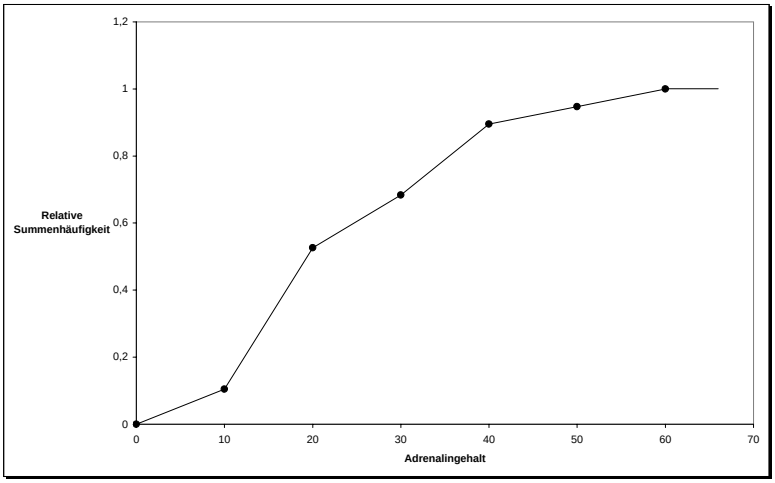
2. Die Gesamtfläche aller Rechtecke eines Histogramms ist gleich 1.

3. Ist die Klasseneinteilung äquidistant (gilt also $\Delta K_1 = \Delta K_2 = ... = \Delta K_k$), so sind die Rechteckshöhen proportional zu den relativen Häufigkeiten.

Frequenzpolygon: Verbindet man die Mittelpunkte der Flächenoberkanten aller Rechtecke eines Histogramms, so erhält man das Frequenzpolygon.



Summenhäufigkeiten und Verteilungsfunktion: Zur graphischen Darstellung der Summenhäufigkeiten von klassifizierten Merkmalen trägt man in der Horizontalen wie beim Histogramm und beim Frequenzpolygon die Klassengrenzen auf, in der Vertikalen werden die kumulativen Häufigkeiten aufgetragen und zwar jeweils über der rechten Klassengrenze (da die Summenhäufigkeit alle Werte bis zu dieser mit einbezieht). Unter der Annahme, daß die Merkmalswerte gleichmäßig über jede Klasse verteilt sind, werden die Punkte hier gradlinig verbunden. Die zugehörige Verteilungsfunktion $F(x)$ gibt wieder an, wie viele Merkmalsträger einen Wert haben, der höchstens so groß wie x ist. Durch die Klassifizierung der Daten entstehen dabei natürlich Ungenauigkeiten im Vergleich zur Urliste. Für Beispiel 2.1.2 ergibt sich folgendes Bild:



Allgemein stellt sich die Verteilungsfunktion $F(x)$ wie folgt dar:

x liege in der Klasse K_i mit unterer Grenze ${x_u}^i$ und oberer Grenze ${x_o}^i$ und ${F_u}^i$ bzw. ${F_o}^i$ bezeichne die kumulierte Klassenhäufigkeit an der Stelle ${x_u}^i$ bzw. ${x_o}^i$, dann gilt:

$$F(x) \quad = \quad {F_u}^i + \frac{{F_o}^i - {F_u}^i}{{x_o}^i - {x_u}^i} \left(x - {x_u}^i\right) \ . \tag{2.1}$$

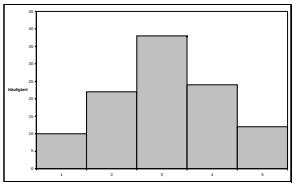
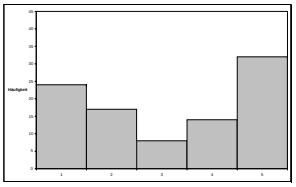
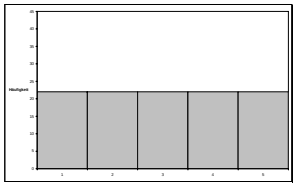
2.4.3 Visuelle Analyse

Im folgenden sollen die Möglichkeiten der visuellen Analyse von eindimensionalen Häufigkeitsverteilungen anhand der Histogramm-Darstellung veranschaulicht werden. Prinzipiell käme auch jeder andere der beschriebenen Darstellungstypen in Frage; das Histogramm bietet sich allerdings besonders an, da aus historischen Gründen ein großer Teil des Beschreibungsvokabulars aus dieser Darstellungsart abgeleitet ist. Im folgenden sei ein systematischen Ansatz zur visuellen Analyse des Histogramms beschrieben.

2.4.3.1 Anzahl der Gipfel

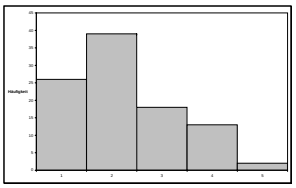
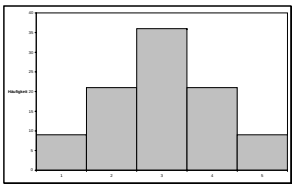
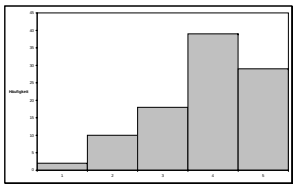
Betrachtet man das Histogramm einer Datenreihe, so ist es zweckmäßig zunächst darauf zu achten, ob die Daten mehr oder weniger gleichmäßig über den Wertebereich verteilt sind oder ob sie irgendwo gehäuft auftreten und dadurch eine Art Gipfel bilden. Die folgenden drei Bilder zeigen drei Grundtypen von Daten-Verteilungen. Das rechte Beispiel veranschaulicht eine Verteilung

der Daten, bei der kein Gipfel zu erkennen ist. Eine solche Datenanordnung wird *gleichförmig* genannt. Sie begegnet einem vor allem in der Theorie der statistischen Spielmodelle; z.B. sind die Gewinnzahlen im Lotto oder im Roulette auf diese Art und Weise verteilt: Jede Zahl kommt im Prinzip gleich häufig vor. Im linken Beispiel häufen sich die Daten an einer Stelle zu einem Gipfel; man spricht in diesem Fall von einer *unimodalen* Verteilung der Werte. Im mittleren Bild dagegen häufen sich die Daten an zwei Stellen des Wertebereich, es handelt sich also um eine Verteilung mit zwei Gipfeln, die auch *bimodal* genannt wird. Bimodale (oder auch *multimodale* Verteilungen mit mehr als zwei Gipfeln) kommen in der Praxis seltener vor als unimodale. Häufig entstehen bimodale Verteilungen dann, wenn zwei deutlich unterschiedliche Datenreihen zusammengeworfen werden (z.B. Körpergröße von Jungen und Mädchen in einer Schulklasse).

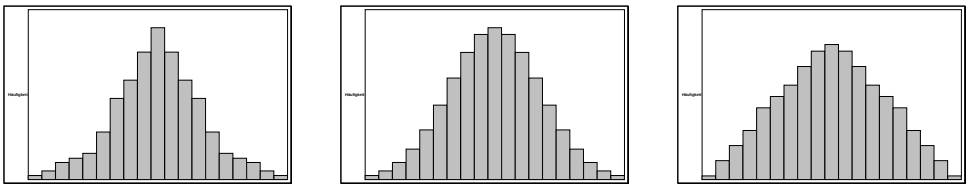
2.4.3.2 Symmetrische vs. schiefe Verteilung

In den folgenden drei Bildern wird der Begriff Schiefe illustriert. Bei schiefen Verteilungen ist der Gipfel nach einer Seite hin verschoben, so daß das Bild unsymmetrisch wird. Ist der Gipfel nach links verschoben, so spricht man von *positiver Schiefe* (linkes Bild), ist der Gipfel nach rechts verschoben, so spricht man von *negativer Schiefe* (rechtes Bild). Das mittlere Diagramm dagegen zeigt eine symmetrische Verteilung, bei der sich die beiden Hälften rechts und links des Gipfels spiegelbildlich zueinander verhalten. Schiefe Verteilungen treten unter anderem auf, wenn sich die Daten nicht "frei" über den den ganzen Wertebereich verteilen können, weil eine natürliche Grenze auf einer der beiden Seiten existiert. So liegt z.B. der Häufigkeitsgipfel für Triglyceride (Fettwerte im Blut) bei ca. 85 mg pro 100 ml Blut. Nach oben hin können ziemlich hohe Werte auftreten, in Einzelfällen bis weit über 1000 mg; nach unten existiert jedoch eine Grenze bei bereits 30 mg, unterhalb derer der Organismus nicht mehr lebensfähig ist. Durch den kurzen Abstand des Gipfels zur linken Grenze erscheint der Gipfel verzogen.

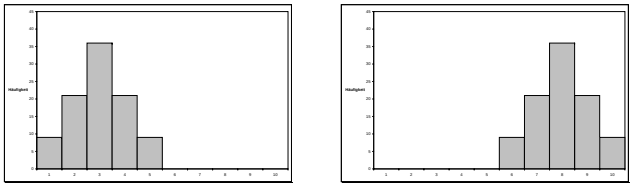
2.4.3.3 Normalverteilung

Der "Bilderbuchfall" einer Verteilung von Daten ist die sogenannte *Normalverteilung*. Die Normalverteilung ist eine unimodale, symmetrische Verteilung, die als Idealfall bei vielen natürlichen Phänomenen vorkommt und deswegen in der Praxis von außerordentlicher Bedeutung ist. Sie zeichnet sich durch ihr charakteristisches Bild aus, das die Gestalt einer Glocke hat (*Gauß'sche Glockenkurve*[1]). Zur besseren Charakterisierung werden dieser Verteilung in den folgenden drei Bildern verwandte Formen gegenübergestellt: Nur das mittlere Beispiel zeigt eine Normalverteilung mit dem charakteristischen Glockenprofil; die linke und die rechte Verteilung sind zwar auch unimodal und symmetrisch, die linke ist allerdings oben zu stark eingebuchtet und die rechte ist oben zu dick. Eine etwas genauere numerische Charakterisierung normalverteilter Daten werden wir in Abschnitt 3.2.4 kennenlernen.



2.4.3.4 Lage des Datenzentrums

Das wichtigste spezielle Merkmal einer Datenreihe ist die Lage des Datenzentrums, das bei sonst gleichem Verteilungstyp das wichtigste Unterscheidungsmerkmal von Datenreihen darstellt. An ihm läßt sich ablesen, an welcher Stelle des Wertebereichs die meisten Daten vorkommen und ob sich die meisten Daten eher bei niedrigen oder bei hohen Werten häufen

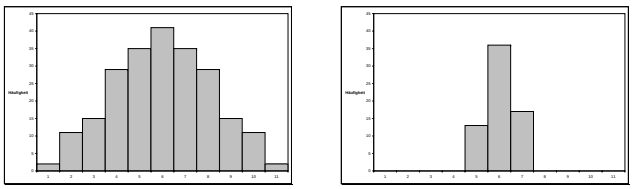


2.4.3.5 Streuungsgrad

Wenn man Sand aus einem Trichter ausschüttet, dann fallen die Sandkörner nicht etwa zu einem "Stab" exakt übereinander, sondern verteilen sich rechts und links vom Aufprallzentrum in der

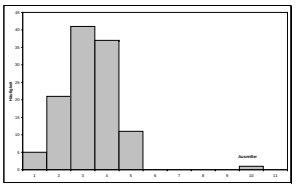
[1] Carl Friedrich GAUSS, 1777-1855, deutscher Naturwissenschaftler und Mathematiker, Professor an der Universität Göttingen. Gauß war ein Universalgenie mit wegweisenden Arbeiten auf vielen verschiedenen Gebieten der Mathematik, Astronomie und Physik.

Form eines “Haufens”. Ähnlich ist es mit Daten; die allgemeine Variation sorgt dafür, daß die Daten um das Zentrum herum streuen. Je größer die Variation ist, umso breiter fällt die Streuung auf der Meßskala aus. Im untenstehenden Bild ist links eine Verteilung mit breiter Streuung, rechts eine Verteilung mit schmaler Streuung dargestellt. Neben der Lage des Zentrums kann auch die Streuung bei statistischen Untersuchungen von besonderen Interesse sein: So äußert sich z.B. die Wirkung von Kaffee auf die Reaktionszeit des Menschen vor allem in einer größeren Streuung der Werte, weniger aber in einer Verschiebung des Zentrums.



2.4.3.6 Ausreißer

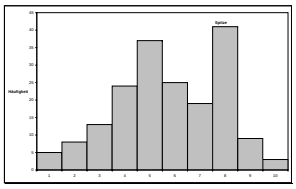
In der Praxis sind so homogene Verteilungen wie in den obigen Beispielen eher die Ausnahme. Besonders an den Rändern sind die Verteilungen häufig “unschön”. Der Grund liegt in einzelnen Extremwerten, die Ausreißer genannt werden. Es kann sich dabei um grobe Beobachtungs- bzw. Meßfehler handeln, allerdings auch um biologische Extravaganzen, die das sonst stimmige Bild stören. Ausreißerwerte sollten als atypische Daten gewertet werden, jedoch nicht einfach in der Analyse unter den Tisch fallen gelassen werden: Nicht selten sind gerade sie Anlaß zu neuen interessanten wissenschaftlichen Fragestellungen.

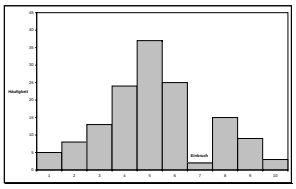


2.4.3.7 Einbrüche oder Spitzen

Im linken Bild ist ein relativ ebenmäßiger Kurvenverlauf mit einer isolierten Spitze zu sehen, die man vielleicht auch als Gipfel bezeichnen könnte, wenn sie nicht so schmal wäre. Das rechte Bild zeigt einen Einbruch. Spitzen und Einbrüche, also ein isoliertes Emporsteigen oder Wegsacken der Häufigkeit an einer Stelle sind im Gegensatz zu den Ausreißerwerten meistens auf Besonderheiten der Daten zurückzuführen und werden seltener durch Meßfehler hervorgerufen. Als Beispiel sei die Messung von Harnsäure aufgeführt: Mißt man bei einer Gruppe von

Versuchspersonen den Harnsäurespiegel im Blut, so findet man auf den ersten Blick näherungsweise eine Normalverteilung mit einem Datenzentrum bei 5 mg Harnsäure pro 100 ml Blut vor. Bei genauerer Analyse fällt eine schmale isolierte Spitze bei 6.4 mg auf, die das ansonsten harmonische Bild stört. Geht man der Ursache nach, so stößt man auf einen interessanten Zusammenhang: Harnsäure ist nur bis zu einer Konzentration von 6.4 mg im Blut löslich. Bei höheren Konzentrationen bildet sie Kristalle, die beim Menschen zu Gichtanfällen führen können. Der Körper kennt offensichtlich diese Löslichkeitsschwelle und versucht den Harnsäureanstieg an dieser Stelle eine Zeitlang aufzufangen. Dadurch häufen sich hier die Werte und bilden die beobachtete isolierte Spitze.





3 Kennzahlen eindimensionaler Häufigkeitsverteilungen

Nach der visuellen Analyse wenden wir uns nun einer weiteren Möglichkeit der Datenanalyse zu: der numerischen Analyse. Durch die Berechnung einer Handvoll Zahlen, der *statistischen Kennwerte*, soll die Datenreihe so gut wie möglich beschrieben werden. Da allerdings in diesen Werten die ursprünglichen Daten nicht mehr sichtbar sind, ist es wichtig, ihr Zustandekommen und ihren jeweiligen Anwendungsbereich gut zu kennen. Denn nicht selten wird insbesondere in der Werbung mit sogenannten "Kennwerten" herumjongliert, deren Bezug zu den Daten mehr als fragwürdig ist.

3.1 Lageparameter

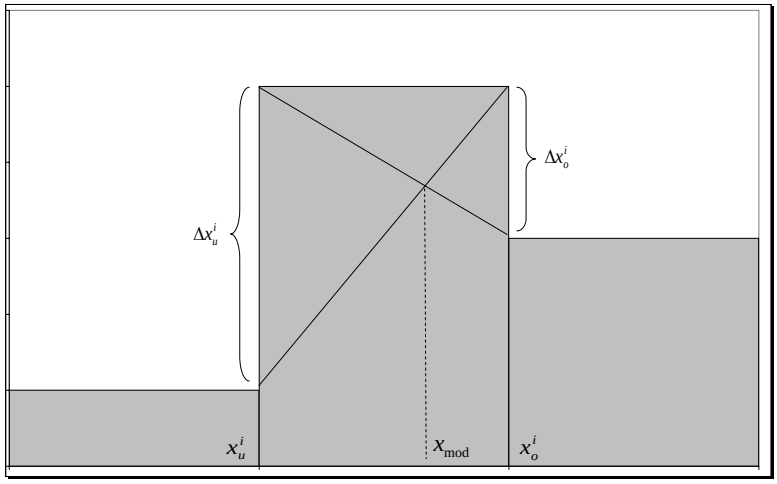
Die Urlisten bzw. Häufigkeitsverteilungen sollen durch einzelne Zahlen bzw. Parameter charakterisiert werden. Der Lageparameter soll möglichst gut beschreiben, **wo** sich das gesamte Datenmaterial auf der Merkmalsachse befindet.

3.1.1 Der Modalwert (Modus)

Im Fall nicht-klassierter Daten ist der Modalwert x_{mod} derjenige Merkmalswert, der am häufigsten beobachtet wurde. Es kann mehrere Werte mit dieser Eigenschaft geben. Einschränkungen für eine sinnvolle Verwendung des Modalwertes als Lageparameter sind:

- Die Verteilung sollte eingipfelig sein
- In der Umgebung des Modalwertes sollte eine erkennbare Konzentration der Merkmalswerte vorliegen.

Im Fall klassierter Daten befindet sich der Modalwert x_{mod} in der Klasse K_i mit der höchsten Histogrammhöhe. Dort wird er gemäß folgender Skizze bestimmt:



Durch geometrische Überlegungen erhält man folgende Formel für x_{mod} mit den Bezeichnungen aus der Skizze:

$$x_{mod} = {x_u}^i + \frac{{\Delta x_u}^i}{{\Delta x_u}^i + {\Delta x_o}^i} \left({x_o}^i - {x_u}^i \right)$$

3.1.2 Der Median

Um den Median bei nicht-klassierten Daten zu definieren, ordnen wir die n Beobachtungswerte $x_1, x_2, ..., x_n$ der Größe nach, um so die geordneten Werte $x_{(1)}, x_{(2)}, ..., x_{(n)}$ mit der Eigenschaft

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

zu erhalten.

Der Median $\tilde{x}$ ist dann der Wert, der genau "in der Mitte" der geordneten Werte steht, d.h.

$$\tilde{x} := \begin{cases} x_{(\frac{n+1}{2})}, & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right), & \text{falls } n \text{ gerade} \end{cases}$$

Der Median $\tilde{x}$ zeichnet sich durch die charakteristische Eigenschaft aus, daß mindestens 50 Prozent aller Merkmalswerte $\leq \tilde{x}$ und mindestens 50 Prozent aller Merkmalswerte $\geq \tilde{x}$ sind.

Im Fall klassierter Daten ist der Median dasjenige x, bei dem die kumulierte Klassenhäufigkeit (gegeben durch die Verteilungsfunktion $F(x)$), den Wert 0.5 annimmt. Dies ergibt unter Zuhilfenahme der am Ende von 2.4.2 ermittelten Formel 2.1:

$$\begin{aligned} 0.5 &= {F_u}^i + \frac{{F_o}^i - {F_u}^i}{{x_o}^i - {x_u}^i}\left(x - {x_u}^i\right) \\ \Leftrightarrow \quad x &= {x_u}^i + \frac{{x_o}^i - {x_u}^i}{{F_o}^i - {F_u}^i}\left(0.5 - {F_u}^i\right) \end{aligned}$$

Dabei bezeichne der Index i diejenige Klasse K_i, in der die relative Summenhäufigkeit F_i das erste Mal den Wert 0.5 annimmt oder überschreitet.

3.1.3 Das arithmetische Mittel

Bei nicht-klassierten Daten ist das arithmetische Mittel $\bar{x}$ der n Merkmalswerte $x_1, x_2, ..., x_n$ definiert als:

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

Liegen für die verschiedenen Ausprägungen y_i, $i = 1, 2, ..., k$, der x_i die absoluten bzw. relativen Häufigkeiten h_i bzw. f_i vor, so gilt auch:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} y_i h_i = \sum_{i=1}^{k} y_i f_i$$

Der Mittelwert hat folgende Eigenschaft, die sich unmittelbar aus seiner Definition ergibt:

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

Im Fall klassierter Daten mit k Klassen und absoluten bzw. relativen Klassenhäufigkeiten f_i bzw. h_i, $i = 1, 2, ..., k$, werden in der obigen Formel die Werte y_i durch die Klassenmitten ${x_i}^*$ ersetzt, d.h. es gilt:

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{k} {x_i}^* h_i = \sum_{i=1}^{k} {x_i}^* f_i$$

Bei äquidistanten Klassen weicht das auf diese Weise approximativ bestimmte arithmetische Mittel maximal um eine halbe Klassenbreite vom tatsächlichen Wert ab.

3.1.4 Das geometrische Mittel

Werden relative Änderungen einer Größe durch Faktoren (z.B. Wachstums- oder Aufzinsungsfaktoren) und damit Prozente beschrieben, so muß der durchschnittliche Faktor und damit die

durchschnittliche Prozentzahl mit dem geometrischen anstelle des arithmetischen Mittels berechnet werden. Das geometrische Mittel x_g für n positive Werte $x_1, x_2, ..., x_n$ ist wie folgt definiert:

$$x_g = \sqrt[n]{(x_1 \cdot x_2 \cdot ... \cdot x_n)}$$

Zur Verdeutlichung betrachten wir ein Beispiel: Die Entwicklung des Jahresumsatzes eines Unternehmens im Zeitraum von 1999 bis 2003 sei durch folgende Tabelle gegeben:

Jahr	Index i	Umsatz in Tausend \$	Wachstumsrate p_i bzgl. Vorjahr	Wachstumsfaktor $x_i = 1 + \frac{p_i}{100}$ bzgl. Vorjahr
1999	1	500	-	-
2000	2	570	+14.0%	1.140
2001	3	680	+19.3%	1.193
2002	4	640	-5.9%	0.941
2003	5	720	+12.5%	1.125

Durch Multiplikation der jährlichen Wachstumsfaktoren x_i erhalten wir den Wachstumsfaktor von 1999 bis 2003, d.h.

$$x_1 \cdot x_2 \cdot x_3 \cdot x_4 = 1.140 \cdot 1.193 \cdot 0.941 \cdot 1.125 = 1.440$$

Der Umsatz ist also von 1999 bis 2003 um insgesamt 44.0 Prozent gewachsen. Der durchschnittlichen jährlichen Wachstumsfaktor ergibt sich dann über das geometrische Mittel x_g von x_1, x_2, x_3 und x_4:

$$x_g = \sqrt[4]{1.140 \cdot 1.193 \cdot 0.941 \cdot 1.125} = \sqrt[4]{1.440} = 1.0954$$

Wegen ${x_g}^4 = 1.440$ liefert eine durchschnittliche Wachstumsrate von 9.54% in 4 Jahren die oben angegebenen Gesamtwachstumsrate von 44.0%.

Die falsche Anwendung des arithmetischen Mittels ergäbe $\bar{x} = 1.100$ und damit eine um 0.46 Prozentpunkte zu hohe durchschnittliche Wachstumsrate. Es läßt sich zeigen, daß das geometrische Mittel stets höchstens so groß wie das arithmetische Mittel ist, d.h. es gilt folgende Ungleichung:

$$x_g \leq \bar{x}$$

3.1.5 Quantile

Der Begriff des Medians läßt sich zum α-Quantil x_α, $\alpha \in (0,1)$, in der folgenden Weise verallgemeinern:

$$x_\alpha := \begin{cases} x_{([n\alpha]+1)}, & \text{falls } n\alpha \text{ keine ganze Zahl ist} \\ \frac{1}{2}\left(x_{(n\alpha)} + x_{(n\alpha+1)}\right), & \text{sonst} \end{cases}$$

Das 0.5-Quantil entspricht demnach dem Median ($[x]$ bezeichnet hierbei die Gaußklammer, d.h. $[x]$ ist die größte ganze Zahl $\leq x$.).

Für das α-Quantil gilt, daß mindestens $100 \cdot \alpha\%$ aller Werte $\leq x_\alpha$ und mindestens $100 \cdot (1-\alpha)\,\%$ aller Werte $\geq x_\alpha$ sind.

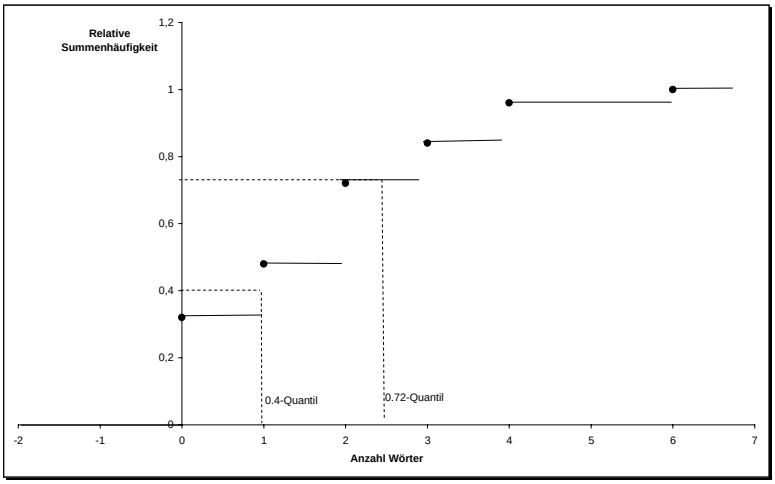
Das 0.25-Quantil und das 0.75-Quantil werden auch *unteres* bzw. *oberes Quartil* genannt.

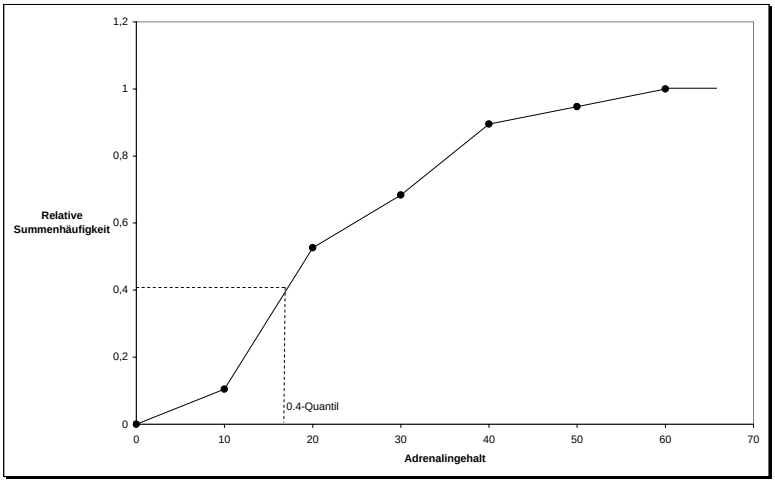
Im Fall klassierter Daten berechnet man das α-Quantil analog zum Median über die Formel:

$$\begin{aligned} \alpha &= {F_u}^i + \frac{{F_o}^i - {F_u}^i}{{x_o}^i - {x_u}^i}\left(x - {x_u}^i\right) \\ \Leftrightarrow \quad x &= {x_u}^i + \frac{{x_o}^i - {x_u}^i}{{F_o}^i - {F_u}^i}\left(\alpha - {F_u}^i\right) \end{aligned}$$

Hierbei bezeichne der Index i wieder diejenige Klasse K_i, in der die relative Summenhäufigkeit F_i das erste Mal den Wert α annimmt oder überschreitet.

Wie sich das α-Quantil im Falle von nicht-klassifierten bzw. klassierten Daten graphisch bestimmen läßt, ist anhand von Beispiel 2.1.1 und Beispiel 2.1.2 in den folgenden zwei Abbildungen dargestellt.





3.2 Streuungsparameter

So interessant die Kennwerte der Lage auch sein mögen, in der Praxis ist es unerläßlich auch die Streuung der Daten zu berücksichtigen. Verläßt man sich z.B. als Nichtschwimmer auf die Angabe “mittlere Tiefe: 1.50 Meter”, so gerät man leicht in Gefahr zu ertrinken, da trotz einer relativ geringen “mittleren” Tiefe das Wasser an einigen Stellen sehr viel tiefer sein kann.

3.2.1 Die Spannweite

Die Spannweite r von n Beobachtungswerten $x_1, x_2, ..., x_n$ ist definiert als

$$\begin{aligned} r &= \text{größter Merkmalswert - kleinster Merkmalswert} \\ &= x_{(n)} - x_{(1)} \end{aligned}$$

Bei klassierten Daten wird der als kleinster Wert die Untergrenze der ersten Klasse und als größter Wert die Obergrenze der letzten Klasse verwendet.

Die Spannweite ist ein sehr einfaches Streuungsmaß, das sehr empfindich auf Extremwerte reagiert; da sie lediglich auf 2 Werten der Beobachtungsreihe beruht, vermittelt sie nur ein sehr vages Bild von der Streuung.

3.2.2 Der Quartilsabstand

Die Differenz q zwischen oberen und unterem Quantil, d.h.

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

heißt (zentraler) Quartilsabstand.

Anschaulich handelt es sich dabei um den Bereich, in dem die mittleren 50% der Beobachtungswerte liegen. Er ist robust gegen Störungen an den Enden der Datenreihe durch Ausreißer, da die oberen und unteren 25% der Werte abgeschnitten werden.

Ausreißer, also Merkmalswerte, die extrem vom Gros der anderen Daten abweichen (vergleiche Abschnitt 2.4.3.6) lassen sich mit Hilfe des des Quartilsabstands auch numerisch identifizieren. Ein Merkmalswert x_i ist demnach ein Ausreißer, wenn er um mindestens das Anderthalbfache des Quartilsabstand unterhalb des unteren bzw. oberhalb des oberen Quartils liegt, d.h. wenn gilt:

$$x_i \in (-\infty, x_{0.25} - 1.5q] \cup [x_{0.75} + 1.5q, \infty)$$

x_i ist ein *extremer Ausreißer*, wenn er sogar um mindestens das Dreifache des Quartilsabstand unterhalb des unteren bzw. oberhalb des oberen Quartils liegt, d.h. wenn gilt:

$$x_i \in (-\infty, x_{0.25} - 3q] \cup [x_{0.75} + 3q, \infty)$$

3.2.3 Mittlere absolute Abweichung

Sei M ein "Mittelwert", also z.B. der Median, Modus, das arithmetische Mittel oder ähnliches, dann ist die mittlere absolute Abweichung a von M wie folgt definiert:

$$a := \frac{1}{n}\sum_{i=1}^{n}|x_i - M| = \frac{1}{n}\sum_{i=1}^{k} h_i|y_i - M| = \sum_{i=1}^{k} f_i|y_i - M|$$

In der Regel wählt man für M den Median, da für ihn die mittlere absolute Abweichung den kleinsten Wert annimmt (*Minimaleigenschaft des Medians*).

Bei klassierten Daten berechnet sich die mittlere absolute Abweichung analog unter Verwendung der Klassenmitten ${x_i}^*$

$$a :== \frac{1}{n}\sum_{i=1}^{k} h_i|{x_i}^* - M| = \sum_{i=1}^{k} f_i|{x_i}^* - M|$$

3.2.4 Varianz und Standardabweichung

Die Varianz bzw. mittlere quadratische Abweichung s^2 von n Merkmalswerten $x_1, x_2, ..., x_n$ vom arithmetischen Mittel $\bar{x}$ ist definiert als:

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

Die Wurzel aus der Varianz ($\sqrt{s^2}$) heißt Standardabweichung s; sie hat gegenüber der Varianz den Vorteil, daß sie dieselbe Dimension besitzt wie die Merkmalswerte.

Der *Verschiebungssatz* für s^2 erlaubt eine leichtere Berechnung von s^2:

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

Diese Formel leitet sich wie folgt her:

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

Bei klassierten Daten errechnet sich die Varianz wieder über die Klassenmitten, d.h.

$$\begin{aligned}
s^2 &:= \frac{1}{n}\sum_{i=1}^{n}({x_i}^*-\bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{k}h_i({x_i}^*-\bar{x})^2 = \sum_{i=1}^{k}f_i({x_i}^*-\bar{x})^2
\end{aligned}$$

Bemerkungen:

- In der Literatur wird die Varianz häufig auch durch

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

  definiert. Dies ist für die Anwendung von Ergebnissen aus der induktiven Statistik von Bedeutung (Schließen von der Varianz einer Stichprobe auf die Varianz der Grundgesamtheit). Für große n ist der Unterschied zwischen beiden Formel aber offensichtlich wegen $\lim\limits_{n\to\infty}\frac{n}{n-1}=1$ vernachlässigbar.

- Für beliebige Häufigkeitsverteilungen gilt, daß von allen Beobachtungswerten

  - mindestens 75% im Intervall $[\bar{x}-2s,\bar{x}+2s]$ und

  - mindestens 88.89% im Intervall $[\bar{x}-3s, \bar{x}+3s]$ liegen.

- Hat die Darstellung der Häufigkeitsverteilung das Aussehen einer Normalverteilung (vergleiche Abschnitt 2.4.3.3), so gilt insbesondere für große n, daß von den Beobachtungswerten

  - ca. 68% im Intervall $[\bar{x}-s, \bar{x}+s]$
  - ca. 95% im Intervall $[\bar{x}-2s, \bar{x}+2s]$ und
  - ca. 99% im Intervall $[\bar{x}-3s, \bar{x}+3s]$ liegen.

- Definiert man für n beliebig vorgegebenen Merkmalswerte $x_1, x_2, ..., x_n$ eine Funktion $s^2 : \mathbb{R} \mapsto \mathbb{R}$ durch

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

  so nimmt diese Funktion ihren kleinsten Wert für $x = \bar{x}$ an. Dies wird *Minimumseigenschaft des Mittelwertes* genannt.

3.2.5 Der Variationskoeffizient

Um Streuungen verschiedener Häufigkeitsverteilungen besser vergleichen zu können, kann man sie auf das arithmetische Mittel des Datensatzes beziehen. Dies führt zur Definition des Variationskoeffizienten v:

$$v := \frac{s}{|\bar{x}|} \cdot 100, \text{ falls } \bar{x} \neq 0$$

Da der Variationskoeffizient eine dimensionslose Größe ist, wird er üblicherweise in Prozent angegeben.

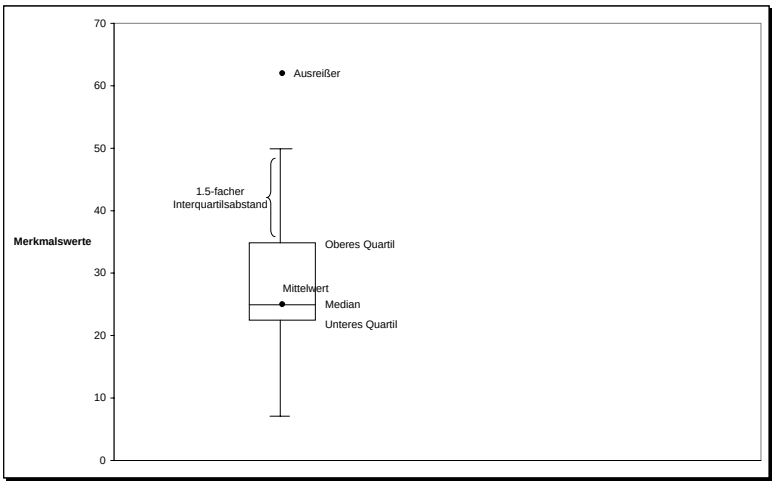
Beispiel: Eine Untersuchung des Durchschnittseinkommens der Bevölkerung zweier Länder hat zu folgenden Ergebnissen geführt:

	$\bar{x}$	s	$v := \frac{s}{\lvert\bar{x}\rvert} \cdot 100$
Land 1	4250 $	500 $	11.76%
Land 2	610 $	120 $	19.67%

Da das Durchschnittseinkommen $\bar{x}$ beider Länder sehr unterschiedlich ist, liefert das *relative* Streuungsmaß v einen besseren Vergleich für die Streuung des Durchschnittseinkommens als das *absolute* Streuungsmaß s.

3.2.6 Der Box-Whisker-Plot

Eine guten Überblick über die allzu oft unübersichtliche Häufigkeitsverteilung bietet der Box-Whisker-Plot. Durch die gleichzeitige graphische Darstellung verschiedener Lage- und Streuungsparameter soll das Datenmaterial überschaubarer und aussagekräftiger gemacht werden.



3.3 Konzentrationsmessung

Bei der Konzentrationsmessung wird untersucht, wie sich die Gesamtmerkmalssumme auf die Merkmalsträger verteilt. Sie wird häufig bei wirtschaftlichen Fragestellungen verwandt, z.B. bei der Untersuchung wie sich der Umsatz (Beschäftigte, Marktanteile) auf einzelne Unternehmen verteilt.

3.3.1 Die Lorenzkurve

Ein graphisches Hilfsmittel zur Veranschaulichung der (relativen) Konzentration ist die Lorenzkurve. Wir erläutern dies zunächst an einem Beispiel.

Beispiel: Es liegen für $n = 8$ Unternehmen einer Branche die folgenden Umsatzzahlen in Mio \$ vor

$$300 \quad 50 \quad 80 \quad 120 \quad 60 \quad 510 \quad 280 \quad 120$$

Es soll untersucht werden, wieviel Prozent des Gesamtumsatzes von 1520 Mio \$ auf welchen prozentualen Anteil der Unternehmen entfällt.

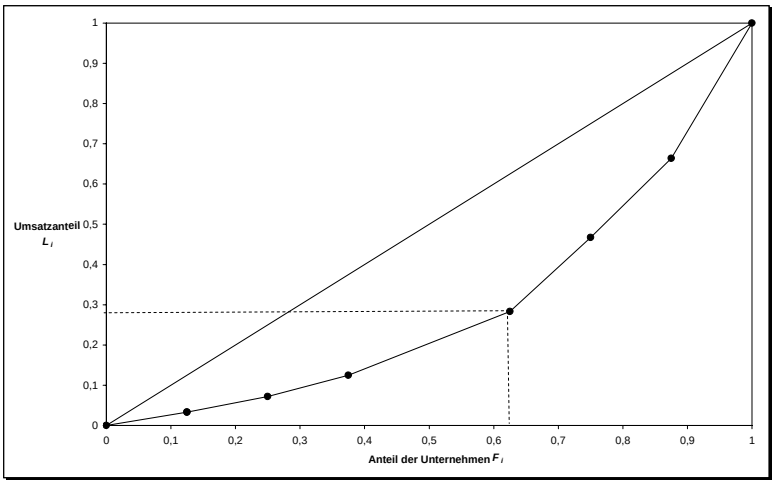
Dazu werden zunächst die $k = 7$ verschiedenen Merkmalsausprägungen y_i, $i = 1,2,\ldots,k$, der Größe nach geordnet und die Verteilung in folgender Häufigkeitstabelle zusammengefaßt (dabei bezeichne wie gehabt h_i bzw. f_i die absolute bzw. relative Häufigkeit der Merkmalsausprägung y_i sowie $F_i := \sum_{j=1}^{i} f_j$ die kumulative relative Häufigkeit der Merkmalsausprägung y_i):

Index i	y_i	h_i	f_i	$y_i h_i$	$l_i := \frac{y_i h_i}{\sum_{j=1}^{k} y_j h_j}$	F_i	$L_i := \sum_{j=1}^{i} l_j$
1	50	1	0.125	50	0.033	0.125	0.033
2	60	1	0.125	60	0.039	0.25	0.072
3	80	1	0.125	80	0.053	0.375	0.125
4	120	2	0.25	240	0.158	0.625	0.283
5	280	1	0.125	280	0.184	0.75	0.467
6	300	1	0.125	300	0.197	0.875	0.664
7	510	1	0.125	510	0.336	1	1
$\sum$		8	1	1520	1		

Die Lorenzkurve entsteht, wenn in einem Koordinatensystem die $k+1$ Punktepaare

$$(0,0) =: (F_0,L_0),(F_1,L_1),\ldots,(F_k,L_k) = (1,1)$$

eingezeichnet und gradlinig verbunden werden. In der Graphik der Lorenzkurve wird auch stets die erste Winkelhalbierende, d.h. die Strecke zwischen den Punkten $(0,0)$ und $(1,1)$ eingetragen.



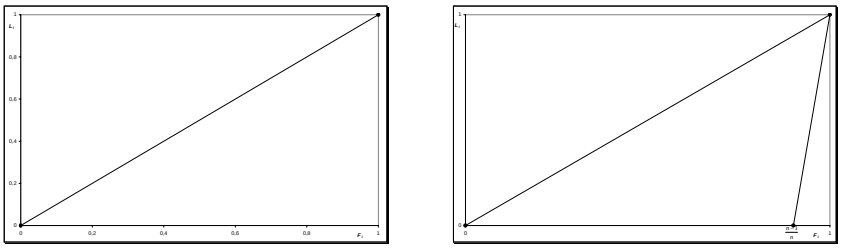
Aus der Lorenzkurve läßt sich z.B. entnehmen, daß auf 62.5% der (umsatzkleinsten) Betriebe nur ca. 28.3% des Gesamtumsatzes fallen.

Die Lorenzkurve ist eine monoton wachsende, konvexe Funktion, die sich stets unterhalb der ersten Winkelhalbierenden befindet. Ihre Lage bewegt sich immer zwischen den folgenden 2 Extremfällen:

1. *Keine Konzentration*: Die Merkmalswerte der n Merkmalsträger sind alle gleich (in obigem Beispiel bedeutete dies, daß alle 8 Unternehmen jeweils 190 Mio \$ Umsatz machten), d.h. die Lorenzkurve ist identisch mit der 1. Winkelhalbierenden.

2. *Vollständige Konzentration*: Die Merkmalssumme konzentriert sich auf genau einen der n Merkmalsträger und die anderen $n-1$ Merkmalträger haben den Wert 0 (in obigem Beispiel bedeutete dies, daß 7 Unternehmen 0 Mio \$ Umsatz und eines 1520 Mio \$ Umsatz machten).

Je näher also die Lorenzkurve an der ersten Winkelhalbierenden liegt, desto gleichmäßiger verteilen sich die Merkmalswerte auf die Merkmalsträger, d.h. desto geringer ist die Konzentration. Umgekehrt, je mehr die Lorenzkurve in Richtung des Punktes $(1,0)$ durchhängt, desto mehr vereinigt sich die Gesamtmerkmalssumme auf wenige Merkmalsträger, d.h. desto größer ist die Konzentration.



Im Fall von klassierten Daten mit unbekannten Merkmalssummen pro Klasse, werden die Merkmalssummen pro Klasse K_i über das Produkt aus Klassenmitte und Klassenhäufigkeit $x_i^* f_i$ näherungsweise berechnet.

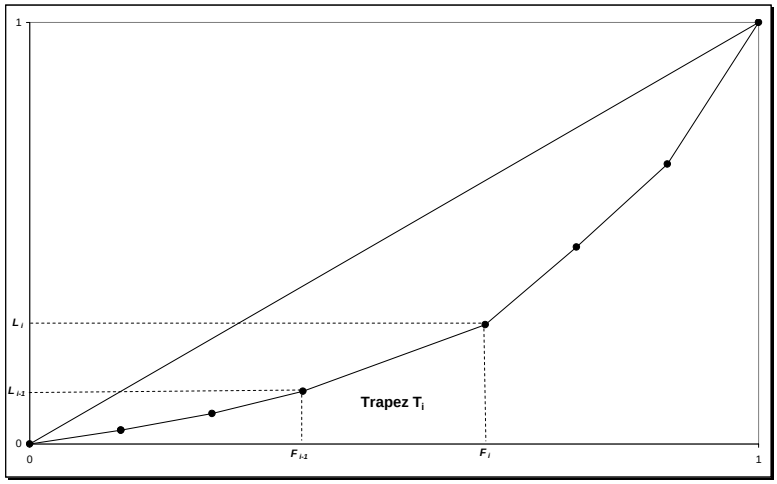
3.3.2 Der Gini-Koeffizient

Wegen der im letzten Abschnitt beschriebenen Eigenschaften der Lorenzkurve bietet es sich an als Maßzahl für die Konzentration die Fläche zwischen der ersten Winkelhalbierenden und der Lorenzkurve zu verwenden. Dieses Konzentrationsmaß heißt Gini-Koeffizient G und ist wie folgt definiert:

$$G := \frac{\text{Fläche zwischen der ersten Winkelhalbierenden und der Lorenzkurve}}{\text{Fläche unter der ersten Winkelhalbierenden}}$$

$$= 2 \cdot \text{Fläche zwischen der ersten Winkelhalbierenden und der Lorenzkurve}$$

G berechnet sich anhand der folgenden Zeichnung mittels der Summation von k Trapezflächen.



Das i-te Trapez T_i hat die Fläche

$$\begin{aligned} F(T_i) &:= (F_i - F_{i-1})\,L_{i-1} + \frac{1}{2}(F_i - F_{i-1})(L_i - L_{i-1}) \\ &= f_i L_{i-1} + \frac{1}{2} f_i (L_i - L_{i-1}) \\ &= f_i \frac{L_{i-1} + L_i}{2} \end{aligned}$$

Demnach errechnet sich die Gesamtfläche F unter der Lorenzkurve als

$$F = \sum_{i=1}^{k} F(T_i) = \sum_{i=1}^{k} f_i \frac{L_{i-1} + L_i}{2}$$

Dies führt zur folgenden Formel für den Ginikoeffizienten G

$$G = 2\left(\frac{1}{2} - F\right) = 1 - 2F = 1 - \sum_{i=1}^{k} f_i (L_{i-1} + L_i)$$

Daraus lassen sich folgenden Eigenschaften von G ableiten:

- Bei keiner Konzentration hat G den Wert 0 .
- Bei vollständiger Konzentration hat G den Wert $1 - \frac{1}{n} = \frac{n-1}{n}$.

- Stets ist $0 \leq G \leq \frac{n-1}{n}$ und der Ginikoeffizient nimmt innerhalb dieser Grenzen einen umso größeren Wert an, je höher die Konzentration der Merkmalssumme ist.

Um Daten mit einer unterschiedlichen Zahl von Merkmalsträgern besser vergleichbar zu machen, definiert man neben dem Ginikoeffizienten G den normierten Ginikoeffzienten G_{norm}

$$G_{norm} := G \cdot \frac{n}{n-1}$$

4 Zweidimensionale Häufigkeitsverteilungen

Hängt das Arbeitseinkommen vom Ausbildungsstand ab, vom Alter oder vom Geschlecht? Beeinflußt der Fernsehkonsum das Leseverhalten bzw. den Schulerfolg von Jugendlichen? Solche und ähnliche Fragestellungen erfordern die Untersuchung von Zusammenhängen und Abhängigkeiten von zwei Merkmalen, die gemeinsam erhoben werden müssen.

4.1 Streuungsdiagramm und gemeinsame Verteilung

Im Unterschied zur den bisherigen Überlegungen betrachten wir also nun gleichzeitig zwei verschiedene Merkmale von n Merkmalsträgern. d.h. es treten jetzt Paare von Merkmalswerten

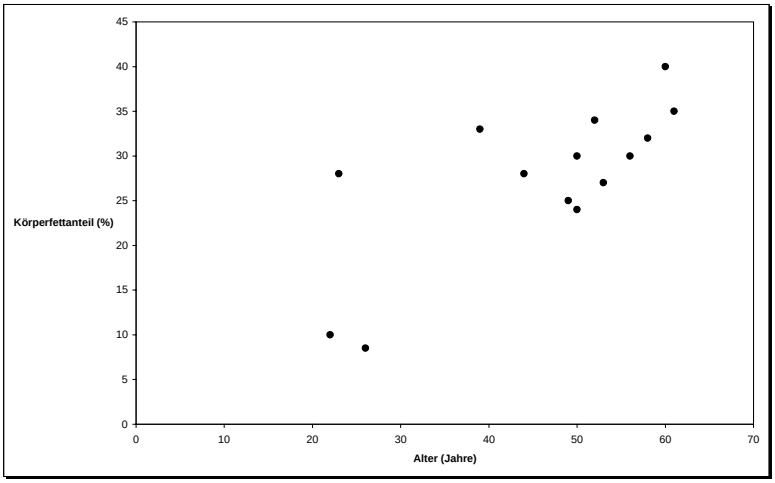
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

auf. Dabei entspricht das i-te Zahlenpaar den beiden Werten, die der i-te Merkmalsträger liefert.

Beispiel 4.1.1: Bei $n = 14$ Erwachsenen werden Alter und Körperfettanteil gemessen.

Versuchsperson	Alter in Jahren	Körperfettanteil in Prozent
1	22	10
2	26	8.5
3	23	28
4	39	33
5	44	28
6	49	25
7	50	24
8	50	30
9	52	34
10	53	27
11	56	30
12	58	32
13	60	40
14	61	35

Diese Daten lassen sich gut durch ein Streuungsdiagramm visualisieren: Dabei werden die n Wertepaare (x_i, y_i) als Koordinaten von Punkten P_i angesehen und ein ein Koordinatensystem eingezeichnet.



Analog zum Fall eindimensionaler Verteilungen lassen sich die absoluten bzw. relativen Häufigkeiten bilden. Sei dazu n die Anzahl aller Merkmalsträger und k die Anzahl der verschiedenen Ausprägungen des Merkmals X und l die Anzahl der verschiedenen Ausprägungen des Merkmals Y. Dann definieren wir für $i = 1,2,\ldots,k$ und $j = 1,2,\ldots,l$

$$\begin{aligned} h_{ij} &:= \text{absolute Häufigkeit des Paares } (x_i, y_j) \\ f_{ij} &:= \text{relative Häufigkeit des Paares } (x_i, y_j) = \frac{h_{ij}}{n} \end{aligned}$$

Beispiel 4.1.2: Anzahl der erfolgreichen Selbstmorde in Kopenhagen (1988), aufgeschlüsselt nach Geschlecht und Diagnose, Tabelle der absoluten und relativen Häufigkeiten

	Absolute Häufigkeiten		Relative Häufigkeiten	
Diagnose	Männer	Frauen	Männer	Frauen
Schizophrenie	2	4	0.018	0.036
Affektive Psychose	1	7	0.009	0.063
Psychogene Psychose	0	2	0	0.018
Neurose	2	7	0.018	0.063
Persönlichkeitsstörung	12	8	0.107	0.071
Alkoholismus	16	12	0.143	0.107
Drogenabhängigkeit	2	1	0.018	0.009
Andere Diagnose	2	7	0.018	0.063
Nicht psychisch krank	13	14	0.116	0.125

Bezeichnet X das Merkmal Diagnose und Y das Merkmal Geschlecht, so läßt sich z.B. aus der obigen Tabelle ablesen, daß die absolute Häufigkeit h_{52} gleich 8 ist, d.h. 8 Frauen mit Persönlichkeitsstörungen haben 1988 erfolgreich Selbstmord in Kopenhagen begangen.

Bei stetigen oder diskreten Merkmalen mit sehr vielen Ausprägungen, werden bei der Anfertigung von zweidimensionalen Häufigkeitstabellen wie bei eindimensionalen Häufigkeitsverteilungen sehr oft Größenklassen gebildet, um das Datenmaterial überschaubar zu machen.

4.2 Randverteilungen

Natürlich kann man auch bei zweidimensionalem Datenmaterial das Augenmerk nur auf eines der beiden Merkmal richten und Zusammenhänge zunächst unbeachtet lassen. Man wird dann diese eindimensionalen Merkmale getrennt behandeln und mit den Verfahren für eindimensionale Häufigkeiten auswerten. Dies geschieht durch Aufsummieren der Zeilen bzw. Spalten in der zweidimensionalen Häufigkeitstabelle, man spricht deswegen auch von Randverteilungen.

Im Beispiel 4.1.2 errechnen sich die Randverteilungen wie folgt:

	Absolute Häufigkeiten			Relative Häufigkiten		
Diagnose	Männer	Frauen	$\sum$	Männer	Frauen	$\sum$
Schizophrenie	2	4	6	0.018	0.036	0.054
Affektive Psychose	1	7	8	0.009	0.063	0.071
Psychogene Psychose	0	2	2	0	0.018	0.018
Neurose	2	7	9	0.018	0.063	0.080
Persönlichkeitsstörung	12	8	20	0.107	0.071	0.179
Alkoholismus	16	12	28	0.143	0.107	0.25
Drogenabhängigkeit	2	1	3	0.018	0.009	0.027
Andere Diagnose	2	7	9	0.018	0.063	0.080
Nicht psychisch krank	13	14	27	0.116	0.125	0.241
$\sum$	50	62	112	0.446	0.554	1

Bezeichnen wir mit $h_{i.}$ bzw. $f_{i.}$ die absolute bzw. relative Häufigkeit der i-ten Ausprägung des Merkmals X und mit $h_{.j}$ bzw. $f_{.j}$ die absolute bzw. relative Häufigkeit der j-ten Ausprägung des Merkmals Y, so lassen sich die absoluten bzw. relativen Häufigkeiten der Randverteilungen in Formeln wie folgt ausdrücken:

$$h_{i.} := \sum_{j=1}^{l} h_{ij} \quad \text{bzw.} \quad f_{i.} := \sum_{j=1}^{l} f_{ij}$$

$$h_{.j} := \sum_{i=1}^{k} h_{ij} \quad \text{bzw.} \quad f_{.j} := \sum_{i=1}^{k} f_{ij}$$

Beschränkt man die Betrachtung allerdings nur auf die Randverteilungen, so geht die wesentliche Information der zweidimensionalen Statistik, nämlich die über das gemeinsame Verhalten

der beiden Merkmale und ihrer Abhängigkeiten untereinander verloren.

4.3 Bedingte Verteilung und statistische Unabhängigkeit

Besonders interessiert bei Paaren von Merkmalsträgern die Verteilung der relativen Häufigkeiten eines Merkmals, während das andere auf einem bestimmten Wert festgehalten wird. Auf diese Weise erhält man einen wichtigen Einblick in die Art des Zusammenhanges beider Merkmale untereinander. Man spricht in diesem Fall von bedingten Verteilungen. In Beispiel 4.1.2 wäre eine solche bedingte Verteilung z.B. die Verteilung der Diagnosen bei den Selbstmörderinnen bzw. die Geschlechterverteilung bei den Selbstmördern, die unter Schizophrenie litten.

Bezeichne $f_{i|y_j}$ die relative Häufigkeit des Merkmalswertes x_i unter allen Merkmalspaaren, bei denen das Merkmal Y den Wert y_j annimmt, und analog $f_{j|x_i}$ die relative Häufigkeit des Merkmalswertes y_j unter allen Merkmalspaaren, bei denen das Merkmal X den Wert x_i annimmt. Dann berechnen sich die bedingten Verteilungen in Formeln wie folgt:

$$\begin{aligned} f_{i|y_j} &:= \frac{f_{ij}}{\sum_{i=1}^{k} f_{ij}} = \frac{f_{ij}}{f_{.j}} \\ f_{j|x_i} &:= \frac{f_{ij}}{\sum_{j=1}^{l} f_{ij}} = \frac{f_{ij}}{f_{i.}} \end{aligned}$$

Für die gemeinsame Verteilung aus Beispiel 4.1.2 gibt es zwei bedingte Verteilungen des Merkmals X Diagnose, je nachdem ob das Merkmal Y Geschlecht den Wert y_1 männlich oder den Wert y_2 weiblich annimmt.

Index i	Diagnose	$f_{i\|y_1}$	$f_{i\|y_2}$
1	Schizophrenie	0.041	0.064
2	Affektive Psychose	0.02	0.113
3	Psychogene Psychose	0	0.032
4	Neurose	0.041	0.113
5	Persönlichkeitsstörung	0.118	0.129
6	Alkoholismus	0.32	0.194
7	Drogenabhängigkeit	0.041	0.016
8	Andere Diagnose	0.041	0.113
9	Nicht psychisch krank	0.259	0.226
	$\sum$	1	1

Analog gibt es neun bedingte Verteilungen des Merkmals Y Geschlecht, je nachdem welchen Wert das Merkmal X Diagnose annimmt.

Anhand der bedingten Verteilungen läßt sich untersuchen, ob zwei Merkmale statistisch unabhängig sind, d.h. anschaulich gesprochen, ob die Ausprägung eines Merkmals keinen Einfluß

Diagnose	$f_{1\|x_i}$	$f_{2\|x_i}$	Σ
Schizophrenie	0.339	0.661	1
Affektive Psychose	0.128	0.885	1
Psychogene Psychose	0	1	1
Neurose	0.227	0.784	1
Persönlichkeitsstörung	0.602	0.398	1
Alkoholismus	0.611	0.429	1
Drogenabhängigkeit	0.69	0.345	1
Andere Diagnose	0.041	0.784	1
Nicht psychisch krank	0.479	0.517	1

auf die Ausprägung des anderen Merkmals besitzt. In diesem Fall läßt sich aus der Kenntnis der Randverteilungen die gemeinsame Verteilung herleiten, da keine Interaktionen zwischen den Merkmalen vorliegen. Dies führt zu folgender mathematischer Definition der statistischen Unabhängigkeit:

Ist die gemeinsame Verteilung f_{ij} zweier Merkmale X und Y gleich dem Produkt der beiden Randverteilungen, d.h. gilt

$$f_{ij} = f_{i.}f_{.j}, \text{ für } i = 1,2,\ldots,k \text{ und } j = 1,2,\ldots,l,$$

so heißen X und Y statistisch unabhängig.

Bei statistisch unabhängigen Variablen X und Y sind die $j = 1,2,\ldots,l$ bedingten Verteilungen von X identisch und jeweils gleich der Randverteilung:

$$f_{i|y_j} = \frac{f_{ij}}{f_{.j}} = \frac{f_{i.}f_{.j}}{f_{.j}} = f_{i.}, \text{ für } i = 1,2,\ldots,k$$

Analog sind auch alle $i = 1,2,\ldots,k$ bedingten Verteilungen von Y identisch und gleich der Randverteilung:

$$f_{j|x_i} = \frac{f_{ij}}{f_{i.}} = \frac{f_{i.}f_{.j}}{f_{i.}} = f_{j.}, \text{ für } j = 1,2,\ldots,l$$

Demnach sind in Beispiel 4.1.2 die Merkmale Diagnose und Geschlecht **nicht** statistisch unabhängig, da sich alle bedingten Verteilungen voneinander unterscheiden. Bei realen Daten ist es allerdings äußerst unwahrscheinlich, daß sich ein lupenreine statistische Unabhängigkeit in der empirischen Verteilung zeigt, selbst wenn keinerlei (theoretische) Abhängigkeit zwischen den zwei Merkmalen besteht. In den folgenden Abschnitten werden wir Möglichkeiten kennenlernen, den Grad der statistischen Abhängigkeit genauer zu quantifizieren.

4.4 Statistischer Zusammenhang von mindestens ordinal skalierten Merkmalen

4.4.1 Die Kovarianz

Die Kovarianz $Cov(X,Y)$ ist ein Maß für die Streuung zweier Merkmale X und Y. Sie ist definiert als das arithmetische Mittel des Produktes des Abweichungen der einzelnen Beobachtungen $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ von ihrem jeweiligen Mitteln, also

$$Cov(X,Y) := \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

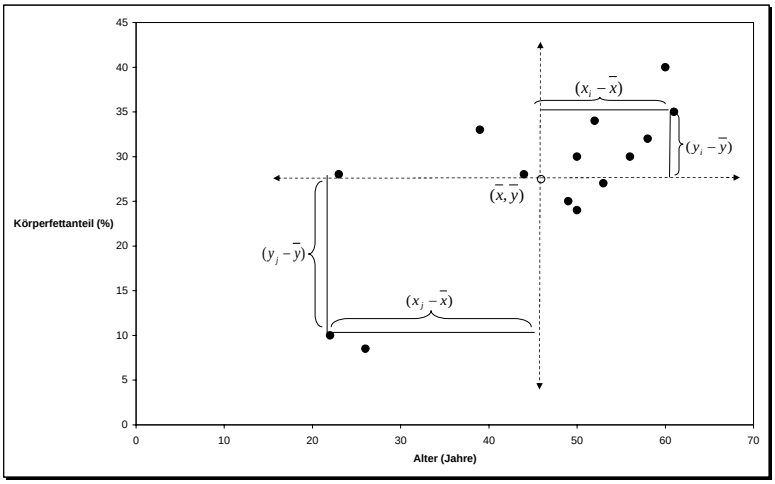
bzw. unter Benutzung der relativen Häufigkeiten f_{ij}

$$Cov(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}(x_i-\bar{x})(y_j-\bar{y})$$

Man beachte, daß die Kovarianz eine Verallgemeinerung der Varianz darstellt, da offensichtlich gilt:

$$Cov(X,X) = s_X^2$$

Zur Illustration der Kovarianz zeichnen wir im Streuungsdiagramm im Beispiel 4.1.1 ein Hilfs-Koordinatensystem ein, das durch den Schwerpunkt $(\bar{x},\bar{y})$ der Punktewolke geht.



In diesem Koordinatensystem werden die Abweichungen der Beobachtungswerte von ihren eigenen arithmetischen Mitteln gemessen. Daher sind seine Achsen mit $x-\bar{x}$ und $y-\bar{y}$ bezeichnet. Die einzelnen Abweichungsprodukte $(x_i-\bar{x})(y_i-\bar{y})$ entsprechen den orientierten Flächen der von den einzelnen Punkte aufgespannten Rechtecken. Sind die Abweichungen groß, entstehen große Rechtecke, sind die Abweichungen klein, entstehen kleine Rechtecke. Die Rechtecksflächen im I. und III. Quadranten entsprechen positiven Abweichungsprodukten, die im II. und IV. Quadranten negativen Abweichungsprodukten. Überwiegen die positiven Abweichungsprodukte, ist die Gesamtsumme der Abweichungsprodukte positiv, überwiegen dagegen die Beobachtungen im II. und IV. Quadranten, so wird sie negativ.

Eine positive Kovarianz beschreibt somit einen gemeinsamen Trend der beobachteten Werte x_i und y_i: Relativ große (kleine) Werte von X gehen im Durchschnitt mit relativ großen (kleinen) Werten von Y einher. Analog zeigt eine negative Kovarianz an, daß große (kleine) Werte der eienn Variable im Durchschnitt eher mit kleinen (großen) Werten der anderen Variable einhergehen.

Ähnlich wie bei der Varianz gibt es auch bei der Kovarianz eine vereinfachte Berechnungsformel:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

Sind zwei Merkmale X und Y statistisch unabhängig, so ist die Kovarianz $Cov(X,Y)$ zwischen ihnen Null, denn in diesem Fall gilt:

$$\begin{aligned}
Cov(X,Y) &= \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}(x_i-\bar{x})(y_j-\bar{y}) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{l} f_{i.}f_{.j}(x_i-\bar{x})(y_j-\bar{y}) \\
&= \left(\sum_{i=1}^{k} f_{i.}(x_i-\bar{x})\right)\left(\sum_{j=1}^{l} f_{.j}(y_j-\bar{y})\right) \\
&= \left(\sum_{i=1}^{k} f_{i.}x_i - \sum_{i=1}^{k} f_{i.}\bar{x}\right)\left(\sum_{j=1}^{l} f_{.j}(y_j-\bar{y})\right) \\
&= \left(\sum_{i=1}^{k} f_{i.}x_i - \bar{x}\sum_{i=1}^{k} f_{i.}\right)\left(\sum_{j=1}^{l} f_{.j}(y_j-\bar{y})\right) \\
&= (\bar{x}-\bar{x}\cdot 1)\left(\sum_{j=1}^{l} f_{.j}(y_j-\bar{y})\right) = 0
\end{aligned}$$

Man beachte, daß dieser Satz nicht umkehrbar ist: Aus der statistischen Unabhängigkeit folgt zwar das Verschwinden der Kovarianz, jedoch liegt keineswegs immer Unabhängigkeit vor, wenn die Kovarianz verschwindet. In der Tat mißt die Kovarianz nur den *linearen Anteil* der statistischen Abhängigkeit.

4.4.2 Der Korrelationskoeffizient von Bravais[1]-Pearson[2]

Da ein großer Zahlenwert der Kovarianz auch allein daher rühren kann, daß die Streuung der beiden Merkmale X und Y für sich genommen schon groß ist, obwohl gar keine allzu große lineare Abhängigkeit zwischen ihnen besteht, definiert man ein normiertes Maß für die Stärke des linearen statistischen Zusammenhanges, indem man die Kovarianz durch die Standardabweichungen der beiden Merkmale dividiert. Damit erhalten wir den Korrelationskoeffizienten von Bravais-Pearson $r(X,Y)$

$$r(X,Y) := \frac{Cov(X,Y)}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$r(X,Y)$ hat folgende Eigenschaften:

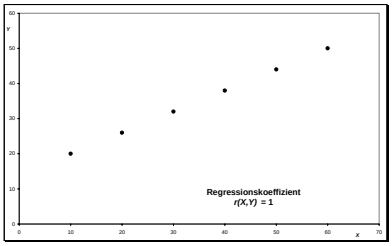
- $r(X,Y)$ hat das gleiche Vorzeichen wie die Kovarianz, liegt aber stets zwischen -1 und 1, d.h.
$$-1 \leq r(X,Y) \leq 1$$
- $r(X,Y)$ bleibt betragsmäßig unverändert, wenn man eine oder beide Variablen linear transformiert.
- Vertauscht man die Merkmale Xund Y, so ändert sich dadurch nichts an der Kovarianz ($Cov(X,Y) = Cov(Y,X)$) und damit auch nichts am Korrelationskoeffizienten, vielmehr ist
$$r(X,Y) = r(Y,X)$$
Beide Merkmale werden also in der Korrelationsrechnung symmetrisch behandelt, keines ist gegenüber dem anderen bevorzugt. Es wird zwar die statistische Abhängigkeit untersucht, aber ohne festzulegen, welches der beiden Merkmale die abhängige bzw. unabhängige Variable ist. Dies ist in der Regressionsrechnung, die in Abschnitt 4.4.4 behandelt werden wird, anders.

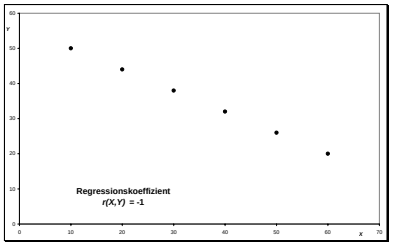
Zur Verdeutlichung betrachten wir im folgenden mehrere Beispiele von Merkmalspaaren (x_1,y_1), $(x_2,y_2),\ldots,(x_6,y_6)$, die in Streuungsdiagrammen dargestellt werden und für die jeweils $r(X,Y)$ berechnet wird. Für das Merkmal X gelte dabei stets:

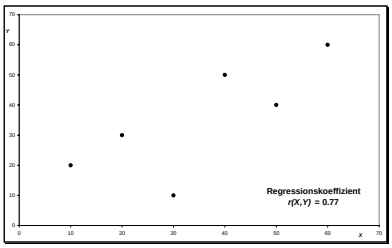
$$x_1 = 10, x_2 = 20, x_3 = 30, x_4 = 40, x_5 = 50, x_6 = 60$$

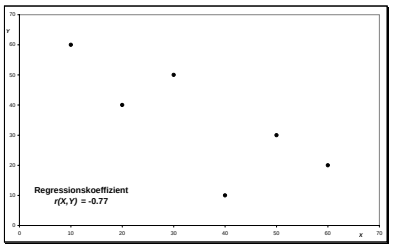
[1] Auguste BRAVAIS, 1811-1863, französischer Physiker, Professor an der Pariser Ecole polytechnique, berühmt durch die Entdeckung der Gitterstruktur der Kristalle (Bravais-Gitter). Wahrscheinlich hat er den Korrelationskoeffizienten "erfunden".

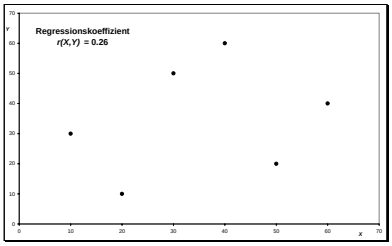
[2] Karl PEARSON, 1857-1936, englischer Mathematiker und Anthropologe am Londoner University College. Er ist einer der Begründer der modernen Statistik. Außerdem war er noch als Rechtsanwalt, Poet und radikaler Politiker tätig.

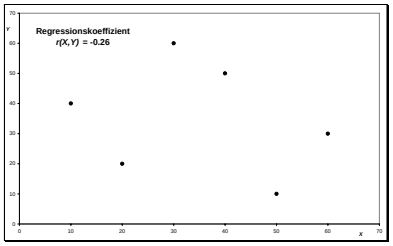


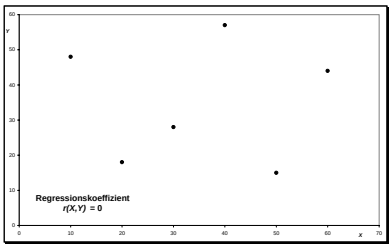


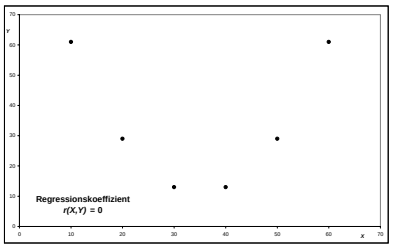












Eine Übersicht über die verschiedenen Merkmalswerte von Y pro Diagramm (geordnet von links oben nach rechts unten) liefert die folgende Tabelle:

	Merkmalswerte von Y					
Diagramm	y_1	y_2	y_3	y_4	y_5	y_6
1	20	26	32	38	44	50
2	50	44	38	32	26	20
3	20	30	10	50	40	60
4	60	40	50	10	30	20
5	30	10	50	60	20	40
6	40	20	60	50	10	30
7	48	18	28	57	15	44
8	61	29	13	13	29	61

Die Diagramme verdeutlichen folgende Eigenschaften von $r(X,Y)$:

- $r(X,Y) = 1$ gilt genau, dann wenn alle Punkte $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ auf einer Geraden mit positiver Steigung liegen, bzw. $r(X,Y) = -1$ gilt genau dann, wenn alle Punkte auf einer Geraden mit negativer Steigung liegen. Den Beweis für diese Aussage werden wir in Abschnitt 4.4.4 liefern.
- $r(X,Y) = 0$ bedeutet, daß kein linearer Zusammenhang zwischen den Merkmalen X und Y besteht (auch nicht näherungsweise). Zusammenhänge anderer Art (z.B. quadratisch oder exponentiell) können allerdings bestehen. Die Merkmale heißen in diesem Fall unkorreliert.
- Im Fall $r(X,Y) > 0$ bzw. $r(X,Y) < 0$ heißen die Merkmale positiv bzw. negativ korreliert.
- Ist $r(X,Y)$ betragsmäßig groß bzw. klein, so spricht man auch noch von stark bzw. schwach korrelierten Merkmalen.

Bemerkungen:

- Der Wert des Korrelationskoeffizienten kann sehr stark durch Extremwerte beeinflusst werden. Das ist leicht nachzuvollziehen, da Extremwerte die Varianz eines Merkmals stark erhöhen, und dann durch die Regression sehr viel von dieser Varianz "erklärt" werden kann (vergleiche auch Abschnitt 4.4.4).
- Die gemeinsame Betrachtung von zwei sehr unterschiedlichen Gruppen kann zu einer hohen Korrelation zwischen Merkmalen führen, obwohl innerhalb jeder Gruppe nur eine geringe oder gar keine Korrelation zwischen den Merkmalen besteht (*Heterogenitätskorrelation*). In Extremfällen kann sich sogar die Richtung der Korrelation umkehren: Ein bekanntes Beispiel ist eine im Handelsblatt unter dem Stichwort "Methusalems machen Kasse" veröffentlichte Statistik, nach der langes Studieren ein hohes Starteinkommen fördert. Die dort gemeldete positive Korrelation von Studiendauer und Gehalt beruhte einfach darauf, daß alle Studienfächer in einen Topf geworfen wurden und die Hochschulabsolventen der langwierigsten Fächer, wie Chemie und Medizin, die höchsten Startgehälter

besaßen. Diese hohen Startgehälter hatten sie aber nicht wegen der Länge, sondern wegen der Schwere des Studiums. Hält man aber die dritte Variable, nämlich das Studienfach konstant, so ist der Zusammenhang zwischen Studiendauer und Gehalt in alle Fächern negativ.

- Der Korrelationskoeffizient liefert **keine** Aussage über einen kausalen Zusammenhang. Viele veröffentlichte Nonsens-Korrelationen beruhen auf übersehenen Hintergrundvariablen: So zeigen zum Beispiel Zeitreihendaten wie Volkseinkommen, Staatsverschuldung, Studentenzahlen und Auslandsurlauber aus verschiedenen Gründen häufig einen monotonen Trend, so daß je zwei dieser Variablen miteinander (positiv oder negativ) korrelieren. Dieser gemeinsame Trend ist z.B. auch für die positive Korrelation von Klapperstörchen und Geburten in der Bundesrepublik veranwortlich, denn beide Variablen nahmen über lange Zeit im Gleichschritt ab.

4.4.3 Der Rangkorrelationskoeffizient von Spearman[3]

Der Rangkorrelationskoeffizient von Spearman mißt monotone Zusammenhänge zwischen zwei Merkmalen, basierend auf den Rängen der Merkmalswerte. Zur Erläuterung des Begriffes Rang betrachten wir die n Merkmalswerte

$$x_1, x_2, \ldots, x_n$$

und die zugehörige geordnetet Liste

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

(vergleiche Abschnitt 3.1.2). Jedem x_i wird als Rang $R(x_i)$ die Platznummer zugewiesen, die x_i in der geordneten Liste einnimmt. Sind alle x_i verschieden, so ist diese Rangzuordnung eindeutig; treten zwei oder mehrere gleich große Werte auf, so numeriert man zunächst einfach durch, ordnet dann aber den gleichen Werten das arithmetische Mittel ihrer Rangplätze zu ("verbundene Ränge").

Beispiel: Es seien $n = 8$ Beobachtungswerte wie folgt vorgegeben:

$$4.1, 18.2, 19, 3.5, 19, 1.8, 10.5, 11.3$$

Die zugehörige geordnete Liste stellt sich dar als

$$1.8, 3.5, 4.1, 10.5, 11.3, 18.2, 19, 19$$

Damit ergeben sich folgende Ränge:

x_i	4.1	18.2	19	3.5	19	1.8	10.5	11.3
$R(x_i)$	3	6	7.5	2	7.5	1	4	5

[3] Charles Edward SPEARMAN, 1863-1945, englischer Psychologe und wie PEARSON Professor am Londoner University College. Er schuf die Ansätze zur objektiven Messung von Intelligenz und anderen menschlichen Fähigkeiten.

Der Rangkorrelationskoeffizient von Spearman $r_S(X,Y)$ von n Merkmalspaaren (x_1,y_1), (x_2,y_2), $\ldots,(x_n,y_n)$ errechnet sich dann, indem zunächst jeweils die Ränge der x_i und der y_i getrennt bestimmt werden und dann auf die Rangpaare

$$(R(x_1),R(y_1)),(R(x_2),R(y_2)),\ldots,(R(x_n),R(y_n))$$

der Korrelationskoeffizient von Bravais-Pearson angewendet wird. Bezeichnen wir mit $\overline{R(x)} := \frac{1}{n}\sum_{i=1}^{n} R(x_i)$ das arithmetische Mittel der Ränge $R(x_i)$ und analog mit $\overline{R(y)} := \frac{1}{n}\sum_{i=1}^{n} R(y_i)$ das arithmetische Mittel der Ränge $R(y_i)$, so gilt also:

$$r_S(X,Y) \quad := \quad \frac{\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)\left(R(y_i)-\overline{R(y)}\right)}{\sqrt{\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2}}$$

Bemerkungen:

- $r_S(X,Y)=1$ ist genau dann erfüllt, wenn gilt:

$$R(x_i)=R(y_i), \text{ für } i=1,2,\ldots,n$$

- $r_S(X,Y)=-1$ ist genau dann erfüllt, wenn

$$R(x_i)=n+1-R(y_i), \text{ für } i=1,2,\ldots,n$$

Monotone Transformationen der beiden Variablen (wie z.B. lineare Transformationen, das Logarithmieren oder Exponentieren) verändern allenfalls das Vorzeichen des Rangkorrelationskoeffizienten von Spearman, da sie die Rangplätze entweder erhalten oder ihre Reihenfolge genau umkehren. Man wird im Einzelfall eher die Rangkorrelation berechnen, wenn man zwar einen monotonen, aber nicht unbedingt linearen Zusammenhang zwischen den beiden Merkmalenvermutet oder man der Qualität der Skala eines oder beider Merkmale nicht traut, also nicht weiß, ob sie abstandstreu ist. Bei Examensnoten z.B. ist sicherlich klar, daß eine 1 besser als eine 2 ist, aber ob die Differenz zwischen der 1 und der 2 genauso viel bedeutet, wie diejenige zwischen der 2 und der 3 ist fraglich.

4.4.4 Lineare Regression

Die lineare Regressionrechnung hat zum Ziel, einen linearen Zusammenhang zwischen zwei stetigen oder zumindest quasistetigen Merkmalen genauer zu beschreiben und hiermit z.B. den Wert der einen Variable vorherzusagen, wenn man nur den Wert der anderen kennt. Der Korrelationskoeffizient von Pearson genügt dieser Aufgabe offensichtlich nicht, da er nur den Grad und die Richtung des linearen Zusammenhangs in einer Zahl mißt, aber keine Auskunft darüber gibt, wie dieser Zusammenhang genau beschaffen ist.

Patientennummer i	Körpergewicht x_i	systolischer Blutdruck y_i	x_i^2	$x_i y_i$
1	54.5	128	2970.25	6976
2	77.0	154	5929	11858
3	78.5	180	6162.25	14130
4	48.0	96	2304	4608
5	90.0	142	8100	12780
6	86.5	170	7482.25	14705
7	54.6	122	2981.16	6661.2
8	61.0	130	3721	7930
9	66.0	118	4356	7788
10	54.0	98	2916	5292
11	85.0	172	7225	14620
12	80.0	149	6400	11920
13	80.5	150	6480.25	12075
14	96.7	181	9350.89	17502.7
15	68.0	170	4624	11560
16	50.0	109	2500	5450
17	71.5	140	5112.25	10010
18	55.0	150	3025	8250
19	78.5	139	6162.25	10911.5
20	94.5	157	8930.25	14836.5
21	68.7	121	4719.69	8312.7
22	97.2	160	9447.84	15552
23	53.0	91	2809	4823
24	84.0	161	7056	13524
$\sum$	1732.7	3388	130764.33	252075.6

Beispiel: Bei 24 zufällig ausgewählte Patienten einer dermatologischen Ambulanz wurden Körpergewicht (in kg) und systolischer Blutdruck (in mm Hg) gemessen.

Die Verteilung der Werte läßt es plausibel erscheinen, daß ein linearer Zusammenhang zwischen Körpergröße (X) und systolischem Blutdruck (Y) besteht. Dabei scheint es sinnvoll, den Blutdruck in Abhängigkeit vom Gewicht und nicht umgekehrt zu betrachten, d.h. wir nehmen an, daß reelle Zahlen a und b existieren, so daß näherungsweise gilt:

$$Y = a + bX$$

a und b sollen jetzt so bestimmt werden, daß die obige Gleichung für die vorliegenden Daten, also die Punktepaare (x_i, y_i), $i = 1, 2, \ldots, n$, möglichst gut erfüllt ist.

Es gibt mehrere Ansätze dieses Problem zu lösen. Die Standardmethode ist die sogenannte *Methode der kleinsten Quadrate* nach Carl Friedrich Gauß, bei der a und b so gewählt werden, daß

die Quadratsumme der einzelnen Abweichungen zwischen den y_i und $a + bx_i$ minimiert wird. Demnach sind a und b so zu bestimmen, daß die Funktion

$$Q(a,b) := \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

ihr Minimum annimmt. Mit Hilfsmitteln aus der mehrdimensionalen Differentialrechnung läßt sich zeigen, daß $Q(a,b)$ für folgenden Werte $\hat{a}$, $\hat{b}$ von a und b minimal wird:

$$\begin{aligned} \hat{b} &= \frac{Cov(X,Y)}{s_X^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} \end{aligned}$$

$\hat{a}$ und $\hat{b}$ sind *Schätzwerte* für die Parameter a und b. Die Gerade $y = \hat{a} + \hat{b}x$ heißt *Regressionsgerade*; die durch die Regressionsgerade bestimmten y-Werte $\hat{y}_i = \hat{a} + \hat{b}x_i$ werden die durch *die Regression erklärten Werte* genannt.

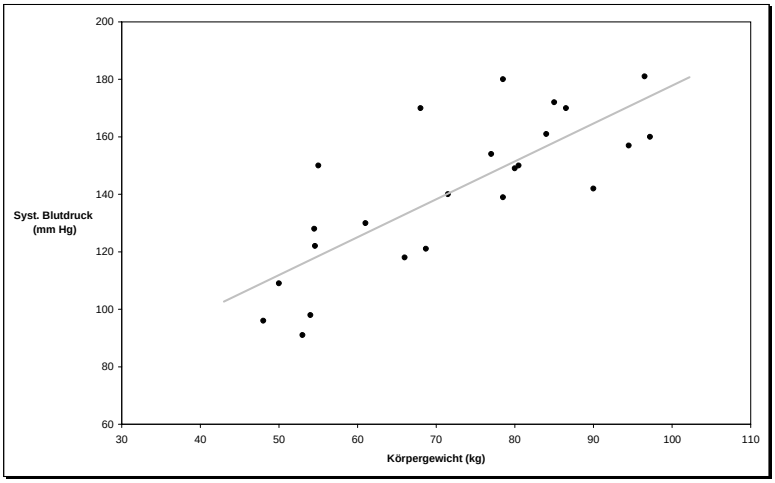
Aus der obigen Gleichung für $\hat{a}$ läßt sich erkennen, daß die Regressionsgerade stets durch den "Schwerpunkt" $(\bar{x}, \bar{y})$ der Punktewolke geht.

In unserem Beispiel errechnet sich die Regressionsgerade wie folgt:

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{24} x_i y_i - 24\bar{x}\bar{y}}{\sum_{i=1}^{24} x_i^2 - 24\bar{x}^2} = \frac{252075.6 - \frac{1732.7 \cdot 3388}{24}}{130764.33 - \frac{1732.7^2}{24}} \approx 1.32 \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} \approx \frac{3388}{24} - 1.32\frac{1732.7}{24} = 45.87 \\ \text{also:} \quad y &\approx 45.87 + 1.32x \end{aligned}$$

Die Geradengleichung zeigt an, daß der Wert des systolischen Blutdrucks im Mittel um ca. 1.32 mm Hg ansteigt, wenn der Wert des Körpergewichts um 1 kg zunimmt. Bei einer 70 kg schweren Person ist mit einem Blutdruck von $70 \cdot 1.32 + 45.87 \approx 138$ mm Hg zu rechnen.

Das Streuungsdiagramm inklusive Regressionsgerade sieht damit folgendermaßen aus:



Zur Quantifizierung der Stärke des linearen Zusammenhangs wird das Quadrat des Korrelationskoeffizienten von Pearson $r^2(X,Y)$ (auch *Bestimmtheitsmaß* genannt) verwendet. Es mißt den Anteil der Streuung der y-Werte, der durch die lineare Regression erklärt werden kann. Dies gilt wegen

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 &= \sum_{i=1}^{n}\left(y_i-\hat{a}-\hat{b}x_i\right)^2 = \sum_{i=1}^{n}\left(y_i-\bar{y}+\hat{b}\bar{x}-\hat{b}x_i\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\bar{y}-\hat{b}\left(x_i-\bar{x}\right)\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 - 2\hat{b}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(x_i-\bar{x}\right) + \hat{b}^2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 - 2\hat{b}\hat{b}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 + \hat{b}^2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 - \hat{b}^2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 - \frac{\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\right]^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \\
&= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\left\{1-\frac{\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\right]^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}\right\} \\
&= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\left(1-r^2\left(X,Y\right)\right)
\end{aligned}$$

Bemerkungen:

- Mit Hilfe der Regressionsanalyse gewonnene Aussagen über Ausprägungen eines Merkmals Y sind stets mit Vorsicht zu bewerten, da die lineare Modellgleichung nur eine (mehr oder weniger) genaue Annäherung an die Realität darstellt. Mit ausgefeilteren mathematischen Methoden ist allerdings eine Quantifizierung des Fehlers möglich.

- Die Regressionsgerade sollte in keinem Fall dazu benutzt werden, Vorhersagen für X-Werte außerhalb der Spannweite der beobachteten Werte $x_1, x_2, \ldots, x_n$ zu machen. Eine solche Extrapolation ist nicht gestattet, da keine Unterlagen über den Zusammenhang zwischen X und Y außerhalb der beobachteten Merkmalswerte vorliegen.

- Ebenso wie die Korrelationskoeffizienten sagt auch die Regressionsgerade **nichts** über einen kausalen Zusammenhang zwischen zwei Merkmalen aus.

Falls zwischen zwei Variablen zwar ein funktionaler, aber nicht linearer Zusammenhang vermutet wird, läßt sich in einigen Fällen die Methode der linearen Regression trotzdem anwenden, und zwar durch eine *Variablentransformation*. Erwartet man z.B. einen exponentiellen Zusammenhang zwischen zwei Merkmalen X und Y, d.h. vermutet man, daß Zahlen $a \in \mathbb{R}_+$, $b \in \mathbb{R}$ existieren, so daß näherungsweise gilt

$$Y = a\exp(bX),$$

so läßt sich das Problem der Bestimmung von a und b durch Logarithmieren auf eine lineare Gleichung zurückführen:

$$\ln(Y) = \ln(a) + bX$$

Auf diese Geradengleichung $\tilde{X} = \alpha + \beta\tilde{Y}$ mit den transformierten Parametern

$$\begin{aligned} \alpha &:= \ln(a) \\ \beta &:= b \end{aligned}$$

und den transformierten Variablen

$$\begin{aligned} \tilde{X} &:= X \\ \tilde{Y} &:= \ln(Y) \end{aligned}$$

läßt sich jetzt die bekannte Methode der kleinsten Quadrate zur Bestimmung von Schätzwerten $\hat{\alpha}, \hat{\beta}$ von α und β anwenden. Aus $\hat{\alpha}$ und $\hat{\beta}$ kann man dann durch Rücktransformation

$$\hat{a} = \exp(\hat{\alpha}) \quad \text{und} \quad \hat{b} = \hat{\beta}$$

Schätzwerte für die ursprünglich gesuchten Parameter der exponentiellen Gleichung ableiten. Einen Überblick über zwei weitere Zusammenhänge, die sich durch geeignete Transformationen in eine lineare Gleichung überführen lassen, liefert die folgende Tabelle:

Vermuteter Zusammenhang	Transformierte Gleichung	Transformierte Variablen	Transformierte Parameter
$Y = \frac{X}{a+bX}$	$\frac{1}{Y} = b + a\frac{1}{X}$	$\tilde{X} := \frac{1}{X}$ und $\tilde{Y} := \frac{1}{Y}$	$\alpha := b$ und $\beta := a$
$Y = aX^b$ $(a, X > 0)$	$\ln(Y) = \ln(a) + b\ln(X)$	$\tilde{Y} := \ln(Y)$ und $\tilde{X} := \ln(X)$	$\alpha := \ln(a)$ und $\beta := b$

Man beachte, daß bei der Durchführung der Regression an transformierten Variablen, alle in der Regression errechneten Parameter wie Varianzen, Kovarianzen und insbesondere auch Bestimmtheitsmaß auf die transformierten und **nicht** auf die ursprünglichen Merkmale Bezug nehmen.

Beispiel: Es seien 5 Merkmalspaare $(x_1, y_1), (x_2, y_2), ..., (x_5, y_5)$ wie folgt vorgegeben:

$$(1.5, 5.9), (2.4, 17.3), (3.7, 39.5), (5.3, 87.0), (10.4, 306.8)$$

Zwischen den zugehörigen Merkmalen X und Y werde ein Zusammenhang der Form

$$Y = aX^b,\ a \in \mathbb{R}_+, b \in \mathbb{R} \text{ geeignet gewählt,}$$

vermutet. Durch die oben beschriebenen Transformationen $\tilde{Y} = \ln(Y)$ und $\tilde{X} = \ln(X)$ errechnen sich die transformierten Merkmalspaare $(\tilde{x_1}, \tilde{y_1}), (\tilde{x_2}, \tilde{y_2}) \ldots, (\tilde{x_5}, \tilde{y_5})$ (gerundet) als

$$(0.41, 1.77), (0.88, 2.85), (1.31, 3.68), (1.67, 4.47), (2.34, 5.73)$$

Mit Hilfe der Methode der kleinsten Quadrate ergeben sich daraus die Schätzwerte $\hat{\alpha}, \hat{\beta}$ der Parameter α, β der transformierten Geradengleichung $\tilde{Y} = \alpha + \beta\tilde{X}$ als

$$\begin{aligned} \hat{\beta} &= \frac{Cov(\tilde{X}, \tilde{Y})}{s^2_{\tilde{X}}} = 2.046 \\ \hat{\alpha} &= \overline{\tilde{Y}} - \hat{\beta}\overline{\tilde{X}} = 0.995 \end{aligned}$$

Rücktransformation mit Hilfe der Exponentialfunktion liefert dann die Schätzwerte $\hat{a}, \hat{b}$ für die gesuchten Parameter a, b der Ausgangsgleichung:

$$\hat{a} = exp(\hat{\alpha}) = 2.705 \quad \text{und} \quad \hat{b} = \hat{\beta} = 2.046$$

Die gesuchte Näherungsfunktion hat demnach die Gestalt

$$Y = 2.705X^{2.046}$$

4.5 Statistischer Zusammenhang zwischen nominal skalierten Merkmalen

4.5.1 Der Kontingenzkoeffizient

Ausgangspunkt für unserer Überlegungen ist der Begriff der statistischen Unabhängigkeit (siehe Abschnitt 4.3). Nach der dort vorgestellten Definition werden zwei Merkmale X und Y als statistisch unabhängig bezeichnet, wenn sich ihre gemeinsame Verteilung aus dem Produkt der beiden Randverteilungen

$$f_{ij} = f_{i.} f_{.j} \text{ für } i = 1,2,\ldots,k \text{ und } j = 1,2,\ldots,l$$

berechnen läßt. In absoluten Häufigkeiten ausgedrückt würde das Unabhängigkeitskriterium

$$h_{ij} = \frac{h_{i.} h_{.j}}{n} =: E_{ij}$$

lauten. Dabei ist zu beachten, daß die *erwarteten Häufigkeiten* E_{ij} hypothetische Werte sind, die keineswegs ganzzahlig zu sein brauchen. Um das Ausmaß der Abhängigkeit zu quantifizieren, ist es plausibel, die Abweichungen

$$h_{ij} - E_{ij}$$

zu betrachten. Im allgemeinen kann man bei empirischen Verteilungen keine wirkliche Unabhängigkeit erwarten, sondern es werden stets mehr oder weniger große Abweichungen vorliegen. Je stärker die beobachteten Häufigkeiten von den erwarteten abweichen, umso größer wird der statistische Zusammenhang sein. Um eine Maßzahl dafür zu gewinnen, definiert man die *quadratischen Kontingenz* bzw. den *Chi-Quadrat-Koeffizienten QK* wie folgt:

$$QK := \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - E_{ij})^2}{E_{ij}}$$

Es sei bemerkt, daß die Summe aller Differenzen $h_{ij} - E_{ij}$ Null ist, da sich sowohl die beobachteten und als auch die erwarteten Häufigkeiten insgesamt zur Gesamtanzahl n aller Merkmalspaare (x_i, y_i) aufaddieren. Die Differenzen werden also vor der Summation quadriert, analog zur Berechnung der Standardabweichung. Im Fall vollkommener Unabhängigkeit ist die quadratische Kontingenz natürlich Null. In allen anderen Fällen ist sie positiv, und sie kann, wenn Abhängigkeit vorliegt, für große n sehr groß werden. Deswegen ist sie als Zusammenhangsmaß nicht besonders geeignet, sondern ein normiertes Maß wäre vorzuziehen. Dies erreicht man in zwei Schritten: Zunächst definiert man den *Kontingenzkoeffizienten*

$$KK := \sqrt{\frac{QK}{QK + n}}$$

KK verschwindet genau dann, wenn QK verschwindet. Mit steigendem QK wächst auch KK, erreicht aber den Wert Eins nicht ganz, sondern maximal den Wert $KK_{\max}$, der von der Größe

der zweidimensionalen Häufigkeitstabellen abhängt:

$$0 \leq KK \leq KK_{\max} = \sqrt{\frac{\min(k,l)-1}{\min(k,l)}} < 1$$

Deswegen korrigiert man den Kontingenzkoeffizienten ein weiteres Mal, um den *korrigierten Kontingenzkoeffizienten* KK^* zu erhalten, dessen Wert stets zwischen Null und Eins liegt:

$$KK^* := \frac{KK}{KK_{\max}} = \sqrt{\frac{QK \cdot \min(k,l)}{(QK+n)(\min(k,l)-1)}}$$

KK^* eignet sich am besten dafür die Stärke des Zusammenhanges verschiedener Kontingenztabellen zu vergleichen.

Bemerkungen:

- Der Kontingenzkoeffizient kann natürlich auch für ordinale Merkmale berechnet und sinnvoll interpretiert werden. Jedoch ist zu beachten, daß er nur angibt, wie stark der Zusammenhang ist, aber im Gegensatz zu den Korrelationskoeffizienten nichts über die Richtung des Zusammenhanges aussagt. Man kann aufgrund eines hohen Kontingenzkoeffizienten demnach nicht schließen, daß große Werte der einen Variablen tendenziell mit großen Werten der anderen Variablen einhergehen. Das liegt daran, daß bei der Berechnung der quadratischen Kontingenz nur das Nominalskalenniveau beachtet wird, Größen und Abstände der Merkmalswerte werden in den Formeln nicht berücksichtigt. Dies wird auch daran sichtbar, daß beliebige Umstellungen von Spalten oder Zeilen in der zweidimensionalen Häufigkeitstabelle die Kontingenzmaße nicht verändern.

- Der korrigierte Kontingenzkoeffizient einer Verteilung ist eins, wenn bei zwei Merkmalen mit je k Ausprägungen in jeder Zeile genau eine Spalte und in jeder Spalte genau eine Zeile mit Häufigkeiten besetzt ist und somit *vollkommene Abhängigkeit* besteht.

- Der Kontingenzkoeffizient sollte nicht benutzt werden, wenn die erwarteten Häufigkeiten teilweise sehr gering sind, da solche Zellen in der zweidimensionalen Häufigkeitstabelle zu einem sehr großen Wert des Kontingenzkoeffizienten führen können, unabhängig vom Wert der anderen Häufigkeiten. Nach einer Faustregel, die dem Statistiker W.G. Cochran zugeschrieben wird, sollten für 80% der Zellen der zweidimensionalen Häufigkeitstabelle die erwartete Häufigkeiten größer als 5 sein, und alle Zellen sollten eine erwartete Häufigkeit von größer als 1 besitzen. Man beachte, daß hier die nur die erwarteten, aber nicht die tatsächlichen Häufigkeiten eine Rolle spielen.

Beispiel: Eine Studie wurde durchgeführt, um einen möglichen Zusammenhang zwischen Koffeinkonsum und Familienstand bei Frauen vor der Geburt zu ermitteln (Martin und Bracken, 1987). Die Daten sind folgender Tabelle zusammengefaßt:

Wäre der Koffeinkonsum unabhängig vom Familienstand, so müßte die gemeinsame Verteilung etwa so aussehen:

	Koffeinkonsum (in mg/Tag)				
Familienstand	0	1-150	151-300	>300	$\sum$
Verheiratet	652	1537	598	242	3029
Geschieden, getrennt oder verwitwet	36	46	38	21	141
Alleinlebend	218	327	106	67	718
$\sum$	906	1910	742	330	3888

	Koffeinkonsum (in mg/Tag)				
Familienstand	0	1-150	151-300	>300	$\sum$
Verheiratet	705.8	1488.0	587.1	257.1	3029
Geschieden, getrennt oder verwitwet	32.9	69.3	26.9	12.0	141
Alleinlebend	167.3	352.7	137.0	60.9	718
$\sum$	906	1910	742	330	3888

Damit errechnet sich der Chi-Quadrat-Koeffizient QK als 51.66. Die untenstehende Tabelle zeigt den Beitrag jeder Zelle zum Chi-Quadrat-Koeffizienten.

	Koffeinkonsum (in mg/Tag)				
Familienstand	0	1-150	151-300	>300	$\sum$
Verheiratet	4.11	1.61	0.69	0.89	7.30
Geschieden, getrennt oder verwitwet	0.30	7.82	4.57	6.82	19.51
Alleinlebend	15.36	1.88	7.02	0.60	24.86
$\sum$	19.77	11.31	12.28	8.31	51.66

Der korrigierte Korrelationskoeffizient KK^* ergibt sich damit als

$$KK^* = \sqrt{\frac{51.66 \cdot 3}{(51.66 + 3888)(3-1)}} \approx 0.140$$

Es scheint also vielleicht eine gewisse, aber auf keinen Fall sehr ausgeprägte, Abhängigkeit von Koffeinkonsum und Familienstand vorzuliegen

5 Wahrscheinlichkeitsrechnung und schließende Statistik

Wie schon in Abschnit 1.1.2 erwähnt, beschäftigt sich die schließende Statistik damit, Erkenntnisse aus den Daten zu gewinnen, die über das Kollektiv, aus dem sie gewonnen worden sind, hinausgehen, d.h. mit den Methoden, die nötig sind, um von einer Stichprobe auf die Grundgesamtheit zurückzuschließen. Damit dies möglich ist, muß die Stichprobe ein unverzertes Abbild der Grundgesamtheit liefern, also *repräsentativ* sein. Diese wichtige Eigenschaft ist nur bei Zufallsstichproben gewährleistet, so daß die Wahrscheinlichkeitsrechnung die Grundlage der schließenden Statistik bildet.

5.1 Einführung in die Wahrscheinlichkeitsrechnung

5.1.1 Zufallsexperiment, Zufallsvariable und Zufallsereignisse

Die Wahrscheinlichkeitstheorie versucht die bei Experimenten mit zufälligem Ausgang (also bei sogenannten *Zufallsexperimenten*) herrschenden Gesetzmäßigkeiten mathematisch zu beschreiben. Beispiele für Zufallsexperimente sind

- das ein- oder mehrmalige Werfen einer Münze oder eines Würfels
- das Lotto- oder Roulette-Spiel

Ein Zufallsexperiment zeichnet sich grundsätzlich dadurch aus, daß das Ergebnis bei seiner Durchführung nicht mit Sicherheit vorhersehbar ist und es außerdem unter den gleichen Rahmenbedingungen beliebig oft wiederholbar ist.

Die Menge aller möglichen Ergebnisse eines Zufallsexperimentes heißt *Ergebnismenge* oder *Ereignisraum* und wird symbolisch in der Regel mit Ω bezeichnet, also

$$\Omega := \{e | e \text{ ist ein mögliches Ergebnis des Zufallsexperimentes}\}$$

Um eine einheitliche Beschreibung von Zufallsexperimenten zu ermöglichen, werden den Ergebnissen eines Zufallsexperimentes Zahlen zugeordnet, d.h. man betrachtet sogenannte *Zufallsvariablen*. Eine Zufallsvariable X ist demnach eine Abbildung der Ergebnismenge Ω in die reellen Zahlen, also

$$X : \Omega \mapsto \mathbb{R}$$

Dem Merkmal in der deskriptiven Statistik entspricht also die Zufallsvariable in der schließenden Statistik.

In der Wahrscheinlichkeitstheorie interessiert man sich für spezielle Ausgänge von Zufallsexperimenten, also für *Ereignisse*, die mathematisch wie folgt definiert werden:

Eine Teilmenge A der Ergebnismenge Ω, $A \subset \Omega$, heißt Ereignis.

Das Ereignis A tritt ein, wenn ein Ergebnis e beobachtet wird, das zu A gehört, wenn also gilt $e \in A$. Ereignisse werden häufig auch über Zufallsvariable definiert und haben dann die Form

$$\begin{aligned} [X = x] &:= \{e \in \Omega | X(e) = x\} \\ [X \in W] &:= \{e \in \Omega | X(e) \in W\} \end{aligned}$$

Beispiel: Das Zufallsexperiment bestehe aus dem zweimaligen Werfen einer Münze, dann gilt:

$$\Omega = \{(\text{Kopf, Kopf}), (\text{Zahl, Zahl}), (\text{Kopf, Zahl}), (\text{Zahl, Kopf})\}$$

Eine zugehörige Zufallsvariable $X : \Omega \mapsto \mathbb{R}$ läßt sich dann z.B. definieren als:

$$\begin{aligned} X(\text{Kopf, Kopf}) &= 1 \\ X(\text{Zahl, Zahl}) &= 2 \\ X(\text{Kopf, Zahl}) &= 3 \\ X(\text{Zahl, Kopf}) &= 4 \end{aligned}$$

Das Ereignis "Bei beiden Würfen dasselbe Ergebnis" läßt sich dann mittels X beschreiben durch

$$[X \in \{1, 2\}]$$

Die folgenden speziellen Ereignisse sind grundlegend:

- sicheres Ereignis, $A = \Omega$
- unmögliches Ereignis, $A = \emptyset$
- Elementarereignis, d.h. das Ereignis besteht nur aus einem Ergebnis, also $A = \{e\}$ für irgendein $e \in \Omega$

Aus mehreren Ereignissen eines Ergebnisraumes lassen sich durch logische Verknüpfungen (und/oder/nicht) neue Ereignisse bilden. Diesen Verknüpfungen sind mengentheoretische Operationen wie folgt zugeordnet:

$A \cap B$ (oder auch AB),	d.h. sowohl A als auch B treten ein
$A \cup B$,	d.h. entweder A oder B treten ein
$\bar{A}$,	d.h. A tritt nicht ein (Komplementärereignis)

Für die Mengenoperationen "$\cup$" und "$\cap$" gelten dieselben Rechenregeln wie für die Addition und Multiplikation, nämlich

$$\begin{array}{rcll} (A\cup B)\cup C = A\cup(B\cup C) & \text{sowie} & (A\cap B)\cap C = A\cap(B\cap C) & \text{(Assoziativgesetz)} \\ A\cup B = B\cup A & \text{sowie} & A\cap B = B\cap A & \text{(Kommutativgesetz)} \\ A\cap(B\cup C) & = & (A\cap B)\cup(A\cap C) & \text{(Distributivgesetz)} \end{array}$$

Zusätzlich gelten die deMorgan'schen Regeln

$$\begin{aligned} \overline{A\cup B} &= \bar{A}\cap\bar{B} \\ \overline{A\cap B} &= \bar{A}\cup\bar{B} \end{aligned}$$

5.1.2 Wahrscheinlichkeiten

Die Einführung des Wahrscheinlichkeitskonzeptes wird durch die Beobachtung des Verhaltens relativer Ereignishäufigkeiten f_n nahegelegt, wenn ein Zufallsexperiment n-mal hintereinander ausgeführt wird. Für ein beliebiges Ereignis $A \in \Omega$ wird also bei jeder Versuchsdurchführung festgehalten, ob es eingetreten ist oder nicht, um die relative Häufigkeit $f_n(A)$ zu erhalten:

$$f_n(A) = \frac{\text{Anzahl der Versuche, bei denen } A \text{ eingetreten ist}}{n}$$

Die relative Ereignishäufigkeit hat folgende Eigenschaften:

$$\begin{aligned} 0 \le\ & f_n(A) && \le 1 \\ f_n(\emptyset) &= && 0 \\ f_n(\Omega) &= && 1 \\ f_n(A\cup B) &= && f_n(A)+f_n(B)-f_n(A\cap B) \end{aligned}$$

Führt man Zufallsexperimente wie etwa das Werfen eines Würfels in sehr langen Versuchsreihen durch, so zeigt sich eine Stabilisierung der relativen Häufigkeiten $f_n(A)$ eines festen Ereignisses A, welche an das Konzept der Konvergenz in der Analysis erinnert (wenn man die Möglichkeit unbeschränkter Wiederholung des Experimentes unterstellt). Dieser Grenzwert scheint außerdem nicht von der speziellen Versuchsserie abzuhängen, sondern bei einer neuen Serie scheint die zugehörige Folge $f_n'(A)$ wieder gegen den gleichen Wert zu konvergieren (*empirisches Gesetz der großen Zahlen*). Es liegt nun nahe, den Grenzwert der Folge $f_n(A)$ die Wahrscheinlichkeit des Ereignisses A zu nennen (*statistischer Wahrscheinlichkeitsbegriff* nach Richard v. Mises). Da die vermutete Konvergenz letzten Endes aber nicht beobachtbar ist, steht dieser Wahrscheinlichkeitsbegriff aber auf tönernen Füßen.

Die Ausdifferenzierung des statistischen Wahrscheinlichkeitsbegriffes war in der Geschichte der Mathematik ein dorniger Weg. Man hat deswegen letztlich einen anderen Zugang gewählt: die *axiomatische Begründung des Wahrscheinlichkeitsbegriffes* nach Kolmogorov. Danach gehört zu jedem Zufallsexperiment eine Wahrscheinlichkeitsverteilung, d.h. eine Funktion P, die jedem

Ereignis $A \in \Omega$ eine reelle Zahl $P(A)$ - seine Wahrscheinlichkeit - zuordnet, wobei P über die folgenden Eigenschaften verfügt:

$$\begin{aligned} P(A) &\geq 0 \text{ für jedes Ereignis } A \\ P(A \cup B) &= P(A) + P(B), \text{ falls } A \cap B = \emptyset \\ P(\Omega) &= 1 \end{aligned}$$

Die axiomatische Definition trifft zunächst keine Aussage darüber, wie man in einer konkreten Fragestellung Wahrscheinlichkeiten von Ereignissen erhält, sondern gibt Methoden an die Hand, wie man mit diesen Wahrscheinlichkeiten (weiter) rechnet. Aus dem Axiomen von Kolomogorov lassen sich folgende weitere Eigenschaften der Wahrscheinlichkeitsverteilung herleiten:

$$\begin{aligned} (1) \qquad P\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} P(A_i), \text{ falls } A_i \cap A_j = \emptyset \text{ für } i \neq j,\ n \in \mathbb{N} \\ (2) \qquad P(\bar{A}) &= 1 - P(A) \\ (3) \qquad P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Formel (1) erhält man durch vollständige Induktion, Formel (2) gilt wegen

$$1 = P(\Omega) = P(A \cup \bar{A}) = P(A) + P(\bar{A}) \,,$$

und Formel (3) leitet sich aus folgenden Überlegungen her:

$$\begin{aligned} A &= (A \cap \bar{B}) \cup (A \cap B) \\ B &= (A \cap B) \cup (\bar{A} \cap B) \\ A \cup B &= (A \cap \bar{B}) \cup (A \cap B) \cup (\bar{A} \cap B) \end{aligned}$$

Da alle Vereinigungen paarweise disjunkt sind, ergibt sich daraus:

$$\begin{aligned} P(A) &= P(A \cap \bar{B}) + P(A \cap B) \\ P(B) &= P(A \cap B) + P(\bar{A} \cap B) \\ P(A \cup B) &= P(A \cap \bar{B}) + P(A \cap B) + P(\bar{A} \cap B) \\ \Rightarrow \qquad P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Die axiomatische Wahrscheinlichkeitsdefinition scheint, wie oben schon erwähnt, zunächst keine Hilfe zu geben, wie man in einer konkreten Situation Wahrscheinlichkeiten von Ereignissen berechnet. Viele Zufallsexperimente mit endlicher Ergebnismenge zeichen sich jedoch durch Symmetriebedingungen in der Hinsicht aus, daß kein Versuchsausgang vor den anderen bevorzugt ist, so daß also intuitiv jeder Versuchsausgang gleich wahrscheinlich scheint (z.B. beim Werfen eines idealen nicht getürkten (!) Würfels). Hier erhält man aus den Wahrscheinlichkeitsaxiomen die *Laplace-Definition der Wahrscheinlichkeit*, nämlich:

Besitzt ein Zufallsexperiment m verschiedene, gleich wahrscheinliche Ausgänge und sind g davon für das Ereignis A "günstig", dann gilt:

$$P(A) = \frac{g}{m} = \frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl der möglichen Fälle}}$$

Diese Formel ergibt sich wie folgt: Seien $A_1, A_2, \ldots, A_m$ Ereignisse mit den folgenden Eigenschaften:

- Die A_i, $i = 1,2,\ldots,m$, sind paarweise disjunkt, also $A_i \cap A_j = \emptyset$, für $i \neq j$.
- Bei der Versuchsdurchführung tritt stets eines der A_i ein.
- Alle Ereignisse A_i sind gleich wahrscheinlich.

Sei nun A ein Ereignis, das sich aus den A_i zusammensetzt, also

$$A = A_{j_1} \cup A_{j_2} \cup \ldots \cup A_{j_g}, \text{ wobei o.B.d.A gelte } A_{j_i} \cap A_{j_l} = \emptyset, \text{ für } j_i \neq j_l.$$

Dann gilt:

$$\begin{aligned} 1 &= P(\Omega) = P(A_1 \cup A_2 \cup \ldots \cup A_m) = \sum_{i=1}^{m} P(A_i) = m \cdot P(A_1) \\ \Rightarrow \quad P(A_1) &= \frac{1}{m} = P(A_2) = \ldots = P(A_m) \\ \Rightarrow \quad P(A) &= P(A_{j_1} \cup A_{j_2} \cup \ldots \cup A_{j_g}) = \sum_{i=1}^{g} P(A_{j_i}) = \frac{g}{m} \end{aligned}$$

Beispiel: Um die Wahrscheinlichkeit zu errechnen, beim einmaligen Werfen eines (idealen) Würfels eine gerade Zahl zu erhalten (Ereignis A), sei mit A_i, $i = 1,2,\ldots,6$, das Ereignis bezeichnet, die Augenzahl i zu werfen. Dann erfüllen die A_i alle obigen Voraussetzungen, und es gilt infolgedessen:

$$\begin{gathered} A = A_2 \cup A_4 \cup A_6 \\ \Rightarrow \quad g = 3,\ m = 6 \text{ und somit } P(A) = \tfrac{3}{6} = \tfrac{1}{2} \end{gathered}$$

Eine "Versöhnung" des statistischen und des axiomatischen Wahrscheinlichkeitsbegriffes liefert das sogenannte *Gesetz der großen Zahlen* von Jakob Bernoulli. Dazu betrachtet man die n-malige Wiederholung eines Zufallsexperimentes, bei dem nur das Eintreten eines einzelnen Ereignisses A von Interesse ist. Vor der Durchführung des Versuches steht die relative Häufigkeit $f_n(A)$ des Eintritts von A eine Zufallsvariable dar, deren konkrete Realisierung noch unbekannt ist. Mit Hilfe von $f_n(A)$ lassen sich daher Ereignisse formulieren und Wahrscheinlichkeiten berechnen. In dieser Situation läßt sich dann mathematisch beweisen, daß für jede noch so kleine Zahl $\varepsilon > 0$ gilt:

$$\begin{gathered} \lim_{n\to\infty} P(|f_n(A) - P(A)| > \varepsilon) = 0 \\ \text{bzw. äquivalent} \\ \lim_{n\to\infty} P(|f_n(A) - P(A)| < \varepsilon) = 1 \end{gathered}$$

Das heißt, die relative Eintrittshäufigkeit von A konvergiert mit Wahrscheinlichkeit 1 gegen die Wahrscheinlichkeit von A.

5.2 Theoretische Wahrscheinlichkeitsverteilungen

Für das Studium von Wahrscheinlichkeitsverteilungen ist es zweckmäßig, die Ausgänge von Zufallsexperimenten durch Zahlen zu beschreiben; dazu dient das bereits eingeführte Konzept von Zufallsvariablen. Für Zufallsvariablen gelten die gleichen Einteilungsprinzipien wie für Merkmale in der deskriptiven Statistik, d.h. in Abhängigkeit von den möglichen Ausprägungen unterscheiden wir zwischen nominal/ordinal/intervallskalierten und absolutskalierten Zufallsvariablen bzw. zwischen diskreten und stetigen Zufallsvariablen. Für diskreten Zufallsvariable wählt man in der Regel $\mathbb{N}$, $\mathbb{N}_0$ oder $\mathbb{Z}$ als Wertebereich, für stetige Zufallsvariable dagegen $\mathbb{R}$ oder $\mathbb{R}_+$.

5.2.1 Wahrscheinlichkeitsverteilungen diskreter Zufallsvariablen

Eine diskrete Wahrscheinlichkeitsverteilung wird durch die Angabe des Wertebereiches $\Omega_X := X(\Omega)$ und der endlich oder abzählbar vielen Elementarwahrscheinlichkeiten vollständig beschrieben. Dabei werden Analoga zu den aus der deskriptiven Statistik bekannten Begriffen der Häufigkeitsverteilung und der empirischen Verteilungsfunktion wie folgt gebildet:

Es sei X eine diskrete Zufallsvariable mit endlichem oder abzählbar unendlichem Wertebereich $\Omega_X = \{x_1, x_2, \ldots, x_i, x_{i+1}, \ldots\}$. Dann heißt

- die Funktion f, die jedem Elementarereignis $[X = x_i]$ (kurz: $\{x_i\}$) seine Wahrscheinlichkeit zuordnet,

  $$f(x_i) = p_i = P([X = x_i]),\ i = 1, 2, 3, \ldots$$

  die *(Wahrscheinlichkeits)dichte* von X

- die Funktion $F : \mathbb{R} \mapsto [0, 1]$, definiert durch

  $$F(x) = P(X \leq x) = \sum_{i: x_i \leq x} p_i$$

  die *(theoretische) Verteilungsfunktion* von X

Es gilt natürlich: $\sum_i p_i = 1$

Beispiel 5.2.1.1: Beim dreimaligen Werfen einer Münze gibt es acht gleichwahrscheinliche Elementarereignisse $e_1, e_2, \ldots, e_8$. Definiert man als Zufallsvariable $X =$ Anzahl Kopf, so erhält man folgende Ausprägungen und Wahrscheinlichkeiten ($Z =$ Zahl, $K =$ Kopf):

Elementarereignis	Wahrscheinlichkeit $P(e_i)$	Anzahl Kopf	Wahrscheinlichkeit p_i
$e_1 = ZZZ$	$P(e_1) = 0.125$	$x_1 = 0$	$f(x_1) = P(X = x_1) = 0.125$
$e_2 = ZZK$	$P(e_2) = 0.125$	$x_2 = 1$	$f(x_2) = P(X = x_2) = 0.375$
$e_3 = ZKZ$	$P(e_3) = 0.125$		
$e_4 = KZZ$	$P(e_4) = 0.125$		
$e_5 = KKZ$	$P(e_5) = 0.125$	$x_3 = 2$	$f(x_3) = P(X = x_3) = 0.375$
$e_6 = ZKK$	$P(e_6) = 0.125$		
$e_7 = KZK$	$P(e_7) = 0.125$		
$e_8 = KKK$	$P(e_8) = 0.125$	$x_4 = 3$	$f(x_4) = P(X = x_4) = 0.125$

Die Verteilungsfunktion F einer diskreten Zufallsvariable X ist also eine monoton wachsende Treppenfunktion mit $0 \leq F(x) \leq 1$, die bei den Werten x_i Sprünge der Höhe $P(X = x_i) = p_i$ aufweist:

$$\begin{aligned} F(x_i) &= P(X \leq x_i) = P(X < x_i) + P(X = x_i) = F(x_{i-1}) + p_i \\ \Rightarrow \quad p_i &= F(x_i) - F(x_{i-1}) \end{aligned}$$

Sie verfügt demnach über dieselben Eigenschaften wie die empirische Verteilungsfunktion eines diskreten Merkmals.

In Anwendungen werden meistens die Wahrscheinlichkeiten für sogenannte Intervallereignisse benötigt, also für Ereignisse der Form

$$[a < X < b],\ [a < X \leq b],\ [a \leq X < b],\ [a \leq X \leq b]$$

Diese Wahrscheinlichkeiten lassen sich mit Hilfe der Wahrscheinlichkeitsfunktion wie folgt berechnen:

$$P(a < X \leq b) = \sum_{i:a<x_i\leq b} p_i = \sum_{i:x_i\leq b} p_i - \sum_{i:x_i\leq a} p_i = F(b) - F(a)$$

Bei den anderen Intervallereignissen ist zu unterscheiden, ob einer der höchstens abzählbar vielen Punkte des Wertebereichs von Xmit dem jeweiligen Intervallendpunkt übereinstimmt. Wenn weder a noch b mit einem der x_i zusammenfällt, so sind die Wahrscheinlichkeiten aller vier oben aufgeführten Intervallereignisse gleich groß, da dann gilt $P(X = a) = P(X = b) = 0$. Wenn aber z.B. b und x_k übereinstimmen, so ist $P(X = x_k) = P(X = b) = p_k > 0$, und es gilt:

$$P(a < X < b) = P(a < X \leq b) - P(b) = F(b) - F(a) - p_k$$

bzw. analog

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X \leq b) + P(a) = F(b) - F(a) + p_j, \text{ falls } a = x_j \\ P(a \leq X < b) &= P(a < X \leq b) + P(a) - P(b) = F(b) - F(a) + p_j - p_k, \text{ falls } a = x_j,\ b = x_k \end{aligned}$$

5.2.2 Wahrscheinlichkeitsverteilungen stetiger Zufallsvariablen

Die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen ist bekannt, wenn man die Wahrscheinlichkeiten $P(a < X \leq b)$ für jedes Intervall $(a,b] \subset \mathbb{R}$ kennt. An die Stelle der Wahrscheinlichkeitsfunktion tritt hier die Dichtefunktion f, die wie folgt definiert ist:

Sei X eine stetige Zufallsvariable und sei $f : \mathbb{R} \mapsto \mathbb{R}$ eine integrierbare Funktion mit der Eigenschaft

$$P(a < X \leq b) = \int_a^b f(x)\mathrm{d}x$$

Dann heißt f Wahrscheinlichkeitsdichte der Verteilung von X.

Die Wahrscheinlichkeitsdichte hat folgende Eigenschaften:

$$f(x) \geq 0 \quad \text{und} \quad \int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

Dichten von in der Praxis auftretenden Verteilungen stetiger Zufallsvariablen sind in der Regel stetig oder zumindestens stückweise stetig.

Die Verteilungsfunktion F einer stetigen Zufallsvariable ist **keine** Treppenfunktion, sondern eine Funktion, die mit der Dichte in der folgenden Weise zusammenhängt:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

Ist f in x stetig, so folgt nach dem Hauptsatz der Integral- und Differentialrechnung:

$$F'(x) = \left(\int_{-\infty}^{x} f(t)\mathrm{d}t\right)' = f(x)$$

In allen Punkten, in denen die Dichtefunktion stetig ist, ergibt sich die Dichte also als erste Ableitung der Verteilungsfunktion. Wie bei diskreten Zufallsvariablen ist die Verteilungsfunktion monoton wachsend und nimmt Werte zwischen 0 und 1 an. Außerdem gilt:

$$\lim_{x\to\infty} F(x) = 1 \text{ und } \lim_{x\to-\infty} F(x) = 0$$

Kennt man die Verteilungsfunktion, so lassen sich die Wahrscheinlichkeiten für Intervallereignisse sofort berechnen, denn:

Wegen $(-\infty,a] \cup (a,b] = (-\infty,b]$ folgt

$$\begin{aligned} & P(X \leq a) + P(a < X \leq b) &&= P(X \leq b) \\ \Rightarrow \quad & P(a < X \leq b) &&= P(X \leq b) - P(X \leq a) = F(b) - F(a) \end{aligned}$$

Im Gegensatz zu diskreten Zufallsvariablen ist es hier allerdings gleichgültig, ob die Intervallgrenzen a und b dazugenommen werden oder nicht, da die Wahrscheinlichkeit einzelner Punkte bei stetigen Verteilungen stets verschwindet:

$$P(X = a) = 0, \text{ für alle } a \in \mathbb{R}$$

Das heißt, wir erhalten:

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

Da sich die Wahrscheinlichkeiten durch die Differenz $F(b) - F(a)$ wesentlich einfacher berechnen lassen als durch das Integral, werden für stetige Verteilungen in der Regel nur die Werte der Verteilungsfunktion tabelliert. Andererseits sind die Formeln für die Dichtefunktion in der Regel einfacher darzustellen, so daß konkrete stetige Verteilungen meistens durch die Angabe der Dichte definiert werden.

Beispiel 5.2.2.1: Die stetige Zufallsvariable X sei definiert als "Verspätung der U-Bahn Linie 3 an der Haltestelle Schweizer Platz" (in Minuten). X habe die folgenden Dichtefunktion

$$f(x) = \begin{cases} \frac{1}{2} - \frac{x}{8}, & \text{für } 0 \leq x \leq 4 \\ 0, & \text{sonst} \end{cases}$$

Wie groß ist die Wahrscheinlichkeit, daß sich die U-Bahn mehr als eine, aber nicht mehr als zwei Minuten verspätet?

Wir prüfen zunächst nach, daß es sich bei f wirklich um eine Dichtefunktion handelt: $f(x) \geq 0$ ist offensichtlich für alle $x \in \mathbb{R}$ erfüllt und außerdem gilt

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_0^4 (0.5 - 0.125x)\mathrm{d}x = \left[0.5x - \frac{0.125}{2}x^2\right]_0^4 = 2 - 1 = 1$$

Für die Verteilungsfunktion F gilt für $x \leq 4$:

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t = \int_0^x (0.5 - 0.125t)dt = \left[0.5t - \frac{0.125}{2}t^2\right]_0^x = 0.5x - 0.0625x^2$$

Bezieht man noch die Bereiche ein, in denen f verschwindet, so erhält man insgesamt:

$$F(x) = \begin{cases} 0, & \text{für } x < 0 \\ 0.5x - 0.0625x^2, & \text{für } 0 \leq x \leq 4 \\ 1, & \text{für } x > 4 \end{cases}$$

Die gesuchte Wahrscheinlichkeit ergibt sich damit als

$$P(1 \leq X \leq 2) = F(2) - F(1) = 0.75 - 0.4375 = 0.3125$$

5.3 Erwartungswert und Varianz von Zufallsvariablen

Wie die empirischen Verteilungen in der deskriptiven Statistik, so lassen sich auch die theoretischen Wahrscheinlichkeitsverteilungen von Zufallsvariablen durch Maßzahlen charakterisieren. Dem arithmetischen Mittel als Lageparameter entspricht dabei der Erwartungswert. Als Streuungsparameter wird bevorzugt die (theoretische) Varianz verwendet, die das Analogon zur empirischen Varianz bildet.

Sei X eine diskrete Zufallsvariable mit Wertebereich $\Omega_X = \{x_1, x_2, \ldots\}$ und zugehöriger Wahrscheinlichkeitsfunktion f. Dann heißt

$$E(X) := \sum_i x_i f(x_i) = \sum_i x_i p_i$$

der Erwartungswert von X. Der Erwartungswert wird in der mathematischen Literatur auch häufig mit dem Buchstaben μ bezeichnet. Die Berechnung des Erwartungswertes erfolgt also ganz analog zu derjenigen des arithmetischen Mittels, wobei nur die relativen Häufigkeiten durch die Wahrscheinlichkeiten ersetzt worden sind und die Summe unter Umständen (im Falle eines abzählbar unendlichen Wertebereiches) aus unendlich vielen Summanden besteht.

Die Varianz $Var(X)$ einer diskreten Zufallsvariable X, die auch häufig mit dem Buchstaben σ^2 bezeichnet wird, ist definiert als:

$$Var(X) := \sum_i (x_i - E(X))^2 f(x_i) = \sum_i (x_i - \mu)^2 p_i$$

Auch hier herrscht große Ähnlichkeit zur entsprechenden Formel für die (empirische) Varianz s^2 aus der deskriptiven Statistik: Das arithmetische Mittel wurde durch den Erwartungswert und die relativen Häufigkeiten wurden durch die Wahrscheinlichkeiten ersetzt. Für die in Beispiel 5.2.1.1 definierte Zufallsvariable errechnen sich Erwartungswert μ und Varianz σ^2 als:

$$\begin{aligned}
\mu &= \sum_{i=1}^{4} x_i p_i = 0 \cdot 0.125 + 1 \cdot 0.375 + 2 \cdot 0.375 + 3 \cdot 0.125 = 1.5 \\
\sigma^2 &= \sum_{i=1}^{4} (x_i - \mu)^2 p_i \\
&= (0 - 1.5)^2 \cdot 0.125 + (1 - 1.5)^2 \cdot 0.375 + (2 - 1.5)^2 \cdot 0.375 + (3 - 1.5)^2 \cdot 0.125 \\
&= 0.75
\end{aligned}$$

Bei stetigen Zufallsvariablen werden Erwartungswert und Varianz über Integrale wie folgt definiert:

$$\begin{aligned}
\mu = E(X) &:= \int_{-\infty}^{\infty} x f(x) \mathrm{d}x \\
\sigma^2 &:= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \mathrm{d}x
\end{aligned}$$

Die Standardabweichung σ erhält man wie in der deskriptiven Statistik als Wurzel der Varianz

$$\sigma = \sqrt{\sigma^2}$$

Für die in Beispiel 5.2.2.1 definierte Zufallsvariable errechnen sich Erwartungswert μ und Varianz σ^2 als:

$$\mu = \int_{-\infty}^{\infty} x f(x) = \int_0^4 x \left(\frac{1}{2} - \frac{1}{8} x \right) \mathrm{d}x = \left[\frac{1}{4} x^2 - \frac{1}{24} x^3 \right]_0^4$$

$$
\begin{aligned}
&= 4-\frac{8}{3}=\frac{4}{3}\\
\sigma^2 &= \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\mathrm{d}x=\int_0^4\left(x-\frac{4}{3}\right)^2\left(\frac{1}{2}-\frac{1}{8}x\right)\mathrm{d}x\\
&= \int_0^4\left(x^2-\frac{8}{3}x+\frac{16}{9}\right)\left(\frac{1}{2}-\frac{1}{8}x\right)\mathrm{d}x=\int_0^4-\frac{1}{8}x^3+\frac{5}{6}x^2-\frac{14}{9}x+\frac{8}{9}\mathrm{d}x\\
&= \left[-\frac{1}{32}x^4+\frac{5}{18}x^3-\frac{7}{9}x^2+\frac{8}{9}x\right]_0^4=-8+\frac{160}{9}-\frac{112}{9}+\frac{32}{9}\\
&= \frac{8}{9}\approx 0.89
\end{aligned}
$$

Auch die folgenden vereinfachten Formeln zur Berechnung der Varianz sind in ähnlicher Weise bereits in der deskriptiven Statistik hergeleitet worden:

Im Falle einer diskreten Zufallsvariable X:

$$
\begin{aligned}
\sigma^2 &= \sum_i (x_i-\mu)^2 f(x_i)\\
&= \sum_i x_i^2 f(x_i)-2\mu\sum_i x_i f(x_i)+\mu^2\sum_i f(x_i)\\
&= \sum_i x_i^2 f(x_i)-2\mu\cdot\mu+\mu^2\cdot 1=\sum_i x_i^2 f(x_i)-\mu^2
\end{aligned}
$$

Im Falle einer stetigen Zufallsvariable X:

$$
\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\mathrm{d}x\\
&= \int_{-\infty}^{\infty}x^2 f(x)\mathrm{d}x-2\mu\int_{-\infty}^{\infty}x f(x)\mathrm{d}x+\mu^2\int_{-\infty}^{\infty}f(x)\mathrm{d}x\\
&= \int_{-\infty}^{\infty}x^2 f(x)\mathrm{d}x-2\mu\cdot\mu+\mu^2\cdot 1=\int_{-\infty}^{\infty}x^2 f(x)\mathrm{d}x-\mu^2
\end{aligned}
$$

Es gilt also stets:

$$Var(X)=E(X^2)-E(X),$$

wobei X^2 die Zufallsvariable bezeichne, die genau dann den Wert x^2 annimmt, wenn X den Wert x annimmt.

5.3.1 Die Normalverteilung $N(\mu,\sigma^2)$ und der zentrale Grenzwertsatz

Wie schon in Kapitel 2.4.3 erwähnt, stellt die Normalverteilung den Prototyp einer eingipfeligen und symmetrischen Verteilung dar. Die Familie der Normalverteilungen gehört zu den mit Abstand wichtigsten statistischen Verteilungsmodellen überhaupt. Die Normalverteilung wurde erstmals im Jahr 1733 in einer Schrift von deMoivre erwähnt; verbunden ist sie aber mit dem Namen von Gauß, der sie bei der Ausgleichung der Fehler astronomischer Messungen systematisch anwendete und um 1810 grundlegende Arbeiten zu diesem Thema veröffentlichte.

Jede Normalverteilung ist eindeutig bestimmt durch ihren Erwartungswert μ und ihre Varianz σ^2 (abgekürzt:$N(\mu,\sigma^2)$-Verteilung); ihre Dichtefunktion f bzw. Verteilungsfunktion F haben nämlich folgende Gestalt:

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2\right) \\ &= \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \text{ wobei } \varphi(x) := \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right) \\ \Rightarrow \quad F(x) &= \int_{-\infty}^{x}\frac{1}{\sigma}\varphi\left(\frac{t-\mu}{\sigma}\right)\mathrm{d}t \\ &= \frac{1}{\sigma}\Phi\left(\frac{x-\mu}{\sigma}\right), \text{ wobei } \Phi(x) := \int_{-\infty}^{x}\varphi(t)\mathrm{d}t \end{aligned}$$

$\varphi(x)$ bzw. $\Phi(x)$ sind hierbei die Dichte- bzw. Verteilungsfunktion der sogenannten *Standardnormalverteilung* $N(0,1)$mit Erwartungswert $\mu = 0$ und Varianz $\sigma^2 = 1$. Da sich jede normalverteilte Zufallsvariable X durch die Transformation $\tilde{X} = \frac{X-\mu}{\sigma}$ in eine standardnormalverteilte Zufallsvariable umwandeln läßt, kann man sich bei der Untersuchung der Normalverteilung in der Regel auf die Standarnormalverteilung beschränken. Da für das Integral der Verteilungsfunktion keine einfache Stammfunktion existiert, liegen die Werte von $\Phi(x)$ tabelliert vor (siehe Appendix A). Hierbei reicht es aus, nur die Werte $\Phi(x)$ für $x \geq 0$ aufzuführen, da wegen der Symmetrie der Standardnormalverteilung um den Erwartungswert $\mu = 0$ gilt:$\Phi(-x) = 1 - \Phi(x)$

Beispiel: Angenommen das Gewicht von Hühnereiern sei eine normalverteilte Zufallsvariable mit Erwartungswert $\mu = 55$g und Standardabweichung $\sigma = 2$g. Gesucht sei die Wahrscheinlichkeit, daß ein zufällig ausgewähltes Hühnerei

1. schwerer als 56 g ist.
2. mindestens 53 g und höchstens 58 g wiegt.

Da $\tilde{X} := \frac{X-55}{2}$ standardnormalverteilt ist, ergeben sich die gesuchten Wahrscheinlichkeiten wie folgt:

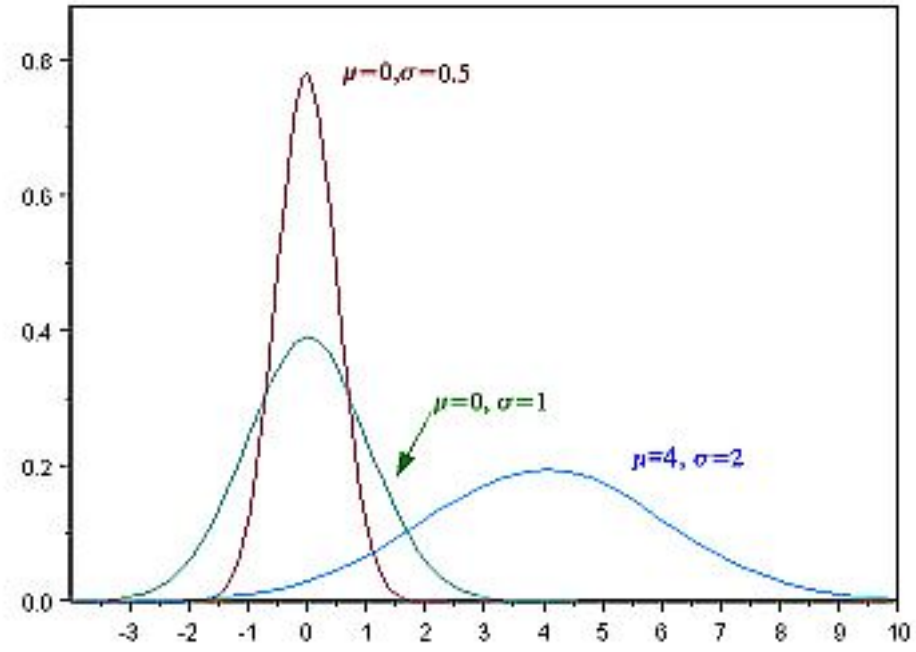
$$\begin{aligned} 1. \qquad P(X > 56) &= P\left(\frac{X-55}{2} > \frac{56-55}{2}\right) = P\left(\tilde{X} > \frac{1}{2}\right) = 1 - P\left(\tilde{X} \leq \frac{1}{2}\right) \\ &= 1 - \Phi\left(\frac{1}{2}\right) = 1 - 0.6915 = 0.3085 \\ 2. \quad P(53 \leq X \leq 58) &= P(X \leq 58) - P(X < 53) = P(X \leq 58) - P(X \leq 53) \\ &= P\left(\tilde{X} \leq \frac{3}{2}\right) - P(\tilde{X} \leq -1) = P\left(\tilde{X} \leq \frac{3}{2}\right) - \left(1 - P(\tilde{X} \leq 1)\right) \\ &= \Phi\left(\frac{3}{2}\right) - 1 + \Phi(1) = 0.9332 - 1 + 0.8413 = 0.7745 \end{aligned}$$

D.h. bei zufälliger Auswahl wird in ca. 31% der Fälle das Hühnerei über ein Gewicht von mehr als 56 g und in ca. 77% der Fälle über ein Gewicht zwischen 53 und 58 g verfügen.

Die Normalverteilung verfügt über folgende charakteristische Eigenschaften:

- Die Dichtefunktion nimmt ihr Maximum beim Erwartungswert$x = \mu$ an und besitzt bei $x = \mu + \sigma$ und $x = \mu - \sigma$ je einen Wendepunkt. Für $x \to \pm\infty$ schmiegt sich die Kurve von oben an die x-Achse an. Die Verteilung ist also durch eine mit der Entfernung von Erwartungswert abnehmende Konzentration gekennzeichnet.
- Die Höhe der Dichtefunktion an der Maximalstelle $x = \mu$ ist proportional zum Quotienten $\frac{1}{\sigma}$, d.h. je größer σ ist, desto flacher verläuft die Dichte und desto größer ist die Streuung.
- Die Verteilungsfunktion hat einen S-förmigen Verlauf mit einem Wendepunkt bei $x = \mu$.

Die untenstehende Graphik zeigt den Verlauf der Dichtefunktion der Normalverteilung für verschiedene Werte von μ und σ.



Die große Bedeutung der Normalverteilung leitet sich daraus ab, daß viele empirische Verteilungen durch sie beschrieben werden können. Wenn die Normalverteilung auch nicht ganz in dem Ausmaß wie früher angenommen den Regelfall einer stetigen Verteilung darstellt, so ist sie doch ohne Frage die wichtigste Verteilung der mathematischen Statistik. Die herausragende Stellung der Normalverteilung liegt insbesondere auch darin begründet, daß sie in vielen Fällen

als Grenzverteilung auftritt. Dies besagt der immens wichtige *Zentrale Grenzwertsatz*. Bevor wir diesen Satz mathematisch formulieren können, benötigen wir zunächst noch folgende Definition der *stochastischen Unabhängigkeit* von Zufallsvariablen, die ganz analog zur statistischen Unabhängigkeit von Merkmalen in der deskriptiven Statistik erfolgt:

Zwei Zufallsvariablen X und Y heißen (stochastisch) unabhängig, wenn für alle Ereignisse $A \in \Omega_X$ und $B \in \Omega_Y$ gilt:

$$P([X \in A] \cap [Y \in B]) = P([X \in A]) \cdot P([Y \in B])$$

Analog zur statistischen Unabhängigkeit von Merkmalen bedeutet dies anschaulich, daß der Wert der Zufallsvariablen X keinerlei Auswirkungen auf den Wert hat, den die Zufallsvariable Y annimmt, sich die beiden Zufallsvariablen also NICHT gegenseitig beeinflussen. Stochastisch unabhängige Variablen zeichnen sich ähnlich wie statistisch unabhängige Merkmale durch die Additivität ihrer Varianzen aus, d.h. sind X und Y zwei unabhängige Zufallsvariable, so gilt:

$$Var(X+Y) = Var(X) + Var(Y)$$

Der Zentrale Grenzwertsatz lautet nun wie folgt:

Sei $X_1, X_2, X_3, \ldots$ eine Folge unabhängiger und identisch verteilter Zufallsvariablen mit Erwartungswert $E(X_i) = \mu$ und Varianz $Var(X_i) = \sigma^2$, $i \in \mathbb{N}$. Dann konvergiert die Verteilung der standardisierten Summen Z_n der X_i

$$Z_n := \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$$

gegen eine Standardnormalverteilung, also:

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x)$$

Für die unstandardisierte Summe folgt daraus für große Werte von n

$$P(\sum_{i=1}^{n} X_i \leq x) \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

Das heißt: Völlig unabhängig davon, wie die eigentliche Verteilung der Zufallsvariablen X_i, $i = 1, 2, \ldots, n$ beschaffen ist, kann man für hinreichend großes n zur Analyse der Summe $X_1 + X_2 + \ldots + X_n$ oder auch des arithmetischen Mittels $\frac{1}{n}(X_1 + X_2 + \ldots + X_n)$ approximativ eine Normalverteilung unterstellen.

Immens wichtig ist auch folgende Invarianzeigenschaft der Normalverteilung:

Seien $X_1, X_2, \ldots, X_n$, $n \in \mathbb{N}$ beliebig, unabhängige, normalverteilte Zufallsvariablen, dabei sei X_i $N(\mu_i, \sigma_i^2)$-verteilt. Dann ist auch die Summe $X_1 + X_2 + \ldots X_n$ normalverteilt und zwar mit Erwartungswert $\mu = \mu_1 + \mu_2 + \ldots \mu_n$ und Varianz $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots \sigma_n^2$.

5.4 Schließende Statistik

5.4.1 Zufallsstichproben und Stichprobenfunktionen

Wir wollen uns jetzt der Frage zuwenden,wie man mit Hilfe der Wahrscheinlichkeitstheorie aus den Ergebnissen einer zufälligen Stichprobe auf die Eigenschaften der Grundgesamtheit zurückschließen kann. Das heißt, wir beschäftigen uns im folgenden stets mit *Zufallsstichproben*, die sich dadurch auszeichnen, daß jedes Element aus der Grundgesamtheit die gleiche Chance hat, gezogen zu werden. Falls die Stichprobenelemente außerdem unabhängig voneinander aus der Grundgesamtheit entnommen werden, heißt die Zufallsstichprobe auch *einfach*.

Ist X ein im Rahmen einer statistischen Untersuchung interessierendes Merkmal, so liefert eine Stichprobe vom Umfang n Stichprobenwerte $x_1, x_2, \ldots, x_n$, die sich als Realisierungen von n identisch verteilten (und im Falle einer einfache Stichprobe auch unabhängigen) Zufallsvariablen $X_1, X_2, \ldots, X_n$ interpretieren lassen. Parameter zur Charakterisierung der empirischen Verteilung des Merkmals wie z.B. Mittelwert, Varianz oder relative Häufigkeit, die aus den Stichprobenwerten ermittelt werden können, sind damit ebenfalls Realisierungen von Zufallsvariablen, die als Funktionen der $X_1, X_2, \ldots, X_n$ gewonnen werden (*Stichprobenfunktionen*).

Bei Anwendung von Stichprobenverfahren wird nun versucht, mit Hilfe geeignet gewählter Stichprobenfunktionen Rückschlüsse auf die entsprechenden Parameter der Grundgesamtheit zu ziehen. Dazu müssen die Verteilungen der Stichprobenfunktionen bekannt sein, die natürlich von der Verteilung der X_i abhängen. Asymptotisch ergibt sich hierbei häufig eine Normalverteilung (hier sei an das Gesetz der großen Zahlen erinnert) oder zumindest eine Verteilung, die mit der Normalverteilung in Verbindung steht (wie z.B. die Lognormal-Verteilung, bei der nicht die Werte selbst, aber die logarithmierten Werte normalverteilt sind).

5.4.2 Hauptsatz und Grundaufgaben der schließenden Statistik

Das Fundament der schließenden Statistik ist der sogenannte Hauptsatz der Statistik, der wie folgt lautet:

Ist X eine Zufallsvariable mit Verteilungsfunktion F, so konvergiert die empirische Verteilungsfunktion $\hat{F}_n$ einer Stichprobe von n unabhängigen und gemäß F identisch verteilten Werten $X_1, X_2, \ldots, X_n$ für jedes $x \in \mathbb{R}$ mit wachsendem Stichprobenumfang n gegen F.

Mit Hilfe der theoretischen Verteilungsfunktion F lassen sich dann - hierin liegt die Bedeutung des Satzes - Wahrscheinlichkeiten berechnen, wobei insbesondere Wahrscheinlichkeiten für Fehler von Interesse sind, die bei Entscheidungen unter Unsicherheit unausweichlich auftreten. Die parametrisierten Verteilungsmodelle bilden hierbei eine Brücke zwischen der deskriptiven und der schließenden Statistik. Sie erlauben es nämlich vielfach, die komplexe Approximation der empirischen Verteilungsfunktion durch eine handhabbare theoretische Verteilungsfunktion, die nach dem Hauptsatz für große n erlaubt ist, zu vermeiden, und sich auf die vergleichsweise einfache Ersetzung von Verteilungsparametern wie z.B. Erwartungswert,Varianz oder Trefferwahrscheinlichkeit durch die entsprechenden empirischen Kennziffern zu beschränken. Diese

Vorgehen stützt sich auf Konvergenzaussagen für die einzelnen aus der Häufigkeitsverteilung abgeleiteten empirischen Kennziffern, wie z.B.

$$\begin{aligned} \text{Arithmetischer Mittelwert} &\Rightarrow \text{Erwartungswert} \\ \text{Empirische Varianz} &\Rightarrow \text{Theoretische Varianz} \\ \text{Relative Häufigkeit} &\Rightarrow \text{Wahrscheinlichkeit} \end{aligned}$$

Auf der Basis des Hauptsatzes der Statistik ergeben sich folgende Grundaufgaben der schließenden Statistik:

- Schätztheorie, die dazu dient, unbekannte Parameter des Modells mit der Angabe von Fehlergrenzen zu schätzen
- Testtheorie, die dazu dient, Hypothesen über das Modell unter Berücksichtigung von Fehlerwahrscheinlichkeiten zu prüfen
- Anpassungstests, die dazu dienen, ein gewähltes Verteilungsmodell zu überprüfen

Die ersten beiden dieser Grundaufgaben wollen wir im folgenden jeweils an einem Beispiel verdeutlichen.

5.4.3 Schätztheorie

Wir betrachten ein Merkmal X der Untersuchungseinheiten und eine einfache Zufallsstichprobe vom Umfang n, beschrieben durch die n unabhängigen und identisch verteilten Zufallsvariablen $X_1, X_2, \ldots, X_n$. Es sei bekannt, daß die X_i normalverteilt sind, aber die die Normalverteilung charakterisierenden Parameter Erwartungswert μ und Varianz σ^2 seien unbekannt. Zum Beispiel solle in einer anthropologischen Untersuchung an einem Volksstamm die Verteilung der Körpergröße anhand der vorliegenden Größendaten von $n = 10$ zufällig ausgesuchten Männern untersucht werden, die in der folgenden Tabelle aufgeführt sind:

Proband	1	2	3	4	5	6	7	8	9	10
Größe in Meter	1.65	1.54	1.49	1.50	1.54	1.48	1.61	1.59	1.45	1.45

Erwartungswert und Varianz der Größenverteilung der Gesamtpopulation sollen nun mit Hilfe einer geeigneten *Schätzfunktion* möglichst gut aus den vorliegenden Stichprobenwerten ermittelt werden. Kriterien für die Güte einer Schätzfunktion sind dabei (der Wichtigkeit nach geordnet):

Konsistenz: Je größer der Stichprobenumfang n ist, desto sicherer soll der aus den Stichprobenwerten ermittelte Schätzwert $\hat{\theta}_n$ dicht beim zu schätzenden Parameter θ liegen, d.h. für beliebiges $\varepsilon > 0$ soll gelten:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta|) = 1$$

Konsistenz ist eine unverzichtbare Eigenschaft jeder Schätzfunktion.

Erwartungstreue: Der Erwartungswert der Schätzgröße sollte gleich θ sein, so daß nicht systematisch zu große der zu kleine Schätzer auftreten können, also

$$E(\hat{\theta}_n) = \theta, \text{ für alle } n \in \mathbb{N}$$

Effizienz: Die Schätzwerte sollen möglichst wenig um θ streuen, d.h. die Varianz der Schätzfunktion sollte möglichst klein sein, also

$$Var(\hat{\theta}_n) = \min(Var(T)),$$

wobei das Minimum über alle konsistenten und erwartungstreuen Schätzfunktionen T von θ gebildet werde.

Asymptotische Normalität: Die Verteilung der standardisierten Schätzfunktion (mit Erwartungswert 0 und Varianz 1) sollte mit wachsendem Stichprobenumfang n gegen eine Standardnormalverteilung $N(0,1)$ konvergieren.

Über alle diese Eigenschaften verfügen in unserem Beispiel die folgenden zwei Schätzfunktionen T_1 und T_2:

$$\begin{aligned} T_1(X_1,X_2,\ldots,X_n) &:= \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ als Schätzfunktion für den Erwartungswert } \mu \\ T_2(X_1,X_2,\ldots,X_n) &:= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \text{ als Schätzfunktion für die Varianz } \sigma^2 \end{aligned}$$

Das heißt in unserem Beispiel gewinnen wir aus den vorliegenden Daten Schätzwerte $\hat{\mu}$ bzw. $\hat{\sigma^2}$ für Erwartungswert und Varianz der als normalverteilt vorausgesetzten Körpergröße wie folgt:

$$\begin{aligned} \hat{\mu} &= \frac{1}{10}(1.65+1.54+\ldots+1.46) = 1.53 \\ \hat{\sigma^2} &= \frac{1}{9}\Big((1.65-1.53)^2+(1.54-1.53)^2+\ldots+(1.46-1.53)^2\Big) \approx 0.0047 \end{aligned}$$

Wir zeigen im folgenden nur die Erwartungstreue beider Schätzfunktionen. Die Erwartungstreue von T_1 ergibt sich unmittelbar aus den Linearitätseigenschaften des Erwartungswertes:

$$E(T_1(X_1,X_2,\ldots,X_n)) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu$$

Für den Nachweis der Erwartungstreue von T_2 sei an zwei die folgenden Eigenschaften der Varianz erinnert:

$$\begin{aligned} Var(aX) &= a^2 Var(X), \text{ für alle } a \in \mathbb{R} \text{ und eine beliebige Zufallsvariable } X \\ Var(X+Y) &= Var(X)+Var(Y), \text{ falls } X \text{ und } Y \text{ unabhängige Zufallsvariablen sind} \end{aligned}$$

Wegen $E(\bar{X}^2) = Var(\bar{X}) + E(\bar{X})^2$ und $E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + \mu$ (vergleiche Abschnitt 5.2.3) gilt:

$$\begin{aligned}
E(\bar{X}^2) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \mu^2 = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) + \mu^2 = \frac{1}{n^2}\cdot n\sigma^2 + \mu^2 \\
&= \frac{\sigma^2}{n} + \mu^2 \\
\Rightarrow \quad E(T_2(X_1, X_2, \ldots, X_n)) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} E(X_i^2) - \frac{n}{n-1}E(\bar{X}^2) \\
&= \frac{n}{n-1}(\sigma^2 + \mu) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu\right) \\
&= \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 \\
&= \sigma^2
\end{aligned}$$

Dies erklärt, warum konträr zur Intuition, bei der Definition von T_2 durch $n-1$ und nicht durch n geteilt wird.

Die Normalverteilungseigenschaft der X_i haben wir in diesem Beweis nicht benötigt, tatsächlich wird sie auch nur für den Nachweis der Effizienz der Schätzfunktionen benötigt, alle anderen Eigenschaften der Schätzfunktionen T_1 und T_2 gelten für jede beliebige Verteilung der X_i.

5.4.4 Testtheorie

Angenommen die Frauenbeauftragte eines großen öffentlichen Betriebes will der Vermutung nachgehen, daß Frauen im Durchschnitt weniger verdienen als Männer bei gleicher Qualifikation und gleichem Alter. In Zusammenarbeit mit dem Personalbüro wählt sie eine feste Kombination von Qualifikationsniveau und Altersgruppe und zieht aus den zugehörigen Frauen und Männer je eine einfache Zufallsstichprobe vom Umfang n. Die Gehälter der Frauen X und die Gehälter der Männer Y seien hierbei als normalverteilt vorausgesetzt mit gleicher Varianz, aber gegebenenfalls unterschiedlichem Mittelwert

$$X \sim N(\mu_X, \sigma^2) \quad \text{und} \quad Y \sim N(\mu_Y, \sigma^2)$$

Um nun anhand der Stichprobendaten zu testen, ob tatsächlich Unterschiede vorhanden sind, geht man wie folgt vor: Man fragt sich, wie wahrscheinlich die aus den Stichprobendaten gewonnenen Ergebnisse sind, unter der sogenannten *Nullhypothese*, daß tatsächlich **kein** Unterschied zwischen Männer und Frauen vorliegt. Das heißt, man geht von dem Gegenteil dessen aus, was man eigentlich zeigen möchte, und fragt sich dann, inwieweit sich diese Hypothese falsifizieren läßt. Je unwahrscheinlicher also die vorliegenden Daten unter der Voraussetzung

gleicher Gehaltsverteilung bei den Geschlechtern sind, desto eher wird man geneigt sein, diese Voraussetzung zu verwerfen. Unsere Nullhypothese H_0 lautet also in diesem Fall:

$$H_0 : \mu_X = \mu_Y \quad \text{bzw.} \quad \mu_X - \mu_Y = 0,$$

mit entsprechender Alternativhypothese H_a:

$$H_a : \mu_Y > \mu_X$$

Aus den Stichprobendaten können wir den Unterschied ω zwischen dem Durchschnittsgehalt der Männer und dem Durchschnittsgehalt der Frauen ermitteln. ω kann dabei als Wert der Zufallsvariable

$$Z := \bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} Y_i$$

angesehen werden, wobei die X_i $N(\mu_X, \sigma^2)$-verteilte und die Y_i $N(\mu_Y, \sigma^2)$-verteilte Zufallsvariablen sind, die alle paarweise unabhängig sind. Da Summen von normalverteilten unabhängigen Zufallsvariablen ebenfalls wieder normalverteilt sind (vergleiche Abschnitt 5.3.1), ist Z ist ebenfalls eine normalverteilte Zufallsvariable mit:

$$Var(Z) = Var(\bar{X}) + Var(\bar{Y}) = \frac{2\sigma^2}{n}$$

Unter Annahme der Nullhypothese gilt außerdem:

$$E(Z) = E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y = 0$$

Mit Hilfe der tabellierten Werte der Normalverteilung, läßt sich nun ermitteln, wie wahrscheinlich es ist, daß sich die mittleren Gehälter der Frauen und Männer in der Stichprobe um den Betrag ω oder gar mehr unterscheiden, obwohl kein Unterschied in der gesamten Gehaltsverteilung zwischen Männern und Frauen vorliegt, d.h. man bestimmt die Wahrscheinlichkeit

$$\alpha = P(Z \geq \omega)$$

Sind z.B. Daten von 50 Frauen und Männern erhoben worden und beträgt das Durchschnittseinkommen in der Stichprobe bei den Frauen 2900 Euro, bei den Männern 3000 Euro und liegt die Standardabweichung σ der Gehälter bei 600 Euro, so folgt:

$$Var(Z) = \frac{2 \cdot 600^2}{50} = 14400$$

und damit unter Annahme der Nullhypothese

$$\begin{aligned} P(Z \geq \omega) &= P(Z \geq 100) \\ &= P\left(\frac{Z}{\sqrt{14400}} \geq \frac{5}{6}\right) \\ &= 1 - P\left(\frac{Z}{120} \leq \frac{5}{6}\right) \\ &\approx 1 - 0.8621 = 0.1379, \text{ da } \tfrac{Z}{120} \text{ standardnormalverteilt ist} \end{aligned}$$

In der Praxis verwirft man in der Regel die Nullhypothese, wenn α unter 5% liegt. In unserem Beispiel stehen die Daten also nicht im Widerspruch zur Nullhypothese, was aber NICHT in dem Sinn zu interpretieren ist, daß die Nullhypothese zutrifft. Sie läßt sich anhand der vorliegenden Daten nur nicht falsizifieren. Dies ist ein gewaltiger Unterschied, der beim statistischen Testen immer zu beachten ist!

6 Literaturverzeichnis

- Douglas G. Altman, PRACTICAL STATISTICS FOR MEDICAL RESEARCH, Chapman & Hall
- P. Armitage/G. Berry/ J.N.S. Matthews, STATISTICAL METHODS IN MEDICAL RESEARCH, Blackwell
- Günther Bourier, BESCHREIBENDE STATISTIK, Gabler
- David Bowers, STATISTICS FROM SCRATCH, Wiley
- Larry Gonick/Woollcott Smith, THE CARTOON GUIDE TO STATISTICS, HarperPerennial
- Walter Krämer, SO LÜGT MAN MIT STATISTIK,Piper
- Walter Krämer, STATISTIK VERSTEHEN, Piper
- Claus-Michael Langenbahn, GRUNDLAGEN DER STATISTIK, Vorlesungsskript FH Darmstadt (WS 2003/2004)
- J. Lehn/T. Müller-Gronbach/S. Retting, EINFÜHRUNG IN DIE DESKRIPTIVE STATISTIK, Gabler
- Deborah Rumsey, STATISTIK FÜR DUMMIES, mitp
- Walter Schneller, STATISTIK FÜR DEN STUDIENGANG MEDIENMANAGEMENT, Vorlesungsskript FH Würzburg-Schweinfurt
- John W. Tukey, EXPLORATORY DATA ANALYSIS, Addison-Wesle

Appendix A: Die Verteilungsfunkion der Standardnormalverteilung $\Phi(x)$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.5000	0.33	0.6293	0.66	0.8212	0.99	0.8389	1.32	0.9066	1.65	0.9505
0.01	0.5040	0.34	0.6331	0.67	0.8238	1.00	0.8413	1.33	0.9082	1.66	0.9515
0.02	0.5080	0.35	0.6368	0.68	0.8264	1.01	0.8438	1.34	0.9099	1.67	0.9525
0.03	0.5120	0.36	0.6406	0.69	0.8289	1.02	0.8461	1.35	0.9115	1.68	0.9535
0.04	0.5160	0.37	0.6443	0.70	0.8315	1.03	0.8485	1.36	0.9131	1.69	0.9545
0.05	0.5199	0.38	0.6480	0.71	0.8340	1.04	0.8508	1.37	0.9147	1.70	0.9554
0.06	0.5239	0.39	0.6517	0.72	0.8365	1.05	0.8531	1.38	0.9162	1.71	0.9564
0.07	0.5279	0.40	0.6554	0.73	0.8389	1.06	0.8554	1.39	0.9177	1.72	0.9573
0.08	0.5319	0.41	0.6591	0.74	0.8413	1.07	0.8577	1.40	0.9192	1.73	0.9582
0.09	0.5359	0.42	0.6628	0.75	0.8438	1.08	0.8599	1.41	0.9207	1.74	0.9591
0.10	0.5398	0.43	0.6664	0.76	0.8461	1.09	0.8621	1.42	0.9222	1.75	0.9599
0.11	0.5438	0.44	0.6700	0.77	0.8485	1.10	0.8643	1.43	0.9236	1.76	0.9608
0.12	0.5478	0.45	0.6736	0.78	0.8508	1.11	0.8665	1.44	0.9251	1.77	0.9616
0.13	0.5517	0.46	0.6772	0.79	0.8531	1.12	0.8686	1.45	0.9265	1.78	0.9625
0.14	0.5557	0.47	0.6808	0.80	0.8554	1.13	0.8708	1.46	0.9279	1.79	0.9633
0.15	0.5596	0.48	0.6844	0.81	0.8577	1.14	0.8729	1.47	0.9292	1.80	0.9641
0.16	0.5636	0.49	0.6879	0.82	0.8599	1.15	0.8749	1.48	0.9306	1.81	0.9649
0.17	0.5675	0.50	0.6915	0.83	0.8621	1.16	0.8770	1.49	0.9319	1.82	0.9656
0.18	0.5714	0.51	0.6950	0.84	0.8643	1.17	0.8790	1.50	0.9332	1.83	0.9664
0.19	0.5753	0.52	0.6985	0.85	0.8665	1.18	0.8810	1.51	0.9345	1.84	0.9671
0.20	0.5793	0.53	0.7019	0.86	0.8686	1.19	0.8830	1.52	0.9357	1.85	0.9678
0.21	0.5832	0.54	0.7054	0.87	0.8708	1.20	0.8849	1.53	0.9370	1.86	0.9686
0.22	0.5871	0.55	0.7088	0.88	0.8729	1.21	0.8869	1.54	0.9382	1.87	0.9693
0.23	0.5910	0.56	0.7123	0.89	0.8749	1.22	0.8888	1.55	0.9394	1.88	0.9699
0.24	0.5948	0.57	0.7157	0.90	0.8770	1.23	0.8907	1.56	0.9306	1.89	0.9706
0.25	0.5987	0.58	0.7190	0.91	0.8790	1.24	0.8925	1.57	0.9418	1.90	0.9713
0.26	0.6026	0.59	0.7224	0.92	0.8810	1.25	0.8944	1.58	0.9429	1.91	0.9719
0.27	0.6064	0.60	0.7257	0.93	0.8830	1.26	0.8962	1.59	0.9441	1.92	0.9726
0.28	0.6103	0.61	0.7291	0.94	0.8849	1.27	0.8980	1.60	0.9452	1.93	0.9732
0.29	0.6141	0.62	0.7324	0.95	0.8869	1.28	0.8997	1.61	0.9463	1.94	0.9738
0.30	0.6179	0.63	0.7357	0.96	0.8888	1.29	0.9015	1.62	0.9474	1.95	0.9744
0.31	0.6217	0.64	0.7389	0.97	0.8907	1.30	0.9032	1.63	0.9484	1.96	0.9750
0.32	0.6255	0.65	0.7422	0.98	0.8925	1.31	0.9049	1.64	0.9495	1.97	0.9756

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1.98	0.9761	2.26	0.9881	2.54	0.9945	2.82	0.9976	3.10	0.9990	3.38	0.9996
1.99	0.9767	2.27	0.9884	2.55	0.9946	2.83	0.9977	3.11	0.999	3.39	0.9997
2.00	0.9772	2.28	0.9887	2.56	0.9948	2.84	0.9977	3.12	0.9991	3.40	0.9997
2.01	0.9778	2.29	0.9890	2.57	0.9949	2.85	0.9978	3.13	0.9991	3.41	0.9997
2.02	0.9783	2.30	0.9893	2.58	0.9951	2.86	0.9979	3.14	0.9992	3.42	0.9997
2.03	0.9788	2.31	0.9896	2.59	0.9952	2.87	0.9979	3.15	0.9992	3.43	0.9997
2.04	0.9793	2.32	0.9898	2.60	0.9953	2.88	0.9980	3.16	0.9992	3.44	0.9997
2.05	0.9798	2.33	0.9901	2.61	0.9955	2.89	0.9981	3.17	0.9992	3.45	0.9997
2.06	0.9803	2.34	0.9904	2.62	0.9956	2.90	0.9981	3.18	0.9993	3.46	0.9997
2.07	0.9808	2.35	0.9906	2.63	0.9957	2.91	0.9982	3.19	0.9993	3.47	0.9997
2.08	0.9812	2.36	0.9909	2.64	0.9959	2.92	0.9982	3.20	0.9993	3.48	0.9997
2.09	0.9817	2.37	0.9911	2.65	0.9960	2.93	0.9983	3.21	0.9993	3.49	0.9998
2.10	0.9821	2.38	0.9913	2.66	0.9961	2.94	0.9984	3.22	0.9994	3.50	0.9998
2.11	0.9826	2.39	0.9916	2.67	0.9962	2.95	0.9984	3.23	0.9994	3.51	0.9998
2.12	0.9830	2.40	0.9918	2.68	0.9963	2.96	0.9985	3.24	0.9994	3.52	0.9998
2.13	0.9834	2.41	0.9920	2.69	0.9964	2.97	0.9985	3.25	0.9994	3.53	0.9998
2.14	0.9838	2.42	0.9922	2.70	0.9965	2.98	0.9986	3.26	0.9994	3.54	0.9998
2.15	0.9842	2.43	0.9925	2.71	0.9966	2.99	0.9986	3.27	0.9995	3.55	0.9998
2.16	0.9846	2.44	0.9927	2.72	0.9967	3.00	0.9987	3.28	0.9995	3.56	0.9998
2.17	0.9850	2.45	0.9929	2.73	0.9968	3.01	0.9987	3.29	0.9995	3.57	0.9998
2.18	0.9854	2.46	0.9931	2.74	0.9969	3.02	0.9987	3.30	0.9995	3.58	0.9998
2.19	0.9857	2.47	0.9932	2.75	0.9970	3.03	0.9988	3.31	0.9995	3.59	0.9998
2.20	0.9861	2.48	0.9934	2.76	0.9971	3.04	0.9988	3.32	0.9995	3.60	0.9998
2.21	0.9864	2.49	0.9936	2.77	0.9972	3.05	0.9989	3.33	0.9996	3.61	0.9998
2.22	0.9868	2.50	0.9938	2.78	0.9973	3.06	0.9989	3.34	0.9996	3.62	0.9999
2.23	0.9871	2.51	0.9940	2.79	0.9974	3.07	0.9989	3.35	0.9996		
2.24	0.9875	2.52	0.9941	2.80	0.9974	3.08	0.9990	3.36	0.9996		
2.25	0.9878	2.53	0.9943	2.81	0.9975	3.09	0.9990	3.37	0.9996		