Inhaltsverzeichnis

Skript für den Kurs
"Übungen zur Biomathematik "

Orientiert am Gegenstandskatalog (GK)
"Medizinische Biometrie"

Institut für Biometrie der Medizinischen Hochschule Hannover

4. Auflage

Hartmut Hecker

1. Einführung

1.1. Gegenstand statistischer Untersuchungen

In der Statistik beschäftigt man sich mit Erscheinungen und Vorgängen, die *nicht exakt vorhersagbar* sind.

***Deskriptive* Statistik:**

In der *deskriptiven* Statistik wird versucht, über die Vielfalt der beobachteten Erscheinungen durch

- **Ordnen nach unterschiedlichen Kriterien** und durch
- **Reduktion auf wenige Kenngrößen**

eine Übersicht zu erhalten.
Die Wahl der dabei benutzten Methoden hängt insbesondere von der *Fragestellung* ab, die der Betrachtung der Daten zugrunde liegt.

***Analytische* Statistik:**

In der *analytischen* Statistik macht man zunächst die **Annahme**, dass die Unvorhersagbarkeit der einzelnen Beobachtungen auf *"zufällige" Erscheinungen* zurückzuführen ist, die in dem Gesamtgeschehen eine Rolle spielen. Es wird weiterhin angenommen, dass dieses "Zufallsgeschehen" mit Hilfe der *Wahrscheinlichkeitsrechnung* angemessen behandelt werden kann. Auf entsprechender mathematischer Grundlage wird dann versucht, ***Regelmäßigkeiten* oder *Gesetzmäßigkeiten* zu entdecken bzw. nachzuweisen**, die den Beobachtungen zugrunde liegen und von den zufälligen Erscheinungen "überlagert" werden.

Die Biomathematik oder "*Medizinische Biometrie*" beschäftigt sich sowohl mit den deskriptiven als auch mit den analytischen Methoden der Statistik, wobei deren Auswahl durch die Anwendungen in der Medizin bestimmt ist. Spezifische Fragestellungen aus der medizinischen Forschung haben darüber hinaus zur Entwicklung eines Methodenspektrums geführt, das über die Statistik selbst hinausgeht und somit auch der Biometrie ihr spezifisches Gepräge gegeben hat.

In den folgenden Kapiteln wird zunächst die *Deskriptive Statistik* behandelt (Kapitel 2). Kapitel 3 bringt dann eine kurze Einführung in die *Wahrscheinlichkeitsrechnung* und stellt insbesondere die Normalverteilung vor. Die beiden restlichen Kapitel 4 und 5 sind der *analytischen Statistik* gewidmet. Dabei werden –mit dem Ziel der *Parameterschätzung* und des *statistischen Testens von Nullhypothesen*– Beziehungen zwischen Stichproben und Grundgesamtheit hergestellt und deren Ergebnisse im Rahmen der medizinischen Fragestellungen interpretiert.

1.2. Lehrangebote: Vorlesung, Übung, Skript

Die bis hierher nur kurz angedeuteten Inhalte, die in den *"Übungen zur Biomathematik für Mediziner"* zu vermittelt sind, "haben es in sich": Im Vergleich zu klinischen

Fächern wird hier eine sehr andere Sichtweise auf Vorgänge und Erscheinungen im medizinischen Bereich vorgestellt und eingeübt. Statt genaue Analysen von Einzelfällen vorzunehmen betrachtet man immer zugleich eine *Vielzahl* von Einzelfällen *in ihrer Gesamtheit.* Auf diese Weise gelingt es häufig, Strukturen zu erkennen und Erkenntnisse zu gewinnen, die erst aus der größeren Distanz heraus sichtbar werden können. Diese Distanzierung bedeutet aber zugleich auch: *Abstraktion in der Denkweise* und *Anwendung mathematischer Methoden.*

Das Verstehen und das Einüben solcher Sichtweisen und das Vertrautwerden mit mathematischen Ansätzen in diesem Bereich erfordert erfahrungsgemäß viel *Zeit*, viele *Beispiele* und wiederholte eigene *Anwendung* dieser Methoden. Aus diesen Gründen haben wir gegenüber früheren Jahren nun wieder die *Vorlesung im Hörsaal* eingeführt, da es nur in dieser Form von Veranstaltung möglich ist, viel Zeit für eine große Zahl von Studierenden zur Verfügung zu stellen. Um trotzdem ausreichend Gelegenheit zur Anwendung statistischer Methoden in kleineren Gruppen und -nicht zuletzt- für Nachfragen, Dialog und Diskussion zu geben, sind die *Übungen in Seminarräumen* -wenn auch zeitlich reduziert- beibehalten worden. Als drittes Lehrangebot liegt Ihnen hier der erste Teil des *Skripts zur Biomathematik* vor. Als Begleittext zur Vorlesung und zu den Übungen hat es den Vorteil, den zu behandelnden Stoff *nachlesbar* präsentieren zu können, so dass man sich individuell auf den Text einrichten kann: Je nach Neigung (und mathematischer Vorbildung) kann man die verschiedenen Abschnitte mit unterschiedlichen Schwerpunkten durcharbeiten. Um dies zu ermöglichen, wurde versucht, die wichtigsten Gedankengänge, Definitionen und Aussagen nicht nur in mathematischer Terminologie, sondern –so weit wie möglich– auch verbal (und dennoch möglichst präzise) wiederzugeben und teilweise auch graphisch zu veranschaulichen.
Vorlesung, Übungen und Skript (oder irgendein anderer geeigneter Text: siehe Literaturanhang) zusammen sollten jede Studentin und jeden Studenten in die Lage versetzen, sowohl die Klausuren zum Erwerb des Übungsscheines zu bestehen als auch das nötige Basiswissen und Grundverständnis für die Biometrie zu erwerben, um damit statistische Gedankengänge und Argumentationen in der Medizin im Ansatz nachvollziehen und beurteilen zu können.
Auf die hier vermittelten Grundkenntnisse wird im 2. Klinischen Abschnitt im Rahmen des Ökologischen Kurses aufgebaut; darüber hinaus aber werden sie insbesondere für die Anfertigung einer eigenen, empirisch ausgerichteten Dissertation wieder benötigt und bilden dort ggf. die Basis für eine spezifische, themenbezogene biometrische Beratung.

Nachtrag zum Skript und Aufruf
Wir möchten das Skript zur Biomathematik ständig verbessern. Es soll in Sprache, Form und Gestaltung so werden, dass es möglichst von allen Studierenden akzeptiert, gelesen und verstanden wird. Auch inhaltliche Modifikationen sollten zu diesem Zweck noch möglich sein. Um dies zu erreichen, sind wir auf Ihre Mithilfe angewiesen: *Sie* können vermutlich am sichersten einschätzen, nach welchen Vorlagen Sie und Ihre Mit-StudentInnen am Besten lernen und verstehen können.
Daher: wenn Sie gute Ideen zur Weiterentwicklung des Skripts haben und Lust und Interesse, diese umzusetzen, dazu über das nötige handwerkliche Geschick verfügen (Textverarbeitung und Graphik am PC), haben Sie Gelegenheit, dies alles als **Studentische Hilfskraft** an unserem Institut zu verwirklichen. Bitte melden Sie sich bald! Schon im laufenden Semester könnten weitere Teile des Skripts auch Ihren Namen tragen!

2. Deskriptive Statistik

2.1. Merkmale (GK 1.3)

Zur systematischen Beobachtung und Beschreibung von (zufälligem) Geschehen ist es nötig festzulegen, *wen* und *was daran* und *mit welchem Ergebnis* man beobachtet:

> Als **Merkmalsträger** (oder auch *Beobachtungseinheit, Fall, subject, case, unit*) bezeichnet man das *"Subjekt"*, an dem eine *Beobachtung* oder *Messung* durchgeführt wird.

Beispiel: Patient, Blutprobe, Tag, Stenose, Station, Praxis, ...

> Ein **Merkmal** (auch: *Variable, Parameter, Meßgröße, item, feature*) ist eines von mehreren Aspekten, das zur Beschreibung eines Merkmalsträgers oder von "Ereignissen an einem Merkmalsträger" dient.

Beispiel: Geschlecht, Beruf, Diagnose, Anzahl Läsionen, Größe, Alter (jeweils: eines Patienten), Riboflavin–Konzentration (einer Blutprobe), SO_2–Gehalt (in Hannover am Tag x), Durchmesser–Reduktion (einer stenosierten Koronararterie), Anzahl Krankenpfleger/innen (einer Station), Jahresumsatz (einer Firma).

> Die **Merkmalsausprägung** (*Wert, value*) ist der numerische *Wert* oder der *Begriff*, der den aktuellen Stand des Merkmals beschreibt, der *Meßwert*.

Beispiel: *weiblich; akuter Infarkt*; *3* (Läsionen); *176* (cm); *54* (Jahre); 15 (μ/ccm); ...
In Statistik-Progammsystemen werden die Daten in Form einer *Rechteckdatei* angeordnet. Darin werden die einzelnen *Zeilen* durch die *Merkmalsträger* und die *Spalten* durch die *Merkmale* gebildet (siehe Abb. 2.1).

Einteilung der Merkmale

Je nach Art der Beobachtung wird das Ergebnis einer Messung auf ganz unterschiedlichen Skalenniveaus festgehalten und kann z.B. (s.o.) "*akuter Infarkt*" oder auch "*0.3* (mg/ccm)" heißen. Man unterscheidet die Merkmale nach den Skalenniveaus, auf denen sie gemessen werden:

Qualitativ versus ***Quantitativ:***

> Ein **qualitatives** Merkmal (auch "**nominal**" oder "**kategoriell**" genannt) hat *begriffliche, nicht–numerische* (oder eben "*qualitative*") Merkmalsausprägungen. Jede der theoretisch möglichen Merkmalsausprägungen bildet eine *Kategorie*. Die einzelnen Kategorien müssen in keiner Beziehung zueinander stehen (insbesondere also nicht geordnet sein).

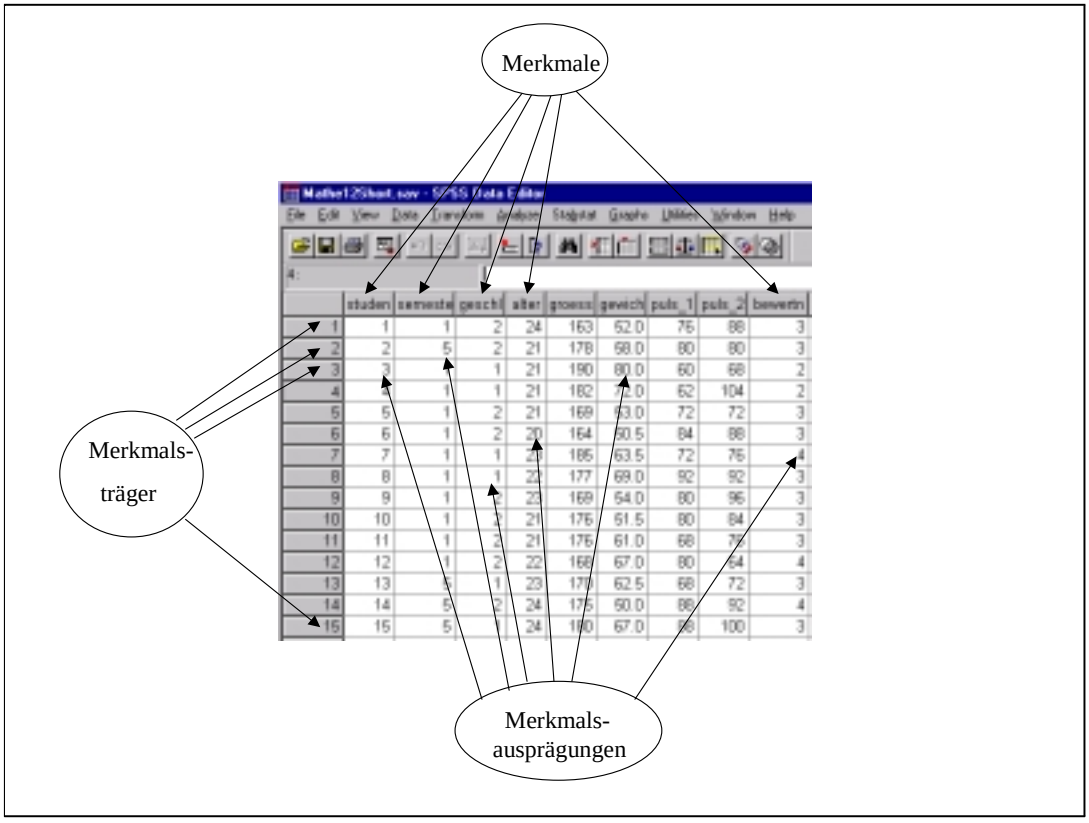


Abbildung 2.1: Beispiel einer Rechteckdatei

Beispiel eines qualitativen Merkmals:
"*Erste auftretende Nebenwirkung*" bei einem Patienten, mit den Kategorien "Keine", "Übelkeit", "Kopfschmerz" und "Sonstige".

Quantitative Merkmale haben ***numerische*** Merkmalsausprägungen.

Beispiele quantitativer Merkmale:
"Anzahl der Nebenwirkungen" bei einem Patienten, mit den möglichen Ausprägungen 0, 1, 2, 3, ...; systolischer Blutdruck in mmHg; Anzahl Kinder; Körpertemperatur in °Celsius etc.
Wichtig ist dabei, dass die numerischen Ergebnisse verschiedener Messungen die Beziehungen der jeweiligen Zahlen zueinander (zumindest teilweise) "erben":

"3" sind *mehr* als "2" Nebenwirkungen ("ordinal", s.u.)

39° Fieber ist *1° höher* als 38° Fieber (man darf Differenzen bilden: "Intervallskala").

"4" Kinder sind *doppelt so viel wie* "2" Kinder (man darf Quotienten bilden: "Rationalskala" . Siehe auch GK Med. Psychologie u. Soziologie).

Ordinal sind solche Merkmale, deren Ausprägungen in eine "natürliche" Reihenfolge gebracht werden können.

Alle quantitativen Merkmale sind ordinal.

Qualitative Merkmale können ordinal sein.
Beispiel: *Schmerzintensität* mit den Kategorien "keine", "leichte", "starke" Schmerzen.

***Anmerkung*:** In der Datenverarbeitung werden häufig auch qualitative und ordinale Merkmalsausprägungen *numerisch gespeichert*, z.B. 1 ="männlich", 2 = "weiblich". Den numerischen Werten sind "Werte–Etiketten" oder "Value–Labels" zugeordnet. Die Zuordungsregel wird als "Schlüssel" oder "Kodierung" bezeichnet. Das Merkmal bleibt trotz der Kodierung natürlich *qualitativ* bzw. *ordinal,* da die Merkmalsausprägungen zwar mit Zahlen versehen sind, nicht aber deren Eigenschaften tragen.
Beispiel (Abb. 2.2):

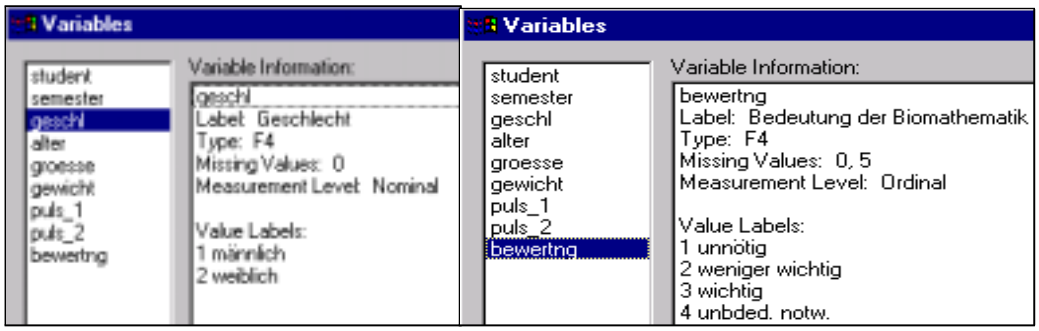


Abbildung 2.2: Kodierung eines qualitativen und eines ordinalen Merkmals

Einteilung quantitativer Merkmale:

Bei *diskreten* Merkmalen sind nur endlich viele (oder abzählbar viele) Werte möglich.
Bei *stetigen* Merkmalen sind –häufig nur in einem bestimmten Intervall– alle reellen Zahlen als Merkmalsausprägung möglich.

Aufgrund der Grenzen der Genauigkeit jedes Messvorganges sind auch stetige Merkmale, wenn sie zusammen mit der Menge der möglichen Messergebnisse gesehen werden, diskret (mit z.B. den Werten 0.0, 0.1, 0.2, ... bei Messgenauigkeit 1/10). Dennoch macht es Sinn, zwischen diskreten und stetigen Merkmalen zu unterscheiden, da sie in der deskriptiven wie auch später in der analytischen Statistik oft unterschiedlich bearbeitet werden.

2.2. Ordnung der Beobachtungen (Eindimensionale Häufigkeitsverteilungen; GK 2.1)

Die einfachste Situation, in der man sich über eine größere Anzahl von Daten einen Überblick verschaffen will, liegt dann vor, wenn man sich zunächst für die Werte nur *eines* Merkmales interessiert:

2.2.1. Statistische Beschreibung diskreter und qualitativer Merkmale

Zu jeder möglichen Ausprägung (Klasse) des Merkmals wird die Anzahl der Einzelbeobachtungen bestimmt, welche diese Ausprägung haben:

> Die ***absoluten Häufigkeiten*** einer Klasse sind die Anzahl der Fälle der Stichprobe, die in diese Klasse fallen.

Beispiel: In einer Untersuchung mit 57 weiblichen und 12 männlichen Patienten sind "57" und "12" die absoluten Häufigkeiten für die Klassen "weiblich" bzw. "männlich" des Merkmales "Geschlecht".
Ein Bezug zur Gesamtheit aller Beobachtungen wird dadurch hergestellt, dass man durch die Anzahl aller Beobachtungen –meist mit "n" bezeichnet– dividiert:

> Die ***relativen Häufigkeiten*** einer Klasse sind deren absoluten Häufigkeiten, dividiert durch die Anzahl der Beobachtungen. Diese werden häufig auch in % angegeben.

Beispiel: Bei 16 weiblichen und 19 männlichen Patienten ist die Gesamtzahl der Fälle $n = 35$. Die relativen Häufigkeiten sind daher $\frac{16}{35} = 0.457 = 45.7$ % für "weiblich" und $\frac{19}{35} = 0.543 = 54.3$ % für "männlich".

***Anmerkung*:** Es kommt vor, dass in einer Stichprobe ein oder mehrere Merkmale nicht in allen Fällen erboben werden können: es gibt "*fehlende Angaben* ("*Missings*"). Dann kann man die absolute Häufigkeit entweder auf die Gesamtzahl aller Fälle beziehen oder nur auf die Anzahl der (für das jeweilige Merkmal) *gültigen Fälle*. Statistikprogramme berechnen meistens beide Versionen.

Beispiel aus einer Untersuchung mit 35 gültigen und 3 fehlenden Angaben zum Geschlecht (Abb. 2.3):

Geschlecht

		Frequency	Percent	Valid Percent
Valid	männlich	19	50.0	54.3
	weiblich	16	42.1	45.7
	Total	35	92.1	100.0
Missing	System	3	7.9	
Total		38	100.0	

Abbildung 2.3: Häufigkeiten bezogen auf *alle* bzw. auf *alle gültigen* Fälle

Vollständiges Beispiel zur Ordnung der Beobachtungen:
Aufnahmediagnose von n = 19105 Aufnahmen an der MHH (siehe Tab. 2.1).

Merkmal: ICD Klassifikation (Internationale Klassifikation der Krankheiten)			
Merkmalsausprägung		Absolute	Relative
Code/Wert/ Kategorie Nr (k)	Value Label, "Werte-Etikett"	Häufig-keit n_k	Häufigk. $h_k(\%)$ $= \frac{n_{k \times 100}}{n}$
1	Infektiöse u. parasitäre Krankheiten	$n_1 = 1979$	$h_1 = 10.4$ %
2	Neubildungen	2516	13.2 %
3	Endokrinopath., Ernähr.- u. Stoffwechselkr.,St.d.Immunsyst.	657	3.4 %
4	Krankht.des Blutes u.d. blutbildenden Organe	486	2.5 %
5	Psychiatrische Krankheiten	2454	12.8 %
6	Krankht. des Nervensystems u. d.Sinnesorgane	1218	6.4%
7	Krankht. d. Kreislaufsystems	2647	13.8%
8	Krankht. d. Atmungsorgane	629	3.3%
9	Krankht. d.Verdauungsorgane	1048	5.5%
10	Krankht. d. Harn-u.Geschlechtsorgane	948	5.0%
11	Komplik.d.Schwangerschaft, bei Entbind.imWochenbett	1	0.0%
12	Krankht.d.Haut u.d.Unterhautzellgewebes	102	0.5%
13	Krankht. d. Skeletts,d. Muskeln u.d.Bindegewebes	529	2.8%
14	Kongenitale Anomalien	93	0.5%
15	Bestimmte Affektionen, d. i. Urspr. i.d. Perinatalzeit haben	6	0.0%
16	Symptome u.schlecht bezeichnete Affektionen	710	3.7%
17	Verletzungen u. Vergiftungen	3082	16.1%
	Summe:	19105	100.0 %

Tabelle 2.1: Häufigkeitstabelle: Aufnahme nach ICD-Klassifikation

2.2.2. Statistische Beschreibung *ordinaler qualitativer* und *diskreter quantitativer* Merkmale

Bei diesen Merkmalen sind die Kategorien bzw. deren Werte *geordnet*, z.B. das Merkmal *Bewertung der Bedeutung der Biomathematik im Studium*, siehe (Abb. 2.3): (1: unnötig) $<$ (2: weniger wichtig) $<$ (3: wichtig) $<$ (4: unbedingt notwendig).
Man bildet dann zusätzlich zu den relativen Häufigkeiten der einzelnen Kategorien die **absoluten *Summenhäufigkeiten* bis zu den einzelnen Kategorien:**

$$\begin{aligned} N_k &= \text{Anzahl der Beobachtungen bis zur Kategorie } k \\ &= n_1 + n_2 + \ldots + n_k \end{aligned}$$

die **relativen *Summenhäufigkeiten* bis zu den einzelnen Kategorien:**

$$\begin{aligned} H_k &= \frac{\text{Anzahl der Beobachtungen bis zur Kategorie } k}{\text{Anzahl aller Beobachtungen}} \\ &= \frac{N_k}{n} \\ &= \frac{n_1 + n_2 + \ldots + n_k}{n} \end{aligned}$$

Beispiel: Auszählung aus Fragebogenaktionen bei MHH-Studenten in Biomathe-Unterricht aus verschiedenen Jahrgängen und graphische Darstellung der absoluten Häufigkeiten (Abb. 2.4):
Man kann aus der letzten Spalte dieser Tabelle ("*Cumulative Percent*" = "*Relative Summenhäufigkeit in Prozent*") z.B. ablesen, dass 71.8 % der befragten StudentInnen die Biomathematik als unnötig oder weniger wichtig einstuften (also 28.2 % sehen sie

Bedeutung der Biomathematik im Studium

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 unnötig	35	7.3	7.4	7.4
	2 weniger wichtig	307	64.1	64.5	71.8
	3 wichtig	117	24.4	24.6	96.4
	4 unbded. notw.	17	3.5	3.6	100.0
	Total	476	99.4	100.0	
Missing	System	3	.6		
Total		479	100.0		

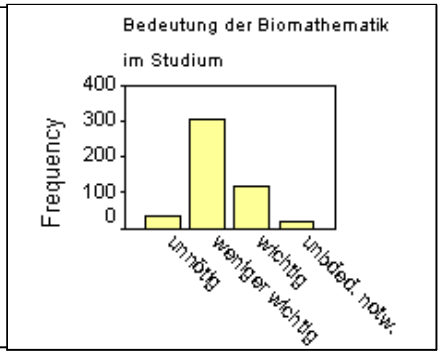


Abbildung 2.4: A: Relative Summenhäufigkeiten B: Balkendiagramm

als wichtig oder sogar unbedingt notwendig an), und 96.4 % bewerteten sie mit einer der Kategorien 1, 2 oder 3. Die *absolute Summenhäufigkeit* wird in dieser Auswertung nicht ausgedruckt. Wie lauten die absoluten Summenhäufigkeiten bis zu den Kategorien 1, 2, 3 und 4 ?

Bei *quantitativen* Merkmalen sind die einzelnen Kategorien durch *numerische Werte* charakterisiert. Man bildet dann in der gleichen Weise für jeden möglichen Wert x (z.B. "Anzahl Kinder" : $x = 1, 2, 3, \ldots$) die (absoluten und) relativen Summenhäufigkeiten "*bis zum Wert x*". Dadurch erhält man eine mathematische *Funktion der möglichen Merkmalswerte* und nennt sie die *Kumulative Verteilungsfunktion F(x) des Merkmals:*

$$F(x) = \frac{\text{Anzahl der Beobachtungen} \leq x}{\text{Anzahl aller Beobachtungen}} \tag{2.1}$$

Die kumulative Verteilungsfunktion an der Stelle x gibt die *relative Anzahl der Beobachtungen bis zum Wert* x an.

Im folgenden **Beispiel** wurden durch die MHH-Unfallforschung 1000 Verkehrsunfälle analysiert. Die "Anzahl beteiligter Personen" als diskretes quantitatives Merkmal weist dabei folgende Verteilung auf (siehe Tabelle 2.2).

Die graphische Darstellung der Verteilungsfunktion ist in Abb. 2.5 dargestellt.

Man kann aus der kumulativen Verteilungsfunktion z.B. unmittelbar ablesen, dass die Anzahl der Unfälle mit *einer* Person bei 17 % lag, mit Beteiligung von 1 oder 2 Personen aber bereits bei 65 % usw. Zugleich fällt auf, dass der maximale Wert der Anzahl beteiligter Personen 29 war.

2.2.3. Statistische Darstellung stetiger Merkmale

Klasseneinteilung und Histogramm

Bei stetigen Merkmalen kommt eine Häufigkeitsauszählung der *einzelnen Werte* nicht in Betracht, da es zu viele verschiedene Einzelwerte gibt. Man hilft sich wie folgt:
Der Wertebereich wird in *Intervalle* (*Klassen)* eingeteilt. *Zu jedem Intervall werden dann die relativen Häufigkeiten dieser Klassen* gebildet.

Anzahl beteiligter Personen

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	166	16.6	16.6	16.6
	2	479	47.9	47.9	64.5
	3	194	19.4	19.4	83.9
	4	73	7.3	7.3	91.2
	5	47	4.7	4.7	95.9
	6	19	1.9	1.9	97.8
	7	11	1.1	1.1	98.9
	8	5	.5	.5	99.4
	9	1	.1	.1	99.5
	10	4	.4	.4	99.9
	29	1	.1	.1	100.0
	Total	1000	100.0	100.0	

Tabelle 2.2: Anzahl der am Unfall beteiligten Personen

Eine graphische Darstellung erfolgt durch *zusammenhängende Säulen* über den Intervallen. Achtung: wenn die gewählten Klassenbreiten *nicht* alle *identisch* sind, darf man die Höhe der Säulen nicht proportional zu den relativen Häufigkeiten wählen! Dadurch würden die *größeren* Klassenbreiten *begünstigt*. Stattdessen wähle man:

$$\text{Höhe der Säule} = \frac{\text{Relative Häufigkeit der Klasse}}{\text{Breite des Intervalls}}$$

Stellt man die Gleichung um, so folgt daraus:

$$\begin{aligned} \text{Relative Häufigkeit der Klasse} &= \text{Höhe der Säule} \times \text{Breite des Intervalls} \\ &= \text{Fläche der Säule} \end{aligned}$$

D.h.:

die *Flächen der Säulen* entsprechen den *relativen Häufigkeiten.*

Die Bildung von Histogrammen hat den Vorteil, dass diese die Verteilung der aufgetretenen Werte eines Merkmales intuitiv leicht nachvollziehbar wiedergeben. Sie hat aber auch Nachteile:

- Informationsverlust: Durch die Klassenbildung wird die Verteilung der aufgetretenen Werte nicht mehr vollständig dargestellt.
- Willkür in der Klasseneinteilung: Das Histogramm hängt wesentlich von der gewählten Klasseneinteilung ab.
  Als Faustregel für die Anzahl K der zu bildenden Klassen (bei identischen Klassenbreiten) wird in der Literatur angegeben:

$$K \approx \begin{cases} \sqrt{n} & \text{für } n \leq 1000 \\ 10 \ \lg n & \text{für } n > 1000 \end{cases}$$

  Beispiele hierzu siehe Abb. 2.6.

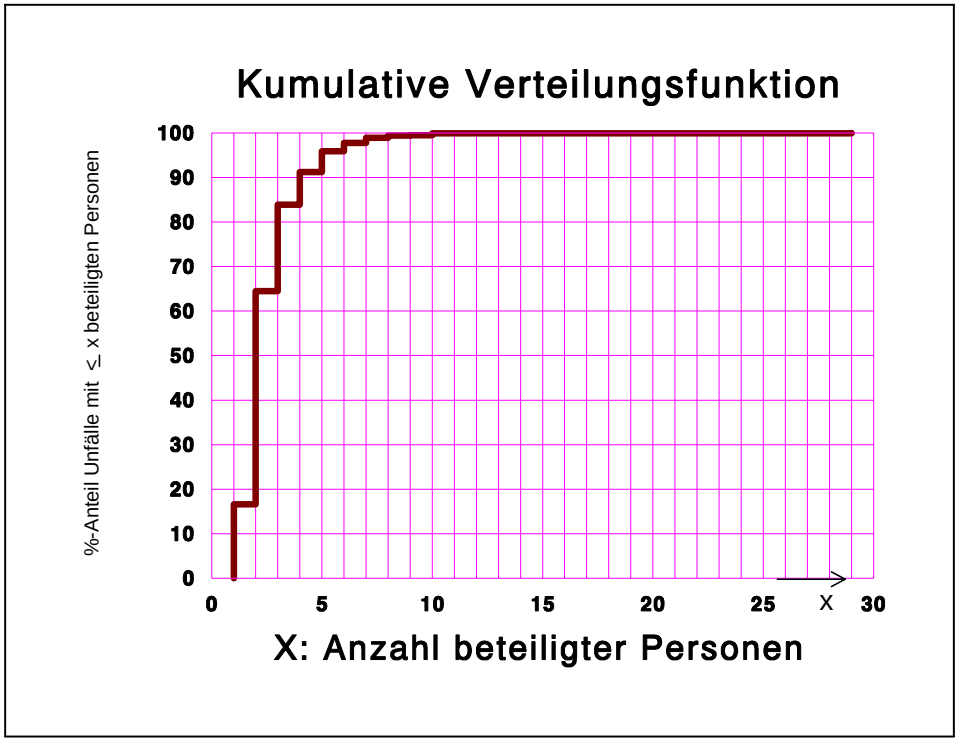
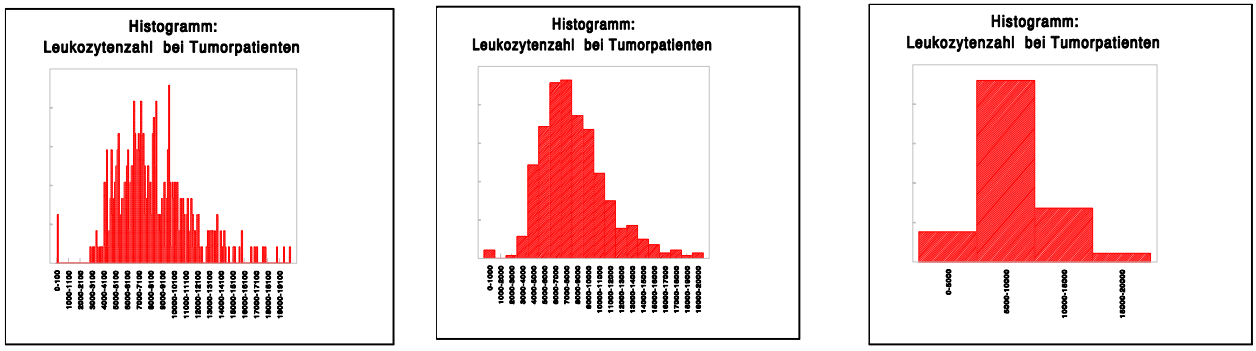


Abbildung 2.5: Anzahl der am Unfall beteiligten Personen



200 Klassen bei n=417 Fällen 20 Kl. bei n = 417 Fällen ($\sqrt{417} \approx 20$)4 Klassen bei n=417 Beobachtungen

Abbildung 2.6: Unterschiedliche Klassenbreiten

(Empirische) kumulative Verteilungsfunktion

Sie ist wie für diskrete Merkmale definiert, siehe (Gleichung 2.1), jedoch können die auftretenden Merkmalswerte x jetzt *jeden numerischen Wert* (meist: in einem bestimmten Intervall) annehmen. Im Extremfall tritt kein Wert mehrfach (d.h. bei mehr als einer Beobachtungseinheit der Stichprobe) auf . Dann macht die kumulative Verteilungsfunktion bei jedem beobachteten Messwert einen Sprung der Höhe $\frac{1}{n}$. Allgemein ist die Sprunghöhe an der Stelle x gleich $\frac{k}{n}$, wenn genau k Fälle den Wert x haben. Zwischen je zwei benachbarten beobachteten Messwerten bleibt sie konstant.

Beispiel 1:
Östrogenwerte bei $n = 10$ Patienten, bereits geordnet: $1,\ 4,\ 5,\ 5,\ 6,\ 6,\ 7,\ 8,\ 8, 10$ (Abb. 2.7):

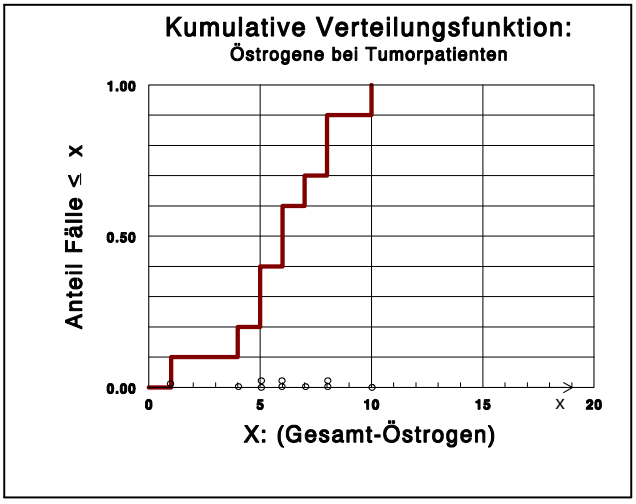


Abbildung 2.7: Kumulative Verteilungsfunktion für n=10 Beobachtungen

Beispiel 2: Leukozytenzahl bei n=417 Tumorpatienten. Die hier dargestellten Daten sind mit denen der Histogramme (s.o.) identisch (Abb. 2.8):

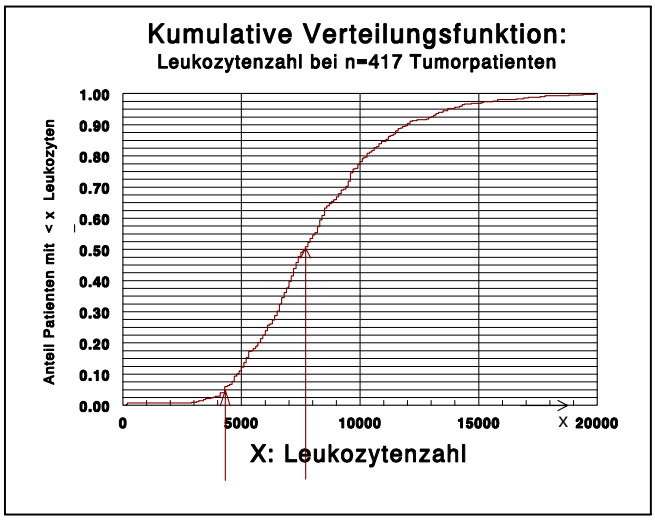


Abbildung 2.8: Kumulative Verteilungsfunktion für n=417

Hieraus ist z.B. abzulesen, dass 50 % der Patienten einen Wert bis zu etwa 7800 hatten, und 5 % einen Wert < 4200.

2.3. Statistische Kenngrößen einer Verteilung

In vielen Zusammenhängen interessiert nicht die *vollständige* Häufigkeitsverteilung eines Merkmals; es reicht dann, die Verteilungen durch eine oder wenige *Kenngrößen* zu charakterisieren. Die wichtigsten Charakteristika der Verteilung eines *quantitativen* Merkmals beziehen sich dabei auf deren *Lage* und ihre *Gestalt.* Die Gestalt wiederum kann durch die *Streuung* und durch ein Maß für eine etwaige *Unsymmetrie* der Verteilung beschrieben werden.

2.3.1. Kenngrößen, die aus der kumulativenVerteilungsfunktion $F(x)$ ablesbar sind

Definitionen. Die kumulative Verteilungsfunktion $F(x)$ gibt die relative Häufigkeit der Fälle mit Werten $\leq x$ an. Ist beispielsweise für das Merkmal "Körpergröße in cm" $F(172) = 0.5 = 50$ %, so sind genau 50 % der gemessenen Personen ≤ 1.72 m groß: In diesem Sinne ist dann 1.72 m die "Mitte" der Beobachtungen und wird als "*Median*" der Verteilung bezeichnet. Allgemeiner definiert man:

- ***Lagemaße:***

  Median: Der erste Wert x, für den $F(x) \geq 0.5$ ist.

  Unteres Quartil (= 25%–Quantil): Der erste Wert x, für den $F(x) \geq 0.25$ ist.

  Oberes Quartil (= 75%–Quantil): Der erste Wert x, für den $F(x) \geq 0.75$ ist.

  q–**Quantil**: Der erste Wert x, für den $F(x) \geq q$ ist, speziell z.B.:

  95%–**Quantil**: Der erste Wert x, für den $F(x) \geq 0.95$ ist.

  Minimum: Kleinster beobachteter Wert.

  Maximum: Größter beobachteter Wert.

- ***Streumaße:***

  Quartilsabstand: Differenz zwischen oberem und unterem Quartil.

  Spannweite (Range): Differenz zwischen Maximum und Minimum.

Alle Definitionen werden in der folgenden Abbildung 2.9 erläutert. In dieser Graphik wird noch einmal die kumulative Verteilungsfunktion der Leukozytenzahl bei n=417 Patienten der Abbildung 2.8 zugrundegelegt (Abb. 2.9):

Beispiele und Anwendungen

Beispiel 1

Größe- und Gewichtsverteilungen von MHH-StudentInnen sind in den beiden Graphiken der Abb. 2.10 kumulativ dargestellt.

Interpretation: Für Körpergröße und -gewicht sind die Verteilungsfunktionen für Studentinnen und Studenten auf den ersten Blick *in ihrer Gestalt* "ähnlich". Bezüglich *der Lage* gibt es in beiden Fällen eine Rechtsverschiebung (also eine Verschiebung zu *größeren* Werten) bei den Studenten.
Läßt sich dieser Eindruck numerisch belegen? Man lese dazu die statistischen Kenngrößen der kumulativen Verteilungsfunktion (so gut wie möglich) ab und vergleiche nach Tabelle 2.3.

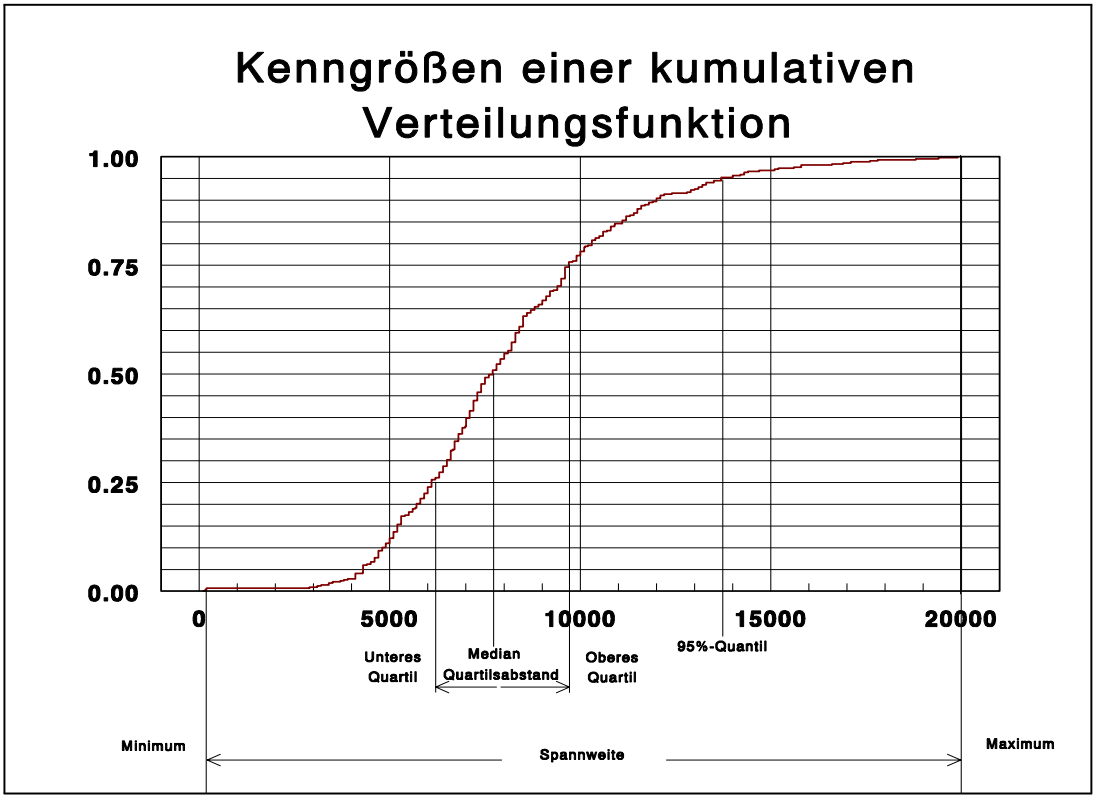


Abbildung 2.9: Kumulative Verteilung der Leukozytenzahl

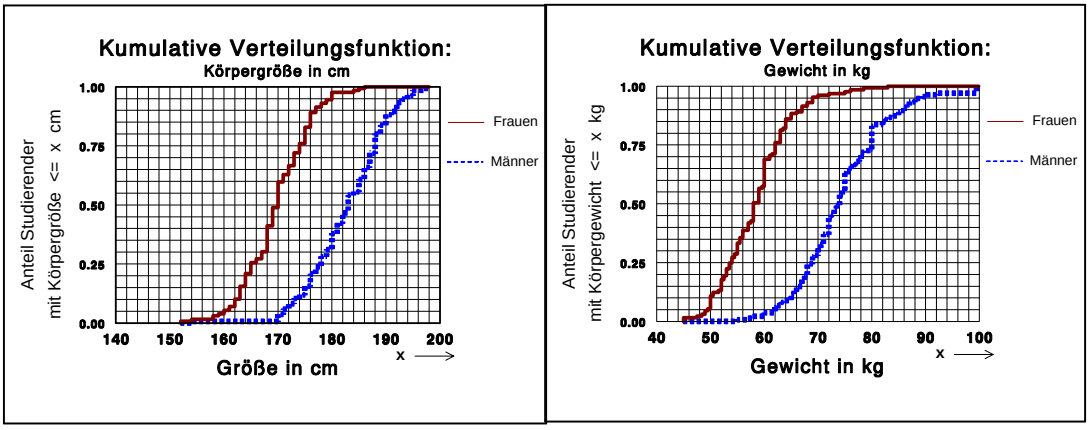
Körpergröße in cm				Körpergewicht in kg			
Kenngröße	Student-innen	Studenten	Vergleich (Diff.)	Kenngröße	Student-innen	Studenten	Vergleich (Diff.)
Median				Median			
Unteres Quartil				Unteres Quartil			
Oberes Quartil				Oberes Quartil			
95%-Quantil				95%-Quantil			
Minimum				Minimum			
Maximum				Maximum			
Quartilsabstand				Quartilsabstand			
Range				(Range)			

Tabelle 2.3: Körpergröße und Gewicht bei Studentinnen und Studenten

Beispiel 2

- Welches ist in Tabelle 2.2 und Abb. 2.5 der *Median* der Anzahl der am Unfall beteiligten Personen?
- Wie ändert sich der Median, wenn an dem einen Unfall statt 29 nur 9 Personen beteiligt gewesen wären?

Graphische Darstellungen. Im gerade behandelten Beispiel (Anzahl der am Unfall beteiligten Personen) war der Wert "29" im Vergleich zu den anderen Werten offenbar



Größe kumulativ Gewicht kumulativ

Abbildung 2.10: Kumulative Verteilungen für Studentinnen und Studenten

sehr extrem, ein "Ausreißer". *Was ist ein Ausreißer?*
Die folgende Definition wird benutzt, wenn man als graphische Darstellung einer Verteilung den "Boxplot" benutzt:

> Ein *Ausreißer* ist ein Wert, der mehr als 1 1/2 Längen des Quartilsabstandes *oberhalb des oberen* oder *unterhalb des unteren Quartils* liegt.

Im *Boxplot* werden alle bisher besprochenen statistischen Kenngrößen und besondere Einzelwerte (incl. "Ausreißer") auf einen Blick dargestellt (Abb. 2.11).

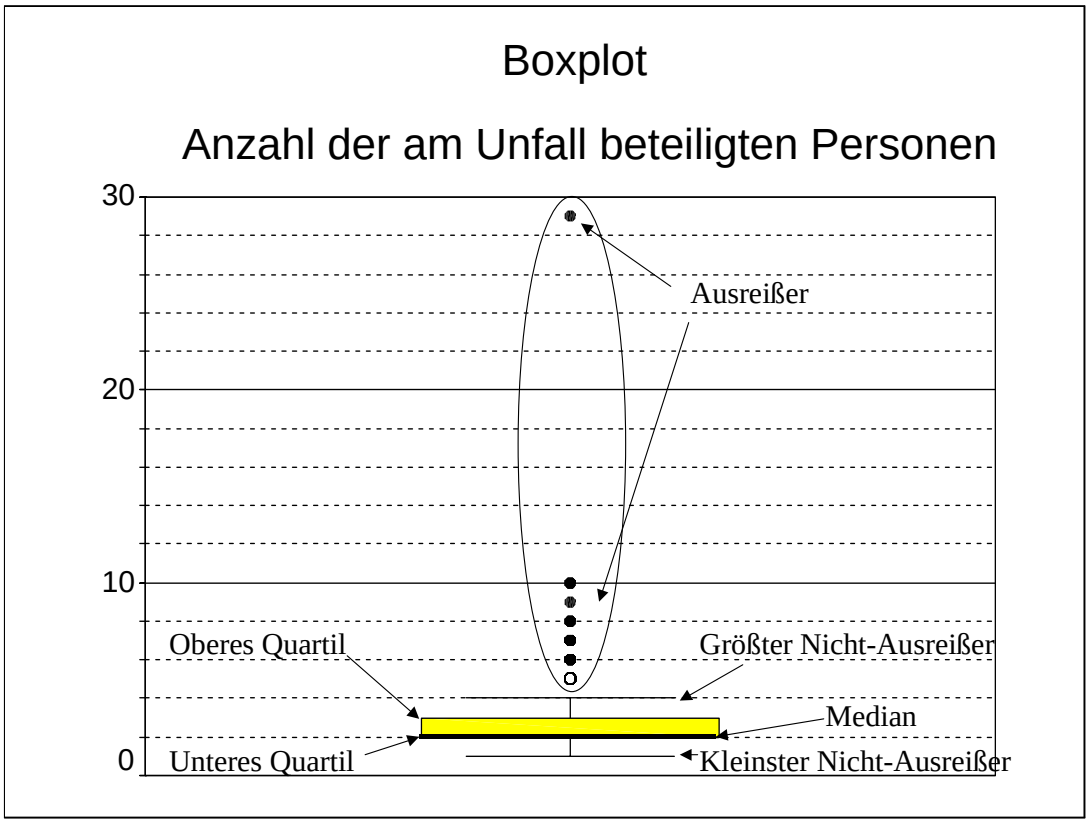


Abbildung 2.11: Definition des *Boxplots* an Hand des Beispiels der Unfallstudie

- Die oberen und unteren Ränder der "Box" werden durch das obere und untere *Quartil* gebildet.
- Die horizontale Linie innerhalb der Box ist der *Median* (im oberen Beispiel identisch mit dem unteren Quartil).
- Die oberen und unteren "Antennen" geben die Werte des größten und des kleinsten "*Nicht-Ausreißers*" wieder.
- Die Ausreißerwerte selber werden einzeln als Punkte dargestellt.

Im hier dargestellten Beispiel ist der Median (= 2) identisch mit dem unteren Quartil (man überprüfe das auf Grund der Daten von Abbildung 2.5!). Das obere Quartil hat denWert 3. Der Quartilsabstand beträgt demnach 1. Anderthalb mal Quartilsabstand vom oberen Quartil nach oben bedeutet: $3 + 1.5 \times 1 = 4.5$. Werte oberhalb 4.5 sind nach dieser Definition also bereits Ausreißer! Man sieht, dass diese Ausreißerdefinition für *unsymmetrische Verteilungen* offenbar nicht gut geeignet ist. Durch den Boxplot wird aber gerade diese Unsymmetrie (relativ häufiges Auftauchen extremer und sehr extremer Werte in nur einer Richtung) anschaulich aufgedeckt.
Weiteres **Beispiel** (Abb. 2.12):

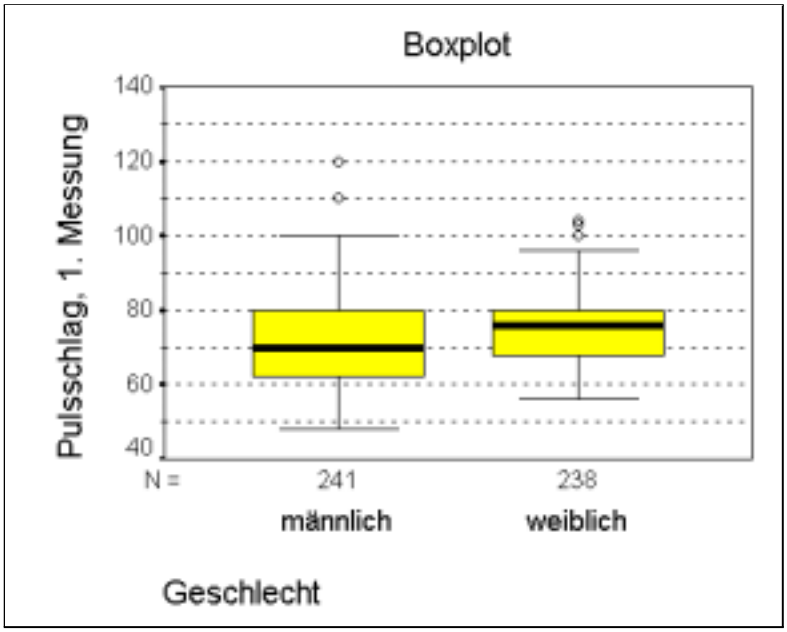


Abbildung 2.12: Pulsschlag bei 479 MHH-Studentinnen und Studenten

Aus dieser Darstellung geht hervor, dass bei den Studentinnen der Pulsschlag im Median um etwa 6 Schläge *höher* lag als bei den Studenten (76 versus 70 /min). - Gibt es eine physiologische Erklärung dafür?
Die Streuung, gemessen am Quartilsabstand (oder auch am Range), ist bei den Studentinnen *kleiner* als bei den Studenten.

2.3.2. Kenngrößen, die ohne vorherige Ordnung der beobachteten Werte berechenbar sind

Mittelwert

Beispiel: In einer Monty Python -Szene sorgt Robin Hood für Gerechtigkeit: Er

überfällt eine Postkutsche, sammelt von allen Mitfahrern deren Bargeld ein und verteilt den Gesamtbetrag wieder, nun aber für jeden der Mitfahrer zu gleichen Teilen. Wieviel erhält dann jeder, wenn die 6 Mitfahrer vorher folgende Beträge hatten: 10, 140, 42, 100, 108, und 2000 Dukaten ?
Lösung: Gesamtsumme $= 10 + 140 + 42 + 100 + 108 + 2000 = 2400$ Dukaten.
Gleichmäßige Verteilung auf 6 Personen ergibt $2400/6 = 400$ Dukaten für jede Person.

Definition:

Die Summe der Einzelwerte, dividiert durch deren Anzahl, ist der ***Mittelwert*** der Einzelwerte.

Allgemeine Bezeichnungen und Definition

1. Die *Merkmale* werden im Folgenden allgemein mit *Großbuchstaben* wie $X, Y, \ldots$bezeichnet.
2. Wird ein Merkmal X der Reihe nach an den Beobachtungseinheiten Nr. $i = 1, 2, \ldots, n$ gemessen, so werden die Merkmals*werte* in der Reihenfolge ihrer *Beobachtung* indiziert und mit *kleinen Buchstaben* bezeichnet: $x_1, x_2, \ldots x_n$

Beispiel (Tabelle 2.4):

Merkmal X: Bargeld		
Beobachtungs-einheit Nr. (i)	Bezeich-nung	Wert der Beobachtung
1	x_1	10
2	x_2	140
3	x_3	42
4	x_4	100
5	x_5	108
6	x_6	2000
Anzahl: n=6;	Summe:	2400
Mittelwert: $(\frac{\text{Summe}}{\text{Anzahl}})$:		$\frac{2400}{6} = 400$

Tabelle 2.4: Mittelwert

Aus den Beobachtungswerten $x_1, x_2, \ldots x_n$ eines Merkmals X bildet man als *Lagemaß* seiner Verteilung den ***Mittelwert*** $\overline{x}$:

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2 + \ldots + x_n}{n} \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i \end{aligned}$$

$\overline{x}$ kann in mehrfacher Weise anschaulich interpretiert werden:

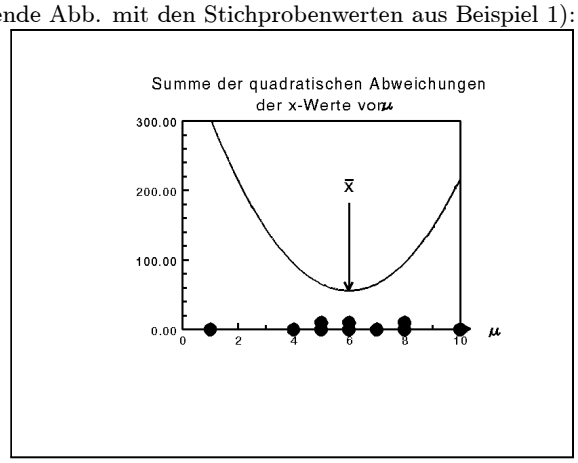
1. (Durchschnitt): Sollen alle Beobachtungswerte identisch sein, ohne dass sich die *Summe der Werte* ändert, müssen alle Werte gleich dem Mittelwert sein.

2. (Schwerpunkt): Befestigt man an einem gewichtslosen und mit einer numerischen Skala versehenen Stab an den Stellen $x_1, x_2, \ldots x_n$ jeweils eine Kugel mit demselben Gewicht, so bleibt der Stab im Gleichgewicht, wenn man ihn an der Stelle $\overline{x}$ unterstützt.

3. (Kleinste quadratische Abweichung): Man wähle probehalber irgendeinen Wert μ und bilde die quadrierten Abweichungen der beobachteten Werte x_i von μ und deren Summe $S(\mu)$:

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

Variiert man den probeweise gewählten Bezugswert μ, so ergibt sich folgende Beziehung (siehe folgende Abb. mit den Stichprobenwerten aus Beispiel 1):



Der Mittelwert minimiert die Summe der quadratischen Abweichungen.
Die Summe der Abweichungsquadrate, $S(\mu)$, ist am kleinsten, wenn man für den Bezugswert μ den Mittelwert $\overline{x}$ einsetzt.

Streuung

Im Beispiel Monty Python gab es vor dem Einsatz von Robin Hood unterschiedlich starke *Abweichungen der Einzelwerte vom Mittelwert.* Um dies zu quantifizieren, bildet man die einzelnen *quadratischen Abweichungen vom Mittelwert.* Dadurch werden Abweichungen nach oben und nach unten gleichbehandelt (Tabelle 2.5):

Merkmal X: Bargeld				
Beobachtungs-einheit Nr. (i)	Bezeich-nung (x_i)	Wert der Beobachtung	Abweichung v. Mittelwert $x_i - \overline{x}$	Quadrat.Abw. v. Mittelwert $(x_i - \overline{x})^2$
1	x_1	10	-390	152100
2	x_2	140	-260	67600
3	x_3	42	-358	128164
4	x_4	100	-300	90000
5	x_5	108	-292	85264
6	x_6	2000	1600	2560000
Anzahl: n=6;	Summe:	2400	0	3083128
Mittelwert: $(\frac{\text{Summe}}{\text{Anzahl}})$:		$\overline{x} = \frac{2400}{6} = 400$		$s_x^2 = \frac{3083128}{6-1} = 616625.6$

Tabelle 2.5: Berechnung von Mittelwert und Varianz

Die *durchschnittliche* quadratische Abweichung vom Mittelwert ist gleich $\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$. Geht man davon aus, dass der Mittelwert $\overline{x}$ "zufallsabhängig" ist und daher in der Regel

von einem hypothetischen "richtigen" Wert (siehe "Analytische Statistik") abweicht, so ist die durchschnittliche quadratische Abweichung von diesem "*richtigen*" Wert also eher etwas größer als die von $\overline{x}$. Man dividiert daher nur durch $(n-1)$ statt durch n und nennt das Ergebnis die ***Varianz*** s_x^2 ***(mittlere quadratische Abweichung)*** von X:

$$s_x^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\ldots+(x_n-\overline{x})^2}{(n-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Durch Wurzelbildung kommt man wieder in die ursprüngliche Skala von X und erhält damit ein Maß für die ***Streuung*** s_x ***(Standardabweichung)***:

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

Im obigen Beispiel ist die Standardabweichung $s_x = \sqrt{s_x^2} = \sqrt{616625.6} = 785.255$. Dieser große Wert für die Streuung kommt wesentlich durch *einen* Extremwert ($x_6 = 2000$) zustande. Dieser Einzelwert bringt auch den Mittelwert sehr hoch. Dies wird durch die kumulative Darstellung der Verteilung verdeutlicht (Abb. 2.13):

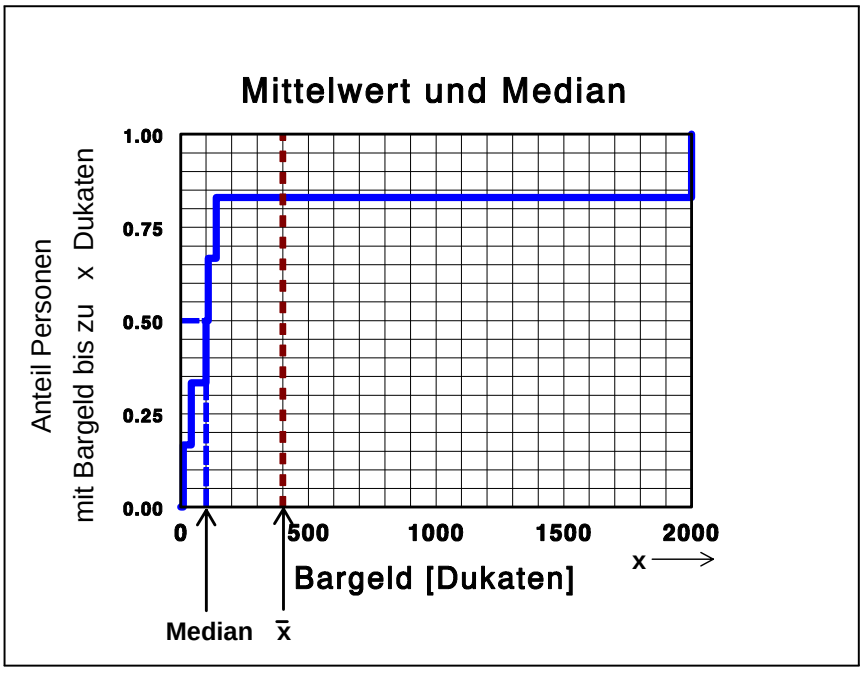


Abbildung 2.13: Mittelwert und Median einer rechtsschiefen Verteilung

Man überlege:

- Wie ändern sich Mittelwert und Streuung, wenn aus dem Extremwert $x_6 = 2000$ der Wert $x_6 = 200$ wird? (Bei der Streuung muss man komplett neu rechnen).
- Wie ändern sich in diesem Fall Median und Quartilsabstand?
- Soll man nun also als Lage- und Streumaß einer Verteilung lieber Mittelwert und Streuung nehmen oder Median und Quartilsabstand?

Hinweise:

1. Die eben dargestellte Verteilung ist *"rechtsschief": Extreme* und *sehr extreme* Werte kommen (*nur* bzw. *überwiegend*) zur *rechten* Seite hin vor.
   Diese *Rechtsschiefe* hat u.a. zur Folge, dass der *Mittelwert deutlich größer ist als der Median.*
2. Sind die Werte x_i um ihren Mittelwert herum zur positiven wie zur negativen Seite hin "etwa gleich" verteilt (wenn also die Spiegelung aller Werte am Mittelwert die Verteilungsfunktion nicht wesentlich verändert), so spricht man von einer *symmetrischen* Verteilung.
3. In vielen Fällen kann man aus einer rechtsschiefen Verteilung (wenn dort nur positive Werte vorkommen) eine *symmetrische* Verteilung erhalten, wenn man alle Beobachtungswerte *logarithmiert.*
   Als Beispiel hierzu wieder das Monty Python-Beispiel (Abb. 2.14):

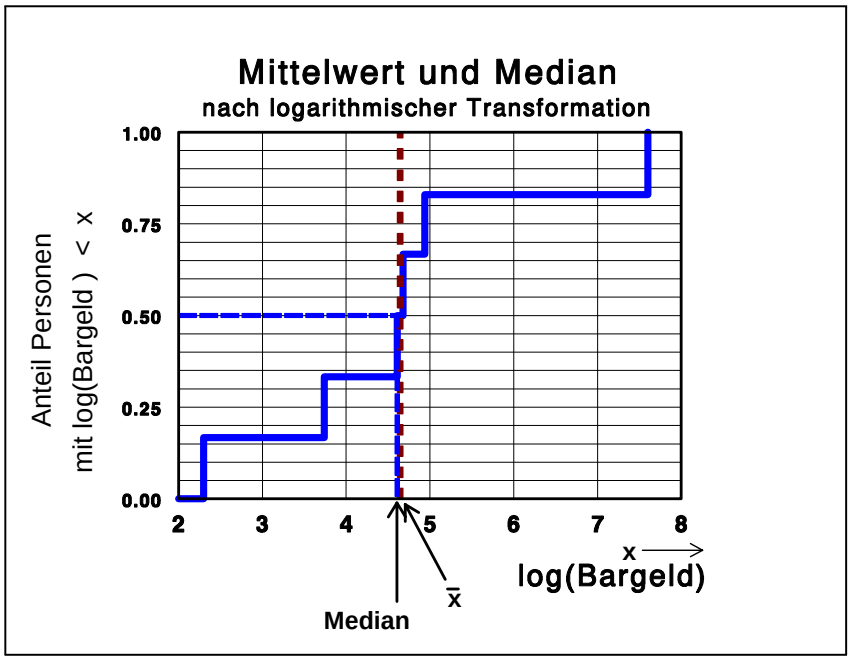


Abbildung 2.14: Mittelwert und Median einer symmetrischen Verteilung

Eine Unsymmetrie ist kaum noch zu erkennen; Mittelwert und Median sind fast identisch.

Um eine etwaige *Abweichung von der Symmetrie* genauer zu definieren, führt man als weitere Kenngröße einer Verteilung deren ***Schiefe*** γ ein:

$$\gamma = \frac{\frac{(x_1-\bar{x})^3+(x_2-\bar{x})^3+\ldots+(x_n-\bar{x})^3}{n}}{s_x^3}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{s_x^3}$$

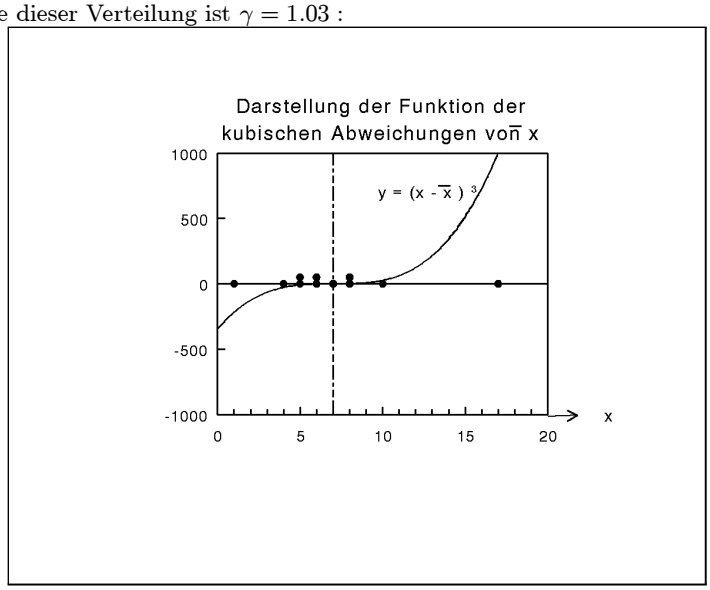
Aus der Definition folgt:

1. Die Definition ist *unabhängig von der gewählten Skala* des Merkmals X und kennzeichnet daher nur die *Form* der Verteilung. (Mißt man z.B. die Körpergröße in *cm* statt in *m*, so werden alle Werte x_i mit 100 multipliziert, aber wegen der Division durch s_x^3 bleibt γ unverändert).

2. Je nach den Meßwerten x_i kann γ *positiv* oder *negativ* sein.

3. Überwiegen in einer Verteilung unter den mittleren und insbesondere unter den extremen Abweichungen vom Mittelwert deutlich die *positiven* Abweichungen, so wird $\gamma \gg 0$. Die Verteilung wird dann *rechts–schief* genannt. Im umgekehrten Fall heißt die Verteilung *links–schief* . Aus einer graphischen Darstellung ist eine Verteilung mit z.B.$\gamma = +1$ meistens, mit $\gamma = +2$ immer leicht als rechts–schief zu erkennen.

4. Diese Eigenschaften des Unsymmetrie–Maßes γ sind dadurch zu erklären, dass in deren Definition die Abweichungen vom Mittelwert *kubisch* (mit der 3. Potenz) in die Berechnung eingehen. Dadurch fallen extreme Abweichungen ganz wesentlich ins Gewicht, wobei auch noch das Vorzeichen berücksichtigt wird.
   Beispiel für eine Verteilung mit ausgeprägter Dominanz einer positiven Abweichung. Die Schiefe dieser Verteilung ist $\gamma = 1.03$:



Die Abweichungen gehen kubisch und mit ihrem Vorzeichen in die Berechnung ein

Beispiel für die Berechnung der Kenngrößen einer Verteilung*:*

Beob. Nr. (i)	Wert x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$
1	4	−2	4	−8
2	5	−1	1	−1
3	6	0	0	0
4	1	−5	25	−125
5	10	4	16	64
6	7	1	1	1
7	6	0	0	0
8	5	−1	1	−1
9	8	2	4	8
10	8	2	4	8
Summe:	60	0	56	−54

$$
\begin{aligned}
n &= 10 \\
\overline{x} &= \frac{60}{10} = 6 \\
s_x^2 &= \frac{56}{9} = 6.222 \\
s_x &= \sqrt{6.222} = 2.4944 \\
\gamma &= \frac{\frac{-54}{10}}{2.4944^3} = -.34793
\end{aligned}
$$

2.4. Zweidimensionale Häufigkeitsverteilungen (GK 2.2)

2.4.1. Kontingenztafeln

Einführung. In einer klinischen Studie wurden Patienten nach Zufallszuteilung ("Randomisation") entweder mit einer *niedrigen* oder mit einer *hohen Dosis* eines Medikamentes behandelt. Neben dem Behandlungsergebnis interessierte u.a. auch das *Auftreten von Nebenwirkungen.*

Ergebnis (Tabelle: 2.6):

Merkmal X: Dosis		
	Häufigkeit	
Ausprägung	absolut	relativ
1=niedrig	184	51.3 %
2=hoch	175	48.7 %
Total	359	100 %

Merkmal Y: Nebenwirkungen		
	Häufigkeit	
Ausprägung	absolut	relativ
1=nein	292	81.3 %
2=ja	67	18.7 %
Total	359	100 %

Tabelle 2.6: Zwei 1-dimensionale Verteilungen

Kann man daraus schließen, dass das Auftreten von Nebenwirkungen etwas mit der Höhe der Dosis zu tun hat? Offenbar nicht, denn die 67 Patienten mit Nebenwirkungen könnten z.B. alle aus der Niedrig- oder alle aus der Hochdosisgruppe stammen; wahrscheinlicher ist eine Mischung.
Um Näheres zu erfahren muss man die Daten erneut auszählen und bei jedem Patienten die *Kombination beider Merkmale* notieren (Abb. 2.15):

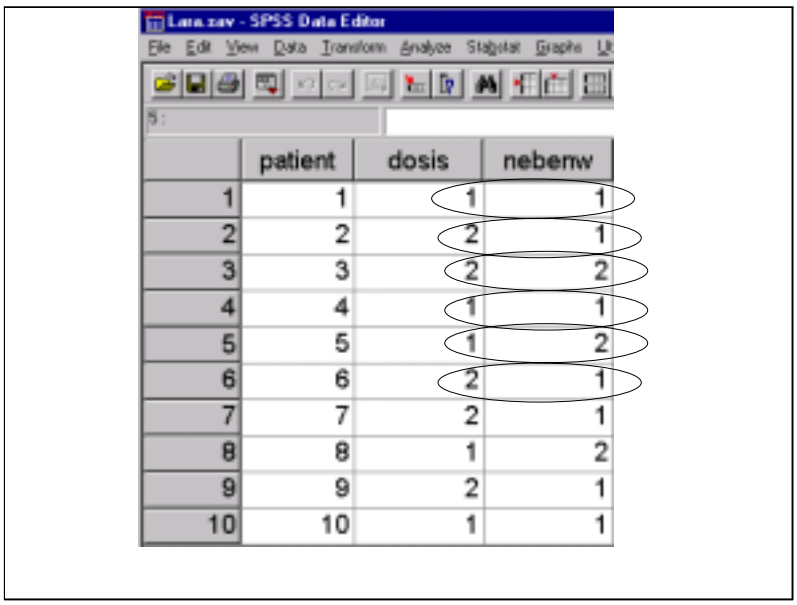


Abbildung 2.15: Wertekombinationen in einer Rechteckdatei

Die Ergebnisse der Auszählung bringt man in eine 2-dimensionale Tabelle. Darin werden die absoluten Häufigkeiten der *Kombinationen* der Werte der beiden Merkmale in die 4 Zellen eingetragen (Tabelle 2.7):

	Merkmal Y : Nebenwirkung eingetreten?		
Merkmal X :	Kategorie Nr.	1 = nein 2 = ja	Summe:
Verabreichte Dosis	1 = niedrig	152 32	184
	2 = hoch	140 35	175
	Summe:	292 67	359

Tabelle 2.7: Kreuztabelle: Absolute Häufigkeiten

Man sieht jetzt: Von den 67 Patienten mit Nebenwirkungen waren 32 in der Niedrig- und 35 in der Hochdosisgruppe.
Um endgültig vergleichen zu können, bezieht man diese Zahlen noch auf die unterschiedlichen *Gruppengrößen* (relative Häufigkeiten *innerhalb jeder Zeile*) (Tabelle 2.8):

	Merkmal Y : Nebenwirkung eingetreten?			
Merkmal X :	Kategorie Nr.	1 = nein	2 = ja	Summe:
Verabreichte Dosis	1 = niedrig	152	32	184
	Zeilen-%	82.6%	17.4%	100%
	2 = hoch	140	35	175
	Zeilen-%	80.0%	20.0%	100%

Tabelle 2.8: Kreuztabelle: Absolute Häufigkeiten und Zeilen-Prozente

Die Nebenwirkungsrate war also mit 20 % in der Hochdosisgruppe (geringfügig) größer als bei niedriger Dosis (17.4 %).

Allgemeine Darstellung.

Die 2-dimensionale Tabelle

	Merkmal Y			
Merkmal X	Kategorie Nr.	1	2	Summe:
	1	a	b	$a+b$
	2	c	d	$c+d$
	Summe:	$a+c$	$b+d$	$n=a+b+c+d$

mit der Auszählung der Kombinationen der Werte zweier qualitativer Merkmale X und Y nennt man eine **Kontingenztafel** oder **Kreuztabelle**.

Im hier behandelten Fall mit jeweils nur 2 Ausprägungen beider Merkmale spricht man von einer 2×2-Kontingenztafel oder von einer **Vierfeldertafel.** An den Rändern der Tabelle findet man die 1-dimensionalen Verteilungen beider Merkmale wieder, die sogenannten "*Randverteilungen*".

Die Kontingenztafel ermöglicht Aufschlüsse über den *Zusammenhang* zweier Merkmale.

Dies wird im folgenden Abschnitt konkretisiert.

2.4.2. Zusammenhangsmaße in Vierfeldertafeln

Relatives Risiko

Im obigen Beispiel wurden die Nebenwirkungsraten als Maß für das *Risiko einer Nebenwirkung* berechnet (20 % versus 17.4 %). Das Verhältnis beider Risiken: $\frac{20.0}{17.4} = 1.149$, wird als relatives Risiko für Nebenwirkungen der Gruppe 2 im Vergleich zur Gruppe 1 interpretiert. Allgemeiner:

$$RR = \frac{c/(c+d)}{b/(a+b)}$$

ist das relative Risiko für das Ergebnis "2" (2. Spalte) der Gruppe 2 im Vergleich zur Gruppe 1.

Anmerkung:
Die Kodierungen beider Merkmale sind willkürlich. Wählt man statt Tabelle 2.8 die Darstellung Tabelle 2.9:

	Merkmal Y : Nebenwirkung eingetreten?			
Merkmal X :	Kategorie Nr.	$1 =$ ja	$2 =$ nein	Summe:
Verabreichte Dosis	$1 =$ hoch	$a = 35$	$b = 140$	175
	$2 =$ niedrig	$c = 32$	$d = 152$	184

Tabelle 2.9: Kreuztabelle in anderer Kodierung

so ist das relative Risiko für Nebenwirkungen der Hochdosis- im Vergleich zur Niedrigdosisgruppe definiert als

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

Wesentlich ist also immer der inhaltliche Bezug auf das interessierende "Ereignis" (hier: Nebenwirkung) und die Vergleichsgruppe (hier: Niedrig-Dosis).

Interpretation: Ein RR von 1.0 bedeutet: gleiche Risiken in beiden Gruppen; RR > 1.0 bedeutet höheres, RR<1.0 niedrigeres Risiko im Vergleich zur Bezugsgruppe.

Verhältis der "Chancen": Odds Ratio. Wie stehen die Chancen für das Auftreten einer Nebenwirkung für die Patienten der Therapiestudie?
Antwort nach der folgenden Tabelle (Tab. 2.10):

	Merkmal Y : Nebenwirkung eingetreten?			
Merkmal X :	Kategorie Nr.	$1 =$ ja	$2 =$ nein	Summe:
Verabreichte Dosis	$1 =$ hoch	$a = 35$	$b = 140$	175
	$2 =$ niedrig	$c = 32$	$d = 152$	184

Tabelle 2.10: Tabelle zum Odds Ratio

Die Chancen stehen 35 zu 140 in der Hochdosis- und 32 zu 152 in der Niedrigdosisgruppe. Um das Verhältnis der Chancen beider Gruppen zu berechnen, bildet man: $\frac{35\ zu\ 140}{32\ zu\ 152} = \frac{35/140}{32/152} = 1.19$
Ergebnis: die Chancen in der Hochdosisgruppe sind 1.19 mal so hoch wie in der Niedrigdosisgruppe.
Das englische Wort für "Chancen" ist "odds", das Verhältnis der Chancen heißt"Odds Ratio" (OR).
Allgemeine Definition des Odds Ratios eines untersuchten Ereignisses in einer Risikogruppe im Vergleich zur Kontrollgruppe:

		Ereignis eingetreten?	
		ja	nein
Risiko vorhanden?	ja	a	b
	nein	c	d

$$OR = \frac{a/b}{c/d} = \frac{a \times d}{b \times c}$$

Interpretation:
$\boxed{OR = 1}$ bedeutet "gleiche Chancen", $\boxed{OR > 1}$ heißt "erhöhte", $\boxed{OR < 1}$ "erniedrigte Chancen" für das Eintreten des untersuchten Ereignisses in der Risikogruppe im Vergleich zur Kontrollgruppe.

Anwendung: Fall-Kontroll-Studie.

In einer Studie über den plötzlichen Kindstod und seine möglichen Risikofaktoren wurde die Datenerhebung *retrospektiv* am Ergebnis orientiert: man bildete die **Gruppe aller Fälle** in einem definierten Beobachtungszeitraum und notierte nachträglich, welche der interessierenden potentiellen Risikofaktoren vorhanden waren. Als **Kontrolle** diente eine **Stichprobe** von Kindern, bei denen kein plötzlicher Kindstod eingetreten war.
Ergebnis bezüglich des Risikofaktors "Rauchen" (Tabelle 2.11):

Plötzlicher Kindstod	Merkmal Y : Fall oder Kontrolle?		
Merkmal X :	Kategorie	Fall	Kontrolle
Rauchen während der	Ja	86	1277
Schwangerschaft?	Nein	104	4643

Tabelle 2.11: Odds Ratio: Plötzlicher Kindstod und Rauchen

Bildet man nun das Odds Ratio, obwohl die Kontrollgruppe nur eine Stichprobe darstellt und in der Gesamtpopulation daher unterrepräsentiert ist, so erhält man:

$$OR = \frac{86 \times 4643}{104 \times 1277} = 3.007$$

Muss das Ergebnis noch korrigiert werden?

Angenommen man wüsste, dass man in der Stichprobe nur jede 10-te Geburt erfasst hätte und dass der Risikofaktor in der Gesamtpopulation genau so wie in der Stichprobe verteilt wäre. Man müsste dann mit der folgenden Tabelle 2.12 rechnen:

Plötzlicher Kindstod	Merkmal Y : Fall oder Kontrolle?		
Merkmal X :	Kategorie	Fall	Kontrolle
Rauchen während der	Ja	86	1277×10
Schwangerschaft?	Nein	104	4643×10

Tabelle 2.12: Odds Ratio und Fall-Kontroll-Studie

Als Odds Ratio ergäbe sich:

$$OR = \frac{86 \times (4643 \times 10)}{104 \times (1277 \times 10)} = \frac{86 \times 4643}{104 \times 1277} = 3.007$$

es bliebe also unverändert! Und dasselbe gälte für jeden anderen Stichprobenfaktor
Folgerung:

Das Odds Ratio ist auch in Fall-Kontroll-Studien berechenbar, selbst wenn die Anteile der Kontrollen (und der Fälle) an der Gesamtpopulation unbekannt sind.

Die Berechnung des Odds Ratio muss also nicht korrigiert werden: Die "Chancen" für plötzlichen Kindstod sind nach dieser Studie 3–fach erhöht bei Kindern, deren Mutter während der Schwangerschaft geraucht hat.

Die Abweichungen der beobachteten Häufigkeiten von den erwarteten Häufigkeiten.
Aufgabe:
Gegeben sind die Randverteilungen einer Kreuztabelle (Tabelle 2.13) :

	Merkmal Y		
	Kategorie Nr.	1 2	Summe:
Merkmal X	1		175
	2		184
	Summe:	67 292	359

Tabelle 2.13: Randverteilungen und Unabhängigkeit

Wie muss man die Tabelle ergänzen, wenn folgende Annahme gültig ist:

Annahme: Die Verteilung von Y sei unabhängig von den Werten von X
Die relativen Häufigkeiten für die Ergebnisse (Y=1) und (Y=2) (Spalte 1 bzw. Spalte 2) wären dann für beide Zeilen (X=1) und (X=2) *identisch*, und zwar so wie in der Gesamtpopulation.
Lösung:

1. Die relative Häufigkeit für das Ergebnis (Y=1) in der *Gesamtpopulation* ist $\frac{67}{359} = 0.186\,63 \approx 18.7\ \%$.

2. Bei 175 Beobachtungen in der *ersten Gruppe* (X=1) würde man unter der gemachten Annahme *in 18.7 % der Fälle das Ergebnis (Y=1) erwarten*, das wären also $175 * 0.187$, genauer: $= \frac{175 \times 67}{359} = 32.66$.

3. Man sieht, dass die Berechnungsformel allgemein so aussieht:

$$Erwartete\ Anzahl = \frac{Zeilensumme \times Spaltensumme}{Gesamtzahl}$$

   Dies gilt für alle 4 Felder der Kreuztabelle. Z.B. für Zeile (X=1), Spalte (Y=2):

$$Erwartete\ Anzahl = \frac{175 \times 292}{359} = 142.34$$

4. Tabelle aller *(unter der Annahme der Unabhängigkeit von X und Y) erwarteten* Häufigkeiten (Tabelle 2.14):

	Merkmal Y			
Erwartete Häufikeiten	Kategorie Nr.	1	2	Summe:
Merkmal X	1	32.66	142.34	175
	2	34.34	149.66	184
	Summe:	67	292	359

Tabelle 2.14: Erwartete Häufigkeiten

Wenn nun die tatsächlich beobachteten Häufigkeiten genau so wären wie vorausberechnet (evtl. auf- bzw. abgerundet auf die nächste ganze Zahl), so würde die beobachtete Tabelle exakt zu der Annahme der Unabhängigkeit passen. Umgekehrt gilt:
Je größer die Abweichungen der beobachteten von den erwarteten Häufigkeiten, desto stärker ist dies ein *Indiz gegen die Annahme der Unabhängigkeit der beiden Merkmale.*
Um diese Abweichung von der Unabhängigkeit über alle 4 Felder der Kreuztabelle *numerisch* zu erfassen, bildet man das Abweichungsmaß X^2 :

$$X^2 = \sum \frac{(\text{Beobachtete Anzahl - Erwartete Anzahl})^2}{\text{Erwartete Anzahl}}$$

wobei die Summation ($\sum$) über alle Felder der Kreuztabelle geht.

Das Abweichungsmaß $\boxed{X^2 \text{ ist stets} \geq 0}$. $\boxed{X^2 = 0}$ bedeutet totale Unabhängigkeit von X und Y. Je größer X^2, desto stärker weichen die beobachteten Daten von der Annahme der Unabhängigkeit beider Merkmale ab.

Beispiele.

- Schwache Abweichung von der Unabhängigkeit:
  Im Beispiel der Therapiestudie war die Abhängigkeit der Nebenwirkungsrate von der Therapie schwach (20 % versus 17.4 %).

		Merkmal Y : Nebenwirkung eingetreten?				
		ja		nein		
Merkmal X :	Kategorie Nr.	beobachtet	erwartet	beobachtet	erwartet	Summe:
Verabreichte Dosis	hoch	35	32.66	140	142.34	175
	niedrig	32	34.34	152	149.66	184
	Summe:	67		292		359

Tabelle 2.15: Beobachtete und erwartete Häufigkeiten

Die beobachteten und erwarteten Häufigkeiten sind bereits bekannt (Tabelle 2.15):

Beispielsweise ist hier die Anzahl beobachteter Nebenwirkungen in der Hochdosisgruppe etwas höher, in der Niedrigdosisgruppe etwas niedriger als erwartet (35 versus 32.66 bzw. 32 versus 34.34).
Das Abweichungsmaß ist

$$\begin{aligned} X^2 &= \frac{(35-32.66)^2}{32.66} + \frac{(140-142.34)^2}{142.34} + \frac{(32-34.34)^2}{34.34} + \frac{(152-149.66)^2}{149.66} \\ &= 0.402 \end{aligned}$$

- Stärkere Abweichung von der Unabhängigkeit:
  In der bereits zitierten Unfallstudie wurde untersucht, ob es Zusammenhänge zwischen der Schwere eines Unfalls und der Tageszeit gibt. Die Schwere wurde nach dem Grad der schwersten Verletzung der Unfallbeteiligten in zwei Klassen eingeteilt. Ergebnis (Abb. 2.16):

Crosstab

			Maximaler Schweregrad der Verletzungen		
			höchstens leicht verletzt	schwer verl.od. tödl.	Total
Tageszeit	Tag incl. Dämmerung	Count	464	285	749
		Expected Count	432.2	316.8	749.0
		% within Tageszeit	61.9%	38.1%	100.0%
	Nacht	Count	113	138	251
		Expected Count	144.8	106.2	251.0
		% within Tageszeit	45.0%	55.0%	100.0%
Total		Count	577	423	1000
		Expected Count	577.0	423.0	1000.0
		% within Tageszeit	57.7%	42.3%	100.0%

Abbildung 2.16: Beobachtete und erwartete Häufigkeiten von Unfällen

Interpretation:
1. Tagsüber lag der Anteil der Unfälle mit schweren bis tödlichen Verletzungen bei 38.1 %, nachts war dieser Anteil auf 55.0 % erhöht.
2. Bei Unabhängigkeit der Unfallschwere von der Tageszeit würde man bei insgesamt 1000 untersuchten Unfällen nachts 106.2 Unfälle mit schweren bis tödlichen Verletzungen erwarten; stattdessen waren es etwa 32 Unfälle mehr, nämlich 138.
3. Der X^2-Wert beträgt 22.076.

Anmerkungen

1. Der X^2-Wert hängt nicht nur von relativen Häufigkeiten innerhalb der Zeilen ab, sondern sehr stark auch von der Größe der Untersuchungspopulation (wie ändert er sich z.B., wenn alle beobachteten Häufigkeiten einer Kreuztabelle sich um den Faktor 10 erhöhen?). Eine begründete *Bewertung* der Größe eines X^2-Wertes in Bezug auf die Frage nach der Unabhängigkeit wird in der analytischen Statistik angegeben (siehe dort unter "$\chi^2 - Test$").

2. Die Formel

$$X^2 = \sum \frac{(\text{Beobachtete Anzahl - Erwartete Anzahl})^2}{\text{Erwartete Anzahl}}$$

ist auch für Kreuztabellen mit mehr als 2 Zeilen und/oder Spalten anwendbar. Im Fall einer Vierfelder–Tafel kann der X^2 –Wert einfacher nach folgender Formel berechnet werden:

$$X^2 = \frac{(ad - bc)^2 \ n}{(a+b)(c+d)(a+c)(b+d)} \tag{2.2}$$

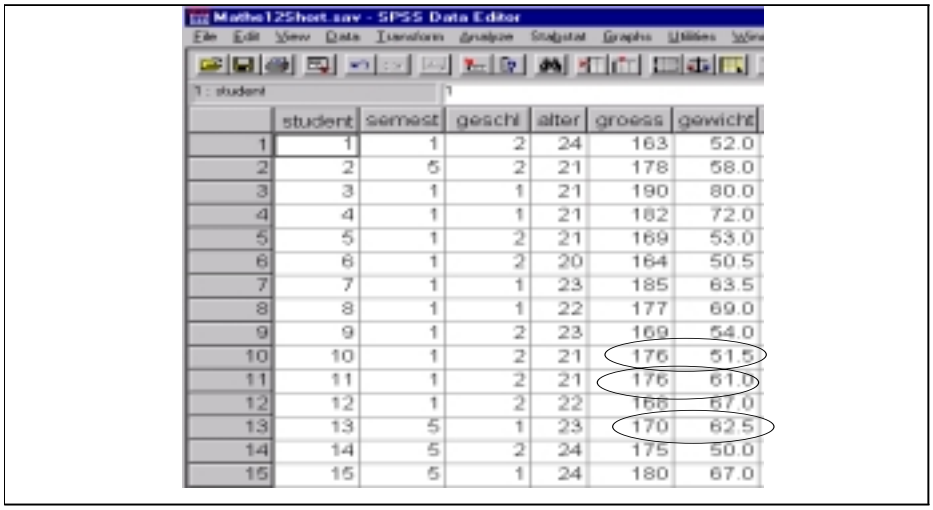
Beispiel für Tabelle 2.10:

$$X^2 = \frac{(35 \times 152 - 140 \times 32)^2 \times 359}{(35+140)(32+152)(35+32)(140+152)} = 0.402$$

2.4.3. Streudiagramme: Korrelation und Regression

Einführung

Beispiel 1. Wie ist der Zusammenhang zwischen Größe und Gewicht?
Das weiß man: Gößere Menschen sind eher auch schwerer als kleinere. Aber um wieviel eigentlich: Wieviel machen z.B. 5 cm Körpergröße im Gewicht aus, und wie sehr hält sich jeder Einzelne an die Regel "Je größer desto schwerer"?
Um das herauszufinden benutzen wir wieder die Daten der Befragung von MHH-Studierenden (Abb. 2.17):



	student	semest	geschl	alter	groess	gewicht
1	1	1	2	24	163	52.0
2	2	5	2	21	178	58.0
3	3	1	1	21	190	80.0
4	4	1	1	21	182	72.0
5	5	1	2	21	169	53.0
6	6	1	2	20	164	50.5
7	7	1	1	23	185	63.5
8	8	1	1	22	177	69.0
9	9	1	2	23	169	54.0
10	10	1	2	21	176	51.5
11	11	1	2	21	176	61.0
12	12	1	2	22	166	57.0
13	13	5	1	23	170	62.5
14	14	5	2	24	175	50.0
15	15	5	1	24	180	67.0

Abbildung 2.17: Größe und Gewicht in einer Rechteckdatei

Man sieht z.B.: "je größer desto schwerer" stimmt bei den ersten 4 Studierenden perfekt. Bei Studentin 11 und 12 gibt es starke Gewichtsunterschiede bei gleicher Größe, und Student 13 ist kleiner aber schwerer als die beiden. Wie verschafft man sich aus den vielen Einzeldaten einen Gesamtüberblick?
Man bildet für alle Studierenden die *Kombinationen* von angegebener Größe und Gewicht und trägt sie als Punkte in ein Koordinatenkreuz ein. Darin sind auf der x-Achse ("Abszisse") die Größe und auf der y-Achse ("Ordinate") das Gewicht aufgetragen. Das dabei entstehende Diagramm wird "*Streudiagramm*" ("*scattergram*") genannt. Man erhält folgendes Bild (Abb. 2.18, linkes Bild):

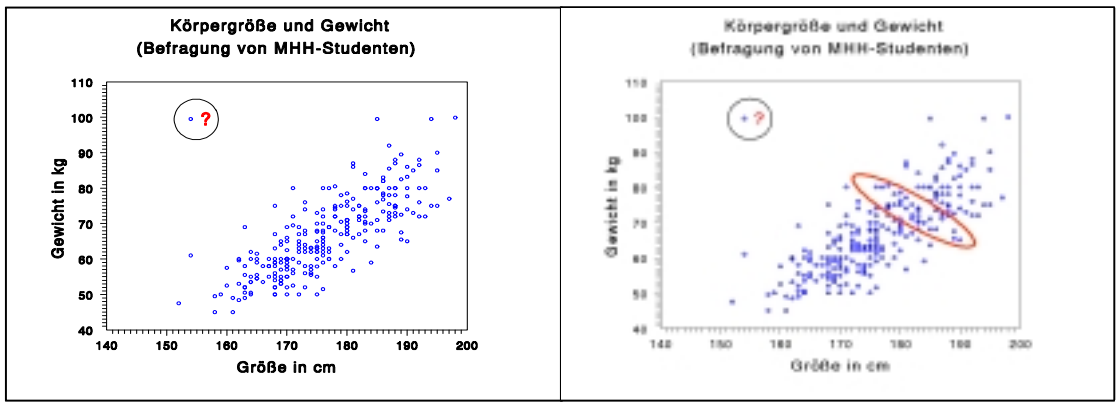


Abbildung 2.18: Größe und Gewicht im Streudiagramm

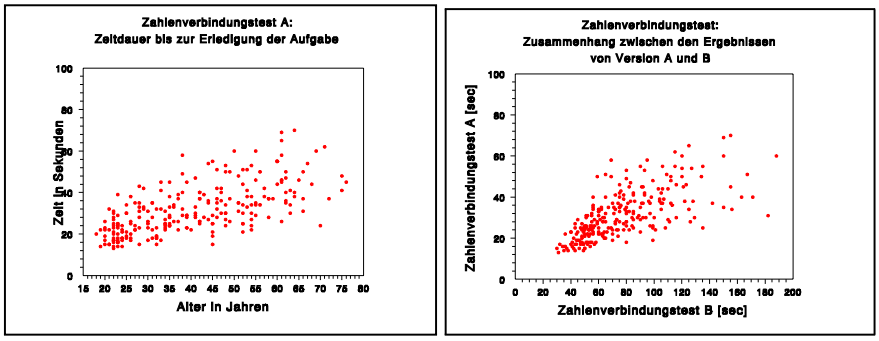
Man erkennt:

1. Die vermutete Beziehung "Je größer desto schwerer" ist als Punktwolke wiederzuerkennen, die sich von links unten nach rechts oben hinzieht.

2. Die Beziehung gilt nur in der Tendenz, nicht für Einzelfälle; sie ist in diesem Sinne *nicht sicher*. Hätte man zufällig nur die Studierenden befragt, die im rechten Bild durch die Ellipse markiert sind, hätte man sogar eine gegenteilige Beziehung vermutet: Je größer desto leichter

3. Es gibt einen Ausreißer in der *Kombination* von Größe und Gewicht (beide Einzelwerte wären noch plausibel, die Kombination ist es nicht; hier muss ein Fehler auf dem Dokumentationsbogen oder bei der Dateneingabe vorliegen).

Beispiel 2. Untersuchungen über den "Zahlenverbindungstest", bei dem die Dauer ermittelt wird, die ein Patient (Proband) braucht, um die auf einem Blatt verteilten Zahlen in aufsteigender Reihenfolge miteinander zu verbinden, zeigten folgende Ergebnisse:



Alter und Testergebnis Testergebnisse von Version A und B

Abbildung 2.19: Zahlenverbindungstest

In beiden Abbildungen erkennt man eine Abhängigkeit zwischen den zwei Merkmalen. Links: Mit zunehmendem Alter (X–Achse) ist die Zeitdauer (Y–Achse), die jemand für den Test braucht –nicht im Einzelfall, aber in der Tendenz– größer. Rechts: Je länger jemand für die Version B des Tests braucht (X–Achse), desto höher "im Schnitt" auch das Ergebnis in Version A. In der rechten Abbildung ist darüber hinaus die Streuung der Y–Werte stärker als in der linken Abbildung davon abhängig, in welchem Bereich der x–Wert liegt. Für beide Abbildungen gilt: Der Zusammenhang ("kleiner x-Wert$\Longleftrightarrow$kleiner y–Wert" und "großer x-Wert$\Longleftrightarrow$großer y–Wert") ist deutlich, aber nicht perfekt.

Beispiel 3. Bilirubin und Thrombozyten sind Merkmale, die beide eine sehr stark rechtsschiefe Verteilung haben. Ihr *Zusammenhang* ist zunächst in Abb. 2.20 (linkes Bild) dargestellt. Wenn man beide Merkmale logarithmiert und in ein Streudiagramm bringt, erhält man wieder eine gestreckte Punktwolke (Abb. 2.20, rechtes Bild).

Im Gegensatz zu den bisherigen Beispielen ist der Zusammenhang hier aber *negativ gerichtet*: *Größere* Bilirubinwerte sind eher mit *kleineren* Thrombozytenzahlen verbunden.

Wie *mißt* man die *Enge des Zusammenhangs* unter der Berücksichtigung der *Richtung* bei dieser Art von (je–desto–) Beziehung?

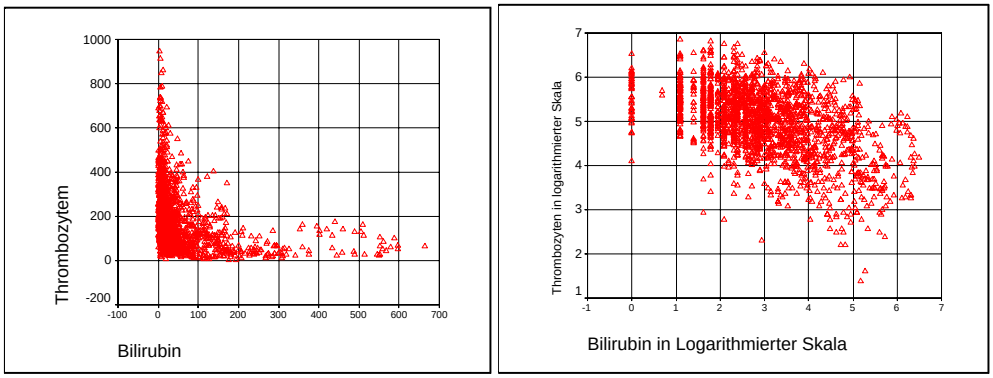


Abbildung 2.20: Bilirubin und Thrombozyten vor und nach Logarithmierung

Der Korrelationskoeffizient.

Man bezeichnet zunächst die beiden Merkmale mit X und Y und die beobachteten Werte in der Reihe ihres Auftretens mit $x_{1,}x_2,\ldots$ bzw. $y_{1,}y_2,\ldots$. Dann bildet man zu jedem Wertepaar (x_i, y_i) die Abweichungen beider Komponenten vom jeweiligen Mittelwert und berechnet das Produkt daraus (das "Kreuzprodukt"):
$(x_i - \bar{x})(y_i - \bar{y})$. Dieses ist z.B. *groß und positiv*, wenn x_i und y_i *beide stark nach oben* oder *beide stark nach unten* von ihren Mittelwerten abweichen. $(x_i - \bar{x})(y_i - \bar{y})$ ist dem Betrag nach *klein*, wenn *beide Werte nahe ihrem Mittelwert* liegen (siehe Abbildung 2.21 und Tabelle 2.16).

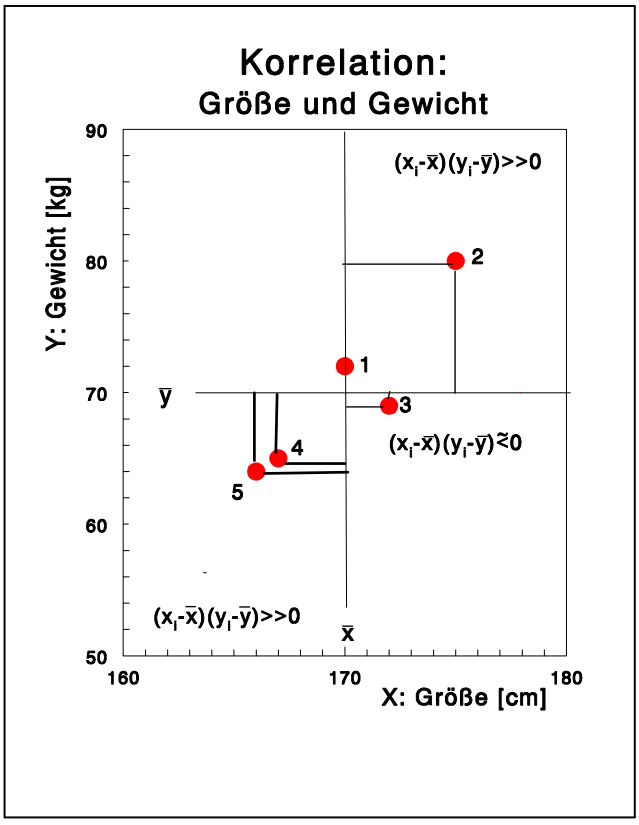


Abbildung 2.21: Die Kreuzprodukte $(x_i - \overline{x})(y_i - \overline{y})$

Daraus folgt: Wenn die "Punktwolke" sehr schmal ist und von links unten nach rechts

Nr.	Größe	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	Gewicht	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	170	0	0	72	2	4	0
2	175	5	25	80	10	100	50
3	172	2	4	69	-1	1	-2
4	167	-3	9	65	-5	25	15
5	166	-4	16	64	-6	36	24
$\sum$:	850		54	350		166	87
$\frac{1}{n}\sum$:	$\bar{x}$ =170			$\bar{y}$ =70			
$\frac{1}{n-1}\sum$:			s_x^2 =13.5			s_y^2 =41.5	s_{xy} =21.75

Tabelle 2.16: Größe und Gewicht: Kreuzprodukte

oben verläuft, überwiegen die stark positiven Werte von $(x_i - \bar{x})(y_i - \bar{y})$: Die "*Stichproben–Kovarianz*" s_{xy}, definiert durch

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{2.3}$$

wird dann stark positiv. (Im Rechenbeispiel der Tabelle 2.16 ist $s_{xy} = 21.75$). Ebenso kann man sich klar machen, dass s_{xy} stark negativ wird, wenn die "Wolke" von links oben nach rechts unten verläuft. Und s_{xy} ist ungefähr gleich Null, wenn keine "je–desto"–Struktur vorhanden ist, d.h. wenn z.B. große x–Werte mit *großen* ebenso wie mit *kleinen* y –Werten gekoppelt sind.
s_{xy} ist daher vom Ansatz her als Zusammenhangsmaß geeignet; es muß aber noch normiert werden, damit es unhabhängig davon ist, in welchen Einheiten X und Y gemessen werden.

Man definiert als *Zusammenhangsmaß* für X und Y:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \; \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2.4}$$

und nennt r_{xy} den *Korrelationskoeffizienten* zwischen X und Y.

Rechenbeispiel für die Tabelle 2.16:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{21.75}{\sqrt{13.5 \times 41.5}} = 0.919$$

Ist $r = 0.92$ eine hohe Korrelation? Zu dieser Frage ist es gut zu wissen, wie hoch eine Korrelation *maximal* sein kann. Dazu überlegt man sich den theoretischen Fall, dass X und Y sogar identisch sind: $x_i = y_i$. Dann würden in Tablle 2.16 die Spalten für $(x_i - \bar{x})$, $(y_i - \bar{y})$ und $(x_i - \bar{x})(y_i - \bar{y})$ alle identisch sein: es wäre also $s_{xy} = s_x^2$ und $s_y = s_x$ und man erhielte $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_x^2}{s_x s_x} = 1$. Bei totalem negativ gerichteten Zusammenhang mit $x_i = -y_i$ käme bei sonst gleicher Rechnung das negative Vorzeichen der Kreuzprodukte zur Geltung und man erhielte: $r_{xy} = -1$. Dies sind die beiden Extremwerte.

- Die Korrelation r_{xy} misst den *linearen Anteil des Zusammenhangs* zweier quantitativer Merkmale.
- Es gilt stets: $\boxed{-1 \leq r_{xy} \leq +1}$
- Bei $\boxed{r_{xy} = -1}$ liegen alle Punkte auf einer Geraden mit *negativer* Steigung.
- Bei $\boxed{r_{xy} = +1}$ liegen alle Punkte auf einer Geraden mit *positiver* Steigung.
- $\boxed{r_{xy} = 0}$ bedeutet, dass kein linearer Zusammenhang zwischen X und Y vorliegt.

Beispiele:

- Größe und Gewicht bei MHH-Studenten (Abb. 2.18): $r = 0.74$
- Zahlenverbindungstest in Abb. 2.19: $r = 0.65$ für das linke und $r = 0.66$ für das rechte Bild. Der lineare Anteil des Zusammenhangs zwischen X und Y wird vom Korrelationskoeffizienten also für beide Bilder etwa gleich bewertet; die optisch erkennbaren Unterschiede in der *Struktur* der Punktwolken werden allerdings durch r_{xy} nicht erfaßt.
- Bilirubin und Thrombozyten in Abb. 2.20: $r = -0.31$ vor Logarithmierung und $r = -0.50$ nach Logarithmierung. Die Logarithmierung hat hier also nicht nur die Schiefe der Verteilungen beseitigt sondern auch zu einer Skala geführt, in der der Zusammenhang beider Laborparameter eher als *linear* erscheint.

Die Regressionsgerade.

Der beschriebene *lineare* Aspekt der Beziehungen zwischen X und Y tritt noch expliziter in den Vordergrund, wenn man fragt:

> *Welches ist der Mittelwert von Y, wenn man nur diejenigen Beobachtungseinheiten betrachtet, bei denen X den vorgegebenen Wert x hat* (z.B: das Durchschnittsgewicht aller Personen mit der Größe 175 cm)? Und *wie ändert sich dies in Abhängigkeit von x?*
>
> Auf den Einzelfall bezogen: *Welchen Wert für Y habe ich bei einer Beobachtungseinheit "zu erwarten", wenn ihr X–Wert bereits bekannt und gleich x* ist*?*

Die Funktion

$$f(x) = \text{Mittelwert von } Y, \text{ falls } X \text{ den Wert } x \text{ annimmt} \qquad (2.5)$$

heißt die "*Regressionsfunktion*".

Im allgemeinen reicht die Anzahl der Beobachtungen nicht aus, um für jeden möglichen x-Wert einer Variablen X einen dazu zugehörigen "stabilen" Mittelwert der Variablen Y

zu erhalten und sich die Regressionsfunktion auf diese Weise zusammenzusetzen. Kann man aber (z.B. aufgrund der Betrachtung des Streudiagramms) annehmen, dass die Regressionsfunktion eine *Gerade* ist, so geht man nicht punktweise, sondern "ganzheitlich" vor:

> *Wähle als Regressionsfunktion diejenige Gerade, zu der die beobachteten Y–Werte in der Summe den kleinsten quadratischen Abstand haben*

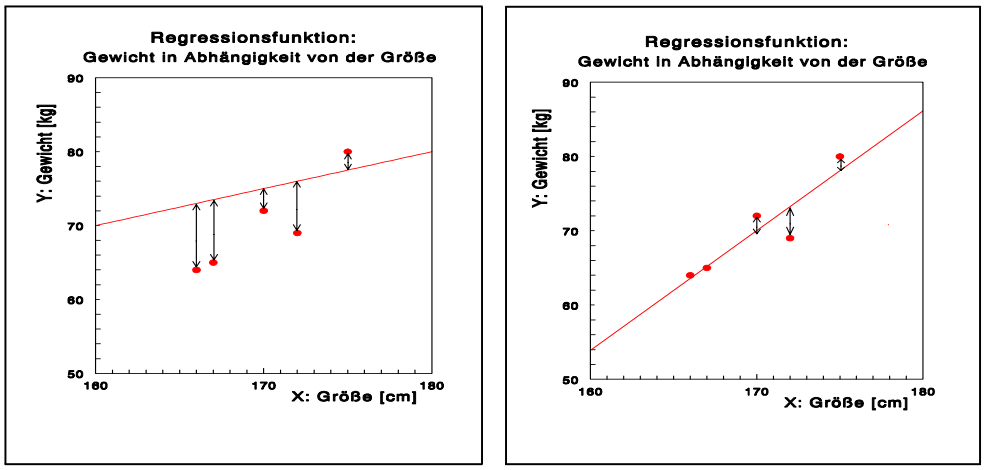
In mathematischer Formulierung:

Wähle den Achsenabschnitt a und die Steigung b der Geraden $y = a + bx$ so, dass

$$s^2_{y|x} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 \tag{2.6}$$

minimal wird.

Beispiel:



Schlechte Anpassung Optimale Anpassung

Abbildung 2.22: Anpassung einer Geraden an die Punktwolke

Die nach diesem Kriterium ("*Methode der kleinsten Quadrate*": *MKQ*) *optimale* Gerade kann aus den Daten direkt ermittelt werden. Die Parameter sind:

$$\begin{aligned} b_{yx} &= \text{Steigung der Regressionsgeraden "}von\ Y\ auf\ X\text{"} \\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ &= \frac{s_{xy}}{s_x^2} \end{aligned} \tag{2.7}$$

$$\begin{aligned} a_{yx} &= \text{Achsenabschnitt der Regressionsgeraden "}von\ Y\ auf\ X\text{"} \\ &= \overline{y} - b_{yx}\overline{x} \end{aligned} \tag{2.8}$$

Aus der Gleichung für den Achsenabschnitt a_{yx} ist auch abzulesen, **dass die Regressionsgerade durch den "Schwerpunkt"** $(\overline{x}, \overline{y})$ **verläuft**: Setzt man in der Gleichung der Regressionsgeraden $a_{yx} + b_{yx}x$ für x den Wert $\overline{x}$ ein, so folgt:

$$\begin{aligned} (\text{Regressionsgerade an der Stelle } \overline{x}) &= a_{yx} + b_{yx}\overline{x} \\ &= (\overline{y} - b_{yx}\overline{x}) + b_{yx}\overline{x} \\ &= \overline{y}, \end{aligned}$$

d.h.: der zum Mittelwert $\overline{x}$ von X gehörige Punkt auf der Regressionsgeraden ist der Mittelwert $\overline{y}$ von Y.

Im folgenden Beispiel sind diese Zusammenhänge noch einmal graphisch dargestellt:

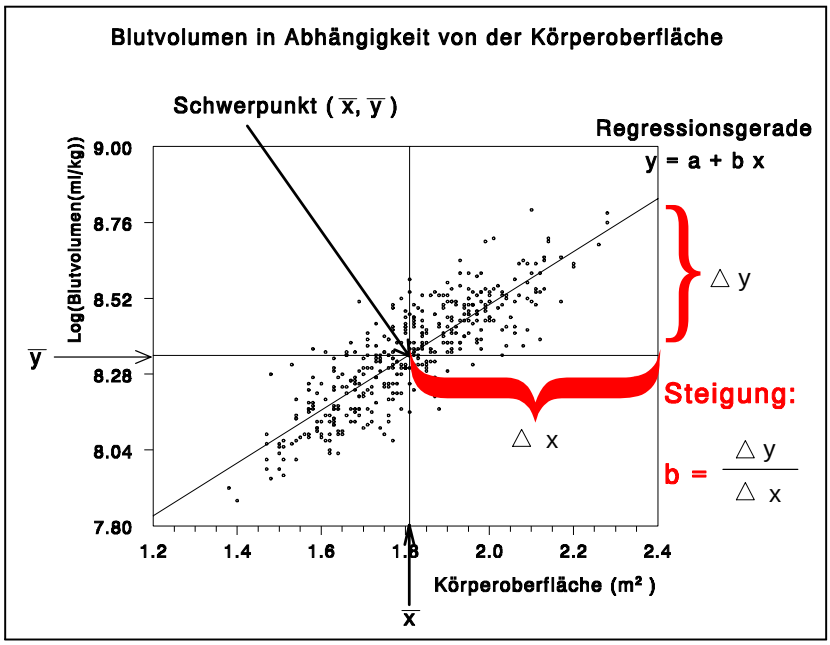


Abbildung 2.23: Schwerpunkt und Steigung einer Regressionsgeraden

Korrelation und Regression hängen eng zusammen: Vergleicht man die Formeln für b_{yx} und r_{xy}, so folgt:

$$\begin{aligned} b_{yx} &= \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x s_y}\frac{s_y}{s_x}, \text{ also} \\ b_{yx} &= r_{xy}\frac{s_y}{s_x} \end{aligned} \tag{2.9}$$

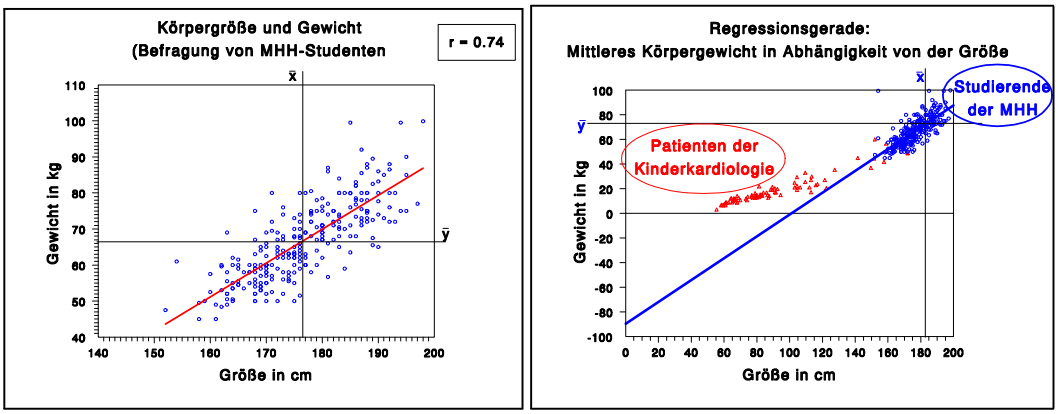
Die Korrelation r_{xy} ist symmetrisch bezüglich X und Y ($r_{xy} = r_{yx} = r$). Bei der Regresssion ist es aber wesentlich, welche der beiden Variablen man im Sinne der Definition 2.5 als *gegeben* ("unabhängig") und welche man als davon "*abhängig*" betrachtet: Die jeweils resultierenden Geraden sind *nicht deckungsgleich.* (Denn dann müßte das Produkt der Steigungen gleich 1 sein. Es gilt aber: $b_{yx} \times b_{xy} = r_{xy}\frac{s_y}{s_x}\; r_{xy}\frac{s_x}{s_y} = r^2$.)

Interpretation der Regressionsgeraden $y = a + bx$:

- Für Beobachtungseinheiten, deren Messwert des Merkmales X bekannt und gleich x ist, ist der mittlere Wert des Merkmals Y gleich $\boxed{a + bx.}$
- Beobachtungseinheiten, deren X–Wert um eine Einheit größer ist als der von anderen Fällen, haben im Mittel einen um b höheren Y–Wert.

Beispiel

Die Regressionsgerade für das Gewicht in Abhängigkeit von der Größe zum Streudiagramm 2.18 von MHH-StudentInnen ist in Abbildung 2.24 dargestellt.



Regressionsgerade im Erhebungsbereich Extrapolation nach links

Abbildung 2.24: Regression: Größe und Gewicht

Ihre Gleichung lautet:

$$y\ (\textit{mittleres Gewicht in kg}) = -99.3 + 0.94 \times \textit{Größe[cm]}$$

StudentInnen, die 1 cm größer sind als ihre KommilitonInnen, sind daher im Schnitt um 0.94 kg schwerer als diese. *5 cm größer* bedeutet $5 \times 0.94 = 4.7$ *kg schwerer.*
Aufgabe: Vergleichen Sie diese Regressiongleichung mit der "Broca-Formel":

$$\boxed{\text{Gewicht[kg]} = \text{Größe}\textit{[cm]} - 100.}.$$

Setzen Sie dazu verschiedene Wert für die Größe x in beide Gleichungen ein. Ist das "Broca-Gewicht" im Vergleich zu den Daten der MHH-Befragung daher eher als das *mittlere* Gewicht oder als eine *Normgrenze* zu interpretieren?

Wichtiger Hinweis: Wenn die Annahme einer linearen Regression gut zu den Daten passt, so darf sie dennoch nicht beliebig in die eine oder die andere Richtung verlängert werden.
Dies wird im rechten Bild der Abbildung 2.24 verdeutlicht: Die Punktwolke der Daten von Kindern passt *gar nicht* zu der Regressionsgeraden, die an Hand der Erwachsenenendaten ermittelt wurde.

Interpretation des Korrelationskoeffizienten r_{xy} im Rahmen der Regressionsanalyse:
Die (quadrierten) Abweichungen der Y–Werte y_i *von der Regressionsgeraden* sind (in der Summe) kleiner als die vom Mittelwert $\overline{y}$, denn die Regressionsgerade minimiert ja diese Abweichungen. Dies wird in der folgenden Abbildung 2.25 noch einmal veranschaulicht:

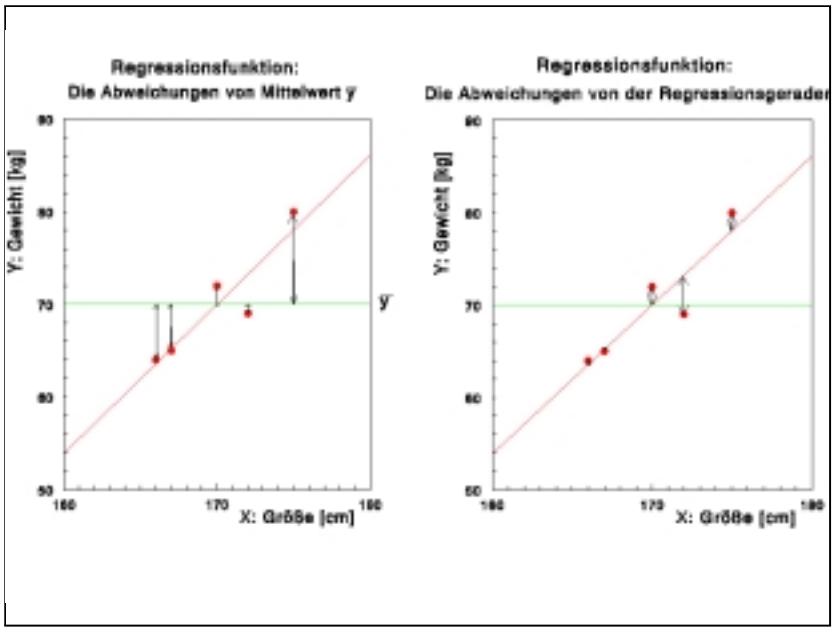


Abbildung 2.25: Die Abweichungen der Beobachtungen Y vom Mittelwert und von der Regressionsgeraden

Die Varianz von Y wird also *durch Bezug auf die Regressionsgerade verringert* und daher teilweise "*durch die Variable X erklärt*". Man kann zeigen, dass die relative Varianzreduktion gerade gleich dem quadrierten Korrelationskoeffizienten ist (vgl.(2.6)):

$$\frac{s_y^2 - s_{y|x}^2}{s_y^2} = r^2 \tag{2.10}$$

> r^2 **wird als Bestimmtheitsmaß bezeichnet und gibt an, welcher Anteil der Varianz von Y durch X erklärt wird**.

Beispiele:

- Größe und Gewicht bei MHH-Studenten (Abb. 2.18): $r = 0.74$. Die Varianz des Körpergewichts unter den MHH-Studierenden wird zu $0.74^2 = 0.55 = 55\ \%$ durch unterschiedliche Körpergrößen erklärt.
  Man berechne und interpretiere in gleicher Weise die Werte für r^2 für die bereits bekannten Beispiele:

- Zahlenverbindungstest in Abb. 2.19: $r = 0.65$ für das linke und $r = 0.66$ für das rechte Bild.
- Bilirubin und Thrombozyten in Abb. 2.20: $r = -0.31$ vor Logarithmierung und $r = -0.50$ nach Logarithmierung.

3. Wahrscheinlichkeitsrechnung (GK 3)

3.1. Einführung

Die Ergebnisse empirischer Untersuchungen (zusammengefasst in Häufigkeitsverteilungen und statistischen Kenngrößen) sind in der Regel nicht exakt reproduzierbar.
Beispiel:
In einer Studie über den Einfluß eines Medikaments auf die Hemmung des Fortschreitens der Koronararteriosklerose im Vergleich zu einem Placebo trat bei den ersten 20 Patienten der Verum-Gruppe in 55 % der Fälle eine Verschlechterung ein, bei den nächsten 20 Fällen dieser Gruppe in 35 %; in der Placebo-Gruppe verschlechterten sich 50 % der ersten 20 Fälle und 45 % der nächsten 20 Patienten

Aber:

> **Je größer der Stichprobenumfang, desto "stabiler" sind die Ergebnisse.**

Beispiel: Therapiestudie.

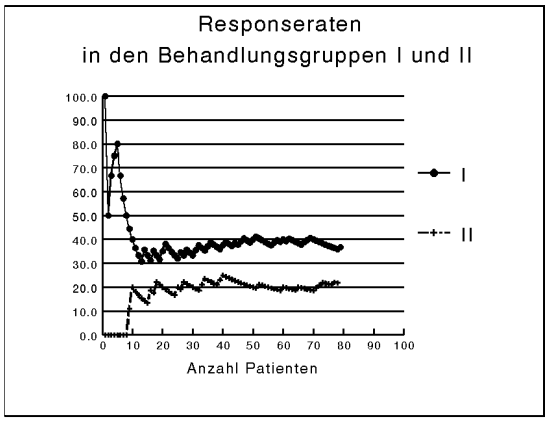


Abb. 3.1: Häufigkeitsraten in Abhängigkeit vom Stichprobenumfang

Hier sind für zwei unterschiedliche Behandlungsgruppen aus einer Therapiestudie die *Anteile der Patienten mit Therapieerfolg* in Abhängigkeit vom Stichprobenumfang aufgetragen. Aufgrund der Abbildung ist die Annahme naheliegend, dass die relative Häufigkeit für das Ergebnis "Therapieerfolg" mit wachsendem Stichprobenumfang in jeder Gruppe einem Grenzwert zustrebt. Dieser liegt etwa bei 38% in Gruppe I und etwa bei 22% in Gruppe II.

Es handelt sich hier in folgendem Sinne um ein *Zufallsexperiment*:

1. Das Experiment wird nach einer genau festgelegten Vorschrift durchgeführt.
2. Das Experiment kann unter den gleichen Bedingungen beliebig oft wiederholt werden.

3. Es sind mehrere Ergebnisse des Experimentes möglich. Sämtliche überhaupt möglichen Ergebnisse können vor Durchführung des Experimentes angegeben werden.

4. Es kann nicht mit Sicherheit angegeben werden, welches Ergebnis sich bei Durchführung des Zufallsexperimentes einstellen wird (das Ergebnis hängt vom Zufall ab).

Grundannahme:

> Wird ein Zufallsexperiment unter gleichen Bedingungen beliebig oft wiederholt, und bezeichnet A ein bestimmtes mögliches Ergebnis jedes Einzelexperimentes, so strebt die **relative Häufigkeit** $h_n(A)$ **für das Ergebnis** A **für wachsenden Stichprobenumfang** n **gegen einen Grenzwert.**

Danach ist jedes mögliche Ergebnis A eines Zufallsexperimentes mit dem *"Grenzwert" der relativen Häufigkeiten* $h_n(A)$ verbunden
Dieser Grenzwert wird als die *Wahrscheinlichkeit für das Ergebnis* A interpretiert. Man bezeichnet ihn mit $P(A)$, wobei P für "Probability" steht.

$$h_n(A) \longrightarrow P(A) \tag{3.1}$$

Die Wahrscheinlichkeit $P(A)$ ist nach diesen Annahmen also eine Eigenschaft des betrachteten Ergebnisses A und wird durch die Art und die Rahmenbedingungen des Experimentes bestimmt. Im Beispiel der Abb. 3.1 etwa ist die Wahrscheinlichkeit für das Ergebnis A = "Therapieerfolg" im Experiment I (Behandlung mit Therapie I) offenbar größer als im Experiment II (Behandlung mit Therapie II): **Therapie I hat eine größere Erfolgswahrscheinlichkeit als Therapie II.**
Folgerung:
Therapien sind in ihrer Anwendung letztlich durch **Wahrscheinlichkeiten** charakterisiert. Eine klinische Studie soll Aufschlüsse über diese **Wahrscheinlichkeiten** liefern. Man nähert sich den verborgenen, unbekannten Wahrscheinlichkeitswerten dadurch, dass man **relative Häufigkeiten** bildet und diese als **Schätzwerte für die Wahrscheinlichkeiten** ansieht.

Das führt zu folgenden Unsicherheiten:

- Wie stark kann die relative Häufigkeit von der dahinter liegenden Wahrscheinlichkeit abweichen?
- In welcher Weise hängt dies vom Stichprobenumfang ab?
- Wie kann man entscheiden, ob die Erfolgs*wahrscheinlichkeiten* zweier Therapien tatsächlich unterschiedlich sind, wenn die beobachteten Erfolgs*raten* sich beispielsweise um 16 %-Punkte unterscheiden wie in Abb. 3.1?

Um solche Fragen beantworten zu können, ist es nötig, sich näher damit zu beschäftigen, was in Zufallsexperimenten "so alles möglich ist": Welche Ergebnisse können mit welcher Wahrscheinlichkeit "rein zufällig" auftreten?

Unter bestimmten Annahmen kann man so etwas *berechnen,* ohne irgendein Experiment durchzuführen, ganz allein mit Hilfe mathematischer Methoden: der "*Wahrscheinlichkeitsrechnung*". Die Regeln für die Wahrscheinlichkeitsrechnung gründen auf der engen Beziehung zwischen den relativen Häufigkeiten $h_n(A)$ und den dahinter liegenden Wahrscheinlichkeiten $P(A)$: **Alles was für das Rechnen mit relativen Häufigkeiten gilt, muss auf das Rechnen mit Wahrscheinlichkeiten übertragbar sein.** Allein auf Grund dieser Forderung lassen sich schon die Grundregeln herleiten.

3.2. Grundregeln der Wahrscheinlichkeitsrechnung:

Bezeichnungen:
In der deskriptiven Statistik wurden bei diskreten Merkmalen X die *beobachteten Werte von* X betrachtet und deren relative Häufigkeiten aus einer Beobachtungsserie berechnet. In der Wahrscheinlichkeitsrechnung steht nun im Vordergrund, dass man ein *Zufallsexperiment* betrachtet (ohne dass es wirklich ausgeführt werden muß) und *dass jedes mögliche Ergebnis* A *dieses Experimentes mit einer Wahrscheinlichkeit* $P(A)$ *versehen* ist. Diese möglichen Ergebnisse werden "*Ereignisse*" genannt. Dabei handelt es sich nicht nur um Einzelwerte eines Merkmales, sondern allgemein auch um zusammengesetzte Charakterisierungen des Ausgangs eines Experimentes wie z.B.: $A =$ "Blutdruck um ≥ 10 mmHg gesenkt", $B =$ "Blutdruck nicht gestiegen, keine Nebenwirkungen" etc. Man geht aber davon aus, dass man immer nur ein "festes System" von Ereignissen betrachtet, d.h.:

1. Mit den Ereignissen A und B gehört auch das Ereignis $A \cup B$: ("Das Ereignis A oder das Ereignis B oder beide Ereignisse treten ein") zu diesem System. Bezeichnung: die *Vereinigung* von A und B.

2. Mit den Ereignissen A und B gehört auch das Ereignis $A \cap B$: ("Das Ereignis A *und* das Ereignis B tritt ein") zum System. Bezeichnung: der *Durchschnitt* von A und B.

3. Mit dem Ereignis A gehört auch das Ereignis "Nicht–A" , Schreibweise $\overline{A}$: ("Das Ergebnis A tritt *nicht* ein") zum System. Bezeichnung: das *Komplement* von A.

4. Die Gesamtheit *aller* möglichen Ausgänge des Experimentes, bezeichnet als das Ereignis Ω, und das Komplement dazu, die "leere Menge", das Ereignis $\emptyset$, gehören zum System.

Wird nun das untersuchte Experiment mehrfach, und zwar n mal durchgeführt, so kann man aus den Ergebnissen für alle betrachteten Ereignis $A, B, \ldots$ die relativen Häufigkeiten $h_n(A), h_n(B), \ldots$ auszählen und daraus auch die relativen Häufigkeiten für $A \cup B$ usw. berechnen. Führt man dies aus, so stellt man fest, dass z.B. die folgende Rechenregel gilt:

$$h_n(\overline{A}) = 1 - h_n(A)$$

d.h. die relative Häufigkeit für die Experimente, in denen A *nicht* eingetreten ist, ist gleich 1 minus der relativen Häufigkeit der Experimente mit dem Ergebnis A.
Aus dem oben beschriebenen **Grundsatz,** wonach die Regeln der Wahrscheinlichkeitsrechnung den Regeln entsprechen müssen, die für relative Häufigkeiten gelten, lässt sich nun herleiten:

1. Für alle "Ereignisse" A gilt
$$0 \leq P(A) \leq 1 \tag{3.2}$$

2. Bezeichnet Ω die Gesamtheit *aller möglichen Ausgänge* eines Experimentes, so gilt:
$$P(\Omega) = 1 \tag{3.3}$$

3. Falls die Ereignisse A und B sich *gegenseitig ausschließen* (man sagt: A und B sind "*disjunkt*", ihr Durchschnitt ist *leer:* $A \cap B = \emptyset$), so gilt:
$$P(A \cup B) = P(A) + P(B) \tag{3.4}$$

4. Bezeichnet $\bar{A}$ das Ereignis "A tritt nicht ein", so gilt:
$$P(\bar{A}) = 1 - P(A) \tag{3.5}$$

5. Für irgend zwei Ereignisse A und B gilt:
$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \tag{3.6}$$
Dabei bezeichnet $A \cap B$ das Ereignis: "Sowohl A als auch B ist eingetreten".

Beispiel: Herleitung der Rechenregel 5

Experiment: Per Zufall einen Probanden aus der Gesamtheit der Bevölkerung der Stadt Hannover im Alter von 50 Jahren wählen und ärztlich untersuchen. Dieses Einzelexperiment n mal durchführen.
Betrachtete "Ereignisse" und relative Häufigkeiten (siehe auch nachfolgende Tabelle):

A: Risikofaktor Rauchen vorhanden; $h(A) = \frac{a+b}{n}$

B: Risikofaktor Hypertonie vorhanden; $h(B) = \frac{a+c}{n}$

C: Mindestens ein Risikofaktor; Rauchen oder Hypertonie (oder beides):
$C = A \cup B$; $\quad h(A \cup B) = \frac{a+b+c}{n}$

D: Beide Risikofaktoren; Rauchen und Hypertonie: $D = A \cap B$; $\quad h(A \cap B) = \frac{a}{n}$

		Hypertonie		
		$B:+$	-	
Rauchen	$A:+$	$A \cap B):\ a$	b	$a+b$
	-	c	d	
		$a+c$		n

$$\begin{aligned} h(A \cup B) &= \frac{a+b+c}{n} \\ &= \frac{(a+b)+(a+c)-a}{n} \\ &= \frac{a+b}{n} + \frac{a+c}{n} - \frac{a}{n} \\ &= h(A) + h(B) - h(A \cap B) \end{aligned}$$

Durch diese Regeln ist die Basis für die gesamte Wahrscheinlichkeitsrechnung gelegt. Streng genommen gehören sogar nur die Regeln 1. bis 3. dazu; die anderen sind hieraus bereits ableitbar.

3.3. Vom Nutzen der Regeln zur Wahrscheinlichkeitsrechnung

Die genannten Regeln (außer Nummer 1) sind *Formeln.* Sie können zu zweierlei verwendet werden:

1. zur numerischen Berechnung "neuer" Wahrscheinlichkeiten, wenn andere Wahrscheinlichkeiten bereits bekannt sind.
   Beispiel (Regel 4): Wenn die Wahrscheinlichkeit für Therapieerfolg $0.38 = 38\,\%$ beträgt, dann ist die Wahrscheinlichkeit, dass die Therapie *keinen* Erfolg hat, gleich $1 - 0.38 = 0.62 = 62\,\%$.

2. zur Ableitung neuer Formeln für komplexere Ereignisse oder *unter zusätzlichen Annahmen.*

Zu Punkt 2 ein Beispiel:
Wie wahrscheinlich ist es, eine "6" zu würfeln?
Spontane Antwort: 1/6. Ja, aber warum? Herleitung:

1. Es gibt 6 mögliche "*Elementarereignisse*" (d.h.: nicht aus anderen zusammengesetzte Ereignisse) des Experimentes "Würfeln": die Einzelergebnisse 1 bis 6.

2. Die Summe der zugehörigen Wahrscheinlichkeiten muss den Wert 1 ergeben (Anwendung von Regel 2 und Regel 3).

3. **Zusatzannahme: Alle 6 Elementarereignisse haben dieselbe Wahrscheinlichkeit.**

4. Wenn man diese mit p bezeichnet, muss also $6 \times p = 1$ sein. Also ist $p = 1/6$.

Für die Berechnung der Wahrscheinlichkeit eines *zusammengesetzten* Ereignisses A braucht man dann nur die Anzahl der Elementarereignisse zu zählen, aus denen sich A zusammensetzt, und diese Anzahl mit $p = 1/6$ zu multiplizieren.
Beispiel: Wahrscheinlichkeit für eine *gerade* Augenzahl. $A = \{2, 4, 6\}$ besteht aus 3 Elementarereignissen. Also: $P(A) = 3 \times 1/6 = 1/2$.
Wesentlich an dieser Herleitung war, dass sie nur unter einer *Zusatzannahme* möglich war: *Der Würfel muss unverfälscht sein!* Dahinter steht ein Grundprinzip, das weiterhin ständig angewendet werden muss:

Zur Herleitung neuer Formeln und zur numerische Berechnung von "neuen" Wahrscheinlichkeiten benötigt man:

- Kenntnisse über den numerischen Wert anderer Wahrscheinlichkeiten
  und/oder
- die Gültigkeit bestimmter *Zusatzannahmen.*

Das Würfelbeispiel kann übrigens ohne Probleme verallgemeinert werden:

Allgemeine Regel zur Berechnung von Wahrscheinlichkeiten
Sind alle Elementarereignisse *gleichwahrscheinlich* und ist

N = Die Anzahl *aller* Elementarereignisse

n_A = Die Anzahl der Elementarereignisse, aus denen sich A zusammensetzt

so gilt:

$$P(A) = \frac{n_A}{N} = \frac{\text{Anzahl "günstiger" Fälle für } A}{\text{Anzahl "möglicher" Fälle}}$$

Dies ist die Regel von LAPLACE. Sie ist denkbar einfach, kann aber in der Anwendung sehr kompliziert werden, wenn N und n_A sich kompliziert zusammensetzen. (Wie wahrscheinlich ist es beispielsweise, dass beim Austeilen der Karten des Skatspiels der erste Spieler mindesten 3 Buben erhält? Wie wahrscheinlich ist es, dass *irgendeiner* der 3 Spieler mindestens 3 Buben erhält? Was ist da ein Elementarereignis? Wie setzt sich A zusammen?)

3.4. Bedingte Wahrscheinlichkeiten

Man betrachte zwei Ereignisse A und B eines Zufallsexperimentes und überlege sich, ob durch das Eintreten von B die Chancen für das Eintreten von A *verändert* werden oder *unverändert* bleiben.
Beispiel:

1. Die Wahrscheinlichkeit für 3 Richtige im Zahlenlotto (Ereignis A) schätzt man *höher* ein, *wenn man bei der Ziehung der Lottozahlen die 1. gezogene Kugel richtig hat* (Ereignis B).

2. Die Wahrscheinlichkeit, dass eine Mutter als *zweites Kind einen Jungen* (Ereignis A) bekommt, schätzt man als *unabhängig* davon ein , dass sie *als erstes Kind ein Mädchen* bekommen hat (Ereignis B).

Ereignisse können demnach *unabhängig* voneinander auftreten, oder es kann eine *Beziehung zwischen Ereignissen* vorhanden sein. So gibt es einen *Zusammenhang* zwischen *Symptom und Diagnose* oder *Diagnose und Prognose*. Man spricht in der Statistik von *abhängigen Ereignissen*, wenn es möglich ist, von einer Größe auf die andere Rückschlüsse zu ziehen:
Zwei Ereignisse A und B sind voneinander *abhängig*, wenn die Wahrscheinlichkeit für das Auftreten des Ereignisses A *verändert* wird, wenn das Ereignis B eingetreten ist.

Zur mathematischen Formulierung solcher Überlegungen benötigen wir zunächst die Definition der "*bedingten Wahrscheinlichkeit*". Dazu gehen wir zunächst wieder zu den relativen Häufigkeiten zurück. Die Analogie zu den Wahrscheinlichkeiten ermöglicht dann die Übertrageung auf die Wahrscheinlichkeitsrechnung.
Beispiel:
Man betrachte ein "Zufallsexperiment", bei welchem einem Patienten aufgrund einer Zufallszuteilung ("Randomisation") ein Verum– oder ein Placebo–Präparat zugeteilt wird. Nach Abschluß der Behandlung wird das Ergebnis der Therapie –zusammengefaßt als "Erfolg" oder "kein Erfolg"– notiert. Eine Darstellung der Gesamtheit aller Elementarereignisse dieses Experimentes in einer Vierfeldertafel, zusammen mit den beobachteten Häufigkeiten nach n-facher Durchführung des Experimentes, kann dann so

aussehen:

		Erfolg		
		B : Erfolg	kein Erfolg	Summe:
Therapie	A :Verum	$A \cap B : a$	b	$a+b$
	Placebo	c	d	$c+d$
	Summe:	$a+c$	$b+d$	n

Ein Therapievergleich wird dann in der Regel auf dem Vergleich der relativen Häufigkeiten für das Ergebnis "Erfolg" innerhalb der beiden Zeilen ("Verum" versus "Placebo") beruhen: Man bildet die relativen Häufigkeiten für "Erfolg" *unter der Bedingung, dass die Therapie das Verum ist* und vergleicht sie mit der relativen Häufigkeit für "Erfolg" *unter der Bedingung, dass Placebo gegeben wurde* (siehe auch Abschnitt 2 Seite 24: die **"Zeilenprozente"**). Entsprechend werden diese Quotienten auch "*bedingte relative Häufigkeiten*" genannt: Sei

A das Ereignis "Zuteilung zu Verum" $\quad (h(A) = \frac{a+b}{n})$

B das Ereignis "Erfolg der Therapie" $\quad (h(B) = \frac{a+c}{n})$, und

$A \cap B$ das Ereignis "Verum–Therapie angewendet und Erfolg der Behandlung"
$(h(A \cap B) = \frac{a}{n})$

Man bildet dann also als Maß für den Erfolg in der Verum-Gruppe die "*bedingte* relative Häufigkeit für B *gegeben A",* bezeichnet als $h(B|A)$, und findet:

$$h(B|A) = \frac{a}{a+b} = \frac{a/n}{(a+b)/n} = \frac{h(A \cap B)}{h(A)}$$

Hierzu als analoges Beispiel für "Dosis und Nebenwirkung" die Tabelle von Seite 24 in neuer Terminologie:

	Merkmal Y : Nebenwirkung eingetreten?			
Merkmal X :	Kategorie .	B : keine Nebenwirkungn	Nebenwirkung	Summe:
Verabreichte Dosis	A : niedrige Dosis	$A \cap B$: 152	32	184
	hohe Dosis	140	35	175
	Summe	292	67	359

$$\begin{aligned} \text{Relative Häufigkeit für } B \text{ gegeben } A &= \frac{152}{184} = \frac{152/359}{184/359} = 0.826 \\ &= \frac{\text{relative Häufigkeit für } (A \cap B)}{\text{relative Häufigkeit für } (A)} \end{aligned}$$

Entsprechend werden nun auch *bedingte Wahrscheinlichkeiten* definiert:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \text{"Bedingte Wahrscheinlichkeit für } B \text{ gegeben } A\text{"}$$

Man schließt hier also einen Teil der möglichen Ergebnisse des Experimentes aus: es werden nur diejenigen Ereignisse betrachtet, in denen jedenfalls auch A eingetreten ist. Dies geschieht in der Formel im *Zähler* durch die *Schnittbildung mit dem Ereignis A*.

Die Division durch P(A) ist dann nötig, um die *Wahrscheinlichkeit wieder auf 1 zu normieren.*

Analog definiert man

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \text{"Bedingte Wahrscheinlichkeit für } A \text{ gegeben } B\text{"}$$

Diese Definition setzt jeweils voraus, dass die im Nenner stehende Wahrscheinlichkeit nicht gleich 0 ist.

3.5. Stochastische Unabhängigkeit von Ereignissen

Die eingangs dargestellten Überlegungen zur Unabhängigkeit von Merkmalen können nun konkreter definiert werden:
In der Tabelle über Behandlung und Therapieergebnis ist der Therapieerfolg (B) von der Zuteilung zur Verumgruppe (A) *unabhängig*, wenn die Wahrscheinlichkeit für Erfolg in der Verumgruppe (A), also $P(B|A)$, genauso groß ist wie in der Gesamtpopulation: $P(B)$, wenn also gilt :

$$\begin{aligned} P(B|A) &= P(B), \qquad \text{d.h. wenn} \\ \frac{P(A \cap B)}{P(A)} &= P(B), \qquad \text{wenn also} \\ P(A \cap B) &= P(A)\, P(B) \end{aligned} \tag{3.10}$$

Dies ist die *Definition* der (stochastischen[1]) *Unabhängigkeit zweier Ereignisse* A und B.

In den Anwendungen der Wahrscheinlichkeitsrechnung wird die Unabhängigkeit zweier Ereignisse oft nicht durch die *Gültigkeit der Gleichung (3.10) nachgewiesen*; vielmehr geht man umgekehrt *aufgrund der Anordnung des Zufallsexperimentes* davon aus, dass A und B unabhängig sind, und nutzt dieses aus, um die Wahrscheinlichkeit für ihr gemeinsames Eintreten, also um $P(A \cap B)$ zu berechnen.

Beispiel 1 (ganz einfach) :

Experiment: 2–maliger Münzwurf:
Ereignis A: Erster Wurf = Zahl
Ereignis B: Zweiter Wurf = Zahl
Ereignis $A \cap B$: Beide Würfe Zahl

$$P(A \cap B) = P(A)\, P(B) = \frac{1}{2}\,\frac{1}{2} = \frac{1}{4}$$

Beispiel 2:
Es sei $p = 0.9$ die Wahrscheinlichkeit für Therapieerfolg einer Einzelbehandlung. Wie groß ist dann die Wahrscheinlichkeit, dass bei der Behandlung von 10 Patienten

- alle Behandlungen erfolgreich sind?

[1] Unter "Stochastik" (aus dem Griechischen: "vermuten", "erraten") wird der Wissenschaftsbereich verstanden, der sich mit der mathematischen Behandlung von Zufallserscheinungen befaßt.

- keine Behandlung erfolgreich ist?
- mindestens eine Behandlung *ohne* Erfolg ist?

Lösung:

$$\begin{aligned} p &= \text{Wahrscheinlichkeit für Therapieerfolg } = 0.9 \\ n &= 10 \text{ unabhängige Behandlungen} \end{aligned}$$

$$P(\text{alle Behandlungen erfolgreich})) = \underbrace{p\ p\ p\ \ldots\ p}_{10 \text{ mal}} = p^{10} = 0.9^{10} = 0.349 = 34.9 \quad \%$$

$$\begin{aligned} P(\text{keine Behandlung erfolgreich})) &= (1-p)\ (1-p)\ (1-p)\ \ldots (1-p) = (1-p)^{10} \\ &= 0.1^{10} = 0.000\ 000\ 000\ 1 \end{aligned}$$

$$\begin{aligned} P(\text{mindestens 1 Behandlung } \textit{ohne} \text{ Erfolg} &= 1\ - P(\text{alle Behandlungen erfolgreich}) \\ &= 1 - 0.349 = 0.651 = 65.1 \quad \% \end{aligned}$$

3.6. Zerlegung von Wahrscheinlichkeiten

Aus dem letzten Beispiel geht hervor, dass man mit Hilfe der Regeln der Wahrscheinlichkeitsrechnung aus bekannten Wahrscheinlichkeiten (meist für "einfache" Ereignisse) auf die Wahrscheinlichkeiten von zusammengesetzten, komplexer strukturierten Ereignissen schließen kann. In diesem Zusammenhang ist auch die folgende Zerlegung von Ereignissen und zugehörigen Wahrscheinlichkeiten zu sehen:
Ein Ereignis B ist entweder mit dem Ereignis A gekoppelt oder mit dem Ereignis "Nicht–A", d.h. mit $\overline{A}$:

$$B \ = \ (B \cap A) \ \cup \ (B \cap \bar{A})$$

und die Ereignisse$(B \cap A)$ und $(B \cap \bar{A})$ sind disjunkt (weil ja schon A und$\bar{A}$ disjunkt sind). Aus dieser Zerlegung von B folgt für die Wahrscheinlichkeiten:

$$\begin{aligned} P(B) &= P(B \cap A) \ + \ P(B \cap \bar{A}) \\ &= \frac{P(B \cap A)}{P(A)} P(A) \ + \frac{P(B \cap \bar{A})}{P(\bar{A})} P(\bar{A}), \qquad \text{also:} \\ P(B) &= P(B|A)P(A) \ + P(B|\bar{A})P(\bar{A}) \end{aligned} \tag{3.11}$$

Folgerung: Die Wahrscheinlichkeit für B läßt sich berechnen, wenn man die *bedingten Wahrscheinlichkeiten* für B zu zwei gegebenen komplementären Ereignissen und deren Wahrscheinlichkeiten kennt.

Beispiel:
Bei Tumorpatienten sei die Wahrscheinlichkeit für Lungenmetastasen (A) gleich 55%.
Bei Patienten *mit* Lungenmetastasen sei die Wahrscheinlichkeit für Erfolg einer Chemotherapie $((B|A))$ gleich 60%:

$$\begin{aligned} P(A) &= 0.55 \\ P(B|A) &= 0.6 \end{aligned}$$

Bei Patienten *ohne* Lungenmetastasen (Warsch. $= 1 - 0.55 = 45\%$) sei die Wahrscheinlichkeit für Erfolg gleich 90%:

$$\begin{aligned} P(\bar{A}) &= 0.45 \\ P(B|\bar{A}) &= 0.9 \end{aligned}$$

Dann ist die Wahrscheinlichkeit für Erfolg *in der Gesamtpopulation*:

$$P(B) = 0.6 \times 0.55 \; + \; 0.9 \times 0.45 = 0.330 \; + 0.405 = 0.735 \; = \; 73.5\%.$$

Die Gleichung (3.11) stellt also die Wahrscheinlichkeit für B als eine "*gewichtete Summe*" der bedingten Wahrscheinlichkeiten für B unter komplementären Bedingungen dar.

3.7. Satz von BAYES: Der Salto rückwärts

Fragestellung: Ein Patient kommt mit Fieber in die Sprechstunde. Wie wahrscheinlich ist es, dass er an einem Grippevirus erkrankt ist?

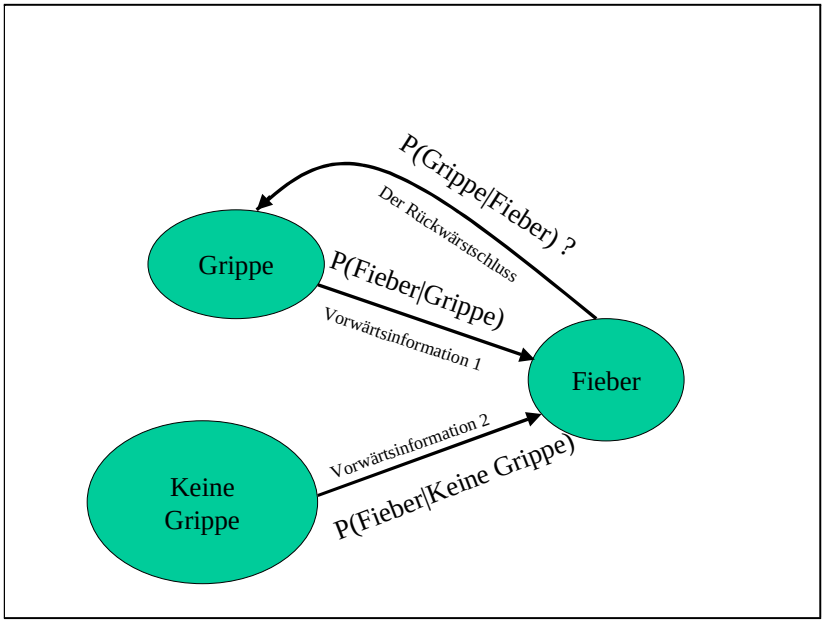
Der Arzt hat hier mehrere Wahrscheinlichkeiten zu bedenken:

1. Wenn eine Grippeerkrankung vorliegt, ist diese oft, aber nicht immer, mit Fieber verbunden. Nehmen wir an, in 80 % der Fälle. Dann ist die bedingte Wahrscheinlichkeit für Fieber, "gegeben Grippe", = 0.8:
   $P(Fieber \mid Grippe) = 0.8$ (Vorwärtsinformation 1)

2. Das Fieber kann auch andere Ursachen haben. Nehmen wir an, dass von den Patienten der Sprechstunde, die *keine* Grippe haben, 10 % mit Fieber erscheinen. Dann ist die bedingte Wahrscheinlichkeit für Fieber, "gegeben *keine* Grippe", = 0.1:
   $P(Fieber \mid keine \quad Grippe) = 0.1$ (Vorwärtsinformation 2)

3. Vielleicht gibt es gerade eine Grippe-Epidemie. Dann ist die Wahrscheinlichkeit ohnehin erhöht, dass es sich um einen Grippepatienten handelt (ob Fieber oder nicht). Nehmen wir an, 40 % der Sprechstundenpatienten haben Grippe:
   $P(\ Grippe) = 0.4$ (A-priori- Wahrscheinlichkeit)

Jetzt ist der Arzt bereit für den Salto rückwärts: die Berechnung der Wahrscheinlichkeit, dass ein Grippevirus vorliegt, wenn der Patient Fieber hat:



$$
\begin{aligned}
P(Grippe|\ Fieber) &= \frac{P(Grippe \cap Fieber)}{P(Fieber)} \times \frac{P(Grippe)}{P(Grippe)} \quad \text{(die Nenner vertauschen:)} \\
&= \frac{P(Grippe \cap Fieber)}{P(Grippe)} \times \frac{P(Grippe)}{P(Fieber)} \\
&= P(Fieber|Grippe) \times \frac{P(Grippe)}{P(Fieber)}, \quad \text{nach der Formel (3.11) also} \\
P(Grippe|Fieber) &= \frac{P(Fieber|Grippe)\ P(Grippe)}{P(Fieber|Grippe)\ P(Grippe) + P(Fieber|k.Grippe)\ P(k.\ Grippe)}
\end{aligned}
$$

Damit ist es gelungen, die gesuchte Wahrscheinlichkeit auf die "Vorwärtsinformationen" (1 und 2) und die a-priori-Wahrscheinlichkeit zurückzuführen. Ergebnis:

$$P(Grippe|Fieber) = \frac{0.8 \times 0.4}{0.8\ \times 0.4\ + 0.1 \times (1 - 0.4)} = 0.842\,11 = 84.2\ \%$$

Wie ändert sich diese Wahrscheinlichkeit in "normalen Zeiten", wenn also keine Grippe-Epidemie herrscht? Angenommen dann haben nur 5 % der Sprechstundenpatienten Grippe. Dann erhält man:

$$P(Grippe|Fieber) = \frac{0.8 \times 0.05}{0.8\ \times 0.05\ + 0.1 \times (1 - 0.05)} = 0.296\,30 = 29.6\ \%$$

die Wahrscheinlichkeit wird also *ganz erheblich* kleiner!
Und wie wäre es, wenn das Fieber in Wirklichkeit ein noch *spezifischeres* Zeichen für Grippe wäre, wenn also Fieber nur bei 5 % (statt 10 %) der Patienten *ohne* Grippe aufträte? Wieder bei einer allgemeinen Grippehäufigkeit von 40 % gerechnet wäre das:

$$P(Grippe|Fieber) = \frac{0.8 \times 0.4}{0.8\ \times 0.4\ + 0.05 \times (1 - 0.4)} =: 0.914\,29 = 91.4\ \%$$

d.h. die Wahrscheinlichkeit wäre noch deutlich erhöht.

Weitere Übung: Wie ändert sich die Wahrscheinlichkeit, wenn die Grippe sogar mit 90% statt mit 80 % Wahrscheinlichkeit mit Fieber verbunden ist? (Lösung: nicht mehr viel; sie wächst von 91.4 % auf 92.3 %.)

Die hier benutzte Formel für den Rückwärtsschluss ist als **Satz von BAYES** bekannt. Die allgemeine Formulierung lautet:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B|A)\ P(A) + P(B|\overline{A})\ P(\overline{A})} \tag{3.12}$$

Es wird deutlich, dass der Satz von BAYES in der Diagnostik eine wichtige Rolle spielt. Er setzt voraus, dass die a-priori-Wahrscheinlichkeit oder auch die "*Prävalenz*" $P(A)$ der Krankheit A in der untersuchten Population bekannt ist und dass man die "Vorwärts"-Wahrscheinlichkeiten, mit denen verschiedene Zustände ("A"bzw. "Nicht–A") das Symptom B "hervorrufen", kennt.

3.8. Gütekriterien eines diagnostischen Tests: Sensitivität, Spezifität und prädiktive Werte

Betrachtet man in diesen Überlegungen nicht allgemein ein "Symptom B", sondern die Reaktion eines gezielt eingesetzten klinischen Tests, welcher auf die Krankheit A reagieren soll, so kann man mit der gleichen Formel folgende Frage beantworten:

> *Angenommen, der klinische Test ist positiv; mit welcher Wahrscheinlichkeit liegt die Krankheit A wirklich vor?*

Man nennt diese bedingte Wahrscheinlichkeit die "*positive Prädiktheit*" des Tests. Bezeichnet man allgemein mit "+" und "−" die positive und die negative Reaktion des Tests bzw. das Vorhandensein oder Nichtvorhandensein der zu untersuchenden Krankheit, so definiert man:

$$\begin{aligned}
\textit{Prävalenz} \text{ der Krankheit} &= P(\text{Krankheit } +) && (3.13)\\
\textit{Sensitivität} \text{ des Tests} &= P(\text{Test } + \mid \text{Krankheit } +) && (3.14)\\
\textit{Spezifität} \text{ des Tests} &= P(\text{Test } - \mid \text{Krankheit } -) && (3.15)\\
\textit{Positive Prädiktheit} \text{ des Tests} &= P(\text{Krankheit } + \mid \text{Test } +) && (3.16)\\
\textit{Negative Prädiktheit} \text{ des Tests} &= P(\text{Krankheit } - \mid \text{Test } -) && (3.17)
\end{aligned}$$

Die *Sensitivität* (die Fähigkeit, eine vorhandene Krankheit anzuzeigen) und die *Spezifität* (die Fähigkeit, "spezifisch" nur dann zu reagieren, wenn die Krankheit vorhanden ist –falls überhaupt–; genauer: *nicht* zu reagieren, wenn die Krankheit *nicht* vorliegt) sind Eigenschaften, die die Qualität eines Tests in "reinen" Populationen (Populationen aus entweder nur kranken oder nur gesunden Probanden) kennzeichnen. Positive und negative *Prädiktheit (*auch positiver und negativer *prädiktiver Wert"* genannt) kennzeichnen dagegen die Eigenschaften eines Tests in *gemischten* Populationen. Sie sind dementsprechend von dem Anteil Kranker in der Population, der "*Prävalenz*" abhängig. Nach der BAYES–Formel gilt nämlich:

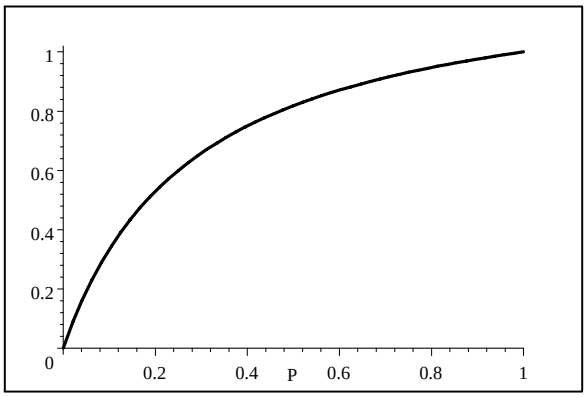
$$
\begin{aligned}
& \textit{Positiver Prädiktiver Wert} \text{ des Tests} \\
= \; & P(\text{Krankh. } + \mid \text{Test } +) \\
= \; & \frac{\text{Sensitivität } \times \text{Prävalenz}}{\text{Sensitivität } \times \text{Prävalenz} + (1 - \text{Spezifität})(1 - \text{Prävalenz})} \qquad (3.18)
\end{aligned}
$$

Analog ist der "*Negative Prädiktive Wert*" eines Test definiert durch die Wahrscheinlichkeit, dass die Krankheit t*atsächlich nicht vorliegt*, wenn der *Test negativ* ist. Er lässt sich wie folgt berechnen:

$$
\begin{aligned}
& \textit{Negativer Prädiktiver Wert} \text{ des Tests} \\
= \; & P(\text{Krankh. } - \mid \text{Test } -) \\
= \; & \frac{\text{Spezifität } \times (1 - \text{Prävalenz})}{\text{Spezifität } \times (1 - \text{Prävalenz}) \; + (1 - \text{Sensitivität}) \times \text{Prävalenz}} \qquad (3.19)
\end{aligned}
$$

Beispiel:
Ein Srceening–Test habe eine Sensitivität von 0.9 und eine Spezifität von 0.8 . Der positive prädiktive Wert ist dann in Abhängigkeit von der Prävalenz P :

$$
\text{Positiver Prädiktiver Wert } = \frac{0.9 \times P}{0.9 \times P + (1 - 0.8)(1 - P)}
$$



Positiver Prädiktiver Wert in Abhängigkeit von der Prävalenz P

Insbesondere sieht man: Auch bei hoher Sensititivät und Spezifität kann der prädiktive Wert **sehr klein** (fast gleich 0) sein, wenn die **Prävalenz der Krankheit nur genügend niedrig ist.** Dieses gilt im Prinzip für jeden Test (es sei denn die Spezifität ist gleich 1).

Woher bekommt man eigentlich die "Vorwärtsinformation", bei diagnostischen Tests also die Werte für Sensitivität und Spezifität?
Man untersuche

- eine Gruppe von Patienten, bei denen die untersuchte Krankheit *sicher* vorhanden ist,

- eine zweite Gruppe von Patienten, bei denen die Krankheit *sicher nicht* vorhanden ist,

und wende in allen Fällen das diagnostische Testverfahren an.
Dazu ist es nötig, dass es ein Verfahren gibt, das die tatsächliche Situation sicher erkennen kann. Man spricht dann häufig von dem "*Goldstandard*". Es handelt sich dabei in der Regel um ein *aufwendigeres*, *belastenderes* oder *nicht immer durchführbares* Verfahren.
Die Ergebnisse der Untersuchung können dann in Form einer Vierfeldertafel wie in der folgenden Abbildung notiert werden.

		Tatsächliche Situation		
Die Diagnose lautet		krank (positiv)	gesund (negativ)	
	krank (positiv)	richtige Entscheidung a	falsch positiv b	Positiver Vorhersagewert a/(a+b)
	gesund (negativ)	falsch negativ c	richtige Entscheidung d	negativer Vorhersagewert d/(d+s)
		Sensitivität a/(a+c)	Spezifität d/(b+d)	

Sensitivität, Spezifität und Prädiktive Werte

Man beachte aber unbedingt:
Die in der Tabelle angegebenen Formeln für den positiven und den negativen prädiktiven Wert (rechte Spalte) **gelten nur, wenn die Anteile der Kranken und der Gesunden der Tabelle denen der Gesamtpopulation entsprechen**. Sind dagegen z.B. die Kranken im Vergleich zur Gesamtpopulation überrepräsentiert, so liefert die angegebenen Formel einen *zu hohen positiven* und einen *zu kleinen negativen prädiktiven Wert.* Zur Berechnung dieser Kenngrößen muss man dann auf die Formeln (3.18) und (3.19) zurückgreifen.

3.9. Haben Wahrscheinlichkeitsverteilungen einen Mittelwert?

Bisher haben wir im Rahmen der Wahrscheinlichkeitsrechnung immer nur von Wahrscheinlichkeiten gesprochen, und zwar jeweils in Analogie zu den *relativen Häufigkeiten* in der deskriptiven Statistik. Von dort kennen wir aber noch anderes: Mittelwert, Streuung, Median und weitere Kenngrößen. Die waren jeweils über die *Werte der Beobachtungen einer Stichprobe* definiert. Aber in der Wahrscheinlichkeitsrechnung gibt es keine Beobachtungen, nur Regeln, Formeln und Gleichungen. Wie kommt man nun zu Definitionen, die den Kenngrößen der deskriptiven Statistik entsprechen? Auch das wird nun in wenigen Schritten über die Analogie von relativen Häufigkeiten und Wahrscheinlichkeiten erreicht:

1. Die *"Ereignisse"* (z.B. "Therapieerfolg") werden als die *möglichen Werte einer Variablen"* $(X, Y, \ldots)$ betrachtet (z.B. der Variablen: "Ergebnis der Therapie"). Das Ergebnis des Experimentes ist (auch) von *Zufälligkeiten* mit beeinflußt. Die Variable wird daher *"Zufallsvariable"* genannt.
   Ist z.B. "10" ein möglicher Wert eines Experimentes, und kennzeichnet die Zufallsvariable Y das Ergebnis eines Experimentes, so wird mit

$$A: \ (Y = 10)$$

   das "Ereignis" bezeichnet, dass das Ergebnis des Experimentes gleich 10 ist.
   Ist Y die "Anzahl Erfolge bei 3 Versuchen" so bedeutet

$$(Y \geq 1) : \text{"Mindestens 1 Erfolg in 3 Versuchen"}$$

2. Die "möglichen Werte" werden bei der Durchführung eines Experimentes mit unterschiedlichen Wahrscheinlichkeiten angenommen.
   Sind $w_1, w_2, w_3, \ldots, w_K$ die *möglichen Werte* einer Zufallsvariablen Y, so bezeichne

$$p_i := \ P(Y = w_i)$$

   die Wahrscheinlichkeit dafür, dass Y den Wert w_i annimmt.
   Unter der *Verteilung einer Zufallsvariablen* Y versteht man die Gesamtheit der Wahrscheinlichkeiten, mit denen diese Zufallsvariable bestimmte Einzelwerte oder Wertebereiche annimmt.

3. Der *Erwartungswert* μ der Verteilung von Y ist definiert durch:

$$\begin{aligned} \mu &= E(Y) = \sum_{i=1}^{K} w_i P(Y = w_i) \\ &= \sum_{i=1}^{K} w_i p_i \end{aligned} \tag{3.20}$$

   Er stellt also eine *gewichtete Summe der möglichen Werte* einer Zufallsvariablen dar, wobei als *Gewicht* eines Wertes w seine Wahrscheinlichkeit $P(Y = w)$ gewählt wird. Dieses entspricht der Definition des *Mittelwertes* in der deskriptiven Statistik: Sind z.B. $1, 2$ und 3 die möglichen Werte eines Merkmales X, und liegen die (geordneten) Beobachtungen $1, 1, 2, 2, 2, 2, 2, 3, 3, 3$ vor, so ist

$$\begin{aligned} \overline{x} &= \frac{1}{10}(\underbrace{1+1}_{2 \text{ mal}} + \underbrace{2+2+2+2+2}_{5 \text{ mal}} + \underbrace{3+3+3}_{3 \text{ mal}}) \\ &= 1 \times \frac{2}{10} \ + \ 2 \times \frac{5}{10} \ + \ 3 \times \frac{3}{10} \\ &= 1 \times h(\text{"2"}) \ + \ 2 \times h(\text{"5"}) \ + \ 3 \times h(\text{"3"}) \end{aligned}$$

   d.h. die möglichen Werte von X werden mit ihren *relativen Häufigkeiten* gewichtet und dann addiert. Hier kommt wieder das Prinzip zur Anwendung, wonach eine Analogie zwischen deskriptiver Statistik und Wahrscheinlichkeitsrechnung über die enge Beziehung zwischen relativer Häufigkeit und Wahrscheinlichkeit hergestellt wird.

4. Analog ist die *Varianz* als die zu erwartende quadratische Abweichung vom Erwartungswert definiert:

$$\begin{aligned} \sigma^2 &= var(Y) = \sum_{i=1}^{K} (w_i - \mu)^2 \; P(Y = w_i) \\ &= \sum_{i=1}^{K} (w_i - \mu)^2 p_i \end{aligned} \tag{3.21}$$

Fertig!

3.10. Große Fallzahl und die Folgen: Verteilungen von großen Summen

Eigentlich will man z.B. ja nur eines wissen: Wie groß ist die Erfolgswahrscheinlichkeit einer Therapie? Eine Einzelbehandlung kann darüber nicht viel aussagen, denn in das Einzelergebnis gehen auch viele unkontrollierbare Effekte, "Zufallseffekte" ein. Also macht man eine *Serie* von Einzelversuchen und zählt die *Erfolge in dieser Serie*. Aber diese *Anzahl der Erfolge in der Serie ist ja ebenfalls (auch) vom Zufall abhängig*. Um herauszufinden, welche Ergebnisse mit welchen Wahrscheinlichkeiten möglich sind, muss man sich also auch mit den *Wahrscheinlichkeitsverteilungen von Gesamtergebnissen aus Versuchsserien* beschäftigen. Jetzt kommt die Wahrscheinlichkeitsrechnung erst richtig zum Zug!

3.10.1. Beispiel: Die Binomialverteilung

Ein "Gesamt"–Experiment bestehe darin, dass ein bestimmtes Einzelexperiment ohne Veränderung der Versuchsanordnung oder der Rahmenbedingungen n mal durchgeführt wird, wobei keines der Einzelexperimente irgendein anderes Einzelexperiment beeinflußt. Jedes Einzelexperiment endet mit dem Ergebnis "Erfolg" oder "kein Erfolg". Gesucht ist die Wahrscheinlichkeit für genau k Erfolge nach n Versuchen.
Bezeichnungen:

$$\begin{aligned} Y_i &= \text{Zufallsvariable zur Kennzeichnung des Ergebnisses von Versuch Nr. } i \\ &\quad (i = 1, 2, \ldots, n) \\ \{0, 1\} &= \text{Menge der möglichen Ausgänge eines Versuches, mit den Interpretationen} \\ &\quad 0 \mathrel{\widehat{=}} \text{"Kein Erfolg"} \\ &\quad 1 \mathrel{\widehat{=}} \text{"Erfolg"} \\ p &= P(Y_i = 1) \text{ die Wahrscheinlichkeit für Erfolg im Einzelversuch } (i = 1, 2, \ldots, n) \\ T_n &= \sum_{i=1}^{n} Y_i \text{ die Anzahl der Erfolge nach } n \text{ Versuchen} \end{aligned} \tag{3.22}$$

Die hier beschriebenen *Einzel*experimente werden "*Bernoulli–Experimente*" genannt, und die zugehörige Verteilung

$$\{P(X = 0) = 1 - p, \qquad P(X = 1) = p\} \tag{3.23}$$

die "*Bernoulli–Verteilung*".
In der folgenden Tabelle wird für $n = 1, 2, 3$ zunächst jedes Ereignis $(T_n = k) \;\; (k \leq n)$ in seine "Elementarereignisse" zerlegt.

Beispiel (Zeilen 4 und 5): "Genau $k = 1$ Erfolg in $n = 2$ Versuchen" setzt sich zusammen aus den Serien $(0, 1)$ und $(1, 0)$:

$$(T_2 = 1) = (Y_1 = 0, Y_2 = 1) \cup (Y_1 = 1, Y_2 = 0)$$

Für diese Elementarereignisse wird in der folgenden Spalte jeweils nach der Regel (3.10) über die Unabhängigkeit von Ereignissen die Wahrscheinlichkeit berechnet.
Z.B. ist

$$\begin{aligned} P(Y_1 = 0, Y_2 = 1) &= P[(Y_1 = 0) \cap (Y_2 = 1)] \\ &= P(Y_1 = 0)\ P(Y_2 = 1) \\ &= (1 - p)\ p \end{aligned}$$

In der folgenden Spalte werden diese Wahrscheinlichkeiten dann addiert (z.B.ist $(1 - p)p + p(1 - p) = 2p(1 - p)$).

Ereignisse und Wahrscheinlichkeiten: Genau k Erfolge in n Versuchen:

n	k	Ereignis	Wahrsch.	Wahrsch. insgesamt
1	0	0	$1 - p$	$1(1 - p)$
	1	1	p	$1p$
2	0	0 0	$(1 - p)(1 - p)$	$1(1 - p)^2$
	1	0 1	$(1 - p)p$	$2p(1 - p)$
		1 0	$p(1 - p)$	
	2	1 1	pp	$1p^2$
3	0	0 0 0	$(1 - p)(1 - p)(1 - p)$	$1(1 - p)^3$
	1	0 0 1	$(1 - p)(1 - p)p$	
		0 1 0	$(1 - p)p(1 - p)$	$3p(1 - p)^2$
		1 0 0	$p(1 - p)(1 - p)$	
	2	0 1 1	$(1 - p)pp$	
		1 0 1	$p(1 - p)p$	$3p^2(1 - p)$
		1 1 0	$pp(1 - p)$	
	3	1 1 1	ppp	$1p^3$

Nach diesem Schema kann man die allgemeine Formel zur Berechnung von $P(T_n = k)$ herleiten. Das Ergebnis ist .

$$P(T_n = k) = \frac{n!}{k!(n - k)!}\ p^k\ (1 - p)^{n-k} \tag{3.24}$$

Darin ist die Erfolgswahrwcheinlichkeit p für den Einzelversuch noch offengelassen und kann durch irgendeinen Wert zwischen 0 und 1 ersetzt werden. In Zähler und Nenner des Bruches stehen die "Fakultäten" von natürlichen Zahlen. Z.B. ist 4! (sprich "Vier Fakultät") gleich $1 \times 2 \times 3 \times 4 = 24$ (und es ist $0! = 1$). Beispiel:

Wahrscheinlichkeit für genau 1 Erfolg in 4 Versuchen,

wenn die Erfolgswahrscheinlichkeit im Einzelversuch p gleich 0.1 ist,

$$\begin{aligned} &= \frac{4!}{1!(4 - 1)!}\ 0.1^1\ (1 - 0.1)^{4-1} \\ &= \frac{1 \times 2 \times 3 \times 4}{1 \times (1 \times 2 \times 3)}\ 0.1^1\ (0.9)^{4-1} \\ &= \frac{24}{6}\ 0.1^1\ (0.9)^3 \\ &= .2916 \end{aligned}$$

Diese Verteilung für die Anzahl der Erfolge aus n unabhängigen Experimenten mit jeweils gleicher Erfolgswahrscheinlichkeit p heißt die *Binomialverteilung B(n,p).*
Es folgen einige Beispiele für verschiedene Kombinationen von n und p.

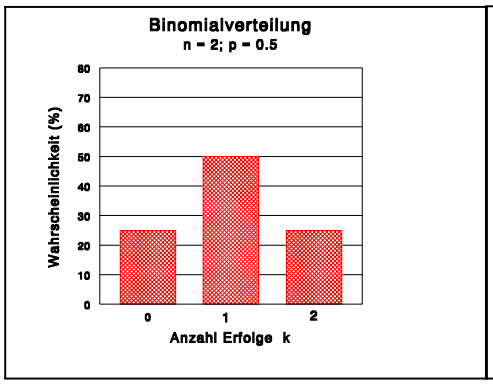


Abb. 3.2: B(2, 0.5)

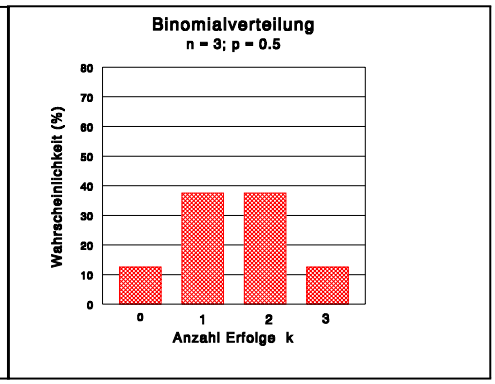


Abb. 3.3: B(3, 0.5)

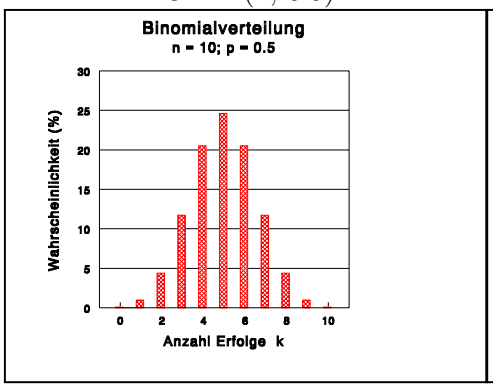


Abb. 3.4: B(10, 0.5)

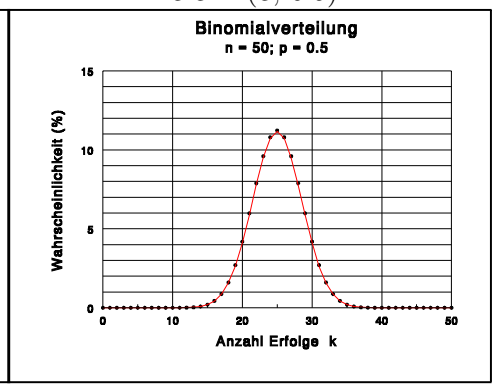


Abb. 3.5: B(50, 0.5)

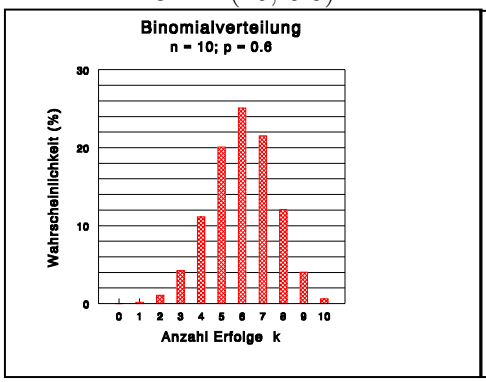


Abb. 3.6: B(10, 0.6)

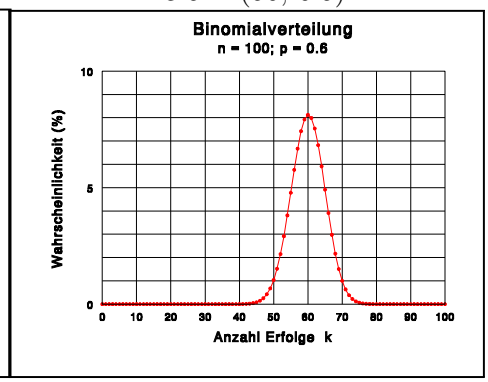


Abb. 3.7: B(100, 0.6)

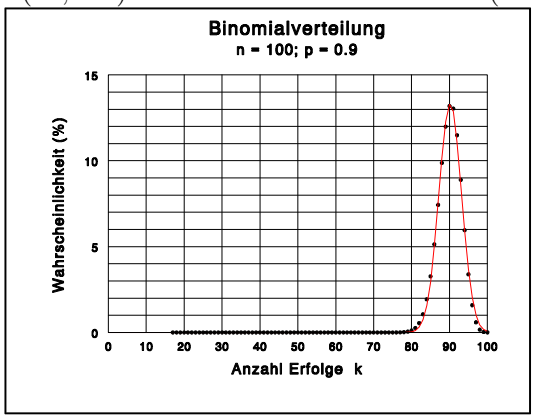


Abb. 3.8: B(100, 0.9)

3.11. Die Normalverteilung

3.11.1. Der zentrale Grenzwertsatz

Aus den Abbildungen 3.2 - 3.8 zur Binomialverteilung folgt:

1. Die Wahrscheinlichkeitsverteilung für die "Anzahl T_n der Erfolge in n Versuchen" kann für wachsendes n immer genauer durch eine stetige, symmetrische Kurve $f_n(x)$ beschrieben werden.

2. Die *Gestalt* dieser Kurve ("glockenförmig") ist unabhängig davon, ob als Erfolgswahrscheinlichkeit für den Einzelversuch gleich $p = 0.5$, 0.6 oder 0.9 gewählt wird.

3. Bildet man zwei Grenzen a und b (in gewissem Mindestabstand voneinander), so ist die Fläche unter der Kurve $f_n(x)$ zwischen diesen beiden Grenzen etwa gleich der Wahrscheinlichkeit dafür, dass die Anzahl der Erfolge T_n zwischen diesen Grenzen liegt (siehe Abb.3.9: jeder Wert $k = 1, 2, \ldots, n = 50$ auf der x-Achse erzeugt eine Säule der Breite 1 und der Höhe $P(T_n = k) \approx f_n(k)$ und damit auch der Fläche $P(T_n = k) \approx f_n(k)$. Die Summe der Einzelwahrscheinlichkeiten $P(T_n = k)$ entspricht daher der Summierung der Flächen der einzelnen Säulen und damit der Bildung der Gesamtfläche zwischen den Grenzen.) Die Fläche unter der "gesamten" Kurve hat daher den Wert 1.

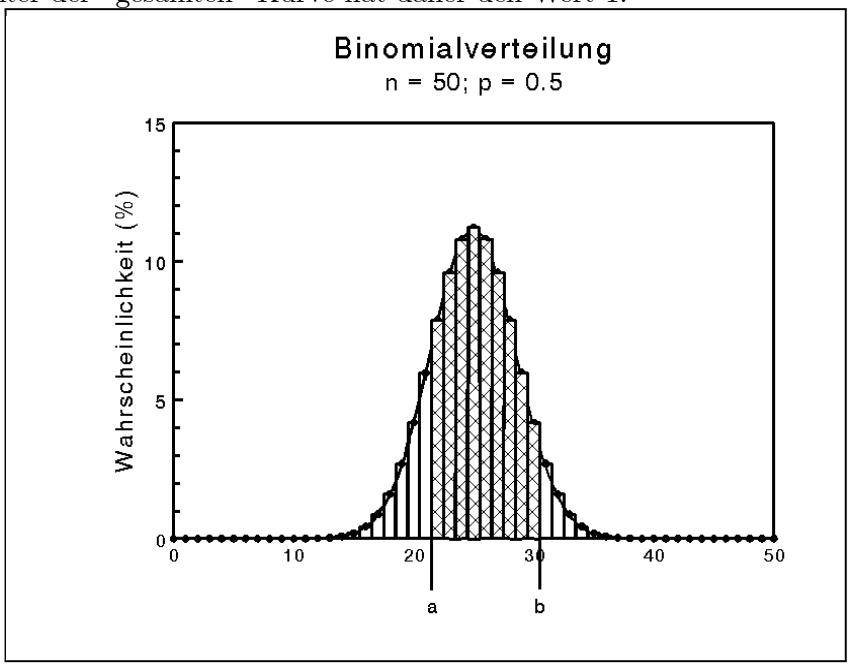


Abb. 3.9: Die Wahrscheinlichkeiten von Intervallen

Mit aufwendigeren mathematischen Methoden kann gezeigt werden, dass die "Grenzfunktion" $f_n(x)$ die folgende mathematische Gestalt hat:

$$f_n(x) = \frac{1}{\sqrt{2\pi}\ \sigma_n} \exp\left(-\frac{(x-\mu_n)^2}{2\sigma_n^2}\right) \tag{3.25}$$

Darin ist

$$\begin{aligned} \mu_n &= np \text{ der Erwartungswert und} \\ \sigma_n^2 &= np(1-p) \text{ die Varianz} \end{aligned}$$

von T_n.
Definition:

1. Die Funktion
$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3.26}$$
ist die *Dichtefunktion* der *Normalverteilung* mit Erwartungswert μ und Varianz σ^2. Das bedeutet:
Eine Zufallsvariable X ist *normalverteilt mit Erwartungswert μ und Varianz σ^2*, wenn die Wahrscheinlichkeit für Intervalle $[a, b]$ gleich der Fläche unter der Kurve $\varphi_{\mu,\sigma}(x)$ zwischen a und b ist:
$$P(X \geq a \text{ und } X \leq b) = \int_a^b \varphi_{\mu,\sigma}(x)dx$$
Bezeichnung:
$$X \sim N(\mu, \sigma^2)$$

2. Durch die sogenannte "*Z–Transformation*" erhält man eine Normalverteilung mit Erwartungswert 0 und Varianz 1:
Man subtrahiert den Erwartungswert (dadurch wird die Verteilung auf den Wert 0 zentriert) und dividiert durch die Standardabweichung (dadurch wird die Verteilung auf die Varianz 1 normiert). Ist X normalverteilt mit Erwartungswert μ und Varianz σ^2 , so ist
$$Z = \frac{X-\mu}{\sigma}$$
"*standard–normalverteilt*":
$$Z \sim N(0, 1)$$
Die Dichtefunktion der Standardnormalverteilung ist nach (3.26):
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) : \tag{3.27}$$

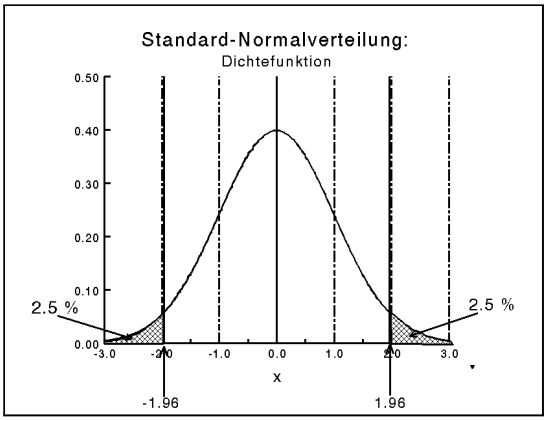


Abb. 3.10: Dichtefunktion der Standard–Normalverteilung

3. Die oben für die Binomialverteilung dargestellte "Annäherung" an die Normalverteilung gilt allgemeiner ("**Zentraler Grenzwertsatz**"):
Sind $Y_1, Y_2, \ldots, Y_n$ unabhängige Zufallsvariable, die alle dieselbe Verteilung mit

dem Erwartungswert μ und der Varianz σ^2 haben (die ansonsten völlig beliebig sein kann!), so ist der Mittelwert daraus (nach Z–Transformation) für große n angenähert standardnormalverteilt:

$$Z = \frac{\bar{Y}_n - \mu}{\frac{1}{\sqrt{n}}\sigma} \longrightarrow N(0,1) \tag{3.28}$$

(Hier geht die Information aus der Wahrscheinlichkeitsrechnung ein, dass die Varianz des Mittelwertes $\bar{Y}_n$ gleich $\frac{1}{n}\sigma^2$ ist, siehe Kapitel 4!)
Die große Bedeutung der Normalverteilung für die Statistik folgt u.a. aus diesem Grenzwertsatz: Zum einen bildet man in den Anwendungen häufig längere "Versuchsserien" von unabhängigen Wiederholungen eines Einzelexperimentes; zum anderen kann das Ergebnis eines Einzelexperimentes selbst das Resultat vieler kleiner "Effekte" sein, die sich addieren (man denke etwa an die Brownsche Molekularbewegung). Im zweiten Fall ist das Einzelexperiment selber schon (und nicht erst der Mittelwert aus vielen Wiederholungen) –angenähert– normalverteilt.
Einen ersten Hinweis dazu, ob eine Zufallsvariable normalverteilt ist, kann man dem Histogramm entnehmen (siehe Abb. 2.6): Dort waren die einzelnen Säulen so gewählt, dass ihre *Flächen gleich den Wahrscheinlichkeiten* für die zugrundeliegenden Intervalle sind. Faßt man die obere Begrenzung der Säulen eines Histogramms insgesamt als "Kurve" auf, so müßte also bei einer normalverteilten Zufallsvariable das Histogramm in diesem Sinne angenähert mit der Dichtefunktion einer Normalverteilung übereinstimmen. Um einen direkten Vergleich zu erhalten, legt man häufig Histogramm und Dichtefunktion der Normalverteilung übereinander, wobei in der Funktionsgleichung (3.26) die unbekannten "Parameter", Erwartungswert μ und Varianz σ^2, durch die Stichprobenwerte $\overline{x}$ und s_x^2 ersetzt werden. Beispiele:

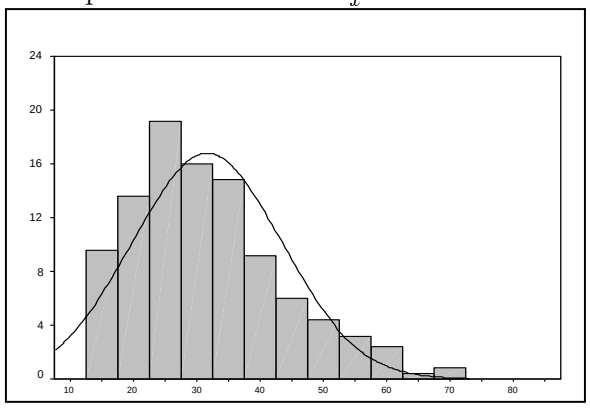


Abb. 3.11: Rechtsschiefe Verteilung

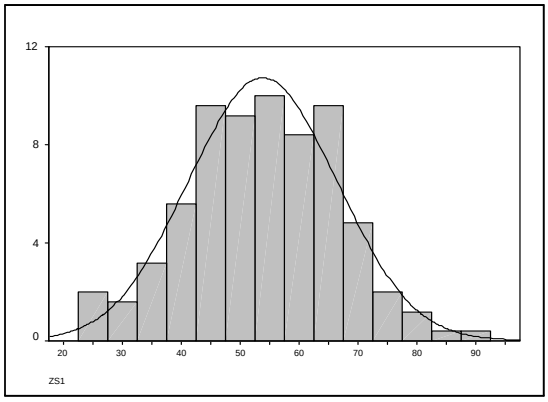


Abb. 3.12: "Ungefähr normalverteilt"

Während im linken Beispiel deutliche Abweichungen von der Normalverteilung erkennbar sind, wird man im rechten Bild die Unterschiede zwischen Histogramm und Normalverteilungskurve auf die Ungenauigkeiten und Schwankungen der Stichprobe zurückführen (ausgenommen eventuell am linken Rand der Kurve?).

3.11.2. Die kumulative Verteilungsfunktion

Die Normalverteilung ist durch ihre Dichtefunktion (3.26) vollständig definiert. Aus dieser Funktion sind aber irgendwelche Wahrscheinlichkeiten nicht unmittelbar ablesbar, vielmehr muß man dazu die "Flächen unter der Kurve" in den gewünschten Intervallen bilden (d.h.: das Integral der Kurve). Einen wichtigen Spezialfall hierzu bildet

die "Fläche bis zur Stelle y". Ist allgemein Y eine Zufallsvariable mit der Dichtefunktion $f(y)$, so ist :

$$\begin{aligned} F(y): &= P(Y \leq y) \\ &= \int_{-\infty}^{y} f(u)du \end{aligned} \tag{3.29}$$

die Wahrscheinlichkeit dafür, dass Y höchstens den Wert y annimmt. $F(y)$ wird *die kumulative Verteilungsfunktion von* Y genannt (siehe Abb. 3.13).

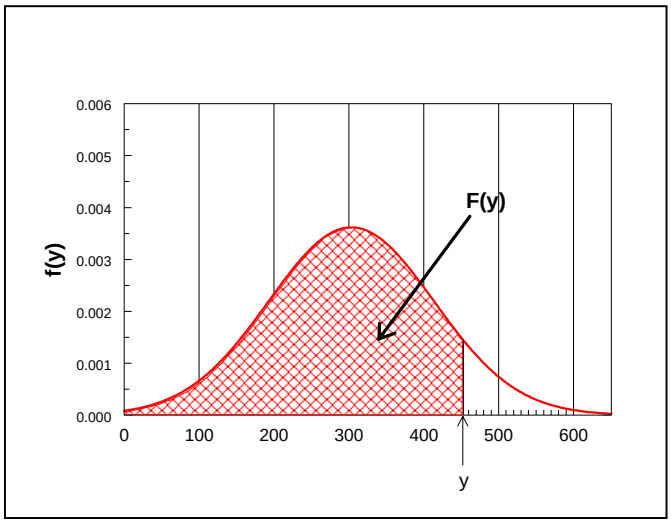


Abb.3.13: Dichtefunktion f und kumulative Verteilungsfunktion F

Diese Definition entspricht derjenigen aus der deskriptiven Statistik. Zur besseren Unterscheidung spricht man dort auch von der "*empirischen*" kumulativen Verteilungsfunktion.
Speziell erhält die Standardnormalverteilung die Bezeichnung $\Phi(y)$:

$$\Phi(y) = \int_{-\infty}^{y} \varphi(u)du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{u^2}{2}\right) du \tag{3.30}$$

Aus der Funktion $F(y)$ kann man leicht auch weitere Wahrscheinlichkeiten ablesen. z.B. ist

$$\begin{aligned} P(a < Y \leq b) &= P(Y \leq b) - P(Y \leq a) \\ &= F(b) - F(a) \qquad \text{für } a \leq b, \quad \text{und} \end{aligned}$$

$$\begin{aligned} P(Y > a) &= 1 - P(Y \leq a) \\ &= 1 - F(a) \end{aligned}$$

Die in der deskriptiven Statistik gebildeten Definitionen für Kenngrößen einer Verteilung, die aus der *empirischen* kumulativen Verteilungsfunktion ablesbar sind, werden nun -bis auf Minimum, Maximum und Range– unverändert auf die kumulative Verteilungsfunktion der *Wahrscheinlichkeitsverteilung* übetragen. Speziell erhält man für die Standard–Normalverteilung die folgenden numerischen Werte (siehe auch Abb. 3.14 und 3.15):

	$N(0,1)$	$N(\mu,\sigma^2)$
2.5%–Quantil	-1.960	$\mu - 1.960\ \sigma$
5%–Quantil	-1.645	$\mu - 1.645\ \sigma$
Median	0	μ
95%–Quantil	1.645	$\mu + 1.645\ \sigma$
97.5%–Quantil	1.960	$\mu + 1.960\ \sigma$
16%–Quantil	-1	$\mu - \sigma$
84%–Quantil	1	$\mu + \sigma$

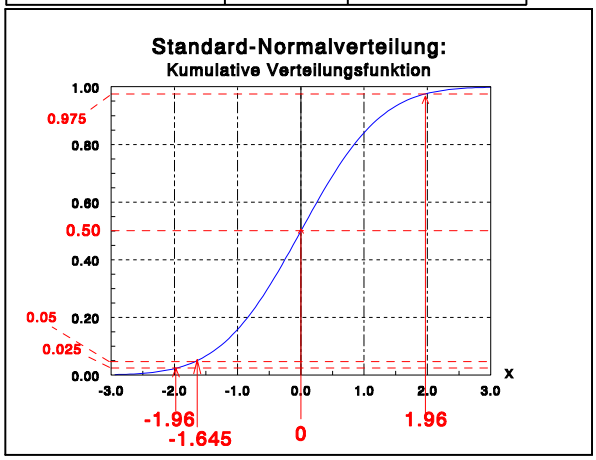


Abb. 3.14: Standardnormalverteilung $N(0,1)$

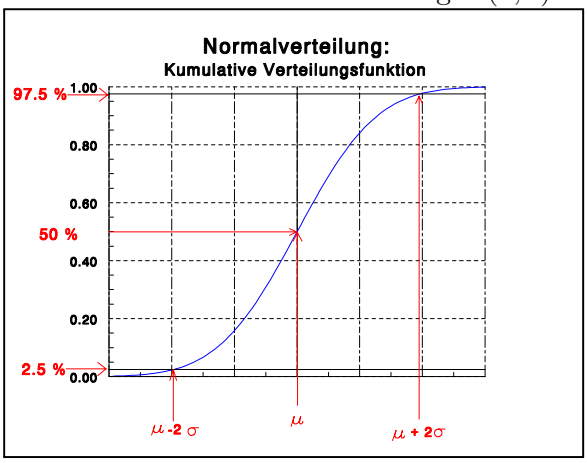


Abb. 3.15: Normalverteilung $\mathrm{N}(\mu,\sigma^2)$

Aus den letzten beiden Zeilen der Tabelle zu den Normalverteilungsquantilen folgt noch, dass zwischen $\mu - \sigma$ und $\mu + \sigma$ (bzw. bei der Standardnormalverteilung zwischen -1 und $+1$) $84\% - 16\% = 68\%$ der Wahrscheinlichkeit liegt. Weiterhin folgt: Will man einen Bereich festlegen, der symmetrisch zum Wert 0 ist und der mit genau 95% Wahrscheinlichkeit angenommen wird, so muß man das Intervall von $\mu - 1.960\ \sigma$ bis $\mu + 1.960\ \sigma$ wählen. In guter Näherung wird daher oft der Bereich $\mu \pm 2\ \sigma$ als "*Normbereich*" eines Merkmals festgelegt.
Die wesentlichen Eigenschaften der Normalverteilung können wie folgt zusammengefasst werden:

1. Die Normalverteilung ist eine stetige symmetrische Verteilung. Das Maximum der Verteilungsdichte liegt beim Erwartungswert μ

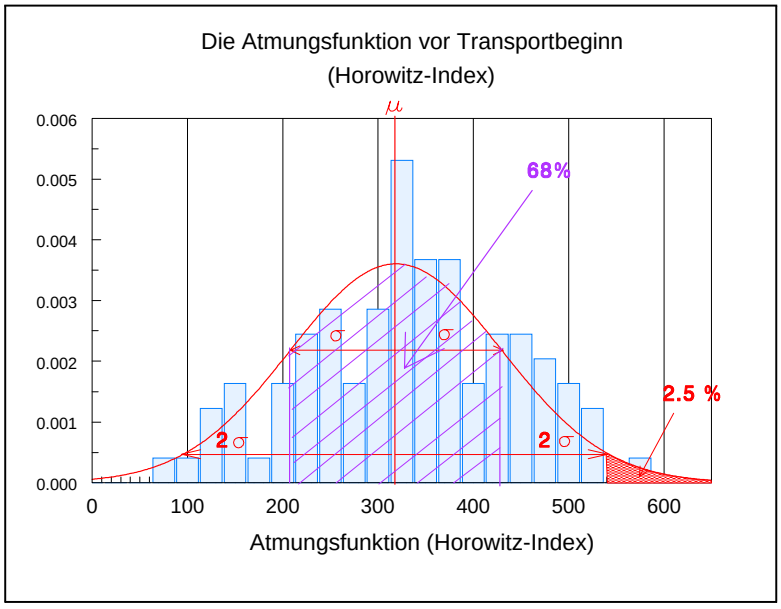
2. Die Fläche, die von der Dichtefunktion der Normalverteilung zwischen $x_1 = -k\sigma$

und $x_2 = \mu - k\sigma$ ($k > 0$ beliebig) eingeschlossen wird, ist für alle Normalverteilungen gleich groß. Zur Verdeutlichung die folgende Tabelle

k	Intervallgrenzen	Anteil der Teilfläche
1	x = μ ± 1*σ	68,27 %
2	x = μ ± 2*σ	95,44 %
3	x = μ ± 3*σ	99,73 %
4	x = μ ± 4*σ	99,99 %
1,64	x = μ ± 1,64*σ	90 %
1,96	x = μ ± 1,96*σ	95 %
2,58	x = μ ± 2,58*σ	99 %
2,81	x = μ ± 2,81*σ	99,5 %

In der Praxis werden die $2\sigma-$, $3\sigma-$ Bereiche häufig zur Berechnung von Normbereichen benutzt.

Im folgenden Beispiel eines angenähert normalverteilten Merkmales sind diese Kenngrößen ablesbar, wobei für das 2.5%– und das 97.5%–Quantil die Näherung $\mu \pm 2\ \sigma$ eingezeichnet ist:



Normalverteilung mit 1- und 2σ - Bereichen

4. Statistisches Schätzen (GK4)

4.1. Einführendes Beispiel:

In vier parallel durchgeführten Kursgruppen zur Biomathematik wurde folgendes Zufallsexperiment durchgeführt:

Alle StudentInnen wurden gebeten, ihren Puls zu messen. Dann sollten sie möglichst lange die Luft anhalten und anschließend erneut den Pulsschlag messen. Die Veränderungen ("Nachher minus Vorher") wurden als Messergebnisse aufgeschrieben.

Man bildet nun -für jede der 4 Kursgruppen getrennt- *nach jedem angegebenen Messwert erneut* den **Mittelwert aus den bis dahin vorliegenden Angaben** über die Pulsänderung. Die Folge dieser Mittelwerte wird in ein Diagramm eingetragen. Man erhält folgendes Bild:

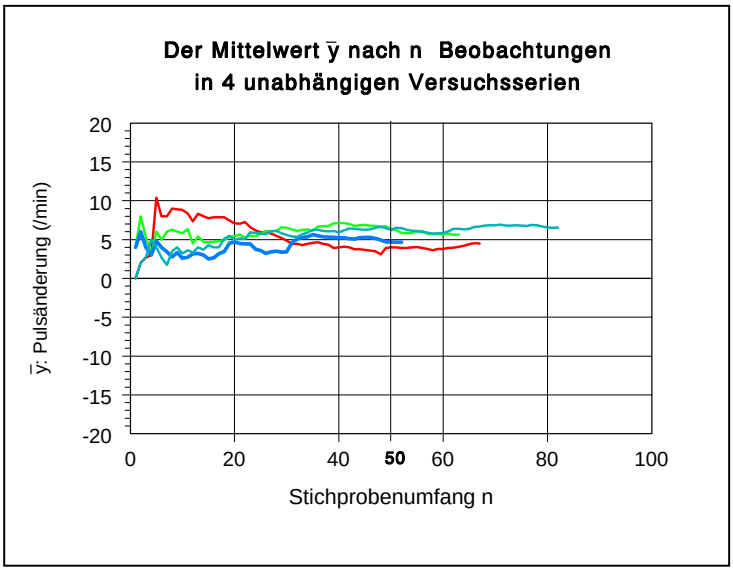


Abb. 4.1: Zufallsschwankungen des Mittelwertes

Man erkennt daraus:

1. *Innerhalb* jeder Versuchsserie zeigen sich *Schwankungen des Mittelwertes.*
2. Die Schwankungen werden *mit zunehmendem Stichprobenumfang kleiner.*
3. Die vier Versuchsserien erscheinen "bis auf zufällige Schwankungen" *ähnlich*, insbesondere:
4. Alle vier Versuchsserien scheinen sich auf *denselben Wert (etwa 5 Schläge/min) hin zu "stabilisieren".*

Diese Beobachtungen entsprechen weitgehend denen aus Kapitel 3.1, in dem die *relativen Häufigkeiten* für "Erfolg" in Abhängigkeit vom Stichprobenumfang betrachtet wurden (siehe auch Abb. 3.1).
Die Abbildung 4.1 macht noch einmal deutlich, dass die *exakten empirischen Ergebnisse einer Untersuchung* eigentlich "nicht so besonders interessant" sein können: Unter denselben Rahmenbedingungen kann nach Abbildung 4.1 der Mittelwert der Pulsänderung nach 50 Versuchen z.B. 4.0, 6.3, 4.7 oder 6.3 Schläge/min betragen.

In diesem Sinne ist jedes numerische Ergebnis eines Zufallsexperiments nur eine **Zufallszahl!**

Interessanter wäre es im vorliegenden Fall vielmehr, *denjenigen Wert zu erfahren, auf den sich alle vier Versuchsserien offenbar "einpendeln".*
Im Rahmen der Wahrscheinlichkeitsrechnung bedeutet dieses:

1. Die Pulsänderung Y ist eine *Zufallsvariable*, deren Wahrscheinlichkeitsverteilung durch die Rahmenbedingungen (Versuchsanordnung und Probandenauswahl) festgelegt ist.

2. Es interessiert der Erwartungswert μ dieser Verteilung.

Diesen Erwartungswert kann man nicht mit Hilfe der Wahrscheinlichkeitsrechnung *berechnen.* (Die Berechnung von Wahrscheinlichkeiten ist immer nur dann möglich, wenn *alle* Basisinformationen über das Experiment vorhanden sind oder entsprechende *Annahmen* gemacht werden wie z.B. beim Münzwurf: Die Münze muss *unverfälscht* sein und die einzelnen Würfe müssen unabhängig voneinander durchgeführt werden. Daher ist man auf die Durchführung von *Experimenten* angewiesen, die unter den genannten Rahmenbedingungen durchgeführt werden. Aus den Daten einer vorliegenden Versuchsserie bildet man dann den *Mittelwert* und interpretiert diesen als eine "*Schätzung*" des Erwartungswertes.
Allgemein gilt nach dem "**Gesetz der großen Zahl**"

> Der *Mittelwert* eines Merkmales Y, gebildet aus n Beobachtungen, strebt mit wachsendem Stichprobenumfang n gegen den *Erwartungswert* der Wahrscheinlichkeitsverteilung von Y.

Dies ist ein Ergebnis der Wahrscheinlichkeitsrechnung. Es macht deutlich: die Erfahrungen aus empirischen Beobachtungen (siehe Abb. 3.1 und 4.1) entsprechen den Ergebnissen der Wahrscheinlichkeitsrechnung. Darüber hinaus liefert das Gesetz der großen Zahl auch eine theoretische Begründung für die Interpretation des *empirischen Mittelwertes als Schätzung für den Erwartungswert.*
Diese Grundgedanken werden in den folgenden Abschnitten noch konkretisiert und weiter ausgebaut.

4.2. Grundgesamtheit und Stichprobe (GK 4.1)

Nach diesen Vorüberlegungen hat der Mittelwert $\overline{y}$ aus den beobachteten Werten y_1, $y_2, \ldots, y_n$ zweierlei Bedeutung:

1. Im Rahmen der deskriptiven Statistik ist $\overline{y}$ ein Verteilungsparameter, der die (mittlere) *Lage der Stichprobenwerte* $y_1, y_2, \ldots, y_n$ beschreibt.

2. Aus Sicht der Wahrscheinlichkeitsrechnung ist $\overline{y}$ ein Wert, von dem man weiß, dass er mit wachsendem n beliebig nahe an den *Erwartungswert einer Wahrscheinlichkeitsverteilung* herankommt.

Voraussetzung hierfür ist aber:

1. Man geht (wie auf Seite 41 beschrieben) von einem wohldefinierten *Zufallsexperiment* aus.

2. Die *möglichen* Ausgänge des Experimentes werden durch die *Wahrscheinlichkeitsverteilung einer Zufallsvariablen* Y beschrieben.

3. Die Beobachtungen $y_1, y_2, \ldots, y_n$ sind die Ergebnisse von n unabhängigen Durchführungen dieses Experimentes. In mathematischer Terminologie:
   Die Beobachtungen $y_1, y_2, \ldots, y_n$ sind *Realisationen von n unabhängigen Zufallsvariablen* $Y_1, Y_2, \ldots, Y_n$, *die alle dieselbe Wahrscheinlichkeitsverteilung (wie* Y*) haben.*

Die Wahrscheinlichkeitsverteilung von Y bzw. das zugrundeliegende Zufallsexperiment mit allen seinen Rahmenbedingungen wird als die "*Grundgesamtheit*" bezeichnet, und die Beobachtungen $y_1, y_2, \ldots, y_n$ bilden eine "*Stichprobe vom Umfang n*".
Der Aussagewert einer statistischen Untersuchung, insbesondere z.B. einer klinischen Studie, hängt wesentlich davon ab, bis zu welchem Grade die genannten Voraussetzungen eines Zufallsexperimentes als gegeben angenommen werden können. Deshalb spielen die *Versuchsplanung* und die *Versuchsüberwachung* eine elementare Rolle in der experimentellen Forschung. (Näheres hierzu im "Ökologischen Kurs".)

4.3. Schätzwerte und ihre Eigenschaften (GK 4.2 und 4.3)

Es hat sich nun herausgestellt (ob man will oder nicht und ob man es weiß oder nicht): In klinischen Untersuchungen ist der Arzt/die Ärztin auf der Suche nach *Kenngrößen ("Parametern") von Wahrscheinlichkeitsverteilungen.*

Beispiele:

- *Die Wirksamkeit einer Behandlung* unter standardisierten Bedingungen ist definiert durch die *Wahrscheinlichkeit* p *für Erfolg der Behandlung* oder durch die Differenz dieser Wahrscheinlichkeit im Vergleich zur Erfolgswahrscheinlichkeit einer Placebo-Behandlung.

- *Der Einfluss des Luftanhaltens auf die Pulsfrequenz* ist definiert durch den *Erwartungswert* μ *der Pulsänderung.*

- Das *Ausmaß der Progression einer Koronarsklerose nach einem Jahr* ist definiert durch den *Erwartungswert* μ *der Differenz der Stenosendurchmesser* zwischen beiden Zeitpunkten.

4.3.1. Schätzung des Erwartungswertes

Im Falle des Erwartungswertes ist nun aufgrund der dargestellten Beziehung zwischen Grundgesamtheit und Stichprobe naheliegend:

Der (empirische) Mittelwert einer Stichprobe wird als "*Schätzwert*" für den unbekannten Erwartungswert μ angesehen.

Ganz allgemein können die in der deskriptiven Statistik gebildeten (empirischen) *Kenngrößen einer Häufigkeitsverteilung* als *Schätzung der korrespondierenden Kenngrößen der Wahrscheinlichkeitsverteilung* (der Grundgesamtheit) aufgefaßt werden.

Wir hatten nun aber mehrfach gesehen, dass der Schätzwert $\overline{y}_n$ *eigentlich immer falsch* ist (denn er ändert sich ständig, nach jedem neuen Einzelversuch und von einer Versuchsserie zur andern). Da es aber keine Alternative zur Schätzung gibt (Rechnen geht nicht: die Voraussetzungen sind nicht exakt bekannt), versucht man wenigstens, herauszufinden, *wie gut denn die Schätzung ist.*

Für die Schätzung des Erwartungswertes durch den Stichprobenmittelwert gilt dazu Folgendes:

1. Bildung des Mittelwertes:
   Die *Zufallsvariablen* $Y_1, Y_2, \ldots, Y_n$ beschreiben die Ergebnisse der n unabhängigen und unter gleichen Bedingungen durchgeführten Einzelexperimente mit zufälligem Ausgang. Der Mittelwert daraus:

   $$\overline{Y}_n = \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

   ist ebenfalls vom Zufall abhängig und daher eine *Zufallsvariable.*

2. Wenn $\mu = E(Y_1) = E(Y_2) = \ldots = E(Y_n)$ den gemeinsamen Erwartungswert aller Einzelexperimente bezeichnet, so gilt für den *Erwartungswert des Mittelwertes* $\overline{Y}_n$:

   $$E(\overline{Y}_n) = \mu$$

   Dies kann so interpretiert werden, dass der Erwartungswert μ vom empirischen Mittelwert $\overline{Y}_n$ –nicht im Einzelfall, aber *"in der Erwartung"– richtig* geschätzt wird. Der Mittelwert $\overline{Y}_n$ wird daher als "*erwartungstreuer Schätzer*" für μ bezeichnet.

3. Bereits erwähnt wurde, dass nach dem "Gesetz der großen Zahl" der Mittelwert $\overline{Y}_n$ für große n beliebig nahe an den gesuchten Parameter μ herankommt. Aufgrund dieser Eigenschaft heißt $\overline{Y}_n$ auch "*konsistenter*" Schätzer für μ.

Daher ist man mit dem *Schätzer* $\overline{Y}_n$ also zumindest *auf dem richtigen Weg* zum Erwartungswert μ !

4.3.2. Schätzung der Varianz

Die *Varianz* σ^2 einer Verteilung war nach (3.21) definiert als die "zu erwartende quadratische Abweichung vom Erwartungswert":

$$\sigma^2 = var(Y) = \sum_{i=1}^{K} (w_i - \mu)^2 \; P(Y = w_i)$$

wobei w_i $(i = 1, 2, \ldots, K)$ die möglichen Werte von Y bezeichnen. Analog dazu war in der deskriptiven Statistik für eine Stichprobe $y_1, y_2, \ldots, y_n$ von Beobachtungen des Merkmals Y die "empirische" oder "Stichproben"–Varianz durch

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2$$

definiert. Beachtet man, dass die Beobachtungen y_1, $y_2, \ldots, y_n$ Realisierungen von n unabhängigen, identisch verteilten Zufallsvariablen $Y_1, Y_2, \ldots, Y_n$ sind, so kann man

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2$$

nun als *Schätzer für die Varianz* σ^2 *der Grundgesamtheit* auffassen. Man kann zeigen, dass s_Y^2 "*erwartungstreu*" ist:

$$E(s_Y^2) = \sigma^2$$

D.h.: "Im Schnitt" trifft die *Stichprobenvarianz* s_Y^2 genau die *Varianz "in der Grundgesamtheit"* σ^2 . Die Division durch $n-1$ statt durch n erhält hiermit also ihre genaue Begründung.

4.3.3. Der Standardfehler des Mittelwertes

Konsistenz und Erwartungstreue von $\overline{Y}_n$ sind schon mal wünschenswerte Eigenschaften des Schätzers $\overline{Y}_n$. Für die jeweilige Anwendung will man aber auch noch wissen, mit welcher Unsicherheit der in einer Studie aktuell gefundene Mittelwert $\overline{y}$ immer noch verbunden sein kann: $\overline{y}$ ist ja nur die Realisierung einer *Zufallsvariablen* $\overline{Y}_n$: Mit welchen Abweichungen vom Erwartungswert μ muss man immer noch rechnen?

Aus der Abbildung 4.1 geht hervor, dass die Streuung des Mittelwertes mit wachsendem Stichprobenumfang kleiner wird. Die Wahrscheinlichkeitsrechnung kann das präzisieren:

Ist $\sigma^2 = var(Y_i)$ die Varianz der Einzelversuche $Y_1, Y_2, \ldots, Y_n$, so hat der Mittelwert $\overline{Y}_n$ aus n Einzelversuchen nur noch die Varianz bzw. die Standardabweichung

$$var(\overline{Y}_n) = \frac{1}{n}\,\sigma^2 \qquad \text{bzw.} \tag{4.1}$$

$$\sigma(\overline{Y}_n) = \sqrt{\frac{1}{n}\,\sigma^2} = \frac{\sigma}{\sqrt{n}} \tag{4.2}$$

(Die "*Sigma durch Wurzel n*" — Regel)

Das wollen wir doch mal sehen!

Dazu haben wir das Eingangsexperiment (Pulsänderung nach Luftanhalten) noch in 9 weiteren Versuchsserien wiederholt, wenn auch nur bis $n \approx 20$ je Versuchsserie:

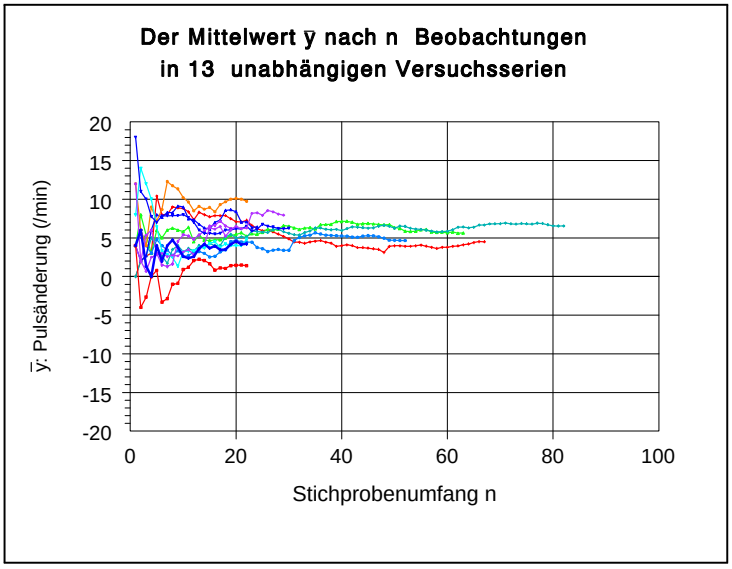


Abbildung 4.2

Sehen wir erst mal nur danach, *wie die Mittelwerte streuen*. Bei jedem Stichprobenumfang wird dazu die Streuung der Mittelwerte aus diesen 13 Versuchsserien berechnet und in das folgende Diagramm eingetragen. Zusätzlich ist noch die theoretisch berechnete "Sigma durch Wurzel n"- Regel eingezeichnet:

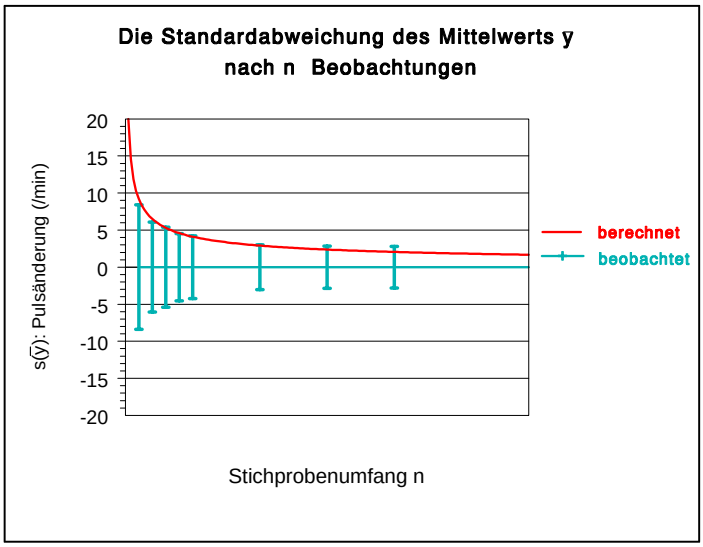


Abb. 4.3

Offenbar stimmen Vorausberechnung und Beobachtung recht gut überein. Das ist prima: dann brauchen wir zukünftig eine Versuchsserie nicht -wie hier getan- mehrfach zu wiederholen, um die Streuung des Mittelwertes der Serie zu berechnen; die *eine* Serie reicht: wir berechnen daraus die Varianz und dividieren durch Wurzel n - fertig!
Nur ein kleines Problem noch: Die Varianz σ^2 ist ja nicht wirklich bekannt. Man muss sie daher durch die *Schätzung der Varianz,* also durch s_Y^2 ersetzen. Die Streuung des Mittelwertes wird also geschätzt durch "S durch Wurzel n" statt durch "Sigma durch Wurzel n". Man nennt dies den "*Standardfehler des Mittelwertes*" (engl. *S.E.M.="Standard Error of the Mean"*):

$$S.E.M. = \frac{s_Y}{\sqrt{n}} \tag{4.3}$$

Einschub: Wie lügt man mit Statistik?

Beispiel: Es soll dargestellt werden, wie die Verteilung des Pulsschlags vor und nach dem Luftanhalten ist und ob es dabei Unterschiede zwischen Frauen und Männern gibt. Erster Versuch: Graphische Darstellung von Mittelwert $\pm$ Standardabweichung $(\overline{x} \pm s)$. Ergebnis in Abbildung 4.4 (s.u.). Interpretation: Zwischen unterem und oberem Balken sind (bei Normalverteilungsannahme) jeweils etwa 68 % der Studierenden vertreten. Es gibt also sehr breite Überlappungen aller Verteilungen. Und es gibt eine Erhöhung der Werte nach dem Luftanhalten. Diese Erhöhung ist bei den Studenten stärker ausgeprägt als bei den Studentinnen. Aber offenbar ist die Erhöhung in keinem Fall "riesig groß", jedenfalls nicht so groß wie z.B. schon allein die Standardabweichung der Werte.

Der Untersucher findet diese Darstellung nicht imposant genug. Aber es soll doch imponieren. Also sucht er nach Wegen, die Balken enger zu kriegen, und erinnert sich an S.E.M. Mit $\overline{x} \pm S.E.M.$ erhält er folgendes Bild (Abb. 4.5):

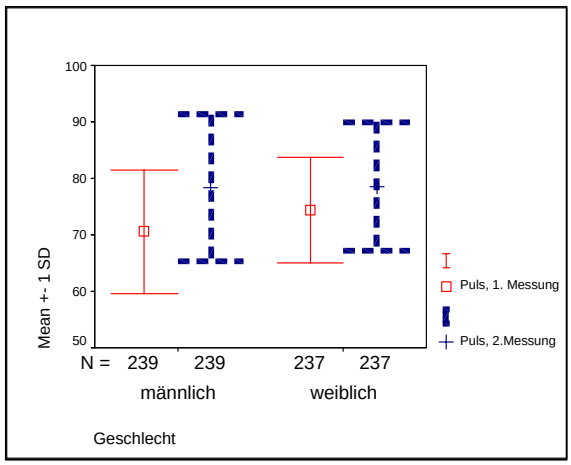


Abb. 4.4: $\overline{x} \pm s$

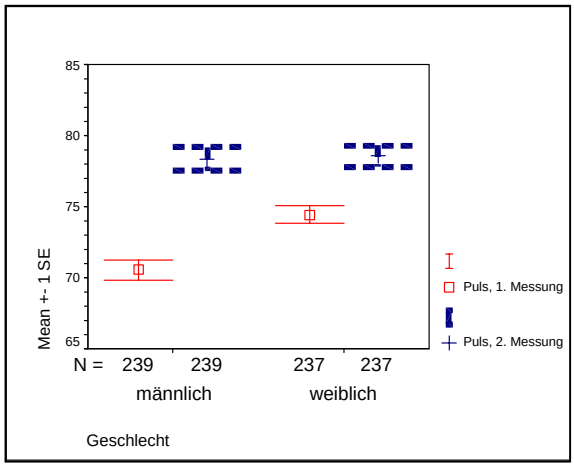


Abb. 4.5: $\overline{x} \pm S.E.M.$

Das sieht schon besser aus: Riesige Unterschiede a) zwischen Männern und Frauen bei der ersten Messung, und b) jeweils zwischen erster und zweiter Messung. Diese Abbildung kommt ins paper!

Ist das nun gelogen oder gemogelt?

Nicht wirklich, alles ist korrekt (sofern das Konstruktionsprinzip für die Fehlerbalken richtig angegeben wird). Es werden aber unterschiedliche Bewertungsmaßstäbe für die Mittelwertsunterschiede herangezogen: Links die *Streuung der Werte in der Untersuchungspopulation,* also die Streuung der Werte zwischen den einzelnen Studierenden; und rechts die *Streuung des Mittelwertes nach 237 bzw. 239 Versuchen!* Die ist natürlich erheblich kleiner, nämlich nur $1/\sqrt{237} \approx 1/15$ -tel der Streuung der Werte in der Population! In dieser Darstellung wird der Stichprobenmittelwert eigentlich "gegen den Zufall" verglichen. Bei genügend großem Stichprobenumfang können da kleine (auch klinisch unbedeutende) Mittelwertsunterschiede bedeutsam erscheinen, wenn sie größer sind als das, was man bei rein zufälligen Mittelwertsschwankungen erwarten würde. Das Mogeln beginnt erst dann, wenn man explizit oder unausgesprochen die Bewertungsmaßstäbe vertauscht und von statistisch erkennbaren Mittelwertsunterschieden zu klinischer Relevanz übergeht.

4.3.4. Die Verteilung des Mittelwertes

1. Nach dem Zentralen Grenzwertsatz (3.28) ist für große Stichprobenumfänge der standardisierte Mittelwert "ungefähr" normalverteilt:

$$Z_n = \frac{\bar{Y}_n - \mu}{\frac{1}{\sqrt{n}}\sigma} \approx N(0,1) \tag{4.4}$$

Geht man davon aus, dass die Varianz σ^2 der Grundgesamtheit "hinreichend" genau durch die empirische Varianz s_Y^2 geschätzt wird, so kann man darin σ durch s_Y ersetzen, d.h. auch

$$\frac{\bar{Y}_n - \mu}{\frac{s_Y}{\sqrt{n}}} = \frac{\bar{Y}_n - \mu}{S.E.M.}$$

ist "etwa" standard–normalverteilt.

2. Kann man eine Normalverteilung nicht erst (nach dem zentralen Grenzwertsatz) für den Mittelwert $\bar{Y}_n$ annehmen, sondern bereits für das Einzelexperiment Y, sind also alle "Einzelversuche" Y_i normalverteilt mit gleichem Erwartungswert und gleicher Varianz:

$$Y_i \sim N(\mu, \sigma^2) \qquad (i = 1, 2, \ldots, n)$$

so kann man mit mathematischen Mitteln zeigen, dass dann auch der Mittelwert (exakt, nicht "ungefähr"!) normalverteilt ist:

$$\begin{aligned} \bar{Y}_n &= \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \frac{\sigma^2}{n}) \qquad \text{und daher} \\ Z_n &= \frac{\bar{Y}_n - \mu}{\frac{1}{\sqrt{n}}\sigma} \sim N(0,1) \end{aligned}$$

Ersetzt man in dieser Situation wieder die (in der Regel unbekannte) Varianz σ^2 durch den Schätzwert s_Y^2, so ist die resultierende Verteilung wieder exakt berechenbar, jedoch stimmt sie –wie sich nach den vorigen Regeln vermuten läßt– erst für große Stichprobenumfänge mit der Standard–Normalverteilung überein: Die Verteilung von

$$T_n := \frac{\bar{Y}_n - \mu}{\frac{s_Y}{\sqrt{n}}} \tag{4.5}$$

heißt die (Student'sche) "*t-Verteilung*". Wie die Standard–Normalverteilung ist die t–Verteilung symmetrisch zum Wert 0. Da bei der Definition von T_n im Nenner statt der konstanten Zahl σ die stichprobenabhängige Streuung s_Y steht, hat die t–Verteilung aber eine etwas größere Varianz. Dadurch sind die Quantile der Verteilung –absolut gesehen– größer als die der Standard–Normalverteilung.
Die Abweichung von der Normalverteilung ist um so kleiner, je größer der Stichprobenumfang n ist (denn mit wachsendem n wird σ immer sicherer durch s_Y geschätzt). Insgesamt führt dies dazu, dass für jedes n eine "eigene" t–Verteilung definiert ist. Sie wird durch die Anzahl der "*Freiheitsgrade*" (df = degrees of freedom) näher charakterisiert. Allgemein:
Der standardisierte Mittelwert aus n Beobachtungen ist t–verteilt mit

$df = n - 1$ **Freiheitsgraden.**
In der folgenden Abbildung 4.6 ist als Beispiel die Dichtefunktion der t-Verteilung

mit10 Freiheitsgraden zusammen mit den Markierungen des 2.5%– und des 97.5%–Quantils dargestellt. Dies ist die Verteilung von T_n für $n = 11$.

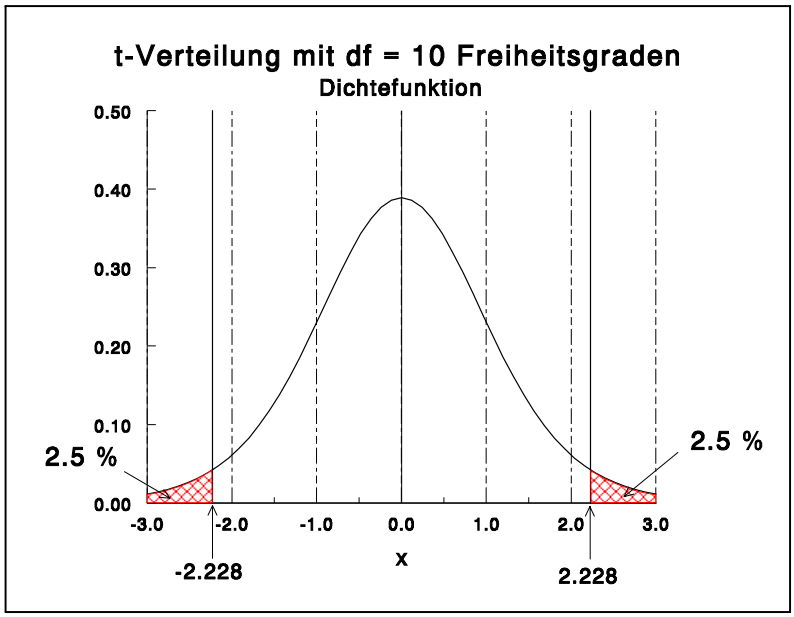


Abb. 4.6: t-Verteilung mit 10 Freiheitsgraden (Dichtefunktion)

3. Aus dieser Abbildung ist beispielsweise abzulesen:
Das 97.5%–Quantil der t–Verteilung mit 10 Freiheitsgraden ist 2.228 (statt 1.96 bei der Standardnormalverteilung). Es wird mit $t_{10,\ 0.975}$ bezeichnet: $t_{10,\ 0.975} = 2.228$. Allgemein bezeichnet $t_{n-1,\ q}$ das q–Quantil der t–Verteilung mit $n - 1$ Freiheitsgraden. Einige weitere Beispiele:

Freiheitsgrade(df):	5	10	50	100	∞
q					
0.5%–Quantil	−4.032	−3.169	−2.678	−2.626	−2.576
1%–Quantil	−3.365	−2.764	−2.403	−2.364	−2.326
2.5%–Quantil	−2.571	**−2.228**	−2.009	−1.984	**−1.960**
5%–Quantil	−2.015	**−1.812**	−1.676	−1.660	**−1.645**
95%–Quantil	2.015	**1.812**	1.676	1.660	**1.645**
97.5%–Quantil	2.571	**2.228**	2.009	1.984	**1.960**
99%–Quantil	3.365	2.764	2.403	2.364	2.326
99.5%–Quantil	4.032	3.169	2.678	2.626	2.576

Aufgrund der Symmetrie der Verteilung zum Wert 0 gilt z.B.(speziell für $q = \alpha/2$) :

$$P(|T_n| \leq t_{n-1,1-\alpha/2}) = 1 - \alpha \ , \tag{4.6}$$

d.h. T_n ist mit der Wahrscheinlichkeit $1 - \alpha$ Werte dem Betrage nach höchstens gleich dem oberen $\alpha/2$–Quantil an.
Umgekehrt ist daher

$$P(|T_n| > t_{n-1,1-\alpha/2}) = \alpha \ , \tag{4.7}$$

und speziell für $\alpha = 0.05 = 5\%$:

$$P(|T_n| > t_{n-1,0.975}) = 0.05 = 5\%: \tag{4.8}$$

T_n ist mit der Wahrscheinlichkeit von 5% dem Betrage nach größer als das obere 97.5%–Quantil der t-Verteilung mit n − 1 Freiheitsgraden.

4.4. Einschub: Das Dilemma der reinen Wahrscheinlichkeitsrechnung

Wahrscheinlichkeitsrechnung ist reine Mathematik. Die numerische Berechnung von Wahrscheinlichkeiten ist immer nur unter bestimmten *Annahmen und Voraussetzungen* möglich (z.B. wenn man den Erwartungswert einer Verteilung kennt). Aber über eben diese Voraussetzungen weiß man in der Regel nichts.

Man kann z.B. berechnen, dass (für n=11): $\left|\frac{\bar{Y}_n-\mu}{\frac{s_Y}{\sqrt{n}}}\right|$ mit 95% Wahrscheinlichkeit ≤ 2.28 ausfallen muß. Aber was nutzt einem das, wenn man μ nicht kennt?

Die 5%–Hürde: Aktuelle Stichprobenwerte versus Wahrscheinlichkeitsrechnung

Grundsatz:

Akzeptiere eine Annahme über die Verteilung von Y dann, aber auch nur dann, wenn die aktuelle Beobachtung aus der Stichprobe unter dieser Annahme nicht als "zu unwahrscheinlich" erscheint!

Übliche Grenze hierfür: 5% Wahrscheinlichkeit (die "*5%–Hürde*").

4.5. Konfidenzbereich: Welcher Erwartungswert einer Verteilung ist mit den Stichprobendaten "verträglich"?

Mit Wahrscheinlichkeit 95% fällt (für n=11) die Testgröße $T_n := \frac{\bar{Y}_n-\mu}{\frac{s_y}{\sqrt{n}}}$ zwischen die Grenzen -2.228 und +2.228 (siehe Abb. oben), d.h.:

$$-2.228 \leq \frac{\overline{Y}_n - \mu}{\frac{s_y}{\sqrt{n}}} \qquad \text{und} \qquad \frac{\overline{Y}_n - \mu}{\frac{s_y}{\sqrt{n}}} \leq 2.228$$

1. Daher ist jeder Wert μ mit den vorliegenden Daten $\overline{y}_n$und s_y (Mittelwert und Streuung) noch "vereinbar", der in diesem Sinne "nahe genug am Stichprobenmittelwert $\overline{y}_n$ liegt", für den also beide Ungleichungen erfüllt sind. Löst man diese nun in 2 Schritten nach μ auf, so ist dies gleichbedeutend mit:

   Schritt 1: $\quad -2.228\,\frac{s_y}{\sqrt{n}} \leq \overline{y}_n - \mu \quad$ und $\quad \overline{y}_n - \mu \leq +2.228\,\frac{s_y}{\sqrt{n}}$

   Schritt 2

   $\mu \leq \overline{y}_n + 2.228\,\frac{s_y}{\sqrt{n}}$ (rechte Grenze) und $\overline{y}_n - 2.228\,\frac{s_y}{\sqrt{n}} \leq \mu$ (linke Grenze)

2. Man erhält also das Intervall mit den Grenzen

$$\overline{y}_n - 2.228\ \frac{s_y}{\sqrt{n}} \qquad \text{und} \qquad \overline{y}_n + 2.228\ \frac{s_y}{\sqrt{n}}\ .$$

Bei dieser Konstruktion passiert es nur mit einer (Fehler-) Wahrscheinlichkeit von 5%, dass man mit dem angegebenen Intervall den "wahren" Parameter μ *nicht* erfasst.

3. Allgemeine Version: Das von der Stichprobe abhängige, also *zufällige Intervall* mit den Grenzen

$$\bar{Y}_n - t_{n-1,1-\alpha/2}\ \frac{s_Y}{\sqrt{n}} \qquad \text{und} \qquad \bar{Y}_n + t_{n-1,1-\alpha/2}\ \frac{s_Y}{\sqrt{n}} \tag{4.9}$$

überdeckt den wahren Erwartungswert μ der Verteilung von Y in der Grundgesamtheit mit der Wahrscheinlichkeit von $1-\alpha$.

Das Intervall (4.9) wird "*Konfidenzintervall für den Erwartungswert μ zum Niveau* $1-\alpha$" genannt.

Wie man aus der Gleichung sieht, ist es symmetrisch zum Mittelwert. Seine Länge ist $2 \times\ t_{n-1,1-\alpha/2} \times \frac{s_Y}{\sqrt{n}}$, ist also proportional zum Standardfehler des Mittelwertes $\frac{s_y}{\sqrt{n}}$, und zum oberen $(1-\alpha/2)$–Quantil der t-Verteilung mit $n-1$ Freiheitsgraden.

Rechenbeispiel:
Bei $n = 13$ Patienten mit peripheren arteriellen Stenosen wurde die Zunahme (Y) der Plaque-Fläche (gemessen in cm^2) innerhalb eines Jahres gemessen. Man erhielt die folgenden Ergebnisse:

$$\begin{aligned} \overline{y} &= 0.21 \\ s_y &= 0.32 \end{aligned}$$

Da die Anzahl der Beobachtungen $n = 13$ war, wird zur Berechnung des 95%−Konfidenzintervalls das 97.5%−Quantil der t−Verteilung mit $n - 1 = 12$ Freiheitsgraden benötigt. Aus einer entsprechenden Tabelle liest man ab:

$$t_{12,\ 97.5\%} = 2.179$$

Daraus erhält man die folgenden Grenzen für das 95%−Konfidenzintervall:

$$0.21 - \frac{0.32}{\sqrt{13}}\ 2.179 \qquad \text{und} \qquad 0.21 + \frac{0.32}{\sqrt{13}}\ 2.179\text{:}$$

also:

0.017 (linke Grenze) und 0.403 (rechte Grenze).

Interpretation: Die mittlere Zunahme betrug 0.21 cm^2. Der Erwartungswert für die Zunahme liegt in der *Grundgesamtheit* (d.h.: falls die hier geltenden Rahmenbedingungen für die Stichprobenauswahl, die Behandlung und die Meßmethoden eingehalten werden) zwischen 0.017 cm^2 und 0.403 cm^2. Und: die Methode, die zu dieser Aussage geführt hat, liefert mit 95% Wahrscheinlichkeit eine richtige Aussage.

Man beachte: die Aussage: "Der Ewartungswert liegt mit 95% Wahrscheinlichkeit zwischen 0.017 cm^2 und 0.403 cm^2" wäre *nicht* richtig, denn der Erwartungswert ist ein fester Wert und ist insbesondere also auch *nicht zufallsabhängig*. Eine *Wahrscheinlichkeitsaussage* kann sich daher auch nicht auf die Lage des *Erwartungswertes* beziehen sondern *nur auf zufällige Ereignisse*, in diesem Fall also auf die *Grenzen des Intervalls*. Dies steht in Übereinstimmung damit, dass in der Formel (4.9) die Zufallsvariablen $\bar{Y}_n$ und s_Y (großes Y) und nicht deren mögliche Realisierungen (kleines y) notiert sind.

5. Statistisches Testen (GK 5)

5.1. Vom Konfidenzbereich zum statistischen Test (GK 5.1)

Am Ende des vorigen Abschnitts wurde die Bildung eines *Konfidenzbereiches* als ein Verfahren vorgestellt, welches *den "wahren Wert" des Parameters einer Verteilung* (z.B. Erwartungswert μ einer Normalverteilung, Wahrscheinlichkeit p einer Bernoulli-Verteilung) *mit vorgegebener Wahrscheinlichkeit* (z.B. 95%) *überdeckt.* Aus der Herleitung ging hervor, dass man einen Konfidenzbereich auch auffassen kann als *die Gesamtheit derjenigen numerischen Werte, die der interessierende Paramter haben könnte, ohne dass die vorliegenden Daten "ernsthaft dagegen sprächen"* (dies war im Fall des Erwartungswertes μ durch die Bildung des 95%–Bereichs des standardisierten Stichprobenmittelwertes $\overline{Y}_n$ präzisiert worden). Kurz gesagt: *Ein Konfidenzbereich enthält alle diejenigen Parameterwerte, die mit den Daten der Stichprobe "verträglich sind".*
Während man mit der Bildung des Konfidenzbereiches also zunächst "offen" an die Untersuchung eines Paramters herangeht, interessiert man sich oft darüber hinaus oder gar ausschließlich für einen ganz bestimmten hypothetischen Wert des Parameters, z.B $\mu = 0$, und möchte wissen, ob *dieser spezielle Wert* "mit den Daten verträglich" ist:

Fragestellung:

Allgemein:	**Speziell:**
Wie groß ist die Wirkung des Medikamentes?	Ist die Wirkung des Medikamentes gleich **null**?
Wie groß sind die Wirkungen der Behandlungen A und B im Vergleich; wie groß ist die Differenz ihrer Wirkungen?	Ist die Differenz der Wirkungen von Behandlung A und B gleich **null**?
Wie groß ist die Korrelation zwischen dem Alter und dem Test–Score im Zahlenverbindungstest?	Ist diese Korrelation zwischen Alter und Zahlenverbindungstest gleich **null**?

Ein **statistischer Test** behandelt die *spezielle* Version der Fragestellung: Es ist ein Verfahren, welches eine *Entscheidung* darüber erlaubt, ob der wahre *Parameter einer Verteilung gleich einem hypothetischen, vorgegebenem Wert ist* (z.B. $\mu = 0$, wenn μ der Erwartungswert der Verteilung der Blutdrucksenkung ist).

Bezeichnung:
In vielen Anwendungen ist der hypothetische Wert, dessen Richtigkeit durch den Test überprüft werden soll, gleich **null**. Allgemein wird daher die zugehörige **Hypothese**, die zu prüfen ist, die **Nullhypothese** H_0 genannt.

Vorgehensweise:
Steht zu dem in Frage stehenden Parameter μ einer Verteilung die Konstruktion eines **95%-Konfidenzbereichs** (C.I.) zur Verfügung, so geht man so vor:

Betrachte die Nullhypothese $H_0 : \mu = \mu_0$ als widerlegt, wenn μ_0 nicht im 95%-Konfidenzintervall C.I. zu μ enthalten ist.

In diesem Fall kommt der hypothetische Wert μ_0 also als "wahrer" Wert des Parameters μ "nicht in Betracht" und wird abgelehnt. Genauer gilt:

Angenommen, die Nullhypothese $H_0 : \mu = \mu_0$ ist richtig. Die Wahrscheinlichkeit, dass H_0 dennoch abgelehnt wird, ist in diesem Fall höchstens gleich 5%[2]

Auf diese Art wird also die Wahrscheinlichkeit kontrolliert, dass das Verfahren zu einem Fehler der folgenden Art führt:

Fehler 1. Art:	**Ablehnung der Nullhypothese, obwohl diese richtig ist**

Wählt man in der obigen Vorgehensweise statt des 95%-Konfidenzintervalls das 99%-Konfidenzintervall, so ist die Wahrscheinlichkeit dafür, dass der Fehler 1. Art (wenn H_0 richtig ist) gleich $0.01 = 1\%$. Wie man sieht, kann man diese Wahrscheinlichkeit **vorgeben** und **bestimmt das Testverfahren so, dass diese Vorgabe eingehalten wird**.

Die (obere Grenze für die) **Wahrscheinlichkeit für den Fehler 1. Art wird mit α bezeichnet**

Üblich für seine (willkürliche) Festlegung sind: $\alpha = 0.05$. 0.01 oder 0.001 ($\hat{=}5\%$, 1% und 0.1%)

Vorläufige Zusammenfassung und Definition:

Ein Signifikanztest zum Niveau α ist ein Verfahren, welches in Abhängigkeit von den Daten einer Stichprobe über die Ablehnung oder die Annahme einer Nullhypothese H_0 entscheidet; ist dabei H_0 tatsächlich richtig, so wird höchstens mit der Wahrscheinlichkeit α *gegen* H_0 entschieden.

Die Kontrolle des Fehlers 1. Art (auch "α–Fehler" genannt) schützt ("bis auf α") vor der fälschlichen Ablehnung einer Nullhypothese, schützt also beispielsweise

[2]Das kann man auch leicht herleiten: Mit P_{μ_0} werden i.f. Wahrscheinlichkeiten bezeichnet, die unter der Annahme gelten, dass μ_0 der wahre Parameter ist. Dann ist:

$$\begin{aligned} & P_{\mu_0}(\text{ Ablehnung der Nullhypothese } \quad H_0 : \mu = \mu_0) \\ = \; & P_{\mu_0}(C.I. \quad \text{überdeckt} \quad \text{nicht} \quad \mu_0) \\ = \; & 1 - P_{\mu_0}(C.I. \quad \text{überdeckt} \quad \mu_0) \\ = \; & 1 - 0.95 = 0.05 = \; 5\% \end{aligned} \tag{5.1}$$

- vor der Zulassung und Verordnung eines Medikamentes, dessen Wirkung (im Vergleich zum Placebo) *gleich null* ist,
- vor der (zusätzlichen) Anwendung eines diagnostischen Eingriffs, dessen Ergebnis mit den untersuchten Krankheiten *keinen Zusammenhang* hat,
- vor der Deklaration eines "Risikofaktors" für Herzinfarkt, der das Infarktrisiko *gar nicht beeinflusst.*

Die Anwendung eines Testes ist in diesem Sinn eine *Sicherheitsmaßnahme.* Tatsächlich aber wird der Test und das damit eingeleitete Entscheidungsverfahrem *mit der Intention durchgeführt, die Nullhypothese zu widerlegen*: Man *sucht*

- nach wirksamen Medikamenten und will ihre Wirksamkeit, *wenn sie vorhanden ist, nachweisen*;
- nach Maßnahmen, welche ein Diagnoseverfahren echt verbessern können;
- nach *tatsächlich vorhandenen Zusammenhängen,* wenn man nach den Ursachen oder Entstehungsbedingungen einer Krankheit forscht.

Während man sich also einerseits davor hüten möchte, bei der Behauptung von Zusammenhängen irgendwelchen Zufälligkeiten aufzusitzen, sollen andererseits aber *tatsächlich vorhandene Zusammenhänge* auch erkannt und nachgewiesen werden. Ein Signifikanztest sollte also im folgenden Sinne eine hohe "Testschärfe" haben:

Testschärfe (Güte, "power") = Wahrscheinlichkeit, eine tatsächlich vorliegende Abweichung von der Nullhypothese zu entdecken

Eine hohe Testschärfe ist damit gleichbedeutend mit einer niedrigen Wahrscheinlichkeit für den folgenden Fehler:

Fehler 2. Art: Beibehaltung der Nullhypothese, obwohl sie falsch ist

Die Wahrscheinlichkeit für den Fehler 2. Art wird i.a. mit β bezeichnet und der Fehler 2. Art dementsprechend auch der "β–Fehler" genannt. Es gilt somit

$$\beta \;=\; 1 \;-\; \text{Güte ("power") des Tests}$$

Während die Wahrscheinlichkeit α für den Fehler 1. Art einzig durch die *Festlegung* des Anwenders bestimmt wird, hängt die power eines Tests von Umständen ab, die teilweise gar nicht und teilweise durch die Versuchsanordnung *beeinflußt* werden können:

Wovon hängt die "power" ab (wenn α festgelegt ist)**?**

1. Vom Ausmaß der tatsächlichen Abweichung von der Nullhypothese (eine starke Wirksamkeit eines Medikamentes wird leichter erkannt als eine schwache Wirksamkeit),

2. vom Stichprobenumfang (je größer der Stichprobenumfang, desto eher wird eine Abweichung von der Nullhypothese erkannt),

3. von der Streuung der Daten in der Grundgesamtheit (eine große Streuung verdeckt die tatsächliche Wirkung),

4. von der speziellen Wahl der statistischen Testprozedur.

Die hier genannten Punkte zur Beeinflussung der power eines Tests sind an dieser Stelle höchstens "plausibel", können aber alle spezifiziert und mit mathematischen Methoden präzisiert werden. Tatsächlich besteht die Aufgabe der mathematischen Statistik im Rahmen der Testanwendung u.a. gerade darin, die oben genannten Abhängigkeiten zu quantifizieren, um dem Anwender Kriterien für die Versuchsplanung und für die Auswahl eines geeigneten Tests zu liefern. Dies wiederum ist im Gesamtzusammenhang zu sehen, der durch die medizinischen Zielsetzungen und Fragestellungen einerseits und ihre partielle Übersetzung in Terminologie und Methoden der mathematischen Statistik andererseits gekennzeichnet ist. Bevor hierauf eingegangen werden kann, muß zunächst eine **Basisaufgabe der Statistik** erledigt werden:

> **Die Bereitstellung von Signifikanztests für unterschiedliche Situationen und Nullhypothesen**.

Nachdem eingangs dieses Kapitels gezeigt wurde, dass man *Konfidenzintervalle* zur Testkonstruktion benutzen kann, soll im nächsten Abschnitt ein weiteres Konstruktionsprinzip vorgestellt werden.

5.2. Abweichungsmaße und Testkonstruktion (GK 5.1, 5.2.1)

5.2.1. Beispiel 1 (Vorzeichentest).

Das folgende hypothetische Beispiel soll die Vorgehensweise bei der Konstruktion und Durchführung eines Signifikanztests verdeutlichen.

Fragestellung: Es soll eine neue Salbe zur Behandlung von Neurodermitis untersucht werden, die nach allen präklinischen Erfahrungen erfolgversprechend erscheint. Durch eine empirische Studie soll nun belegt werden, dass die neue Salbe wirksamer ist als eine Standardsalbe.
Der Erfolgsnachweis:
Es werden 10 Patienten mit Neurodermitis in die Studie aufgenommen, bei denen beide Arme etwa gleich stark betroffen sind.
Bei jedem Patienten wird auf dem einen Arm stets die alte, auf dem anderen stets die neue Salbe angewendet. Die Zuteilung von alter und neuer Salbe auf den rechten bzw. linken Arm erfolgt per Zufall.
Nach 3 Monaten werden von einem unabhängigen Arzt (der nicht weiß, welcher der beiden Arme mit der Standard– und welcher mit der neuen Salbe behandelt wurde) miteinander verglichen. Der Vergleich wird mit "+" notiert, wenn sich anschließend herausstellt, dass der Zustand des Arms mit der neuen Salbe als *besser* bewertet wurde, im umgekehrten Fall mit "−" (der Fall "unentschieden" soll nicht vorkommen). Das Ergebnis:

Paar Nr.	1	2	3	4	5	6	7	8	9	10
Ergebnis	+	−	+	+	+	+	+	−	+	+

8 mal + , 2 mal −
Ist damit der Nachweis gelungen?
Argumentation:

1. Die Ergebnisse zeigen, dass das Ergebnis der neuen Salbe auch *schlechter* als das der alten Salbe sein kann.

2. Das Einzelergebnis eines Patienten ist also jedenfalls *nicht sicher* vorhersagbar, d.h. die Einzelergebnisse sind *zufallsabhängig*.

3. Wahrscheinlichkeitsmodell:
   Das Ergebnis "+" wird mit der (unbekannten) Wahrscheinlichkeit p erhalten.

4. Wenn die neue Salbe *nicht besser* ist als die Standardsalbe, sind die Ergebnisse "+" und "-" *gleichwahrscheinlich*; dann ist also

$$p = \frac{1}{2} = 0.5$$

   Dies ist die *Nullhypothese* des Tests.

5. Dann kann man die Wahrscheinlichkeit für genau k positive Ergebnisse bei $n = 10$ Patienten genau berechnen (siehe Binomialverteilung (3.22)) :

$$\begin{aligned} P(X = k) &= \frac{10!}{k!(10-k)!}\left(\frac{1}{2}\right)^k\left(1-\frac{1}{2}\right)^{10-k} \\ &= \frac{10!}{k!(10-k)!}\left(\frac{1}{2}\right)^{10} \end{aligned}$$

   Also kann man z.B. auch berechnen, mit welcher Wahrscheinlichkeit das zusammengesetzte Ergebnis "*mindestens 8 mal positiv*" auftritt:

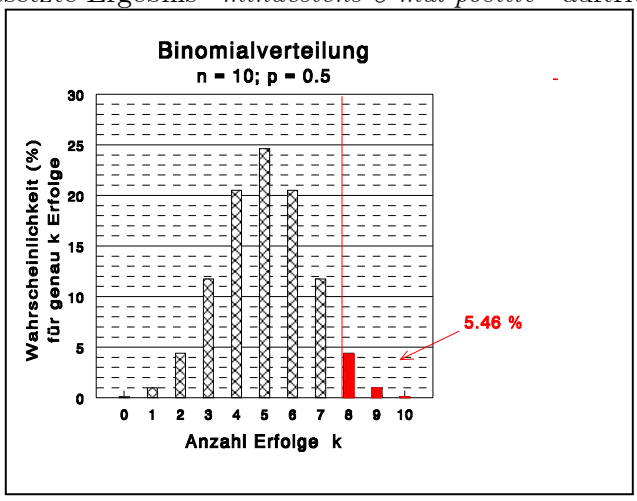


   Abb. 5.1: Die Wahrscheinlichkeit für 8 oder mehr Erfolge

   Die Wahrscheinlichkeit hierfür beträgt 5.46 %.

6. Vereinbarung:
   Die Überlegenheit der neuen Salbe soll als *nachgewiesen* gelten, wenn die Anzahl der *positiven* Ergebnisse "sehr groß" ist. Genauer: *Lege eine "kritische*

Zahl" k_0 fest, so dass diese Zahl höchstens mit einer Wahrscheinlichkeit von $\alpha = 0.05 = 5\%$ "rein per Zufall" erreicht oder übertroffen werden kann (wenn also die Nullhypothese richtig ist, d.h. wenn die Wahrscheinlichkeit für ein positives Ergebnis im Einzelfall gleich 0.5 ist). Im übrigen soll k_0 aber möglichst klein sein, damit ein Nachweis im anderen Fall auch möglichst leicht gelingen kann.

7. Die kleinste Zahl k_0, die diese Bedingung erfüllt, ist $k_0 = 9$:

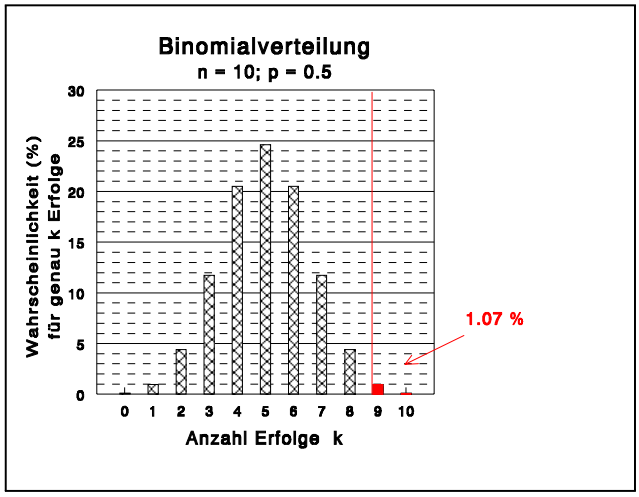


Abb. 5.2: Wahrscheinlichkeit f ür 9 oder mehr Erfolge

Der positive Effekt der neuen Salbe gegenüber dem Standard gilt daher als nachgewiesen, wenn bei *mindestens 9* Patienten der *mit der neuen Salbe* behandelte Arm einen *besseren* Zustand aufweist.

8. Allgemeine Folgerung dieser Vereinbarung:
   Die Wahrscheinlichkeit, dass eine positive Wirkung "nachgewiesen" wird, wenn sie in Wirklichkeit gar nicht vorhanden ist, ist $\leq \alpha = 5\%$.

9. Folgerung im vorliegenden Fall:
   Eine Überlegenheit der neuen Salbe kann nach diesen Vorgaben *nicht* als nachgewiesen gelten, denn das Ergebnis lautet nur "8 mal positiv", die Nachweisgrenze $k_0 = 9$ wurde also nicht erreicht.

Bei der Durchführung dieses Tests wurden nur die *Vorzeichen* der einzelnen Paarvergleiche verwendet; er wird daher als "*Vorzeichentest*" bezeichnet. In der angegebenen Form ist er auf viele verschiedene Situationen übertragbar. Für größere Stichprobenumfänge kann man dabei die Berechnung der Binomial–Wahrscheinlichkeiten durch die Normalverteilungsapproximation ersetzen (siehe (3.26) und Abb. 3.9).
Hinter der *Konstruktion* dieses Tests verbirgt sich ein allgemeines Prinzip, das im folgenden Abschnitt im einzelnen dargestellt wird:

5.2.2. Allgemeines Konstruktionsprinzip

1. Es seien $X, Y, \ldots$ Zufallsvariable, die die Ausgänge von einem oder mehreren Experimenten beschreiben.

2. Die Wahrscheinlichkeitsverteilung von $X, Y, \ldots$ sei "im Wesentlichen" bekannt, lediglich der Wert eines oder mehrerer *Parameter der Verteilung sind unbekannt*

(aber mindestens einen von diesen Werten will man gerade untersuchen!)

Unter der *Nullhypothese* H_0 wird eine bestimmte Bedingung für den interessierenden Parameter formuliert.
Die *Alternativhypothese* H_1besagt, dass die Nullhypothese *nicht richtig* ist.

3. Aus einer Stichprobe vom Umfang n zum Experiment $X, Y, \ldots$ wird eine Testgröße S gebildet, aus der man die "*Distanz der vorliegenden Stichprobendaten zu der Nullhypothese*" ablesen kann.
   Dann sprechen *dem Betrage nach* große Werte von S *gegen die Nullhypothese.*

4. Diese Testgröße oder "Teststatistik" S muß so beschaffen sein, dass ihre Wahrscheinlichkeitsverteilung zumindest dann *genau berechnet werden kann, wenn die Nullhypothese gültig ist.*

5. Lege dann zu vorgegebenem Signifikanzniveau α einen "*Ablehnbereich*" für die Testgröße S so fest, dass die Testgröße S bei Gültigkeit der Nullhypothese höchstens mit der Wahrscheinlichkeit α in diesen Bereich fällt:

$$P_{H_0}(\ S \text{ fällt in den Ablehnbereich}) \leq \alpha \tag{5.2}$$

   wobei P_{H_0} die Wahrscheinlichkeit unter der Annahme bezeichnet, dass H_0 richtig ist.
   Der Ablehnbereich wird in der Regel durch "kritische Werte" für die Testgröße definiert, etwa in der Form:

$$\begin{aligned} &\text{"Ablehnung von H}_0\text{,wenn } S &&> \lambda\text{"} \qquad \text{oder} \\ &\text{"Ablehnung von H}_0\text{,wenn } |S| &&> \lambda\text{"} \end{aligned}$$

Aus der letzten Gleichung (5.2) wird unmittelbar klar, dass bei diesem Konstruktionsprinzip die Wahrscheinlichkeit für den Fehler 1. Art höchstens gleich α ist, dass also das vorgegebene "Signifikanzniveau" eingehalten wird. Das eigentliche Problem bei diesem Prinzip liegt in Punkt 5: Auch unter der Nullhypothese ist die Klasse der Verteilungen der Basisvariablen $X, Y, \ldots$ häufig nicht vollständig bekannt (Beispiel: Ist unter H_0 der Erwartungswert einer normalverteilten Zufallsvariablen gleich 0, so ist wegen der unbekannten Varianz σ^2 weder die Verteilung von X noch die vom Mittelwert $\overline{X}_n$ bekannt). Diese Aufgabe wird in den einzelnen Fällen unterschiedlich gelöst, wie aus der Beschreibung einzelner Test hervorgehen wird.

5.2.3. Beispiel 2 (Z–Test):

Wenn die Stichprobe $X_1, X_2, \ldots, X_n$ von Messungen der Blutdruckveränderung ("vor Behandlung" minus "nach Behandlung") aus einer normalverteilten Grundgesamtheit stammen mit jeweils dem (unbekannten) Erwartungswert μ und der (bekannten) Streuung $\sigma = 10$ mmHg, so ist der Stichprobenmittelwert

$$S(X_1, X_2, \ldots, X_n) := \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

ebenfalls normalverteilt mit demselben Erwartungswert μ und der Streuung $\frac{\sigma}{\sqrt{n}}$ (siehe (4.1)). Wenn man dann annimmt, dass *die Nullhypothese* $H_0 : \mu = 0$ *richtig* wäre (angenommen also: Keine Veränderung des Blutdrucks zu erwarten), so kann man (rein mathematisch, ohne weitere Daten erheben zu müssen) *vorausberechnen*, mit welchen Wahrscheinlichkeiten die Testgröße $S = \overline{X}$ nach z.B. $n = 49$ Versuchen welche Werte annehmen wird: $S \sim N(0, \left(\frac{10}{\sqrt{49}}\right)^2)$, d.h. S ist normalverteilt mit Erwartungswert 0 und Standardabweichung $\frac{10}{\sqrt{49}} = \frac{10}{7} = 1.4286$. Um die Berechnung dieser Verteilung auf die *Standard*normalverteilung zurückzuführen, transformiert man S noch mit Hilfe der Z–Transformation und erhält dann die standard–normalverteilte Testgröße Z:

$$Z = \frac{S - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}_n}{1.4286} \sim N(0,1)$$

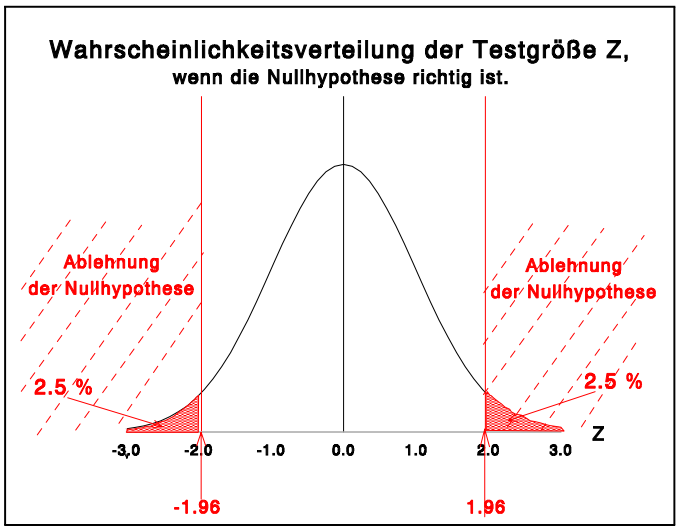


Abb.5.3: Ablehnbereich und "kritischer Wert" einer Testgöße

Es wird nun vorab festgelegt, dass die Nullhypothese abgelehnt wird, wenn Z "stark positiv" oder "stark negativ" ausfällt, genauer: wenn $Z < -1.96$ oder $Z > 1.96$ wird. Damit ist der *Ablehnbereich* festgelegt:

$$\text{Lehne die Nullhypothese ab, wenn } |Z| = \frac{|\overline{X} - \mu|}{\frac{\sigma}{\sqrt{n}}} = \frac{|\overline{X}_n|}{1.4286} \geq 1.96$$

Der sogenannte "*kritische Wert*" $\lambda = 1.96$ für $|Z|$ ist dabei so gewählt, dass die Forderung (5.2) erfüllt ist, wenn die Fehler–I–Wahrscheinlichkeit α auf 5% festgelegt ist: Wenn H_0 richtig ist, ist die Wahrscheinlichkeit für eine Ablehnung von H_0 nicht größer als 5%.

Beispiel (Z–Test für $n = 49$, $\sigma = 10$, $\alpha = 0.05 = 5\%$; Nullhypothese $H_0 : \mu = 0$):

Mittelw. $\overline{X}$	Testgröße Z	Kriterium:	Entscheidung:
-4.3	-3.01	$\lvert -3.01\rvert = 3.01 \geq \mathbf{1.96}$	**Ablehnung der Nullhypothese**
-1.5	-1.05	$\lvert -1.05\rvert = 1.05 < 1.96$	Nullhypothese nicht widerlegt
1.8	1.26	$\lvert 1.26\rvert = 1.26 < 1.96$	Nullhypothese nicht widerlegt
2.3	1.75	$\lvert 1.75\rvert = 1.75 < 1.96$	Nullhypothese nicht widerlegt
3.3	2.31	$\lvert 2.31\rvert = 2.31 \geq \mathbf{1.96}$	**Ablehnung der Nullhypothese**

Im Fall der Ablehnung der Nullhypothese wird dies so interpretiert, dass eine Veränderung des Blutdrucks unter der Behandlung als "*statistisch gesichert*" oder "*signifikant*" gilt. (Die Veränderung muß aber ihre *Ursache* nicht in der Behandlung selber haben, sie könnte auch als Placebo–Effekt oder "natürlicher Verlauf" auftreten. Um hierüber Klarheit zu gewinnen, müßte auch eine *Kontrollgruppe* in die Untersuchung einbezogen werden.)

5.2.4. Der P–Wert einer Testanwendung.

Aus dem vorigen Beispiel ging hervor: Die Durchführung eines Tests geschieht so, dass man die aus den Stichprobenwerten berechnete *Testgröße* mit einem "*kritischen Wert*" λ vergleicht. Dieser Wert λ wird *vor der Durchführung des Tests* aufgrund von Vorgaben und Annahmen (Stichprobenumfang n; Signifikanzniveau α; Annahme einer zugrundeliegenden Verteilung, z.B. Normalverteilung) *berechnet* bzw. aus statistischen Tabellen abgelesen. Der Vergleich von Testgröße und kritischem Wert führt dann zur Beibehaltung oder Ablehnung der Nullhypothese.
Wenn man ein entsprechendes Rechenprogramm hat, kann man auch anders vorgehen: Angenommen, im vorliegenden Beispiel ist der Stichprobenmittelwert $\overline{x} = 3.3$ (wie in der letzten Zeile der Tabelle), und der Wert der Testgöße daher gleich 2.31. Man berechne dann, *wie groß (unter der Nullhypothese) die Wahrscheinlichkeit dafür ist, dass diese oder eine noch größere Abweichung von der Nullhypothese auftreten kann.* Das heißt *hier*: Wie wahrscheinlich ist es, dass $|Z| > 2.31$ wird? Ist diese Wahrscheinlichkeit kleiner als $\alpha = 5\%$, wird die Nullhypothese abgelehnt, im anderen Fall wird sie beibehalten. Im vorliegenden Fall beträgt die Wahrscheinlichkeit $2 \times 1.044\% = 2.088\%$, die Nullhypothese wird also abgelehnt. In der folgenden Abb. 5.4 sind diese Beziehungen graphisch dargestellt:

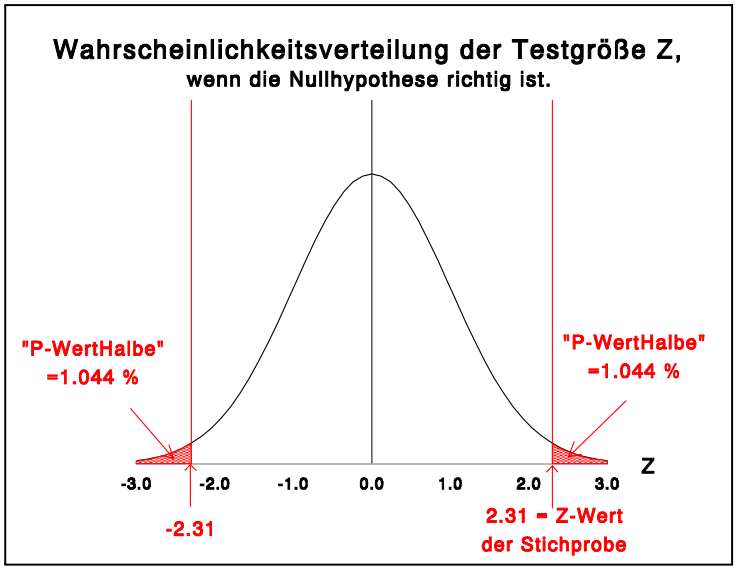


Abb. 5.4: Der P-Wert einer Stichprobe im Z–Test

Die so berechnete Wahrscheinlichkeit wird als der "*P-Wert*" des Tests oder als "*Abweichwahrscheinlichkeit*" bezeichnet. In den gängigen Statistik–Programmsystemen wird üblicherweise dieser P–Wert (häufig auch unter der Bezeichnung "Signifikanz") ausgedruckt. Dadurch wird die Kenntnis der "kritischen Werte" λ überflüssig. Die zum vorliegenden Beispiel entsprechende Tabelle sieht dann so aus:

Mittelw. $\overline{X}$	Testgröße Z	P–Wert	Kriterium:	Entscheidung:
-4.3	-3.01	**0.26 %**	$0.26\ \% \leq 5\ \%$	**Ablehnung der Nullhypothese**
-1.5	-1.05	29.37 %	$29.37\ \% > 5\ \%$	Nullhypothese nicht widerlegt
1.8	1.26	20.77 %	$20.77\ \% > 5\ \%$	Nullhypothese nicht widerlegt
2.3	1.75	8.01 %	$8.01\ \% > 5\ \%$	Nullhypothese nicht widerlegt
3.3	2.31	**2.09 %**	$2.09\ \% \leq 5\ \%$	**Ablehnung der Nullhypothese**

Über seine Funktion als Entscheidungskriterium hinaus ist der P-Wert aber auch von eigenem Interesse: Aufgrund seiner Konstruktion kann er als "Abweichwahrscheinlichkeit" interpretiert werden:

> **Der P-Wert gibt an, mit welcher Wahrscheinlichkeit bei Gültigkeit der Nullhypothese H_0 eine Abweichung von H_0 auftreten kann, die so groß ist oder noch größer als aktuell beobachtet.**

Der P-Wert ist daher die *wahrscheinlichkeitstheoretische Bewertung der beobachteten Differenz zur Nullhypothese*: Je kleiner der P-Wert, desto weniger ist die Abweichung von H_0 durch Zufall zu erklären. Da aber der P–Wert unter anderem vom Stichprobenumfang abhängt, ist er *kein Maß für die tatsächliche Abweichung von der Nullhypothese.* Für diese Abweichungen und die Bewertung ihrer (klinischen) Relevanz sind unverändert die im vorigen Kapitel beschriebenen statistischen Schätzungen mit den zugehörigen Konfidenzintervallen zuständig.

5.2.5. Einseitige Tests.

Im vorigen Beispiel sollte die Nullhypothese "keine Änderung der Blutdruckwerte zu erwarten" dann abgelehnt werden, wenn die mittlere Änderung "stark von 0 verschieden" ausfällt, also bei deutlichen Senkungen wie auch bei deutlichen *Erhöhungen* des Blutdrucks. Handelt es sich aber um die Untersuchung der Wirksamkeit eines Mittels zur *Blutdrucksenkung* (und nicht um die etwaige Nebenwirkung einer anderen Behandlung), so wird man eine Erhöhung des Mittelwertes von vornherein gar nicht in Betracht ziehen und eine Wirksamkeit der Behandlung nur dann als nachgewiesen ansehen, wenn die Änderung des Blutdrucks ("Vorher minus Nachher") *stark positiv* ist. Der Ablehnbereich liegt jetzt ganz im Bereich *positiver* Werte. Er kann dort gegenüber vorher aber noch *vergrößert* werden, damit er insgesamt wieder 5% Wahrscheinlichkeit der Testgröße ausmacht:

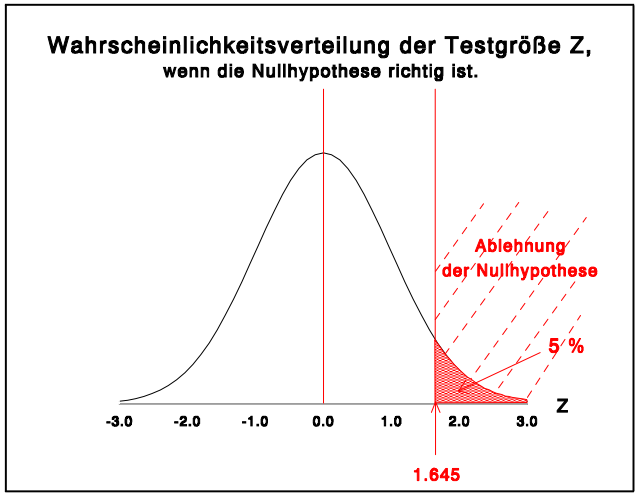


Abb. 5.5: Z–Test bei einseitiger Fragestellung

Man sieht, dass der kritische Wert jetzt nur noch 1.645 beträgt statt 1.96:

$$\text{Lehne die Nullhypothese ab, wenn } Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq 1.645$$

Dadurch wird, wenn die Behandlung tatsächlich wirksam ist, die Wahrscheinlichkeit für den Nachweis der Veränderung, also die "Power" des Tests, größer.
Im umgekehrten Fall aber gibt es überhaupt keine Chance:
Eine Abweichung von der Nullhypothese entgegen der vorausgesagten Richtung kann überhaupt nicht entdeckt werden (da sie beim einseitigen Testen gar nicht in Betracht gezogen wird).
In den folgenden Abschnitten werden die gebräuchlichsten Testverfahren beschrieben.

5.3. Spezielle Testverfahren (GK 5.2.2)

5.3.1. Mittelwertvergleiche bei normalverteilten Grundgesamtheiten

t-Test für *verbundene* Stichproben ("gepaarter t-Test")

1. **Zufallsvariable**: V und W seien Zufallsvariablen, die den Meßwert eines Parameters zu zwei verschiedenen Bedingungen an jeweils *derselben Beobachtungseinheit* festhält (Beispiel: Körpergewicht vor (V) und 8 Wochen nach einer Diät (W)). Mit $Y = W - V$ werde die Differenz der Werte bezeichnet.
   Bezeichnung: Da je zwei Werte (V und W) über dieselbe Beobachtungseinheit miteinander verbunden oder gepaart sind, spricht man von "verbundener" oder "paariger" Stichprobe.

2. **Verteilung in der Grundgesamtheit:** Es wird angenommen, dass die Differenz $Y = W - V$ normalverteilt ist:

   $$Y = W - V \sim N(\mu, \sigma^2)$$

   (Diese Annahme ist z.B. immer dann erfüllt, wenn V und W "gemeinsam" normalverteilt sind)

3. **Nullhypothese**: Es soll überprüft werden, ob V und W denselben Erwartungswert haben, ob also $E(Y) = E(W - V) = 0$ ist:

   $$H_0 : \mu = 0$$

4. **Stichprobe und Abweichungsmaß:** Gepaarte Beobachtungen $(V_1, W_1), (V_2, W_2), \ldots, (V_n, W_n)$ mit den zugehörigen Differenzen $Y_1, Y_2, \ldots, Y_n$
Abweichungsmaß: Der Mittelwert dieser Differenzen:

$$S = \frac{1}{n}\sum_{i=1}^{n}(W_i - V_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}_n$$

5. **Teststatistik**: Als Transformation in eine Zufallsvariable, deren Verteilung unter $H_0 : \mu = 0$ vollständig bekannt ist, wähle

$$\begin{aligned} T &= \frac{\bar{Y}_n}{\frac{s}{\sqrt{n}}} \\ &= \frac{\bar{Y}_n}{s}\sqrt{n} \end{aligned} \tag{5.3}$$

wobei

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{5.4}$$

die Varianz der Differenzen $Y_i = W_i - V_i$ ist.
Nach (4.5) ist die Verteilung von T bekannt, wenn Y den Erwartungswert $\mu = 0$ hat: Dann ist T nach der (zentralen) t–Verteilung mit $n-1$ Freiheitsgraden verteilt.

6. **Zweiseitiger Test:**

   - **Kritischer Wert**: Zu gegebenem Signifikanzniveau α wähle man (aus einer Tabelle der t-Verteilung) das obere $\alpha/2$–Quantil (das ist das $(1-\alpha/2)$–Quantil) der t–Verteilung mit $n-1$ Freiheitsgraden, bezeichnet als $t_{n-1,1-\alpha/2}$ (vgl. Abb. 4.2 und zugehörige Tabelle). Dann gilt:

     $$P(|T| > t_{n-1,1-\alpha/2}) = \alpha$$

     Wähle also das obere $\alpha/2$–Quantil der t–Verteilung mit $n-1$ Freiheitsgraden als kritischen Wert für die Testgröße $T = \frac{\bar{Y}_n}{s}\sqrt{n}$.
   - **Testentscheidung**: Berechne aus einer aktuell vorliegenden Stichprobe den Wert von $|T|$ und entscheide genau dann gegen $H_0 : \mu = 0$, wenn dieser Wert größer ist als $t_{n-1,1-\alpha/2}$

7. **Einseitiger Test**:

   - **Vorgabe der "Richtung"**: Beispielsweise sei für den Erwartungswert μ der Differenz $Y = W - V \sim N(\mu, \sigma^2)$ grundsätzlich $\mu \geq 0$ vorausgesetzt (rechtsseitige Fragestellung). Man geht also (aufgrund von Kenntnissen und Vorinformationen über den Versuch) davon aus, dass eine *Verringerung* des Erwartungswertes von V auf W (also $\mu < 0$) *grundsätzlich ausgeschlossen* werden kann.
   - **Kritischer Wert**: Zu gegebenem Signifikanzniveau α wähle man dann (aus einer Tabelle der t-Verteilung) das obere α–Quantil (das ist das $(1-\alpha)$–Quantil) der t–Verteilung mit $n-1$ Freiheitsgraden, bezeichnet als $t_{n-1,1-\alpha}$. Dann gilt

     $$P(T > t_{n-1,1-\alpha}) = \alpha$$

Bei "linksseitiger" Fragestellung wähle man das *untere* α–Quantil $t_{n-1,\alpha}$. Nach Definition des Quantils ist für dieses die Gleichung $P(T \leq t_{n-1,\alpha}) = \alpha$ erfüllt.

- **Testentscheidung**: *Rechtsseitige* Fragestellung: Berechne aus einer aktuell vorliegenden Stichprobe den Wert von T und entscheide genau dann gegen H_0 :, wenn dieser Wert größer ist als $t_{n-1,1-\alpha}$. Bei *linksseitiger* Fragestellung ist H_0 abzulehnen, wenn $T \leq t_{n-1,\alpha}$ ist.

Beispiel:
Es soll überprüft werden, ob nach der Awendung einer bestimmten Diät eine Gewichtsveränderung eintritt.

- $\mu =$ Erwartungswert der Gewichtsveränderung
- Nullhypothese H_0 : Keine Gewichtsveränderung: $\mu = 0$
- Alternative H_1 zweiseitig: Gewichtsveränderung: $\mu \neq 0$
  Alternative H_1 Einseitig: *Verringerung* des Gewichts: $\mu \leq 0$

Dazu werden aus der Zielpopulation (definierte Ein- und Ausschlusskriterien!) 10 Patienten zufällig ausgewählt und in die Studie aufgenommen. Bei jedem Patienten wurde das Gewicht vor und nach der Diät gemessen. Die Gewichtsveränderung kann als normalverteilt angesehen werden (die Streuung dieser Werte ist aber unbekannt). Der Test soll auf dem Signifikanzniveau $\alpha = 5\%$ durchgeführt werden.

Pat.(i)	Vorher(V_i)	Nachher(W_i)	Diff.(Y_i)	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
1	98	92	-6	-3	9
2	101	98	-3	0	0
3	88	87	-1	2	4
4	79	81	2	5	25
5	110	105	-5	-2	4
6	91	92	1	4	16
7	106	96	-10	-7	49
8	111	109	-2	1	1
9	92	94	2	5	25
10	96	88	-8	-5	25
$\sum$	972	942	-30	0	158
$\frac{1}{n}\sum$	97.2	94.2	-3		
$\frac{1}{n-1}\sum$					17.556

$$
\begin{aligned}
s &= \sqrt{17.556} = 4.1899 \\
t &= \frac{\bar{y}}{s}\sqrt{n} = \frac{-3}{4.1899}\sqrt{10} = -2.2642 \\
t_{n-1,1-\alpha/2} &= t_{9,\ 0.975} = 2.262 \quad \text{(aus Tabelle zur t-Verteilung!)}
\end{aligned}
$$

Der aus der Stichprobe berechnete Testgrößenwert $|t| = 2.2642$ ist *größer als der kritische Wert* $t_{9,\ 0.975} = 2.262$. *Die Nullhypothese* $H_0 : \mu = 0$ *wird daher abgelehnt. Die Annahme, dass nach der Diät keine Gewichtsänderung eintritt, ist damit widerlegt.*

Bei *einseitiger* Fragestellung wird hier vorab angenommen, dass die Diät –falls sie überhaupt wirkt– tendenziell zu einer Gewichts*reduktion* führt, dass also grundsätzlich $\mu \leq 0$ angenommen werden kann (linksseitige Fragestellung). In diesem Fall ist T mit dem Wert $t_{n-1,\ \alpha} = t_{9,\ 0.05} = -1.833$ zu vergleichen. Da $t = -2.2642 < -1.833 = t_{9,\ 0.05}$ ist, wird die Nullhypothese abgelehnt. Die Gewichtsreduktion zwischen erster und zweiter Messung "ist signifikant auf dem 5%–Niveau".
Man beachte: Bei dieser Studienform kann nicht geklärt werden, ob die Reduktion tatsächlich auf die Diät zurückzuführen ist. Die Veränderung könnte z.B. auch daher kommen, dass allein schon die Teilnahme an der Studie bei den Patienten eine solche Wirkung zeigt. Um in dieser Hinsicht Klarheit zu erhalten, müßte eine Studie mit einer *Kontrollgruppe* durchgeführt werden, wobei diese sich von der Diät–Gruppe weder in der Zusammensetzung noch in der Behandlung (Gewichtskontrolle etc.) unterscheidet, außer eben in der speziellen Diätanweisung.

t-Test für *unverbundene* Stichproben

1. **Zufallsvariable**: X und Y seien Zufallsvariablen, die den Meßwert eines Parameters zu zwei verschiedenen Bedingungen festhalten; dabei kommt jede Beobachtungseinheit nur *einmal*, also nur unter *einer* der beiden Bedingungen in die Stichprobe. Es entstehen also *zwei unverbundene Stichproben:* eine Stichprobe vom Umfang n_X für die Variable X und eine zweite vom Umfang n_Y für die Variable Y. Man spricht von *unabhängigen Stichproben.*

2. **Verteilung in der Grundgesamtheit:** Es wird angenommen, dass beide Zufallsvariablen, X und Y, *normalverteilt* sind und *dieselben Varianzen* haben:
$$X \sim N(\mu_X, \sigma^2)$$
$$Y \sim N(\mu_Y, \sigma^2)$$

3. **Nullhypothese**: Es soll überprüft werden, ob X und Y denselben Erwartungswert haben, ob also $E(X) = \mu_X = E(Y) = \mu_Y$ ist:
$$H_0 : \mu_X = \mu_Y$$

4. **Stichprobe und Abweichungsmaß:** Zu den beiden Bedingungen werden die Stichproben $X_1, X_2, \ldots, X_{n_X}$ und $Y_1, Y_2, \ldots, Y_{n_Y}$ gebildet. Alle Variablen $X_1, X_2, \ldots$, $Y_1, Y_2, \ldots$ sind stochastisch unabhängig voneinander. Als Abweichungsmaß wählt man zunächst den Absolutbetrag der Differenz der Mittelwerte beider Stichproben, also den Absolutbetrag von:
$$\frac{1}{n_X}\sum_{i=1}^{n_X} X_i - \frac{1}{n_Y}\sum_{j=1}^{n_Y} Y_j = \overline{X}_{n_X} - \overline{Y}_{n_Y}$$

5. **Teststatistik**: Um diese Mittelwertsdifferenz in eine Zufallsvariable zu transformieren, deren Verteilung unter $H_0 : \mu_X = \mu_Y$ vollständig bekannt ist, dividiert man sie durch den *Standardfehler der Mittelwertsdifferenz*:
$$\begin{aligned} T &= \frac{\overline{X}_{n_X} - \overline{Y}_{n_Y}}{s_{(\overline{X}_{n_X} - \overline{Y}_{n_Y})}} \\ &= \frac{\overline{X}_{n_X} - \overline{Y}_{n_Y}}{\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}}} \end{aligned} \tag{5.5}$$

wobei

$$s_X^2 = \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(X_i - \overline{X})^2 \qquad \text{und} \qquad s_Y^2 = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2 \qquad (5.6)$$

die Varianzen in den beiden Stichproben bezeichnen.
Unter der Nullhypothese, d.h. wenn $\mu_X - \mu_Y = 0$ ist, ist auch hier wieder T nach der (zentralen) t–Verteilung verteilt. Die Anzahl der Freiheitsgrade ist aber $n-2$ wenn $n = n_X + n_Y$ den Gesamt-Stichprobenumfang bezeichnet.

6. **Zweiseitiger Test**

   - **Kritischer Wert**: Zu gegebenem Signifikanzniveau α wähle man (aus einer Tabelle der t-Verteilung) das obere $\alpha/2$–Quantil (das ist das $(1-\alpha/2)$–Quantil) der t–Verteilung mit $n-2$ Freiheitsgraden, bezeichnet als $t_{n-2,\ 1-\alpha/2}$ (vgl. Abb. 4.2 und zugehörige Tabelle). Dann gilt:

     $$P(|T| > t_{n-2,\ 1-\alpha/2}) = \alpha$$

     Wähle also das obere $\alpha/2$–Quantil der t–Verteilung mit $n-2$ Freiheitsgraden als kritischen Wert für die Testgröße $T = \frac{\overline{X}_{n_X} - \overline{Y}_{n_Y}}{s_{(\overline{X}_{n_X} - \overline{Y}_{n_Y})}}$.
   - **Testentscheidung**: Berechne aus einer aktuell vorliegenden Stichprobe den Wert von $|T|$ und entscheide genau dann gegen $H_0 : \mu_X = \mu_Y$, wenn dieser Wert größer ist als $t_{n-2,\ 1-\alpha/2}$

7. **Einseitiger Test**:

   - **Vorgabe der "Richtung"**: Beispielsweise sei grundsätzlich $\mu_X - \mu_Y \geq 0$ angenommen (rechtsseitige Fragestellung). Man geht also davon aus, dass eine *Verringerung* des Erwartungswertes in Gruppe 1 gegenüber Gruppe 2 *grundsätzlich ausgeschlossen* werden kann. (Bei linksseitiger Fragestellung wird $\mu_X - \mu_Y \leq 0$ vorausgesetzt).
   - **Kritischer Wert**: Zu gegebenem Signifikanzniveau α wähle man dann (aus einer Tabelle der t-Verteilung) das obere α–Quantil der t–Verteilung mit $n-2$ Freiheitsgraden, bezeichnet als $t_{n-2,\ 1-\alpha}$. Dann gilt

     $$P(T > t_{n-2,\ 1-\alpha}) = \alpha$$

     Bei "linksseitiger" Fragestellung wähle man das *untere* α–Quantil $t_{n-2\ ,\alpha}$. Nach Definition des Quantils ist für dieses die Gleichung $P(T \leq t_{n-2,\ \alpha}) = \alpha$ erfüllt.
   - **Testentscheidung**: *Rechtsseitige* Fragestellung: Berechne aus einer aktuell vorliegenden Stichprobe den Wert von T und entscheide genau dann gegen $H_0 : \mu = 0$, wenn dieser Wert größer ist als $t_{n-2,\ 1\ -\alpha}$. Bei *linksseitiger* Fragestellung ist H_0 abzulehnen, wenn $T \leq t_{n-2,\ \alpha}$ ist.

Rechenbeispiel:

Gruppe 1			
Pat.(i)	X_i	$X_i - \overline{X}$	$(X_i - \overline{X})^2$
1	30	-1	1
2	27	-4	16
3	25	-6	36
4	36	5	25
5	33	2	4
6	29	-2	4
7	35	4	16
8	34	3	9
9	26	-5	25
10	35	4	16
$\sum$	310		152
$\frac{1}{n}\sum$	31		
$\frac{1}{n-1}\sum$			16.889

Gruppe 2			
Pat.(i)	Y_i	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
1	40	4	16
2	37	1	1
3	29	-7	49
4	31	-5	25
5	41	5	25
6	44	8	64
7	31	-5	25
8	35	-1	1
$\sum$	288		206
$\frac{1}{n}\sum$	36		
$\frac{1}{n-1}\sum$			29.429

$$
\begin{aligned}
n_X &= 10, \quad n_Y = 8, \quad n_X + n_Y - 2 = 16 \\
s_x^2 &= \frac{1}{(n_X - 1)} \sum_{i=1}^{n_X} (x_i - \overline{x})^2 = \frac{152}{9} = 16.889 \\
s_y^2 &= \frac{1}{(n_Y - 1)} \sum_{i=1}^{n_X} (y_i - \overline{y})^2 = \frac{206}{7} = 29.429 \\
\overline{X}_{n_X} - \overline{Y}_{n_Y} &= 31 - 36 = -5 \\
t &= \frac{\overline{x}_{n_X} - \overline{y}_{n_Y}}{\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right) \frac{(n_X - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}} \\
&= \frac{-5}{\sqrt{\left(\frac{1}{10} + \frac{1}{8}\right) \frac{9 \times 16.889 + 7 \times 29.429}{16}}} \\
&= -2.2284
\end{aligned}
$$

Für $\alpha = 0.05$ ist das obere $\alpha/2$–Quantil der t–Verteilung mit 16 Freiheitsgraden gleich $t_{16,0.975} = 2.1199$.
Folgerung: Da der aktuell beobachtete Wert t der Teststatistik T dem Betrage nach den kritischen Wert überschreitet:

$$|t| = 2.2284 \quad \geq \quad t_{16,\ 0.975} = \ 2.1199,$$

wird die Nullhypothese abgelehnt: Die Annahme gleicher Erwartungswerte in den beiden Grundgesamtheiten gilt als widerlegt. Die Unterschiede sind "*signifikant*", "*statistisch gesichert auf dem* 5 *%–Niveau*". Der "P–Wert" zu diesen Daten ist "$P = 0.0405$", d.h.

$$P_{H_0}\left(|T| \geq 2.2284\right) = 0.0405 :$$

Die Wahrscheinlichkeit dafür, dass die Testgröße T (dem Betrage nach) einen Wert so groß wie aktuell beobachtet oder noch größer annimmt, ist *unter der Nullhypothese* gleich 4.05 %.
Auswertung für *einseitige* Fragestellung:

Für $\alpha = 0.05$ ist das obere α–Quantil der t–Verteilung mit 16 Freiheitsgraden gleich $t_{16,\ 0.95} = 1.746$.
Bei *rechtsseitiger* Fragestellung wird die Nullhypothese *nicht abgelehnt* (denn dafür müßte das Vorzeichen von t positiv sein. Es ist aber $t = -2.2284$).
Bei *linksseitiger* Fragestellung wird die Nullhypothese *abgelehnt*, da $t = -2.2284 \leq -1.746 = t_{16,\ 0.05}$.

> Hinweis: In diesem Beispiel sind die Stichprobenvarianzen s_x^2 und s_y^2 gleich 16.889 bzw. 29.429. Es erscheint deshalb zumindest fraglich, ob die unter Punkt 2 gemachte Annahme, wonach die Varianzen in der Grundgesamtheit identisch (gleich einem unbekannten Wert σ^2) sind, gerechtfertigt erscheint. Ist diese Annahme in Wirklichkeit nicht erfüllt, so stimmt auch die Folgerung nicht, dass die Verteilung der Testgröße t , definiert in (5.5), die t–Verteilung ist mit $n_X + n_Y - 2$ Freiheitsgraden. In diesem Fall ist ein (nach WELCH) modifizierter t–Test anzuwenden, bei dem sowohl die Testgröße als auch deren Verteilung unter der Nullhypothese neu berechnet werden muß. Dieser "t-Test bei ungleichen Varianzen" ist in guten Statistikprogrammen neben dem "üblichen" t–Test verfügbar.

5.3.2. Kontingenztafel–Analyse.

Unabhängigkeit in Kreuztabellen In der deskriptiven Statistik haben wir bei der Analyse von Kreuztabellen ("Kontingenztafeln") bereits ein Abweichungsmaß kennengelernt, ohne damit schon so recht was anfangen zu können:
Der dort mit X^2 bezeichnete Wert

$$X^2 = \sum \frac{(\text{Beobachtete Anzahl - Erwartete Anzahl})^2}{\text{Erwartete Anzahl}}$$

(S. 28) war so konstruiert, dass er anzeigt, *wie stark die beobachteten Häufigkeiten in einer Kreuztabelle von der Annahme der Unabhängigkeit der beiden Merkmale abweichen.* Mit Hilfe der Wahrscheinlichkeitsrechnung ist es nun möglich zu berechnen, welche *Werte X^2 mit welcher Wahrscheinlichkeit annehmen kann, wenn die beiden Merkmale der Kreuztabelle tatsächlich unabhänging sind.* Damit können wir also schon wieder einen Test konstruieren:

1. **Zufallsvariable**: X und Y seien diskrete Zufallsvariable. X und Y sind jeweils an derselben Beobachtungseinheit zu messen (gepaarte Beobachtung).

2. **Verteilung in der Grundgesamtheit:** Die gemeinsame Verteilung von X und Y wird mit
$$p_{ij} = P(X = i, Y = j) \tag{5.7}$$
bezeichnet.

3. **Nullhypothese**: Es soll überprüft werden, ob X und Y unabhängig sind. Nach der Definition der Unabhängigkeit lautet die Nullhypothese also
$$H_0 : \ P(X = i, Y = j) \ = \ P(X = i)\ P(Y = j) \tag{5.8}$$

4. **Stichprobe und Abweichungsmaß:** Aus n Beobachtungspaaren $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ bildet man –wie in der deskriptiven Statistik– die Häufigkeitstabelle für alle Kombinationen und erhält somit eine Kontingenztafel. Die Randhäufigkeiten sind die Zeilen– und Spaltensummen.

5. Wenn X und Y *unabhängig* sind, ist für jede Zelle der Kreuztabelle "erwartete Häufigkeit" gleich

$$Erwartete\ Anzahl\ = \frac{Zeilensumme \times Spaltensumme}{Gesamtzahl}$$

6. **Teststatistik**: Bilde die Summe der quadrierten und standardisierten Differenzen zwischen beobachteten und erwarteten Häufigkeiten:

$$X^2 = \sum \frac{(\text{Beobachtete Anzahl - Erwartete Anzahl})^2}{\text{Erwartete Anzahl}}$$

Unter der Nullhypothese, d.h. bei Unabhängigkeit von X und Y, ist die Verteilung dieser Testgröße angenähert gleich der sogenannten "χ^2*–Verteilung*".
Hinweis 1: Wie für die t-Verteilung gibt es auch für die χ^2*–Verteilung* Versionen mit unterschiedlichen *"Freiheitsgraden"*. Die Anzahl der Freiheitsgrade wird hier aber nicht durch den Stichprobenumfang festgelegt sondern durch die *Anzahl der Zeilen und Spalten der Kreuztabelle.* Sie kennzeichnen die Anzahl der Felder, deren Werte man *frei* wählen kann, wenn die Randsummen als fest gegeben angesehen werden. Für die Vierfeldertafel kann nur 1 Feld frei gewählt werden; die anderen ergeben sich daraus als Ergänzung zur jeweiligen Randsumme. Allgemeine Formel:

$$Freiheitsgrade = (Zeilen - 1) \times (Spalten - 1)$$

Hinweis 2: Die χ^2–Verteilung entsteht durch Quadrierung einer normalverteilten Zufallsvariablen: ist $X \sim N(0,1)$ standardnormalverteilt, so wird die Verteilung von X^2 als die "χ^2*–Verteilung mit einem Freiheitsgrad*" bezeichnet. Sie ist also aus der Standard-Normalverteilung berechenbar. Ebenso die χ^2–Verteilung mit (allgemein) k Freiheitsgraden: Sind $X_1, X_2, \ldots X_k$ unabhängige, standard–normalverteilte Zufallsvariablen, so heißt die Verteilung von $X^2 := X_1^2 + X_2^2 + \ldots + X_k^2$ die "χ^2*–Verteilung mit k Freiheitsgraden*".

7. **Kritischer Wert**: Zu gegebenem Signifikanzniveau α wähle man (aus einer Tabelle der χ^2-Verteilung) das obere α–Quantil (das ist das $(1-\alpha)$–Quantil) der χ^2-Verteilung mit f Freiheitsgraden, bezeichnet als $\chi^2_{f,1-\alpha}$. Nach der allgemeinen Definition eines Quantils gilt dann

$$\begin{aligned} P(X^2 > \chi^2_{f,1-\alpha}) &= 1 - P(X^2 \leq \chi^2_{f,1-\alpha}) \\ &= 1 - (1-\alpha) \\ &= \alpha \end{aligned}$$

Wähle also das obere α–Quantil der χ^2-Verteilung mit $f := (Zeilen - 1) \times (Spalten - 1)$ Freiheitsgraden als kritischen Wert für die Testgröße X^2.

8. **Testentscheidung**: Berechne aus einer aktuell vorliegenden Stichprobe den Wert der Teststatistik X^2 und entscheide genau dann gegen die Nulhypothese H_0 der Unabhängigkeit, wenn dieser Wert größer ist als $\chi^2_{f,1-\alpha}$

Der (4-Felder) χ^2–Test Dies ist der Spezialfall einer Kontingenztafelanalyse für jeweils nur 2 mögliche Ergebnisse der Zufallsvariablen X und Y (X und Y sind "*binäre*" Zufallsvariable). Wie in Kapitel 2 beschrieben, wählt man dann häufig die Buchstaben a, b, c und d zur Bezeichnung der absoluten Häufigkeiten.

Die Testgröße X^2 kann in diesem Fall einfacher nach der Formel

$$X^2 = \frac{(ad - bc)^2 n}{(a + c)(b + d)(a + b)(c + d)}$$

berechnet werden. X^2 ist angenähert χ^2–verteilt mit einem Freiheitsgrad. Die oberen α–Quantile $\chi^2_{1,1-\alpha}$ und damit die kritischen Werte für die Teststatistik X^2 für die üblichen Signifikanzniveaus α sind:

α	0.05	0.01	0.001
$\chi^2_{1,1-\alpha}$	3.8414	6.6349	10.8276

Rechenbeispiel (aus Kapitel 2):

Das "Auftreten von mindestens einer Nebenwirkung" und "die verabreichte Dosis" sind die beiden Variablen, deren Zusammenhang überprüft werden soll. Die in der Studie beobachteten Häufigkeiten waren

	Ergebnis		
Gruppe	"negativ"	"positiv"	Total
A	a=32	b=152	184
B	c=35	d=140	175
Total	67	292	$n = 359$

Berechnung der Testgröße:

$$\begin{aligned} X^2 &= \frac{(32 \times 140 - 152 \times 35)^2 \times 359}{67 \times 292 \times 184 \times 175} \\ &= 0.4021 \end{aligned}$$

Für $\alpha = 0.05$ ist der kritische Wert gleich 3.8414. Da der aktuelle Wert der Testgröße, $X^2 = 0.4021$, kleiner ist als der kritische Wert, wird die Nullhypothese *nicht* abgelehnt: Ein Zusammenhang zwischen der Dosis und dem Eintreten von Nebenwirkungen ist durch diese Ergebnisse *nicht* nachgewiesen (die beobachteten Unterschiede können – auch bei Unabhängigkeit von Dosis und Nebenwirkung– allein durch Zufallsschwankungen erklärt werden). Dies wird in der folgenden Abbildung 5.6 noch einmal graphisch veranschaulicht:

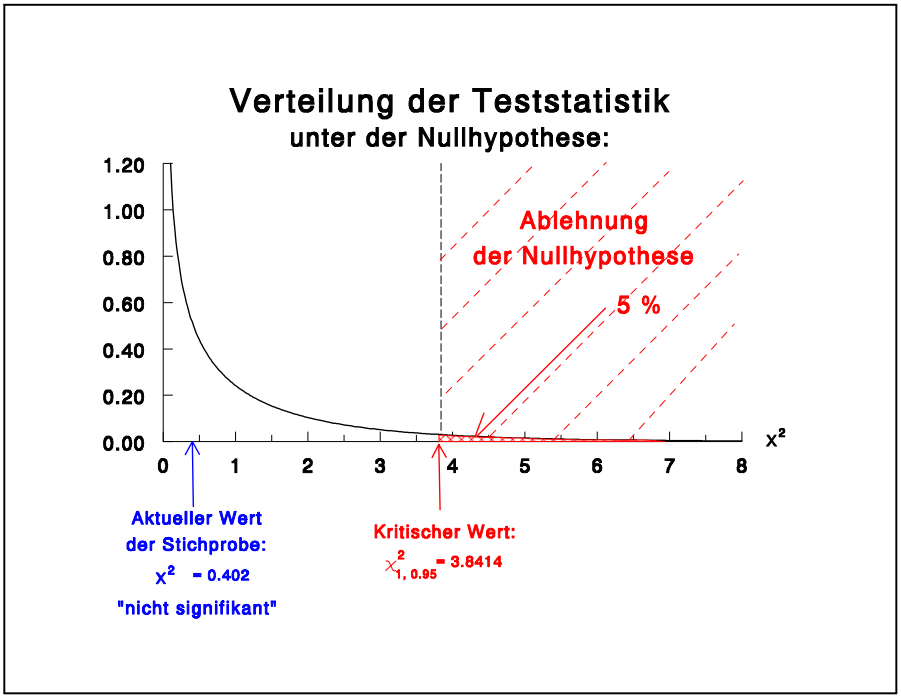


Abb. 5.6: Der 4–Felder– χ^2–Test

Zur Anwendbarkeit des χ^2–Tests
Wie bereits angedeutet, ist die für Kontingenztafeln definierte Testgröße X^2 *nur angenähert nach der χ^2-Verteilung verteilt.* Das entscheidende Kriterium für die Anwendbarkeit des Tests sind die *Erwartungshäufigkeiten* der einzelnen Zellen der Kreuztabelle (vgl. Seite 28):

$$\begin{aligned} E_{ij} &= \text{ErwartungsHäufigkeit der Zelle mit Zeile Nr. i und Spalte Nr. j} \\ &= \frac{\text{Zeilensumme} \times \text{Spaltensumme}}{\text{Gesamtzahl der Fälle}} \end{aligned}$$

Sind alle diese Erwartungshäufigkeiten ≥ 5, so kann man den χ^2-Test bedenkenlos anwenden. Die Angaben darüber, unter welchen Voraussetzungen der χ^2-Test auch sonst noch angewendet werden darf, sind in der Literatur nicht einheitlich. Am häufigsten wird die Regel von Cochran (1954) herangezogen:

- Kreuztabellen mit mehr als 2 Zeilen oder Spalten: "If relatively few expectations are less than 5 (say in 1 cell out of 5 or more, or 2 cells out of 10 or more), a minimum expectation of 1 is allowable in computing χ^2 "
- Die 2x2-Kreuztabelle: "Use Fisher's exact test (i) if the total n of the table < 20, (ii) if $20 < n < 40$ and the smallest expectation is less than 5.
  If $n > 40$ use χ^2, corrected for continuity."

Man entnimmt dieser Regel, dass für die 2x2-Tabelle zwei weitere Tests in Frage kommen:

- Der χ^2-Test mit Stetigkeitskorrektur. Diese Korrektur (häufig auch mit "YATES-Korrektur" bezeichnet), verbessert die Approximation der Testgröße an die χ^2-Verteilung. Man sollte überprüfen, ob das Statistikprogramm, mit welchem man

arbeitet, entsprechend der Regel von Cochran grundsätzlich diese Korrektur anwendet. Sie ist nämlich ziemlich restriktiv und z.B. bei ErwartungsHäufigkeiten alle ≥ 5 sicher nicht nötig. Es gibt auch einige Autoren, die die Anwendung dieser Korrektur generell ablehnen. Numerische Untersuchungen zeigen, dass die Anwendung der Yates-Korrektur fast die gleichen Ergebnisse liefert wie der "exakte Test von Fisher".

Statistikprogrammpakete wie z.B. SPSS geben die minimale Erwartungshäufigkeit der Zellen sowie den Prozentsatz von Zellen mit einer Erwartungshäufigkeit < 5 aus, so dass man mit Hilfe dieser Angaben die Cochran-Regel anwenden kann. Bei SPSS findet man für 2×2–Tabellen ebenfalls die Yates–Korrektur berechnet sowie den exakten Test nach Fisher. Bei diesem werden die Randsummen der Kreuztabelle als gegeben angesehen und auf dieser Basis unter der Nullhypothese die Wahrscheinlichkeiten für die verschiedenen Möglichkeiten berechnet, die Beobachtungen auf die 4 Zellen der Kreuztabelle zu verteilen. Die Wahrscheinlichkeiten für die beobachtete und die noch "extremeren" solcher Belegungen bilden dann den P–Wert des Tests.

5.3.3. Weitere spezielle Tests und Hinweise für Doktoranden

Mit den hier dargestellten statistischen Tests haben Sie ganz wichtige, aber nur sehr wenige Tests kennengelernt. Das Skript war eher darauf angelegt, statistische Denk- und Schlussweisen darzustellen und diese an Hand einiger spezieller Tests zu verdeutlichen. Dazu war es nötig, sich mit den Grundlagen der Wahrscheinlichkeitsrechnung vertraut zu machen, denn alle "Signifikanzaussagen", die in jeder empirischen Dissertation erwartet werden, sind Aussagen über Wahrscheinlichkeiten und gehen von Annahmen über Verteilungen einer Variablen "in der Grundgesamtheit" aus. Speziell bei Signifikanztests werden Verteilungen einer Testgröße "unter der Annahme der Nullhypothese" zu Grunde gelegt und mit den aktuell gefundenen Werten einer Untersuchung verglichen. Dieses Prinzip wird unverändert auf viele andere Situationen übertragen. Wenn aber die *Schlussweise* des statistischen Testens deutlich geworden ist, macht es nur noch wenig Mühe, auch weitere statistische Tests, die in entsprechenden Statistik-Programmsystemen angeboten werden, richtig anzuwenden. Es wird dann nicht mehr nötig sein, die jeweils dahinterstehenden Formeln zu kennen und nachzuvollziehen. Wichtiger ist dann vielmehr, die für die jeweilige Fragestellung interessierende *Nullhypothese zu formulieren* und einen zur statistischen Prüfung *geeigneten Test auszuwählen.* Ein kurzer Leitfaden hierzu ist auf der homepage des Instituts für Biometrie unter der Adresse

http://www.mh-hannover.de/institut/biometrie/Scripte/Tests/swtest7.pdf

abgelegt. Beigefügt ist eine ZIP-Datei mit einer SPSS-Beispiel-Datei und Syntax-Dateien, in der alle angeführten Tests enthalten und durch Markieren und Mouse-Klick aufzurufen sind. Wenn Sie sich dies lieber erst mal vorführen lassen wollen, besuchen Sie einfach den "Vertiefungskurs SPSS", der in jedem Semester angeboten wird. Nachsehen unter:

http://www.mh-hannover.de/institute/biometrie/veranstaltung.html

Literatur zum Kurs "Biomathematik":

E. Walter: *Biomathematik für Mediziner* Teubner, 1980

V. Harms: *Biomathematik, Statistik und Dokumentation* Harms-Verlag, Kiel, 6. Auflage, 1992

A.Heinecke, E.Hultsch, R. Repges: *Medizinische Biometrie* Springer, 1992

J.Adam: *Einführung in die medizinische Biometrie* Gustav Fischer UTB, 1992

D.Renner: *GK2. Medizinische Biometrie (mit 88 Seiten Kurzlehrbuch)* Chapman & Hall, 1995 ("Schwarze Reihe")

Literatur zur Medizinischen Statistik:

J.Werner: *Biomathematik und medizinische Statistik: eine praktische Anleitung für Studierende, Doktoranden, Ärzte und Biologen* 2. Auflage, München, Urban und Schwarzenberg, 1992

I.Guggenmoos-Holzmann, K.-D- Wernecke: *Medizinische Statistik* Blackwell Wissenschaftsverlag, Berlin-Wien, 1996

H.J.Trampisch, J. Windeler (Hrsg.), B.Ehle, S.Lange: *Medizinische Statistik* Springer, 1997

M.Bland: *An Introduction to Medical Statistics* Oxford University Press, 1995

R.Beaglehole, R.Bonita, T.Kjellström: *Basic Epidemiology* World Health Organization, Genf, 1993

INTERNET:

Achim Heinecke und Wolfgang Köpcke: *JUMBO - Java-unterstützte Münsteraner Biometrie-Oberfläche*
http://medweb.uni-muenster.de/institute/imib/lehre/skripte/biomathe/jumbo.html

H.-P. Altenburg, unter Mitarbeit von S. Büttner, M. Reichenbach und Th. Floren: *Grundlegende Begriffe und Verfahren der medizinischen Biometrie*
http://www.biom.uni-heidelberg.de/bioskrip/biomscri.html

Uwe Feldmann, Stefan Gräber, Heiko Buchinger, Michael Noll-Hussong und Axel Quattländer: *Tutorsystem Medizinische Informationsverarbeitung*
http://www.med-rz.uni-sb.de/med_fak/imbei/projekt/index.html