

Skript mit statistischen Grundlagen

Unterlagen zur Veranstaltung

Version 7.3

Dipl.-Handelslehrer Michael Sauer
sauermichael@gmx.net
FON: 06436-2850773
Mobil: 0176-78468313

Inhaltsverzeichnis

VORWORT	1
1 EINFÜHRUNG	2
1.1 Motivation.....	2
1.2 Fachbegriffe	3
1.3 Skalenniveaus und Merkmalsarten	5
1.4 Aufgaben.....	6
1.5 Lösungen.....	7
2 DATENARBEIT	8
2.1 Die erste Datenerfassung	8
2.2 Lösung.....	10
2.3 Übung zur Datenarbeit I.....	11
2.4 Lösung.....	12
2.5 Übung zur Datenarbeit II	13
2.6 Lösung.....	14
2.7 Quantile.....	15
2.8 Boxplots	17
2.9 Histogramme	19
2.10 Übungen.....	20
2.11 Lösungen.....	21
3 KENNZAHLEN EINES MERKMALS	23
3.1 Lagemaße	23
3.1.1 Erklärung	23
3.1.2 Übungen	25
3.1.3 Lösungen	27
3.2 Streuungsmaße	29
3.2.1 Erklärung	29
3.2.2 Übung.....	32
3.2.3 Lösung.....	33
3.2.4 Varianzverschiebungssatz („Satz von Steiner“).....	34
3.2.5 Übungen	35
3.2.6 Lösungen	36
4 KENNZAHLEN KLASSIERTER DATEN	38
4.1 Quantile.....	38
4.2 Lagemaße	39
4.2.1 Modus.....	39
4.2.2 Median.....	39
4.2.3 Arithmetisches Mittel.....	39
4.3 Streuungsmaße	39
4.3.1 Spannweite	39
4.3.2 Interquartilsabstand	39
4.3.3 Mittlere absolute Abweichung	39
4.3.4 Varianz und Standardabweichung.....	39
4.4 Übungen.....	40
4.5 Lösungen.....	41
5 ZUSAMMENHÄNGE ZWEIER MERKMALE	42
5.1 Zusammenhang zweier nominaler Merkmale.....	42
5.1.1 Beispiel	42

5.1.2	Auswertung	43
5.1.3	Übung	43
5.1.4	Lösung	43
5.2	Zusammenhang zweier metrischer Merkmale	43
5.2.1	Kovarianz	43
5.2.2	Korrelationskoeffizient nach Bravais-Pearson	44
5.3	Zusammenhang zweier ordinaler Merkmale	46
5.4	Aufgaben	48
5.5	Lösungen	49
6	REGRESSION	51
6.1	Einführungsbeispiel	51
6.2	Lösung	54
6.3	Kurzübungen zur Selbstkontrolle	55
6.3.1	Übung 1	55
6.3.2	Übung 2	55
6.4	Lösungen	56
6.4.1	Übung 1*	56
6.4.2	Übung 2	56
6.5	Übungen	57
6.6	Lösungen	58
6.7	Weitere Übung (bei Unsicherheit)	61
6.7.1	Aufgabe	61
6.7.2	Aufgabe	61
6.8	Lösungen	62
6.8.1	Aufgabe	62
6.8.2	Aufgabe 2	62
7	GRUNDLAGEN DER WAHRSCHEINLICHKEITSRECHNUNG	63
7.1	Empirisches Gesetz der großen Zahlen	63
7.2	Definition Wahrscheinlichkeit	63
7.3	Die Axiome von Kolmogorov	64
7.3.1	Grundlagen	64
7.3.2	Übungen	66
7.3.3	Lösung	68
7.4	4-Felder-Tafeln	69
7.4.1	Erklärung	69
7.4.2	Übungen	70
7.4.3	Lösungen	71
7.5	Vertiefung: Bedingte und totale Wahrscheinlichkeiten	73
7.5.1	Erklärung	73
7.5.2	Übungen und Lösungen	74
7.5.3	Totale Wahrscheinlichkeit	77
7.6	Stochastische Unabhängigkeit	78
7.6.1	Erklärung	78
7.6.2	Übungen	79
7.6.3	Lösungen	80
7.7	Abschlussaufgabe	82
7.8	Lösung	82
7.9	Vierfeldertafel vs. Baumdiagramm	83
8	ERWARTUNGSWERT UND VARIANZ	84
8.1	Erklärung	84

8.2	Übungen.....	85
8.3	Lösungen.....	86
8.4	Ausblick: Market for lemons	87
9	DISKRETE WAHRSCHEINLICHKEITSVERTEILUNGEN.....	89
9.1	Zufallsvariablen	89
9.1.1	Erklärung	89
9.1.2	Übung: Formulierung von Ereignissen bei ZV	89
9.1.3	Lösung.....	90
9.2	Übersicht über Verteilungen	91
9.3	Binomialverteilung	92
9.3.1	Definition	92
9.3.2	Beispielaufgabe	94
9.3.3	Lösung.....	95
9.3.4	Übungen	96
9.3.5	Lösungen	99
9.4	Hypergeometrische Verteilung	103
9.4.1	Definition	103
9.4.2	Beispiel	103
9.4.3	Übungen	104
9.4.4	Lösungen	105
9.5	Poissonverteilung	108
9.5.1	Definition	108
9.5.2	Beispiel	108
9.5.3	Übungen	109
9.5.4	Lösungen	111
9.6	Vermischte Übung I: Verteilungen	113
9.7	Lösung.....	114
9.8	Vermischte Übung II: Verteilungen.....	115
9.9	Lösung.....	116
9.10	Verteilungen in der Qualitätskontrolle	117
10	NORMALVERTEILUNG	119
10.1	Motivation.....	119
10.2	Beispiel.....	122
10.3	Lösung des Beispiels.....	122
10.4	Annäherung diskreter Verteilungen	123
10.5	Übungen.....	125
10.6	Lösungen.....	126
11	SCHÄTZEN – KONFIDENZINTERVALLE	128
11.1	Motivation.....	128
11.2	Beispiel.....	129
11.3	Konfidenzintervall bei unbekannter Varianz	130
11.3.1	Erklärung	130
11.3.2	Beispiel	131
11.4	Übungen.....	133
11.5	Lösungen.....	134
11.6	Schätzen eines Anteilswerts	135
11.7	Übungen zu Anteilswerten.....	136
11.8	Lösungen.....	136
12	TESTEN VON HYPOTHESEN	137

12.1	Test auf Erwartungswert	137
12.2	Übungen.....	139
12.3	Lösungen.....	140
LITERATURVERZEICHNIS		142

Tabellenverzeichnis

Tabelle 1:	Quantile der Standardnormalverteilung.	121
Tabelle 2:	Quantile der t-Verteilung.....	132

Vorwort

Willkommen in meiner Veranstaltung!

Ich freue mich, Ihnen Statistik und Wahrscheinlichkeitsrechnung näher bringen zu können. Einige organisatorische Dinge sollten vorab vereinbart werden:

1. Ziel der Veranstaltung

Ziel ist nicht, dass jemand die Klausur nicht besteht, dass „gesiebt“ wird oder Ähnliches. Ziel ist vielmehr, Sie für die Arbeit mit größeren Datenmengen mit ein paar wichtigen Instrumenten vertraut zu machen. Im Idealfall (jeder versteht und ist geübt mit den Aufgaben) sollte dann auch jeder die Klausur mit einer guten oder sehr guten Note abschließen.

2. Organisation

Zur Erreichung des Ziels gliedert sich die Veranstaltung in Vorlesungen und Übungen. In den Vorlesungen stelle ich Ihnen die wichtigsten Inhalte vor (wobei Sie Fragen stellen können – und auch stellen sollen!), in der Übung werden Beispiele gerechnet, d. h. Sie bekommen die Zeit, selbst zu rechnen und anschließend vergleichen und besprechen wir die Ergebnisse und evtl. aufgetretene Probleme. Ggf. müssen auch Aufgaben zuhause gerechnet werden, um in den Übungsstunden mehr Zeit für Fragen zu haben. Sie erhalten die Lösungen aller Übungen aber in gedruckter Form, damit Sie, wenn Sie eine Aufgabe nicht lösen konnten, nicht abschreiben müssen und mehr Zeit zum Denken haben. Manchmal können eigene Notizen bei den Lösungen aber sinnvoll ein, um Konzepte besser zu verinnerlichen!

3. Literatur

Ich habe Ihnen alle wichtigen Inhalte in diesem Skript zusammengefasst; Sie finden aber auch am Ende im Literaturverzeichnis Bücher, in denen Sie alles noch einmal selbst nachlesen können.

4. Weiteres

Ich bin 39 Jahre alt, arbeite als Lehrer (BWL/VWL, Mathematik, Informatik) an einem beruflichen Gymnasium und war als Dozent 3 Jahre bei der FOM in Frankfurt/Main und bin nun seit 2012 an der Frankfurt University of Applied Sciences und seit 2018 bei StudiumPlus der Technischen Hochschule Mittelhessen tätig. Ich bin durchaus noch lernfähig, also: „Nur wer meckert, kann verbessern!“

1 Einführung

1.1 Motivation

Die „International Federation of Football History & Statistics“ (IFFHS) bestimmt jedes Jahr den erfolgreichsten Fußballclub der Welt.

Jedes Jahr findet sich hier eine andere Aufstellung.

Wie kann das sein?

Wie kann denn ständig ein anderer Club bester Club der Welt sein?

Informationen gibt es auf einer separaten Seite der Website zur Berechnung der Plätze:

„Die Club-Weltrangliste berücksichtigt stets nur die Resultate der letzten 12 Monate von den (inter-) kontinentalen Wettbewerben, der nationalen Liga-Meisterschaft (einschließlich Play-offs) und dem nationalen Pokalwettbewerb.“¹

Aha!

Klar – Wenn die Kriterien irgendwie gelegt werden können...

„Statistik ist für Politiker häufig das, was für Betrunkene die Laterne ist. Sie dient nicht zur Erleuchtung, aber man klammert sich daran fest.“ (Herbert Salcher (*1929))

Gilt das nicht auch für Nicht-Politiker: Statistik ist dann...

...einzig und allein dafür notwendig, um seinen eigenen (schwankenden) Standpunkt zu stützen.

Wie Statistik wirklich funktioniert, was man mit Statistik alles machen kann, das sollte man sich mal genauer anschauen...

¹ vgl. Website der IFFHS, <https://www.iffhs.com/>, Abrufdatum: 25.08.2026.

1.2 Fachbegriffe

Wir benötigen ein paar Fachbegriffe für die weitere Arbeit

Hierzu wieder unser Fall mit dem WM-Bus der Fußball-Weltmeisterschaft:

In einem Bus vom Hauptbahnhof zum Stadion gab es 50 Plätze.

30 wurden von Deutschen besetzt, 10 von Franzosen, 4 von Briten, 5 von US-Amerikanern und auch ein begeisterter Japaner fuhr mit.

Ein Schaffner kontrolliert nun bei zwei Personen die Fahrkarten. Es interessiert ihn nur, ob eine Person eine Fahrkarte hat oder nicht.

Wir schreiben nun einfach einmal folgendes:

H = Eine Fahrkarte haben

K = Keine Fahrkarte haben

Grundgesamtheit

Alle Objekte, über die man etwas wissen möchte (alle Personen im Zug)

Stichprobe

Auswahl aus der Grundgesamtheit (zwei Personen).

Urliste

Liste mit allen Ergebnissen der Befragung.

Merkmal

Eigenschaft, die untersucht bzw. betrachtet wird. Hier: Merkmal „Besitz einer Fahrkarte“.

Wertebereich

Werte, die das Merkmal annehmen kann: H oder K (s. o.).

Merkmalsträger/Erhebungseinheit

Merkmalsträger ist ein *untersuchtes* Objekt (also ein Element der Stichprobe), das Merkmale hat. Hier: Die kontrollierte Person.

Merkmalsausprägung

Konkreter Wert aus dem Wertebereich, der bei einem Merkmalsträger vorliegt.

Hier: Die kontrollierte Person hat eine Fahrkarte.

Skalenniveau

Nominal: Nur Wert für sich wichtig

Ordinal: Reihenfolgen können gebildet werden.

Metrisch: Intervallskaliert: Abstände können interpretiert werden („3 mehr als“).

Der Abstand von 20°C bis 40°C ist genau so groß wie der von 50 °C bis 70 °C.

Verhältnisskala: Verhältnisse können interpretiert werden („doppelt so“).

40 kg ist doppelt so schwer wie 20 kg.

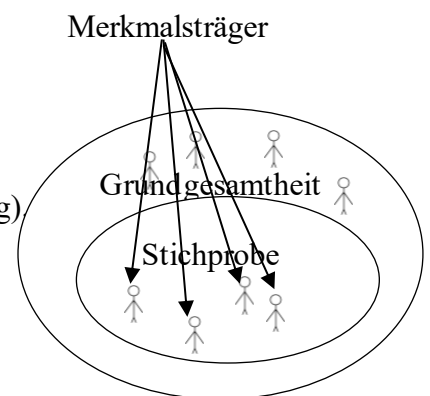
Merkmalsart

Qualitativ: Nicht-messbare Daten (nominal oder ordinal)

Quantitativ: Messbare Daten (metrisch)

Diskrete Merkmale: Endliche/ abzählbar unendliche viele mögliche Ergebnisse.

Stetige Merkmale: Ergebnisse sind kontinuierlich (überabzählbar), bspw. Länge, Temperatur.



Repräsentativität

Eine Stichprobe lässt Rückschlüsse auf die Grundgesamtheit zu, wenn sie repräsentativ ist. Hierfür müssen zufällig Objekte aus der Grundgesamtheit zur Bildung der Stichprobe ausgewählt werden.

Deskriptive Statistik

Die Daten werden nur beschrieben: Häufigkeiten, Mittelwerte, Streuung

Induktive Statistik

Es werden von einer Stichprobe Rückschlüsse auf die Grundgesamtheit gemacht.

Explorative Datenanalyse

Übergang zwischen deskriptiver und induktiver Statistik: Es wird nach Strukturen und Zusammenhängen in den gesucht.

Univariate Statistik

Es wird ein Merkmal des Merkmalsträgers untersucht.

Bivariate Statistik

Es werden zwei Merkmale des Merkmalsträgers untersucht.

Multivariate Statistik

Es werden viele Merkmale des Merkmalsträgers untersucht.

Ergebnis

Dies ist ein möglicher Ausgang eines Zufallsexperiments, bspw. eine Antwort bei einer Umfrage.

Ereignis

Zusammenfassung von Ergebnissen, bspw. Alter im Bereich von 20 bis 24.

1.3 Skalenniveaus und Merkmalsarten

Die in Fragebögen Daten ermittelten Daten sind sehr verschieden von ihrer Art her.

Beim Alter können wir noch einen Durchschnitt bilden... was tun wir aber beim Geschlecht?

Es gibt drei Niveaustufen, welche Skala man für Werte verwenden kann (Skalenniveaus):

1. Nominal skaliert: Es zählt nur die Ausprägung der Werte. Es kann bspw. nur gesagt werden „A hat die gleiche Farbe wie B“ oder „A ist das gleiche Modell wie B“. $= / \neq$

Beispiele sind: Farbe, Automodell, ...

2. Ordinal skaliert: Sie können der Reihenfolge nach geordnet werden („A ist besser als B“). $= / \neq$

Beispiele sind: Beliebtheitsskala, Plätze in Musik-Charts, ... $> / <$

3. Metrisch skaliert: Sie können der Reihenfolge nach geordnet werden und gleiche Abstände haben gleiche Bedeutung. („Der Abstand zwischen 1,90 cm und 1,80 cm ist halb so groß wie der zwischen 1,90 cm und 1,70 cm“). Es muss eine Maßeinheit dabei sein.

Beispiele sind: Größe (in cm), Temperatur (in °C).

Nicht Noten: Abstand zwischen 5 und 6 ist nicht so groß wie zwischen 4 und 5

(von den Prozenten).

Können nur Reihenfolge & Abstand interpretiert werden, sind es Intervallskalen.

Verhältnisskala: Intervallskala und das Verhältnis ist interpretierbar („A ist doppelt so ... wie B“). Notwendig hierzu: Ein absoluter Nullpunkt (kleinster Wert; Größe (0 cm), Alter (0 Jahre)) Bei Temperatur in Grad Celsius ist der Nullpunkt aber nicht der kleinstmögliche Wert, daher sind 20 °C nicht halb so warm wie 40 °C! Die Kelvinskala hat für die Temperatur einen absoluten Nullpunkt, ist also eine Verhältnisskala. Bei Größe aber: 180 cm ist doppelt so groß wie 90 cm.

Zusammenfassung: Kreuze an!

		Reihenfolge	Abstand	Verhältnis
Nominal				
Ordinal				
Metrisch	Intervallskala			
	Verhältnisskala			

Zahl der Ausprägungen

Die Zahl der Ausprägungen kann entweder diskret (abzählbar: 1-2-3-...)

oder stetig (nicht abzählbar) sein.

Messziel

Eine weitere Unterscheidung kann nach quantitativen/qualitativen Merkmalen erfolgen:

1. Quantitative Merkmale sind konkret messbare, also metrische Merkmale

(Größe, Alter, Gewicht, ...).

Sie unterscheiden sich in der Größe der Werte.

2. Qualitative Merkmale sind nicht konkret messbare, also nominale oder ordinale Merkmale

(Farbe, Geschlecht, Beliebtheitsskalen, ...).

Sie unterscheiden sich in der Art (mehr/wenig, weiblich/männlich).

1.4 Aufgaben

1. Merkmalsarten und Skalenniveaus.

Kreuzen Sie an, welcher Art die Merkmale sind! Begründe auch diene Antwort!

Merkmal	Nomi- nal	Ordi- nal	Metrisch		Qualita- tiv	Quantita- tiv
			Intervall	Verhält- nis		
Beliebtheit (kaum-mittel-sehr)						
Haarfarbe						
Kleidergröße (...-S-M-L-...)						
Tägliche Loginzeit in Minuten auf Instagram						
Zeugnisnote Deutsch als Note						
Zeugnisnote Deutsch in Punkten						
Luftdruck in Pascal						

- Erklären Sie die Begriffe Grundgesamtheit, Erhebungseinheit, Stichprobe, Merkmal, Ausprägung, Wertebereich am Beispiel: Studenten werden nach der Zufriedenheit mit ihrer Studiensituation befragt.
- Was bedeutet „Repräsentativität“ einer Stichprobe?
- Erklären Sie den Unterschied zwischen deskriptiver und induktiver Statistik sowie explorativer Datenanalyse. Machen Sie auch jeweils ein Beispiel.
- Erklären Sie den Unterschied zwischen uni-, bi- und multivariater Statistik.

1.5 Lösungen

1. Merkmalsarten und Skalenniveaus.

Merkmal	Nominal	Ordinal	Metrisch		Qualitativ	Quantitativ
			Intervall	Verhältnis		
Beliebtheit (kaum-mittel-sehr)		X			X	
Haarfarbe	X				X	
Kleidergröße (...-S-M-L-...)		X			X	
Tägliche Loginzeit in Minuten auf Facebook				X		X
Zeugnisnote Deutsch als Note		X			X	
Zeugnisnote Deutsch in Notenpunkten		X			X	
Luftdruck in Pascal				X		X

6. Erklären Sie die Begriffe Grundgesamtheit, Erhebungseinheit, Stichprobe, Merkmal, Ausprägung, Wertebereich am Beispiel: Studenten werden nach der Zufriedenheit mit ihrer Studiensituation befragt.

Erhebungseinheit: Jeder einzelne Student.

Grundgesamtheit: Alle Studenten.

Stichprobe: Befragte Studenten.

Merkmal: Zufriedenheit mit Studiensituation.

Wertebereich: sehr zufrieden, eher zufrieden, eher unzufrieden, unzufrieden (oder andere Einteilung).

Ausprägung: Wert des Merkmals bei einer bestimmten Erhebungseinheit.

3. **Stichprobe erlaubt Verallgemeinerung auf Grundgesamtheit**

4. **Deskriptive Statistik: Beschreibung und Darstellung vorliegender Daten (Diagramme). Beispiel: Stichprobe in Deutschland mit 100 Männern nach Bierkonsum, dann Durchschnitt berechnen etc.**

Induktive Statistik: Wenn Stichprobe repräsentativ: Rückschluss auf Grundgesamtheit.

Beispiel: Mit Stichprobe auf Trinkverhalten aller deutschen Männer schließen.

Explorative Datenanalyse: Daten auf Strukturen untersuchen, Hypothesen aufstellen.

Beispiel: Untersuchung, ob die Daten einem Trend folgen o. Ä.

5. **Univariat: eine Variable wird untersucht,**

bivariat: zwei Variablen,

multivariat: mehrere Variablen.

2 Datenarbeit

2.1 Die erste Datenerfassung

Grundlage jeder Statistik sind systematische Datenerhebungen. Ein paar Vokabeln hierzu:

Merkmal: Die Eigenschaft, die untersucht wird (Geschlecht, Alter, ...).

Merkmalsausprägung: Welchen „Wert“ das Merkmal bei einem Befragten hat (Alter: 17).

Merkmalsträger: Die Person, die bzgl. ihrer Merkmale befragt wird.

Bei großen Datenmengen werden in der Regel Fragebögen eingesetzt, die anonym oder anonymisiert (jede Person bekommt einen Code) ausgewertet werden. Die so gewonnenen Ergebnisse bereitet man mit den Methoden der Statistik auf und stellt sie möglichst anschaulich dar. Wir erheben hierzu unsere erste Statistik mit einem Fragebogen...

Ermitteln Sie die Werte unter 2.-4., wenn Ihnen die folgende Urliste schon gegeben ist!

1. Urliste

Geschlecht	Alter	Pizza-Beliebtheit	Geschlecht	Alter	Pizza-Beliebtheit
M	17	1	M	17	2
M	18	2	M	18	2
W	16	2	M	17	1
M	18	1	M	18	5
M	18	1	M	17	2
W	17	2	M	18	2
W	18	1	M	17	5
W	17	1	M	17	1
M	17	2	M	18	2
M	18	1	W	17	2
W	17	3	W	18	2
W	18	2	W	17	1
W	17	2			

2. Strichliste

	Männlich	Weiblich
Geschlecht		

	16	17	18
Alter			

	1 Sehr gern	2 Gern	3 Mittel	4 Nicht gut	5 Gar nicht
Pizza-Beliebtheit					

3. absolute Häufigkeiten

Die absolute Häufigkeit gibt an, wie oft etwas vorkommt, also die Anzahl der Striche.

Hinweis: Die Formel hierfür ist n_i (i ist die „Nummer“ der Ausprägung, bspw. 1 = männlich, 2 = weiblich. Die Zusammenfassung aller Personen wird „Grundgesamtheit“ genannt, ihre Anzahl wird mit n bezeichnet.)

Geschlecht	Männlich	Weiblich
Anzahl	$n_1 =$	$n_2 =$

Alter	16	17	18
Anzahl	$n_{16} =$	$n_{17} =$	$n_{18} =$

Pizza-Beliebtheit	1	2	3	4	5
Anzahl	$n_1 =$	$n_2 =$	$n_3 =$	$n_4 =$	$n_5 =$

4. relative Häufigkeiten

Die relative Häufigkeit gibt den Anteil an der Gesamtzahl (n) wieder.

Sie wird in der Regel in Prozent (%) angegeben, bspw. „60% sind männlich“.

Formel: $h_i =$

Geschlecht	Männlich	Weiblich
Anteil	$h_1 =$	$h_2 =$

Alter	16	17	18
Anteil	$h_{16} =$	$h_{17} =$	$h_{18} =$

Pizza-Beliebtheit	1	2	3	4	5
Anteil	$h_1 =$	$h_2 =$	$h_3 =$	$h_4 =$	$h_5 =$

2.2 Lösung

2. Strichliste

	Männlich	Weiblich
Geschlecht	++++ +++ +++	++++ +++

	16	17	18
Alter		++++ +++	++++ +++

	1 Sehr gern	2 Gern	3 Mittel	4 Nicht gut	5 Gar nicht
Pizza-Beliebtheit	+++	++++ +++			

3. absolute Häufigkeiten (Urliste 2. Teil): n_i

Geschlecht	Männlich	Weiblich
Anzahl	$n_1 = 15$	$n_2 = 10$

Alter	16	17	18
Anzahl	$n_{16} =$ 1	$n_{17} =$ 13	$n_{18} =$ 11

Pizza-Beliebtheit	1	2	3	4	5
Anzahl	$n_1 =$ 9	$n_2 =$ 13	$n_3 =$ 1	$n_4 =$ 0	$n_5 =$ 2

4. relative Häufigkeiten: $h_i = \frac{n_i}{n}, n = \frac{1}{n} \sum_{i=1}^k n_i$

Geschlecht	Männlich	Weiblich
Anteil	$h_1 =$ 60%	$h_2 =$ 40%

Alter	16	17	18
Anteil	$h_{16} =$ 4%	$h_{17} =$ 52%	$h_{18} =$ 44%

Pizza-Beliebtheit	1	2	3	4	5
Anteil	$h_1 =$ 36%	$h_2 =$ 52%	$h_3 =$ 4%	$h_4 =$ 0%	$h_5 =$ 8%

2.3 Übung zur Datenarbeit I

„Don Giovanni“, eine lokale Eisdiele, macht eine Umfrage, deren Ergebnisse im Folgenden zusammengefasst sind:

Wie viele männliche, wie viele weibliche Gäste besuchen die Eisdiele?

Männlich	Weiblich
32	43

Wie alt waren die Befragten?

15	16	17	18	19
18	10	16	13	18

Was sind die Lieblingseissorten der Gäste?

Vanille	Erdbeere	Schokolade	Nuss	Snickers	Exotisches
17	15	8	12	10	13

Bestimmen Sie die relativen Häufigkeiten der Geschlechts-Ausprägungen.

Männlich	Weiblich

Bestimmen Sie die relativen Häufigkeiten der Alters-Ausprägungen.

15 Jahre	16 Jahre	17 Jahre	18 Jahre	19 Jahre

Bestimmen Sie die relativen Häufigkeiten der Lieblingssorten.

Vanille	Erdbeere	Schokolade	Nuss	Snickers	Exotisches

Nun wollen wir die Altersverteilung in der Eisdiele herausfinden.

Bestimmen Sie jeweils absolute und relative Häufigkeit der Personen, die höchstens ... Jahre alt sind!

Alter x	15	16	17	18	19
$n(x \leq x_i)$					
$h(x \leq x_i)$					



2.4 Lösung

Relative Häufigkeiten Geschlecht

Männlich	Weiblich
$h_1 = 42,67\%$	$h_2 = 57,33\%$

Relative Häufigkeiten Alter

15 Jahre	16 Jahre	17 Jahre	18 Jahre	19 Jahre
$h_{15} = 24\%$	$h_{16} = 13,33\%$	$h_{17} = 21,33\%$	$h_{18} = 17,33\%$	$h_{19} = 24\%$

Relative Häufigkeiten Lieblingssorten

Vanille	Erdbeere	Schokolade	Nuss	Snickers	Exotisches
$h_1 = 22,67\%$	$h_2 = 20\%$	$h_3 = 10,67\%$	$h_4 = 6\%$	$h_5 = 13,33\%$	$h_6 = 17,33\%$

Altersverteilung

Absolute und relative Häufigkeit der Personen, die höchstens ... Jahre alt sind!

Alter	15	16	17	18	19
Anzahl $\leq$	18	28	44	57	75
Anteil $\leq$	24%	37,33%	58,67%	76%	100%

2.5 Übung zur Datenarbeit II

Wir machen eine Befragung zum Getränkeverhalten bei jungen Erwachsenen.

Geschlecht	Männlich	Weiblich

Alter	17	18	19	20	21	22
	I					

Ge- tränk	Kein Al- kohol	Bier	Wodka Energy	Wodka- O	Passoa- O	Sex on the beach	Caipi- rinha

Um die Auswertung einer Umfrage zu üben, sollt ihr das nun übernehmen!

Als Grundlage hierzu dient euch die Strichliste, die wir angefertigt haben. Die folgenden Aufgaben sollt ihr nun lösen.

1. Bestimmen Sie die absoluten Häufigkeiten

	Männlich	Weiblich
Geschlecht		

	17	18	19	20	21	22
Alter						

	Kein Al- kohol	Bier	Wodka Energy	Wodka-O	Passoa-O	Sex on the Beach	Caipirinha
Getränk							

2. Bestimmen Sie die relativen Häufigkeiten

	Männlich	Weiblich
Geschlecht		

	17	18	19	20	21	22
Alter						

Ge- tränk	Kein Al- kohol	Bier	Wodka Energy	Wodka-O	Passoa-O	Sex on the Beach	Caipirinha

2.6 Lösung

1. Absolute Häufigkeiten

Geschlecht	Männlich	Weiblich
	$n_1 = 39$	$n_2 = 22$

Alter	17	18	19	20	21	22
	$n_1 = 1$	$n_1 = 30$	$n_3 = 18$	$n_4 = 7$	$n_5 = 2$	$n_6 = 3$

Ge-tränk	Kein Al-kohol	Bier	Wodka Energy	Wodka-O	Passoa-O	Sex on the Beach	Caipirinha
	$n_1 = 11$	$n_2 = 19$	$n_3 = 12$	$n_4 = 3$	$n_5 = 5$	$n_6 = 7$	$n_7 = 4$

2. Relative Häufigkeiten, Grundgesamtheit: 61 Personen

	Männlich	Weiblich
Ge-schlecht	$h_1 = 63,9\%$	$h_2 = 36,1\%$

Al-ter	17	18	19	20	21	22
	$h_{17} = 1,6\%$	$h_{18} = 49,2\%$	$h_{19} = 29,5\%$	$h_{20} = 11,5\%$	$h_{21} = 3,3\%$	$h_{22} = 4,9\%$

Ge-tränk	Kein Al-kohol	Bier	Wodka Energy	Wodka-O	Passoa-O	Sex on the Beach	Caipirinha
	$h_1 = 18\%$	$h_2 = 31,1\%$	$h_3 = 19,7\%$	$h_4 = 4,9\%$	$h_5 = 8,2\%$	$h_6 = 11,5\%$	$h_7 = 6,6\%$

2.7 Quantile

„Die Arbeit ist sehr schlecht ausgefallen! Mindestens 40% haben höchstens 50 Punkte!“

In einer Klassenarbeit ergaben sich die folgenden Noten:

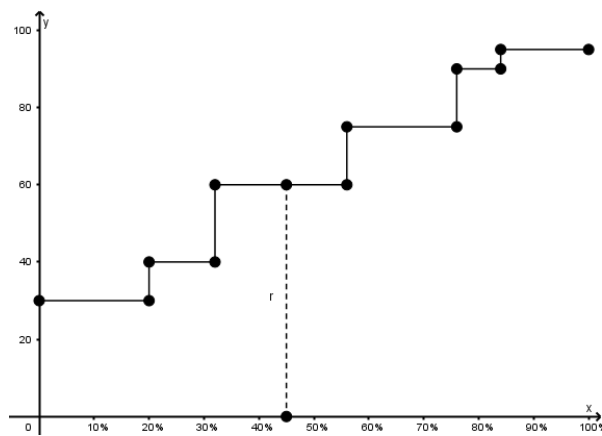
Punkte	30	40	60	75	90	95
Anzahl	5	3	6	5	2	4
Anteil (rel. Häufigkeiten)	20%	12%	24%	20%	8%	16%
Anteil, wo Punkte $\leq$	20%	32%	56%	76%	84%	100%

Stimmt diese Behauptung?

Man kann auch allgemein sagen:

Ein p -Quantil $x_{p\%}$ ist der kleinste Wert von x , bei dem mind. $p\%$ der Werte kleiner oder gleich diesem Wert sind. („ $p\%$ sind kleiner oder gleich x “, also höchstens x)

In der folgenden Grafik, die die kumulierten Wahrscheinlichkeiten wiedergibt, können die Quantile direkt abgelesen werden, also bspw. das 45%-Quantil liegt bei 60.



Die Zehntel-Quantile werden „Dezile“, die Viertel-Quantile „Quartile“ genannt.

Bestimmen Sie nun die folgenden Quantile:

$$x_{10\%} = x_{0,1} =$$

$$x_{25\%} = x_{0,25} =$$

$$x_{40\%} = x_{0,4} =$$

Bestimmen Sie anschließend alle Dezile und alle Quartile!

Kurzlösung

Punkte	30	40	60	75	90	95
Anzahl	5	3	6	5	2	4
Anteil	20%	12%	24%	20%	8%	16%
Anteil, wo Punkte $\leq$	20%	32%	56%	76%	84%	100%

$$x_{10\%} = x_{0,1} = \boxed{30},$$

$$x_{25\%} = x_{0,25} = \boxed{40},$$

$$x_{40\%} = x_{0,4} = \boxed{60}$$

$$x_{0,1} = x_{0,2} = 30, x_{0,3} = 40, x_{0,4} = x_{0,5} = 60, x_{0,6} = x_{0,7} = 75, x_{0,8} = 90, x_{0,9} = x_1 = 95$$

$$x_{0,25} = 40, x_{0,5} = 60, x_{0,75} = 75, x_1 = 95$$

2.8 Boxplots

Wir betrachten als Werte die Noten einer Klausur (10 Studenten): 1, 1, 2, 2, 2, 3, 4, 4, 4, 5.

Note	1	2	3	4	5
Anzahl	2	3	1	3	1
Anteil	20%	30%	10%	30%	10%
Anteil, wo Note $\leq$	20%	50%	60%	90%	100%

Einige Quantile:

$$x_{0,25} = 2, x_{0,50} = 2$$

$$\tilde{x} = 2,5, x_{0,75} = 4$$

$$x_{1,00} = 5$$

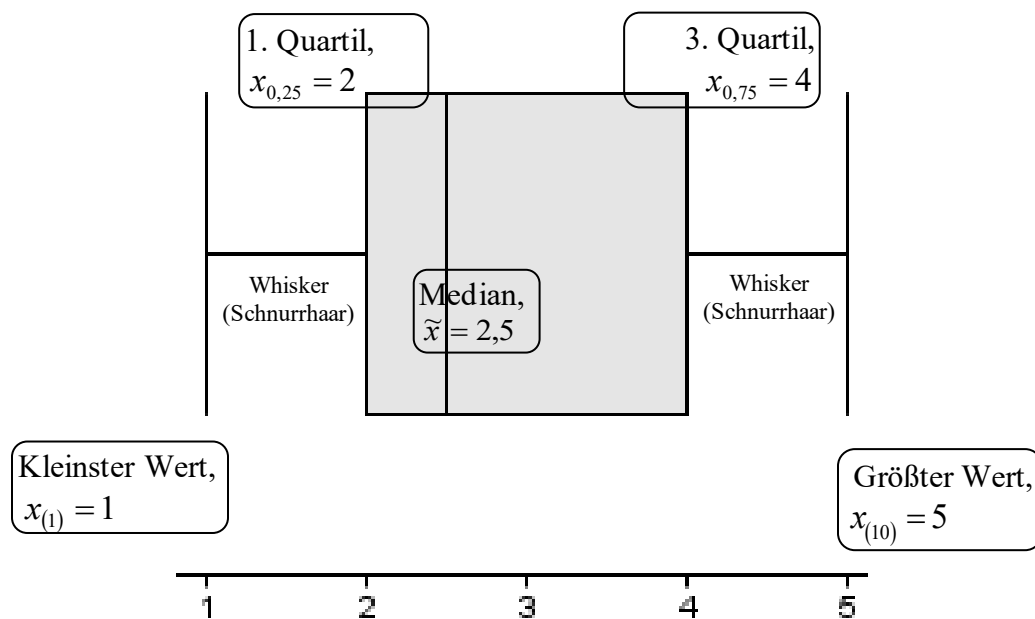
$\tilde{x} = 2,5$ ist der „Median“; dies ist der Wert, der die Stichprobe in genau 2 gleich große Hälften teilt.

Die eine Hälfte der Werte der Stichprobe ist größer, die andere kleiner als der Median.

(Hier sieht man auch: Die häufig in der Literatur aufgestellte Behauptung, dass $\tilde{x} = x_{0,5}$ sei, ist falsch!)

Einfacher ist die Ermittlung der Quantile natürlich, wenn bereits Prozentwerte vorgegeben sind (bspw. Anteile an der Gesamtbevölkerung, der Kursmitglieder etc.). Hier entfällt die zweite Zeile.

Eine schöne Zusammenfassung wichtiger Daten liefert der Boxplot (5-Zahlen-Zusammenfassung):



Was brauchen wir hierfür?

Es kann auch sein, dass einige Werte (bspw. 1. Quartil und Median) identisch sind. Dann werden die entsprechenden Teile der Box zu einem Bereich.

Schauen wir uns die Kiste noch einmal genauer an...



- Der Abstand der Schnurrhaare ist die Spannweite (Unterschied zwischen bester und schlechtester Note).
- Der Abstand des linken Schnurrhaars zum ersten Quartil ist der Unterschied innerhalb der Gruppe der besten 25% der Schüler. (Es sind allerdings mind. 25% der Schüler! In unserem Beispiel sind es sogar 50%!)
- Der Abstand der Boxenden („Interquartilsabstand“) ist der Bereich, in dem die mittleren 50% der Noten sind. (Wieder müssen es nicht genau 50% sein – Hier sind es $90\% - 50\% = 40\%$.)
- Der Abstand des 3. Quartils zum rechten Schnurrhaar ist der Unterschied innerhalb der Gruppe der schlechtesten 25% der Schüler. (Wieder nicht genau: Hier sind es $100\% - 90\% = 10\%$.)

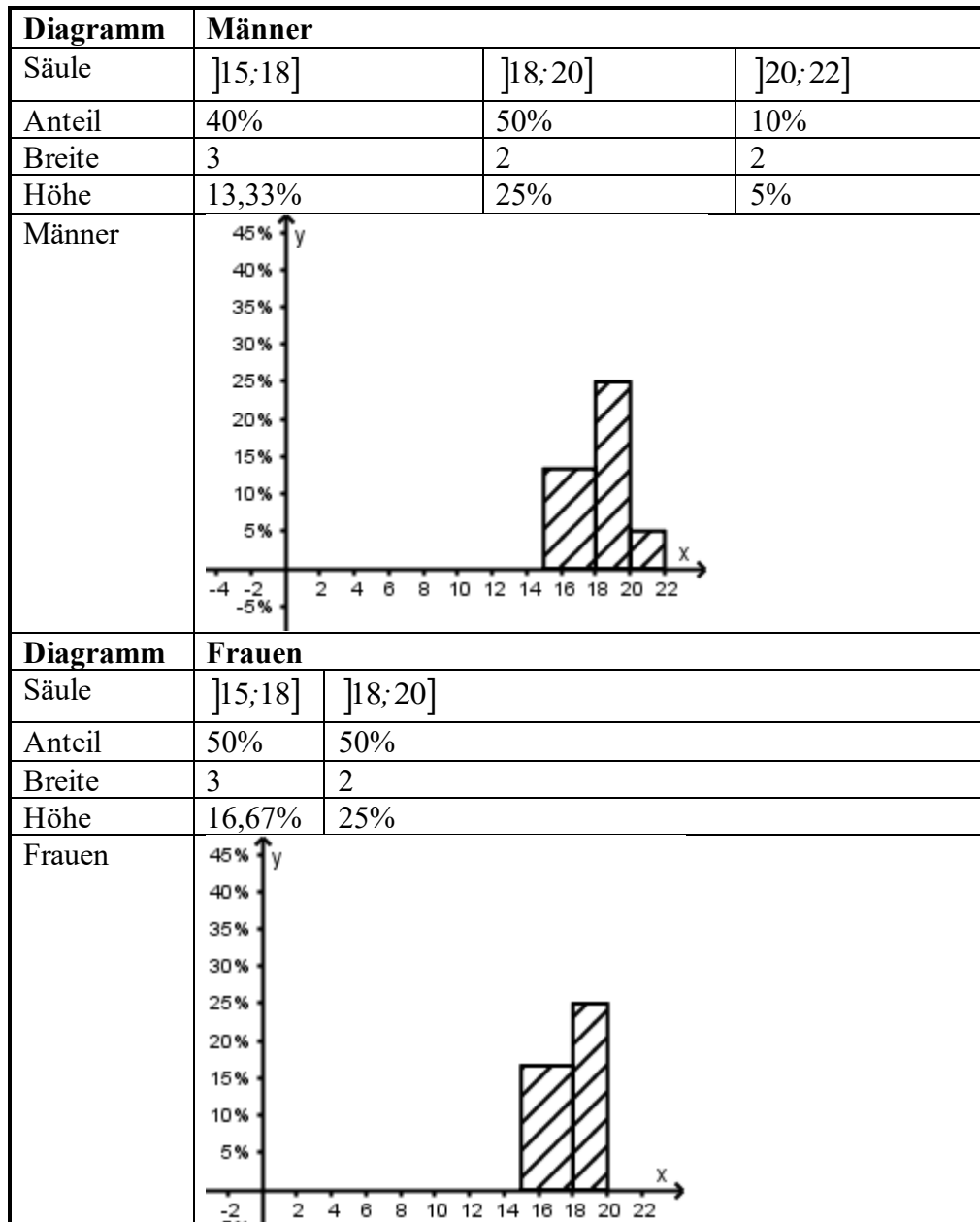
2.9 Histogramme

Ein Histogramm ist ein Säulendiagramm.

Das Problem hier: Die Säulen haben unterschiedliche Breite (3 bzw. 2 Jahre.)

Die Lösung ist ganz einfach: Wenn die Fläche die Gesamtwahrscheinlichkeit ist, dann gilt:

$$\text{Höhe} = \frac{\text{Wahrscheinlichkeit}}{\text{Breite}}$$



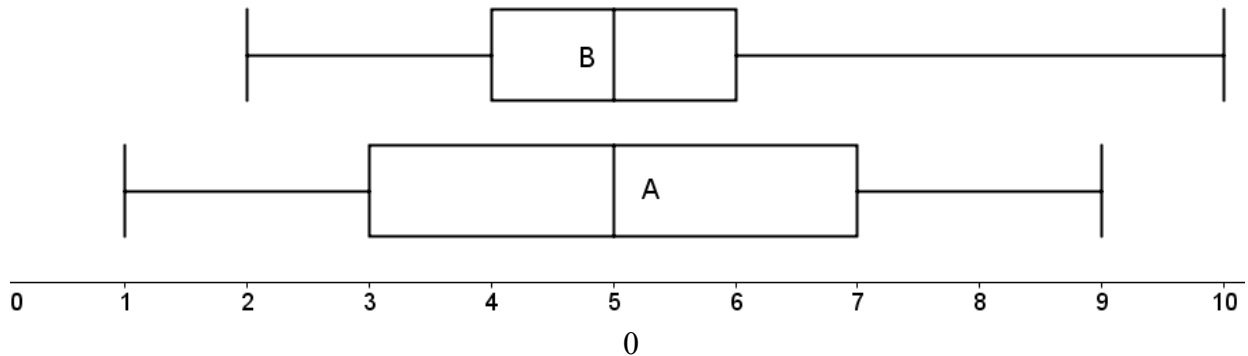
2.10 Übungen

1. Gegeben sind die Ergebnisse einer Umfrage zur Wichtigkeit von Hobbies:

	1	2	3	4
Urliste	+++ +++	+++		
Anzahl				
Anteil				
Anteil $\leq$				

Füllen Sie die Tabelle aus, zeichnen Sie abschließend den Boxplot.

2. Gegeben sind die Monatseinkommen der Absolventen zweier Universitäten (jeweils 100 Studenten) im ersten Jahr nach dem Abschluss. Welche Aussagen stimmen? Begründen Sie!



- ☐ Das mittlere Einkommen ist bei beiden Absolventengruppen gleich groß.
- ☐ Der Einkommensunterschied zwischen bestem und schlechtestem Einkommen ist bei Uni A höher.
- ☐ Eine Person, die weniger als 2.000 € verdient, stammt nicht von der Uni B.
- ☐ Mindestens 75% aller Absolventen haben ein Einkommen bis 7.000 €.
- ☐ Höchstens 25% aller Absolventen haben ein Einkommen bis 4.000 €.

3. Gegeben ist die folgende Einkommensverteilung der Absolventen eines Jahrgangs:

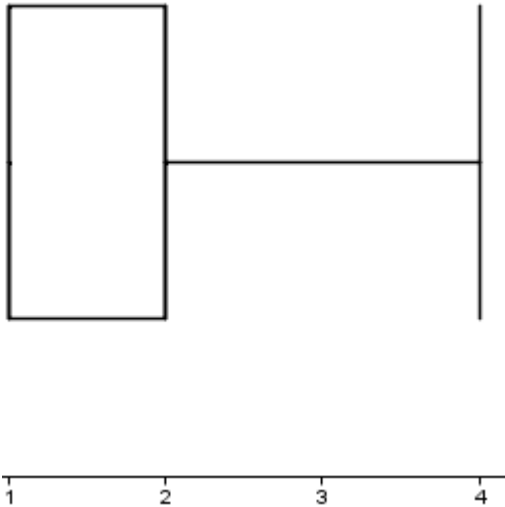
Einkommen in 1.000 €	Anteil
]0;40]	20%
]40;90]	50%
]90;150]	30%

Zeichnen Sie das Histogramm

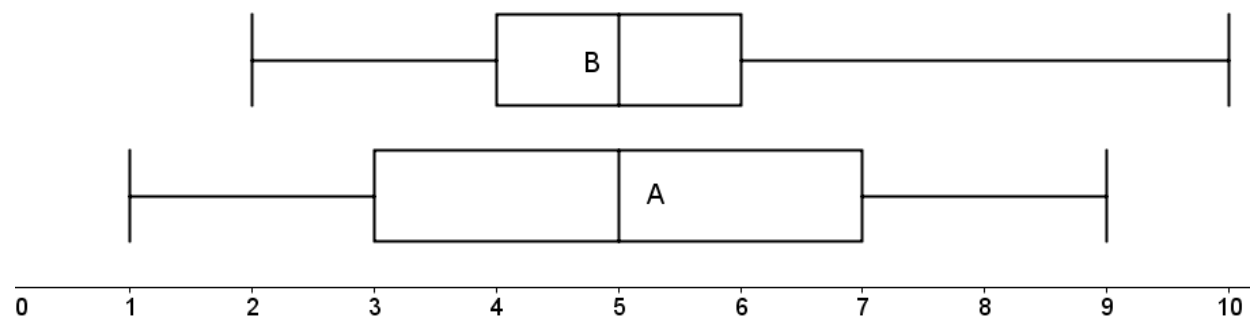
2.11 Lösungen

1. Boxplot

Wert	1	2	3	4
Anteil $\leq$	44%	80%	88%	100%



2. Gegeben sind die Monatseinkommen der Absolventen zweier Universitäten (jeweils 100 Studenten) im ersten Jahr nach dem Abschluss. Welche Aussagen stimmen? Begründen Sie!

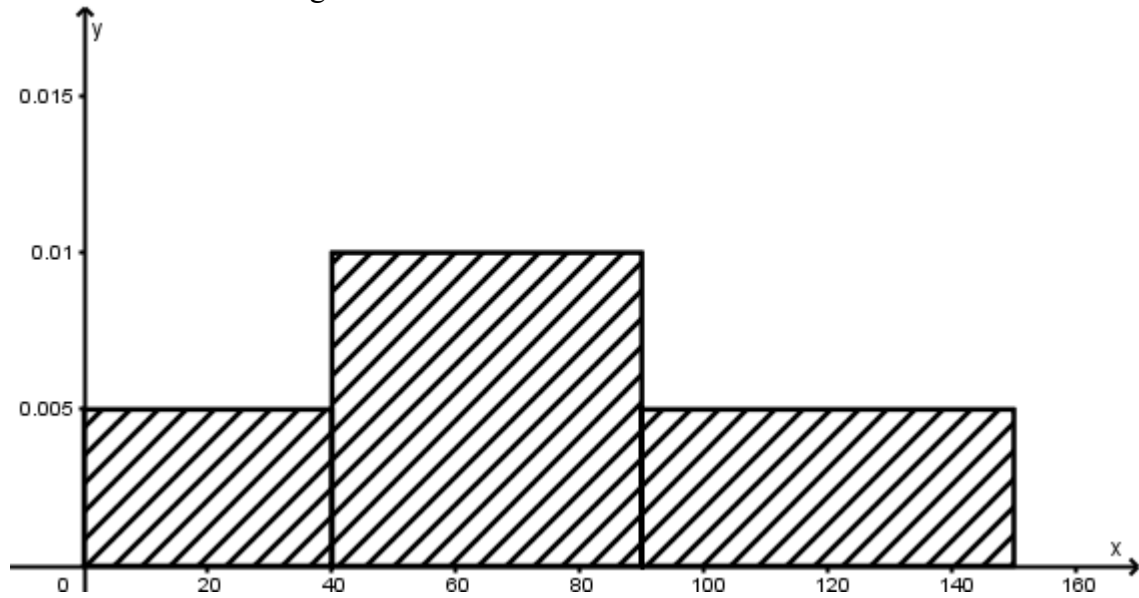


- W** Das mittlere Einkommen ist bei beiden Absolventengruppen gleich groß.
Das Einkommen, das in der Mitte liegt, ist der Median.
- F** Der Einkommensunterschied zwischen bestem und schlechtestem Einkommen ist bei Uni A höher.
Der Unterschied ist bei beiden gleich groß! (Whisker-Abstand jeweils 8.000 €!)
- W** Eine Person, die weniger als 2.000 € verdient, stammt nicht von der Uni B.
- W** Mindestens 75% aller Absolventen haben ein Einkommen bis 7.000 €.
Wenn bei den A-Absolventen das 75%-Quantil bei 7.000 € ist und bei B bei 6.000 €, sind bei den B-Absolventen sogar noch mehr als 75% in der Gruppe „bis 7.000 €.“
- F** Höchstens 25% aller Absolventen haben ein Einkommen bis 4.000 €.
Wenn in der B-Gruppe schon mindestens 25% ein Einkommen von bis 4.000 € haben und in der A-Gruppe mindestens 25% ein Einkommen von bis 3.000 € haben, haben insgesamt mehr als 25% ein Einkommen bis 4.000 €.

3. Gegeben ist die folgende Einkommensverteilung der Absolventen eines Jahrgangs:

Einkommen in 1.000 €	Anteil
$]0;40]$	20%
$]40;90]$	50%
$]90;150]$	30%

Zeichnen Sie das Histogramm.



3 Kennzahlen eines Merkmals

3.1 Lagemaße

3.1.1 Erklärung

Angenommen im Frankfurt gäbe es insgesamt 30 Eisdiele, aber die Kugeln hätten verschiedene Größen. Bei 13 Eisdiele hätten sie einen Durchmesser von 8 cm, bei den 5 weiteren 6 cm, 10 Eisdiele haben 4 cm und bei den letzten 2 einen Durchmesser von 3 cm.

Um das zu vereinfachen, stellen wir die Werte in einer Tabelle dar:

Durchmesser [cm]	8	6	3	4
Anzahl der Eisdiele	13	5	2	10

a. Modus / Modalwert

Ein schnell zu berechnender Mittelwert ist der **Modalwert** (Modus). Das ist der Wert, der am häufigsten vorkommt, im obigen Beispiel der Durchmesser, der von den meisten Eisdiele angeboten wird. Dieser Mittelwert aller Kugeln liegt also bei 8 cm: $\bar{x} = 8$. Es gibt immer mindestens einen Modalwert; es kann auch mehrere Modalwerte geben.

b. Median

Eine weitere Möglichkeit für den Statistiker ist es, den **Median** zu verwenden.

Der Median stellt den mittleren Wert dar. Er ist eine Weiterentwicklung der 50%-Quantils, da dieses nicht immer genau in der Mitte liegt. Um den Median zu berechnen, müssten die Eisdiele nach dem Kugel-Durchmesser (davon suchen wir ja den Median!) geordnet werden. Hierbei ist egal, ob der größte oder der kleinste Wert zuerst kommt. Wir fangen mit dem kleinsten Wert an:

Durchmesser	3	4	6	8
Eisdiele	2	10	5	13

Wenn wir die Werte der Größe nach geordnet haben, können wir die Werte neu durchnummerieren: $x_{(1)}$ ist der erste, $x_{(2)}$ ist der zweite, ... Wert.

Bei einer ungeraden Zahl an Werten ist der Wert $x_{\left(\frac{n+1}{2}\right)}$ der Median.

(Gäbe es bspw. 3 Eisdiele, so wäre der $\left(\frac{3+1}{2}\right) = 2$. Wert der Median.)

Da die Anzahl der Eisdiele gerade ist, gibt es keinen Wert in der Mitte. Wir brauchen folgende Formel, um den Median zu berechnen (Mittelwert der zwei mittleren Werte):

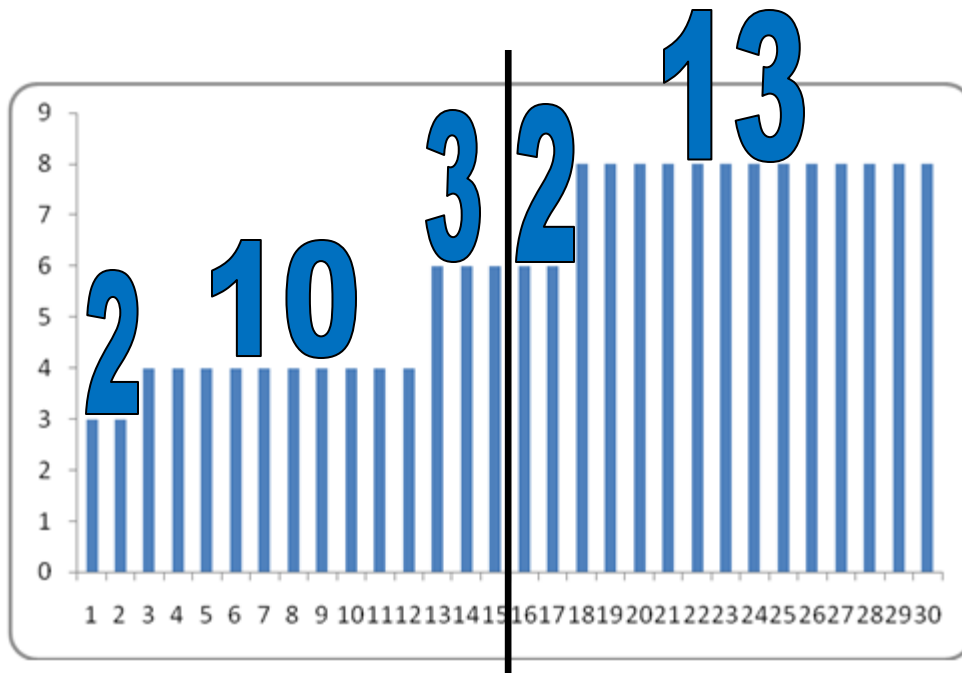
$$\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Für „n“ setzen wir die Gesamtzahl der Eisdiele ein. Hier: 30. Es ergibt sich:

$$\tilde{x} = \frac{x_{\left(\frac{30}{2}\right)} + x_{\left(\frac{30}{2}+1\right)}}{2} = \frac{x_{(15)} + x_{(16)}}{2}$$

Der Median liegt also zwischen der 15. und 16. Eisdiele. Der Durchmesser der 15. und 16. Eisdiele beträgt 6 cm.

Graphisch dargestellt:



Der Median ist in dem Fall also deutlich bei 6 cm. In einer Formel dargestellt sieht das Ganze so

$$\text{aus: } \tilde{x} = \frac{6 + 6}{2} = 6$$

Der Statistiker würde also behaupten, dass der mittlere Wert bei 6 cm liegt.

c. arithmetisches Mittel

Eine andere Berechnung erfolgt beim **arithmetisches Mittel**. Für dieses Mittel muss er den Wert des Durchmessers mal der Zahl der Eisdien mit diesem Durchmesser nehmen und am Ende durch die Anzahl der Eisdien teilen. Dies ist wie bei der Berechnung des Durchschnitts einer Klassenarbeit!

$$\text{Hier: } \bar{x} = \frac{\sum_{k=0}^l x_k \cdot n_k}{n} = \frac{8 \cdot 13 + 6 \cdot 5 + 3 \cdot 2 + 4 \cdot 10}{30} = 6$$

(l ist hier die Anzahl der verschiedenen Durchmesser (Gruppen), hier: 4)

Er könnte auch mit relativen Häufigkeiten berechnet werden:

$$\bar{x} = \sum_{k=0}^l x_k \cdot \frac{n_k}{n} = \sum_{k=0}^l x_k \cdot h_k = 8 \cdot \frac{13}{30} + 6 \cdot \frac{5}{30} + 3 \cdot \frac{2}{30} + 4 \cdot \frac{10}{30} = 6$$

Also liegt das arithmetische Mittel bei 6 cm. Dann würde der Statistiker behaupten, dass sich der Durchschnittsdurchmesser bei den Eisdien bei 6 cm befindet.

Das arithmetische Mittel minimiert die quadratischen Abstände:

$$\sum_{k=0}^n (x_k - a)^2 \rightarrow \min \Leftrightarrow a = \bar{x}$$

Denn:

$$S = \sum_{k=0}^n (x_k - a)^2 = \sum_{k=0}^n (x_k^2 - 2x_k a + a^2)$$

$$S' = \sum_{k=0}^n (2x_k - 2a) = 0 \Rightarrow \sum_{k=0}^n 2x_k - \sum_{k=0}^n 2a = 0 \Rightarrow \sum_{k=0}^n 2x_k = \sum_{k=0}^n 2a$$

$$\Rightarrow \sum_{k=0}^n x_k = \sum_{k=0}^n a \Rightarrow \sum_{k=0}^n x_k = na \Rightarrow a = \frac{\sum_{k=0}^n x_k}{n} = \bar{x}$$

Hier sieht man: Abstände müssen interpretierbar sein! → metrische Skalierung!)
Zudem liegt das arithmetische Mittel wirklich genau in der Mitte:

$$\sum_{k=0}^n (x_k - \bar{x}) = \sum_{k=0}^n x_k - \sum_{k=0}^n \bar{x} = \sum_{k=0}^n x_k - n \cdot \bar{x} = \sum_{k=0}^n x_k - \sum_{k=0}^n x_k = \sum_{k=0}^n x_k - n \cdot \frac{\sum_{k=0}^n x_k}{n} = 0$$

Mittelwert	Modus	Median	Arithmetisches Mittel
Skalenniveau	nominal, ordinal, metrisch	ordinal, metrisch	metrisch, Intervallskala

Bei metrischen und verhältnisskalierten Werten wird das geometrische Mittel verwendet.

Beispiel: Wirtschaftswachstum in drei Jahren: 2%, 4%, -1%

$$x_g = \sqrt[3]{1,02 \cdot 1,04 \cdot 0,99} \approx 1,016, \text{ also circa } 1,6\% \text{ Wachstum pro Jahr.}$$

3.1.2 Übungen

Aufgabe 1

Gegeben sind die Geburtenzahlen in den Orten Villariba und Villabajo in 10 aufeinanderfolgenden Wochen.

Berechnen Sie alle Mittelwerte. Welcher Wert erscheint Ihnen hier am aussagekräftigsten für die beiden Datensätze? Welcher beschreibt also die Datensätze am besten?

Villariba	0	1	0	2	8	2	1	0	1	0
Villabajo	2	2	0	1	2	1	1	0	0	1

Aufgabe 2

Bei einer Studie zur Entfernung vom Arbeitsplatz ergaben sich folgende Werte:

Befragter Nr.	1	2	3	4	5
Entfernung	55	100	5	35	5

Bestimmen Sie den Modus, den Median und das arithmetische Mittel.

Aufgabe 3

In einer anderen Gruppe ergaben sich folgende Werte:

Befragter Nr.	1	2	3	4
Entfernung	24	32	12	2

Bestimmen Sie den Modus, den Median und das arithmetische Mittel.

Aufgabe 4

Die Statistik über die Schüler einer Schulklasse liefert folgende Werte:

Haarfarbe:	Blond: 10	Brünett: 10	Schwarz: 5
Alter:	17 Jahre: 10	18 Jahre: 8	19 Jahre: 7
Geschlecht:	Männlich: 14	Weiblich: 11	

Was ist das arithmetische Mittel, was der Median und was der Modus der Daten, wo sie berechnet werden können?

Aufgabe 5

Manchmal sind gar nicht die absoluten Häufigkeiten, sondern nur die relativen Häufigkeiten gegeben, mit denen verschiedene Werte vorkommen. Bei einer Messung ergaben sich folgende Werte:

Größe	160	165	170	175	180	185	190
Anteil	5%	10%	10%	30%	20%	10%	15%

Berechnen Sie das arithmetische Mittel, den Median und den Modalwert der Körpergröße.

Aufgabe 6

Die Renditen eines Aktienfonds werden über 5 Jahre hinweg notiert.

Jahr	1	2	3	4	5
Rendite	5%	4%	-2%	1%	3%

Zu Beginn stand ein Fondsanteil bei 120 €.

Wie hoch ist die durchschnittliche Rendite? Wie wird sie berechnet?

Wie hoch steht ein Fondsanteil am Ende der 5 Jahre?

Welche Werte haben Median, Modus und arithmetisches Mittel der Rendite?

Aufgabe 7

Für metrische Daten haben wir unter anderem Median und arithmetisches Mittel kennengelernt. Welche Vorteile hat welcher dieser Mittelwerte?

Betrachten Sie hierfür die folgenden Fälle und berechne sowohl Median als auch arithmetisches Mittel:

1. In unserem Dönerladen verkauften wir in einer Maiwoche an einem Tag (Feiertag) 1.500 Döner, an den anderen sechs Tagen jeweils 100 Döner.
2. Die Kunden, die wir in der Feiertags-Woche neu hinzugewonnen haben, bleiben bei uns, kaufen aber über die Woche verteilt:
Wir verkaufen nun an zwei Tagen jeweils 200, an zwei Tagen 250 und an den restlichen drei Tagen 400 Döner. Es wird also insgesamt die gleiche Menge gekauft wie im ersten Fall, allerdings anders verteilt.

Aufgabe 8***

Beweisen Sie die Ungleichung vom arithmetischen und geometrischen Mittel.

Im einfachsten Fall, bei zwei Werten, gilt: Wenn a und b beide das gleiche Vorzeichen haben, so gilt:

$$\frac{a+b}{2} \geq \sqrt{a \cdot b}$$

3.1.3 Lösungen

Aufgabe 1

Villariba: $\bar{x} = 1,5, \tilde{x} = 1, \dot{x} = 0, x_g = 0$

Villabajo: $\bar{x} = 1, \tilde{x} = 1, \dot{x} = 1, x_g = 0$

Vor Allem Einsen in beiden Datensätzen $\Rightarrow$ Median am besten, da er dies als Mittelwert ausweist.

Aufgabe 2

$$\dot{x} = 5, \text{sortieren} : 5 - 5 - 35 - 55 - 100 \Rightarrow \tilde{x} = 35, \bar{x} = \frac{2 \cdot 5 + 35 + 55 + 100}{5} = \frac{200}{5} = 40$$

Aufgabe 3

$$\dot{x} = 24 \text{ oder } \dot{x} = 32 \text{ oder } \dot{x} = 12 \text{ oder } \dot{x} = 2, \\ \text{sortieren} : 2 - 12 - 24 - 32 \Rightarrow \tilde{x} = \frac{12 + 24}{2} = 18, \bar{x} = \frac{2 + 12 + 24 + 32}{4} = \frac{70}{4} = 17,5$$

Aufgabe 4

	Arithmet. Mittel	Median	Modus
Haarfarbe:	nicht möglich	Nicht möglich	Blond und Brünett
Alter:	17,88	18	17
Geschlecht:	Nicht möglich	Nicht möglich	Männlich

Aufgabe 5

$$\bar{x} = 160 \cdot 5\% + 165 \cdot 10\% + 170 \cdot 10\% + 175 \cdot 30\% + 180 \cdot 20\% + 185 \cdot 10\% + 190 \cdot 15\% = 177 \\ \dot{x} = 175, \tilde{x} = 175$$

Aufgabe 6

Wie hoch ist die durchschnittliche Rendite? Wie wird sie berechnet?

$$x_g = \sqrt[5]{1,05 \cdot 1,04 \cdot 0,98 \cdot 1,01 \cdot 1,03} \approx 1,022$$

Wie hoch steht ein Fondsanteil am Ende der 5 Jahre?

$$120 \cdot 1,02^5 \approx 133,79\text{€}$$

Die Werte der anderen Mittelwerte sind:

$$\dot{x} = \langle \langle \text{alle Werte} \rangle \rangle, \tilde{x} = 3\%, \bar{x} = \frac{5\% + 4\% + (-2\%) + 1\% + 3\%}{5} = 2,2\%$$

Aufgabe 7

$$1. \bar{x} = \frac{2.100}{7} = 300, \tilde{x} = 100 \quad 2. \bar{x} = \frac{2.100}{7} = 300, \tilde{x} = 250$$

	Median	Arithmetisches Mittel
Vorteil	Robust gegenüber Ausreißern	Robust gegenüber internen Verschiebungen
Nachteil	Anfällig gegenüber internen Verschiebungen	Anfällig gegenüber Ausreißern

Aufgabe 8***

$$\frac{a+b}{2} - \sqrt{a \cdot b} \geq 0 \Rightarrow \frac{a+b}{2} - \frac{2 \cdot \sqrt{a \cdot b}}{2} \geq 0 \Rightarrow \frac{a - 2 \cdot \sqrt{a \cdot b} + b}{2} \geq 0 \Rightarrow \frac{(\sqrt{a} - \sqrt{b})^2}{2} \geq 0$$

Somit ist die Behauptung bewiesen, da ein Quadrat nie negativ wird.

3.2 Streuungsmaße

3.2.1 Erklärung

Al- ter	Anzahl in Kurs 1	Anzahl in Kurs 2
16	0	5
17	2	5
18	21	5
19	2	5
20	0	5
Diagramm	A bar chart for Kurs 1. The vertical axis has 6 horizontal grid lines. The horizontal axis represents age. There are three bars: a small bar at age 17 (height 2), a tall bar at age 18 (height 21, reaching the 5th grid line), and a small bar at age 19 (height 2).	A bar chart for Kurs 2. The vertical axis has 6 horizontal grid lines. The horizontal axis represents age. There are five bars, all of equal height 5, located at ages 16, 17, 18, 19, and 20.

Bei gleichem arithmetischem Mittel können die Werte offensichtlich sehr unterschiedlich verteilt sein.

Wie kann das passieren?

Es zählt nicht nur der Mittelwert, sondern auch die Streuung, die durchschnittliche Abweichung vom Mittelwert.

Maße hierfür sind Varianz, Standardabweichung, Interquartilsabstand und Spannweite.

Varianz

Diese durchschnittliche quadratische Abweichung wird Varianz bezeichnet. Die Formel lautet

$$s_X^2 = \text{Var}(X) = \frac{\sum_{k=1}^l n_k (x_k - \bar{x})^2}{n} \quad (l \text{ ist die Anzahl verschiedener Werte.})$$

Wir benötigen hier das arithmetische Mittel – die Daten müssen metrisch skaliert sein!

Es muss also von jedem Wert der Mittelwert abgezogen, dann quadriert werden.

Die Quadrate werden mit der Anzahl gewichtet, addiert und abschließend durch die Gesamtzahl geteilt.

Für den ersten Kurs ergibt sich:

Alter	Anzahl in Kurs 1	$(\text{Alter} - \text{Mittel})^2$	Gewichtet mit Anzahl, also $(\text{Alter} - \text{Mittel})^2 \cdot \text{Anzahl}$
16	0	$(16 - 18)^2 = 4$	$4 \cdot 0 = 0$
17	2	$(17 - 18)^2 = 1$	$1 \cdot 2 = 2$
18	21	$(18 - 18)^2 = 0$	$0 \cdot 21 = 0$
19	2	$(19 - 18)^2 = 1$	$1 \cdot 2 = 2$
20	0	$(20 - 18)^2 = 4$	$4 \cdot 0 = 0$
	$\bar{x} = 18$	8	Summe: 4
			$s_{X_1}^2 = \frac{\text{Summe}}{\text{Gesamtzahl}} = \frac{4}{25} = 0,16$

$$\text{Kurz: } s_{X_2}^2 = \frac{2 \cdot (17 - 18)^2 + 21 \cdot (18 - 18)^2 + 2 \cdot (19 - 18)^2}{25} = \frac{4}{25} = 0,16$$

Berechnen Sie auf dem gleichen Weg die Varianz des Alters in dem zweiten Kurs!

Lösung

Alter	Anzahl in Kurs 2	$(\text{Alter} - \text{Mittel})^2$	Gewichtet mit Anzahl, also $(\text{Alter} - \text{Mittel})^2 \cdot \text{Anzahl}$
16	5	$(16 - 18)^2 = 4$	$4 \cdot 5 = 20$
17	5	$(17 - 18)^2 = 1$	$1 \cdot 5 = 5$
18	5	$(18 - 18)^2 = 0$	$0 \cdot 5 = 0$
19	5	$(19 - 18)^2 = 1$	$1 \cdot 5 = 5$
20	5	$(20 - 18)^2 = 4$	$4 \cdot 5 = 20$
	$\bar{x} = 18$		Summe: 50
			$s_X^2 = \frac{50}{25} = 2$

Das Problem ist aber noch, dass die Varianz die quadratische Abweichung angibt.
 Das Ergebnis ist nur schwer interpretierbar.
 (Was bedeutet schon „100 Quadratstunden Abweichung“?)

Standardabweichung

Die durchschnittliche Abweichung der Merkmalswerte vom Mittelwert ist die Standardabweichung s . Sie ist die Wurzel der Varianz.

Als Formel ergibt sich also: $s_X = \sqrt{\text{Var}(X)} = \sqrt{\frac{\sum_{k=1}^I n_k (x_k - \bar{x})^2}{n}}$

Bedeutung: Mittlere Abweichung vom Mittelwert.

Berechnen Sie die Standardabweichung des Alters in den beiden Kursen!

$$s_{X_1} = \sqrt{0,16} = 0,4, s_{X_2} = \sqrt{2} \approx 1,41$$

Wo ist die Streuung also größer?

In der 2. Gruppe

Eine andere Größe zur Beschreibung der Streuung ist der Interquartilsabstand:

Interquartilsabstand

Der Interquartilsabstand gibt den Abstand des 3. vom 1. Quartil wieder.

Es ist also die Größe des Bereichs, in dem die inneren 50% der Werte liegen: $Q = x_{0,75} - x_{0,25}$

Welchen Teil des Boxplots betrachten wir beim Interquartilsabstand?

Abstand der Grenzen der Box

Berechnen Sie den Interquartilsabstand des Alters in den beiden Kursen!

$$\begin{aligned} 1. \text{ Kurs : } x_{0,25} &= 18, x_{0,75} = 18 \Rightarrow Q_1 = 0 \\ 2. \text{ Kurs : } x_{0,25} &= 17, x_{0,75} = 19 \Rightarrow Q_1 = 2 \end{aligned}$$

Die einfachste Größe für die Streuung bei metrischen Daten ist die Spannweite

Spannweite (Range)

Die Spannweite ist der Unterschied zwischen größtem und kleinstem Wert

Die Formel lautet folglich:

$$R = x_{(n)} - x_{(1)}$$

Welchen Teil des Boxplots betrachten wir bei der Spannweite?

Abstand der Schnurrhaare

Berechnen Sie die Spannweite des Alters in den beiden Kursen!

1. Kurs : $x_{(25)} = 19, x_{(1)} = 17 \Rightarrow R_1 = 19 - 17 = 2$
 2. Kurs : $x_{(25)} = 20, x_{(1)} = 16 \Rightarrow R_2 = 20 - 16 = 4$

Mittlere absolute Abweichung

Dies ist einfach die durchschnittliche Abweichung vom arithmetischen Mittel:

$$MAD_X = \frac{\sum_{k=1}^l n_k \cdot |x_k - \bar{x}|}{n} \quad (\text{MAD steht hier für „mean absolute deviation“})$$

Für den ersten Kurs ergibt sich:

Alter	Anzahl in Kurs 1	Alter-Mittel	Gewichtet mit Anzahl, also Alter-Mittel · Anzahl
16	0	$ 16 - 18 = 2$	$2 \cdot 0 = 0$
17	2	$ 17 - 18 = 1$	$1 \cdot 2 = 2$
18	21	$ 18 - 18 = 0$	$0 \cdot 21 = 0$
19	2	$ 19 - 18 = 1$	$1 \cdot 2 = 2$
20	0	$ 20 - 18 = 2$	$2 \cdot 0 = 0$
	$\bar{x} = 18$		Summe: 4
			$MAD_X = \frac{\text{Summe}}{\text{Gesamtzahl}} = \frac{4}{25} = 0,16$

3.2.2 Übung

Berechnen Sie auf dem gleichen Weg die MAD des Alters in dem zweiten Kurs!

3.2.3 Lösung

Alter	Anzahl in Kurs 2	$ \text{Alter-Mittel} $	Gewichtet mit Anzahl, also $ \text{Alter-Mittel} \cdot \text{Anzahl}$
16	5	$ 16-18 =2$	$2 \cdot 5 = 10$
17	5	$ 17-18 =1$	$1 \cdot 5 = 5$
18	5	$ 18-18 =0$	$0 \cdot 5 = 0$
19	5	$ 19-18 =1$	$1 \cdot 5 = 5$
20	5	$ 20-18 =2$	$2 \cdot 5 = 10$
	$\bar{x} = 18$		Summe: 30
			Geteilt durch Anzahl $MAD_x = 1,2$

3.2.4 Varianzverschiebungssatz („Satz von Steiner“)

Die quadratischen Abweichungen lassen sich auf dem Taschenrechner mit Klammern nur sehr schwer eintippen. Es gibt aber einen einfacheren Weg, der zu den gleichen Ergebnissen führt... Nehmen wir uns als Beispiel das Alter (X) in einem Kurs:

k	1	2
x_k	58	18
n_k	1 (Lehrer)	19 (Schüler)

$$\bar{x} = 20, \text{Var}(X) = \frac{1 \cdot (58 - 20)^2 + 19 \cdot (18 - 20)^2}{20} = 76$$

Wenn es sehr viele Altersgruppen gibt, wird die Berechnung aber ein großer Aufwand. Wir können die Formel der Varianz aber vereinfachen – Zur Vereinfachung leiten wir die Formel für die ungewichtete Varianz her. Sie gilt aber genauso für die gewichtete Varianz... Es geht in mehreren Schritten los...

$$1. \text{Var}(X) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$$

$$2. \text{ Binomische Formel auflösen: } \sum_{k=1}^n \frac{1}{n} \cdot (x_k - \bar{x})^2 = \sum_{k=1}^n \left(\frac{1}{n} \cdot x_k^2 - \frac{1}{n} \cdot 2x_k \bar{x} + \frac{1}{n} \cdot \bar{x}^2 \right)$$

3. Statt der Gesamtsumme die Summe jedes einzelnen Teils:

$$\sum_{k=1}^n \left(\frac{1}{n} \cdot x_k^2 - \frac{1}{n} \cdot 2x_k \bar{x} + \frac{1}{n} \cdot \bar{x}^2 \right) = \sum_{k=1}^n \frac{1}{n} \cdot x_k^2 - \sum_{k=1}^n \frac{1}{n} \cdot 2x_k \bar{x} + \sum_{k=1}^n \frac{1}{n} \cdot \bar{x}^2$$

$$\text{Es ist } \sum_{k=1}^n \frac{1}{n} \cdot x_k^2 = \overline{x^2} \text{ und } \sum_{k=1}^n \frac{1}{n} \bar{x}^2 = \frac{1}{n} \sum_{k=1}^n \bar{x}^2 = \frac{1}{n} \cdot n \cdot \bar{x}^2 = \bar{x}^2.$$

$$\text{Betrachte nun auch noch } \frac{1}{n} \sum_{k=1}^n 2x_k \cdot \bar{x} = 2\bar{x} \cdot \frac{1}{n} \sum_{k=1}^n x_k = 2\bar{x} \cdot \bar{x} = 2\bar{x}^2$$

$$4. \text{ Insgesamt ergibt sich somit: } \text{Var}(X) = \overline{x^2} - 2 \cdot \bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2$$

Satz von Steiner

$$\text{Es gilt die Formel } \sigma_X^2 = \text{Var}(X) = \overline{x^2} - \bar{x}^2$$

Zur Statistik:

k	x_k	x_k^2	n_k
1	18	324	1
2	58	3364	19
$\emptyset$	$\bar{x} = 20$	$\overline{x^2} = 476$	

$$\text{Var}(X) = 476 - 20^2 = 76$$

3.2.5 Übungen

1. Die Polizei kontrolliert an zwei aufeinanderfolgenden Wochenenden eine Bar, da es eine Anzeige gab, dass Minderjährige diese häufig bis spät in die Nacht besuchen.

Es ergaben sich die folgenden Werte:

Alter	Wochenende 1	Wochenende 2
16	20	10
18	10	5
20	20	35

Berechnen Sie die Spannweite, den Quartilsabstand, die mittlere absolute Abweichung, Varianz und Standardabweichung des Alters am ersten und am zweiten Wochenende.

2. Es wurde bei einer Besuchergruppe der Party eine Alterskontrolle gemacht. Es ergaben sich die folgenden Ergebnisse:

Alter	20	18	14	16	12
-------	----	----	----	----	----

Berechnen Sie die Spannweite, Varianz und Standardabweichung des Alters.

3. Erneut ist die Statistik der Größenverteilung gegeben:

Größe	160	165	170	175	180	185	190
Anteil	5%	10%	10%	30%	20%	10%	15%

Es war $\bar{x} = 177$.

Berechnen Sie die Spannweite, die mittlere absolute Abweichung, Varianz und Standardabweichung.

4. Eine abschließende Übung noch... Eine erneute Alterskontrolle, diesmal bei fünf Personen:

Alter	26	18	18	18
-------	----	----	----	----

Berechnen Sie Varianz und Standardabweichung.

5. Eine Befragung zur Geschwisterzahl ergab bei einer Stichprobe von

Geschwister	0	0	2	0	3
-------------	---	---	---	---	---

Berechnen Sie Varianz und Standardabweichung. Ermitteln Sie, welcher Anteil der Werte im Intervall von 1,5 Standardabweichungen um das arithmetische Mittel liegt.

3.2.6 Lösungen

1. Die Polizei kontrolliert an zwei aufeinanderfolgenden Wochenenden eine Bar, da es eine Anzeige gab, dass Minderjährige diese häufig bis spät in die Nacht besuchen.

Es ergaben sich die folgenden Werte:

Alter	Wochenende 1	Wochenende 2
16	20	10
18	10	5
20	20	35

$$R_1 = 20 - 16 = 4$$

$$x_{0,25} = 16, x_{0,75} = 20 \Rightarrow Q_1 = 20 - 16 = 4$$

$$\bar{x}_1 = 18$$

$$MAD_1 = \frac{20 \cdot |16 - 18| + 10 \cdot |18 - 18| + 20 \cdot |20 - 18|}{50} = 1,6$$

$$s_1^2 = \frac{20 \cdot (16 - 18)^2 + 10 \cdot (18 - 18)^2 + 20 \cdot (20 - 18)^2}{50} = 3,2$$

$$\text{Varianzverschiebungssatz: } \overline{x^2} = \frac{20 \cdot 16^2 + 10 \cdot 18^2 + 20 \cdot 20^2}{50} = 327,2 \Rightarrow s_1^2 = 327,2 - 18^2 = 3,2$$

$$s_1 = \sqrt{3,2} \approx 1,788$$

$$R_2 = 20 - 16 = 4$$

$$x_{0,25} = 18, x_{0,75} = 20 \Rightarrow Q_1 = 20 - 18 = 2$$

$$\bar{x}_2 = 19$$

$$MAD_2 = \frac{10 \cdot |16 - 19| + 5 \cdot |18 - 19| + 35 \cdot |20 - 19|}{50} = 1,4$$

$$s_2^2 = \frac{10 \cdot (16 - 19)^2 + 5 \cdot (18 - 19)^2 + 35 \cdot (20 - 19)^2}{50} = 2,6$$

$$\text{Varianzverschiebungssatz: } \overline{x^2} = \frac{10 \cdot 16^2 + 5 \cdot 18^2 + 35 \cdot 20^2}{50} = 363,6 \Rightarrow s_1^2 = 363,6 - 19^2 = 2,6$$

$$s_2 = \sqrt{2,6} \approx 1,612$$

2. Alterskontrolle

$$R = 20 - 12 = 8$$

$$\bar{x} = 16$$

$$MAD = 2,4$$

$$s^2 = \frac{(12 - 16)^2 + (14 - 16)^2 + (16 - 16)^2 + (18 - 16)^2 + (20 - 16)^2}{5} = 8 \Rightarrow s = \sqrt{8} \approx 2,828$$

Varianzverschiebungssatz

$$\bar{x} = 16, \overline{x^2} = \frac{12^2 \cdot 1 + 14^2 \cdot 1 + 16^2 \cdot 1 + 18^2 \cdot 1 + 20^2 \cdot 1}{5} = 264$$

$$s^2 = 264 - 16^2 = 8$$

3. Erneut ist die Statistik der Größenverteilung gegeben:

Größe	160	165	170	175	180	185	190
-------	-----	-----	-----	-----	-----	-----	-----

Anteil	5%	10%	10%	30%	20%	10%	15%
--------	----	-----	-----	-----	-----	-----	-----

Es war $\bar{x} = 177$.

$$R = 190 - 160 = 30$$

$$MAD = 5\% \cdot |160 - 177| + 10\% \cdot |165 - 177| + \dots + 15\% \cdot |190 - 177| = 6,7$$

$$s^2 = 5\% \cdot (160 - 177)^2 + 10\% \cdot (165 - 177)^2 + \dots + 15\% \cdot (190 - 177)^2 = 68,5$$

$$s = \sqrt{68,5} \approx 8,28$$

Varianzverschiebungssatz

$$\overline{x^2} = \frac{160^2 \cdot 5\% + 165^2 \cdot 10\% + 170^2 \cdot 10\% + \dots + 190^2 \cdot 15\%}{1} = 31397,5$$

$$s^2 = 31397,5 - 177^2 = 68,5$$

4. Eine abschließende Übung noch... Eine erneute Alterskontrolle, diesmal bei fünf Personen:

Alter | 26 | 18 | 18 | 18

$$\bar{x} = \frac{80}{4} = 20, \overline{x^2} = \frac{26^2 + 3 \cdot 18^2}{4} = \frac{1648}{4} = 412$$

$$s^2 = 412 - 20^2 = 12 \Rightarrow s = \sqrt{12} \approx 3,46$$

5. Eine Befragung zur Geschwisterzahl ergab bei einer Stichprobe von

Geschwister | 0 | 0 | 2 | 0 | 3

Berechnen Sie Varianz und Standardabweichung. Ermitteln Sie, welcher Anteil der Werte im Intervall von 1,5 Standardabweichungen um das arithmetische Mittel liegt.

						Arit. Mittel
x	0	0	2	0	3	1
x ²	0	0	4	0	9	2,6

$$s^2 = 2,6 - 1^2 = 1,6 \Rightarrow s = \sqrt{1,6} \approx 1,26$$

$$[\bar{x} \pm 1,5s] = [1 \pm 1,5 \cdot 1,26] = [-0,89; 2,89]$$

$$[0; 2] \Rightarrow \text{Anteil} = \frac{4}{5} = 80\%$$

4 Kennzahlen klassierter Daten

Bei intervallskalierten Daten liegt nicht fest, wie oft ein konkreter Wert vorkommt, sondern lediglich, wie häufig der Wert in einem bestimmten Bereich lag.

Als Beispiel verwenden wir Werte des Alters der Männer aus Kapitel 2.9, S. 19:

Bereich	]15;18]	]18;20]	]20;22]
Anteil	40%	50%	10%

Es wird davon ausgegangen, dass die Werte innerhalb der Klassen gleichmäßig verteilt sind.

4.1 Quantile

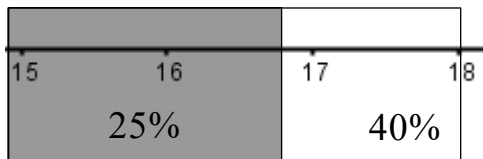
Wenn die Quantile bestimmt werden, nutzen wir die Gleichverteilung aus.

Beispiel:

Das 1. Quartil (25%-Quantil) wird bestimmt, indem wir bei der ersten Säule von den 40% nur 25% nutzen. Es wird von der Strecke 15-18 also nur $\frac{25\%}{40\%} = 0,625$ verwendet.

Das erste Quartil ist also:

$$x_{0,25} = 15 + \frac{25}{40} \cdot (18 - 15) = 16,875$$



Wird ein Quantil gesucht, das sich nicht mehr in der ersten Klasse befindet, wird analog vorgegangen, bspw.

$$x_{0,50} = 18 + \frac{10}{50} \cdot (20 - 18) = 18,4$$

(die 10 ergeben sich daher, dass in der ersten Klasse bereits 40% der Werte liegen, aus der zweiten Klassen also nur noch 10% verwendet werden.)

4.2 Lagemaße

Da wir davon ausgehen, dass die Daten innerhalb jeder Klasse gleichmäßig verteilt sind, wird zur Bestimmung der Lagemaße die Klassenmitte verwendet.

Bereich	]15;18]	]18;20]	]20;22]
Anteil	40%	50%	10%

4.2.1 Modus

Der Modus ist die Klassenmitte der Klasse, die am häufigsten vorkommt, also $\hat{x} = 19$.

4.2.2 Median

Der Median wird mit dem 50%-Quantil gleichgesetzt, liegt also bei (siehe oben)

$$x_{0,50} = 18 + \frac{10}{50} \cdot (20 - 18) = 18,4$$

4.2.3 Arithmetisches Mittel

Beim arithmetischen Mittel werden wiederum die Klassenmitten verwendet, es ergibt sich daher:

$$\bar{x} = \frac{16,5 \cdot 40\% + 19 \cdot 50\% + 21 \cdot 10\%}{100\%} = 16,5 \cdot 40\% + 19 \cdot 50\% + 21 \cdot 10\% = 18,2$$

4.3 Streuungsmaße

Bei der Streuung wird analog vorgegangen. Wir gehen weiterhin von dem obigen Beispiel aus.

4.3.1 Spannweite

Die Spannweite ist der Unterschied zwischen größtem und kleinstem Wert:

$$R = 22 - 15 = 7$$

4.3.2 Interquartilsabstand

Der Interquartilsabstand ist der Abstand zwischen erstem und drittem Quartil:

$$Q = x_{0,75} - x_{0,25} = \left(18 + \frac{35\%}{50\%} \cdot 2\right) - \left(15 + \frac{25\%}{40\%} \cdot 3\right) = 19,4 - 16,875 = 2,525$$

4.3.3 Mittlere absolute Abweichung

Zur Berechnung der mittleren absoluten Abweichung werden wieder die Klassenmitten verwendet:

$$MAD = 40\% \cdot |16,5 - 18,2| + 50\% \cdot |19 - 18,2| + 10\% \cdot |21 - 18,2| = 1,36$$

4.3.4 Varianz und Standardabweichung

Auch an dieser Stelle werden die Klassenmitten verwendet:

$$s^2 = 40\% \cdot (16,5 - 18,2)^2 + 50\% \cdot (19 - 18,2)^2 + 10\% \cdot (21 - 18,2)^2 = 2,26$$

$$s = \sqrt{2,26} \approx 1,503$$

4.4 Übungen

Übung 1

In der Statistik-Vorlesung wird eine Umfrage zum Monatseinkommen der Vorlesungsteilnehmer gemacht. Es werden folgende Antworten gegeben:

Einkommen	$[0; 500[$	$[500; 1.500[$	$[1.500; 2.500[$
Anteil	30%	60%	10%

Bestimmen Sie

1. alle Quartile,
 2. alle Mittelwerte,
 3. Spannweite,
 4. Interquartilsabstand,
 5. Varianz und Standardabweichung
- dieser Werte.

Übung 2

Im Vorjahr wurden 50 Personen befragt. Es ergaben sich die folgenden Ergebnisse:

Einkommen	$[0; 500[$	$[500; 1.500[$	$[1.500; 2.500[$
Anzahl	10	25	15

Bestimmen Sie auch hier

1. alle Quartile,
 2. alle Mittelwerte,
 3. Spannweite,
 4. Interquartilsabstand,
 5. Varianz und Standardabweichung
- dieser Werte.

4.5 Lösungen

Übung 1

1. alle Quartile

$$x_{0,25} = \frac{25\%}{30\%} \cdot 500 \approx 416,67, x_{0,50} = 500 + \frac{20\%}{60\%} \cdot 1.000 \approx 833,33,$$

$$x_{0,75} = 500 + \frac{45\%}{60\%} \cdot 1.000 = 1.250, x_{1,00} = 2.500$$

2. alle Mittelwerte,

$$\dot{x} = 1.000,$$

$$\tilde{x} = x_{0,5} \approx 833,33,$$

$$\bar{x} = 30\% \cdot 250 + 60\% \cdot 1.000 + 10\% \cdot 2.000 = 875$$

3. Spannweite

$$R = 2.500 - 0 = 2.500$$

4. Interquartilsabstand

$$Q = x_{0,75} - x_{0,25} = 1.250 - 416,67 = 833,33$$

5. Varianz und Standardabweichung

$$s^2 = 30\% \cdot (250 - 875)^2 + 60\% \cdot (1.000 - 875)^2 + 10\% \cdot (2.000 - 875)^2 = 253.125$$

$$s = \sqrt{253.125} \approx 503,12$$

Übung 2

1. alle Quartile,

$$x_{0,25} = 500 + \frac{5\%}{50\%} \cdot 1.000 = 600, x_{0,50} = 500 + \frac{30\%}{50\%} \cdot 1.000 = 1.100,$$

$$x_{0,75} = 1.500 + \frac{5\%}{30\%} \cdot 1.000 \approx 1.666,67, x_{1,00} = 2.500$$

2. alle Mittelwerte,

$$\dot{x} = 1.000,$$

$$\tilde{x} = x_{0,5} = 1.100,$$

$$\bar{x} = 20\% \cdot 250 + 50\% \cdot 1.000 + 30\% \cdot 2.000 = 1.150$$

3. Spannweite

$$R = 2.500 - 0 = 2.500$$

4. Interquartilsabstand

$$Q = x_{0,75} - x_{0,25} = 1.666,67 - 600 = 1.066,67$$

5. Varianz und Standardabweichung

$$s^2 = 20\% \cdot (250 - 1.150)^2 + 50\% \cdot (1.000 - 1.150)^2 + 30\% \cdot (2.000 - 1.150)^2 = 390.000$$

$$s = \sqrt{390.000} \approx 624,5$$

5 Zusammenhänge zweier Merkmale

5.1 Zusammenhang zweier nominaler Merkmale

5.1.1 Beispiel

Es soll untersucht werden, inwiefern es bei Studenten einen Zusammenhang zwischen der guten Note in Statistik und der Motivation für Statistik gibt.

These: Wer eine hohe Motivation hat, schreibt auch eine gute Note und wer eine gute Note schreibt, hat wiederum eine hohe Motivation.

Aber dies sollte überprüft werden. Hierzu wurde eine Umfrage unter 100 Studenten gemacht, die zu folgendem Ergebnis führt:

	Hohe Motivation	Geringe Motivation	Summe
Gute Note	65	10	75
Schlechte Note	5	20	25
Summe	70	30	100

Hätten Motivation und Note nichts miteinander zu tun (wir werden dies später noch „stochastisch unabhängig“ nennen), dann würden Personen mit hoher Motivation mit einer Wahrscheinlichkeit von $\frac{75}{100} = 75\%$ eine gute Note schreiben, also wäre die Wahrscheinlichkeit für eine

gute Note und hohe Motivation $52,5\%$ ($75\% \cdot 70\% = 52,5\%$) und auch bei Personen mit geringer Motivation wäre die Wahrscheinlichkeit für gute Noten 75% , also würden mit einer Wahrscheinlichkeit von $22,5\%$ ($75\% \cdot 30\% = 22,5\%$) Personen mit geringer Motivation eine gute Note schreiben. Die Randwahrscheinlichkeiten werden also miteinander multipliziert.

Es ergibt sich als Berechnung das Zusammenhangsmaß Chi-Quadrat:

$$\chi^2 = N \cdot \left(\sum_{i=1}^n \sum_{j=1}^m \left(\frac{n_{ij}^2}{n_i \cdot n_j} \right) - 1 \right), \text{ wobei } i \text{ die Zeilen und } j \text{ die Spalten zählt.}$$

N ist die Gesamtzahl der Befragten, hier also 100.

In unserem Beispiel wäre dies also:

$$\chi^2 = 100 \cdot \left(\frac{65^2}{70 \cdot 75} + \frac{10^2}{30 \cdot 75} + \frac{5^2}{70 \cdot 25} + \frac{20^2}{30 \cdot 25} - 1 \right) = 100 \cdot (0,805 + 0,044 + 0,014 + 0,533 - 1) = 39,6$$

Cramers V normiert dieses Zusammenhangsmaß

$$V = \sqrt{\frac{\chi^2}{N \cdot (\min(n, m) - 1)}}, 0 \leq V \leq 1$$

Für unser Beispiel ergibt sich also:

$$V = \sqrt{\frac{39,6}{100 \cdot (2 - 1)}} \approx 0,629$$

Cramers V ist normiert und nimmt nur Werte im Intervall $[0;1]$ an.

$V = 0$	kein Zusammenhang
$0 < V \leq 0,3$	schwacher Zusammenhang
$0,3 < V \leq 0,7$	mittelstarker Zusammenhang
$0,7 < V < 1$	starker Zusammenhang
$V = 1$	völliger Zusammenhang

5.1.2 Auswertung

Es besteht also ein mittelstarker Zusammenhang zwischen Motivation und Note.
(Man sieht schon, dass es bei hoher Motivation eher gute Noten gibt als bei geringer Motivation.)

5.1.3 Übung

Unser Personalleiter gibt uns folgende Werte bzgl. unserer Mitarbeiter:

	Ausbildung	Hochschulabschluss	Summe
Mitglied der Geschäftsleitung	0	4	4
Kein Mitglied der Geschäftsleitung	15	1	16
Summe	15	5	20

Besteht ein Zusammenhang zwischen Ausbildung und der Chance, in die Geschäftsleitung zu kommen?

5.1.4 Lösung

$$\chi^2 = 20 \cdot \left(\frac{0^2}{15 \cdot 4} + \frac{4^2}{5 \cdot 4} + \frac{15^2}{15 \cdot 16} + \frac{1^2}{5 \cdot 16} - 1 \right) = 20 \cdot \left(\frac{7}{4} - 1 \right) = 15$$

$$V = \sqrt{\frac{15}{20 \cdot (2-1)}} = \sqrt{\frac{3}{4}} \approx 0,866$$

Es besteht also ein starker Zusammenhang.

5.2 Zusammenhang zweier metrischer Merkmale

5.2.1 Kovarianz

„Thirsty Donkey“ – Ein Kiosk in Frankfurt.

Wie ist der Zusammenhang zwischen Preis und abgesetzter Menge für unseren Softdrink „Frosty Star“?

Wir führen ein Preisexperiment durch:

X: Preis	2,5	3	3,5	4	4,5
Y: Menge	50	49	45	46	40

Um den Zusammenhang zwischen Preis und Menge zu berechnen, wird die „Kovarianz“ berechnet:

Idee: Schau, ob die Veränderung immer in die gleiche Richtung geht.

(Steigt der Preis, steigt auch die Menge oder nicht? Anders gesagt: Ist die Menge größer als der Mittelwert, wenn auch der Preis größer als der Mittelwert ist?)

Kovarianz: Zusammenhang zwischen zwei Merkmalen

$$\text{Cov}(X, Y) = s_{XY} = \frac{\sum_{k=1}^n n_k (x_k - \bar{x})(y_k - \bar{y})}{n}$$

Aber auch hier gibt es eine Art „Varianzverschiebungssatz“:

$$Cov(X, Y) = s_{XY} = \frac{\sum_{k=1}^n n_k (x_k - \bar{x})(y_k - \bar{y})}{n}$$

$$= \sum_{k=1}^n \frac{n_k}{n} (x_k \cdot y_k - x_k \cdot \bar{y} - \bar{x} \cdot y_k + \bar{x} \cdot \bar{y}) = \overline{xy} - \bar{x} \cdot \bar{y} - \bar{x} \cdot \bar{y} + \bar{x} \cdot \bar{y} = \overline{xy} - \bar{x} \cdot \bar{y}$$

Wie groß ist nun der Wert der Kovarianz? Ist die Kovarianz positiv oder negativ?

$$s_{XY} = \overline{xy} - \bar{x} \cdot \bar{y}$$

$$\bar{x} = 3,5, \bar{y} = 46, \overline{xy} = 158,7 \Rightarrow s_{XY} = \overline{xy} - \bar{x} \cdot \bar{y} = -2,3$$

negativ (klar!)

Es gilt:

$$Cov(X, Y) \begin{cases} > 0 & \text{positiv} \quad (x \text{ steigt} \Leftrightarrow y \text{ steigt}, x \text{ fällt} \Leftrightarrow y \text{ fällt}) \\ = 0 & \text{keine Abhängigkeit} \\ < 0 & \text{negativ} \quad (x \text{ steigt} \Leftrightarrow y \text{ fällt}, x \text{ fällt} \Leftrightarrow y \text{ steigt}) \end{cases}$$

Stellen Sie sich nun vor, es wären nicht nur die Leute in diesem Kurs sondern alle Schüler der Schule befragt worden und es hätten sich die gleichen Ergebnisse ergeben – allerdings hätte jeweils die 10-fache Anzahl an Personen die Cola gekauft. Der Zusammenhang zwischen Preis und Käuferverhalten wäre identisch, aber welcher Wert ergibt sich für die Kovarianz?

$$s_{XY} \approx 10 \cdot (-2,3) = -23 \quad (\text{da jeder } y\text{-Wert verzehnfacht})$$

Wenn man einschätzen möchte, ob die Kovarianz stark oder schwach ausgeprägt ist, muss man, wie bei der Temperaturskala, Maximum und Minimum festlegen.

Hierfür muss sie zuerst umgerechnet werden.

Das Ergebnis ist der Korrelationskoeffizient.

Vorteil: Es macht keinen Unterschied, ob bspw. das Einkommen in € oder in 1000 € gemessen wird.

5.2.2 Korrelationskoeffizient nach Bravais-Pearson

Auf eine Skala von -1 bis +1 normierter Zusammenhang zwischen zwei Merkmalen

$$Korr(X, Y) = r_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{s_{XY}}{s_X \cdot s_Y}, -1 \leq r_{XY} \leq 1$$

Berechnen Sie den Korrelationskoeffizienten für die Beziehung zwischen Preis und verkauften Stück, wenn die Beispielzahlen verwendet werden.

$$\begin{aligned}s_X^2 &= 0,5, s_Y^2 = 12,4 \\ s_X &\approx 0,707, s_Y \approx 3,521 \\ r_{XY} &\approx \frac{-2,3}{12,4 \cdot 3,521} \approx -0,924\end{aligned}$$

Man kann nun also sagen: Je näher der Korrelationskoeffizient an -1 ist, desto stärker negativ, je näher er an +1 ist, desto stärker positiv ist der Zusammenhang. Ist der Korrelationskoeffizient gleich 0, so sind die Merkmale komplett unabhängig voneinander.

Ist die Korrelation stark, schwach, positiv oder negativ?

Sehr stark, da sehr nahe am Minimalwert von -1.

-

5.3 Zusammenhang zweier ordinaler Merkmale

Manchmal soll auch der Zusammenhang ordinaler Merkmale untersucht werden.

Eigentlich darf das arithmetische Mittel bei diesen Merkmalen nicht bestimmt werden, da es ja der Wert ist, der zu allen ermittelten Merkmalswerten den kürzesten Abstand hat.

Um aber eine Aussage über den Zusammenhang ordinaler Merkmale machen zu können, wird mit diesen „Rängen“ trotzdem eine Korrelationsuntersuchung gemacht. Nach ihrem Erfinder wird das Ergebnis „Spearman'scher² Korrelationskoeffizient genannt.

Es findet eine Sportveranstaltung statt.

In der Vorauswahl treten vier Personen gegeneinander an.

Es ergeben sich die folgenden Platzierungen

Teilnehmer	1	2	3	4
X: Springen	2	1	3,5	3,5
Y: Werfen	4	3	2	1

Im Springen haben Teilnehmer drei und vier identische Weiten erzielt, daher ist nicht eine Person auf Platz drei und die andere auf Platz vier, sondern beide bekommen den Platz 3,5 zur Berechnung.

Die Berechnung des Korrelationskoeffizienten nach Spearman erfolgt in den folgenden Schritten:

1. Berechnung der arithmetischen Mittel:

$$\bar{x} = \frac{2+1+3,5+3,5}{4} = 2,5, \quad \bar{y} = \frac{4+3+2+1}{4} = 2,5$$

2. Berechnung der Standardabweichungen:

$$s_x = \sqrt{\frac{(2-2,5)^2 + (1-2,5)^2 + (3,5-2,5)^2 + (3,5-2,5)^2}{4}} = \sqrt{1,125}$$

$$s_y = \sqrt{\frac{(4-2,5)^2 + (3-2,5)^2 + (2-2,5)^2 + (1-2,5)^2}{4}} = \sqrt{1,25}$$

3. Berechnung der Kovarianz zwischen den beiden Merkmalen:

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

$$= \frac{(2-2,5) \cdot (4-2,5) + (1-2,5) \cdot (3-2,5) + (3,5-2,5) \cdot (2-2,5) + (3,5-2,5) \cdot (1-2,5)}{4}$$

$$= \frac{-3,5}{4} = -0,875$$

Alternativ via Verschiebungssatz

$$\bar{x} = \bar{y} = \frac{10}{4} = 2,5$$

$$\overline{x^2} = 7,375, \quad \overline{y^2} = 7,5$$

$$\overline{xy} = 5,375$$

$$s_x \approx 1,061, \quad s_y \approx 1,118$$

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = -0,875$$

4. Berechnung des Korrelationskoeffizienten **nach Spearman**:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{-0,875}{\sqrt{1,125} \cdot \sqrt{1,25}} \approx -0,738$$

5. Interpretation:

$r_{xy} < 0$: Negative Korrelation: Steigt das eine, fällt das andere und umgekehrt.

$r_{xy} = 0$: Unkorreliertheit: Die Entwicklungen haben keinen Zusammenhang.

$r_{xy} > 0$: Positive Korrelation: Steigt das eine, steigt auch das andere; sinkt das eine, sinkt auch das andere.

² Charles Spearman (1863-1945), englischer Psychologe, der mit seinen Arbeiten in der Statistik bekannt wurde.

Hinweis: Es wird hier das arithmetische Mittel genommen, obwohl die Werte nur ordinal skaliert sind. Dies geschieht, da der Median die Interpretation von Abständen nicht zulässt, man aber hier die Abstände interpretieren muss, obwohl dies theoretisch nicht ganz korrekt ist.

5.4 Aufgaben

1. In einer Umfrage wurden 5 Personen nach ihrem Alter und ihrer wöchentlichen Konsummenge von Schokolade (Tafeln/Woche) befragt.

Es ergaben sich die folgenden Ergebnisse

Person Nr.	1	2	3	4	5
X: Alter	18	54	36	85	46
Y: Konsummenge	4	2	2	1	2

Bestimmen Sie den Korrelationskoeffizienten nach Bravais-Pearson.

2. Bei der Umfrage aus Aufgabe 1 taucht noch ein weiteres Befragungsergebnis auf. Somit ergibt sich die folgende Übersicht:

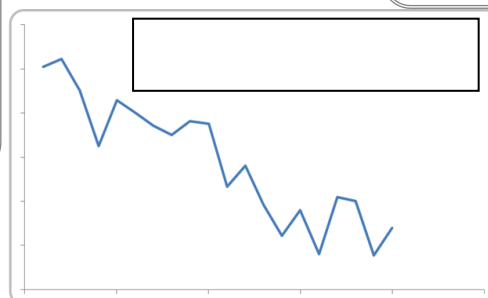
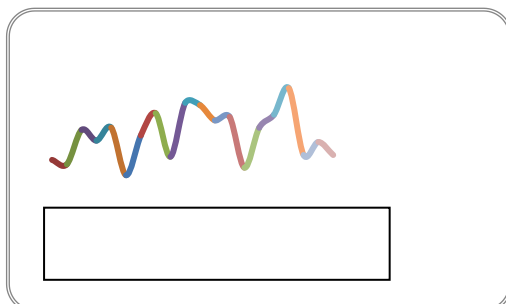
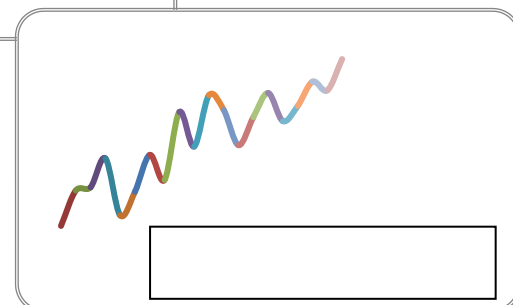
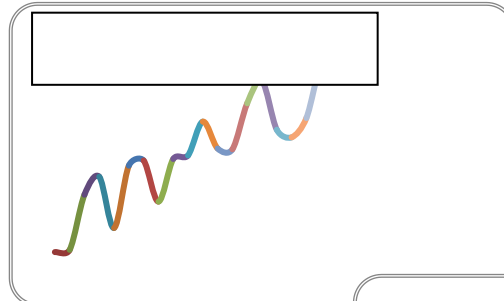
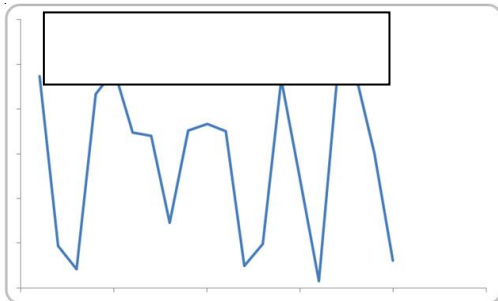
Person	1	2	3	4	5	6
X: Alter	18	54	36	85	46	89
Y: Konsummenge	4	2	2	1	2	1

Bestimmen Sie erneut den Korrelationskoeffizienten nach Bravais-Pearson.

3. Alles verstanden? Hier noch eine Aufgabe zum Check:

Person	1	2	3	4
X: Alter	21	25	29	25
Y: Kinderzahl	0	1	3	0

4. Welche der folgenden Diagramme stellt positive oder negative Korrelation dar? Welche stellen Unkorreliertheit dar? Eine Größe ist auf der x-, die andere auf der y-Achse abgetragen.



5.5 Lösungen

1. Schokoladenkonsum I

Person Nr.	1	2	3	4	5
X: Alter	18	54	36	85	46
Y: Konsummenge	4	2	2	1	2

$$\bar{x} = 47,8; \overline{x^2} = 2775,4; \bar{y} = 2,2; \overline{y^2} = 5,8$$

$$\overline{xy} = 85,8$$

$$s_{XY} = \overline{xy} - \bar{x} \cdot \bar{y} = -19,36$$

$$s_X = \sqrt{\overline{x^2} - \bar{x}^2} \approx 22,149; s_Y = \sqrt{\overline{y^2} - \bar{y}^2} \approx 0,98$$

$$r_{XY} \approx \frac{-19,36}{22,149 \cdot 0,98} \approx -0,892$$

2. Schokoladenkonsum II

Person	1	2	3	4	5	6
X: Alter	18	54	36	85	46	89
Y: Konsummenge	4	2	2	1	2	1

$$\bar{x} \approx 54,667; \overline{x^2} = 3633$$

$$\bar{y} = 2; \overline{y^2} = 5$$

$$\overline{xy} \approx 86,333; s_{XY} = \overline{xy} - \bar{x} \cdot \bar{y} = -23$$

$$s_X = \sqrt{\overline{x^2} - \bar{x}^2} \approx 25,388; s_Y = \sqrt{\overline{y^2} - \bar{y}^2} \approx 1$$

$$r_{XY} \approx \frac{-23}{25,388 \cdot 1} \approx -0,906$$

3. Kinderzahl

Person	1	2	3	4
X: Alter	21	25	29	25
Y: Kinderzahl	0	1	3	0

$$\bar{x} = 25; \overline{x^2} = 633$$

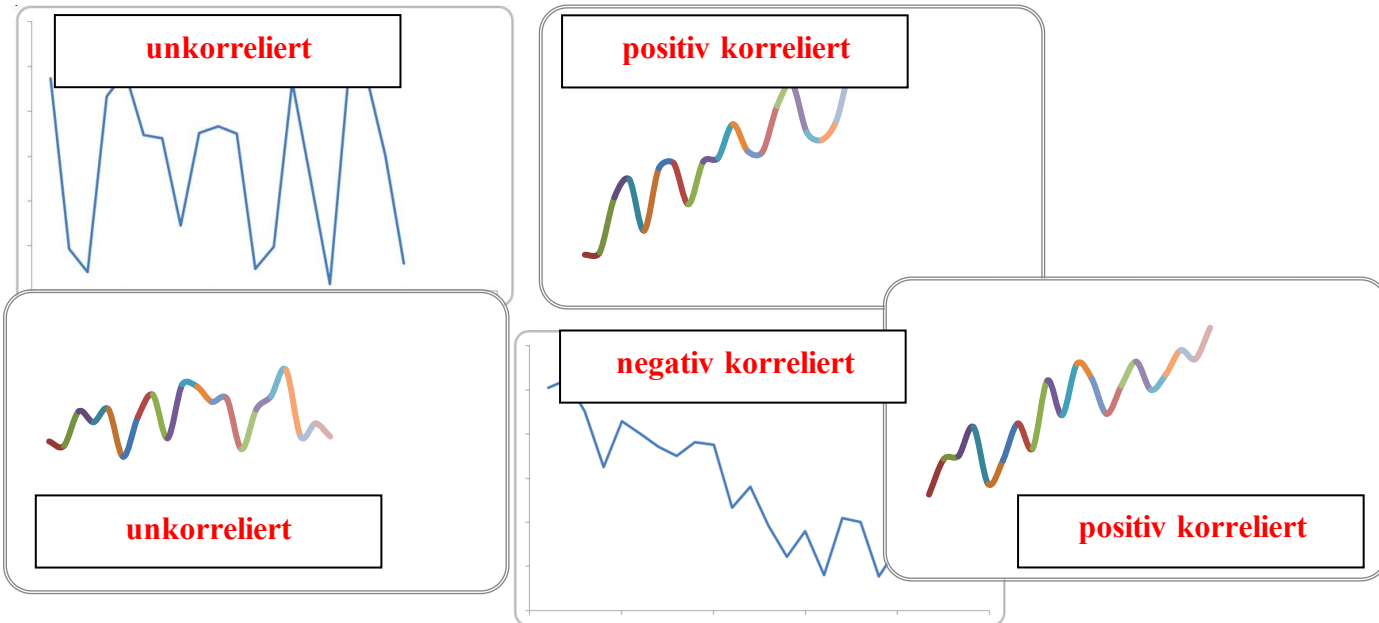
$$\bar{y} = 1; \overline{y^2} = 2,5$$

$$\overline{xy} = 28; s_{XY} = \overline{xy} - \bar{x} \cdot \bar{y} = 3$$

$$s_X = \sqrt{\overline{x^2} - \bar{x}^2} \approx 2,828; s_Y = \sqrt{\overline{y^2} - \bar{y}^2} \approx 1,22$$

$$r_{XY} \approx \frac{3}{2,828 \cdot 1,22} \approx 0,87$$

4. Welche der folgenden Diagramme stellt positive oder negative Korrelation dar? Welche stellen Unkorreliertheit dar? Eine Größe ist auf der x-, die andere auf der y-Achse abgetragen.



6 Regression

6.1 Einführungsbeispiel

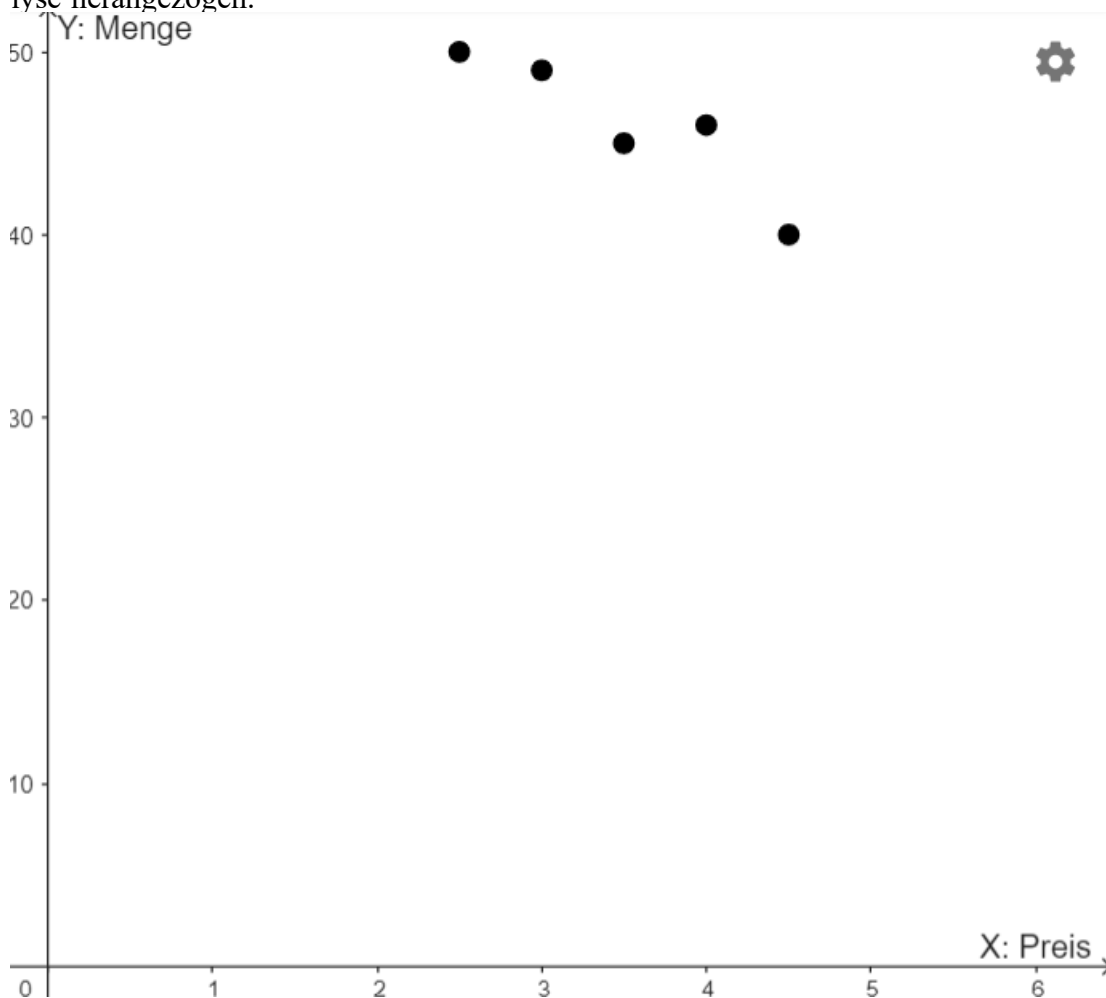
Zurück zu unserem Beispiel aus dem Abschnitt Korrelation:

X: Preis	2,5	3	3,5	4	4,5
Y: Menge	50	49	45	46	40

Frage: Welche Menge ist bei einem Preis von 5 € zu erwarten?

Ziel ist es zu erfahren, wie stark sich eine Preiserhöhung auf den Umsatz auswirkt.

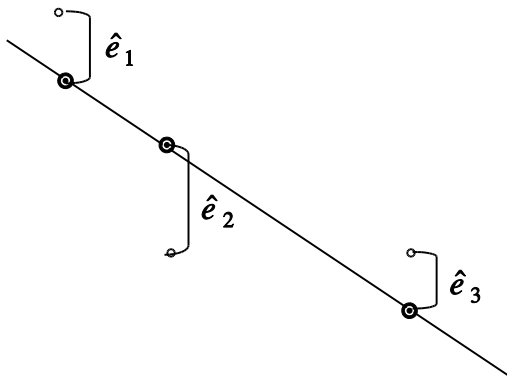
Um die genaue Abhängigkeit zwischen einer Einflussgröße auf eine andere Größe zu untersuchen, um also eine lineare Funktionsgleichung aufstellen zu können, wird die Regressionsanalyse herangezogen.



In einem ersten Schritt müssen wir die ermittelten Werte in ein Koordinatensystem eintragen.

Um uns die Sache möglichst einfach zu machen, sollten wir versuchen, eine lineare Funktion durch die Punkte zu legen, die diese im Mittel gut beschreibt (also die Summe aller Abstände minimiert).

Nennen wir den Abstand des i . Punkts von der geschätzten Geraden einfach $\hat{e}_i$ (error), dann ergibt sich das folgende Bild:



Wie können wir aber eine Gerade bestimmen, die genau durch die Mitte geht?
Raten und ausprobieren sind eindeutig keine gute Lösung.

Die Abstände sollen möglichst klein sein. Da es zum Einen Abstände nach oben (positives Vorzeichen) und Abstände nach unten (negatives Vorzeichen) sind, müssen wir eine andere Lösung finden.

Es bieten sich zwei Lösungen an:

1. Beträge der Abstände
2. Quadrate der Abstände

Der zweite Fall hat den Charme, dass eine quadratische Funktion gut darstellbar ist, also:

$$\min S = \sum_{i=1}^n (y_i - a - bx_i)^2$$

Bestimme nun das Minimum von S !

Hinweise:

Die erste Ableitung nach jeder Variablen bilden und Nullsetzen (Tiefpunkt gesucht!).

Es gilt:
$$\frac{d \sum_{i=1}^n (y_i - a - bx_i)^2}{da} = \sum_{i=1}^n \frac{d(y_i - a - bx_i)^2}{da} \quad (\text{Summenregel})$$

$$\sum_{i=1}^n \frac{d(y_i - a - bx_i)^2}{da} = \sum_{i=1}^n 2(y_i - a - bx_i)(-1) = 0$$

$$\sum_{i=1}^n \frac{d(y_i - a - bx_i)^2}{db} = \sum_{i=1}^n \sum_{i=1}^n 2(y_i - a - bx_i)(-x_i) = 0$$

Diese Gleichungen können wir nun noch vereinfachen:

1. Gleichung mit Ableitung nach a durch n teilen und nach a auch nach a auflösen:

$$\begin{aligned} \Rightarrow \sum_{i=1}^n (y_i - a - bx_i) &= 0 \quad / : n \\ \Rightarrow \bar{y} - a - b\bar{x} &= 0 \Rightarrow \boxed{a = \bar{y} - b\bar{x}} \end{aligned}$$

2. Ableitung nach b : Klammer auflösen, vereinfachen und Ergebnis der Ableitung nach a einsetzen

$$\begin{aligned} \sum_{i=1}^n 2(y_i - a - bx_i)(-x_i) &= 0 \quad / : 2 \\ \sum_{i=1}^n (-x_i y_i + ax_i + bx_i^2) &= 0 \Rightarrow -\sum_{i=1}^n x_i y_i + \sum_{i=1}^n ax_i + \sum_{i=1}^n bx_i^2 = 0 \end{aligned}$$

Einsetzen von $a = \bar{y} - b\bar{x}$

$$-\sum_{i=1}^n x_i y_i + \sum_{i=1}^n (\bar{y} - b\bar{x})x_i + \sum_{i=1}^n bx_i^2 = 0$$

3. Nach b auflösen

$$\begin{aligned} -\sum_{i=1}^n x_i y_i + \sum_{i=1}^n \bar{y} x_i - \sum_{i=1}^n b\bar{x} x_i + \sum_{i=1}^n bx_i^2 &= 0 \\ \Rightarrow \sum_{i=1}^n (x_i \bar{y} - x_i y_i) - b \sum_{i=1}^n (x_i \bar{x} - x_i^2) &= 0 \\ \Rightarrow b \sum_{i=1}^n (x_i \bar{x} - x_i^2) &= \sum_{i=1}^n (x_i \bar{y} - x_i y_i) \\ \Rightarrow b &= \frac{\sum_{i=1}^n (x_i \bar{y} - x_i y_i)}{\sum_{i=1}^n (x_i \bar{x} - x_i^2)} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i \bar{y} - x_i y_i)}{\frac{1}{n} \sum_{i=1}^n (x_i \bar{x} - x_i^2)} = \frac{\bar{x} \cdot \bar{y} - \overline{xy}}{\bar{x} \cdot \bar{x} - \overline{x^2}} \\ &= \frac{\bar{x} \cdot \bar{y} - \overline{xy}}{\bar{x}^2 - \overline{x^2}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{Cov(X, Y)}{Var(X)} \Rightarrow b = \frac{s_{XY}}{s_X^2} \quad \boxed{b = \frac{s_{XY}}{s_X^2}} \\ \text{Wir sehen auch: } r_{XY} &= \frac{s_{XY}}{s_X \cdot s_Y} = b \cdot \frac{s_X}{s_Y} \end{aligned}$$

Benutzen Sie die Zahlen (Preise & Mengen) aus unserer Statistik: Welchen Wert hat a , welchen hat b ?

Welche Menge ist bei einem Preis von 5 € zu erwarten?

6.2 Lösung

$$\bar{x} = 3,5, \bar{y} = 46, \overline{x^2} = 12,75, \overline{xy} = 158,7$$

$$s_{xy} = 158,7 - 3,5 \cdot 46 = -2,3$$

$$s_x^2 = 12,75 - 3,5^2 = 0,5$$

$$b = \frac{-2,3}{0,5} = -4,6, a = 46 - (-4,6) \cdot 3,5 = 62,1$$

$$y = 62,1 - 4,6x$$

$$y(5) = 62,1 - 4,6 \cdot 5 = 39,1$$

Was bedeuten die Werte von a und b inhaltlich in unserem Beispiel?

a =Wie hoch ist die Menge bei einem Preis von 0 € (Sättigungsmenge)

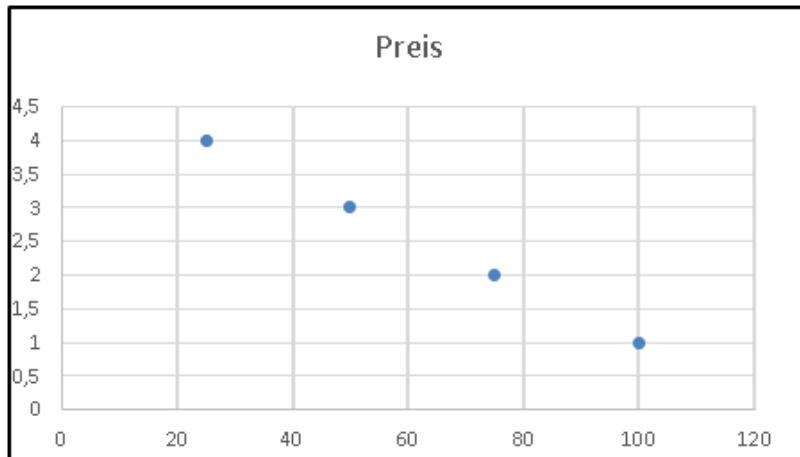
b =Um wie viel sinkt die Menge, wenn der Preis um 1 € erhöht wird? Genauer: $b=dy/dx$

6.3 Kurzübungen zur Selbstkontrolle

Bestimmen Sie jeweils die Regressionsgerade bei folgenden gegebenen Werten:
Welcher Preis ist anzusetzen, um eine Menge von 90 Stück verkaufen zu können?

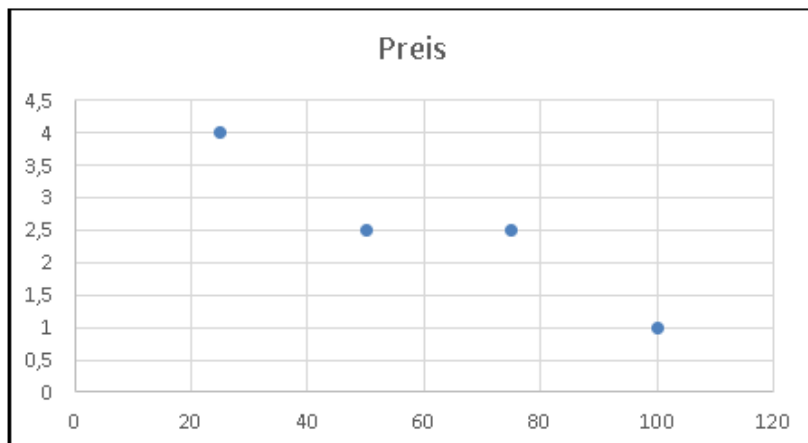
6.3.1 Übung 1

X: Menge	100	50	75	25
Y: Preis	1	3	2	4



6.3.2 Übung 2

X: Menge	100	50	75	25
Y: Preis	1	2,5	2,5	4



6.4 Lösungen

6.4.1 Übung 1*

$$\bar{y} = 2,5, \bar{x} = 62,5, \overline{x^2} = 4687,5, \overline{xy} = 125$$

$$s_{XY} = -31,25$$

$$s_X^2 = 781,25$$

$$b = -0,04, a = 5$$

$$y = 5 - 0,04x$$

$$y = 5 - 0,04 \cdot 90 = 1,40 \text{ €}$$

6.4.2 Übung 2

$$\bar{y} = 2,5, \bar{x} = 62,5, \overline{x^2} = 4.687,5, \overline{xy} = 128,125$$

$$s_{XY} = -28,125$$

$$s_X^2 = 781,25$$

$$b = -0,036, a = 4,75$$

$$y = 4,75 - 0,036x$$

$$y = 4,75 - 0,036 \cdot 90 = 1,51 \text{ €}$$

6.5 Übungen

1. Es wurde eine Umfrage zur Zahlungsbereitschaft von Kunden für Bio-Döner gemacht („rein ökologischer Gemüseanbau und Fleisch aus artgerechter Haltung“). Hierbei ergaben sich die folgenden Ergebnisse:

Preis	4	3	5	2	4	4	5	5
Menge	100	120	80	150	90	100	90	70

Bestimmen Sie die Gleichung der Regressionsgeraden. Welcher Absatz wird bei einem Preis von 4,20 € erwartet?

Wir können den Preis frei setzen, er ist also unsere Variable „x“.

2. Gegeben sind die Ergebnisse einer Untersuchung zum Einfluss der Konsummenge an Kaffee (Tassen/Tag) auf die von Tee (Tassen/Tag) der untersuchten Personen.

X: Kaffee	1	0	6	5	3	5	1	3
Y: Tee	6	7	1	2	5	2	6	3

Welcher Teekonsum ist bei einem Konsum von einer Tasse Kaffee zu erwarten?

3. Gegeben ist eine Umfrage aus der Vergangenheit zur Absatzmenge von Schokoriegeln in Abhängigkeit vom Preis. Bestimme die Regressionsgerade.

Preis	0,20	0,40	0,60	0,80	1,00	1,20	1,40	1,60	1,80	2,00
Menge	25	20	16	13	9	8	6	4	2	1

Tragen Sie die Werte in ein Koordinatensystem ein.

Woran erkennt man, dass eine Regressionsgerade nur im Mittel richtig ist, aber in einzelnen Punkten durchaus daneben liegen kann?

Glauben Sie, dass die Stichprobe tatsächlich diese Zahlen aufwies – oder wurde hier gefuscht?

Nur für ökonomisch Interessierte (nicht klausurrelevant):

Gehen Sie von Fixkosten in Höhe von 200 € und variablen Kosten in Höhe von 20 ct. aus.

Wie hoch ist in diesem Fall der gewinnmaximierende Preis?

4. Welche Bedingung gibt an, dass die Regressionsgerade durch den Schwerpunkt der Punktwolke $(\bar{x}, \bar{y})$ verläuft! Überprüfen sie mit dieser Beziehung Ihre Ergebnisse aus den Aufgaben 1-3.

6.6 Lösungen

1. Es wurde eine Umfrage zur Zahlungsbereitschaft von Kunden für Bio-Döner gemacht („rein ökologischer Gemüseanbau und Fleisch aus artgerechter Haltung“). Hierbei ergaben sich die folgenden Ergebnisse:

X: Preis	4	3	5	2	4	4	5	5
Y: Menge	100	120	80	150	90	100	90	70

Bestimmen Sie die Gleichung der Regressionsgeraden. Welcher Absatz wird bei einem Preis von 4,20 € erwartet?

Mittelwerte:

Preis (X)	4
Menge (Y)	100
X^2	17,000
$X \cdot Y$	377,500

$$y = 190 - 22,5x$$

$$y(4,20) = 95,5$$

Var(X)	1,0
Cov(X,Y)	-22,5

a=	190
b=	-22,5

2. Gegeben sind die Ergebnisse einer Untersuchung zur Konsummenge an Kaffee (Tassen/Tag) und Tee (Tassen/Tag) der untersuchten Personen.

Kaffee	1	0	6	5	3	5	1	3
Tee	6	7	1	2	5	2	6	3

Mittelwerte:

Kaffee (X)	3,00
Tee (Y)	4,00
X^2	13,25
$X \cdot Y$	7,75

Var(X)	4,250
Cov(X,Y)	-4,250

a=	7
b=	-1

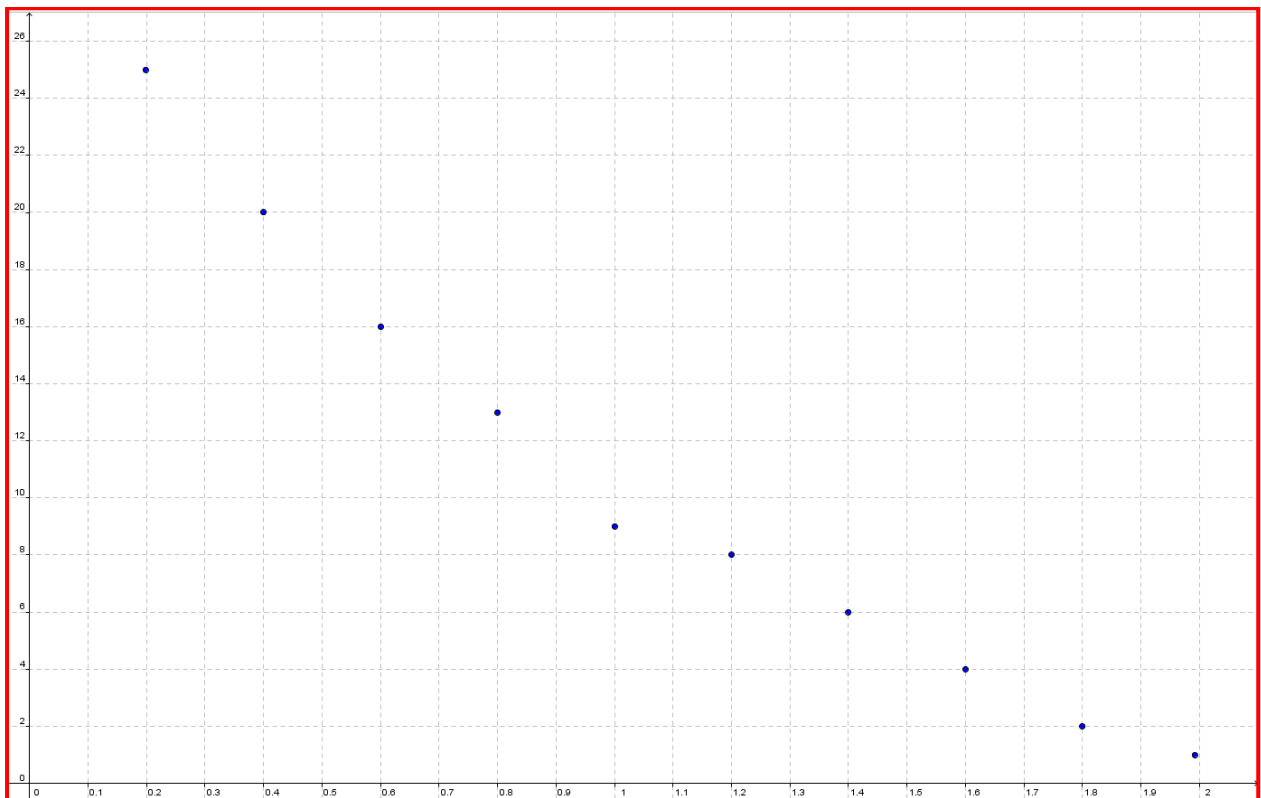
$$y = 7 - x$$

$$y(1) = 6$$

3. Gegeben ist eine Umfrage aus der Vergangenheit zur Absatzmenge von Schokoriegeln in Abhängigkeit vom Preis. Bestimme die Regressionsgerade.

Preis	0,20	0,40	0,60	0,80	1,00	1,20	1,40	1,60	1,80	2,00
Menge	25	20	16	13	9	8	6	4	2	1

Tragen Sie die Werte in ein Koordinatensystem ein.



Glauben Sie, dass die Stichprobe tatsächlich diese Zahlen aufwies – oder wurde hier gefuscht?
„Zu“ ebenmäßig... wahrscheinlich gefuscht...

X: Preis	0,2	0,4	0,6	0,8	1	1,2	1,4	1,6	1,8	2	Mittelwert:	1,1
Y: Menge	25	20	16	13	9	8	6	4	2	1	Mittelwert:	10,4
X^2:	0,04	0,16	0,36	0,64	1	1,44	1,96	2,56	3,24	4	Mittelwert:	1,54
X*Y:	5	8	9,6	10,4	9	9,6	8,4	6,4	3,6	2	Mittelwert:	7,2
b=	-12,848		a=	24,5333								

Hier wird auch deutlich: Die Regressionsgerade ist „im Mittel“ richtig, allerdings in einzelnen Punkten durchaus falsch (Menge von 25 bei Preis von 0,20 €, in der Stichprobe war bei einem Preis von Null eine kleinere Menge!)

$$y = 24,533 - 12,848p$$

Nur für ökonomisch Interessierte (nicht klausurrelevant):

Gehe von Fixkosten in Höhe von 200 € und variablen Kosten in Höhe von 20 ct. aus.

Wie hoch ist in diesem Fall der gewinnmaximierende Preis?

$$\begin{aligned}
 G &= py - 0,2y - 200 \\
 G &= p(24,533 - 12,848p) - 0,2(24,533 - 12,848p) - 200 \\
 G &= -12,848p^2 + 24,533p - 4,907 + 2,570p - 200 \\
 G &= -12,848p^2 + 27,103p - 204,907 \\
 \frac{dG}{dp} &= -25,696p + 27,103 = 0 \Rightarrow p \approx 1,055
 \end{aligned}$$

4. Welche Bedingung gibt an, dass die Regressionsgerade durch den Schwerpunkt der Punktwolke $(\bar{x}, \bar{y})$ verläuft!

Diese Tatsache ergibt sich direkt aus der Bedingung

$$\bar{y} - a - b\bar{x} = 0 \Rightarrow \bar{y} = a + b\bar{x}$$

Überprüfen Sie mit dieser Beziehung Ihre Ergebnisse aus den Aufgaben 1-3.

Aufgabe 1

$$\bar{x} = 4,25, \bar{y} = 902,1$$

$$y = 1213,61 - 73,30p \Rightarrow 902,1 \stackrel{!}{=} 1213,61 - 73,30 \cdot 4,25 \approx 902,1$$

Aufgabe 2

$$\bar{x} = 1,1, \bar{y} = 1.610,9$$

$$y = 2258,73 - 588,94p \Rightarrow 1610,90 \stackrel{!}{=} 2258,73 - 588,94 \cdot 1,1 \approx 1610,93$$

Aufgabe 3

$$\bar{x} = 1,1, \bar{y} = 10,4$$

$$y = 24,533 - 12,848p \Rightarrow 10,4 = 24,533 - 12,848 \cdot 1,1 \approx 10,4$$

6.7 Weitere Übung (bei Unsicherheit)

Bestimmen Sie jeweils die Gleichung Regressionsgerade, die den Zusammenhang zwischen den angegebenen Größen wiedergibt. Bestimmen Sie zur Übung auch jeweils den Korrelationskoeffizienten nach Bravais-Pearson.

6.7.1 Aufgabe

Untersucht wird der Teekonsum in Abhängigkeit vom Alter der Person.

Person	1	2	3	4	5	6
X: Alter	20	21	24	20	20	24
Y: Teetassen/Tag	4	6	8	3	3	6

Welcher Konsum ist bei einer Person mit 30 Jahren zu erwarten?

6.7.2 Aufgabe

Untersucht wird der Zusammenhang zwischen Masse und Benzinverbrauch von Autos.

Fahrzeug	1	2	3	4	5	6	7	8	9	10
X: Masse [100 kg]	8	7	9	11	13	15	10	18	16	15
Y: Treibstoffverbrauch	9	6	10	12	12,5	13	8	18	16	14,5

Welcher Verbrauch ist bei einer Masse von 2,0 Tonnen zu erwarten?

6.8 Lösungen

6.8.1 Aufgabe

	1	2	3	4	5	6	Mittelwert
X	20	21	24	20	20	24	21,500
Y	4	6	8	3	3	6	5,000
X ²	400	441	576	400	400	576	465,500
Y ²	16	36	64	9	9	36	28,333
X*Y	80	126	192	60	60	144	110,333

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = 2,833$$

$$s_x^2 = \overline{x^2} - \bar{x}^2 = 465,5 - 21,5^2 = 3,25$$

$$s_y^2 = \overline{y^2} - \bar{y}^2 = 28,33 - 5^2 = 3,33$$

$$b = \frac{s_{xy}}{s_x^2} \approx 0,871 \Rightarrow a = \bar{y} - b \cdot \bar{x} = -13,7265$$

$$y = -13,7265 + 0,871 \cdot x$$

$$y(30) = -13,7265 + 0,871 \cdot 30 = 12,4$$

$$r_{xy} = \rho_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{2,833}{\sqrt{3,25} \cdot \sqrt{3,33}} = 0,861$$

6.8.2 Aufgabe 2

	1	2	3	4	5	6	7	8	9	10	Mittelwert
X	8	7	9	11	13	15	10	18	16	15	12,200
Y	9	6	10	12	12,5	13	8	18	16	14,5	11,900
X ²	64	49	81	121	169	225	100	324	256	225	161,400
Y ²	81	36	100	144	156,25	169	64	324	256	210,25	154,050
X*Y	72	42	90	132	162,5	195	80	324	256	217,5	157,100

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = 157,1 - 12,2 \cdot 11,9 = 11,92$$

$$s_x^2 = \overline{x^2} - \bar{x}^2 = 161,4 - 12,2^2 = 12,56$$

$$s_y^2 = \overline{y^2} - \bar{y}^2 = 154,05 - 11,9^2 = 12,44$$

$$b = \frac{s_{xy}}{s_x^2} \approx 0,949 \Rightarrow a = \bar{y} - b \cdot \bar{x} = 11,9 - 0,949 \cdot 12,2 \approx 0,322$$

$$y = 0,322 + 0,949 \cdot x$$

$$y(20) = 0,322 + 0,949 \cdot 20 = 19,30$$

$$r_{xy} = \rho_{xy} = \frac{s_{xy}}{s_x \cdot s_y} \approx \frac{11,92}{\sqrt{12,56} \cdot \sqrt{12,44}} \approx 0,954$$

[illegible]

7.3 Die Axiome von Kolmogorov

7.3.1 Grundlagen

Der russische Mathematiker Kolmogorov (je nach Schreibweise auch „Kolmogorow“, „Kolmogoroff“) führte ein System von Gesetzen (Axiomen) über Wahrscheinlichkeiten ein.

Wir wollen uns heute Ereignisse und Wahrscheinlichkeiten genauer anschauen.

1. Kolmogorov-Axiom

Wahrscheinlichkeiten haben immer einen Wert größer oder gleich 0.

Wie groß ist die Wahrscheinlichkeit, dass ein beliebiges Ergebnis eintritt?

Anders gefragt: Wie wahrscheinlich ist es, dass wir entweder eine 1 oder eine 2 oder eine 3 oder eine 4 oder eine 5 oder eine 6 würfeln??



$$P(\{1,2,3,4,5,6\}) = 100\%$$

Ω

1 2 3 4 5 6

Wir halten fest:

2. Kolmogorov-Axiom

Die Wahrscheinlichkeit des Ergebnisraums Ω ist gleich Eins.

Wie groß ist die Wahrscheinlichkeit, dass man entweder eine gerade Zahl oder eine ungerade Zahl würfelt? Bestimme zuerst die Wahrscheinlichkeit der Ereignisse

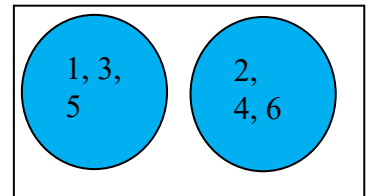
E_1 : gerade Zahl, E_2 : ungerade Zahl



$$E_1 : \text{gerade Zahl}, E_1 = \{2,4,6\} \Rightarrow P(E_1) = \frac{3}{6} = 50\%$$

$$E_2 : \text{ungerade Zahl}, E_2 = \{1,3,5\} \Rightarrow P(E_2) = \frac{3}{6} = 50\%$$

$$E_1 \cup E_2 = \{1,2,3,4,5,6\} \Rightarrow P(E_1 \cup E_2) = 50\% + 50\% = 100\%$$



Wir halten fest:

3. Kolmogorov-Axiom

Die Wahrscheinlichkeit der Vereinigung von Ereignissen $E_i \cup E_j$ ist gleich

$$P(E_i \cup E_j) = P(E_i) + P(E_j), \text{ wenn } E_i \cap E_j = \{ \}$$

Dies gilt max. auch für abzählbar unendlich viele Ereignisse.

Dieses Axiom hat weitreichende Folgen.

Beispiel: Wie wahrscheinlich ist es, dass wir keine 6 würfeln (Gegenereignis dazu, dass wir eine 6 würfeln)?



$$E = \{6\}, P(\bar{E}) = P(\{1,2,3,4,5\}) = \frac{5}{6} \approx 83,33\%$$

(gesprochen: „Wahrscheinlichkeit von Nicht-E“. $\bar{E}$, ist das Gegenereignis von E)

Wir können aber schneller indirekt berechnen, denn:

Es ist sicher, dass wir entweder eine 6 oder keine 6 würfeln, also:

$$P(6) + P(\bar{6}) = 1 \Rightarrow P(\bar{6}) = 1 - P(6) = 1 - \frac{1}{6} = \frac{5}{6} \approx 83,33\%$$

(Daraus folgt auch: Die Wahrscheinlichkeit des unmöglichen Ereignisses (keine 1, 2, 3, 4, 5, 6, also beim Würfel $E = \{ \}$) ist gleich Null, da ja nichts mehr übrigbleibt.)

Es kann sein, dass Ereignisse nicht unabhängig voneinander sind.

In diesem Fall kann man die Wahrscheinlichkeiten der Ereignisse addieren, muss aber die Wahrscheinlichkeit der Schnittmenge (doppelt gezählt: bei jedem Ereignis) wieder abziehen.

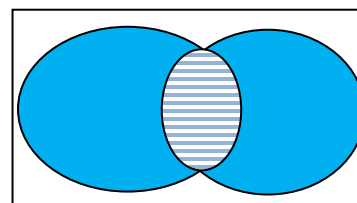
Wie groß ist bspw. die Wahrscheinlichkeit eine gerade Zahl oder eine durch drei teilbare Zahl zu würfeln?

Bestimme zuerst die Wahrscheinlichkeit der Ereignisse

E_1 : gerade Zahl, E_2 : durch 3 teilbare Zahl

$$E_1 = \{2,4,6\}, E_2 = \{3,6\}$$

$$P(E_1) = \frac{3}{6} = 50\% \quad P(E_2) = \frac{2}{6} \approx 33,33\%$$



Die Ereignisse haben ein gemeinsames Element:

$$P(\text{gerade} \cap \text{durch 3 teilbar}) = \{6\}$$

Da wir dies doppelt gezählt haben, müssen wir die Wahrscheinlichkeit der Schnittmenge einmal von der Vereinigungsmenge abziehen.



$$P(\text{gerade} \cup \text{durch 3 teilbar}) =$$

$$P(\{2,3,4,6\}) = \frac{4}{6} \approx 66,67\%$$

$$= P(\text{gerade}) + P(\text{durch 3 teilbar}) - P(\text{gerade} \cap \text{durch 3 teilbar})$$

$$= \frac{1}{2} + \frac{2}{6} - P(\{6\}) = \frac{1}{2} + \frac{2}{6} - \frac{1}{6} = \frac{2}{3} \approx 66,67\%$$

Additionssatz / Satz von Sylvester

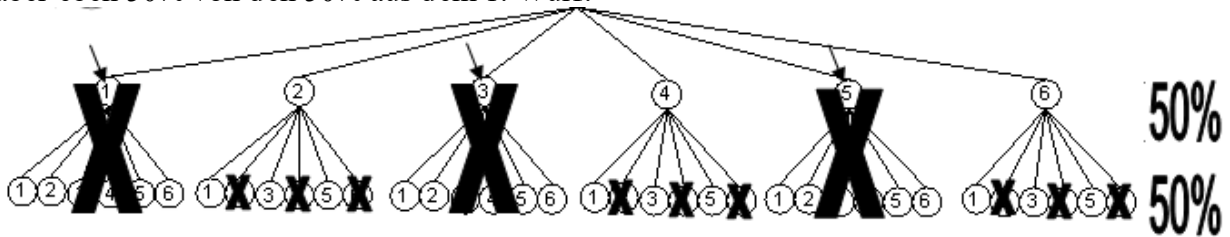
$$\text{Ist } E_1 \cap E_2 \neq \{ \}, \text{ so gilt: } P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

(da die Schnittmenge andernfalls doppelt gezählt würde)

Wie groß ist die Wahrscheinlichkeit, dass, wenn man zweimal würfelt, beim ersten Mal eine gerade und beim zweiten Mal eine ungerade Zahl erscheint?

Hierbei wird „gesiebt“: Zuerst wird die Wahrscheinlichkeit ermittelt, dass man eine gerade Zahl erwischt, anschließend wird ermittelt, wie groß die Wahrscheinlichkeit ist, dass man eine ungerade Zahl erwischt. Nach dem ersten Wurf ist die Wahrscheinlichkeit bei 50%. Von diesen 50%

wird der Anteil berechnet, dass beim zweiten Wurf eine ungerade Zahl erscheint – wieder 50% - aber eben 50% von den 50% aus dem 1. Wurf.



$$P((\text{gerade}; \text{ungerade})) = P(\{2,4,6\}) \cdot P(\{1,3,5\}) = \frac{3}{6} \cdot \frac{3}{6} = \frac{1}{4} = 25\%$$

Die Wahrscheinlichkeit, dass das Ereignis E_i und das Ereignis E_j eintreffen, berechnet sich – wenn die Ereignisse voneinander unabhängig sind - mittels

$$P(E_i \cap E_j) = P(E_i) \cdot P(E_j)$$

7.3.2 Übungen

Aufgabe 1

Ein Skatspiel besteht aus 32 Karten

	7	8	9	10	As	Bube	Dame	König	Gesamt
Kreuz	1	1	1	1	1	1	1	1	8
Pik	1	1	1	1	1	1	1	1	8
Herz	1	1	1	1	1	1	1	1	8
Karo	1	1	1	1	1	1	1	1	8

Bestimmen Sie die folgenden Wahrscheinlichkeiten für das Ziehen einer Karte:

1. $P(\text{Kreuz}) =$
2. $P(\text{Zahl}) =$
3. $P(\bar{7}) =$
4. $P(\text{Kreuz} \cup \text{Herz}) =$
5. $P(\text{Kreuz} \cap \text{Herz}) =$
6. $P(\text{Zahl} \cup \text{Herz}) =$
7. $P(\text{Zahl} \cap \text{Herz}) =$

Aufgabe 2

In einer Lostrommel sind 20 Lose, die von 1-20 nummeriert sind.

Es wird ein Los gezogen.

Betrachten wir nun die folgenden Ereignisse:

E_1 : Die Zahl ist gerade.

E_2 : Die Zahl ist durch drei teilbar.

E_3 : Die Zahl ist eine Primzahl.

$E_4 = E_1 \cap E_2$.

$E_5 = E_2 \cap E_3$.

$E_6 = E_1 \cup E_2$, überprüfe auch mit dem Satz von Sylvester

1. Beschreibe jedes Ereignis durch die zugehörigen Ergebnisse und bestimme die Wahrscheinlichkeit des Ereignisses!
2. Bestimme ein unmögliches Ereignis!
3. Bestimme das sichere Ereignis!

7.3.3 Lösung

Aufgabe 1

$$1. P(\text{Kreuz}) = \frac{8}{32} = 25\% \quad 2. P(\text{Zahl}) = \frac{16}{32} = 50\% \quad 3. P(\bar{7}) = 1 - \frac{4}{32} = 87,5\%$$

$$4. P(\text{Kreuz} \cup \text{Herz}) = \begin{array}{l} \text{unabhängig} \\ = P(\text{Kreuz}) + P(\text{Herz}) \\ = \frac{8}{32} + \frac{8}{32} = 50\% \end{array}$$

$$5. P(\text{Kreuz} \cap \text{Herz}) = \begin{array}{l} 0\% \\ (\text{Kreuz} \cap \text{Herz} = \{\}) \end{array}$$

$$6. P(\text{Zahl} \cup \text{Herz}) = \begin{array}{l} P(\text{Zahl}) + P(\text{Herz}) - P(\text{Zahl} \cap \text{Herz}) \\ = 50\% + 25\% - P(\{7,8,9,10\} \cap \text{Herz}) = 62,5\% \end{array} \quad 4$$

$$7. P(\text{Zahl} \cap \text{Herz}) = \frac{4}{32} = 12,5\%$$

Aufgabe 2

1. Einfach die zugehörigen Werte bestimmen.

$$E_1 = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\} \Rightarrow P(E_1) = \frac{10}{20} = 50\%$$

$$E_2 = \{3, 6, 9, 12, 15, 18\} \Rightarrow P(E_2) = \frac{6}{20} = 30\%$$

$$E_3 = \{2, 3, 5, 7, 11, 13, 17, 19\} \Rightarrow P(E_3) = \frac{8}{20} = 40\%$$

$$E_4 = E_1 \cap E_2 = \{6, 12, 18\} \Rightarrow P(E_4) = \frac{3}{20} = 15\%$$

$$E_5 = E_1 \cup E_2 = \{2, 3, 4, 6, 8, 9, 10, 12, 14, 15, 16, 18, 20\} \Rightarrow P(E_5) = \frac{13}{20} = 65\%$$

$$E_6 = E_1 \cup E_2 = \{2, 3, 4, 6, 8, 9, 10, 12, 14, 15, 16, 18, 20\} \Rightarrow P(E_6) = \frac{13}{20} = 65\%$$

$$\text{Sylvester: } P(E_6) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 50\% + 30\% - 15\% = 65\%$$

2. Bestimme ein unmögliches Ereignis!

Das unmögliche Ereignis ist das Gegenteil des sicheren Ereignisses.

$E_{\text{unmöglich}}$: Es erscheint die Zahl außer 1-20.

3. Bestimme ein sicheres Ereignis!

Das sichere Ereignis tritt immer ein, also mit absoluter Sicherheit.

(Wahrscheinlichkeit = 100%)

E_{sicher} : Es erscheint eine gerade oder eine ungerade Zahl.

7.4 4-Felder-Tafeln

7.4.1 Erklärung

Es kann leicht passieren, dass man in der Arbeit mit Wahrscheinlichkeiten „...vor lauter Bäumen den Wald nicht mehr sieht“.

Übersichtstabellen können Ordnung in den Zahlenschungel bringen.

Hierzu habe ich in einem früheren Kurs eine Statistik gemacht: Wer würde seinem Partner einen Seitensprung verzeihen?

Anzahl der Männer im Kurs	15
Anzahl der Frauen im Kurs	10
Gesamt	25

Anzahl der Personen, die verzeihen	16
Anzahl der Frauen, die verzeihen	4

Eine Hilfe kann die folgende Übersicht sein. Sie wird „4-Felder-Tafel“ genannt, da sie vier innenliegende Felder hat (ohne Summenfelder). Trage die absoluten Zahlen ein! (In den Schnittfeldern von Zeile und Spalte steht jeweils die Zahl der Schnittmenge der Merkmale.)

	Männer	Frauen	Summe
Verzeihen	12	4	16
Nicht verzeihen	3	6	9
Summe	15	10	25

Tragen Sie nun die relativen Häufigkeiten ein:

(Sie geben an, wie wahrscheinlich es ist, dass, wenn man eine beliebige Person anspricht, sie zum entsprechenden Typ gehört.)

	M	$\bar{M}$	Σ
V	48%	16%	64%
$\bar{V}$	12%	24%	36%
Σ	60%	40%	100%

Wie wahrscheinlich ist es, dass wir, wenn wir eine beliebige Person ansprechen, einen Mann erwischen, der einen Seitensprung verzeihen würde?

$$P(M \cap V) = 48\%$$

Wir treffen einen Mann aus dem Kurs. Mit welcher Wahrscheinlichkeit würde er seiner Frau einen Seitensprung verzeihen?

Dieser Fall ist die „bedingte Wahrscheinlichkeit“; man besitzt Vorwissen.

Man schreibt:

$$P(V|M) = \frac{48\%}{60\%} = 80\%$$

Wie groß ist die Wahrscheinlichkeit, eine Person zu befragen, die Mann ist oder verzeiht?

$$\begin{aligned} \text{direkt : } P(M \cup V) &= 48\% + 12\% + 16\% = 76\% \\ \text{alles : } P(M \cup V) &= 100\% - 24\% = 76\% \\ \text{Sylvester : } P(M \cup V) &= 60\% + 64\% - 48\% = 76\% \end{aligned}$$

	M	$\bar{M}$	Σ
V	48%	16%	64%
$\bar{V}$	12%	24%	36%
Σ	60%	40%	100%

7.4.2 Übungen

Beantworten Sie die folgenden Fragen mit 4-Felder-Tafeln.

(Tipp: Es kann helfen, mit absoluten Zahlen zu rechnen, sich also einfach eine Menge von 100 Leuten vorzustellen und alle Prozentzahlen entsprechend umzurechnen!)

1. An einer Uni sind 48% der Absolventen eines Jahrgangs männlich. Von den Absolventen treten 75% der Alumni-Community bei. Bei den Frauen ist der Anteil höher als bei den Männern: 60% der Männer tritt der Community bei. Es werden nur Absolventen als Mitglieder der Community gezählt.

Wir treffen eine Person. Mit welcher Wahrscheinlichkeit ist sie ein männliches, mit welcher Wahrscheinlichkeit ein nicht-männliches Community-Mitglied?

Mit welcher Wahrscheinlichkeit ist eine Person weiblich oder ein Community-Mitglied?

2. Eine Umfrage hat ergeben, dass nach dem Abitur 70% der Absolventen studieren und 30% eine Ausbildung absolvieren. Zudem ist bekannt, dass 62% der guten Absolventen studieren. Es gilt weiterhin, dass 48% der Absolventen nicht gut sind.

Es wird eine Person befragt. Mit welcher Wahrscheinlichkeit ist es kein guter Absolvent, der studiert? Mit welcher Wahrscheinlichkeit ist es kein guter Absolvent oder ein Student?

3. Die Hans-Böckler-Stiftung schrieb 2009:

„In mitbestimmenden Kontrollgremien sind insgesamt 11,7 Prozent der Mitglieder weiblich.“
Schauen wir uns das einmal genauer an:

Dies bedeutet also gerundet: Ca. 12% der Führungskräfte in Unternehmen sind weiblich.

Insgesamt arbeiten nur 5% der Menschen in Führungspositionen.

Wenn wir mit einbeziehen, dass 35% der Arbeitnehmer weiblich und 65% männlich waren, wie groß ist die Wahrscheinlichkeit, dass, wenn wir eine beliebige Person befragen, sie eine weibliche Führungskraft ist? Mit welcher Wahrscheinlichkeit ist sie weiblich oder eine Führungskraft?

4. In einer Klasse mit 30 Schülern sind 50% der schlaunen Leute weiblich. Insgesamt sind nur 40% der Schüler schlau. Man muss aber wissen, dass 30% der Klasse weiblich sind.

Wir sprechen eine Person an. Mit welcher Wahrscheinlichkeit ist es eine schlaue Frau?

Mit welcher Wahrscheinlichkeit ist es ein nicht-schlauer Mann?

5. Faule Lehrer? (Nachdenk-Aufgabe)

In einer Umfrage gaben 60% der Schüler an, dass sie glauben, ihr Lehrer würde sich auf jede Stunde super vorbereiten. 30% der Schüler meinten, ihre Lehrer wären meistens gut vorbereitet, der Rest meinte, ihre Lehrer würden sich überhaupt nicht vorbereiten. 65% der Schüler hatten gute Noten.

Von diesen Schülern glaubten 20% nicht, dass ihre Lehrer immer gut vorbereitet wären.

Wir sprechen einen Schüler an.

- 5.1 Mit welcher Wahrscheinlichkeit hat er keine gute Note und glaubt nicht, dass sein Lehrer immer gut vorbereitet wäre?

- 5.2 Mit welcher Wahrscheinlichkeit ein Schüler hat eine gute Note oder glaubt, sein Lehrer wäre immer gut vorbereitet?

7.4.3 Lösungen

1. An einer Uni sind 48% der Absolventen eines Jahrgangs männlich. Von den Absolventen treten 75% der Alumni-Community bei. Bei den Frauen ist der Anteil höher als bei den Männern: 60% der Männer tritt der Community bei. Es werden nur Absolventen als Mitglieder der Community gezählt.

Wir treffen eine Person. Mit welcher Wahrscheinlichkeit ist sie ein männliches, mit welcher Wahrscheinlichkeit ein weibliches Community-Mitglied?

Mit welcher Wahrscheinlichkeit ist eine Person weiblich oder ein Community-Mitglied?

	M	$\bar{M}$	Σ	
A	28,8%	46,2%	75%	$P(M \cap A) = 48\% \cdot 60\% = 28,8\%$
$\bar{A}$	19,2%	5,8%	25%	$P(\bar{M} \cap A) = 75\% - 28,8\% = 46,2\%$
Σ	48%	52%	100%	$P(\bar{M} \cup A) = 28,8\% + 46,2\% + 5,8\% = 80,8\%$
				oder: $P(\bar{M} \cup A) = 1 - P(M \cap \bar{A}) = 1 - 19,2\% = 80,8\%$
				oder: $P(\bar{M} \cup A) = P(\bar{M}) + P(A) - P(\bar{M} \cap A)$ $= 52\% + 75\% - 46,2\% = 80,8\%$

2. Eine Umfrage hat ergeben, dass nach dem Abitur 70% der Absolventen studieren und 30% eine Ausbildung absolvieren. Zudem ist bekannt, dass 62% der guten Absolventen studieren. Es gilt weiterhin, dass 48% der Absolventen nicht gut sind.

Es wird eine Person befragt. Mit welcher Wahrscheinlichkeit ist es kein guter Absolvent, der studiert? Mit welcher Wahrscheinlichkeit ist es kein guter Absolvent oder ein Student?

	Gut	Nicht gut	Σ	
Studium	32,24%	37,76%	70%	$P(\bar{G} \cap S) = 37,76\%$
Ausbildung	19,76%	10,24%	30%	$P(\bar{G} \cup S) = 48\% + 70\% - 37,76\% = 80,24\%$
Σ	52%	48%	100%	oder: $P(\bar{G} \cup S) = 32,24\% + 37,76\% + 10,24\%$
				oder: $P(\bar{G} \cup S) = 100\% - 19,76\%$

3. Die Hans-Böckler-Stiftung schrieb 2009:

„In mitbestimmenden Kontrollgremien sind insgesamt 11,7 Prozent der Mitglieder weiblich.“
Schauen wir uns das einmal genauer an:

Dies bedeutet also gerundet: Ca. 12% der Führungskräfte in Unternehmen sind weiblich.

Insgesamt arbeiten nur 5% der Menschen in Führungspositionen.

Wenn wir mit einbeziehen, dass 35% der Arbeitnehmer weiblich und 65% männlich waren, wie groß ist die Wahrscheinlichkeit, dass, wenn wir eine beliebige Person befragen, sie eine weibliche Führungskraft ist? Mit welcher Wahrscheinlichkeit ist sie weiblich oder eine Führungskraft?

		Männer	Frauen	Summe	
Führung		4,4%	0,6%	5%	$P(\bar{M} \cap F) = 0,6\%$
Normal		60,6%	34,4%	95%	$P(\bar{M} \cup F) = 5\% + 35\% - 0,6\% = 39,4\%$
Summe		65%	35%	100%	oder: $P(\bar{M} \cup F) = 4,4\% + 0,6\% + 34,4\%$
					oder: $P(\bar{M} \cup F) = 100\% - 60,6\%$

4. In einer Klasse mit 30 Schülern sind 50% der schlaue Leute weiblich. Insgesamt sind nur 40% der Schüler schlaue. Man muss aber wissen, dass 30% der Klasse weiblich sind. Wir sprechen eine Person an. Mit welcher Wahrscheinlichkeit ist es eine schlaue Frau? Mit welcher Wahrscheinlichkeit ist es ein nicht-schlauer Mann?

	M	$\bar{M}$	Summe
S	20%	20%	40%
$\bar{S}$	50%	10%	60%
Summe	70%	30%	100%

$$P(S \cap \bar{M}) = 20\%, P(\bar{S} \cap M) = 50\%$$

5. Faule Lehrer? (Nachdenk-Aufgabe)

In einer Umfrage gaben 60% der Schüler an, dass sie glauben, ihr Lehrer würde sich auf jede Stunde super vorbereiten. 30% der Schüler meinten, ihre Lehrer wären meistens gut vorbereitet, der Rest meinte, ihre Lehrer würden sich überhaupt nicht vorbereiten. 65% der Schüler hatten gute Noten.

Von diesen Schülern glaubten 20% nicht, dass ihre Lehrer immer gut vorbereitet wären.

Wir sprechen einen Schüler an.

- 5.1 Mit welcher Wahrscheinlichkeit hat er keine gute Note und glaubt nicht, dass sein Lehrer immer gut vorbereitet wäre?
- 5.2 Mit welcher Wahrscheinlichkeit hat er eine gute Note oder glaubt, dass sein Lehrer immer gut vorbereitet wäre?

	Immer gut vorbereitet (V)	Nicht immer gut vorbereitet ($\bar{V}$)	Summe
Gute Note (N)	52%	13%	65%
Schlechte Noten ($\bar{N}$)	8%	27%	35%
Summe	60%	40%	100%

5.1 $P(\bar{N} \cap \bar{V}) = 27\%$

5.2 $P(N \cup V) = 65\% + 60\% - 52\% = 73\%$

7.5 Vertiefung: Bedingte und totale Wahrscheinlichkeiten

7.5.1 Erklärung

In einer Statistik eines Abi-Jahrgangs (40% Raucher, 60% Nichtraucher) sind 30% der Raucher regelmäßige Alkoholkonsumenten. Insgesamt sagen 20% aller Abiturienten, dass sie regelmäßig Alkohol konsumieren.

Andere trinken selten oder gar keinen Alkohol...

Wie wahrscheinlich ist es, dass ein beliebig angesprochener Alkoholtrinker Raucher, wie groß, dass er Nichtraucher ist?

Wie groß ist der Anteil der Raucher unter den Alkoholtrinkern, wie groß der Anteil der Nichtraucher?

(Tipp: Es kann helfen, mit absoluten Zahlen zu rechnen, sich also einfach eine Menge von 100 Leuten vorzustellen und alle Prozentzahlen entsprechend umzurechnen!)

Bei der Beantwortung hilft, wieder mal, eine 4-Felder Tafel:

	A	$\bar{A}$	Summe
R	12%	28%	40%
$\bar{R}$	8%	52%	60%
Summe	20%	80%	100%

Man kann sich nun also auf einen Teil der Tafel beschränken, den man braucht.

Markieren wir diesen... Wie groß ist also der Anteil der Raucher *unter den regelmäßigen Alkoholtrinkern* bzw. *wie groß ist die Wahrscheinlichkeit, wenn ein Alkoholkonsument gefragt wird, dass er auch Raucher ist?*

	A	$\bar{A}$	Summe
R	12%	28%	40%
$\bar{R}$	8%	52%	60%
Summe	20%	80%	100%

Man schreibt: $P(R | A) = \frac{12\%}{20\%} = 60\%$

Allgemein kann man sagen:

Bedingte Wahrscheinlichkeit

...ist der Wert der Wahrscheinlichkeit, wenn wir Vorwissen besitzen, in unserem Fall: wenn wir wissen, dass eine Person Alkohol trinkt.

Man schreibt $P(R | A) = \frac{P(R \cap A)}{P(A)}$ oder $P(\bar{R} | A) = \frac{P(\bar{R} \cap A)}{P(A)}$

für die Wahrscheinlichkeit, einen Raucher bzw. Nichtraucher vor sich zu haben, wenn man weiß, dass es ein Alkoholtrinker ist.

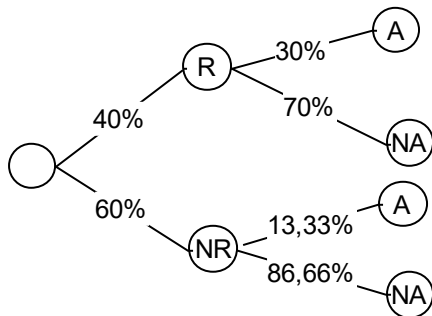
Geteilt wird immer durch die Randzahl, die wir kennen (hier: Die Personen sind Alkoholiker)!

Wie wahrscheinlich ist es, dass ein beliebig angesprochener Alkoholtrinker Raucher, wie groß, dass er Nichtraucher ist?

Wie wahrscheinlich ist es, dass ein beliebig angesprochener Raucher Alkoholtrinker, wie groß, dass er kein Alkoholtrinker ist?

$$P(R | A) = \frac{12}{20}, \quad P(\bar{R} | A) = \frac{8}{20} = 40\%, \quad P(A | R) = \frac{12}{40} = 30\%, \quad P(A | \bar{R}) = \frac{8}{60} = 13,3\%$$

Bei Baumdiagrammen kann die bedingte Wahrscheinlichkeit direkt am Ast abgelesen werden, in unserem Fall hier also:



Der Nachteil: Wenn die Wahrscheinlichkeit gesucht ist und bekannt ist, ob eine Person Alkoholiker ist oder nicht, muss man ein neues Baumdiagramm zeichnen...

7.5.2 Übungen und Lösungen

Übung 1

Wir haben endlich einen Weg gefunden, zwischen fleißigen und faulen Studenten zu unterscheiden – einen medizinisch fundierten Test!

Wenn jemand faul ist, zeigt der Test es mit einer Sicherheit von 90% auch an.

Wenn jemand nicht faul ist, zeigt der Test trotzdem mit 2% Wahrscheinlichkeit an, dass er faul sei, obwohl dies nicht stimmt.

In jedem Jahrgang mit Bewerbern für die Hochschule (100 Personen) gibt es in der Regel einen Faulpelz. Faulheit tritt also in 1% der Fälle auf.

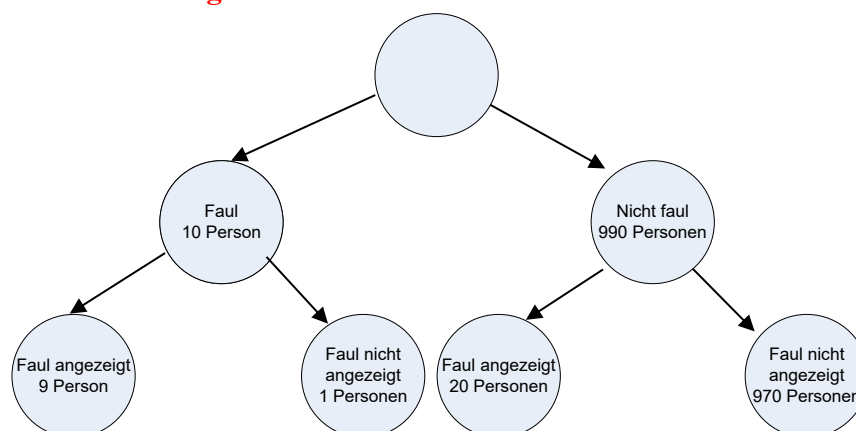
Wir haben den Test beim Sohn eines reichen lokalen Unternehmers durchgeführt und der Test zeigt „faul“ an.

Sollten wir ihn zulassen? Sein Vater bietet uns viel Geld, wenn wir ihn nähmen...³

1. Argumentieren Sie intuitiv!
2. Zeichnen Sie ein Baumdiagramm oder nutzen Sie eine 4-Felder-Tafel, gehen Sie von 1000 Personen aus. Wie hoch ist die Wahrscheinlichkeit, dass der Kandidat tatsächlich faul ist, wenn der Test „faul“ anzeigt?

Lösung 1

Da es viele Nicht-Faule gibt, ist die Zahl der Personen, die fälschlicherweise als faul eingestuft werden, verhältnismäßig hoch.



³ In Anlehnung an: Gigerenzer, G.: „Das Einmaleins der Skepsis- Über den richtigen Umgang mit Zahlen und Risiken“, Berlin, 2002, S. 287.

	Angezeigt	Nicht angezeigt	Summe
Faul	0,9%	0,1%	1%
Nicht faul	1,98%	97,02%	99%
Summe	2,88%	97,12%	100%

$$P(F|A) = \frac{0,9\%}{2,88\%} = 31,25\%$$

⇒ Wenn der Test „faul“ anzeigt, ist die Testperson es nur mit einer Wahrscheinlichkeit von 31,25% tatsächlich.

Übung 2

Schwanger oder nicht schwanger?

Der Schwangerschafts-Bluttest zeigt nach 10 Tagen bereits mit 95%iger Wahrscheinlichkeit bei einer schwangeren Frau an, dass sie schwanger ist.

Die Wahrscheinlichkeit, beim (ungeschützten) Sex schwanger zu werden, beträgt weniger als 30%.⁴ Rechnen wir zur Vereinfachung mit genau 30%.

Gehen wir einmal davon aus, dass alle Frauen, die ungeschützten Sex hatten, auch einen Schwangerschaftstest machen. Wie groß ist die Wahrscheinlichkeit, dass man, wenn der Test „schwanger“ anzeigt, auch tatsächlich schwanger ist?

Nehmen Sie hierfür einmal an, dass der Test auch bei 20% der Nicht-Schwangeren anzeigt, dass sie schwanger wären.

Benutzen Sie eine 4-Felder-Tafel oder ein Baumdiagramm, um das Folgende zu berechnen:

1. Mit welcher Wahrscheinlichkeit ist eine Person schwanger, wenn der Test „nicht schwanger“ anzeigt?
2. Mit welcher Wahrscheinlichkeit ist eine Person nicht schwanger, wenn der Test „schwanger“ anzeigt?

Wenn Sie lieber mit absoluten Häufigkeiten umgehen: Gehen Sie von 1.000 Frauen aus, die den Test machen.



Lösung 2

	S	$\bar{S}$	Σ
A	285	140	425
$\bar{A}$	15	560	575
Σ	300	700	1000

	S	$\bar{S}$	Σ
A	28,5%	14%	42,5%
$\bar{A}$	1,5%	56%	57,5%
Σ	30%	70%	100%

Erläuterung:

$$S \cap A = 285 = 95\% \cdot 300$$

$$\bar{S} \cap A = 140 = 20\% \cdot 700$$

$$P(S \cap \bar{A}) = 1,5\%, P(\bar{S} \cap A) = 14\%$$

⇒

$$1. P(S|\bar{A}) = \frac{15}{575} \approx 2,61\%$$

$$2. P(\bar{S}|A) = \frac{140}{425} \approx 32,94\%$$

⁴ vgl. „Die fruchtbaren Tage“, http://www.wdr.de/tv/quarks/global/pdf/Q_Wunschkind.pdf, Abrufdatum: 20.09.2013.

Übung 3 (Achtung: Erhöhter Schwierigkeitsgrad)

In einer Studie wurde untersucht, wie sich regelmäßiger Alkoholkonsum auf die Zahl der Krankheitstage auswirkt. Es wurden folgende Kategorien gebildet:

Kategorie	Anteil der Personen in der Studie
A: Normale Anzahl Krankheitstage	60%
B: Leicht erhöhte Anzahl Krankheitstage	25%
C: Stark erhöhte Anzahl Krankheitstage	15%

Es konnte auch festgestellt werden, wie hoch der Anteil der Personen mit regelmäßigem Alkoholkonsum in den einzelnen Kategorien ist

Kategorie	Anteil der Personen mit regelmäßigem Alkoholkonsum
A	30%
B	35%
C	35%

Wie groß ist die Wahrscheinlichkeit für eine erhöhte Zahl an Krankheitstagen, wenn eine Person beginnt, regelmäßig Alkohol zu trinken? Ist es schlimmer, als wenn die Person nicht viel Alkohol trinkt?

Tipp: Ereignisse auf zwei Ereignispaare reduzieren!

Lösung 3

	Normal (N)	Erhöht (E)	Summe
Alkoholkonsument (A)	18%	14%	32%
Nichtalkoholkonsument ($\bar{A}$)	42%	26%	68%
Summe	60%	40%	100%
$P(E A) = \frac{14}{32} = 43,75\% \quad P(E \bar{A}) = \frac{26}{68} \approx 38,24\%$			

Bei Alkoholtrinkern gibt es ein höheres Risiko zu mehr Krankheitstagen.

7.5.3 Totale Wahrscheinlichkeit

Manchmal liegen nur bedingte Wahrscheinlichkeiten vor, bspw.:

In einer Erhebung ergab sich, dass 20% der Personen hin und wieder Alkohol trinken.

Von diesen Personen rauchen 60% gelegentlich.

Von den Nicht-Alkoholtrinkern rauchen 35%.

Wie hoch ist die Wahrscheinlichkeit, bei Befragung einer beliebigen Person, einen Raucher zu erwischen?

Wieder hilft eine Vierfeldertafel:

	A	$\bar{A}$	Summe
R	12%	28%	40%
$\frac{n_{Ri}}{n_{i\cdot}}$			
Summe	20%	80%	100%

Ohne Vierfeldertafel könnte man auch direkt ausrechnen:

$$P(R) = 60\% \cdot 20\% + 80\% \cdot 35\% = 12\% + 28\% = 40\%$$

Allgemein

Sind keine inneren Wahrscheinlichkeiten in der Vierfeldertafel gegeben, so ergibt sich eine gesuchte Wahrscheinlichkeit durch

$$P(A) = P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2) + \dots + P(A|B_n) \cdot P(B_n)$$

$$\text{Denn: } P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)}$$

7.5.4 Eine moralische Frage

Beim Test auf Trisomie 21 kommt ein Test zum Einsatz, den wir uns genauer ansehen müssen. Schauen wir uns die Daten für 20-jährige Schwangere an.

Die Krankheit hat eine Prävalenz von $P(\text{Positiv}) = \frac{1}{1177} \approx 0,085\%$.

Die Spezifität ist $P(\overline{\text{Angezeigt}} | \overline{\text{Positiv}}) \approx 99,91\%$.

Die Sensitivität ist $P(\text{Angezeigt} | \text{Positiv}) \approx 96\%$.

Eine Frau bekommt einen positiven Befund und treibt ab. Mit welcher Wahrscheinlichkeit war dies die richtige Entscheidung? (Tatsächlich treiben fast alle Frauen mit positivem Befund ab.) Altersgruppe: 20-jährige Frauen.

Spezifität: Wenn nicht positiv, wird dies in 99,91% korrekt angezeigt.

Sensitivität: Wenn positiv, wird es in 96% richtig angezeigt (Test- und Durchführungsfehler).

Prävalenz: Eine von 1/1177, also circa 0,085%

Bezeichnung: P=Tatsächlich positiv, A=Positiv angezeigt.

	P	$\bar{P}$	Σ
A	0,082%	0,09%	0,172%
$\bar{A}$	0,003%	99,825%	99,828%
Σ	0,085%	99,915%	100%

Ergebnis: $P(P|A) \approx \frac{0,082\%}{0,172\%} \approx 47,67\%$, also weniger als 50%!

Da mit steigendem Alter auch die Positiv-Rate steigt, ergibt sich ein besserer Wert, allerdings nur, weil ein höherer Anteil der Frauen tatsächlich positiv ist. (35 Jahre: 75,2% bei Prävalenz 1/352, 40 Jahre: 92% bei Prävalenz von 1/86). W

7.6 Stochastische Unabhängigkeit

7.6.1 Erklärung

Eine wichtige Voraussetzung vieler Zufallsexperimente ist stochastische Unabhängigkeit.

Man kann es ganz einfach erklären, wenn wir noch einmal das folgende Beispiel betrachten:

Schlaueheit \ Geschlecht	Schlau (S)	Nicht schlau ($\bar{S}$)	Summe
Männlich (M)	20%	50%	70%
Weiblich ($\bar{M}$)	20%	10%	30%
Summe	40%	60%	100%

Ist Schlaueheit abhängig vom Geschlecht?

→ Dies ist dann der Fall, wenn die Wahrscheinlichkeit, eine schlaue Person zu erwischen, unter den Männern anders ist als die Wahrscheinlichkeit, in der gesamten Gruppe eine schlaue Person zu treffen.

Bezeichne nun: M = männlicher Befragter
 S = schlauer Befragter

Es muss also gelten, dass $P(S|M) = P(S)$,

wenn es egal ist, ob die Person männlich oder weiblich ist (also das Geschlecht nichts ausmacht).

Es gilt aber: $P(S|M) = \frac{P(S \cap M)}{P(M)} \Rightarrow P(S) = \frac{P(S \cap M)}{P(M)}$.

Somit folgt als Bedingung also: $P(S \cap M) = P(S) \cdot P(M)$.

Dies ist genau die Bedingung, die sich auch anschaulich ergibt:

Schlaueheit \ Geschlecht	Schlau	Nicht schlau	Summe
Männlich	20%	50%	70%
Weiblich	20%	10%	30%
Summe	40%	60%	100%

$$P(S \cap M) = P(S) \cdot P(M)$$

$$20\% \neq 40\% \cdot 70\% = 28\%$$

$$\Rightarrow \text{stochastisch abhängig}$$

Es kann also immer nur *paarweise* Unabhängigkeit geprüft werden!

Wenn männlich und schlau unabhängig sind, sind es auch weiblich und dumm etc., da es um Intelligenz und Geschlecht geht, nicht um die Ausprägungen!

Rechnerisch ist es klar: Wenn Unabhängigkeit vorliegt, also $P(S \cap M) = P(S) \cdot P(M)$ ist, dann gilt das auch für alle anderen Felder, da sich diese dann automatisch ausfüllen lassen:

Schlaueheit \ Geschlecht	Schlau	Nicht schlau	Summe
Männlich	28%		70%
Weiblich			
Summe	40%		100%

Nun könnte der Rest ausgefüllt werden. Es gilt also auch $P(\bar{S} \cap \bar{M}) = P(\bar{S}) \cdot P(\bar{M})$

Wenn sie nicht unabhängig sind und es passt in diesem einen Feld nicht, dann passt es auch wieder für die ganze Tabelle ist, weil man einfach ergänzen kann!

Eine andere Aufgabe: Unser Lieferant liefert uns Fleisch und Gemüse. Wir haben nun untersucht, wann Fleisch und Gemüse faul bzw. in Ordnung sind.

Fleisch Gemüse	OK (FOK)	Faul ($\overline{FOK}$)	Summe
OK (GOK)	97,5%	1,5%	99%
Faul ($\overline{GOK}$)	0,5%	0,5%	1%
Summe	98%	2%	100%

Sind die Qualität von Fleisch und Gemüse stochastisch unabhängig?

$$P(GOK \cap FOK) = 97,5\% \neq 97,02\% = P(GOK) \cdot P(FOK) = 98\% \cdot 99\%$$

...oder $2\% \cdot 1\% \neq 0,5\%$ etc.

7.6.2 Übungen

1. Abi-Jahrgang: Alkoholkonsum

In einer Statistik eines Abi-Jahrgangs (40% männlich, 60% weiblich) geben 30% der Männer an, 4regelmäßig Alkohol zu trinken. Insgesamt sagen 24% der Abiturienten, dass sie regelmäßig Alkohol konsumieren. Ist Alkoholneigung stochastisch abhängig vom Geschlecht?

2. Faulheitstest

Ein Test soll anzeigen, ob eine Person faul ist.

Wenn jemand faul ist, zeigt der Test es mit einer Sicherheit von 90% an.

Wenn jemand nicht faul ist, zeigt der Test trotzdem mit 2% Wahrscheinlichkeit an, dass er faul sei, obwohl dies nicht stimmt. In jedem Jahrgang mit Bewerbern für die Uni (100 Personen) gibt es in der Regel einen Faulpelz. Sind Testergebnis und Faulheit stochastisch abhängig?

3. Schwangerschaftstest

Die Wahrscheinlichkeit, beim (ungeschützten) Sex schwanger zu werden, beträgt weniger als 30%. Rechnen wir einmal mit 30%. Der Schwangerschafts-Bluttest zeigt bei einer Schwangeren mit 95%-Wahrscheinlichkeit die Schwangerschaft an.

Gehen wir einmal davon aus, dass alle Frauen, die ungeschützten Sex hatten, auch einen Schwangerschaftstest machen. Sind Schwangerschaft und Testanzeige stochastisch unabhängig? (Dies wäre ja schlimm. – Dann würde der Test nichts mit Schwangerschaft zu tun haben!)

Nehmen hierfür einmal an, dass der Test auch bei 20% der Nicht-Schwangeren anzeigt, dass sie schwanger wären. Sind Testergebnis und Schwangerschaft stochastisch abhängig?

(Anders gefragt: Hat das Testergebnis überhaupt etwas mit tatsächlicher Schwangerschaft zu tun?

4. Uni-Mitarbeiter

Es wird eine Stelle an der Uni ausgeschrieben.

Aus der Vergangenheit ist bekannt, dass 40% der Bewerber aus Osteuropa, 20% aus den USA und 40% aus Deutschland kommen. Zudem ist bekannt, dass 60% der deutschen Bewerber ihren Abschluss mit Prädikat (Auszeichnung) gemacht haben, ausländische Mitarbeiter hingegen haben zu 70% einen Prädikatsabschluss. Hat die Tatsache, einen Prädikatsabschluss zu haben,

etwas damit zu tun, ob man Deutscher ist? Wenn ja: Ist die Wahrscheinlichkeit bei Deutschen oder bei Ausländern höher?

7.6.3 Lösungen

1. Abi-Jahrgang: Alkoholkonsum:

In einer Statistik eines Abi-Jahrgangs (40% männlich, 60% weiblich) geben 30% der Männer an, regelmäßig Alkohol zu trinken. Insgesamt sagen 24% der Abiturienten, dass sie regelmäßig Alkohol konsumieren. Ist Alkoholneigung abhängig vom Geschlecht?

	M	$\bar{M}$	Summe
A	12%	12%	24%
$\bar{A}$	28%	48%	76%
Summe	40%	60%	100

$$P(M \cap A) = 12\% \neq P(M) \cdot P(A) = 40\% \cdot 24\% = 9,6\% \Rightarrow \text{stochastisch abhängig}$$

Männer trinken häufiger regelmäßig Alkohol als Frauen. (Konsum ist geschlechtsabhängig.)

2. Faulheitstest

Wenn jemand faul ist, zeigt der Test es mit einer Sicherheit von 90% auch an.

Wenn jemand nicht faul ist, zeigt der Test trotzdem mit 2% Wahrscheinlichkeit an, dass er faul sei, obwohl dies nicht stimmt. In jedem Jahrgang mit Bewerbern fürs Wirtschaftsgymnasium (100 Personen) gibt es in der Regel einen Faulpelz.

	Angezeigt	Nicht angezeigt	Summe
Faul	0,9%	0,1%	1%
Nicht faul	1,98%	97,02%	99%
Summe	2,88%	97,12%	100%

$$P(F \cap A) = 90\% \cdot 1\% = 0,9\% \neq P(F) \cdot P(A) = 1\% \cdot 2,88\% = 0,0288\% \Rightarrow \text{stochastisch abhängig}$$

**Wenn der Test faul anzeigt, ist man tendenziell eher faul, als wenn er es nicht anzeigt.
 $\Rightarrow$ stochastisch abhängig!**

3. Schwangerschaftstest

Die Wahrscheinlichkeit, beim (ungeschützten) Sex schwanger zu werden, beträgt weniger als 30%. Rechnen wir einmal mit 30%. Der Schwangerschafts-Bluttest zeigt bei einer Schwangeren mit 95%-Wahrscheinlichkeit die Schwangerschaft an.

Gehen wir einmal davon aus, dass alle Frauen, die ungeschützten Sex hatten, auch einen Schwangerschaftstest machen. Sind Schwangerschaft und Testanzeige stochastisch unabhängig? (Dies wäre ja schlimm. – Dann würde der Test nichts mit Schwangerschaft zu tun haben!)

Nimm hierfür einmal an, dass der Test auch bei 20% der Nicht-Schwangeren anzeigt, dass sie schwanger wären. Ist ein positiver Test tatsächlich ein Hinweis, dass jemand schwanger ist? (Anders gefragt: Hat das Testergebnis überhaupt etwas mit tatsächlicher Schwangerschaft zu tun?)

$$P(S \cap A) = 30\% \cdot 95\% = 28,5\%$$

$$P(S \cap A) = 28,5\% \neq P(S) \cdot P(A) = 30\% \cdot 42,5\% = 12,75\% \Rightarrow \text{stochastisch abhängig}$$

	S	$\bar{S}$	Summe
A	28,5%	14%	42,5%
$\bar{A}$	15%	56%	57,5%
Summe	30%	70%	100%

Wenn der Test schwanger anzeigt, ist man tendenziell eher schwanger, als wenn er es nicht anzeigt. $\Rightarrow$ stochastisch abhängig!

4. Uni-Mitarbeiter

Aus der Vergangenheit ist bekannt, dass 40% der Bewerber aus Osteuropa, 20% aus den USA und 40% aus Deutschland kommen. Zudem ist bekannt, dass 60% der deutschen Bewerber ihren Abschluss mit Prädikat (Auszeichnung) gemacht haben, ausländische Mitarbeiter hingegen haben zu 70% einen Prädikatsabschluss. Hat die Tatsache, einen Prädikatsabschluss zu machen, etwas damit zu tun, ob man Deutscher ist? Wenn ja: Ist es gut oder schlecht für seine Chancen, Deutscher zu sein?

Es interessiert nur, ob Ausländer oder nicht, unabhängig von der exakten Herkunft.

	Prädikat	Kein Prädikat	Summe
Deutscher	24%	16%	40%
Ausländer	42%	18%	60%
Summe	66%	34%	100%

$$P(D \cap P) = 24\% \neq P(D) \cdot P(P) = 40\% \cdot 66\% = 26,4\% \Rightarrow \text{stochastisch abhängig}$$

Deutsche haben seltener einen Prädikatsabschluss als ihre ausländischen Kollegen.

7.7 Abschlusssaufgabe

Eine beliebige Vorlesung.

60% der Teilnehmer sind weiblich.

70% der Teilnehmer finden die Vorlesung interessant.

80% der Frauen finden die Vorlesung interessant.

1. Erstellen Sie die Vierfeldertafel, um die folgenden Fragen beantworten zu können.
2. Wir sprechen eine beliebige Person an. Mit welcher Wahrscheinlichkeit ist sie männlich?
3. Mit welcher Wahrscheinlichkeit ist eine Person männlich und findet die Vorlesung interessant?
4. Mit welcher Wahrscheinlichkeit ist eine Person männlich oder findet die Vorlesung interessant?
5. Wir sprechen einen Mann an. Mit welcher Wahrscheinlichkeit findet er die Vorlesung interessant?
6. Wir sprechen eine Person an, die die Vorlesung interessant findet. Mit welcher Wahrscheinlichkeit ist sie weiblich?
7. Sind Geschlecht und Interesse an der Vorlesung stochastisch unabhängig?
→ In diesem Fall ist das Interesse unabhängig vom Geschlecht, d. h. der Anteil der Interessierten ist bei Männern und Frauen gleich.

7.8 Lösung

1. Vierfeldertafel

	W	$\bar{W}$	Σ
I	48%	22%	70%
$\bar{I}$	12%	18%	30%
Σ	60%	40%	100%

oder

	I	$\bar{I}$	Σ
W	48%	12%	60%
$\bar{W}$	22%	18%	40%
Σ	70%	30%	100%

2. $P(\bar{W}) = 40\%$
3. $P(\bar{W} \cap I) = 22\%$
4. $P(\bar{W} \cup I) = 100\% - 12\% = 88\%$
5. $P(I | \bar{W}) = \frac{22\%}{40\%} = 55\%$
6. $P(W | I) = \frac{48\%}{70\%} = 68,57\%$
7. $P(W \cap I) = 48\% \neq P(W) \cdot P(I) = 60\% \cdot 70\% = 42\%$

⇒ Stochastisch abhängig:

Wären sie unabhängig, müssten 70% der Frauen die Vorlesung interessant finden, also 42% weiblich sein und die Vorlesung interessant finden.

Tatsächlich sind aber 48% weiblich und finden die Vorlesung interessant, also mehr als erwartet, folglich interessieren sich Frauen eher für die Vorlesung als Männer.

7.9 Vierfeldertafel vs. Baumdiagramm

Betrachten wir erneut das „Verzeihen“-Beispiel aus der Einführung von 4-Felder-Tafeln (s. Seite 69):

	Männer	Frauen	Summe
Verzeihen	48%	16%	64%
Nicht verzeihen	12%	24%	36%
Summe	60%	40%	100%

 $\Rightarrow$

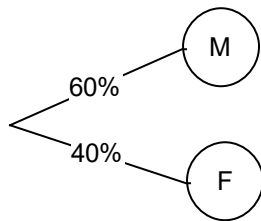
	M	$\bar{M}$	Σ
V	48%	16%	64%
$\bar{V}$	12%	24%	36%
Σ	60%	40%	100%

In den Feldern im Inneren stehen die Schnittmengen aus Zeile und Spalte, also bspw.

$$P(V \cap M) = 48\%$$

Bauen wir nun das zugehörige Baumdiagramm:

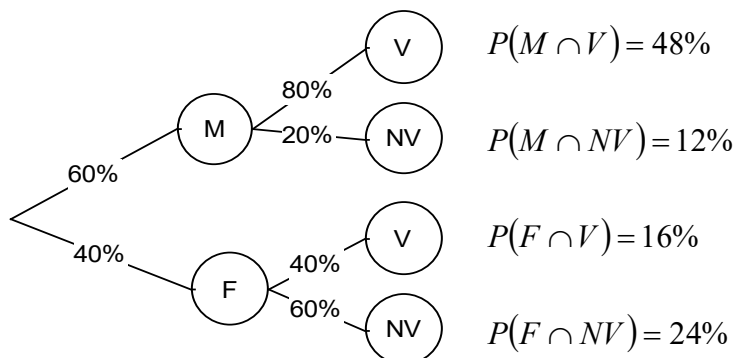
Zuerst einmal die Unterscheidung in Männer und Frauen:



Als nächsten Schritt kommt die Unterscheidung nach verzeihen oder nicht.

Von den Männern verzeihen $\frac{48\%}{60\%} = 80\%$, von den Frauen $\frac{16\%}{40\%} = 40\%$.

Von den Männern verzeihen $\frac{12\%}{60\%} = 20\%$ nicht, von den Frauen $\frac{24\%}{40\%} = 60\%$ nicht.



Die Wahrscheinlichkeiten an den Folgeästen eines Baums beziehen sich immer auf den vorangegangenen Knoten, man schreibt dies folgendermaßen:

$$P(V | M) = 80\%$$

Gesprochen: „Die Wahrscheinlichkeit zu verzeihen bedingt auf die Männer beträgt 80%“.

Dies bedeutet: Wenn man weiß, dass es ein Mann ist, ist die Wahrscheinlichkeit, dass die Person verzeiht. 80%

8 Erwartungswert und Varianz

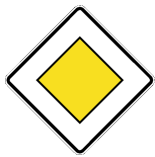
8.1 Erklärung

Die Zukunft ist ungewiss...

Mit einer Wahrscheinlichkeit von 20% wird die „Tecstar GmbH“ einen Gewinn von 1 Mio. € machen, mit einer Wahrscheinlichkeit von 50% einen Gewinn von 2 Mio. € und mit einer Wahrscheinlichkeit von 30% einen Gewinn von 4 Mio. €.

Wie hoch ist der zu erwartende mittlere Gewinn??

⇒ Relevant ist der sogenannte „Erwartungswert“ $E(X)$ oder μ der Zufallsvariable X (Mü wie Mittelwert).



Erwartungswert

$$E(X) = \mu = \sum_{k=1}^n x_k \cdot p_k$$

$$= 20\% \cdot 1 + 50\% \cdot 2 + 30\% \cdot 4 = 2,4 \text{ [Mio. €]}$$

(kleines Mü, griechischer Buchstabe für „m“)

Im Durchschnitt verliert man!

Wichtig ist auch noch das „Verlierer-Risiko“, die mittlere Abweichung vom Erwartungswert! Wie groß ist das durchschnittliche Risiko? Hier hilft die Standardabweichung weiter.



Varianz: $Var(X) = \sigma^2 = \sum_{k=1}^n (x_k - \mu)^2 \cdot p_k \Rightarrow \text{Std.-Abw.: } \sigma = \sqrt{\sigma^2}$

(kleines Sigma, griechischer Buchstabe für „s“)

Die Varianz haben wir schon einmal berechnet:

x_k	p_k	$(x_k - \mu)^2 \cdot p_k$
1	20%	$(1 - 2,4)^2 \cdot 0,2 = 0,392$
2	50%	$(2 - 2,4)^2 \cdot 0,5 = 0,08$
4	30%	$(4 - 2,4)^2 \cdot 0,3 = 0,768$
$\sigma^2 = \sum_{k=1}^n (x_k - \mu)^2 \cdot p_k = 1,24$		
$\sigma = \sqrt{\sigma^2} \approx 1,114$		

Man kann auch wieder mit dem Varianzverschiebungssatz rechnen:

Varianzverschiebungssatz: $Var(X) = \sigma^2 = E(X^2) - E(X)^2$

x_k	x_k^2	p_k
1	1	20%
2	4	50%
4	16	30%
$E(X) = 2,4$	$E(X^2) = 7$	

$$Var(X) = \sigma^2 = E(X^2) - E(X)^2 = 7 - 2,4^2 = 1,24 \Rightarrow \sigma \approx 1,114 \text{ (Standardabweichung)}$$

8.2 Übungen

Aufgabe 1: Schuhe

In einem großen Sack sind 20 Paar Schuhe der Größe 38. Hiervon sind 25% der Paare 100 € wert, 35% der Paare haben einen Wert von 30 € und 15% der Paare sind 10 € wert. 25% der Paare sind nur noch für die Mülltonne geeignet (Wert 0 €).

In einer Verkaufsaktion („You don’t know which pair you get“) sollen diese Schuhe verkauft werden. Welcher Preis muss verlangt werden, damit kein Kunde im Mittel zu viel bezahlt, wir aber trotzdem den Wert der Schuhe wieder hereinbekommen? Wie hoch ist die durchschnittliche Abweichung von diesem mittleren Preis?

Aufgabe 2: Gewinne!

Ein Schreibwarenladen will mit unterschiedlichen Werbeaktionen auf einem Fest neue Kunden anlocken. Ziel dieser Aktion ist also, potenzielle neue „Bücherwürmer“ zu gewinnen.

Während des Festes kaufen einige Kunden unterschiedlich viele Lose.

Der Schreibwarenhändler bietet außerdem unterschiedliche Preise an, wie zum Beispiel Radiergummis, Taschenbücher, Terminplaner und Büchergutscheine.

Insgesamt werden 100 Lose ausgehändigt:

- 10% der Lose sind 1€-Gewinne
- 50% der Lose sind 4€-Gewinne
- 40% der Lose sind 6€-Gewinn

2.1 Welchen Preis sollte ein Los haben, damit der Schreibwarenhändler keinen Gewinn mit der Verlosung macht aber auch keinen Verlust? Wie hoch ist die durchschnittliche Abweichung des Gewinns?

2.2 Bei einem fairen Spiel macht kein Teilnehmer im Mittel einen Gewinn oder Verlust.

Welchen Preis sollte ein Los haben, damit das Spiel fair ist?

2.3 Bisher war der Preis 4 €. Um wieviel muss der höchste Gewinn geändert werden, damit das Spiel fair ist?

Aufgabe 3: Laptopbewertung

„Computer-Journal“, eine Zeitschrift für PC-Hardware, führt Laptopbewertungen durch.

Da die Jury keine Ahnung hat, hängt die Note, die ein Laptop bekommt, letztendlich vom Zufall ab.

Die Wahrscheinlichkeit, dass ein Laptop von „Ultimate*Tech“ mit 10 Punkten (Bestnote) bewertet wird, liegt bei 10%.

Die Wahrscheinlichkeit, dass er mit 7 Punkten bewertet wird, liegt bei 50%.

Die Wahrscheinlichkeit, dass er mit 3 Punkten bewertet wird, liegt bei 40%.

Ihr Laptop ist kaputt gegangen und du möchtest dir ein neues Gerät zulegen. Wenn Sie sich das „Powerbook 3000“ von „Ultimate*Tech“ anschauen, welche Bewertung erwarten Sie dann im „Computer-Journal“? Wie viel liegen Sie damit im Mittel daneben?

8.3 Lösungen

Aufgabe 1: Schuhe

X : Wert eines Schuhpaares

Wahrscheinlichkeit	Wert pro Paar
25%	100 €
35%	30 €
15%	10 €
25%	0 €
100%	= 740 €

$$\mu = E(X) = 0,25 \cdot 100 € + 0,35 \cdot 30 € + 0,15 \cdot 10 € + 0,25 \cdot 0 € = 37 [€]$$

$$\sigma^2 = Var(X) = 0,25 \cdot (100 - 37)^2 + 0,35 \cdot (30 - 37)^2 + 0,15 \cdot (10 - 37)^2 + 0,25 \cdot (0 - 37)^2 = 1461$$

Alternativ: Varianzverschiebungssatz: $\sigma^2 = E(X^2) - E(X)^2$

$$E(X^2) = 0,25 \cdot 100^2 + 0,35 \cdot 30^2 + 0,15 \cdot 10^2 + 0,25 \cdot 0^2 = 2.830$$

$$\sigma^2 = E(X^2) - E(X)^2 = 2.830 - 37^2 = 1.461 \Rightarrow \sigma = \sqrt{\sigma^2} \approx 38,22 [€]$$

Der Preis, der zu einem fairen Spiel führt, ist der Erwartungswert des möglichen Losgewinns, damit es im Mittel weder einen Gewinn noch einen Verlust gibt.

Aufgabe 2: Gewinne

X : Höhe des Gewinns

$$2.1 \mu = E(X) = \frac{1 \cdot 10 + 4 \cdot 50 + 6 \cdot 40}{100} = 4,50 [€], E(X^2) = \frac{1^2 \cdot 10 + 4^2 \cdot 50 + 6^2 \cdot 40}{100} = 22,50 [€]$$

$$\sigma^2 = Var(X) = 0,1 \cdot (1 - 4,50)^2 + 0,5 \cdot (4 - 4,50)^2 + 0,4 \cdot (6 - 4,50)^2 = 2,25$$

Alternativ: Varianzverschiebungssatz: $\sigma^2 = E(X^2) - E(X)^2 = 22,5 - 4,5^2 = 2,25$

$$\sigma = \sqrt{\sigma^2} = 1,50 [€]$$

$$2.2 p = \mu = 22,50 [€]$$

Aufgabe 3: Laptopbewertung

P : Bewertung in Punkten

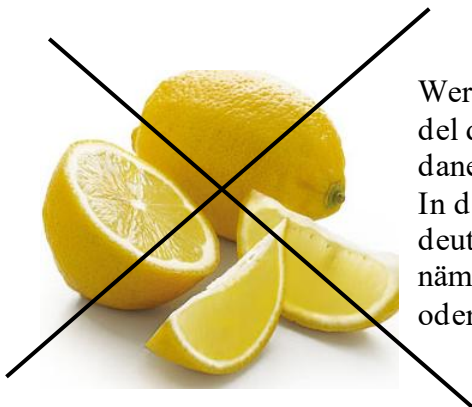
$$\mu = E(P) = 10 \cdot 0,1 + 7 \cdot 0,5 + 3 \cdot 0,4 = 5,7$$

$$\sigma^2 = Var(P) = \sigma^2 = 0,1 \cdot (10 - 5,7)^2 + 0,5 \cdot (7 - 5,7)^2 + 0,4 \cdot (3 - 5,7)^2 = 5,61$$

Alternativ: Varianzverschiebungssatz: $\sigma^2 = E(X^2) - E(X)^2 = 38,1 - 5,7^2 = 5,61$

$$\sigma = \sqrt{5,61} \approx 2,37$$

8.4 Ausblick: Market for lemons



Wer jetzt an einen Obsthandel denkt, liegt ganz schön daneben!
In der Umgangssprache bedeutet das Wort „lemon“ nämlich so etwas wie Schrott oder Müll.



Die 3 Wirtschaftsprofessoren

George A. Akerlof, University of California at Berkeley, Andrew Michael Spence, Stanford University, und Joseph E. Stiglitz, Columbia University, erhielten 2001 den Nobelpreis für Wirtschaftswissenschaften. Es ging um das Problem, wie sich Märkte verhalten, wenn Informationen unterschiedlich verteilt sind.



(Quelle: <http://exame.abril.com.br>)

Ein einfaches Beispiel, an dem sie dies erläuterten, können wir uns schnell vorstellen:

Auf einem Markt gibt es verschiedene Anbieter von Gebrauchtwagen.

Ein Freund von uns möchte ein „gutes, gebrauchtes Auto“, zum Beispiel einen Golf, kaufen.

Wer kann schon in ein Auto hineinsehen? Wenn die Autos nicht über einen Händler verkauft werden, so kennt nur der bisherige Eigentümer ihren wahren Wert.

60% der Wagen, die für 10.000 € angeboten werden, sind tatsächlich 10.000 € wert.

30% der Wagen, die für 10.000 € angeboten werden, sind nur 8.000 € wert.

10% der Wagen, die für 10.000 € angeboten werden, sind sogar nur 5.000 € wert.

Jetzt das Problem: Unser Freund weiß zwar von diesen Verhältnissen, kann aber nicht erkennen, welcher Wagen nun tatsächlich 10.000 € wert ist.

(Dieses Problem nennt man „asymmetrische Informationsverteilung“ (ASIV), da der Verkäufer über mehr Informationen als der Käufer verfügt.)

Da er also den Wert nicht kennt, sagt er sich: Dafür kaufe ich kein Auto.

Er will also nur den Mittelwert bezahlen.

Da er nicht weiß, welches Auto er bekommt, ist das also der Wert eines Autos, den er erwartet zu bekommen. Mehr will er auch nicht bezahlen! Wie hoch ist dieser?

$$E(X) = 0,6 \cdot 10.000 + 0,3 \cdot 8.000 + 0,1 \cdot 5.000 = 8.900 \text{ €}$$

Wieviel weicht der Wert der Autos im Mittel von diesem erwarteten Wert ab?

$$\sqrt{\text{Var}(X)} = \sqrt{0,6 \cdot (10.000 - 8.900)^2 + 0,3 \cdot (8.000 - 8.900)^2 + 0,1 \cdot (5.000 - 8.900)^2} \approx 1577,97 \text{ €}$$

Die Autoverkäufer, deren Autos noch 10.000 € wert sind, wollen nicht für weniger verkaufen, als ihr Auto noch wert ist und fahren es lieber noch etwas, bis es auch weniger wert ist. Zurück bleiben also nur noch die Wagen, die 8.000 € wert sind und die, die 5.000 € wert sind. Wenn die Verteilungen gleich sind (also dreimal so viele für 8.000 € wie für 5.000 €), sind nun

75% der Wagen, die für diesen Preis angeboten werden, sind nur 8.000 € wert.
25% der Wagen, die für diesen Preis angeboten werden, sind nur 5.000 € wert.

Wieder wissen nur die Autobesitzer, wie viel ihr Auto noch wert ist. Unser Freund hört von Bekannten, dass die 10.000 €-Wagen jetzt draußen sind und überlegt sich erneut, wie viel er nun noch für ein Auto ausgeben würde.

Wie viel ist er nun noch bereit, für ein Auto zu bezahlen?

$$E(X) = 0,75 \cdot 8.000 + 0,25 \cdot 5.000 = 7.250 \text{ €}$$

Wieviel weicht der Wert der Autos im Mittel von diesem erwarteten Wert ab?

$$\sqrt{\text{Var}(X)} = \sqrt{0,75 \cdot (8.000 - 7.250)^2 + 0,25 \cdot (5.000 - 7.250)^2} \approx 1299,03 \text{ €}$$

Die Autoverkäufer, deren Autos noch 8.000 € wert sind, sagen sich nun: Sie fahren ihr Auto auch lieber noch etwas. Zurück bleiben nur noch die Verkäufer, deren Auto 5.000 € wert ist. Es sind die einzigen Autos, die man noch kaufen kann.

Wenn unser Freund nun davon hört, wie viel ist er dann noch bereit, für ein Auto zu bezahlen?

$$E(X) = 1,00 \cdot 5.000 = 5.000 \text{ €}$$

Wie viel weicht der Wert der Autos im Mittel von diesem Preis ab?

$$\sqrt{\text{Var}(X)} = \sqrt{1,00 \cdot (5.000 - 5.000)^2} = 0 \text{ €}$$

...und das ist der market for lemons! - Mit Zitronen hat es nichts zu tun!

Zentrales Thema der Nobelpreisempfänger: Wie verhalten sich Märkte, wenn Informationen ungleich verteilt sind?

Ergebnis: Dies kann dazu führen, dass der Markt zusammenbricht und nur noch schlechte Ware angeboten wird! Der „market for lemons“ wird Wirklichkeit, es gibt nur noch Schrott...

Dieses Phänomen nannten die Nobelpreisträger „Adverse Selection“ (nachteilige Auslese), da nicht die Guten die Schlechten, sondern die Schlechten die Guten verdrängen.

Was verhindert auf dem Gebrauchtwagenmarkt, dass er zusammenbricht?

9 Diskrete Wahrscheinlichkeitsverteilungen

9.1 Zufallsvariablen

9.1.1 Erklärung

Zufallsvariablen ordnen jedem Ergebnis eines Zufallsexperiments eine Zahl zu.

Beispielsweise ein Tipp beim Lotto.

Zufallsvariable X : Anzahl der richtig getippten Zahlen

Wir können die Wahrscheinlichkeit bestimmen, dass eine Zufallsvariable einen bestimmten Wert annimmt.

Beispiel: Die Zufallsvariable X gibt das Einkommen einer Person an.

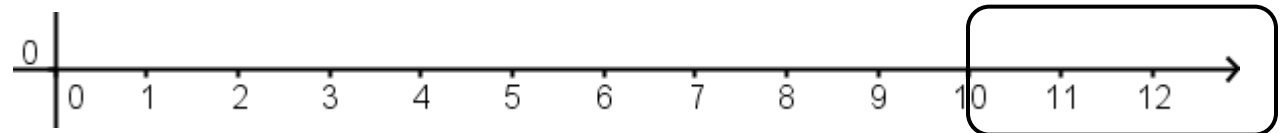
Dann bezeichnet $P(X = 1.000)$ die Wahrscheinlichkeit, dass eine Person 1.000 € pro Monat verdient.

Häufig treten im Alltag Aussagen auf wie „Es gab mindestens 10 Opfer“ oder „Höchstens 100 Einwohner verdienen weniger als 3.000 €/Monat“.

Doch was bedeutet dies?

1. Mindestens

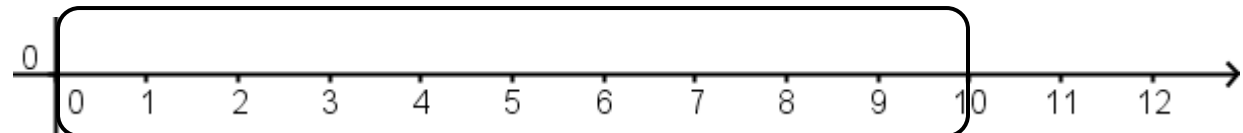
Betrachten wir die Arbeitszeit: Jemand arbeitet mindestens 10 Stunden am Tag.



Mathematisch formuliert: $P(S \geq 10)$

2. Höchstens

Kleine Änderung der Redaktion... Er arbeitet höchstens 10 Stunden am Tag.



Mathematisch formuliert: $P(S \leq 10)$

(Negative Werte für Z sind hier nicht möglich.)

9.1.2 Übung: Formulierung von Ereignissen bei ZV

Schreiben Sie die folgenden Sätze in mathematischer Schreibweise.

1. Die Wahrscheinlichkeit für 10 Stunden beträgt 40%.
2. Die Wahrscheinlichkeit, eine Stunde zu arbeiten, ist höchstens 30%.
3. Die Wahrscheinlichkeit, mindestens drei Stunden zu arbeiten, ist mindestens 30%.
4. Die Wahrscheinlichkeit für mindestens 1 Stunde ist 100% minus der Wahrscheinlichkeit für keine Stunde.
5. Die Wahrscheinlichkeit für höchstens eine Stunde ist mindestens 20%.
6. Wie groß muss die Stundenzahl angegeben werden, dass mit mindestens 70% Wahrscheinlichkeit nicht mehr Stunden arbeiten?

9.1.3 Lösung

1. Die Wahrscheinlichkeit für 5 Stunden beträgt 40%.

$$P(S=10)=40\%$$

2. Die Wahrscheinlichkeit, ein Stunden zu treffen, ist höchstens 30%.

$$P(S=1)\leq 30\%$$

3. Die Wahrscheinlichkeit, mindestens 3 Stunden zu haben, ist mindestens 30%.

$$P(S\geq 3)\geq 30\%$$

4. Die Wahrscheinlichkeit für mindestens 1 Stunden ist 100% minus der Wahrscheinlichkeit für kein Stunden.

$$P(S\geq 1)=100\%-P(S=0)$$

5. Die Wahrscheinlichkeit für höchstens ein Stunden ist mindestens 20%.

$$P(S\leq 1)\geq 20\%$$

6. Wie groß muss die Stundenzahl angegeben werden, dass mit mindestens 70% Wahrscheinlichkeit nicht mehr Stunden auftreten?

Mit mind. 70%iger Wahrscheinlichkeit ist die Zahl der Opfer höchstens z: $P(S\leq z)\geq 70\%$

9.2 Übersicht über Verteilungen

$$P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$$

Kleine Grundgesamtheit und Ziehen ohne Zurücklegen
Hypergeometrische Verteilung

Luxusartikel

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Große Grundgesamtheit oder Ziehen mit Zurücklegen
Binomialverteilung

Standardware

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

Binomialverteilung bei großer Stichprobe und geringer
 Erfolgswahrscheinlichkeit ($n \geq 100$ und $p \leq 5\%$)
Poissonverteilung

Fehler bei
 Massenware

Natürliche Prozesse bzw. Annäherung aller Verteilungen für
 große Grundgesamtheiten.
Normalverteilung

Körpergröße

$$\text{Dichte: } f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, P(X = x) = 0$$

9.3 Binomialverteilung

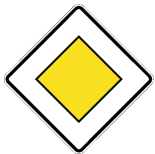
9.3.1 Definition

In der Produktion der Fußball-Trikots ist ein Fehler passiert: Die Maschine hat $\frac{1}{4}$ (25%) der Trikots nicht beflockt (sagt zumindest unser Mitarbeiter!). Wir müssen eine Kontrolle machen! Wir entnehmen 2 Trikots... Mit welcher Wahrscheinlichkeit sind beide nicht beflockt?



Definieren wir in einem ersten Schritt eine Zufallsvariable!
 X : Zahl der nicht beflockten Trikots

Wir suchen also die Wahrscheinlichkeit



$$P(X = 2)$$

In der ersten Stufe entnehmen wir ein Trikot. Die Wahrscheinlichkeit, dass dieses nicht beflockt ist, beträgt $\frac{1}{4} = 25\%$

In der zweiten Stufe entnehmen wir wieder ein Trikot. Die Wahrscheinlichkeit, dass dieses nicht beflockt ist, beträgt $\frac{1}{4} = 25\%$.

Der Wert der Wahrscheinlichkeit ändert sich nicht, da so viele Trikots gefertigt werden, dass die Entnahme von einem im Prinzip nichts am Anteil weiß/orange ändert.

Nach der Pfadregel aus unserem Baumdiagramm ist die Gesamtwahrscheinlichkeit für zweimal nicht beflockt also



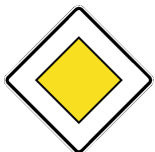
$$P(X = 2) = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16} = 6,25\%$$

Wie groß ist beim zweimaligen Ziehen die Wahrscheinlichkeit, dass nur eines nicht beflockt ist? Wenn das nicht beflockte im ersten und das beflockte Trikot im zweiten Ziehen kommen, ist die Wahrscheinlichkeit



$$\frac{1}{4} \cdot \frac{3}{4} = 0,1875 = 18,75\%$$

Wenn das nicht Beflockte erst beim zweiten Ziehen kommt, ist die Wahrscheinlichkeit



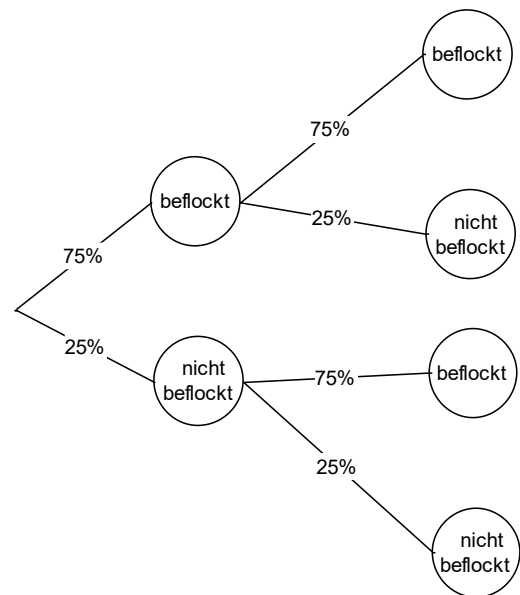
$$\frac{3}{4} \cdot \frac{1}{4} = 0,1875 = 18,75\%$$

Folglich ist die Gesamtwahrscheinlichkeit



$$P(X=1) = 2 \cdot \frac{3}{4} \cdot \frac{1}{4} = 37,5\%$$

Dies kann auch an einem Baum gezeigt werden:



Allgemein mathematisch: Es gibt $\binom{n}{k}$ Kombinationen, die k „Gewinner“-Kugeln in einer Stichprobe von n Kugeln zu haben (Möglichkeiten „n aus k“).
Auf dem Taschenrechner heißt die Taste nCr („from n Choose r“).

Allgemein lautet die Formel für die Wahrscheinlichkeit, bei $p\%$ Kugeln mit „Gewinnen“ und $1-p\%$ Kugeln mit „Verlusten“ bei einer Ziehung von n Kugeln k „Gewinne“ mit Zurücklegen zu ziehen

Formel von Bernoulli

Ist X binomialverteilt (man schreibt $X \sim B(n, p)$),

dann gilt für die Wahrscheinlichkeit des Ereignisses $X = k$:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Wichtig: 2 Merkmalsausprägungen, Ziehen mit Zurücklegen oder große Grundgesamtheit (konstante Wahrscheinlichkeit p).

Dies ist intuitiv auch klar:

$p^k \cdot (1-p)^{n-k}$ ergibt sich aus dem Baumdiagramm.

Da die Kugeln immer wieder zurückgelegt werden, bleiben die Wahrscheinlichkeiten für „Gewinn“/„Verlust“ gleich.

$\binom{n}{k}$

gibt die Zahl der möglichen Reihenfolgen an, in der die Erfolge vorkommen können.

Der Binomialkoeffizienten ist definiert als

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ und gibt die Zahl der Kombinationen an, aus n Kugeln k zu ziehen.

Die Fakultät berechnet sich durch: $n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 1$

Nachtrag: Es kann auch sein, dass ohne Zurücklegen gezogen wird. Wenn die Zahl der Kugeln sehr groß ist, ändern sich die Wahrscheinlichkeiten aber kaum, weshalb die Formel dennoch angewendet wird. Bei einer kleinen Grundgesamtheit ohne Zurücklegen aber nicht!

9.3.2 Beispielaufgabe

Wir haben Schrauben gekauft, um die Dekoration für unser Restaurant aufzuhängen. Der Hersteller behauptet, dass in der Kiste 92% in Ordnung und 8% fehlerhaft sind. Wir entnehmen für eine Qualitätskontrolle 4 Stück. Definieren Sie zuerst eine Zufallsvariable:

Wie groß ist die Wahrscheinlichkeit, dass wir

- a. ausschließlich mangelhafte Exemplare erwischen?
- b. 3 mangelhafte Exemplare erwischen?
- c. 2 mangelhafte Exemplare erwischen?
- d. 1 mangelhaftes Exemplar erwischen?
- e. Kein mangelhaftes Exemplar erwischen?
- f. Mindestens ein mangelhaftes Exemplar erwischt?
- g. Höchstens ein mangelhaftes Exemplar erwischt?
- h. Mindestens zwei mangelhafte Exemplare erwischt?

Wir können einfach überprüfen, ob wir alles richtig gerechnet haben, denn die Summe der Wahrscheinlichkeiten aller möglichen Ergebnisse ist, von Rundungsfehlern abgesehen, gleich Eins (100%).

Dies ist zwar ein Ziehen ohne Zurücklegen, allerdings ist hier die Grundgesamtheit (Zahl der hergestellten Schrauben) so groß, dass es nichts ausmacht, dass wir nicht zurücklegen.

Entscheidend ist hier: In der Schachtel ist genau der gleiche Anteil defekt wie in der Gesamtproduktion. Somit ist die Wahrscheinlichkeit, beim ersten Ziehen eine fehlerhafte Schraube zu erwischen 2%, beim zweiten Ziehen ist sie ebenfalls 2%, ...

9.3.3 Lösung

Wir haben Schrauben gekauft, um die Dekoration für unser Restaurant aufzuhängen. Der Hersteller behauptet, dass in der Kiste 92% in Ordnung und 8% fehlerhaft sind. Wir entnehmen für eine Qualitätskontrolle 4 Stück. Definiere zuerst eine Zufallsvariable:

X: Zahl der fehlerhaften Schrauben

Wie groß ist die Wahrscheinlichkeit, dass wir,

- a. ausschließlich mangelhafte Exemplare erwischen?

$$P(X = 4) = \binom{4}{4} \cdot 0,08^4 \cdot 0,92^0 \approx 0,004\%$$

- b. 3 mangelhafte Exemplare erwischen?

$$P(X = 3) = \binom{4}{3} \cdot 0,08^3 \cdot 0,92^1 \approx 0,188\%$$

- c. 2 mangelhafte Exemplare erwischen?

$$P(X = 2) = \binom{4}{2} \cdot 0,08^2 \cdot 0,92^2 \approx 3,25\%$$

- d. 1 mangelhaftes Exemplar erwischen?

$$P(X = 1) = \binom{4}{1} \cdot 0,08^1 \cdot 0,92^3 \approx 24,918\%$$

- e. Kein mangelhaftes Exemplar erwischen?

$$P(X = 0) = \binom{4}{0} \cdot 0,08^0 \cdot 0,92^4 \approx 71,639\%$$

- f. Mindestens ein mangelhaftes Exemplar erwischt?

Hier kann mit der Gegenwahrscheinlichkeit gearbeitet werden; es fehlt nur der Fall von null mangelhaften Exemplaren, also:

$$P(X \geq 1) = 1 - P(X = 0) \approx 100\% - 71,639\% = 28,361\%$$

- g. Höchstens ein mangelhaftes Exemplar erwischt?

$$P(X \leq 1) = P(X = 0) + P(X = 1) \approx 71,639\% + 24,918\% = 96,557\%$$

- h. Mindestens zwei mangelhafte Exemplare erwischt?

$$P(X \geq 2) = 1 - P(X \leq 1) \approx 1 - 96,557\% = 3,443\%$$

9.3.4 Übungen

Beantworten Sie die folgenden Aufgaben, indem Sie in einem ersten Schritt zu jeder Aufgabe eine geeignete Zufallsvariable definieren. Suchen Sie dann eine Aufgabe aus und erstellen Sie das Diagramm der Wahrscheinlichkeits- und Verteilungsfunktion!



Wahrscheinlichkeitsfunktion:

Die Wahrscheinlichkeitsfunktion ordnet jedem Ergebnis die zugehörige Wahrscheinlichkeit zu: $P(X = k)$. Es ergeben sich also Punkte der Form

$$\left(\underbrace{k}_{x\text{-Wert}}, \underbrace{P(X = k)}_{y\text{-Wert}} \right).$$

Die Punkte dürfen nicht verbunden werden, da es keine Zwischenwerte gibt.



Verteilungsfunktion

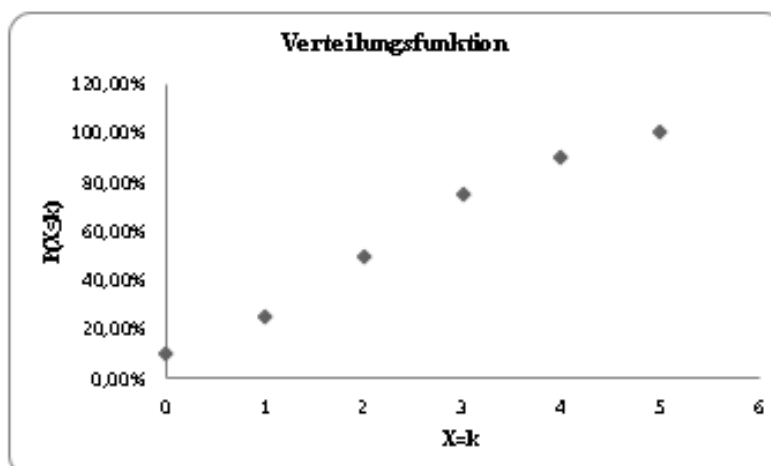
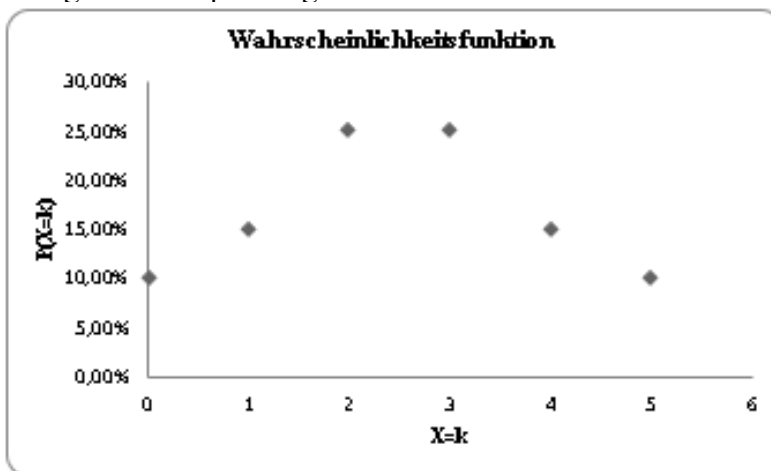
Die Verteilungsfunktion ordnet jedem Ergebnis die Wahrscheinlichkeit zu, höchstens diesen Wert zu erzielen: $P(X \leq k)$. Die Punkte haben also die Form

$$\left(\underbrace{k}_{x\text{-Wert}}, \underbrace{P(X \leq k)}_{y\text{-Wert}} \right)$$

(Wir kennen dieses Konzept bereits von den Quantilen.)

Die Punkte dürfen nicht verbunden werden.

Es ergibt sich bspw. folgendes Aussehen:



Aufgabe 1: Ohrringe

Nina kauft sich zu ihrem 18. Geburtstag einige Ohrringe, um diese je nach Gelegenheit wechseln zu können. Sie entscheidet sich dafür, eine große Schachtel voller Ohrringe zu kaufen, da diese als einzelnes Paar zu teuer sind. Der Hersteller behauptet, dass bei der Produktion 6% fehlerhafte Objekte (Ausschuss) entstehen und in den Schachteln demnach auch im Mittel 6% fehlerhafte Objekte ausgeliefert werden.

Wie groß ist die Wahrscheinlichkeit, dass in der Schachtel mit 4 Ohrringen

1. nur fehlerhafte Ohrringe enthalten sind?
2. 3 fehlerhafte Ohrringe enthalten sind?
3. 2 fehlerhafte Ohrringe enthalten sind?
4. 1 fehlerhafter Ohrring enthalten ist?
5. keine fehlerhaften Ohrringe enthalten sind?
6. höchstens ein fehlerhafter Ohrring enthalten ist?
7. mindestens zwei fehlerhafte Ohrringe enthalten sind?
8. höchstens drei fehlerhafte Ohrringe enthalten sind?
9. mindestens zwei und höchstens drei fehlerhafte Ohrringe enthalten sind?

Aufgabe 2: Kleidung

Fünf Freundinnen beschließen, eine Großbestellung bei „P-Z“, einem neuen Trendversand, aufzugeben. Alle bestellen zusammen, um Porto zu sparen. Da sie nicht viel Geld übrighaben, können sie nur ein Kleid pro Person bestellen. Durch bisherige Erfahrungen der Einzelnen haben sie festgestellt, dass von 10 bestellten Teilen in der Regel bei einem nicht der richtige Artikel oder ein fehlerhafter Artikel geliefert wird. Somit ist jeder einzelne Artikel mit einer Wahrscheinlichkeit von 90% zufriedenstellend.

P-Z

Wie groß ist die Wahrscheinlichkeit, dass

1. die komplette Bestellung fehlerhaft ist?
2. die komplette Bestellung fehlerfrei ist?
3. 3 falsche Artikel vorhanden sind?
4. 2 falsche Artikel vorhanden sind?
5. mindestens ein falscher Artikel vorhanden ist?
6. mindestens ein und maximal drei falsche Artikel vorhanden sind?

Aufgabe 3: Pralinen

Eine Süßwarenfabrik stellt Pralinen her. Bei der Produktion sind etwa 97% der Pralinen fehlerfrei und 3% fehlerhaft. Diese müssen bei der Produktkontrolle aussortiert werden. Wie groß ist die Wahrscheinlichkeit, dass bei der Überprüfung von 5 Pralinen

1. alle 5 Stück fehlerhaft sind?
2. nur 4 Stück fehlerhaft sind?
3. nur 3 Stück fehlerhaft sind?
4. nur 2 Stück fehlerhaft sind?
5. nur 1 Praline fehlerhaft ist?
6. keine Praline fehlerhaft ist?
7. mindestens eine Praline fehlerhaft ist?
8. höchstens vier Pralinen fehlerhaft sind?
9. *Zusatzaufgabe*

Wie viele Pralinen müssen mindestens überprüft werden, dass mit mindestens 95% Wahrscheinlichkeit mindestens eine fehlerhafte Praline dabei ist?

10. *Zusatzaufgabe*

Wenn du eine Vorhersage machen solltest, auf wie viele max. fehlerhafte Stück würdest du tippen, um mit mindestens 90% Wahrscheinlichkeit sicher zu sein, dass es nicht mehr fehlerhafte gibt?

Aufgabe 4: Äpfel

Wir haben einen großen Korb mit Äpfeln geschenkt bekommen. Nach Aussage des Obsthändlers sind 3 % der Äpfel von Würmern befallen und 2% aus einem anderen Grund ungenießbar sein.

Wie groß ist die Wahrscheinlichkeit, dass wir, wenn wir 5 Äpfel entnehmen,

1. nur ungenießbare Äpfel finden?
2. nur genießbare Äpfel finden?
3. 4 ungenießbare Äpfel finden?
4. 3 ungenießbare Äpfel finden?
5. 2 ungenießbare Äpfel finden?
6. 1 ungenießbaren Apfel finden?
7. höchstens 2 ungenießbare Äpfel finden?

Aufgabe 5: Fußball

Die SG Westerburg hat sich dazu entschlossen, in der Sportabteilung der Firma Koch Fußbälle zu kaufen. Der Hersteller der Fußbälle behauptet, dass 1% leichte und 3% schwere Mängel aufweisen. Wie groß ist die Wahrscheinlichkeit, dass wenn wir 5 Fußbälle entnehmen,

1. ausschließlich mangelhafte Fußbälle herausnehmen?
2. 4 mangelhafte Fußbälle herausnehmen?
3. 3 mangelhafte Fußbälle herausnehmen?
4. 2 mangelhafte Fußbälle herausnehmen?
5. 1 mangelhafte Fußbälle herausnehmen?
6. keinen mangelhaften Fußball herausnehmen?
7. mindestens einen mangelhaften Fußball herausnehmen?
8. mindestens 3 mangelhafte Fußbälle herauszunehmen?

9.3.5 Lösungen

Aufgabe 1: Ohringe

Zufallsvariable A = Anzahl fehlerhafter Ohringe

$$1. P(A=4) = \binom{4}{4} \cdot (6\%)^4 \cdot (1-6\%)^{4-4} = \binom{4}{4} \cdot 0,06^4 \cdot 0,94^0 \approx 0,001\%$$

$$2. P(A=3) = \binom{4}{3} \cdot 0,06^3 \cdot 0,94^1 \approx 0,081\%$$

$$3. P(A=2) = \binom{4}{2} \cdot 0,06^2 \cdot 0,94^2 \approx 1,909\%$$

$$4. P(A=1) = \binom{4}{1} \cdot 0,06^1 \cdot 0,94^3 \approx 19,934\%$$

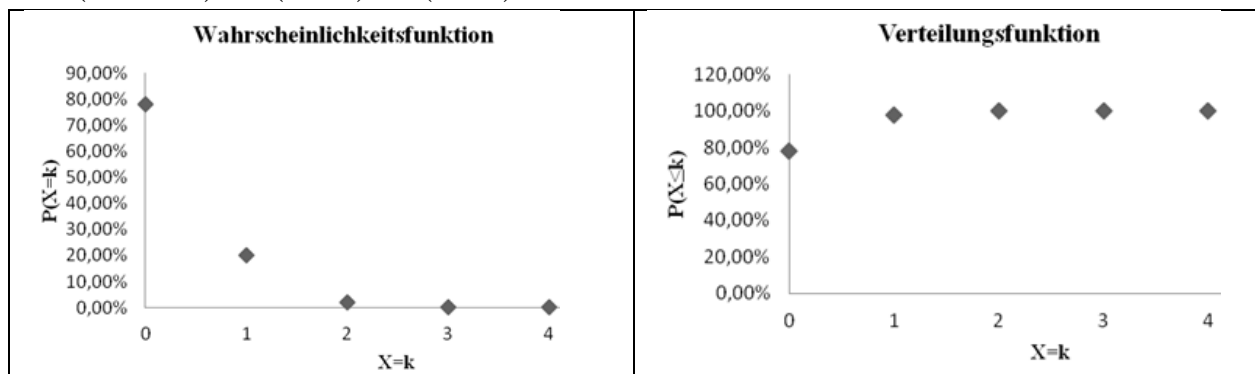
$$5. P(A=0) = \binom{4}{0} \cdot 0,06^0 \cdot 0,94^4 \approx 78,075\%$$

$$6. P(A \leq 1) = P(X=0) + P(X=1) \approx 78,075\% + 19,934\% = 98,009\%$$

$$7. P(A \geq 2) = 1 - P(A < 2) = 1 - P(A \leq 1) \approx 1 - 98,009\% = 1,991\%$$

$$8. P(A \leq 3) = P(A=0) + P(A=1) + P(A=2) + P(A=3) \\ \approx 78,075\% + 19,934\% + 1,909\% + 0,081\% = 99,999\% = 1 - P(X=4)$$

$$9. P(2 \leq A \leq 3) = P(A=2) + P(A=3) = 1,909\% + 0,081\% = 1,99\%$$



Aufgabe 2: Kleidung

$$1. P(B=5) = \binom{5}{5} \cdot 0,1^5 \cdot 0,9^0 = 0,001\%$$

$$2. P(B=0) = \binom{5}{0} \cdot 0,1^0 \cdot 0,9^5 = 59,049\%$$

$$3. P(B=3) = \binom{5}{3} \cdot 0,1^3 \cdot 0,9^2 = 0,81\%$$

$$4. P(B=2) = \binom{5}{2} \cdot 0,1^2 \cdot 0,9^3 = 7,29\%$$

$$5. P(B \geq 1) = 1 - P(B=0)$$

$$P(B \geq 1) = 1 - P(B=0) = 1 - 59,049\% = 40,951\%$$

(Gegenwahrscheinlichkeit)

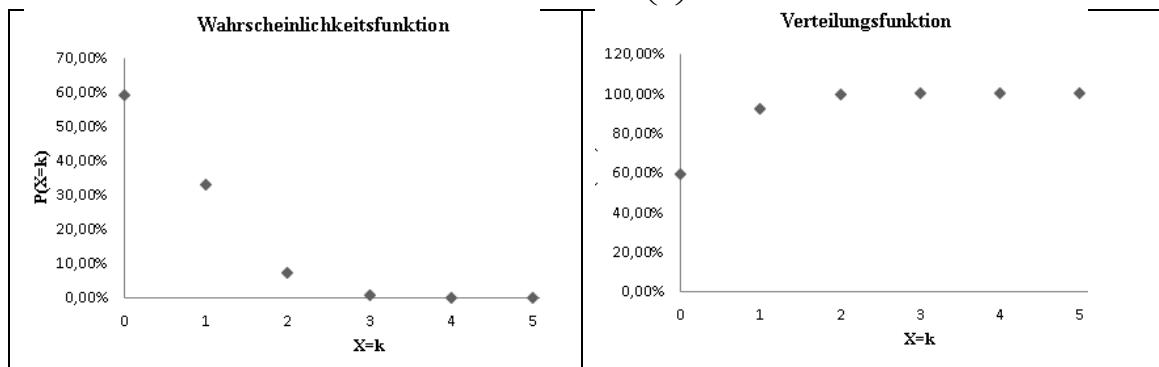
$P(X \text{ höchstens } k)$

$$= P(X \leq k) = P(X=0) + P(X=1) + \dots + P(X=k)$$

$$P(X \text{ mind. } k) = P(X=k) + P(X=k+1) + \dots + P(X=n)$$

$$= P(X \geq k) = 1 - P(X \leq k-1)$$

$$6. P(1 \leq B \leq 3) = P(B=1) + P(B=2) + P(B=3) = \binom{5}{1} \cdot 0,1^1 \cdot 0,9^4 + 7,29\% + 0,81\% = 40,905\%$$



Aufgabe 3: Pralinen

Zufallsvariable: C : Zahl der fehlerhaften Pralinen

$$1. P(C=5) = \binom{5}{5} \cdot 0,03^5 \cdot 0,97^0 \approx 0,00000243\% \approx 0\%$$

$$2. P(C=4) = \binom{5}{4} \cdot 0,03^4 \cdot 0,97^1 \approx 0,000393\% \approx 0\%$$

$$3. P(C=3) = \binom{5}{3} \cdot 0,03^3 \cdot 0,97^2 \approx 0,025\%$$

$$4. P(C=2) = \binom{5}{2} \cdot 0,03^2 \cdot 0,97^3 \approx 0,82\%$$

$$5. P(C=1) = \binom{5}{1} \cdot 0,03^1 \cdot 0,97^4 \approx 13,27\%$$

$$6. P(C=0) = \binom{5}{0} \cdot 0,03^0 \cdot 0,97^5 \approx 85,87\%$$

$$7. P(C \geq 1) = 1 - P(C=0) \approx 13,13\%$$

$$8. P(C \leq 4) = 1 - P(C=5) \approx 99,99999757\% \approx 100\%$$

$$9. P(C \geq 1) \geq 95\%, \quad n = ?$$

$1 - P(C < 1) \geq 95\%$ (Gegenwahrscheinlichkeit, hier: $P(C < 1) = P(C=0)$)

$$\Rightarrow -P(C=0) \geq -5\% \Rightarrow P(C=0) \leq 5\%$$

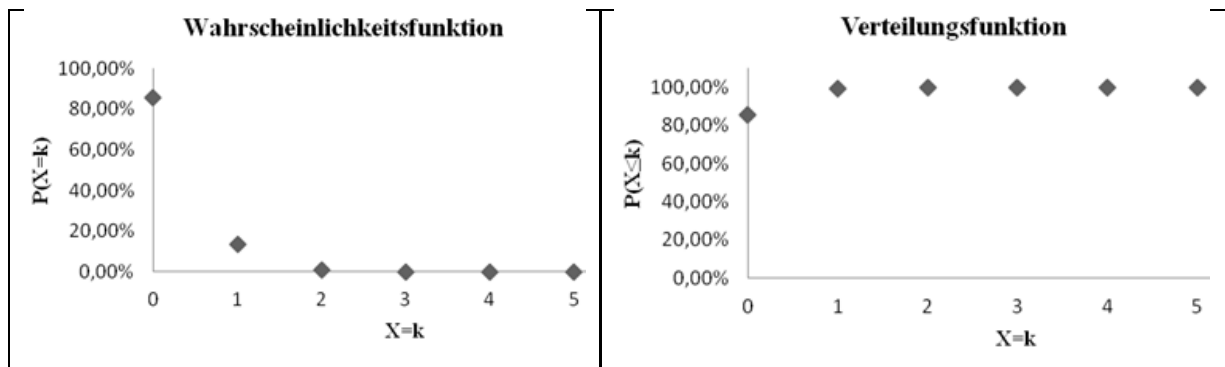
$$\binom{n}{0} \cdot 0,03^0 \cdot 0,97^n \leq 5\%, \quad \binom{n}{0} = 1, 0,03^0 = 1$$

$$\Rightarrow 0,97^n \leq 5\% \Rightarrow n \cdot \log 0,97 \leq \log 5\% \Rightarrow n \geq 99$$

99 Pralinen müssen mindestens überprüft werden!

$$10. P(X \leq k) \geq 90\%, P(X \leq 0) \approx 85,87\% < 90\%, P(X \leq 1) \approx 99,14\% > 90\%$$

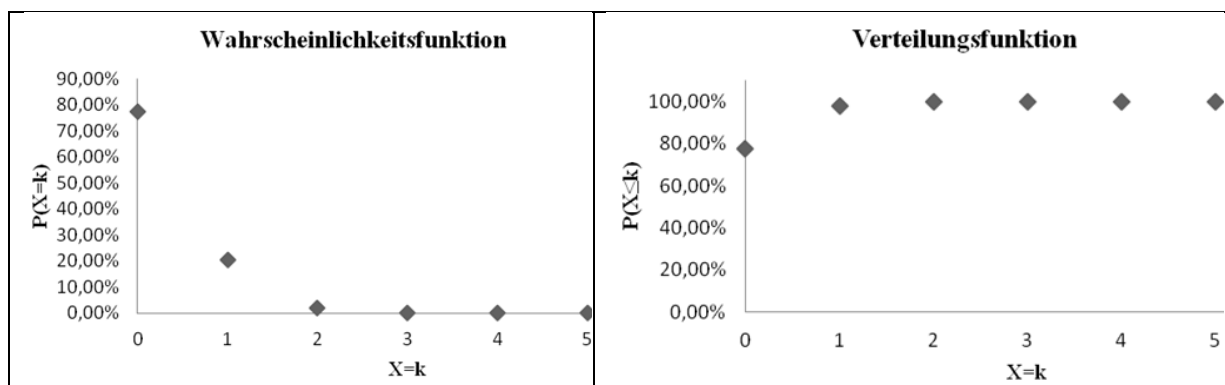
Höchstens eins!

**Aufgabe 4: Äpfel**

Zufallsvariable X = Zahl der ungenießbaren Äpfel

Insgesamt sind also 5% der Äpfel fehlerhaft.

1. $P(X = 5) = \binom{5}{5} \cdot 0,05^5 \cdot 0,95^0 \approx 0,00003125\%$
2. $P(X = 0) = \binom{5}{0} \cdot 0,05^0 \cdot 0,95^{5-0} \approx 77,38\%$
3. $P(X = 4) = \binom{5}{4} \cdot 0,05^4 \cdot 0,95^{5-4} \approx 0,00297\%$
4. $P(X = 3) = \binom{5}{3} \cdot 0,05^3 \cdot 0,95^{5-3} \approx 0,113\%$
5. $P(X = 2) = \binom{5}{2} \cdot 0,05^2 \cdot 0,95^{5-2} \approx 2,14\%$
6. $P(X = 1) = \binom{5}{1} \cdot 0,05^1 \cdot 0,95^4 \approx 20,36\%$
7. $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 99,88\%$



Aufgabe 5: Fußball

1% leichte und 3% schwere Mängel $\Rightarrow$ 4% fehlerhaft.

Zufallsvariable X = Zahl der mangelhaften Fußbälle

$P(X = k)$ = Gibt die Wahrscheinlichkeit von k mangelhaften Fußbällen an.

$$k \mapsto P(X = k)$$

$$1. P(X = 5) = \binom{5}{5} \cdot 0,04^5 \cdot 0,96^{5-5} \approx 0,0000001024 \approx 0$$

$$2. P(X = 4) = \binom{5}{4} \cdot 0,04^4 \cdot 0,96^{5-4} \approx 0,0000123 \approx 0$$

$$3. P(X = 3) = \binom{5}{3} \cdot 0,04^3 \cdot 0,96^2 \approx 0,0005898 \approx 0$$

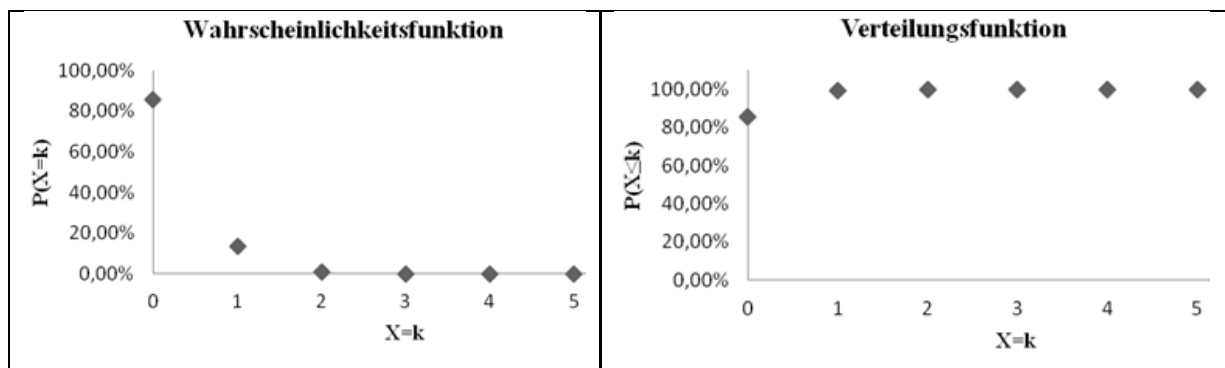
$$4. P(X = 2) = \binom{5}{2} \cdot 0,04^2 \cdot 0,96^3 \approx 1,42\%$$

$$5. P(X = 1) = \binom{5}{1} \cdot 0,04^1 \cdot 0,96^4 \approx 0,1699 = 16,99\%$$

$$6. P(X = 0) = \binom{5}{0} \cdot 0,04^0 \cdot 0,96^5 \approx 0,8154 = 81,54\%$$

$$7. P(X \geq 1) = 1 - P(X = 0) \approx 1 - 0,8154 = 18,46\%$$

$$8. P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) \approx 0,000602$$



9.4 Hypergeometrische Verteilung

9.4.1 Definition

Die Binomialverteilung war in zwei Fällen anwendbar:

1. Es wird mit Zurücklegen gezogen (erkennbar an konstanten Wahrscheinlichkeiten) oder
2. Die Grundgesamtheit ist derart groß, dass es nicht darauf ankommt, ob das Gezogene zurückgelegt wird, bspw. bei der Produktion aller Gummibärchen und einer Stichprobe von fünf Stück.

Eine weitere interessante diskrete Verteilung⁵ ist die hypergeometrische Verteilung.

Eigenschaften der hypergeometrischen Verteilung

1. Es wird ohne Zurücklegen gezogen *und*
2. Die Grundgesamtheit ist eher gering im Vergleich zur Größe der entnommenen Stichprobe.

$$P(X = k) = \frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$$

9.4.2 Beispiel

Verändern wir also die Trikotaufgabe aus der Binomialverteilung geringfügig: Es werden nur sehr wenige Trikots hergestellt.

Wir stellen jeden Tag nur 200 Luxus-Trikots in Handarbeit her. Durch Qualitätskontrollen wissen wir, dass in der Produktion im Mittel 10 Trikots fehlerhaft sind.

Wir entnehmen eine Stichprobe von 5 Stück. Mit welcher Wahrscheinlichkeit ist ein fehlerhaftes Trikot dabei? Diese Frage wollen wir nun beantworten...

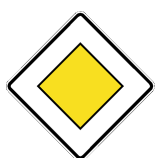
Wie bei der Binomialverteilung definieren wir zuerst eine Zufallsvariable:

X: Zahl fehlerhafter Trikots



Wir berechnen die Zahl aller möglichen Kombinationen, dass von den 10 fehlerhaften Trikots einer und von den 190 ordentlich hergestellten Trikots 4 (Stichprobe (n=5) – 1 fehlerhaft) in der Stichprobe sind. (Wie berechnet man die Zahlen, dass aus 10 eines *und* aus 190 vier gezogen werden?) und teilen wir noch durch die Anzahl aller Möglichkeiten, von 200 Stück genau 5 Stück zu ziehen.

Es ergibt sich somit als Formel für die Wahrscheinlichkeit für das Eintreffen des Ereignisses:



$$P(X=1) = \frac{\binom{10}{1} \cdot \binom{190}{4}}{\binom{200}{5}}$$

Als Wert für die Wahrscheinlichkeit ergibt sich: $P(X=1) \approx 20,75\%$

Berechne nun auch die Wahrscheinlichkeiten für 0-5 fehlerhafte Trikots!

$X = k$	0	1	2	3	4	5
$P(X = k)$	77,17%	20,75%	2,00%	0,08%	0,00%	0,00%

⁵ Diskret: Es gibt nur endlich viele Ergebnisse oder die Zahl der Ergebnisse ist zumindest abzählbar (1, 2, 3, ...)

9.4.3 Übungen

Aufgabe 1: Ohrringe

Ein Schmuckhersteller fertigt jeden Monat 300 Ohrringe aus Handarbeit her. Durch Qualitätskontrollen ist bekannt, dass in der Produktion 20 Ohrringe fehlerhaft sind.

Es wird eine Stichprobe von 7 Stück entnommen. Mit welcher Wahrscheinlichkeit sind 0 - 7 fehlerhafte Ohrringe dabei?

Aufgabe 2: Kleidung

Fünf Freundinnen bestellen zusammen jeweils ein Kleid für eine Party.

Sie haben festgestellt, dass die Lieferungen teilweise fehlerhafte Waren oder falsche Artikel beinhalten.

Aus der Zeitung haben sie erfahren, dass in diesem Monat in der Produktion (500 Kleider) genau 2% fehlerhaft waren. Mit welcher Wahrscheinlichkeit erhalten sie 0-5 fehlerhafte Kleider?

Wie hoch ist die Wahrscheinlichkeit für mindestens ein, wie für höchstens zwei fehlerhafte Kleider?

Aufgabe 3: Pralinen

Eine Süßwarenfabrik stellt pro Tag 200 Pralinen in Handfertigung her. Davon sind erfahrungsgemäß 3% fehlerhaft. Die Pralinen werden kontrolliert und aussortiert, wenn sie fehlerhaft sind, jedoch ist es nicht garantiert, dass nach der Kontrolle alle Pralinen fehlerfrei sind.

Wie groß ist die Wahrscheinlichkeit, dass bei der Überprüfung von 6 Pralinen

- 3.1 alle 6 Stück fehlerhaft sind?
- 3.2 5 Stück fehlerhaft sind?
- 3.3 4 Stück fehlerhaft sind?
- 3.4 3 Stück fehlerhaft sind?
- 3.5 2 Stück fehlerhaft sind?
- 3.6 1 Praline fehlerhaft ist?
- 3.7 keine Praline fehlerhaft ist?
- 3.8 mindestens eine Praline fehlerhaft ist?
- 3.9 höchstens vier Pralinen fehlerhaft sind?

Zusatzaufgabe

In der Fabrik wird eine besondere Sorte hergestellt, von der nur 20 Stück produziert werden.

4 Stück hiervon sind in der Regel fehlerhaft.

- 3.9 Wie viele Pralinen müssen mindestens überprüft werden, dass mit mindestens 95% Wahrscheinlichkeit mindestens eine fehlerhafte Praline dabei ist?
- 3.10 Wenn Sie eine Vorhersage machen sollten, auf wie viele fehlerhafte Stück würden Sie mindestens tippen, um mit mindestens 80% Wahrscheinlichkeit sicher zu sein, dass es nicht mehr fehlerhafte gibt?

Aufgabe 4: Äpfel

Wir haben einen großen Korb mit Äpfeln geschenkt bekommen, in dem sich 350 Äpfel befinden. Wir haben gehört, dass in so einem Früchtekorb normalerweise 10 Äpfel von Würmern befallen sind und sie somit ungenießbar werden. Wir nehmen 5 Äpfel heraus.

Berechne die Wahrscheinlichkeit, wenn von diesen 5 Äpfeln...

- 4.1 0 Äpfel ungenießbar sind?
- 4.2 1 Apfel ungenießbar ist?
- 4.3 2 Äpfel ungenießbar sind?
- 4.4 3 Äpfel ungenießbar sind?
- 4.5 4 Äpfel ungenießbar sind?
- 4.6 5 Äpfel ungenießbar sind?

9.4.4 Lösungen**Aufgabe 1: Ohrhinge**

k	0	1	2	3	4	5	6	7
$P(X = k)$	61,38%	31,36 %	6,50%	0,71	0,04%	0,00%	0,00%	0,00%

$$P(X=0) = \frac{\binom{20}{0} \cdot \binom{280}{7}}{\binom{300}{7}} \approx 61,38\%$$

$$P(X=1) = \frac{\binom{20}{1} \cdot \binom{280}{6}}{\binom{300}{7}} \approx 31,36\%$$

$$P(X=2) = \frac{\binom{20}{2} \cdot \binom{280}{5}}{\binom{300}{7}} \approx 6,5\%$$

$$P(X=3) = \frac{\binom{20}{3} \cdot \binom{280}{4}}{\binom{300}{7}} \approx 0,71\%$$

$$P(X=4) = \frac{\binom{20}{4} \cdot \binom{280}{3}}{\binom{300}{7}} \approx 0,043\%$$

$$P(X=5) = \frac{\binom{20}{5} \cdot \binom{280}{2}}{\binom{300}{7}} \approx 0,0015\%$$

$$P(X=6) = \frac{\binom{20}{6} \cdot \binom{280}{1}}{\binom{300}{7}} \approx 0,000027\%$$

$$P(X=7) = \frac{\binom{20}{7} \cdot \binom{280}{0}}{\binom{300}{7}} \approx 0,00000019\%$$

Aufgabe 2: Kleidung

X = Zahl der fehlerhaften Kleider in der Lieferung, 2% von 500 = 10 fehlerhafte Stück

$$P(X=0) = \frac{\binom{10}{0} \cdot \binom{490}{5}}{\binom{500}{5}} \approx 90,35\%$$

$$P(X=1) = \frac{\binom{10}{1} \cdot \binom{490}{4}}{\binom{500}{5}} \approx 9,30\%$$

$$P(X=2) = \frac{\binom{10}{2} \cdot \binom{490}{3}}{\binom{500}{5}} \approx 0,34\%$$

$$P(X=3) = \frac{\binom{10}{3} \cdot \binom{490}{2}}{\binom{500}{5}} \approx 0,01\%$$

$$P(X=4) = \frac{\binom{10}{4} \cdot \binom{490}{1}}{\binom{500}{5}} \approx 0,00\%$$

$$P(X=5) = \frac{\binom{10}{5} \cdot \binom{490}{0}}{\binom{500}{5}} \approx 0,00\%$$

$$P(X \geq 1) = 1 - P(X=0) \approx 1 - 90,35\% = 9,65\%$$

$$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) \approx 90,35\% + 9,30\% + 0,34\% = 99,99\%$$

Aufgabe 3: Pralinen

X = Zahl der fehlerhaften Pralinen

$$3.1 \quad P(X=6) = \frac{\binom{194}{0} \cdot \binom{6}{6}}{\binom{200}{6}} \approx 0\%$$

$$3.2 \quad P(X=5) = \frac{\binom{194}{1} \cdot \binom{6}{5}}{\binom{200}{6}} \approx 0\%$$

$$3.3 \quad P(X=4) = \frac{\binom{194}{2} \cdot \binom{6}{4}}{\binom{200}{6}} \approx 0,00034\%$$

$$3.4 \quad P(X=3) = \frac{\binom{194}{3} \cdot \binom{6}{3}}{\binom{200}{6}} \approx 0,029\%$$

$$3.5 \quad P(X=2) = \frac{\binom{194}{4} \cdot \binom{6}{2}}{\binom{200}{6}} \approx 1,04\%$$

$$3.6 \quad P(X=1) = \frac{\binom{194}{5} \cdot \binom{6}{1}}{\binom{200}{6}} \approx 15,8\%$$

$$3.7 \quad P(X=0) = \frac{\binom{194}{6} \cdot \binom{6}{0}}{\binom{200}{6}} \approx 83,1\%$$

$$3.8 \quad P(X \geq 1) = 1 - P(X=0) \approx 1 - 83,1\% = 16,9\%$$

$$3.9 \quad P(X \leq 4) = 1 - P(X=5) - P(X=6) \approx 100\%$$

3.9 In der Fabrik wird eine besondere Sorte hergestellt, von der nur 20 Stück produziert werden. 4 Stück hiervon sind in der Regel fehlerhaft.

$$P(X \geq 1) = 1 - P(X=0) \geq 95\% \Rightarrow -P(X=0) \geq -5\% \Rightarrow P(X=0) \leq 5\%$$

$$P(X=0) = \frac{\binom{16}{n} \cdot \binom{4}{0}}{\binom{20}{n}} \leq 5\% \Rightarrow P(X=0) = \frac{\binom{16}{n}}{\binom{20}{n}} \leq 5\%$$

Ausprobieren:

n	1	2	3	4	...	8	9	10
$P(X=0)$	80%	63,16%	49,12%	37,56%		10,22%	6,81%	4,34%

Mindestens 10 Pralinen müssen überprüft werden!

$$3.10 \quad P(X \leq k) \geq 80\%$$

Tippen auf null Fehlerhafte, da hierfür die Wahrscheinlichkeit bereits über 83%.

Aufgabe 4: Äpfel

$$4.1 \quad P(X=0) = \frac{\binom{10}{0} \cdot \binom{340}{5}}{\binom{350}{5}} \approx 86,43\%$$

$$4.2 \quad P(X=1) = \frac{\binom{10}{1} \cdot \binom{340}{4}}{\binom{350}{5}} \approx 12,86\%$$

$$4.3 \quad P(X=2) = \frac{\binom{10}{2} \cdot \binom{340}{3}}{\binom{350}{5}} \approx 0,69\%$$

$$4.4 \quad P(X=3) = \frac{\binom{10}{3} \cdot \binom{340}{2}}{\binom{350}{5}} \approx 0,02\%$$

$$4.5 \quad P(X=4) = \frac{\binom{10}{4} \cdot \binom{340}{1}}{\binom{350}{5}} \approx 0,00\%$$

$$4.6 \quad P(X=5) = \frac{\binom{10}{5} \cdot \binom{340}{0}}{\binom{350}{5}} \approx 0,00\%$$

9.5 Poissonverteilung

9.5.1 Definition

Wenn die Stichprobe sehr groß ist, wird statt der Binomialverteilung die Poissonverteilung verwendet. Ihren Namen hat sie von ihrem Entdecker, Siméon Poisson.⁶

Wichtig ist auch, dass die Erfolgswahrscheinlichkeit relativ gering ist. Aus diesem Grund wird die Poissonverteilung auch „Verteilung der seltenen Ereignisse“ genannt.

Beispiel: Wie groß ist die Wahrscheinlichkeit, dass 5 Einwohner einer Stadt in einem Jahr vom Blitz getroffen werden? Dies ist zwar für jeden einzelnen Einwohner unwahrscheinlich, aber es werden sehr viele Personen untersucht (Stichprobe).

Ist X poissonverteilt (man schreibt $X \sim Po(\lambda)$), dann gilt für die Wahrscheinlichkeit des Ereignisses $X = k$:

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \text{ hierbei ist } \lambda = \mu = E(X).$$

Diese Näherung gilt für $n \geq 100$ und $p \leq 5\%$.

Gerade für Versicherungen ist die Verteilung von zentraler Bedeutung.

9.5.2 Beispiel

Gehen wir einmal zurück zum Beispiel mit der Trikot-Produktion.

In der Produktion von Trikots ist ein Fehler passiert: Die Maschine hat 2,5% der Trikots nicht beflockt (sagt zumindest unser Mitarbeiter!). Wir müssen eine Kontrolle machen! Wie wahrscheinlich ist es, dass von 1.000 Trikots eines Jahrgangssatzes 30 Stück nicht beflockt wurden?

Antwort: Sehr große Stichprobe, also Poissonverteilung:

X : Zahl der unbedruckten Trikots

$$\mu = n \cdot p = 1.000 \cdot 2,5\% = 25 \Rightarrow \lambda = 25$$

$$k = 30$$

$$P(X = 30) = e^{-25} \cdot \frac{25^{30}}{30!} \approx 4,54\%$$

Die Binomialverteilung liefert hier: $P(X = 30) = \binom{1000}{30} \cdot (2,5\%)^{30} \cdot (97,5\%)^{970} \approx 4,55\%$

Für einen anderen Wert der Zufallsvariable, bspw. $k = 20$ ergibt sich:

Poissonverteilung: $P(X = 20) = e^{-25} \cdot \frac{25^{20}}{20!} \approx 5,19\%$

Binomialverteilung: $P(X = 20) = \binom{1000}{20} \cdot (2,5\%)^{20} \cdot (97,5\%)^{980} \approx 5,18\%$

Man sieht also, dass die Poissonverteilung eine gute Annäherung ist.

Für die ersten beiden Momente gilt: $\mu = \sigma^2 = n \cdot p = \lambda$

⁶ Siméon Poisson (1781-1840) war ein französischer Mathematiker und Physiker. Er war ein Schüler von Laplace. 1838 veröffentlichte er seine Erkenntnisse in der Wahrscheinlichkeitsrechnung, unter anderem die Poissonverteilung.

9.5.3 Übungen

Aufgabe 1: Alkoholtest

In einem großen Unternehmen: Dem Chef sind auffällig viele Fehltage der Mitarbeiter durch angebliche Krankheitsfälle aufgefallen. Ein vertrauenswürdiger Mitarbeiter sagt ihm, dass er schon öfter bestimmte Mitarbeiter beim Genuss hochprozentiger alkoholischer Getränke gesehen hätte. Daraufhin lässt der Chef an unangekündigtem Tag 110 seiner Angestellten zu Beginn der Schicht einen Alkoholtest machen. Der ehemalige Mathestudent, welcher ihm den Hinweis auf die Alkoholiker gab, versicherte ihm, dass 4% der Mitarbeiter in seiner Branche in der Regel Alkoholiker seien.

Berechnen Sie nun, wie hoch die Wahrscheinlichkeit ist, dass er

1. 8 Alkoholiker erwischt.
2. 5 Alkoholiker erwischt.
3. 2 Alkoholiker erwischt.
4. 6 Alkoholiker erwischt.
5. mindestens 1 Alkoholiker erwischt.
6. höchstens 3 Alkoholiker erwischt.

Aufgabe 2: Tischtennisbälle

Ein Unternehmen, welches Tischtennisbälle herstellt, erhält einen Auftrag. Es sollen 500.000 Bälle für die ITF produziert werden. Diese Bälle werden mit einem speziellen Gas gefüllt, um ein besseres Sprungverhalten aufzuweisen. Bei diesem Produktionsprozess kommt es vor, dass 3% der Bälle nicht richtig versiegelt werden, wodurch das Gas wieder austritt. Diese Bälle sind fehlerhaft.

Es werden 100 Bälle überprüft. Bestimmen Sie die Wahrscheinlichkeit, dass

1. 2 Bälle nicht korrekt versiegelt worden sind.
2. 9 Bälle nicht korrekt versiegelt worden sind.
3. mindestens 1 Bälle nicht korrekt versiegelt worden sind.
4. höchstens 2 Bälle nicht richtig versiegelt worden sind.
5. höchstens 3 Bälle nicht richtig versiegelt worden sind.
6. mindestens 3 Bälle nicht richtig versiegelt worden sind.

Aufgabe 3: Chronographen

Bei der Produktion von Chronographen ist der Frühschicht ein Fehler unterlaufen:

Der Qualitätsmanager schätzt, dass bei 1,5% der Chronographen die Zahnräder im Uhrwerk vertauscht wurden. Aus diesem Grund muss eine Kontrolle durchgeführt werden.

Wie hoch ist die Wahrscheinlichkeit, dass unter den 500 überprüften Uhren der Frühschicht:

- 1 1 fehlerhafter
- 2 2 fehlerhafte
- 3 3 fehlerhafte
- 4 4 fehlerhafte
- 5 mindestens 1 fehlerhafter
- 6 höchstens 3 fehlerhafte Chronographen sind?
- 7 mindestens 3 und höchstens 5 fehlerhafte Chronographen sind?

Aufgabe 4: Schadensfälle

Du arbeitest für eine Versicherungsgesellschaft, die viele Schadensfälle täglich bearbeitet. Aufgrund historischer Daten weißt du, dass die Wahrscheinlichkeit eines Schadensfalles für einen einzelnen Versicherten sehr gering ist. Die durchschnittliche Anzahl von Schadensfällen pro Tag beträgt 1. Berechnen Sie die Wahrscheinlichkeit, dass

1. genau 2 Schadensfälle an einem Tag auftreten.
2. genau 5 Schadensfälle an einem Tag auftreten.
3. an einem Tag keine Schadensfälle auftreten.

4. mindestens 1 Schadensfall an einem Tag auftritt.
5. mindestens 3 Schadensfälle an einem Tag auftreten.
6. höchstens 2 Schadensfälle an einem Tag auftreten.

Aufgabe 5: Medikamente

Bei der Herstellung von „Xozis“, einem neuen Mittel zur Bekämpfung von Kater am Morgen, ist ein Fehler unterlaufen. Ein Mitarbeiter vermutet, dass 2% der Ware fehlerhaft ist.

1. Wie viele Stück müssen entnommen werden, um im Mittel 5 fehlerhafte Exemplare zu erhalten?
2. Wir entnehmen 20 Stück. Wie viele fehlerhafte Exemplare müssen mindestens toleriert werden, wenn die Vermutung des Mitarbeiters mit einer Wahrscheinlichkeit von mindestens 80% korrekt sein soll?

9.5.4 Lösungen

Aufgabe 1: Alkoholtest

X : Zahl der erwischten Alkoholiker

Die allgemeine Formel lautet: $P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$, hierbei ist $\lambda = \mu = E(X)$

Es gilt für den Parameter λ : $\lambda = \mu = n \cdot p = 110 \cdot 4\% = 4,4 \Rightarrow \lambda = 4,4$

$$1. \ k = 8 \Rightarrow P(X = 8) = e^{-4,4} \cdot \frac{4,4^8}{8!} \approx 4,28\%$$

$$2. \ k = 5 \Rightarrow P(X = 5) = e^{-4,4} \cdot \frac{4,4^5}{5!} \approx 16,87\%$$

$$3. \ k = 2 \Rightarrow P(X = 2) = e^{-4,4} \cdot \frac{4,4^2}{2!} \approx 11,88\%$$

$$4. \ k = 6 \Rightarrow P(X = 6) = e^{-4,4} \cdot \frac{4,4^6}{6!} \approx 12,37\%$$

$$5. \ P(X \geq 1) = 100\% - P(X = 0)$$

$$P(X = 0) = e^{-4,4} \cdot \frac{4,4^0}{0!} \approx 1,23\% \Rightarrow P(X \geq 1) \approx 100\% - 1,23\% = 98,77\%$$

$$6. \ P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$P(X = 1) = e^{-4,4} \cdot \frac{4,4^1}{1!} \approx 5,40\%, \quad P(X = 3) = e^{-4,4} \cdot \frac{4,4^3}{3!} \approx 17,43\%$$

$$\Rightarrow P(X \leq 3) \approx 1,23\% + 5,40\% + 11,88\% + 17,43\% = 35,94\%$$

Aufgabe 2: Tischtennisbälle

X = Anzahl der fehlerhaften Bälle

Für die Wahrscheinlichkeit des Ereignisses gilt: $X = k : P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$

Für λ gilt: $\mu = n \cdot p = 100 \cdot 3\% = 3 \Rightarrow \lambda = 3$

$$1. \ k = 2 \Rightarrow P(X = 2) = e^{-3} \cdot \frac{3^2}{2!} \approx 22,40\%$$

$$2. \ k = 9 \Rightarrow P(X = 9) = e^{-3} \cdot \frac{3^9}{9!} \approx 0,27\%$$

$$3. \ k = 0 \Rightarrow P(X = 0) = e^{-3} \cdot \frac{3^0}{0!} \approx 4,98\% \Rightarrow P(X \geq 1) = 100\% - 4,98\% = 95,02\%$$

4. Es gilt:

$$k = 1 \Rightarrow P(X = 1) = e^{-3} \cdot \frac{3^1}{1!} \approx 14,94\%$$

$$k = 2 \Rightarrow P(X = 2) = e^{-3} \cdot \frac{3^2}{2!} \approx 22,4\%$$

$$\Rightarrow P(X \leq 2) = 4,97\% + 14,94\% + 22,40\% = 42,31\%$$

$$5. \text{ Es gilt: } P(X \leq 3) = 4,97\% + 14,94\% + 22,4\% + 22,4\% = 64,71\%$$

$$6. \text{ Es gilt: } P(X \geq 3) = 100\% - P(X \leq 2) = 100\% - 4,97\% - 14,93\% - 22,40\% = 57,7\%$$

Aufgabe 3: Chronographen X = Anzahl der fehlerhaften Chronographen

$$\mu = n \cdot p = 500 \cdot 1,5\% = 7,5 \Rightarrow \lambda = 7,5$$

$$1. \ k = 1 \Rightarrow P(X = 1) = e^{-7,5} \cdot \frac{7,5^1}{1!} \approx 0,41\%$$

$$2. \ k = 2 \Rightarrow P(X = 2) = e^{-7,5} \cdot \frac{7,5^2}{2!} \approx 1,56\%$$

$$3. \ k = 3 \Rightarrow P(X = 3) = e^{-7,5} \cdot \frac{7,5^3}{3!} \approx 3,89\%$$

$$4. \ k = 4 \Rightarrow P(X = 4) = e^{-7,5} \cdot \frac{7,5^4}{4!} \approx 7,29\%$$

$$5. \ k \geq 1 \Rightarrow P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-7,5} \cdot \frac{7,5^0}{0!} = 1 - 0,06\% = 99,94\%$$

$$6. \ k \leq 3 \Rightarrow P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \approx 5,92\%$$

$$7. \ 3 \leq k \leq 5 \Rightarrow P(3 \leq X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5) \approx 22,12\%$$

$$P(X = 5) = e^{-7,5} \cdot \frac{7,5^5}{5!} \approx 10,94\%$$

Aufgabe 4: Schadensfälle X = Anzahl der fehlerhaften Chronographen $\lambda = 1$ (laut Aufgabenstellung)

$$1. \ P(X = 2) = e^{-1} \cdot \frac{1^2}{2!} \approx 18,394\%$$

$$2. \ P(X = 5) = e^{-1} \cdot \frac{1^5}{5!} \approx 0,307\%$$

$$3. \ P(X = 0) = e^{-1} \cdot \frac{1^0}{0!} \approx 36,788\%$$

$$4. \ P(X \geq 1) = 1 - P(X = 0) \approx 1 - 36,788\% = 63,212\%$$

$$5. \ P(X \geq 3) = 1 - P(X \leq 2) \approx 1 - 91,97\% = 8,03\%$$

$$6. \ P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) \approx 91,97\%$$

$$P(X = 1) = e^{-1} \cdot \frac{1^1}{1!} \approx 36,788\%$$

Aufgabe 5: Medikamente X = Anzahl der fehlerhaften Exemplare

$$1. \ \mu = n \cdot p \Rightarrow n = \frac{\mu}{p} = \frac{5}{2\%} = 250$$

$$2. \ P(X \leq k) \geq 80\%, P(X = k) = e^{-5} \cdot \frac{5^k}{k!}$$

k	0	1	2	3	4	5	6	7
$P(X = k)$	0,67%	3,37%	8,42%	14,04%	17,55%	17,55%	14,62%	10,44%
$P(X \leq k)$	0,67%	4,04%	12,47%	26,50%	44,05%	61,60%	76,22%	86,66%

7 fehlerhafte Stück müssen mindestens toleriert werden.

9.6 Vermischte Übung I: Verteilungen

Firma Teknick hat drei verschiedene Abteilungen, die unterschiedliche Produkte herstellen:

Aufgabe 1

Abteilung 1 stellt pro Tag 300 Handys her. Aus vorangegangenen Qualitätskontrollen ist bereits bekannt, dass von diesen Handys jeweils 25 fehlerhaft sind.

Es wird eine Stichprobe von 8 Geräte entnommen.

- a) Mit welcher Wahrscheinlichkeit sind 0 Handys fehlerhaft?
- b) Mit welcher Wahrscheinlichkeit sind 1 Handys fehlerhaft?
- c) Mit welcher Wahrscheinlichkeit ist höchstens 1 Handy fehlerhaft?
- d) Mit welcher Wahrscheinlichkeit sind mindestens 2 Handys fehlerhaft?
- e) Mit welcher Wahrscheinlichkeit sind höchstens 6 Handys fehlerhaft?

Aufgabe 2

Abteilung 2 stellt pro Woche eine sehr große Zahl an Handyhüllen her. Bei dieser Herstellung entstehen 15% Fehlproduktionen, die Handyhüllen passen also nicht auf die Handys. Für eine Qualitätskontrolle werden 6 Hüllen entnommen.

- a) Mit welcher Wahrscheinlichkeit erwischt man kein fehlerhaftes Exemplar?
- b) Mit welcher Wahrscheinlichkeit sind 3 Hüllen fehlerhaft?
- c) Mit welcher Wahrscheinlichkeit sind mindestens 5 Exemplare fehlerhaft?
- d) Mit welcher Wahrscheinlichkeit sind höchstens 4 Hüllen fehlerhaft?

Aufgabe 3

Abteilung 3 stellt über 50.000 Kleinteile pro Woche in Massenware her. Jetzt soll die Qualität überprüft werden, da diese wichtigen Teile eines Smartphones darstellen. Es wird eine Stichprobe von 100 Teilen entnommen. Erfahrungsgemäß sind 4% der produzierten Teile Ausschuss.

- a) Mit welcher Wahrscheinlichkeit sind 0 Exemplare fehlerhaft?
- b) Mit welcher Wahrscheinlichkeit sind 1 Exemplare fehlerhaft?
- c) Mit welcher Wahrscheinlichkeit sind höchstens 2 Exemplare fehlerhaft?
- d) Mit welcher Wahrscheinlichkeit sind mindestens 3 Teile fehlerhaft?

9.7 Lösung

Aufgabe 1: Hypergeometrische Verteilung

Kleine Grundgesamtheit, Ziehen ohne Zurücklegen

X: Zahl der fehlerhaften Handys

$$\text{a) } P(X=0) = \frac{\binom{25}{0} \cdot \binom{275}{8}}{\binom{300}{8}} \approx 49,42\% \quad \text{b) } P(X=1) = \frac{\binom{25}{1} \cdot \binom{275}{7}}{\binom{300}{8}} \approx 36,883\%$$

$$\text{c) } P(X \leq 1) = P(X=0) + P(X=1) \approx 49,42\% + 36,88\% = 86,3\%$$

$$\text{d) } P(X \geq 2) = 1 - P(X \leq 1) \approx 1 - 86,3\% = 13,7\%$$

$$\text{e) } P(X \leq 6) = 1 - P(X \geq 7) = 1 - \left[\underbrace{\frac{\binom{25}{7} \cdot \binom{275}{1}}{\binom{300}{8}}}_{\approx 0,00\%} + \underbrace{\frac{\binom{25}{8} \cdot \binom{275}{0}}{\binom{300}{8}}}_{\approx 0,00\%} \right] \approx 100\%$$

Aufgabe 2: Binomialverteilung

Unbekannte Größe der Grundgesamtheit, große Fehlerwahrscheinlichkeit, kleine Stichprobe

X: Zahl fehlerhafter Handyhüllen

$$\text{a) } P(X=0) = \binom{6}{0} \cdot 0,15^0 \cdot 0,85^6 \approx 37,72\% \quad \text{b) } P(X=3) = \binom{6}{3} \cdot 0,15^3 \cdot 0,85^3 \approx 4,15\%$$

$$\text{c) } P(X \geq 5) = P(X=5) + P(X=6) = \binom{6}{5} \cdot 0,15^5 \cdot 0,85^1 + \binom{6}{6} \cdot 0,15^6 \cdot 0,85^0 \approx 0,0399\%$$

$$\text{d) } P(X \leq 4) = 1 - P(X \geq 5) \approx 99,96\%$$

Aufgabe 3: Poissonverteilung

Große Stichprobe, kleine Fehlerwahrscheinlichkeit

X: Zahl fehlerhafter Kleinteile

$$\lambda = 100 \cdot 4\% = 4$$

$$\text{a) } P(X=0) = e^{-4} \cdot \frac{4^0}{0!} \approx 1,83\% \quad \text{b) } P(X=1) = e^{-4} \cdot \frac{4^1}{1!} \approx 7,326\%$$

$$\text{c) } P(X=2) = e^{-4} \cdot \frac{4^2}{2!} \approx 14,653\% \Rightarrow P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) \approx 23,809\%$$

$$\text{d) } P(X \geq 3) = 1 - P(X \leq 2) \approx 1 - 23,809\% = 76,191\%$$

9.8 Vermischte Übung II: Verteilungen

In den Medien wird häufig behauptet, dass die Noten inzwischen viel zu gut seien.

Bestimmen Sie jeweils mit einer geeigneten Verteilung die Wahrscheinlichkeiten. Notieren Sie auch jeweils den Namen der verwendeten Verteilung.

Aufgabe 1

In Singapur sind die Bewerter sehr streng und nur 0,5% der Absolventen erreichen an einer betrachteten Schule die Spitzenbewertung. Es werden 200 Personen befragt.

Mit welcher Wahrscheinlichkeit hat bzw. haben hiervon

1. keine Personen die Note 1,0 erreicht?
2. eine Person die Note 1,0 erreicht?
3. höchstens eine Person die Note 1,0 erreicht?
4. mindestens eine Person die Note 1,0 erreicht?
5. mindestens zwei Personen die Note 1,0 erreicht?
6. höchstens zwei Personen die Note 1,0 erreicht?

Aufgabe 2

An einer Hochschule haben in einem Jahr 10 der 250 Absolventen die Note 1,0 erreicht.

Auf der Abschlussfeier werden 5 Personen befragt.

Mit welcher Wahrscheinlichkeit hat bzw. haben hiervon

1. keine Personen die Note 1,0 erreicht?
2. eine Person die Note 1,0 erreicht?
3. höchstens eine Person die Note 1,0 erreicht?
4. mindestens eine Person die Note 1,0 erreicht?
5. mindestens zwei Personen die Note 1,0 erreicht?
6. höchstens zwei Personen die Note 1,0 erreicht?

Aufgabe 3

An allen Hochschulen eines Landes in Europa haben 6% der Absolventen die Note 1,0 erreicht.

Es werden 10 Personen befragt.

Mit welcher Wahrscheinlichkeit hat bzw. haben hiervon

1. keine Personen die Note 1,0 erreicht?
2. eine Person die Note 1,0 erreicht?
3. höchstens eine Person die Note 1,0 erreicht?
4. mindestens eine Person die Note 1,0 erreicht?
5. mindestens zwei Personen die Note 1,0 erreicht?
6. höchstens zwei Personen die Note 1,0 erreicht?

9.9 Lösung

Aufgabe 1: Poissonverteilung

X : Anzahl der Personen mit Note 1,0

$$\lambda = 200 \cdot 0,5\% = 1$$

$$1. P(X=0) = e^{-1} \cdot \frac{1^0}{0!} \approx 36,788\%$$

$$2. P(X=1) = e^{-1} \cdot \frac{1^1}{1!} \approx 36,788\%$$

$$3. P(X \leq 1) = P(X=0) + P(X=1) \approx 73,576\%$$

$$4. P(X \geq 1) = 1 - P(X=0) \approx 63,212\%$$

$$5. P(X \geq 2) = 1 - P(X \leq 1) = 1 - 73,576\% \approx 26,424\%$$

$$6. P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$$

$$P(X=2) = e^{-1} \cdot \frac{1^2}{2!} \approx 18,394\% \Rightarrow P(X \leq 2) \approx 91,970\%$$

Aufgabe 2: Hypergeometrische Verteilung

$$1. P(X=0) = \frac{\binom{10}{0} \cdot \binom{240}{5}}{\binom{250}{5}} \approx 81,40\% \quad 2. P(X=1) = \frac{\binom{10}{1} \cdot \binom{240}{4}}{\binom{250}{5}} \approx 17,246\%$$

$$3. P(X \leq 1) = P(X=0) + P(X=1) = 98,646\%$$

$$4. P(X \geq 1) = 1 - P(X=0) \approx 18,60\%$$

$$5. P(X \geq 2) = 1 - P(X \leq 1) \approx 1,354\%$$

$$6. P(X \leq 2) = P(X=0) + P(X=1) + P(X=2), P(X=2) = \frac{\binom{10}{2} \cdot \binom{240}{3}}{\binom{250}{5}} \approx 1,31\%$$

$$\Rightarrow P(X \leq 2) \approx 99,956\%$$

Aufgabe 3: Binomialverteilung

$$1. P(X=0) = \binom{10}{0} \cdot 0,06^0 \cdot 0,94^{10} \approx 53,862\% \quad 2. P(X=1) = \binom{10}{1} \cdot 0,06^1 \cdot 0,94^9 \approx 34,38\%$$

$$3. P(X \leq 1) = P(X=0) + P(X=1) = 88,242\%$$

$$4. P(X \geq 1) = 1 - P(X=0) \approx 46,138\%$$

$$5. P(X \geq 2) = 1 - P(X \leq 1) \approx 11,758\%$$

$$6. P(X \leq 2) = P(X=0) + P(X=1) + P(X=2), P(X=2) = \binom{10}{2} \cdot 0,06^2 \cdot 0,94^8 \approx 9,875\%$$

$$\Rightarrow P(X \leq 2) \approx 98,116\% \quad \text{bzw.} \quad P(X \leq 2) \approx 98,116\% \text{ (gerundet)}$$

9.10 Verteilungen in der Qualitätskontrolle

Ein interessanter Anwendungsfall der Binomialverteilung ist die Qualitätskontrolle.

Wichtige Daten hierzu liefert der „Editio Cantor“-Verlag.

Mit der Zeit bauen sie [die Gründer] das Verlagsprogramm aus, dessen inhaltlicher Fokus auf dem Arzneimittel liegt. Die Themen umfassen das Auffinden neuer Wirksubstanzen, sämtliche Prüfstadien bis zur Zulassung eines Präparates, die pharmazeutisch-technologischen Aspekte in der Produktion, alle gesundheitspolitischen sowie marktrelevanten und regulatorischen Hintergründe.

Freundlicherweise hat der Verlag uns einige Daten zur Überprüfung überlassen.

Auf Seite 64 der „Allgemeinen Grundlagen“ („General principles“, Band 12) zur Qualitätssicherung pharmazeutischer und kosmetischer Packmittel findet sich die folgende Tabelle:

Losgröße N	Stichproben- Umfang n	AQL									
		0,1	0,25	0,4	0,65	1,0	1,5	2,5	4,0	6,5	8,0
		Annahmezahl c (Anzahl von in der Stichprobe jeweils zulässigen fehlerhaften Teile)									
bis 500	N bzw. 80	0	1	1	1	2	3	4	6	9	11
501-1200	80	0	1	1	1	2	3	4	6	9	11
1201-3200	125	0	1	2	2	3	4	6	8	12	15
3201-10.000	200	1	1	2	3	4	6	8	12	18	22
10.001-35.000	315	1	2	3	4	6	8	12	18	27	33
35.001-150.000	500	1	3	5	6	9	12	18	27	44	55
150.001-500.000	800	2	4	6	9	13	17	27	41	63	77
über 500.000	1250	3	6	10	13	18	26	40	61	96	110

Stichprobengrundtabelle für qualitative Prüfungen (Annahmewahrscheinlichkeit 95%)

Der AQL (Acceptable Quality Level) ist die Wahrscheinlichkeit, dass nicht mehr als c fehlerhafte Teile in der Stichprobe sind. Er entspricht also dem Anteil fehlerhafter Teile in der Stichprobe.

Der Wert für die Annahmezahl ergibt sich, indem man die Zahl sucht, bei der die Wahrscheinlichkeit, dass nicht mehr fehlerhafte Teile als diese Zahl in der Lieferung sind, gerade noch größer als 95% (Annahmewahrscheinlichkeit) ist. (Es wird also das 95%-Quantil gesucht.)

Bei 800 Stück müssen wir in der Stichprobe 2 Defekte tolerieren, wenn wir mit einer Sicherheit von 95% behaupten möchten, insgesamt in unserer Los nur 0,1% defekte Teile zu haben.



Die Werte wurden von der Norm DIN-ISO 2859-1 abgeleitet. Man kann sie aber rechnerisch überprüfen.

Wir suchen eigentlich „nur“ das 95%-Quantil.

Beachten Sie die folgende Regel

Für

- $n/N < 0,05$ ist aus Gründen des einfachen Rechnens die Binomialverteilung zu verwenden.

Welche Verteilung sollte eigentlich für diesen Fall verwendet werden?

- Fälle von großen Stichproben und sehr kleinen AQL (defekten Teilen) ist die Poissonverteilung zu verwenden.

$$\lambda = 800 \cdot 0,001 = 0,8$$

$$P(X \leq 1) = \sum_{k=0}^1 e^{-0,8} \cdot \frac{0,8^k}{k!} \approx 80,87\%$$

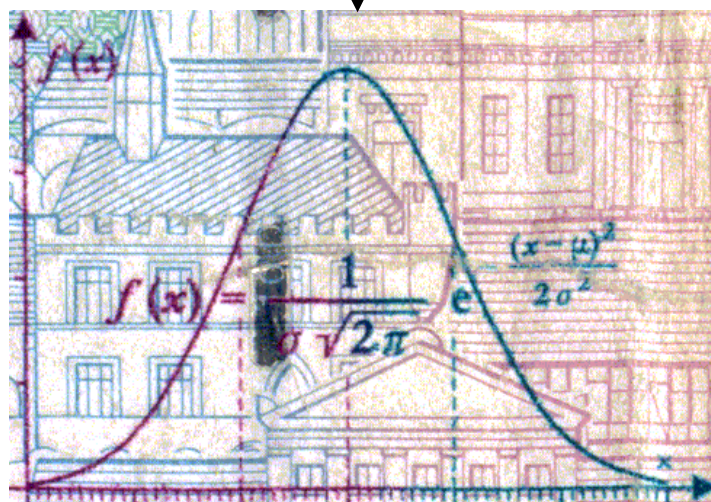
$$P(X \leq 2) = \sum_{k=0}^2 e^{-0,8} \cdot \frac{0,8^k}{k!} \approx 95,25\%$$

Nehmen Sie für die Auswahl, wann welche Verteilung verwendet wird, für N immer den größten der Werte des angegebenen Intervalls (1. Spalte) an.

10 Normalverteilung

10.1 Motivation

Die Normalverteilung ist DIE Grundlage der gegenwärtigen Wahrscheinlichkeitsrechnung und findet sich beinahe überall... Sie ist so bedeutend, dass ein Sonderfall von ihr, neben ihrem Entdecker, Carl Friedrich Gauß, den 10-DM-Schein zierte:



Dies ist die Dichtefunktion der „Standardnormalverteilung“.

Da unendlich viele mögliche Ergebnisse existieren (und daher jedes einzelne Ergebnis im Prinzip die Wahrscheinlichkeit Null hat), gibt die Dichtefunktion einer stetigen Verteilung an, wie dicht die einzelnen Werte in den einzelnen Bereichen liegen.

Die Quantile der Standardnormalverteilung können aus Tabellen abgelesen werden.

Da es aber unendlich viele Normalverteilungen gibt (weil es unendlich viele Werte für Erwartungswert und Varianz geben kann), muss die gegebene Binomialverteilung umgeformt werden, um sie zu vereinheitlichen.

Es gibt unendlich viele Kombinationen von Erwartungswert und Varianz bei Verteilungen.

Allerdings gibt es einen Ausdruck, mit dem immer gerechnet werden kann.

$Z = \frac{X - \mu}{\sigma}$ ist die allgemeine Bezeichnung hierfür. Wenn die Zahl der Durchführungen groß genug ist, ist sie immer normalverteilt.

Sie steht für den Wert der *Standardnormalverteilung*, da $Z \sim N(0,1)$ ist.

Die Quantile dieser Verteilung können aus Tabellen abgelesen werden:

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520

$P(Z \leq 1,96)$
= 97,5%

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4,0	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998

Tabelle 1: Quantile der Standardnormalverteilung.

Man sieht direkt, dass die Fläche auf der linken Seite genau so groß ist wie auf der rechten. Es ist also $P(Z \geq X) = P(Z \leq -X)$. Mit der Gegenwahrscheinlichkeit erhält man

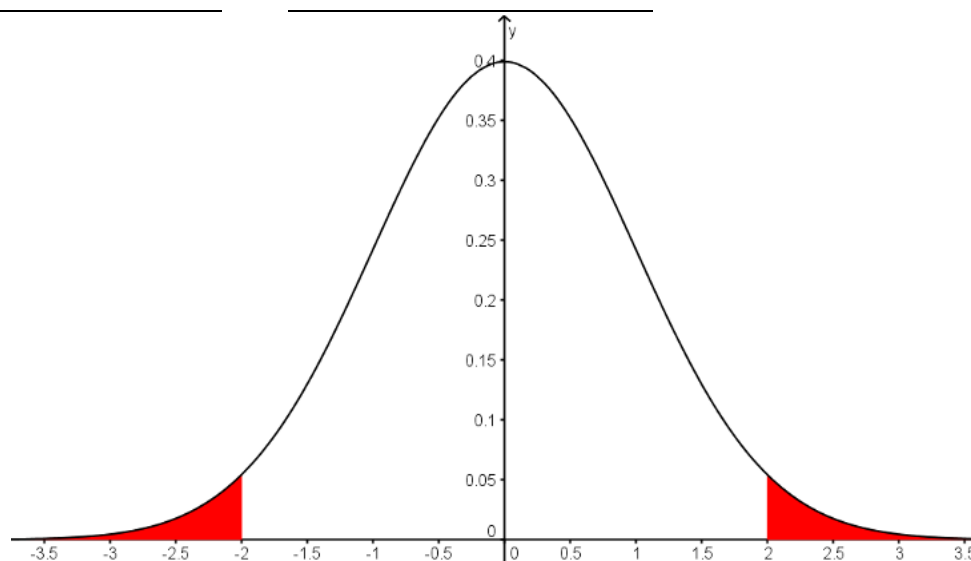
$$P(Z \geq X) = 1 - P(Z < X), \text{ es ist also } 1 - P(Z < X) = P(Z \leq -X) \text{ bzw.}$$

$$P(Z < X) = 1 - P(Z \leq -X).$$

Da es bei stetigen Verteilungen aber über abzählbar unendlich viele Ergebnisse gibt, ist

$$P(Z = X) = 0. \text{ Somit gilt also}$$

$$P(Z \geq X) = 1 - P(Z \leq X) \text{ und } P(Z \leq -X) = 1 - P(Z \leq X)$$



Betrachten wir hierzu ein Beispiel.

10.2 Beispiel

Unser Döner-Restaurant hat einen mittleren wöchentlichen Umsatz von 2.000 €.

Die mittlere Abweichung von diesem Mittel beträgt 150 €. Der Umsatz kann als normalverteilt angenommen werden.

Mit welcher Wahrscheinlichkeit liegt der Umsatz

- unter 1.700 €?
- unter 1.850 €?
- zwischen 1.700 € und 2.300 €?
- zwischen 1.850 € und 2.150 €?
- über 2.150 €?
- über 2.300 €?

Transformieren Sie die Verteilung und benutzen Sie die Tabelle der Standardnormalverteilung zur Lösung!

10.3 Lösung des Beispiels

- a. unter 1.700 €?

$$P\left(Z < \frac{1700 - 2.000}{150}\right) = P(Z < -2) = 1 - P(Z \leq 2) = 1 - 0,97725 = 0,02275 = 2,275\%$$

- b. unter 1.850 €?

$$P\left(Z < \frac{1850 - 2.000}{150}\right) = P(Z < -1) = 1 - P(Z \leq 1) = 1 - 0,84134 = 0,15866 = 15,866\%$$

- c. zwischen 1.700 € und 2.300 €?

$$\begin{aligned} P(1700 < X < 2300) &= P(X < 2300) - P(X < 1700) \\ P\left(Z < \frac{2300 - 2.000}{150}\right) &= P(Z \leq 2) = 0,97725 = 97,725\% \\ P\left(Z < \frac{1700 - 2.000}{150}\right) &= P(Z \leq -2) = 1 - 0,97725\% = 2,275\% \\ P(-2 < Z < 2) &= P(Z \leq 2) - P(Z \leq -2) = P(Z \leq 2) - (1 - P(Z \leq 2)) \\ &= 97,725\% - (1 - 97,725\%) = 97,725\% - 2,275\% = 95,45\% \end{aligned}$$

- d. zwischen 1.850 € und 2.150 €?

$$\begin{aligned} P(1850 < X < 2150) &= P(X < 2150) - P(X < 1850) \\ &= P\left(Z < \frac{2150 - 2.000}{150}\right) - P\left(Z < \frac{1850 - 2.000}{150}\right) \\ &= P(Z \leq 1) - P(Z \leq -1) = 0,84134 - 0,15866 = 0,68268 = 68,268\% \end{aligned}$$

- e. über 2.150 €?

$$\begin{aligned} P(X > 2150) &= 1 - P(X \leq 2150) \\ &= 1 - P\left(Z \leq \frac{2150 - 2.000}{150}\right) = 1 - P(Z \leq 1) = 1 - 0,84134 = 0,15866 = 15,866\% \end{aligned}$$

- f. über 2.300 €?

$$\begin{aligned} P(X > 2300) &= 1 - P(X \leq 2300) \\ &= 1 - P\left(Z \leq \frac{2300 - 2.000}{150}\right) = 1 - P(Z \leq 2) = 1 - 0,97725 = 0,02275 = 2,275\% \end{aligned}$$

10.4 Annäherung diskreter Verteilungen

Mit der Normalverteilung können diskrete Verteilungen „approximiert“ werden. Das bedeutet, dass statt der diskreten Verteilung die Normalverteilung verwendet wird, bspw. wenn es keine Tabelle gibt, aus der die Werte abgelesen werden können.

Beispielsweise nähert sich die Binomialverteilung der Normalverteilung an, wenn die Zahl der Durchführungen steigt. Wir kennen dieses Vorgehen bereits aus der Differentialrechnung, es heißt „Konvergenz“.

In diesem Fall handelt es sich um Konvergenz in Verteilung, man schreibt:

$B(n, p) \xrightarrow[n \rightarrow \infty]{d} N(\mu, \sigma)$, wobei d für das englische Wort für Verteilung, distribution, steht.

Betrachten wir beispielsweise den Fall, dass eine faire Münze 100 Mal geworfen wird. Wie groß ist die Wahrscheinlichkeit, dass höchstens 60-mal Kopf erscheint?

Diese Frage kann mit der Binomialverteilung und Tabellen beantwortet werden, denn wir suchen die Wahrscheinlichkeit für das Ereignis $E_1 : X \leq 60$.

Wenn man nur eine Tabelle für alle Fälle verwenden möchte, hilft die Normalverteilung.

Die gesuchte Wahrscheinlichkeit ist also: $P(X \leq 60)$.

Die Zufallsvariable nennen wir einfach X , sie ist binomialverteilt. $X \sim B(n=100, p=0,5)$

Erwartungswert ist: $\mu = 0,5 \cdot 100 = 50$

Die Varianz ist: $\sigma^2 = 0,5 \cdot 0,5 \cdot 100 = 25$

1. Die sich ergebende Wahrscheinlichkeit, wenn wir bei $P(X \leq 60)$ auf beiden Seiten der Ungleichung den Erwartungswert abziehen, ist $P(X - 50 \leq 10)$.

Die ersten beiden Momente sind:

$$E(X - 50) = E(X) - E(50) = 50 - 50 = 0$$

$$\text{Var}(X - 50) = \text{Var}(X) = \sigma^2 = 25$$

2. Nun müssen wir bei $P(X - 50 \leq 10)$ auf beiden Seiten durch die Standardabweichung teilen.

3. Es ergibt sich: $P\left(\frac{X - 50}{5} \leq 2\right)$

4. Für Erwartungswert und Varianz von $\frac{X - 50}{5} = \frac{1}{5}(X - 50)$ gilt:

$$E\left(\frac{X - 50}{5}\right) = \frac{1}{5}(E(X) - E(50)) = \frac{1}{5}(50 - 50) = 0$$

$$\text{Var}\left(\frac{X - 50}{5}\right) = \text{Var}\left(\frac{X}{5} - 10\right) = \left(\frac{1}{5}\right)^2 \text{Var}(X) = \frac{1}{25} \sigma^2 = \frac{1}{25} \cdot 25 = 1$$

Noch einmal: $Z = \frac{X - \mu}{\sigma}$ ist die allgemeine Bezeichnung hierfür. Wenn die Zahl der Durchführungen groß genug ist, ist sie immer normalverteilt.

Sie steht für den Wert der *Standardnormalverteilung*, da $Z \sim N(0,1)$ ist.

Die Standardnormalverteilung ist die Funktion, die wir schon auf dem 10-DM-Schein gesehen haben. Wir haben also einen Weg gefunden, aus allen Verteilungen schnell die Standardnormalverteilung zu berechnen:

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

$\varphi(Z)$: Dichtefunktion, kleiner Buchstabe Phi

und zudem gilt $\Phi(Y) = P(Z \leq Y)$ (großes Phi)

Für jede beliebige Verteilung ist $\frac{X - \mu}{\sigma / \sqrt{n}} \overset{a}{\sim} N(0,1)$. ($\overset{a}{\sim}$: approximativ: für große Zahlen)

Eine Annäherung (Approximation) der Binomial- durch die Normalverteilung ist nur zulässig, falls $\text{Var}(X) \geq 9$

Normalverteilungen sind stetige Verteilungen.

Eine diskrete Verteilung ordnet nur diskreten (endlich oder abzählbar unendlich vielen) Werten eine Wahrscheinlichkeit zu.

Beispiel: Wahrscheinlichkeit für ein fehlerhaftes Teil bei Überprüfung von fünf Teilen.

Eine stetige Verteilung ordnet stetigen Werten (überabzählbar unendlich vielen) Werten eine Wahrscheinlichkeit zu.

Beispiel: Wahrscheinlichkeit einer bestimmten Körpergröße in der Bevölkerung.

Hinweis: Ist eine diskrete Verteilung (z. B. die Binomialverteilung) als Grundlage gegeben und man berechnet zur Annäherung die Wahrscheinlichkeiten mit der Normalverteilung, muss man das Intervall auf jeder Seite jeweils um 0,5 verbreitern. (Dies wird „Stetigkeitskorrektur“ genannt und dient dazu, die Annäherung genauer zu machen, da immer eine halbe Säule abgeschnitten wird.)

Also:

$$\begin{aligned} P(0 \leq X \leq 60) | BV &= P(X \leq 60) - P(X \leq 0) \approx P\left(Z \leq \frac{60 + 0,5 - 50}{5}\right) - P\left(Z \leq \frac{0 - 0,5 - 50}{5}\right) \\ &= P(Z \leq 2,1) - P(Z \leq -10,1) \\ &= \Phi(2,1) - \Phi(-10,1) \approx 0,9821 - 0 = 98,21\% \end{aligned}$$

Ohne Stetigkeitskorrektur ergäbe sich: also:

$$P\left(\frac{0 - 50}{5} \leq Z \leq \frac{60 - 50}{5}\right) = P(Z \leq 2) - P(Z \leq -10) = \Phi(2) - \Phi(-10) \approx 0,9772 - 0 = 97,72\%$$

(Die Quantile der Standardnormalverteilung sind in der folgenden Tabelle dargestellt.)

Da $\Phi(-Z) = 1 - \Phi(Z)$, sind sowohl $\Phi(-10) \approx \Phi(-10,1) \approx 0$.

In der Praxis wird häufig das Abziehen der Untergrenze daher weggelassen.

10.5 Übungen

Übung 1

Ein Unternehmen stellt Nägel her, die im Mittel 10 cm lang sind mit einer Standardabweichung von 0,5 cm. Die Größe ist normalverteilt.

Mit welcher Wahrscheinlichkeit liegt die Größe eines willkürlich herausgegriffenen Nagels

- bei genau 10 cm?
- unter 10,8 cm?
- über 9,2 cm?
- zwischen 9,5 und 10,5 cm?

Übung 2

Die durchschnittliche Nutzungsdauer eines Smartphones beträgt 18 Monate mit einer Standardabweichung von 6 Monaten. Die Nutzungsdauer ist normalverteilt.

Mit welcher Wahrscheinlichkeit liegt die Nutzungsdauer

- unter 18 Monaten?
- über 20 Monaten?
- zwischen 18 und 24 Monaten?
- zwischen 12 und 24 Monaten?

Übung 3

Die durchschnittliche Laufleistung eines Dienstwagens in einem Unternehmen beträgt 50.000 km/Jahr mit einer Standardabweichung von 20.000 km/Jahr.

Mit welcher Wahrscheinlichkeit liegt die Nutzungsdauer eines Wagens

- über 60.000 km/Jahr?
- unter 30.000 km/Jahr?
- im Bereich von einer Standardabweichung um den Erwartungswert?
- im Bereich von zwei Standardabweichungen um den Erwartungswert?

Übung 4

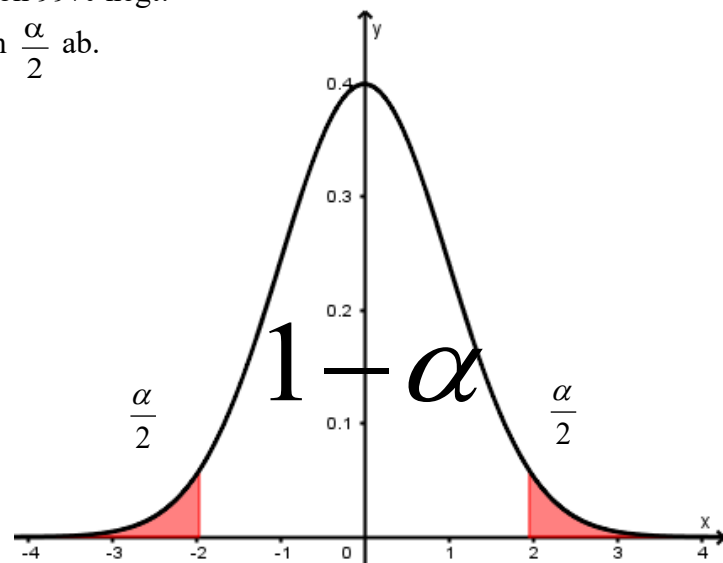
Aus einem anderen Unternehmen:

Mittlere Laufleistung: 50.000 km/Jahr, Standardabweichung: 10.000 km/Jahr.

Bestimmen Sie einen symmetrischen Bereich um den Erwartungswert, in dem

- die Laufleistung mit einer Wahrscheinlichkeit von 90% liegt.
- die Laufleistung mit einer Wahrscheinlichkeit von 95% liegt.
- die Laufleistung mit einer Wahrscheinlichkeit von 99% liegt.

Zu beachten: Sie schneiden links und rechts folglich $\frac{\alpha}{2}$ ab.



10.6 Lösungen

Übung 1

- a. bei genau 10 cm?

$$P\left(Z = \frac{10-10}{0,5}\right) = 0\%$$

- b. unter 10,8 cm?

$$P\left(Z < \frac{10,8-10}{0,5}\right) = P(Z < 1,6) = 94,52\%$$

- c. über 9,2 cm?

$$P\left(Z > \frac{9,2-10}{0,5}\right) = P(Z \geq -1,6) = P(Z \leq 1,6) = 94,52\%$$

- d. zwischen 9,5 und 10,5 cm?

$$\begin{aligned} P\left(\frac{9,5-10}{0,5} < Z < \frac{10,5-10}{0,5}\right) &= P(Z \leq 1) - P(Z \leq -1) = P(Z \leq 1) - (1 - P(Z \leq 1)) \\ &= 2 \cdot P(Z \leq 1) - 1 = 0,6826 = 68,26\% \end{aligned}$$

Übung 2

- a. unter 18 Monaten?

$$P\left(Z < \frac{18-18}{6}\right) = P(Z < 0) = 50\%$$

- b. über 20 Monaten?

$$P\left(Z > \frac{20-18}{6}\right) = 1 - P\left(Z \leq \frac{20-18}{6}\right) = 1 - P(Z \leq 0,3333) = 1 - 0,6293 = 37,07\%$$

- c. zwischen 18 und 24 Monaten?

$$\begin{aligned} P\left(\frac{18-18}{6} < Z < \frac{24-18}{6}\right) &= P\left(Z \leq \frac{24-18}{6}\right) - P\left(Z \leq \frac{18-18}{6}\right) = P(Z \leq 1) - P(Z \leq 0) \\ &= 0,8413 - 0,5 = 34,13\% \end{aligned}$$

- d. zwischen 12 und 24 Monaten?

$$\begin{aligned} P\left(\frac{12-18}{6} < Z < \frac{24-18}{6}\right) &= P\left(Z \leq \frac{24-18}{6}\right) - P\left(Z \leq \frac{12-18}{6}\right) = P(Z \leq 1) - P(Z \leq -1) \\ &= 0,8413 - (1 - 0,8413) = 68,26\% \end{aligned}$$

Übung 3

- a. über 60.000 km/Jahr?

$$P\left(Z > \frac{60.000-50.000}{20.000}\right) = P(Z > 0,5) = 1 - P(Z \leq 0,5) = 1 - 0,6915 = 30,85\%$$

- b. unter 30.000 km/Jahr?

$$P\left(Z < \frac{30.000-50.000}{20.000}\right) = P(Z < -1) = 1 - P(Z \leq 1) = 1 - 0,8413 = 15,87\%$$

- c. im Bereich von einer Standardabweichung um den Erwartungswert?

$$P(30.000 \leq X \leq 70.000)$$

$$P\left(\frac{30.000 - 50.000}{20.000} \leq Z \leq \frac{70.000 - 50.000}{20.000}\right) = P(Z \leq 1) - P(Z \leq -1) = P(Z \leq 1) - (1 - P(Z \leq 1)) \\ = 0,8413 - 0,1587 = 68,26\%$$

- d. im Bereich von zwei Standardabweichungen um den Erwartungswert?

$$P(10.000 \leq X \leq 90.000)$$

$$P\left(\frac{10.000 - 50.000}{20.000} \leq Z \leq \frac{90.000 - 50.000}{20.000}\right) = P(Z \leq 2) - P(Z \leq -2) \\ = P(Z \leq 2) - (1 - P(Z \leq 2)) = 0,9772 - 0,0228 = 95,44\%$$

Übung 4

Aus einem anderen Unternehmen:

Laufleistung: 50.000 km/Jahr, Standardabweichung: 10.000 km/Jahr.

Bestimmen Sie einen symmetrischen Bereich um den Erwartungswert, in dem

- a. die Laufleistung mit einer Wahrscheinlichkeit von 90% liegt.

$$z_{1-\frac{\alpha}{2}} = z_{0,95} = 1,645 \Rightarrow [50.000 \pm 1,645 \cdot 10.000] = [33.550; 66.450]$$

- b. die Laufleistung mit einer Wahrscheinlichkeit von 95% liegt.

$$z_{1-\frac{\alpha}{2}} = z_{0,975} = 1,96 \Rightarrow [50.000 \pm 1,96 \cdot 10.000] = [30.400; 69.600]$$

- c. die Laufleistung mit einer Wahrscheinlichkeit von 99% liegt.

$$z_{1-\frac{\alpha}{2}} = z_{0,995} = 2,575 \Rightarrow [50.000 \pm 2,575 \cdot 10.000] = [24.250; 75.750]$$

11 Schätzen – Konfidenzintervalle

11.1 Motivation

„In welchem Bereich ist der Erwartungswert mit 95% Sicherheit?“

Die wenigsten Menschen würden auf diese Frage einen genauen Wert nennen. Es werden wohl eher Antworten kommen wie...

„Ein Großteil der Frauen ist zwischen 1,68 m und 1,74 m, bei den Männern sind die meisten zwischen 1,80 m und 1,86 m groß.“

Statistiker geben gern genauere Antworten, bspw.

„Mit 95% Sicherheit ist der Erwartungswert zwischen 1,84 m. und 1,86 m.“

Einige Vokabeln

Grundgesamtheit:

Alle Objekte, die gewisse Eigenschaften aufweisen (bspw. alle Menschen).

Stichprobe:

Teil der Grundgesamtheit, der untersucht wird (bspw. die Personen, die befragt werden).

Doch wie kommen sie darauf?

Das ist ganz einfach... Hier entfaltet die Standardnormalverteilung ihre Macht!

Was wir schon wissen: $z = \frac{x - \mu}{\sigma}$

Formen wir doch mal um... $\Rightarrow x = \mu + z\sigma \Rightarrow \mu = x - z\sigma$

Da wir einen Bereich (links und rechts um den Erwartungswert suchen), brauchen wir also $x - z\sigma \leq \mu \leq x + z\sigma$

Für uns gilt also, da wir uns für das arithmetische Mittel interessieren:

$$\bar{x} - z\sigma_{\bar{x}} \leq \mu_{\bar{x}} \leq \bar{x} + z\sigma_{\bar{x}}$$

Wir interessieren uns für einen Bereich um den Erwartungswert...

Nur links und rechts werden gleich große

Stücke der Normalverteilung abgeschnitten.

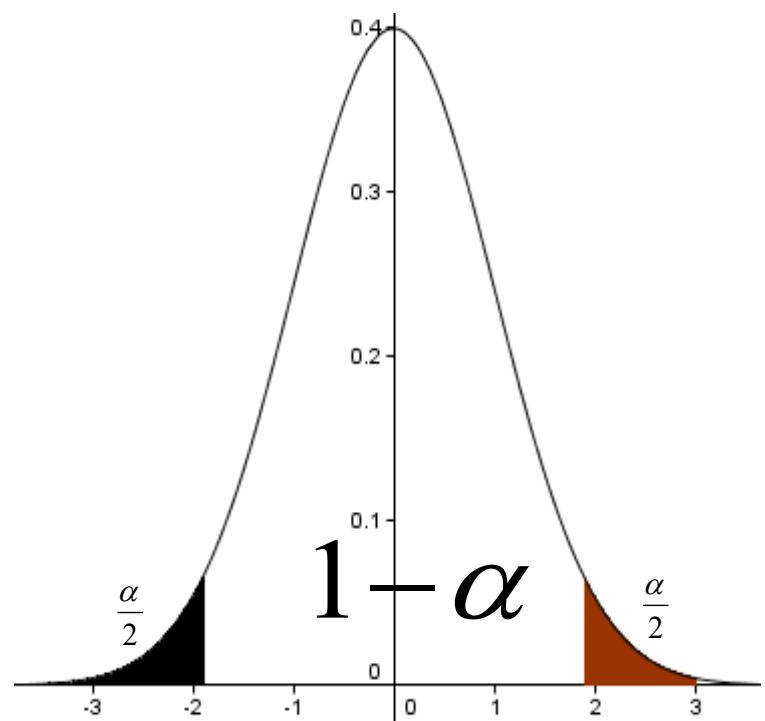
Wenn wir also mit 95%iger Sicherheit (einer Irrtumswahrscheinlichkeit („Signifikanzniveau“) von 5%) eine Aussage treffen wollen, so schneiden wir auf jeder Seite 2,5% Fläche von der Verteilung ab und benötigen also:

$$z_{0,025} \text{ bzw. } z_{1-0,025} = z_{0,975}.$$

Die beiden Werte unterscheiden sich nur im Vorzeichen ($z_{0,025} = -z_{0,975}$), was man auch in der Grafik sieht.

Diese Irrtumswahrscheinlichkeit wird in der Regel mit α bezeichnet.

„Irrtumswahrscheinlichkeit“ heißt es deshalb, weil wir auf Basis der Stichprobe ausschließen, dass der Erwartungswert derart weit außen ist. Je größer Alpha wird (je unsicherer die Vorhersage), desto kleiner das Konfidenzintervall und umgekehrt.



Der Erwartungswert liegt mit einer Wahrscheinlichkeit von $1 - \alpha$ in dem Konfidenzintervall (Vertrauensbereich).

Nochmal zurück: Wir wissen jetzt, wie wir z bestimmen.

Allerdings kennen wir unseren Erwartungswert der Größe der Frauen bzw. Männer nicht. Wir wissen aber, wie die Größe im Allgemeinen schwankt, unser Wert σ ist also bekannt.

(Dieser Fall kommt zwar eher selten vor, ist aber wichtig, um das Prinzip zu verstehen!)

Wir haben nur eine Stichprobe, von, sagen wir, 9 Männern, die wir gemessen haben und die mittlere Größe ist 1,82 m. Die Standardabweichung der Größe beträgt in der männlichen Bevölkerung 0,05 m.

Das Vorgehen ist nun immer das gleiche bei Konfidenzintervallen:

Wir müssen zuerst die Verteilung des Gesuchten (hier: arithmetisches Mittel) bestimmen.

- Das arithmetische Mittel ist in einer großen Grundgesamtheit normalverteilt, da jeder Wert normalverteilt ist. (Dinge aus der Natur werden i. d. R. als normalverteilt angenommen.)
- Der Erwartungswert des arithmetischen Mittels beträgt

$$E\left(\frac{1}{n} \sum_{k=1}^n x_k\right) = \frac{1}{n} \sum_{k=1}^n E(x_k) = \frac{1}{n} \sum_{k=1}^n \mu = \frac{1}{n} \cdot n\mu = \mu$$

Hierbei sind die Werte x_k Zufallsvariablen, die $\sim (\mu; \sigma)$, aufgrund der Normalverteilung sogar $\sim N(\mu; \sigma)$, sind.

- Die Varianz des arithmetischen Mittels ist:

$$\sigma^2_{\bar{x}} = \text{Var}\left(\frac{1}{n} \sum_{k=1}^n x_k\right) = \frac{1}{n^2} \text{Var}\left(\sum_{k=1}^n x_k\right) = \frac{1}{n^2} \left(\sum_{k=1}^n \text{Var}(x_k)\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{1}{n} \sigma^2 \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(Klar: Je mehr Werte wir zur Schätzung verwenden, desto genauer wird die Schätzung!)

Es ergibt sich also für das Konfidenzintervall (Vertrauensbereich), geschätzt mit der Sicherheit α , wenn wir die Formel $\mu - z\sigma_{\bar{x}} \leq x \leq \mu + z\sigma_{\bar{x}}$ zur Bestimmung der Grenzen verwenden:

$$KI_{1-\alpha} = \left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

Je mehr Fische ich mit einem Netz fangen möchte, desto größer sollte das Netz sein!

Der Wert $\frac{\alpha}{2}$ stammt von der Tatsache, dass es ein beidseitiges Intervall ist (s. Grafik).

Diese Formel gilt aber nur, wenn die Standardabweichung der Grundgesamtheit (hier: der Körpergröße aller Männer in diesem Alter) bekannt ist!

11.2 Beispiel

Für unseren Fall ($\bar{x} = 1,82, \sigma = 0,05, n = 9$) ergibt sich:

$$KI_{1-0,05} = KI_{0,95} = \left[1,82 - 1,96 \frac{0,05}{\sqrt{9}}; 1,82 + 1,96 \frac{0,05}{\sqrt{9}} \right] \approx [1,79; 1,85]$$

Rechnen Sie selbst nach!

Aussage: Auf Basis der Stichprobe kann gesagt werden, dass der Erwartungswert der Größe der Männer mit 95% Wahrscheinlichkeit 1,79m-1,85m ist.

11.3 Konfidenzintervall bei unbekannter Varianz

11.3.1 Erklärung

Wenn wir die Varianz nicht kennen, müssen wir sie schätzen. Wir dürfen dann aber auch nicht mehr die Normalverteilung verwenden

Jetzt ist das Konfidenzintervall gegeben durch

$$KI_{1-\alpha} = \left[\bar{x} - t_{1-\frac{\alpha}{2}}(n-1) \frac{\hat{s}}{\sqrt{n}}; \bar{x} + t_{1-\frac{\alpha}{2}}(n-1) \frac{\hat{s}}{\sqrt{n}} \right]$$

Die einzigen Unterschiede zum Konfidenzintervall bei bekannter Varianz ist die Verwendung der aus Daten geschätzten Varianz statt einer bereits bekannten Varianz und die Verwendung der t-Verteilung statt der Normalverteilung.

Eine geschätzte Varianz wird berechnet durch $\hat{s}^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$.

(Bei der geschätzten Varianz wird nur durch $n-1$ geteilt, weil auch das arithmetische Mittel ja nicht bekannt ist, sondern auf Basis der Daten ermittelt wurde, von denen auch die Varianz geschätzt wurde.

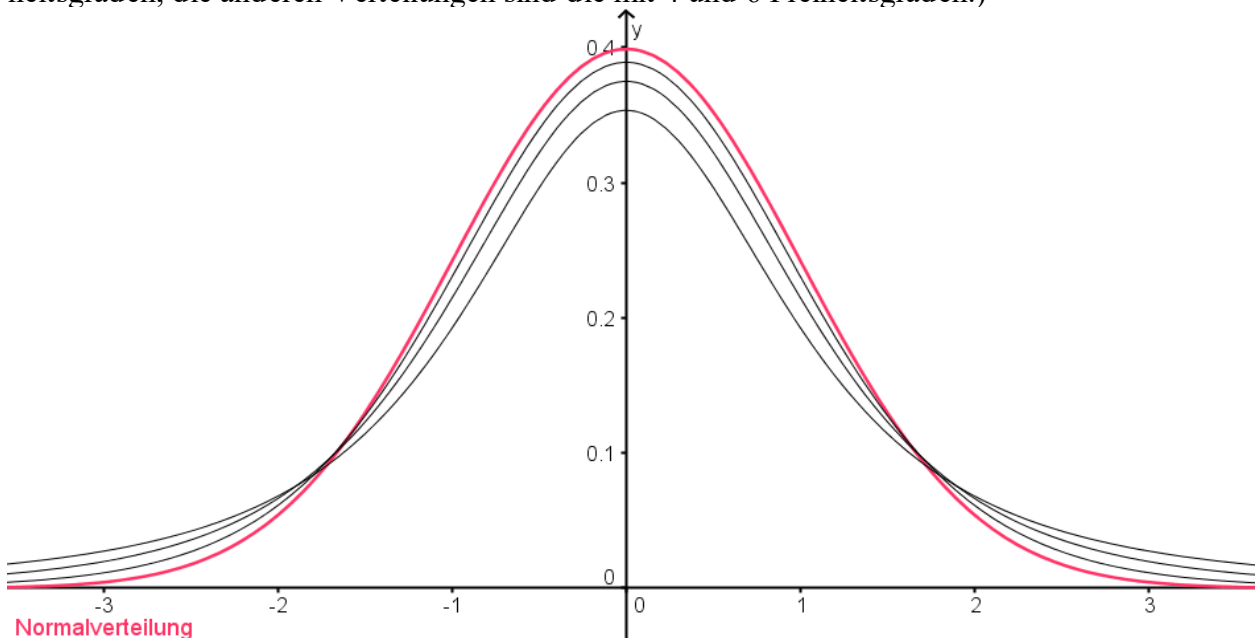
Der Wert $n-1$ wird verwendet, da das arithmetische Mittel auch nur aus der Stichprobe berechnet wird und somit unsicher ist, ob dies der tatsächliche Durchschnitt der Grundgesamtheit (hier: aller Menschen) ist.

Die t-Verteilung hat $\nu = n-1$ sogenannte „Freiheitsgrade“.

Was bedeuten diese Freiheitsgrade?

Schauen wir uns einmal t- und Standardnormalverteilung im Vergleich an:

(Die an den Rändern höhere und in der Mitte niedrigste Verteilung ist die t-Verteilung mit 2 Freiheitsgraden, die anderen Verteilungen sind die mit 4 und 6 Freiheitsgraden.)



Da n die Größe der Stichprobe bezeichnet, liegt die Vermutung nahe, dass die t-Verteilung gegen die Standardnormalverteilung konvergiert, dass also gilt: $T(n-1) \xrightarrow[n \rightarrow \infty]{d} N(0;1)$. Hier haben wir

also wieder einen Fall von Konvergenz in Verteilung – wie bei der hypergeometrischen und Binomialverteilung sowie der Binomial- und der Normalverteilung.

Wenn wir also die Normalverteilung statt der t-Verteilung verwenden, sind die Basisdaten normalverteilt oder die Stichprobe ist sehr groß.

Aus diesem Grund wird zur Vereinfachung bei $\nu = n - 1 \geq 30$ bei Konfidenzintervallen und unbekannter Varianz wieder die Normalverteilung verwendet.

Die t-Verteilung ist symmetrisch wie die Normalverteilung, es gilt also $t_\alpha = -t_{1-\alpha}$

11.3.2 Beispiel

Schauen wir uns ein Beispiel an:

Wie lautet das Konfidenzintervall der mittleren Körpergröße bei einer Irrtumswahrscheinlichkeit von 5%? Hierzu wurden in einem Jahrgang die folgenden Daten gemessen:

Person	1	2	3	4	5	6	7	8	9
Größe	182	178	190	175	199	182	185	180	185

Es gilt hier: $\bar{x} = \frac{1656}{9} = 184, s^2 = \frac{1}{9-1} \cdot 404 = 50,5 \Rightarrow \hat{s} \approx 7,11, t_{1-\frac{0,05}{2}}(9-1) = t_{0,975}(8) = 2,306$

Somit ergibt sich für das Konfidenzintervall:

$$KI_{0,95} = \left[184 - 2,306 \frac{7,11}{3}; 184 + 2,306 \frac{7,11}{3} \right] = [179; 189]$$

Die Quantile der t-Verteilung können aus der folgenden Tabelle abgelesen werden:

	$1 - \alpha/2$ bzw. $1 - \alpha$							
Freiheitsgrade	0,75	0,875	0,90	0,95	0,975	0,99	0,995	0,999
1	1,000	2,414	3,078	6,314	12,706	31,821	63,657	318,309
2	0,816	1,604	1,886	2,920	4,303	6,965	9,925	22,327
3	0,765	1,423	1,638	2,353	3,182	4,541	5,841	10,215
4	0,741	1,344	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,301	1,476	2,015	2,571	3,365	4,032	5,893
6	0,718	1,273	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,254	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,240	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,230	1,383	1,833	2,262	2,821	3,250	4,297
10	0,700	1,221	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,214	1,363	1,796	2,201	2,718	3,106	4,025
12	0,695	1,209	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,204	1,350	1,771	2,160	2,650	3,012	3,852
14	0,692	1,200	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,197	1,341	1,753	2,131	2,602	2,947	3,733
16	0,690	1,194	1,337	1,746	2,120	2,583	2,921	3,686
17	0,689	1,191	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,189	1,330	1,734	2,101	2,552	2,878	3,610
19	0,688	1,187	1,328	1,729	2,093	2,539	2,861	3,579
20	0,687	1,185	1,325	1,725	2,086	2,528	2,845	3,552
21	0,686	1,183	1,323	1,721	2,080	2,518	2,831	3,527
22	0,686	1,182	1,321	1,717	2,074	2,508	2,819	3,505
23	0,685	1,180	1,319	1,714	2,069	2,500	2,807	3,485
24	0,685	1,179	1,318	1,711	2,064	2,492	2,797	3,467
25	0,684	1,178	1,316	1,708	2,060	2,485	2,787	3,450
26	0,684	1,177	1,315	1,706	2,056	2,479	2,779	3,435
27	0,684	1,176	1,314	1,703	2,052	2,473	2,771	3,421
28	0,683	1,175	1,313	1,701	2,048	2,467	2,763	3,408
29	0,683	1,174	1,311	1,699	2,045	2,462	2,756	3,396
30	0,683	1,173	1,310	1,697	2,042	2,457	2,750	3,385
40	0,681	1,167	1,303	1,684	2,021	2,423	2,704	3,307
50	0,679	1,164	1,299	1,676	2,009	2,403	2,678	3,261
60	0,679	1,162	1,296	1,671	2,000	2,390	2,660	3,232
70	0,678	1,160	1,294	1,667	1,994	2,381	2,648	3,211
80	0,678	1,159	1,292	1,664	1,990	2,374	2,639	3,195
90	0,677	1,158	1,291	1,662	1,987	2,368	2,632	3,183
100	0,677	1,157	1,290	1,660	1,984	2,364	2,626	3,174
500	0,675	1,152	1,283	1,648	1,965	2,334	2,586	3,107

Tabelle 2: Quantile der t-Verteilung.

11.4 Übungen

Übung 1

In einer Stichprobe von 20 Männern ist die mittlere Körpergröße 1,82 m. Es ist bekannt, dass die Größe in der Bevölkerung im Mittel um 0,05 m um den Erwartungswert schwankt.

In welchem Körpergröße-Bereich liegt der Erwartungswert mit einer Sicherheit von 90% bzw. 95%?

Übung 2

Wir befragen eine Stichprobe von 20 Männern bzgl. ihrer Körpergröße.

Die mittlere Größe in der Stichprobe beträgt 1,80 m.

Die Standardabweichung der Größe in der männlichen Bevölkerung beträgt 0,08 m

Bestimmen Sie den Körpergröße-Bereich, in dem der Erwartungswert sich mit einer Sicherheit von 90% bzw. 95% befindet.

Übung 3

Zurück zu unserem Beispiel der Nägel:

Ein Unternehmen stellt Nägel her.

In einer Stichprobe vom Umfang $n = 100$ ergibt sich eine mittlere Größe von 9,8 cm.

Aus Erfahrung wissen wir, dass die Standardabweichung in der Produktion bei $\sigma = 1$ liegt.

Bestimmen Sie das Konfidenzintervall zum Signifikanzniveau

3.1 $\alpha = 10\%$,

3.2 $\alpha = 5\%$,

3.3 $\alpha = 1\%$.

Übung 4

Wir befragen eine Stichprobe von 5 Männern bzgl. ihrer Körpergröße.

Die mittlere Größe in der Stichprobe beträgt 1,80 m, die Standardabweichung der Körpergröße in der Stichprobe ist 0,08.

Bestimmen Sie den Körpergröße-Bereich, in dem der Erwartungswert sich mit einer Sicherheit von 90% bzw. 95% befindet.

Übung 5

Der Schokoladenriegel „Star-Choc“ hat aufgrund der enthaltenen ganzen Haselnüsse leichte Schwankungen im Gewicht. Zur Qualitätskontrolle werden 100 Riegel entnommen.

Das durchschnittliche Gewicht beträgt 26 g, die Standardabweichung in dieser Stichprobe 2,5 g.

Bestimmen Sie den Bereich, in dem sich der Erwartungswert des Gewichts mit einer Sicherheit von 95% bzw. 99% befindet.

Übung 6: Zum Nachdenken

Was passiert mit dem Konfidenzintervall, wenn

6.1 die Menschen mit der Zeit größer werden?

6.2 die Größenunterschiede in der Bevölkerung zunehmen?

6.3 das Intervall einen größeren Teil der Bevölkerung umfassen soll?

6.4 die Stichprobe vergrößert wird?

11.5 Lösungen

Übung 1

$$\alpha = 0,1 \Rightarrow z_{1-0,05} = z_{0,95} = 1,6449, \alpha = 0,05 \Rightarrow z_{1-0,025} = z_{0,975} = 1,96$$

$$KI_{1-0,1} = \left[1,82 \pm 1,645 \cdot \frac{0,05}{\sqrt{20}} \right] \approx [1,82 \pm 0,0184] \approx [1,8016; 1,8384]$$

$$KI_{1-0,05} = \left[1,82 \pm 1,96 \cdot \frac{0,05}{\sqrt{20}} \right] = [1,82 \pm 0,0219] \approx [1,7981; 1,8419]$$

Übung 2

$$\alpha = 0,1 \Rightarrow z_{1-0,05} = z_{0,95} = 1,645, \alpha = 0,05 \Rightarrow z_{1-0,025} = z_{0,975} = 1,96$$

$$KI_{1-0,1} = \left[1,80 \pm 1,645 \cdot \frac{0,08}{\sqrt{20}} \right] \approx [1,80 \pm 0,0294] \approx [1,7706; 1,8294]$$

$$KI_{1-0,05} = \left[1,80 \pm 1,96 \cdot \frac{0,08}{\sqrt{20}} \right] = [1,80 \pm 0,0351] \approx [1,7649; 1,8351]$$

Übung 3

$$3.1 \quad \alpha = 10\% \Rightarrow z_{0,95} \approx 1,645 \Rightarrow KI_{0,9} = \left[9,8 \pm 1,645 \cdot \frac{1}{\sqrt{100}} \right] \approx [9,6355; 9,9645]$$

$$3.2 \quad \alpha = 5\% \Rightarrow z_{0,975} \approx 1,96 \Rightarrow KI_{0,95} = \left[9,8 \pm 1,96 \cdot \frac{1}{\sqrt{100}} \right] \approx [9,604; 9,996]$$

$$3.3 \quad \alpha = 1\% \Rightarrow z_{0,995} \approx 2,575 \Rightarrow KI_{0,99} = \left[9,8 \pm 2,575 \cdot \frac{1}{\sqrt{100}} \right] \approx [9,5425; 10,0575]$$

Übung 4

$$\alpha = 0,1 \Rightarrow t_{1-0,05} (5-1) = t_{0,95} (4) = 2,1319, \alpha = 0,05 \Rightarrow t_{1-0,025} (5-1) = t_{0,975} (4) = 2,7765$$

$$KI_{1-0,1} = KI_{0,9} = \left[1,80 \pm 2,1319 \cdot \frac{0,08}{\sqrt{5}} \right] \approx [1,80 \pm 0,0763] \approx [1,725; 1,876]$$

$$KI_{1-0,05} = KI_{0,95} = \left[1,80 \pm 2,7765 \cdot \frac{0,08}{\sqrt{5}} \right] = [1,80 \pm 0,0993] \approx [1,701; 1,899]$$

Übung 5

$$\alpha = 5\% \Rightarrow t_{1-0,025} (100-1) = t_{0,975} (99) = 1,984, \alpha = 0,01 \Rightarrow t_{1-0,005} (100-1) = t_{0,995} (99) = 2,626$$

$$KI_{1-0,05} = \left[26 \pm 1,984 \cdot \frac{2,5}{\sqrt{100}} \right] \approx [26 \pm 0,496] \approx [25,504; 26,496]$$

$$KI_{1-0,01} = \left[26 \pm 2,626 \cdot \frac{2,5}{\sqrt{100}} \right] \approx [26 \pm 0,6565] \approx [25,344; 26,656]$$

Übung 6

6.1 Intervall verschiebt sich nach rechts, da der Erwartungswert (Mitte) steigt.

6.2 Intervall wird breiter, da sich die Streuung erhöht.

6.3 Intervall wird größer, da die Sicherheit der Vorsage erhöht werden soll.

6.4 Intervall wird kleiner, da n erhöht wird und somit die Sicherheit steigt.

11.6 Schätzen eines Anteilswerts

Beispiel: Eine Supermarktkette behauptet, dass nur 30% der Kunden biologisch essen. Das Sortiment in diesem Bereich müsste nicht erweitert werden, sondern mehr auf herkömmliche Landwirtschaftsprodukte gesetzt werden.

In einer Umfrage unter den Kunden fragte eine Tierschutzorganisation nach, ob die Kunden gern biologisch angebaute Lebensmittel essen (mögliche Antworten: ja/nein).

Von 400 Befragten gaben 240 an, gern solche Lebensmittel zu essen.

Es gibt nur zwei Antworten, gleiche Personen werden nicht noch einmal gefragt, es handelt sich somit um eine Binomialverteilung.

Wissen wir schon: Wenn viele Personen befragt werden, kann die Binomialverteilung durch die Normalverteilung approximiert werden.

Wie beim Test auf Erwartungswert ergibt sich das Konfidenzintervall als

$$KI_{1-\alpha} = \left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

Es gibt hier nur die Ausprägungen 0 (nein) und 1 (ja).

Somit ergibt sich als Mittelwert:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{n_i}{n} = \hat{p} \quad (\text{Anteil aus der Stichprobe})$$

$$\sigma^2 = x^2 - \bar{x}^2$$

Da es nur die Ausprägungen 0 und 1 gibt, gilt: $\overline{x^2} = \bar{x}$

Es ergibt sich also:

$$\sigma^2 = \hat{p} - \hat{p}^2 = \hat{p} \cdot (1 - \hat{p})$$

Setzen wir also ein, so ergibt sich:

$$KI_{1-\alpha} = \left[\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}; \hat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \right]$$

Kommen wir zurück zu unserem Ausgangsbeispiel:

$$\hat{p} = \frac{240}{400} = 60\% = 0,6$$

Nehmen wir $\alpha = 5\%$.

Somit ergibt sich mit $z_{1-\frac{0,05}{2}} = z_{0,975} = 1,96$

$$KI_{1-0,05} = \left[0,6 - 1,96 \cdot \sqrt{\frac{0,6 \cdot (1-0,6)}{400}}; 0,6 + 1,96 \cdot \sqrt{\frac{0,6 \cdot (1-0,6)}{400}} \right]$$

$$KI_{0,95} = [0,6 \pm 1,96 \cdot 0,024] \approx [0,6 \pm 0,048] = [0,552; 0,648]$$

11.7 Übungen zu Anteilswerten

1. Die Umsätze des Unternehmens „Startech“ sind gesunken, obwohl der Marketingleiter behauptet, dass das Unternehmen sehr bekannt sei. Eine Befragung ergab, dass von 200 Personen 160 das Unternehmen kennen. In welchem Bereich liegt die Bekanntheit mit einer Sicherheit von 95%?
2. Bei der Produktion von Displays für die neuen Smartphones passieren teilweise technische Fehler.
Zur Überprüfung wird eine Stichprobe von 500 Stück entnommen. 10 Stück hiervon sind fehlerhaft. In welchem Bereich liegt die Fehlerrate mit einer Sicherheit von 90%?

11.8 Lösungen

1. Bekanntheitsgrad

$$KI_{1-0,05} = \left[0,8 - z_{0,975} \sqrt{\frac{0,8 \cdot (1-0,8)}{200}}; 0,8 + z_{0,975} \sqrt{\frac{0,8 \cdot (1-0,8)}{200}} \right]$$

$$= [0,8 - 1,96 \cdot 0,028; 0,8 + 1,96 \cdot 0,028] \approx [0,745; 0,855]$$

2. Fehlerfreiheit

$$KI_{1-0,1} = \left[0,02 - z_{0,95} \sqrt{\frac{0,02 \cdot (1-0,02)}{500}}; 0,02 + z_{0,95} \sqrt{\frac{0,02 \cdot (1-0,02)}{500}} \right]$$

$$= [0,02 - 1,64 \cdot 0,006; 0,02 + 1,64 \cdot 0,006] \approx [0,0101; 0,0298] \approx [0,01; 0,03]$$

12 Testen von Hypothesen

12.1 Test auf Erwartungswert

„Erwachsene Männer sind im Schnitt 1,83 m groß.“

So ungefähr könnte die Aussage in einer Zeitung stehen, am besten noch in der Form „Das Institut XY hat durch eine statistische Untersuchung ermittelt, dass erwachsene Männer im Schnitt 1,83 groß sind.“

OK, aber stimmt das auch?

Wie wahrscheinlich ist es, dass die mittlere Größe wirklich 1,83 m ist?

Anders gefragt – wir sind ja Experten: Kann ich mit einer Irrtumswahrscheinlichkeit von 5% ($\alpha = 0,05$) behaupten, dass Männer 1,83 m groß sind?

Wir hatten schon Umfrageergebnisse beim Konfidenzintervall: $\bar{x} = 1,86, \sigma = 0,05, n = 10$.

Wir prüfen nun im Prinzip, ob der angegebene Wert von 1,83 m im Konfidenzintervall liegt.

Das wäre ein mögliches Vorgehen.

Zuerst legen wir einmal fest, was gesagt wurde:

Die Hypothese (Annahme), die wir prüfen, wird als „Nullhypothese“ oder kurz H_0 bezeichnet.

Hier gilt: $H_0 : \mu = 1,83$

Die Gegenannahme ist die „Alternativhypothese“ und wird mit H_1 bezeichnet.

Hier gilt: $H_1 : \mu \neq 1,83$.

(Hängt davon ab, was andere eben behaupten.)

1. zweiseitiger Test

$$H_1 : \mu \neq 1,83$$

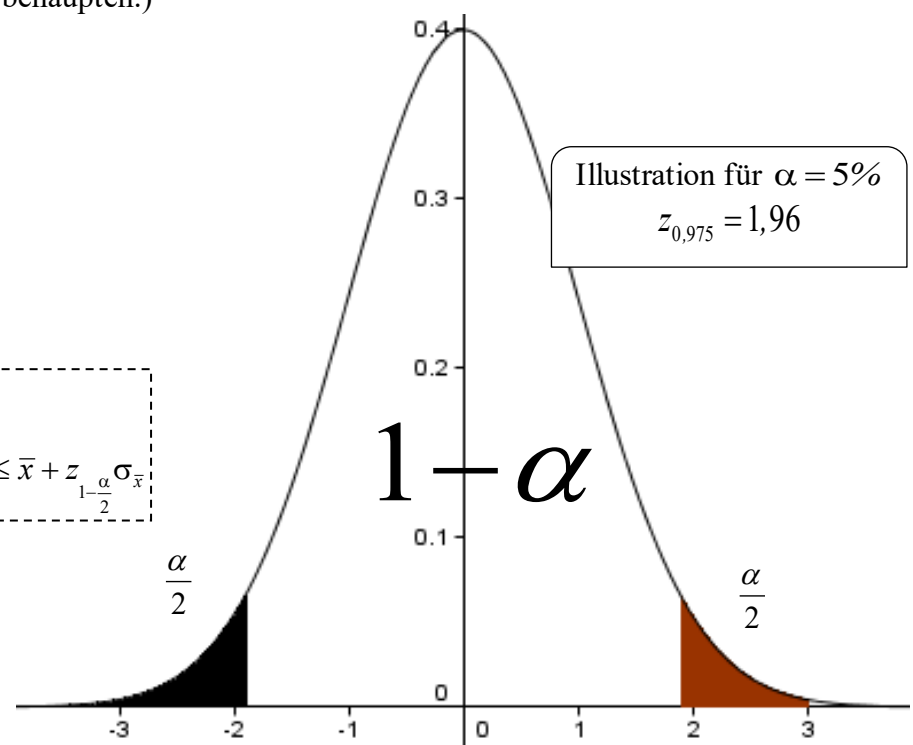
Wir wissen, dass wir einen beidseitigen Bereich um den Erwartungswert suchen (von... bis):

$$\mu - z\sigma_{\bar{x}} \leq \bar{x} \leq \mu + z\sigma_{\bar{x}}$$

$$\begin{aligned} \bar{x} \leq \mu + z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}} &\Rightarrow \mu \geq \bar{x} - z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}} \\ \bar{x} \geq \mu + z_{\frac{\alpha}{2}}\sigma_{\bar{x}} &\Rightarrow \mu \leq \bar{x} - z_{\frac{\alpha}{2}}\sigma_{\bar{x}} \Rightarrow \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}} \end{aligned}$$

Also ist der Annahmebereich

$$\mu_0 \in \left[\bar{x} \pm z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}} \right] = \mu_0 \in \left[\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$



Dies ist der größtmögliche Bereich, in dem sich der Erwartungswert auf Basis der Daten bei der vorgegebenen Irrtumswahrscheinlichkeit α befinden kann.

μ_0 ist der Wert aus der Hypothese, hier 1,83.

Also, zurück zum Fall: $H_0 : \mu = 1,83$

Ein anderes Institut behauptet, die Menschen wären im Schnitt größer oder kleiner als 1,83, also

$H_1 : \mu \neq 1,83$.

Wir hatten schon gegeben, dass $\bar{x} = 1,85, \sigma = 0,05, n = 10$.

Stimmt die Nullhypothese auf einem Niveau von $\alpha = 5\%$?

Lösung: Wir nehmen die Nullhypothese an, falls:

$$\mu_0 \in \left[\bar{x} \pm z_{1-\frac{0,05}{2}} \sigma_{\bar{x}} \right] \Rightarrow \mu_0 \in \left[1,86 \pm 1,96 \cdot \frac{0,05}{\sqrt{10}} \right] = [1,86 \pm 0,031] = [1,829; 1,891]$$

$\mu_0 = 1,83 \in [1,829; 1,891]$, also wird die Nullhypothese angenommen.

Erwachsene Männer sind also im Mittel 1,83 m groß.

2. einseitiger Test:

$H_0 : \mu \geq 1,83, H_1 : \mu < 1,83$ Annahmebereich: $z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \geq -z_{1-\alpha} \Rightarrow \bar{x} - \mu_0 \geq -z_{1-\alpha} \sigma_{\bar{x}}$ $\Rightarrow \mu_0 \leq \bar{x} + z_{1-\alpha} \sigma_{\bar{x}}$ $\Rightarrow \mu_0 \in]-\infty; \bar{x} + z_{1-\alpha} \sigma_{\bar{x}}]$ $z_{1-0,05} = z_{0,95} = 1,645, \bar{x} = 1,86, \sigma_{\bar{x}} = \frac{0,05}{\sqrt{10}} \approx 0,016$ $\Rightarrow \mu_0 \in]-\infty; 1,86 + 1,645 \cdot 0,016]$ $\Rightarrow \mu_0 \in]-\infty; 1,86 + 1,645 \cdot 0,016]$ $\Rightarrow 1,83 \in]-\infty; 1,886]$ $H_0 \text{ annahmen!}$	$H_0 : \mu \leq 1,83, H_1 : \mu > 1,83$ Annahmebereich: $z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \leq z_{1-\alpha} \Rightarrow \bar{x} - \mu_0 \leq z_{1-\alpha} \sigma_{\bar{x}} \Rightarrow \mu_0 \geq \bar{x} - z_{1-\alpha} \sigma_{\bar{x}}$ $\Rightarrow \mu_0 \in [\bar{x} - z_{1-\alpha} \sigma_{\bar{x}}; +\infty[$ $z_{1-0,05} = z_{0,95} = 1,645, \bar{x} = 1,86, \sigma_{\bar{x}} = \frac{0,05}{\sqrt{10}} \approx 0,016$ $\Rightarrow \mu_0 \in [1,86 - 1,645 \cdot 0,016; +\infty[$ $\Rightarrow 1,83 \notin [1,834; +\infty[$ $H_0 \text{ ablehnen!}$
---	--

Testfehler

Wir können nun noch folgende Fälle unterscheiden:

		In Realität liegt vor	
		H_0	H_1
Entscheidung für	H_0	Richtig entschieden!	Fehler 2. Art
	H_1	Fehler 1. Art (Wahrscheinlichkeit des Fehlers ist $= \alpha$)	Richtig entschieden!

Der Fehler 1. Art ist also der schlimmere Fehler, weil die Nullhypothese (Normalzustand, wovon wir ausgehen) abgelehnt wird, obwohl sie richtig ist! Daran müssen wir denken, wenn wir die Höhe von α festlegen. Wenn Alpha zu groß wird, lehnen wir H_0 zu oft ab, wenn α zu klein ist, akzeptieren wir die Nullhypothese im Prinzip immer. Es hat sich ein gängiger Wert von 1%–10% für α eingebürgert.

Wichtig: In der Nullhypothese muss immer ein spezieller Wert auch enthalten sein, wir können also nicht $H_0 : \mu < 1,83$ testen, da wir dann nicht wüssten, welchen Wert wir mit dem Konfidenzintervall vergleichen sollen einsetzen sollen. (1,83 gehört ja nicht zur Nullhypothese!)

Je nach Lage des Ablehnungsbereichs spricht man von „links-“, „rechts-“, „beidseitigem“ Test.

12.2 Übungen

Übung 1

Chris behauptet, die mittlere Körpergröße der Bevölkerung sei max. 1,75 m. Sie meinen, die mittlere Körpergröße wäre deutlich höher.

In einer Stichprobe von 20 Personen ist die mittlere Körpergröße 1,80 m. Es ist bekannt, dass die Größe in der männlichen Bevölkerung dieses Alters um 0,1 m schwankt.

Stellen Sie die Hypothesen auf. Hat Chris auf einem Signifikanzniveau von

1.1 $\alpha = 1\%$ 1.2 $\alpha = 5\%$ 1.3 $\alpha = 10\%$ 1.4 $\alpha = 20\%$ 1.5 $\alpha = 50\%$ 1.6 $\alpha = 90\%$

Recht?

Übung 2

Ein Unternehmen verkauft Kugelschreiber und wirbt damit, dass sie „statistisch alle mindestens 5.000 Blätter beschreiben“.

Wir wissen, dass bei der Herstellung der Minen eine Standardabweichung von $\sigma = 100$ Blatt vorliegt und überprüfen die Behauptung anhand einer Stichprobe von 400 Stück, wobei sich eine durchschnittliche Blattzahl von 4.950 ergibt.

Können wir auf Basis der Daten der Behauptung des Unternehmens auf einem Signifikanzniveau von $\alpha = 5\%$ zustimmen?

Übung 3

Zurück zu unserem Beispiel der Nägel:

Ein Unternehmen stellt Nägel her und behauptet, dass diese im Mittel alle 10 cm lang sind.

Ein Prüfer behauptet, dass dies nicht der Fall ist.

In einer Stichprobe vom Umfang $n = 100$ ergibt sich eine mittlere Größe von 9,8 cm.

Aus Erfahrung wissen wir, dass die Standardabweichung in der Produktion bei $\sigma = 1$ cm liegt.

Überprüfen Sie die Behauptung des Unternehmens auf einem Signifikanzniveau α von

3.1 $\alpha = 5\%$.

3.2 $\alpha = 1\%$.

Übung 4 (Achtung: Verteilung beachten!)

Wir befragen eine Stichprobe von 5 Männern bzgl. ihrer Körpergröße.

Wir behaupten, dass Männer im Alter 18-30 Jahre im Mittel genau 1,80 m groß sind.

Die mittlere Größe in der Stichprobe beträgt 1,75 m, die Standardabweichung der Körpergröße in der Stichprobe ist 0,08. Das Signifikanzniveau beträgt 5%.

Achtung: Schauen Sie noch einmal im Kapitel zum Schätzen, wie vorgegangen wird, wenn die Standardabweichung auf Basis der Stichprobe geschätzt wird.

Wiederholungsfragen

1. Rekapitulieren Sie erneut: Was bedeutet das Signifikanzniveau, bspw. $\alpha = 1\%$?

2. Wiederholung: Erklären Sie, warum $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$.

12.3 Lösungen

Übung 1

Körpergröße

$$H_0 : \mu \leq 1,75, H_1 : \mu > 1,75$$

$$1.1 \alpha = 1\% \Rightarrow z_{0,99} = 2,33 \Rightarrow \text{Annahmebereich} : \left[1,80 - 2,33 \cdot \frac{0,1}{\sqrt{20}}; +\infty \right] = [1,80 - 0,052; +\infty[\\ \Rightarrow 1,75 \in [1,748; +\infty[\Rightarrow H_0 \text{ annehmen.}$$

$$1.2 \alpha = 5\% \Rightarrow z_{0,95} = 1,645 \Rightarrow \text{Annahmebereich} : \left[1,80 - 1,645 \cdot \frac{0,1}{\sqrt{20}}; +\infty \right] = [1,80 - 0,037; +\infty[\\ \Rightarrow 1,75 \notin [1,763; +\infty[\Rightarrow H_0 \text{ ablehnen.}$$

$$1.3 \alpha = 10\% \Rightarrow z_{0,9} = 1,28 \Rightarrow \text{Annahmebereich} : \left[1,80 - 1,28 \cdot \frac{0,1}{\sqrt{20}}; +\infty \right] = [1,80 - 0,029; +\infty[\\ \Rightarrow 1,75 \notin [1,771; +\infty[\Rightarrow H_0 \text{ ablehnen.}$$

$$1.4 \alpha = 20\% \Rightarrow z_{0,8} = 0,84 \Rightarrow \text{Annahmebereich} : \left[1,80 - 0,84 \cdot \frac{0,1}{\sqrt{20}}; +\infty \right] = [1,80 - 0,019; +\infty[\\ \Rightarrow 1,75 \notin [1,771; +\infty[\Rightarrow H_0 \text{ ablehnen.}$$

$$1.5 \alpha = 50\% \Rightarrow z_{0,5} = 0 \Rightarrow \text{Annahmebereich} : \left[1,80 - 0 \cdot \frac{0,1}{\sqrt{20}}; +\infty \right] = [1,80; +\infty[\\ \Rightarrow 1,75 \notin [1,80; +\infty[\Rightarrow H_0 \text{ ablehnen.}$$

$$1.6 \alpha = 90\%: \text{ Unsinn: Niemand prüft einen Test auf einem Niveau, dass er mit 90\% falsch ablehnt!}$$

Übung 2

Ein Unternehmen verkauft Kugelschreiber und wirbt damit, dass sie „statistisch alle mindestens 5.000 Blätter beschreiben“.

Wir wissen, dass bei der Herstellung der Minen eine Standardabweichung von $\sigma = 100$ Blatt vorliegt und überprüfen die Behauptung anhand einer Stichprobe von 400 Stück, wobei sich eine durchschnittliche Blattzahl von 4.950 ergibt.

Können wir auf Basis der Daten der Behauptung des Unternehmens auf einem Signifikanzniveau von $\alpha = 5\%$ zustimmen?

$$H_0 : \mu \geq 5.000, H_1 : \mu < 5.000$$

$$\text{Annahmebereich: } \left[-\infty; 4.950 + z_{0,95} \cdot \frac{100}{\sqrt{400}} \right], z_{0,95} = 1,645 \Rightarrow]-\infty; 4958,225]$$

$$5000 \notin]-\infty; 4958,22] \Rightarrow H_0 \text{ ablehnen}$$

Übung 3

Zurück zu unserem Beispiel der Nägel:

Ein Unternehmen stellt Nägel her und behauptet, dass diese im Mittel genau 10 cm lang seien.

Ein Prüfer behauptet, dass dies nicht der Fall ist.

In einer Stichprobe vom Umfang $n = 100$ ergibt sich eine mittlere Größe von 9,8 cm.

Aus Erfahrung wissen wir, dass die Standardabweichung in der Produktion bei $\sigma = 1$ liegt.

Überprüfen Sie die Behauptung des Unternehmens auf einem Signifikanzniveau von

3.1 $\alpha = 5\%$, 3.2 $\alpha = 1\%$.

$$H_0 : \mu = 10, H_1 : \mu \neq 10$$

$$\bar{x} = 9,8, n = 100, \sigma = 1$$

3.1 Annahmebereich:

$$\left[9,8 \pm z_{1-\frac{0,05}{2}} \cdot \frac{1}{\sqrt{100}} \right], z_{0,975} = 1,96 \Rightarrow [9,8 \pm 0,196] = [9,604; 9,996]$$

$$10 \notin [9,604; 9,996] \Rightarrow H_0 \text{ ablehnen}$$

3.2 Annahmebereich:

$$\left[9,8 \pm z_{1-\frac{0,01}{2}} \cdot \frac{1}{\sqrt{100}} \right], z_{0,995} = 2,575 \Rightarrow [9,8 \pm 0,2575] = [9,5425; 10,0575]$$

$$10 \in [9,5425; 10,0575] \Rightarrow H_0 \text{ annehmen}$$

Übung 4

Wir befragen eine Stichprobe von 5 Männern bzgl. ihrer Körpergröße.

Wir behaupten, dass Männer im Alter 18-30 Jahre im Mittel genau 1,80 m groß sind.

Die mittlere Größe in der Stichprobe beträgt 1,75 m, die Standardabweichung der Körpergröße in der Stichprobe ist 0,08. Das Signifikanzniveau α beträgt 5%.

$$H_0 : \mu = 1,80, H_1 : \mu \neq 1,80$$

$$\bar{x} = 1,75, n = 5, \hat{s} = 0,08$$

Annahmebereich:

$$\left[1,75 \pm t(5-1)_{0,975} \cdot \frac{0,08}{\sqrt{5}} \right], t(4)_{0,975} = 2,776 \Rightarrow [1,75 \pm 0,099] = [1,651; 1,849]$$

$$1,8 \in [1,651; 1,849] \Rightarrow H_0 \text{ annehmen}$$

Wiederholungsfragen

1. Rekapitulieren Sie erneut: Was bedeutet das Signifikanzniveau α ?

Das Signifikanzniveau α ist die Wahrscheinlichkeit, mit der wir die Nullhypothese ablehnen, obwohl sie richtig ist. Je größer α ist, desto häufiger wird abgelehnt, da wir uns außerhalb des normalen Bereichs befinden (s. Zeichnung auf Seite 1).

2. Erklären Sie, warum $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$.

Normalverteilung: am linken Ende ist die Fläche darunter genau so groß wie am rechten Ende, da die Verteilung achsensymmetrisch ist.

Somit ist auf der linken Seite von bspw. $z_{0,1}$ genau so viel Fläche wie auf der rechten Seite von $z_{0,9}$

Literaturverzeichnis

- Büchler, A. & Henn, H.-W., 2007. *Elementare Stochastik - Eine Einführung in die Mathematik der Daten und des Zufalls*. Berlin: Springer.
- Duller, C., 2013. *Einführung in die Statistik mit EXCEL und SPSS*. Berlin: Springer Gabler.
- Fahrmeier, L. et al., 2016. *Statistik: Der Weg zur Datenanalyse*. s.l.:Springer - Spektrum.
- Gigerenzer, G., 2002. *Das Einmaleins der Skepsis*. Berlin: s.n.
- Krämer, W., 2001. *Denkste!*. München: Piper Verlag.
- Krämer, W., 2011. *So lügt man mit Statistik*. München: Piper.
- Schlittgen, R., 2000. *Einführung in die Statistik: Analyse und Modellierung von Daten (Lehr- und Handbücher der Statisti*. s.l.:Oldenbourg.
- Sibbertsen, P. & Lehne, H., 2015. *Statistik: Einführung für Wirtschafts- und Sozialwissenschaftler*. Berlin: Springer.
- Stiefl, J., 2008. *Wirtschaftsstatistik*. s.l.:Oldenbourg.