1 EINLEITUNG

1.1 Historisches und Grundsätzliches

1. Suchen Sie Beispiele für „Statistiken“ (Tageszeitungen, Zeitschriften, Lehrbücher, ...), erläutern Sie die tabellarische und/oder graphische Darstellung und recherchieren Sie den Hintergrund bzw. den Zweck der Untersuchung. Interessieren Sie sich auch dafür, woher die Daten stammen bzw. auf welche Weise sie erhoben wurden.

   Lösung: Im folgenden einige Beispiele aus ganz verschiedenen Bereichen (zwei aus einer Tageszeitung, eines aus dem *Statistischen Jahrbuch Österreich 2006*, eines auf Basis von Daten aus dem Internet, und ein historisches Beispiel).

   (a) Die Analyse von *Glücksspielen* stand am Beginn der Wahrscheinlichkeitsrechnung. Die folgende Aufstellung (aus dem STANDARD) zeigt die Ergebnisse und Gewinne von Lotto, Joker, Toto und Torwette vom 6. August 2006. Während die ersten beiden reine Glücksspiele sind, kann man bei den zwei letzten durch sorgfältige Analyse der bisherigen Begegnungen der diversen Mannschaften die Gewinnchancen zumindest leicht erhöhen. Die Quantifizierung dieses „Expertenwissens“ ist allerdings schwierig und ein gutes Beispiel für *subjektive* Wahrscheinlichkeiten. Die Ergebnisse der beiden anderen Spiele hingegen stehen in gutem Einklang mit der *frequentistischen* (oder *objektivistischen*) Wahrscheinlichkeitsdefinition.

Die Quoten vom 06. 08. 2006

LOTTO 6 aus 45

1 Sechser zu	723.475,80
6 Fünfer mit Zusatzzahl zu je	22.967,40
134 Fünfer zu je	1.156,90
7.511 Vierer zu je	36,60
119.230 Dreier zu je	3,60

TOTO

3 Zwölfer zu je	16.613,00
62 Elfer zu je	247,10
768 Zehner zu je	19,90

JOKER

3 Joker zu je	78.700,90
6mal	7.700,–
90mal	770,–
1.040mal	77,–
10.671mal	7,–

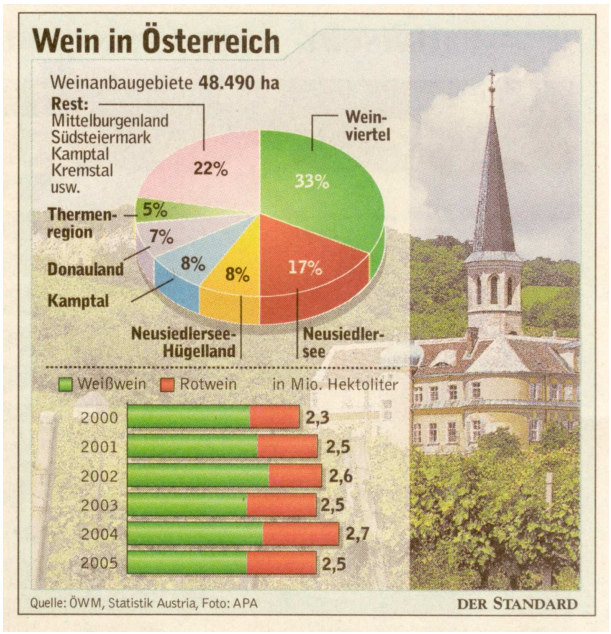
TORWETTE

1.Rang: keinmal	9fach-Jackpot
2.Rang: 21mal	157,00
Hattrick:	106.390,00

ALLE ANGABEN OHNE GEWÄHR

   (b) Ein Problem der europäischen *Weinwirtschaft* ist die Überproduktion, insbesondere von Wein niedriger Qualität („Tafelwein“). Dies hängt u.a. mit dem insgesamt sinkenden Verbrauch zusammen, aber auch damit, daß in Ländern mit Zuwächsen (Großbritannien, skandinavische Länder) Importe aus Australien und USA an Boden gewinnen. Maßnahmen zur Eindämmung der Überproduktion (u.a. die Stilllegung von Anbauflächen) werden seit Jahren diskutiert. Ö befindet sich dabei in einer vergleichsweise priviligierten Situation; dies ist eine Folge der in den letzten Jahren stark gestiegenen Qualität, aber auch der nur geringen Überschüsse.

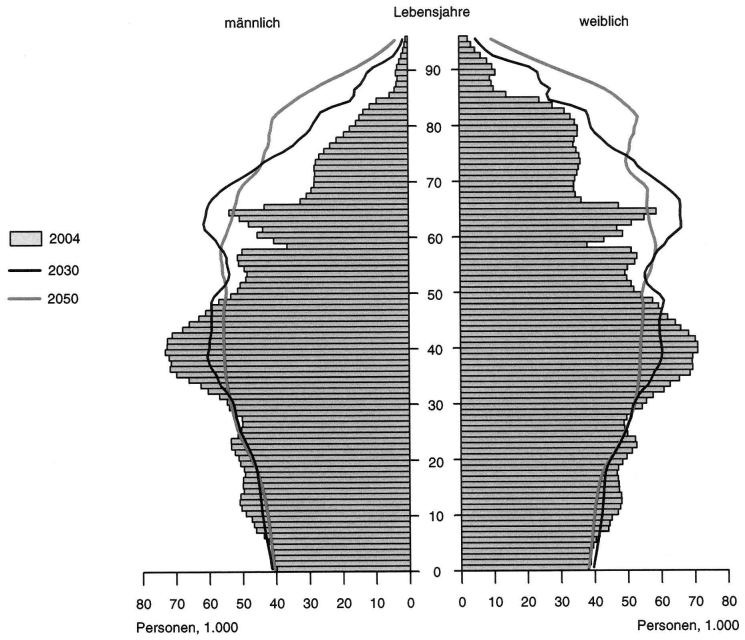
   Die folgende Graphik veranschaulicht den Statusquo in Ö; die angegebenen Werte (Anteile) sind dabei keine „genauen“ Zahlen sondern nur mehr oder weniger grobe Schätzungen. Diese „Fuzzyness“ beinträchtigt jedoch nicht den Zweck der Zusammenstellung. *Kreisdiagramme* („Tortendiagramme“) zur Veranschaulichung von *nominellen* Merkmalen (hier: Anbaugebiet) sind zwar sehr beliebt aber nicht ganz unkritisch; durch entsprechende Farbgebung, perspektivische Darstellung oder durch Herausziehen einzelner Sektoren läßt sich leicht eine manipulative Wirkung erzielen.



(c) Die *Bevölkerungspyramide* eines Landes zeigt die geschlechtsspezifische Altersverteilung der Bevölkerung. Die Darstellung (*Histogramm*) basiert im wesentlichen auf den Volkszählungen (in Ö seit 1869, etwa alle 10 Jahre, zuletzt 2001), wobei noch eine Reihe von Definitions– bzw. Abgrenzungsfragen zu klären sind. Wie man sieht, wird die Altersverteilung durch den Ausdruck „Pyramide“ nicht (mehr) adäquat beschrieben (eher durch den Ausdruck „spindelförmig“); dies betrifft inzwischen alle „(post)modernen“ Gesellschaften. Außerdem sieht man, daß die Zahl der Frauen in den höheren Altersgruppen viel größer als die der Männer ist.

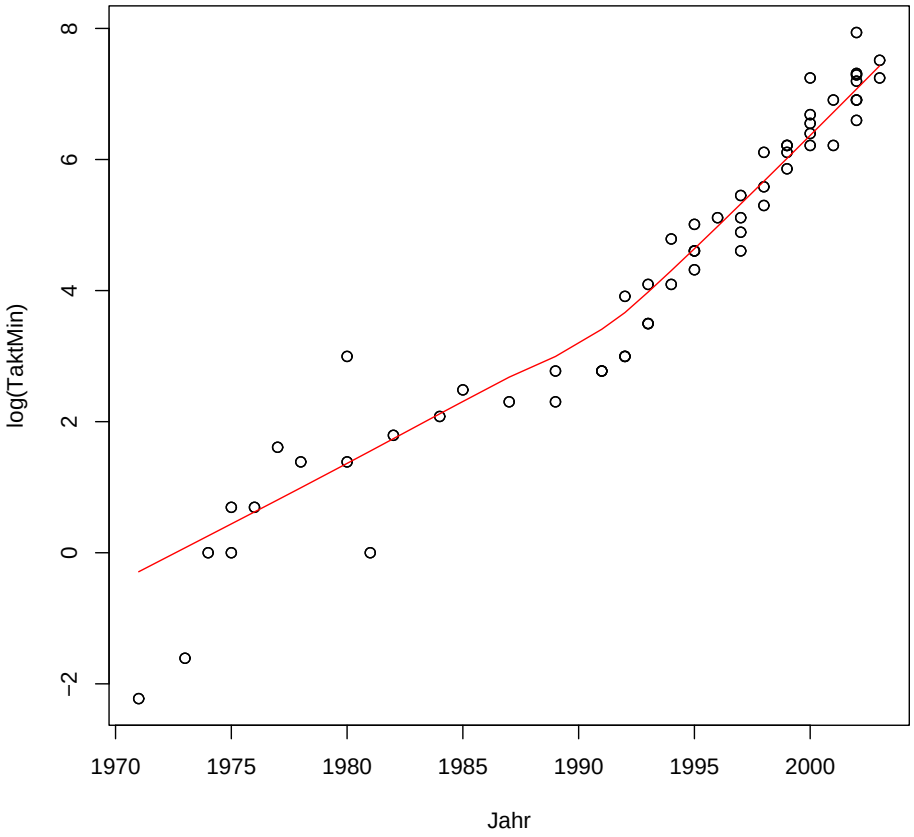
Die Prognosen für die Jahre 2030 und 2050 wiederum basieren auf bestimmten *Annahmen* (oder statistischen *Modellen*) über die Entwicklung der Bevölkerung; derartige Prognosen sind naturgemäß mit mehr oder weniger großen *Unsicherheiten* behaftet.

2.02 **Bevölkerungspyramide 2004, 2030 und 2050 (mittlere Variante)**
Population pyramids in 2004, 2030 and 2050



Q: STATISTIK AUSTRIA.

(d) In diesem Beispiel betrachten wir die Entwicklung der *Taktraten* von CPU's von 1971 bis 2003. Der Datensatz wurde aus Angaben auf `http://www.pc-erfahrung.de` zusammengestellt und enthält neben der (minimalen/maximalen) Taktrate auch Angaben zu Hersteller, Modell und Zahl der Transistoren. Ein Blick auf die (minimalen) Taktraten (1971: 108 KHz; 2002: 2.8 MHz) genügt, um zu erkennen, daß ein einfaches *Streudiagramm* von `TaktMin` gegen `Jahr` – außer der rapiden Entwicklung ab etwa 1990 – wenig zeigen wird; aussagekräftiger (insbesondere für Zwecke der *Prognose*) ist ein Plot von `log(TaktMin)` gegen `Jahr`. Letzterer wurde mit einer „glatten" Entwicklungsschätzung überlagert. Statistisch gesehen handelt es sich um eine *Zeitreihe*.

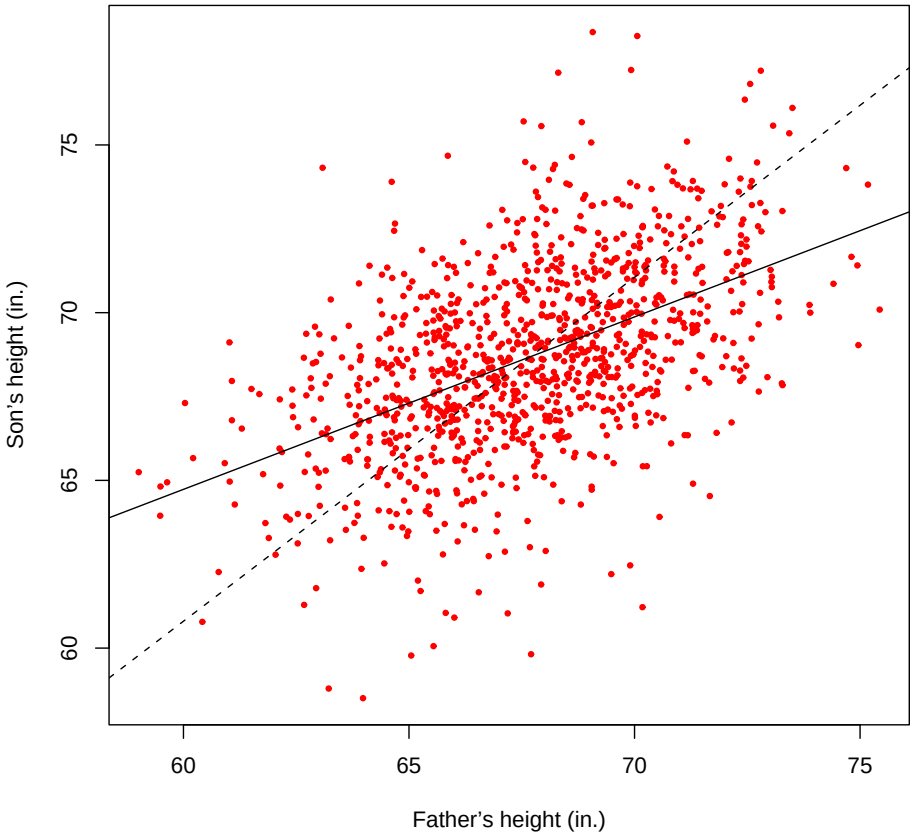


(e) Gegen Ende des 19. Jahrhunderts beschäftigte sich der britische Naturforscher Francis Galton (1822–1911) u.a. mit Fragen der Vererbung. Dabei entdeckte er ein Phänomen, das er mit dem Ausdruck *regression* bezeichnete:

> *Each pecularity in a man is shared by his kinsman but on the average in a less degree.*
> (F. Galton, *Natural Inheritance*, Macmillan, London, 1889.)

D.h., jede vom „Normalen" abweichende Eigenschaft eines Menschen wird – nach Galton's Ansicht – von der folgenden Generation zwar übernommen, aber im *Durchschnitt* in einem geringeren Ausmaß. Es kommt also bezüglich dieser Eigenschaft zu einem „Rückschritt".

Zur Untermauerung seiner Theorie(n) führte er zahlreiche Untersuchungen durch, u.a. auch zusammen mit seinem jüngeren Kollegen Karl Pearson (1857–1936). Letzterer untersuchte die Körpergrößen von Vater und Sohn an 1078 Familien und publizierte seine Ergebnisse im Jahre 1903. Die folgende Abbildung zeigt das Streudiagramm der Daten, zusammen mit der „Regressionsgeraden" (Die strichlierte Gerade ist eine Art Referenz, sie verläuft durch den Mittelpunkt der Daten und hat den Anstieg Streuung[Sohn]/Streuung[Vater].) Statistisch gesehen handelt es sich bei den Untersuchungen von Galton und Pearson um frühe Beispiele einer *Regressions–* bzw. *Korrelationsanalyse*.



1.2 Beschreibende Statistik

1.2.1 Diskrete Merkmale

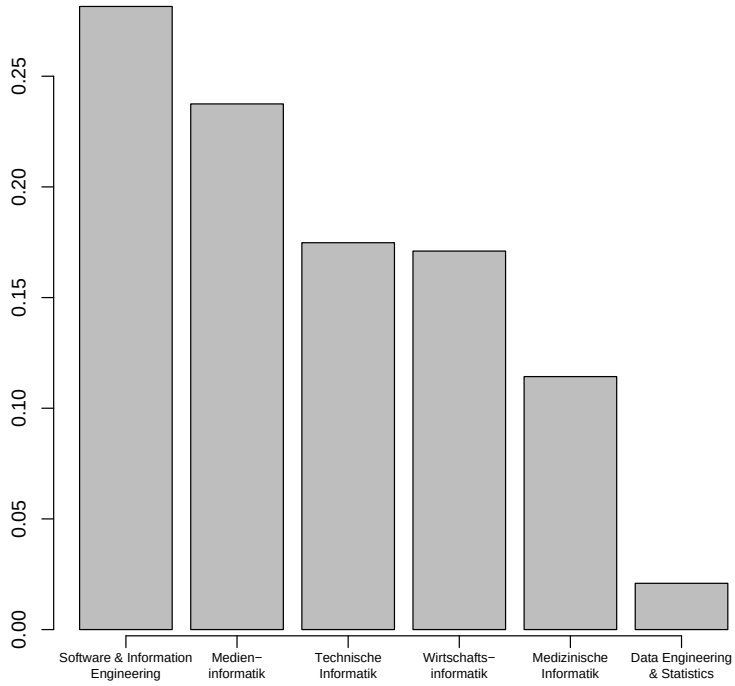
2. Im WS 2005 verteilten sich die Inskriptionszahlen der Studienrichtungen, für die die VO/UE *Statistik u. Wahrscheinlichkeitstheorie* anrechenbar bzw. verpflichtend ist, wie unten angegeben. Erstellen Sie ein Balkendiagramm und ein Kreisdiagramm für die Gesamtzahlen.

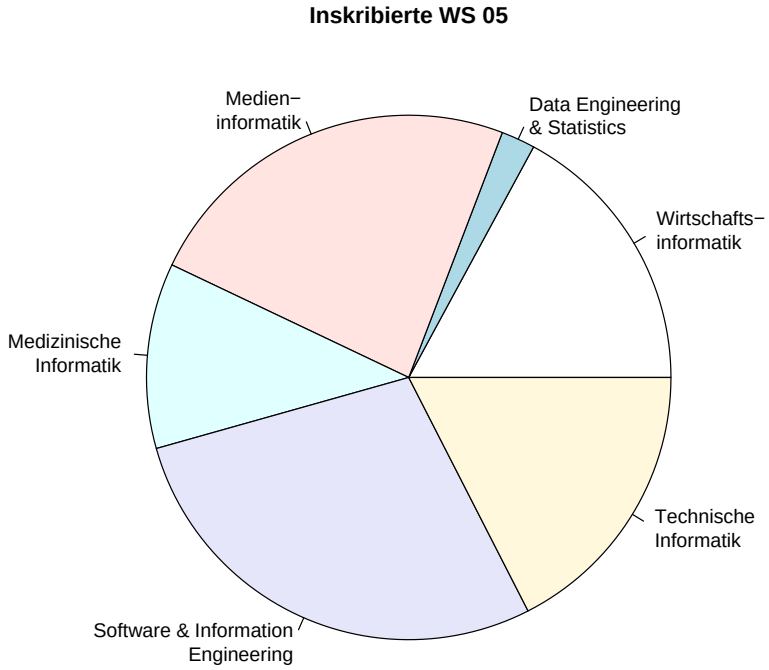
		Inländer		Ausländer		
KNR	Studienrichtung	Forts.	Anf.	Forts.	Anf.	Summe
526	Wirtschaftsinformatik	497	106	142	24	769
531	Data Egineering & Statistics	55	14	22	3	94
532	Medieninformatik	645	210	180	33	1068
533	Medizinische Informatik	285	104	101	24	514
534	Software & Information Engineering	702	167	336	61	1266
535	Technische Informatik	427	117	205	37	786

Quelle: TUWIEN

[inf05.r]

Lösung: Auch wenn die Studienrichtungen keine Rangordnung aufweisen (Artmerkmal), so ist es doch sinnvoll, die Balken nach den Inskriptionszahlen zu ordnen.





3. Betrachten Sie die Zahl der Inversionen einer zufälligen Permutation von $1, 2, \ldots, n$. Das Paar (i, j) ist eine *Inversion*, wenn $i < j$ aber j in der Permutation vor i liegt. Ist $n = 5$ und lautet die Permutation beispielsweise:

$$2 \quad 4 \quad 1 \quad 3 \quad 5$$

so gibt es drei Inversionen: $(1, 2)$, $(1, 4)$ und $(3, 4)$. Die folgende Tabelle enthält die Anzahl der Inversionen von 200 zufälligen (simulierten) Permutationen der Zahlen $1, 2, \ldots, 10$ (UE–Homepage: `inversions.dat`). Erstellen Sie eine Strichliste (Häufigkeitstabelle) und bereiten Sie den Datensatz mittels geeigneter Diagramme graphisch auf (Stabdiagramm, Summenkurve, etc.).

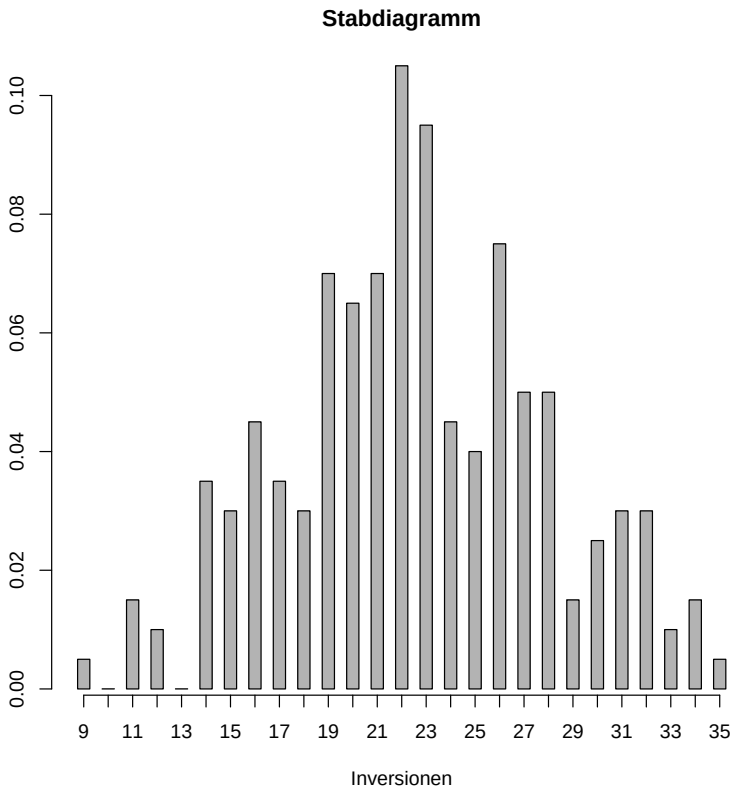
Inversionen

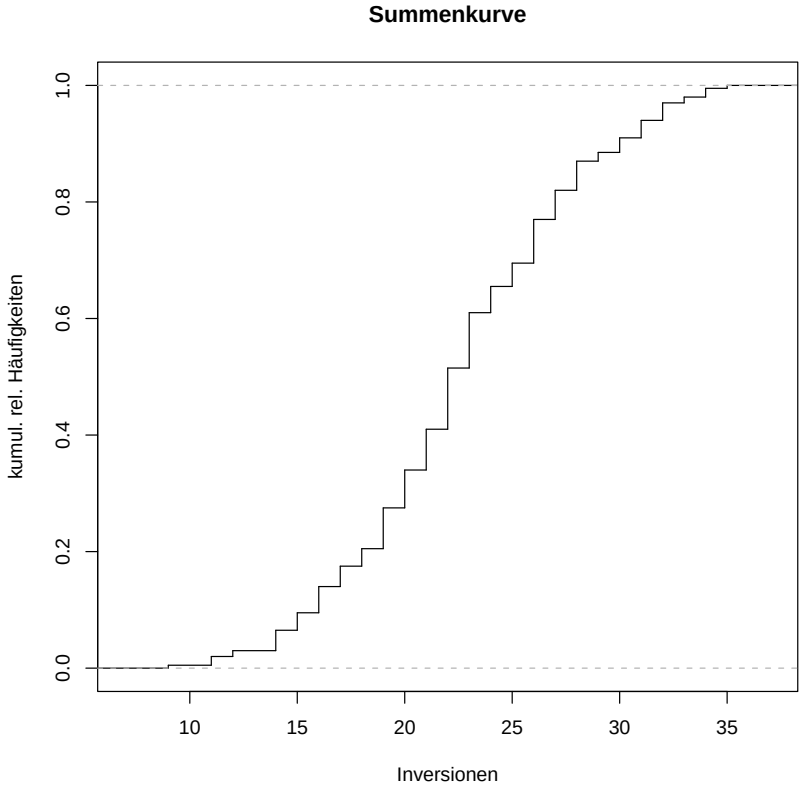
25	18	15	17	27	22	21	23	14	22	11	34	23	19	20	14	31	23	15	19
30	27	22	35	30	27	17	23	34	19	27	22	25	19	19	19	19	19	28	28
24	33	18	28	15	17	23	26	32	23	16	14	22	22	26	27	20	26	21	23
28	29	21	16	21	22	20	31	16	22	17	31	23	21	32	21	24	23	26	23
20	24	24	30	21	21	20	29	26	22	9	21	19	18	25	26	14	32	18	28
32	26	18	21	16	26	23	21	21	22	23	28	15	16	25	17	21	24	20	27
28	26	22	15	17	22	22	15	26	16	30	21	14	24	33	26	29	20	24	28
25	22	25	16	34	16	23	22	31	27	14	32	31	28	23	22	11	26	25	16
32	23	26	26	19	22	12	18	22	19	23	17	28	26	20	24	23	25	20	20
19	31	19	19	20	22	20	27	11	30	12	22	20	27	27	23	24	23	22	14

[`inversions.r`]

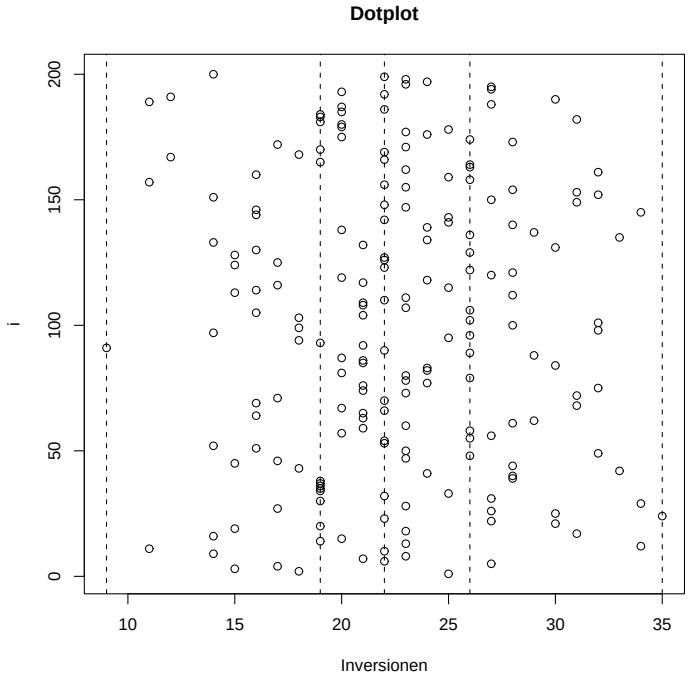
Lösung: Häufigkeitstabelle:

i	x.i	n.i	h.i	F.i
1	9	1	0.005	0.005
2	10	0	0.000	0.005
3	11	3	0.015	0.020
4	12	2	0.010	0.030
5	13	0	0.000	0.030
6	14	7	0.035	0.065
7	15	6	0.030	0.095
8	16	9	0.045	0.140
9	17	7	0.035	0.175
10	18	6	0.030	0.205
11	19	14	0.070	0.275
12	20	13	0.065	0.340
13	21	14	0.070	0.410
14	22	21	0.105	0.515
15	23	19	0.095	0.610
16	24	9	0.045	0.655
17	25	8	0.040	0.695
18	26	15	0.075	0.770
19	27	10	0.050	0.820
20	28	10	0.050	0.870
21	29	3	0.015	0.885
22	30	5	0.025	0.910
23	31	6	0.030	0.940
24	32	6	0.030	0.970
25	33	2	0.010	0.980
26	34	3	0.015	0.995
27	35	1	0.005	1


Stabdiagramm
0.00
0.02
0.04
0.06
0.08
0.10
9
11
13
15
17
19
21
23
25
27
29
31
33
35
Inversionen


Summenkurve
kumul. rel. Häufigkeiten
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
25
30
35
Inversionen

Über die Verteilung der Datenpunkte gibt auch ein einfacher „Dotplot“ Auskunft. Dabei werden die Datenpunkte x_i in der Reihenfolge ihrer Beobachtung gegen den Index i aufgetragen. Die vertikalen Linien entsprechen dem Minimum, 1. Quartil, Median (= 2. Quartil), 3. Quartil und dem Maximum.



4. Ist $x_1, \ldots, x_n$ eine Binärfolge, so heißt die Teilfolge $(x_{i+1}, \ldots, x_{i+k})$ ein *Lauf* der Länge k, wenn:
(1) $i = 0$ oder $x_i \neq x_{i+1}$
(2) $x_{i+1} = \cdots = x_{i+k}$
(3) $i + k = n$ oder $x_{i+k} \neq x_{i+k+1}$
Ist beispielsweise $n = 10$ und lautet die Binärfolge:

0 0 1 0 0 0 1 1 0 0

so gibt es 5 Läufe mit den Längen 2, 1, 3, 2, 2. Die folgende Tabelle enthält für 250 (simulierte) Binärfolgen der Länge 60 (mit gleicher Wahrscheinlichkeit für 0 und 1) die Zahl der Läufe (UE–Homepage: `numberruns.dat`). Bilden Sie eine Häufigkeitstabelle und erstellen Sie ein Stabdiagramm und die Summenkurve.

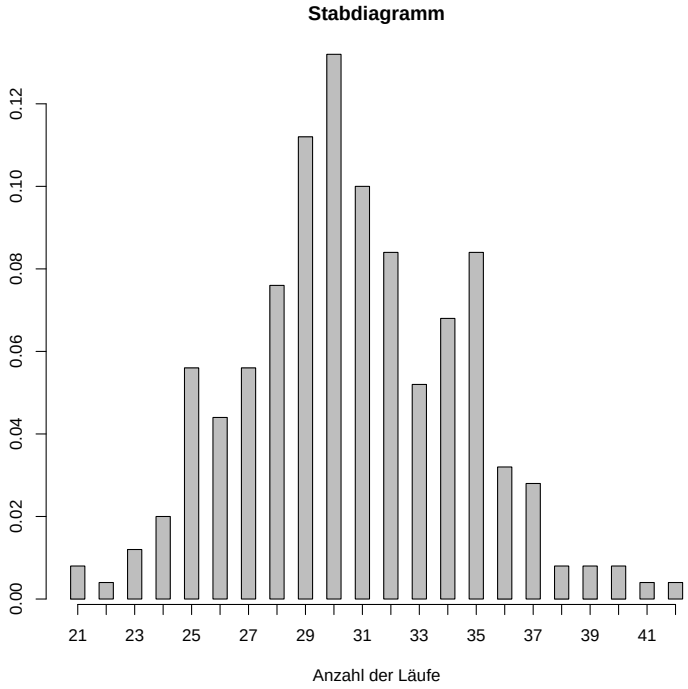
Zahl der Läufe

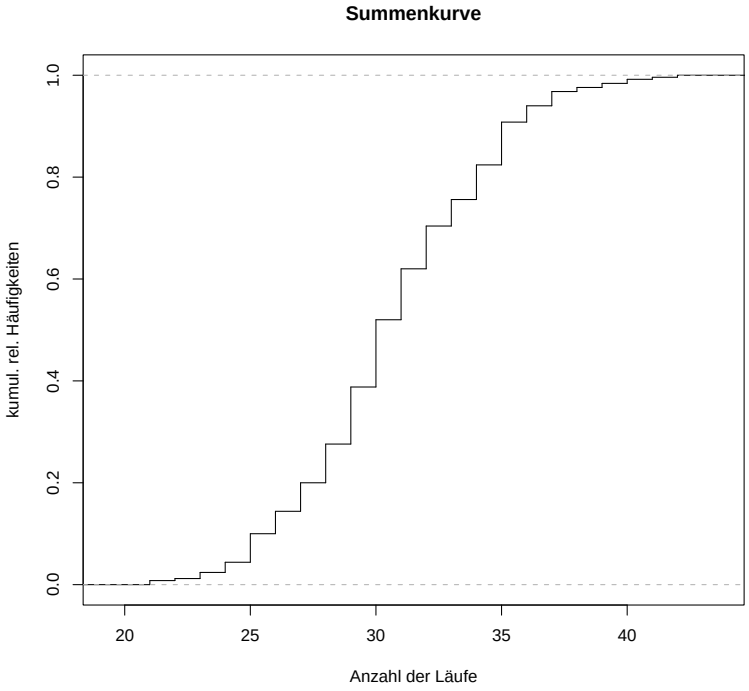
```
------------------------------------------------------------
28 34 42 23 27 37 31 31 32 29 33 31 34 35 37 28 30 24 38 30
32 34 35 22 37 37 25 31 30 28 38 31 28 31 31 26 21 31 35 30
29 26 28 24 30 35 26 32 27 34 25 30 37 32 34 30 36 28 32 33
29 29 30 29 32 31 30 26 25 35 33 34 29 30 29 29 35 32 30 30
33 25 32 27 29 34 32 35 32 29 30 29 31 41 32 28 29 35 31 40
25 31 30 31 26 32 28 31 33 30 28 34 32 25 36 24 29 33 32 29
33 29 29 33 25 28 29 27 27 40 30 30 30 37 34 35 29 30 29 30
35 29 35 31 29 31 27 28 35 32 35 27 32 25 33 25 28 26 31 35
29 32 36 27 36 31 24 33 37 26 25 28 30 33 36 30 30 30 31 34
30 36 33 27 35 31 34 29 39 31 23 25 39 28 31 27 26 34 26 29
30 27 32 34 34 31 36 35 31 30 33 34 32 30 34 28 35 29 28 32
27 30 35 28 29 28 31 27 23 32 30 30 36 27 25 25 25 24 30 35
35 35 29 29 28 21 26 26 30 34
```

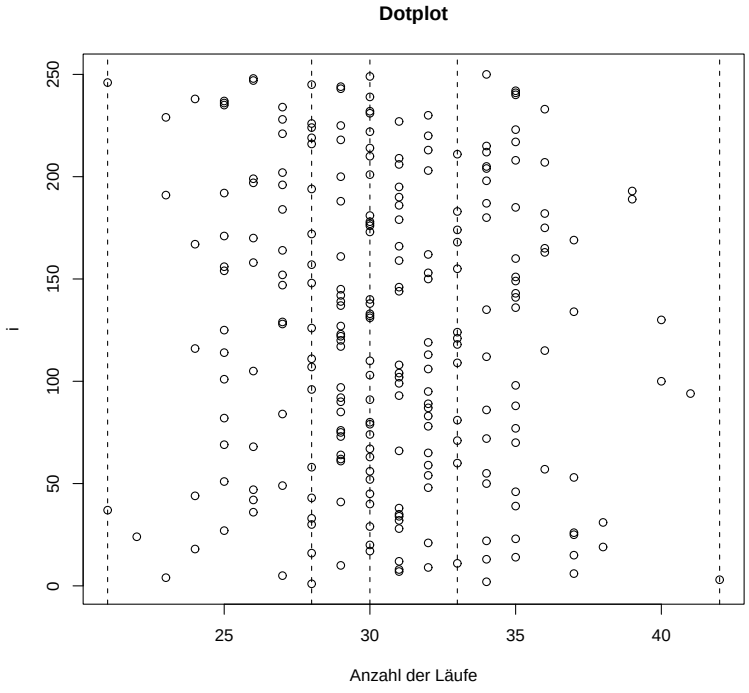
[`runs.r`]

Lösung: Häufigkeitstabelle:

i	x.i	n.i	h.i	F.i
1	21	2	0.008	0.008
2	22	1	0.004	0.012
3	23	3	0.012	0.024
4	24	5	0.020	0.044
5	25	14	0.056	0.100
6	26	11	0.044	0.144
7	27	14	0.056	0.200
8	28	19	0.076	0.276
9	29	28	0.112	0.388
10	30	33	0.132	0.520
11	31	25	0.100	0.620
12	32	21	0.084	0.704
13	33	13	0.052	0.756
14	34	17	0.068	0.824
15	35	21	0.084	0.908
16	36	8	0.032	0.940
17	37	7	0.028	0.968
18	38	2	0.008	0.976
19	39	2	0.008	0.984
20	40	2	0.008	0.992
21	41	1	0.004	0.996
22	42	1	0.004	1.000







Wie die Abbildungen zeigen, gruppiert sich die Verteilung mehr oder weniger symmetrisch um etwa 30.

5. [Fortsetzung des vorhergehenden Beispiels] Neben der Anzahl der Läufe wurde auch die Länge des *längsten* Laufs ermittelt (UE–Homepage: `maxruns.dat`). Bilden Sie eine Häufigkeitstabelle und erstellen Sie ein Stabdiagramm und die Summenkurve.

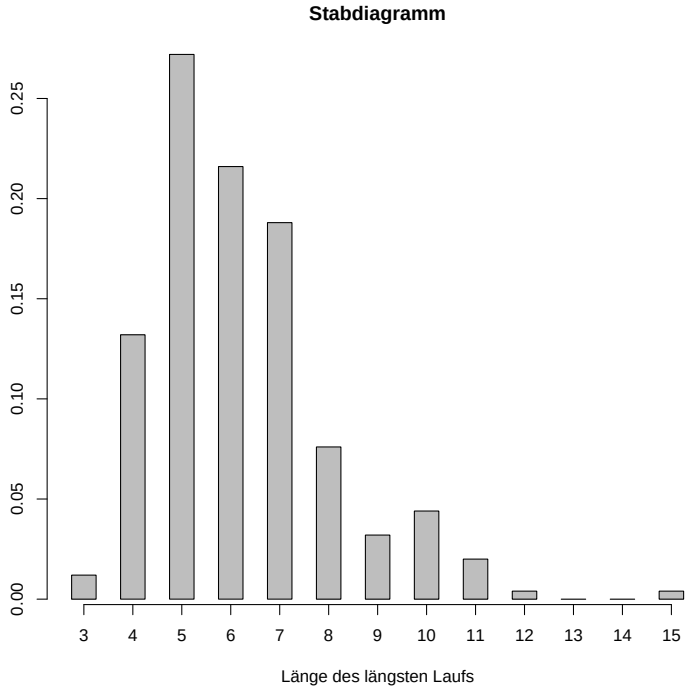
Länge des längsten Laufs

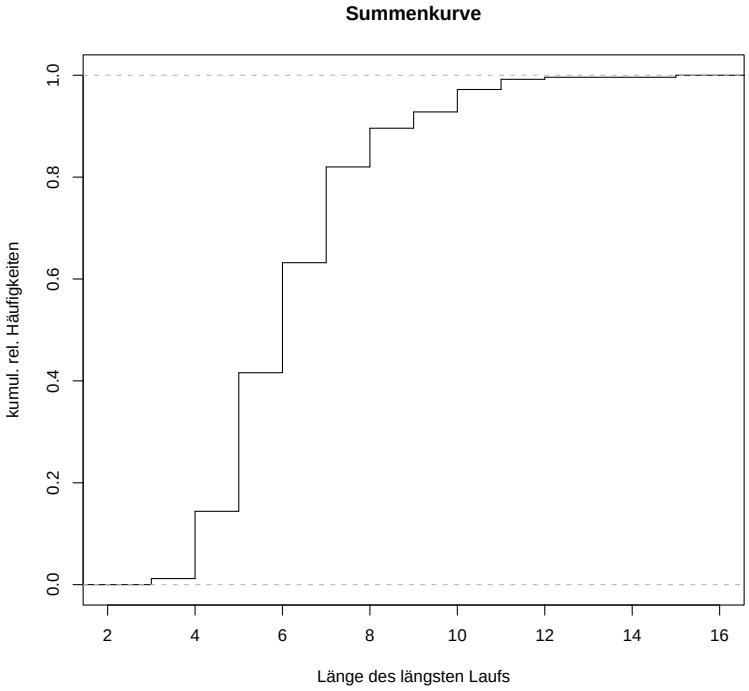
```
 6 10  5  6  7  5  6  6  5  6  6  7  6  4  5  5  5  9  4  6
 6  5  7 10  5  8 10  5  7  6  5  5  5  5 11  6  8  4  4  6
 7  8  6 10  6  4  7  5  6  7  7  6  5  6  6  5  4  8  5  6
 7  7  8  6  4  4  6  9  9  3  4  5  6  8  6  6  6  6  7  8
 7  6  5  8  7  4  6  6  5 10  8 10  5  3  6  7  7  5  7  5
 7  7  5  5  8  5  4  5  5  5  6  4  5  7  5  7  8  4  5  7
 7 15  5  6  8  6 10  6  6  4  5  6  6  5  7  4  7  4  6  5
 5  6  4  9  7  5  9  4  5  7  5 11  5  6  5  6 11  5 11  5
 6  5  5  7  5  7  7  8  3  5  8  5  7  4  4  7  6  5  4  5
10  7  5  5  6  5  4  7  4  4  7 10  4  5  7  7  8  5  5  7
 6  4  8  5  7  9  4  5  6  7  4  4  6  9  6  7  7  5  6  6
 9  7  4  8 12  6  8 11 10  5  4  5  5  6  8  5  7  7  6  7
 4  4  6  5  7 10  7  5  5  5
```

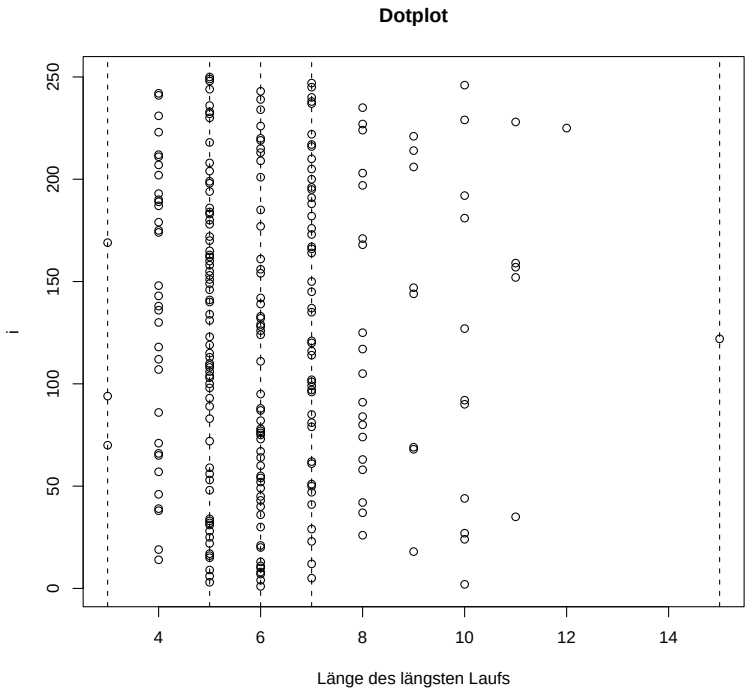
[runs.r]

Lösung: Häufigkeitstabelle:

```
 i x.i n.i   h.i   F.i
----------------------
 1   3   3 0.012 0.012
 2   4  33 0.132 0.144
 3   5  68 0.272 0.416
 4   6  54 0.216 0.632
 5   7  47 0.188 0.820
 6   8  19 0.076 0.896
 7   9   8 0.032 0.928
 8  10  11 0.044 0.972
 9  11   5 0.020 0.992
10  12   1 0.004 0.996
11  13   0 0.000 0.996
12  14   0 0.000 0.996
13  15   1 0.004 1.000
```







Wie die Abbildungen zeigen, ist die Verteilung eindeutig schief; letzteres ist notgedrungen bei Verteilungen eines auf irgendeine Weise „extremen“ Wertes (hier: längster Lauf) der Fall.

Bem.: Die Ergebnisse dieses und des vorhergehenden Beispiels sind insofern bemerkenswert, als man bei zufälligen Binärfolgen (Münzwurf) gefühlsmäßig dazu tendiert, die Zahl der Läufe zu überschätzen und – damit zusammenhängend – die Häufigkeit des Auftretens von langen Läufen zu unterschätzen.

*6. Eine Möglichkeit zur automatischen Erkennung, in welcher Sprache ein Text abgefaßt ist, ist der Vergleich der (relativen) Buchstabenhäufigkeiten des Textes mit den bekannten (relativen) Häufigkeiten der jeweiligen Sprache. Lesen Sie als Beispiel einen längeren deutschen Text ein, verwandeln Sie Groß– in Kleinbuchstaben, separieren Sie den Text in die einzelnen Zeichen (Buchstaben, Satzzeichen, Sonderzeichen, etc.) und ermitteln Sie die relativen Häufigkeiten der Buchstaben des Alphabets (inklusive Umlaute und Zwischenraum; ß=ss). Stellen Sie die ermittelten Häufigkeiten den für die deutsche Sprache charakteristischen Häufigkeiten gegenüber. Die Angaben bezüglich letzterer schwanken etwas; *eine* Liste finden Sie unter `lettersd.dat` auf der UE–Homepage.

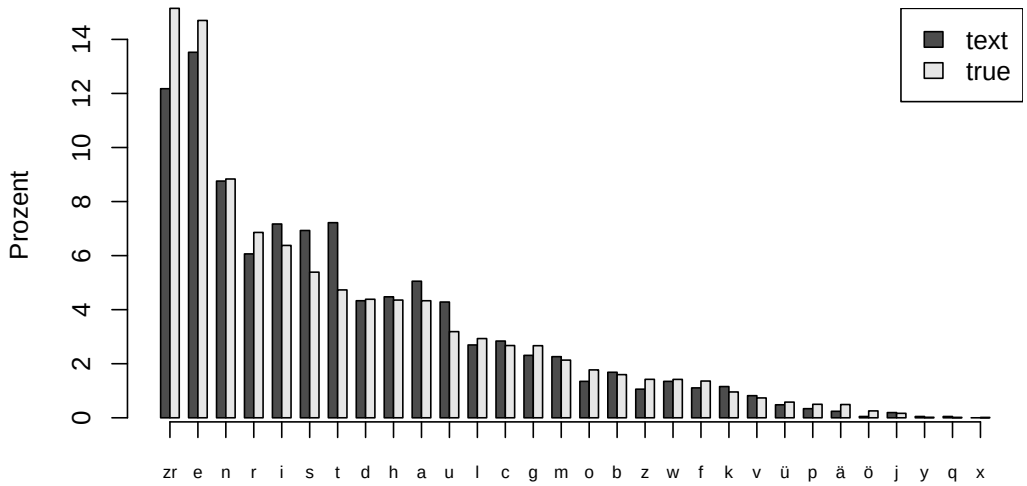
[`lettersd.r`]

Lösung: Als Beispieltext wurde ein Ausschnitt aus einem bekannten Lehrbuch – R. Storm: *Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualtitätskontrolle*, 10. Auflage, 1995 – gewählt:

> *Die mathematische Statistik hat die Aufgabe, Massenerscheinungen in der Natur, Wirtschaft und Technik mit den Mitteln der Wahrscheinlichkeitsrechnung zu untersuchen und wissenschaftlich zu beurteilen. Ihr Anwendungsgebiet reicht heute von der Medizin über die Biologie, Landwirtschaftswissenschaft, Physik, Wirtschaftswissenschaft und Technik bis zur Industrie, wo vor allem die statistische Qualitätskontrolle mit speziellen statistischen Verfahren zur Überwachung und Verbesserung der Erzeugnisse beiträgt. Überall da, wo man Versuchsergebnisse auszuwerten hat, sind die Methoden der mathematischen Statistik zu einem unentbehrlichen Hilfsmittel geworden.*
>
> *Der Ursprung der mathematischen Statistik liegt im Sammeln und Zusammenstellen von Daten und in der grafischen Darstellung der gewonnenen Messergebnisse. So entstanden die ersten Geburts-, Heirats-, und Sterberegister in der Bevölkerungsstatistik. Schon früh war man bemüht, das umfangreiche und damit unübersichtliche Untersuchungsmaterial zu ordnen und auf einzelne charakteristische Werte zu reduzieren. Das führte zu Tabellen, grafischen Darstellungen und zur Bildung von Summen, Prozenten, Mittelwerten und Streuungen. Diese Methoden der beschreibenden oder deskriptiven Statistik wurden später auch auf naturwissenschaftlichem und technischem Gebiet angewendet. Sie sind heute in Industrie und Wirtschaft weit verbreitet und finden ihren Niederschlag u.a. in den Statistischen Jahrbüchern.*
>
> *Mit Hilfe der beschreibenden Statistik erhält man aber immer nur Aussagen über das vorliegende konkrete Datenmaterial. Das wesentliche Anliegen bei den Untersuchungen besteht jedoch darin, ausgehend von den konkreten Daten allgemeingültige Aussagen zu erhalten.*
>
> *In den zwanziger Jahren unseres Jahrhunderts wurden vor allem in den USA und in England derartige mathematisch-statistische Verfahren zur wissenschaftlichen Beurteilung der Messergebnisse entwickelt, die auf den Gesetzmässigkeiten der Wahrscheinlichkeitsrechnung beruhen. Man spricht heute von beurteilender, schliessender oder induktiver Statistik, auch von der eigentlichen mathematischen Statistik.*

Wie die folgende Abbildung zeigt, stimmen die beobachteten Häufigkeiten insgesamt recht gut mit den theoretischen Häufigkeiten überein. Die Beurteilung, ob einzelne größere Abweichungen (etwa bei zr = Zwischenraum oder bei t und *s*; bei letzterem gibt es möglicherweise ein Problem mit dem eigentlichen s und ss=ß) „signifikant“ oder „zufälliger“ Natur sind, bleibt der schließenden Statistik überlassen (statistische Tests).

Bem.: Die Buchstaben wurden nach ihrer theoretischen Häufigkeit absteigend geordnet; dies erhöht die Übersichtlichkeit der Darstellung. Eine andere Möglichkeit wäre z.B. die Beschränkung auf die Vokale allein, oder auf die 10 häufigsten Buchstaben, etc.



1.2.2 Kontinuierliche Merkmale

7. Der Datensatz `resistor.dat` (UE–Homepage) enthält Messungen (in Ohm) von 80 Widerständen:

```
74.4 77.3 77.4 74.3 69.6 71.3 72.5 77.4 73.5 76.1
75.6 75.5 73.3 74.3 77.2 75.1 75.1 79.3 75.8 77.1
76.9 75.8 73.3 74.8 74.7 76.4 78.7 78.5 77.4 74.0
73.2 72.3 76.9 76.2 76.3 74.1 77.0 74.1 72.5 72.7
73.6 78.4 77.7 71.6 73.2 76.4 71.8 78.1 75.6 74.0
75.0 77.5 76.4 72.5 72.8 71.8 73.1 75.0 77.7 77.9
76.2 75.0 76.4 76.3 73.1 74.7 76.0 75.6 75.3 79.6
74.6 77.0 72.1 75.2 75.7 74.7 73.6 75.2 76.6 74.6
```

   (a) Ermitteln und zeichnen Sie die empirische Verteilungsfunktion.

   (b) Stellen Sie die Verteilung der relativen Klassenhäufigkeiten in Form eines Histogramms dar; nehmen Sie dazu beispielsweise die folgende äquidistante Klasseneinteilung:

$$[69.55,\ 70.65],\ (70.65,\ 71.75],\ \ldots,\ (79.45,\ 80.55]$$

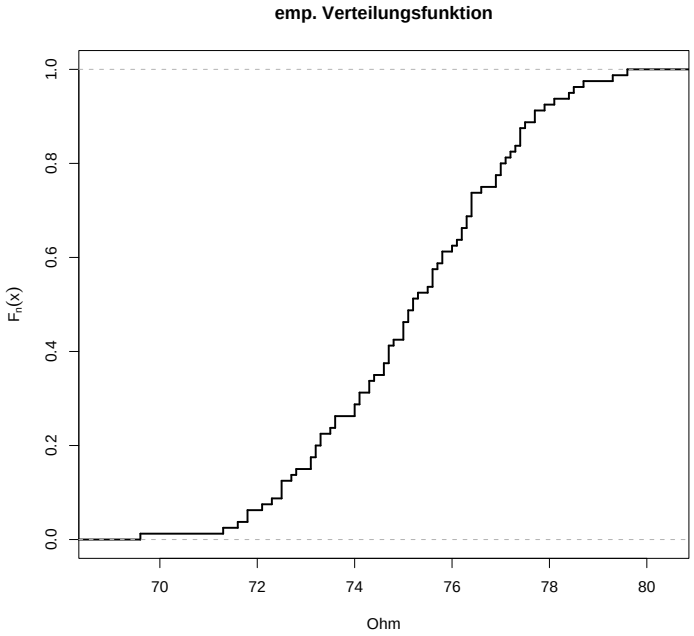
   Wählen Sie die Darstellung so, daß die Fläche unter dem Histogramm gleich 1 ist („Dichtehistogramm“ oder „flächentreues“ Histogramm genannt).

   (c) Zeichnen Sie auf Basis der obigen Klasseneinteilung das Summenpolygon.

[`resistor.r`]

Lösung:

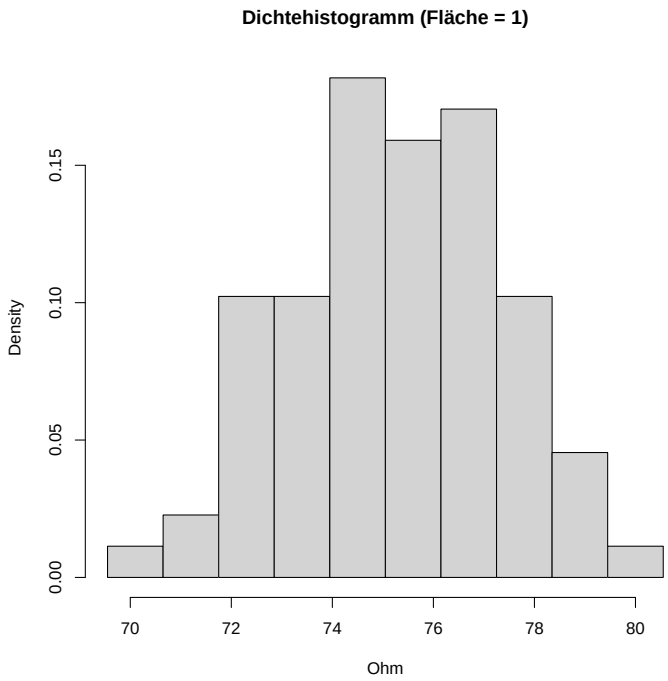
   (a) Die empirische Verteilungsfunktion $F_n^*(x)$ ist eine Treppenfunktion mit Sprüngen der Höhe $1/n$ an den Stellen der geordneten Stichprobe, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Sind mehrere Beobachtungen identisch, beträgt die Sprunghöhe das entsprechende Vielfache von $1/n$.



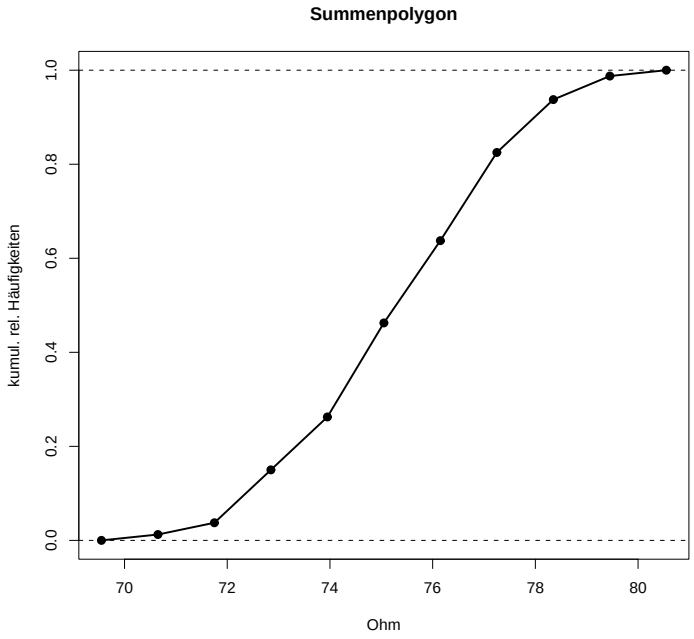
(b) Beim Histogramm der relativen Häufigkeiten beträgt die Höhe der Balken $h_n(K_j)$ (= relative Klassenhäufigkeit); beim Dichtehistogramm teilt man diesen Wert noch durch die Klassenbreite b_j. Auf diese Weise erreicht man, daß die Fläche unter dem Histogramm den Wert 1 hat (das Histogramm also eine – stufige – „Dichte“ repräsentiert):

$$\text{Fläche unter dem Histogramm} = \sum_{j=1}^{m} b_j \, \frac{h_n(K_j)}{b_j} = \sum_{j=1}^{m} h_n(K_j) = 1$$

Bem.: Haben alle Klassen (so wie hier) die gleiche Breite, unterscheidet sich das Histogramm der relativen Häufigkeiten hinsichtlich seiner Form nicht vom Dichtehistogramm. Bei unterschiedlichen Klassenbreiten allerdings ist – um einen falschen optischen Eindruck zu vermeiden – eine flächentreue Darstellung unumgänglich.



(c) Die kumulierten relativen Klassenhäufigkeiten werden mit einem Polygonzug verbunden:



8. Die folgende Tabelle enthält Angaben zu Geschlecht (GE), Körpergröße (GR) und Körpergewicht (GW) für Studierende einer medizinischen Universität (erweiterter Datensatz auf der UE–Homepage: `meddat.dat`):

	GE	GR	GW		GE	GR	GW		GE	GR	GW
1	M	178	70	24	M	170	66	47	W	170	54
2	M	178	75	25	M	173	68	48	W	170	60
3	M	179	60	26	M	174	74	49	W	175	65
4	M	179	71	27	M	177	81	50	W	176	65
5	M	187	75	28	M	178	72	51	W	161	56
6	M	178	67	29	M	183	72	52	W	168	54
7	M	179	76	30	M	190	83	53	W	170	57
8	M	181	73	31	M	178	66	54	W	172	56
9	M	186	70	32	M	180	72	55	W	174	64
10	M	186	75	33	M	183	69	56	W	176	60
11	M	188	77	34	M	198	90	57	W	178	65
12	M	189	80	35	M	179	70	58	W	164	60
13	M	189	89	36	M	180	79	59	W	165	52
14	M	190	75	37	M	187	82	60	W	173	62
15	M	190	80	38	M	192	80	61	W	165	58
16	M	192	96	39	M	182	98	62	W	177	60
17	M	170	68	40	W	158	46	63	W	160	48
18	M	175	75	41	W	166	63	64	W	163	65
19	M	178	64	42	W	168	63	65	W	167	72
20	M	180	67	43	W	170	57	66	W	168	56
21	M	184	88	44	W	165	50	67	W	169	59
22	M	186	73	45	W	166	49	68	W	162	57
23	M	192	94	46	W	168	56	69	W	164	65

Betrachten Sie das Merkmal „Körpergröße“:

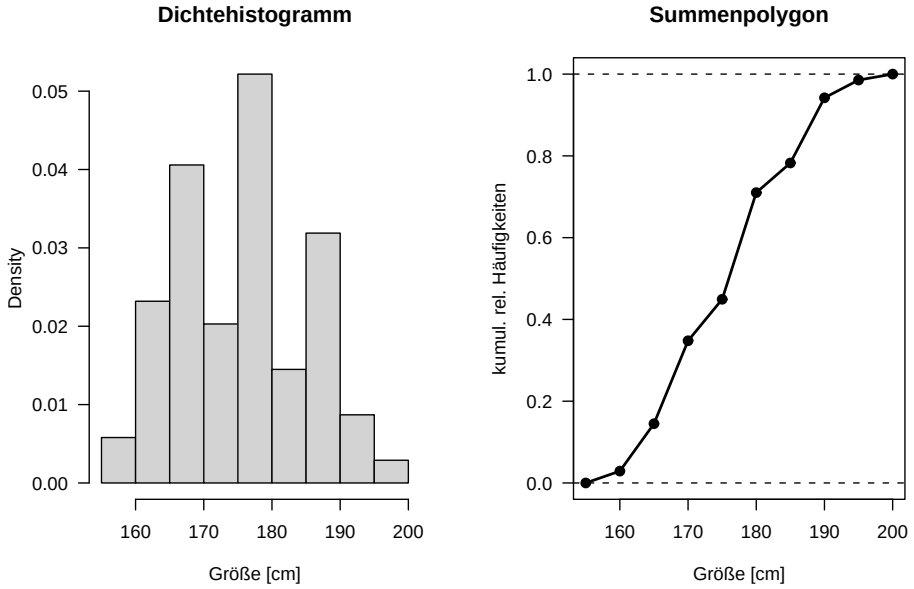
(a) Stellen Sie die Verteilung in Form eines (Dichte–) Histogramms dar. (Die Klasseneinteilung sollte dabei nicht zu fein aber auch nicht zu grob sein.)

(b) Zeichnen Sie auf Basis der obigen Klasseneinteilung das Summenpolygon.

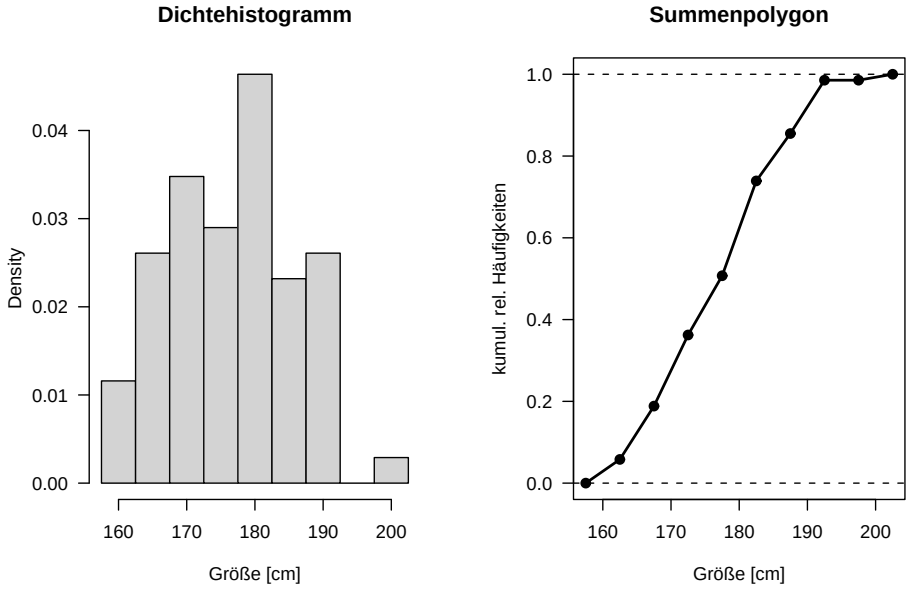
Zeichnen Sie Histogramm und Summenpolygon sowohl für alle Beobachtungseinheiten zusammen als auch getrennt nach Geschlecht.

[`meddat.r`]

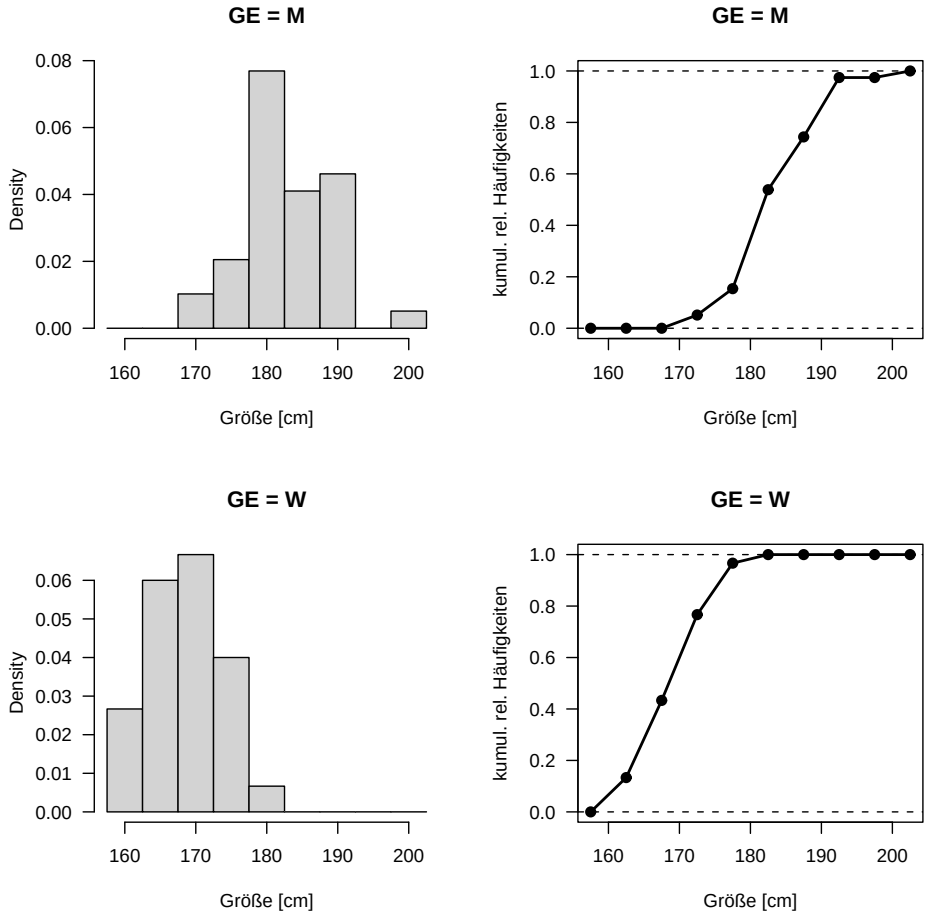
Lösung: Die Klasseneinteilung (Anzahl, Grenzen) beeinflußt die Form eines Histogramms. Um die nicht ganz unkritische Wahl der Klassen zu erleichtern, gibt es eine Reihe von Regeln, etwa die *Wurzel–Regel*, Anzahl $\approx \sqrt{n}$, oder die „klassische“ *Sturges–Regel*, Anzahl $= \lceil 1 + \log_2(n) \rceil$. Die automatische Anwendung derartiger Regeln führt allerdings nicht immer zu befriedigenden Ergebnissen. Die Sturges–Regel (in R der Default) führt zu folgendem Ergebnis:



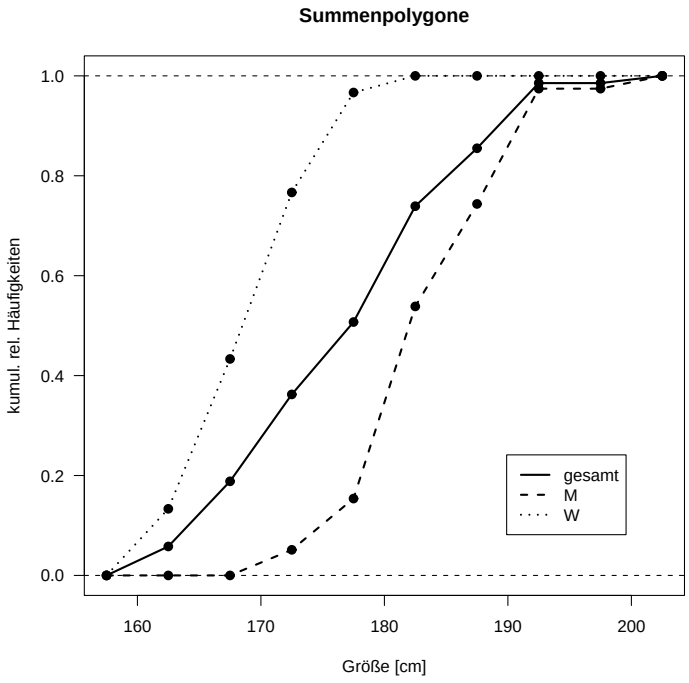
Die vielen „Peaks“ irritieren etwas; behält man Anzahl und Breite der Klassen bei, verschiebt aber den Anfangspunkt von 155[cm] auf 157.5[cm], ändert sich der Eindruck:



Im folgenden behalten wir die zuletzt verwendete Klasseneinteilung bei und schlüsseln die Daten nach Geschlecht auf.

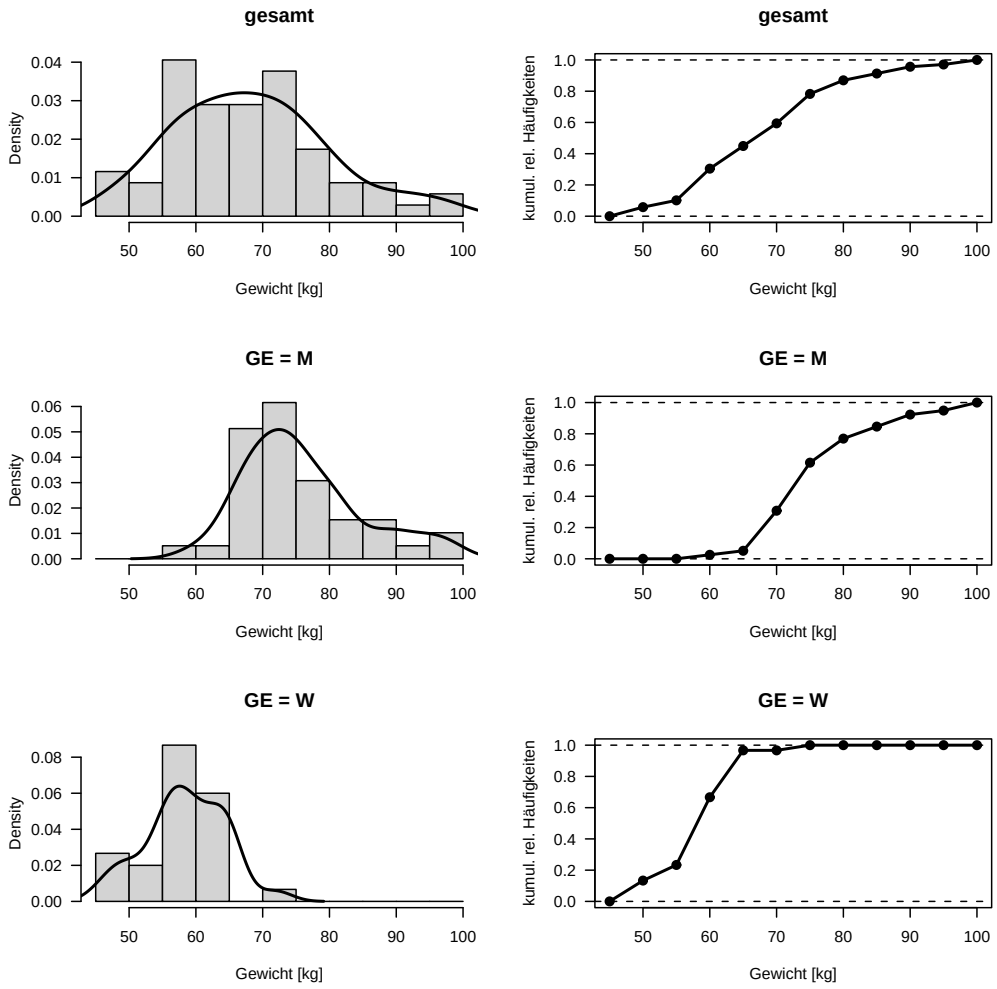


Direkter Vergleich der Summenpolygone:



9. [Fortsetzung des vorhergehenden Beispiels] Wiederholen Sie die Aufgabe für das Merkmal „Körpergewicht“. Zeichnen Sie Histogramm und Summenpolygon sowohl für alle Beobachtungseinheiten zusammen als auch getrennt nach Geschlecht.
[`meddat.r`]

Lösung:



Bem.: Die „Stufigkeit“ eines Histogramms – zumal für ein stetiges Merkmal – beeinträchtigt etwas den optischen Eindruck. Die über die Histogramme gezeichneten glatten Kurven wurden mittels *Kernschätzung* (in der VO nicht behandelt) ermittelt; die Fläche unter diesen Kurven ist – ebenso wie beim Dichtehistogramm – gleich Eins.

10. Ein wichtiges Qualitätskriterium von Wasser ist die Konzentration [ppm] von schwebenden festen Teilchen. Die folgende Tabelle enthält 40 Messungen dieser Größe für einen Badesee (UE–Homepage: `solid.dat`):

```
3.41 0.45 0.29 1.15 4.40 1.98 1.08 0.99 1.23 1.52
0.98 2.47 0.30 1.79 0.44 0.37 1.66 3.49 0.15 0.82
1.93 0.12 0.95 0.44 0.16 6.94 0.61 3.47 0.36 2.12
1.31 1.40 0.29 0.51 0.86 0.55 0.85 1.32 0.57 0.27
```

   (a) Ermitteln und zeichnen Sie ein Histogramm sowie die empirische Verteilungsfunktion.

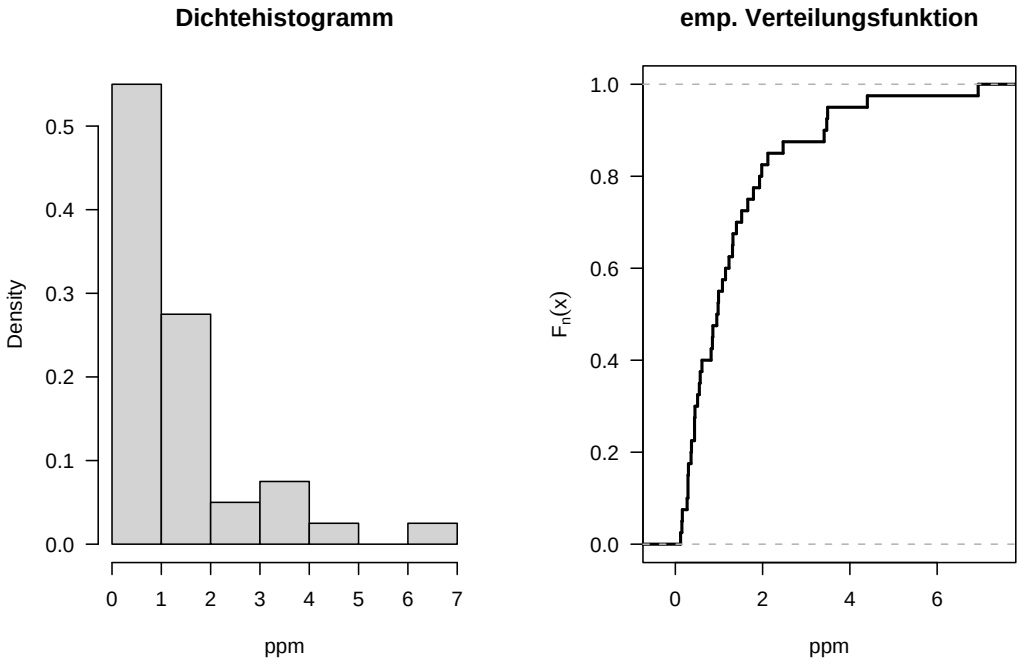
   *(b) Wie sich zeigt, ist die Verteilung stark (rechts–) schief. Mittels einer geeigneten Transformation kann man versuchen, die Verteilung zu symmetrisieren. Welche der folgenden Transformationen eignet sich dafür am besten?

$$x^* = \sqrt{x}, \quad x^* = \ln x, \quad x^* = 1/x$$

[`solid.r`]

Lösung:

(a) Die Verteilung ist stark asymmetrisch (rechtsschief, „J–förmig“):



(b) Zur Beurteilung der Schiefe einer Verteilung sind – neben dem Augenschein – eine Reihe von Maßzahlen gebräuchlich; der im folgenden verwendete γ–Koeffizient wird wie folgt berechnet (Vorgriff auf die folgenden Abschnitte):

$$\gamma = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \overline{x})^3 \quad \text{mit} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Positive/Negative Werte von γ deuten auf eine rechts–/linksschiefe Verteilung hin; für symmetrische Verteilungen ist γ annähernd gleich 0.

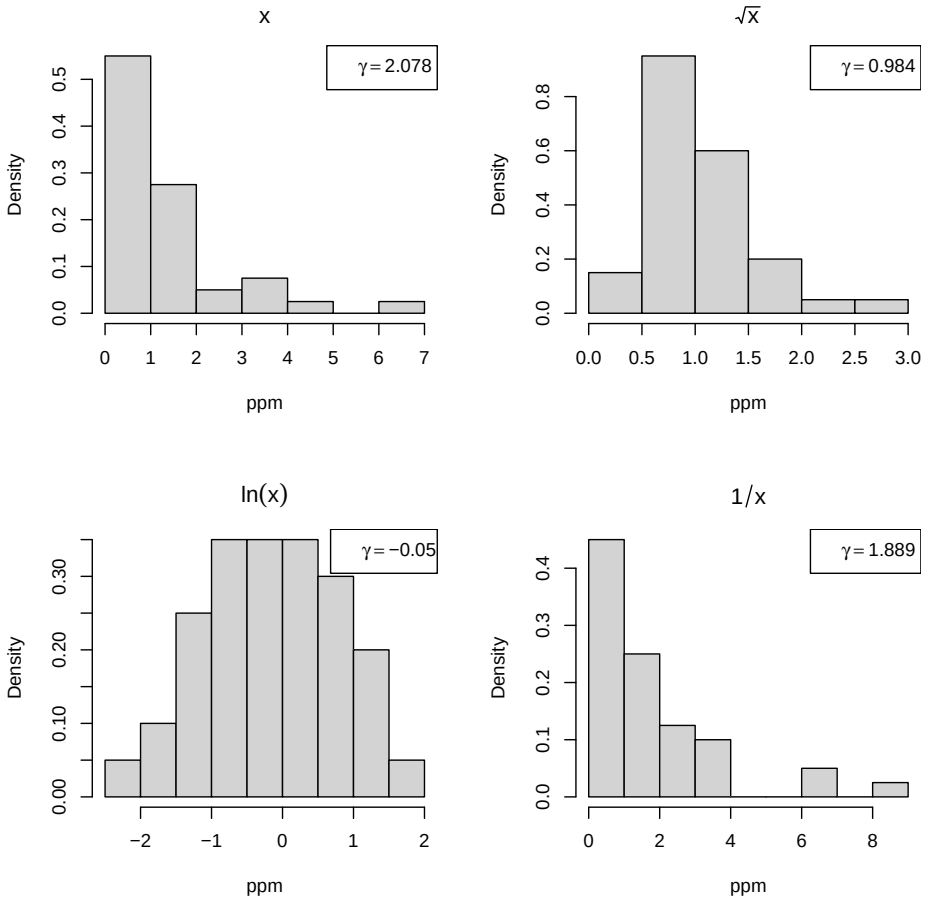
Wie die folgenden Abbildungen zeigen, wirkt die Wurzeltransformation zwar in die richtige Richtung, ist aber etwas zu „schwach“; die logarithmische Transformation führt zu einer nahezu symmetrischen Verteilung; die reziproke Transformation hingegen hat (überraschenderweise?) hier kaum eine Wirkung.

Bem.:

(1) Eine gewisse Schwierigkeit mit transformierten Daten ergibt sich daraus, daß sie häufig nicht mehr direkt interpretierbar sind. Abgesehen von „guten“ statistischen Eigenschaften wird man bei der Wahl einer Transformation auch darauf achten, daß die Daten nach der Transformation interpretierbar bleiben (ist beispielsweise x eine Länge, ist x^2 eine Fläche). Wann immer möglich, wird man aber versuchen, die ursprünglichen (gemessenen) Einheiten beizubehalten.

(2) Klarerweise muß man sich nicht auf die obigen drei Transformationen beschränken. Eine besonders wichtige Klasse in dieser Hinsicht sind die *Potenztransformationen* (die drei obigen Transformationen sind Spezialfälle); meist verwendet man sie in einer etwas abgewandelten Form (*Box-Cox–Transformationen*):

$$x^{(\lambda)} = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases}$$



11. Der Datensatz `bulb.dat` (UE–Homepage) umfaßt die Lebensdauern [Stunden] von 200 Glühbirnen.
    (a) Ermitteln Sie eine Häufigkeitstabelle; nehmen Sie als Klassengrenzen: 500(100)1500.
    (b) Zeichnen Sie auf Basis der obigen Klasseneinteilung das Histogramm und das Summenpolygon.
    (c) Zeichnen Sie die empirische Verteilungsfunktion.

    [`bulb.r`]

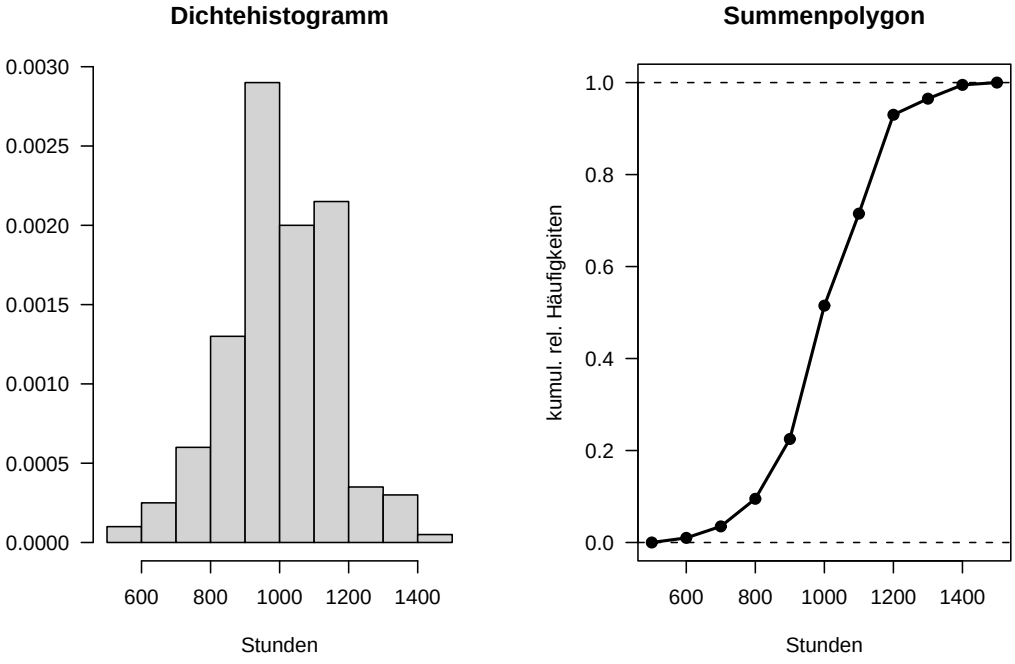
    Lösung:

    (a) Häufigkeitstabelle:

u.i	o.i	z.i	b.i	H.i	h.i	F.i
500	600	550	100	2	0.010	0.010
600	700	650	100	5	0.025	0.035
700	800	750	100	12	0.060	0.095
800	900	850	100	26	0.130	0.225
900	1000	950	100	58	0.290	0.515
1000	1100	1050	100	40	0.200	0.715
1100	1200	1150	100	43	0.215	0.930
1200	1300	1250	100	7	0.035	0.965
1300	1400	1350	100	6	0.030	0.995
1400	1500	1450	100	1	0.005	1.000

    Bem.: `u.i, o.i` = untere, obere Klassengrenzen; `z.i` = Klassenmitten; `b.i` = Klassenbreiten; `H.i` = Klassenhäufigkeiten; `h.i` = relative Klassenhäufigkeiten; `F.i` = kumulierte relative Klassenhäufigkeiten.

(b)



1.2.3 Lageparameter

12. Bestimmen Sie für den Datensatz `inversions.dat` die folgenden Lageparameter: Mittelwert, Median, Modalwert (Modus).
[`invers.r`]

Lösung: Der Mittelwert $\overline{x}_n$ ist das arithmetische Mittel der Daten:

$$\overline{x}_n = \frac{1}{200}\sum_{i=1}^{200} x_i = \frac{1}{200}(25 + 18 + 15 + \cdots + 22 + 14) = 22.59$$

Der Median $\widetilde{x}$ ist der mittlere Wert der geordneten Daten; da es für $n = 200$ aber zwei mittlere Werte gibt, nimmt man deren arithmetisches Mittel:

$$\widetilde{x} = \frac{x_{(100)} + x_{(101)}}{2} = \frac{22 + 22}{2} = 22$$

Der Modalwert ist bei diskreten Daten der am häufigsten vorkommende Wert:

$$x_{\mathrm{mod}} = 22$$

Bem.:

(1) Als Folge der (annähernden) Symmetrie der Verteilung sind die drei Lageparameter nahezu identisch.

(2) Weder $\overline{x}_n$ noch $\widetilde{x}$ sind notwendigerweise Elemente der Menge aller möglichen Beobachtungen (hier: $M = \{0, 1, 2, \ldots, n(n-1)/2\}$); als – auf unterschiedliche Weise typische – Repräsentanten ihres Datensatzes sind sie als *fiktive* Werte zu betrachten.

(3) Wegen seiner meist nur geringen Repräsentativität spielt der empirische Modalwert nur eine untergeordnete Rolle; bei theoretischen Verteilungen allerdings repräsentiert er die Ausprägungen größter Wahrscheinlichkeit.

13. Bestimmen Sie für die Datensätze (a) `NumberRuns.dat` und (b) `MaxRuns.dat` die folgenden Lageparameter: Mittelwert, Median, Modalwert.
[`runs.r`]

Lösung:

(a) Anzahl der Läufe:

$$\overline{x}_n = 30.616, \qquad \widetilde{x} = 30, \qquad x_{\text{mod}} = 30$$

Als Folge der (annähernden) Symmetrie der Verteilung sind die drei Lageparameter nahezu identisch.

(b) Länge des längsten Laufs:

$$\overline{x}_n = 6.2, \qquad \widetilde{x} = 6, \qquad x_{\text{mod}} = 5$$

Hier gilt: $x_{\text{mod}} < \widetilde{x} < \overline{x}_n$; diese Reihenfolge ist typisch für rechtsschiefe Verteilungen.

14. Bestimmen Sie für den Datensatz `resistor.dat` die folgenden Lageparameter: Mittelwert (unklassierte und klassierte Daten), Modalwert (klassierte Daten).
[`resistor.dat`]

Lösung: Der Mittelwert der unklassierten Daten ist:

$$\overline{x} = \frac{1}{80}\sum_{i=1}^{80} x_i = 75.19 \text{ [Ohm]}$$

Stützt man sich auf die Klasseneinteilung 69.55(1.1)80.55, so ergibt sich (z_j = Klassenmitten; $H_n(K_j)$ = Klassenhäufigkeiten):

$$\overline{x}_g = \frac{1}{n}\sum_{j=1}^{m} z_j H_n(K_j) = \frac{1}{80}(70.1 \cdot 1 + 71.2 \cdot 2 + \cdots + 78.9 \cdot 4 + 80.0 \cdot 1) = 75.26 \text{ [Ohm]}$$

Wie man am Histogramm sieht, ist $K_5 = (73.95, 75.05]$ die Modalklasse; deren Mittelpunkt betrachtet man als Modalwert:

$$x_{\text{mod}} = 74.5$$

Bem.:

(a) Man überlegt sich leicht, daß der Abstand zwischen dem Mittelwert $\overline{x}$, berechnet auf Basis der unklassierten Daten, und dem Mittelwert $\overline{x}_g$, berechnet auf Basis einer Klasseneinteilung (mit äquidistanten Klassen der Breite b), maximal die halbe Klassenbreite betragen kann:

$$|\overline{x} - \overline{x}_g| \leq \frac{b}{2}$$

(b) Die Bestimmung von Modalwerten auf Basis von unklassierten Daten ist bei stetigen Merkmalen nicht sinnvoll.

15. Bestimmen Sie für den Datensatz `bulb.dat` die Quartile, d.h. das 25%–, das 50%– (= Median) und das 75%–Quantil (Fraktil); stützen Sie sich dabei auf die klassierten Daten. *Wie lautet ein allgemeiner Ausdruck für das p–Quantil (Fraktil) auf Basis eines klassierten Datensatzes?
[`bulb.r`]

Lösung: Nach Definition der VO ist das p–Quantil jener Wert, für den das Summenpolygon den Wert p hat. Für die Bestimmung dieser Werte muß man zuerst die entsprechende Klasse ausfindig machen. Auf Basis der Häufigkeitstabelle sieht man, daß das 1. Quartil in die 5. Klasse, $K_5 = (900; 1000]$, fällt (die kumulierten relativen Klassenhäufigkeiten sind dort erstmals größer oder gleich 0.25); daher gilt (lineare Interpolation; vgl. die Abbildung):

$$x_{0.25} = 900 + \frac{0.25 - 0.225}{0.290} \cdot 100 \doteq 908.62 \text{ [Stunden]}$$

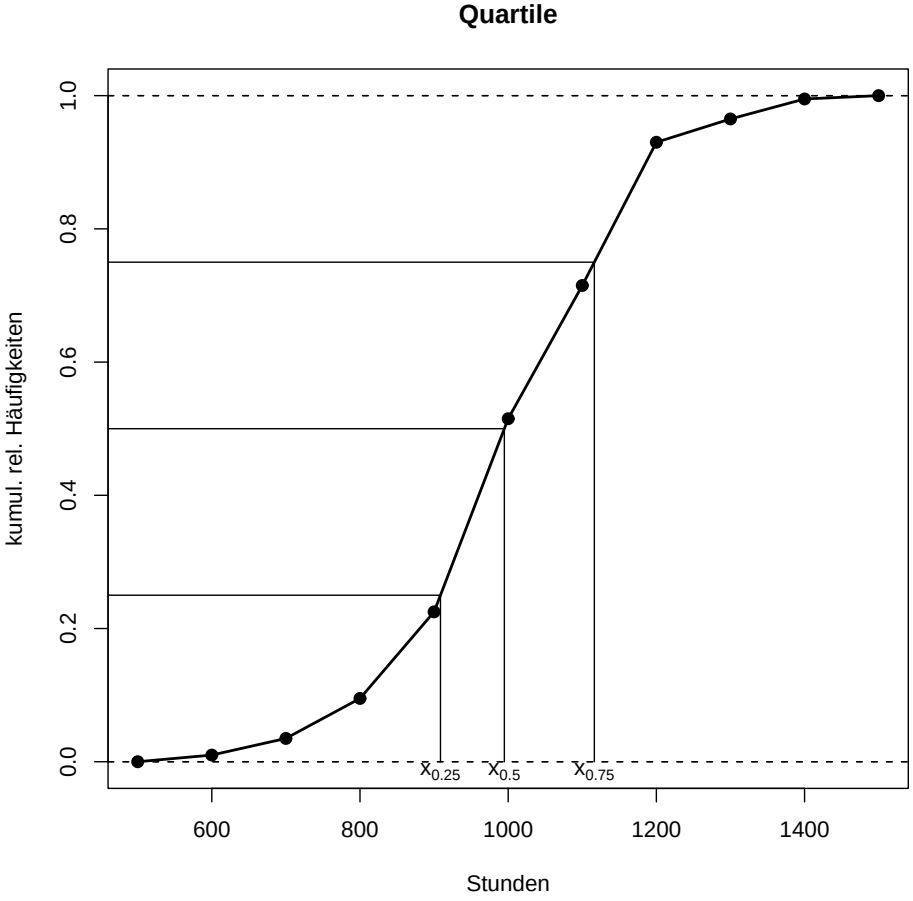
Analog bestimmt man die anderen Quartile; der Median liegt ebenfalls in der 5. Klasse:

$$x_{0.50} = 900 + \frac{0.50 - 0.225}{0.290} \cdot 100 \doteq 994.83 \text{ [Stunden]}$$

Das 3. Quartil liegt in der 7. Klasse, $K_7 = (1100; 1200]$:

$$x_{0.75} = 1100 + \frac{0.75 - 0.715}{0.215} \cdot 100 \doteq 1116.23 \text{ [Stunden]}$$

Die folgende Abbildung zeigt die graphischen Verhältnisse:



Allgemein: Ist $K_i = (u_i, o_i]$ diejenige Klasse, in die das p–Quantile hineinfällt, so gilt (lineare Interpolation):

$$x_p = u_i + \frac{p - \sum_{j=1}^{i-1} H_j/n}{H_i/n} (o_i - u_i) = u_i + \frac{np - \sum_{j=1}^{i-1} H_j}{H_i} (o_i - u_i)$$

16. Zeigen Sie für einen beliebigen Datensatz (diskret oder stetig) $x_1, \ldots, x_n$, daß:

    (a) der Mittelwert $\overline{x}_n$ die folgende Minimumeigenschaft hat:

$$\sum_{i=1}^{n} (x_i - \overline{x}_n)^2 \leq \sum_{i=1}^{n} (x_i - c)^2, \quad c \in \mathbb{R}$$

*(b) der Median $\widetilde{x}$ die folgende Minimumeigenschaft hat:

$$\sum_{i=1}^{n} |x_i - \widetilde{x}| \leq \sum_{i=1}^{n} |x_i - c|, \quad c \in \mathbb{R}$$

Lösung:

(a) Die Funktion $g(c) = \sum_{i=1}^{n}(x_i - c)^2$ ist überall differenzierbar:

$$g'(c) = -2\sum_{i=1}^{n}(x_i - c) = 0 \quad \longrightarrow \quad \sum_{i=1}^{n} x_i = nc \quad \longrightarrow \quad c = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n$$

Daß $\overline{x}_n$ die Stelle eines lokalen (hier auch globalen) Minimums ist, sieht man mit der 2. Ableitung:

$$g''(c) = 2 > 0, \quad c \in \mathbb{R}$$

(b) Die Betragsfunktion $|x|$ ist zwar an der Stelle $x = 0$ nicht differenzierbar, aber es gilt:

$$\frac{d}{dx}|x| = \operatorname{sgn}(x) \quad \text{für} \quad x \neq 0$$

Damit folgt für $h(c) = \sum_{i=1}^{n} |x_i - c|$:

$$h'(c) = \sum_{i=1}^{n} \operatorname{sgn}(x_i - c) = 0 \quad \longrightarrow \quad c = \text{Median}\{x_1, \ldots, x_n\}$$

Der Median bewirkt, daß die eine Hälfte der Summanden gleich -1 und die andere gleich $+1$ ist.

1.2.4 Streuungsparameter

17. Bestimmen Sie für den Datensatz `inversions.dat` die mittlere quadratische Abweichung (empirische Varianz) und die empirische Streuung.
[`inversions.r`]

Lösung: Die mittlere quadratische Abweichung berechnet man wie folgt:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{200}\sum_{i=1}^{200}(x_i - 22.59)^2 = 27.8219$$

Rechnet man mit der Hand, empfiehlt sich die Anwendung des *Verschiebungssatzes* (vgl. unten):

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}_n^2 = 538.13 - (22.59)^2 = 27.8219$$

Die Streuung ist die (positive) Wurzel aus der Varianz:

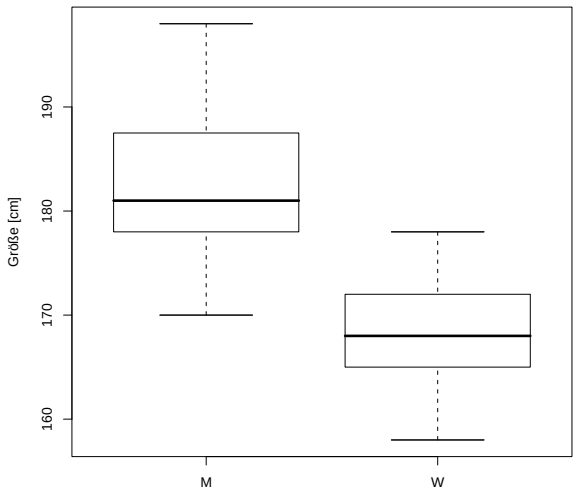
$$s = +\sqrt{27.8219} = 5.2746$$

18. Bestimmen und vergleichen Sie für das Merkmal „Größe“ (Datensatz: `meddat.dat`) die Varianz, die Streuung und den Variationskoeffizienten für Männer und Frauen.
[`meddat.r`]

Lösung: Die folgenden Werte wurden mit Hilfe von R berechnet:

	M	W
Mittelwert	182.5128	168.2667
Varianz	41.7370	26.3956
Streuung	6.4604	5.1377
Variationskoeffizient	0.0354	0.0305

Bem.: Eine gute Möglichkeit, einen Datensatz darzustellen aber auch Lage und Streuung von mehreren Datensätzen miteinander zu vergleichen ist der *Boxplot*: Die dicken Linien repräsentieren den Median; die Box entspricht den mittleren 50% der Daten; die strichlierten Linien erstrecken sich bis zu den äußersten Punkten, die vom Rand der Box noch innerhalb von $1.5 \times \text{QA}$ liegen. Gibt es Punkte, die außerhalb davon liegen, werden sie eigens dargestellt („Ausreißer“).

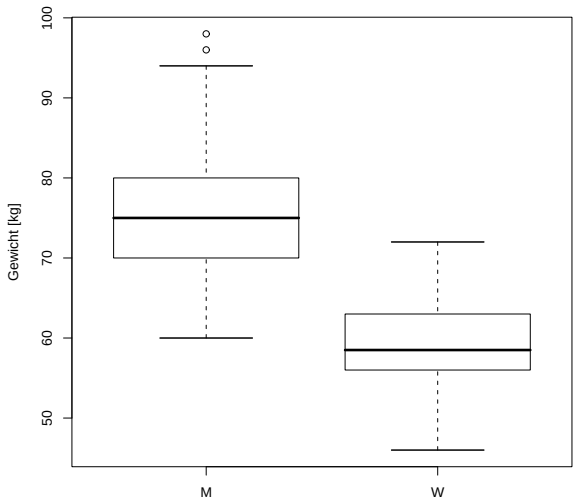


19. [Fortsetzung des vorhergehenden Beispiels] Wiederholen Sie die Aufgabe für das Merkmal „Gewicht“. [`meddat.r`]

Lösung: Die folgenden Werte wurden mit Hilfe von R berechnet:

	M	W
Mittelwert	75.8974	58.4667
Varianz	77.7331	34.3156
Streuung	8.8166	5.8579
Variationskoeffizient	0.1162	0.1002

Boxplots:



Bem.: Die zwei größten Punkte bei den Männern (96, 98) werden eigens dargestellt, d.h. es handelt sich um Punkte, die etwas abseits des Datenbulks liegen. Solche Punkte bezeichnet man als (potentielle) „Ausreißer". Eine exakte Charakterisierung derartiger Punkte ist allerdings schwierig. Üblicherweise führt man – bei Verdacht auf Ausreißer – statistische Analysen mit und ohne diese(n) Punkte(n) durch und vergleicht die Ergebnisse. Eine andere Möglichkeit ist die Verwendung von „robusten" statistischen Methoden.

20. Ermitteln Sie für den Datensatz `bulb.dat` (UE–Homepage) die folgenden Streuungsparameter: Spannweite, Quartilabstand, MAD, Varianz, Streuung.

Zusatz: Wie in der VO kurz diskutiert, versucht man in der schließenden Statistik den empirisch gegebenen Verteilungen (Histogrammen) theoretische Verteilungen (Dichten) anzupassen. Man versuche dies hier mit der Anpassung einer *Normaldichte*; für die beiden Parameter dieser Verteilung (μ, σ^2) nehme man die entsprechenden empirischen Größen ($\overline{x}$, s^2).

[`bulb.r`]

Lösung: Für die Spannweite der Daten ergibt sich:

$$\text{SW} \ = x_{(n)} - x_{(1)} = 1425 - 521 = 905 \text{ [Stunden]}$$

Die Quartile wurden bereits in einem früheren Beispiel bestimmt:

$$\text{QA} \ = x_{0.75} - x_{0.25} = 1116.28 - 908.62 = 207.66 \text{ [Stunden]}$$

Mittlere absolute Abweichung:

$$\text{MAD} \ = \frac{1}{n}\sum_{i=1}^{n} |x_i - \widetilde{x}| = 124.51 \text{ [Stunden]}$$

Mittlere quadratische Abweichung (empirische Varianz):

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = 24795.95 \text{ [Stunden}^2\text{]}$$

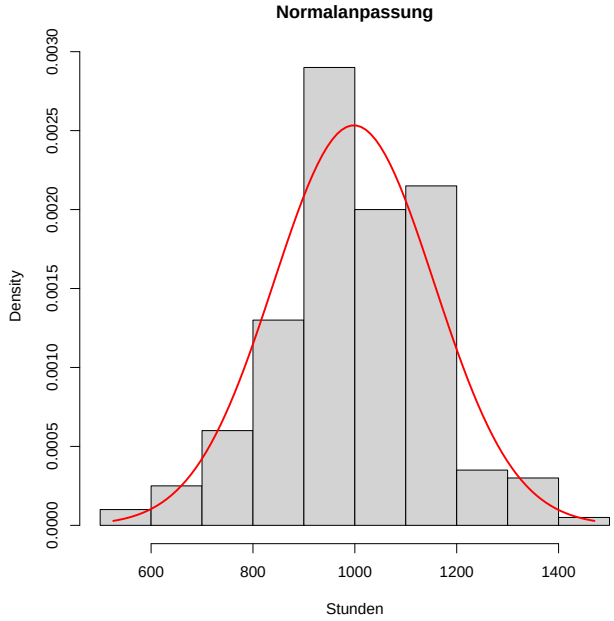
Die empirische Streuung ist die Quadratwurzel aus der Varianz:

$$s = +\sqrt{s^2} = 157.47 \text{ [Stunden]}$$

Bem.: Die z.T. sehr unterschiedlichen Werte (und Einheiten) für die verschiedenen Streuungsparameter erklären sich daraus, daß der Begriff „Streuung" mehrdeutig ist und viele Aspekte hat; ähnliches gilt auch für die „Lage" eines Datensatzes.

Zusatz: Nimmt man für die Parameter der Normalverteilung die empirischen Gegenstücke, so lautet die angepaßte Dichte wie folgt:

$$f(x) = \frac{1}{\sqrt{2\pi}\, s} \exp\left[-\frac{(x-\overline{x})^2}{2s^2}\right], \quad -\infty < x < \infty$$



21. Zeigen Sie den *Verschiebungssatz* für die Varianz (Daten: $x_1, \ldots, x_n$):

$$\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \sum_{i=1}^{n}(x_i - c)^2 - n(c - \overline{x}_n)^2, \quad c \in \mathbb{R}$$

Speziell für $c = 0$ gilt:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}_n^2$$

Lösung: Dies zeigt man durch Subtrahieren und Addieren von c:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 &= \sum_{i=1}^{n}(x_i - c + c - \overline{x}_n)^2 \\
&= \sum_{i=1}^{n}(x_i - c)^2 + n(c - \overline{x}_n)^2 + 2(c - \overline{x}_n)\sum_{i=1}^{n}(x_i - c) \\
&= \sum_{i=1}^{n}(x_i - c)^2 + n(c - \overline{x}_n)^2 + 2(c - \overline{x}_n)\left(\sum_{i=1}^{n} x_i - nc\right) \\
&= \sum_{i=1}^{n}(x_i - c)^2 + n(c - \overline{x}_n)^2 + 2(c - \overline{x}_n)(n\overline{x}_n - nc) \\
&= \sum_{i=1}^{n}(x_i - c)^2 + n(c - \overline{x}_n)^2 - 2n(c - \overline{x}_n)^2 \\
&= \sum_{i=1}^{n}(x_i - c)^2 - n(c - \overline{x}_n)^2
\end{aligned}$$

22. Angenommen, es gibt Daten (Meßwerte) $x_1, \ldots, x_j$ und Sie benötigen den (empirischen) Mittelwert $\overline{x}_j$ und die (empirische) Varianz s_j^2. Nun kommt eine weitere Beobachtung x_{j+1} dazu. Im Sinne einer *Realtime*–Berechnung ist es nicht notwendig, den Mittelwert und die Varianz für alle $j+1$ Daten neu zu berechnen, sondern man kann auf die bereits vorhandenen Werte zurückgreifen. Zeigen Sie:

(a) Rekursion für den Mittelwert:

$$\overline{x}_{j+1} = \overline{x}_j + \frac{1}{j+1}(x_{j+1} - \overline{x}_j), \quad j = 1, 2, \ldots; \ \overline{x}_1 = x_1$$

*(b) Rekursion für die Varianz:

$$s_{j+1}^2 = \left(1 - \frac{1}{j+1}\right) s_j^2 + j(\overline{x}_{j+1} - \overline{x}_j)^2, \quad j = 1, 2, \ldots;\ s_1^2 = 0$$

Überprüfen Sie die Gültigkeit der beiden Rekursionen konkret an den Werten: 3, 4, 7, 2, 9, 6.

Lösung:

(a) Dies folgt einfach aus der Definition:

$$\begin{aligned}
\overline{x}_{j+1} &= \frac{1}{j+1} \sum_{i=1}^{j+1} x_i \\
&= \frac{1}{j+1} \left(\sum_{i=1}^{j} x_i + x_{j+1} \right) \\
&= \frac{1}{j+1} (j\overline{x}_j + x_{j+1}) \\
&= \frac{1}{j+1} \left[(j+1)\overline{x}_j + x_{j+1} - \overline{x}_j \right] \\
&= \overline{x}_j + \frac{1}{j+1} (x_{j+1} - \overline{x}_j)
\end{aligned}$$

(b) Zunächst gilt nach dem Verschiebungssatz (vgl. das vorige Bsp.) für $c = \overline{x}_j$:

$$\begin{aligned}
(j+1)s_{j+1}^2 &= \sum_{i=1}^{j+1} (x_i - \overline{x}_{j+1})^2 \\
&= \sum_{i=1}^{j+1} (x_i - \overline{x}_j)^2 - (j+1)(\overline{x}_j - \overline{x}_{j+1})^2 \\
&= \sum_{i=1}^{j} (x_i - \overline{x}_j)^2 + (x_{j+1} - \overline{x}_j)^2 - (j+1)(\overline{x}_j - \overline{x}_{j+1})^2 \\
&= j s_j^2 + (x_{j+1} - \overline{x}_j)^2 - (j+1)(\overline{x}_j - \overline{x}_{j+1})^2
\end{aligned}$$

Nun wissen wir von (a):

$$x_{j+1} - \overline{x}_j = (j+1)(\overline{x}_{j+1} - \overline{x}_j) \quad \longrightarrow \quad (x_{j+1} - \overline{x}_j)^2 = (j+1)^2 (\overline{x}_{j+1} - \overline{x}_j)^2$$

Setzt man dies oben ein, so ergibt sich:

$$\begin{aligned}
(j+1)s_{j+1}^2 &= j s_j^2 + (j+1)^2 (\overline{x}_{j+1} - \overline{x}_j)^2 - (j+1)(\overline{x}_j - \overline{x}_{j+1})^2 \\
&= j s_j^2 + (j+1)(\overline{x}_{j+1} - \overline{x}_j)^2 (j+1-1) \\
&= j s_j^2 + j(j+1)(\overline{x}_{j+1} - \overline{x}_j)^2
\end{aligned}$$

Daraus folgt die Behauptung:

$$s_{j+1}^2 = \frac{j}{j+1} s_j^2 + j(\overline{x}_{j+1} - \overline{x}_j)^2 = \left(1 - \frac{1}{j+1}\right) s_j^2 + j(\overline{x}_{j+1} - \overline{x}_j)^2$$

Bsp.: Die direkte Berechnung (mittels R) ergibt:

$$\overline{x}_6 = 5.1667, \qquad s_6^2 = 5.8056$$

Nun rekursiv:

i	$\overline{x}_i$	s_i^2
1	3	0
2	$3+\frac{1}{2}(4-3)=\frac{7}{2}$	$\left(1-\frac{1}{2}\right)0+1\left(\frac{7}{2}-3\right)^2=\frac{1}{4}$
3	$\frac{7}{2}+\frac{1}{3}\left(7-\frac{7}{2}\right)=\frac{14}{3}$	$\left(1-\frac{1}{3}\right)\frac{1}{4}+2\left(\frac{14}{3}-\frac{7}{2}\right)^2=\frac{26}{9}$
4	$\frac{14}{3}+\frac{1}{4}\left(2-\frac{14}{3}\right)=4$	$\left(1-\frac{1}{4}\right)\frac{26}{9}+3\left(4-\frac{14}{3}\right)^2=\frac{7}{2}$
5	$4+\frac{1}{5}(9-4)=5$	$\left(1-\frac{1}{5}\right)\frac{7}{2}+4(5-4)^2=\frac{34}{5}$
6	$5+\frac{1}{6}(6-5)=\frac{31}{6}=5.1667$	$\left(1-\frac{1}{6}\right)\frac{34}{5}+5\left(\frac{31}{6}-5\right)^2=\frac{209}{36}=5.8056$

1.3 Wahrscheinlichkeitsbegriffe

1.3.1 Klassische Wahrscheinlichkeitsdefinition

23. (a) Bei einem Picknick von 50 Personen hatten 30 Hamburger, 25 hatten Hotdogs und 15 hatten beides. Wieviele hatten weder noch?

    (b) Auf wieviele Arten kann ein 20 Fragen wahr/falsch–Test beantwortet werden?

    (c) In einer Lade sind 12 schwarze und 12 weiße Socken. Welche minimale Anzahl von Socken muß man (zufällig) entnehmen, um zumindest ein passendes Paar zu bekommen?

    (d) In einem Behälter befinden sich 3 rote, 4 weiße und 5 blaue Kugeln. Auf wieviele Arten kann man 4 Kugeln entnehmen, sodaß von jeder Farbe eine Kugel darunter ist?

    Lösung:

    (a) Additionsprinzip: Jede Person gehört zu genau einer der folgenden disjunkten Mengen:

    $$A = \text{nur Hamburger}, \quad B = \text{nur Hotdogs}, \quad C = \text{beides}, \quad D = \text{weder noch}$$

    $$|A|+|B|+|C|+|D| = (30-15)+(25-15)+15+|D| = 50 \quad \Longrightarrow \quad |D| = 10$$

    (b) Multiplikationsprinzip: Jede Frage kann auf 2 Arten beantwortet werden:

    $$2^{20} = 1048576$$

    (c) „Postfachprinzip“: Es gibt zwei Fächer (weiß, schwarz) und 24 Socken. Die minimale Zahl ist also 3.

    (d) Additions– und Multiplikationsprinzip:

    $$\binom{3}{2}\binom{4}{1}\binom{5}{1}+\binom{3}{1}\binom{4}{2}\binom{5}{1}+\binom{3}{1}\binom{4}{1}\binom{5}{2} = 60+90+120 = 270$$

24. In einer Schuhablage gibt es 5 verschiedene Paare und 5 Schuhe werden willkürlich herausgegriffen. Mit welcher Wahrscheinlichkeit befindet sich darunter kein Paar, 1 Paar, 2 Paare?

    Lösung: Es gibt 5 Schuhtypen (etwa v w x y z); gibt es kein Paar, sind alle Typen beteiligt:

    $$P\{\text{kein Paar}\} = \frac{\binom{5}{5}\binom{2}{1}^5}{\binom{10}{5}} = \frac{32}{252}$$

Bei 1 Paar (z.B. ww x y z) sind 4 Typen beteiligt, einer davon bildet das Paar:

$$P\{\text{1 Paar}\} = \frac{\binom{5}{4}\binom{4}{1}\binom{2}{2}\binom{2}{1}^2}{\binom{10}{5}} = \frac{160}{252}$$

Bei 2 Paaren (z.B. xx yy z) sind 3 Typen beteiligt, zwei davon bilden die Paare:

$$P\{\text{2 Paare}\} = \frac{\binom{5}{3}\binom{3}{2}\binom{2}{2}^2\binom{2}{1}}{\binom{10}{5}} = \frac{60}{252}$$

Die Summe dieser Wahrscheinlichkeiten ist Eins:

$$\frac{32}{252} + \frac{160}{252} + \frac{60}{252} = 1$$

25. Angenommen, alle 365 Tage eines Jahres kommen mit gleicher Wahrscheinlichkeit als Geburtstage in Frage. Wie groß ist die Wahrscheinlichkeit, daß in einer Gruppe von 60 Personen genau 5 Personen am selben Tag geboren sind?

Lösung: Betrachtet man die Aufgabe als „Kugel/Fächer–Problem“, so gibt es 60 Kugeln (Personen) und 365 Fächer (Tage) mit unbegrenztem Fassungsvermögen; die Zahl der möglichen Aufteilungen der Geburtstage beträgt also $m = 365^{60}$. Die Zahl der günstigen Aufteilungen beträgt $g = \binom{60}{5}(365)_{56}$; letztere Zahl kommt wie folgt zustande: Wähle zunächst 5 Personen aus den 60 aus; zu jeder dieser $\binom{60}{5}$ Möglichkeiten gibt es $365 \cdot 364 \cdots 310 = (365)_{56}$ Möglichkeiten von 56 verschiedenen Geburtstagen (einer für die 5 Personen und 55 weitere für die restlichen Personen). Die gesuchte Wahrscheinlichkeit beträgt also:

$$p = \frac{\binom{60}{5}(365)_{56}}{365^{60}} = \frac{\binom{60}{5}}{365^4}\prod_{i=0}^{55}\frac{365-i}{365} \approx 3.6 \cdot 10^{-6}$$

(*Bem.*: $(n)_k := n(n-1)(n-2)\cdots(n-k+1)$)

26. Sieben Personen betreten einen Lift im Erdgeschoß eines 11–stöckigen Gebäudes. Angenommen, die Personen steigen unabhängig voneinander und zufällig auf einem der Stockwerke 1 bis 11 aus. Mit welcher Wahrscheinlichkeit steigen alle auf verschiedenen Stockwerken aus?

Lösung: Betrachtet man die Aufgabe als „Kugel/Fächer–Problem“, so gibt es 7 Kugeln (Personen), die in 10 Fächer (Stockwerke) fallen:

$$p = \frac{(10)_7}{10^7} = \frac{10!}{10^7\,3!} = \frac{189}{3125} = 0.065$$

(*Bem.*: $(n)_k := n(n-1)(n-2)\cdots(n-k+1)$)

27. In einem Feld der Länge n werden zufällig k ($\leq n$) Daten abgelegt. Mit welcher Wahrscheinlichkeit kommt es dabei zu Kollisionen (d.h. Mehrfachbelegungen)? Wie groß muß k konkret für $n = 100$ mindestens sein, damit diese Wahrscheinlichkeit größer als 0.5 (0.9) ist?

Lösung: Hier ist es einfacher, zunächst das komplementäre Ereignis zu betrachten:

$$W\{\text{keine Kollison}\} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k} = \frac{(n)_k}{n^k} = \prod_{i=0}^{k-1}\left(1 - \frac{i}{n}\right)$$

Die gesuchte Wahrscheinlichkeit beträgt also:

$$W\{\text{Kollision}\} = 1 - \prod_{i=0}^{k-1} \left(1 - \frac{i}{n}\right)$$

Durch Probieren findet man schnell heraus, daß letztere Wahrscheinlichkeit für $n = 100$ bereits für $k = 13$ größer als 0.5 und für $k = 22$ größer als 0.9 ist.

Bem.: Bezeichnet man mit p die Mindestwahrscheinlichkeit einer Kollision, so gilt annähernd für großes n:

$$k(n) \approx \sqrt{-2n \ln(1-p)}$$

28. (a) Wieviele Möglichkeiten gibt es, r unterscheidbare Kugeln auf n Fächer so zu verteilen, daß das i–te Fach r_i Kugeln enthält ($i = 1, \ldots, n$, $\sum_{i=1}^{n} r_i = r$)? Bestimmen Sie die Wahrscheinlichkeit für dieses Ereignis.

*(b) Wieviele Möglichkeiten gibt es, r *ununterscheidbare* Kugeln auf n Fächer zu verteilen? Sind diese Möglichkeiten alle gleichwahrscheinlich?

Lösung:

(a) Zuerst wählt man die r_1 Kugeln, die in das 1. Fach fallen; dafür gibt es $\binom{r}{r_1}$ Möglichkeiten. Dann wählt man aus den verbleibenden $r - r_1$ Kugeln die r_2, die in das 2. Fach fallen; dafür gibt es $\binom{r-r_1}{r_2}$ Möglichkeiten, usw. Nach dem allgemeinen Multiplikationsprinzip gilt:

$$g = \binom{r}{r_1}\binom{r-r_1}{r_2} \cdots \underbrace{\binom{r-r_1-\cdots-r_{n-1}}{r_n}}_{=1} = \frac{r!}{r_1!\, r_2! \cdots r_n!} =: \binom{r}{r_1,\, r_2, \ldots,\, r_n}$$

Insgesamt gibt es $m = n^r$ Möglichkeiten der Aufteilung; die gesuchte Wahrscheinlichkeit beträgt also:

$$\frac{g}{m} = \frac{r!}{r_1!\, r_2! \cdots r_n!\, n^r}$$

(b) Die n Fächer bestehen insgesamt aus $n + 1$ „Trennwänden“; 2 davon stehen an den Rändern. Eine mögliche Aufteilung für beispielsweise $r = 10$ und $n = 5$ schaut wie folgt aus:

$$|xx||xxx|x|xxxx|$$

Die r Kugeln und die mittleren $n-1$ Trennwände können in jeder beliebigen Anordnung aufeinander folgen; dafür gibt es die folgende Zahl von Möglichkeiten:

$$\binom{n+r-1}{n-1} = \binom{n+r-1}{r}$$

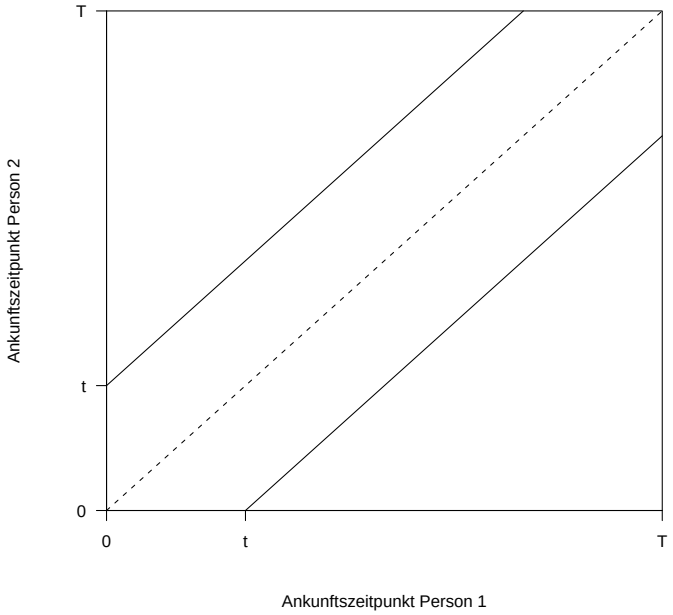
Diese sind (i.a.) nicht gleichwahrscheinlich.

1.3.2 Geometrische Wahrscheinlichkeiten

29. Zwei Personen haben die gleiche Wahrscheinlichkeit an einem bestimmten Ort zu einem beliebigen Zeitpunkt im Intervall $[0, T]$ einzutreffen. Die zuerst eintreffende Person wartet auf die andere. Ermitteln Sie die Wahrscheinlichkeit, daß keine Person länger als t auf die andere warten muß.

Lösung: Eine Person muß länger als t warten, wenn der Abstand zwischen ihren Eintreffzeitpunkten größer als t ist, $|t_1 - t_2| > t$. Die Wahrscheinlichkeit dafür ist $[(T-t)/T]^2$; die gesuchte Wahrscheinlichkeit ist also:

$$p = 1 - \left(\frac{T-t}{T}\right)^2 = 1 - \left(1 - \frac{t}{T}\right)^2$$



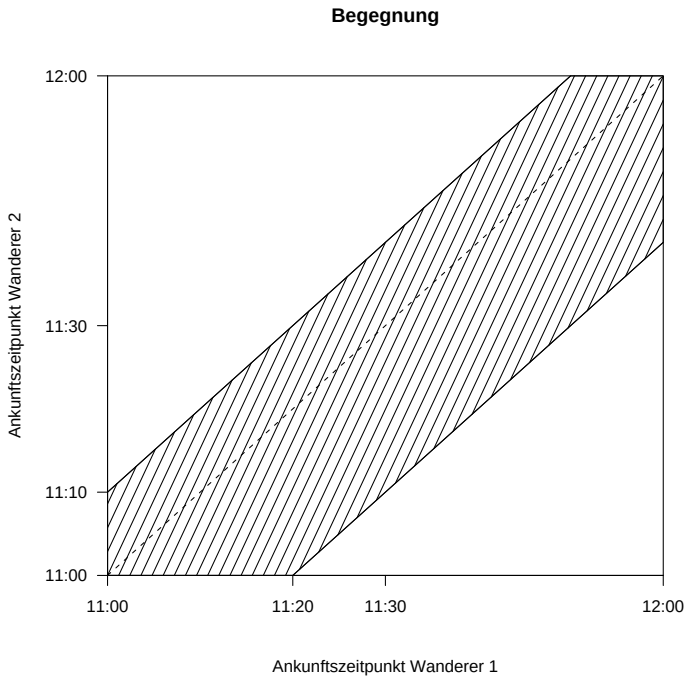
30. Zwei Wanderer erreichen aus unterschiedlichen Richtungen einen Aussichtspunkt und halten sich dort 10 Minuten (Wanderer 1) bzw. 20 Minuten (Wanderer 2) auf. Ihre Ankunftszeitpunkte liegen – unabhängig voneinander – zufällig zwischen 11 und 12 Uhr.

    (a) Mit welcher Wahrscheinlichkeit begegnen sie einander am Aussichtspunkt?

    (b) Wie groß ist die Wahrscheinlichkeit, daß sich um 11:30 (1) keiner, (2) genau einer, (3) beide am Aussichtspunkt befinden?

Lösung:

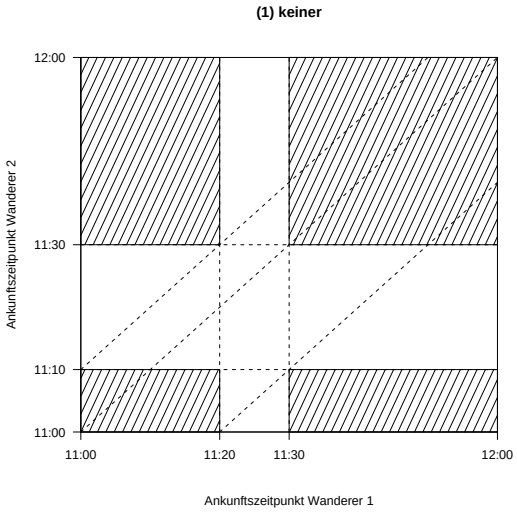
(a) Der schraffierte Bereich entspricht dem Ereignis „Begegnung“:



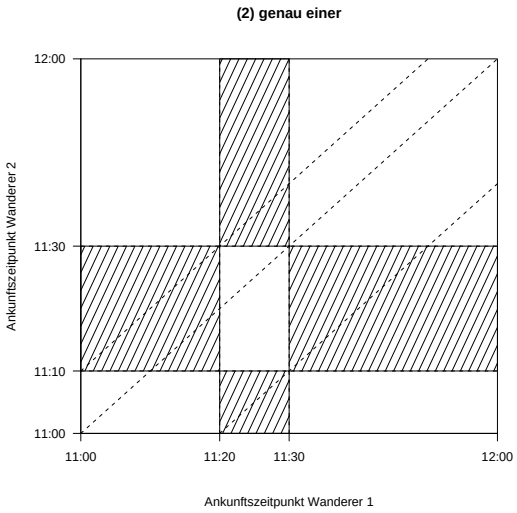
Die Wahrscheinlichkeit einer Begegnung beträgt (Rechnung in Minuten):

$$W\{\text{Begegnung}\} = 1 - \frac{40^2/2 + 50^2/2}{60^2} = 1 - \frac{41}{72} = \frac{31}{72} \doteq 0.43$$

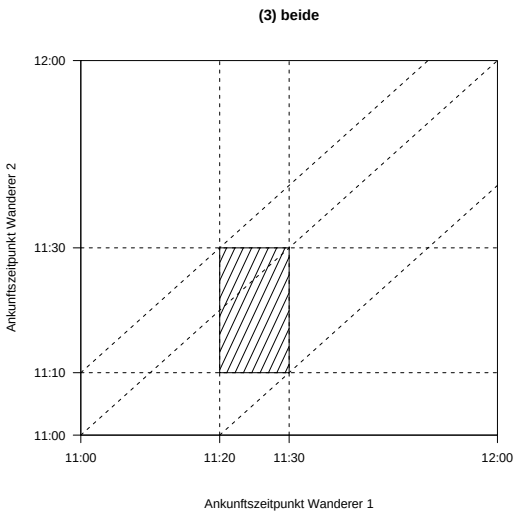
(b) Die Ereignisse teilen den Raum der möglichen Versuchausausgänge in drei disjunkte Bereiche:



$$p_1 = \frac{200 + 600 + 300 + 900}{3600} = \frac{10}{18}$$



$$p_2 = \frac{100 + 300 + 400 + 600}{3600} = \frac{7}{18}$$



$$p_3 = \frac{200}{3600} = \frac{1}{18}$$

31. M verläßt zufällig zwischen 15 und 17 Uhr seinen/ihren Arbeitsplatz und begibt sich zur U–Bahn. Seine/Ihre Mutter lebt am einen Ende der Stadt, seine/ihr Freundin/Freund am anderen. Er/Sie will fair sein und nimmt jeweils diejenige U–Bahn, welche als erste eintrifft. Nach einiger Zeit beklagt sich die Mutter darüber, daß er/sie nur ganz selten zum Abendessen kommt; an den letzten 20 Arbeitstagen nur zweimal. Kommt dieses Ungleichgewicht zufällig zustande oder gibt es eine andere Erklärung dafür?

    Lösung: Auch wenn das Arrangement auf den ersten Blick fair wirkt, so steckt die Ursache für das Mißverhältnis doch im Fahrplan. Angenommen, die Züge Richtung Freundin/Freund fahren ganz regelmäßig alle 10 Minuten um 15:00, 15:10, 15:20, ...; die Züge Richtung Mutter aber regelmäßig um 15:01, 15:11, 15:21, ... Um den Zug Richtung Mutter zu nehmen, muß M also genau in den 1–Minuten Intervallen zwischen den beiden Zügen eintreffen; die Wahrscheinlichkeit dafür beträgt aber nur 1/10. Fair wäre das Arrangement also nur, wenn beide Züge zugleich eintreffen (hier müßte er/sie noch eine Münze werfen) oder genau 5 Minuten dazwischen liegen.

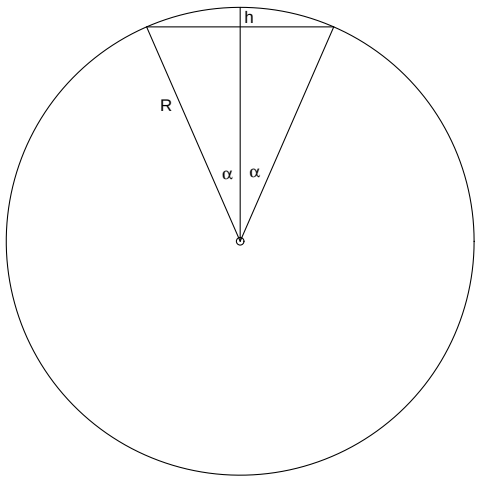
32. Zwei Punkte werden willkürlich auf der Oberfläche einer Kugel mit Radius R markiert. Wie groß ist die Wahrscheinlichkeit, daß der (Groß–) Kreisbogen, der die beiden Punkte miteinander verbindet, einen Winkel kleiner als α ($< \pi$) einschließt?

    Lösung: Betrachtet man einen Punkt als fest, so muß der andere innerhalb einer Kugelkappe (Kalotte) um den ersten als Mittelpunkt und Öffnungswinkel 2α liegen. Die Fläche dieser Kappe ist $2R\pi h$ (Formelsammlung!), wobei:

$$\cos\alpha = \frac{R-h}{R} \quad \longrightarrow \quad h = R(1-\cos\alpha)$$

    Die gesuchte Wahrscheinlichkeit beträgt also:

$$p = \frac{2R^2\pi(1-\cos\alpha)}{4R^2\pi} = \frac{4R^2\pi\sin^2(\alpha/2)}{4R^2\pi} = \sin^2\left(\frac{\alpha}{2}\right)$$



33. Bestimmen Sie die Wahrscheinlichkeit, daß die Wurzeln der quadratischen Gleichung $x^2 + 2ax + b = 0$ reell sind, wenn bekannt ist, daß die Koeffizienten mit gleicher Wahrscheinlichkeit aus dem Rechteck $|a| \leq A$, $|b| \leq B$ stammen. Bestimmen Sie unter diesen Bedingungen auch die Wahrscheinlichkeit dafür, daß die Wurzeln beide positiv sind.

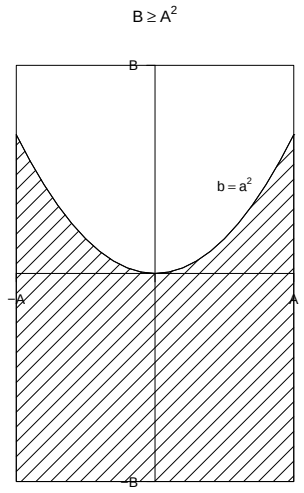
    Lösung: Nach der Lösungsformel für quadratische Gleichungen sind die Lösungen genau dann reell, falls $b \leq a^2$. Um die Wahrscheinlichkeit dafür zu bestimmen, muß man zwei Fälle unterscheiden (vgl. die Abbildungen):

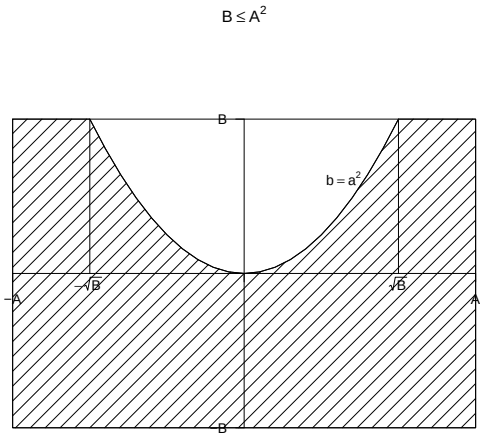
$$B \geq A^2: \quad p = \frac{1}{4AB}\left[2AB + 2\int_0^A a^2\,da\right] = \frac{1}{4AB}\left[2AB + \frac{2a^2}{3}\Big|_0^A\right] = \frac{1}{2} + \frac{A^2}{6B}$$

$$B \leq A^2: \quad p = \frac{1}{4AB}\left[4AB - 2\int_0^B \sqrt{b}\,db\right] = \frac{1}{4AB}\left[4AB - \frac{4b^{3/2}}{3}\Big|_0^B\right] = 1 - \frac{\sqrt{B}}{3A}$$

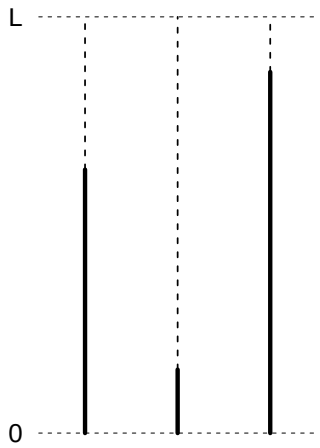
Die Lösungen sind $x_{1,2} = -a \pm \sqrt{a^2 - b}$; damit beide positiv sind, muß $a \leq 0$ und $b \geq 0$. Die Wahrscheinlichkeiten dafür sind:

$$B \geq A^2: \quad p = \frac{A^2}{12B} \qquad\qquad B \leq A^2: \quad p = \frac{1}{4} - \frac{\sqrt{B}}{6A}$$





*34. Drei Geradensegmente, deren Länge nicht größer als L ist, werden zufällig gewählt. Wie groß ist die Wahrscheinlichkeit, daß man mit ihnen ein Dreieck bilden kann?



Lösung: Die Längen der Segmente seien x, y und z. Möglich ist $0 \leq x, y, z \leq L$, also ein Würfel mit Kantenlänge L. Damit sich aus diesen Segmenten ein Dreieck bilden läßt, muß die Summe zweier Längen größer als die dritte sein:

$$x + y \geq z, \quad x + z \geq y, \quad y + z \geq x$$

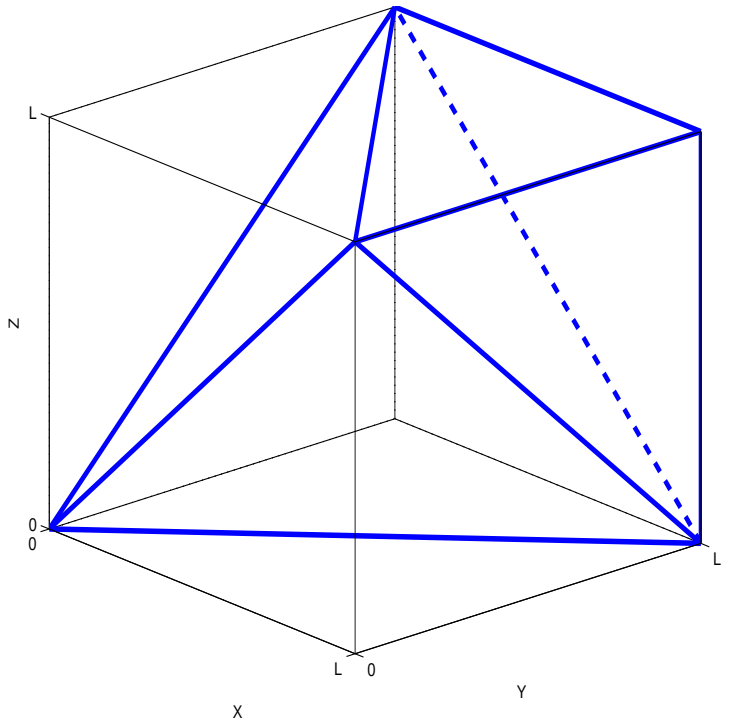
Die Punkte (x, y, z), die die erste Bedingung erfüllen, entsprechen dem Würfel ohne „z–Eck“ (dreieckige Pyramide mit den Eckpunkten $(0, 0, 0), (L, 0, L), (0, L, L), (0, 0, L)$); das Volumen dieses Ecks beträgt:

$$\frac{\text{Grundfläche} \times \text{Höhe}}{3} = \frac{L^2/2 \times L}{3} = \frac{L^3}{6}$$

Analog entspricht die zweite Bedingung dem Würfel ohne y–Eck und die dritte Bedingung dem Würfel ohne x–Eck. Da sich diese Ecken nicht überschneiden, gilt für die gesuchte Wahrscheinlichkeit:

$$p = \frac{L^3 - 3L^3/6}{L^3} = \frac{1}{2}$$

Der günstige Bereich entspricht den Punkten innerhalb des blauen „Drahtgitters“ (Tetraeder mit auf einer Seite aufgesetzter dreieckiger Pyramide):



1.3.3 Grenzwert von Häufigkeiten

35. Der französische Offizier und Schriftsteller Chevalier de Méré (1607 – 1684) wandte sich im Jahre 1654 mit der folgenden Frage an Blaise Pascal (1623 – 1662): Was ist vorteilhafter, beim Spiel mit einem Würfel auf das Eintreten mindestes eines Sechsers in vier Würfen oder beim Spiel mit zwei Würfeln auf das Eintreten eines Doppelsechsers in 24 Würfen zu setzen? Er wußte aus Erfahrung, daß die erste Wette für ihn vorteilhaft ist; bei der zweiten Wette, von der er annahm, daß sie nur eine Variante der ersten sei, gestalteten sich die Einnahmen aber nicht nach seinen Vorstellungen.

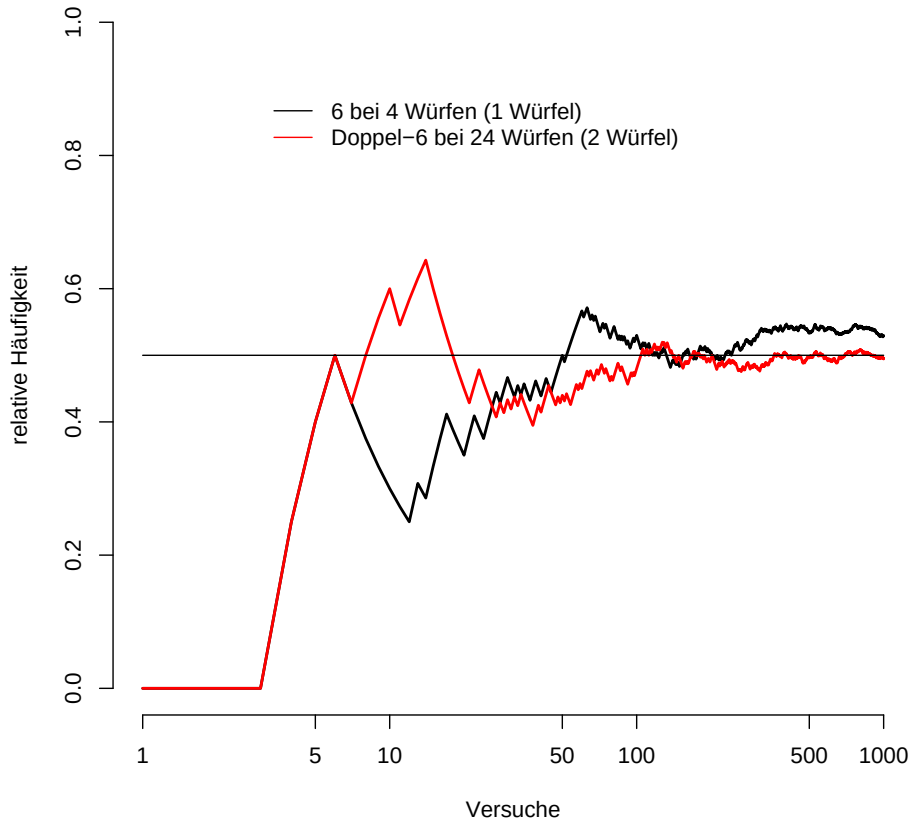
[`demere.r`]

Lösung: Aus heutiger Sicht läßt sich die Frage einfach beantworten:

$$P\{\text{mindestens ein Sechser in 4 Würfen}\} = 1 - \left(\frac{5}{6}\right)^4 = 0.581$$

$$P\{\text{mindestens ein Doppelsechser in 24 Würfen}\} = 1 - \left(\frac{35}{36}\right)^{24} = 0.491$$

Zur Zeit von De Méré gab es die Voraussetzungen dafür noch nicht, sie wurden vielmehr durch Probleme dieser und ähnlicher Art erst nach und nach geschaffen.



36. Ein ähnliches (Wett–) Problem hatte der englische Abgeordnete und berühmte Tagebuchschreiber („Die geheimen Tagebücher“) Samuel Pepys (1633 – 1703). Welches der folgenden Ereignisse ist am wahrscheinlichsten?

    (1) Mindestens 1 Sechser beim Werfen von 6 Würfeln.

    (2) Mindestens 2 Sechser beim Werfen von 12 Würfeln.

    (3) Mindestens 3 Sechser beim Werfen von 18 Würfeln.

    Als Präsident der *Royal Society* gehörte u.a. Isaac Newton (1643 – 1727) zu seinem Bekanntenkreis. Letzterem legte er sein Problem in einem kompliziert formulierten Brief dar.
    [`pepys.r`]

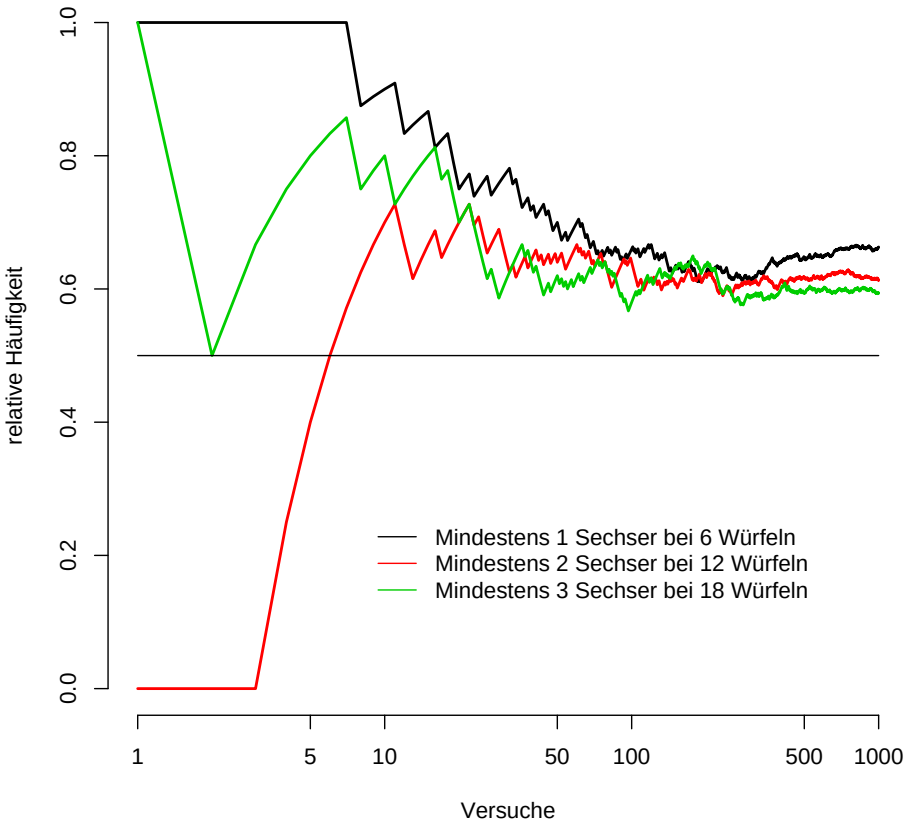
    Lösung: Pepys vermutete, daß alle drei Ereignisse gleich wahrscheinlich sind (mit $W = 1/2$). Er begründete dies damit, daß man bei 6 Würfeln durchschnittlich 1 Sechser erwarten kann, bei 12 Würfeln durchschnittlich 2, usw. Dies ist zwar korrekt, doch verwechselte er dabei „Wahrscheinlichkeit“ mit „Durchschnittswert“.
    Aus heutiger Sicht läßt sich die Frage einfach beantworten. Die Wahrscheinlichkeit, beim Werfen allgemein von $6n$ Würfeln mindestens n Sechser zu bekommen, ist gegeben durch:

$$\sum_{k=n}^{6n} \binom{6n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{6n-k} = 1 - \sum_{k=0}^{n-1} \binom{6n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{6n-k}$$

Einige konkrete Wahrscheinlichkeiten:

$6n$	n	W
6	1	0.6651
12	2	0.6187
18	3	0.5973
24	4	0.5845
30	5	0.5757
96	16	0.5424
600	100	0.5170
900	150	0.5139

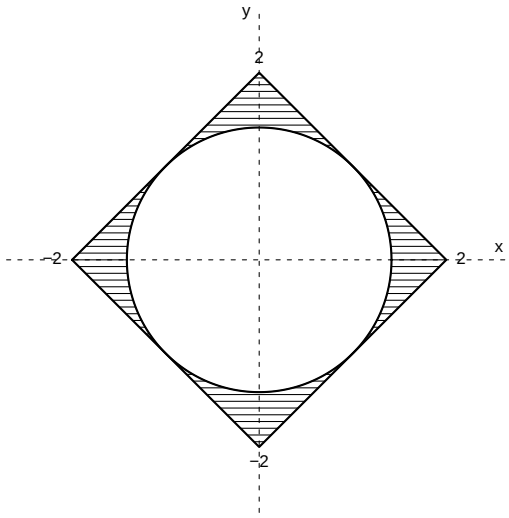
Bem.: Mit wachsendem n wird die Verteilung der Zahl der Sechser immer symmetrischer um den Mittelwert $(= n)$ und die Wahrscheinlichkeit, n oder mehr Sechser zu bekommen, nähert sich dem Wert 1/2.



1.3.4 Axiomatische Wahrscheinlichkeiten

37. Bestimmen Sie das komplementäre Ereignis $A^\perp$ zu $A = \{(x, y) : x^2 + y^2 < 2\}$ in Bezug auf das sichere Ereignis $e = \{(x, y) : |x| + |y| \leq 2\}$.

Lösung: Das komplementäre Ereignis entspricht der schraffierten Fläche (inklusive Randpunkte):



38. Bestimmen Sie $\bigvee_{k=1}^{\infty} A_k$ für die folgenden Ereignisse:

    (a) $A_k = \{x : 1/k \le x \le 3 - 1/k\},\ k = 1, 2, 3, \ldots$

    (b) $A_k = \{(x, y) : 1/k \le x^2 + y^2 \le 4 - 1/k\},\ k = 1, 2, 3, \ldots$

    Lösung: (a) $\{x : 0 < x < 3\}$; (b) $\{(x, y) : 0 < x^2 + y^2 < 4\}$

39. Bestimmen Sie $\bigwedge_{k=1}^{\infty} A_k$ für die folgenden Ereignisse:

    (a) $A_k = \{x : 2 - 1/k < x \le 2\},\ k = 1, 2, 3, \ldots$

    (b) $A_k = \{x : 2 < x \le 2 + 1/k\},\ k = 1, 2, 3, \ldots$

    (c) $A_k = \{(x, y) : 0 \le x^2 + y^2 \le 1/k\},\ k = 1, 2, 3, \ldots$

    Lösung: (a) $\{x : x = 2\}$; (b) o (unmögliches Ereignis); (c) $\{(x, y) : (x, y) = (0, 0)\}$

40. Für jedes (eindimensionale) Ereignis A sei eine Wahrscheinlichkeit wie folgt definiert:

$$W(A) = \sum_A f(x) \quad \text{mit} \quad f(x) = \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^x,\ x = 0, 1, 2, \ldots \quad (f(x) = 0 \text{ sonst})$$

    Wenn $A_1 = \{x : x = 0, 1, 2, 3\}$ und $A_2 = \{x : x = 0, 1, 2, \ldots\}$, bestimmen Sie $W(A_1)$ und $W(A_2)$.

    Lösung: Endliche geometrische Reihe:

$$W(A_1) = \frac{2}{3}\left[1 + \left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^3\right] = \frac{2}{3} \cdot \frac{1 - (1/3)^4}{1 - (1/3)} = \frac{80}{81}$$

    Unendliche geometrische Reihe:

$$W(A_2) = \frac{2}{3}\left[1 + \left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)^2 + \cdots\right] = \frac{2}{3} \cdot \frac{1}{1 - 1/3} = 1$$

41. Für jedes (eindimensionale) Ereignis A sei eine Wahrscheinlichkeit wie folgt definiert:

$$W(A) = \int_A f(x)\, dx \quad \text{mit} \quad f(x) = 6x(1 - x)\, I_{(0,1)}(x)$$

    (Existiert das Integral nicht, ist $W(A)$ nicht definiert.) Wenn $A_1 = \{x : \frac{1}{4} < x < \frac{3}{4}\}$, $A_2 = \{\frac{1}{2}\}$ und $A_3 = \{x : 0 < x < 10\}$, bestimmen Sie $W(A_1)$, $W(A_2)$ und $W(A_3)$.

Lösung:

$$W(A_1) = \int_{1/4}^{3/4} 6x(1-x)\,dx = 6\left(\frac{x^2}{2} - \frac{x^3}{3}\right)\Big|_{1/4}^{3/4} = \frac{11}{16}$$

$$W(A_2) = \int_{1/2}^{1/2} 6x(1-x)\,dx = 0$$

$$W(A_3) = \int_{0}^{10} 6x(1-x)I_{(0,1)}(x)\,dx = \int_{0}^{1} 6x(1-x)\,dx = 1$$

1.3.5 Subjektive Wahrscheinlichkeiten

42. Wahrscheinlichkeiten – insbesondere subjektive – werden häufig in Form von Verhältnissen angegeben. So sagt man beispielsweise, die *Chancen* (engl. *odds*) für den Eintritt von A stehen 1 zu 10. Heißt das nun, daß die Wahrscheinlichkeit $W(A)$ von A gleich $1/9$, $1/10$ oder $1/11$ ist? Umgekehrt, wenn beispielsweise $W(A) = 2/3$, stehen dann die Chancen (für den Eintritt von A) $2:3$, $3:2$, oder $2:1$?

Lösung: Bezeichnet man die Chancen für den Eintritt von A mit $\mathcal{O}(A)$ (und gegen den Eintritt mit $\mathcal{O}(\overline{A})$), so gelten die folgenden Beziehungen:

$$W(A) = \frac{\mathcal{O}(A)}{1+\mathcal{O}(A)}, \qquad \mathcal{O}(A) = \frac{W(A)}{1-W(A)}$$

Die richtige Antwort lautet also im ersten Fall $W(A) = 1/11$, und im zweiten Fall $\mathcal{O}(A) = 2:1$.

Bem.: Wenn $W(A) = p$ und $W(\overline{A}) = 1-p = q$, so schreibt man die Chancen häufig in der folgenden Form:

$$\mathcal{O}(A) = \frac{p}{q} : 1, \qquad \mathcal{O}(\overline{A}) = \frac{q}{p} : 1$$

43. Wenn die Chancen $2:3$ für A, $3:7$ für B und $1:4$ für $A \cap B$ stehen, wie stehen dann die Chancen für $A \cup B$?

Lösung: Aus den Angaben folgt:

$$W(A) = \frac{2}{5}, \quad W(B) = \frac{3}{10}, \quad W(A \cap B) = \frac{1}{5}$$

Damit folgt:

$$W(A \cup B) = W(A) + W(B) - W(A \cap B) = \frac{1}{2}$$

Die Chancen für $A \cup B$ stehen also $1:1$.

44. Ein Buchmacher offeriert in einem bestimmten Rennen für *Lucky Star* eine Quote von $99:1$ auf Sieg. Bedeutet das, daß die Siegeschance des Pferdes in diesem Rennen (nach Einschätzung des Buchmachers) größer, kleiner oder gleich $1/100$ ist ?

Lösung: Bei einer *fairen* Wette auf den Eintritt eines Ereignisses A entsprechen die Chancen $a:1$ der Auszahlungsquote $r:1$:

$$W(A) = \frac{1}{a+1} = \frac{1}{r+1}$$

Alle gewinnorientierten Spiele sind in diesem Sinn „unfair“, d.h. es gilt:

$$W(A) < \frac{1}{r+1}$$

Die Siegeschance von Lucky Star ist also – nach Einschätzung des Buchmachers – kleiner als $1/100$.

45. In einem Rennen mit n Pferden offeriert ein Buchmacher Quoten von $r_i : 1$, $i = 1, \ldots, n$.

    (a) Betrachten Sie die Summe:

    $$S = \sum_{i=1}^{n} \frac{1}{r_i + 1}$$

    Ist S größer, kleiner oder gleich 1 ?

    *(b) Diskutieren Sie den Fall $S < 1$. Könnte man (und wenn ja, wie) einen Vorteil daraus ziehen?

Lösung: A_i sei das Ereignis, daß Pferd i das Rennen gewinnt.

(a) Nach der Lösung des vorigen Beispiels gilt:

$$W(A_i) < \frac{1}{r_i + 1}, \quad i = 1, \ldots, n$$

Daraus folgt:

$$\underbrace{\sum_{i=1}^{n} W(A_i)}_{=1} < \sum_{i=1}^{n} \frac{1}{r_i + 1}$$

D.h., S wird größer als 1 sein.

(b) Der Fall $S < 1$: Durch gleichzeitiges Wetten auf alle n Pferde könnte man in diesem Fall einen sicheren Gewinn erzielen („Dutch Book“). Der Gesamteinsatz betrage $\sum_{j=1}^{n} a_j$; gilt nun $S < 1$, könnte man die Einsätze so verteilen, daß:

$$a_i(r_i + 1) - \sum_{j=1}^{n} a_j > 0, \quad i = 1, \ldots, n$$

D.h., unabhängig davon, welches Pferd gewinnt, der Gesamtgewinn ist größer als Null. Daß eine derartige Aufteilung der Einsätze möglich wäre, sieht man wie folgt:

$$S < 1 \quad \longrightarrow \quad \frac{1}{r_i + 1} < 1, \quad i = 1, \ldots, n$$

D.h., es gibt a_i's mit:

$$\frac{1}{r_i + 1} < \frac{a_i}{\sum_{j=1}^{n} a_j}, \quad i = 1, \ldots, n$$

Bsp.: 10 Pferde, $r_i = 10$, $i = 1, \ldots, 10$, $S = 10/11 < 1$

$$a_i = 1, \ i = 1, \ldots, 10 : \quad \frac{1}{11} < \frac{a_i}{\sum_{j=1}^{10} a_j} = \frac{1}{10}, \ i = 1, \ldots, 10$$

Der Gesamtgewinn beträgt in diesem Fall $1 \cdot 11 - 10 \cdot 1 = 1$.

46. Im Wetterbericht hören Sie, daß für morgen in Wien die Wahrscheinlichkeit für Regen 30% beträgt. Was bedeutet das?

    (a) In 30% des Stadtgebiets wird es morgen regnen.

    (b) Zu 30% der Zeit wird es morgen regnen.

    (c) An 30% der Tage wie morgen regnet es.

Lösung: (a) und (b) sind sicher nicht gemeint; (c) ist die objektivistische (frequentistische) Interpretation. Genauer müßte man sagen, daß es an 30% der Tage in der Vergangenheit mit einer ähnlichen meteorologischen Konstellation (Jahreszeit, Luftdruck, Temperatur, etc.) wie heute am darauf folgenden Tag geregnet hat. Es ist allerdings zweifelhaft, ob die Bewertung tatsächlich auf diese Weise zustande gekommen ist und die Wetterarchive dahingehend durchforstet wurden. Eher kann man vermuten, daß neben dem ohne

Zweifel vorhandenen Erfahrungswissen (und meteorologisch-physikalischen Modellen, etc.) auch subjektive Faktoren eine Rolle spielen („Expertenwissen“); letztere lassen sich meist in Form einer Wette beschreiben:

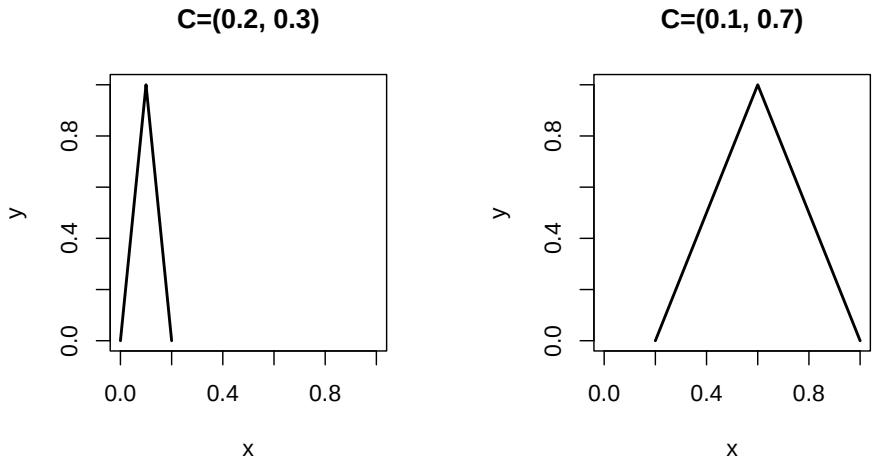
Der/Die Meteorolog(e)in ist bereit, sich (z.B.) auf die folgende Wette einzulassen: Sollte es morgen (für eine bestimmte Zeit) regnen, bekommt er/sie EUR 70,-, andernfalls verliert er/sie EUR 30,-.

Dies entspricht der subjektiven Wahrscheinlichkeitsinterpretation.

1.3.6 Unscharfe Wahrscheinlichkeiten

*47. Ein Experiment bestehe in der Wahl eines Punktes aus dem Intervall $[0,1]$. Für eine Menge $C \subseteq [0,1]$ sei eine unscharfe Wahrscheinlichkeit wie folgt definiert: Die Zugehörigkeitsfunktion von $P^*(C)$ ist dreiecksförmig mit den Eckpunkten ($|C|$ = Länge von C) $(0,0)$, $(|C|,1)$, $(2|C|,0)$, falls $|C| \leq 0.5$, und den Eckpunkten $(2|C|-1,0)$, $(|C|,1)$, $(1,0)$, falls $|C| > 0.5$. Zeigen Sie, daß es sich um eine unscharfe Wahrscheinlichkeitsverteilung handelt.

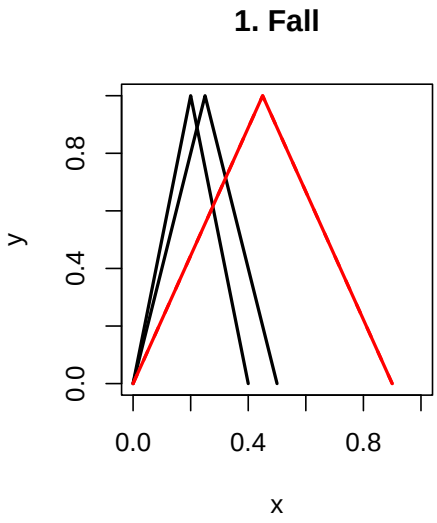
Lösung: Die Zugehörigkeitsfunktionen sind gleichschenkelige Dreiecke mit der Spitze an der Stelle $|C|$; beispielsweise:

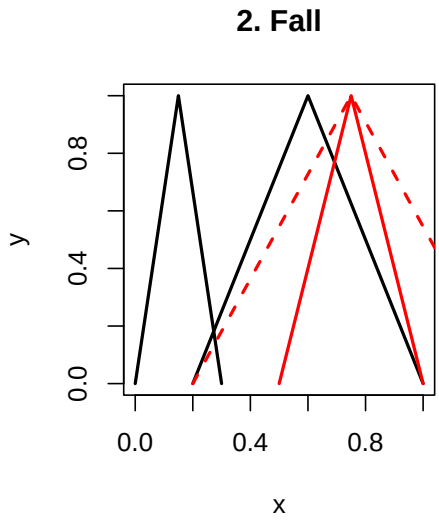


Die δ–Schnitte sind Intervalle und für das unmögliche (sichere) Ereignis ist die Zugehörigkeitsfunktion $I_{\{0\}}(x)$ ($I_{\{1\}}(x)$). Nun zur Additivität: Sind C_1 und C_2 zwei disjunkte Ereignisse, so gibt es zwei Möglichkeiten: Die Summe beider Längen $|C_1| + |C_2| = |C_1 \cup C_2|$ ist kleiner als 0.5, oder größer als 0.5. Im ersten Fall ist die Summe der (unteren/oberen) δ–Schnitte von $P^*(C_1)$, $P^*(C_2)$ identisch mit den (unteren/oberen) δ–Schnitten von $P^*(C_1 \cup C_2)$; im zweiten Fall bildet die Summe der Schnitte ein „Dach“ über der Zugehörigkeitsfunktion von $P^*(C_1 \cup C_2)$. In beiden Fällen gilt also:

$$\overline{P}_\delta(C_1 \cup C_2) \leq \overline{P}_\delta(C_1) + \overline{P}_\delta(C_2), \quad \underline{P}_\delta(C_1 \cup C_2) \geq \underline{P}_\delta(C_1) + \underline{P}_\delta(C_2), \quad \delta \in (0,1]$$

In den folgenden Abbildungen ist die ausgezogene rote Linie die Zugehörigkeitsfunktion von $P^*(C_1 \cup C_2)$, die strichlierte rote Linie die Summe der δ–Schnitte von $P^*(C_1)$ und $P^*(C_2)$.


1. Fall
y
0.0
0.4
0.8
0.0
0.4
0.8
x


2. Fall
y
0.0
0.4
0.8
0.0
0.4
0.8
x