Data Mining / Intelligente Datenanalyse WS 2017/2018
PD Dr.-Ing. habil. Christian Borgelt

Lösung des 1. Übungsblattes

Aufgabe 1 Datenanalyse und Statistik im Alltag

• [Huff 1954] schreibt zu dem Zitat aus dem *Time Magazine*:

Well, good for him!

But wait a minute. What does this impressive figure mean? Is it, as it appears to be, evidence that if you send your boy to Yale you won't have to work in your old age and neither will he?

Two things about the figure stand out at first suspicious glance. It is surprisingly precise. It is quite improbably salubrious.[1]

There is small likelihood that the average income of any far-flung group is ever going to be known down to the dollar. It is not particularly probable that you know your own income for last year so precisely as that unless it was all derived from salary. And $25,000 incomes are not often all salary; people in that bracket are likely to have well-scattered investments.

Furthermore, this lovely average is undoubtedly calculated from the amounts the Yale men *said* they earned. Even if they had the honor system in New Haven in '24, we cannot be sure that it works so well after a quarter of a century that all these reports are honest ones. Some people when asked their incomes exaggerate out of vanity or optimism. Others minimize, especially, it is to be feared, on income-tax returns; and having done this may hesitate to contradict themselves on any other paper. Who knows what the revenuers may see? It is possible that these two tendencies, to boast and to understate, cancel each other out, but it is unlikely. One tendency may be far stronger than the other, and we do not know which one.

We have begun then to account for a figure that common sense tells us can hardly represent the truth. Now let us put our finger on the likely source of the biggest error, a source that can produce $25,111 as the "average income" of some men whose actual average may well be nearer half that amount.

This is the sampling procedure, which is the heart of the greater part of the statistics you meet on all sorts of subjects. Its basis is simple enough, although its refinements in practice have led into all sorts of by-ways, some less than respectable. If you have a barrel of beans, some red and some white, there is only one way to find out exactly how many of each color you have: Count 'em. However, you can find out approximately how many are red in much easier fashion by pulling out a handful of beans and counting just those, figuring that the proportion will be the same all through the barrel. If your sample is large enough and selected properly, it will represent the whole well enough for most purposes. If it is not, it may be far less accurate than an intelligent guess and have nothing to recommend it but a spurious air of scientific precision. It is sad truth that conclusions from such samples, biased or too small or both, lie behind much of what we read or think we know.

The report on the Yale men comes from a sample. We can be pretty sure of that because reason tells us that no one can get hold of all the living members of that class of '24. There are bound to be many whose addresses are unknown twenty-five years later.

And, of those whose addresses are known, many will not reply to a questionaire, particularly a rather personal one. With some kinds of mail questionaires, a five or ten per cent

[1] salubrious: heilsam, gesund, zuträglich, bekömmlich

response is quite high. This one should have done better than that, but nothing like one hundred per cent.

So we find that the income figure is based on a sample composed of all class members whose addresses are known and who replied to the questionaire. Is this a representative sample? That is, can this group be assumed to be equal in income to the unrepresented group, those who cannot be reached or who do not reply?

Who are the little lost sheep down in the Yale rolls as "address unknown"? Are they the big-income earners—the Wall Street men, the corporation directors, the manufacturing and utility executives? No; the addresses of the rich will not be hard to come by. Many of the most prosperous members of the class can be found through *Who's who in America* and other reference volumes even if they have neglected to keep in touch with the alumni office. It is good guess that the lost names are those of the men who, twenty-five years or so after becoming Yale bachelors of arts, have not fulfilled any shining promise. They are clerks, mechanics, tramps, unemployed alcoholics, barely surviving writers and artists... people of whom it would take half a dozen or more to add up to an income of $25,111. These men do not so often register at class reunions, if only because they cannot afford the trip.

Who are those who chucked the questionaire into the nearest wastebasket? We cannot be so sure about these, but it is at least a fair guess that many of them are just not making enough money to brag about. They are a little like the fellow who found a note clipped to his first pay check suggesting that he consider the amount of his salary confidential and not material for the interchange of office confidences. "Don't worry," he told the boss. "I'm just as ashamed of it as you are."

It becomes pretty clear that the sample has omitted two groups most likely to depress the average. The $25,111 figure is beginning to explain itself. If it is a true figure for anything it is merely for that special group of the class of '24 whose addresses are known and who are willing to stand up and tell how much they earn. Even that requires the assumption that the gentlemen are telling the truth.

• [Krämer 1996] schreibt zu dem Zitat aus der *ADAC-Motorwelt*:

Was sagt uns diese Nachricht? Oder: Was will uns diese Nachricht sagen?

Gemeint war offensichtlich folgendes: „Leute, schnallt Euch an! Sonst ist Eure Überlebenschance bei einem Unfall weit geringer!"

Das Dumme ist nur: Um dieses Argument zu stützen, sind diese Zahlen völlig ungeeignet; genauso könnte man daraus auch schließen, daß Gurte höchst gefährlich sind. Denn hatten sich nicht sechs von zehn tödlich verunglückten Autofahrern vorher angeschnallt...

Zehn Autofahrer sterben, sechs mit Gurt und vier ohne. Wenn wir weiter nichts von dem Verkehrsgeschehen wissen, bleibt nur diese Überlegung übrig: Zehn Autofahrer sterben, die meisten davon angeschnallt, also Hände weg von diesen Teufelsdingen...

Dieser Schluß ist falsch, wie wir alle wissen, denn worauf es hier ganz offensichtlich ankommt, ist: Wieviele Unfallbeteiligte tragen einen Sicherheitsgurt, und wieviele tragen keinen? Für beide Gruppen würde man dann gern den Anteil derer wissen, die den Unfall überleben, wobei vermutlich herauskommt, daß dieser Anteil für die Angeschnallten weitaus größer ist. Aber dazu sagt die ADAC-Statistik überhaupt nichts aus...

Wir sind hier Zeugen einer globalen Konfusion, nämlich der Neigung vieler Menschen inklusive vieler Journalisten, bei bedingten Wahrscheinlichkeiten die Bedingung und das bedingte Ereignis zu verwechseln. Zur Erinnerung: Die bedingte Wahrscheinlichkeit eines Ereignisses A, gegeben ein Ereignis B ist eingetreten, ist nichts anderes als die „normale" Wahrscheinlichkeit für A, wenn wir die Menge aller zulässigen Möglichkeiten auf B beschränken. Wenn wir beim Würfeln wissen, daß eine gerade Zahl gefallen ist, so ändert das die Wahrscheinlichkeit für „Zahl größer 3". Vorher war die Wahrscheinlichkeit $\frac{1}{2}$ — es gibt sechs

Möglichkeiten, alle gleich wahrscheinlich, davon drei größer als drei. Wenn wir aber wissen, daß eine gerade Zahl gefallen ist, so bleiben nur noch drei Möglichkeiten, nämlich 2, 4 und 6. Von diesen drei Möglichkeiten sind zwei größer als drei, also ist die bedingte Wahrscheinlichkeit von „gerade“, gegeben eine Zahl größer als drei ist gefallen, jetzt $\frac{2}{3}$.

Im Prinzip ist also das Rechnen mit bedingten Wahrscheinlichkeiten nicht besonders schwer. Wer „normale“ Wahrscheinlichkeiten berechnen kann, kann auch bedingte Wahrscheinlichkeiten berechnen. Das Problem ist nicht das Berechnen, sondern das Interpretieren. Beim Pro und Contra Sicherheitsgurte etwa ist die in der *ADAC-Motorwelt* genannte bedingte Wahrscheinlichkeit von vierzig Prozent, keinen Gurt zu tragen, gegeben ein tödlicher Unfall ist geschehen, völlig irrelevant. Sie mag korrekt sein oder nicht — das interessiert hier nicht. Interessant ist doch nur die bedingte Wahrscheinlichkeit zu sterben, gegeben wir sind angeschnallt, also das Ganze quasi umgedreht. Und die würden wir dann gerne vergleichen mit der bedingten Wahrscheinlichkeit zu sterben, gegeben wir sind *nicht* angeschnallt (mit dem voraussichtlichen Ergebnis, daß letztere beträchtlich größer ist). Aber zu diesen Wahrscheinlichkeiten sagt die eingangs zitierte Schlagzeile überhaupt nichts aus.

• [Krämer 1997] zu dem Zitat aus dem *Stern*:

[Das] stimmt [...] einfach nicht. Fußballspieler verursachen vor allem deswegen mehr Unfälle als andere, weil es mehr Fußballspieler gibt. Die größten Bruchpiloten sind eher die alpinen Skifahrer, die mit 8 Prozent der Unfälle weit mehr Schaden anrichten, als ihrem Anteil an den Sportlern der Nation entspricht.

• [Krämer 1997] zu „Im Alter wirst Du glücklicher“:

In Wahrheit ist jedoch genau das Gegenteil der Fall. Die Selbstmorde steigen mit höherem Alter an, von weniger als fünf pro hunderttausend in der Gruppe der unter 20-jährigen bis auf fast fünfzig pro hunderttausend bei den über 70-jährigen. Je älter wir werden, desto eher scheiden wir aus freien Stücken aus dem Leben, und zwar zu allen Zeiten und in allen Ländern. Daß dennoch die Selbstmorde gerade bei Jugendlichen eine solch prominente Rolle spielen, liegt allein daran, daß Jugendliche eben generell nur selten sterben. Sie haben selten Krebs und Kreislaufleiden, sie haben keine Altersschwäche und keinen Alzheimer, und auch Schlaganfälle oder Leberschäden kommen unter Jugendlichen nicht sehr häufig vor. Mit anderen Worten, in diesem Alter sind Unfall, Mord und Selbstmord fast die einzigen Todesursachen, die noch übrig bleiben, so daß der hohe Anteil an Selbstmördern unter den Verstorbenen in dieser Altersklasse nicht überrascht.

Literatur

[Huff 1954] D. Huff. *How to Lie with Statistics.* W.W. Norton, New York, NY, USA 1954

[Krämer 1996] W. Krämer. *Denkste!.* Campus-Verlag, Frankfurt, Germany 1996

[Krämer 1997] W. Krämer. *So lügt man mit Statistik (7. Auflage).* Campus-Verlag, Frankfurt, Germany 1997

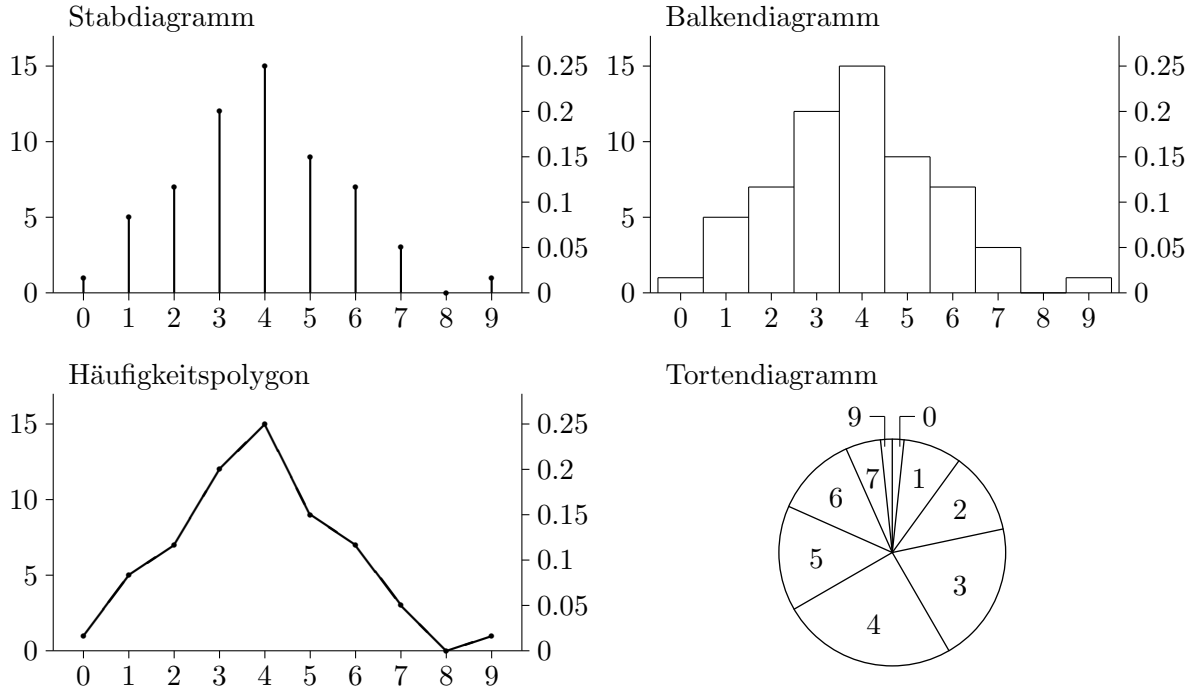
Aufgabe 2 Tabellarische und graphische Darstellung von Daten

Häufigkeitstabelle:

a_k	h_k	r_k	$\sum_{i=1}^{k} h_i$	$\sum_{i=1}^{k} r_i$
0	1	$\frac{1}{60} = 0.01\bar{6}$	1	$\frac{1}{60} = 0.01\bar{6}$
1	5	$\frac{5}{60} = 0.08\bar{3}$	6	$\frac{6}{60} = 0.1$
2	7	$\frac{7}{60} = 0.11\bar{6}$	13	$\frac{13}{60} = 0.21\bar{6}$
3	12	$\frac{12}{60} = 0.02$	25	$\frac{25}{60} = 0.41\bar{6}$
4	15	$\frac{15}{60} = 0.25$	40	$\frac{40}{60} = 0.\bar{6}$
5	9	$\frac{9}{60} = 0.15$	49	$\frac{49}{60} = 0.81\bar{6}$
6	7	$\frac{7}{60} = 0.11\bar{6}$	56	$\frac{56}{60} = 0.93\bar{3}$
7	3	$\frac{3}{60} = 0.05$	59	$\frac{59}{60} = 0.98\bar{3}$
8	0	$\frac{0}{60} = 0$	59	$\frac{59}{60} = 0.98\bar{3}$
9	1	$\frac{1}{60} = 0.01\bar{6}$	60	$\frac{60}{60} = 1$

a_k	Merkmalswerte
h_k	absolute Häufigkeiten
r_k	relative Häufigkeiten
$\sum_{i=1}^{k} h_i$	absolute Summenhäufigkeiten
$\sum_{i=1}^{k} r_i$	relative Summenhäufigkeiten

Graphische Darstellungen:



Auf Flächen- und Volumendiagramm, die weniger anschaulich sind, sei hier verzichtet. Wir weisen lediglich darauf hin, daß bei dem Flächendiagramm die Seitenlänge des Quadrates die Quadratwurzel aus der Häufigkeit, bei dem Volumendiagramm die Kantenlänge des Würfels der dritten Wurzel aus der Häufigkeit entsprechen muß, damit die Flächen- bzw. Volumenverhältnisse die Häufigkeitsverhältnisse wiedergeben.

Aufgabe 3 Irreführende graphische Darstellungen

Die Aussage „Jeder zweite lebt allein“ nimmt eine falsche Basis des Diagramms an. Das Diagramm zeigt die Häufigkeiten von *Haushalten* verschiedener Größe und nicht die Häufigkeiten von *Personen*, die in Haushalten dieser Größen leben. Rechnet man die Häufigkeiten auf Personen um (was einfach durch Gewichtung der Prozentzahlen mit der Zahl der Personen, die

in den jeweils betrachteten Haushalten leben, geschehen kann), so zeigt sich, daß nur etwa 26.3% — also etwa jeder vierte — allein lebt.

Das Diagramm ist zwar formal korrekt, kann jedoch leicht (wie durch die Süddeutsche Zeitung) fehlinterpretiert werden, da wir mit Personen als Zählgrundlage eher vertraut sind als mit der abstrakten Zähleinheit „Haushalt“. Man sollte daher stets sorgfältig prüfen, mit welcher Zählbasis man Häufigkeiten darstellt (so es denn Wahlmöglichkeiten gibt), um die Gefahr solcher Fehlinterpretationen zu mindern.

Aufgabe 4 Mittelwerte

Wenn diese Statistik richtig ist, so ist zwar das Risiko, mit einem Flugzeug auf einer k-stündigen Reise zu verunglücken, mehr als dreimal so groß wie auf einer k-stündigen Bahnfahrt zu verunglücken, doch benötigt man mit einem Flugzeug i.a. auch sehr viel weniger Zeit, um einen gegebenen Zielort zu erreichen. Man ist also der höheren Gefährdung eine kürzere Zeit lang ausgesetzt. Mit dieser Statistik ist eine Einschätzung des relativen Risikos schwierig, wenn man nichts über die relativen Reisedauern weiß.

Sinnvoller ist daher folgender Ansatz: Gesetzt den Fall, ich möchte von Stadt A nach Stadt B reisen. Ich habe die Wahl, mit der Bahn zu fahren oder ein Flugzeug zu nehmen. Mit welchem Transportmittel ist das Risiko zu verunglücken geringer? In beiden Fällen muß ich die gleiche Strecke zurücklegen. Also interessiert mich die durchschnittliche Zahl von Verkehrstoten pro zurückgelegter Längeneinheit, also z.B. pro Passagierkilometer, für die beiden Verkehrsmittel. In dem Buch „So lügt man mit Statistik“ von Walter Krämer finden sich dazu die folgenden Angaben:

Bahn:	0.9 Verkehrstote pro 1 Milliarde Passagierkilometer,
Flugzeug:	0.3 Verkehrstote pro 1 Milliarde Passagierkilometer.

Das Einschätzung hat sich gerade umgekehrt. Bei dieser Rechnung ist das Risiko, mit der Bahn zu verunglücken, dreimal größer als das Risiko, mit einem Flugzeug zu verunglücken. Ähnlich wie in Aufgabe 5, wo die Wahl der Zähleinheit eine falsche Interpretation nahelegte, beeinflußt hier die Wahl der Basisgröße stark das Risikoeinschätzung.

Aufgabe 5 Mittelwerte

Natürlich ist die in der Aufgabe angegebene Mittelung der Prozentzahlen durch Berechnung des arithmetischen Mittels Unfug, denn die Prozentzahlen beziehen sich auf verschiedene Basisgrößen, nämlich den Aktienwert am Beginn des jeweiligen Jahres. Sie können folglich nicht durch arithmetische Durchschnittsbildung gemittelt werden. Auch der Median ist hier nicht anwendbar, und zwar aus den gleichen Gründen. Um eine korrekte Mittelung durchführen zu können, müssen wir die Prozentzahlen auf Faktoren umrechnen. Es bedeutet

eine Steigerung um	60%	eine Multiplikation mit	1.6,
ein Fallen um	50%	eine Multiplikation mit	0.5,
eine Steigerung um	70%	eine Multiplikation mit	1.7,
eine Fallen um	40%	eine Multiplikation mit	0.6.

Um den durchschnittlichen Faktor zu berechnen, müssen wir das *geometrische Mittel* dieser Faktoren bilden, also

$$s = \sqrt[4]{1.6 \cdot 0.5 \cdot 1.7 \cdot 0.6} = \sqrt[4]{0.816} \approx 0.95.$$

Der Aktienwert ist folglich im Mittel nicht um 10% pro Jahr *gestiegen*, sondern vielmehr um 5% pro Jahr *gefallen*. Daß man das geometrische Mittel benötigt, um die Änderungen zu

mitteln, sieht man übrigens wie folgt: Sei x der Wert der Aktie zu Beginn des ersten Jahres. Am Ende des ersten Jahres ist ihr Wert folglich $1.6x$, am Ende des zweiten $0.5 \cdot 1.6x = 0.8x$, am Ende des dritten $1.7 \cdot 0.8x = 1.36x$ und am Ende des vierten Jahres $0.6 \cdot 1.36x = 0.816x$. Gesucht ist nun ein Faktor s, der, in jedem Jahr hinzumultipliziert, x in $0.816x$ überführt, für den also gilt:

$$s \cdot s \cdot s \cdot s \cdot x = s^4 x = 0.816x \qquad \text{oder} \qquad s = \sqrt[4]{0.816}.$$

Aufgabe 6 Statistische Kenngrößen

Modalwert	x^*	$= 4$
Median	$\tilde{x}$	$= 4$
$\frac{1}{3}$-Quantil		$= 3$
Mittelwert	$\bar{x}$	$= \frac{58}{15} = 3.8\overline{6}$
Spannweite	R	$= 9 - 0 = 9$
Interquartilbereich	$Q_3 - Q_1$	$= 5 - 3 = 2$
Varianz	s^2	$= \frac{2834}{885} \approx 3.20$
Standardabweichung	s	$= \sqrt{\frac{2834}{885}} \approx 1.79$
Schiefe	α_3	≈ -0.23
Wölbung	α_4	≈ 2.92
bzw.	α_4'	≈ -0.08

Kastendiagramm (Box-Plot):