978-3-7910-3393-8 Schuldenzucker, Prüfungstraining Deskriptive Statistik

SCHÄFFER
POESCHEL

Einleitung

Die deskriptive Statistik dient der systematischen Erfassung und Darstellung von Daten, die bestimmte Zustände oder Entwicklungen aufzeigen. Sehr viele Entscheidungen des Alltags, in Wirtschaftsunternehmen oder etwa bei der Entwicklung von Medikamenten basieren auf der Erhebung von Daten: Kommt die Straßenbahn, die jemand benötigt, häufig zu spät, muss er genügend Reserve einplanen, um pünktlich zu sein. Führt der DVBT-Empfang beim Fernsehen häufig zu Störungen, wird man überlegen, ob man die Verstärkung des Signals erhöht oder auf Satellitenempfang umstellt. Zur Beurteilung eines Investitionsprojekts werden Daten erhoben, um zu klären, welche Rückflüsse zu welchen Zeitpunkten zu erwarten sind. Für die Fortentwicklung eines Autos wird man Daten bezüglich des technischen Stands vergleichbarer Autos und bezüglich der Kundenwünsche erheben. Ein neu entwickeltes Medikament muss auf Wirksamkeit und Verträglichkeit überprüft werden; hat es häufige oder starke Nebenwirkungen, wird man weiter entwickeln müssen.

Die Erfassung von Daten erfordert zunächst einige Vorbereitungen, um sicherzustellen, dass mit Hilfe der Daten die gewünschten Ziele tatsächlich erreicht werden können. Nach der Erfassung werden die Daten tabellarisch oder grafisch aufgearbeitet, um einen ersten Überblick zu erhalten. Anschließend wird eine Analyse der Daten durchgeführt, aussagekräftige Parameter werden ermittelt und Schlussfolgerungen gezogen, soweit dies möglich ist. Diese Ergebnisse werden in geeigneter Form dargestellt. Häufig dient das Erheben und Analysieren von Daten der Vorbereitung und Absicherung von Entscheidungen, sodass der Bezug der Erkenntnisse aus den Daten zu diesen Entscheidungen herausgearbeitet werden muss.

Bei der Erhebung eines vollständigen Datensatzes erfasst man alle in einem Bereich existierenden Werte. Häufig werden allerdings Daten als Stichprobe einer größeren Grundgesamtheit erhoben, da es nicht möglich ist, alle interessierenden Werte zu erfassen. In diesem Fall werden manche Parameter gegenüber einem vollständigen Datensatz leicht abgewandelt, um der Unsicherheit bezüglich der nicht erhobenen Werte Rechnung zu tragen. Beide Sichtweisen werden im vorliegenden Buch sorgfältig besprochen.

Methoden der deskriptiven Statistik reichen über eine Vielfalt von Feldern. Die Methoden, die hier angesprochen werden, beginnen nach einer einführenden Darstellung der Klassifizierung von Merkmalen mit der Behandlung eindimensionaler Datenreihen, sodass zum Beispiel charakteristische Daten eines Betriebs zusammengestellt werden können. Zusammenhänge zwischen zwei Merkmalen werden untersucht. Konzentrationseffekte wie etwa Einkommenskonzentration können grafisch wie rechnerisch erfasst werden. Parameter zur Messung von Inflation werden vorgestellt.

1 Einführung und Grundbegriffe

1.1 Begriff Statistik

Die Statistik befasst sich mit dem Sammeln und Aufbereiten von Wissen über bestimmte interessierende Merkmale. Solche Merkmale können von sehr unterschiedlicher Natur sein. Zum Beispiel die Beschreibungen der Merkmale »Haarfarbe«, »Vorliebe bei Mineralwasser«, »Körpergröße«, »Alter in Jahren« unterscheiden sich sehr stark: Haarfarben werden nur durch Worte beschrieben; Vorlieben werden ebenfalls nur durch Worte beschrieben, enthalten aber eine Präferenz. Die Körpergröße kann jeden Wert innerhalb eines bestimmten Intervalls annehmen, während beim Alter in Jahren als Werte nur natürliche Zahlen in Frage kommen. Daten, die etwa in einem Betrieb erhoben werden, weisen diese Vielfalt auf, denn sie reichen von Geschlecht und Familienstand des Personals bis hin zum Materialverbrauch oder zur Qualität einer Ware.

Die deskriptive Statistik befasst sich mit der reinen Erhebung und Analyse von Daten. Ein Datensatz wird beschrieben, tabellarisch oder grafisch dargestellt und es werden Kenngrößen ermittelt.

In der induktiven Statistik nutzt man einen Datensatz als Stichprobe; das Ziel ist, aufgrund von erhobenen Daten Schlussfolgerungen über Grundgesamtheiten zu ziehen, die größer sind als der Datensatz.

1.2 Statistische Einheiten und deren Merkmale

Eine in der deskriptiven Statistik interessierende Größe wird ein *Merkmal* genannt. Die möglichen Ergebnisse beim Erheben von Daten zu diesem Merkmal heißen *Merkmalsausprägungen*; sie werden an sogenannten *Merkmalsträgern* oder *statistischen Einheiten* erhoben.

Etwa beim Merkmal »Geschlecht« werden die Ausprägungen »männlich« und »weiblich« beobachtet, beim Merkmal »Körpergröße« liegen bei Erwachsenen die Ausprägungen meist zwischen 1.00 m und 2.20 m.

Man unterscheidet zwischen unterschiedlichen *Merkmalstypen*:

1. *Qualitative* Merkmale sind solche, die nur durch Worte beschrieben werden können.
   Diese werden weiter unterschieden:

   (a) *Nominale* Merkmale sind solche ohne natürliche Rangordnung wie etwa Haarfarbe, Beruf oder Familienstand.

   (b) *Ordinal* nennt man qualitative Merkmale, die eine natürliche Rangordnung aufweisen. Hier sind zum Beispiel Güteklassen bei Lebensmitteln, Tabellenplätze einer Fußballiga oder Noten zu nennen.

2. *Quantitative* (kategoriale) Merkmale sind solche, deren Ausprägungen durch Zahlen beschrieben werden können. Sie können natürlich insbesondere der Größe nach geordnet werden.
   Unter den quantitativen Merkmalen gibt es folgende Unterscheidungen:
   (a) *Diskrete* Merkmale besitzen nur endlich viele oder abzählbar viele verschiedene Ausprägungen. Hier sind Anzahlen oder Häufigkeiten oder etwa Stunden pro Tag zu nennen.
   (b) *Stetige* Merkmale können Ausprägungen annehmen, die ein ganzen Intervall ausfüllen. Etwa Gewicht oder Körpergröße zählen dazu.

1.3 Messbarkeitseigenschaften

Der Typ eines Merkmals legt fest, auf welcher Skala es gemessen werden kann:

A) Eine *Nominalskala* beschreibt nur die Verschiedenheit der Ausprägungen.
B) Eine *Ordinalskala* bringt Merkmalsausprägungen in eine Rangordnung.
C) Eine *Metrische Skala* (*Kardinalskala*) ermöglicht rechnerische Vergleiche zwischen Merkmalsausprägungen und deren Interpretation.
   – Bei einer *Intervallskala* sind Abstände sinnvoll; zum Beispiel Temperatur [°C], Breiten- und Längengrade fallen hierunter.
   – Im Fall einer *Verhältnisskala* existiert darüber hinaus ein absoluter Nullpunkt, sodass die Ausprägungen zueinander ins Verhältnis gesetzt werden können: Beispielsweise bei der Temperatur [°K], bei Längen, Gewichten oder Einkommen ist es möglich, etwa von Verdoppelung zu sprechen.
   – Eine *Absolutskala* ist eine Verhältnisskala, die über eine natürlich gegebene Maßeinheit (»Stück«) verfügt, zum Beispiel die Zahl der Studierenden an einer Hochschule.

1.4 Rezeptartige Lösungswege

Aufgabe: Merkmalstypen und Skalen erkennen
Gegeben: Unterschiedliche Merkmale
Gesucht: Zugehörige Merkmalstypen und Skalen
Lösungsweg:
Ein Merkmal ist qualitativ, wenn die Ausprägungen nur durch Worte beschrieben werden können.
Dann ist es nominal, wenn seine Ausprägungen keine natürliche Rangfolge haben.
Zugehörige Skala: Nominalskala
Es ist ordinal, wenn seine Ausprägungen eine natürliche Rangfolge haben.
Zugehörige Skala: Ordinalskala
Ein Merkmal ist quantitativ, wenn die Ausprägungen durch Zahlen beschrieben werden können.
Zugehörige Skala: Metrische Skala
Dann ist es diskret, wenn es endlich viele oder höchstens abzählbar viele Ausprägungen gibt.
Es ist stetig, wenn die Ausprägungen ein ganzes Intervall füllen.
Zugehörige Skalen:
Intervallskala, falls kein absoluter Nullpunkt existiert
Verhältnisskala, falls ein absoluter Nullpunkt existiert; dann ist es auch möglich, etwa davon zu sprechen, dass eine Ausprägung doppelt so groß ist wie eine andere.
Absolutskala, falls zusätzlich eine natürliche Maßeinheit vorgegeben ist

s. Aufgabe 1.1, S. 6

1.5 Übungsaufgaben

Merkmalstypen und Skalen

Aufgabe 1.1
Gegeben sind die Merkmale
- Beruf
- Leistungsbeurteilung
- Kinderzahl
- Temperatur in °C
- Länge

(a) Geben Sie an, ob diese Merkmale qualitativ oder quantitativ sind.
(b) Bei qualitativen Merkmalen bestimmen Sie, ob sie nominal oder ordinal sind. Bei quantitativen Merkmalen geben Sie an, ob sie diskret oder stetig sind.
(c) Geben Sie die zugehörige Skala an.

Aufgabe 1.2
(a) Beschreiben Sie, was qualitative von quantitativen Merkmalen unterscheidet.
(b) Welche Eigenschaft eines qualitativen Merkmals bewirkt, dass es ordinal ist?
(c) Was unterscheidet bei quantitativen Merkmalen diskrete von stetigen?
(d) Was zeichnet ein Merkmal aus, das man auf einer Absolutskala misst?

1.6 Lösungen

Merkmalstypen und Skalen

Lösung 1.1

- Beruf: Qualitativ, nominal. Nominalskala
- Leistungsbeurteilung: Qualitativ, ordinal. Ordinalskala
- Kinderzahl: Quantitativ, diskret. Metrisch, Absolutskala
- Temperatur in $°C$: Quantitativ, stetig. Metrisch, Intervallskala. In der Regel wird das Merkmal diskretisiert.
- Länge: Quantitativ, stetig. Metrisch, Verhältnisskala

Lösung 1.2

(a) Qualitative Merkmale können nur durch Worte beschrieben werden, quantitative Merkmale werden durch Zahlen beschrieben.

(b) Bei einem ordinalen Merkmal können die Ausprägungen angeordnet werden, es gibt eine Hierarchie.

(c) Ein diskretes Merkmal besitzt Ausprägungen, die durchnummeriert werden können. Bei einem stetigen Merkmal füllen die Ausprägungen ein ganzes Intervall.

(d) Ein Merkmal wird auf einer Absolutskala gemessen, wenn es quantitativ ist, einen absoluten Nullpunkt besitzt und es nur eine natürliche Maßeinheit gibt.

2 Eindimensionale Datenreihen

2.1 Datensatz/Stichprobe

2.1.1 Absolute und relative Häufigkeiten, empirische Verteilungsfunktion

Beim Erfassen von Daten wird die Anzahl der Daten mit n abgekürzt. Für den Fall, dass die Daten als Stichprobe einer größeren Grundgesamtheit dienen, heißt diese Anzahl auch Stichprobenlänge. Daten werden mit dem Buchstaben x bezeichnet, der Datensatz wird als n-Tupel der Messwerte durchnummeriert in der Form

$$(x_1, \ldots, x_n).$$

Die möglichen oder gemessenen verschiedenen Merkmalsausprägungen werden bezeichnet als

$$a_1, \ldots, a_m.$$

Bemerkung:
Stellvertretend für die Indizes, also die Zahlen, die die Messwerte oder Ausprägungen durchnummerieren, wählt man einen Buchstaben. Häufig ist dieser allgemeine Index i oder j oder k.

Zur allgemeinen Beschreibung für die durchnummerierten Ausprägungen $a_1, \ldots, a_m$ wird ein *Laufindex* j benutzt. Die Daten $x_1, \ldots, x_n$ werden mit dem Index i nummeriert.

Die absolute Häufigkeit, mit der eine Merkmalsausprägung a_j im Datensatz vorkommt, ist

$$h_j = h_n(a_j).$$

Die relative Häufigkeit, also der Anteil, zu dem eine Merkmalsausprägung a_j im Datensatz vorkommt, ist

$$r_j = r_n(a_j) = \frac{h_j}{n}.$$

Die Summe der absoluten Häufigkeiten ist $\sum_{j=1}^{m} h_j = n$; die relativen Häufigkeiten summieren sich zu $\sum_{j=1}^{n} r_j = 1$. Hierbei wird das große griechische Sigma Σ als Symbol für das Aufsummieren verwendet. Unten am $\sum$ notiert man, bei welchem Index die Summation beginnt, oben notiert man den größten verwendeten Index.
Als absolute Häufigkeitsverteilung bezeichnet man die Zusammenstellung der Paare (a_j, h_j) der Ausprägungen und ihrer absoluten Häufigkeiten. Die relative Häufigkeitsverteilung besteht aus den Paaren (a_j, r_j). Beide Verteilungen können in Form von Tabellen dargestellt werden.

Beispiel:
Sie möchten einen Eindruck von der Lohnverteilung in einem Niedriglohnsektor von 1500 bis 2000 Euro bekommen, um substantiiert über Mindestlöhne diskutieren zu können.
Vorbereitende Überlegungen:
– Sie identifizieren Ihre Ziele: Sie möchten den Niedriglohnsektor kennenlernen.
– Die interessierende Gesamtheit besteht aus allen Beschäftigten in diesem Sektor.
– Erhebbar ist nur eine Stichprobe.
– Das interessierende Merkmal ist das Monatseinkommen.
– Dieses Merkmal ist quantitativ und diskret.
– Die Skala ist metrisch.

Erfassen von Daten:
Es wurden die Monatseinkommen von $n = 50$ Personen erhoben:

$x_1 = 1600$	$x_2 = 1900$	$x_3 = 1800$	$x_4 = 1950$	$x_5 = 1850$
$x_6 = 1600$	$x_7 = 2000$	$x_8 = 1950$	$x_9 = 2000$	$x_{10} = 1900$
$x_{11} = 1950$	$x_{12} = 1900$	$x_{13} = 1800$	$x_{14} = 1950$	$x_{15} = 1950$
$x_{16} = 1850$	$x_{17} = 1850$	$x_{18} = 1950$	$x_{19} = 2000$	$x_{20} = 1950$
$x_{21} = 1900$	$x_{22} = 1900$	$x_{23} = 1850$	$x_{24} = 2000$	$x_{25} = 1800$
$x_{26} = 1900$	$x_{27} = 1850$	$x_{28} = 1600$	$x_{29} = 1500$	$x_{30} = 1900$
$x_{31} = 1850$	$x_{32} = 1800$	$x_{33} = 1850$	$x_{34} = 1950$	$x_{35} = 1900$
$x_{36} = 1800$	$x_{37} = 1850$	$x_{38} = 1750$	$x_{39} = 2000$	$x_{40} = 1800$
$x_{41} = 1850$	$x_{42} = 1900$	$x_{43} = 1850$	$x_{44} = 1950$	$x_{45} = 1600$
$x_{46} = 1500$	$x_{47} = 1850$	$x_{48} = 1800$	$x_{49} = 1950$	$x_{50} = 1650$

Aus den Daten erstellt man zunächst eine Strichliste:

Monatliches Einkommen	Absolute Häufigkeit
$a_1 = 1500$	||
$a_2 = 1550$	
$a_3 = 1600$	||||
$a_4 = 1650$	|
$a_5 = 1700$	
$a_6 = 1750$	|
$a_7 = 1800$	𝍸 ||
$a_8 = 1850$	𝍸 𝍸 |
$a_9 = 1900$	𝍸 ||||
$a_{10} = 1950$	𝍸 𝍸
$a_{11} = 2000$	𝍸

Datenaufbereitung:

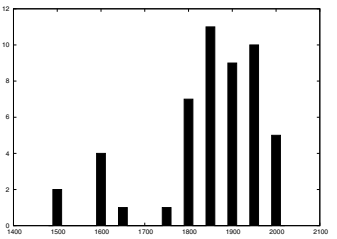
(a) Tabellarische Darstellung:

Monatliches Einkommen	Absolute Häufigkeit	Relative Häufigkeit
$a_1 = 1500$	$h_1 = 2$	$r_1 = 0.04$
$a_2 = 1550$	$h_2 = 0$	$r_2 = 0.00$
$a_3 = 1600$	$h_3 = 4$	$r_2 = 0.08$
$a_4 = 1650$	$h_4 = 1$	$r_4 = 0.02$
$a_5 = 1700$	$h_5 = 0$	$r_4 = 0.00$
$a_6 = 1750$	$h_6 = 1$	$r_4 = 0.02$
$a_7 = 1800$	$h_7 = 7$	$r_7 = 0.14$
$a_8 = 1850$	$h_8 = 11$	$r_8 = 0.22$
$a_9 = 1900$	$h_9 = 9$	$r_9 = 0.18$
$a_{10} = 1950$	$h_{10} = 10$	$r_{10} = 0.20$
$a_{11} = 2000$	$h_{11} = 5$	$r_{10} = 0.10$

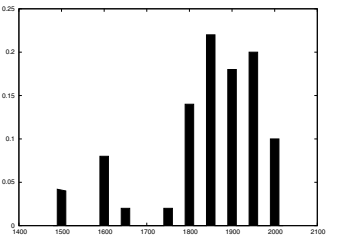
Bei der tabellarischen Darstellung werden in der ersten Spalte die *verschiedenen* Merkmalsausprägungen aufgelistet.

(b) Grafische Darstellung durch Balkendiagramme:

Auf der waagerechten Achse (Abszisse) sind die Merkmalsausprägungen aufgetragen, auf der senrechten (Ordinate) die Häufigkeiten.



Absolute Häufigkeiten



Relative Häufigkeiten

Zur Beantwortung von Fragen zu Anzahlen oder Anteilen der Daten mit Messwerten bis zu einer gewissen Grenze oder ab einer gewissen Grenze erweitert man die Tabelle um die absolute Summenhäufigkeit $H_n(a_j)$ und die relative Summenhäufigkeit $F_n(a_j)$, die auch *empirische Verteilungsfunktion* genannt wird. Diese beiden Funktionen entstehen durch sukzessives Addieren der absoluten beziehungsweise relativen Häufigkeiten. Diese kumulierten Häufigkeiten können in jeder beliebigen

Zahl x ausgewertet werden, denn sie spiegeln die Anzahl beziehungsweise den Anteil der Daten wider, die kleinergleich x sind:

$$H_n(x) = \sum_{j \text{ mit } a_j \le x} h_j$$
$$F_n(x) = \sum_{j \text{ mit } a_j \le x} r_j = \frac{1}{n} \sum_{j \text{ mit } a_j \le x} h_j = \frac{1}{n} H_n(x)$$

Am Beispiel:

(a) Bestimmen Sie die Anzahl der Befragten mit einem Monatskeinkommen von höchstens 1800 Euro.
(b) Berechnen Sie den Anteil der Befragten mit einem Monatseinkommen von höchstens 1700 Euro.
(c) Errechnen Sie den Prozentsatz der Befragten mit einem Monatseinkommen von mindestens 1750 Euro.

Lösung:
Erweiterung der Tabelle um die absolute Summenhäufigkeit und die empirische Verteilungsfunktion:

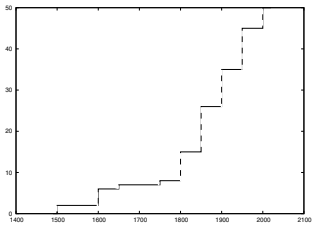
a_j	$h(a_j)$	$H_n(a_j)$	$r(a_j)$	$F_n(a_j)$
1500	2	2	0.04	0.04
1550	0	2	0.00	0.04
1600	4	6	0.08	0.12
1650	1	7	0.02	0.14
1700	0	7	0.00	0.14
1750	1	8	0.02	0.16
1800	7	15	0.14	0.30
1850	11	26	0.22	0.52
1900	9	35	0.18	0.70
1950	10	45	0.20	0.90
2000	5	50	0.10	1.00

(a) 15 Befragte habe ein Monatseinkommen von höchstens 1800 Euro.
(b) Der Anteil der Befragten mit einem Monatseinkommen von höchstens 1700 Euro liegt bei 14 %.
(c) Der Prozentsatz der Befragten mit einem Monatseinkommen von mindestens 1750 ist

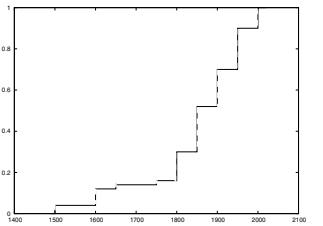
$$(100 - 0.14 \cdot 100)\% = 86\,\%$$

Die Graphen von absoluter Summenhäufigkeitsfunktion und empirischer Verteilungsfunktion sind monoton steigende Treppenfunktionen.

Am Beispiel:



Absolute Summenhäufigkeitsfunktion



Empirische Verteilungsfunktion

2.1.2 Klasseneinteilung

Bei manchen Datensätzen ist es günstig, die Messwerte in Klassen zusammenzufassen. Das wird etwa bei einem stetigen Merkmal wegen der Messungenauigkeit nötig sein, aber auch bei diskreten Merkmalen können Klassen die Daten übersichtlicher darstellen.

Beispiele:

- Einkommensklassen
- Mietspiegel (abhängig von der Wohnungsgröße)
- Unfallstatistik, etwa bezüglich der Zeit bis zum ersten Unfall nach dem Führerscheinerwerb

Bei der Erstellung von Klassen muss immer festgelegt werden, zu welchen Klassen die Klassengrenzen gehören. Häufig werden halboffene, nach oben geschlossene Klassen gewählt. Alternativ sind auch halboffene, nach unten geschlossene Klassen möglich. Die unterste Klasse wird meist nach unten geschlossen gewählt, die oberste nach oben geschlossen.

Bei der Darstellung von Klassen ist es üblich, dass ein Endpunkt einer Klasse zur Klasse gehört, wenn die zugehörige Klammer sich der Klasse zuwendet: $[a, b]$. Ein Endpunkt gehört nicht zur Klasse, wenn die zugehörige Klammer sich von der Klasse wegwendet. Halboffene nach oben geschlossene Klassen sind also von der Gestalt $]a, b]$; halboffene nach unten geschlossene Klassen sind von der Gestalt $[a, b[$.

Beispiele:
- Einkommen: $[0, 500],]500, 1000],]1000, 2000], \ldots$ sind halboffene, nach oben geschlossene Klassen.
  Hier gehören also die Zahlen 0 und 500 zur ersten Klasse, die Zahl 1000 gehört zur zweiten Klasse, die Zahl 2000 zur dritten. Zu beachten ist, dass etwa in der zweiten Klasse jede Zahl liegt, die größer als 500 und kleinergleich 1000 ist; so gehört die Zahl 500.00000001 zur zweiten Klasse.
- Wohnungsgröße: $[10, 30],]30, 50],]50, 70],]70, 100],]100, 150],]150, 300]$ sind ebenfalls halboffene, nach oben geschlossene Klassen.
- Körpergröße auf 2 cm genau, mit halboffenen, nach unten geschlossenen Klassen: $[160, 162[, [162, 164[, [164, 166[, \ldots$

Beispiel:
Bei einer Kontrolle in einer Tempo-30-Zone wurden folgende Geschwindigkeitsüberschreitungen festgestellt:

Geschwindigkeit:	31	33	35	37	39	41	43	45	47	49	55
Häufigkeit:	2	5	3	8	3	1	9	3	3	2	1

Die Höhe des Verwarngeldes ist an Grenzen gebunden:

Geschwindigkeitsklasse:	$]30, 35]$	$]35, 50]$	$]50, 70]$	$\ldots$
Verwarngeld:	10 Euro	20 Euro	50 Euro	$\ldots$

(a) Auf welchen Betrag belaufen sich die Einnahmen der Stadt?
(b) Wie viele Fahrer zahlten das geringste Bußgeld?
(c) Welcher Prozentsatz der Fahrer fuhr höchstens 50 km/h?
(d) Welcher Anteil der Fahrer fuhr mehr als 35 km/h?

Lösung:

Klasse	Verwarngeld	Absolute Häufigkeit	Relative Häufigkeit	Absolute Summenhäufigkeit	Empirische Verteilungsfunktion
$]30, 35]$	10	10	0.25	10	0.25
$]35, 50]$	20	29	0.725	39	0.975
$]50, 70]$	50	1	0.025	40	1

(a) Einnahmen der Stadt:
$10 \cdot 10 + 29 \cdot 20 + 1 \cdot 50 = 730$ Euro
(b) Das geringste Bußgeld bezahlen 10 Fahrer.
(c) Bis einschließlich 50 km/h fuhren 39 Fahrer, entsprechend 97.5 %.
(d) Der Anteil der Fahrer, die mehr als 35 km/h fuhren, liegt bei
$1 - 0.25 = 0.75$.