Aufgabensammlung zur deskriptiven Statistik

Mit ausführlichen Lösungen und Erläuterungen

Von
Univ.-Prof. Dr. Martin Missong

7., durchgesehene Auflage

R. Oldenbourg Verlag München Wien

Bibliografische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
www.oldenbourg.de

Gedruckt auf säure- und chlorfreiem Papier
Gesamtherstellung: AZ Druck und Datentechnik GmbH, Kempten

ISBN 3-486-57845-6

Inhaltsverzeichnis

Vorwort

Die vorliegende Aufgabensammlung ist gedacht als ideale Ergänzung zum deskriptiven Teil in

W. SCHNEIDER, J. KORNRUMPF, W. MOHR:
"Statistische Methodenlehre - Definitions- und Formelsammlung
zur deskriptiven und induktiven Statistik mit Erläuterungen",
München, Wien: Oldenbourg Verlag, 1993.

Diese Definitions- und Formelsammlung entwickelt in Definitionen, Bemerkungen und Sätzen ein formales Gerüst der statistischen Methodenlehre. Ergänzend dazu soll die Aufgabensammlung Anwendungsbeispiele liefern, Lösungswege aufzeigen, den Gehalt der dabei verwendeten Formeln verständlich machen und an gegebener Stelle auf die Gefahr möglicher Fehlschlüsse hinweisen. Besonders ausführlich werden dabei jene Analysemethoden behandelt, deren korrekte Handhabung erfahrungsgemäß einem großen Teil der Studierenden in den ersten Semestern schwerfällt. Da die grundlegenden Formeln und Definitionen bei den Lösungsvorschlägen zumeist kurz wiederholt und erklärt werden und viele Aufgaben Beispielcharakter haben, eignet sich diese Aufgabensammlung durchaus auch zum Selbststudium und als Begleitmaterial zu anderen Lehrbüchern zur beschreibenden Statistik.

Zur Beantwortung der einzelnen Aufgaben ist eine Kenntnis der vorhergehenden Aufgaben und Lösungen eventuell hilfreich aber nicht unbedingt notwendig. Dadurch sollen die Lösungswege auch bei einer selektiven Bearbeitung einzelner Aufgaben oder einzelner Kapitel problemlos nachvollziehbar bleiben. Diese Konzeption bedingt notwendigerweise Wiederholungen in den Lösungsvorschlägen zu verschiedenen Aufgaben.

Der Aufgabensammlung vorangestellt ist eine Tabelle mit Kurzbeschreibungen der in den jeweiligen Aufgaben angesprochenen Bereiche der Methodenlehre der beschreibenden Statistik. Der Anhang enthält eine knappe Zusammenstellung der verwendeten Symbolik und der grundlegenden Rechenregeln, deren Kenntnis zur Bearbeitung der Aufgaben vonnöten ist. Sämtliche in eckige Klammern eingeschlossenen Referenzen in den Aufgaben und Lösungen beziehen sich auf die Definitionen, Bemerkungen und Sätze in der obengenannten Definitions- und Formelsammlung.

Vorläufer des vorliegenden Bandes ist eine Aufgabensammlung zur deskriptiven Statistik, die in verschiedenen Versionen in den Jahren 1979 bis 1991 von den Mitarbeitern des Instituts für Statistik und Ökonometrie an der Christian-Albrechts-Universität zu Kiel zusammengestellt wurde und als Übungsmaterial in der statistischen Grundausbildung diente. Die stetige Entwicklung der Definitions- und Formelsammlung sowie der von studentischer Seite immer häufiger geäußerte Wunsch nach aktuelleren Anwendungsbeispielen machten eine grundlegende Überarbeitung und weitgehende Neugestaltung dieser Aufgabensammlung erforderlich. Dennoch konnten einige (im Text besonders gekennzeichnete) Aufgaben in überarbeiteter Form übernommen werden. Eine erste Version der vorliegenden Aufgabensammlung wurde während des Sommersemesters 1992 in den Tutorien und Übungen zur deskriptiven Statistik an der Universität Kiel und der Fachhochschule Flensburg erprobt.

Ich möchte mich bedanken bei Herrn Prof. Dr. H. Lütkepohl und Herrn Prof. Dr. G. Hansen, die mein Interesse an der beschreibenden Statistik weckten und förderten, sowie bei meinen Kollegen, den Tutoren und Studenten, die mich auf Fehler in der Vorabversion dieser Aufgabensammlung aufmerksam gemacht haben. Über Hinweise auf die sicherlich noch verbliebenen Fehler würde ich mich sehr freuen.
Mein besonderer Dank gilt Herrn Dr. Ingo Klein für die Beratung in allen fachlichen Fragen und Herrn Diplom–Wirtschaftsingenieur Michael Wagner für seine Hilfe in sämtlichen EDV–Angelegenheiten. Beider geduldige Unterstützung hat wesentlich zur Entstehung dieser Aufgabensammlung beigetragen.

Martin Missong

Vorwort zur vierten Auflage

Für die vierte Auflage wurde die Aufgabensammlung völlig überarbeitet. Die Änderungen lassen sich in drei Kategorien zusammenfassen:

Einige Aufgaben, die auf fiktiven Datensätzen aufbauten, wurden durch Übungsaufgaben zur gleichen Thematik ersetzt, in denen die Verteilung "realer Daten" analysiert wird. Andere Aufgaben wurden gestrichen, neue Aufgaben sind hinzugekommen.

Die Nummern der Sätze, Bemerkungen und Definitionen in der "Statistischen Methodenlehre" von Schneider, Kornrumpf und Mohr, auf die in den Aufgaben jeweils Bezug genommen wird, sind nicht mehr explizit angegeben. Statt dessen sind alle verwendeten Begriffe, die in der "Statistischen Methodenlehre" definiert sind, in den Übungsaufgaben jetzt in Fettdruck gesetzt. Dadurch sollen Mißverständnisse vermieden werden, die auftreten könnten, wenn sich in erweiterten Neuauflagen der Formelsammlung von Schneider, Kornrumpf und Mohr Änderungen in der Numerierung der Sätze, Bemerkungen und Definitionen ergeben.

Schließlich wurden in den früheren Auflagen vorhandene Fehler und Ungenauigkeiten korrigiert, und ich hoffe, daß sich nicht allzu viele neue Fehler oder Unklarheiten eingeschlichen haben.

Martin Missong

Vorwort zur siebten Auflage

Die Vorauflagen waren derart rasch vergriffen, dass ich mich darauf beschränken konnte, den gesamten Text kritisch durchzusehen.

Martin Missong

Aufgabe	Kurzbeschreibung
Kapitel 1: Grundbegriffe der beschreibenden Statistik	
1.1	Massen, Merkmale
1.2	Merkmalsarten, Skalierung
1.3	Merkmalsausprägungen und Mengen
Kapitel 2: Datenorganisation für qualitative Merkmale	
2.1	Ein Merkmal, insbesondere Entropie
2.2	Zwei Merkmale, insbesondere bedingte Verteilungen
2.3	Zwei Merkmale, insbesondere Transinformation
2.4	Zwei Merkmale, insbesondere Randverteilungen
2.5	Zwei Merkmale, bedingte relative Häufigkeiten
2.6	Drei Merkmale, insbesondere bedingte Verteilungen
Kapitel 3: Datenorganisation für komparative Merkmale	
3.1	Ein Merkmal, insbesondere Rangfunktionen
3.2	Ein Merkmal, insbesondere Streuungsmessung, Summenhäufigkeitsentropie
3.3	Zwei Merkmale, insbesondere Assoziationsmessung
3.4	Zwei Merkmale, insbesondere bedingte (kumulierte) Häufigkeiten, Assoziationsmessung anhand konkordanter und diskordanter Paare
3.5	Zwei Merkmale, insbesondere Streuungs- und Assoziationsmessung
3.6	Zwei Merkmale, insbesondere zeitlicher Vergleich von Verteilungen
3.7	Drei Merkmale, Simpsons Paradoxon
Kapitel 4: Datenorganisation für quantitative Merkmale	
4.1	Ein Merkmal, Einzeldaten, Verteilungsfunktion, Variablentransformation
4.2	Ein Merkmal, Einzeldaten, insbesondere Varianz, Varianzzerlegung
4.3	Ein Merkmal, klassierte Daten, insbesondere graphische Darstellung
4.4	Ein Merkmal, Einzeldaten, Mittelwertberechnung, insbesondere geometrisches Mittel
4.5	Ein Merkmal, Einzeldaten, Mittelwertberechnung, insbesondere harmonisches Mittel
4.6	Ein Merkmal, Einzeldaten, Lorenzkurve und Konzentrationsmessung
4.7	Ein Merkmal, klassierte Daten, insbesondere approximierende Anteilsfunktion, Lorenzkurve und Konzentrationsmessung
4.8	Ein Merkmal, klassierte Daten, Konzentrationsmessung, Gegenüberstellung gegebener und approximierter Klassenmerkmalssummen
4.9	Ein Merkmal, klassierte Daten, Lorenzkurve und Konzentrationsmessung
4.10	Zwei Merkmale, Einzeldaten, Abhängigkeitsmessung, Datenkompensierung
4.11	Zwei Merkmale, klassierte Daten, gemeinsame und bedingte Verteilungen, Variablentransformation
Kapitel 5: Regressionsrechnung	
5.1	Zwei Merkmale, klassierte Daten, verschiedene Regressionsverfahren, Regressionskurven
5.2	Zwei Merkmale, klassierte Daten, verschiedene Regressionsverfahren, Regressionskurven

Aufgabe	Kurzbeschreibung
5.3	Zwei Merkmale, Einzeldaten, Kleinstquadrat–Regression
5.4	Theorie der Kleinstquadrat–Methode
5.5	Zwei Merkmale, Einzeldaten, Kleinstquadrat–Regression, Umkehrregression
5.6	Zwei Merkmale, Einzeldaten, Kleinstquadrat–Regression, Einfluß einzelner Beobachtungen
5.7	Drei Merkmale, Einzeldaten, trivariate Kleinstquadrat–Regression, Fehlspezifikation linearer Modelle
5.8	Zwei Merkmale, Einzeldaten, Schätzung einer quadratischen Regressionskurve, Variablentransformation
5.9	Drei Merkmale, Einzeldaten, Schätzung einer Cobb–Douglas–Funktion
Kapitel 6: Deskriptive Zeitreihenanalyse	
6.1	Trendschätzung, insbesondere logistischer Trend
6.2	Trendschätzung, insbesondere exponentieller Trend
6.3	Trendschätzung, insbesondere quadratischer Trend
6.4	Exponentielle Glättung, gleitende Durchschnitte
6.5	Additives Komponentenmodell, gleitende Durchschnitte, Saisonnormale
6.6	Additives Komponentenmodell, linearer Trend, gleitende Durchschnitte, Saisonnormale
6.7	Saisonschätzung, Saisonnormale
6.8	Schätzung eines autoregressiven Modells zweiter Ordnung
Kapitel 7: Meß– und Indexzahlen	
7.1	Kategorisierung von Verhältniszahlen
7.2	Zeitliche Meßzahlen, Wachstumsraten
7.3	Verkettung von Meßzahlen, Umbasierung von Zeitreihen
7.4	Gewichtung von Meßzahlen, Paasche– und Laspeyres–Indizes
7.5	Paasche– und Laspeyres–Indizes
7.6	Ländervergleich von Lebenshaltungskosten (Kaufkraftparität)
7.7	Paasche– und Laspeyres–Indizes, insbesondere Gewichtsform
7.8	Paasche– und Laspeyres–Indizes

Kapitel 1

Grundbegriffe der deskriptiven Statistik

1.1 Aufgaben

Aufgabe 1.1:[1]

Im Januar 1996 gibt das Bundesministerium für Wirtschaft eine Umfrage in Auftrag, die Aufschluß über die wirtschaftliche Lage der deutschen Unternehmen im produzierenden Gewerbe geben soll. Im Rahmen dieser Erhebung werden 300 Unternehmen des produzierenden Gewerbes ausgewählt und befragt. Dabei werden u.a. unternehmensspezifische Daten wie Rechtsform, Bilanzsumme, Anzahl der Beschäftigten etc. erhoben. Die Unternehmen sollen auch ihre derzeitige Ertragslage beurteilen, ihre Beschäftigtenzahl am Jahresende prognostizieren und angeben, welche Wachstumsrate sie für ihre Produktion im Laufe des Jahres erwarten.

a) Wie lautet die Grundgesamtheit und warum handelt es sich dabei um eine Bestandsmasse? Definieren Sie die korrespondierenden Ereignismassen.

b) Stimmt die Erhebungsgesamtheit mit der Grundgesamtheit überein? Um welche Art der Erhebung handelt es sich folglich?

c) Wie lauten die Merkmalsträger und wie die Merkmale?

d) Die erwartete Wachstumsrate der Produktion der Unternehmen bezieht sich auf das laufende Jahr, also auf einen Zeitraum. Dennoch handelt es sich bei den erwarteten Änderungsraten um Bestandsdaten. Warum?

[1] Diese Aufgabe geht zurück auf eine ähnliche Frage in ARBEITSGRUPPE DES INSTITUTS FÜR STATISTIK UND ÖKONOMETRIE (1979–1991).

Aufgabe 1.2:

Im Wintersemester 1995/96 wurde die Wohnsituation von Studenten in Kiel untersucht. Dazu wurden 100 Studierende nach folgenden Merkmalen befragt:

Merkmal A: "Geschlecht"
Merkmal B: "Studienfach (Hauptfach)"
Merkmal C: "Semesterzahl (einschließlich laufendes Semester)"
Merkmal D: "Zufriedenheit mit der Wohnung"
Merkmal E: "Wohnfläche"
Merkmal F: "Art der Wohnung"
Merkmal G: "Monatliche Mietzahlung"
Merkmal H: "Monatliches Einkommen"

a) Nennen Sie für jedes Merkmal zwei mögliche Ausprägungen und geben Sie an, ob es sich um ein qualitatives, komparatives oder quantitatives Merkmal handelt. Geben Sie für die quantitativen Merkmale an, ob diese mit einer Intervall-, Verhältnis- oder Absolutskala gemessen werden.

b) Welches Problem taucht auf, wenn Sie "Studienfach" messen wollen? Wie können die Merkmalsausprägungen definiert werden?

c) Warum stellt "Semesterzahl" ein statistisches Merkmal dar? Ist dieses Merkmal diskret oder stetig? Handelt es sich bei Merkmal C um
   - ein extensives Merkmal?
   - ein intensives Merkmal?

d) Das Merkmal G wird in DM erhoben. Um die Situation der Kieler Studenten mit der Lage von Kommilitonen in anderen europäischen Universitätsstädten vergleichen zu können, sollen die Merkmalswerte in ECU umgerechnet werden (1 ECU = 2,04 DM). Wie lautet das transformierte Merkmal und um welchen Transformationstyp handelt es sich? Ist diese Transformation zulässig (relationstreu)?

e) Für das Merkmal D: "Zufriedenheit mit der Wohnung" wurden in der Befragung folgende (Urbild-)Klassen vorgegeben:

   D_1 : "unzufrieden"
   D_2 : "im Großen und Ganzen zufrieden"
   D_3 : "hochzufrieden"

   und folgendermaßen gemessen:

$$d = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad .$$

   Begründen Sie, welche der drei folgenden Merkmalstransformationen zulässig wäre(n):

$$i) \quad x(d) = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 100 \\ 1000 \end{bmatrix} \qquad ii) \quad y(d) = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 12 \\ 30 \end{bmatrix}$$

$$iii) \quad z(d) = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 12 \\ 11 \end{bmatrix}$$

Aufgabe 1.3:

Um Informationen über das Freizeitverhalten von Kieler Studenten und Studentinnen zu gewinnen, wurden 10 Studierenden (I-X) folgende Fragen gestellt:

1.) Haben Sie sich während Ihres ersten Semesters in Kiel ein Rennrad gekauft?
2.) Haben Sie sich während Ihres ersten Semesters in Kiel ein Surfbrett gekauft?

Das Ergebnis der Befragung sah folgendermaßen aus:

		Antwort auf Frage 1.)	Antwort auf Frage 2.)
Studentin/ Student	I	ja	nein
	II	nein	nein
	III	ja	nein
	IV	ja	ja
	V	ja	nein
	VI	nein	ja
	VII	nein	ja
	VIII	ja	ja
	IX	nein	nein
	X	ja	nein

a) Handelt es sich um eine Primär- oder Sekundärstatistik? Wie lautet die Urliste?

b) Bezeichnen Sie die Antwort auf Frage 1 als Merkmal A, die Antwort auf Frage 2 als Merkmal B. Definieren Sie die Merkmalsausprägungen wie folgt:

$$a_1 = 1 \quad (\text{“ja”}) \qquad a_2 = 0 \quad (\text{“nein”}) \quad ,$$

entsprechend für Merkmal B. Erstellen Sie die Datenmatrix für diesen bivariaten Datensatz.

c) Die Merkmale A und B sollen zu einem Merkmal zusammengefaßt werden. Warum erfüllt C: “Anschaffungen während des ersten Semesters” mit den Ausprägungen $c_1 = 1$ (“Kauf eines Rennrads”) und $c_2 = 2$ (“Kauf eines Surfbretts”) nicht die Anforderungen an ein statistisches Merkmal? Verdeutlichen Sie Ihre Aussage anhand eines Venn–Diagramms (vgl. Anhang A.IV).

d) Fassen Sie die beiden Merkmale A und B in einem Merkmal D: “Anschaffungen während des ersten Semesters” mit vier Ausprägungen zusammen. Geben Sie die relativen und absoluten Häufigkeiten der Ausprägungen von D an.

1.2 Lösungen

Lösung zu Aufgabe 1.1:

a) Die Unternehmen des produzierenden Gewerbes Deutschlands bilden die **Grundgesamtheit**[2], sie sind sachlich (Unternehmen des produzierenden Gewerbes), räumlich (Deutschland) und zeitlich (Januar 1996) eindeutig abgegrenzt.
Es handelt sich um eine **Bestandsmasse**, da sie zu einem Zeitpunkt betrachtet wird (auch wenn das genaue Datum der Untersuchung im Januar 1996 nicht angegeben ist).
Die Zugänge und Abgänge dieser Bestandsmasse bilden die korrespondierenden **Ereignismassen**. Zugänge wären z.B. Firmenneugründungen, Abgänge z.B Konkurse im produzierenden Gewerbe, jeweils innerhalb eines geeignet definierten Zeitintervalls.

b) Die dreihundert befragten Unternehmen machen die **Erhebungsgesamtheit** aus. Sie bilden eine Teilmenge der Grundgesamtheit, es handelt sich also um eine **Teilerhebung**.

c) **Merkmalsträger** sind die Unternehmen des produzierenden Gewerbes .
Folgende für das Untersuchungsziel bedeutsamen Merkmale sind im Aufgabentext genannt:

Merkmal A:	"Rechtsform"
Merkmal B:	"Bilanzsumme"
Merkmal C:	"Beschäftigtenzahl"
Merkmal D:	"Derzeitige Ertragslage"
Merkmal E:	"Erwartete Beschäftigtenzahl am Jahresende"
Merkmal F:	"Erwartete Produktionsänderungsrate für das laufende Jahr."

d) Die Erwartungen werden zu einem bestimmten Zeitpunkt, nämlich dem Tag der Befragung, erhoben. Es handelt sich also um einen "Bestand an Erwartungen". Dabei ist es unerheblich, ob sich der Gegenstand der Erwartungen auf einen Bestand (z.B. Beschäftigtenzahl am Jahresende) oder einen Strom (z.B. Produktionsänderung) bezieht.

Lösung zu Aufgabe 1.2:

a) Die Merkmalsausprägungen können stets verbal eindeutig definiert werden, z.B. für Merkmal A:

$$a_1 : \text{"weiblich"}, \qquad a_2 : \text{"männlich"} \quad .$$

Im folgenden werden die **Merkmalsausprägungen** jedoch stets als reelle Zahlen definiert, z. B.:

$$a_1 = 1\,, \qquad a_2 = 2 \quad .$$

Dadurch werden Transformationen und insbesondere die Speicherung der Daten in elektronischen Datenträgern erheblich erleichtert.

[2]Zu den im folgenden Text fettgedruckten Begriffen findet man an entsprechender Stelle in SCHNEIDER/KORNRUMPF/MOHR (1995) die zugehörigen Definitionen und gegebenenfalls zusätzliche Erläuterungen.

Mögliche Ausprägunen sind für Merkmal

A: 1 ("weiblich"), 2 ("männlich")
B: 1 ("Medizin"), 2 ("Betriebswirtschaftslehre"), ...
C: 1 ("1 Semester"), 6 ("6 Semester"), ...
D: 4 ("einigermaßen zufrieden"), 0 ("äußerst unzufrieden"), ...
E: 12 ("12 qm"), 23 ("23 qm"), ...
F: 100 ("Wohnheimplatz"), 200 ("Zimmer in einer Wohngemeinschaft"), ...
G: 260 ("260 DM"), 385 ("385 DM"), ...
H: 1150 ("1150 DM"), 870 ("870 DM"), ...

Bei den Merkmalen A, B und F sind weder die Reihenfolgen noch die Abstände der Ausprägungen sinnvoll interpretierbar. Es handelt sich deshalb um qualitative Merkmale, die eine **Nominalskala** arstellen.

Die Ausprägungen des Merkmals D lassen sich zwar der Reihe nach ordnen, die Abstände können aber nicht (objektiv) gemessen werden. Es handelt sich um ein komparatives Merkmal, das eine **Ordinalskala** ist.

Die Merkmale E, G und H sind quantitativ und werden mit einer **Kardinalskala** gemessen, denn sowohl die Reihenfolge als auch die Abstände zwischen den Beträgen besitzen inhaltliche Bedeutung. Da die Skala einen natürlichen Nullpunkt, aber keine natürliche Einheit besitzt (so ist die Währungseinheit willkürlich gewählt, vgl. d)), handelt es sich um eine **Verhältnisskala**.

Merkmal C ist ebenfalls quantitativ und metrisch skaliert. Da diese Skala neben dem natürlichen Nullpunkt auch eine natürliche Einheit aufweist, handelt es sich um eine **Absolutskala**.

b) Es könnte Befragte geben, die mehrere Fächer zugleich studieren (sogenannte Häufbarkeit). Um trotzdem eine Meßvorschrift zu erhalten, können die Merkmalsausprägungen wie folgt definiert werden:

$$\begin{array}{rcl} b_1 & : & \text{"(ausschließlich) Medizin"} \\ b_2 & : & \text{"(ausschließlich) Betriebswirtschaftslehre"}, \\ \vdots & & \vdots \\ b_k & : & \text{"(ausschließlich) } \mathbf{ein} \text{ anderes Fach"} \\ b_{k+1} & : & \text{"mehrere Fächer gleichzeitig"} \end{array} \quad , \quad \begin{array}{rcl} b_1 & = & 1 \\ b_2 & = & 2 \\ \vdots & & \vdots \\ b_k & = & k \\ b_{k+1} & = & k+1 \end{array} .$$

Studenten mit mehreren Hauptfächern fallen jetzt in die Klasse B_{k+1}, jeder Befragte besitzt nur eine Merkmalsausprägung $b_i, i = 1, \ldots k+1$. Die Ausprägung b_k = "(ausschließlich) **ein** anderes Fach" stellt sicher, daß das Merkmal B vollständig ist, auch wenn die $k-1$ ersten Ausprägungen nicht alle in Kiel angebotenen Studiengänge enthalten.

c) Jedes Erhebungselement nimmt genau eine Ausprägung von C an. Die Merkmalsausprägungen ordnen jedem Merkmalsträger genau eine natürliche Zahl zu. Die Menge aller natürlichen Zahlen ist abzählbar, das Merkmal demnach **diskret**.

Die Summe der Merkmalswerte (Semesterzahlen) ist nicht sinnvoll interpretierbar, wohl aber der Durchschnitt (mittlere Semesterzahl der Befragten). Es handelt sich daher allein um ein **intensives Merkmal**.

d) Die transformierte Variable lautet y=T(g)=$\frac{1}{2.04}$g, es handelt sich um eine Maßstabsänderung y=T(g)=cg mit $c = 0,49 > 0$, und damit um eine **monotone (und affine) Transformation**. Für eine Verhältnisskala stellt die lineare Abbildung $y = cx$ eine **relationstreue Transformation** dar, da die sinnvollen Relationen (Äquivalenzrelation, Ordnungsrelation, Abstandsrelation, Verhältnisrelation) erhalten bleiben.

e) Die Transformationen unter *i*) und *ii*) sind zulässig, da es sich bei x und y um monotone Transformationen von d handelt:

$$d_i \leq d_j \implies x(d_i) = x_i \leq x(d_j) = x_j$$
$$d_i \leq d_j \implies y(d_i) = y_i \leq y(d_j) = y_j$$

und monotone Abbildungen auf einer Ordinalskala **relationstreu** sind. (Die Abstände zwischen den Merkmalsausprägungen sind zwar unterschiedlich groß, sie können aber ohnehin nicht sinnvoll interpretiert werden.)
Die Transformation unter *iii*) ist dagegen nicht zulässig: Es gilt

$$d_1 < d_2 \quad \wedge \quad z(d_1) = z_1 < z(d_2) = z_2 \ ,$$
$$\text{gleichzeitig aber} \quad d_2 < d_3 \quad \wedge \quad z(d_2) = z_2 > z(d_3) = z_3 \ .$$

Bei $z(d)$ handelt es sich demnach nicht um eine monotone Transformation.

Lösung zu Aufgabe 1.3:

a) Es handelt sich um eine **Primärstatistik**, die eigens für den Untersuchungszweck erhoben wurde. Die in der Aufgabe angegebene Tabelle ist die **Urliste**, sie enthält die Rohdaten der Erhebung.

b) Die Urliste stellt bereits eine Datenmatrix dar. Es ist grundsätzlich gleich, ob die Merkmalsausprägungen als a_1 : "ja", a_2 : "nein" oder $a_1 = 1$, $a_2 = 0$ definiert werden. Im folgenden werden jedoch als **Merkmalsausprägungen** stets reelle Zahlen verwendet. Die **Datenmatrix** sieht bei der angegebenen Skalierung folgendermaßen aus:

		Merkmal	
		A	B
Unter-	I	1	0
suchungs-	II	0	0
einheiten	III	1	0
	IV	1	1
	V	1	0
	VI	0	1
	VII	0	1
	VIII	1	1
	IX	0	0
	X	1	0

c) Ein **statistisches Merkmal** weist *jedem* Element e_ν der Erhebungsgesamtheit E *genau eine* reelle Zahl zu. C mit den Ausprägungen c_1 und c_2 erfüllt diese Anforderung nicht:

Zum einen ist es nicht vollständig, d.h. nicht jedem Befragten wird eine Merkmalsausprägung zugewiesen: Für die Studierenden, die weder ein Rennrad noch ein Surfbrett erwarben, existiert keine Klasse. Für ein statistisches Merkmal A mit k Klassen gilt:

$$A_1 \cup A_2 \cup \ldots \cup A_k = E \qquad \text{bzw.} \qquad \overline{A_1 \cup A_2 \cup \ldots \cup A_k} = \emptyset \quad .$$

Im vorliegenden Beispiel ist jedoch

$$\overline{C_1 \cup C_2} = \{II, IX\} \neq \emptyset \quad .$$

Entsprechend läßt sich die Vollständigkeit eines Merkmals anhand der absoluten Häufigkeiten ausdrücken:

$$n(A_1 \cup A_2 \cup \ldots \cup A_k) = n(E) \quad ,$$

hier gilt jedoch

$$n(C_1 \cup C_2) = |\{I, III, IV, V, VI, VII, VIII, X\}| = 8 \neq n(E) = 10 \quad .$$

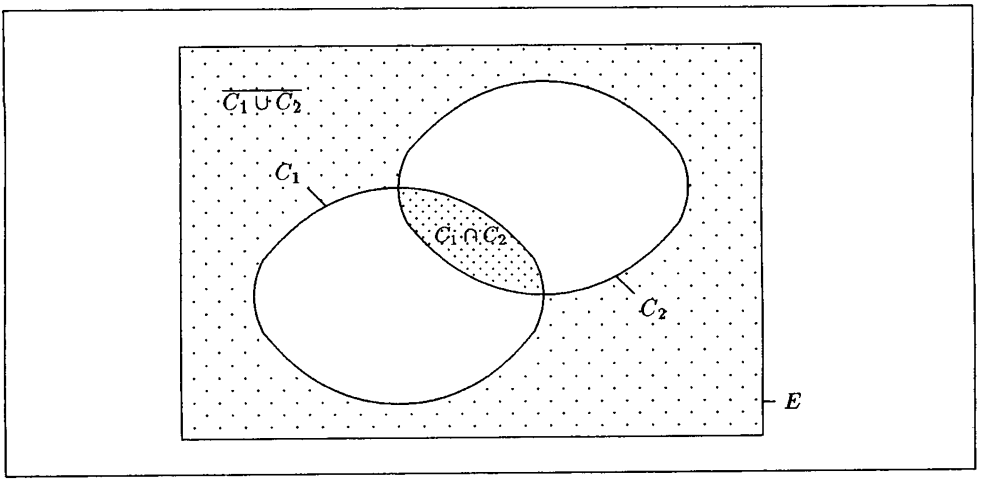
Zum anderen ist C häufbar: Ein Erhebungselement kann gleichzeitig mehrere Ausprägungen von C annehmen, d.h. die Befragten können sowohl ein Fahrrad als auch ein Surfbrett erworben haben. Für ein statistisches Merkmal A mit k Klassen gilt:

$$A_i \cap A_j = \emptyset \quad \text{bzw.} \quad n(A_i \cap A_j) = n(\emptyset) = 0 \quad \forall\, i \neq j \quad .$$

Demgegenüber erhält man in der vorliegenden Aufgabe

$$C_1 \cap C_2 = \{IV, VIII\} \neq \emptyset \quad \text{bzw.} \quad n(C_1 \cap C_2) = |\{IV, VIII\}| = 2 \neq n(\emptyset) \quad .$$

Im Venn–Diagramm zeigt sich die Häufbarkeit und die Unvollständigkeit von C durch die nichtleeren Mengen $\overline{C_1 \cup C_2}$ und $C_1 \cap C_2$:



d) Definiert man für das Merkmal D zunächst die Ausprägungen

d_1 : "Kauf eines Rennrads"

d_2 : "Kauf eines Surfbretts" ,

so kann die Häufbarkeit durch Einführung der Klasse

D_3 : "Kauf eines Rennrads und eines Surfbretts"

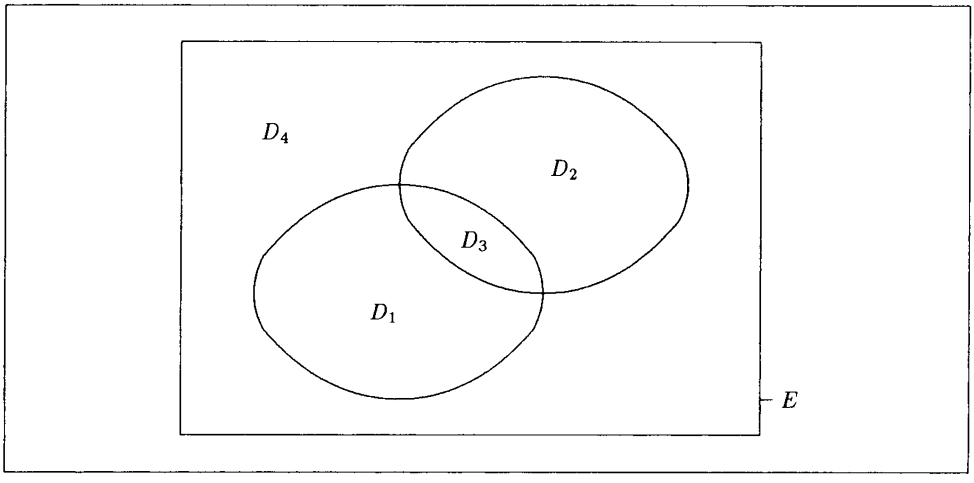
ausgeschlossen und die Vollständigkeit durch die Klasse

D_4 : "keine Anschaffung"

sichergestellt werden.

Die so definierte Klasseneinteilung bewirkt eine **Partition** der Erhebungsgesamtheit:

$$D_1 \cup D_2 \cup D_3 \cup D_4 = E \ , \quad D_i \cap D_j = \emptyset \quad \forall\, i \neq j \quad ,$$



und bildet ein statistisches Merkmal:

Merkmals-ausprägung	(Urbild-)Klasse	absolute Häufigkeit	relative Häufigkeit
$d_1 = 1$	$D_1 = \{I, III, V, X\}$	$h(d_1) = n(D_1) = 4$	$f(d_1) = \frac{n(D_1)}{N} = 4/10 = 0,4$
$d_2 = 2$	$D_2 = \{VI, VII\}$	$h(d_2) = n(D_2) = 2$	$f(d_2) = \frac{n(D_2)}{N} = 2/10 = 0,2$
$d_3 = 3$	$D_3 = \{IV, VIII\}$	$h(d_3) = n(D_3) = 2$	$f(d_3) = \frac{n(D_3)}{N} = 2/10 = 0,2$
$d_4 = 4$	$D_4 = \{II, IX\}$	$h(d_4) = n(D_4) = 2$	$f(d_4) = \frac{n(D_4)}{N} = 2/10 = 0,2$

Kapitel 2

Datenorganisation für qualitative Merkmale

2.1 Aufgaben

Aufgabe 2.1:

Die Auszubildenden in Schleswig–Holstein wurden am 31.12.1988 gezählt und nach dem Merkmal A: "Ausbildungsbranche" untergliedert. Das Ergebnis der Erhebung ist in folgender Tabelle wiedergegeben:

Industrie und Handel	A_1	32.350
Handwerk	A_2	25.950
Landwirtschaft	A_3	3.050
Öffentlicher Dienst	A_4	3.250
Sonstige	A_5	7.050

Quelle: Statistisches Taschenbuch Schleswig–Holstein 1989, Tab. 26.

a) Berechnen und interpretieren Sie den Erhebungsumfang N.

b) Berechnen Sie die relativen Häufigkeiten und interpretieren Sie $f(a_4)$. Stellen Sie den Erhebungsbefund in einem Kreissektorendiagramm dar.

c) Erstellen Sie ein Stabdiagramm der relativen Häufigkeiten und nehmen Sie Stellung zu folgender Aussage:

"Die Verteilung des Merkmals A ist 'linkssteil' ".

d) Geben Sie den Modus von A an und berechnen Sie die Entropie.

e) Zeigen Sie, daß die maximale Entropie einer Häufigkeitsverteilung $\mathbf{f}$ mit k Merkmalsklassen $H_k(\mathbf{f}) = \log_2 k$ beträgt. (Hinweis: Die Entropie erreicht ihr Maximum bei einer Gleichverteilung des Merkmals.)

f) Berechnen Sie für eine Zweipunktverteilung W, die die Ausprägungen W_1 und W_2 mit den relativen Häufigkeiten $f(w_1)$ und $f(w_2)$ annimmt, die Entropie für verschiedene Werte von $f(w_1)$. Tragen Sie Ihre Ergebnisse in eine Skizze ein und ermitteln Sie anhand dieser Skizze, welche Zweipunktverteilung dieselbe Streuung besitzt wie die Häufigkeitsverteilung des Merkmals A (wenn die Entropie als Streuungsmaß zugrunde gelegt wird). Überprüfen Sie Ihr Ergebnis, indem Sie die Entropie für die so erhaltene Zweipunktverteilung berechnen.

Aufgabe 2.2:

Die Waldschadeninventur 1988 in Schleswig-Holstein kam zu folgendem Ergebnis:

			Merkmal B: Baumartenfläche nach Schadstufen (in ha)		Baumartenfläche insgesamt (in ha)
			nicht geschädigt (B_1)	geschädigt (B_2)	
Merkmal A: Baumart	Fichte	(A_1)	16	19	35
	Kiefer	(A_2)	9	4	13
	Buche	(A_3)	8	26	34
	Eiche	(A_4)	8	9	17
	Sonstige	(A_5)	31	11	42
				Σ	141

Quelle: Statistisches Taschenbuch Schleswig-Holstein 1989, Tab. 66.

a) Geben Sie die modalen Klassen der Randverteilungen der Merkmale A und B sowie den Modus der gemeinsamen Verteilung an.

b) Berechnen und interpretieren Sie $f(a_1, b_1)$, $f(a_1|b_1)$, $f(b_1|a_1)$ und $f(a_3)$.

c) Wie lautet der Modus der bedingten Häufigkeitsverteilung $f(a_i|b_2)$ und wie ist er zu interpretieren? Berechnen Sie diese Häufigkeitsverteilung und interpretieren Sie $f(\mathbf{mod}(\mathbf{f}_{A|b_2}))$.

d) Nehmen Sie Stellung zu folgender Aussage:

> *"Der Kiefernbestand ist (im Verhältnis zum gesamten Baumbestand in Schleswig-Holstein) relativ gesund."*

Welche Häufigkeiten müssen Sie dazu miteinander vergleichen?

e) Überprüfen Sie die Abhängigkeit beider Merkmale mit Hilfe der PRE-Maße $\lambda_{A|B}$ und $\lambda_{B|A}$ und des Cramérschen Kontingenzmaßes.

f) In welche Richtung änderte sich die mittlere quadratische Kontingenz, wenn alle Felder der Kontingenztabelle die doppelten absoluten Häufigkeiten aufweisen würden?

Aufgabe 2.3:

1988 wurden 262 Abiturientinnen und Abiturienten (Merkmal G mit G_1 : "weiblich", G_2 : "männlich") bezüglich ihrer Studienabsicht (Merkmal S mit S_1 : "studienwillig", S_2 : "unentschlossen", S_3 : "ohne Studienabsicht") befragt. 61 der Befragten waren noch unentschlossen. Die Zahl der unentschlossenen Abiturientinnen war um eins geringer als die Zahl der unentschlossenen Abiturienten. 11 Abiturienten hatten keine Studienabsicht. Der Anteil der weiblichen Befragten betrug 45%. 61% der Studienwilligen waren männlich. (Die Prozentzahlen sind gerundet.)
(Datengrundlage: Statistisches Bundesamt (Hrsg.): Datenreport 1989, S. 64.)

a) Tragen Sie die genannten absoluten Häufigkeiten in eine Kontingenztabelle der Merkmale G und S ein. Ergänzen Sie die fehlenden absoluten (gemeinsamen und Rand-)Häufigkeiten.

b) Stellen Sie das Befragungsergebnis graphisch dar.

c) Berechnen und interpretieren Sie $f(g_1|s_2)$. Können Sie anhand dieses Wertes und den Angaben im Aufgabentext bereits eine Aussage über die Abhängigkeit der Merkmale S und G machen?

d) Berechnen und interpretieren Sie das PRE-Maß $\lambda_{S|G}$. Welchen entscheidenden Nachteil besitzen die PRE-Maße gegenüber der Cramérschen Kontingenz?

e) Überprüfen Sie den Wahrheitsgehalt folgender Aussage:

> *"Der Anteil der bezüglich ihres Studiums unentschlossenen Befragten ist bei den Schülerinnen größer als bei den Schülern. Dennoch ist bei den Unentschlossenen der Anteil der Frauen geringer als der der Männer"*

f) Erstellen Sie die Kontingenztabelle der relativen Häufigkeiten.

g) Berechnen Sie die Entropien der Randverteilungen sowie die Entropien der bedingten Verteilungen von S bei gegebenem G=g_1 und G=g_2. Vergleichen Sie $H_l(\mathbf{f}_{S|g_1})$ mit $H_l(\mathbf{f}_{S|g_2})$ und interpretieren Sie diesen Vergleich. Berechnen Sie die durchschnittliche Entropie der bedingten Verteilung $\mathbf{f}_{S|G}$.

h) Berechnen Sie die Entropie der gemeinsamen Verteilung von S und G

   i) über die gemeinsamen relativen Häufigkeiten

   ii) über den Additionssatz der Entropie.

i) Berechnen und interpretieren Sie die normierte Transinformation.

j) Zeigen Sie, daß im Falle der Unabhängigkeit der beiden Merkmale S und G $H_{kl}(\mathbf{f}_{GS}) = H_k(\mathbf{f}_G) + H_l(\mathbf{f}_S)$ und $T(\mathbf{f}_{GS}) = 0$ gelten würde.

Aufgabe 2.4:

Folgende Tabelle gibt den Benzinabsatz, gemessen in 1000 Tonnen, in der Bundesrepublik Deutschland wieder:

Benzinsorte (Merkmal A)	Jahr	Insgesamt	Verbleiung (Merkmal B)	
			unverbleit (B_1)	verbleit (B_2)
Normalbenzin (A_1)	1987	10138	4220	5918
	1988	7363	7350	13
Superbenzin (A_2)	1987	14898	2215	12683
	1988	18656	4221	14435
Insgesamt	1987	25036	6435	18601
	1988	26019	11571	14448

Quelle: Statistisches Bundesamt (Hrsg.): Datenreport 1989, S. 354.

a) Wie groß war

   i) 1988 der Anteil des verbleiten Superbenzins am gesamten Superbenzinabsatz?

   ii) 1987 der Anteil des Normalbenzins am verbleiten Kraftstoff?

   iii) in beiden Jahren zusammen der Anteil des verbleiten Superbenzins am gesamten Benzinabsatz?

b) Hat sich

   i) die modale Klasse der gemeinsamen Verteilung der Merkmale A und B

   ii) der Modus der Randverteilung von A

   iii) der Modus der bedingten Verteilung des Merkmals B bei gegebener Ausprägung a_1 des Merkmals A

   von 1987 auf 1988 geändert?

c) Stellen Sie die Randverteilungen des Merkmals B: "Verbleiung" jeweils für 1987 und 1988 in einem Stabdiagramm dar. Begründen Sie allein anhand der Zeichnung, ob die Entropie dieser Randverteilung 1988 größer, kleiner oder gleich der entsprechenden Entropie im Jahr 1987 war.

d) Erstellen Sie für beide Jahre die Kontingenztabelle der relativen Häufigkeiten und berechnen Sie jeweils das Cramérsche Kontingenzmaß. (Benutzen Sie zur Berechnung von ϕ^2 die vereinfachte Formel für Vierfeldertafeln.) Vergleichen Sie die beiden Werte und interpretieren Sie diesen Vergleich.

e) Zeigen Sie die Gültigkeit der Gleichung

$$\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(f_{ij}-f_{i\cdot}f_{\cdot j})^2}{f_{i\cdot}f_{\cdot j}} = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{f_{ij}^2}{f_{i\cdot}f_{\cdot j}} - 1 \qquad \left(=\phi^2(\mathbf{f})\right) \quad .$$

Aufgabe 2.5:

• Zimmermann: Senioren am Steuer zuverlässiger

Ältere Autofahrer sind sicherere Verkehrsteilnehmer als junge Fahrer. Das meint Bundesverkehrsminister Zimmermann (CSU). Während Fahrer über 65 Jahre nur vier Prozent der Unfallverursacher ausmachten, gingen 27,5% aller Unfälle auf Fehlverhalten von Fahrern zwischen 18 und 25 Jahren zurück.

Quelle: Kieler Nachrichten vom 11.09.1990.

Wird die Einschätzung der Autofahrer seitens des Verkehrsministers durch die im Artikel genannten relativen Häufigkeiten bestätigt?

Aufgabe 2.6:[1]

Die folgende Tabelle zeigt die Aufteilung der Beschäftigten eines Betriebes nach den Merkmalen Geschlecht (A), Arbeitsverhältnis (B) und Ausbildung (C). Dabei treten folgende Merkmalsklassen auf:

A_1 : "männlich" A_2 : "weiblich"
B_1 : "Angestellte(r)" B_2 : "Arbeiter(in)"
C_1 : "mit Fachausbildung" C_2 : "um- oder angelernt"

$C \rightarrow$	c_1			c_2			$\Sigma(C)$		
$B \rightarrow$ / $A \downarrow$	b_1	b_2	$\Sigma(B)$	b_1	b_2	$\Sigma(B)$	b_1	b_2	$\Sigma(B)$
a_1	10				50				
a_2		10		10				40	
$\Sigma(A)$	30							150	200

Bemerkung: $\Sigma(\cdot)$ soll die Summation über das jeweilige Merkmal andeuten

[1] Diese Aufgabe geht zurück auf eine ähnliche Frage in ARBEITSGRUPPE DES INSTITUTS FÜR STATISTIK UND ÖKONOMETRIE (1979–1991).

a) Interpretieren Sie die angegebenen absoluten Häufigkeiten.

b) Ergänzen Sie die fehlenden Werte in der dreidimensionalen Kontingenztafel.

c) Berechnen und interpretieren Sie folgende relativen Häufigkeiten: $f(a_1, c_2)$, $f(b_2)$, $f(a_2, b_2, c_2)$, $f(a_2|b_1)$, $f(a_2|b_1, c_2)$, $f(b_1, c_1|a_2)$.

d) Bestimmen und interpretieren Sie die modalen Klassen folgender Verteilungen: $f(a_i, b_j, c_h)$, $f(a_i, c_h)$, $f(b_j)$, $f(c_h|a_2)$, $f(c_h|a_2, b_1)$, $f(a_i, b_j|c_1)$.

e) Nehmen Sie Stellung zu folgenden Aussagen:

i) *"Die Hälfte der Beschäftigten sind männliche Arbeiter."*

ii) *"Die meisten Arbeiterinnen haben eine Fachausbildung."*

iii) *"Die um- oder angelernten Beschäftigten sind überwiegend als Arbeiter oder Arbeiterinnen tätig."*

f) Überprüfen Sie die Abhängigkeit der Merkmale B und C anhand der normierten Transinformation der bivariaten Häufigkeitsverteilung dieser beiden Merkmale.

2.2 Lösungen

Lösung zu Aufgabe 2.1:

a) A stellt ein statistisches Merkmal dar, d.h. die Klasseneinteilung von A induziert eine **Partition** der Erhebungsgesamtheit. Für die absoluten Häufigkeiten des Merkmals A gilt der **Additionssatz**:

$$N = \sum_{i=1}^{5} n(A_i) = 32350 + 25950 + 3050 + 3250 + 7050 = 71650 ,$$

d.h. am 31.12.1988 waren in Schleswig-Holstein 71.650 Auszubildende beschäftigt.

b) Es gilt: $f(a_i) = n(A_i)/N$, z.B. $f(a_1) = 32350/71650 = 0,452$. Entsprechend ergeben sich die übrigen relativen Häufigkeiten:

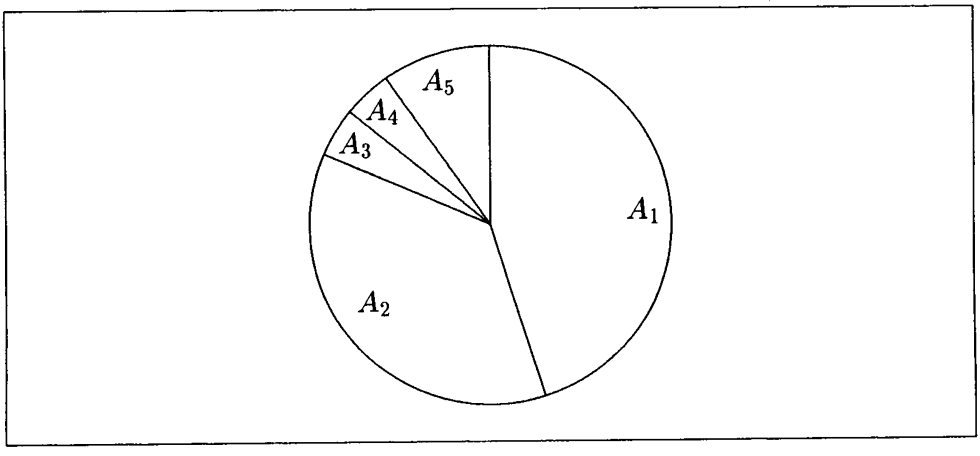
i	1	2	3	4	5	Σ
$f(a_i)$	0,452	0,362	0,043	0,045	0,098	1

$f(a_4) = 0,045$, d.h. am 31.12.1988 waren 4,5% der Auszubildenden in Schleswig-Holstein im öffentlichen Dienst beschäftigt.

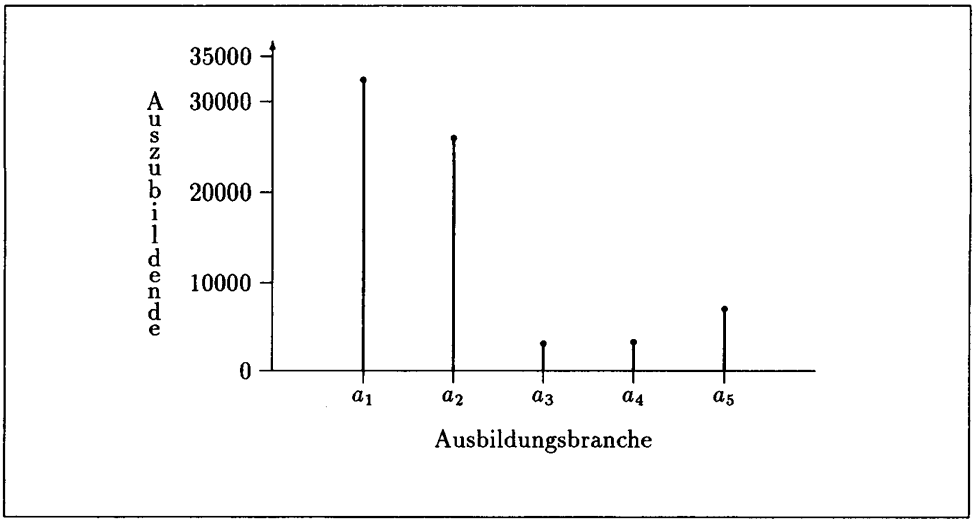
Im **Kreissektorendiagramm** sind die Zentriwinkel α_i der Kreisausschnitte den Häufigkeiten proportional, z.B. $\alpha_1 = 0,452 \cdot 360° = 162,72° \doteq 163°$:

i	1	2	3	4	5	Σ
α_i	163°	130°	15°	16°	35°	360°

Damit ergibt sich das Kreisdiagramm:



c) Im **Stabdiagramm** entspricht die Höhe der Stäbe den jeweiligen Häufigkeiten:



Die Aussage ist unzulässig: A ist ein qualitatives Merkmal (ohne Ordnungsstruktur). Die Anordnung der Ausprägungen im Stabdiagramm ist willkürlich. Allein durch Umordnung der Ausprägungen läßt sich aus demselben Urmaterial eine "rechtssteile" Verteilung erzeugen.

d) Der **Modus** a_{i^*} ist diejenige Merkmalsausprägung, die am häufigsten vorkommt. Es gilt

$$\mathbf{max}_i[f(a_i)] = f(a_1) \Rightarrow a_{i^*} = a_1 : \text{"Industrie und Handel"} \quad ,$$

d.h. die meisten Lehringe wurden im Industrie– und Handelssektor ausgebildet.
Die **Entropie** beträgt:

$$\begin{aligned} H_k(\mathbf{f}_A) &= -\sum_{i=1}^{5} f_i \log_2 f_i = -\sum_i f_i \frac{\ln f_i}{\ln 2} = -1,443 \sum_i f_i \ln f_i \\ &= -1,443(0,452 \ln 0,452 + 0,362 \ln 0,362 + 0,043 \ln 0,043 + 0,045 \ln 0,045 \\ &\quad +0,098 \ln 0,098) = -1,443(-1,229) = 1,773 \quad . \end{aligned}$$

e) Die **maximale Entropie** ergibt sich, wenn die Häufigkeitsverteilung eines Merkmals A die größtmögliche Streuung aufweist, d.h. wenn alle k Merkmalsausprägungen die gleiche absolute Häufigkeit $n(A_i) = N/k$ bzw. relative Häufigkeit $f(a_i) = 1/k$ besitzen. Die Entropie beträgt in diesem Fall

$$\begin{aligned} H_k(\mathbf{f}_A) &= -\sum_{i=1}^{k} \frac{1}{k} \log_2 \frac{1}{k} = -k \left(\frac{1}{k} \log_2 \frac{1}{k} \right) \\ &= -(\log_2 1 - \log_2 k) = 0 + \log_2 k = \log_2 k \quad . \end{aligned}$$

f) Bei der Zweipunktverteilung gilt $f(w_1) + f(w_2) = 1$ bzw. $f(w_2) = 1 - f(w_1)$. Die Entropie

$$H(\mathbf{f}_W) = -1,443\,(f(w_1) \ln f(w_1) + (1 - f(w_1)) \ln(1 - f(w_1)))$$