Gliederung

Teil I: Grundlagen

1. Gegenstand
2. Grundbegriffe
   2.1. Statistische Einheit und statistische Masse
   2.2. Merkmalsarten
3. Datengewinnung
   3.1. Datenerhebung
   3.2. Datenaufbereitung

Aufgaben und Lösungen zu Teil I

Teil II: Univariate Datenanalyse

4. Eindimensionale Häufigkeitsverteilung
   4.1. Häufigkeitsverteilung bei gruppierten Daten
   4.2. Häufigkeitsverteilung bei klassierten Daten
5. Lageparameter
   5.1. Modus (Modalwert)
   5.2. Median (Zentralwert)
   5.3. Arithmetisches Mittel (Mittelwert)
6. Streuungsparameter
   6.1. Spannweite (Range)
   6.2. Quartilsabstand
   6.3. Varianz und Standardabweichung
   6.4. Variationskoeffizient

Aufgaben und Lösungen zu Teil II

Teil III: Bivariate Datenanalyse

7. Zweidimensionale Häufigkeitsverteilung
8. Zusammenhang bei nominalen Merkmalen
9. Zusammenhang bei metrischen Merkmalen
10. Zusammenhang bei ordinalen Merkmalen
11. Zusammenhang bei Merkmalen mit unterschiedlichem Skalenniveau

Aufgaben und Lösungen zu Teil III

Teil IV: Prognoseverfahren

12. Regressionsanalyse
    12.1. Einfache lineare Regressionsrechnung
    12.2. Einfache nichtlineare Regressionsrechnung
    12.3. Multiple lineare Regressionsrechnung
13. Zeitreihenanalyse
    13.1. Komponenten einer Zeitreihe und ihre Modellierung
    13.2. Methoden zur Trendermittlung
    13.3. Ermittlung der Saisonkomponente

Aufgaben und Lösungen zu Teil IV

1. Gegenstand

Statistik ist das methodische Vorgehen bei der Beschaffung von Informationen, um vernünftige Entscheidungen treffen zu können. Ihre grundlegenden Aufgaben liegen zum einen in der Bereitstellung von Daten und zum anderen in der Bereitstellung von Methoden zur Erhebung, Aufbereitung und Analyse dieser Daten.

Diese Daten werden häufig durch Befragungen gewonnen. Abbildung 1 zeigt hierzu beispielhaft einen Auszug aus einem Fragebogen einer Jugendherberge, bei dem es darum geht, konkrete Informationen zur Gästestruktur, zur Erwartungshaltung und zur Zufriedenheit dieser Gäste zu erhalten. Die so gewonnenen Informationen bilden dann die Basis, um in einem zweiten Schritt die strategische Ausrichtung dieses Unternehmens zu überprüfen und gegebenenfalls neu zu justieren. Dieser Prozess soll im Folgenden zumindest in Ansätzen vorgestellt werden. Hierzu wird ein repräsentativer Datensatz aus dieser konkreten Befragung gezogen, um hieran die in den nachfolgenden Kapiteln vorzustellenden statistischen Verfahren zu testen. Die ursprünglichen Daten wurden im Rahmen eines Forschungsprojektes mit dieser Jugendherberge gewonnen und mussten hier aus Gründen der zugesicherten Vertraulichkeit etwas verzerrt werden.

Natürlich sind die Aussagen der einzelnen Objekte im Zuge der vorgenommenen Analysen nicht mehr erkennbar, sie sind dann auch nicht mehr von Interesse. Vielmehr versucht die Statistik aus der Untersuchung einer Vielzahl anonymisierter Objekte (= Masse) grundsätzliche Aussagen zu dieser Masse herauszufiltern und so wichtige Informationen und Zusammenhänge sichtbar zu machen, z.B. Konsumentscheidungen von Haushalten oder Meinungen der Gesamtbevölkerung zu politischen Fragestellungen.

Die Methodenlehre der Statistik stellt hierfür ein breites Instrumentarium zur Verfügung. Sie beschränkt sich dabei häufig allein aus Zeit- und Kostengründen auf eine stichprobenartige Auswahl von Objekten aus der entsprechenden Gesamtheit. Neben der komprimierenden Beschreibung des gewonnenen Zahlenmaterials müssen dann in einem zweiten Schritt mit Hilfe wahrscheinlichkeitstheoretischer Überlegungen Rückschlüsse von der Stichprobe auf die - unbekannte - Gesamtheit gezogen werden. Aus diesen Überlegungen lassen sich dann die beiden Teilgebiete der Statistik ableiten:

Die **beschreibende (deskriptive) Statistik** wendet Methoden zur Erfassung, Auswertung und übersichtlichen Darstellung von Daten an, die von einer Gruppe von Objekten erhoben wurden. Diese Gruppe von Objekten wird dabei als Gesamtheit und nicht als eine Teilmenge (Stichprobe) angesehen. Es müssen folglich keine Hochrechnungen von den Ergebnissen einer Stichprobe auf die Gesamtheit erfolgen.

Die **schließende (induktive) Statistik** setzt das Instrumentarium der deskriptiven Statistik voraus, geht aber genau diesen Schritt weiter und sieht in der betrachteten Gruppe von Objekten die Stichprobe einer (unbekannten) Gesamtheit. Die Ergebnisse einer solchen Stichprobe müssen also unter Berücksichtigung von Zufallseinflüssen auf die Gesamtheit übertragen werden. Die induktive Statistik vermittelt daher Methoden zum Treffen von Aussagen auf der Grundlage von Hypothesen, Schätzungen und Prognosen. Dazu nutzt sie die Erkenntnisse der Wahrscheinlichkeitsrechnung.

Gegenstand der Lehrveranstaltungen und damit auch der nachfolgenden Kapitel ist schwerpunktmäßig die deskriptive Statistik. Sie wird jedoch punktuell um induktive Verfahren ergänzt werden, um letztendlich fundiertere Aussagen treffen zu können.

1. Gegenstand

Abbildung 1

Gästebefragung einer Jugendherberge (Auszug)

1. Kennen Sie unsere Jugendherberge bereits von früheren Aufenthalten?
   ❑ nein (0) ❑ ja (1)
2. Wie sind Sie erstmalig auf uns aufmerksam geworden?
   ❑ DJH-Verzeichnis (1) ❑ Empfehlung (2) ❑ Web-Seite/Internet (3)
3. Mit wem verbringen Sie hier bei uns Ihren Aufenthalt?
   ❑ Schulklasse (1) ❑ Junior = allein / Paar unter 27 Jahre (2)
   ❑ Gruppe / Verein (3) ❑ 27P = allein / Paar ab 27 Jahre (4)
   ❑ Familie (5)
4. Wie lange bleiben Sie hier in der Jugendherberge? ____ Nächte
5. Wie bewerten Sie auf einer Skala von 1 (= sehr schlecht) bis 5 (= sehr gut) die ...
   ... Einrichtung und Ausstattung der Jugendherberge
   ❑ 1 ❑ 2 ❑ 3 ❑ 4 ❑ 5 = sehr gut
   ... Lage und Umgebung der Jugendherberge
   ❑ 1 ❑ 2 ❑ 3 ❑ 4 ❑ 5 = sehr gut
   ... Service und Kompetenz des Personals
   ❑ 1 ❑ 2 ❑ 3 ❑ 4 ❑ 5 = sehr gut
   ... Verpflegung in der Jugendherberge
   ❑ 1 ❑ 2 ❑ 3 ❑ 4 ❑ 5 = sehr gut
6. Wie bewerten Sie zusammenfassend die Jugendherberge im Intervall von 0 (= sehr schlecht) bis 50 Punkte (= sehr gut)? ____ Punkte
7. Können Sie sich vorstellen, unsere Jugendherberge demnächst wieder zu besuchen?
   ❑ nein (0) ❑ ja (1)
8. Würden Sie unsere Jugendherberge auch weiterempfehlen?
   ❑ 1 = nein ❑ 2 ❑ 3 = jein ❑ 4 ❑ 5 = ja
9. Dürfen wir Sie abschließend noch nach Ihrem Alter fragen? ____ Jahre

...

2. Grundbegriffe

2.1. Statistische Einheit und statistische Masse

Vor dem Hintergrund der oben bereits vorgestellten Gästebefragung sollen in diesem Abschnitt die zentralen statistischen Grundbegriffe geklärt werden. Sie sind deshalb von so elementarer Bedeutung, da die Art des vorliegenden Datenmaterials die richtige Auswahl und den Einsatz statistischer Analyseverfahren ganz entscheidend determiniert.

Die Grundlage statistischer Untersuchungen bilden die **statistischen Einheiten**. Hierunter sind Personen, Objekte oder auch Ereignisse zu verstehen, die durch bestimmte Eigenschaften oder Merkmale gekennzeichnet sind. Sie werden daher alternativ auch als Merkmalsträger oder einfach als Elemente bezeichnet.

Die Menge aller statistischen Einheiten, die in sachlicher, räumlicher und zeitlicher Form abgegrenzt ist, heißt **statistische Masse** oder Grundgesamtheit der Untersuchung. Die Anzahl N der statistischen Einheiten bezeichnet man als den Umfang der betreffenden statistischen Masse.

Die statistischen Einheiten stellen die **Merkmalsträger** der Untersuchung dar, deren Eigenschaften festgestellt werden sollen. Dabei muss es sich um reale, klar voneinander abgrenzbare und damit auch abzählbare Elemente handeln. Die Merkmale der statistischen Einheiten dienen einerseits zur Abgrenzung der statistischen Masse (Unterscheidungsmerkmale), zum anderen zur Gewinnung von Informationen über diese Einheiten bezüglich des Untersuchungsgegenstandes (Erhebungsmerkmale).

Weiterhin muss zwischen **Merkmal** (charakteristische Eigenschaft) und **Merkmalsausprägung** (Merkmalswerte) unterschieden werden. Jedes Merkmal besitzt in der Regel mehrere Merkmalsausprägungen. Dabei muss aber die für die Untersuchung relevante Ausprägung auch feststellbar sein.

Beispiele:

- Statistische Masse: Anzahl an Haushalten in Deutschland am 31.12.2013
  Merkmal: Haushaltsgröße
  Merkmalsausprägungen: 1 Person, 2 Personen, 3 Personen, 4 und mehr Personen

- Statistische Masse: Bestand an Pkw in Deutschland am 22.9.2013
  Merkmal: Ausstattung mit dem ABS-Bremssystem
  Merkmalsausprägungen: vorhanden, nicht vorhanden

- Statistische Masse: befragte Gäste in einer Jugendherberge im Jahr 2015
  Merkmal: Gesamtbewertung der Jugendherberge
  Merkmalsausprägungen: 0, 1, 2, ..., 20, ..., 40, 41, 42, ..., 46, 47, 48, 49, 50 Punkte

Durch ein Merkmal wird eine statistische Masse in Teilmengen zerlegt. Jede statistische Einheit kommt in genau einer Teilmenge vor; dadurch entsteht eine disjunkte Klasseneinteilung der Einheiten der statistischen Masse. Die Teilmengen können selbst wieder eine statistische Masse bilden.

2.2. Merkmalsarten

An den Beispielen Haushaltsgröße, ABS-Ausstattung und Gesamtbewertung der Jugendherberge wird bereits deutlich, dass Merkmale und ihre Ausprägungen von recht unterschiedlicher Art sein können.

Statistische Merkmale lassen sich daher in verschiedener Hinsicht klassifizieren. Eine Möglichkeit wäre die Einteilung in **qualitative** und **quantitative** Merkmale.

- **Qualitative Merkmale** zeichnen sich durch eine verbale Beschreibung der jeweiligen Eigenschaften aus. Ihre Ausprägungen unterscheiden sich durch ihre Art, beispielsweise die Merkmale Geschlecht (männlich, weiblich), Familienstand (ledig, verheiratet, geschieden, verwitwet) oder Weiterempfehlung der Jugendherberge (nein, vermutlich nein, jein, vermutlich ja, ja).

- **Quantitative Merkmale** nehmen dagegen ihre Ausprägungen originär in reellen Zahlen (Alter, Umsatz, Einkommen, Produktionshöhe) an. Sie unterscheiden sich folglich durch ihre Größe. Es lassen sich daher Merkmalssummen sowie Mittelwerte, Streuungsmaße und andere Kennzahlen berechnen.

Während qualitative Merkmale nur bestimmte, voneinander isolierte Werte annehmen können, sind quantitative Merkmale unter diesem Aspekt differenzierter zu betrachten. Hier gibt es Merkmale, für die innerhalb eines vorgegebenen Intervalls jede beliebige reelle Zahl und damit unendlich viele Ausprägungen realisierbar wären. Vor diesem Hintergrund werden **diskrete, stetige und quasi-stetige** Merkmale unterschieden:

- **Diskrete Merkmale** können nur einzelne, klar voneinander getrennte Ausprägungen annehmen. Die Anzahl der verschiedenen Ausprägungen ist dabei meist überschaubar, vor allem dann, wenn es sich um qualitative Merkmale handelt (z.B. Geschlecht, Familienstand, Gästeart, Weiterempfehlung der Jugendherberge).

- **Stetige Merkmale** können dagegen in einem bestimmten Bereich alle reelle Zahlen annehmen und zwar unabhängig davon, ob die Messgenauigkeit ausreicht oder nicht (z.B. Körpergröße, Körpergewicht, Alter, Temperatur).

- **Quasi-stetige Merkmale** sind eigentlich diskrete Merkmale, die aber zahlreiche mögliche Ausprägungen in einem begrenzten Intervall besitzen. Es handelt sich dabei i.d.R. um quantitative Merkmale mit sehr vielen, aber eben nicht unendlich differenzierbaren Ausprägungen (z.B. Geldgrößen, Arbeitslosenzahl, Lagerbestände, Alter in Jahren, Bewertung der Jugendherberge in Punkten).

Da bei stetigen und quasi-stetigen Merkmalen eine Vielzahl verschiedener Merkmalsausprägungen vorliegt, werden diese Werte zwecks übersichtlicherer Darstellung in Klassen zusammengefasst. Auf diesen Prozess der Datenverdichtung wird dann im nachfolgenden Kapitel näher eingegangen.

2. Grundbegriffe

Bei den bislang beispielhaft genannten Merkmalen mit ihren Ausprägungen ist bereits deutlich geworden, dass sie sich hinsichtlich Art, Qualität und Anzahl ganz grundlegend unterscheiden können. Merkmalsausprägungen, deren Werte als natürliche oder gar reelle Zahlen angegeben werden können, lassen sich dann auch mathematisch miteinander verknüpfen. Bei qualitativen Merkmalen macht dies jedoch keinen Sinn. Hier lassen die Ausprägungen nur eine Festlegung einer Rangordnung oder vielleicht sogar nur eine Unterscheidung nach ihrer Art zu.

Die Ausprägungen der Merkmale können also ein ganz unterschiedliches Informationsniveau besitzen. Dieses Informationsniveau wird als Skalenniveau bezeichnet und in **nominal, ordinal und metrisch** unterschieden. Die Unterscheidung der Merkmale in verschiedene Skalen ist deshalb von herausragender Bedeutung, weil sie die Art und damit den Umfang der anwendbaren statistischen Verfahren bestimmen:

- **Nominal skalierte** Merkmale zeichnen sich lediglich durch eine Verschiedenartigkeit ihrer Ausprägungen aus. Dabei ist es nicht relevant, ob sie originär verbal oder in Zahlen dargestellt werden, da die Zahlen in diesem Fall lediglich für die Codierung eben dieser verbalen Ausprägungen stehen. Nominal skalierte Merkmale sind z.B. Telefonnummern, Beruf, Geschlecht, Gästeart oder Familienstand.

- **Ordinal skalierte** Merkmale bringen neben einer Verschiedenheit eine natürliche Rangordnung zum Ausdruck. Ihre Ausprägungen können verbal (z.B. sehr gut oder gut) oder zahlenmäßig verschlüsselt (z.B. 1 für sehr gut oder 2 für gut) dargestellt werden. Ihre Abstände sind allerdings nicht quantifizierbar, die Codierungen sagen also nichts über die Abstände zwischen den Merkmalsausprägungen aus. Ordinal skalierte Merkmale sind z.B. Prüfungsnoten, Güteklassen bei Lebensmitteln oder die Platzierung bei Ligatabellen.

- **Metrisch-intervall skalierte** Merkmale werden grundsätzlich mit Zahlen beschrieben. Die Zahlen sind dabei als solche auch interpretierbar, neben die Rangordnung tritt also noch die Möglichkeit, die Abstände zwischen den Merkmalsausprägungen zu berechnen und zu vergleichen. Der Nullpunkt ist eher willkürlich festgelegt (z.B. Temperatur in Celsius und Kalenderjahre); es ist daher nicht sinnvoll, die Merkmalsausprägungen ins Verhältnis zueinander zu setzen.

- **Metrisch-verhältnis skalierte Merkmale** besitzen darüber hinaus einen natürlichen Nullpunkt. Dadurch ist es möglich und sinnvoll, den Quotienten zweier Merkmalsausprägungen zu bilden. So ist ein Mensch mit 180 cm eben doppelt so groß als einer mit 90 cm. Weitere Beispiele für metrisch-verhältnis skalierte Merkmale sind die Temperatur in Kelvin, die Körpergröße, das Einkommen oder das Alter.

3. Datengewinnung

Die Durchführung einer statistischen Erhebung hängt wesentlich von der speziellen Aufgabenstellung ab. Im Vorfeld ist daher zu konkretisieren, welche statistische Masse (= Grundgesamtheit) analysiert werden soll; sie ist dazu in sachlicher, zeitlicher und räumlicher Hinsicht einzugrenzen. Ist über den Umfang der Merkmalsträger (Untersuchungseinheiten oder Objekte) entschieden, sind in einem zweiten Schritt die zu erhebenden Merkmale mit ihren Ausprägungen festzulegen. Je sorgfältiger diese Schritte geplant werden, desto aussagefähiger ist später die gewonnene Datenbasis.

Ablaufphasen einer statistischen Erhebung:

- Datenerhebung
- Datenaufbereitung
- Analyse des Datenmaterials

3.1. Datenerhebung

Erhebungsart

Daten können auf verschiedene Weise erhoben werden. Gemeinhin wird nach dem Umfang und nach der Zugriffsebene unterschieden:

- Bei einer **Totalerhebung** werden alle Einheiten der betreffenden statistischen Masse erfasst. Im Gegensatz hierzu findet bei einer **Teilerhebung** nur ein Teil aller statistischen Einheiten Berücksichtigung. Die Teilerhebung verfolgt meist das Ziel, aus den Informationen über wenige Einheiten der statistischen Masse zu Aussagen über die Gesamtheit der statistischen Masse zu gelangen. Bei der Unterscheidung in deskriptive und induktive Statistik wurde auf diesen Zusammenhang bereits hingewiesen. Die induktive Statistik nutzt folglich den Vorteil einer kostengünstigeren Teilerhebung.

- Werden die Einheiten und die zu diesen Einheiten gewünschten Daten direkt und durch eine eigens hierfür durchgeführten Untersuchung erfasst, wird von einer **primärstatistischen Erhebung** gesprochen. Die Produzenten primärstatistischer Daten lassen sich unterteilen in die Träger der amtlichen Statistik (Statistisches Bundesamt, Statistische Landesämter, Kommunalstatistische Ämter, Ministerien, Deutsche Bundesbank, OECD oder Europäische Zentralbank) und in die Träger der nichtamtlichen Statistik (Wirtschaftsverbände, Wirtschafts- und Marktforschungsinstitute usw.).
  Wird dagegen keine eigene Erhebung organisiert, sondern auf die Daten dieser Träger zurückgegriffen, die eigentlich für andere Zwecke ermittelt wurden, liegt eine **sekundärstatistische Erhebung vor**. Als typisches Beispiel gilt der Kauf von GfK-Marktstudien durch den Einzelhandel.

3. Datengewinnung

Erhebungstechnik

Die Merkmalsausprägungen der vorab eingegrenzten statistischen Einheiten können grundsätzlich durch **Befragung, Beobachtung oder Experiment** erfasst werden.

- Bei der **Befragung** sind die Erhebungsformulare so zu gestalten, dass eine einwandfreie Datenqualität und eine hohe Antwortquote erwartet werden können.
  Fragen nach leicht feststellbaren Tatsachen ergeben zuverlässigere Antworten als Fragen nach Meinungen oder Wünschen. Letztere sollten daher über geschlossene Fragen erfasst werden. Hierzu werden einer Person auf eine Frage die Antwortmöglichkeiten bereits vorgegeben, das Spektrum an Antworten ist folglich vom Auftraggeber bereits eingegrenzt worden. Geschlossene Fragen verlangen deshalb einen beträchtlichen Vorbereitungsaufwand, sie erleichtern aber dem Befragten die Beantwortung und auch ganz erheblich die Auswertung. Demgegenüber erlauben offene Fragen vollkommen freie Aussagen.
  Die Befragung kann schriftlich (durch den Befragten selbst) oder mündlich (durch den Interviewer im Dialog mit dem Befragten) durchgeführt werden. Beide Vorgehensweisen weisen im konkreten Einzelfall abzuwägende Vor- und Nachteile auf.
  So sind schriftliche Befragungen im Normalfall kostengünstiger, führen aber meist zu einer geringeren Antwortquote. Bei mündlichen Befragungen verhält es sich genau umgekehrt, darüber hinaus können auch schwierigere Fragen formuliert werden.

- Ein **Experiment** dient vorwiegend dem Testen von Annahmen (Beispiel: Geschmackstest von Cola-Getränken). Ein Experiment lässt sich als Befragung oder Beobachtung innerhalb einer kontrollierten Versuchsanordnung definieren. Bei vorgegebenen künstlichen Rahmenbedingungen wird von einem Laborexperiment gesprochen, bei natürlichen von einem Feldexperiment.

Durch Befragung, Beobachtung oder Experiment werden an N statistischen Einheiten ein oder mehrere Merkmale erfasst. Die Zusammenstellung der Ausprägungen x_i (i = 1,...,N) eines Merkmals X in der Reihenfolge ihrer Erhebung wird Urliste genannt. Die Daten liegen unsortiert vor.

3.2. Datenaufbereitung

Speicherung statistischer Daten

Bei jeder Datenerhebung wird für jede statistische Einheit eine entsprechende Merkmalsausprägung erfasst. Die Ausprägung kann danach in einen 'Zahlenwert' des Merkmals transformiert werden (= Codierung). Diese Zahlenwerte bilden die Basis für weitere statistische Analysen. Allerdings darf dabei nicht vergessen werden, dass mathematische Rechenoperationen nur bei metrisch skalierten Merkmalen durchgeführt werden dürfen, da nur diesen Merkmalen numerische Daten zugrunde liegen!!!

Das Beispiel unserer Gästebefragung soll an dieser Stelle wieder aufgegriffen werden, um diesen Prozess der Datenaufbereitung zu konkretisieren. Hierzu finden Sie in der nachfolgenden Tabelle fünfzig repräsentativ ausgewählte Datensätze zu den in Abbildung 1 vorgestellten Fragen. Aus diesen Datensätzen sollen dann im weiteren Verlauf immer wieder auch Beispielrechnungen rekrutiert werden.

3. Datengewinnung

Tabelle 1

Ausgewählte Datensätze (50) zur Gästebefragung

Nr.	Afrü	Info	Gart	Nach	EiAu	LaUm	SeKo	Verp	Gbew	Awie	Empf	Alte
	F1	F2	F3	F4	F5a	F5b	F5c	F5d	F6	F7	F8	F9
1	1	1	4	4	4	4	5	5	47	1	5	58
2	1	2	4	2	5	5	5	5	45	1	4	44
3	0	3	4	2	4	4	4	5	43	1	4	38
4	1	2	4	1	4	4	5	5	43	0	4	28
5	0	3	4	1	4	4	4	4	40	0	4	35
6	0	1	4	1	4	4	4	4	41	1	4	32
7	1	2	5	4	5	4	5	5	46	1	4	47
8	0	1	5	5	5	4	5	4	47	1	5	54
9	1	3	5	3	5	4	4	5	48	1	5	23
10	0	3	5	3	5	4	5	4	45	1	5	43
11	0	3	5	4	4	4	5	4	45	0	5	44
12	1	1	5	2	4	4	4	5	43	1	5	36
13	0	3	5	2	4	4	4	4	41	0	4	33
14	0	1	3	2	5	5	5	5	47	1	5	56
15	1	3	3	2	3	4	5	5	46	1	5	48
16	1	3	3	3	5	5	5	5	46	1	4	52
17	1	2	3	1	4	4	5	5	45	0	4	44
18	1	3	3	2	5	4	5	4	44	1	4	44
19	0	2	3	3	4	3	5	5	45	1	4	44
20	0	2	3	2	4	4	4	5	41	0	3	31
21	1	3	3	3	3	4	4	4	42	0	5	29
22	0	3	3	4	5	4	5	5	46	1	4	24
23	0	3	3	4	3	4	5	5	44	0	4	27
24	0	3	2	2	5	4	4	5	46	1	5	21
25	0	1	2	2	4	4	4	4	44	0	5	26
26	1	3	2	1	4	3	4	5	45	0	5	25
27	0	3	2	1	4	4	5	4	42	0	4	26
28	0	3	2	1	4	4	4	4	41	0	4	20
29	0	2	1	3	4	4	5	5	44	0	4	14
30	0	2	1	5	5	4	5	5	46	0	4	13
31	0	1	1	3	4	4	5	5	45	0	4	15
32	0	2	1	5	5	4	5	5	46	0	4	11
33	0	3	1	5	4	4	5	5	45	0	4	14
34	0	3	1	5	5	3	5	5	46	1	4	10
35	1	2	1	3	3	4	5	5	44	0	4	13
36	0	2	1	5	5	4	5	5	46	0	4	10
37	0	2	1	5	4	4	5	5	45	0	4	13
38	0	2	1	5	4	4	5	5	46	0	4	11
39	1	2	1	2	5	5	5	5	50	1	5	7
40	1	2	1	3	5	5	5	4	49	1	5	9
41	0	3	1	2	5	5	5	3	47	1	5	8
42	0	1	1	5	5	5	5	5	50	1	5	10
43	0	2	1	2	4	4	4	4	44	0	5	12
44	0	2	1	3	4	4	5	4	42	0	4	15
45	1	2	1	2	5	4	4	4	42	0	5	14
46	0	1	3	3	4	4	5	5	47	1	5	58
47	1	3	3	2	5	4	5	5	46	1	4	49
48	1	3	3	2	4	4	5	5	44	0	4	42
49	1	2	3	1	4	4	4	5	41	1	3	30
50	0	3	3	2	4	4	5	4	43	0	5	34

3. Datengewinnung

Merkmal ... mit	Merkmalsausprägungen und Codierung (in Klammer)
F1: Afrü	Früherer Aufenthalt x_1 = nein (0); x_2 = ja (1)
F2: Info	Erstmalige Information über die Jugendherberge x_1 = DJH-Verzeichnis (1); x_2 = Empfehlung (2); x_3 = Internet (3)
F3: Gart	Gästeart bzw. -kategorie x_1 = Schulklasse (1); x_2 = Junior (2); x_3 = Gruppe/Verein (3); x_4 = 27P (4); x_5 = Familie (5)
F4: Nach	Anzahl Übernachtungen x_1 = 1 (1); x_2 = 2 (2); x_3 = 3 (3); x_4 = 4 (4); x_5 = 5 (5)
F5a: EiAu	Bewertung Einrichtung und Ausstattung der Jugendherberge x_1 = sehr schlecht (1); x_2 = schlecht (2); x_3 = mittelmäßig (3); x_4 = gut (4); x_5 = sehr gut (5)
F5b: LaUm	Bewertung Lage und Umgebung der Jugendherberge x_1 = sehr schlecht (1); x_2 = schlecht (2); x_3 = mittelmäßig (3); x_4 = gut (4); x_5 = sehr gut (5)
F5c: SeKo	Bewertung Service und Kompetenz des Personals x_1 = sehr schlecht (1); x_2 = schlecht (2); x_3 = mittelmäßig (3); x_4 = gut (4); x_5 = sehr gut (5)
F5d: Verp	Bewertung Verpflegung in der Jugendherberge x_1 = sehr schlecht (1); x_2 = schlecht (2); x_3 = mittelmäßig (3); x_4 = gut (4); x_5 = sehr gut (5)
F6: Gbew	Gesamtbewertung der Jugendherberge x_1 = 40 (40); x_2 = 41 (41); x_3 = 42 (42); x_4 = 43 (43); x_5 = 44 (44); x_6 = 45 (45); x_7 = 46 (46); x_8 = 47 (47); x_9 = 48 (48); x_{10} = 49 (49); x_{11} = 50 (50);
F7: Awie	Erneuter Besuch x_1 = nein (0); x_2 = ja (1)
F8: Empf	Weiterempfehlung x_1 = nein (1); x_2 = vermutlich nein (2); x_3 = jein (3); x_4 = vermutlich ja (4); x_5 = ja (5)
F9: Alte	Alter x_1 = 7 (7); x_2 = 8 (8); x_3 = 9 (9); x_4 = 10 (10); x_5 = 11 (11); x_6 = 12 (12); x_7 = 13 (13); x_8 = 14 (14); x_9 = 15 (15); ... x_{20} = 31 (31); x_{21} = 32 (32); x_{22} = 33 (33); x_{23} = 34 (34); ... x_{33} = 52 (52); x_{34} = 54 (54); x_{35} = 56 (56); x_{36} = 58 (58);

3. Datengewinnung

Datenverdichtung

Werden die Merkmalsausprägungen in einer Urliste fortlaufend ausgewiesen, fällt es schwer, sich einen Überblick über die Verteilung und deren Schwerpunkt zu verschaffen. Es macht daher Sinn, die vorliegenden Daten zu verdichten. In Abhängigkeit von der Anzahl unterschiedlicher Ausprägungen bietet sich hierfür die **Gruppierung** oder die **Klassierung** der Daten an:

- Bei der Datenverdichtung durch **Gruppierung** werden gleiche Merkmalswerte unter Berücksichtigung der Anzahl ihres Auftretens zusammengefasst.
  Für nominal skalierte Merkmale ist die Gruppierung die einzige Möglichkeit der Datenverdichtung. Da dieses Vorgehen bei einer etwas größeren Anzahl verschiedener Merkmalsausprägungen schnell an die Grenzen der Übersichtlichkeit stößt, müssen verschiedene Ausprägungen zu einer übergeordneten Gruppe zusammengefasst werden. In diesem Fall wird von einer mehrstufigen Klassifikation gesprochen (z.B. Unterscheidung der Nationalität nicht nach Ländern, sondern nach Kontinenten).

  Beispiel:

  Erstmalige Information über die Jugendherberge (F2: Info)

  Urliste:
  1, 2, 3, 2, 3, 1, 2, 1, 3, 3, 3, 1, 3, 1, 3, 3, 2, 3, 2, 2, 3, 3, 3, ..., 2, 2, 2, 3, 1, 2

  Gruppierung:

Info	**Anzahl**
1 (DJH-Verzeichnis)	9
2 (Empfehlung)	19
3 (Internet)	22
∑	**50**

- Bei der Datenverdichtung durch **Klassierung** werden benachbarte Merkmalswerte zu einer größeren Einheit zusammengefasst. Dieses Verfahren bietet sich bei einer Vielzahl von Ausprägungen (stetig oder quasi-stetig) eines ordinal oder metrisch skalierten Merkmals an; also bei Merkmalen, die mit Blick auf eine sinnvolle Klasseneinteilung über eine natürliche Rangfolge ihrer Ausprägungen verfügen. Daneben sind folgende Hinweise zu beachten:

  - möglichst Klassen mit konstanter Klassenbreite bilden; lediglich bei großem Wertebereich unterschiedlicher Klassenbreiten verwenden
  - Anzahl der Klassen sollte sich zwischen 5 und maximal 15 bewegen
  - x_i^u = Untergrenze der Klasse i (mit i = 1, …, k)
  - x_i^o = Obergrenze der Klasse i, wobei gilt: $x_i^o = x_{i+1}^u$
  - jede Merkmalsausprägung befindet sich in genau einem halboffenen Intervall (von der Untergrenze bis unter die Obergrenze) $x_i^u \leq x < x_i^o$
  - die Breite einer Klasse Δx_i errechnet sich dann aus $\Delta x_i = x_i^o - x_i^u$
  - die Klassenmitte x_i' errechnet sich dann aus $x_i' = (x_i^u + x_i^o)/2$

3. Datengewinnung

Beispiel:

Alter (F9: Alte)

Urliste:
58, 44, 38, 28, 35, 32, 47, 54, 23, 43, 44, 36, ..., 15, 14, 58, 49, 42, 30, 34

Klassierung:

Alter	Anzahl
[6; 10)	3
[10; 18)	14
[18; 27)	7
[27;40)	11
[40; 50)	10
[50; 61)	5
∑	50

Die durch eine Erhebung ermittelten Daten haben zunächst wenig Aussagekraft. Ihre Bedeutung lässt sich schon eher erkennen, wenn sie in einem ersten Schritt nach Merkmalen und deren Ausprägungen geordnet und übersichtlich in Tabellen oder Grafiken dargestellt werden. Sowohl mit der tabellarischen und grafischen Darstellung der Daten als auch mit der Berechnung von Lage- und Streuungsparametern werden sich die nachfolgenden Kapitel beschäftigen.

Aufgaben und Lösungen

Aufgabe I.1:

Charakterisieren Sie die nachfolgenden Merkmale:

a) Temperatur in Celsius
b) Temperatur in Kelvin
c) Staatsangehörigkeit
d) Aktienkurs
e) Klausurnoten

Aufgabe I.2:

Welchem Skalentyp sind die folgenden Merkmale zugeordnet und welches sind die statistischen Einheiten zur Erfassung dieser Merkmale?

a) erlernter Beruf von derzeit Arbeitslosen im Kreis Nordwest-Mecklenburg
b) Alter von Maschinen eines Unternehmens
c) Körpergröße von Schülern der Klasse 11a am Gymnasium in Schönberg
d) Lebensdauer von Kfz-Reifen
e) Abfüllmenge von Flaschen eines Automaten zur Bierabfüllung

Aufgabe I.3:

Nennen Sie statistische Einheit, statistische Masse, Merkmal und Skalenniveau.

a) Eine Befragung von 10 Wismarer Studierenden nach ihrer Semesterzahl brachte am 02.06.2005 folgendes Ergebnis: …
b) An einer Hauptverkehrsstraße in Rostock wurden bei einer Polizeikontrolle am 05.03.05 zwischen 11 und 12 Uhr an 100 Fahrzeugen folgende Geschwindigkeiten gemessen: …
c) Der Besitzer einer Erdbeerplantage notierte an den vergangenen 100 Tagen, wie viele Stunden die Sonne täglich auf seine Erdbeeren schien.
d) 7 Forschungsinstitute schätzen die reale Wachstumsrate in Deutschland für 2010.
e) 1000 repräsentativ ausgewählte Bäume in Mecklenburg-Vorpommern lassen sich in Abhängigkeit vom Alter in fünf Schadensklassen einteilen.

Lösung Aufgabe I.1:

a) quantitativ / stetig (klassiert) / metrisch-intervall
b) quantitativ / stetig (klassiert) / metrisch-verhältnis
c) qualitativ / diskret (gruppiert) / nominal
d) quantitativ / quasi-stetig (klassiert) / metrisch-verhältnis
e) qualitativ / diskret (gruppiert) / ordinal

Lösung Aufgabe I.2:

a) nominal / Arbeitsloser
b) metrisch-verhältnis / Maschine
c) metrisch-verhältnis / Schüler
d) metrisch-verhältnis / Reifen
e) metrisch-verhältnis / Flasche

4. Eindimensionale Häufigkeitsverteilung

Die Häufigkeitstabelle ist die wichtigste (Darstellungs-)Form des Datenmaterials im Rahmen der univariaten Datenanalyse. Dabei wird den einzelnen Merkmalsausprägungen die Häufigkeit ihres Auftretens zugeordnet. Darüber hinaus wird die tabellarische um eine grafische Darstellung ergänzt.

4.1. Häufigkeitsverteilungen bei gruppierten Daten

absolute und relative Häufigkeiten

Ein Merkmal, das in k verschiedenen Ausprägungen vorkommen kann, wird an N statistischen Einheiten beobachtet.

Definition: absolute Häufigkeiten h_i

Besitzt ein Merkmal die k verschiedenen Merkmalsausprägungen x_1, .., x_k, so ist h_i (i = 1, 2, .., k) die Anzahl der Merkmalsträger mit der Merkmalsausprägung x_i; h_i heißt absolute Häufigkeit der Ausprägung x_i. Die Summe aller absoluten Häufigkeiten entspricht dem Umfang der statistischen Masse N:

$$\sum_{i=1}^{k} h_i = N$$

Werden allerdings verschiedene statistische Massen miteinander verglichen, sind absolute Häufigkeiten nicht sehr aussagekräftig, da sie sehr stark vom Umfang N ihrer Gesamtheit abhängig sind. Es ist daher sinnvoll, die absoluten Häufigkeiten in Relation zu ihrer Summe N zu setzen; hieraus errechnen sich dann die relativen Häufigkeiten.

Definition: relative Häufigkeiten f_i

Werden die absoluten Häufigkeiten durch die Gesamtzahl der Elemente N dividiert, so ergeben sich hieraus die relativen Häufigkeiten f_i:

$$f_i = \frac{h_i}{N} \quad (i = 1, 2, ..., k) \quad \text{mit} \quad \sum_{i=1}^{k} f_i = 1$$

Grafisch lässt sich die Häufigkeitsverteilung eines gruppierten Merkmals am besten durch ein Stab- bzw. Balkendiagramm veranschaulichen.

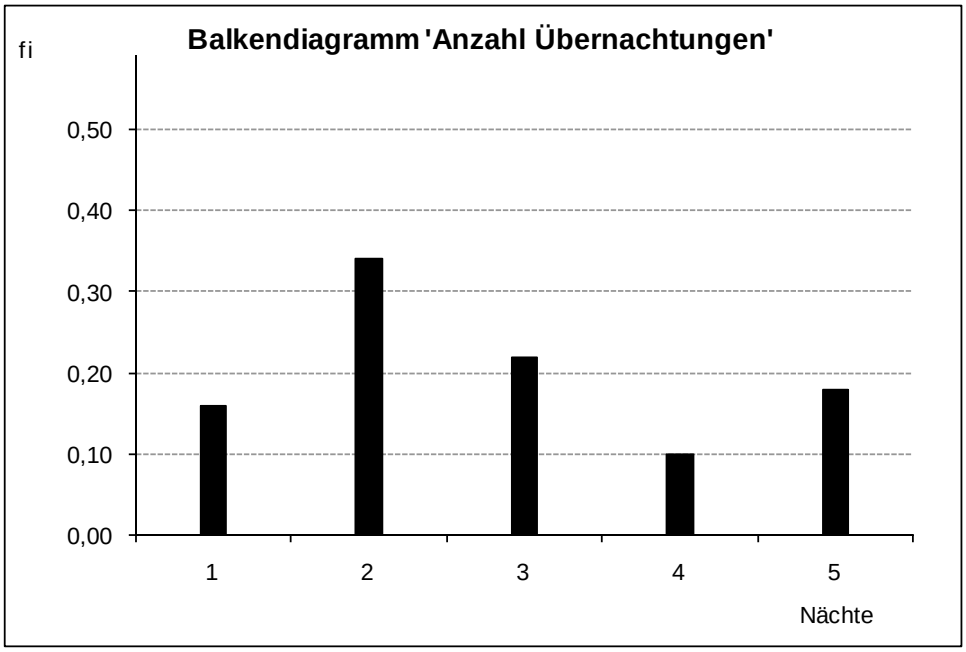
4. Eindimensionale Häufigkeitsverteilung

Beispiel:

Sehen wir uns an dieser Stelle das Merkmal 'Anzahl Übernachtungen' (F4: Nach) etwas genauer an. Es handelt sich hierbei um ein metrisches Merkmal (verhältnis-skaliert) mit einer Vielzahl möglicher, aber diskreter Ausprägungen, also quasi-stetig. Naheliegend wäre demnach eine klassierte Form der Datenaufbereitung. Der Blick auf unsere Urliste in Tabelle 1 zeigt jedoch, dass konkret nur fünf verschiedene Merkmalsausprägungen von einer bis zu fünf Nächten auftreten. Es spricht daher nichts gegen eine gruppierte Häufigkeitsverteilung, die um die Berechnung relativer Häufigkeiten erweitert wird:

x_i (Nach)	h_i	f_i
1	8	0,16
2	17	0,34
3	11	0,22
4	5	0,10
5	9	0,18
∑	**50**	**1,00**

Zur grafischen Veranschaulichung dieser relativen Häufigkeiten bietet sich dann ein Stab- oder ein Balkendiagramm an. In beiden Fällen wird die relative Häufigkeit einer Merkmalsausprägung durch die Höhe des Stabes bzw. des Balkens visualisiert:



absolute und relative Summenhäufigkeiten

Häufig ist nicht nur von Interesse, wie viele statistische Einheiten eine bestimmte Merkmalsausprägung besitzen, sondern auch, wie viele der statistischen Einheiten höchstens einen bestimmten Wert erreichen. Um diese Frage zu beantworten, müssen die Ausprägungen der Größe nach geordnet werden. Betrachtet werden daher nur ordinal und metrisch skalierte Merkmale, da nur diese eine natürliche Rangfolge besitzen.

4. Eindimensionale Häufigkeitsverteilung

Definition: Summenhäufigkeiten H_i und F_i _ gruppiert

Gegeben sei ein gruppiertes, mindestens ordinal skaliertes Merkmal X. Die Summe der absoluten Häufigkeiten der Merkmalsausprägungen, die den Wert x_i nicht überschreiten, heißt absolute Summenhäufigkeit H_i :

$$H_i = h(x \le x_i) = h_1 + ... + h_i = \sum_{j=1}^{i} h_j \quad (i = 1, 2, ..., k)$$

Die Summe der relativen Häufigkeiten der Merkmalsausprägungen, die den Wert x_i nicht überschreiten, heißt relative Summenhäufigkeit F_i :

$$F_i = f(x \le x_i) = f_1 + ... + f_i = \sum_{j=1}^{i} f_j \quad (i = 1, 2, ..., k)$$

Aus den relativen Summenhäufigkeiten lässt sich die **Funktion der relativen Summenhäufigkeiten bzw. die Verteilungsfunktion F(x)** ableiten. F(x) gibt dann den Anteil der Elemente mit einem Merkmalswert an, der kleiner oder gleich x ist.

Beispiel

Wir erweitern also unser ursprüngliches Beispiel zum Merkmal 'Anzahl Übernachtungen' (F4: Nach), indem wir in einem ersten Schritt absolute und relative Summenhäufigkeiten berechnen. Da die ursprüngliche Häufigkeitsverteilung ordinal oder metrisch skalierter Merkmale deren Ausprägungen bereits aufsteigend sortiert erfasst, errechnen sich die Summenhäufigkeiten einfach aus der zeilenweisen Addition der entsprechenden (absoluten oder relativen) Häufigkeiten:

x_i (Nach)	h_i	f_i	H_i	F_i
1	8	0,16	8	0,16
2	17	0,34	25	0,50
3	11	0,22	36	0,72
4	5	0,10	41	0,82
5	9	0,18	50	1,00
∑	**50**	**1,00**		

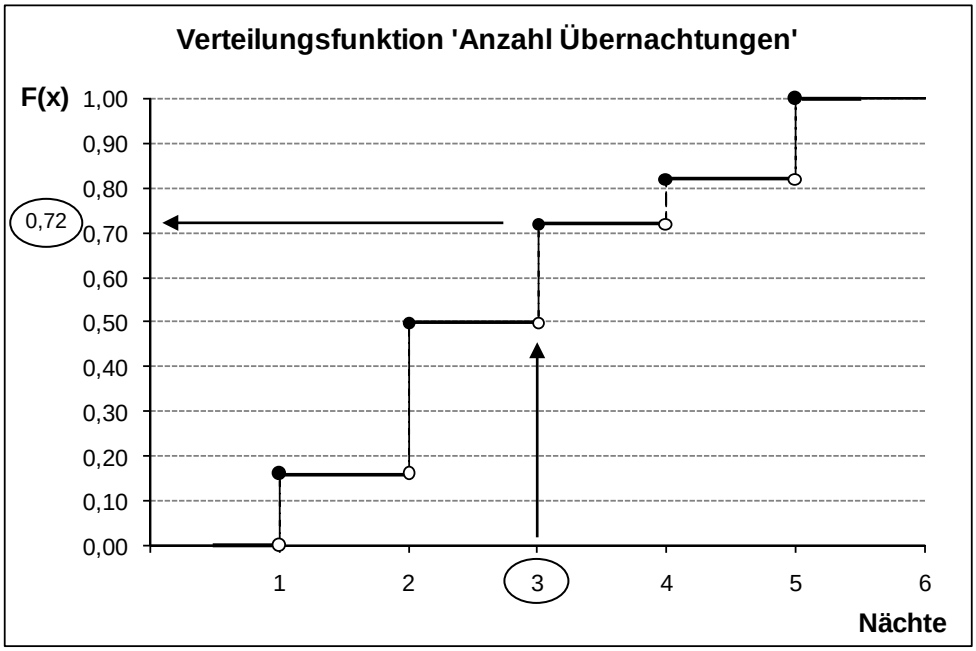
36 Personen oder umgerechnet 72% aller befragter Gäste übernachten also in der Jugendherberge maximal (oder: bis zu) 3 Nächte. Die formale Schreibweise hierfür lautet dann:

$H_{i=3} = h(x \le x_3) = h(x \le 3) = 36$ mit: i = 3 = 3. Gruppe

$F_{i=3} = f(x \le x_3) = f(x \le 3) = 0{,}75$ mit: i = 3 = 3. Gruppe

Die grafische Darstellung der relativen Summenhäufigkeiten zeigt uns dann einen treppenartigen Verlauf. Da die statistischen Einheiten bei gruppierter Aufbereitung nur einzelne (isolierte) Ausprägungen annehmen, kann es zwischen diesen Ausprägungen auch keinen Zuwachs der Summenhäufigkeiten geben. Die Verteilungsfunktion 'springt' daher nur an der nächsten erfassten Merkmalsausprägung auf das nächsthöhere Niveau:

4. Eindimensionale Häufigkeitsverteilung



Auch grafisch lässt sich der Anteil der befragten Personen bestimmen, die bspw. bis zu drei Nächte in der Jugendherberge bleiben. Die eingezeichneten Pfeile verdeutlichen den Lösungsweg, der natürlich zum gleichen Ergebnis von 72% führen muss.

Der Gesamtverlauf dieser Verteilungsfunktion über alle 5 Gruppen hinweg lässt sich dann formal in der folgenden Weise beschreiben:

$$F(x) = \begin{cases} 0 & \text{für } x < 1 \\ 0{,}16 & \text{für } 1 \le x < 2 \\ 0{,}50 & \text{für } 2 \le x < 3 \\ 0{,}72 & \text{für } 3 \le x < 4 \\ 0{,}82 & \text{für } 4 \le x < 5 \\ 1 & \text{für } x \ge 5 \end{cases}$$

Anhand des vorangestellten Beispiels sollte dann auch die verallgemeinerte Form der Verteilungsfunktion bei gruppierten Merkmalen nachvollziehbar sein.

Definition: Verteilungsfunktion F(x) _ gruppiert

Aus den relativen Summenhäufigkeiten lässt sich die relative Summenhäufigkeitsfunktion F(x) ableiten. F(x) wird auch als Verteilungsfunktion bezeichnet. Ihre Werte geben den Anteil der Elemente mit einem Merkmalswert kleiner oder gleich x an und errechnen sich bei gruppierten Daten wie folgt:

$$F(x) = \begin{cases} 0 & \text{für } x < x_1 \\ F_i & \text{für } x_i \le x < x_{i+1} \\ 1 & \text{für } x \ge x_k \end{cases}$$

4.2. Häufigkeitsverteilungen bei klassierten Daten

Tritt ein Merkmal in sehr vielen Ausprägungen auf, wie dies bei stetigen oder quasi-stetigen Merkmalen der Fall ist, dann wird eine gruppierte Häufigkeitsverteilung sehr unübersichtlich. Es ist praktisch unmöglich, jeder einzelnen Merkmalsausprägung einen Häufigkeitswert zuzuordnen.

Die Klassenbildung ist eine Voraussetzung für die Bestimmung der Häufigkeitsverteilung. Die Klassen werden durch ihre Klassengrenzen charakterisiert. Im Allgemeinen strebt man Klassen konstanter Klassenbreite an; allerdings kann es bei einem großen Variationsbereich des Datenmaterials sinnvoll sein, unterschiedliche Klassenbreiten zu verwenden. Die Klasseneinteilung soll die Struktur der untersuchten Gesamtheit möglichst deutlich abbilden.

absolute und relative Häufigkeiten

Die Berechnung der relativen Häufigkeiten wie auch der absoluten und relativen Summenhäufigkeiten erfolgt bei klassierten Daten analog zum gruppierten Datenmaterial. Die Häufigkeiten beziehen sich dann auf eine Klasse und eben nicht mehr auf eine Gruppe. Auf eine gesonderte formale Darstellung kann daher an dieser Stelle verzichtet werden. Vielmehr konzentrieren sich die nachfolgenden Ausführungen auf die grafische Aufbereitung der relativen Häufigkeiten und der daraus abgeleiteten Häufigkeitssummenfunktion.

Die Häufigkeitsverteilung eines klassierten Merkmals weist im Vergleich zur gruppierten Datenaufbereitung eine identische Struktur auf. An die Stelle der k diskreten Ausprägungen treten hier lediglich die k Klassen. Allerdings könnten durch eine geschickte Wahl der Klassengrenzen die Häufigkeiten in einer Klasse leicht manipuliert werden. Wird eine Klasse sehr breit angelegt, fallen mehr Merkmalsträger mit ihren Ausprägungen in diese Klasse und sie erscheint damit auf den ersten Blick gewichtiger.

Das Säulen- bzw. Balkendiagramm würde genau diese Manipulationsmöglichkeit auch optisch weitergeben. Es wird daher bei der grafischen Darstellung von Häufigkeiten klassierter Daten durch ein Histogramm ersetzt.
Dabei bringt nicht die Höhe, sondern die Fläche der Rechtecke die Häufigkeit einer Klasse zum Ausdruck, d.h. die absoluten bzw. relativen Klassenhäufigkeiten sind den Flächeninhalten der einzelnen Rechtecke proportional. Deren Höhe h_i^* (absolute Häufigkeitsdichte) bzw. f_i^* (relative Häufigkeitsdichte) kann dann einfach aus dem Quotienten von Fläche h_i bzw. f_i und Breite der jeweiligen Klasse Δx_i errechnet werden.
Die aus der klassischen eindimensionalen Häufigkeitsverteilung entwickelte Arbeitstabelle wird also um die Spalten Klassenbreite Δx_i und relative Häufigkeitsdichte f_i^* erweitert.

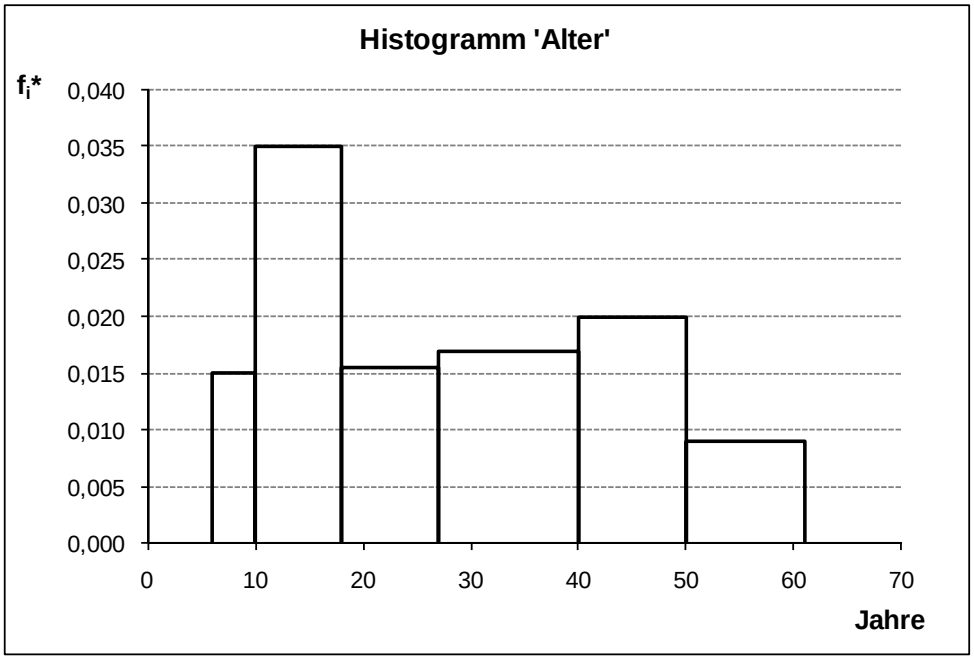
4. Eindimensionale Häufigkeitsverteilung

Beispiel:

Es bietet sich natürlich an, hier auf das Merkmal 'Alter' (F9: Alte) zurückzugreifen, für das in Abschnitt 3.2. bereits eine Klasseneinteilung vorgenommen wurde:

x_i (Alte)	h_i	f_i	H_i	F_i	Δx_i	f_i^*
[6; 10)	3	0,06	3	0,06	4	0,0150
[10; 18)	14	0,28	17	0,34	8	0,0350
[18; 27)	7	0,14	24	0,48	9	0,0156
[27;40)	11	0,22	35	0,70	13	0,0169
[40; 50)	10	0,20	45	0,90	10	0,0200
[50; 61)	5	0,10	50	1,00	11	0,0091
$\sum$	**50**	**1,00**				

Aus den in dieser Arbeitstabelle vorliegenden Informationen lassen sich die relativen Häufigkeiten problemlos in Form eines Histogramms darstellen:



Deutlich ist zu erkennen, dass sich mit Blick auf die relativen Häufigkeitsdichten f_i^* die Gewichtung der einzelnen Klassen (bei objektiver Betrachtung) verschiebt.
Obwohl zahlenmäßig am kleinsten, hat die erste Klasse nach der relativen Häufigkeitsdichte mit 0,015 zur dritten Klasse fast aufgeschlossen und liegt auch nur knapp hinter der vierten Klasse. Die relative Häufigkeitsdichte ist ein Indikator dafür, wie eng die statistischen Einheiten in einer Klasse beieinander liegen. Die relative Häufigkeitsdichte mit 0,015 gibt also an, dass durchschnittlich in jeder Einheit dieser Klasse (in unserem Beispiel ein Jahr) 0,015 oder 1,5% aller Merkmalsträger liegen, bei einer statistischen Masse von 50 somit 0,75 Personen. Dieser Durchschnittswert ergibt sich dann auch für die dritte Klasse.
Deutlich hervorgehoben wird in diesem Beispiel die zweite Klasse. Aus Sicht der Jugendherbergsleitung stellen damit die 10- bis unter 18-jährigen die mit Abstand wichtigste Altersgruppe dar.

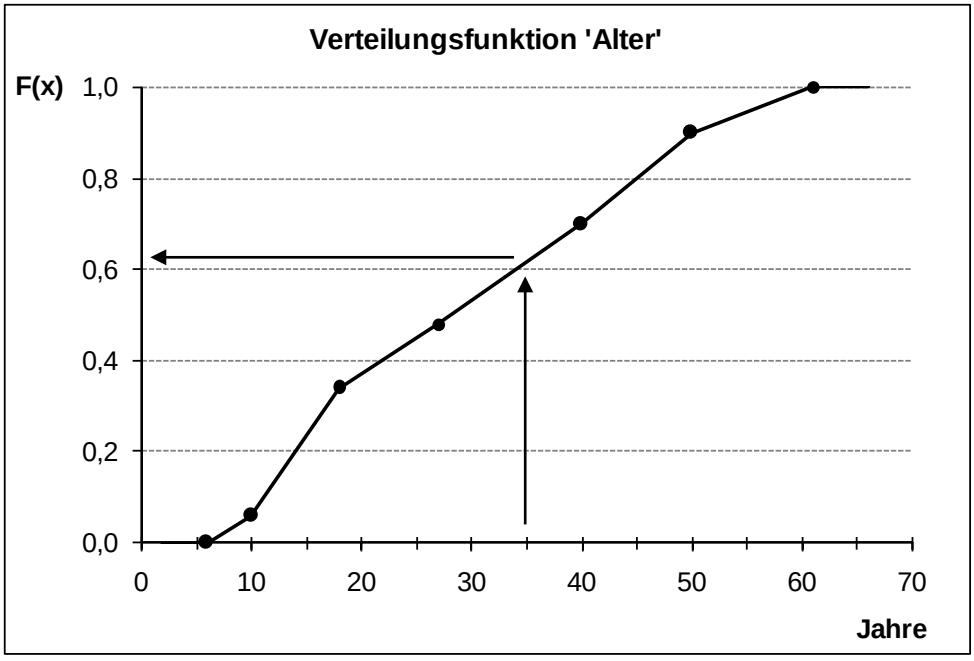
4. Eindimensionale Häufigkeitsverteilung

absolute und relative Summenhäufigkeiten

Für klassierte Daten lässt sich durch Kumulierung ebenfalls die Summenhäufigkeitsfunktion bilden. Allerdings ist deren exakte Bestimmung nur für die Klassengrenzen x_i^u und x_i^o möglich, da unbekannt ist, welche Merkmalswerte die einzelnen Elemente innerhalb einer Klasse besitzen.
Unter der Annahme, dass sich die Merkmalsausprägungen möglichst gleichmäßig über eine Klasse verteilen, kann jedoch von einem kontinuierlichen Anstieg der Summenhäufigkeiten innerhalb einer Klasse ausgegangen werden. Zur grafischen Darstellung der Verteilungsfunktion sind daher die bekannten Summenhäufigkeiten an den Klassengrenzen linear miteinander zu verbinden, wodurch sich ein stetiger Linienzug (Summenpolygon) ergibt.

Beispiel:

Betrachten wir auch hier das Merkmal 'Alter' (F9: Alte) in der bereits bekannten klassierten Form. Die relativen Summenhäufigkeiten am Ende der jeweiligen Klasse können wir direkt aus der Arbeitstabelle im vorangegangenen Beispiel ablesen. Sie sind in der Grafik als Punkte markiert und werden - wie oben ausgeführt - nur noch linear miteinander verbunden:



Analog zur Verteilungsfunktion gruppierter Merkmale lässt sich für jeden Merkmalswert von X (Alter) die zugehörige relative Summenhäufigkeit - zumindest ungefähr - ablesen. Den eingezeichneten Pfeilen folgenden ergibt sich demnach für ein Alter von 35 Jahren eine relative Summenhäufigkeit von etwa 0,62. Somit sind 62% der befragten Besucher der Jugendherberge maximal 35 Jahre alt.
Die Verteilungsfunktion zeigt auch Geradenabschnitte, die sich in ihrem Anstieg deutlich unterscheiden. Während dieser Anstieg in der ersten und dritten Klasse nahezu gleich ausfällt, sticht der vergleichsweise hohe Anstieg in der zweiten Klasse hervor, in der sechsten und letzten Klasse fällt er dagegen am geringsten aus. Diese Unterschiede zwischen den Klassen korrespondiert mit den oben bereits diskutierten Unterschieden in den relativen Häufigkeitsdichten, was wiederum nicht weiter verwundern sollte, da ja der Quotient aus relativer Häufigkeit und Klassenbreite genau der Steigung $\Delta y/\Delta x$ der einzelnen Geradenabschnitte entspricht.

4. Eindimensionale Häufigkeitsverteilung

Mit Hilfe dieser Überlegungen lässt sich die Verteilungsfunktion auch formal exakt beschreiben. Sie setzt sich aus den Geradengleichungen der verschiedenen Klassen zusammen:

$$F(x) = \begin{cases} 0 & \text{für } x < 6 \\ 0 + \frac{0{,}06}{4} * (x - 6) & \text{für } 6 \le x < 10 \\ 0{,}06 + \frac{0{,}28}{8} * (x - 10) & \text{für } 10 \le x < 18 \\ 0{,}34 + \frac{0{,}14}{9} * (x - 18) & \text{für } 18 \le x < 27 \\ 0{,}48 + \frac{0{,}22}{13} * (x - 27) & \text{für } 27 \le x < 40 \\ 0{,}70 + \frac{0{,}20}{10} * (x - 40) & \text{für } 40 \le x < 50 \\ 0{,}90 + \frac{0{,}10}{11} * (x - 50) & \text{für } 50 \le x < 61 \\ 1 & \text{für } x \ge 61 \end{cases}$$

Versuchen wir nun die grafische Lösung mit Hilfe der exakten Verteilungsfunktion zu überprüfen, so errechnet sich der Anteil der Besucher mit einem Alter von maximal 35 Jahren wie folgt:

$$f(x \le 35) = F(35) = 0{,}48 + \frac{0{,}22}{13} * (35 - 27) = 0{,}615$$

Aus dieser beispielhaften Darstellung der Verteilungsfunktion lässt sich dann wieder problemlos deren allgemeine Definition ableiten.

Definition: Verteilungsfunktion F(x) _ klassiert

Gegen sei ein klassiertes Merkmal X. Unter der Annahme gleich verteilter Merkmalsausprägungen innerhalb einer Klasse ergibt sich die Verteilungsfunktion F(x) aus einer linearen Verbindung der relativen Summenhäufigkeiten an den Klassenobergrenzen x_i^o bzw. an den Klassenuntergrenzen x_i^u:

$$F(x) = \begin{cases} 0 & \text{für } x < x_1^u \\ F(x_i^u) + \frac{f_i}{\Delta x_i} * (x - x_i^u) & \text{für } x_i^u \le x < x_i^o \\ 1 & \text{für } x \ge x_k^o \end{cases}$$

5. Lageparameter

Tabellen und Diagramme geben einen Überblick und eine erste Auskunft über das Wesen der Häufigkeitsverteilung in dem Beobachtungsmaterial. Besonders bei einer umfangreichen Datenmasse ist nicht vorrangig die Angabe der vollständigen Häufigkeitsverteilung interessant, sondern die Gesamtheit der beobachteten empirischen Daten in einer stärker verdichteten Form. Gesucht wird nach einigen wenigen Zahlenwerten, die das Beobachtungsmaterial so komprimieren, dass auf eine detaillierte Wiedergabe aller erhobenen Zahlenwerte verzichtet werden kann.

Diese repräsentativen Maßzahlen sollen die Struktur der Häufigkeitsverteilung so charakterisieren, dass alle Besonderheiten deutlich hervorgehoben werden und zugleich Vergleiche mit anderen Verteilungen ermöglicht werden. Hierzu zählen auch die Lageparameter, die in diesem Kapitel vorgestellt werden. Ihre Aufgabe ist es, das Zentrum der Merkmalswerte einer statistischen Masse zu lokalisieren.

5.1. Modus (Modalwert)

Der Modalwert einer Verteilung entspricht der Merkmalsausprägung, die am häufigsten auftritt. Er kann bei allen Merkmalen unabhängig von ihrem Skalenniveau berechnet werden, hat andererseits jedoch eine vergleichsweise geringe Aussagekraft.

Definition: Modalwert x_D

Der Modalwert x_D eines gruppierten Merkmals ist derjenige Merkmalswert x_i, für den die relative Häufigkeit f_i ihr Maximum annimmt:

$$x_D = x_i \quad \text{mit} \quad f_i \rightarrow \text{Max.!}$$

Erfüllen mehrere Ausprägungen diese Bedingung, so ist der Modalwert nicht eindeutig. Es ist in diesem Fall nicht sinnvoll, ihn als Lagemaß zu verwenden.

Der Modalwert x_D eines klassierten Merkmals entspricht der Mitte der modalen Klasse. Die modale Klasse ist diejenige mit der größten Häufigkeitsdichte f_i^*:

$$x_D = \frac{x_i^u + x_i^o}{2} = x_i' \quad \text{mit} \quad f_i^* \rightarrow \text{Max.!}$$

5.2. Median (Zentralwert)

Der Median ist die in der Mitte liegende Merkmalsausprägung eines geordneten Datensatzes. Er besitzt die Eigenschaft, dass oberhalb und unterhalb von ihm gleich viele Merkmalswerte liegen, d.h. er halbiert eine der Größe nach geordnete Rangfolge von Merkmalswerten. Wegen der notwendigen Rangfolge sind zur Bestimmung des Medians mindestens ordinal skalierte Merkmale erforderlich.

Definition: Median x_Z _ gruppiert

Gegeben sei ein gruppiertes, mindestens ordinal skaliertes Merkmal X. Der Median x_Z ist dann der zentrale Wert einer geordneten Folge vom Merkmalswerten und wird in Abhängigkeit vom Umfang N der statistischen Masse wie folgt berechnet:

$$x_Z = (1\text{-nka}) \cdot x_{(k)} + \text{nka} \cdot x_{(k+1)}$$

mit: k = ganze Zahl ((N+1) · 0,5)

nka = Nachkommaanteil ((N+1) · 0,5)

$x_{(k)}$ symbolisiert dabei die Merkmalsausprägung an der k-ten Stelle einer geordneten Folge von Ausprägungen.

Beispiel:

Wenn die eingangs vorgestellte Gästebefragung als Diskussionsgrundlage einer strategischen Neuausrichtung der Jugendherberge dienen soll, so sind gerade die Fragen nach der Zufriedenheit ihrer Gäste von hervorgehobener Bedeutung. Es bietet sich daher an, bspw. einen genaueren Blick auf die Bewertung der Einrichtung und Ausstattung (F5a: EiAu) der Jugendherberge zu werfen; dies umso mehr, als es sich dabei um ein ordinal skaliertes Merkmal handelt. Die Voraussetzung für eine sinnvolle Berechnung des Medians ist damit erfüllt.

x_i (EiAu)	h_i	f_i	H_i	F_i
3	4	0,08	4	0,08
4	26	0,52	30	0,60
5	20	0,40	50	1,00
∑	**50**	**1,00**		

Der vorliegenden Tabelle ist zu entnehmen, dass sehr schlechte (1) und schlechte (2) Bewertungen nicht vergeben wurden. Aufgrund weniger diskreter Ausprägungen wurde das Merkmal 'EiAu' in gruppierter Form aufbereitet, bei der Bestimmung des Medians kann daher nach der obigen Definition vorgegangen werden.

Die statistische Masse hat in unserem Beispiel den Umfang von 50 Einheiten, es ist also nach der mittleren von insgesamt 50 Merkmalsausprägungen zu suchen. Die Mitte zwischen der ersten und fünfzigsten Merkmalsausprägung liegt bei 25,5; also einer Position, der keine Merkmalsausprägung zugeordnet ist. Gemäß obiger Definition werden in die-

sem Fall die beiden Merkmalsausprägungen ausgewählt, die rechts und links von der Mitte liegen; in unserem Beispiel also die an der 25. und 26. Stelle.:

k = ganze Zahl (51·0,5) = ganze Zahl (25,5) = 25
nka = Nachkommaanteil (25,5) = 0,5

$x_z = (1\text{-}0{,}5) \cdot x_{(25)} + 0{,}5 \cdot x_{(26)} = 0{,}5 \cdot 4 + 0{,}5 \cdot 4 = 4$

Die Codierung 4 entspricht im vorliegenden Fragebogen in etwa einer guten Bewertung. 50% der Gäste bewerten also Einrichtung und Ausstattung der Jugendherberge maximal bzw. mindestens mit gut.

Da die '4' im vorliegenden Beispiel keiner Zahl im mathematischen Sinne, sondern der Codierung einer ordinalen Merkmalsausprägung (gut) entspricht, ist die hier durchgeführte Berechnung von Durchschnitten eigentlich nicht zulässig. Sie erfordert metrisch skalierte Merkmale. Da jedoch gleiche Merkmalsausprägungen vorlagen, konnte dieses Problem einfach umgangen werden.

Liegt das Zahlenmaterial in klassierter Form vor, so lässt sich der Median x_Z mit Hilfe der Verteilungsfunktion F(x) aus der Beziehung $F(x_Z) = 0{,}5$ ableiten. Der Median entspricht somit genau der Merkmalsausprägung, die von 50% aller Merkmalsträger maximal erreicht wird. Hierbei ist zuerst die Einfallsklasse i des Medians (= Medianklasse) festzustellen. Danach erfolgt die Feinjustierung über die Verteilungsfunktion.

Definition: Median x_Z _ klassiert

Gegeben sei ein klassiertes, mindestens ordinal skaliertes Merkmal X. Der Median x_Z entspricht dann derjenigen Merkmalsausprägung, für die die Verteilungsfunktion F(x) den Wert von 0,5 annimmt:

$$F(x_z) = F(x_i^u) + \frac{f_i}{\Delta x_i} * (x_z - x_i^u) = 0{,}5$$

$$\Updownarrow$$

$$x_z = x_i^u + \frac{0{,}5 - F(x_i^u)}{f_i} * \Delta x_i$$

Beispiel:

Greifen wir an dieser Stelle kurz das klassierte Merkmal 'Alter' aus Abschnitt 4.2. auf. Aus der dortigen Tabelle ist abzulesen, dass die gesuchte Ausprägung in der vierten Klasse, also zwischen 27 und unter 40 Jahren liegen muss, da in dieser Klasse die Verteilungsfunktion den Wert von 0,5 erreicht (und zwar relativ früh nach dem Beginn dieser Klasse).

$$x_z = x_4^u + \frac{0{,}5 - F(x_4^u)}{f_4} \cdot \Delta x_4 = 27 + \frac{0{,}5 - 0{,}48}{0{,}22} \cdot 13 = 28{,}182 \text{ [Jahre]}$$

50% der Gäste sind folglich maximal 28,182 Jahre alt.

5.3. Arithmetisches Mittel (Mittelwert)

Der arithmetische Mittelwert ist der gebräuchlichste Lageparameter, er wird häufig einfach nur als Durchschnitt bezeichnet. Das arithmetische Mittel wird aus der Summe aller Merkmalsausprägungen einer statistischen Masse berechnet und durch deren Umfang N dividiert. Wegen der additiven Verknüpfung der Merkmalswerte ist das arithmetische Mittel auch nur für metrisch skalierte Merkmale geeignet.
Im Gegensatz zu den anderen Lageparametern kann der arithmetische Mittelwert auch direkt aus den Originalwerten der Urliste errechnet werden. Aufsteigend sortierte Merkmalsausprägungen sind hier nicht erforderlich, da diese Werte summiert werden. In welcher Reihenfolge die Merkmalsausprägungen addiert werden, spielt dabei ja keine Rolle. Wenn jedoch bereits vorab eine Häufigkeitsverteilung erstellt wurde, macht es natürlich Sinn, aus Zeitgründen gleich hierauf zuzugreifen. In Abhängigkeit von der Aufbereitung des Datenmaterials ergeben sich daher mehrere Rechenvarianten.

Definition: arithmetischer Mittelwert $\bar{x}$

Gegeben sei ein metrisch skaliertes Merkmal X. Der arithmetische Mittelwert $\bar{x}$ errechnet sich, indem die Summe aller Merkmalswerte durch die Anzahl aller Merkmalsträger N dividiert wird.

- bei (unsortierten) Originalwerten:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- bei gruppierten Daten:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i \cdot h_i = \sum_{i=1}^{k} x_i \cdot f_i$$

- bei klassierten Daten:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i' \cdot h_i = \sum_{i=1}^{k} x_i' \cdot f_i$$

Bei klassierten Merkmalen wird im Vergleich zu gruppierten Daten die konkrete Merkmalsausprägung einer Gruppe durch die Klassenmitte x_i' der jeweiligen Klasse ersetzt.

Das arithmetische Mittel besitzt dabei vier wesentliche Eigenschaften:

a) Die Summe der Abweichungen der Einzelwerte von ihrem arithmetischen Mittel ist gleich Null:

$$\sum_{i=1}^{N}(x_i - \bar{x}) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \bar{x} = N \cdot \frac{1}{N} \cdot \sum_{i=1}^{N} x_i - N \cdot \bar{x} = N \cdot \bar{x} - N \cdot \bar{x} = 0$$

b) Die Summe der Abweichungsquadrate der Einzelwerte von ihrem arithmetischen Mittel ist kleiner als von einem beliebigen anderen Wert:

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 < \sum_{i=1}^{N}(x_i - x^*)^2 \quad \text{mit} \quad x^* \neq \bar{x}$$

c) Werden die Einzelwerte einer linearen Transformation $x_i^* = a + b \cdot x_i$ (z.B. Umrechnung von EUR in $) unterzogen, so unterliegt das arithmetische Mittel der gleichen Transformation wie die Einzelwerte.

d) Ist eine Grundgesamtheit in mehrere Teilgesamtheiten aufgeteilt, dann ergibt sich das arithmetische Mittel der Grundgesamtheit aus den gewogenen arithmetischen Mitteln der Teilgesamtheiten (Additionssatz für Mittelwerte):

$$\bar{x} = \frac{N_1 \overline{x_1} + N_2 \overline{x_2}}{N_1 + N_2}$$

Beispiel:

Bei der Anwendung des Medians im vorletzten Beispiel wurde bereits auf die Bedeutung der Kundenzufriedenheit im Zuge der strategischen Ausrichtung von Unternehmen (Stichwort: Kundenanalyse) hingewiesen. Wir sehen uns daher das Merkmal 'Gesamtbewertung der Jugendherberge' (F6: Gbew) etwas genauer an. Das Merkmal ist auf einer Skala von 0 bis 50 Punkten erfasst und daher im Gegensatz zu den Einzelbewertungen der Fragen 5a bis 5d metrisch-verhältnis skaliert. Und metrisch skalierte Merkmale sind ja die Voraussetzung für eine sinnvolle Berechnung arithmetischer Mittelwerte.

x_i (Gbew)	h_i	f_i	H_i	F_i	$x_i \cdot fi$
40	1	0,02	1	0,02	0,800
41	5	0,10	6	0,12	4,100
42	4	0,08	10	0,20	3,360
43	4	0,08	14	0,28	3,440
44	7	0,14	21	0,42	6,160
45	9	0,18	30	0,60	8,100
46	11	0,22	41	0,82	10,120
47	5	0,10	46	0,92	4,700
48	1	0,02	47	0,94	0,960
49	1	0,02	48	0,96	0,980
50	2	0,04	50	1,00	2,000
∑	**50**	**1,00**			44,720

Da nur Bewertungen im höheren Punktebereich abgegeben wurden, ist eine gruppierte Datenaufbereitung noch sinnvoll. Aus der erweiterten Arbeitstabelle lässt sich der arithmetische Mittelwert dann problemlos berechnen:

5. Lageparameter

$$\bar{x} = \sum_{i=1}^{k} x_i \cdot f_i = 44{,}720\ [\text{Punkte}]$$

Alternativ hätte der arithmetische Mittelwert auch direkt aus den unsortierten Werten der Urliste (vgl. Tabelle 1) ermittelt werden können:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{50} \cdot (47 + 45 + 43 + ... + 44 + 41 + 43) = \frac{1}{50} \cdot 2236 = 44{,}720\ [\text{Punkte}]$$

Natürlich muss sich auf beiden Rechenwegen das gleiche Ergebnis einstellen. Die Jugendherberge wird danach von allen befragten Gästen im Durchschnitt mit 44,72 Punkten bewertet, was auf einer Skala bis maximal 50 Punkten einem fast sehr guten Ergebnis entspricht.

Interessant für die Jugendherbergsleitung ist allerdings nicht nur eine allgemeine Auswertung dieses Merkmals, sondern auch eine differenzierte Betrachtung nach den fünf Gästearten. Die Gästearten (vgl. Frage 3) Schulklasse, Junior, Gruppe/Verein, 27P und Familie bilden zumindest grobe Zielgruppen und damit die Basis einer differenzierten strategischen Ausrichtung des Unternehmens.

Wir nehmen daher in einem zweiten Schritt eine Auswertung der Gesamtbewertung getrennt für jede dieser fünf Gästearten vor:

nur Schulklasse

x_i (Gbew)	h_i	f_i	H_i	F_i	$x_i \cdot fi$
42	2	0,118	2	0,118	4,956
44	3	0,176	5	0,294	7,744
45	3	0,176	8	0,471	7,920
46	5	0,294	13	0,765	13,524
47	1	0,059	14	0,824	2,773
49	1	0,059	15	0,882	2,891
50	2	0,118	17	1,000	5,900
∑	**17**	**1,000**			45,708

nur Junior

x_i (Gbew)	h_i	f_i	H_i	F_i	$x_i \cdot fi$
41	1	0,200	1	0,200	8,200
42	1	0,200	2	0,400	8,400
44	1	0,200	3	0,600	8,800
45	1	0,200	4	0,800	9,000
46	1	0,200	5	1,000	9,200
∑	**5**	**1,000**			43,600

5. Lageparameter

nur Gruppe/Verein

x_i (Gbew)	h_i	f_i	H_i	F_i	$x_i \cdot fi$
41	2	0,133	2	0,133	5,453
42	1	0,067	3	0,200	2,814
43	1	0,067	4	0,267	2,881
44	3	0,200	7	0,467	8,800
45	2	0,133	9	0,600	5,985
46	4	0,267	13	0,867	12,282
47	2	0,133	15	1,000	6,251
∑	**15**	**1,000**			44,466

nur 27P

x_i (Gbew)	h_i	f_i	H_i	F_i	$x_i \cdot fi$
40	1	0,167	1	0,167	6,680
41	1	0,167	2	0,333	6,847
43	2	0,333	4	0,667	14,319
45	1	0,167	5	0,833	7,515
47	1	0,167	6	1,000	7,849
∑	**6**	**1,001**			43,210

nur Familie

x_i (Gbew)	h_i	f_i	H_i	F_i	$x_i \cdot fi$
41	1	0,143	1	0,143	5,863
43	1	0,143	2	0,286	6,149
45	2	0,286	4	0,571	12,870
46	1	0,143	5	0,714	6,578
47	1	0,143	6	0,857	6,721
48	1	0,143	7	1,000	6,864
∑	**7**	**1,001**			45,045

Die arithmetischen Mittelwerte für die Gesamtbewertung in Abhängigkeit von der jeweiligen Gästeart können nun direkt aus den 5 erweiterten Häufigkeitstabellen entnommen werden:

$\overline{x}$ (Gbew|Schulklasse) = 45,708 [Punkte]

$\overline{x}$ (Gbew|Junior) = 43,600 [Punkte]

$\overline{x}$ (Gbew|Gruppe/Verein) = 44,466 [Punkte]

$\overline{x}$ (Gbew|27P) = 43,210 [Punkte]

$\overline{x}$ (Gbew|Familie) = 45,045 [Punkte]

Auch hier könnten die Mittelwerte alternativ aus den unsortierten Werten der Urliste bestimmt werden und nicht nach der Formel für gruppierte Daten gemäß Arbeitstabellen. Insbesondere für die Teilmenge der Junioren würde sich dies anbieten, da jede Merkmalsausprägung eh nur einmal auftritt und eine tabellarische Erfassung keinen Zusatznutzen stiftet. Ähnlich verhält es sich für die Teilmenge der Senioren (27P) und Familien.

Welche Rückschlüsse lassen sich nun aus den gewonnenen Ergebnissen ziehen? Bezogen auf die fünf Gästearten halten sich die Bewertungsunterschiede in Grenzen. Positiv zu bewerten ist sicherlich, dass die Schulklassen mit durchschnittlich 45,708 Punkten die höchste Bewertung abgeben. Die Schulklassen sind unter den Gästen dieser Jugendherberge am stärksten vertreten und stellen daher eine interessante Zielgruppe dar.

Allerdings sollten wir mit einer abschließenden Aussage zurückhaltend sein, die Bewertungsunterschiede zwischen den Gästearten könnten möglicherweise zufallsbedingt sein und uns dann zu voreiligen Schlussfolgerungen verleiten. Wir werden daher an späterer Stelle nochmals auf diesen Sachverhalt zurückkommen, um dann mit Hilfe des Instrumentariums der induktiven Statistik eine gesichertere Antwort geben zu können.

Exkurs: Geometrisches Mittel

Sind Merkmalsausprägungen nicht additiv, sondern multiplikativ verknüpft, dann muss anstelle des arithmetischen Mittels das geometrische Mittel zur Berechnung eines sinnvollen Durchschnittswertes verwendet werden.

Definition: geometrisches Mittel x_G

Gegeben sei ein metrisch skaliertes Merkmal X. Sind dessen (nicht negative) Merkmalsausprägungen multiplikativ verknüpft, wird das geometrische Mittel x_G als adäquater Mittelwert berechnet, indem aus dem Produkt der N Merkmalsausprägungen die N-te Wurzel gezogen wird:

$$x_G = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$$

Beispiel:

Ein einfaches Beispiel soll diesen Sachverhalt verdeutlichen: Ein Anleger unterhält ein Wertpapierportfolio im Kurswert von 100.000 Euro. Nach einer erfreulichen Wertsteigerung um 20% im ersten Jahr erleidet er im zweiten Jahr einen Rückschlag um 20%. Die Wertsteigerung um 20% entspricht einem Wachstumsfaktor von 1,2; der entsprechende Rückgang im nachfolgenden Jahr dann einem Faktor von 0,8.

Jahr	Wachstumsrate	Wachstumsfaktor	Kapital-Endbestand
0			100000 EUR
1	20 %	1 + 0,2 = 1,2	120000 EUR
2	-20 %	1 - 0,2 = 0,8	96000 EUR

Arithmetisch errechnet sich hieraus für das Merkmal Wachstumsfaktor ein Durchschnittswert von 1, der Wert des Portfolios hätte sich demnach per Saldo nicht verändert. Tatsächlich ist sein Wert aber auf 100000 Euro · 1,2 · 0,8 = 96000 Euro gefallen, woran auch die multiplikative Verknüpfung der Merkmalswerte sichtbar wird.

Für das obige Beispiel beträgt also der durchschnittliche jährliche Wachstumsfaktor auf der Grundlage des geometrischen Mittels

$$x_G = \sqrt[2]{1{,}20 * 0{,}80} = 0{,}9798$$

was einer durchschnittlichen jährlichen Wachstumsrate von

$\overline{wr}$ = 0,9798-1 = -0,0202 = -2,02 %

und damit dem Rückgang des eingesetzten Kapitals von 100000 EUR auf

100000 · 0,9798 · 0,9798 = 96000 EUR

entspricht.

Alternativ kann die durchschnittliche jährliche Wachstumsrate auch über den Quotienten von End- und Anfangsniveau bestimmt werden, da dieser Quotient dem Produkt der Wachstumsfaktoren entspricht:

$$x_G = \sqrt[2]{\frac{96000}{100000}} = \sqrt[2]{1{,}20 * 0{,}80} = 0{,}9798$$

6. Streuungs-Parameter

Mittelwerte sind die in der Praxis am häufigsten verwendeten Verteilungsparameter. Sie liefern allerdings nur eine sehr unvollständige Beschreibung einer Häufigkeitsverteilung, da sie keine Aussage über die Abweichung der einzelnen Merkmalswerte von dem betreffenden Lageparameter geben.
Maßzahlen, die eine Aussage über die Abweichung der einzelnen Merkmalswerte von ihrem Lageparameter zulassen, werden als Streuungsparameter bezeichnet. Sie lassen sich allerdings nur für metrische Merkmale sinnvoll interpretieren.

6.1. Spannweite (Range)

Eine recht grobe Maßzahl, die vor allem bei statistischen Massen kleinen Umfangs angewandt werden kann, ist die Spannweite R. Sie misst den Abstand zwischen der größten und der kleinsten Merkmalsausprägung.

Definition: Spannweite R

Gegeben sei ein metrisch skaliertes Merkmal X. Die Spannweite einer Häufigkeitsverteilung bei gruppierten Daten errechnet sich aus der Differenz zwischen dem größten und dem kleinsten Merkmalswert:

$$R = x_k - x_1$$

Bei klassierten Merkmalen entspricht die Spannweite der Differenz zwischen der Obergrenze der höchsten und der Untergrenze der niedrigsten Klasse:

$$R = x_k^o - x_1^u$$

6.2. Quartilsabstand

Wenn der kleinste oder größte Wert einer Beobachtungsreihe sehr stark von den übrigen Werten abweicht, d.h. ein so genannter Ausreißer ist, dann beeinflusst dieser den Wert der Spannweite sehr stark.

Ein robusterer Streuungsparameter ist dagegen der Quartilsabstand Q. Er gibt die Größe des Bereichs zwischen dem oberen und unteren Quartil an, konzentriert sich also auf die mittleren 50% der Merkmalsträger und blendet damit die möglichen Ausreißer am oberen und unteren Rand der Häufigkeitsverteilung aus. Zur Berechnung des Quartilsabstandes Q ist es daher vorab nötig, das untere und das obere Quartil zu ermitteln.

6. Streuungsparameter

Definition: Quartilsabstand Q

Gegeben sei ein metrisch skaliertes Merkmal X. Der Quartilsabstand bestimmt die Differenz zwischen dem oberen und dem unteren Quartil:

$$Q = x_{0,75} - x_{0,25}$$

Die beiden Quartile wie auch der Median gehören zur großen Familie der Quantile. In Anlehnung an den Median kann allgemein nach dem Merkmalswert gefragt werden, der von einem bestimmten Anteil p aller Merkmalsträger nicht überschritten wird. Dieser Merkmalswert x_p wird als p-Quantil bezeichnet. Das p-Quantil gibt also den Merkmalswert an, der von dem Anteil p aller Merkmalsträger maximal erreicht wird. Analog zur Interpretation des Medians gilt für das p-Quantil folglich, dass $100 \cdot p$ der Beobachtungswerte kleiner oder gleich x_p sind und $100 \cdot (1-p)$ der Beobachtungswerte größer oder gleich x_p.

Das 0,5-Quantil entspricht somit dem Median. Das 0,25-Quantil wird auch unteres (bzw. erstes) Quartil genannt, das 0,75-Quantil entsprechend oberes (bzw. drittes) Quartil. Aus der Differenz von oberem und unterem Quartil ergibt sich dann gemäß obiger Definition der Quartilsabstand.

Nachdem der Median also als Spezialfall aus der Familie der Quantile anzusehen ist, kann auch die Berechnung der Quantile analog zur der des Medians erfolgen. An die Stelle des Anteils von 0,5 tritt verallgemeinert lediglich der Anteil p.

Definition: Quantil x_p

Gegeben sei ein gruppiertes, mindestens ordinal skaliertes Merkmal X. Das Quantil x_p wird in Abhängigkeit vom Produkt $(N+1) \cdot p$ wie folgt berechnet:

$$x_p = (1\text{-nka}) \cdot x_{(k)} + \text{nka} \cdot x_{(k+1)}$$

mit: k = ganze Zahl $((N+1) \cdot p)$

nka = Nachkommaanteil $((N+1) \cdot p)$

Bei einem klassierten, mindestens ordinal skalierten Merkmal X entspricht dann das Quantil x_p derjenigen Merkmalsausprägung, für die die Verteilungsfunktion F(x) innerhalb der Quantilsklasse i den Wert von p annimmt:

$$x_p = x_i^u + \frac{p - F(x_i^u)}{f_i} * \Delta x_i$$

6. Streuungsparameter

Beispiel:

Um Quantile berechnen zu können, müssen die Merkmale mindestens ordinal skaliert sein. Der Quartilsabstand erfordert wegen der Bildung von Differenzen sogar metrisch skalierte Merkmale. Wir greifen daher auf die beiden bereits bekannten Merkmale 'Alter' (F9: Alte) und 'Gesamtbewertung der Jugendherberge' (F6: Gbew) zurück. Beide Merkmale sind metisch (verhältnis) skaliert und deshalb für die Berechnung sowohl von Quantilen als auch des Quartilsabstandes geeignet.

Für das gruppierte Merkmal 'Gesamtbewertung' ergeben sich dann auf der Basis der hier nochmal abgebildeten Häufigkeitsverteilung bzw. Arbeitstabelle die folgenden Quartilswerte:

x_i (Gbew)	h_i	f_i	H_i	F_i	$x_i \cdot fi$
40	1	0,02	1	0,02	0,800
41	5	0,10	6	0,12	4,100
42	4	0,08	10	0,20	3,360
43	4	0,08	14	0,28	3,440
44	7	0,14	21	0,42	6,160
45	9	0,18	30	0,60	8,100
46	11	0,22	41	0,82	10,120
47	5	0,10	46	0,92	4,700
48	1	0,02	47	0,94	0,960
49	1	0,02	48	0,96	0,980
50	2	0,04	50	1,00	2,000
∑	**50**	**1,00**			44,720

- erstes (unteres) Quartil $x_{0,25}$:

  → k = ganze Zahl ((50+1)·0,25) = ganze Zahl (12,75) =12

  nka = Nachkommaanteil (12,75) = 0,75

  $x_{0,25} = (1-0,75) \cdot x_{(12)} + 0,75 \cdot x_{(13)} = 0,25 \cdot 43 + 0,75 \cdot 43 = 43$ [Punkte]

- zweites Quartil $x_{0,50}$ (Median x_z):

  → k = ganze Zahl ((50+1)·0,50) = ganze Zahl (25,5) =25

  nka = Nachkommaanteil (25,5) = 0,50

  $x_z = (1-0,50) \cdot x_{(25)} + 0,50 \cdot x_{(26)} = 0,50 \cdot 45 + 0,50 \cdot 45 = 45$ [Punkte]

- drittes (oberes) Quartil $x_{0,75}$:

  → k = ganze Zahl ((50+1)·0,75) = ganze Zahl (38,25) =38

  nka = Nachkommaanteil (38,25) = 0,25

  $x_{0,75} = (1-0,25) \cdot x_{(38)} + 0,25 \cdot x_{(39)} = 0,75 \cdot 46 + 0,25 \cdot 46 = 46$ [Punkte]

25% der Gäste bewerten die Jugendherberge also mit maximal 43 Punkten und ebenso 25% mit 46 Punkten und mehr. Umgekehrt bewegen sich dann 50% der Gäste in einer Spanne von 43 bis 46 Punkten. Der Quartilsabstand beträgt also 3 Punkte; oder anders formuliert: die 50% der Gäste, die mit ihrer Bewertung in der Mitte der Verteilung liegen, unterscheiden sich dabei um maximal 3 Punkte.

6. Streuungsparameter

In gleicher Reihenfolge können wir nun für das klassierte Merkmal 'Alter' vorgehen:

x_i (Alte)	h_i	f_i	H_i	F_i
[6; 10)	3	0,06	3	0,06
[10; 18)	14	0,28	17	0,34
[18; 27)	7	0,14	24	0,48
[27;40)	11	0,22	35	0,70
[40; 50)	10	0,20	45	0,90
[50; 61)	5	0,10	50	1,00
∑	**50**	**1,00**		

- erstes (unteres) Quartil $x_{0,25}$:

  → Quartilsklasse i = 2, da $F(x_2^u) = 0,06 < 0,25 < F(x_2^o) = 0,34$

  $x_{0,25}$ muss also zwischen 10 und 18 Jahren liegen

$$x_{0,25} = x_2^u + \frac{0,25 - F(x_2^u)}{f_2} \cdot \Delta x_2 = 10 + \frac{0,25 - 0,06}{0,28} \cdot 8 = 15,429 \text{ [Jahre]}$$

- zweites Quartil $x_{0,50}$ (Median x_z):

  → Quartilsklasse i = 4, da $F(x_4^u) = 0,48 < 0,50 < F(x_4^o) = 0,70$

  $x_{0,50}$ muss also zwischen 27 und 40 Jahren liegen

$$x_z = x_4^u + \frac{0,50 - F(x_4^u)}{f_4} \cdot \Delta x_4 = 27 + \frac{0,50 - 0,48}{0,22} \cdot 13 = 28,182 \text{ [Jahre]}$$

- drittes (oberes) Quartil $x_{0,75}$:

  → Quartilsklasse i = 5, da $F(x_5^u) = 0,70 < 0,75 < F(x_5^o) = 0,90$

  $x_{0,75}$ muss also zwischen 40 und 50 Jahren liegen

$$x_{0,75} = x_5^u + \frac{0,75 - F(x_5^u)}{f_5} \cdot \Delta x_5 = 40 + \frac{0,75 - 0,70}{0,20} \cdot 10 = 42,50 \text{ [Jahre]}$$

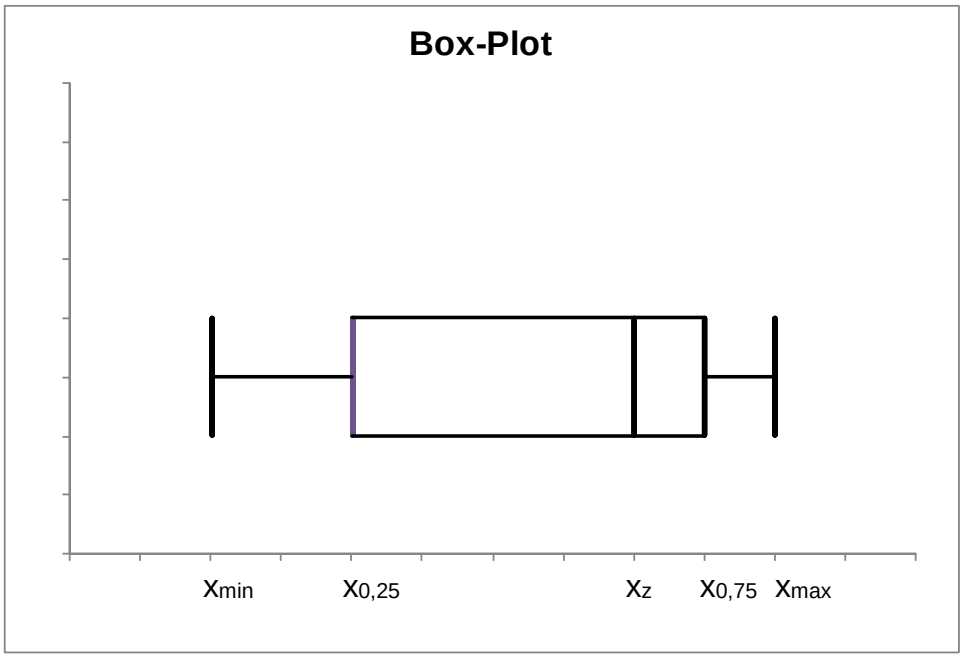
25% der Gäste sind demnach maximal 15,429 Jahre alt und weitere 25% mindestens 42,50 Jahre alt. Demzufolge liegen dann die anderen 50% der Gäste in einem Bereich von 15,429 bis 42,50 Jahren. Der Quartilsabstand beträgt hier 27,071 Jahre, d.h. die 50% der Gäste, die in der Mitte der Altersverteilung liegen, unterscheiden sich dabei um maximal 27, 071 Jahre.

Im Vergleich der beiden Merkmale 'Gesamtbewertung' und 'Alter' fällt (hoffentlich) auf, dass beim Alter der Median fast exakt in der Mitte zwischen dem unteren und oberen Quartil liegt. Bei dem Merkmal 'Gesamtbewertung' ist dies nicht der Fall. Mit seinen 45 Punkten liegt der Median dort näher am oberen Quartil. Die Verteilung fällt in diesem Fall also schiefer aus.

Solche Aspekte zur Verteilung eines Merkmals sollen in dem folgenden Exkurs noch etwas genauer unter die Lupe genommen werden.

Exkurs: Boxplot

Aus dem Vergleich verschiedener Lage- und Streuungsparameter lassen sich - wie oben bereits vorsichtig andiskutiert - wichtige Rückschlüsse auf die Verteilung eines Merkmals ziehen. Als wichtiges Hilfsmittel hierfür dient der sog. Boxplot (auch Box-Whisker-Plot), der das Zusammenwirken dieser Parameter grafisch aufbereitet.
Er stützt sich dabei auf die kleinste Merkmalsausprägung x_{min}, die größte Merkmalsausprägung x_{max}, das erste Quartil $x_{0,25}$, das dritte Quartil $x_{0,75}$ und auf den Median x_z. Im Zentrum der Grafik liegt die namensgebende Box bzw. Rechteck, die vom unteren bis zum oberen Quartil reicht. Darüber hinaus ist mit senkrechten Strichen die kleinste und größte Merkmalsausprägung sowie der Median markiert. Außerdem werden die Striche der kleinsten und größten Merkmalsausprägung durch waagrechte Linien mit der Box verbunden und erhalten so das Aussehen von Antennen (Whisker).



Das Zentrum der Verteilung ist durch den Median gegeben. Den gesamten Wertebereich der Merkmalsausprägungen markieren die kleinste und die größte Merkmalsausprägung. Ihre Differenz entspricht der in Abschnitt 6.1. kurz angesprochenen Spannweite. Die Box selbst zeigt dann den Wertebereich, in dem die mittleren 50% der Merkmalsausprägungen liegen. Die Breite der Box repräsentiert dann den Quartilsabstand.

Aus der genauen Positionierung der fünf Lage- bzw. Streuungsparameter lassen sich nun wichtige Rückschlüsse auf die gesamte Verteilung des Merkmals ziehen, ohne die detaillierte Häufigkeitstabelle kennen zu müssen.

Liegt bspw. der Median exakt in der Mitte der Box, so sind zumindest die mittleren 50% der Merkmalsausprägungen symmetrisch um den Median angeordnet. Sind zudem auch noch beide Antennen des Box-Plots gleich lang, dann gilt die Symmetrie nicht nur für die mittleren 50% der Merkmalsausprägungen, sondern für die gesamte Verteilung. In diesem Fall würden Median und arithmetischer Mittelwert exakt übereinstimmen.

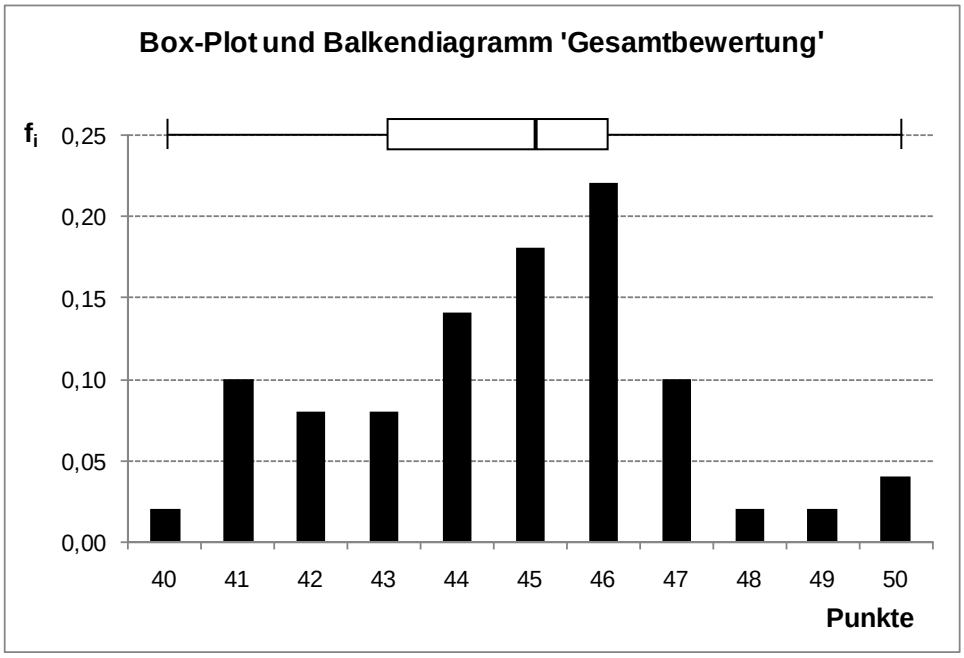
Liegt dagegen der Median näher am oberen Quartil und fällt der obere Whisker wie in der obigen Grafik kurz aus, dann wird von einer schiefen Verteilung gesprochen. In diesem Fall von einer linksschiefen oder rechtssteilen Verteilung, da sich die 25% der Merkmalsausprägungen zwischen Median und oberem Quartil bzw. die 50% der Merkmalsausprägungen zwischen Median und dem Maximalwert (also rechts vom Median) auf 'engerem Raum zusammenquetschen' und folglich ihre relativen Häufigkeiten höher ausfallen. Links vom Median bis zum unteren Quartil verteilen sich die anderen 50% der Merkmalsausprägungen in einem breiteren Intervall. Im Gegensatz zu dem steilen Anstieg der relativen Häufigkeiten im rechten Teil verlaufen die relativen Häufigkeiten im linken Teil relativ flach und verursachen so eine schiefe Verteilung auf der linken Seite (daher linksschief).

Genau umgekehrt verhält es sich, wenn der Median näher am unteren Quartil liegt und der untere Whisker kurz ausfällt. Die hohen relativen Häufigkeiten liegen nun bei den Merkmalsausprägungen links vom Median, die Verteilung ist dann linkssteil bzw. rechtsschief.

Die Schiefe einer Verteilung spiegelt sich auch im Größenvergleich von Median und arithmetischem Mittelwert wider. Bei einer linkssteilen Verteilung bewegen sich die Merkmalsausprägungen unterhalb des Medians in einem engeren Wertebereich als die oberhalb des Medians liegenden. Da die Abstände der Merkmalsausprägungen vom Median bei den größeren Merkmalsausprägungen dann höher ausfallen, fällt auch der arithmetische Mittelwert höher aus als der Median. Bei linkssteilen bzw. rechtsschiefen Verteilungen liegt also der arithmetische Mittelwert rechts vom Median und vice versa.

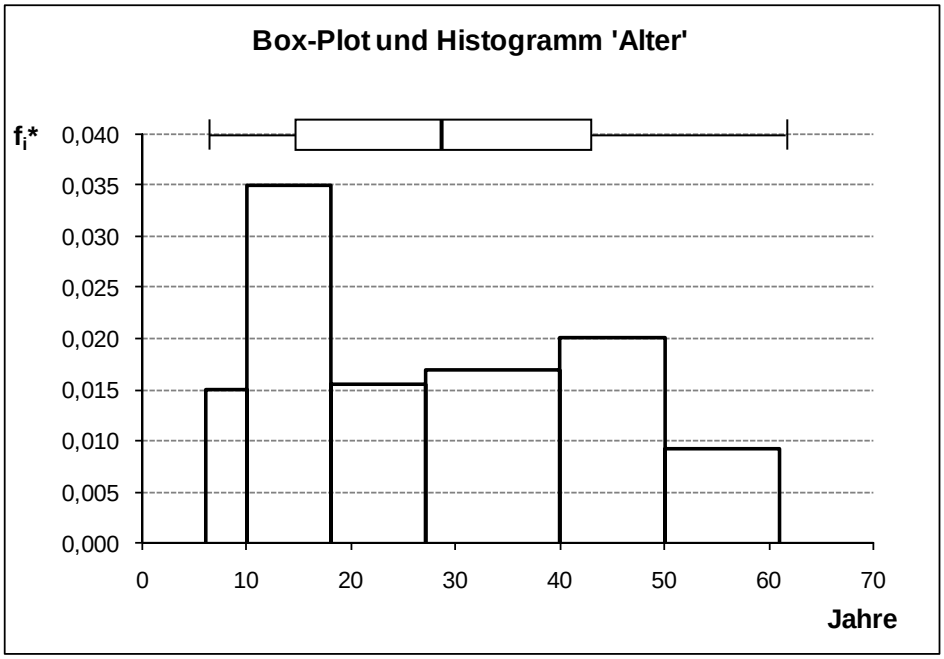
Beispiel:

Um die obigen Ausführungen zum Box-Plot zu konkretisieren, greifen wir die beiden Merkmale 'Gesamtbewertung' (F6: Gbew) und 'Alter' (F9: Alte) hier wieder auf. Zu beiden Merkmalen liegen die notwendigen Parameter bereits vor und können diese direkt in die grafische Darstellung der relativen Häufigkeiten übertragen.



6. Streuungsparameter

Auffällig ist zuallererst, dass das Merkmal 'Gesamtbewertung' durch eine kompakte Verteilung in der Mitte auffällt. Im Vergleich zur Spannweite liegen die mittleren 50% der Merkmalsausprägungen in der Box dicht beieinander. Die Verteilung ist zwar rechtssteil, allerdings nicht eindeutig ausgeprägt. So liegen die oberen 25% der Merkmalsausprägungen in der Box enger zusammen als die unteren 25%. Außerhalb der Box kehrt sich diese Relation jedoch um. Dort ist der untere Whisker etwas kleiner als der obere. Es ist daher nicht sehr überraschend, dass der arithmetische Mittelwert mit 44,72 Punkten zwar links vom Median liegt, allerdings nicht sehr stark abweicht.



Im Vergleich zur Gesamtbewertung sind bei der Verteilung des Merkmals 'Alter' deutliche Unterschiede zu erkennen. So ist zu erkennen, dass die Box auch in Relation zur gesamten Spannweite breiter ausfällt. Die Merkmalsausprägungen sind vor allem in der Mitte breiter gestreut, was ja auch an den beiden nahezu gleich gewichtigen Klassen 3 und 4 abzulesen ist. Daher liegt auch der Median fast exakt in der Mitte der Box. Wären darüber hinaus auch die beiden Antennen des Plots gleich lang, würde sich eine vollständig symmetrische Verteilung des Merkmals 'Alter' ergeben. Der untere Whisker ist jedoch kürzer als der obere. Dadurch stellt sich - am Histogramm auch gut zu sehen - eine leicht linkssteile bzw. rechtsschiefe Verteilung ein. Der arithmetische Mittelwert liegt mit einem Durchschnittswert von 29,47 Jahren etwas rechts vom Median.

Mit Hilfe des Box-Plots ist es somit gelungen, eine grundsätzliche Vorstellung über die Streuung und Verteilung der Merkmalsausprägungen zu bekommen. Nachfolgend werden klassische Streuungsparameter vorgestellt, die die Streuung von Merkmalsausprägungen noch pointierter beschreiben.

6.3. Varianz und Standardabweichung

Varianz und Standardabweichung sind die am häufigsten verwendeten Streuungsparameter. Sie nutzen zum einen das hohe Informationsniveau metrisch skalierter Merkmale optimal aus, zum anderen liefern sie wichtige Informationen für die Durchführung statistischer Testverfahren, auf die dann in späteren Abschnitten immer wieder eingegangen werden soll.

Definition: Varianz s_x^2

Gegeben sei ein metrisch skaliertes Merkmal X. Die Varianz s_x^2 entspricht der mittleren quadratischen Abweichung der Merkmalsausprägungen vom arithmetischen Mittelwert:

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot h_i \quad = \quad \frac{1}{N}\sum_{i=1}^{k} x_i^{\,2} \cdot h_i - \bar{x}^2 \quad = \quad \sum_{i=1}^{k} x_i^{\,2} \cdot f_i - \bar{x}^2$$

Bei klassierten Merkmalen rückt in bekannter Weise an die Stelle der Merkmalsausprägung x_i wieder die Mitte der jeweiligen Klasse x_i':

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i' - \bar{x})^2 \cdot h_i \quad = \quad \frac{1}{N}\sum_{i=1}^{k} x_i'^2 \cdot h_i - \bar{x}^2 \quad = \quad \sum_{i=1}^{k} x_i'^2 \cdot f_i - \bar{x}^2$$

Analog zum arithmetischen Mittelwert kann auch die Varianz aus (unsortierten) Originalwerten einer Urliste errechnet werden:

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^{\,2} - \bar{x}^2$$

Zentraler Bestandteil dieser Terme ist das Quadrat der Merkmalsausprägung, das mit der entsprechenden absoluten oder relativen Häufigkeit multipliziert und über alle Gruppen bzw. Klassen aufsummiert wird. Sinnvoll ist es, diese Berechnungen in einer Arbeitstabelle zu erfassen.

Beispiel:

Wir strapazieren hierzu nochmals die beiden Merkmale 'Gesamtbewertung' (F6: Gbew) und 'Alter' (F9: Alte). Beide Merkmale sind metrisch skaliert und liegen in gruppierter bzw. klassierter Form vor.

Um die in der Definition zur Varianz vorgegebenen Rechenschritte umsetzen zu können, greifen wir auf die an früherer Stelle bereits entwickelten Häufigkeits- und Arbeitstabellen zurück und erweitern diese um die zur Varianzberechnung notwendigen Informationen.

6. Streuungsparameter

x_i (Gbew)	h_i	f_i	H_i	F_i	$x_i \cdot fi$	$x_i^2 \cdot fi$
40	1	0,02	1	0,02	0,800	32,000
41	5	0,10	6	0,12	4,100	168,100
42	4	0,08	10	0,20	3,360	141,120
43	4	0,08	14	0,28	3,440	147,920
44	7	0,14	21	0,42	6,160	271,040
45	9	0,18	30	0,60	8,100	364,500
46	11	0,22	41	0,82	10,120	465,520
47	5	0,10	46	0,92	4,700	220,900
48	1	0,02	47	0,94	0,960	46,080
49	1	0,02	48	0,96	0,980	48,020
50	2	0,04	50	1,00	2,000	100,000
∑	**50**	**1,00**			44,720	2005,200

In der vorletzten Spalte der Tabelle findet sich nochmals die 'Herleitung' des arithmetischen Mittelwertes sowie anschließend die Summe der mit den relativen Häufigkeiten gewichteten Quadrate der Merkmalsausprägungen. Auf Basis dieser Informationen errechnet sich für die Varianz des gruppierten Merkmals 'Gesamtbewertung' ein Wert von

$s_x^2 = 2005{,}200 - 44{,}720^2 = 5{,}322$ Punkte^2.

Analog verfahren wir bei dem Merkmal 'Alter'. Da es sich um ein klassiert aufbereitetes Datenmaterial handelt, muss in der Arbeitstabelle die Klassenmitte als zusätzliche Information bereitgestellt werden:

x_i (Alte)	h_i	f_i	H_i	F_i	x_i'	$x_i' \cdot fi$	$x_i'^2 \cdot fi$
[6; 10)	3	0,06	3	0,06	8,0	0,480	3,840
[10; 18)	14	0,28	17	0,34	14,0	3,920	54,880
[18; 27)	7	0,14	24	0,48	22,5	3,150	70,875
[27;40)	11	0,22	35	0,70	33,5	7,370	246,895
[40; 50)	10	0,20	45	0,90	45,0	9,000	405,000
[50; 61)	5	0,10	50	1,00	55,5	5,550	308,025
∑	**50**	**1,00**				29,470	1089,515

$s_x^2 = 1089{,}515 - 29{,}470^2 = 221{,}034$ Jahre^2

Beide Beispiele verdeutlichen, dass die Varianz die quadratische Dimension der Maßeinheit des Merkmals besitzt, was die Interpretation der Varianz auch im Zusammenhang mit dem arithmetischen Mittelwert erschwert.

In einem zweiten Schritt wird daher die Standardabweichung berechnet. Sie hat als Quadratwurzel aus der Varianz dann die gleiche Dimension wie der arithmetische Mittelwert.

6. Streuungsparameter

Definition: Standardabweichung s_x

Gegeben sei ein metrisch skaliertes Merkmal X. Die Standardabweichung s_x errechnet sich aus der Quadratwurzel der Varianz:

$$s_x = \sqrt{s_x^2}$$

Beispiel:

Nach dieser Formel ergibt sich für das Merkmal 'Gesamtbewertung'

$s_x^2 = 2005{,}200 - 44{,}720^2 = 5{,}322$ Punkte^2

$s_x = 2{,}307$ Punkte

und für das Merkmal 'Alter'

$s_x^2 = 1089{,}515 - 29{,}470^2 = 221{,}034$ Jahre^2

$s_x = 14{,}867$ Jahre.

Die befragten Gäste weichen demnach bei ihrer Gesamtbewertung der Jugendherberge im Durchschnitt um 2,307 Punkte vom Mittelwert in Höhe von 44,72 Punkten ab. Die durchschnittliche Abweichung vom Alters-Mittelwert in Höhe von 29,47 Jahren beträgt 14,867 Jahre.
Auf weitergehende Interpretationen sollte allerdings tunlichst verzichtet werden. So führen häufig Streuungsvergleiche auf Basis der Standardabweichung zu objektiv nicht haltbaren Aussagen. Auch Aussagen dazu, wie viele Merkmalsträger sich im einfachen, zweifachen oder dreifachen Intervall der Standardabweichung um den arithmetischen Mittelwert bewegen, sind ohne ergänzende wahrscheinlichkeitstheoretische Überlegungen nur schwer haltbar.

Analog zum arithmetischen Mittel besitzt auch die Varianz zentrale Eigenschaften:

a) Bei linearer Transformation der Einzelwerte $x_i^* = a + bx_i$ ergibt sich die Varianz zu

$$s^{*2} = \frac{1}{N}\sum_{i=1}^{N}(x_i^* - \overline{x^*})^2 = b^2 s^2$$

b) Setzt sich eine Grundgesamtheit aus zwei Teilgesamtheiten im Umfang N_1 und N_2 und jeweils bekannten Mittelwerten und Varianzen zusammen, so lässt sich hieraus die Varianz der Grundgesamtheit errechnen:

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} + \frac{N_1(\overline{x_1} - \overline{x})^2 + N_2(\overline{x_2} - \overline{x})^2}{N_1 + N_2}$$

6.4. Variationskoeffizient

Ist das arithmetische Mittel einer Verteilung beispielsweise 1000, so wird man eine Standardabweichung von 10 als relativ klein interpretieren, dagegen ist diese als relativ groß anzusehen, wenn das arithmetische Mittel nur 1 beträgt.

Daher erweist es sich als sinnvoll, die Standardabweichung auf das arithmetische Mittel zu beziehen. Da die Standardabweichung die gleiche Maßeinheit wie das arithmetische Mittel besitzt, ist die Verhältniszahl dimensionslos. Sie quantifiziert die relative Streuung eines Merkmals und eignet sich zum Vergleich verschiedener Merkmale und/oder verschiedener statistischer Massen.

Definition: Variationskoeffizient V

Gegeben sei ein metrisch skaliertes Merkmal X. Der Variationskoeffizient V gibt die relative Streuung eines Merkmals wider und errechnet sich aus dem Quotienten von Standardabweichung und arithmetischem Mittelwert:

$$V = \frac{s_x}{\bar{x}}$$

Der Variationskoeffizient eignet sich damit zum Vergleich der Streuungen zweier quantitativer Merkmale. Er ist nicht zu empfehlen, wenn die Merkmalswerte auch negative Werte annehmen können; die Merkmale sollten folglich metrisch-verhältnis skaliert sein.

Beispiel:

Die im bisherigen Beispiel aufgeworfene, aber unbeantwortet gebliebene Frage nach der vergleichsweise geringeren Streuung der beiden Merkmale 'Gesamtbewertung' und 'Alter' lässt sich durch Anwendung des Variationskoeffizienten lösen:

s_x^2 (Gbew) = 2005,200 - 44,720^2 = 5,322 Punkte^2

s_x (Gbew) = 2,307 Punkte

V (Gbew) = 2,307 / 44,72 = 0,052

s_x^2 (Alte) = 1089,515 - 29,470^2 = 221,034 Jahre^2

s_x (Alte) = 14,867 Jahre.

V (Alte) = 14,867 / 29,470 = 0,504

Werte des Variationskoeffizienten von unter 0,2 weisen auf eine vergleichsweise geringe Streuung der Merkmalsausprägungen hin, Werte von über 0,5 auf eine hohe. Beide Merkmale unterscheiden sich in ihrem Streuungsverhalten also erheblich. Erinnern wir uns an ihre beiden Box-Plots. Auch dort konnten wir bereits Anhaltspunkte gewinnen, dass bei der Gesamtbewertung der Jugendherberge äußerst homogene Einschätzungen abgegeben wurden; jetzt konkreter formuliert: die Unterschiede in der Gesamtbewertung belaufen sich im Durchschnitt nur auf 5,2% des Mittelwertes, der bei 44,72 Punkten liegt.

Aufgabe II.1:

Eine Baugesellschaft hat 1250 Wohnungen zu folgenden Konditionen vermietet:

Monatliche Miete in EUR	Anzahl Wohnungen
[100; 200)	150
[200; 300)	275
[300; 350)	525
[350; 500)	175
[500; 650)	125

a) Berechnen Sie die relative Häufigkeitsfunktion und stellen Sie diese grafisch dar.

b) Bestimmen Sie die gesamten jährlichen Mieteinnahmen der Baugesellschaft. Welche Annahme treffen Sie dabei?

c) Berechnen Sie das arithmetische Mittel, den Modus und den Median. Interpretieren Sie die berechneten Werte und geben Sie die Form der hier gegebenen Verteilung an.

d) Wie viel Prozent der Wohnungen erbringen Mieteinnahmen zwischen 250 und 400 EUR? Welche Miethöhe wird von 60 % der Mieter nicht überschritten?

Lösung:

Arbeitstabelle

x_i in Euro	x_i'	Δx_i	h_i	f_i	f_i^*	F_i	$xi' \cdot hi$
[100; 200)	150	100	150	0,12	0,00120	0,12	22500
[200; 300)	250	100	275	0,22	0,00220	0,34	68750
[300; 350)	325	50	525	0,42	0,00840	0,76	170625
[350; 500)	425	150	175	0,14	0,00093	0,90	74375
[500; 650)	575	150	125	0,10	0,00067	1,00	71875
$\sum$			1250				408125

a) siehe f_i in Arbeitstabelle, Grafik als Histogramm mit f_i^*

b) jährliche Mieteinnahmen = $12 \cdot (\Sigma x_i' \cdot h_i) = 12 \cdot 408125 = 4897500$ EUR

c) $\overline{x} = 408125 / 1250 = 326{,}50$ Euro
$x_D = 325$ Euro Modalklasse = 3. Klasse
$x_Z = 300+(0{,}5-0{,}34)/0{,}42 \cdot 50 = 319{,}05$ Euro Medianklasse = 3. Klasse
Mittelwert und Median fast gleich, Tendenz zu symmetrischer Verteilung

d) $f(250<x<400) = F(400) - F(250) = 0{,}577$
$F(400) = 0{,}76+(400-350)/150 \cdot 0{,}14 = 0{,}807$
$F(250) = 0{,}12+(250-200)/100 \cdot 0{,}22 = 0{,}23$
$x_{0,6} = 300+(0{,}6-0{,}34)/0{,}42 \cdot 50 = 330{,}95$ Euro Quantilklasse = 3. Klasse

Aufgabe II.2:

Der Besitzer einer Erdbeerplantage notierte sich in der vergangenen Saison, wie viele Stunden die Sonne täglich auf seine Erdbeeren schien:

Sonnenstunden pro Tag	Anzahl Tage
[0 ; 2)	20
[2 ; 3)	15
[3 ; 5)	20
[5 ; 8)	35
[8 ; 12)	10

a) Beschreiben Sie die statistische Einheit, Merkmal, Ausprägung und Skalierung.

b) Stellen Sie die relativen Häufigkeiten grafisch dar.

c) An wie vielen Tagen der Saison schien die Sonne mindestens 4 Stunden?

d) Wie lange höchstens schien die Sonne an den 40 sonnenärmsten Tagen der Saison?

e) Wie viele Stunden schien die Sonne durchschnittlich pro Tag?

Lösung:

Arbeitstabelle

x_i in Stunden	x_i'	Δx_i	h_i	H_i	f_i	f_i^*	F_i	$x_i' \cdot f_i$
[0; 2)	1	2	20	20	0,20	0,100	0,20	0,200
[2; 3)	2,5	1	15	35	0,15	0,150	0,35	0,375
[3; 5)	4	2	20	55	0,20	0,100	0,55	0,800
[5; 8)	6,5	3	35	90	0,35	0,117	0,90	2,275
[8; 12)	10	4	10	100	0,10	0,025	1,00	1,000
∑			100					4,650

a) Tag / Sonnenstunden pro Tag / 0 … 24 Stunden / metrisch-verhältnis skaliert

b) Grafik als Histogramm mit f_i^* aus Arbeitstabelle

c) h(x>=4) = 100 - h(x<4) = 100 - H(4) = 100 - [35+(4-3)/2 · 20] = 100-45 = 55 Tage

d) $x_{0,4}$ = 3+(0,4-0,35)/0,2 · 2 = 3,5 Stunden Quantilklasse = 3. Klasse

e) siehe letzte Spalte Arbeitstabelle: $\overline{x}$ = 4,65 Stunden

Aufgabe II.3:

Die PKW-Produktion in der deutschen Automobilindustrie hat sich von 5,6 Mio. Fahrzeugen im Jahre 2003 auf 5,85 Mio. im Jahr 2008 erhöht. Die nachfolgende Tabelle zeigt die Aufteilung auf die einzelnen Hersteller:

Hersteller	PKW-Produktion (in 1000)	
	2003	2008
BMW	342	376
Ford	616	622
Mercedes	549	597
Opel	879	896
Sonstige	1713	1790
VW/Audi	1501	1569

a) Welcher Hersteller kann die höchste jährliche Wachstumsrate der PKW-Produktion von 2003 bis 2008 vorweisen? Berechnen Sie diese.

b) Sind die Unterschiede in der PKW-Produktion zwischen den Herstellern von 2003 bis 2008 größer oder kleiner geworden? Berechnen Sie hierzu eine geeignete Kennzahl.

c) Wie würden sich die Ergebnisse unter a) und b) verändern, wenn VW und Audi als zwei getrennte Unternehmen im Verhältnis von 2:1 der gemeinsamen PKW-Produktion agieren würden? Begründen Sie Ihre Antwort, keine Berechnung erforderlich.

Lösung:

a) BMW: Wachstumsfaktor 03-08 = 376/342 = 1,0994
durchschnittlicher Wachstumsfaktor: $x_G = \sqrt[5]{1{,}0994} = 1{,}019$; Wachstumsrate = 1,9%

b) Arbeitstabelle

x_i für 2003	h_i	f_i	$x_i \cdot h_i$	$x_i^2 \cdot h_i$	x_i für 2008	h_i	$x_i \cdot h_i$	$x_i^2 \cdot h_i$
342	1	0,167	342	116964	376	1	376	141376
549	1	0,167	549	301401	597	1	597	356409
616	1	0,167	616	379456	622	1	622	386884
879	1	0,167	879	772641	896	1	896	802816
1501	1	0,167	1501	2253001	1569	1	1569	2461761
1713	1	0,167	1713	2934369	1790	1	1790	3204100
∑	6		5600	6757832	∑	6	5850	7353346

$\overline{x}$ (2003) = 5600 / 6 = 933,333 [1000 Stück] s_x^2(2003) = 255194,844 [1000 Stück]^2
s_x(2003) = 505,168 [1000 Stück] V(2003) = 505,168 / 933,333 = 0,541

$\overline{x}$ (2008) = 975 [1000 Stück]
s_x(2008) = 524,340 [1000 Stück] V(2008) = 0,538
da nahezu konstante Variationskoeffizienten, keine Veränderung

c) Konstante Aufteilung der PKW-Produktion auf VW und Audi im Verhältnis 2:1 lässt die Wachstumsraten unverändert; bei jetzt 7 Unternehmen und konstanter PKW-Produktion sinkt der Mittelwert, die Streuung ebenfalls, der Variationskoeffizient ändert sich kaum.

Aufgabe II.4:

Eine Befragung von 10 Wismarer Studenten nach ihrem Monatseinkommen brachte folgendes Ergebnis:

Einkommen (EUR) 500 400 400 450 500 200 350 450 500 650

a) Stellen Sie die relative Häufigkeitssummenfunktion des Einkommens grafisch dar.

b) Berechnen Sie Median und arithmetischen Mittelwert des Einkommens und interpretieren Sie beide Werte. Welche der beiden Kennzahlen würden Sie im vorliegenden Beispiel vorziehen? Begründen Sie Ihre Antwort.

c) Berechnen Sie Quartilsabstand und Standardabweichung des Merkmals Einkommen und interpretieren Sie auch diese Ergebnisse. Welche Erkenntnisse über die Einkommensverteilung lassen sich hieraus ableiten?

d) Angenommen, jeder der 10 Studenten erzielt monatlich ein um 100 EUR höheres Einkommen. Wie würden sich der Quartilsabstand und Standardabweichung verändern? Begründen Sie Ihre Antwort ausführlich, ohne Berechnung.

Lösung:

Arbeitstabelle nach Gruppierung der Merkmalsausprägungen:

x_i	h_i	f_i	F_i	$x_i \cdot h_i$	$x_i^2 \cdot f_i$
200	1	0,1	0,1	200	4000
350	1	0,1	0,2	350	12250
400	2	0,2	0,4	800	32000
450	2	0,2	0,6	900	40500
500	3	0,3	0,9	1500	75000
650	1	0,1	1,0	650	42250
∑	10			4400	206000

a) Grafik als Treppenfunktion mit F_i

b) $x_z = 0{,}5 \cdot x_{(5)} + 0{,}5 \cdot x_{(6)} = 0{,}5 \cdot 450 + 0{,}5 \cdot 450 = 450$ Euro
mit: k = ganze Zahl $(11 \cdot 0{,}5) = 5$ / nka = Nachkommaanteil (5,5) = 0,5
$\overline{x} = 4400/10 = 440$ Euro (siehe Arbeitstabelle)
Bei metrischen Merkmalen (ohne Ausreißer) ist arithmetischer Mittelwert vorzuziehen.

c) $x_{0,25} = 0{,}25 \cdot 350 + 0{,}75 \cdot 400 = 387{,}5$ Euro mit: k = ganze Zahl (2,75) = 2 / nka = 0,75
$x_{0,75} = 0{,}75 \cdot 500 + 0{,}25 \cdot 500 = 500$ Euro mit: k = ganze Zahl (8,25) = 8 / nka = 0,25
Q = 112,5 Euro; d.h. die Einkommensunterschiede der mittleren 50% Einkommensbezieher betragen maximal 112,50 Euro;
$s_x^2 = 206000 - 440^2 = 12400$ Euro^2 $s_x = 111{,}36$ Euro
Die Einkommen weichen durchschnittlich 111,36 Euro vom Mittelwert mit 440 Euro ab.

d) Die Quartile steigen dann auch um 100 Euro, der Quartilsabstand bleibt folglich konstant. Mit den Merkmalsausprägungen steigt auch der arithmetische Mittelwert um 100 Euro, die Standardabweichung als durchschnittliche Abweichung hiervon bleibt dann konstant.

Aufgabe II.5:

Gehaltszahlungen eines Betriebes:

Gehaltsklasse	Monatsgehalt in EUR	Anzahl der Gehaltsempfänger
1	[700; 750)	200
2	[750; 1050)	180
3	[1050; 1150)	190
4	[1150; 1300)	150
5	[1300; 1400)	120
6	[1400; 1600)	160

a) Berechnen Sie verschiedene Lageparameter.

b) Im Anschluss an einen Lehrgang in EDV erreichen 90 Beschäftigte der dritten Gehaltsklasse die nächst höhere Klasse. Begründen Sie, ob sich der Median verändert und wenn ja, in welche Richtung.

c) Wie verändert sich der Median, wenn 40 Beschäftigte der vorletzten Gehaltsklasse die letzte Klasse erreichen? Berechnen Sie für diesen Fall das arithmetische Mittel.

Lösung:

Arbeitstabelle:

x_i	x_i'	Δx_i	h_i	f_i	f_i^*	F_i	$xi' \cdot fi$
[700; 750)	725	50	200	0,2	0,0040	0,20	145,00
[750; 1050)	900	300	180	0,18	0,0006	0,38	162,00
[1050; 1150)	1100	100	190	0,19	0,0019	0,57	209,00
[1150; 1300)	1225	150	150	0,15	0,0010	0,72	183,75
[1300; 1400)	1350	100	120	0,12	0,0012	0,84	162,00
[1400; 1600)	1500	200	160	0,16	0,0008	1,00	240,00
$\sum$			1000				1101,75

a) x_D = 725 Euro Modalklasse = 1. Klasse
x_Z = 1050+(0,5-0,38)/0,19 · 100 = 1113,16 Euro Medianklasse = 3. Klasse
$\overline{x}$ = 1101,75 Euro (siehe Arbeitstabelle)

b) ja, der Median ändert sich, da die Veränderung über die Medianklasse hinweg erfolgt, der Median wird steigen, die Verteilungsfunktion erreicht jetzt erst zu Beginn der vierten Klasse einen Wert von 0,5, der Median wird also knapp über 1150 Euro liegen

c) siehe b), da keine Veränderung über Medianklasse hinweg, bleibt Median konstant,
$\overline{x}$ (neu) = 1101,75 - 1350 · 0,04 + 1500 · 0,04 = 1107,75 Euro

Aufgabe II.6:

Eine Untersuchung von 40.000 Landwirtschaftsbetrieben zur Größe der landwirtschaftlich genutzten Fläche ergab folgende Häufigkeitstabelle:

Klasse i	landwirtschaftlich genutzte Fläche in ha	Anzahl der Betriebe in 1000
1	[0; 20)	5
2	[20; 35)	6
3	[35; 50)	12
4	[50; 100)	11
5	[100; 300)	6

a) Berechnen Sie die relative Häufigkeitsverteilung sowie die absoluten und die relativen Summenhäufigkeiten.

b) Stellen Sie die relative Häufigkeitsverteilung und die Verteilungsfunktion grafisch dar.

c) Berechnen Sie den Wert der empirischen Verteilungsfunktion an der Stelle x = 200, überprüfen Sie Ihr Ergebnis durch die grafische Bestimmung des Funktionswertes und interpretieren Sie dieses.

d) Bestimmen Sie das erste (untere) Quartil zunächst grafisch, berechnen Sie danach dieses Quartil und interpretieren Sie diesen Wert.

Lösung:

Arbeitstabelle:

x_i	Δx_i	h_i	f_i	f_i^*	H_i	F_i
[0; 20)	20	5	0,125	0,0063	5	0,125
[20; 35)	15	6	0,150	0,0100	11	0,275
[35; 50)	15	12	0,300	0,0200	23	0,575
[50; 100)	50	11	0,275	0,0055	34	0,850
[100; 300)	200	6	0,150	0,0008	40	1,000
$\sum$		40				

a) siehe Arbeitstabelle zu f_i, H_i und F_i

b) Grafik als Histogramm mit f_i^* aus Arbeitstabelle und Grafik als Summenpolygon mit F_i

c) $F(200) = 0{,}85 + 0{,}15/200 \cdot (200\text{-}100) = 0{,}925$

d) $x_{0,25} = 20 + 15/0{,}15 \cdot (0{,}25\text{-}0{,}125) = 32{,}5$ ha

25% aller Landwirtschaftsbetriebe bewirtschaften eine Fläche von maximal 32,5 ha.

Aufgabe II.7:

Zur Planung eines Einkaufszentrums wurden für den in Frage kommenden Standort die Anreisewege der potentiellen Kunden ermittelt:

Entfernung zum Standort in km	Kunden in 1000
[0; 1)	60
[1; 3)	140
[3; 10)	350
[10; 20)	150
[20; 50)	300

a) Stellen Sie die relativen Häufigkeiten in geeigneter Weise grafisch dar. (Histogramm)

b) Wie groß ist die durchschnittliche Entfernung der potentiellen Kunden zu dem Standort? ($\overline{x}$ = 15,335 km)

c) 50% aller potentiellen Kunden sollen angeblich nicht weiter als 5 km vom Standort wohnen. Überprüfen Sie die Behauptung. (F(5) = 0,3)

d) Ermitteln Sie grafisch und rechnerisch den Quartilsabstand und interpretieren Sie diesen. ($x_{0,25}$ = 4 km / $x_{0,75}$ = 25 km)

Aufgabe II.8:

Aus dem Wirtschaftsbericht eines Landkreises sind folgende Informationen zu entnehmen:

Beschäftigte	Anzahl Unternehmen	Gesamtumsatz in Mio. EUR
[20; 50)	200	300
[50; 100)	140	700
[100; 200)	80	600
[200; 500)	50	1150
[500; 1000)	20	1000
[1000; 2600)	10	1500

a) Bestimmen Sie für das Merkmal 'Umsatz pro Unternehmen' den Mittelwert.
($\overline{x}$ = 10,5 Mio. EUR)

b) Stellen Sie die Verteilung der Beschäftigten in geeigneter Weise grafisch dar.
(Histogramm)

c) Ermitteln Sie für das Merkmal 'Beschäftigte' den Median und Quartilsabstand und interpretieren Sie diese.
(x_z = 67, 857 Pers. / $x_{0,25}$ = 38,75 Pers. / $x_{0,75}$ = 143,75 Pers.)

d) Klein- und mittelständische Unternehmen beschäftigen maximal 250 Personen. Wie viele der erfassten Unternehmen würden folglich in diese Kategorie fallen?
(H(250) = 428,333 ≈ 428)

Aufgabe II.9:

An einer Rostocker Hauptverkehrsstraße (zulässige Höchstgeschwindigkeit 50 km/h) wurden bei einer Polizeikontrolle innerhalb von 30 Minuten folgende Geschwindigkeiten registriert:

Geschw. (km/h)	hi
[40; 45)	16
[45; 50)	44
[50; 55)	16
[55; 60)	12
[65; 70)	10
[70; 75)	2

a) Nennen Sie die in der Aufgabe betrachtete statistische Einheit bzw. statistische Masse sowie das Merkmal mit dem zugehörigen Skalenniveau.

b) Stellen Sie die relative Häufigkeitssummenfunktion des Merkmals grafisch dar.

c) Geben Sie die gemessene Durchschnittsgeschwindigkeit an. ($\overline{x}$ = 51,2 km/h)

d) Welche Geschwindigkeit wurde von der Hälfte aller Fahrzeuge nicht überschritten? (x_z = 48,864 km/h)

e) Bei Überschreiten der zulässigen Höchstgeschwindigkeit um mehr als 3 km/h wird ein Bußgeld erhoben. Wie viele Fahrzeuge betrifft das? (h(x>53) = 30,4 ≈ 30)

f) Innerhalb welcher Tempogrenzen liegen die mittleren 50% der Fahrzeuge? ($x_{0,25}$ = 46,023 km/h / $x_{0,75}$ = 54,688 km/h)

Aufgabe II.10:

Bundesschatzbriefe werden von der Deutschen Bundesbank in zwei Varianten angeboten. Bei Typ B werden die Zinsen dem Anlagebetrag zugerechnet und im Folgejahr mitverzinst (Zinseszins-Effekt). Die Laufzeit des Typs B beträgt 7 Jahre. Zum Anfang eines Jahres gilt folgendes Angebot:

Jahr	1.	2.	3.	4.	5.	6.	7.
Zins p.a. in %	2,75	3,25	3,75	4,00	4,25	4,50	5,00

a) Welcher Zinsertrag wird einem Anleger bei einem Anfangskapital von 5.000 EUR nach 7 Jahren ausbezahlt? Wie hoch ist der durchschnittliche jährliche Zinssatz? (Zinsertrag = 1547,04 EUR / $\overline{w}f$ = 1,03926 / $\overline{w}r$ = 0,03926)

b) Alternativ bietet die SPAR-Bank eine vergleichbare 6-jährige Anlage mit einem durchschnittlichen jährlichen Zinssatz von 3,8%. Zu welchem Zinssatz müsste der Anleger seine Einlage von 5.000 EUR nach dem 6. Jahr im 7. Jahr neu investieren, um insgesamt den gleichen Zinsertrag zu erhalten wie bei Typ B? (Zinssatz = 4,69%)

Aufgabe II.11:

Im Rahmen der Cost-Averaging-Methode wird beispielsweise jeden Monat über einen längeren Zeitraum ein fester Betrag in Fondsanteile investiert:

Monat	Anlagesumme in EUR	Preis des Fondsanteils in EUR
Januar	250	20
Februar	250	20
März	250	10
April	250	5
Mai	250	10

Berechnen Sie den durchschnittlichen Preis für einen Fondsanteil. ($\bar{x}$ = 10 EUR)
Wie hoch wäre der durchschnittliche Preis gewesen, wenn in jedem Monat eine konstante Zahl von Fondsanteilen erworben worden wäre? ($\bar{x}$ = 13 EUR)

Aufgabe II.12:

Fast-Food-Essen ist zwar beliebt, befindet sich aber in einem starken Wandel. Der schnelle Imbiss wird edler. Dies ist die Chance junger Gastro-Ketten, die den etablierten Unternehmen die Kunden wegzunehmen versuchen (Umsatzangaben in Mio. EUR):

Unternehmen	Umsatz 2006	Umsatz 2013	Zahl der Filialen 2013
Aral	150	190	1100
Autobahn Tank+Rast	490	600	390
Burger King	460	820	700
IKEA	150	180	50
Lufthansa Service	720	750	50
Mc Donalds	2280	3100	1470
Nordsee	310	290	330
SSP	170	180	380
Subway	170	190	590
Yum	230	230	160

a) Beschreiben Sie die statistische Masse und das Skalenniveau der betrachteten Merkmale.

b) Welches Unternehmen weist für den Zeitraum von 2006 bis 2013 das größte Umsatzwachstum aus? Berechnen Sie hierzu die durchschnittliche jährliche Wachstumsrate. ($x_{G;\ Burger\ King}$ = 8,609%)

c) Erstellen Sie für die Umsätze der beiden Jahre 2006 und 2013 jeweils eine eindimensionale Häufigkeitsverteilung und stellen Sie deren relative Summenhäufigkeiten in einer Grafik dar.

d) Berechnen Sie einen geeigneten Lageparameter für den Umsatz von 2006 und 2013. ($x_{z,\ 2006}$ = 270 Mio. EUR / $x_{z,\ 2013}$ = 260 Mio. EUR)

e) Sind die Umsatzzahlen der 10 Unternehmen für 2013 gegenüber 2006 homogener oder deren Unterschiede größer geworden? Begründen Sie Ihre Antwort durch Berechnung einer geeigneten Kennzahl. Wie lässt sich dieses Ergebnis begründen? (V_{2006} = 1,192 / V_{2013} = 1,301)

Aufgabe II.13:

Die Mitglieder eines Sportverbandes mit ca. 100000 Personen weisen folgende Altersstruktur auf:

Alter	Mitglieder in 1000
[6; 14)	16
[14; 20)	20
[20; 35)	22
[35; 50)	16
[50; 65)	14
ab 65	12

a) Stellen Sie die relativen Häufigkeiten grafisch dar. (Histogramm)

b) Wie viele Mitglieder des Verbandes haben ein Alter von 15 bis unter 25 Jahre? ($h(15 \leq x < 25) = 24$)

c) Wie viele Mitglieder sind genau 25 Jahre alt? ($h(x=25) = 0$)

d) Bestimmen und interpretieren Sie Modalwert, Median und arithmetisches Mittel des Alters der Mitglieder unter der Annahme, dass kein Mitglied 80 Jahre und älter ist. Welchen der Lageparameter würden Sie hier vorziehen? Begründen Sie Ihre Antwort. (x_D = 17 J. / x_z = 29,545 J. / $\overline{x}$ = 34,60 J. / arithmetisches Mittel)

e) Ermitteln Sie grafisch und rechnerisch den Quartilsabstand. (Q = 34,371 J.)

f) Beträgt die relative Abweichung des Alters vom arithmetischen Mittel mehr als 50%? (V = 0,596 = 59,6% / also ja)

Aufgabe II.14:

Für 50 ausgewählte Unternehmen der Bauwirtschaft sind für das abgelaufene Geschäftsjahr folgende Daten zur Gewinnsituation bekannt:

Gewinn in 1000 EUR	Anzahl Unternehmen
[-500; -250)	6
[-250; -50)	8
[-50; 50)	15
[50; 1000)	21

a) Berechnen Sie die relativen Summenhäufigkeiten und zeichnen Sie die Verteilungsfunktion. (Summenpolygon)

b) Wie hoch ist der Anteil der Unternehmen mit Verlust in diesem Geschäftsjahr? ($f(x<0) = 0,43$)

c) Welchen Gewinn erzielte ein Unternehmen im Durchschnitt? ($\overline{x}$ = 151,5 TEUR)

d) Nehmen Sie zur obigen Klasseneinteilung Stellung und machen Sie Verbesserungsvorschläge. (schlechte Einteilung)

e) Begründen Sie (keine Berechnung), wie sich das Ergebnis unter c) verändert, wenn bekannt ist, dass 5 Unternehmen einen Verlust zwischen 0 und 50 TEUR erlitten haben. (da weniger Unternehmen mit Verlust als angenommen, steigt der Mittelwert)

Aufgabe II.15:

Die KarstadtQuelle AG plant eine Bestandsprüfung in ihrem zentralen Auslieferungslager. Um zu gesicherten Aussagen zu gelangen, wurden bspw. die Nachfragemengen von Waschmaschinen jeweils für die letzten 36 Monate ermittelt und in nachfolgender Tabelle zusammengefasst:

Menge in 1000	Anzahl Monate
[150; 170)	4
[170; 190)	7
[190; 200)	7
[200; 210)	7
[210; 230)	7
[230; 250)	4

a) Bestimmen Sie Merkmal, Skalenniveau des Merkmals und statistische Einheit.

b) Stellen Sie die relative Verteilung der Nachfragemengen in geeigneter Weise grafisch dar und beschreiben Sie kurz die vorliegende Verteilung.

c) In wie viel Prozent aller Monate übersteigt die Nachfragemenge 215000 Stück?

d) Wie hoch war die durchschnittliche Nachfrage in den letzten 36 Monaten. Begründen Sie (keine Berechnung!), welchen Wert der Median in diesem Fall annimmt.

e) Das Unternehmen geht auch für die nahe Zukunft von vergleichbaren Nachfragemengen aus. Mit welchem monatlichen Lagerbestand an Waschmaschinen muss geplant werden, wenn die Nachfrage in 95% aller Fälle gedeckt werden soll (= Servicegrad)?

Aufgabe II.16:

Nach Angaben des Instituts der deutschen Wirtschaft stellen sich die tatsächliche und die geschätzte Altersverteilung in Deutschland für die Jahre 1999 und 2030 wie folgt dar:

Alter	Anzahl Personen in Millionen	
in Jahren	1999	2030 (geschätzt)
[0; 15)	12,9	10,4
[15; 21)	5,5	4,4
[21; 40)	22,9	17,0
[40; 60)	21,9	19,8
[60; 100)	18,9	24,0

a) Stellen Sie die relative Häufigkeitssummenfunktion für die Bevölkerung 1999 grafisch dar.

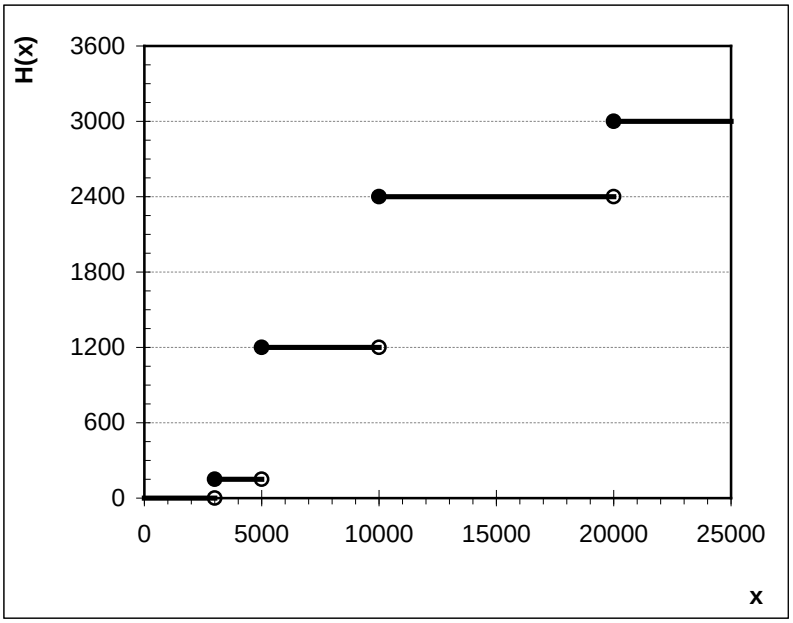
b) Berechnen Sie das Durchschnittsalter der Bevölkerung für 1999 und 2030.

c) Um welchen Prozentsatz wird das Durchschnittsalter von 1999 bis 2030 pro Jahr steigen?

d) Berechnen Sie den Altersquotienten als Anteil der [65;100)-Jährigen an den [15;65)-Jährigen für die beiden Jahre. Auf welcher Annahme basiert diese Berechnung? Auf Diskutieren Sie die Problematik, die sich aus dieser Entwicklung ergibt.

Aufgabe II.17:

Das Autohaus Hansa-Nord erzielte im Geschäftsjahr 2008 einen Absatz (Basis: Verkaufspreis x in Euro) an Fahrzeugen, der durch die folgende absolute Verteilung der Summenhäufigkeiten beschrieben werden kann:



a) Leiten Sie für das Autohaus aus der Grafik die absolute und relative Häufigkeitsverteilung des Fahrzeugpreises ab. Wie ist das Merkmal skaliert und wie ist es aufbereitet?

b) Berechnen Sie den durchschnittlichen Verkaufspreis der Fahrzeuge und ein adäquates Streuungsmaß. Welche Aussage lässt sich hieraus ableiten?

c) Wie hoch ist der Anteil der Fahrzeuge mit einem Verkaufspreis von mindestens 5000 EUR bzw. von über 5000 EUR? Wie erklären sich die unterschiedlichen Ergebnisse?

Aufgabe II.18:

Zu einem Unternehmen mit vier Betriebsstätten liegen folgende Informationen (Umsatz in 1000 EUR) vor:

Betriebsstätte	Umsatz 2000	Umsatz 2006	Umsatz pro Beschäftigten 2006	Umsatzrentabilität 2006
1	200	300	150	0,016
2	260	320	180	0,020
3	280	240	110	0,005
4	320	340	100	-0,006

a) Berechnen Sie für das Gesamtunternehmen den durchschnittlichen Umsatz pro Beschäftigten sowie die durchschnittliche Umsatzrentabilität für 2006.

b) Geben Sie für die vier Betriebe getrennt und für das Gesamtunternehmen an, um wie viel Prozent sich die Umsätze im Jahresdurchschnitt verändert haben.

c) In welchem Jahr wies der Umsatz der Betriebsstätten die größere Streuung auf?

Aufgabe II.19:

Die deutschen Wirtschaftsforschungsinstitute und der Sachverständigenrat schätzen die reale Wachstumsrate des Bruttoinlandsproduktes (BIP) in Deutschland für 20xx wie folgt:

Institut	reale Wachstumsrate des BIP für 20xx in %
DIW	1,4
Ifo	1,8
IfW	1,8
IWH	1,6
HWWA	1,7
RWI	1,8
Sachverständigenrat	1,6

a) Beschreiben Sie die statistische Masse und das Skalenniveau des Merkmals.

b) Berechnen Sie die Summenhäufigkeiten und zeichnen Sie die Verteilungsfunktion.

c) Berechnen Sie Median und arithmetischen Mittelwert für die reale Wachstumsrate. Begründen Sie, welchen der beiden Lageparameter Sie favorisieren würden.

d) Erstellen Sie für das Merkmal einen Box-Plot und erläutern Sie, welche Erkenntnisse sich über die Verteilung des Merkmals hieraus ableiten lassen.

e) Berechnen Sie ein Maß für die relative Streuung des Merkmals. Lassen sich mit diesem Ergebnisse ihre Überlegungen aus d) bestätigen?

Aufgabe II.20:

In einer Studie zur Kursentwicklung des Deutschen Aktienindex (DAX) werden aus den Indexständen zum Monatsende die Renditen der letzten 400 Monate ermittelt. Sie verteilen sich wie folgt:

Monatsrendite des DAX-Index (in %)	Anzahl Monate
[-30; -10)	14
[-10; -2)	92
[-2; 0)	58
[0; 2)	70
[2; 10)	151
[10; 25)	14
[25; 100)	1

a) Bestimmen Sie Merkmal, Skalenniveau des Merkmals und die statistische Masse.

b) Stellen Sie die Verteilungsfunktion grafisch dar und leiten Sie direkt aus der Funktion die Parameter ab, die Sie zur Erstellung des Box-Plots benötigen. Zeichnen Sie diesen auch in die Grafik ein und beschreiben Sie die Verteilung des Merkmals.

c) Berechnen Sie das arithmetische Mittel und den Median der DAX-Monatsrenditen. Welchen der beiden Lageparameter würden Sie vorziehen? Begründen sie Ihre Antwort.

d) Wie hoch ist die Standardabweichung der Monatsrenditen, wenn der Variationskoeffizient 7,599 beträgt. Welche Informationen liefern die Ergebnisse der beiden Kennzahlen? Halten Sie die Berechnung der beiden Kennzahlen in diesem Beispiel für sinnvoll?

Aufgabe II.21:

Die Fritz Bikes GmbH stellt individuelle Fahrräder her. In der Branche ist es üblich, eine Lieferung innerhalb von 5 Tagen nach Auftragserteilung zu garantieren und bei Überschreiten der Frist an den Kunden 20 EUR pro Tag zu zahlen. Bevor sich das Unternehmen an dieser Aktion beteiligt, sollen hierzu folgende Informationen ausgewertet werden:

Lieferzeit in Stunden	Anzahl Aufträge
[48; 72)	18
[72; 96)	61
[96; 120)	88
[120; 144)	49
[144; 168)	27
[168; 192)	7

a) Wie hoch ist die durchschnittliche Lieferzeit?

b) Berechnen Sie eine geeignete Maßzahl zur absoluten Streuung.

c) Wie viel Prozent der erstellten Fahrräder wären nicht rechtzeitig ausgeliefert worden und wie viel Euro hätte das Unternehmen an seine Kunden auszahlen müssen?

d) Wie viele Stunden müsste bei der gegenwärtigen Situation die garantierte Lieferzeit pro Auftrag betragen, damit 70% der Fahrräder rechtzeitig beim Kunden eintreffen?

Aufgabe II.22:

Die Singulus AG hat im Jahr 2015 für Bildaufnahmegeräte folgende Aufträge erhalten:

Anzahl der Aufträge	Wert des Auftrags (in 1000 EUR)
100	[0; 50)
250	[50; 150)
200	[150; 350)
300	[350; 425)
150	[425; 550)

Die Hamatech AG hat dieses Geschäftsfeld erst 2015 bearbeitet und für den Wert der 100 bisher akquirierten Aufträge folgende Kennzahlen ermittelt (in 1000 EUR):

$$x_{min} = 5 \;/\; x_{max} = 500 \;/\; x_{0,25} = 25 \;/\; x_{0,75} = 125 \;/\; x_z = 75 \;/\; \bar{x} = 225 \;/\; s_x = 100$$

a) Bestimmen Sie für die beiden Unternehmen jeweils einen geeigneten Lageparameter und vergleichen Sie beide Ergebnisse.

b) Stellen Sie die Boxplots für die beiden Unternehmen in geeigneter Weise in einer Grafik dar. Welche Rückschlüsse lassen sich hieraus ableiten?

c) Welches der beiden Unternehmen weist für 2015 die geringeren Unterschiede in den Auftragswerten auf? Belegen Sie Ihre Antwort durch die Berechnung einer geeigneten Kennzahl.

d) Für die Hamatech AG wurden zwei weitere Aufträge über 50.000 EUR und 300.000 EUR nachträglich erfasst. Lassen sich die Auswirkungen auf die oben angegebenen Parameter quantifizieren? Bestimmen Sie - soweit möglich - die neuen Werte.

7. Zweidimensionale Häufigkeitsverteilung

Zweidimensionale Häufigkeitsverteilungen werden verwendet, um die Zusammenhänge zwischen zwei Merkmalen von Elementen einer statistischen Masse zu beschreiben. Dabei sind folgende Fragen zu klären:

- Existiert ein Zusammenhang zwischen den beiden Merkmalen?
- Wenn ja, wie stark ist dieser Zusammenhang?
- Lässt sich ein bestehender Zusammenhang funktional beschreiben?

Auch der Blick auf den Fragebogen in Abbildung 1 wirft Fragen nach der Abhängigkeit von Merkmalen auf: Fällt die Zufriedenheit mit der Jugendherberge in Anhängigkeit vom Alter bzw. von der Art der Gäste unterschiedlich aus? Und gilt dies in vergleichbarer Weise auch für die unterschiedliche Nutzung von Informationsquellen? Hängt die Übernachtungsdauer vielleicht auch von der Zufriedenheit der Gäste ab? Wird die Jugendherberge verstärkt von Stammgästen genutzt? Wollen also vor allem die Gäste die Jugendherberge wieder besuchen, die schon mehrmals hier übernachtet haben? Inwieweit beeinflusst die Bewertung von Einrichtung und Ausstattung, von Lage und Umgebung, von Service und Kompetenz bzw. der Verpflegung die Gesamtbewertung der Jugendherberge? Hat diese Bewertung wiederum einen Einfluss auf die Besuchshäufigkeit und die Empfehlungsrate der Jugendherberge?

Die Liste ließe sich fast beliebig verlängern. Damit lassen sich genügend Beispiele finden, um die verschiedenen Verfahren zur Messung solcher bivariater Zusammenhänge vorzustellen. Entscheidend für den Einsatz eines Verfahrens ist auch hier das Skalenniveau der Merkmale. Mit Ausnahme der nachfolgenden Ausführungen zur zweidimensionalen Häufigkeitsverteilung stellen daher alle nachfolgenden Kapitel zu Teil III auf ein bestimmtes Skalenniveau der Merkmale ab.

Im vierten Kapitel wurde die eindimensionale Häufigkeitstabelle als erste Möglichkeit vorgestellt, die Verteilung eines Merkmals übersichtlich aufzubereiten. Wird diese Darstellungsform auf zwei Merkmale übertragen, so müssen in den Zeilen die Merkmalsausprägungen des einen und in den Spalten die Ausprägungen des anderen Merkmals angetragen werden. Hieraus resultiert dann eine zweidimensionale Häufigkeitstabelle. In den Zellen der Tabelle stehen dann die gemeinsamen absoluten bzw. relativen Häufigkeiten.

Beispiel:

Am Beispiel der beiden Merkmale 'Gästeart' (F3: Gart) und 'Informationsquelle' (F2: Info) lässt sich die grundlegende Struktur dieser zweidimensionalen Häufigkeitstabelle leicht verdeutlichen. Dabei wurden für das Merkmal 'Gästeart' sowohl die Ausprägungen Gruppe/Verein und Junior als auch Familie und 27P jeweils zu einer Gruppe zusammengefasst:

h_{ik}	Informationsquelle			
Gästeart	**DJH-Verzeichnis**	**Empfehlung**	**Internet**	**Summe**
Schulklasse	2	12	3	17
Gruppe/Verein/Junior	3	4	13	20
Familie/27P	4	3	6	13
Summe	9	19	22	50

7. Zweidimensionale Häufigkeitsverteilung

Aus der Kombination der jeweils drei Merkmalsausprägungen resultieren neun Felder der Matrix, die in der Mitte der Häufigkeitstabelle liegen. Danach verreisen beispielsweise 12 Personen im Klassenverband und greifen bei der Auswahl der Jugendherberge bevorzugt auf Empfehlungen zurück. 13 Personen sind dagegen als Gruppe, Verein oder Junior unterwegs und beziehen ihre Informationen bevorzugt aus dem Internet.

Neben diesen zentralen Feldern der Matrix werden an den Rändern der Tabelle die Häufigkeiten (daher Randhäufigkeiten) notiert, die sich aus der Addition der jeweiligen Spalten oder Zeilen ergeben. Von den insgesamt 50 befragten Personen reisen also 17 als Schulklasse, 20 in Gruppe, Verein oder Junior und 13 mit Familie bzw. als Senior. Auf der anderen Seite bevorzugen 9 Personen das DJH-Verzeichnis als Informationsquelle zu Jugendherbergen, 19 Personen direkte Empfehlungen und 22 Gäste das Internet.

Definition: Zweidimensionale Häufigkeitstabelle

Gegeben seien zwei gruppierte Merkmale X und Y mit den Merkmalsausprägungen x_i mit i=1, 2, …, m sowie den Merkmalsausprägungen y_k mit k=1, 2, …, n. Die absolute Häufigkeit $h(x_i; y_k)$ entspricht der Anzahl an Merkmalsträgern, die sowohl die Merkmalsausprägung x_i also auch die Merkmalsausprägung y_k aufweisen und wird kurz mit h_{ik} bezeichnet. In der zweidimensionalen Häufigkeitstabelle werden die Merkmalsausprägungen von X dann zeilenweise und die von Y spaltenweise angetragen.

Am rechten und am unteren Rand der Tabelle können die Verteilungen abgelesen werden, die sich ergeben, wenn jedes Merkmal für sich alleine betrachtet wird. Die Zeilensummen

$$h_{i.} = \sum_{k=1}^{n} h_{ik} \quad (i = 1, 2, ..., m)$$

entsprechen dann der Häufigkeitsverteilung des Merkmals X und die Spaltensummen

$$h_{.k} = \sum_{i=1}^{m} h_{ik} \quad (k = 1, 2, ..., n)$$

der Häufigkeitsverteilung von Y. Dabei ersetzt der Punkt den Summationsindex.

Sowohl die Summe der jeweiligen Randhäufigkeiten wie auch die Summe aller zweidimensionalen (bivariaten) Häufigkeiten ergibt dann die Anzahl der Merkmalsträger N:

$$\sum_{i=1}^{m}\sum_{k=1}^{n} h_{ik} = \sum_{i=1}^{m} h_{i.} = \sum_{k=1}^{n} h_{.k} = N$$

Die im vorangegangenen Beispiel dargestellte Häufigkeitsverteilung des Merkmals 'Gästeart' mit m=3 verschiedenen Ausprägungen und des Merkmals 'Informationsquelle' mit n=3 verschiedenen Ausprägungen bekäme dann auf der Grundlage der hier definierten Symbolik folgendes formales Aussehen:

7. Zweidimensionale Häufigkeitsverteilung

h_{ik}	Merkmal Y			
Merkmal X	$y_{k=1}$	$y_{k=2}$	$y_{k=3}$	**Zeilensumme**
$x_{i=1}$	h_{11}	h_{12}	h_{13}	$h_{1.}$
$x_{i=2}$	h_{21}	h_{22}	h_{23}	$h_{2.}$
$x_{i=3}$	h_{31}	h_{32}	h_{33}	$h_{3.}$
Spaltensumme	$h_{.1}$	$h_{.2}$	$h_{.3}$	N

Analog zur eindimensionalen Datenanalyse lassen sich die absoluten Häufigkeiten durch Division mit N in relative Häufigkeiten umrechnen. Daraus ergibt sich dann eine zweidimensionale Verteilung mit relativen Häufigkeiten:

f_{ik}	Merkmal Y			
Merkmal X	$y_{k=1}$	$y_{k=2}$	$y_{k=3}$	**Zeilensumme**
$x_{i=1}$	f_{11}	f_{12}	f_{13}	$f_{1.}$
$x_{i=2}$	f_{21}	f_{22}	f_{23}	$f_{2.}$
$x_{i=3}$	f_{31}	f_{32}	f_{33}	$f_{3.}$
Spaltensumme	$f_{.1}$	$f_{.2}$	$f_{.3}$	1

Beispiel:

Unsere Verteilung aus dem letzten Beispiel hätte dann folgendes Aussehen:

f_{ik}	Informationsquelle			
Gästeart	**DJH-Verzeichnis**	**Empfehlung**	**Internet**	**Summe**
Schulklasse	0,04	0,24	0,06	0,34
Gruppe/Verein/Junior	0,06	0,08	0,26	0,40
Familie/27P	0,08	0,06	0,12	0,26
Summe	0,18	0,38	0,44	1,00

Des Weiteren wurde bei den bisherigen Ausführungen lediglich auf gruppierte Merkmale abgestellt. Die Häufigkeitstabellen können jedoch problemlos an klassierte Merkmale angepasst werden, die konkrete Ausprägung eines gruppierten Merkmals wird dann lediglich durch die Klassenbezeichnung ersetzt. Damit können sogar gruppierte und klassierte Merkmale in einer Tabelle kombiniert werden.

Auch das Skalenniveau der Merkmale spielt - ganz im Gegensatz zu den nachfolgenden Abschnitten - keine Rolle, lediglich die Bezeichnung der Häufigkeitstabelle wird in Abhängigkeit vom Skalenniveau der Merkmale unterschieden. Liegt mindestens ein nominal skaliertes Merkmal vor, so wird von der zweidimensionalen Häufigkeitstabelle als Kontingenztabelle gesprochen, bei mindestens zwei ordinal skalierten Merkmalen von der Korrelationstabelle.

8. Zusammenhang bei nominalen Merkmalen

Bedingte relative Häufigkeiten und statistische Unabhängigkeit

Werden nur einzelne Zeilen der zweidimensionalen Tabelle betrachtet, so zerfällt das Merkmal X in die marginalen Bereiche (X_1), (X_2), ... (X_m).
Es lässt sich nun jede marginale Klasse wiederum als eine Grundgesamtheit auffassen, die durch das Merkmal Y in n Teilmengen zerlegt wird. Jedem Bereich X_i lässt sich somit wieder eine Verteilung zuordnen. Hieraus lassen sich dann die bedingten relativen Häufigkeiten errechnen.

Definition: bedingte relative Häufigkeiten

Gegeben seien zwei Merkmale X und Y. In Abhängigkeit von den Ausprägungen des Merkmals x_i mit i = 1, 2, …, m lässt sich die gesamte statistische Masse in m Teilmengen zerlegen und für jede dieser Teilmengen die Verteilung der y_k bestimmen. Die sich hieraus ergebenden relativen Häufigkeiten werden dann bedingte relative Häufigkeiten genannt:

$$f(y_k \mid x_i) = \frac{h_{ik}}{h_{i.}} \quad \begin{array}{l} \Leftarrow \text{absolute Häufigkeit von } x_i \text{ und } y_k \\ \Leftarrow \text{Randhäufigkeit von } x_i \end{array}$$

Eine analoge Betrachtung ist grundsätzlich auch umgekehrt für die Verteilung der x_i unter der Bedingung der y_k möglich. Allerdings wird gemeinhin ein Einfluss des Merkmals X auf das Merkmal Y unterstellt. Der Variablen Y wird daher im Normalfall das Merkmal zugeordnet, das vom Merkmal X abhängig ist; Y ist die abhängige und X die unabhängige Variable.

Beispiel:

Um die Berechnung bedingter relativer Häufigkeiten zu konkretisieren, greifen wir erneut den Zusammenhang der beiden Merkmale 'Gästeart' (F3: Gart) und 'Informationsquelle' (F2: Info) auf. Dieser wurde in einer zweidimensionalen Häufigkeitstabelle erfasst:

h_{ik}	Informationsquelle			
Gästeart	**DJH-Verzeichnis**	**Empfehlung**	**Internet**	**Summe**
Schulklasse	2	12	3	17
Gruppe/Verein/Junior	3	4	13	20
Familie/27P	4	3	6	13
Summe	9	19	22	50

Wenn nun eine Abhängigkeit der Informationsquelle von der Gästeart vorliegen soll, dann muss sich zeigen lassen, dass sich das Informationsverhalten der verschiedenen Gästearten grundsätzlich unterscheidet.
Hierzu wird die Gesamtheit aller Gäste in die drei Teilmengen Schulklasse, Gruppe/Verein/Junior und Familie/27P zerlegt und mit Hilfe der bedingten relativen Häufigkeiten die relative Verteilung innerhalb dieser drei Teilmengen betrachtet:

8. Zusammenhang bei nominalen Merkmalen

$f(y_k \mid x_i)$	Informationsquelle			
Gästeart	**DJH-Verzeichnis**	**Empfehlung**	**Internet**	**Summe**
Schulklasse	0,118	0,706	0,176	1,000
Gruppe/Verein/Junior	0,150	0,200	0,650	1,000
Familie/27P	0,308	0,231	0,462	1,000
Summe	0,180	0,380	0,440	1,000

70,6% der Gäste aus Schulklassen beziehen danach ihre Informationen zur Jugendherberge über Empfehlungen, die relative Häufigkeit für $y=y_2$ (Empfehlung) unter der Bedingung dass $x=x_1$ (Schulklasse) beträgt also:

$$f(y_2 \mid x_1) = \frac{h_{12}}{h_{1.}} = \frac{12}{17} = 0{,}706$$

Demgegenüber liegt der Anteil der Gäste aus Gruppe/Verein/Junior bzw. Familie/27P bei dieser Merkmalsausprägung mit

$$f(y_2 \mid x_2) = \frac{h_{22}}{h_{2.}} = \frac{4}{20} = 0{,}200$$

$$f(y_2 \mid x_3) = \frac{h_{32}}{h_{3.}} = \frac{3}{13} = 0{,}231$$

deutlich niedriger.

Je unterschiedlicher die Verteilung zwischen diesen Teilmengen der X-Variablen ausfällt, desto stärker ist die Abhängigkeit zwischen den beiden betrachteten Merkmalen. Umgekehrt gilt dann natürlich, dass zwischen zwei Merkmalen keine Abhängigkeit existiert, wenn die Verteilung bei allen Teilmengen der X-Variablen identisch ist. Oder anders formuliert: Alle bedingten relativen Häufigkeiten für eine konkrete Merkmalsausprägung y_k wären bei Unabhängigkeit der beiden Merkmale gleich. In diesem Fall würden die bedingten relativen Häufigkeit auch mit der relativen Randhäufigkeit von y_k übereinstimmen.

Definition: statistische Unabhängigkeit

Das Merkmal Y heißt statistisch unabhängig von Merkmal X, wenn alle bedingten Verteilungen von y_k gleich sind:

$$f(y_k \mid x_1) = f(y_k \mid x_2) = \ldots = f(y_k \mid x_m) = f_{.k} \quad \text{für } k = 1, 2, \ldots, n$$

Wegen $f(y_k \mid x_i) = h_{ik} / h_{i.} = f_{.k} = h_{.k} / N$ gilt dann bei Unabhängigkeit:

$$h_{ik} = \frac{h_{i.} \cdot h_{.k}}{N} \quad \text{bzw.} \quad f_{ik} = f_{i.} \cdot f_{.k}$$

8. Zusammenhang bei nominalen Merkmalen

Sind die beiden Merkmale X und Y voneinander statistisch unabhängig, dann ist die zweidimensionale Häufigkeitsverteilung also durch das Produkt aus den beiden Randverteilungen von X und Y eindeutig bestimmt.

Beispiel:

Bei statistischer Unabhängigkeit und den gegebenen Randhäufigkeiten hätte die zweidimensionale Häufigkeitstabelle der Merkmale 'Gästeart' (F3: Gart) und 'Informationsquelle' (F2: Info) folgendes Aussehen:

h_{ik}	Informationsquelle			
Gästeart	**DJH-Verzeichnis**	**Empfehlung**	**Internet**	**Summe**
Schulklasse	3,060	6,460	7,480	17,0
Gruppe/Verein/Junior	3,600	7,600	8,800	20,0
Familie/27P	2,340	4,940	5,720	13,0
Summe	9,0	19,0	22,0	50,0

Von den 17 Gästen aus Schulklassen würden folglich bei statistischer Unabhängigkeit der Merkmale

$$h_{11} = \frac{h_{1.} \cdot h_{.1}}{N} = \frac{17 \cdot 9}{50} = 3{,}060 \text{ Personen}$$

das DJH-Verzeichnis als zentrale Informationsquelle nutzen, was einem Anteil (= bedingte relative Häufigkeit) von

$$f(y_1 \mid x_1) = \frac{h_{11}}{h_{1.}} = \frac{3{,}060}{17{,}0} = 0{,}180$$

entspricht.
Auch der Anteil der Gruppen/Vereine/Junioren und der Anteil der Familien/Senioren, die das DJH-Verzeichnis als zentrale Informationsquelle nutzen,

$$f(y_1 \mid x_2) = \frac{h_{21}}{h_{2.}} = \frac{3{,}600}{20{,}0} = 0{,}180$$

$$f(y_1 \mid x_3) = \frac{h_{31}}{h_{1.}} = \frac{2{,}340}{13{,}0} = 0{,}180$$

liegt dann bei exakt 18%. Den gleichen Wert nimmt dann logischerweise auch die relative Randhäufigkeit $y_{.1}$ an. Natürlich muss dieser Zusammenhang auch für die beiden anderen Merkmalsausprägungen von y (Empfehlung und Internet) gelten:

$f(y_k \mid x_i)$	Informationsquelle			
Gästeart	**DJH-Verzeichnis**	**Empfehlung**	**Internet**	**Summe**
Schulklasse	0,180	0,380	0,440	1,000
Gruppe/Verein/Junior	0,180	0,380	0,440	1,000
Familie/27P	0,180	0,380	0,440	1,000
Summe	0,180	0,380	0,440	1,000

8. Zusammenhang bei nominalen Merkmalen

Maß von Cramér

Sind zwei Merkmale statistisch unabhängig, dann entsprechen ihre zweidimensionalen Häufigkeiten dem Produkt der Randhäufigkeiten dividiert durch N. Je stärker ihre tatsächlichen gemeinsamen Häufigkeiten von diesem Wert abweichen, desto stärker ist folglich ihre Abhängigkeit. Genau auf diesen Überlegungen basiert das Maß von Cramér.

Definition: Maß von Cramér C

Gegeben seien zwei mindestens nominal skalierte Merkmale X und Y. Das **Maß von Cramér C** misst die Stärke der statistischen Abhängigkeit zwischen diesen Merkmalen:

$$C = \sqrt{\frac{1}{N} \cdot \frac{\chi^2}{\min((m-1);(n-1))}}$$

mit

$$\chi^2 = \sum_{i=1}^{m}\sum_{k=1}^{n} \frac{\left(h_{ik} - \frac{h_{i.} \cdot h_{.k}}{N}\right)^2}{\frac{h_{i.} \cdot h_{.k}}{N}} = N \cdot \left[\left(\sum_{i=1}^{m}\sum_{k=1}^{n} \frac{h_{ik}^2}{h_{i.} \cdot h_{.k}}\right) - 1\right]$$

Die Maßzahl C bewegt sich im Wertebereich von Null bis Eins. Bei vollständiger statistischer Abhängigkeit nimmt das Maß von Cramér den Wert von Eins an, bei vollständiger statistischer Unabhängigkeit den Wert von Null.

Aufgabe:

Für die Leitung der Jugendherberge ist es sicherlich von Bedeutung, inwieweit sie sich auf Stammgäste stützen kann und ob diese bei den verschiedenen Gästearten unterschiedlich stark vertreten sind. Im weitesten Sinne ist hier eine Person als Stammgast einzuschätzen, wenn sie nach dem aktuellen Aufenthalt beabsichtigt, die Jugendherberge in naher Zukunft wieder zu besuchen (vgl. Frage 7: Awie). Eine weitere Differenzierung der Stammgäste unter Berücksichtigung von Frage 1 (früherer Aufenthalt) soll zumindest an dieser Stelle nicht weiter verfolgt werden. Die Abhängigkeit der Stammgäste von der Gästeart lässt sich damit über eine zweidimensionale Häufigkeitsverteilung der Merkmale 'Gästeart' (Frage 3: Gart) und 'Erneuter Besuch' (Frage 7: Awie) beschreiben:

h_{ik}	Stammgast		
Gästeart	**Nein**	**Ja**	**Summe**
Schulklasse	12	5	17
Gruppe/Verein/Junior	10	10	20
Familie/27P	4	9	13
Summe	26	24	50

Berechnen Sie hieraus ein geeignetes Zusammenhangsmaß für die Abhängigkeit der Stammgäste von der Gästeart und interpretieren Sie das Ergebnis.

8. Zusammenhang bei nominalen Merkmalen

Exkurs: Test auf statistische Unabhängigkeit (Unabhängigkeitstest)

Für die Abhängigkeit der Stammgäste von der Gästeart aus der vorangegangenen Aufgabe errechnet sich ein Maß von Cramér in Höhe von 0,307. Demnach existiert eine statistische Abhängigkeit, auch wenn sie eher schwach ausgeprägt ist.

Diese Aussage ist allerdings mit großer Vorsicht zu 'genießen'. Erinnern wir uns an die zugrunde liegende Datenbasis. Das Ergebnis wurde aus der Analyse einer statistischen Masse von 50 Gästen gewonnen. Hätten wir exakt dieses Ergebnis aus der Analyse der Grundgesamtheit, also der Gesamtzahl aller Gäste der Jugendherberge innerhalb eines Jahres, gewonnen, dann ließe sich das Ergebnis der Cramérschen Maßzahl genau so interpretieren. Unser Ergebnis kann jedoch zufallsbedingten Verzerrungen unterliegen, da aus der Gesamtzahl von mehreren Tausend Gästen eine mehr oder weniger willkürliche Teilmenge (= Stichprobe) von nur 50 Personen untersucht wurde.

Um genau diesen Zufallseinfluss zu quantifizieren, benötigen wir das Instrumentarium der induktiven Statistik. Selbst wenn in der Grundgesamtheit absolut keine statistische Abhängigkeit zwischen zwei Merkmalen besteht, kann die zufällige Auswahl von Einheiten aus dieser Grundgesamtheit und - wie im vorliegenden Beispiel - ein geringer Umfang der Stichprobe zu einem Ergebnis führen, das eine statistische Abhängigkeit signalisiert.
Es ist daher weiter zu untersuchen, wie wahrscheinlich ein solches Stichprobenergebnis ist, wenn von der Annahme ausgegangen wird, dass in der Grundgesamtheit keine statistische Abhängigkeit vorliegt. Tendiert diese Wahrscheinlichkeit gegen Null, so liegt umgekehrt die Vermutung nahe, dass die getroffene Annahme falsch ist. Die Annahme einer statistischen Unabhängigkeit der beiden Merkmale wäre folglich zu verwerfen, das Stichprobenergebnis wäre dann fundiert. Die Statistik spricht in diesem Fall von einem signifikanten Stichprobenergebnis, das eine Verallgemeinerung zulässt, sich also von der Stichprobe auf die unbekannte Grundgesamtheit übertragen lässt.

Auf diesen allgemeinen Überlegungen basiert nun der statistische Unabhängigkeitstest. Danach wird eine Grundgesamtheit unterstellt, in der die beiden Merkmale X und Y statistisch unabhängig sind. Hieraus resultiert die zu überprüfende Nullhypothese:

H_0: Die Merkmale X und Y sind statistisch unabhängig

(H_1: Die Merkmale X und Y sind voneinander abhängig)

Zur Überprüfung der Nullhypothese wird aus dieser Grundgesamtheit eine Stichprobe im Umfang von N statistischen Einheiten gezogen und die Merkmalsausprägungen von X und Y in der bekannten zweidimensionalen Häufigkeitstabelle erfasst. Sind die beiden Merkmale in der Grundgesamtheit nun tatsächlich statistisch unabhängig, so wird dies näherungsweise auch für die Stichprobe gelten. Die erfassten (empirischen) Häufigkeiten der zweidimensionalen Häufigkeitstabelle weichen dann nur minimal von denen bei statistischer Unabhängigkeit ab. Der Chi-Quadrat-Wert, der genau diese Differenzen erfasst, eignet sich folglich ideal zur Überprüfung der Nullhypothese:

$$T = \chi^2 = \sum_{i=1}^{m}\sum_{k=1}^{n} \frac{\left(h_{ik} - \frac{h_{i.} \cdot h_{.k}}{N}\right)^2}{\frac{h_{i.} \cdot h_{.k}}{N}} = N \cdot \left[\left(\sum_{i=1}^{m}\sum_{k=1}^{n} \frac{h_{ik}^2}{h_{i.} \cdot h_{.k}}\right) - 1\right]$$

8. Zusammenhang bei nominalen Merkmalen

Würden nun aus dieser Grundgesamtheit mit voneinander unabhängigen Merkmalen X und Y eine Vielzahl solcher Stichproben im Umfang N gezogen, dann würden die relativen Häufigkeiten der T - Werte aus diesen Stichproben dem Verlauf einer Chi-Quadrat - Verteilung entsprechen. Oder anders (statistischer) formuliert: Gilt die oben formulierte Nullhypothese, so ist die Prüf- oder Testgröße T Chi-Quadrat - verteilt mit (m-1)·(n-1) - Freiheitsgraden df. Die Variable m entspricht dabei der Anzahl der Merkmalsausprägungen von X und die Variable n der Anzahl der Merkmalsausprägungen von Y.

$$T \sim X^2_{df} \quad \text{mit: } df = (m-1) \cdot (n-1)$$

Bei Kenntnis der konkreten Verteilung der Testgröße T lässt sich eine Entscheidungsregel formulieren, die dann zu einer Ablehnung oder Nichtablehnung der Nullhypothese führt. Hohe Werte der Testgröße sind bei Gültigkeit der Nullhypothese eher unwahrscheinlich. Übersteigt also der Wert der Testgröße T aus der Stichprobe einen festgelegten (kritischen) Wert und liegt damit im sog. kritischen Bereich (KB), dann ist die Nullhypothese abzulehnen.

$$KB_\alpha = [X^2_{df;\,1-\alpha} \,;\, +\infty)$$

Der kritische Wert $X^2_{df,\,1-\alpha}$ entspricht dabei dem (1-α) - Quantil der Chi-Quadrat - Verteilung mit df = (m-1) (n-1) - Freiheitsgraden (vgl. Tabelle 2). Hierbei wird der kritische Wert so festgesetzt, dass gemäß unterstellter Chi-Quadrat - Verteilung 90%, 95% oder gar 99% aller Chi-Quadrat - Werte darunter liegen. Dies impliziert natürlich, dass selbst bei Gültigkeit der Nullhypothese mit einer Wahrscheinlichkeit von 10%, 5% oder nur 1% noch Werte der Testgröße T in einer Stichprobe auftreten können, die über dem kritischen Wert liegen. In diesem Fall würde der Anwender des Testverfahrens eine Fehlentscheidung treffen. Er würde dann die Nullhypothese ablehnen, obwohl sie richtig ist. Diese Wahrscheinlichkeit wird daher als Irrtumswahrscheinlichkeit α (bzw. Signifikanzniveau) bezeichnet:

α = Irrtumswahrscheinlichkeit bzw. Signifikanzniveau = 0,05 (evtl. auch 0,10 oder 0,01)

Je geringer das Signifikanzniveau, desto geringer ist natürlich auch die Wahrscheinlichkeit, genau diesen Fehler zu begehen, also die Nullhypothese abzulehnen, obwohl sie richtig ist. In Anlehnung an die Variablenbezeichnung wird hier vom α-Fehler (oder Fehler 1. Art) gesprochen. Soll diese Fehlerquelle möglichst minimiert werden (ein vollständiges Ausschalten gibt es nicht!!!), muss das Signifikanzniveau mit 0,01 oder sogar 0,001 sehr klein gewählt werden. Allerdings geht damit zwangsläufig eine zweite Fehlerquelle, der Fehler 2. Art bzw. β-Fehler, einher, nämlich die Nullhypothese nicht abzulehnen, obwohl sie eigentlich falsch ist. Sozusagen als Kompromiss, um beide Fehlerquellen nicht allzu groß werden zu lassen, hat sich ein Signifikanzniveau α in Höhe von 5% etabliert. Wir werden uns daher im folgenden an diesem Niveau orientieren. Der kritische Wert zur Überprüfung der Nullhypothese entspricht dann dem 0,95-Quantil der Chi-Quadrat - Verteilung bei vorgegebenen Freiheitsgraden. Tabelle 2 zeigt auf der nachfolgenden Seite ausgewählte Quantile der Chi-Quadrat - Verteilung bei Freiheitsgraden zwischen 1 und 35; die Spalte mit dem 0,95-Quantil, die ja mit einem Signifikanzniveau von 0,05 korrespondiert, ist dabei entsprechend hervorgehoben.

Darüber hinaus ist auch zu beachten, dass die Qualität eines Testverfahrens entscheidend vom Umfang der Stichprobe abhängt. Ein hinreichend großer Stichprobenumfang ist gegeben, wenn die bei Unabhängigkeit realisierten absoluten Häufigkeiten in den Zellen einen Wert von 5 nicht unterschreiten:

$$\frac{h_{i\cdot} \cdot h_{\cdot k}}{N} \geq 5$$

8. Zusammenhang bei nominalen Merkmalen

Tabelle 2

Quantile (1-α) der Chi-Quadrat - Verteilung

df	0,10	0,30	0,50	0,70	0,90	0,95	0,99
1	0,0158	0,1485	0,4549	1,0742	2,7055	3,8415	6,6349
2	0,2107	0,7133	1,3863	2,4079	4,6052	5,9915	9,2103
3	0,5844	1,4237	2,3660	3,6649	6,2514	7,8147	11,3449
4	1,0636	2,1947	3,3567	4,8784	7,7794	9,4877	13,2767
5	1,6103	2,9999	4,3515	6,0644	9,2364	11,0705	15,0863
6	2,2041	3,8276	5,3481	7,2311	10,6446	12,5916	16,8119
7	2,8331	4,6713	6,3458	8,3834	12,0170	14,0671	18,4753
8	3,4895	5,5274	7,3441	9,5245	13,3616	15,5073	20,0902
9	4,1682	6,3933	8,3428	10,6564	14,6837	16,9190	21,6660
10	4,8652	7,2672	9,3418	11,7807	15,9872	18,3070	23,2093
11	5,5778	8,1479	10,3410	12,8987	17,2750	19,6751	24,7250
12	6,3038	9,0343	11,3403	14,0111	18,5493	21,0261	26,2170
13	7,0415	9,9257	12,3398	15,1187	19,8119	22,3620	27,6882
14	7,7895	10,8215	13,3393	16,2221	21,0641	23,6848	29,1412
15	8,5468	11,7212	14,3389	17,3217	22,3071	24,9958	30,5779
16	9,3122	12,6243	15,3385	18,4179	23,5418	26,2962	31,9999
17	10,0852	13,5307	16,3382	19,5110	24,7690	27,5871	33,4087
18	10,8649	14,4399	17,3379	20,6014	25,9894	28,8693	34,8053
19	11,6509	15,3517	18,3377	21,6891	27,2036	30,1435	36,1909
20	12,4426	16,2659	19,3374	22,7745	28,4120	31,4104	37,5662
21	13,2396	17,1823	20,3372	23,8578	29,6151	32,6706	38,9322
22	14,0415	18,1007	21,3370	24,9390	30,8133	33,9244	40,2894
23	14,8480	19,0211	22,3369	26,0184	32,0069	35,1725	41,6384
24	15,6587	19,9432	23,3367	27,0960	33,1962	36,4150	42,9798
25	16,4734	20,8670	24,3366	28,1719	34,3816	37,6525	44,3141
26	17,2919	21,7924	25,3365	29,2463	35,5632	38,8851	45,6417
27	18,1139	22,7192	26,3363	30,3193	36,7412	40,1133	46,9629
28	18,9392	23,6475	27,3362	31,3909	37,9159	41,3371	48,2782
29	19,7677	24,5770	28,3361	32,4612	39,0875	42,5570	49,5879
30	20,5992	25,5078	29,3360	33,5302	40,2560	43,7730	50,8922
31	21,4336	26,4397	30,3359	34,5981	41,4217	44,9853	52,1914
32	22,2706	27,3728	31,3359	35,6649	42,5847	46,1943	53,4858
33	23,1102	28,3069	32,3358	36,7307	43,7452	47,3999	54,7755
34	23,9523	29,2421	33,3357	37,7954	44,9032	48,6024	56,0609
35	24,7967	30,1782	34,3356	38,8591	46,0588	49,8018	57,3421

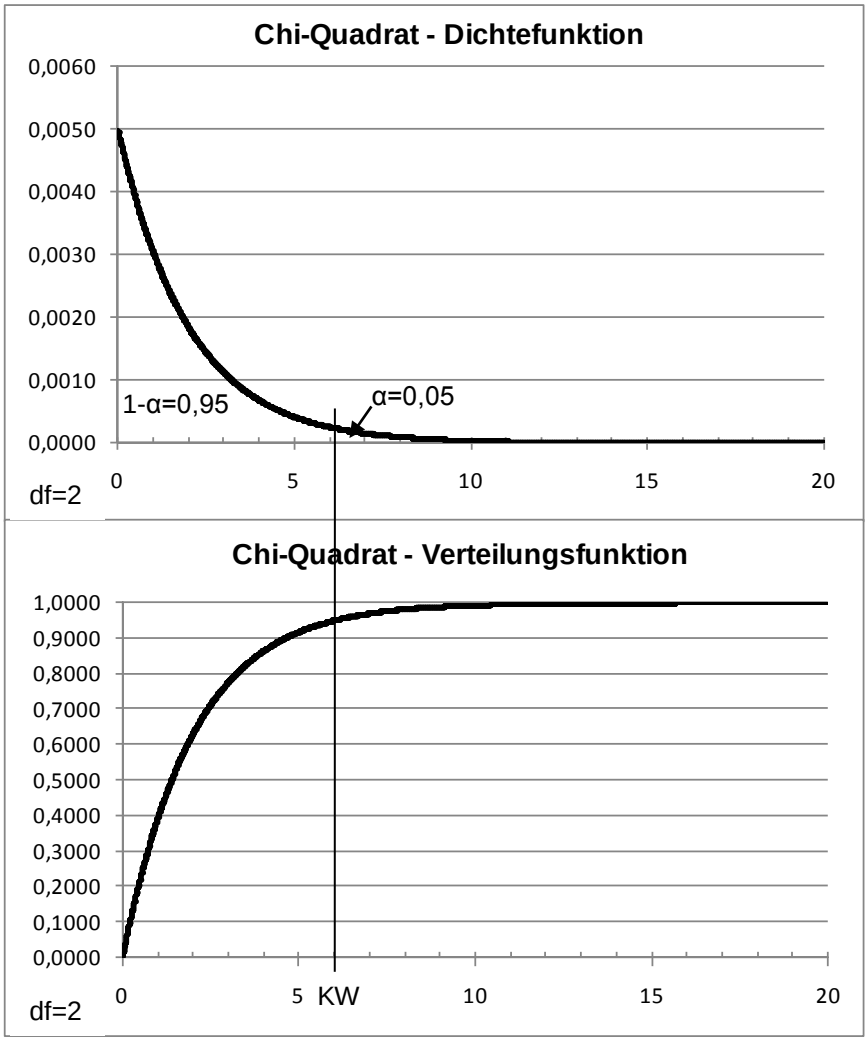
Lesebeispiel: $1-\alpha = 0,95$ und $df = 2 \rightarrow \chi^2_{0,95;2} = 5,9915$

$\rightarrow f(\chi^2 \leq 5,9915) = F(5,9915) = 0,95$

8. Zusammenhang bei nominalen Merkmalen

Beispiel:

Bevor also das Ergebnis von Cramér als Maßzahl für die Stärke der statistischen Abhängigkeit zweier Merkmale im Detail interpretiert werden kann, ist das Ergebnis auf Signifikanz zu überprüfen.
Es stellt sich folglich die Frage, ob der errechnete Chi-Quadrat - Wert in Höhe von 4,70 ausreicht, die Nullhypothese abzulehnen. Die Nullhypothese geht in diesem Fall davon aus, dass die beiden Merkmale Gästeart und Stammgast statistisch unabhängig voneinander sind. Bei m=3 Ausprägungen des Merkmals 'Gästeart' und n=2 Ausprägungen des Merkmals 'Stammgast' ist die Testgröße T Chi-Quadrat - verteilt mit df=(3-1)·(2-1)=2 Freiheitsgraden. Dieser Verlauf lässt sich wie folgt skizzieren:



Die Dichtefunktion beschreibt die relativen Häufigkeiten (induktiv: Wahrscheinlichkeiten) einer Chi-Quadrat - verteilten Variablen, die Verteilungsfunktion in bekannter Weise die kumulierten relativen Häufigkeiten (Wahrscheinlichkeiten). Diese entsprechen der jeweiligen Fläche unter der Dichtefunktion.

Somit können - zumindest bei genauer Betrachtung - sämtliche Informationen zur Durchführung des Signifikanztestes direkt aus den beiden Grafiken abgelesen werden. Der kritische Wert, der die untere Grenze des kritischen Bereichs markiert, beginnt bei einem unterstellten Signifikanzniveau von 0,05 dort, wo die Verteilungsfunktion F exakt den Wert von 0,95 (=1-0,05) erreicht.

Danach liegt der kritische Wert bei etwa 6,0. Der Blick in Tabelle 2 liefert für das 0,95-Quantil bei 2 Freiheitsgraden ein exaktes Ergebnis von 5,9915 (vgl. auch Wert der Excel-Funktion 'CHIINV(0,05; 2)' mit 0,05=α und 2=df).

Auf der Grundlage der so gewonnenen Informationen lässt sich nun der Chi-Quadrat - Unabhängigkeitstest durchführen:

H_0: Die Merkmale Gästeart und Stammgast sind statistisch unabhängig

$T = 4{,}70$

$\alpha = 0{,}05$

$df = (3\text{-}1) \cdot (2\text{-}1) = 2$

$T \sim X^2_2$

$KB_{0,05} = [X^2_{2;\,0,95} ; +\infty) = [5{,}9915; +\infty)$

→ Da die Testgröße T nicht im kritischen Bereich liegt, kann die Nullhypothese nicht abgelehnt werden. Weitergehende Aussagen zur Maßzahl von Cramér sind demnach statistisch nicht gesichert (nicht signifikant).

An diesem Beispiel lässt sich zugleich der Einfluss des Stichprobenumfangs auf das Testergebnis zeigen. Würde sich bspw. der Stichprobenumfang auf N=100 verdoppeln und in gleicher Weise auch alle bisherigen absoluten Häufigkeiten der zweidimensionalen Häufigkeitstabelle, dann würde sich auch der Wert der Prüf- und Testgröße T verdoppeln:

$$T = \chi^2 = \sum_{i=1}^{m}\sum_{k=1}^{n} \frac{\left(h_{ik} - \frac{h_{i.} \cdot h_{.k}}{N}\right)^2}{\frac{h_{i.} \cdot h_{.k}}{N}} = N \cdot \left[\left(\sum_{i=1}^{m}\sum_{k=1}^{n} \frac{h_{ik}^2}{h_{i.} \cdot h_{.k}}\right) - 1\right] = 100 \cdot (1{,}094 - 1) = 9{,}4$$

Bei ansonsten konstanten Testbedingungen würde T jetzt über dem kritischen Wert von 5,9915 und damit im kritischen Bereich liegen. Die Nullhypothese würde nun abgelehnt werden, was bedeutet, dass von einer signifikanten Abhängigkeit zwischen Gästeart und Stammgast ausgegangen werden kann.

Da die relativen Häufigkeiten konstant geblieben wären, hätte sich das Maß von Cramér

$$C = \sqrt{\frac{1}{100} \cdot \frac{9{,}4}{\min((3-1);(2-1))}} = 0{,}307$$

ebenfalls nicht geändert. Bei abgelehnter Nullhypothese könnte folglich zumindest eine schwache statistische Abhängigkeit zwischen Gästeart und Stammgast festgestellt werden.

9. Zusammenhang bei metrischen Merkmalen

Je höher das Skalenniveau der Merkmale X und Y, desto größer sind die statistischen Auswertungsmöglichkeiten. Auf die Analyseverfahren metrischer Merkmale soll daher im Folgenden detailliert eingegangen werden.

Im Rahmen der Gästebefragung (vgl. Abbildung 1) sind drei Merkmale erfasst worden, die metrisch skaliert sind: die Anzahl an Übernachtungen (F4: Nach), die Gesamtbewertung der Jugendherberge in Punkten (F6: Gbew) und das Alter (F9: Alte). Der paarweise Zusammenhang zwischen diesen drei Merkmalen soll daher in diesem Kapitel näher betrachtet werden. Es geht also konkret um die Frage, ob das Alter der Gäste in irgendeiner Form einen Einfluss auf deren Übernachtungsdauer bzw. deren Gesamtbewertung hat oder ob die Gesamtbewertung vielleicht von der vorausgegangenen Übernachtungsdauer abhängt.

Um dabei die Analyseverfahren metrischer Merkmale mit ihrer Vielzahl verschiedener Ausprägungen möglichst überschaubar zu halten, soll die statistische Masse auf 20 ausgewählte Gäste reduziert werden. In der nachfolgenden Tabelle 3 sind die 20 Datensätze zu den drei metrischen Merkmalen erfasst.

Tabelle 3

Ausgewählte Datensätze (20)

Nach	**Gbew**	**Alte**
F4	F6	F9
4	47	58
2	43	38
1	41	32
4	46	47
5	47	54
3	45	43
2	41	33
3	46	52
2	44	44
3	45	44
2	41	31
1	42	26
1	41	20
3	44	14
5	46	13
5	46	11
3	49	9
5	50	10
3	42	15
2	43	34

Alternativ könnten die Informationen zu den Merkmalen zumindest paarweise auch in einer zweidimensionalen Häufigkeitstabelle (bei metrischen Merkmalen: Korrelationstabelle) dargestellt werden.

9. Zusammenhang bei metrischen Merkmalen

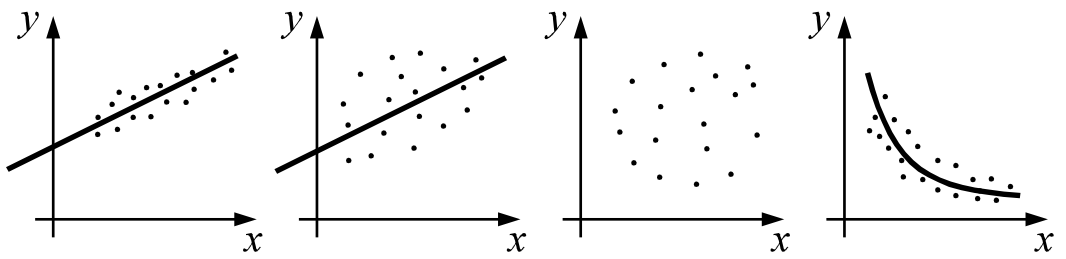
Allerdings macht dies bei metrischen Merkmalen mit einer Vielzahl unterschiedlicher Ausprägungen meist keinen Sinn. Zum einen resultieren hieraus eine Vielzahl von Zeilen und Spalten, was die Lesbarkeit dieser Tabellen erheblich einschränkt. Zum anderen stimmt kaum ein Wertepaar in beiden Ausprägungen mit anderen Wertepaaren überein, so dass in den Zellen solcher Tabellen nur absolute Häufigkeiten von Null oder Eins ausgewiesen werden. Zur Analyse metrischer Merkmale wird daher meist auf die Liste mit den originären Wertepaaren zurückgegriffen.

Streudiagramm

Erste wichtige Anhaltspunkte über die Abhängigkeit von Merkmalen lassen sich relativ einfach aus dem so genannten Streudiagramm gewinnen. Da beide Merkmale über eine natürliche Rangordnung verfügen, macht es Sinn, die beobachteten Wertepaare in Form von Punkten in ein Koordinatensystem einzutragen. Die Merkmalsausprägungen von X finden sich dabei auf der Abszisse, die von Y auf der Ordinate. Die hieraus hervorgehende Punktewolke gibt dann wichtige Hinweise nicht nur auf die Stärke, sondern auch auf die Art des Zusammenhangs.

Abbildung 2

Beispielhafte Darstellung verschiedener Streudiagramme

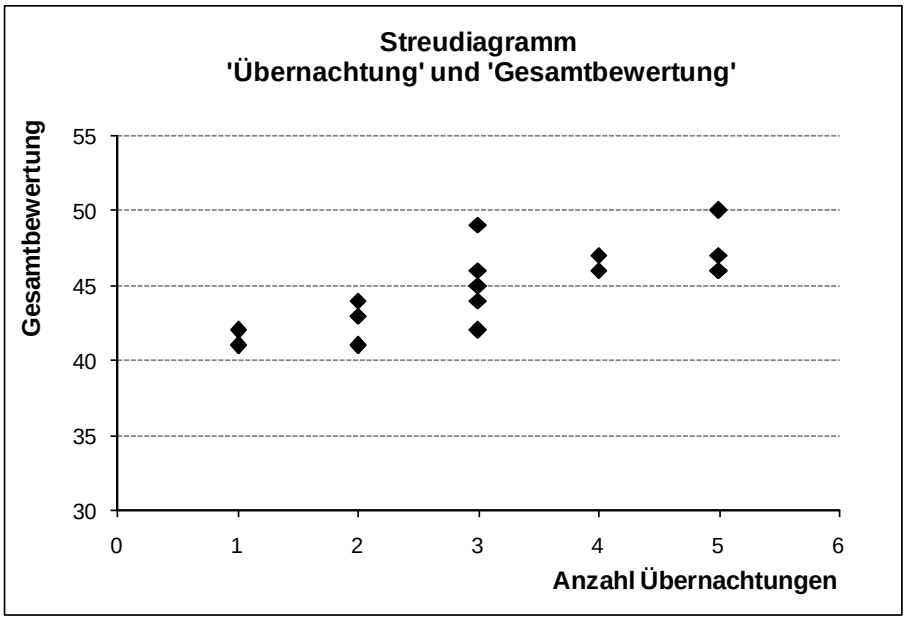


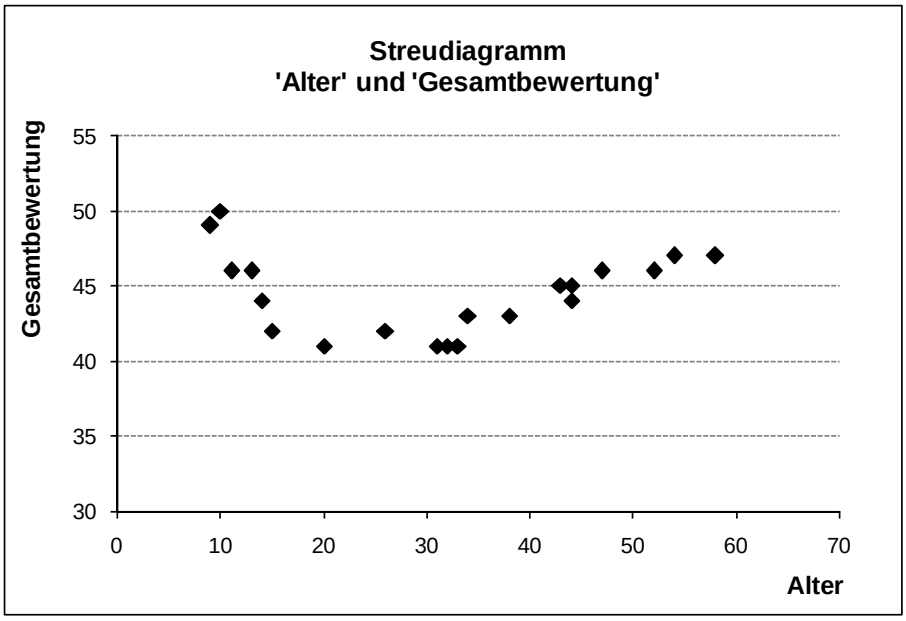
So zeigen die beiden ersten Beispiele in Abbildung 2 einen positiven linearen Zusammenhang, der im ersten Beispiel stärker ausgeprägt ist, weil die Wertepaare wesentlich enger um eine lineare Funktion (Gerade) angeordnet sind. Während im dritten Beispiel kein Zusammenhang erkennbar ist, deutet das vierte dagegen wieder auf einen starken, allerdings nichtlinearen Zusammenhang hin.

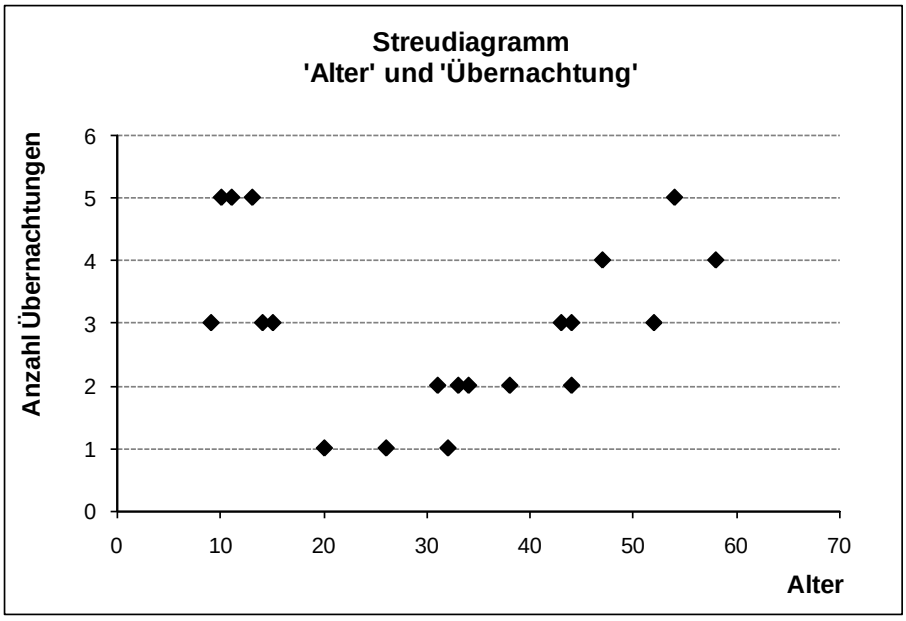
Beispiel:

In den Streudiagrammen zu den drei ausgewählten Merkmalen auf der nächsten Seite sind vergleichbare Verlaufsmuster zu beobachten. Jedoch nur eine Abbildung, nämlich die zwischen den Merkmalen 'Übernachtung' und 'Gesamtbewertung', weist auf einen linearen Zusammenhang hin und ähnelt im weitesten Sinne der ersten Modellvariante aus Abbildung 2.
Wie wichtig eine einleitende Analyse mittels Streudiagramm sein kann, demonstrieren die beiden anderen Beispiele zum Zusammenhang von Alter und Gesamtbewertung bzw. Alter und Übernachtung. In beiden Fällen ist offensichtlich, dass kein durchgängig linearer Zusammenhang vorliegt.

9. Zusammenhang bei metrischen Merkmalen







9. Zusammenhang bei metrischen Merkmalen

In beiden Beispielen wäre daher ein differenzierteres Vorgehen notwendig. Zwei Varianten bieten sich hierbei an. Zum einen ließe sich die Beschreibung des Gesamtzusammenhangs in zwei (gegenläufige) Komponenten zerlegen. Sichtbar ist in beiden Streudiagrammen ein ausgeprägter negativer linearer Zusammenhang für Gäste unter 20 Jahren und ein ausgeprägter positiver linearer Zusammenhang für Gäste ab 20 Jahren. Zum anderen könnten beide Komponenten über einen nichtlinearen Ansatz als Polynom 2. Grades (quadratische Funktion) erfasst werden. Nichtlineare Funktionen werden im Rahmen der Regressionsanalyse in Teil IV ausführlicher diskutiert.

Da nur der Zusammenhang zwischen den Merkmalen 'Übernachtung' und 'Gesamtbewertung' ein klassisch lineares Verlaufsmuster zeigt, soll genau dieses Beispiel nachfolgend genauer betrachtet werden.

Kovarianz

Betrachtet werden also - wie in den bisherigen Beispielen - zwei metrische (quantitative) Merkmale X und Y. Zu diesen Merkmalen liegen N Beobachtungspaare von Merkmalsausprägungen (x_1,y_1), (x_2,y_2), ..., (x_N,y_N) vor. Ein linearer Zusammenhang liegt dann vor, wenn die Erhöhung von X um genau eine Einheit zu einer Veränderung bei Y um einen konstanten Wert führt. Die Schwankung von Y wird also durch die Schwankung von X ausgelöst. In Anlehnung an die Varianz, die die Schwankungsbreite eines metrischen Merkmals quantifiziert, misst die Kovarianz die gemeinsame Variabilität der beiden Merkmale X und Y.

Definition: Kovarianz Cov

Gegeben seien die beiden gemeinsam auftretenden metrisch skalierten Merkmale X und Y mit ihren arithmetischen Mittelwerten. Bei gegebener zweidimensionaler Häufigkeitsverteilung mit den absoluten Häufigkeiten $h(x_i; y_k)$ bzw. den relativen Häufigkeiten $f(x_i; y_k)$ mit i =1,..., m und k =1,..., n errechnet sich die Kovarianz Cov wie folgt:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{m}\sum_{k=1}^{n}(x_i - \bar{x})(y_k - \bar{y})\cdot h_{ik} = \frac{1}{N}\sum_{i=1}^{m}\sum_{k=1}^{n} x_i \cdot y_k \cdot h_{ik} - \bar{x}\cdot\bar{y}$$

bzw.

$$Cov(X,Y) = \sum_{i=1}^{m}\sum_{k=1}^{n}(x_i - \bar{x})(y_k - \bar{y})\cdot f_{ik} = \sum_{i=1}^{m}\sum_{k=1}^{n} x_i \cdot y_k \cdot f_{ik} - \bar{x}\cdot\bar{y}$$

Bei klassierten Werten werden in bekannter Weise die konkreten Merkmalsausprägungen x_i und y_k durch die jeweiligen Klassenmitten x_i' und y_k' ersetzt.

Wird die Kovarianz aus den Paaren von Beobachtungswerten $(x_i; y_i)$ mit i =1,..., N berechnet, vereinfacht sich die Formel zu:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i \cdot y_i - \bar{x}\cdot\bar{y}$$

9. Zusammenhang bei metrischen Merkmalen

Allerdings sind den Erkenntnissen, die aus der Berechnung der Kovarianz gezogen werden können, enge Grenzen gesetzt. Daher sind folgende Aussagen zu beachten:

- Ist die Kovarianz gleich Null, folgt nicht zwingend, dass kein Zusammenhang zwischen den Merkmalen X und Y besteht. Es ist lediglich ein linearer Zusammenhang auszuschließen.
- Die Kovarianz wird umso größer, je häufiger zu einer Beobachtung $(x_i;y_i)$ die Differenzen der beiden Merkmalsausprägungen x_i und y_i von ihren jeweiligen Mittelwerten das gleiche Vorzeichen haben.
- Die Kovarianz wird umso kleiner (negativ), je häufiger zu einer Beobachtung $(x_i;y_i)$ die Differenzen der beiden Merkmalsausprägungen x_i und y_i von ihren jeweiligen Mittelwerten das entgegen gesetzte Vorzeichen haben.

Beispiel:

Mit Blick auf Tabelle 3 fällt auf, dass sich allein 10 der 20 Wertepaare für die Merkmale 'Übernachtung' und 'Gesamtbewertung' unterscheiden. Lediglich 5 Wertepaare - nämlich (1; 41), (2, 41), (2; 43), (3; 45) und (5; 46) - treten jeweils zweimal auf.
Eine Datenaufbereitung im Rahmen einer zweidimensionalen Häufigkeitstabelle macht folglich nur wenig Sinn. Bei fünf verschiedenen Merkmalsausprägungen von X und sogar neun verschiedenen Ausprägungen von Y ergäbe sich eine 5 x 9 - Matrix, deren Zellen in fünf Fällen mit einer 2, in 10 Fällen mit einer 1 und in den restlichen 30 Fällen mit einer 0 belegt werden. Das Ausmaß der Datenkomprimierung hält sich damit in Grenzen.

Es macht daher - wie meist bei metrischen Merkmalen - Sinn, die Berechnung der Kovarianz von Übernachtung und Gesamtbewertung nach der vereinfachten Formel direkt aus den einzelnen Wertepaaren vorzunehmen:

$$\mathrm{Cov}(X,Y) = \frac{1}{20} \cdot (4 \cdot 47 + 2 \cdot 43 + \ldots + 2 \cdot 43) - 2{,}95 \cdot 44{,}45 = 2{,}772 \ \text{[Nächte} \cdot \text{Punkte]}$$

Die Kovarianz weist demnach auf einen gleichgerichteten (positiven) linearen Zusammenhang hin.

Weitergehende Aussagen sind allerdings nicht möglich. Die Kovarianz ist keine standardisierte Größe, ihr Wertebereich also weder nach oben noch nach unten beschränkt. Dadurch wird eine Einschätzung der Stärke des linearen Zusammenhangs zwischen zwei Merkmalen erschwert oder gar unmöglich gemacht.

Korrelationskoeffizient von Pearson/Bravais

Die Kovarianz fällt tendenziell umso größer aus, je größer auch die Varianzen der beiden Merkmale X und Y sind. Die Standardabweichungen von X und Y müssen daher bei einem standardisierten Abhängigkeitsmaß berücksichtigt werden. Genau diese Überlegungen greift der Korrelationskoeffizient r von Pearson/Bravais auf und setzt die Kovarianz ins Verhältnis zu den beiden Standardabweichungen von X und Y.

9. Zusammenhang bei metrischen Merkmalen

Definition: Korrelationskoeffizient r_{xy}

Gegeben sei die gemeinsame Verteilung zweier metrisch skalierter Merkmale X und Y mit der Cov(X, Y) und den Standardabweichungen s_X und s_Y.

$$r_{xy} = \frac{Cov(X,Y)}{s_x \cdot s_y}$$

heißt dann Korrelationskoeffizient und ist ein Maß für die Höhe des linearen Zusammenhanges von X und Y. Er bewegt sich im Intervall zwischen -1 und 1, dessen Grenzen er bei vollständig negativer bzw. vollständig positiver linearer Abhängigkeit erreicht.

Aufgabe:

Berechnen Sie für die beiden Merkmale 'Übernachtung' und 'Gesamtbewertung' den Korrelationskoeffizienten als geeignete Maßzahl für den linearen Zusammenhang bei metrischen Merkmalen. Nutzen Sie hierzu die nachfolgende Arbeitstabelle und interpretieren Sie Ihr Ergebnis.

	Nächte	**Bewertung**			
Gast	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	4	47	188	16	2209
2	2	43	86		
3	1	41	41		
4	4	46	184		
5	5	47	235		
6	3	45	135		
7	2	41	82		
8	3	46	138		
9	2	44	88		
10	3	45	135		
11	2	41	82		
12	1	42	42		
13	1	41	41		
14	3	44	132		
15	5	46	230		
16	5	46	230		
17	3	49	147		
18	5	50	250		
19	3	42	126		
20	2	43	86		
Summe:	**59**	**889**	**2678**		
Mittelwert:	**2,950**	**44,450**			

9. Zusammenhang bei metrischen Merkmalen

Exkurs: Test auf lineare Abhängigkeit (Signifikanz des Korrelationskoeffizienten)

Analog zur Maßzahl von Cramér lässt sich auch für den Korrelationskoeffizienten r_{xy} ein Testverfahren konstruieren, das ein signifikantes Abweichen von Null untersucht. Nur wenn der Korrelationskoeffizient aus der gewählten Stichprobe signifikant von Null abweicht, ist von einem tatsächlichen linearen Zusammenhang zwischen X und Y in der Grundgesamtheit auszugehen und die Stärke dieses Zusammenhangs kann anhand des errechneten Wertes eingeschätzt werden.

Es wird also eine Grundgesamtheit unterstellt, in der die beiden Merkmale X und Y linear unabhängig sind:

H_0: Die Merkmale X und Y sind linear unabhängig

(H_1: Die Merkmale X und Y sind linear abhängig)

Bei Gültigkeit der Nullhypothese und bei normalverteilten Variablen X und Y lässt sich zeigen (hier keine Beweisführung), dass die Größe

$$T = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \cdot \sqrt{N-2}$$

mit: N-2=df Freiheitsgraden

einer t - Verteilung folgt.

$T \sim t_{df}$ mit: $df = n-2$

Liegt nun die über den Stichproben-Korrelationskoeffizienten errechnete Testgröße T im kritischen Bereich, dann ist die Nullhypothese abzulehnen. Die Grenze des kritischen Bereichs errechnet sich aus dem (über das Signifikanzniveau) festgelegte Quantil der t - Verteilung.
Dabei ist allerdings zu berücksichtigen, dass der Korrelationskoeffizient sowohl positive Werte bis +1 als auch negative Werte bis -1 annehmen kann. Der kritische Bereich ist demnach in einen oberen (positiven) und einen unteren (negativen) Bereich aufzuteilen (= zweiseitiger Test). Sie sind so zu wählen, dass sie jeweils die Hälfte der festgelegten Irrtumswahrscheinlichkeit α abdecken.

$KB_\alpha = (-\infty ; -t_{df;\, 1-\alpha/2}] + [t_{df;\, 1-\alpha/2} ; +\infty)$

Beispiel:

Für die beiden Merkmale 'Übernachtung' (F4: Nach) und 'Gesamtbewertung' (F6: Gbew) hat sich aus der vorliegenden Stichprobe im Umfang von 20 Gästen ein Korrelationskoeffizent in Höhe von 0,795 ergeben. Er weist auf einen starken positiven linearen Zusammenhang hin, ist allerdings vorab noch auf Signifikanz zu überprüfen.

Die Nullhypothese

H_0: Die Merkmale X und Y sind linear unabhängig

kann nur dann abgelehnt werden, wenn die Testgröße

9. Zusammenhang bei metrischen Merkmalen

$$T = \frac{0{,}795}{\sqrt{1-0{,}795^2}} \cdot \sqrt{20-2} = 5{,}560$$

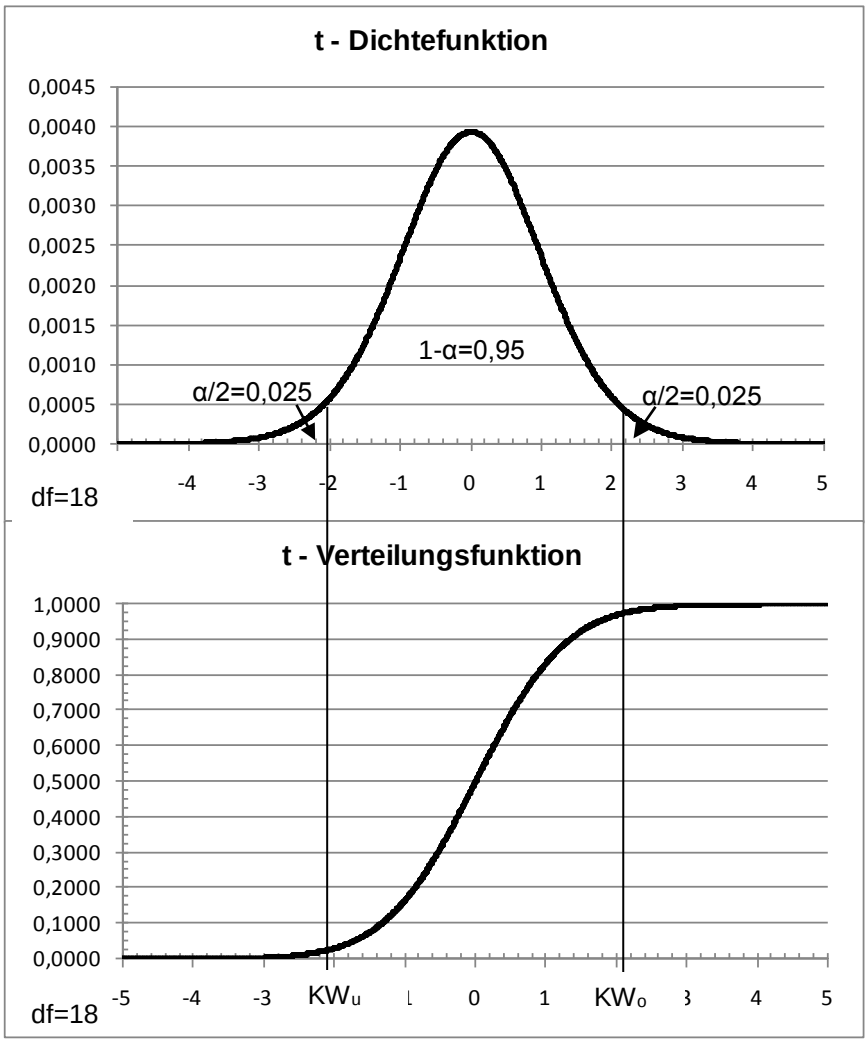
im kritischen Bereich liegt. Da bekannt ist, dass diese Testgröße grundsätzlich einer t - Verteilung mit

df = n-2 = 20-2 = 18 Freiheitsgraden

folgt, lassen sich bei vorgegebener Irrtumswahrscheinlichkeit (Signifikanzniveau)

α = 0,05

die Grenzen des kritischen Bereichs über die t - Verteilung bestimmen.



Danach liegt der kritische Wert zu Beginn des oberen Bereichs bei etwa 2. Tabelle 4 liefert für das 0,975-Quantil bei 18 Freiheitsgraden einen exakten Wert von 2,1009 (vgl. auch Wert der Excel-Funktion 'TINV(0,05; 18)' mit 0,05=α und 18=df). Der kritische Wert des unteren Bereichs liegt dann wegen der Symmetrie der t - Verteilung bei -2,1009.

Da der Wert der Testgröße mit 5,56 eindeutig im oberen kritischen Bereich liegt, muss die Nullhypothese abgelehnt werden. Es ist also von einem signifikanten linearen Zusammenhang zwischen der Anzahl an Übernachtungen und der Gesamtbewertung der Jugendherberge auszugehen, der nach dem Ergebnis des Stichproben-Korrelationskoeffizienten stark positiv ausgeprägt ist.

Tabelle 4

Quantile (1-α/2) der t - Verteilung (Studentverteilung)

df	0,500	0,750	0,800	0,900	0,950	0,975	0,995
1	0,0000	1,0000	1,3764	3,0777	6,3138	12,7062	63,6567
2	0,0000	0,8165	1,0607	1,8856	2,9200	4,3027	9,9248
3	0,0000	0,7649	0,9785	1,6377	2,3534	3,1824	5,8409
4	0,0000	0,7407	0,9410	1,5332	2,1318	2,7764	4,6041
5	0,0000	0,7267	0,9195	1,4759	2,0150	2,5706	4,0321
6	0,0000	0,7176	0,9057	1,4398	1,9432	2,4469	3,7074
7	0,0000	0,7111	0,8960	1,4149	1,8946	2,3646	3,4995
8	0,0000	0,7064	0,8889	1,3968	1,8595	2,3060	3,3554
9	0,0000	0,7027	0,8834	1,3830	1,8331	2,2622	3,2498
10	0,0000	0,6998	0,8791	1,3722	1,8125	2,2281	3,1693
11	0,0000	0,6974	0,8755	1,3634	1,7959	2,2010	3,1058
12	0,0000	0,6955	0,8726	1,3562	1,7823	2,1788	3,0545
13	0,0000	0,6938	0,8702	1,3502	1,7709	2,1604	3,0123
14	0,0000	0,6924	0,8681	1,3450	1,7613	2,1448	2,9768
15	0,0000	0,6912	0,8662	1,3406	1,7531	2,1314	2,9467
16	0,0000	0,6901	0,8647	1,3368	1,7459	2,1199	2,9208
17	0,0000	0,6892	0,8633	1,3334	1,7396	2,1098	2,8982
18	0,0000	0,6884	0,8620	1,3304	1,7341	2,1009	2,8784
19	0,0000	0,6876	0,8610	1,3277	1,7291	2,0930	2,8609
20	0,0000	0,6870	0,8600	1,3253	1,7247	2,0860	2,8453
21	0,0000	0,6864	0,8591	1,3232	1,7207	2,0796	2,8314
22	0,0000	0,6858	0,8583	1,3212	1,7171	2,0739	2,8188
23	0,0000	0,6853	0,8575	1,3195	1,7139	2,0687	2,8073
24	0,0000	0,6848	0,8569	1,3178	1,7109	2,0639	2,7969
25	0,0000	0,6844	0,8562	1,3163	1,7081	2,0595	2,7874
26	0,0000	0,6840	0,8557	1,3150	1,7056	2,0555	2,7787
27	0,0000	0,6837	0,8551	1,3137	1,7033	2,0518	2,7707
28	0,0000	0,6834	0,8546	1,3125	1,7011	2,0484	2,7633
29	0,0000	0,6830	0,8542	1,3114	1,6991	2,0452	2,7564
30	0,0000	0,6828	0,8538	1,3104	1,6973	2,0423	2,7500
40	0,0000	0,6807	0,8507	1,3031	1,6839	2,0211	2,7045
50	0,0000	0,6794	0,8489	1,2987	1,6759	2,0086	2,6778
60	0,0000	0,6786	0,8477	1,2958	1,6706	2,0003	2,6603
70	0,0000	0,6780	0,8468	1,2938	1,6669	1,9944	2,6479
100	0,0000	0,6770	0,8452	1,2901	1,6602	1,9840	2,6259

Lesebeispiel: $1-\alpha/2 = 0{,}975$ und $df = 18 \rightarrow t_{0{,}975;\,18} = 2{,}1009$

$\rightarrow f(t \leq 2{,}1009) = F(2{,}1009) = 0{,}975$

10. Zusammenhang bei ordinalen Merkmalen

Ordinal skalierte Merkmale weisen zwar - wie metrische Merkmale auch - eine natürliche Rangordnung auf, ihre Abstände sind jedoch nicht interpretierbar. Kennzahlen zur Berechnung der Stärke des Zusammenhangs zwischen zwei metrisch skalierten Merkmalen, die die unterschiedlichen Abstände zwischen den Ausprägungen erfassen, lassen sich also nicht auf ordinale Merkmale übertragen.
Vielmehr sind vorab deren Ausprägungen so zu transformieren, dass die Abstände aller benachbarten und vorher sortierten Merkmalswerte konstant bleiben. Dies gelingt am einfachsten durch die fortlaufende Vergabe von Rangzahlen von 1 bis N für jede Merkmalsausprägung eines geordneten Datensatzes. Kommen gleiche Merkmalsausprägungen vor, dann wird eine mittlere (identische) Rangzahl für diese Werte vergeben (= Bindungen).

Die gebildeten Rangzahlen können nun wie metrisch skalierte Merkmale behandelt werden, für sie lässt sich dann problemlos der bereits bekannte Korrelationskoeffizient nach Pearson/ Bravais berechnen. Im Vergleich zu alternativen Verfahren ist die Anwendung des klassischen Korrelationskoeffizienten zwar etwas rechenintensiver, er führt jedoch bei der Vergabe von durchschnittlichen Rangzahlen (Bindungen) zu unverzerrten Ergebnissen und soll daher den Vorzug erhalten.

Definition: Rangkorrelationskoeffizient R

Gegeben seien zwei gemeinsam erhobene und mindestens ordinal skalierte Merkmale X und Y. Den aufsteigenden Beobachtungen x_i und y_i lassen sich Rangzahlen $R(x_i)$ und $R(y_i)$ von 1 bis N zuordnen. Bei gleichen Merkmalsausprägungen werden durchschnittliche Rangzahlen vergeben.

Der Rangkorrelationskoeffizient R lässt sich dann auf der Grundlage der für X und Y gebildeten Rangzahlen nach dem Korrelationskoeffizienten von Person/Bravais berechnen:

$$R = \frac{Cov(R(x),R(y))}{s_{R(x)} \cdot s_{R(y)}}$$

mit:

$$Cov(R(x),R(y)) = \frac{1}{N}\sum_{i=1}^{N} R(x_i)\cdot R(y_i) - \overline{R(x)}\cdot\overline{R(y)}$$

$$s^2_{R(x)} = \frac{1}{N}\sum_{i=1}^{N} (R(x_i))^2 - \overline{R(x)}^2$$

$$s^2_{R(y)} = \frac{1}{N}\sum_{i=1}^{N} (R(y_i))^2 - \overline{R(y)}^2$$

$$\overline{R(x)} = \overline{R(y)} = \frac{N+1}{2}$$

10. Zusammenhang bei ordinalen Merkmalen

Der Rangkorrelationskoeffizient bewegt sich ebenfalls im Intervall zwischen -1 und 1, wobei er diese Grenzen bei vollständig negativer bzw. vollständig positiver monotoner Abhängigkeit erreicht.

Beispiel:

Auf ordinalem Skalenniveau werden im Rahmen der Gästebefragung die Informationen zur Detailbewertung der Jugendherberge sowie zur Weiterempfehlung erfasst. Aus der Vielzahl der sich hieraus ergebenden Untersuchungsmöglichkeiten sollen die beiden Merkmale 'Bewertung der Verpflegung in der Jugendherberge' (F5d: Verp) und 'Bewertung von Service und Kompetenz des Personals' (F5c: SeKo) herausgegriffen werden. Es wird also mit Hilfe des Rangkorrelationskoeffizienten untersucht, ob die Bewertung der Verpflegung einen Einfluss auf die Bewertung von Service und Kompetenz des Personals hat. Um das Beispiel überschaubarer zu halten, werden lediglich die Wertepaare von 10 Gästen herangezogen:

Verp	**SeKo**					
X	**Y**	**R(X)**	**R(Y)**	**R(X) · R(Y)**	**$R(X)^2$**	**$R(Y)^2$**
5	5	8,0	7,5	60,00	64,00	56,3
4	4	3,0	2,5	7,50	9,00	6,3
4	5	3,0	7,5	22,50	9,00	56,3
5	5	8,0	7,5	60,00	64,00	56,3
5	4	8,0	2,5	20,00	64,00	6,3
4	4	3,0	2,5	7,50	9,00	6,3
5	5	8,0	7,5	60,00	64,00	56,3
5	5	8,0	7,5	60,00	64,00	56,3
4	5	3,0	7,5	22,50	9,00	56,3
4	4	3,0	2,5	7,50	9,00	6,3
45	**46**	**55,0**	**55,0**	**327,50**	**365,00**	**362,5**

$Cov(R(X), R(Y)) = 327{,}50/10 - 5{,}5 \cdot 5{,}5 = 2{,}500$

$s^2_{R(X)} = 365{,}0/10 - 5{,}5^2 = 6{,}250$

$s_{R(X)} = 2{,}50$

$s^2_{R(Y)} = 362{,}5/10 - 5{,}5^2 = 6{,}00$

$s_{R(Y)} = 2{,}449$

$$R = \frac{Cov(R(x),R(y))}{s_{R(x)} \cdot s_{R(y)}} = \frac{2{,}500}{2{,}50 \cdot 2{,}449} = 0{,}408$$

10. Zusammenhang bei ordinalen Merkmalen

Der Rangkorrelationskoeffizient weist somit auf einen mittleren positiven monotonen Zusammenhang hin. Verbessert sich die Bewertung der Verpflegung, dann verbessert sich tendenziell auch die Bewertung für den gesamten Service und Kompetenz des Personals.

Um zu einer solchen verallgemeinernden Aussage zu gelangen, sollte natürlich auch dieses Ergebnis vorab auf Signifikanz getestet werden. Da der Rangkorrelationskoeffizient in der hier verwendeten Variante dem Korrelationskoeffizienten entspricht, kann auch das in Kapitel 9 vorgestellte Testverfahren genutzt werden. In Zweifel zu ziehen sind allerdings - insbesondere bei dem hier vorliegenden geringen Stichprobenumfang - normalverteilte Variablen, hier also normalverteilte Rangzahlen. Hierüber werden wir an dieser Stelle großzügig hinwegsehen.

So kann Nullhypothese

H_0: Die Rangzahlen der Merkmale X und Y sind monoton unabhängig

nur dann abgelehnt werden, wenn die Testgröße

$$T = \frac{0{,}408}{\sqrt{1-0{,}408^2}} \cdot \sqrt{10-2} = 1{,}264$$

im kritischen Bereich liegt.

Da bekannt ist, dass diese Testgröße grundsätzlich einer t - Verteilung mit

df = n-2 = 10-2 = 8 Freiheitsgraden

folgt, lassen sich bei vorgegebener Irrtumswahrscheinlichkeit (Signifikanzniveau)

$\alpha = 0{,}05$

die Grenzen des kritischen Bereichs über die t - Verteilung bestimmen.

$KB_{0,05} = (-\infty ; -t_{8;\ 1-0,025}] + [t_{8;\ 1-0,025} ; +\infty) = (-\infty ; -2{,}3060] + [2{,}3060 ; +\infty)$

Da der Wert der Testgröße mit 1,264 eindeutig außerhalb der kritischen Bereiche liegt, kann die Nullhypothese nicht abgelehnt werden. Der in der Stichprobe gemessene monotone Zusammenhang ist statistisch nicht gesichert und sollte daher auch nicht verallgemeinert werden.

11. Zusammenhang bei Merkmalen mit unterschiedlichem Skalenniveau

Nachdem in den letzten drei Kapiteln Zusammenhangsmaße von zwei nominalen Merkmalen, von zwei metrischen Merkmalen und von zwei ordinalen Merkmalen vorgestellt wurden, stellt sich jetzt natürlich die berechtigte Frage, wie bei zwei Merkmalen mit unterschiedlichem Skalenniveau zu verfahren ist.

Da eine detaillierte Darstellung dieser Verfahren, den Rahmen dieses Skriptes 'leicht' sprengen würde, soll hier nur auf die grundlegenden Aspekte eingegangen werden. Dabei wird nach den drei Kombinationsmöglichkeiten nominal-ordinal, nominal-metrisch und ordinal-metrisch unterschieden und soweit notwendig nach der Zuordnung des Skalenniveaus zur X- oder Y-Variablen.

Kombination eines nominalen und eines ordinalen Merkmals

Wenn der Zusammenhang eines nominalen und eines ordinalen Merkmals bestimmt werden soll, wird meist auf die Berechnung der Maßzahl von Cramér zurückgegriffen. Dies ist allerdings nur sinnvoll, wenn die Ausprägungen des ordinalen Merkmals überschaubar sind.

In der Praxis kommen häufig auch statistische Testverfahren wie U-Test oder H-Test zum Einsatz. Dabei geht es im Wesentlichen darum, Unterschiede zwischen den nominalen Ausprägungen bezüglich der ordinalen Variablen zu erkennen.

Liegt die nominale Variable in dichotomer Form vor, existieren also nur zwei Merkmalsausprägungen zu dieser Variablen, dann lässt sich der in Kapitel 10 vorgestellte Rangkorrelationskoeffizient anwenden. Dieser wird dann als biserialer Rangkorrelationskoeffizient R_{bis} bezeichnet.

Kombination eines nominalen und eines metrischen Merkmals

Auch für die Kombination eines nominalen und eines metrischen Merkmals gibt es kein adäquates Zusammenhangsmaß. Im Notfall muss - analog zum vorangegangenen Abschnitt - die Maßzahl von Cramér 'herhalten'. Aufgrund einer Vielzahl von Ausprägungen des metrischen Merkmals ist i.d.R. vorab eine Klassenbildung erforderlich. Die Klassen sind dann lediglich als unterschiedliche Ausprägungen des metrischen Merkmals zu interpretieren.

Häufig wird auch hier auf statistische Testverfahren zurückgegriffen, um die Unterschiede zwischen den Gruppen des nominalen Merkmals hinsichtlich einer metrischen Variable zu überprüfen. Diese Unterschiede können dann mit Hilfe des T-Tests auf Signifikanz überprüft werden. Allerdings müssen dazu alle Kombinationen der nominalen Gruppen untersucht werden. Die Varianzanalyse bzw. Diskriminanzanalyse fassen diese T-Test - Kombinationen zu einem (umfangreicheren) multivariaten Verfahren zusammen. Bei der Varianzanalyse wird der Einfluss von mindestens einer nominalen X-Variablen auf eine metrische Y-Variable auf Signifikanz getestet. Bei der Diskriminanzanalyse verhält es sich genau umgekehrt, es wird der Einfluss von mindestens einer metrischen X-Variablen auf eine nominale Y-Variable untersucht.

Alternativ kommt mitunter der punktbiseriale Korrelationskoeffizient r_{pb} zum Einsatz. Analog zu den Ausführungen bei nominalen und ordinalen Merkmalen muss auch hier eine dichotome Variable, also ein nominales Merkmal mit nur zwei Ausprägungen, vorliegen. Diese Ausprägungen werden in eine 0-1 - Variable codiert und unter Anwendung des klassischen Korrelationskoeffizienten verarbeitet.

Kombination eines ordinalen und eines metrischen Merkmals

Auch hier besteht die einfachste Vorgehensweise darin, anstelle des Korrelationskoeffizienten r_{xy} das niedrigere Zusammenhangsmaß, eben den Rangkorrelationskoeffizienten R, anzuwenden. Alternative Vorgehensweisen hierzu haben sich in der wissenschaftlichen Anwendung nicht durchgesetzt. Dies liegt schlicht und einfach daran, das der Rangkorrelationskoeffizient in dieser Situation ausreichend gute Lösungen liefert.

Aufgaben und Lösungen

Aufgabe III.1.:

200 männliche Personen wurden nach ihrem Berufsstand und dem ihres Vaters befragt. Unter den 200 Personen waren 60 Söhne von Angestellten:

Sohn (Y)	Vater (X) Arbeiter	Angestellter	Beamter	Selbständiger
Arbeiter	40	10	0	
Angestellter	40		5	10
Beamter	10	25	25	0
Selbständiger	0	0	0	10

a) Ergänzen Sie die fehlenden Angaben in der Tabelle. Berechnen und interpretieren Sie außerdem die Randhäufigkeiten.

b) Beantworten Sie die Frage, ob der Berufsstand des Sohnes von dem des Vaters abhängig ist. Berechnen Sie hierzu eine geeignete Kennzahl und interpretieren Sie diese.

c) Formulieren Sie für den vorliegenden Sachverhalt eine entsprechende Nullhypothese und testen Sie diese auf einem Signifikanzniveau von 5%.

Lösung:

a)

Vater (X)	Sohn (Y) Arbeiter	Angestellter	Beamter	Selbständiger	Summe
Arbeiter	40	40	10	0	90
Angestellter	10	25	25	0	60
Beamter	0	5	25	0	30
Selbständiger	0	10	0	10	20
Summe	50	80	60	10	200

$h_{2.} = 60$

b) bei Unabhängigkeit gilt: $h_{ik} = (h_{i.} \cdot h_{.k}) / N$
da z.B. $h_{41} = 0 <> (20 \cdot 50) / 200 = 5$ besteht Abhängigkeit

$$\chi^2 = 200 \cdot \left[\left(\frac{40^2}{90 \cdot 50} + ... + \frac{10^2}{60 \cdot 50} + \frac{25^2}{60 \cdot 80} + ... + \frac{10^2}{20 \cdot 10} + \right) - 1 \right] = 170{,}72$$

$$C = \sqrt{\frac{1}{200} \cdot \frac{170{,}72}{\min((4-1);(4-1))}} = 0{,}533$$

Der Berufsstand des Sohnes hängt also nicht unerheblich von dem des Vaters ab.

c) H_0: Berufsstand Vater und Berufsstand Sohn sind statistisch unabhängig

$\alpha = 0{,}05$ / df = $(4\text{-}1) \cdot (4\text{-}1) = 9$

$T = 170{,}72 \sim X^2_9$

$KB_{0{,}05} = [X^2_{9;\,0{,}95} ; +\infty) = [16{,}9190; +\infty)$

→ eindeutige Ablehnung der Nullhypothese, da T im kritischen Bereich liegt

Aufgaben und Lösungen

Aufgabe III.2.:

Neben monatlichen Einkommen (in Euro) wurden von Wismarer Studenten Informationen zu ihren monatlichen Ausgaben (in Euro) und ihrem häufigsten Fortbewegungsmittel (1 = zu Fuß bzw. Fahrrad, 2 = Auto) erfragt:

Einkommen	500	400	400	450	500	200	350	450	500	650
Ausgaben	400	200	300	450	300	200	350	450	500	500
Fortbewegung	2	1	2	2	1	2	2	2	2	2

a) Nennen Sie die in der Aufgabe betrachtete statistische Einheit bzw. statistische Masse sowie die Merkmale mit dem zugehörigen Skalenniveau.

b) Stellen Sie die Abhängigkeit der Ausgaben vom Einkommen grafisch dar und geben Sie eine begründete Schätzung für die Stärke des linearen Zusammenhangs.

c) Überprüfen Sie Ihre Einschätzung unter b), indem Sie eine geeignete Kennzahl für diesen Zusammenhang berechnen.

d) Lässt sich am vorliegenden Datenmaterial ein Einfluss des Fortbewegungsmittels auf die Ausgaben erkennen? Berechnen Sie ein Zusammenhangsmaß, das den vorliegenden Skalenniveaus gerecht wird und interpretieren Sie das Ergebnis.

Lösung:

Arbeitstabelle:

Studis	Eink. (X)	Ausg. (Y)	x_i^2	y_i^2	$x_i \cdot y_i$
1	500	400	250000	160000	200000
2	400	200	160000	40000	80000
3	400	300	160000	90000	120000
4	450	450	202500	202500	202500
5	500	300	250000	90000	150000
6	200	200	40000	40000	40000
7	350	350	122500	122500	122500
8	450	450	202500	202500	202500
9	500	500	250000	250000	250000
10	650	500	422500	250000	325000
	4400	3650	2060000	1447500	1692500

a) 1 Student, 10 Studierende in Wismar,
Einkommen (metrisch-verhältnis),
Ausgaben (metrisch-verhältnis),
Fortbewegungsmittel (nominal)

b) eindeutige Richtung der Punktewolke von links unten nach rechts oben, allerdings liegen nicht alle Punkte auf einer Geraden, starker positiver linearer Zusammenhang

c) $r_{xy} = 0{,}724$

d) ja, da bei Fortbewegung zu Fuß/Fahrrad die niedrigsten Ausgaben von 200 bzw. 300 EUR; da Fortbewegung nominal und Ausgaben metrisch, punktbiserialer Korrelationskoeffizient oder Varianzanalyse

Aufgabe III.3.:

Bei der Veröffentlichung von Unternehmenszahlen für 2001 hat VW auch auf den Einfluss des PKW-Absatzes (in Mio. Stück) auf das Betriebsergebnis (in Mio. EUR) verwiesen. Für ausgewählte Jahre liegen die folgenden Ergebnisse vor:

Jahr	1992	1993	1996	1997	1998	1999	2000
Betriebsergebnis	75	-992	347	696	1147	844	2062
PKW-Absatz	3,4	3,0	4,0	4,3	4,8	4,9	5,2

a) Stellen Sie die Abhängigkeit des Betriebsergebnisses vom PKW-Absatz grafisch dar. Welche Art und welche Richtung des Zusammenhangs wird hieran deutlich?

b) Berechnen Sie eine geeignete Kennzahl für den unter a) festgestellten Zusammenhang und begründen Sie Ihr Ergebnis ökonomisch.

c) Testen Sie Ihr unter b) gewonnenes Ergebnis auf Signifikanz.

Lösung:

a) Streudiagramm mit Absatz = X und Betriebsergebnis = Y;
mit steigendem Absatz nimmt das Betriebsergebnis zu, dabei bewegen sich die Wertepaare eng um eine gedachte Gerade, folglich ein positiver linearer Zusammenhang

b) Arbeitstabelle:

Jahr	Absatz (X)	Ergebnis (Y)	x_i^2	y_i^2	$x_i \cdot y_i$
1992	3,4	75	11,56	5625	255,0
1993	3,0	-992	9	984064	-2976,0
1996	4,0	347	16	120409	1388,0
1997	4,3	696	18,49	484416	2992,8
1998	4,8	1147	23,04	1315609	5505,6
1999	4,9	844	24,01	712336	4135,6
2000	5,2	2062	27,04	4251844	10722,4
Summe:	29,6	4179,0	129,1	7874303	22023,4
Mittelwert:	4,229	597,000			
Varianz:	0,568	768491,429			
Stan.abw.:	0,754	876,636			

Cov(X, Y) = 22023,4 / 7 - 4,229 · 597 = 621,743 [Mio.Stück · Mio.Euro]
r_{xy} = 621,743 / (0,754 · 876,636) = 0,941
d.h. sehr starke lineare Abhängigkeit zwischen PKW-Absatz und dem Betriebsergebnis von VW; bei positivem Deckungsbeitrag je PKW erhöht sich auch das Betriebsergebnis

c) H_0: Absatz und Betriebsergebnis sind linear unabhängig

α = 0,05 / df = 7-2 = 5

$$T = \frac{0{,}941}{\sqrt{1-0{,}941^2}} \cdot \sqrt{7-2} = 6{,}218$$

$KB_{0,05} = (-\infty ; -t_{5;\ 1-0,025}] + [t_{5;\ 1-0,025} ; +\infty) = (-\infty ; -2{,}5706] + [2{,}5706 ; +\infty)$

→ Ablehnung der Nullhypothese, linearer Zusammenhang ist trotz kleiner Stichprobe signifikant

Aufgabe III.4.:

Aus der Erhebung zum Verkehrsverhalten (KONTIV) 2005 stammen die folgenden Angaben über die Anzahl der PKW im Haushalt (in 1000) nach Haushaltsgrößen:

PKW-Zahl	Haushaltsgröße			
	1 Person	2 Personen	3 Personen	4 Personen
kein PKW	5128	1703	315	155
ein PKW	4443	5101	2562	1767
zwei PKW	3	1416	1574	1221

a) Berechnen Sie zunächst eine allgemeine Maßzahl für die Stärke des Zusammenhangs zwischen beiden Merkmalen (ohne Annahme eines bestimmten Skalenniveaus).

b) Begründen Sie, warum die Berechnung des Rangkorrelationskoeffizienten von Spearman hier möglich wäre. Welches Problem tritt hierbei jedoch auf?

c) Berechnen Sie ein nach dem Skalenniveau der Merkmale adäquates Maß für den Zusammenhang der beiden Merkmale und interpretieren Sie das Ergebnis. Lässt sich dieses Ergebnis verallgemeinern?

Lösung:

a) Maß von Cramér

$$\chi^2 = 25388 \cdot \left[\left(\frac{5128^2}{9574 \cdot 7301} + \ldots + \frac{4443^2}{9574 \cdot 13873} + \ldots + \frac{3^2}{9574 \cdot 4214} + \right) - 1 \right] = 7257{,}557$$

$$C = \sqrt{\frac{1}{25388} \cdot \frac{7257{,}557}{\min((4-1);(3-1))}} = 0{,}378$$

nach Maß von Cramér ein relativ schwacher Zusammenhang zwischen Haushaltsgröße und PKW-Anzahl

b) mindestens ordinal skalierte Merkmale; Problem: sehr große Zahl an Bindungen

c) $\overline{y}_{PKW} = (0 \cdot 7301 + 1 \cdot 13873 + 2 \cdot 4214) / 25388 = 0{,}878$ PKW
$s_{PKW}^2 = (0^2 \cdot 7301 + 1^2 \cdot 13873 + 2^2 \cdot 4214) / 25388 - 0{,}878^2 = 0{,}439$ PKW^2
$s_{PKW} = 0{,}662$ PKW

$\overline{x}_{HH} = (1 \cdot 9574 + 2 \cdot 8220 + 3 \cdot 4451 + 4 \cdot 3143) / 25388 = 2{,}046$ Personen
$s_{HH}^2 = 1{,}046$ Personen^2
$s_{HH} = 1{,}023$ Personen

Cov (X, Y) = $(0 \cdot 1 \cdot 5128 + \ldots + 1 \cdot 1 \cdot 4430 + \ldots) / 25388 - 0{,}878 \cdot 2{,}046 = 0{,}342$ PKW·Pers.

rxy = 0,342 / (0,662 · 1,023) = 0,503

H_0: Haushaltsgröße und PKW-Zahl sind linear unabhängig

$$T = \frac{0{,}503}{\sqrt{1-0{,}503^2}} \cdot \sqrt{25388-2} = 92{,}737$$

$KB_{0{,}05} = (-\infty ; -t_{25386;\ 1-0{,}025}] + [t_{25386;\ 1-0{,}025} ; +\infty) = (-\infty ; -1{,}96] + [1{,}96 ; +\infty)$

→ eindeutige Ablehnung der Nullhypothese, Ergebnis lässt sich also verallgemeinern

Aufgabe III.5.:

Ein Händler erhält Informationen zum Flaschenpreis und zur abgesetzten Menge einer bestimmten Sektmarke sowie zum Wetter an diesen Tagen:

Tag	Flaschen	Preis in EUR	Wetter
1. Tag	150	4,00	schlecht
2. Tag	140	5,00	wechselhaft
3. Tag	120	5,00	schön
4. Tag	80	7,00	wechselhaft
5. Tag	50	6,00	schön

a) Die Kovarianz zwischen dem Flaschenpreis und der abgesetzten Menge ergibt einen Wert von -31,2. Welche Schwierigkeiten treten bei der Interpretation dieser Größe auf?

b) Welche verbesserte Maßzahl lässt sich aus der Kovarianz unter a) ableiten? Berechnen und interpretieren Sie deren Wert.

c) Berechnen Sie eine geeignete Maßzahl für die Stärke des Zusammenhangs zwischen Wetter und abgesetzter Menge und bewerten Sie das Ergebnis.

Lösung:

a) Nur Interpretation der Richtung, aber nicht der Stärke des linearen Zusammenhangs

b) r_{xy} = -31,2 / (1,020 · 37,630) = -0,813 d.h. starker negativer linearer Zusammenhang; mit steigenden Preisen nimmt der Flaschenabsatz ab

c)

Tag	Wetter (X)	Flaschen (Y)	R(X)	R(Y)	R(X) · R(Y)	$R(X)^2$	$R(Y)^2$
1. Tag	schlecht	150	1,0	5	5,0	1,00	25
2. Tag	wechselhaft	140	2,5	4	10,0	6,25	16
3. Tag	schön	120	4,5	3	13,5	20,25	9
4. Tag	wechselhaft	80	2,5	2	5,0	6,25	4
5. Tag	schön	50	4,5	1	4,5	20,25	1
Summe:			15,0	15	38,0	54,00	55
Mittelwert:			3,00	3,00			
Varianz:			1,80	2,00			
Stand.abw.:			1,342	1,414			

Cov(R(X), R(Y)) = 38,0/5 - 3,0·3,0 = -1,4

$$R = \frac{-1{,}4}{1{,}342 \cdot 1{,}414} = -0{,}738$$

d.h. zwischen der Nachfrage nach Sekt und dem Wetter besteht ein starker negativer monotoner Zusammenhang; je schlechter das Wetter, desto mehr Sekt wird gekauft

H_0: Wetter und Flaschenabsatz sind monoton unabhängig

$$T = \frac{-0{,}738}{\sqrt{1-0{,}738^2}} \cdot \sqrt{5-2} = -1{,}894$$

$KB_{0,05} = (-\infty ; -t_{3;\ 1-0,025}] + [t_{3;\ 1-0,025} ; +\infty) = (-\infty ; -3{,}1824] + [3{,}1824 ; +\infty)$

→ keine Ablehnung der Nullhypothese, das Ergebnis ist also nicht signifikant

Aufgabe III.6.:

Im Jahr 2005 wurden bundesweit an allen Hochschulen getrennt nach Fachbereichen die vier Merkmale A = Betreuungsintensität (wissenschaftliches Personal / Anzahl Studierende), B = Studentenbudget (Ausgaben der Hochschule in TEUR / Anzahl Studierende), C = Absolventenquote (Absolvent / Anzahl Studierende) und D = Berufsaussichten der Absolventen erhoben:

Fachbereich	A	B	C	D
Sprachwissenschaften	0,05	5,0	0,035	schlecht
Wirtschaftswissenschaften	0,03	3,3	0,050	gut
Naturwissenschaften	0,10	13,4	0,050	gut
Humanmedizin	0,20	84,3	0,110	unsicher
Ingenieurwissenschaften	0,06	8,2	0,030	sehr gut

a) Beschreiben Sie das Skalenniveau der betrachteten Merkmale.

b) Berechnen Sie für die Stärke des Zusammenhangs zwischen Betreuungsintensität und der Absolventenquote eine Maßzahl, die das Skalenniveau der beiden Merkmale voll ausnutzt und interpretieren Sie diese. Gehen Sie bei Ihren Berechnungen von einer Gleichverteilung der Studierendenzahlen über alle fünf Fachbereiche aus.

c) Es wird oft gefordert, dass die verschiedenen Wissenschaften entsprechend der Berufsaussichten der Absolventen gefördert werden sollten. Kann diese Forderung hier zumindest näherungsweise als erfüllt betrachtet werden? Unterstellen Sie auch hier eine Gleichverteilung der Studierendenzahlen und berechnen Sie zur Begründung Ihrer Aussage eine geeignete Kennzahl.

Lösung:

a) metrisch-verhältnis skaliert:
Betreuungsintensität (A), Studentenbudget (B), Absolventenquote (C)
ordinal skaliert:
Berufsaussichten (D)

b) $\overline{x_A} = 0{,}088$ $s_A^2 = 0{,}057/5 - 0{,}088^2 = 0{,}003656$ $s_A = 0{,}060$

$\overline{x_C} = 0{,}055$ $s_C^2 = 0{,}019225/5 - 0{,}055^2 = 0{,}00082$ $s_C = 0{,}029$

Cov (X, Y) = 0,03205/5 - 0,088 · 0,055 = 0,00157

rxy = 0,907 d.h. zwischen Betreuungsintensität und Absolventenquote besteht ein sehr starker linearer Zusammenhang

c)

	Aussichten	**Budget**					
Fachbereich	**D (X)**	**B (Y)**	**R(X)**	**R(Y)**	**R(X) · R(Y)**	**R(X)²**	**R(Y)²**
Sprachwissenschaften	schlecht	5,0	1,0	2	2,0	1,00	4
Wirtschaftswissenschaften	gut	3,3	3,5	1	3,5	12,25	1
Naturwissenschaften	gut	13,4	3,5	4	14,0	12,25	16
Humanmedizin	unsicher	84,3	2,0	5	10,0	4,00	25
Ingenieurswissenschaften	sehr gut	8,2	5,0	3	15,0	25,00	9
Summe:			15,0	15	44,5	54,50	55
Mittelwert:			3,00	3,00			
Varianz:			1,90	2,00			
Stand.abw.:			1,378	1,414			

R = -0,051 d.h. kein monotoner Zusammenhang zwischen Aussichten und Budget

Aufgabe III.7.:

17 Studierende wurden nach ihren Prüfungsergebnissen in den Fächern EBWL und VWL befragt. Die Häufigkeitstabelle hat folgendes Aussehen:

Note EBWL	VWL 1	2	3	4	5
1	-	-	1	-	-
2	-	2	-	-	-
3	-	1	4	1	1
4	-	-	1	4	-
5	-	-	1	-	1

a) Welches Skalennivau besitzen die beiden betrachteten Merkmale? Beschreiben Sie dieses kurz. (ordinal skaliert)

b) Berechnen Sie für beide Merkmale die zulässigen Lageparameter. (x_D(EBWL)=3; x_D(VWL)=3; x_Z(EBWL)=$x_{(9)}$=3; x_Z(VWL)=3)

c) Zeigen Sie durch geeignete Berechnung bedingter relativer Häufigkeiten, dass zwischen der Note in VWL und der in EBWL eine Abhängigkeit besteht. Beschreiben Sie kurz die Richtung der Abhängigkeit zwischen den beiden Merkmalen. (f(y2|x1)=0, f(y2|x2)=1 → statistische Abhängigkeit; gleichgerichtete Abhängigkeit)

d) Begründen Sie, ob sich der Korrelationskoeffizient von Pearson/Bravais als Maß für die Stärke des Zusammenhanges zwischen den beiden vorliegenden Merkmalen sinnvoll berechnen lässt. (nein, da die Merkmale nicht metrisch skaliert sind)

Aufgabe III.8.:

Im Auftrag der Coca-Cola GmbH Deutschland testet die Consiglio-Marktforschung die Produktvariante Cola-Light. Folgende Angaben liegen Ihnen getrennt nach Geschlechtern über das Interesse an diesem Produkt vor:

Geschlecht	Interesse	kein Interesse
männlich	5	15
weiblich	65	15

a) Begründen Sie (keine Berechnung), ob die beiden Merkmale 'Geschlecht' und 'Interesse' an Cola-Light statistisch unabhängig sind. (abhängig)

b) Berechnen Sie alternativ eine geeignete Größe für den Zusammenhang zwischen den beiden Merkmalen. Welche Aussage können Sie hieraus ableiten? (C=0,491)

c) Wie müsste die zweidimensionale Häufigkeitstabelle aussehen, wenn zwischen den beiden Merkmalen eine vollständige statistische Abhängigkeit bestehen würde? (h_{11} = 0; h_{12} = 20; h_{21} = 80; h_{22} = 0)

Aufgabe III.9.:

Die Wismarer Übernachtungsbetriebe sind 2010 gegenüber 2008 deutlich schlechter bewertet worden. Eine Vermutung besteht darin, dass 2010 vermehrt ältere Personen in Wismar übernachtet haben, diese Gruppe allerdings eine höhere Erwartungshaltung hat und deshalb die Unterkunft kritischer bewertet. Zur Überprüfung dieser These liegen Ihnen Informationen von 10 Touristen zu deren Alter (in Jahren) und Bewertung der Unterkunft (mit Schulnoten von 1 bis 6) vor:

Tourist	1	2	3	4	5	6	7	8	9	10
Alter	45	45	70	65	50	24	45	65	75	36
Bewertung	4	3	5	6	4	1	3	5	6	2

a) Berechnen Sie für die betrachteten Merkmale jeweils einen geeigneten Lageparameter und interpretieren Sie diesen.

b) Erstellen Sie die zweidimensionale Häufigkeitstabelle für die beiden Merkmale. Lassen sich hieraus bereits Erkenntnisse über deren Abhängigkeit ableiten?

c) Berechnen Sie nun abschließend eine geeignete Größe zur Messung dieser Abhängigkeit und nehmen Sie zur obigen These der Wismarer Übernachtungsbetriebe Stellung.

Aufgabe III.10.:

Die Hochschule Wismar hat 300 Touristen in Abhängigkeit von ihrem Alter danach befragt, ob sie eher die Backsteingotik oder die Ostsee mit der Hansestadt in Verbindung bringen (Image). Die gewonnenen Informationen liegen leider nur in unvollständiger Form vor:

	Image	
Alter	Backsteingotik	Ostsee
[10; 35)	30	
[35; 45)		
[45; 90)	124	36

Zumindest ist aber bekannt, dass der Anteil der 10- bis 35-jährigen bei 16,67% liegt. Außerdem bringen 22,22% der 35- bis 45-jährigen die Hansestadt mit der Ostsee in Verbindung.

a) Vervollständigen Sie unter Berücksichtigung der zusätzlichen Informationen die obige Häufigkeitstabelle.

b) Begründen Sie mit Hilfe der bedingten relativen Häufigkeiten, ob zwischen beiden Merkmalen ein Zusammenhang erkennbar ist und schätzen Sie die Stärke dieses Zusammenhangs.

c) Überprüfen Sie Ihre Aussage unter b), indem Sie eine geeignete Größe für den Zusammenhang von Alter und Image berechnen. Interpretieren Sie das Ergebnis.

d) Irrtümlicherweise wurden für die Gruppe der 45- bis 90-jährigen die Häufigkeiten von 124 und 36 vertauscht. Welche Auswirkung hätte die Korrektur auf den Zusammenhang der beiden Merkmale. Begründen Sie Ihre Antwort (keine Berechnung!).

Aufgabe III.11.:

1000 repräsentativ ausgewählte Bäume in Mecklenburg-Vorpommern lassen sich in Abhängigkeit vom Alter in die folgenden vier Schadensklassen einteilen:

Alter	ohne Schäden	schwache Schäden	mittlere Schäden	starke Schäden
[0; 60)	288	168	16	3
[60; 120)	173	271	78	3

a) Berechnen Sie für die beiden Merkmale jeweils den aussagefähigsten Lageparameter.

b) Zeigen Sie durch geeignete Berechnung bedingter relativer Häufigkeiten, dass zwischen dem Alter der Bäume und ihrer Schadensklasse eine Abhängigkeit besteht.

c) Berechnen Sie das Maß von Cramér und interpretieren Sie diese Kennzahl.

d) Berechnen Sie alternativ zu c) andere Zusammenhangsmaße und vergleichen Sie die Ergebnisse.

Aufgabe III.12.:

Die Carl-von-Weinberg - Sportschule in Frankfurt wendet zur Talentsichtung den Jump and Reach - Test (JART) an. Mit diesem Verfahren wird bei Schülern in der fünften Jahrgangsstufe deren Sprungkraft gemessen und dann im siebten Schuljahr mit deren Hochsprungleistungen (HSL) verglichen. Zusätzlich wurde die Trainingsintensität (T-Intensität) im Hochsprung in diesen beiden Jahren erfasst .Hierzu liegen Ihnen nun folgende Informationen vor:

Schüler	T-Intensität	JART (cm)	HSL (cm)	Geschlecht
1	selten	36	128	w
2	selten	28	112	w
3	häufig	35	142	m
4	selten	33	122	w
5	häufig	27	110	m
6	selten	34	124	w
7	häufig	30	122	m
8	häufig	37	150	m
9	häufig	26	106	m
10	häufig	29	118	m

a) Stellen Sie den Einfluss des Jump and Reach - Tests auf die späteren Hochsprungleistungen in geeigneter Weise grafisch dar. Was fällt Ihnen auf?

b) Halten Sie den Einsatz des Jump and Reach - Tests für sinnvoll, um Talentsichtung zu betreiben? Berechnen Sie hierzu eine geeignete Kennzahl und überprüfen Sie, ob es sich hierbei um ein fundiertes Ergebnis handelt.

c) Halten Sie es darüber hinaus für notwendig, den Einfluss der Trainingsintensität bzw. des Geschlechts zu berücksichtigen? Führen Sie hierzu geeignete statistische Berechnungen durch.

Aufgabe III.13.:

Im Rahmen einer Verkehrsstudie wurde untersucht, ob die Einhaltung der Geschwindigkeit in einer Tempo-30-Zone davon abhängt, ob die Autofahrer selbst Anlieger sind oder nicht. Hieraus wurde die folgende Häufigkeitsverteilung erstellt:

	Geschwindigkeit	
Anlieger	bis 30 km/h	über 30 km/h
Anlieger	36	204
Nicht-Anlieger	6	154

a) Wie viel Prozent der Anlieger halten sich an die Geschwindigkeitsbeschränkung?

b) Wie viel Prozent der Verkehrssünder sind Anlieger?

c) Berechnen Sie eine geeignete Größe für den Zusammenhang von Anlieger und Geschwindigkeit und interpretieren Sie diese.

d) Die Verkehrsstudie wurde in einem zweiten Schritt um 200 Einheiten erweitert; drei Viertel hiervon waren Anlieger, wobei sich sowohl die Anlieger als auch die Nicht-Anlieger zu 40% an die Geschwindigkeitsvorgabe hielten.
Wie verändert sich hierdurch die Abhängigkeit der beiden Merkmale? Welchen Einfluss hat die Erweiterung der Verkehrsstudie auf die Aussagekraft der Ergebnisse?

Aufgabe III.14.:

In einem mittelständischen Unternehmen mit 100 Beschäftigten ergibt sich für die Merkmale Bruttogehalt und Geschlecht folgende relative Häufigkeitsverteilung:

	Bruttogehalt in EUR			
Geschlecht	[0; 2000)	[2000; 4000)	[4000; 6000)	[6000; 10000)
Männer	0,1	0,1	0,2	0,2
Frauen	0,05	0,1	0,2	0,05

a) Wie viel Prozent der Männer bzw. der Frauen verdienen jeweils über 5000 EUR?

b) Wie hoch ist der Frauenanteil unter den Beschäftigten mit einem Bruttogehalt unter 4000 EUR bzw. unter den Beschäftigten mit einem Bruttogehalt über 6000 EUR? Lassen sich hieraus bereits Rückschlüsse über die Abhängigkeit des Einkommens vom Geschlecht ableiten? Begründen Sie Ihre Antwort.

c) Berechnen Sie nun eine geeignete Größe für den Zusammenhang der beiden Merkmale und interpretieren Sie das Ergebnis.

d) Wie verändert sich Ihr Ergebnis unter c), wenn sich bei 200 Beschäftigten die gleiche relative Häufigkeitsverteilung ergeben würde? Wird auch die Aussagekraft der Ergebnisse hierdurch beeinflusst?

12. Regressionsrechnung

Während bei der Korrelationsrechnung die Stärke des linearen Zusammenhangs zwischen zwei Merkmalen untersucht wurde, wird bei der Regressionsrechnung ein funktionaler Zusammenhang zwischen dem Merkmal Y als abhängige und dem Merkmal X als unabhängige Variable bestimmt. Dabei bedarf der Zusammenhang zwischen beiden Merkmalen, die grundsätzlich auch metrisch skaliert sein müssen, zuallererst einer theoretisch fundierten Annahme. Der unterstellte Zusammenhang ist also vorab sachlogisch zu begründen.

Ist dieser funktionale Zusammenhang zwischen X und Y bestimmt und sind die Koeffizienten dieser Funktion aus dem vorliegenden Datenmaterial geschätzt, dann lassen sich mit Hilfe dieser Regressionsfunktion für beliebige Ausprägungen des Merkmals X die zu erwartenden Werte der abhängigen Variablen Y berechnen. In Erweiterung der Korrelationsanalysen ist es hier also möglich, unter Vorgabe eines konkreten X-Wertes einen Prognosewert für die Y-Variable zu berechnen. Zu beachten ist allerdings, dass die Genauigkeit dieser Schätzung umso stärker abnimmt, je weiter die Ausprägungen des Merkmals X vom Intervall der tatsächlichen Beobachtungswerte (Stützzeitraum) entfernt liegen.

Aus den Optionen, dass eine X-Variable oder auch mehrere X-Variablen zur Beschreibung des Einflusses auf die Y-Variable herangezogen und lineare oder nicht-linearer Zusammenhänge zwischen den X- und Y-Variablen unterstellt werden können, ergeben sich verschiedene Modellansätze. Diese sollen in den nachfolgenden Abschnitten vorgestellt werden.

Tabelle 5

Entwicklung des Tourismus von 2001 bis 2016

Jahr	ÜZ_JH in 1000	TI_D in Prozent	TI_G1020 in Prozent	TI_DJH in Prozent	MA_JH in TEUR
2001	15,1	102,7	101,2	99,5	9
2002	14,9	102,7	101,3	97,4	10
2003	14,6	100,0	98,7	98,3	12
2004	15,1	100,0	100,0	100,0	8
2005	15,7	100,1	102,1	101,6	8
2006	16,0	101,6	101,7	104,7	8
2007	16,4	103,8	102,9	107,3	6
2008	14,9	106,9	104,2	110,0	12
2009	16,1	109,2	105,9	115,2	8
2010	16,8	108,9	113,2	122,2	6
2011	17,1	112,4	112,0	124,7	6
2012	16,0	116,2	116,7	126,0	8
2013	16,2	120,3	119,7	132,5	12
2014	16,4	121,7	117,1	135,1	9
2015	16,8	125,3	119,4	138,3	8
2016	17,2	128,9	119,6	139,3	6

12. Regressionsanalyse

Analog zum bisherigen Procedere sollen auch hier Datensätze zur Jugendherberge herangezogen werden, um die vorgestellten Verfahren zumindest fallweise zu konkretisieren. Standen in den letzten Kapiteln vor allem Daten zur Kundenanalyse im Mittelpunkt, so soll nachfolgend auf Zeitreihen zur Markt- und Konkurrenzanalyse zurückgegriffen werden. Hierzu sind in Tabelle 5 über einen Zeitraum von 2001 bis 2016 die Übernachtungszahlen dieser Jugendherberge (ÜZ_JH), der Tourismusindex für Gesamtdeutschland (TI_D), für deutsche Gemeinden zwischen 10000 und 20000 Einwohnern (TI_G1020) und für die deutschen Jugendherbergen zusammen (TI_DJH) erfasst, darüber hinaus das jährliche Budget für Marketingausgaben der Jugendherberge selbst (MA_JH). Es bedarf in diesem Fall sicher keiner weiteren Ausführungen, dass die in diesen Indizes erfasste Geschäftslage des Tourismusgewerbes einen (positiven) Einfluss auf die Übernachtungszahlen der Jugendherberge haben könnte. In besonderer Weise sollte dies natürlich für das verausgabte Werbebudget und den Tourismusindex für kleinere Gemeinden gelten, da die Jugendherberge in einer Gemeinde dieser Größenordnung liegt.

12.1. Einfache lineare Regressionsrechnung

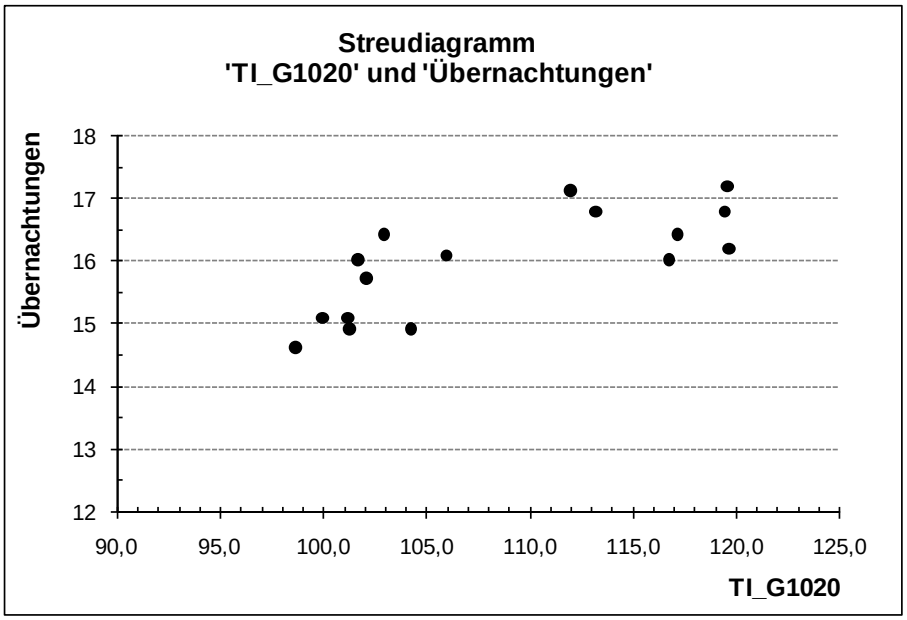
Die einfachste Form stellt ein linearer Zusammenhang dar, der durch die Regressionsgerade

$$\hat{Y} = a_0 + a_1 X$$

beschrieben wird. Ob die Annahme eines linearen Funktionstyps der Realität tatsächlich gerecht wird, lässt sich neben den oben bereits angesprochenen sachlogischen Überlegungen auch anhand des bekannten Streudiagramms beurteilen.

Schätzung der linearen Regressionsfunktion

Würden alle Wertepaare in einem Streudiagramm auf einer Geraden liegen, könnten die Koeffizienten der Regressionsgeraden direkt hieraus bestimmt werden.



12. Regressionsanalyse

Insbesondere ökonomische Variablen weichen von diesem Idealfall mehr oder weniger stark ab, was beispielhaft im Streudiagramm von 'Tourismusindex von Gemeinden zwischen 10000 und 20000 Einwohnern' und 'Übernachtungszahlen der Jugendherberge' sichtbar wird. Ein steigender und auch linearer Zusammenhang ist sicherlich zu erkennen, die eingezeichneten Wertepaare liegen jedoch nicht auf einer Gerade, sondern weisen zumindest 'leichte' Abweichungen hiervon auf.

In einer solchen Situation ist unter einer Vielzahl möglicher Geraden diejenige zu wählen, die die Punktewolke der Beobachtungswerte am besten repräsentiert, die also möglichst im Zentrum dieser Wertepaare liegt.

Definition: Residuum e_i

Die Entfernung eines Punktes der Schätzgeraden ($\hat{y}_i$; x_i) von dem beobachteten Wertepaar (y_i; x_i) wird durch den vertikalen Abstand e_i der Punkte y_i und $\hat{y}_i$ gemessen. Dieser Abstand e_i wird Restgröße oder auch Residuum genannt.

Zur Bestimmung der Regressionsgeraden wird dann auf die Methode der kleinsten Quadrate zurückgegriffen. Die Lage der Regressionsgeraden mit ihren Koeffizienten a_o und a_1 werden nach dieser Methode so bestimmt, dass die Summe der Abweichungsquadrate minimiert wird. Bei N Beobachtungspaaren resultiert hieraus die folgende formale Zielfunktion:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - a_0 - a_1 x_i)^2 = f(a_0, a_1) \rightarrow \text{Min.!}$$

Als notwendige Bedingung für die Existenz eines Extremwertes ist es erforderlich, die erste partielle Ableitung der Zielfunktion nach den Koeffizienten a_o und a_1 zu bilden und gleich Null zu setzen:

$$\frac{\delta f(a_0, a_1)}{\delta a_0} = \sum_{i=1}^{N} 2 (y_i - a_0 - a_1 x_i) (-1) = 0$$

$$\frac{\delta f(a_0, a_1)}{\delta a_1} = \sum_{i=1}^{N} 2 (y_i - a_0 - a_1 x_i) (-x_i) = 0$$

Aus den Normalgleichungen

$$\sum_{i=1}^{N} y_i = a_0 \cdot N + a_1 \cdot \sum_{i=1}^{N} x_i$$

$$\sum_{i=1}^{N} x_i \cdot y_i = a_0 \cdot \sum_{i=1}^{N} x_i + a_1 \cdot \sum_{i=1}^{N} x_i^2$$

lassen sich dann die Regressionskoeffizienten a_0 und a_1 berechnen.

12. Regressionsanalyse

Definition: Einfache lineare Regressionsfunktion

Aus den beiden Normalgleichungen lassen sich die Koeffizienten a_0 und a_1 einer einfachen linearen Regressionsfunktion ableiten. Die beiden Bestimmungsgleichungen lauten:

$$a_1 = \frac{N\sum x_i\, y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2} = \frac{Cov(X,Y)}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}$$

$$a_0 = \frac{\sum y_i}{N} - a_1 \cdot \frac{\sum x_i}{N} = \bar{y} - a_1 \cdot \bar{x}$$

Durch Umformung der Bestimmungsgleichung von a_0 zu

$$\bar{y} = a_0 + a_1 \cdot \bar{x}$$

wird außerdem deutlich, dass jede lineare Einfachregression mit der Steigung a_1 durch den Punkt der arithmetischen Mittelwerte der beiden Merkmale X und Y verläuft.

Der Regressionskoeffizient a_1 bringt die Steigung der Regressionsgerade zum Ausdruck. Er quantifiziert den Einfluss der Variablen X auf die abhängige Variable Y und gibt dabei an, um wie viele Einheiten sich der Wert des Merkmals Y durchschnittlich verändert, wenn der Wert des Merkmals X um eine Einheit steigt. Die Konstante a_0 erfasst demgegenüber den von X unabhängigen Einfluss auf Y, Y nimmt somit den Wert a_0 an, wenn X = 0. Dies entspricht dem Schnittpunkt der Regressionsgerade mit der y-Achse (Ordinate).

Beispiel:

Auf der Grundlage der in Tabelle 5 gegebenen Wertepaare lässt sich somit der im obigen Streudiagramm dargestellte Einfluss des Tourismusindex für Gemeinden zwischen 10000 und 20000 Einwohnern (X) auf die Übernachtungszahlen der Jugendherberge (Y) quantifizieren:

$$Cov(X,Y) = \frac{1}{16} \cdot 27770{,}59 - 108{,}481 \cdot 15{,}956 = 4{,}739\ [\% \cdot 1000\ \text{Gäste}]$$

$$s_x^2 = \frac{1}{16} \cdot 18925{,}13 - 108{,}481^2 = 60{,}068\ \%^2$$

$$a_1 = \frac{Cov(X,Y)}{s_x^2} = \frac{4{,}739\ \% \cdot 1000\,\text{ÜN}}{60{,}068\ \%^2} = 0{,}0789\ \frac{1000\,\text{ÜN}}{\%}$$

$$a_0 = \bar{y} - a_1 \cdot \bar{x} = 15{,}956 - 0{,}0789 \cdot 108{,}481 = 7{,}3968\ [1000\,\text{ÜN}]$$

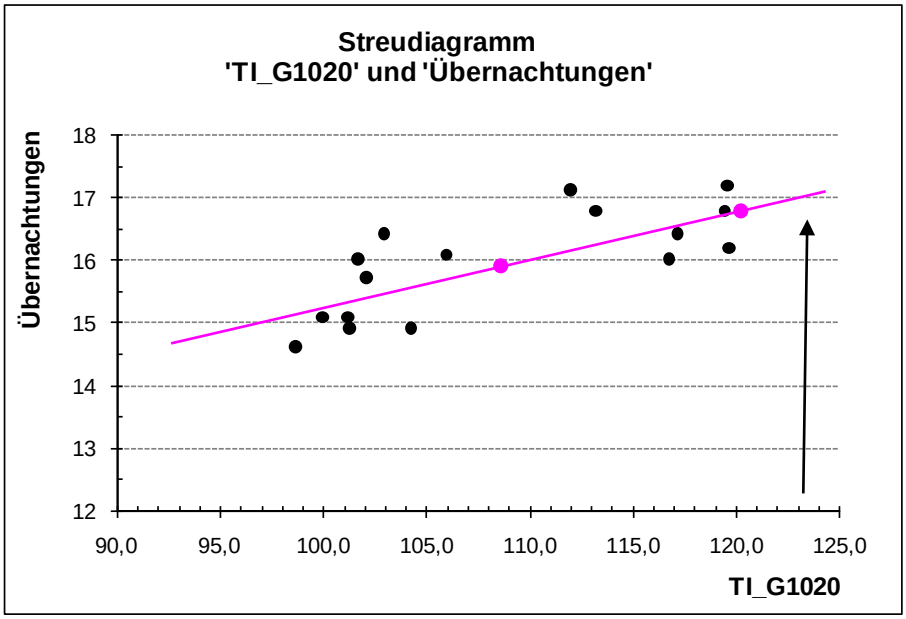
12. Regressionsanalyse

$$\hat{U}Z_JH = a_0 - a_1 \cdot TI_G1020 = 7{,}3968 + 0{,}0789 \cdot TI_G1020$$

Erhöht sich also der Tourismusindex der Gemeinden zwischen 10000 und 20000 Einwohnern um einen Prozentpunkt, dann steigen die Übernachtungszahlen der Jugendherberge im Durchschnitt um 78,9 pro Jahr (a_1).
Fällt dieser Index - aus welchen Gründen auch immer - auf ein Niveau von Null Prozentpunkten, dann könnte die Jugendherberge im Durchschnitt immer noch 7396,8 Übernachtungen pro Jahr verbuchen (a_0). Da der X-Wert von Null extrem weit außerhalb des Intervalls der X-Werte liegt, die zur Schätzung der Regressionskoeffizienten herangezogen wurden (zwischen 98,7 und 119,7; vgl. Tabelle 4), ist das Ergebnis jedoch mit allergrößter Vorsicht zu genießen.

Um den Verlauf der linearen Abhängigkeit der Y-Variablen von X grafisch darzustellen, kann die errechnete Regressionsfunktion direkt in das Streudiagramm eingezeichnet werden. Benötigt werden zwei Funktionswerte, die den Verlauf der Geraden dann eindeutig determinieren. In der obigen Definition wurde hergeleitet, dass jede lineare Einfachregression durch den Punkt der Mittelwerte der beiden Merkmale X und Y verläuft, im vorliegenden Beispiel also durch (108,481; 15,956). Auch der zweite Punkt ist im Normalfall gegeben, da die Regressionsfunktion die y-Achse im Punkt (0; a_0) schneidet. Diese Information kann allerdings nur dann genutzt werden, wenn das Streudiagramm im Ursprung des Koordinatensystems beginnt. Im vorliegenden Beispiel beginnt die x-Achse jedoch bei 90 Indexpunkten, die y-Achse bei 12000 Übernachtungen. Es muss folglich ein zweiter Punkt aus der Regressionsgleichung bestimmt werden:

z.B. $\hat{y}(120) = 7{,}3968 + 0{,}0789 \cdot 120 = 16{,}8648$ [1000 Ü.]



Aus der ermittelten Regressionsgerade lassen sich nun - zumindest im Umfeld der vorliegenden Wertepaare - eine Vielzahl von Prognosen für das Merkmal Y formulieren. Steigt bspw. der betrachtete Tourismusindex auf 123 Punkte, dann ist für die Jugendherberge mit durchschnittlich rund 17000 Übernachtungen in diesem Jahr zu rechnen. Der exakte Wert bestimmt sich wieder aus der Regressionsgleichung:

$\hat{y}(123) = 7{,}3968 + 0{,}0789 \cdot 123 = 17{,}1015$ [1000 Ü.]

12. Regressionsanalyse

Bewertung der Schätzqualität der linearen Regressionsfunktion

Bereits mehrfach wurde darauf hingewiesen, dass die Qualität einer Prognose umso geringer ist, je weiter der eingesetzte X-Wert vom Intervall der vorgegebenen Wertepaare entfernt liegt. Dies bedeutet im Umkehrschluss allerdings nicht, dass automatisch von einer hohen Prognosequalität ausgegangen werden kann, wenn der zugrundeliegende X-Wert in oder eng an diesem Intervall liegt. Vielmehr ist zu prüfen, in welchem Ausmaß die Schwankung der Y-Werte durch die Regressionsfunktion erklärt werden kann.

Grobe Anhaltspunkte liefert bereits der Blick auf das Streudiagramm mit der dort eingezeichneten Regressionsfunktion. Wird die Schwankung der Y-Werte durch die Regressionsfunktion sehr gut erklärt, so liegen die Wertepaare eng an dieser Funktion, im Extremfall sogar alle auf dieser Funktion. Umgekehrt streuen sie sehr stark um den Verlauf der Funktion.

Die Schwankungen der Y-Werte zeigen sich also je nach Qualität der Schätzfunktion in einer entsprechenden Schwankung der Schätzwerte oder in einer Schwankung der Residuen. Da die Schwankung von Merkmalsausprägungen bekanntlich über die Varianz quantifiziert werden kann, setzt sich folglich die Varianz der Y-Werte aus der Varianz der Schätzwerte und der Varianz der Residuen zusammen:

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

Die Varianz der Schätzwerte kann letztendlich - wie die nachfolgende Gleichung für den Fall einer einfachen linearen Regressionsfunktion zeigen soll - auf die Varianz des Merkmals X zurückgeführt werden:

$$s_{\hat{y}}^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{y} - \bar{y})^2 = \frac{1}{N}\sum_{i=1}^{N}(a_0 + a_1 x_i - a_0 - a_1\bar{x})^2 = a_1^2 \cdot \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = a_1^2 \cdot s_x^2$$

Aus der Relation dieser Varianzen zueinander lässt sich das Bestimmtheitsmaß als Kriterium für die Schätzqualität von Regressionsfunktionen ableiten. Da die Streuungszerlegung auch für nichtlineare Regressionsfunktionen gilt, kann das Bestimmtheitsmaß sowohl für lineare als auch an späterer Stelle für nichtlineare Funktionen verwendet werden.

Definition: Bestimmtheitsmaß B^2

Gegeben seien zwei metrisch skalierte Merkmale X und Y mit einer in den Regressionskoeffizienten linearen Regressionsfunktion $\hat{y} = f(x)$.

$$B^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

heißt dann Bestimmtheitsmaß (der Regressionsfunktion) und gibt den Anteil der durch X erklärten Varianz von Y an. B wird als Bestimmtheitskoeffizient bezeichnet.
B^2 gibt den Erklärungswert der Regressionsfunktion im Intervall von 0 bis 1 an; er bringt zum Ausdruck, wie gut die Variable Y durch die Regressionsfunktion beschrieben wird. Liegen alle Beobachtungswerte auf der Regressionsfunktion, so gilt $B^2 = 1$.

12. Regressionsanalyse

Für lineare Einfachregressionen lässt sich der oben festgestellte Zusammenhang nutzen, das Bestimmtheitsmaß vereinfacht sich dadurch zu:

$$B^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2} = \frac{a_1^2 \cdot s_x^2}{s_y^2}$$

Zudem stimmen bei den (nicht nur einfachen) linearen Regressionsfunktionen das Bestimmtheitsmaß und das Quadrat des Korrelationskoeffizienten überein.

Beispiel:

Um nun auf der Basis des Bestimmtheitsmaßes eine Bewertung der Schätzqualität für die Regressionsfunktion

$\hat{y}(x) = 7{,}3968 + 0{,}0789 \cdot x$

vornehmen zu können, muss gemäß obiger Formel noch die Varianz der Y-Werte berechnet werden:

$$s_y^2 = \frac{1}{16} \cdot 4083{,}99 - 15{,}956^2 = 0{,}655\,[1000\,\text{Ü.}]^2$$

Hieraus ergibt sich dann ein Bestimmtheitsmaß in Höhe von:

$$B^2 = \frac{0{,}0789^2 \cdot 60{,}068}{0{,}655} = 0{,}571$$

Das Bestimmtheitsmaß bringt in diesem Beispiel zum Ausdruck, dass 57,1% der Schwankung von Y durch die Regressionsfunktion und damit durch die Schwankung von X erklärt werden kann.

In der Literatur ist allerdings umstritten, welcher Mindestwert des Bestimmtheitsmaßes gegeben sein muss, um von einer ausreichenden Prognosequalität der Regressionsfunktion sprechen zu können. Verbreitet ist zumindest ein Grenzwert von 0,64. Dieser Vorgabe folgend müsste nach einem verbesserten Regressionsansatz gesucht werden.

Natürlich sollte auch der Wert des Bestimmtheitsmaßes vor einer eingehenden Bewertung auf Signifikanz überprüft werden. Analog zum bisherigen Vorgehen wird folglich eine Grundgesamtheit unterstellt, in der das Bestimmtheitsmaß exakt den Wert von Null annimmt. Mit dem ausgewählten Regressionsmodell lässt sich die Schwankung der Y-Variable also nicht erklären, da keine linearen Abhängigkeiten zwischen der Y-Variablen und der X-Variablen bzw. (bei multiplen Modellen) den X-Variablen vorliegen.

H_0: Das Bestimmtheitsmaß des ausgewählten Regressionsmodells mit p verschiedenen X-Variablen ist Null.

12. Regressionsanalyse

Bei Gültigkeit der Nullhypothese und bei normalverteilten Variablen X und Y folgt die Testgröße

$$T = \frac{n-p-1}{p} \cdot \frac{B^2}{1-B^2} \sim F_{df1;\, df2}$$

einer F - Verteilung mit df1 = p und df2 = n-p-1 Freiheitsgraden.

Liegt nun die über das Bestimmtheitsmaß der Stichprobe errechnete Testgröße T im kritischen Bereich, dann ist die Nullhypothese abzulehnen. Die Grenze des kritischen Bereichs errechnet sich aus dem (über das Signifikanzniveau) festgelegte Quantil der F - Verteilung:

$KB_{\alpha} = [F_{df1;\, df2;\, 1-\alpha} ; +\infty)$

Beispiel:

Für die einfache lineare Abhängigkeit der Übernachtungszahlen (ÜZ_JH) vom Tourismusindex (TI_G1020) wurde aus einer Stichprobe mit 16 Wertepaaren die Regressionsfunktion

$\hat{y}(x) = 7{,}3968 + 0{,}0789 \cdot x$

mit einem Bestimmtheitsmaß $B^2 = 0{,}571$ errechnet.

Die Nullhypothese

H_0: Das Bestimmtheitsmaß in der Grundgesamtheit ist Null

kann nur dann abgelehnt werden, wenn die Testgröße

$$T = \frac{16-1-1}{1} \cdot \frac{0{,}571}{1-0{,}571} = 18{,}634$$

im kritischen Bereich liegt. Da bekannt ist, dass diese Testgröße grundsätzlich einer F - Verteilung mit

df1 = p = 1 und df2 = n-p-1 = 16-1-1 = 14 Freiheitsgraden

folgt, lassen sich bei vorgegebener Irrtumswahrscheinlichkeit (Signifikanzniveau) α in Höhe von 0,05 die Grenzen des kritischen Bereichs über die F - Verteilung bestimmen.

$KB_{0{,}05} = [F_{1;\, 14;\, 0{,}95} ; +\infty) = [4{,}600; +\infty)$

Tabelle 6 liefert für das 0,95-Quantil bei 1 und 14 Freiheitsgraden einen Wert von 4,600 (vgl. auch Wert der Excel-Funktion 'FINV(0,05; 1; 14)'). Da die Testgröße mit 18,634 ganz eindeutig in diesem kritischen Bereich liegt, kann die Nullhypothese abgelehnt werden. Es ist also von einem signifikanten Regressionsansatz auszugehen, angesichts des zu niedrigen Bestimmtheitsmaßes muss dieser jedoch nachjustiert werden.

12. Regressionsanalyse

Tabelle 6

0,95 - Quantile der F - Verteilung

	df1							
df2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768	238,883
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266
31	4,160	3,305	2,911	2,679	2,523	2,409	2,323	2,255
32	4,149	3,295	2,901	2,668	2,512	2,399	2,313	2,244
33	4,139	3,285	2,892	2,659	2,503	2,389	2,303	2,235
34	4,130	3,276	2,883	2,650	2,494	2,380	2,294	2,225
35	4,121	3,267	2,874	2,641	2,485	2,372	2,285	2,217
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180
50	4,034	3,183	2,790	2,557	2,400	2,286	2,199	2,130

Lesebeispiel: df1 = 1 und df2 = 14 $\rightarrow F_{1;\,14;\,0,95} = 4,600$

$\rightarrow f(F \leq 4,600) = F(4,600) = 0,95$

12.2. Einfache nichtlineare Regressionsrechnung

Bereits in Kapitel 9 wurde darauf hingewiesen, dass Zusammenhänge zwischen zwei Merkmalen auch nichtlineare Strukturen aufweisen können. Solche nichtlinearen Strukturen lassen sich dann häufig mit Hilfe quadratischer Funktionen, Potenzfunktionen, Exponentialfunktionen oder Hyperbelfunktionen beschreiben.

Problematisch ist allerdings die Schätzung der Koeffizienten dieser Funktionstypen, da die unter 12.1. vorgestellte Methode der Kleinsten Quadrate nur bei linearen Funktionen optimale Ergebnisse liefert. Bevor also die Regressionskoeffizienten aus den Normalgleichungen errechnet werden können, müssen nichtlineare Funktionen durch geeignete Transformationen linearisiert werden. Für das linearisierte Problem wird dann eine Optimallösung erreicht (Minimierung der Residuenquadrate), allerdings sollte nicht verschwiegen werden, dass für das ursprüngliche Problem unter Umständen nur eine Näherungslösung erzielt wird.

Dieser Transformationsprozess soll nun für die oben genannten Funktionstypen kurz vorgestellt werden.

Quadratische Funktion: $\hat{y} = a_0 + a_1 \cdot x + a_2 \cdot x^2$

Die Schätzung der Regressionskoeffizienten einer quadratischen Funktion gestaltet sich eigentlich unproblematisch. Es handelt sich zwar um eine nichtlineare Funktion, dennoch sind ihre Koeffizienten a_0, a_1 und a_2 linear miteinander verknüpft. Die Schätzmethode der Kleinsten Quadrate ist folglich ohne Umwege anwendbar. Aus der partiellen Ableitung der Funktion der Abweichungsquadrate nach diesen drei Koeffizienten ergeben sich die drei Normalgleichungen,

$$\Sigma y_i = a_0 N + a_1 \Sigma x_i + a_2 \Sigma x_i^2$$

$$\Sigma x_i y_i = a_0 \Sigma x_i + a_1 \Sigma x_i^2 + a_2 \Sigma x_i^3$$

$$\Sigma x_i^2 y_i = a_0 \Sigma x_i^2 + a_1 \Sigma x_i^3 + a_2 \Sigma x_i^4$$

aus denen dann die Werte der Koeffizienten bestimmt werden können.

Potenzfunktion: $\hat{y} = a_0 \cdot x^{a_1}$

Aus der Potenzfunktion $a_0 \cdot x^{a_1}$ ergibt sich durch Logarithmieren die linearisierte Regressionsgleichung

$$\ln \hat{y} = \ln a_0 + a_1 \cdot \ln x \, .$$

Die beiden Regressionskoeffizienten $\ln(a_0)$ und a_1 können dann entweder aus den beiden Normalgleichungen

12. Regressionsanalyse

$$\sum_{i=1}^{N} \ln y_i = \ln a_0 \cdot N + a_1 \cdot \sum_{i=1}^{N} \ln x_i$$

$$\sum_{i=1}^{N} \ln x_i \cdot \ln y_i = \ln a_0 \cdot \sum_{i=1}^{N} \ln x_i + a_1 \cdot \sum_{i=1}^{N} (\ln x_i)^2$$

oder aber direkt aus den Bestimmungsgleichungen

$$a_1 = \frac{N \sum \ln x_i \ln y_i - \sum \ln x_i \sum \ln y_i}{N \sum (\ln x_i)^2 - (\sum \ln x_i)^2}$$

$$\ln a_0 = \frac{\sum \ln y_i}{N} - a_1 \cdot \frac{\sum \ln x_i}{N}$$

errechnet werden.
Abschließend wird noch a_0 durch

$$a_0 = e^{\ln a_0}$$

'rücktransformiert'.

Exponentialfunktion: $\hat{y} = a_0 \cdot a_1^x$ bzw. $\hat{y} = a_0 \cdot e^{a_1 \cdot x}$

Die Vorgehensweise ist analog zu der bei Potenzfunktionen. Durch Logarithmieren wird der ursprüngliche Exponentialansatz in eine linearisierte Form gebracht und die Regressionskoeffizienten nach Anwendung der Methode der Kleinsten Quadrate aus den Normalgleichungen geschätzt.

Für $\hat{y} = a_0 \cdot a_1^x$

$$\ln \hat{y} = \ln a_0 + x \cdot \ln a_1$$

$$\ln a_1 = \frac{N \sum x_i \ln y_i - \sum x_i \sum \ln y_i}{N \sum (x_i)^2 - (\sum x_i)^2}$$

$$\ln a_0 = \frac{\sum \ln y_i}{N} - a_1 \cdot \frac{\sum x_i}{N}$$

$$a_1 = e^{\ln a_1}$$

$$a_0 = e^{\ln a_0}$$

12. Regressionsanalyse

Für $\hat{y} = a_0 \cdot e^{a \cdot x}$

$$\ln \hat{y} = \ln a_0 + a_1 \cdot x$$

$$a_1 = \frac{N \sum x_i \ln y_i - \sum x_i \sum \ln y_i}{N \sum (x_i)^2 - (\sum x_i)^2}$$

$$\ln a_0 = \frac{\sum \ln y_i}{N} - a_1 \cdot \frac{\sum x_i}{N}$$

$$a_0 = e^{\ln a_0}$$

Hyperbelfunktion: $\hat{y} = a_0 + a_1 / x$

Im Vergleich zu den beiden Vorgängerfunktionen weist die Hyperbelfunktion bereits linearisierte Koeffizienten auf. Eine Transformation ist daher nicht notwendig, die Bestimmungsgleichungen lassen sich also direkt aus der Ursprungsfunktion ableiten:

$$a_1 = \frac{N \sum \frac{1}{x_i} y_i - \sum \frac{1}{x_i} \sum y_i}{N \sum (\frac{1}{x_i})^2 - (\sum \frac{1}{x_i})^2}$$

$$a_0 = \frac{\sum y_i}{N} - a_1 \cdot \frac{\sum \frac{1}{x_i}}{N}$$

12.3. Multiple lineare Regressionsrechnung

Im Normalfall beschreiben monokausale Erklärungsmodelle die Realität nur unzureichend, so dass es Sinn macht, die bisherigen Regressionsfunktionen zu erweitern. Hierbei wird dann der Einfluss einer Vielzahl von Variablen X_1 bis X_P auf das Merkmal Y unterstellt. In der einfachsten Form lässt sich diese Abhängigkeit wieder über eine lineare Regressionsgleichung (Lineare Mehrfachregression)

$$\hat{y} = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + .. + a_P \cdot x_P$$

beschreiben.

Schätzung der linearen Regressionsfunktion

Die Schätzung der Regressionskoeffizienten erfolgt analog zur bisherigen Vorgehensweise unter Anwendung der Kleinst-Quadrate-Methode. Allerdings erhöht sich mit jeder weiteren X-Variablen auch die Anzahl der Normalgleichungen, die sich aus der partiellen Ableitung der Zielfunktion nach den p+1 - Regressionskoeffizienten ergeben.

Um die Bestimmungsgleichungen der Regressionskoeffizienten bei multiplen Modellen dennoch in übersichtlicher Form darstellen zu können, soll der Lösungsalgorithmus mit Hilfe von Vektoren und Matrizen vorgestellt werden.

Die Zielfunktion f

$$f(a_0, a_1, ..., a_p) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - a_0 - a_1 x_{1i} - ... - a_p x_{pi})^2 \rightarrow \text{Min.!}$$

kann dann für alle statistischen Einheiten (v.a. Zeitpunkte) von 1 bis N in ausführlicher Weise

$$e_1^2 = (y_1 - a_0 \cdot 1 - a_1 \cdot x_{11} - a_2 \cdot x_{21} - ... - a_p \cdot x_{p1})^2 +$$

$$e_2^2 = (y_2 - a_0 \cdot 1 - a_1 \cdot x_{12} - a_2 \cdot x_{22} - ... - a_p \cdot x_{p2})^2 +$$

$$e_3^2 = (y_3 - a_0 \cdot 1 - a_1 \cdot x_{13} - a_2 \cdot x_{23} - ... - a_p \cdot x_{p3})^2 +$$

...

$$e_N^2 = (y_N - a_0 \cdot 1 - a_1 \cdot x_{1N} - a_2 \cdot x_{2N} - ... - a_p \cdot x_{pN})^2 \rightarrow \text{Min!}$$

dargestellt und in Vektoren- bzw. Matrizen-Schreibweise

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ ... \\ e_N \end{bmatrix}^2 = \left[\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ y_N \end{pmatrix} - \begin{pmatrix} 1 & x_{11} & x_{21} & ... & x_{p1} \\ 1 & x_{12} & x_{22} & ... & x_{p2} \\ 1 & x_{13} & x_{23} & ... & x_{p3} \\ ... & ... & ... & ... & ... \\ 1 & x_{1N} & x_{2N} & ... & x_{pN} \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ ... \\ a_p \end{pmatrix} \right]^2 = (y - X \cdot a)^2 \rightarrow \text{MIN.!}$$

überführt werden.

12. Regressionsanalyse

Analog zum Procedere in Kapitel 12.1. werden nun die partiellen Ableitungen der Zielfunktion nach den Regressionskoeffizienten a0, a1, a2, ..., ap bestimmt

$$\delta(\mathbf{y} - \mathbf{Xa})^2 / \delta\mathbf{a} = 2\,(\mathbf{y} - \mathbf{Xa})\,(-\mathbf{X})$$

und diese (als notwendige Voraussetzung für die Existenz eines Minimums) gleich Null gesetzt:

$$2\,(\mathbf{y} - \mathbf{Xa})\,(-\mathbf{X}) = 2\,(-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{Xa}) = 0$$

Aus der Bedingung

$$\mathbf{X}'\mathbf{Xa} = \mathbf{X}'\mathbf{y}$$

resultiert dann der Vektor der geschätzten Regressionskoeffizienten

$$\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}$$

des multiplen Regressionsmodells.

Hierzu ist also in einem ersten Schritt die Matrix der X-Werte zu transponieren (X'), diese mit der ursprünglichen X-Matrix zu multiplizieren (X'X) und von diesem Produkt die Inverse zu bilden. Die neue Matrix $(X'X)^{-1}$ wird dann in einem zweiten Schritt mit dem Produkt von transponierter X-Matrix und y-Vektor (X'y) multipliziert. Das Ergebnis ist dann der gesuchte Koeffizientenvektor des multiplen Regressionsmodells.

Beispiel:

In Kapitel 12.1. wurde die lineare Abhängigkeit der jährlichen Übernachtungszahlen vom Tourismusindex für Gemeinden zwischen 10000 und 20000 Einwohnern über eine einfache lineare Regressionsgleichung geschätzt, deren Bestimmtheitsmaß mit 0,571 nicht unbedingt zufriedenstellend ausfiel. Es liegt daher nahe, dieses Regressionsmodell hier wieder aufzugreifen und um weitere Einflussgrößen zu ergänzen.
Grundsätzlich bieten sich hierfür alle anderen Variablen aus Tabelle 5 an, also der Tourismusindex für Gesamtdeutschland, der Tourismusindex für alle Jugendherbergen in Deutschland zusammen, aber auch die Marketingausgaben der Jugendherberge selbst. Insbesondere die letzte Größe dürfte sich auf das Bestimmtheitsmaß positiv auswirken, da sie andere Informationen erfasst als die beiden Tourismusindizes, die vermutlich dem Gemeinde-Index sehr ähnlich sind.

Das einfache lineare Regressionsmodell soll daher um die Marketingausgaben als zweite Einflussgröße erweitert und die oben vorgestellten Matrixberechnungen an diesem Beispiel konkretisiert werden.

$$\hat{y} = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2$$

mit x_1 = TI_G1020

x_2 = MA_JH

12. Regressionsanalyse

Koeffizientenmatrix X und Vektor der y-Werte

X =				y =	
	1	101,2	9		15,1
	1	101,3	10		14,9
	1	98,7	12		14,6
	1	100,0	8		15,1
	1	102,1	8		15,7
	1	101,7	8		16,0
	1	102,9	6		16,4
	1	104,2	12		14,9
	1	105,9	8		16,1
	1	113,2	6		16,8
	1	112,0	6		17,1
	1	116,7	8		16,0
	1	119,7	12		16,2
	1	117,1	9		16,4
	1	119,4	8		16,8
	1	119,6	6		17,2

Transponierte Koeffizientenmatrix X'

X' =				…			
	1	1	1	…	1	1	1
	101,2	101,3	98,7	…	117,1	119,4	119,6
	9	10	12	…	9	8	6

Produktmatrix (X'X)

X'X =			
	16,00	1735,70	136,00
	1735,70	189251,13	14701,50
	136,00	14701,50	1222,00

Inverse Matrix $(X'X)^{-1}$

$(X'X)^{-1}$ =			
	15,58563	-0,12528	-0,22740
	-0,12528	0,00109	0,00086
	-0,22740	0,00086	0,01583

Vektor (X'y)

X'y =	
	255,300
	27770,590
	2151,500

Vektor der Regressionskoeffizienten $a = (X'X)^{-1} (X'y)$

a =	
	10,73769
	0,06605
	-0,22907

12. Regressionsanalyse

Die Regressionsgleichung erhält damit folgendes Aussehen:

$\hat{y} = 10{,}73769 + 0{,}06605 \cdot x_1 - 0{,}22907 \cdot x_2$

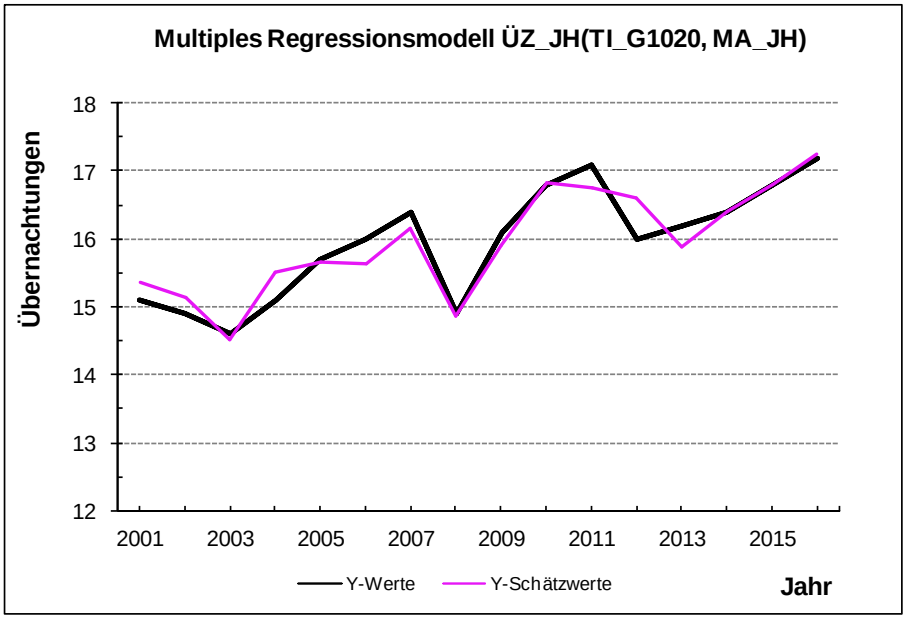
bzw.

Schätzwert ÜZ_JH = 10,73769 + 0,06605 · TI_G1020 - 0,22907 · MA_JH

Auf der Grundlage dieses Regressionsmodells können nun für die betrachteten 16 statistischen Einheiten (= Jahre) die Schätzwerte für die jährlichen Übernachtungszahlen der Jugendherberge berechnet

Jahr i	ÜZ_JH y_i	TI_G1020 x_{1i}	MA_JH x_{2i}	$\hat{y}_i = 10{,}73769 + 0{,}06605 \cdot x_{1i} - 0{,}22907 \cdot x_{2i}$
2001	15,1	101,2	9	15,3603
2002	14,9	101,3	10	15,1379
2003	14,6	98,7	12	14,5080
2004	15,1	100,0	8	15,5101
2005	15,7	102,1	8	15,6488
2006	16,0	101,7	8	15,6224
2007	16,4	102,9	6	16,1598
2008	14,9	104,2	12	14,8713
2009	16,1	105,9	8	15,8998
2010	16,8	113,2	6	16,8401
2011	17,1	112,0	6	16,7609
2012	16,0	116,7	8	16,6132
2013	16,2	119,7	12	15,8950
2014	16,4	117,1	9	16,4105
2015	16,8	119,4	8	16,7915
2016	17,2	119,6	6	17,2629

und natürlich auch grafisch den tatsächlichen Y-Werten gegenüber gestellt werden:



12. Regressionsanalyse

Bewertung der Schätzqualität der multiplen Regressionsfunktion

Dass die Schätzqualität des multiplen Regressionsmodell gegenüber der einfachen linearen Regressionsgleichung deutlich besser ausfällt ist an der Grafik unschwer zu erkennen. Die Schwankungen der Y-Werte können zu einem großen Teil durch die Schwankungen der beiden X-Variablen (TI_G1020 und MA_JH) erklärt werden.

Um zu einer konkreteren Aussage über die Prognosequalität zu gelangen, kann natürlich auch für multiple Modelle das (multiple) Bestimmtheitsmaß berechnet werden:

$$B^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Allerdings ist hier eine vereinfachte Vorgehensweise nicht möglich. Das multiple Bestimmtheitsmaß ergibt sich also nach einer etwas aufwändigen Berechnung der Varianz der Schätzwerte und der Varianz der tatsächlichen Y-Werte aus dem Quotienten dieser beiden Größen. Die Wurzel aus dem Bestimmtheitsmaß entspricht dann dem multiplen Korrelationskoeffizient zwischen der Y-Variable und den X-Variablen.

Beispiel:

Jahr i	ÜZ_JH y_i	TI_G1020 x_{1i}	MA_JH x_{2i}	$\hat{y}_i = 10{,}73769 + 0{,}06605 \cdot x_{1i} - 0{,}22907 \cdot x_{2i}$	$\hat{y}_i^2$	y_i^2
2001	15,1	101,2	9	15,3603	235,939	228,010
2002	14,9	101,3	10	15,1379	229,155	222,010
2003	14,6	98,7	12	14,5080	210,482	213,160
2004	15,1	100,0	8	15,5101	240,564	228,010
2005	15,7	102,1	8	15,6488	244,886	246,490
2006	16,0	101,7	8	15,6224	244,060	256,000
2007	16,4	102,9	6	16,1598	261,140	268,960
2008	14,9	104,2	12	14,8713	221,154	222,010
2009	16,1	105,9	8	15,8998	252,804	259,210
2010	16,8	113,2	6	16,8401	283,590	282,240
2011	17,1	112,0	6	16,7609	280,927	292,410
2012	16,0	116,7	8	16,6132	275,997	256,000
2013	16,2	119,7	12	15,8950	252,652	262,440
2014	16,4	117,1	9	16,4105	269,305	268,960
2015	16,8	119,4	8	16,7915	281,954	282,240
2016	17,2	119,6	6	17,2629	298,006	295,840
Summe	255,30			255,293	4082,616	4083,990
Mittelwert	15,956			15,956		

Bei einer Varianz der Schätzwerte in Höhe von 0,5696 Ü^2 und einer Varianz der tatsächlichen Y-Werte mit 0,6554 Ü^2 errechnet sich ein Bestimmtheitsmaß von 0,869. Es weist damit auf eine deutliche Verbesserung gegenüber dem einfachen Modellansatz und gleichzeitig auf eine ausreichend gute Prognosequalität des multiplen Regressionsmodells hin.

12. Regressionsanalyse

Mit dem Ziel, die Anpassungsgüte eines Regressionsmodells zu maximieren, könnte versucht werden, möglichst viele erklärende X-Variablen in ein Modell aufzunehmen. Das Bestimmtheitsmaß würde im schlechtesten Fall - wenn die neue X-Variable keine neuen Informationen liefert - konstant bleiben, ansonsten vielleicht auch nur minimal ansteigen.
Werden eine Vielzahl solcher Variablen mit einem geringen Erklärungsnutzen aufgenommen, so besteht die Gefahr des sog. 'Overfitting'. Die Modelle werden instabil und neigen zu Überreaktionen.

Das Ziel einer Modellspezifikation sollte es daher sein, die Schwankung der Y-Variablen mit möglichst wenigen, dafür aber entscheidenden Einflussvariablen erklären zu können. Diesem Ziel dient das korrigierte Bestimmtheitsmaß:

$$B^2_{korr} = 1 - (1 - B^2) \cdot \frac{n-1}{n-p-1}$$

Das korrigierte Bestimmtheitsmaß steigt also nur, wenn der Zuwachs des 'normalen' Bestimmtheitsmaßes durch die Hinzunahme einer weiteren X-Variablen größer ist als der hierdurch verursachte Rückgang des Quotienten (n-1)/(n-p-1).
Das korrigierte Bestimmtheitsmaß kann also durch die Aufnahme zusätzlicher Erklärungsvariablen geringer werden, im Extremfall sogar negative Werte annehmen. Auf alle Fälle sollten Regressionsmodelle nur erweitert werden, wenn das korrigierte Bestimmtheitsmaß gesteigert werden kann.

Beispiel:

Da der Anstieg des Bestimmtheitsmaßes von 0,571 auf 0,869 durch die Aufnahme einer zweiten X-Variablen (MA_JU) recht deutlich ausfiel, ist es unwahrscheinlich, dass das korrigierte Bestimmtheitsmaß eine gegenläufige Entwicklung zeigt. Die Berechnungen für die beiden Modelle bestätigen diese Vermutung.

$\hat{y}$(T_G1020) = 7,3968 + 0,0789 · TI_G1020

$B^2 = 0{,}571$

$B^2_{korr} = 1 - (1-0{,}571) \cdot (16-1) / (16-1-1) = 0{,}540$

$\hat{y}$(TI_G1020; MA_JH) = 10,73769 + 0,06605 · TI_G1020 - 0,22907 · MA_JH

$B^2 = 0{,}869$

$B^2_{korr} = 1 - (1-0{,}869) \cdot (16-1) / (16-2-1) = 0{,}849$

Das korrigierte Bestimmtheitsmaß steigt von 0,540 auf 0,849 und unterstreicht damit, dass es sinnvoll war, das einfache lineare Regressionsmodell um eine Variable zu erweitern. Denkbar wäre es auch, noch eine weitere oder sogar die beiden anderen Variablen in das Regressionsmodell aufzunehmen. Diese Überlegungen sollen an dieser Stelle jedoch nicht mehr weiter verfolgt werden.

13. Zeitreihenanalyse

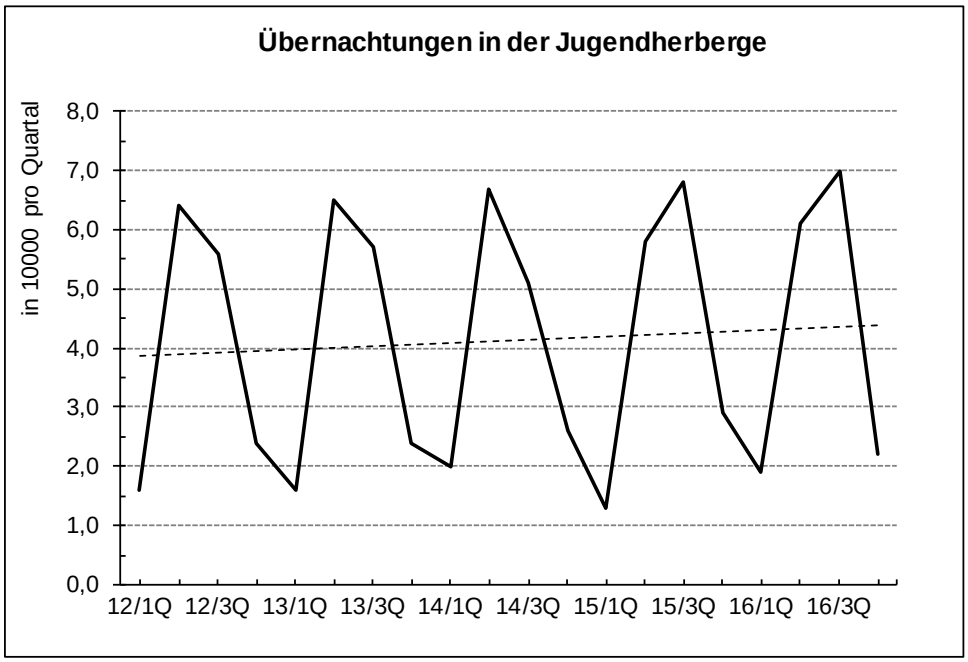
Ist im Gegensatz zur Regressionsrechnung die Gesamtheit der Einflussgrößen oder die Art ihres Einflusses nicht bekannt oder die mathematische Beschreibung nur schwer möglich, wird nur das zeitliche Verhalten der beobachteten Größe selbst untersucht. Die Zeit tritt als Ursache an die Stelle der Gesamtheit der tatsächlichen Einflussfaktoren.

Um das Verhalten eines Merkmals im Zeitablauf zu analysieren, werden die Merkmalsausprägungen für aufeinander folgende Zeitpunkte oder Zeitintervalle erhoben. Eine solche Reihe von Beobachtungswerten heißt Zeitreihe. Für die Zeitreihenanalyse lassen sich dann folgende Ziele und Aufgaben ableiten:

- Identifikation der Gesetzmäßigkeiten einer Zeitreihe
- Vergleich verschiedener Zeitreihen
- Schätzung von unbekannten Zeitreihenwerten innerhalb des Beobachtungsintervalls (**Interpolation**)
- Schätzung von Zeitreihenwerten außerhalb des Beobachtungsintervalls (**Extrapolation, Prognose**)

13.1. Komponenten einer Zeitreihe und ihre Modellierung

Mögliche Einflussfaktoren bzw. Komponenten einer Zeitreihe lassen sich meist schon mit Blick auf ihren grafischen Verlauf erkennen. Hierzu sind für die bereits bekannte Jugendherberge in der nachfolgenden Abbildung deren Quartalszahlen an Übernachtungen seit 2012 dargestellt.



Die gestrichelte Linie verdeutlicht, dass die durchschnittlichen Übernachtungszahlen in dem betrachteten Zeitraum ab März 2012 leicht angestiegen sind, allerdings sind in den jeweiligen Quartalen auch deutliche Abweichungen zu erkennen. So liegen die Höchstwerte innerhalb eines Jahres im zweiten oder dritten Quartal und die Tiefstpunkte immer im ersten Quartal.

Aus diesen ersten Überlegungen lassen sich bereits die zentralen Bewegungskomponenten einer Zeitreihe ableiten. Dabei werden unterschieden:

- eine **Trend-Komponente** m_T,
  zur Erfassung einer langfristigen Grundrichtung der zeitlichen Entwicklung
- eine **zyklische Komponente** z_T,
  zur Erfassung von mittelfristigen, sich periodisch wiederholenden Einflüssen, insbesondere konjunkturelle Schwankungen, die mit einer mehrjährigen, aber nicht völlig konstant bleibenden Periode den Trend überlagern
- eine **saisonale Komponente** s_T,
  zur Erfassung jahreszeitlicher Ereignisse, die mit einer konstanten jährlichen Periode auftreten und von dem Trend und der zyklischen Komponente überlagert sind
- eine **zufällige Restkomponente** r_T,
  zur Erfassung aller Schwankungen der beobachteten Zeitreihe, die nicht durch die drei systematischen Komponenten abgedeckt und gedeutet worden sind; bei diesen zufälligen Schwankungen wird davon ausgegangen, dass ihr Mittelwert Null ist

Da eine klare Trennung von Trend und zyklischer Komponente oftmals nicht exakt möglich ist, werden beide meistens zur **glatte**n **Komponente** g_T zusammengefasst.

Die Zeitreihenwerte lassen sich dann entweder über eine additive Verknüpfung der Bewegungskomponenten

$$\mathbf{y_T = g_T + s_T + r_T \text{ (= additives Modell)}}$$

oder über eine multiplikative Verknüpfung

$$\mathbf{y_T = g_T \cdot s_T \cdot r_T \text{ (= multiplikatives Modell)}}$$

beschreiben. Das multiplikative Modell eignet sich vor allem zur Analyse von Zeitreihen mit einer im Zeitablauf kontinuierlich steigenden oder fallenden Amplitude. In allen anderen Fällen ist eine additive Verknüpfung der Komponenten vorzuziehen.

Während im obigen Schaubild die Amplitude in den ersten drei Jahren annähernd konstant bleibt, nimmt sie im vierten Jahr, also 2015, zu, um dann 2016 wieder auf das vorherige Niveau zurück zu fallen. In dem Beispiel zu den Übernachtungszahlen ist folglich von einem additiven Modell auszugehen.

Gerade für ökonomische Zeitreihen ist dieser Verlauf nicht untypisch. In der Regel zeigt sich also eine relativ konstante Schwingungsbreite zwischen den Jahres-Hoch- und Tiefpunkten der Zeitreihe, die dann nur temporär von steigenden oder fallenden Amplituden abgelöst werden. Kontinuierlich fallende oder steigende Amplituden, die für ein Zeitreihenmodell mit multiplikativ verknüpften Komponenten sprechend würden, sind dagegen äußerst selten. Es soll daher im Folgenden nur auf additive Modelle näher eingegangen werden.

13.2. Methoden zur Trendermittlung

Die Trendkomponente gibt die Grundrichtung einer Zeitreihe vor und hat damit zentrale Bedeutung. Dazu wird nachfolgend sowohl ein funktionsspezifischer Ansatz nach der aus Kapitel 12 bekannten Methode der kleinsten Quadrate als auch die Methode gleitender Durchschnitte beschrieben.

Methode gleitender Durchschnitte

Um die glatte Komponente einer Zeitreihe herauszufiltern, müssen die saisonalen und die irregulären Schwankungen ausgeschaltet werden. Saisonschwankungen haben die konstante Periode eines Jahres und gleichen sich in ihrer Wirkung innerhalb eines Jahres gegenseitig aus. Folglich lassen sich die Saisoneinflüsse dadurch neutralisieren, dass aus allen Zeitreihenwerten einer Periode ein arithmetischer Durchschnitt berechnet wird. Gleichzeitig wird hierdurch auch die Restkomponente eliminiert. Mit der fortlaufenden Berechnung von Durchschnitten lassen sich also 'zwei Fliegen mit einer Klappe schlagen', durch das Ausschalten von Saison- und Restkomponente entspricht der Durchschnittswert dem Trendwert.

Definition: Gleitende Durchschnitte D

Gleitende Durchschnitte D der Ordnung k werden als arithmetisches Mittel fortlaufend aus k benachbarten Zeitreihenwerten berechnet und dem mittleren der bei der Durchschnittsbildung berücksichtigten Zeitpunkte T zugeordnet.
Dabei muss zwischen Durchschnitten ungerader Ordnung mit k = 3, 5, 7, …

$$D_T(k) = \frac{1}{k} \cdot \sum_{h=-i}^{i} y_{T+h} \quad \text{mit} \quad i = \frac{k-1}{2} \quad \text{und} \quad T = i+1, i+2, ..., n-i$$

und Durchschnitten gerader Ordnung mit k = 2, 4, 6, …

$$D_T(k) = \frac{1}{k} \cdot \left(\frac{y_{T-i}}{2} + \sum_{h=-i+1}^{i-1} y_{T+h} + \frac{y_{T+i}}{2} \right) \quad \text{mit} \quad i = \frac{k}{2} \quad \text{und} \quad T = i+1, i+2, ..., n-i$$

unterschieden werden.

Beispiel:

Bei den hier vorliegenden Zeitpunkten im Abstand von drei Monaten umfasst eine Jahres-Periode vier Zeitreihenwerte. Die Zeitreihenwerte einer Periode werden als Phasen bezeichnet und entsprechen in unserem Beispiel der Übernachtungszahlen im ersten, zweiten, dritten und vierten Quartal.
Für die ersten vier Zeitreihenwerte des Jahres 2012 errechnet sich dann aus den Quartalswerten 1,6 / 6,4 / 5,6 und 2,4 ein Mittelwert von 4,0 [1000 Übernachtungen].

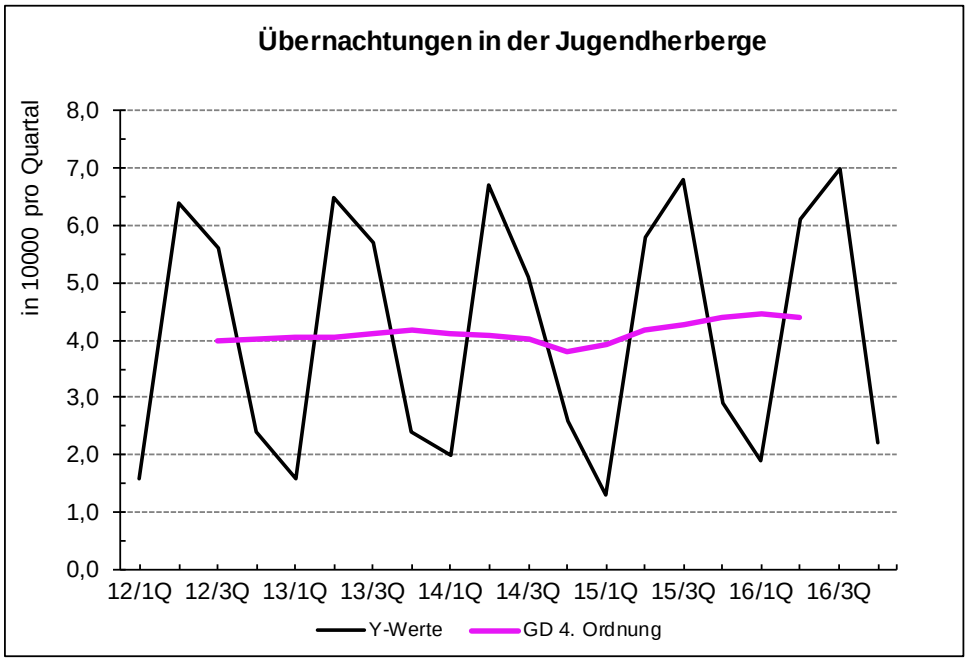
Problematisch ist allerdings - wie bei allen geradzahligen Durchschnittsberechnungen - die Zuordnung dieses Wertes, da der mittlere Zeitpunkt des betrachteten Zeitraums Ende Juni 2012 zwischen dem zweiten und dritten Quartal liegt und damit keinem Zeitpunkt der vorliegenden Zeitreihe entspricht.

Eine Zeitreihen adäquate Zuordnung des Mittelwertes ist also nur möglich, wenn dieser aus einer ungeraden Anzahl von Zeitreihenwerten errechnet wird. Um dabei die Saisoneinflüsse der verschiedenen Phasen nicht zu verzerren, wird - wie aus der obigen Definition ersichtlich - bei einem geradzahligen Durchschnitt zwar ein weiterer Zeitreihenwert herangezogen, der erste und der letzte Wert gehen jedoch nur mit halbem Gewicht in die Berechnung ein, da sie ja die gleiche Phase betreffen:

$$D_{2012/3.Q.}(4) = \frac{1{,}6/2 + 6{,}4 + 5{,}6 + 2{,}4 + 1{,}6/2}{4} = 4{,}000$$

Die ursprüngliche Durchschnittsrechnung vierter Ordnung D(4) wird folglich auf das erste Quartal 2013 ausgedehnt, der mittlere Zeitpunkt entspricht dann dem dritten Quartal 2012. Dabei wird der Einfluss des ersten Quartals jeweils zur Hälfte über den des Jahres 2012 und den des Jahres 2013 erfasst und somit vermieden, dass dieser Effekt gegenüber den anderen drei Quartalen doppelt gewichtet wird.

Die nachfolgende Abbildung stellt dem Verlauf der Aufenthaltsdauer die auf diese Weise berechneten gleitenden Durchschnitte vierter Ordnung gegenüber. Ausgehend von dem oben berechneten Trendwert mit 4,0 im 3. Quartal 2012 haben sich die Durchschnittswerte bis zum zweiten Quartal 2016 auf knapp 4,4 (4400 Übernachtungen) erhöht. Gut sichtbar ist dabei der differenzierte Trendverlauf. Während die Trendwerte in den ersten drei Jahren nahezu konstant blieben, konnten sie erst in den beiden letzten Jahren leicht zulegen.



Die Gleitenden Durchschnitte konnten in den bisherigen Beispielrechnungen als äußerst flexibles Verfahren zur Trendermittlung vorgestellt werden. Sie setzen keinen bestimmten Funktionstyp voraus und signalisieren sofort eine Änderung im Trendverlauf, selbst wenn diese nur gering ausfallen sollte. Durch die Variation der Ordnung, die allerdings immer ein Vielfaches der Phasenabschnitte ausmachen sollte, können sowohl kurzfristige Trends (bei niedriger Ordnung) als auch langfristige Trends (bei höherer Ordnung) analysiert werden.
Mit dieser Vorgehensweise geht jedoch ein gravierender Nachteil einher, der mit Blick auf die letzte Abbildung offensichtlich wird. So können für die ersten und letzten i Zeitpunkte (hier: i=2) des Beobachtungszeitraums keine Gleitenden Durchschnitte der Ordnung k (hier: k=2·i) errechnet werden. Es fehlen damit nicht nur wichtige Informationen gerade am aktuellen Rand des Beobachtungszeitraums, es können darüber hinaus natürlich auch keine Prognosen zur weiteren Trendentwicklung abgegeben werden. Aus diesem Grund soll im nächsten Unterkapitel eine Alternative zur Methode der Gleitenden Durchschnitte vorgestellt werden.

Trendfunktion nach der Methode der kleinsten Quadrate

Wie eben ausgeführt lässt die Methode der gleitenden Durchschnitte weder eine Trendinterpolation noch eine Trendextrapolation am bzw. über den aktuellen Rand des Beobachtungszeitraumes hinaus zu. Dieser Nachteil wird mit der Festlegung auf eine bestimmte Trendfunktion vermieden, deren Typus sich aus dem bisherigen Verlauf der Zeitreihenwerte ableiten lässt. Gesucht ist also eine Funktion, die frei von saisonalen und irregulären Schwankungen durch die gegebenen Zeitreihenwerte verläuft. Ihre Parameter sind so festzulegen, dass die Summe der quadrierten Abweichungen zwischen Zeitreihenwert und Trendwert minimal werden. Die Vorgehensweise zeigt damit enge Parallelen zur Regressionsrechung, an die Stelle der exogenen Variablen X tritt lediglich die Zeit t.

Definition: Lineare Trendfunktion

Die Lage der linearen Trendfunktion

$$\hat{y}(t) = a_0 + a_1 \cdot t$$

wird nach der Methode der Kleinsten Quadrate so bestimmt, dass die Summe der Abweichungsquadrate minimiert wird:

$$a_1 = \frac{N\sum t_i\, y_i - \sum t_i \sum y_i}{N\sum t_i^2 - (\sum t_i)^2} = \frac{Cov(t, Y)}{s_t^2}$$

$$a_0 = \frac{\sum y_i}{N} - a_1 \cdot \frac{\sum t_i}{N} = \bar{y} - a_1 \cdot \bar{t}$$

Um die Koeffizienten der Trendfunktion letztendlich berechnen zu können, ist es notwendig, die Zeitpunkte zu codieren. Dazu wird die kleinste Zeitspanne innerhalb einer Zeitreihe gleich Eins gesetzt. Mit welcher Zahl der erste Zeitpunkt codiert wird, ist unerheblich.

13. Zeitreihenanalyse

Beispiel:

In der nachfolgenden Arbeitstabelle wird bspw. die erste Codiervariante t_1 mit Eins begonnen und - da zwischen allen Zeitpunkten die identische Zeitspanne von einem Quartal liegt - fortlaufend um eine Einheit erhöht.

Zeitpunkt	y_i	Codierung Variante t_{1i}	Codierung Variante t_{2i}	t_{2i}^2	$t_{2i} \cdot y_i$
12/1Q	1,6	1	-9,5	90,25	-15,20
12/2Q	6,4	2	-8,5	72,25	-54,40
12/3Q	5,6	3	-7,5	56,25	-42,00
12/4Q	2,4	4	-6,5	42,25	-15,60
13/1Q	1,6	5	-5,5	30,25	-8,80
13/2Q	6,5	6	-4,5	20,25	-29,25
13/3Q	5,7	7	-3,5	12,25	-19,95
12/4Q	2,4	8	-2,5	6,25	-6,00
14/1Q	2,0	9	-1,5	2,25	-3,00
14/2Q	6,7	10	-0,5	0,25	-3,35
14/3Q	5,1	11	0,5	0,25	2,55
14/4Q	2,6	12	1,5	2,25	3,90
15/1Q	1,3	13	2,5	6,25	3,25
15/2Q	5,8	14	3,5	12,25	20,30
15/3Q	6,8	15	4,5	20,25	30,60
15/4Q	2,9	16	5,5	30,25	15,95
16/1Q	1,9	17	6,5	42,25	12,35
16/2Q	6,1	18	7,5	56,25	45,75
16/3Q	7,0	19	8,5	72,25	59,50
16/4Q	2,2	20	9,5	90,25	20,90
Summe	82,6	210,0	0,0	665,0	17,50

Gerade im Fall identischer Zeitspannen (sog. äquidistante Zeitpunkte) bietet sich eine alternative Codierung (siehe t_2) an, die zu einer Codiersumme von Null führt und damit die Bestimmungsgleichungen der beiden Koeffizienten erheblich vereinfacht:

$$a_1 = \frac{17{,}50}{665{,}0} = 0{,}0263 \left[\frac{1000\,\text{Ü.}}{\text{Quartal}}\right] \quad \text{und} \quad a_0 = \frac{82{,}6}{20} = 4{,}130\,[1000\,\text{Ü.}]$$

$$\hat{y}(t) = 4{,}130 + 0{,}0263 \cdot t$$

Danach liegt die durchschnittliche Anzahl an Übernachtungen im Zeitpunkt t=0 (entspricht nach der vorgenommenen Codierung dem Zeitpunkt Ende Juni 2014) bei 4130. Dieser Wert erhöht sich zudem um durchschnittlich 26,3 Übernachtungen je Quartal. Nach der linearen Trendfunktion ist also über den gesamten betrachteten Zeitraum von einem leicht steigenden Trendverlauf auszugehen.

Analog zur einfachen Regressionsrechnung können auch nichtlineare Trendfunktionen verwendet werden. In Abschnitt 12.2. wurden verschiedene Funktionstypen bereits vorgestellt. Sie lassen sich einfach auf Trendfunktionen übertragen, indem die exogene Variable X durch den Einfluss der Zeit t ersetzt wird.

12.3. Ermittlung der Saisonkomponente

Für eine aussagefähige Beschreibung des zeitlichen Verlaufes einer ökonomischen Größe reicht die Trendbestimmung im Allgemeinen nicht aus, da der Trend oft durch periodische oder auch nichtperiodische (irreguläre) Schwankungen überlagert wird. Zur Ermittlung der periodischen Schwankungen wird angenommen, dass der Gesamtzeitraum aus m Perioden (z.B. Jahre) besteht und jede Periode in n Phasen (Teilzeiträume, z.B. Woche, Monat, Quartal) zerlegbar ist. Sind die periodischen Schwankungen jahreszeitlich bedingt, so heißen sie Saisonschwankungen. Die Kennziffern zur Charakterisierung dieser Schwankungen werden als Saisonindizes bezeichnet.

Nachfolgend soll nun das sog. **Phasendurchschnittsverfahren** vorgestellt werden, das genau diese Saisonschwankungen quantifizieren, also die Saisonindizes berechnen soll. Das Verfahren lässt sich dabei in fünf Schritte zerlegen, unterstellt wird dabei aus den weiter oben erläuterten Gründen ein additives Zeitreihenmodell:

- 1. Schritt: Berechnung der Trendwerte y^T_{ik}
- 2. Schritt: Berechnung der trendbereinigten Werte y^*_{ik}
- 3. Schritt: Berechnung durchschnittlicher trendbereinigter Werte
- 4. Schritt: Berechnung der Saisonindizes I_k
- 5. Schritt: Berechnung der Schätzwerte $\hat{y}_{ik}$

Im Normalfall werden hierzu die Zeitreihenwerte von Y in einer zweidimensionalen Tabelle mit m Zeilen für die Perioden (i = 1, …, m) und n Spalten für die Phasen (k = 1, …, n) angeordnet. Die Zeitreihe der Übernachtungszahlen, auf die das Phasendurchschnittsverfahren angewendet werden soll, erhält somit folgendes Aussehen:

y_{ik}	Phase			
Periode	1. Quartal (k=1)	2. Quartal (k=2)	3. Quartal (k=3)	4. Quartal (k=4)
2012 (i=1)	1,6	6,4	5,6	2,4
2013 (i=2)	1,6	6,5	5,7	2,4
2014 (i=3)	2,0	6,7	5,1	2,6
2015 (i=4)	1,3	5,8	6,8	2,9
2016 (i=5)	1,9	6,1	7,0	2,2

Alternativ lassen sich die Schritte eins bis fünf des Phasendurchschnittsverfahrens auch aus einer eindimensionalen Tabelle entwickeln, in der die Y-Zeitreihenwerte chronologisch zeilenweise notiert sind. Die ursprüngliche Form der Datenerfassung kann also einfach beibehalten werden. Zudem erlaubt diese Variante eine übersichtlichere Darstellung des Phasendurchschnittsverfahrens und soll daher den folgenden Ausführungen zugrunde gelegt werden:

13. Zeitreihenanalyse

Zeitpunkt	Periode i	Phase k	y_{ik}	y_{ik}
1	2012 (i=1)	1. Q. (k=1)	y_{11} (= $y_{12/1.Q.}$)	1,6
2	2012 (i=1)	2. Q. (k=2)	y_{12} (= $y_{12/2.Q.}$)	6,4
3	2012 (i=1)	3. Q. (k=3)	y_{13} (= $y_{12/3.Q.}$)	5,6
4	2012 (i=1)	4. Q. (k=4)	y_{14} (= $y_{12/4.Q.}$)	2,4
5	2013 (i=2)	1. Q. (k=1)	y_{21} (= $y_{13/1.Q.}$)	1,6
6	2013 (i=2)	2. Q. (k=2)	y_{21} (= $y_{13/2.Q.}$)	6,5
7	2013 (i=2)	3. Q. (k=3)	y_{21} (= $y_{13/3.Q.}$)	5,7
8	2013 (i=2)	4. Q. (k=4)	y_{21} (= $y_{13/4.Q.}$)	2,4
9	2014 (i=3)	1. Q. (k=1)	y_{31} (= $y_{14/1.Q.}$)	2,0
10	2014 (i=3)	2. Q. (k=2)	y_{32} (= $y_{14/2.Q.}$)	6,7
11	2014 (i=3)	3. Q. (k=3)	y_{33} (= $y_{14/3.Q.}$)	5,1
12	2014 (i=3)	4. Q. (k=4)	y_{34} (= $y_{14/4.Q.}$)	2,6
13	2015 (i=4)	1. Q. (k=1)	y_{41} (= $y_{15/1.Q.}$)	1,3
14	2015 (i=4)	2. Q. (k=2)	y_{42} (= $y_{15/2.Q.}$)	5,8
15	2015 (i=4)	3. Q. (k=3)	y_{43} (= $y_{15/3.Q.}$)	6,8
16	2015 (i=4)	4. Q. (k=4)	y_{44} (= $y_{15/4.Q.}$)	2,9
17	2016 (i=5)	1. Q. (k=1)	y_{51} (= $y_{16/1.Q.}$)	1,9
18	2016 (i=5)	2. Q. (k=2)	y_{52} (= $y_{16/2.Q.}$)	6,1
19	2016 (i=5)	3. Q. (k=3)	y_{53} (= $y_{16/3.Q.}$)	7,0
20	2016 (i=5)	4. Q. (k=4)	y_{54} (= $y_{16/4.Q.}$)	2,2

1. Schritt: Berechnung der Trendwerte y^T_{ik}

Variante A: Gleitende Durchschnitte der Ordnung k = n = Anzahl der Phasen
Variante B: lineare Trendfunktion

Um im weiteren Verlauf auch Prognoserechnungen vornehmen zu können, sollen die Trendwerte aus der linearen Trendfunktion $\hat{y}(t) = 4{,}130+0{,}0263 \cdot t$ errechnet werden:

Zeitpunkt	y_{ik}	y_{ik}	t	y^T_{ik}	y^T_{ik}
1	y_{11}	1,6	-9,5	= 4,130 + 0,0263 · (-9,5)	3,883
2	y_{12}	6,4	-8,5	= 4,130 + 0,0263 · (-8,5)	3,909
3	y_{13}	5,6	-7,5	= 4,130 + 0,0263 · (-7,5)	3,935
4	y_{14}	2,4	-6,5	= 4,130 + 0,0263 · (-6,5)	3,961
5	y_{21}	1,6	-5,5	= 4,130 + 0,0263 · (-5,5)	3,987
6	y_{22}	6,5	-4,5	= 4,130 + 0,0263 · (-4,5)	4,013
7	y_{13}	5,7	-3,5	= 4,130 + 0,0263 · (-3,5)	4,039
8	y_{24}	2,4	-2,5	= 4,130 + 0,0263 · (-2,5)	4,065
9	y_{31}	2,0	-1,5	= 4,130 + 0,0263 · (-1,5)	4,091
10	y_{32}	6,7	-0,5	= 4,130 + 0,0263 · (-0,5)	4,117
11	y_{33}	5,1	0,5	= 4,130 + 0,0263 · (0,5)	4,143
12	y_{34}	2,6	1,5	= 4,130 + 0,0263 · (1,5)	4,169
13	y_{41}	1,3	2,5	= 4,130 + 0,0263 · (2,5)	4,195
14	y_{42}	5,8	3,5	= 4,130 + 0,0263 · (3,5)	4,221
15	y_{43}	6,8	4,5	= 4,130 + 0,0263 · (4,5)	4,247
16	y_{44}	2,9	5,5	= 4,130 + 0,0263 · (5,5)	4,273
17	y_{51}	1,9	6,5	= 4,130 + 0,0263 · (6,5)	4,299
18	y_{52}	6,1	7,5	= 4,130 + 0,0263 · (7,5)	4,325
19	y_{53}	7,0	8,5	= 4,130 + 0,0263 · (8,5)	4,351
20	y_{54}	2,2	9,5	= 4,130 + 0,0263 · (9,5)	4,377

2. Schritt: Berechnung der trendbereinigten Werte y^*_{ik}

Additives Modell: Subtraktion der Trendwerte von den Messwerten → $y^*_{ik} = y_{ik} - y^T_{ik}$

(Multiplikatives Modell: Division der Messwerte durch die Trendwerte → $y^*_{ik} = y_{ik} / y^T_{ik}$)

Bei dem hier unterstellten additiven Modellansatz ergeben sich die trendbereinigten Werte durch Subtraktion der Trendwerte von den erfassten Zeitreihenwerten. Um die Berechnung der Phasendurchschnitte im nächsten Schritt zu erleichtern, werden die trendbereinigten Werte gleich der betreffenden Phase (unterjähriger Zeitpunkt) zugeordnet:

Zeitpunkt	y_{ik}		y^T_{ik}	y^*_{ik}	1. Quartal	2. Quartal	3. Quartal	4. Quartal
1	y_{11}	1,6	3,883	= 1,6 - 3,883	-2,283			
2	y_{12}	6,4	3,909	= 6,4 - 3,909		2,491		
3	y_{13}	5,6	3,935	= 5,6 - 3,935			1,665	
4	y_{14}	2,4	3,961	= 2,4 - 3,961				-1,561
5	y_{21}	1,6	3,987	= 1,6 - 3,987	-2,387			
6	y_{22}	6,5	4,013	= 6,5 - 4,013		2,487		
7	y_{13}	5,7	4,039	= 5,7 - 4,039			1,661	
8	y_{24}	2,4	4,065	= 2,4 - 4,065				-1,665
9	y_{31}	2,0	4,091	= 2,0 - 4,091	-2,091			
10	y_{32}	6,7	4,117	= 6,7 - 4,117		2,583		
11	y_{33}	5,1	4,143	= 5,1 - 4,143			0,957	
12	y_{34}	2,6	4,169	= 2,6 - 4,169				-1,569
13	y_{41}	1,3	4,195	= 1,3 - 4,195	-2,895			
14	y_{42}	5,8	4,221	= 5,8 - 4,221		1,579		
15	y_{43}	6,8	4,247	= 6,8 - 4,247			2,553	
16	y_{44}	2,9	4,273	= 2,9 - 4,273				-1,373
17	y_{51}	1,9	4,299	= 1,9 - 4,299	-2,399			
18	y_{52}	6,1	4,325	= 6,1 - 4,325		1,775		
19	y_{53}	7,0	4,351	= 7,0 - 4,351			2,649	
20	y_{54}	2,2	4,377	= 2,2 - 4,377				-2,177

Die trendbereinigten Werte lassen durch ihre spaltenweise Zuordnung nach Phasen bereits deutliche Saisoneinflüsse erkennen. Sie nehmen sowohl im ersten als auch im letzten Quartal durchweg negative Werte an. Die Zeitreihenwerte liegen also unter den jeweiligen Trendwerten, was auf einen negativen Saisoneinfluss hindeutet. In den beiden anderen Quartalen verhält es sich genau umgekehrt, hier liegt also ein positiver Saisoneinfluss vor.
Allerdings fallen die Abweichungen innerhalb einer Phase (Quartal) für die verschiedenen Perioden (Jahre) unterschiedlich hoch aus. Hieran zeigt sich die Überlagerung der Saisonkomponente durch Rest- oder Störgrößen, deren Einfluss nicht weiter beschrieben werden kann. Die Durchschnittsberechnungen in den beiden nächsten Schritten des Phasendurchschnittsverfahrens sollen diese Erkenntnisse noch weiter verdichten.

13. Zeitreihenanalyse

3. Schritt: Berechnung durchschnittlicher trendbereinigter Werte

→ Berechnung der Spaltendurchschnitte = Phasendurchschnitt = $\overline{y^*}_{.k}$

→ Berechnung des Durchschnitts aller Spaltendurchschnitte = Gesamtdurchschnitt

$$\overline{y^*} = \frac{1}{n}\sum_{k=1}^{n}\overline{y^*}_{.k}$$

Zeitpunkt	y_{ik}		y^T_{ik}	y^*_{ik}	1. Quartal	2. Quartal	3. Quartal	4. Quartal
1	y_{11}	1,6	3,883	= 1,6 - 3,883	-2,283			
2	y_{12}	6,4	3,909	= 6,4 - 3,909		2,491		
3	y_{13}	5,6	3,935	= 5,6 - 3,935			1,665	
4	y_{14}	2,4	3,961	= 2,4 - 3,961				-1,561
5	y_{21}	1,6	3,987	= 1,6 - 3,987	-2,387			
6	y_{22}	6,5	4,013	= 6,5 - 4,013		2,487		
7	y_{13}	5,7	4,039	= 5,7 - 4,039			1,661	
8	y_{24}	2,4	4,065	= 2,4 - 4,065				-1,665
9	y_{31}	2,0	4,091	= 2,0 - 4,091	-2,091			
10	y_{32}	6,7	4,117	= 6,7 - 4,117		2,583		
11	y_{33}	5,1	4,143	= 5,1 - 4,143			0,957	
12	y_{34}	2,6	4,169	= 2,6 - 4,169				-1,569
13	y_{41}	1,3	4,195	= 1,3 - 4,195	-2,895			
14	y_{42}	5,8	4,221	= 5,8 - 4,221		1,579		
15	y_{43}	6,8	4,247	= 6,8 - 4,247			2,553	
16	y_{44}	2,9	4,273	= 2,9 - 4,273				-1,373
17	y_{51}	1,9	4,299	= 1,9 - 4,299	-2,399			
18	y_{52}	6,1	4,325	= 6,1 - 4,325		1,775		
19	y_{53}	7,0	4,351	= 7,0 - 4,351			2,649	
20	y_{54}	2,2	4,377	= 2,2 - 4,377				-2,177
	Phasendurchschnitte $\overline{y^*}_{.k}$				-2,411	2,183	1,897	-1,669
	Gesamtdurchschnitt $\overline{y^*}$				0,000			

4. Schritt: Berechnung der Saisonindizes I_k (Normierung der Phasenmittelwerte)

Additives Modell: $I_k = \overline{y^*}_{.k} - \overline{y^*}$ mit $\Sigma I_k = 0$

(Mulitiplikatives Modell: $I_k = \overline{y^*}_{.k} / \overline{y^*}$ mit $\Sigma I_k = n$)

Die Berechnung der Saisonindizes ergibt sich im vorliegenden Beispiel eines additiven Zeitreihenmodells aus der Subtraktion des Gesamtdurchschnitts von den jeweiligen Phasendurchschnitten, also:

I(1. Quartal) = $I_1 = \overline{y^*}_{.1} - y^*$ = -2,411 - 0,000 = -2,411

I(2. Quartal) = $I_2 = \overline{y^*}_{.2} - y^*$ = 2,183 - 0,000 = 2,183

I(3. Quartal) = $\bar{I}_3 = \bar{y}^*_{.3} - y^* = 1{,}897 - 0{,}000 = 1{,}897$

I(4. Quartal) = $\bar{I}_4 = \bar{y}^*_{.4} - y^* = -1{,}669 - 0{,}000 = -1{,}669$

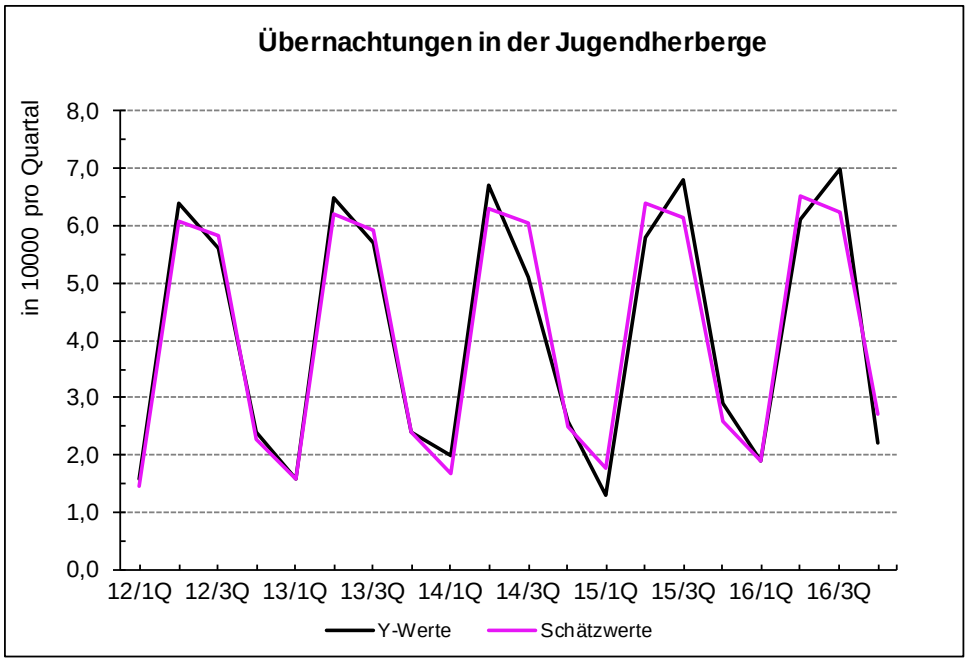
Natürlich wäre hier eine Normierung nicht nötig gewesen, da der Gesamtdurchschnitt der trendbereinigten Werte ja exakt bei Null liegt. Dies ist jedoch nicht zwangsläufig der Fall, sondern hängt von der Stärke und der Struktur zufälliger Einflüsse (Restkomponente) auf die Zeitreihenwerte ab. Ist diese Restkomponente nur schwach ausgeprägt, so lässt sie sich - wie in unserem Beispiel - bereits durch die Berechnung der Phasendurchschnitte vollständig eliminieren. In den meisten Fällen ist allerdings eine zweite Korrektur nötig, nämlich die hier dargestellte Normierung der Phasenmittelwerte.

Spätestens nach dieser Normierung sind die Phasenmittelwerte vollständig von Trend- und Zufallseinflüssen befreit und spiegeln ausschließlich den durchschnittlichen Saisoneinfluss innerhalb einer Periode (i.d.R. Jahr) wider. Dies zeigt sich auch an den Werten der Saisonindizes, die sich bei der hier unterstellten additiven Modellstruktur exakt zu Null addieren.
Positive Saisonindizes signalisieren dann einen positiven Einfluss der Jahreszeit auf die betrachtete Zeitreihe und umgekehrt. So liegen die Übernachtungen im zweiten und dritten Quartal um durchschnittlich 2183 bzw. 1897 über den jeweiligen Trendwerten. Dagegen bewirkt der Saisoneinfluss im ersten und vierten Quartal einen Rückgang gegenüber dem Trendwert um durchschnittlich 2411 bzw. 1669 Übernachtungen. Über das gesamte Jahr betrachtet, also über eine Periode, heben sich diese Saisoneinflüsse gegenseitig auf, die Summe der Saisonindizes ist daher Null.

Schritt 5: Berechnung der Schätzwerte $\hat{y}_{ik}$

Additives Modell: $\hat{y}_{ik} = y^T_{ik} + I_k$

(Mulitiplikatives Modell: $\hat{y}_{ik} = y^T_{ik} \cdot I_k$)



13. Zeitreihenanalyse

Zeitpunkt	y_{ik}		y^T_{ik}	I_k	$\hat{y}_{ik} = y^T_{ik} + I_k$
1	y_{11}	1,6	3,883	-2,411	1,472
2	y_{12}	6,4	3,909	2,183	6,092
3	y_{13}	5,6	3,935	1,897	5,832
4	y_{14}	2,4	3,961	-1,669	2,292
5	y_{21}	1,6	3,987	-2,411	1,576
6	y_{22}	6,5	4,013	2,183	6,196
7	y_{13}	5,7	4,039	1,897	5,936
8	y_{24}	2,4	4,065	-1,669	2,396
9	y_{31}	2,0	4,091	-2,411	1,680
10	y_{32}	6,7	4,117	2,183	6,300
11	y_{33}	5,1	4,143	1,897	6,040
12	y_{34}	2,6	4,169	-1,669	2,500
13	y_{41}	1,3	4,195	-2,411	1,784
14	y_{42}	5,8	4,221	2,183	6,404
15	y_{43}	6,8	4,247	1,897	6,144
16	y_{44}	2,9	4,273	-1,669	2,604
17	y_{51}	1,9	4,299	-2,411	1,888
18	y_{52}	6,1	4,325	2,183	6,508
19	y_{53}	7,0	4,351	1,897	6,248
20	y_{54}	2,2	4,377	-1,669	2,708

Der Vergleich von Originalwerten und ihren Schätzwerten in der obigen Abbildung macht deutlich, dass das additive Modell den Verlauf der Übernachtungszahlen der Jugendherberge recht gut beschreibt.

Natürlich lässt sich die Schätzqualität der Zeitreihenmodelle analog zu den Regressionsmodellen mit Hilfe des Bestimmtheitsmaßes quantifizieren. Der Wert fällt mit über 0,96 sehr eindeutig aus. Es ist daher auch nicht überraschend, dass der Signifikanztest auf der Grundlage des F-Testes (vgl. Kapitel 12.1.) zu einer eindeutigen Ablehnung der Nullhypothese führt.

Angesichts der gegebenen Anpassungsgüte dieses Zeitreihenmodells ist es nun möglich, eine Prognose zur Entwicklung der Übernachtungszahlen für 2017 abzugeben:

$\hat{y}_{2017/1.Q} = \hat{y}_{61} = y^T_{61} + I_1 = (4{,}130 + 0{,}0263 \cdot 10{,}5) - 2{,}411 = 1{,}995$

$\hat{y}_{2017/2.Q} = \hat{y}_{62} = y^T_{62} + I_2 = (4{,}130 + 0{,}0263 \cdot 11{,}5) + 2{,}183 = 6{,}615$

$\hat{y}_{2017/3.Q} = \hat{y}_{63} = y^T_{63} + I_3 = (4{,}130 + 0{,}0263 \cdot 12{,}5) + 1{,}897 = 6{,}356$

$\hat{y}_{2017/4.Q} = \hat{y}_{64} = y^T_{64} + I_4 = (4{,}130 + 0{,}0263 \cdot 13{,}5) - 1{,}669 = 2{,}816$

Für das Jahr 2017 kann die Leitung der Jugendherberge also mit einer Gesamtzahl an Übernachtungen in Höhe von 17782 rechnen.

Aufgaben und Lösungen

Aufgabe IV.1.:

(Ergänzung zu Aufgabe III.2.)
Neben monatlichen Einkommen (in Euro) wurden von Wismarer Studenten Informationen zu ihren monatlichen Ausgaben (in Euro) und ihrem häufigsten Fortbewegungsmittel (1 = zu Fuß bzw. Fahrrad, 2 = Auto) erfragt:

Einkommen	500	400	400	450	500	200	350	450	500	650
Ausgaben	400	200	300	450	300	200	350	450	500	500
Fortbewegung	2	1	2	2	1	2	2	2	2	2

e) Berechnen Sie für die Abhängigkeit der Ausgaben vom Einkommen eine lineare Regressionsfunktion und interpretieren Sie die Regressionskoeffizienten. Welche Ausgaben würden Sie erwarten, wenn ein Student über ein Monatseinkommen von 430 EUR verfügt?

f) Wie würde das Regressionsmodell aussehen, wenn Sie neben dem Einkommen das Fortbewegungsmittel als zweiten Regressor berücksichtigen würden? Formulieren Sie hierzu die entsprechende Regressionsgleichung (keine Berechnung!).

Lösung:

e) $a_1 = Cov(X, Y) / s_x^2 = 8650{,}0 / 12400{,}0 = 0{,}69758$ [EUR-Ausg./EUR-Eink.]
$\hat{y}(x) = 58{,}065 + 0{,}69758 \cdot x$
$\hat{y}(430) = 358{,}02$ EUR

f) $\hat{y}(x_1, x_2) = a_0 + a_1 \cdot x_1$ (=Einkommen) $+ a_2 \cdot x_2$ (= 0, wenn zu Fuß; = 1, wenn Auto)

Aufgabe IV.2.:

(Ergänzung zu Aufgabe III.3.)
Bei der Veröffentlichung von Unternehmenszahlen für 2001 hat VW auch auf den Einfluss des PKW-Absatzes (in Mio. Stück) auf das Betriebsergebnis (in Mio. EUR) verwiesen. Für ausgewählte Jahre liegen die folgenden Ergebnisse vor:

Jahr	1992	1993	1996	1997	1998	1999	2000
Betriebsergebnis	75	-992	347	696	1147	844	2062
PKW-Absatz	3,4	3,0	4,0	4,3	4,8	4,9	5,2

d) Berechnen Sie für den skizzierten Zusammenhang eine lineare Regressionsfunktion und interpretieren Sie beide Regressionskoeffizienten. Kann mit dieser Funktion das Betriebsergebnis für die beiden Jahre 1994 und 1995 geschätzt werden? (Begründung!)

e) Bewerten Sie die Schätzqualität des unter d) berechneten Regressionsmodells.

Lösung:

d) $a_1 = 621{,}743 / 0{,}568 = 1094{,}618$ [Euro/Stück]
$a_0 = 597 - 1094{,}618 \cdot 4{,}229 = -4033{,}666$ [Mio. Euro]
keine Schätzung für 1994 und 1995 möglich, da Absatz-Informationen fehlen

e) $B^2 = 0{,}941^2 = 0{,}885$ d.h. falls signifikant, ausreichende Schätzqualität
$F = 46{,}174 \sim {}_{Fdf1=1;\ df2=5}$ / $KB_{0,05} = [6{,}608; +\infty)$ → eindeutige Ablehnung von H_0

Aufgabe IV.3.:

Das Kieler Institut für Weltwirtschaft hat in den Jahren 2004 bis 2007 die Auswirkung von Lohnzurückhaltung (= Differenz zwischen jährlicher Wachstumsrate des nominalen Bruttosozialprodukts und jährlichem Lohnzuwachs) auf die Zunahme der Beschäftigung in der EU untersucht:

Land	Lohnzurückhaltung (in % pro Jahr)	Beschäftigungszuwachs (in % pro Jahr)
Belgien	2,1	0,0
Deutschland	0,6	-0,9
Finnland	2,3	1,5
Frankreich	1,0	0,2
Irland	6,9	3,9
Italien	1,7	-0,5
Niederlande	2,7	0,8
Spanien	3,2	1,4

a) Stellen Sie die Wertepaare grafisch dar und begründen Sie den erkennbaren Zusammenhang.

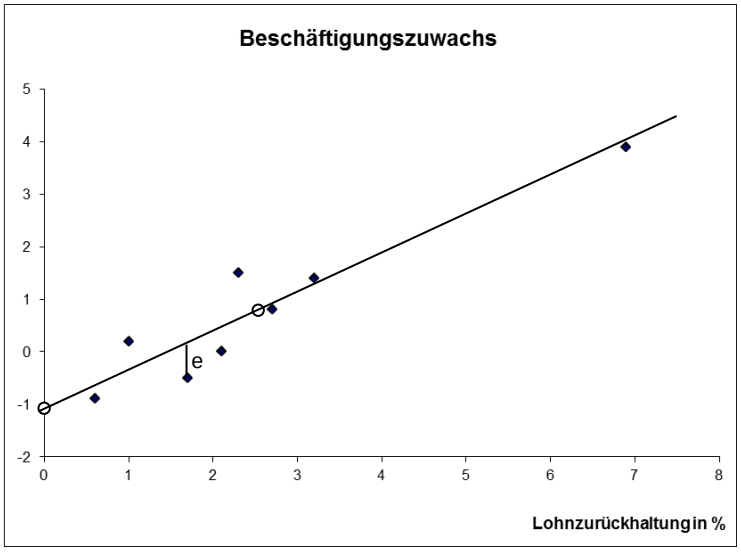
b) Berechnen Sie hierfür einen geeigneten Regressionsansatz und tragen Sie die Regressionsfunktion in die Grafik unter a) ein. Interpretieren Sie die Regressionskoeffizienten.

c) Verdeutlichen Sie exemplarisch anhand der Grafik die Methode der Kleinsten Quadrate und geben Sie eine begründete Schätzung für die Höhe des Bestimmtheitsmaßes.

d) Welche Lohnpolitik müsste folglich in Deutschland praktiziert werden, wenn bei einem nominalen Wirtschaftswachstum von jährlich 4% die Beschäftigtenzahl in den nächsten vier Jahren von 34 auf 35 Millionen gesteigert werden soll? (Lohn%=1,538%)

Lösung:

a) Streudiagramm mit X = Lohnzurückhaltung und Y = Beschäftigungszuwachs



Starker positiver linearer Zusammenhang; bei starker Lohnzurückhaltung wird der Produktionsfaktor Arbeit relativ günstiger und wird daher stärker nachgefragt.

Lösung Aufgabe IV.3. (Fortsetzung):

b) $s_x^2 = 79{,}09/8 - 2{,}563^2 = 3{,}317$ [%-LZ]^2
Cov(X, Y) = 35,81/8 - 2,563 · 0,8 = 2,426 [%-LZ · %-BZ]

$a_1 = 2{,}426 / 3{,}317 = 0{,}731$ [%-BZ / %-LZ]; d.h. nimmt die Lohnzurückhaltung um einen Prozentpunkt zu, dann erhöht sich die Beschäftigung um durchschnittlich 0,731% p.a.;
$a_0 = 0{,}800 - 0{,}731 \cdot 2{,}563 = -1{,}074$ [%-BZ]; d.h. liegt die Lohnzurückhaltung bei 0%, dann nimmt die Beschäftigung im Durchschnitt pro Jahr um 1,074% ab;

$\hat{y}(x) = -1{,}073 + 0{,}731 \cdot x$

c) Darstellung der Residuen e (siehe Grafik), entspricht dem vertikalen Abstand der Wertepaare von der Regressionsgeraden; diese Abstände fallen insgesamt sehr gering aus, daher hohes Bestimmtheitsmaß mit einem Wert über 0,8

d) Lohnzurückhaltung x = nom.BSP% - Lohn%
Beschäftigungszuwachs pro Jahr = (35/34^1/4 - 1) · 100 = 0,727%
$\hat{y}(x) = -1{,}073 + 0{,}731 \cdot (4 - \text{Lohn\%}) = 0{,}727$
Lohn% = 1,538%

Aufgabe IV.4.:

In einem Unternehmen mit 8 Beschäftigten liegen Daten über Alter und monatlichem Bruttoeinkommen vor:

Alter in Jahren	16	17	22	25	27	28	32	38
monatliches Bruttoeinkommen in EUR	1800	1900	2600	3600	3200	3500	4200	4800

a) Berechnen Sie eine Regressionsgleichung, mit der Sie das monatliche Bruttoeinkommen in Abhängigkeit vom Alter schätzen können.

b) Ermitteln Sie ein Maß für die Güte der Anpassung und begründen Sie, wie der Signifikanztest zu dieser Größe ausfallen würde (keine Berechnung!).

c) Interpretieren Sie die Regressionskoeffizienten und bewerten Sie diese Aussagen. Begründen Sie, inwieweit Sie diesen Regressionsansatz zur Schätzung des Bruttoeinkommens einsetzen würden.

Lösung:

a) Cov(X, Y) = 709500/8 - 25,625 · 3200 = 6687,50 [Jahre · EUR]
$s_x^2 = 5635/8 - 25{,}625^2 = 47{,}734$ Jahre^2
$\hat{y} = -390{,}037 + 140{,}099 \cdot x$

b) $s_y^2 = 89740000/8 - 3200^2 = 977500{,}0$ EUR^2
$B^2 = 0{,}958$; angesichts des hohen Bestimmtheitsmaßes nahe 1 wird der Signifikanztest zu einer Ablehnung der Nullhypothese führen

c) $a_1 = 140{,}099$ [Euro/Jahr]; d.h. mit jedem Jahr erhöht sich das monatliche Bruttoeinkommen eines Beschäftigten um 140,099 Euro (nachvollziehbar);
$a_0 = -390{,}037$ [Euro]; d.h. ein Beschäftigter mit Null Jahren erhält ein monatliches Bruttoeinkommen in Höhe von -390,037 Euro (unsinniges Ergebnis);
Linearität über das gesamte Altersspektrum ist problematisch

Aufgabe IV.5.:

Zur Prognose des DAX-Index liegt Ihnen die Entwicklung der Auftragseingänge in Deutschland in halbjährlichen Abständen vor:

Zeitpunkt (MoJa)	12/04	06/05	12/05	06/06	12/06	06/07
DAX-Index	4256	4586	5408	5683	6596	8007
Auftragseingang	100,8	101,3	108,9	108,6	114,0	117,6

a) Berechnen Sie eine Regressionsgleichung, mit der Sie den DAX-Index in Abhängigkeit von der Entwicklung der Auftragseingänge schätzen können.
($\hat{y}$=-15985,873+200,325·x)

b) Ermitteln Sie ein Maß für die Güte der Anpassung und bewerten Sie die ermittelte Regressionsgleichung. Interpretieren Sie zudem die Regressionskoeffizienten.
(B^2=0,946 d.h. sehr hohe Anpassungsgüte oder Schätzqualität)

c) Schätzen Sie mit Hilfe der Regressionsgleichung die weitere Entwicklung des DAX-Index, wenn bis Jahresende mit einem Anstieg der Auftragseingänge auf 136 Punkte zu rechnen ist. ($\hat{y}$(136)=11258,327 DAX-Punkte)

Aufgabe IV.6.:

Ein Ziel der Kreisgebietsreform in Mecklenburg-Vorpommern besteht in der Einsparung von Verwaltungspersonal. Der Vergleich mit anderen Bundesländern bringt folgende Ergebnisse:

Bundesland	Bevölkerung in Mio. Einwohner	Beschäftigte im öffentlichen Dienst in 1000
Berlin	3,395	259,7
Brandenburg	2,599	139,2
Bremen	0,663	39,8
Hamburg	1,744	107,2
Niedersachsen	7,994	431,6
Sachsen	4,274	235,0
Sachsen-Anhalt	2,470	145,8
Schleswig-Holstein	2,833	166,8

a) Stellen Sie die Beschäftigten in Abhängigkeit von der Einwohnerzahl grafisch dar.

b) Berechnen Sie ein geeignetes Maß für die Stärke des linearen Zusammenhangs.
(r_{xy} = 0,977)

c) Berechnen Sie eine lineare Regressionsfunktion und interpretieren Sie die Koeffizienten. Zeichnen Sie die Regressionsgerade in die Grafik ein.
(a_0 = 18,076 [1000 BöD]; a_1 = 53,145 [1000 BöD / 1 Mio. Einw.])

d) Bei einer Bevölkerung von 1,707 Mio. Einwohnern gibt es in Mecklenburg-Vorpommern 106300 Beschäftigte im öffentlichen Dienst. Wäre nach den Bedingungen der anderen Bundesländer eine geringere Anzahl an Beschäftigten zu erwarten?
($\hat{y}$(1,707) = 108,795 [1000 BöD]; also nein)

Aufgaben und Lösungen

Aufgabe IV.7.:

In einem Windkanal wird eine Windkraftanlage auf die Abhängigkeit der erzeugten Leistung (in kW/h) von der Windgeschwindigkeit (in m/s) getestet. Folgende Messdaten liegen vor:

Windgeschwindigkeit in m/s	2	3	4	6	10
Leistung in kW/h	30	52	72	98	100

a) Berechnen Sie eine geeignete Maßzahl für den Zusammenhang zwischen Leistung und Windgeschwindigkeit und interpretieren Sie Ihr Ergebnis. (r_{xy}=0,874 d.h. starker positiver linearer Zusammenhang von Windgeschwindigkeit und Leistung)

b) Berechnen Sie die Koeffizienten der einfachen linearen Regressionsfunktion und stellen Sie diese zusammen mit den Wertepaaren in geeigneter Weise grafisch dar. Verdeutlichen Sie an dieser Grafik die beiden Regressionskoeffizienten. (a_0=28,90 [kw/h]; a_1=8,30 [kw/h / m/s])

c) Interpretieren Sie die beiden Regressionskoeffizienten. Welche Kritikpunkte sind folglich gegen diese Funktion vorzubringen? Begründen Sie Ihre Antwort. (unsinnige Ergebnisse außerhalb des Schätzbereichs, vgl. a_0)

d) Schlagen Sie einen alternativen Regressionsansatz vor und berechnen Sie dessen Koeffizienten. (z.B. quadratische Funktion)

Aufgabe IV.8.:

Das Essener Marktforschungsunternehmen 'Im Pott' soll im Auftrag von Apple für das Produkt 'IPod' eine sinnvolle Preis-Absatz-Funktion ermitteln. Hierzu wurden auf den verschiedenen Testmärkten im Ruhrgebiet die Absatzzahlen bei unterschiedlicher Preisgestaltung festgestellt:

Preis in Euro	30	55	130	170	195	250
Absatz in Stück	17000	7500	2500	1200	500	400

a) Stellen Sie den Absatzverlauf in Abhängigkeit des Preises grafisch dar und begründen Sie, warum sich der Zusammenhang zwischen Preis und Absatz mit Hilfe einer Hyperbelfunktion recht gut beschreiben lässt.

b) Berechnen sie die Koeffizienten dieser Hyperbelfunktion. Mit welchem Absatz an IPods kann das Unternehmen auf einem vergleichbaren Markt bei einem Preis von 75 Euro rechnen? (a_0 = -2128,850; a_1 = 564175,4377; $\hat{y}(75)$ = 5393,489 Stück)

c) Welches Vertrauen bringen Sie der Absatzprognose bei 75 Euro entgegen? Begründen Sie Ihre Antwort auch dadurch, dass Sie eine geeignete Kennzahl für die Schätzqualität dieser nichtlinearen Regressionsfunktion berechnen. (B2 = 0,989)

d) Welche Probleme wären aufgetreten, wenn Sie in diesem Fall einen linearen Regressionsansatz gewählt hätten?

Aufgabe IV.9.:

Ein Unternehmen hat über 8 Quartale Preis, Absatzmenge und Werbeausgaben eines Gutes gegenübergestellt:

Quartal	Absatzmenge in 1000 Stück	Preis in EUR	Werbeausgaben in 1000 EUR
01/15	700	12	100
02/15	800	9	100
03/15	1100	5	100
04/15	1200	4	100
01/16	1000	12	200
02/16	1200	9	200
03/16	1400	6	200
04/16	1600	3	200

a) Welcher Zusammenhang ist zwischen Preis und Absatzmenge bzw. zwischen Werbeausgaben und Absatzmenge zu erwarten? Begründen Sie diese Zusammenhänge.

b) Stellen Sie die Abhängigkeit der Absatzmenge vom Preis in geeigneter Weise grafisch dar. Welche Erkenntnisse für die Modellspezifikation lassen sich hieraus ableiten?

c) Berechnen Sie auf der Grundlage Ihrer Überlegungen aus b) ein geeignetes Regressionsmodell und testen Sie dieses auf Signifikanz.

Aufgabe IV.10.:

In kaum einer anderen Industrienation sind Arbeitnehmer so gut vor Entlassungen geschützt wie in Deutschland. OECD-Ökonomen sehen darin den Hemmschuh für mehr Beschäftigung. Für 7 Länder stellen sie hierzu dem Index der Beschäftigungsschutzgesetzgebung (hoher Indexwert = hoher Kündigungsschutz) die konjunkturunabhängige Arbeitslosenquote gegenüber:

Land	D	F	I	NL	GB	Jap	USA
Index Beschäftigungsschutz	2,6	2,8	3,4	2,2	0,9	2,3	0,7
Arbeitslosenquote in %	7,1	9,7	9,9	5,0	6,1	3,5	5,3

a) Quantifizieren Sie die Behauptung 'Weniger Kündigungsschutz führt zu mehr Beschäftigung', indem Sie eine geeignete lineare Regressionsfunktion berechnen.

b) Zeigen Sie formal, dass aus der Höhe des Steigungskoeffizienten nicht auf die Schätzqualität der Regressionsgleichung geschlossen werden kann. Wie groß ist Ihr Vertrauen in den errechneten Zusammenhang? Berechnen Sie dazu eine geeignete Maßzahl.

c) Stellen Sie grafisch der geschätzten Regressionsfunktion die tatsächlichen Wertepaare gegenüber. Begründen Sie, in welcher Weise sich die Lage der Regressionsfunktion verändert, wenn nur die europäischen Länder betrachtet werden.

Aufgaben und Lösungen

Aufgabe IV.11.:

Die Henkel AG & Co. KGaA möchte ein neues Hairstyling-Produkt für die jugendliche Zielgruppe auf den Markt bringen und testet den Absatz vorab in 8 verschiedenen Städten:

Stadt	Preis des Produktes (in EUR)	Einwohnerzahl (in 1000)	Absatz (in 1000)
1	2,40	50	3
2	2,50	40	2
3	2,80	65	3
4	2,60	38	1
5	2,50	50	2
6	2,30	60	5
7	3,00	55	1
8	2,50	32	1

a) Bestimmen Sie die Gleichung der Regressionsgeraden für die Preisabhängigkeit des Absatzes und interpretieren Sie die Regressionskoeffizienten.

b) Berechnen Sie je eine Maßzahl für die Güte der Anpassung der in a) berechneten Regressionsgeraden sowie für die Stärke des linearen Zusammenhangs der Merkmale Preis und Absatz und interpretieren Sie beide.

Zur Verbesserung der Schätzung wurde die Einwohnerzahl als zweiter Regressor hinzugenommen und die neue Regressionsfunktion $y = 8{,}97 - 4{,}559 \cdot x_1 + 0{,}103 \cdot x_2$ errechnet.

c) Welches geometrisches Gebilde stellt diese Gleichung dar?

d) Schätzen Sie auf der Grundlage der neuen Regressionsfunktion den Absatz, den Henkel in Wismar mit 46000 Einwohnern erzielen kann, wenn dort ein Preis von 2,59 EUR verlangt wird. Hat sich die Anpassungsgüte durch die neue Funktion verbessert?

Aufgabe IV.12.:

Für Anlageentscheidungen am deutschen Aktienmarkt wird die Entwicklung am amerikanischen Aktienmarkt als wichtiger Indikator angesehen. Um diese Einschätzung zu überprüfen, liegen Ihnen folgende Notierungen (in 1000 Punkten) vor:

Zeitpunkt	06/12	12/12	06/13	12/13	06/14	12/14	06/15
Dow Jones	12,9	13,1	14,9	16,6	16,8	17,8	17,6
DAX	6,4	7,6	8,0	9,5	9,8	9,8	10,9

a) Erfassen Sie die Abhängigkeit des DAX vom Dow Jones - Index in einer zweidimensionalen Häufigkeitstabelle. Welche Erkenntnis über die Abhängigkeit des DAX vom Dow Jones - Index lässt sich aus dieser Tabelle ableiten? Ist diese Darstellung sinnvoll?

b) Berechnen Sie nun eine geeignete Regressionsfunktion und interpretieren Sie deren Koeffizienten. Vergleichen Sie diese Ergebnisse mit den unter a) gewonnenen Erkenntnissen. Was fällt Ihnen dabei auf?

c) Aktuell weist der Dow Jones Index einen Stand von 17800 Punkten auf, der DAX von 11300 Punkten. Hätten Sie den Anstieg des DAX bei dieser Kursentwicklung des Dow Jones Index erwartet? Begründen Sie Ihre Antwort durch geeignete Berechnungen.

Aufgaben und Lösungen

Aufgabe IV.13.:

Die Henkel AG & Co. KGaA untersucht für die Geschäftsjahre 2010 bis 2015 die Umsatzentwicklung in Abhängigkeit von den Marketing- und Vertriebsausgaben (in Mrd. EUR):

Geschäftsjahr	2010	2011	2012	2013	2014	2015
Umsatz	14,1	13,3	15,1	15,6	16,5	16,3
Marketing und Vertrieb	3,8	3,6	4,0	4,2	4,8	4,5

a) Stellen Sie die Wertepaare in geeigneter Weise grafisch dar und begründen Sie den erkennbaren Zusammenhang.

b) Warum ist der Korrelationskoeffizient von Pearson/Bravais eine geeignete Kennzahl zur Berechnung des Zusammenhangs zwischen den beiden Merkmalen? Berechnen Sie diesen und interpretieren Sie das Ergebnis.

c) Berechnen Sie darüber hinaus eine geeignete lineare Regressionsfunktion und tragen Sie diese in die Grafik unter a) ein. Interpretieren Sie die Koeffizienten dieser Regressionsfunktion und schätzen Sie den Umsatz für das Geschäftsjahr 2016, wenn bekannt ist, dass sich die Marketing- und Vertriebsausgaben auf 4,8 Mrd. EUR erhöht haben.

   Wie hoch ist Ihr Vertrauen in diese Schätzung? Begründen Sie Ihre Antwort, indem Sie hierzu eine geeignete Kennzahl berechnen. Überprüfen Sie diese auch auf eventuelle Zufallseinflüsse.

Aufgabe IV.14.:

Bei der Veröffentlichung der Unternehmenszahlen für 2015 hat VW auch auf den Einfluss von Absatz (in Mio. PKW) und Umsatz (in Mrd. EUR) auf das Betriebsergebnis (in Mrd. EUR) verwiesen. Für zurückliegende Geschäftsjahre liegen die folgenden Ergebnisse vor:

Jahr	2005	2006	2007	2008	2009	2011	2013	2014
Betriebsergebnis	1,1	2,7	4,1	0,9	4,7	7,2	9,1	11,1
Umsatz	94	105	109	114	105	159	197	202
Absatz	5,2	5,7	6,2	6,3	6,3	8,3	9,7	10,2

Zudem sind folgende Kennzahlen bekannt:

Cov(Umsatz; Betriebsergebnis) = 132,817 [Mrd. EUR · Mrd. EUR]

Cov(Umsatz; Absatz) = 73,159 [Mrd. EUR · Mio. PKW]

a) Stellen Sie die Abhängigkeit des VW-Betriebsergebnisses vom Umsatz in geeigneter Weise grafisch dar. Welcher Zusammenhang ist hieran zu erkennen?

b) Berechnen Sie die Stärke der linearen Abhängigkeit des Betriebsergebnisses vom Umsatz und interpretieren Sie das Ergebnis. Wie lässt sich das Ergebnis ökonomisch begründen?

c) Berechnen Sie für die Abhängigkeit aus b) eine lineare Regressionsfunktion und interpretieren Sie die Regressionskoeffizienten. Begründen Sie, ob mit dieser Funktion das Betriebsergebnis für die Jahre 2010 und 2012 geschätzt werden kann.

d) Ließe sich die Schätzqualität der Regressionsfunktion aus c) deutlich verbessern, wenn der Absatz an Fahrzeugen als weitere erklärende Variable in das Regressionsmodell aufgenommen würde? Führen Sie hierzu geeignete Berechnungen durch.

Aufgabe IV.15.:

Der Produktionsindex in der deutschen Investitionsgüterindustrie zeigt für die Jahre 2003 bis 2006 folgenden Verlauf:

	April	August	Dezember
2003	115,6	103,7	125,5
2004	128,0	114,9	142,6
2005	133,1	117,1	137,0
2006	130,1	116,1	135,2

a) Stellen Sie die Zeitreihe grafisch dar und begründen Sie, weshalb in diesem Fall ein additives Zeitreihenmodell geeignet ist.

b) Bestimmen Sie den Trendverlauf nach der Methode geeigneter gleitender Durchschnitte und zeichnen Sie diesen in die Grafik von a) ein.

c) Warum ist die Methode gleitender Durchschnitte gegenüber einer linearen Trendfunktion besser geeignet, Trendwechsel frühzeitig zu identifizieren? Ist für das Jahr 2006 ein solcher Trendwechsel am Produktionsindex der deutschen Investitionsgüterindustrie zu erkennen? Begründen Sie Ihre Antwort.

Lösung:

a) additives Modell, da schwankende Amplituden:

b)

Datum	**Produktion**	**$D_T(3)$**
Apr 03	115,6	
Aug 03	103,7	114,933
Dez 03	125,5	119,067
Apr 04	128,0	122,800
Aug 04	114,9	128,500
Dez 04	142,6	130,200
Apr 05	133,1	130,933
Aug 05	117,1	129,067
Dez 05	137,0	128,067
Apr 06	130,1	127,733
Aug 06	116,1	127,133
Dez 06	135,2	

c) keine Vorgabe einer festen Trendfunktion; fortlaufende Durchschnittsberechnung für eine begrenzte Anzahl an Zeitreihenwerten (Ordnung) führt dazu, dass alte Werte herausfallen und neue Werte hinzukommen, Trendveränderungen werden dadurch sofort wirksam; seit April 2004 fallende Trendwerte, Trendwechsel noch nicht sichtbar

Aufgabe IV.16.:

Für einen internationalen Flughafen liegen von 2005 bis 2008 die folgenden Passagierzahlen vor:

	Fluggäste (in 1000)			
Zeit	1. Quartal	2. Quartal	3. Quartal	4. Quartal
2005	3581	3630	3877	4069
2006	3982	4061	4360	4485
2007	4389	4554	4831	4908
2008	4795	4848	5224	5332

a) Stellen Sie die Zeitreihe graphisch dar und erläutern Sie hieran die Komponenten einer Zeitreihe.

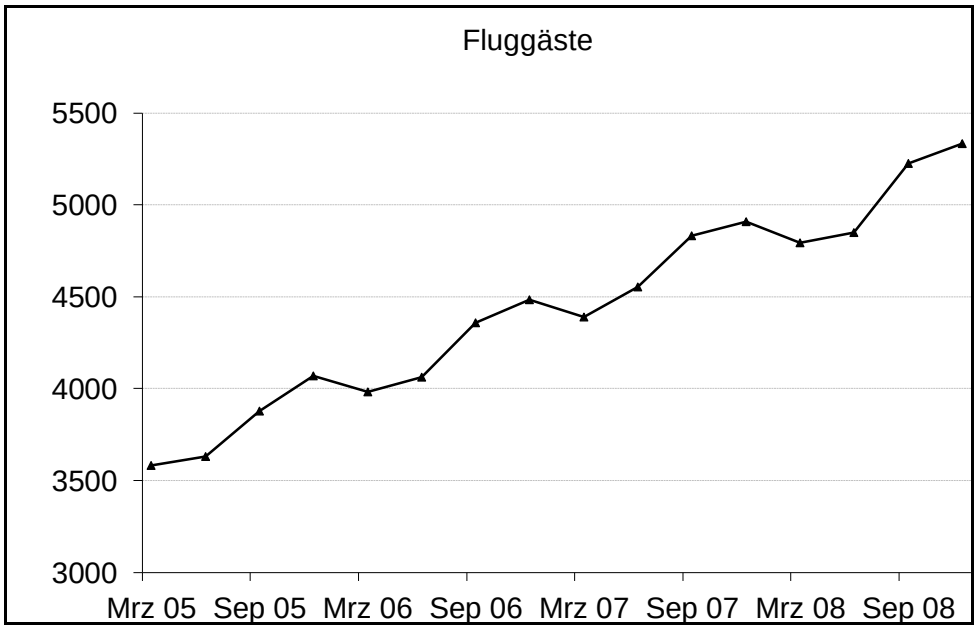
b) Berechnen Sie mit Hilfe des Phasendurchschnittsverfahrens unter Verwendung geeigneter gleitender Durchschnitte die Saisonindizes und interpretieren Sie diese.

c) Verwenden Sie die jährlichen Passagierzahlen als Grundlage für die Berechnung einer linearen Trendfunktion. Interpretieren Sie die Koeffizienten der Trendfunktion.

d) Welche Anzahl war auf Basis der Trendfunktion für das Jahr 2009 zu erwarten? Gliedern Sie diese Anzahl auf die Quartale auf.

Lösung:

a)



glatte Komponente: Trend und zyklische Schwankungen;
eindeutig steigender Verlauf

Saisonkomponente: unterjährige Schwankungen;
eindeutige Tiefpunkte im 1. und 2. Quartal, über dem Trend liegende Werte im 3. und 4. Quartal

Restkomponente: sonstige zufällige (nicht erklärte) Schwankungen;
schwankende Veränderung zwischen den Quartalen

Aufgaben und Lösungen

Lösung Aufgabe IV.16. (Fortsetzung):

b)

			$y^*_{ik} = y_{ik} - y^T_{ik}$			
y	**y_{ik}**	**$y^T_{ik} = D_T(4)$**	1. Q.	2. Q.	3. Q.	4. Q.
y_{11} (= $y_{1.Q2005}$)	3581					
y_{12} (= $y_{2.Q2005}$)	3630					
y_{13} (= $y_{3.Q2005}$)	3877	3839,375			37,625	
y_{14} (= $y_{4.Q2005}$)	4069	3943,375				125,625
y_{21} (= $y_{1.Q2006}$)	3982	4057,625	-75,625			
y_{22} (= $y_{2.Q2006}$)	4061	4170,000		-109,000		
y_{23} (= $y_{3.Q2006}$)	4360	4272,875			87,125	
y_{24} (= $y_{4.Q2006}$)	4485	4385,375				99,625
y_{31} (= $y_{1.Q2007}$)	4389	4505,875	-116,875			
y_{32} (= $y_{2.Q2007}$)	4554	4617,625		-63,625		
y_{33} (= $y_{3.Q2007}$)	4831	4721,250			109,750	
y_{34} (= $y_{4.Q2007}$)	4908	4808,750				99,250
y_{41} (= $y_{1.Q2008}$)	4795	4894,625	-99,625			
y_{42} (= $y_{2.Q2008}$)	4848	4996,750		-148,750		
y_{43} (= $y_{3.Q2008}$)	5224					
y_{44} (= $y_{4.Q2008}$)	5332					
Phasendurchschnitte $\overline{y^*_{.k}}$			-97,375	-107,125	78,1667	108,1667
Gesamtdurchschnitt $\overline{y^*}$			-4,5417			
Saisonindex $I_k = \overline{y^*_{.k}} - \overline{y^*}$			-92,8333	-102,5833	82,7084	112,7084
ΣI_k			0,0002			

im 3. und 4. Quartal liegen die Fluggastzahlen im Durchschnitt um rund 100000 über ihren Trendwerten, im 1. und 2. Quartal um etwa 100000 darunter

c)

Jahr	**gesamt (y)**	**t**	**t * t**	**t * y**
2005	15157	-1,5	2,25	-22735,5
2006	16888	-0,5	0,25	-8444,0
2007	18682	0,5	0,25	9341,0
2008	20199	1,5	2,25	30298,5
Summe	70926	0	5,00	8460,00

a_1 = 8460/5 = 1692 [Fluggäste in 1000 / Jahr]; Anstieg der Fluggastzahlen pro Jahr
a_0 = 70926/4 = 17731,5 [Fluggäste in 1000]; zum Zeitpunkt t=0 = Jahresmitte 2007

d)

$\hat{y}(2009) = \hat{y}(t=2,5) = 17731,5 + 1692 \cdot 2,5 = 21961,5$ [Fluggäste in 1000]

1. Quartal 2009 = 21961,5/4 - 92,8333 = 5397,542
2. Quartal 2009 = 21961,5/4 - 102,5833 = 5387,792
3. Quartal 2009 = 21961,5/4 + 82,7084 = 5573,083
4. Quartal 2009 = 21961,5/4 + 112,7084 = 5603,083

Aufgabe IV.17.:

Die Deutschen sind nicht nur empört über die Entwicklung der Benzinpreise, auch Bier wird immer teurer. Für München liegen Daten zum Bierpreis jeweils für April und Oktober (Oktoberfest) vor:

Jahr	1999		2000		2001		2002		2003	
Halbjahr	Apr	Okt	Apr	Okt	Apr	Okt	Apr	Okt	Apr	Okt
Bierpreis in EUR/Liter	5,70	6,00	6,00	6,40	6,20	6,70	6,50	6,80	6,70	7,00

a) Stellen Sie die Entwicklung des Bierpreises grafisch dar und begründen Sie, weshalb in diesem Fall ein additives Zeitreihenmodell geeignet ist.

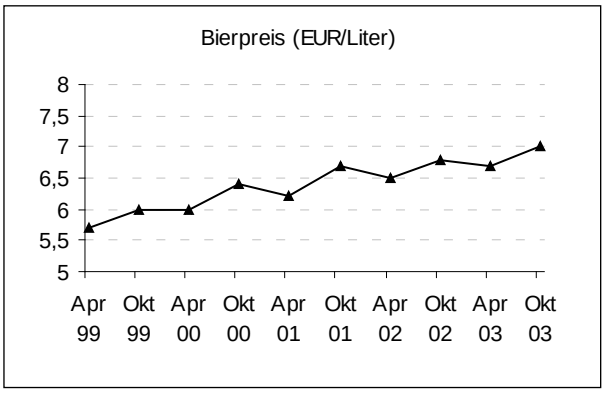
b) Ziel ist es, den Bierpreis für 2004 zu prognostizieren. Warum macht es Sinn, den Trendverlauf nach der Methode der kleinsten Quadrate zu bestimmen? Berechnen sie daher die Koeffizienten einer linearen Trendfunktion und interpretieren Sie diese.

c) Ermitteln Sie mit Hilfe des Phasendurchschnittsverfahren die Saisonindizes und interpretieren Sie diese.

d) Schätzen Sie abschließend den zu erwartenden Bierpreis auf dem Oktoberfest 2004.

Lösung:

a)



konstante Schwankungsbreite, daher additives Zeitreihenmodell

b) da keine Trend- bzw. Strukturbrüche vorliegen, ist der Einsatz einer Trendfunktion wegen deren Prognosetauglichkeit sinnvoll

Datum	**gesamt (y)**	**t**	**t * t**	**t * y**
Apr 99	5,70	-4,5	20,25	-25,7
Okt 99	6,00	-3,5	12,25	-21,0
Apr 00	6,00	-2,5	6,25	-15,0
Okt 00	6,40	-1,5	2,25	-9,6
Apr 01	6,20	-0,5	0,25	-3,1
Okt 01	6,70	0,5	0,25	3,4
Apr 02	6,50	1,5	2,25	9,8
Okt 02	6,80	2,5	6,25	17,0
Apr 03	6,70	3,5	12,25	23,5
Okt 03	7,00	4,5	20,25	31,5
Summe	64	0	82,50	10,70

Lösung Aufgabe IV.17. (Fortsetzung):

b)

a_1 = 10,70 / 82,50 = 0,130 [EUR / Halbjahr]
d.h. der Bierpreis steigt pro Halbjahr um durchschnittlich 0,130 EUR;

a_0 = 64,0 / 10 = 6,40 [EUR]
d.h. der Bierpreis beträgt zum Zeitpunkt t=0 = Ende Juli 2001 6,40 EUR;

c)

				$y^*_{ik} = y_{ik} - y^T_{ik}$	
y	**y_{ik}**	**t**	**$y^T_{ik} = y(t) = 6{,}4 + 0{,}130 \cdot t$**	Apr	Okt
y_{11} (= $y_{Apr1999}$)	5,70	-4,5	5,815	-0,115	
y_{12} (= $y_{Okt1999}$)	6,00	-3,5	5,945		0,055
y_{21} (= $y_{Apr2000}$)	6,00	-2,5	6,075	-0,075	
y_{22} (= $y_{Okt2000}$)	6,40	-1,5	6,205		0,195
y_{31} (= $y_{Apr2001}$)	6,20	-0,5	6,335	-0,135	
y_{32} (= $y_{Okt2001}$)	6,70	0,5	6,465		0,235
y_{41} (= $y_{Apr2002}$)	6,50	1,5	6,595	-0,095	
y_{42} (= $y_{Okt2002}$)	6,80	2,5	6,725		0,075
y_{51} (= $y_{Apr2003}$)	6,70	3,5	6,855	-0,155	
y_{52} (= $y_{Okt2003}$)	7,00	4,5	6,985		0,015

Phasendurchschnitte $\overline{y^*_{.k}}$	-0,115	0,115
Gesamtdurchschnitt $\overline{y^*}$	0,0000	
Saisonindex $I_k = \overline{y^*_{.k}} - \overline{y^*}$	-0,1150	0,1150
ΣI_k	0,0000	

Im April liegt der Bierpreis um durchschnittlich 0,115 EUR unter dem jeweiligen Trendwert und umgekehrt im Oktober (zum Oktoberfest) um genau diesen Betrag darüber.

d) $\hat{y}$(Okt 2004) = y^T(Okt 2004) +I(Okt) =
$\hat{y}^T_{62} + I_2 = \hat{y}(t=6{,}5) + I_2 =$
$6{,}40 + 0{,}13 \cdot 6{,}5 + 0{,}115 = 7{,}36$ EUR

Aufgabe IV.18.:

In einer Tourismusregion wurden folgende Übernachtungszahlen registriert:

Periode Jahr	Phasen I. Quartal	II. Quartal	III. Quartal	IV. Quartal	gesamt
2012	1089	1631	2498	1304	6522
2013	1161	1747	2671	1389	6968
2014	1201	1781	2734	1428	7144
2015	1298	1956	2985	1547	7786
Summe	4749	7115	10888	5668	28420

a) Stellen Sie diese Zeitreihe graphisch dar!

b) Berechnen Sie die gleitenden Durchschnitte aus einer geeigneten Anzahl von Gliedern.

c) Bestimmen Sie die trendbereinigten Werte (unterstellen Sie ein multiplikatives Modell).

d) Ermitteln Sie die Saisonindizes. Welche Erkenntnis können Sie hieraus ableiten?

Lösung:

Periode	**Phase**		y_{ik}^T	$y_{ik}^* = y_{ik} / y_{ik}^T$			
i	**k**	y_{ik}	**D(4)**	1. Quartal	2. Quartal	3. Quartal	4. Quartal
2012	I. Q	1089					
	II. Q	1631					
	III. Q	2498	1639,500			1,5236	
	IV. Q	1304	1663,000				0,7841
2013	I. Q	1161	1699,125	0,6833			
	II. Q	1747	1731,375		1,0090		
	III. Q	2671	1747,000			1,5289	
	IV. Q	1389	1756,250				0,7909
2014	I. Q	1201	1768,375	0,6792			
	II. Q	1781	1781,125		0,9999		
	III. Q	2734	1798,125			1,5205	
	IV. Q	1428	1832,125				0,7794
2015	I. Q	1298	1885,375	0,6885			
	II. Q	1956	1931,625		1,0126		
	III. Q	2985					
	IV. Q	1547					
Phasendurchschnitte $\overline{y_{.k}^*}$				0,6837	1,0072	1,5243	0,7848
Gesamtdurchschnitt $\overline{y^*}$				1,000			
Saisonindizes I_k				0,6837	1,0072	1,5243	0,7848
Normierung Phasendurchschnitte $\overline{y_{.k}^*} / \overline{y^*}$							
Saisonindizes ΣI_k				4,000			

d.h. eindeutiger touristischer Schwerpunkt im 3. Quartal, also in den Monaten Juli bis September, mit rund 50% über den Trendwerten

Aufgabe IV.19.:

Über die Entwicklung der Weltbevölkerung liegen Ihnen folgende Informationen der Deutschen Stiftung Weltbevölkerung vor:

Jahr	1950	1960	1970	1980	1990	1994	1998	2000
Weltbevölkerung in Mrd.	2,5	3,0	3,7	4,4	5,3	5,6	5,9	6,1

a) Berechnen Sie die Koeffizienten einer linearen Trendfunktion und stellen Sie diese im Vergleich zur tatsächlichen Entwicklung grafisch dar.
($\hat{y}(t) = -1{,}357 + 0{,}074 \cdot t$)

b) Interpretieren Sie die Koeffizienten der linearen Trendfunktion. Welche Gründe sprechen gegen eine lineare Trendfunktion zur Schätzung der Weltbevölkerung?
(siehe a_0, d.h. negative Schätzwerte bei kleinem t)

c) Unterstellen Sie alternativ einen exponentiellen Funktionsverlauf und zeichnen Sie auch diese Funktion in die Grafik unter a) ein. Um wie viel Prozent wächst die Weltbevölkerung nach diesem Modell pro Jahr?
($\hat{y}(t) = 1{,}036656 \cdot e^{0{,}0179 \cdot t}$; siehe a_1, d.h. konstante Wachstumsrate von 1,79%)

d) Wie groß wäre die Weltbevölkerung im Jahr 2100 nach dem linearen bzw. exponentiellen Modell?
($\hat{y}(200)_{lin} = 13{,}396$ Mrd.; $\hat{y}(200)_{exp} = 37{,}189$ Mrd.; wobei Jahr 1900 = t = 0)

Aufgabe IV.20.:

Ein gemeinnütziger Verein erhält Zuwendungen aus der Glücksspirale. Für die Jahre 2001 bis 2008 liegen hierzu folgende Informationen vor:

Jahr	2001	2002	2003	2004	2005	2006	2007	2008
Zuwendungen in TEUR	7418	6650	5874	6337	6720	5092	4690	5000

a) Stellen Sie die Entwicklung der Zuwendungen in geeigneter Weise grafisch dar.

b) Bestimmen Sie hierzu eine lineare Trendfunktion. Interpretieren Sie deren Koeffizienten und zeichnen Sie diese in die Grafik unter a) ein. Prognostizieren Sie mit Hilfe der Trendfunktion die Zuwendungen für das Jahr 2010.
($a_0 = 5972{,}625$ [TEUR]; $a_1 = -341{,}536$ [TEUR / Jahr]; $\hat{y}(5{,}5) = 4094{,}179$ [TEUR])

c) Überprüfen Sie, ob die Funktion $\hat{y} = (a_0 \cdot t) / (a_1 + t)$ mit a_0=4500 und a_1=-0,9 eine bessere Anpassung an die Originaldaten liefert. Codieren Sie dazu das Jahr 2001 mit t=2. Welche Zuwendungen wären nach dieser Funktion für 2010 zu erwarten? Zu welcher der beiden Prognosen hätten Sie größeres Vertrauen?
($\hat{y}(11) = 4900{,}990$ [TEUR])

Aufgaben und Lösungen

Aufgabe IV.21.:

In der folgenden Tabelle ist für die Jahre 2000 bis 2013 die Umsatzentwicklung (in Mio. EUR) für DVD - Videos in Deutschland angegeben:

Jahr	2000	2002	2004	2006	2008	2010	2011	2012	2013
Umsatz DVD	210	560	910	1150	1273	1200	1100	1000	950

a) Stellen Sie die Verkaufszahlen grafisch dar und prognostizieren Sie die Verkaufszahlen bis Ende 2018 mit Hilfe einer linearen Trendfunktion.

b) Was würden Sie gegen diese Vorgehensweise einwenden und welchen Modellansatz schlagen Sie demzufolge alternativ vor? Berechnen Sie diesen.

c) Berechnen Sie auf der Basis der Trendfunktion unter b), zu welchem Zeitpunkt DVDs in Deutschland erstmals auf den Markt gebracht wurden und wann mit einer vollständigen Verdrängung durch neuere Datenträger zu rechnen ist.

Aufgabe IV.22:

In der IT- und Unterhaltungsbranche kämpfen Apple und Samsung um die Marktführerschaft. Für die Jahre 2010 bis 2016 weisen die beiden Unternehmen folgende Gewinnentwicklung aus (Gewinn je Aktie in EUR):

Geschäftsjahr	Samsung Gewinn je Aktie	Apple Gewinn je Aktie
2010	71	1,68
2011	62	3,06
2012	109	4,83
2013	136	4,16
2014	116	5,36
2015	99	8,55
2016	124	7,93

Hinweis: Der Gewinn je Aktie errechnet sich aus dem Gesamtgewinn des Unternehmens dividiert durch die Anzahl aller ausgegebenen Aktien und ist bereits in Euro umgerechnet.

a) Existiert eine starke Abhängigkeit in der Gewinnentwicklung je Aktie zwischen beiden Unternehmen? Berechnen Sie hierzu eine geeignete Kennzahl und interpretieren Sie das Ergebnis.

b) Berechnen Sie für beide Unternehmen jeweils eine geeignete lineare Trendfunktion und schätzen Sie hiermit den Gewinn je Aktie für das Geschäftsjahr 2017.

c) Welche Unterschiede in der Gewinnsituation der beiden Unternehmen lassen sich aus den Trendfunktionen unter b) ableiten?

d) Lassen sich die Gewinnschätzungen in dem vorliegenden Beispiel durch den Einsatz des Phasendurchschnittsverfahrens noch weiter verbessern? Begründen Sie Ihre Antwort (keine Berechnung!).

Aufgabe IV.23:

Der Benzinpreis (EUR/Liter Super) weist in Deutschland in den letzten Jahren starke Schwankungen auf. Die Entwicklung der Benzinpreise soll daher in der nachfolgenden Tabelle mit der Entwicklung der Ölpreise (USD/Liter) und der Entwicklung der PKW-Neuzulassungen in Deutschland (in Mio.) verglichen werden:

Jahr	2008	2009	2010	2011	2012	2013	2014	2015
Benzinpreis	1,11	1,33	1,34	1,66	1,55	1,45	1,40	1,37
Ölpreis	0,22	0,50	0,60	0,72	0,69	0,66	0,38	0,28
PKW-Neuzulassungen	2,8	2,9	2,8	3,2	3,1	3,0	3,1	3,2

mit:

$\sum benzpr = 11{,}210$ [EUR/L]
$\sum ölpr = 4{,}050$ [USD/L]
$\sum neuzu = 24{,}100$ [Mio.]

$\sum benzpr^2 = 15{,}896$ [EUR/L]^2
$\sum ölpr^2 = 2{,}310$ [USD/L]^2
$\sum neuzu^2 = 72{,}790$ [Mio.]^2

benzpr = Benzinpreis in Euro pro Liter
ölpr = Ölpreis in US-Dollar pro Liter
neuzu = PKW-Neuzulassungen in Deutschland in Millionen Fahrzeuge

a) Stellen Sie die Abhängigkeit des Benzinpreises vom Ölpreis grafisch dar.

b) Berechnen Sie für die Abhängigkeit des Benzinpreises vom Ölpreis eine lineare Regressionsfunktion und zeichnen Sie diese in die Grafik unter a) ein.

c) Um den Benzinpreis Ende 2016 über die Regressionsfunktion aus b) schätzen zu können, benötigen Sie den Wert des Ölpreises für 2016. Dieser Wert soll nun über eine Trendfunktion ermittelt werden.
Bestimmen Sie eine geeignete Trendfunktion und berechnen Sie den Schätzwert für den Ölpreis 2016. Auf welchem Niveau wird sich dann Ende 2016 der Benzinpreis bewegen?

d) Begründen Sie, weshalb es sinnvoll wäre, den Regressionsansatz aus b) um die PKW-Neuzulassungen als zweiten Regressor zu erweitern.

e) Berechnen Sie nun die Koeffizienten dieses erweiterten linearen Regressionsmodells. Wie hoch hätten Sie nach diesem Modell den Benzinpreis Ende 2015 geschätzt?

Aufgabe IV.24.:

Die Kursentwicklung des Deutschen Aktienindex (DAX) zeigt für die letzten 24 Quartalsmonate folgenden Verlauf (Angaben in 1000 Punkten):

Jahr	Mär	Jun	Sep	Dez
2011	7,0	7,3	6,0	6,0
2012	7,0	6,0	7,3	7,6
2013	7,6	8,0	8,6	9,6
2014	9,6	9,8	9,5	9,8
2015	11,9	10,9	9,6	10,7
2016	10,0	9,8	10,5	11,5

a) Erfassen Sie die Aktienkurse als diskretes Merkmal in einer Häufigkeitsverteilung.

b) Berechnen Sie für die DAX-Kurse einen geeigneten Lage- und Streuungsparameter. Welche Erkenntnisse über die Verteilung des Merkmals lassen sich hieraus ableiten?

c) Stellen Sie die Zeitreihe grafisch dar. Was spricht für ein additives Zeitreihenmodell?

d) Bestimmen Sie den Trendverlauf nach der Methode geeigneter gleitender Durchschnitte. Begründen Sie, warum diese hier einer linearen Trendfunktion vorzuziehen sind.

e) Berechnen Sie unter Verwendung der gleitenden Durchschnitte die Saisonindizes. Ist danach die Börsenregel zutreffend, dass Sie grundsätzlich Ende April Ihre Aktien verkaufen und erst wieder im Oktober an der Börse einsteigen sollten?
Begründen Sie Ihre Antwort.

f) Erstellen Sie auf der Grundlage der gegebenen Daten eine Kursprognose für den DAX-Index zum Jahresende 2017.