4 DESKRIPTIVE STATISTIK UND STATISTISCHE AUSWERTUNG

Reinhold S. Jäger

Unsere Welt ist an Zahlen und Bewertungen orientiert. Diese Tatsache ist nicht nur an eher sarkastischen Bemerkungen über die Olympischen Spiele auszumachen (schneller, höher, weiter), sondern auch an der Tatsache, dass in vielen Bereichen des Alltags sowohl *Bestandsaufnahmen* durchgeführt als auch *Prognosen* abgegeben werden. Beispiele hierfür sind:

- die Ausstattung einer Stadt mit Kindertagesstätten, Räumen und Spielmaterialien,
- die Verkehrszählungen an bestimmten Stellen,
- Bevölkerungszählungen und Bevölkerungsprognosen etc.

Die jeweils gewonnenen Daten werden anschließend weiter verarbeitet. Diese Verarbeitung erfolgt in aller Regel unter Zuhilfenahme der *Statistik* sowie *grafischer Darstellungen*, die ihrerseits der Illustration dienen.

Die Statistik leistet dabei zwei Beiträge:

Deskriptive Statistik

- Es werden Daten mit Hilfe von (statistischen) Kennwerten beschrieben; man spricht hierbei von der *deskriptiven* oder *beschreibenden Statistik*;

Schlussfolgernde Statistik

- es werden Daten darauf hin beurteilt, ob sie bestimmten Erwartungen oder Hypothesen entsprechen; hierbei spricht man von der *schlussfolgernden Statistik*.

Nachfolgend wird lediglich auf die deskriptive Statistik abgehoben. Sie wird in ihren gebräuchlichsten grafischen Darstellungen und Kennziffern erläutert. Im einzelnen wird auf die folgenden Sachverhalte näher eingegangen:

- *Maße zur Charakterisierung von Verteilungsformen*: Daten darüber, welche Form eine Verteilung hat.
- *Maße der zentralen Tendenz*: Hierzu gehören Informationen über solche Messwerte, die eine Tendenz auf der Basis aller Personen berücksichtigen und eine gewissermaßen durchschnittliche Angabe erlauben.
- *Maße der Variabilität*: Informationen darüber, wie die Daten um einen zentralen Wert streuen.

- *Maße zur Beschreibung von Zusammenhängen*: Informationen über den Zusammenhang zwischen zwei Merkmalen. Man spricht hier auch von *Korrelationen*.

4.1 Zwei einführende Beispiele

Beispiel 1: Sie interessiert, welche Jugendlichen eine Freizeiteinrichtung Ihrer Stadt aufsuchen. Hierzu haben sie einen Kurzfragebogen entwickelt, den jeder Jugendliche nur einmal bei einem Besuch der Freizeiteinrichtung ausfüllen soll. Im Bogen werden unter anderem folgende Fragen gestellt (s. u.):

??? Jahr in neuer Zeile?

Besucher der Freizeiteinrichtung A der Stadt X

Wie alt sind Sie?

_____ Jahre

Ihr Geschlecht: □ männlich □ weiblich

Diese Daten werden von Ihnen anschließend ausgewertet. Eine einfache Auswertung erfolgt bereits dadurch, dass eine Strichliste hergestellt wird. In diesem Falle wird für jedes Alter eine Zählliste angefertigt (s. Tab. 4.1):

Tab. 4.1: Einfache Zählliste: Verteilung der Besucher der Freizeiteinrichtung A nach Alter

Alter	Anzahl	Zahl
12	𝍸	5
13	|||	3
14	\\\|	3
15	𝍸\	6
16	|||	3
17	𝍸||	8
18	𝍸||	5
19	/	1
20	||	2

Diese Zählliste hat bereits einen gewissen Informationswert. So kann beispielsweise aus Tabelle 4.1 erschlossen werden, dass die am häufigsten vertretene Gruppe die der 17-jährigen ist. Entsprechend könnte man verfahren, um zugleich auch das Geschlecht zu erfassen. Aus der Kombination von Alter und Geschlecht entsteht dann folgende Tabelle 4.2.

Es ist leicht nachvollziehbar, dass Tabelle 4.2 einen höheren Informationswert besitzt als Tabelle 4.1. Aber es fällt zunächst schwer, eine *zusammenfassende Aussage* über die beiden Variablen Alter und Geschlecht zu machen. Diese Aussage kann mit Hilfe bestimmter *grafischer Darstellungsformen* und *statistischer Kennziffern* getroffen werden, die in den Kapiteln 4.2ff dargestellt werden.

Tab. 4.2: Verteilung der Befragten nach Alter und Gechlecht

Alter	männlich	weiblich	Gesamt
12	3	2	5
13	3	0	3
14	2	1	3
15	3	3	6
16	3	0	3
17	4	4	8
18	2	3	5
19	1	0	1
20	0	2	2

Beispiel 2: Sie interessieren sich für das Freizeitverhalten von Jugendlichen und führen hierzu eine Befragung durch. Die Befragung wird anhand des in Kapitel 3 entwickelten Fragebogens realisiert. Die Angaben von 16 Jugendlichen sind diesem Buch als farbige Beilage beigefügt.

Dieses Beispiel 2 wird nachfolgend immer wieder aufgegriffen werden, um an ihm zu demonstrieren, wie man vorgehen kann, um eine sachgerechte Auswertung von vorliegenden Daten durchzuführen. Die Originaldaten sind in Tabelle 4.3 dargestellt.

Hierbei werden im Einzelnen erfasst:

- Klassenstufe
- Schultyp
- Geschlecht
- Zeugnisnoten in vier Fächern (vier Fächer)
- Beliebtheit von vorgegebenen Freizeittätigkeiten (neun Bereiche)
- Taschengeld pro Woche

Die angegebenen Bereiche werden im Fragebogen durch ein Kreuz beantwortet bzw. es wird eine Zahl angegeben. Die Zahl drückt aus, welche Note in einem Fach erreicht wurde bzw. wie groß die Summe an Taschengeld ist, die pro Woche zur Verfügung steht. Sofern durch ein Kreuz angegeben wird, welcher Sachverhalt zutrifft, kann die betreffende Information entsprechend in eine Zahl umgesetzt werden.

Die resultierenden Zahlen haben aber jeweils einen unterschiedlichen Informationswert. So sei beispielhaft auf Folgendes hingewiesen:

Informationswert von Zahlen

- Der Bereich Schulart (abgekürzt als Schule in Tabelle 4.3) wird in der Tabelle 4.3 wie folgt in Zahlen ausgedrückt: Hauptschule = 1, Realschule = 2, Gymnasium = 3, Gesamtschule = 4. Diese Zahlen deuten lediglich darauf hin, dass die einzelnen Schulen zu unterscheiden sind, nicht aber, dass auf der Grundlage der jeweiligen Zahl auf eine Wertigkeit der Schule geschlossen werden kann. Es ist demnach auch nicht vorstellbar, dass zwischen den Zahlen Übergänge bestehen. Sie sind also als nicht-kontinuierlich verteilte (=diskrete) Werte anzusehen. Gleiches trifft auch auf das Geschlecht zu (abgekürzt in Tabelle 4.3 durch *Gesch*).

Diskrete Werte

Stetige Werte

- Bei den Zeugnisnoten ist dieser Sachverhalt schon anders zu bewerten: Die angegebenen Zahlen sind jeweils auf- oder abgerundet. Denn die Praxis in der Schule ist durchaus die, dass Zwischenwerte vergeben werden. Die Noten gelten daher als annähernd kontinuierlich verteilt, auch wenn die Angaben in Zahlen erfolgen, die ihrerseits *diskret* angegeben sind. Gleiches trifft auch auf die Angaben über die Freizeittätigkeiten zu, durch die Jugendlichen angegeben haben, wie gerne sie die jeweilige Tätigkeit ausüben. Wären die Werte kontinuierlich verteilt, so spräche man auch von stetigen Variablen.

Fehlende Werte

Alle erhobenen Daten zu den genannten Bereichen sind in Tabelle 4.3 wiedergegeben. Für den Fall, dass ein Jugendlicher eine Angabe nicht gemacht hat, ist dies ersichtlich. So hat der Jugendliche mit der Nr. 4 keine Mathematiknote angegeben. Folglich findet sich in Tabelle 4.3 auch keine Eintragung. Weil dieser Wert fehlt, spricht man auch von *missing datum*. Ein fehlender Wert muss bei der Bestimmung von statistischen Kennwerten in besonderer Weise berücksichtigt werden.

Variable

Alle Werte in Tabelle 4.3 entsprechen demnach den Angaben der Jugendlichen aus der Befragung. Diese Tabelle ist dabei so aufgebaut, dass in den Spalten die Informationen über die Bereiche wiedergegeben sind, man spricht auch hier von *Variablen*. Diese Variablen werden durch Kürzel gekennzeichnet. In jeder Zeile der Tabelle 4.3 findet sich der vollständige Satz von Variablen mit den Angaben aus dem Fragebogen jeweils eines Jugendlichen. Um die Jugendlichen

voneinander zu unterscheiden, wurden laufende Nummern vergeben. Sie sind in Tabelle 4.3 durch das Kürzel *NR* und eine fortlaufende Zahl zu unterscheiden und aus dem beigelegten Blatt zu entnehmen.

Was lässt sich auf der Basis des Beispiels 2 und mit Blick auf die unterschiedlichen Variablen nun statistisch aussagen und wie können verschiedene Angaben grafisch veranschaulicht werden? Diesen beiden Fragen wird anschließend nachgegangen.

Statistik wird dafür verwendet, gewonnene Daten zu beschreiben. Man spricht deshalb auch von der deskriptiven Statistik. Ausgangspunkt von Statistik sind quantitative Daten. Sie beschreiben einzelne Personen in bestimmten Charakteristika. Diese Charakteristika werden als Variablen bezeichnet.

Ausgangspunkt der Bestimmung von statistischen Kennwerten ist eine Tabelle, in der die erhobenen Variablen mit ihren quantitativen Ausprägungen für alle Personen angeführt werden. Eine solche Tabelle wird auch als Matrix bezeichnet.

Tab. 4.3: Daten, die Beispiel 2 zugrunde liegen

Nr	Klasse	Schule	Gesch	Note_Dt	Note_Ma	Note_En	Note_Re	Int_Jug	Int_Kino	Int_Musi	Int_Fern	Int_Rum	Int_Jh	Int_Disk	Int_Spor	Int_Comp	Geld
1	9	3	1	2	3	2	1	3	1	1	2	4	2	4	2	1	20,00
2	8	2	1	3	3	2	2	2	2	2	3	3	4	3	3	5	15,00
3	8	3	2	1	1	2	1	4	3	1	4	5	2	5	4	4	16,00
4	9	1	1	4		4	3	4	2	2	2	2	3	3	4	1	16,00
5	7	3	2	1	3	1	2	3	2	2	4	4	2	2	1	4	12,00
6	8	3	1	3	1	3	2	3	2	1	3	3	2	3	1	2	17,00
7	8	1	1	4	3	5	4	4	1	1	2	2	2	3	5	4	12,50
8	8	4	2	2	4	2	1	2	4	2	3	4	3	1	2	5	15,00
9	8	3	2	2	4	2	2	3	3	2	3	4	5	1	2	5	18,00
10	5	2	1	3	2	3	4	3	4	3	1	1	4	5	4	1	10,00
11	8	2	2	1	2	1	1	1	3	1	4	4	5	4	4	4	15,00
12	10	2	2	3	1	2	1	3	2	1	4	4	4	2	4	5	50,00
13	7	3	2	2	2	2	3	2	2	3	2	3	4	2	2	3	12,00
14	9	2	2	1	5	3	1	2	2	2	4	3	3	5	5	5	17,00
15	9	1	2	4	3	5	3	5	4	5	5	5	2	2	2	3	0,00
16	8	3	2	1	2	1	1	2	1	2	1	4	2	2	5	2	35,00

4.2 Darstellung von Daten

4.2.1 Tabellarische Darstellung von Daten

Gehen wir von einem Beispiel aus: Wir gehen der Frage nach, welche Mathematiknote wie viele Jugendliche erreicht haben. Es resultiert hieraus die folgende Tabelle 4.4.

Tab. 4.4: Verteilung der Mathematiknote (Note_Ma) bei den 16 Beispiel-Jugendlichen

Note	Häufig-keit	kum. Häuf.	% der validen An-gaben	kum. % der valid. Angaben	% aller Fälle	kum. % aller Fälle
1	3	3	20,00	20,00	18,75	18,75
2	4	7	26,66	46,66	25,00	43,75
3	5	12	33,33	80,00	31,25	75,00
4	2	14	13,33	93,33	12,50	87,50
5	1	15	6,66	100,00	6,25	93,75
missing	1	16			6,25	100,00

Aus ihr ist zu ersehen, dass es einen Jugendlichen gibt, der keine Angabe gemacht hat. Bei der Überprüfung der Tabelle 4.3 ist zu entnehmen, dass es sich dabei um die Person mit der Nr. 4 handelt. Dieser Sachverhalt ist in der Zeile *„missing“* mitgeteilt.

Die Verteilung der Noten aller 16 Jugendlichen ist aus der Spalte *Häufigkeit* zu entnehmen.

Kumulation

Die Überprüfung in Tabelle 4.4. ergibt, dass alle 16 Jugendlichen der Befragung berücksichtigt wurden, denn in der Spalte mit der Bezeichnung *kum.* (= kumuliert, d. h. hier werden alle Werte über die Zeilen hinweg summiert) ergibt sich als Prüfwert 16. Das entspricht der Anzahl aller Befragten.

Der Begriff Kumulation besagt, dass die Werte über die verschiedenen Stufen addiert, also summiert werden. Zur Veranschaulichung gehen wir nochmals in Tabelle 4.4 zurück.

Hieraus ist Folgendes zu ersehen:

In der Spalte *Häufigkeit* ist die jeweilige Anzahl von Personen zu entnehmen, die eine bestimmte Note in Mathematik erreicht haben. Die Note 1 wurde von 3 Personen angegeben, die Note 2 von 4 Personen usw. Wird nunmehr kumuliert, so wird die Anzahl jeweils addiert. Das Beispiel aus Tabelle 4.4 erläutert den Zusammenhang:

Tab. 4.5: Berechnung von kumulierten Häufigkeiten

Note	Häufigkeit	kum. Häufigkeit
1	3	3
2	4	3 + 4 = 7
3	5	3 + 4 + 5 = 12
4	2	3 + 4 + 5 + 2 = 14
5	1	3 + 4 + 5 + 2 + 1 = 15
missing	1	3 + 4 + 5 + 2 + 1 + 1 = 16

Worin liegt nun der Vorteil von kumulierten Werten? Mit den kumulierenden Werten lässt sich darstellen, wie viele Personen bei einer gegebenen Anzahl N bis zu einem bestimmten Wert „angesiedelt" sind. So kann man aus dem vorangegangenen Beispiel erkennen, dass 14 von insgesamt 15 die Note 4 und besser im Fach Mathematik erreicht haben.

Eine ähnliche Interpretation der Daten ist in anderen Spalten der Tabelle 4.4 gegeben: Die Spalte mit der Bezeichnung *kum. % aller Fälle* zeigt dann an, dass die Noten 1 bis 3 von 75% aller Personen erreicht wurden, die befragt worden sind. Es werden demnach in dieser Spalte *kumulierte* Häufigkeiten angegeben.

Fehlende Messwerte

Nunmehr muss man bei der weiteren Auswertung berücksichtigen, dass eine Angabe fehlt. Berechnet man daher den Prozentanteil derjenigen Personen, die eine bestimmte Mathematiknote erreicht haben, so ist jetzt nur noch die Anzahl N = 15 Personen zu berücksichtigen. N = 15 Personen entspricht den validen Angaben. In der Spalte *% der valid. Angaben* ist daher berücksichtigt, wie viel Prozent der Personen bei der jeweiligen Note überhaupt eine Angabe gemacht haben. Bei der Note 4 entspricht dies einem Prozentanteil von 13,33. Keine Angabe (= missing) entspricht einem Prozentanteil von 6,66% aller Personen.

Kumuliert man aber nunmehr über alle Noten, so muss der Jugendliche Nr. 4 unberücksichtigt bleiben, die Spalte *kum. % der valid. Angaben* endet daher mit 100% bei der Note 5. Analog wird in den anderen Spalten verfahren.

Übung 4.1

Stellen Sie Verteilungstabellen für die folgenden Variablen zusammen:

a) Schule (Schule)

b) Note in Religion (Not_Re)

c) Klassenstufe (Klasse)

Die Lösungen finden Sie im Anhang.

4.2.2 Grafische Darstellung von Daten

Die grafische Darstellung von Daten wird gewählt, um bestimmte Sachverhalte zu illustrieren. Es versteht sich von selbst, dass eine solche Darstellung auf das Datenniveau (vgl. Kapitel 3) Rücksicht nehmen muss. Wir werden daher in den nachfolgenden Ausführungen diesen Sachverhalt mit in Betracht ziehen.
Die beiden obigen Beispiele lassen erkennen, dass jede Darstellung in Tabellenform auch als *Grafik* präsentiert werden kann. Nachfolgend werden einige dieser Darstellungsformen beschrieben.

4.2.2.1 Säulendiagramm

Säulendiagramm

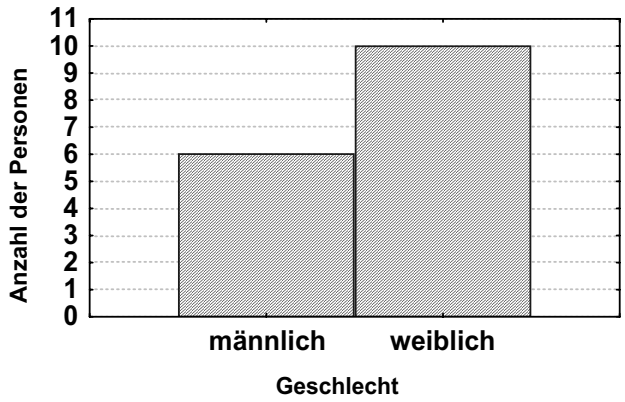
Geht man wiederum vom obigen Beispiel 2 aus, so kann man der Frage nachgehen, wie viele der Befragten männlichen und wie viele weiblichen Geschlechts sind. Es liegt zunächst nahe, die jeweilige Anzahl zu bestimmen und die Häufigkeit abzubilden. Hierzu bedient man sich eines *Säulendiagramms*.

Bei dieser Art der Darstellung wird folgendermaßen vorgegangen:

- Die Anzahl der männlichen Personen und die der weiblichen Personen wird bestimmt.
- Es wird eine Position auf einer X-Achse bestimmt, auf der Säulen etabliert werden. Deren Höhe ist mit der jeweiligen Anzahl von Personen identisch. Die Höhe der Säule wird auf der Y-Achse abgebildet.

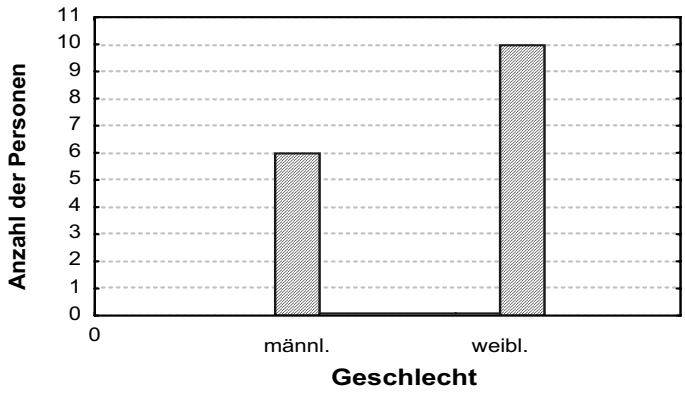
Die nachfolgende Abbildung 4.1 stellt ein Resultat der vorherigen Beschreibung dar. Es ist dabei aus der X-Achse (= *Geschlecht*) die Anordnung der Geschlechter zu entnehmen und aus der Y-Achse (= *Anzahl der Personen*) die entsprechende Anzahl von Jugendlichen des jeweiligen Geschlechts.

Abb. 4.1: Verteilung der Probanden nach Geschlechtszugehörigkeit



Die Tatsache, dass beide Klassen der Geschlechter nicht getrennt nebeneinander abgebildet sind, hat keine besondere Bedeutung. Es ist auch aus der Abbildung 4.1 nicht zu entnehmen, dass ein kontinuierlicher Übergang zwischen den Geschlechtern besteht (s. Abschnitt 4.1). Zwar hätte dieser Sachverhalt auch durch eine Art der Abbildung wie die nachfolgende (Abbildung 4.2) dargestellt werden können, doch ändert sich nichts am eigentlichen Wert der Information.

Abb. 4.2: Verteilung der Probanden nach Geschlechtszugehörigkeit

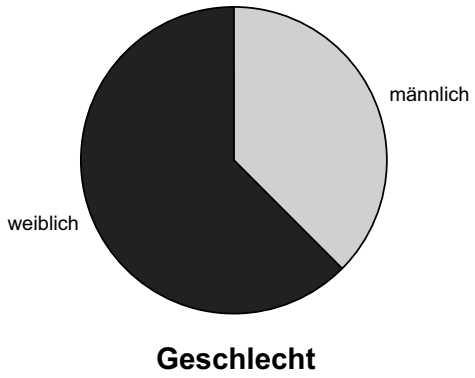


4.2.2.2 Kreisdiagramm

Kreisdiagramm

Die Ausgangsdaten aus Tabelle 4.3 können auch für eine andere Darstellungsform verwendet werden. Eine solche Darstellungsform ist in Abbildung 4.3 gewählt. Hierbei wird die Anzahl der Personen, die eine Angabe über ihr Geschlecht gemacht haben, in Form von *Kreissektoren* dargestellt. Die Größe des jeweiligen Kreissektors ermöglicht eine Aussage über den *relativen Anteil an der Gesamtanzahl von Personen*, die zur Verfügung stehen.

Abb. 4.3:
Verteilung der Personen nach Geschlecht



4.2.2.3 Verteilungskurve

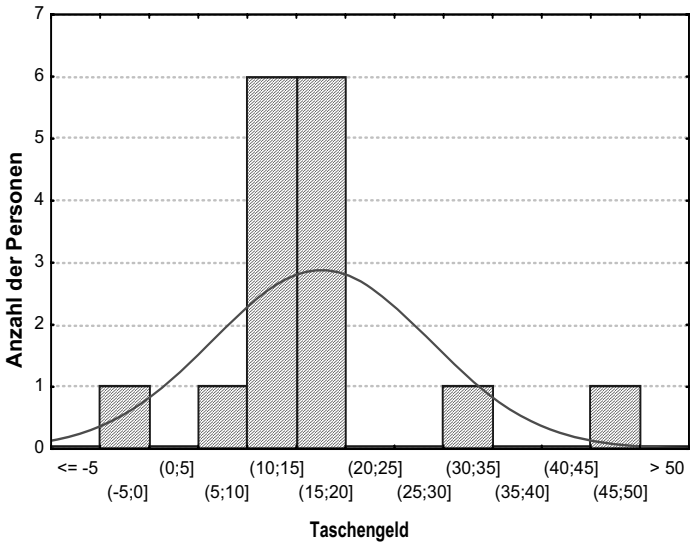
Bei vielen Untersuchungen interessiert nicht nur, welche konkrete Angabe eine Person gemacht hat, sondern wo sie bezüglich ihrer Angabe im Gesamt der quantitativen Angaben angesiedelt ist. Zur Erläuterung beziehen wir uns auf die Variable Taschengeld (abgekürzt in Tabelle 4.3 mit *Geld*).

In Abbildung 4.4 wird auf zwei Sachverhalte hingewiesen:

- Es ist wiederum ein Säulendiagramm wiedergegeben. Dabei sind, der Einfachheit halber, bestimmte Größenordnungen des Taschengelds zusammengefasst: Es wurde in Fünferschritten vorgegangen. Alle Jugendlichen im gleichen (Taschengeld-) Intervall – wie in Abbildung 4.4 angegeben – werden als gleich gezählt. Es entsteht dann eine Säulendiagramm der angegebenen Art.
- In das Säulendiagramm ist eine stetige Verteilung eingezeichnet. Diese Verteilungskurve entsteht aus der Überlegung, welche Verteilung zu erwarten wäre, wenn die Anzahl der Jugendlichen zugrunde gelegt und die angegebene Taschengeldhöhe berücksichtigt werden. Diese stetige Verteilung ist eine so genannte *Verteilungskurve*.

Auf der Basis dieser Verteilungskurve lässt sich dann nachfragen, ob der Erwartung nach die Anzahl der Jugendlichen mit einem Taschengeld von € 10-15 eher zu häufig vorkommt. Dies ist der Fall. Zugleich lässt sich auch lokalisieren, an welcher Stelle jeder Jugendliche mit seinem Taschengeld einzuordnen ist, ob er nämlich – gemessen am Durchschnitt – eher mehr oder weniger als andere Jugendliche erhält.

Abb. 4.4: Verteilungskurve: Variable Taschengeld



Zur Darstellung von Daten bedient man sich unterschiedlicher Zugangsweisen. Grafische Darstellungen dienen der Visualisierung, tabellarische der mehr quantitativen Repräsentation.

Aus einem Säulendiagramm lassen sich unmittelbar die Werte ablesen, so dass man weiß, wie viele Personen mit welcher Ausprägung Angaben gemacht haben. Ein Kreisdiagramm gibt in den meisten Fällen die Informationen in relativen Anteilen von Teilflächen an einer Gesamtfläche wieder.

Während das Säulendiagramm und das Kreisdiagramm auf der Basis von nichtkontinuierlichen Daten zustande kommen, werden bei der Verteilungskurve kontinuierliche Werte vorausgesetzt.

Übung 4.2

a) Zeichnen Sie ein Säulendiagramm, das die Verteilung der Jugendlichen über die Schulformen darstellt.

b) Zeichnen Sie ein Kreisdiagramm aus dem man den Anteil der Jugendlichen in Bezug zu ihren Englischnoten ablesen kann.

Die Lösungen finden Sie im Anhang.

4.3 Statistische Kennwerte

Mit den beschriebenen tabellarischen oder grafischen Darstellungsformen erhält man einen ersten Einblick in die mit Hilfe der verwendeten Erhebungsmethoden gewonnenen Daten. Darüber hinaus ist es insbesondere bei sehr umfangreichem Datenmaterial notwendig, gewissermaßen verdichtete Aussagen über einzelne Variablen und somit Antworten auf bestimmte Fragestellungen zu erhalten. Solche Antworten lassen sich mit statistischen Kennwerten finden.

Die statistischen Kennwerte werden – soweit es sich um solche handelt, die lediglich auf eine einzige Variable Bezug nehmen – als

a) Maße der zentralen Tendenz und

b) Variablitätsmaße

bezeichnet.

4.3.1 Maße der zentralen Tendenz

Maße der zentralen Tendenz

Sehr oft möchte man mit einem einzigen Wert eine Gruppe kennzeichnen. Beispiele für eine solche Vorgehensweise finden sich im Alltag sehr oft:

- Im Durchschnitt betrachtet entspricht der Intelligenzquotient der Bevölkerung der Bundesrepublik Deutschland einem Wert von 100.
- Das Durchschnittsalter der Kumpel im Ruhrgebiet betrug im Jahre 1997 38 Jahre.
- Jeder Arbeitsplatz bei der Kohleförderung in der BRD wurde früher mit durchschnittlich DM 130 000 pro Jahr subventioniert.

Was steckt hinter solchen Aussagen, wie werden diese statistisch gewonnen? Welche statistischen Kennziffern kommen in Frage, um entsprechende Aussagen zu treffen?

Nachfolgend werden einige wesentliche statistische Kennwerte der *zentralen Tendenz* dargestellt.

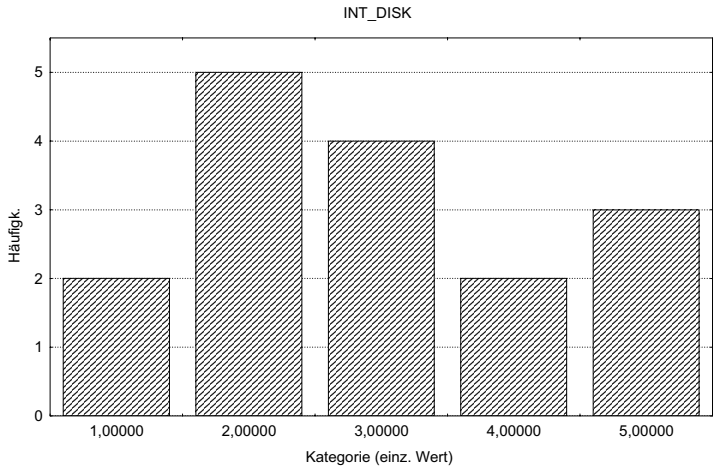
4.3.1.1 Der Modalwert

Modalwert

Der *Modalwert* (man spricht auch von Modus: $\bar{x}_{Mo}$) ist das am einfachsten zu bestimmende Maß der *zentralen Tendenz*. Es entspricht dem Wert, der am *häufigsten* gewählt wurde. Dieser Wert soll anhand des Beispielfragebogens veranschaulicht werden. Dort wurde (siehe Beilageblatt zu diesem Buch) im Zusammenhang mit dem Freizeitverhalten unter anderem danach gefragt, wie gerne die Jugendlichen *In die Diskothek gehen*.

Die Einschätzungen der Jugendlichen sind wiederum aus Tabelle 4.3 und dem Beilageblatt zu entnehmen. Eine grafische Darstellung dieser Einschätzungen ist aus Abbildung 4.5 zu entnehmen:

Abb. 4.5: Einschätzung des Freizeitverhaltens „in die Disko gehen“



Berechnung des Modalwerts

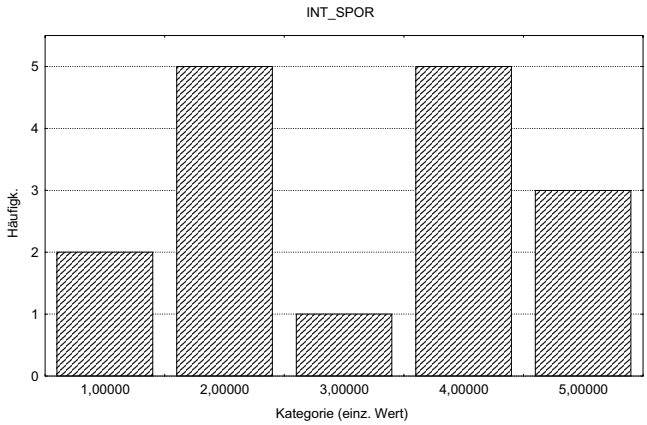
Aus Abbildung 4.5 ist zu ersehen, dass die Einschätzung 2 am häufigsten angekreuzt wurde. Damit ist 2 der entsprechende Modalwert hinsichtlich der Einschätzung des *Interesses, in der Freizeit Diskotheken zu besuchen.* Der Zahlenwert 2 ist zugleich eine Information über den *Grad des Interesses* an dieser Freizeitbetätigung.

Nicht immer ist die Bestimmung des Modalwertes so einfach wie im genannten Beispiel aus Abbildung 4.5. Hierzu soll ein weiteres Beispiel herangezogen werden.

Modalwert bei mehrgipfligen Verteilungen

Bei der Auswertung der Daten zum Freizeitverhalten *Sport treiben* wird deutlich, dass es zwei Abstufungen bei der Einschätzung gibt, nämlich die 2 und die 4, die aus der Häufigkeitsverteilung herausragen (siehe Abbildung 4.6). Eine solche Verteilung nennt man zweigipflig. Es wird durch dieses Beispiel deutlich, dass durchaus nicht nur ein einziger Modalwert existieren muss.

Abb. 4.6: Einschätzung des Freizeitverhalten „Sport treiben“



4.3.1.2 Der Median

Median

Ein weiterer statistischer Kennwert ist der Median ($= \bar{x}_{Me}$). Er ist derjenige Wert, oberhalb und unterhalb dessen genau die Hälfte der Personen angesiedelt ist. Oder anders ausgedrückt: 50% aller Personen haben einen niedrigeren Wert und 50% einen höheren Wert. Dieser Wert entspricht auch der kumulativen prozentualen Häufigkeit aller validen Angaben (vgl. Tabelle 4.4: *kum. % der valid. Angaben*).

Wiederum soll dieser Wert anhand eines Beispiels erläutert werden. Hierbei wird auf das *Interesse am Diskothekenbesuch* zurückgegriffen. Die Hälfte der Personenzahl (= 50%) entspräche genau 8 Personen, denn insgesamt wurden N = 16 Personen befragt.

In der Spaltenangabe von Tabelle 4.5 lässt sich nunmehr in zwei Spalten eine Angabe entnehmen:

Berechnung des Medians

- In der Spalte *kum. Häufig*. wird die Anzahl der Personen genannt, die bis zu einer bestimmten Einschätzung aufaddiert wurde. Die Anzahl 8 wird zwar nicht de facto erreicht, wohl aber die Anzahl 11. In diesem Bereich ist auch die 8. Person enthalten. Demnach entspricht die dazugehörige Einschätzung dem Median. Der Median ist deshalb $\bar{x}_{Me} = 3$.
- Ein anderer Zugang findet sich in der Spalte *kum. % aller Fälle*. Hierbei wird nach dem Wert 50% gesucht. Dieser Wert wird nicht erreicht, wohl aber ein Wert 68,75, der 50% einschließt. Bis zu diesem Wert haben demnach 68,75% aller Personen eine Angabe gemacht. Folgerichtig kann aus der gleichen Zeile auch der Median entnommen werden: Der hierzu gehörende Wert der Einschätzung ist wiederum 3.

Der Median ist deshalb $\bar{x}_{Me} = 3$.

Tab. 4.5: Einschätzung des Freizeitverhalten „Diskotheken besuchen“

Einschätzung	Häufigkeiten	kum. Häufig.	% der validen Angaben	kum. % aller Fälle
1	2	2	12,50	12,5
2	5	7	31,25	43,75
3	4	11	25,00	68,75
4	2	13	12,50	81,25
5	3	16	18,75	100,00
missing	0		0,00	

Halten wir fest: Der $\bar{x}_{Me}$ ist dann mit dem Wert 3 identisch, denn er vereinigt den Wert, unterhalb bzw. oberhalb dessen genau 50% der Personen mit ihren Einschätzungen angesiedelt sind.

Median und Extremwerte

Interessant ist auch, dass der $\bar{x}_{Me}$ gegenüber Extremwerten nicht anfällig ist. Ein solcher Extremwert läge vor, wenn eine ungewöhnliche Angabe gemacht würde. Ein Beispiel hierfür wäre beispielsweise ein Jugendlicher, der pro Monat 1 000 € zur Verfügung hätte, während alle anderen Angaben sich in einem Bereich bewegen würden, wie dies den Angaben in Tabelle 4.3 entspricht. Dies ist anders beim nächsten Maß der zentralen Tendenz, dem arithmetischen Mittel.

4.3.1.3 Arithmetisches Mittel

Arithmetisches Mittel

Die beiden bereits beschriebenen Werte zur Bestimmung der zentralen Tendenz deuten an, dass mit ihnen keine erschöpfende Charakterisierung einer Verteilung erzielt werden kann.

Das *arithmetische Mittel* ist das wohl bekannteste Maß der zentralen Tendenz. Es handelt sich dabei um den im Volksmund als *Durchschnitt* bezeichneten Kennwert. Das arithmetische Mittel wird abgekürzt durch $\bar{x}$ (gelesen als x-quer) gekennzeichnet.

Die Berechnung des arithmetischen Mittels soll wiederum am Beispiel des Diskothekenbesuchs erläutert werden.

Für jeden Jugendlichen aus der Befragung liegen Angaben zum Freizeitinteresse *am Diskothekenbesuch* vor. Die Berechnung ist denkbar einfach. Bei der Berechnung von $\bar{x}$ werden alle einzelnen Messwerte addiert und dann durch die Anzahl der validen Fälle geteilt. Konkret läßt sich dieser Sachverhalt an unserem Beispiel wie folgt nach vollziehen (vgl. Musterfragebogen und Tabelle 4.3):

Berechnung des arithmetischen Mittels

$$\bar{x} = \frac{4+3+5+3+2+3+3+1+1+5+4+2+2+5+2+2}{16} = 2{,}9$$

Es ergibt sich ein Durchschnittsinteresse am Diskothekenbesuch von 2,9.

Zusammenfassend lässt sich die Berechnung von $\bar{x}$ durch folgende Formel bestimmen:

Formel 1

$$\bar{x} = \frac{\sum_i^n x_i}{n} \quad \textit{oder} \quad \bar{x} = \frac{1}{n}\sum_i^n x_i$$

(wobei gilt: n = die Anzahl der validen Fälle, x_i = die einzelne Einschätzung des Jugendlichen i)

Berechnung des arithmetischen Mittels aus Häufigkeitstabellen

Das *arithmetische Mittel* lässt sich aber auch auf der Grundlage einer Häufigkeitstabelle berechnen. Für unser Beispiel müsste man dann Tabelle 4.5 zugrunde legen. In diesem Fall wäre ein anderer Berechnungsmodus notwendig, der in Formel 2 beschrieben ist. f_i entspricht dabei der Häufigkeit, mit der eine bestimmte Einschätzung genannt wurde.

Formel 2

$$\bar{x} = \frac{\sum_i^n f_i x_i}{n} \quad \textit{oder} \quad \bar{x} = \frac{1}{n}\sum_i^n f_i x_i$$

Die entsprechenden Angaben lassen sich aus Tabelle 4.5 entnehmen. Konkret bedeutet dies für das Beispiel:

$$\bar{x} = \frac{(2\times 1)+(5\times 2)+(4\times 3)+(2\times 4)+(3\times 5)}{16} = 2{,}9$$

Übung 4.3

Berechnen Sie Median, Modalwert und arithmetisches Mittel für die folgenden Variablen:

a) Interesse an Computern

b) Mathematiknote

Die Lösungen finden Sie im Anhang.

4.3.2 Maße der Variabilität

Auf der Basis der obigen Berechnungen und Darstellungen verschiedener Maße der zentralen Tendenz lässt sich für die Einschätzungen des Freizeitinteresses *Diskothekenbesuch* zunächst folgendes festhalten (siehe Tab. 4.6):

Tab. 4.6:
Gegenüberstellung von Maßen der zentralen Tendenz bei der Einschätzung des Freizeitinteresses „Diskobesuch“

Maße der zentralen Tendenz	Wert
$\bar{x}_{Mo}$	2
$\bar{x}_{Me}$	3
$\bar{x}$	2,9

Unsymmetrische Verteilung

Diese Gegenüberstellung macht an dieser Stelle nicht nur deutlich, dass die einzelnen Maße jeweils eine eigene Bedeutung haben, sondern auch dass sie darüber hinaus geeignet sind, zur Charakterisierung von Verteilungen (vgl. Abbildung 4.6) herangezogen zu werden. Je weiter diese einzelnen Maße nämlich voneinander abweichen, desto unsymmetrischer ist die Verteilung. Der Begriff unsymmetrisch entspricht dabei einer Verteilung, bei der die hohen und niedrigen Werte nicht gleich häufig auftreten.

Variabilität

Da die Angaben aus Tabelle 4.6 nicht nur etwas über eine Nicht-Symmetrie der Verteilung aussagen, sondern auch darüber, dass offensichtlich die einzelnen Einschätzungen der Jugendlichen untereinander abweichen, ist es daher auch wichtig, etwas über die Variabilität der Einschätzungen der Werte zu erfahren und entsprechende Kennwerte zu bestimmen.

Für eine hinreichende Charakterisierung einer Messwerteverteilung reicht es demnach nicht aus, die oben dargestellten Maße der zentralen Tendenz zu beschreiben. Diese „repräsentieren“ zwar die gesamten Messwerte der Gruppe, d. h. sie stehen stellvertretend für diese. Sie erlauben jedoch keinen Rückschluss auf die Verteilung der Messwerte in der Gesamtstichprobe und lassen somit Unterschiede außer ausser Acht, die zwischen den Werten der Personen der Gruppe bestehen. Variabilitätsmaße, wie die Spannweite, die Standardabweichung oder die Varianz, ermöglichen hingegen einen Rückschluss auf die Streuung oder Variabiltität der Werte der Stichprobe.

4.3.2.1 Die Spannweite

Spannweite

Die *Spannweite* (SP) ist ein Variabilitätsmaß, das direkt aus den Messwerten bestimmt werden kann. Es ist definiert als die Differenz zwischen dem Maximal- und dem Minimalwert und bildet somit ein sehr grobes Maß, da sie nur sehr wenig über die Variationsverhältnisse innerhalb der Verteilung aussagt.

Berechnung der Spannweite

Für die Berechnung der Spannweite gilt die folgende Formel:

Formel 3 $SP = x_{max} - x_{min}$

Aus dem Beispiel „Interesse an Diskothekenbesuchen“ (Tab. 4.5) resultiert demnach eine Spannweite von 4, die wie folgt berechnet wird:

Minimalwert = 1 Maximalwert = 5

$$SP = 5 - 1 = 4$$

4.3.2.2 Standardabweichung und Varianz

Die *Varianz* und die *Standardabweichung* beschreiben die mittlere Abweichung aller Messwerte vom Mittelwert der Stichprobe, allerdings in je unterschiedlicher Art und Weise.

Zur Berechnung der Standardabweichung und der Varianz wird wie folgt vorgegangen:

Bildet man die Differenz

Formel 4 $\sum \left[x_i - \bar{x} \right]$

über alle einzelnen Werte, so lässt sich zeigen, dass die Summe der Abweichungen Null wird. Dieses Maß ist demnach nicht geeignet, um die Variabilität der Werte abzubilden. Dies gelingt dann, wenn man wie folgt vorgeht:

Formel 5 $$s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Varianz

s^2 wird auch als *Varianz* beschrieben. Sie ist – und das geht aus Formel 5 hervor – ein Maß, das als quadrierte mittlere Abweichung zu bezeichnen ist. Der Begriff quadrierte mittlere Abweichung besagt dabei, dass ein Vergleich mit einem Bezugswert hergestellt wird. Dieser Bezugswert ist das arithmetische Mittel $\bar{x}$. Jeder einzelne Wert einer Person i wird dabei in Relation zu $\bar{x}$ gesetzt.

Für s gilt dabei:

Formel 6 $s = \sqrt{s^2}$

Standardabweichung

s wird auch als *Standardabweichung* bezeichnet.

Berechnet man s für die Variable „Taschengeld pro Woche“ (in Tabelle 4.3 = GELD), heißt das konkret:

$\bar{x} = 17{,}53$

$s^2 = 1/16 \text{ x } ((20-17{,}53)^2+(15-17{,}53)^2+(16-17{,}53)^2+$

$(16-17{,}53)^2+(12-17{,}53)^2+(17-17{,}53)^2+(12{,}50-17{,}53)^2+$

$(15-17{,}53)^2+(18-17{,}53)^2+(10-17{,}53)^2+(15-17{,}53)^2+$

$(50-17{,}53)^2+(12-17{,}53)^2+(17-17{,}53)^2+(0-17{,}53)^2+$

$(35-17{,}53)^2) = 116{,}99$

Die Varianz beträgt demnach $s^2 = 116{,}99$.

Die aus der Varianz 116,99 resultierende Standardabweichung (s) entspricht demnach

$s = \sqrt{116{,}99} = 10{,}82$

bei einem $\bar{x} = 17{,}53$.

Was sagt ein solcher Wert aus?

Die Antwort ist: Die mittlere Abweichung (= s) vom arithmetischen Mittel (= $\bar{x}$) beträgt € 10,82. Und durchschnittlich erhalten die befragten Jugendlichen € 17,53 an Taschengeld pro Woche. Gemessen am arithmetischen Mittel, weichen die Angaben der Jugendlichen um 10,82 Einheiten ab. Die Einheiten sind in diesem Fall EURO-Beträge. Demnach besagt dieses Ergebnis: Bei den Befragten variieren die Taschengeldangaben um das arithmetische Mittel im Durchschnitt um € 10,82.

Aus der Zusammenfassung lässt sich erkennen, dass eine einfache Beschreibung von Daten mit Hilfe der Maße der zentralen Tendenz nicht ausreicht, weil eher zu erwarten ist, dass diese Maße untereinander variieren.

Als Maße der Variabilität wurden dargestellt: die Spannweite, die Varianz und die Standardabweichung. Standardabweichung und Varianz sind direkt aufeinander bezogen. Die Quadratwurzel aus der Varianz ist mit der Standardabweichung identisch.

Übung 4.4

Berechnen Sie Spannweite, Varianz und Standardabweichung für die folgenden Variablen:

a) Interesse an Computern

b) Mathematiknote

c) Interesse am Kinobesuch

Die Lösungen finden Sie im Anhang.

4.4 Maße zur Bestimmung des Zusammenhangs

Manchmal ist es notwendig, mehrere Daten gleichzeitig zu betrachten, um Zusammenhänge abzuleiten, aber auch um Prognosen zu machen. In beiden Fällen müssen die Daten parallel betrachtet werden.

Aus der Menge der Daten sollen zwei Variablen näher analysiert werden, um an ihnen ein Zusammenhangsmaß darzustellen. Wir greifen hierzu die Variablen *Taschengeld pro Woche* (= GELD) und *Interesse an der Freizeittätigkeit, Kinofilme ansehen* heraus. Aus der Tabelle 4.3 und dem Beilageblatt ist ersichtlich, dass von allen befragten Jugendlichen Messwertpaare vorliegen, d. h. jeder Jugendliche hat jeweils eine Angabe über sein Interesse gemacht, ins Kino zu gehen, und es liegt gleichzeitig eine Angabe darüber vor, wie hoch das wöchentliche Taschengeld ist.

Aus diesen Angaben lässt sich nachfolgende Grafik bestimmen (siehe Abbildung 4.7):

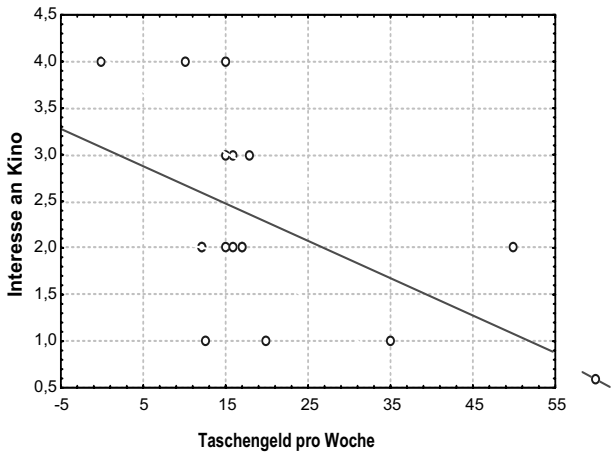


Abb. 4.7: Darstellung des Zusammenhangs zwischen Taschengeld und Kinobesuch

Wenn man einzelne Personen herausgreift, so fällt auf, dass der Jugendliche, der angibt, über kein Taschengeld zu verfügen, sein Kinointeresse mit 4 markiert hat, wogegen der Jugendliche mit einem Taschengeld von € 50 sein Interesse durch eine 2 quantifiziert hat. Entsprechend lassen sich alle anderen Jugendlichen in der zweidimensionalen Darstellung von Abbildung 4.7 identifizieren. Wichtig ist dabei die Feststellung, dass jeder Jugendliche mit zwei quantitativen Werten in der zweidimensionalen Abbildung vertreten ist. Der Schnittpunkt dieser zwei Werte – bedingt durch die Senkrechten auf den beiden Achsen – ist identisch mit dem kleinen Kreis in der Abbildung 4.7. Die kleinen Kreise entsprechen auch Punkten, deshalb spricht man folgerichtig auch von einem Punkteschwarm, wenn alle Punkte im Diagramm berücksichtigt werden.

Punkteschwarm

Aus Abbildung 4.7 ist nun zu ersehen, dass eine Gerade eingezeichnet ist. Diese Gerade zeigt den Verlauf des Punkteschwarms an. Sie zeigt von oben links nach unten rechts. Was bedeutet dieser Verlauf?

Erst eine Inspektion der Fragebögen lässt zu, dass dieser Sachverhalt richtig interpretiert werden kann: Kleine Werte bei der Einschätzung des Freizeitverhaltens zeigen an, dass das *Interesse an Kinobesuchen* groß ist. Höhere Werte beim Taschengeld deuten an, dass die Jugendlichen auch über eine größere Geldmenge verfügen. Demnach deutet die Gerade an, dass Jugendliche mit mehr Taschengeld auch ein größeres Interesse an Kinobesuchen haben. Der angedeutete statistische Zusammenhang ist also demnach sinnvoll interpretierbar. Dieser Zusammenhang kann bei Werten, die stetig verteilt sind, über eine so genannte Produkt-Moment-Korrelation (r_{xy}) bestimmt werden.

Korrelationskoeffizient: Interpretation

Dieser Korrelationskoeffizient ist ein so genanntes normiertes Maß. Der Begriff *Normierung* sagt dabei folgendes aus.

- Resultiert ein $r_{xy} = 0$, so liegt kein Zusammenhang vor;
- Ist $r_{xy} = +1$, so ist ein perfekter Zusammenhang gegeben; der Verlauf der Geraden im Punkteschwarm ist derart, dass hohe Werte bei der einen Variablen mit hohen Werten bei der anderen Variablen einhergehen bzw. niedrige Werte bei der einen Variablen mit niedrigen Werten bei der anderen Variablen.
- Ist $r_{xy} = -1$, so ist zwar auch ein perfekter Zusammenhang gegeben, doch ist der Verlauf der Geraden im Punkteschwarm derart, dass hohe Werte bei der einen Variablen mit niedrigen Werten bei der anderen Variablen bzw. niedrige Werte bei der einen Variablen mit hohen Werten bei der anderen Variablen einhergehen.

Im Regelfall wird kein perfekter Zusammenhang gefunden.

Berechnung des Korrelations-koeffizienten

Der Zusammenhang zwischen den Variablen bestimmt sich bei der Produkt-Moment-Korrelation wie folgt:

Formel 7

$$r_{xy} = \frac{\left[s_x \cdot s_y\right]}{s_{xy}}$$

Kovarianz

Aus der Formel 7 wird deutlich, dass Standardabweichungen von zwei Variablen parallel betrachtet werden; eine neue Größe ist dagegen $s_{xy.}$ Sie ist die sogenannte Kovarianz und wird wie folgt bestimmt:

Formel 8

$$s_{xy} = \frac{\sum\left[\left(x_i - \bar{x}\right) \cdot \left(y_i - \bar{y}\right)\right]}{n}$$

Formel 8 zeigt dabei an, dass eine Differenz zwischen dem arithmetischen Mittel einerseits und dem individuellen Wert andererseits bestimmt wird, nur jetzt mit dem Unterschied, dass die Differenzen für zwei Werte gleichzeitig berechnet werden.

Signifikanz

Im obigen Beispiel (siehe Abbildung 4.7) resultiert dabei ein Korrelationskoeffzient von r_{xy} = -.43. Das bedeutet zwar, dass dieser Wert sehr weit von der Maximalgrenze entfernt ist, doch ist der Zusammenhang als moderat zu bezeichnen. Mit Hilfe von sogenannten Signifikanztests lässt sich bestimmen, ob ein gegebener Korrelationskoeffizient bedeutsam von einer Korrelation r_{xy} = .00 abweicht. Wird ein solches Ergebnis bestätigt, so spricht man auch von einem signifikanten Wert.

Wie lassen sich die dargestellten Sachverhalte weiter erläutern? Offensichtlich handelt es sich bei der Bestimmung einer Korrelation inhaltlich um die quantitative Bestimmung eines Zusammenhangs, statistisch besehen aber – so Formel 7 – um das Verhältnis von Kovarianz (s. o.) und dem Produkt aus den Standardabweichungen der beiden Variablen x und y. Dieses Verhältnis kann in einer Abbildung veranschaulicht werden: In Tabelle 4.3 sind die Einschätzungen der beiden Freizeittätigkeiten *„Rumhängen“* (INT_RUM) und *„Kinofilme ansehen“* (INT_KINO) wiedergegeben. Die Verteilung der beiden Variablen ist in Abbildung 4.8 jeweils getrennt voneinander in den beiden Häufigkeitsverteilungen oben rechts (INT_KINO) und unten links (INT_RUM) abgebildet. Die gemeinsame Verteilung der beiden Variablen ist unten rechts in einem Diagramm dargestellt. Hier spricht man auch von einer bivariaten Verteilung.

Die Korrelation muss, folgt man der Geraden, die im Diagramm eingezeichnet ist, eher klein sein. Sie verläuft flach und leicht ansteigend von unten links nach oben rechts. Dieser Sachverhalt bedeutet, dass

die Korrelation auf keinen Fall maximal sein kann. Denn maximal wäre nach der obigen Definition ein Zusammenhang von $r_{xy} = +1$ oder -1. Der rechnerische Zusammenhang ist in diesem Fall $r_{xy} = +.14$.

Wann aber wäre ein Zusammenhang maximal? Ein solcher Fall ist in Abbildung 4.9 angegeben. Dort verläuft eine Gerade als Winkelhalbierende im Diagramm vom Nullpunkt des Achsensystems ausgehend von unten links noch oben rechts. Zugleich sind im gleichen Diagramm, jeweils von der x- und y-Achse ausgehend, zwei Senkrechte eingezeichnet, die andeuten sollen, dass die x- und y-Abstände zur Winkelhalbierenden jeweils gleich groß sind. Damit wird auch deutlich, wie die Konstruktion der Geraden im Punkteschwarm zustande kommt. Die Gerade wird nämlich so gewählt, dass die Summe der quadrierten Abstände (siehe Abbildung 4.9) ein Minimum beträgt. Unter dieser Voraussetzung erfolgt eine Bestimmung der Korrelation $r_{xy.}$

Ein Korrelationskoeffizient r_{xy} kann im übrigen auch anders interpretiert werden. Berechnet man das Quadrat des Korrelationskoeffizienten, so wird eine Varianz bestimmt:

Formel 9 $$D = \left(r_{xy}\right)^2$$

Determinationskoeffizient

Wie Formel 9 andeutet, wird dieses Quadrat als D abgekürzt. D steht für Determinationskoeffizient. D drückt aus, wie viel gemeinsame Varianz zwischen den beiden Variablen besteht. Multipliziert man den Ausdruck in Formel 9 mit 100, so resultiert

Formel 10 $$D\,\% = \left(r_{xy}\right)^2 \cdot 100$$

Dieser Wert drückt dann aus, wie viel Prozent an gemeinsamer Varianz zwischen beiden Variablen besteht. Bei einer Korrelation von $r_{xy} = .40$ liegt demnach ein gemeinsamer Varianzanteil von 16% vor; es verbleiben 84% nicht gemeinsame Varianz und somit 84% unerklärt.

Abb. 4.8:
Darstellung des Zusammenhangs zwischen zwei Variablen

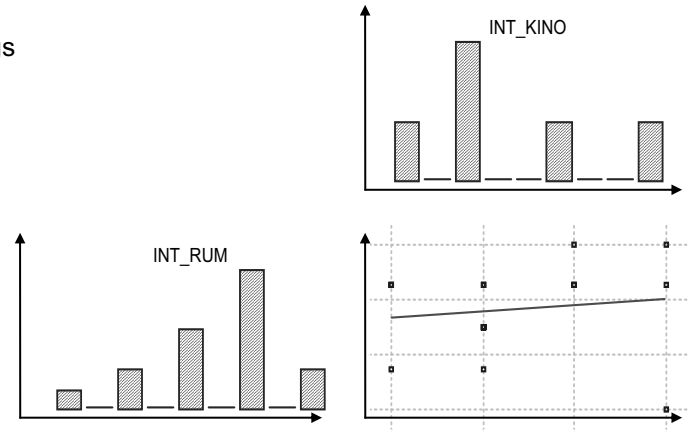
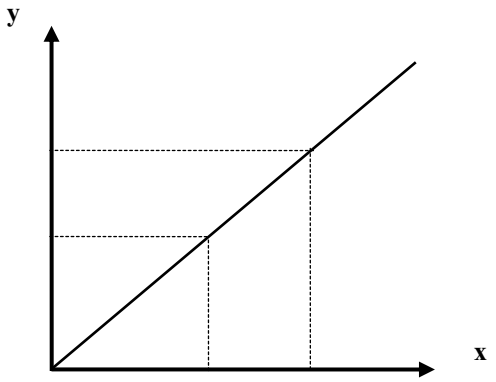


Abb. 4.9:
Darstellung eines maximalen Zusammenhangs bei zwei Variablen x und y



Wie lässt sich nun ein Korrelationskoeffizient interpretieren? Die Interpretionsmöglichkeiten sind zahlreich. Es lassen sich mindestens vier verschiedene Arten der statistischen Abhängigkeit interpretieren:

- Es liegt eine einseitige Steuerung vor: x bewirkt y: x ⇨ y.
- Es ist eine gegenseitige Steuerung gegeben: x bewirkt y und y bewirkt x: x ⇔ y.
- Es liegt eine Beeinflussung durch eine dritte – und möglicherweise unbekannte Größe vor: x und y werden von z bewirkt.
- Es liegt eine komplexe Beeinflussung vor: a + b + c + ... + x bewirken y.

Eine absolut richtige Interpretation kann nur im Zusammenhang mit einer theoretischen Begründung geliefert werden.

Es wurde ein Maß zur Bestimmung des Zusammenhangs dargestellt. Hierbei wurde auf den so genannten Produkt-Moment-Korrelationskoeffizienten zurückgegriffen. Dieser beschreibt den linearen Zusammenhang. Voraussetzung zu seiner Bestimmung ist das Vorliegen kontinuierlich verteilter Messwerte. Ist diese Voraussetzung nicht gegeben, so soll auf andere Maße zur Bestimmung dieses Zusammenhangs zurückgegriffen werden.

Der Korrelationskoeffizient beruht auf einem normierten Maß. Ein perfekter Zusammenhang ist durch den Wert 1 gegeben. Liegt kein Zusammenhang vor, so ist $r_{xy} = 0$.

4.5 Zusammenfassung

In den vorangegangenen Abschnitten wurden verschiedene statistische Maße dargestellt, die eine prinzipielle Eignung dafür haben, bei einer Vielzahl von Fragestellungen zur Anwendung zu kommen. Zur Vereinheitlichung wurde von einem Beispiel ausgegangen, bei dem die Werte für eine größere Gruppe von Personen vorliegen, so dass der Leser in die Lage versetzt wird, die Gedankengänge anhand von Befragungsdaten nachzuvollziehen.

Dabei war es das Ziel, lediglich in einige Grundprinzipien einzuführen, ohne dass weder die notwendigen mathematischen Ableitungen dargelegt, noch dass der Text mit zu vielen Kennwerten überfrachtet wurde.