Betriebsstatistik

beschreibende Statistik

Inhalt

Abbildungen:

Tabellen:

Peter Schmidt: „Betriebsstatistik"; in: Dey und Grauvogel (Hrsg.): "Praxishandbuch – Wirtschaftswissen von A-Z für die erfolgreiche Betriebsratspraxis", Kissing, 2004

1 QUANTITATIVE METHODEN; STATISTIK

Statistik ist ein Gebiet, das mit vielen Vorbehalten und Vorurteilen behaftet ist. Sie dies die Sorge vor zu viel Mathematik, Formeln und anderem schwer verständlichem. Oder seien es Redensarten, die ein unbedarftes Herangehen an dieses Gebiet erschweren, wie der berühmte Ausspruch (Winston Churchill zugeschrieben) „Ich traue keiner Statistik, die ich nicht selbst gefälscht habe“ oder der beliebten Steigerung „Lüge – gemeine Lüge – Statistik“.
Trotzdem begegnet uns Statistik an vielen Stellen des täglichen (betrieblichen) Lebens und es ist wichtig, damit umgehen zu können. Es ist *nicht* nötig, anhand kompliziert klingender Begriffe davon auszugehen, dass das Gegenüber „schon Recht haben wird“, wenn man in der Lage ist, fachkundig nachzufragen und Aussagen kritisch zu hinterfragen. Nicht alle statistischen Modelle und Kennzahlen sind in allen Zusammenhängen und für alle Arten von Daten anwendbar.
Aber nicht nur die Situation, vorgelegte statistische Auswertungen verstehen und (kritisch) interpretieren zu müssen, kann in der täglichen Praxis auftauchen, sondern auch der Wunsch, vorhandene Daten selbst auszuwerten und anschaulich darzustellen. Dies kann die (grafische) Aufbereitung zur Präsentation der Daten sein, aber auch die Analyse von statistischen Zusammenhängen bzw. Unterschieden von Daten oder Sachverhalten.

Begleitende Excel-Datei

Daher werden in diesem Artikel wichtige betriebsstatistische Methoden nicht nur vorgestellt, sondern können mit dem PC selbst nachvollzogen werden, da sich auf der Webseite http://www.fbw.hs-bremen.de/pschmidt unter → MBA/KM eine Excel-Datei befindet, mit der die Beispiele aus dem Text nachvollzogen werden können.
Es wird hier beispielhaft das Tabellenkalkulationsprogramm Excel (aus dem Office Paket der Firma Microsoft) verwendet, da dieses eine sehr große Verbreitung hat. In anderen Tabellenkalkulationen können die dargestellten Methoden ebenso verwendet werden. Darüber hinaus gibt es spezielle Statistik-Programme, die die Verarbeitung von Daten zwar erleichtern, für den täglichen Gebrauch jedoch i.d.R. nicht notwendig sind, z.B. SPSS, SAS, Statgraphics, u.v.m. Auf diese wird hier nicht eingegangen.

Der vorliegende Artikel will mehr bieten als nur die Aufzählung verschiedener Methoden und deren kurze verbale Beschreibung. Ziel ist eine lesbare und alltagstaugliche Übersicht über gängige Methoden und nicht eine mathematisch umfassende Darstellung der Statistik. Auf Formeln wird weitgehend verzichtet; zur Vergleichbarkeit mit bzw. Orientierung in Nachschlagewerken werden die üblichen Symbole (Buchstaben, Abkürzungen) verwendet. Am Ende des Artikels findet sich ein Schlagwortindex, der das Auffinden einzelner Begriffe erleichtern soll. Für tiefergehende Fragen sind am Schluss einige Literaturhinweise zusammengestellt.

Ziel dieses Artikels ist es, zu zeigen, dass quantitative Methoden in der täglichen Arbeit – v.a. durch den Einsatz von EDV-Programmen – einfach zu erstellen und dadurch praktisch und nutzbringend einsetzbar sind.

1.1 Definitionen – Was ist Statistik?

Statistik ist ein Hilfsmittel, ein Werkzeug zur systematischen Darstellung und Auswertung von Zahlenmaterial, meist kurz als „Daten“ bezeichnet. Mit statistischen Methoden werden Kennzahlen gebildet, die dabei helfen, vorliegendes Datenmaterial - vor allem aber die entsprechenden Sachverhalte - möglichst objektiv zu bewerten.
Es gibt zwei grundlegende Ziele statistischer Analysen:

beschreibende Statistik

- Beschreibung vorhandener Daten: *Beschreibende* oder *Deskriptive Statistik*
  Es liegen Daten (Zahlen) vor, die ausgewertet werden sollen: z.B. Alter und Einkommen von 20 Mitarbeitern oder 100 Gewichtsangaben von Werkstücken oder Umsatzzahlen in 16 Quartalen, ... usw.

schließende Statistik

- Ableiten allgemeiner Aussagen aus einer kleinen Auswahl von Daten: *Schließende* oder *Induktive Statistik*
  Es liegt nur eine (kleine) *Stichprobe* von Daten vor, aus diesen sollen allgemeingültige Schlüsse über die Grundgesamtheit aller Daten gezogen werden: z.B.: Aus den Angaben über Alter, Geschlecht und Provision von 50 Angestellten soll auf die entsprechenden Werte aller 800 Mitarbeiter geschlossen werden oder aus den Umsatzentwicklungen von 20 Betrieben soll die Branchenentwicklung abgeschätzt werden.

Dieser Artikel behandelt die beschreibende Statistik.

Verdichtung von Informationen – abhängig von der Skalierung der Daten

Verdichtung von Datenmaterial

Eine Hauptaufgabe statistischer Methoden ist es, die oft sehr große Fülle von Informationen auf wenige (Kenn-) Zahlen zu *verdichten*. Beispiel: Von 500 Beschäftigten mögen z.B. die Dauer der Betriebszugehörigkeit und die Ausbildung vorliegen, dies sind 1.000 Zahlen. Statistisch sprechen wir von **Merkmalen** (z.B. Alter, Geschlecht) und deren **Ausprägungen** (z.B. 20 Jahre, 44 Jahre bzw. männl., weibl.).
Durch Auszählung von Häufigkeiten oder Angabe eines Mittelwertes können diese z.B. auf drei Häufigkeitsangaben (z.B. 100 angelernte Arbeiter (Ar), 250 Facharbeiter (F) und 150 Angestellte (An)) oder im Fall der Betriebszugehörigkeit sogar auf einen Mittelwert (z.B. Durchschnitt von 8,5 Jahren (J)).

Abbildung 1 Verdichtung von Information

Ursprungsdaten:

$$\begin{bmatrix} (Ar,4J) & (An,6J) & (An,2J) & (F,3J) & \ldots \\ (F,1J) & (Ar,10J) & (Ar,9J) & (An,3J) & \ldots \\ (An,6J) & (F,6J) & (Ar,2J) & (Ar,6J) & \ldots \\ (Ar,5J) & (An,1J) & (Ar,1J) & (Ar,9J) & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}$$

Verdichtungen, z.B.:

Mittelwert:
8,5 Jahre
Betriebszugehörigkeit

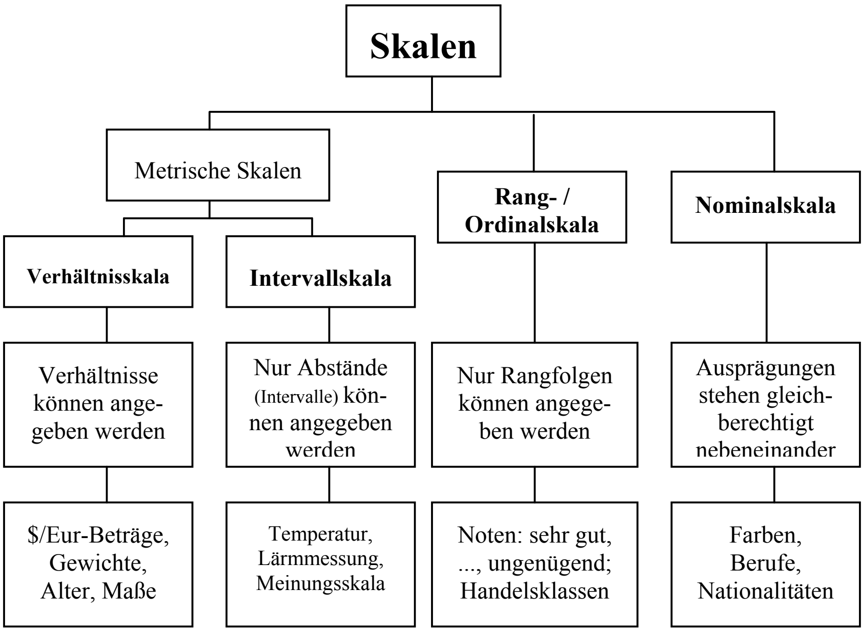
Häufigkeiten:
100 Angelernte
250 Facharbeiter
150 Angestellte

Während in den Ursprungsdaten also alle Personen mit allen Eigenschaften enthalten sind, enthalten Verdichtungen nur einzelne ausgewertete Kennzahlen.

Es zeigt sich jedoch, dass nicht alle Maßzahlen für alle Merkmale möglich sind, so würde ein Mittelwert beim Mitarbeiterstatus keinen Sinn machen. Genauer gesagt ist die Auswahl der statistischen Maßzahlen von der **Skalierung des Merkmals** abhängig. Abbildung 2 zeigt die vier Skalen, die üblicherweise unterschieden werden.

Skalierung von Daten als Basis für die Anwendbarkeit statistischer Methoden

Abbildung 2 Skalierung von Merkmalen



Verhältnisskalierte Daten beinhalten die meiste Information, nominal skalierte die wenigste. Entsprechend stehen mehr oder weniger statistische Methoden zur Auswertung der Daten zur Verfügung

Merkmale können in diskreter oder stetiger Form vorliegen. Diskrete Merkmale können nur abzählbar viele Ausprägungen annehmen, wie z.B. oben der Berufsstatus, das Geschlecht oder Farben. Stetige Merkmale hingegen können beliebig viele Ausprägungen annehmen, oft werden sie in Dezimalzahlen gemessen, z.B. Geldbeträge, Gewichte oder Mengen.

Die Unterscheidungen von Typen und Skalen werden im folgenden wichtig sein, wenn die Methoden zur Auswertung beschrieben werden.

1.2 Häufigkeiten und grafische Darstellungen

Wie oben gesehen, ist auch die Auszählung von Häufigkeiten ein Mittel zur Verdichtung von Daten, gerne wird diese grafisch dargestellt.

1.2.1 Eindimensionale und mehrdimensionale Häufigkeitsdarstellungen

Die Ursprungsdaten (oder Rohdaten - vgl. Abbildung 1) werden oft in Tabellen dargestellt, die aus Zeilen und Spalten bestehen. Dabei stellt jede Zeile eine statistische Einheit (Person, Werkstück, Summe, ...) und jede Spalte ein bestimmtes Merkmal dar.

Diese Darstellung wird auch in Tabellenkalkulationsprogrammen verwendet. Hier kann dann jede Zelle (z.B. Zeile 3, Spalte 4) einzeln angesteuert bzw. berechnet werden. Beispiele hierzu finden sich in den Excel-Dateien. Die laufende Nummer wird auch als (Lauf-)

Index bezeichnet, daher die übliche Abkürzung i. Es können dann alle Angaben anhand dieses Index angegeben werden. Z.B. ist die 3. Person seit 2 Jahren im Betrieb: $B_3 = 2$.

Ursprungsdaten

Tabelle 1 Personaldaten (Beispiel für Ursprungsdaten)

↓ *Spalte*

	lfd. Nummer	Geschlecht	Betriebszugehörigkeit	Berufs-Status	Note im Eignungstest	
	i	G_i	B_i	S_i	N_i	
	1	w	10	Ar	2	
	2	w	5	An	1	
Zeile →	3	m	2	Ar	3	
	4	w	18	F	8	
	5	m	22	Ar	1	*Zelle*
	6	m	9	An	9	
	7	m	14	F	2	

Die einfachste Verdichtung von Daten ist die Angabe von Häufigkeiten, oft ebenfalls in tabellarischer Form, wie Tabelle 2 für den Fall einer einfachen Häufigkeitstabelle für das Merkmal „Berufsstatus" zeigt.

Häufigkeitstabelle

Tabelle 2 Eindimensionale Häufigkeitstabelle – Anzahl Befragte nach Berufsstatus

	Kürzel	Status	Anzahl (Häufigkeit n_i)	(Bezeichnung)
Status:	Ar	angelernte Arbeiter	100	$= n_1$
	F	Facharbeiter	250	$= n_2$
	An	Angestellte	50	$= n_3$
	Summe:		**400**	= **n** Beschäftigte

Der Buchstabe „n" als Symbol für „Anzahl der Beobachtungen" wird in der Statistik sehr häufig verwendet. Wenn es sich auf die Grundgesamtheit aller statistischen Einheiten bezieht, wird auch ein großes „N" verwendet.

Interessanter ist die Aufbereitung mehrerer Dimensionen, etwa die Auszählung der Anzahl der Beschäftigten, diesmal nach Alter *und* Geschlecht, wie sie in Tabelle 3 vorgenommen wird.

zweidimensionale Häufigkeiten

Tabelle 3 Zweidimensionale Häufigkeitstabelle – Befragte nach Alter und Geschlecht

Betriebszugehörigkeit	Geschlecht weiblich	männlich	**alle Personen**	
unter 10 Jahre	80	40	**120**	*Randsummen*
10 - 20 Jahre	100	80	**180**	
über 20 Jahre	120	80	**200**	
alle Personen	**300**	**200**	**500**	*Gesamtsumme*

Randsummen

In dieser Tabelle 3 sind zum einen die Einzelhäufigkeiten für die Kombinationen bestimmter Eigenschaften (z.B. haben 40 Männer eine Betriebszugehörigkeit unter 10 Jahren), aber auch - in den „Randsummen“ die Häufigkeitsauszählungen für die einzelnen Merkmale (z.B. insgesamt gibt es 180 Personen mit einer Betriebszugehörigkeit zwischen 10 und 20 Jahren). Für das Merkmal Betriebszugehörigkeit wurden *Klassen* (von ... bis ...) gebildet. Dies ist sinnvoll, wenn viele Ausprägungen vorhanden sind, so dass diese nicht mehr übersichtlich in einer Tabelle oder Grafik dargestellt werden können.
Üblich ist auch die Darstellung von *relativen* oder *prozentualen* Häufigkeiten.

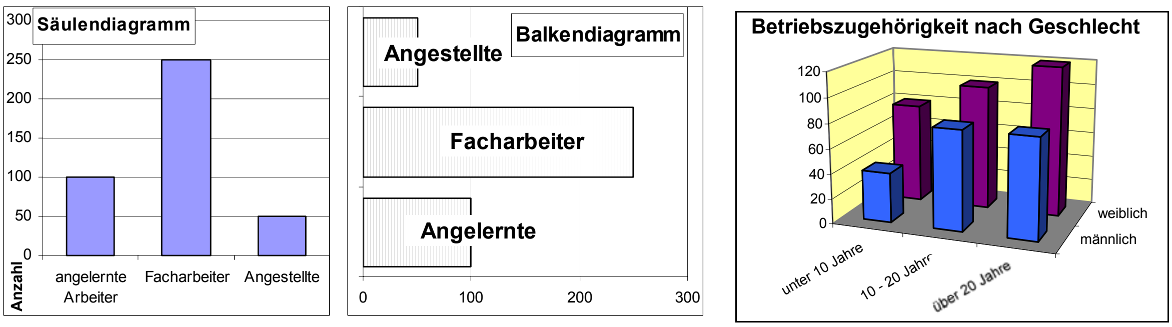
relative Häufigkeiten

Tabelle 4 Prozentuale Häufigkeiten – Befragte nach Alter und Geschlecht

	Zeilenprozente			**Spalten**prozente		
Betriebs-zugehörigkeit	weib-lich	männ-lich	**alle Personen**	weib-lich	männ-lich	**alle Personen**
unter 10 Jahre	66,7%	33,3%	100%	26,7%	20,0%	**24,0%**
10 - 20 Jahre	55,6%	44,4%	100%	33,3%	40,0%	**36,0%**
über 20 Jahre	60,0%	40,0%	100%	40,0%	40,0%	**40,0%**
alle Personen	**60,0%**	**40,0%**	100%	100%	100%	100%

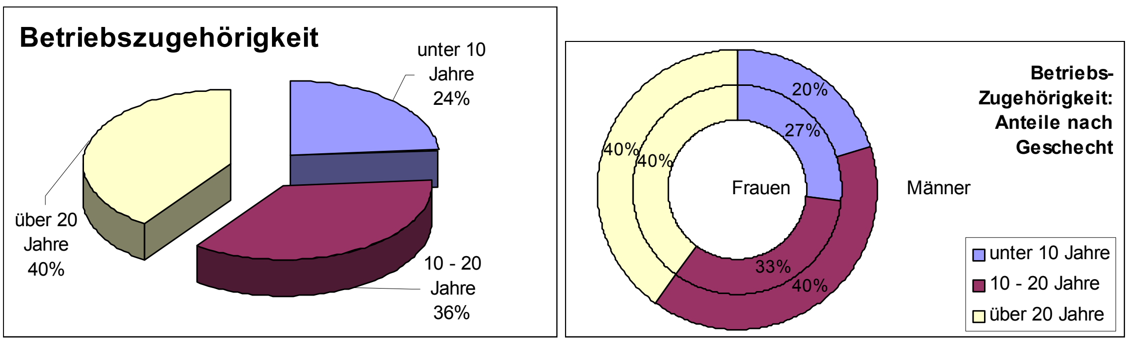
Diese Häufigkeitsdarstellungen, ob in absoluten Zahlen oder relativen Anteilen gemessen, werden oft grafisch dargestellt. So lassen sich die Zahlen aus Tabelle 2 z.B. in einem Balken- oder Säulendiagramm darstellen, wie in Abbildung 3 linke dargestellt ist.

Abbildung 3 Absolute Häufigkeiten und zweidimensionales Säulendiagramm



Auch die zweidimensionalen Häufigkeiten aus Tabelle 3 lassen sich grafisch veranschaulichen (z.B. wie in Abbildung 3 recht oder Abbildung 4).

Abbildung 4 Kreisdiagramme zur Darstellung von relativen Häufigkeiten



Grafiken wie die hier beispielhaft vorgestellten lassen sich mit Hilfe von Computerprogrammen relativ einfach erzeugen. Es gibt eine sehr große Anzahl von Darstellungsmöglichkeiten und es sollte jeweils aus dem konkreten Zusammenhang entschieden werden, welche Darstellung hilfreich „für den Transport der Botschaft“ ist.

Die Daten und die hier dargestellten Beispiele finden sich in der begleitenden Excel-Datei.

1.2.2 Besondere Häufigkeitskonzepte

Histogramme und Summenhäufigkeiten

Histogramme werden verwendet, wenn die Ausprägungen wie oben in Klassen eingeteilt werden und diese unterschiedlich breit sind. Säulendiagramme würden in diesem Fall falsche Häufigkeiten vermuten lassen, so dass die Häufigkeiten als Fläche dargestellt werden. **Summenhäufigkeitsfunktionen** zeigen, wie viel (Prozent der) Ausprägungen *höchstens* einem bestimmten Wert annehmen (*bis zu* ...). Abbildung 5 zeigt dieses Häufigkeitskonzept neben einem einfachen Säulendiagramm.

Abbildung 5 Einfache Häufigkeiten und Summenhäufigkeiten

Werte:

Körpergröße:	
186	157
164	160
187	164
168	196
190	180
178	196
195	186
172	160
163	189
183	178
180	188
180	188
196	178
186	174
182	178
161	186

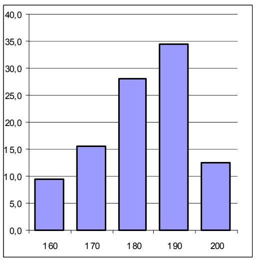
f_i: relative Häufigkeit

Klasse Obergrenze	Häufigkeiten: n_i	f_i %	F_i
160	3	9,4	9,4
170	5	15,6	25,0
180	9	28,1	53,1
190	11	34,4	87,5
200	4	12,5	100,0
Summe:	32		

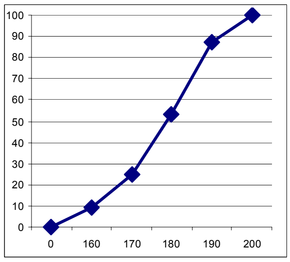
F_i: Summen-Häufigkeit: ... Prozent der Personen sind *höchstens* ... groß

32 befragte Personen

Relative Häufigkeiten (%)



Summenhäufigkeiten



1.3 Lagemaße (Mittelwerte) und Streuungsmaße

Die bisherige Beschreibung zeigte Möglichkeiten der Darstellung, die deutlich anschaulicher sind als die Betrachtung von Ursprungsdaten, aber die Tabellen und Abbildungen in Punkt 1.2 müssen jeweils ihrerseits interpretiert werden: „Wie unterscheiden sich zwei Grafiken?“; „Was ist das wichtige an diese Tabelle?“ ... mögen die Fragen lauten. Daraus ergibt sich der Wunsch, nach noch knapperen statistischen (bzw. betrieblichen) *Kennzahlen*.

Mittelwerte: auf die Skala achten

Die sicherlich bekanntesten statistischen Maße sind Mittelwerte, unter Ihnen der „prominenteste“ das arithmetische Mittel, oft einfach als „Mittelwert“ bezeichnet. Aber nicht für alle Daten (jeder Skalierung) kann ein arithmetische Mittel errechnet werden. In der betrieblichen Praxis sind die in Tabelle 5 angegebenen Mittelwerte relevant.

Wichtig ist hier, dass falsche Verwendung der Mittelwerte eben auch zu falschen (oder verfälschten) Ergebnissen führt. So ist das arithmetische Mittel i.d.R. größer als das geometrische Mittel. Würde man im letzten Beispiel (fälschlicherweise) ein arithmetisches Mittel errechnen, so hätte dies einen Wert von 9 Prozent (10+20+5+1=36 / 4). Die tatsächliche Lohnsteigerung der letzten vier Jahre würde also höher angegeben als sie tatsächlich war.

Mittelwerte werden auch als Lagemaße bezeichnet, da sie die Lage einer Verteilung (auf der waagrechten Achse) angeben. So haben die Verteilungen von Abbildung 3 oder Abbildung 5 ihren jeweiligen Schwerpunkt in ihren Mittelwerten.

Tabelle 5 Mittelwerte

Mittelwert	für Skalen	Definition	Beispiel
Modus (Modalwert)	alle Skalen	häufigster Wert	In Tabelle 2 ist F der Modalwert, da Facharbeiter (mit 250 Personen) die größte Einzelhäufigkeit aufweisen
Median oder Zentralwert	Ordinalskalen und metrische Skalen	Mitte aller geordneten Ausprägungen	Alter: 44, 19, 24, 60, 21, 42, 11 geordnet: 11,19,21,**24**,42,44,60 Median = 24, da mittlerer Wert, es stehen rechts und links davon je drei Zahlen
arithmetisches Mittel („*Mittelwert*“)	metrische Skalen	Summe aller Werte geteilt durch die Anzahl der Beobachtungen (n)	Alter: 44, 19, 24, 60, 21, 42, 11 $\frac{44+19+24+60+21+42+11}{7} =$ $\mathit{Durchschnittsalter}\ \bar{x} = 31{,}6\ \mathit{Jahre}$
geometrisches Mittel	Steigerungsraten von Wachstumsdaten (nur Verhältnisskalen)	n-te Wurzel aus dem Produkt aller Werte	Lohnsteigerung in 4 Jahren: 10 %, 20 %, 5 %, 1 % $\sqrt[4]{1{,}10 \cdot 1{,}20 \cdot 1{,}05 \cdot 1{,}01}$ $GM = 1{,}0877 \quad \mathit{oder}$ rund **8,77 Prozent**

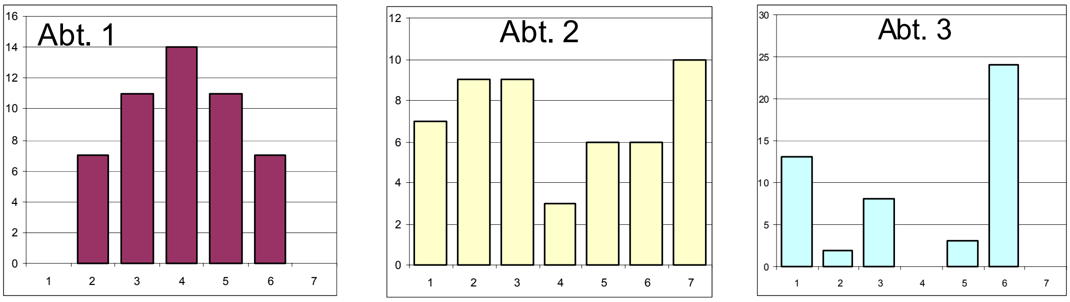
Die einzelnen Berechnung können in der Excel-Datei nachvollzogen werden.

Streu-ungs-maße

Streuungsmaße

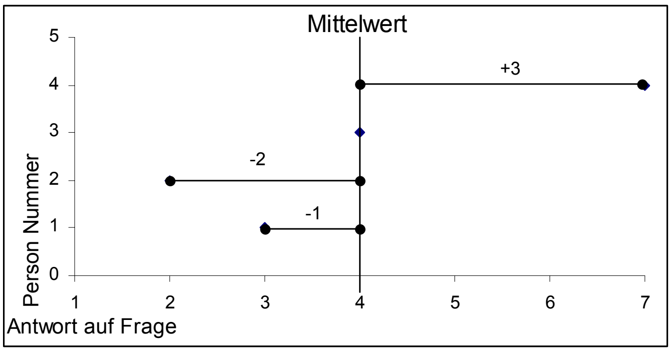
Neben der Lage einer Verteilung ist diese durch ihr Aussehen, etwa ihre „Breite“ bezeichnet: Wie verteilen sich die Ausprägungen des Merkmals um den Mittelwert?. Statistisch wird hier von *Schwankung* oder *Streuung* der Werte gesprochen, so dass die entsprechenden Kennzahlen als **Streuungsmaße** bezeichnet werden. Abbildung 6 zeigt ein Beispiel für 3 verschiedene Verteilungen (Häufigkeitsauszählungen als Säulendiagramm). In allen drei Fällen wurden 50 Personen (aus drei Abteilungen) befragt: „Bewerten Sie die Arbeitszeitregelung im Betrieb“. Die Antworten in allen drei Abteilungen ergaben denselben arithmetischen Mittelwert von 4,0 aufweisen, die Verteilungen sehen aber unterschiedlich aus.

Abbildung 6 Verteilungen mit unterschiedlicher Streuung



Mit Streuungsmaßen kann das unterschiedliche Aussehen dieser drei Verteilungen, statistisch gesagt die unterschiedliche Schwankung gemessen werden: Das einfachste Streuungsmaß ist die ***Spannweite***. Sie wird ermittelt, indem die kleinste Ausprägung von der größten abgezogen wird (Abt.1: 6–2= 4; Abt.2: 7–1= 6; Abt.3: 6–1= 5) und spiegelt damit die Breite der Verteilung wider.
Üblichere Streuungsmaße messen die Abweichung der einzelnen Werte vom Mittelwert, was in Abbildung 7 veranschaulicht wird (am Beispiel von drei Personen aus Abt. 1). Als statistische Kennzahl dient wiederum ein Mittelwert dieser Abweichungen. Die „Durchschnittliche Absolute Abweichung (DAA)" ist ein mögliches Maß, weitaus bekannter jedoch ist die ***Standardabweichung***. Sie wird ermittelt aus der Hilfsgröße „Varianz" (dem Durchschnitt aller quadrierten Abweichungen).

Abbildung 7 Abweichungen der Einzelbeobachtungen vom Mittelwert



Antworten auf einer Ratingskala (Werte von 1 bis 7 konnten angegeben werden)

In Tabelle 6 sind die vorgestellten Maßzahlen für das obige Beispiel (Befragung von je 50 Personen in drei Abteilungen) zusammengefasst.

Tabelle 6 Beispiel für Lage- und Streuungsmaße

Abteilung	Median	arithm. Mittel	Spannweite	Standardabweichung
1	4	4,0	4 (von 2 bis 6)	1,26
2	3	4,0	6 (von 1 bis 7)	2,16
3	5	4,0	5 (von 1 bis 6)	2,18

Es mag sich angesichts dieses einfachen Beispiels die Frage stellen, welchen Sinn solche recht aufwendigen Maßzahlen haben. Dieser liegt vor allem in der *Verarbeitung großer Datenmengen*. Sind nicht nur drei Abteilungen, sondern z.B. 16 Bereiche und nicht nur eine Frage sondern z.B. 35 zu bewerten und so zu verdichten, dass eine Orientierung „auf einen Blick" (oder zumindest wenige Blicke) möglich ist, so geht dies nur mit Hilfe von Kennzahlen. Nicht alle sind an jeder Stelle geeignet. So zeigt hier das arithmetische Mittel eine Übereinstimmung der drei Abteilungen an, was aber angesichts Abbildung 6 nicht zu überzeugen vermag. Schon der Median, vor allem aber die Streuungsmaße zeigen - auch ohne den Blick auf die Grafiken - dass die Antworten in Abteilung 1 recht einheitlich verteilt sind, wogegen diejenigen in Abteilung 2 und 3 größere Schwankungen aufweisen. Dort gehen also die Meinungen weiter auseinander, in unserem Beispiel könnte hiermit ein Anhaltspunkt dafür gegeben sein, dass der Betriebsrat dort - etwa in Einzelgesprächen - klären sollte, ob größere Unzufriedenheit unter den Angestellten herrscht, als in anderen Abteilungen.

1.4 Zusammenhänge zwischen mehreren Merkmalen

Oft ist das Ziel statistischer Analysen nicht nur, *ein einzelnes* Merkmal zu beschreiben, sondern es interessiert die Wirkung verschiedener Merkmale aufeinander. so wurde bereits in Tabelle 4 und Abbildung 3 eine *zweidimensionale* Betrachtung angestellt.
Wiederum können zwei unterschiedliche Fragestellungen unterschieden werden:

- Wird ein (zufälliger) Zusammenhang zwischen zwei Merkmalen untersucht, in dem Sinne, dass die Merkmale sich gegenseitig beeinflussen; oder
- wird ein ursächlicher Zusammenhang vermutet, in dem Sinne, dass bestimmte Merkmale ein anderes beeinflussen bzw. steuern (Kausalität)?

Für die erste Frage eignen sich Zusammenhangmaße, die zweite kann mit Regressionsmodellen untersucht werden.

1.4.1 Zusammenhangmaße

Messen der Stärke eines Zusammenhangs

Zusammenhangmaße beschreiben die *Stärke* eines Zusammenhangs. Die Geschäftsleitung macht auf den steigenden Energieverbrauch einer Abteilung aufmerksam. Der Betriebsrat vermutet, dass dies durch die wachsende Anzahl von Überstunden verursacht wird und vergleicht die beiden Zahlenreihen X: Anzahl der Überstunden pro Woche und Y: Energieverbrauch miteinander. Es werden sechs Wochen ($i = 1, \ldots, 6$), entsprechend sechs Wertepaare (x_i, y_i) miteinander verglichen. Diese sind in Tabelle 7 angegeben.

Tabelle 7 Wertetabelle für Zusammenhang Überstunden / Energieverbrauch

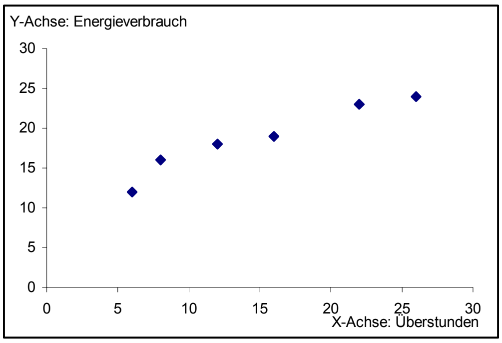
Woche	Überstunden	Energieverbrauch	Woche	Überstunden	Energieverbrauch
i	X_i	Y_i	i	X_i	Y_i
1	6	12	4	16	19
2	8	16	5	22	23
3	12	18	6	26	24

Da es sich um zwei metrisch skalierte Merkmale handelt, können sie - in Abbildung 8 - als xy-Diagramm (Streu- oder Punktdiagramm) dargestellt werden.

In der folgenden Grafik ist ein Zusammenhang zu erkennen: Je höher die Anzahl der Überstunden, desto höher ist auch der Energieverbrauch. Das statistische Maß, welches die Stärke eines solchen Zusammenhanges metrischer Merkmale misst, heißt ***Korrelationskoeffizient*** (nach Bravais-Pearson), üblicherweise mit dem Buchstaben r (oder dem griechischen ρ) bezeichnet. r gibt sowohl die Richtung des Zusammenhanges als auch dessen Stärke an, denn er kann Werte zwischen -1 und 1 annehmen. Im vorliegenden Falle ergibt sich ein

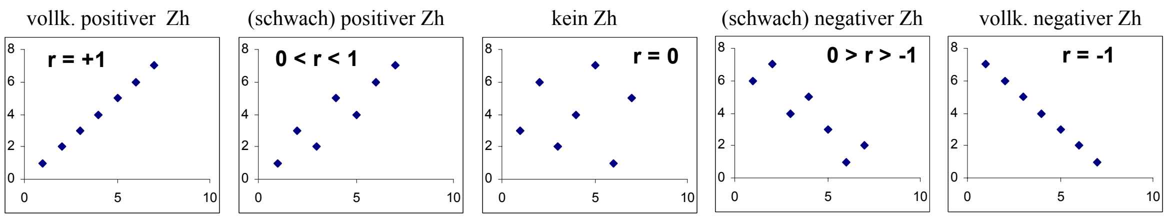
Wert von r=0,97 und damit ein starker statistischer Zusammenhang zwischen Überstunden (X) und Energieverbrauch (Y).

Abbildung 8 xy-Diagramm für Zusammenhang Überstunden / Energieverbrauch



Ein positiver Zusammenhang (r größer als 0) heißt, dass je größer die Ausprägung des einen Merkmals (X), desto *größer* auch die des anderen Merkmals (Y); Ein negativer Zusammenhang (r kleiner als 0) heißt, dass je größer die Ausprägung des einen Merkmals (X), desto *kleiner* die des anderen Merkmals (Y), wobei der Zusammenhang desto stärker ist, je näher r an 1 bzw. -1. Ein Korrelationskoeffizient nahe oder gleich Null bedeutet, dass es *keinen* Zusammenhang zwischen X und Y gibt. Der Korrelationskoeffizient r ist also eine Kennzahl, die eine große Menge an Informationen verdichten kann, indem das Verhältnis beliebig vieler Wertepaare in einer Maßzahl r zusammengefasst wird.

Abbildung 9 Stärke von Zusammenhängen und Werte des Korrelationskoeffizienten



Auch hier gilt wieder, dass ein solches Maß besonders dann nützlich ist, wenn große Mengen von Daten betrachtet werden und nicht für jedes Merkmalspaar ein solches xy-Diagramm erstellt werden kann. Es können dann mittels des Korrelationskoeffizienten schnell diejenigen Merkmale herausgefunden werden, die einen starken Zusammenhang aufweisen und diese näher untersucht werden. Korrelationskoeffizienten sind ein in der betrieblichen Praxis sehr gebräuchliches Maß.

Allerdings kann der Korrelationskoeffizient r nach Bravais-Pearson nur für *metrische* Merkmale ermittelt werden. Bei *ordinal* skalierten Merkmalen muss auf den Rangkorrelationskoeffizienten r_s zurückgegriffen werden; bei *nominal* skalierten Daten steht nur der Kontingenzkoeffizient zur Verfügung. Tabelle 8 zeigt die Zusammenhangmaße für die verschiedenen Skalenniveaus.

Je größer der Informationsgehalt der Skala (vgl. Abbildung 2), desto höher ist auch die Aussagekraft des Zusammenhangmaßes. Der Koeffizient r_s kann nur Sortierungen vergleichen, aber keine Zahlenwerte, der Kontingenzkoeffizient C beinhaltet keine Richtung des Zusammenhanges, bezüglich der Größe von C gilt ebenfalls, dass ein Wert von 0 keinen

Zusammenhang bedeutet und je näher C sich dem Wert 1 nähert, desto stärker ist der untersuchte Zusammenhang zwischen den Merkmalen X und Y.

Tabelle 8 Zusammenhangmaße für unterschiedliche Skalenniveaus

Zusammenhangmaß	für Skalen	Wertebereich	Beispiele
r Korrelationskoeffizient nach Bravais-Pearson	metrische Skalen	$-1 \le r \le 1$	- Produktionsmenge und Kosten - Alter und Einkommen
r_s Korrelationskoeffizient nach Spearman	Ordinalskalen	$-1 \le r_s \le 1$	- Schulnote und Altersklasse - Schulabschluss Leistungsklasse
C Kontingenzkoeffizient	Nominalskalen	$0 \le C \le 1$	- Geschlecht und Beruf

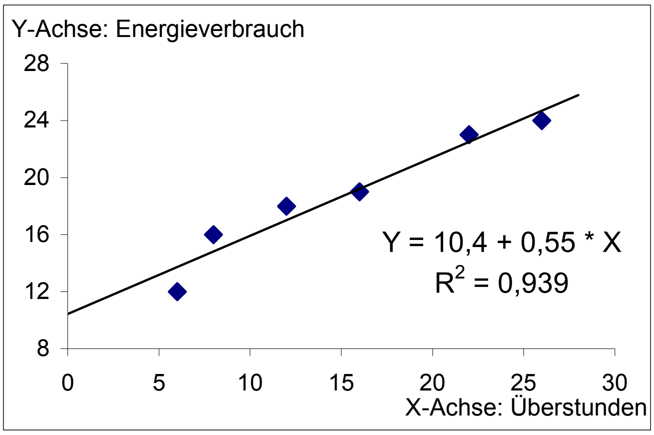
1.4.2 Regressionsanalyse

Bei metrisch skalierten Merkmalen wurden in Abbildung 8 und Abbildung 9 „Punktewolken" betrachtet, also die Verteilung der xy-Wertepaare in einem Koordinatensystem. Als Referenz für die Messung von Stärke eines Zusammenhanges dient dabei eine gedachte Linie durch die Punktewolke und die Betrachtung, wie die Bebachtungspunkte zu dieser Linie liegen. Bei Korrelationskoeffizienten r = 1 und r = -1 liegen die Punkte *auf* dieser gedachten Grade bzw. bilden diese Grade.

Regressionsanalyse mißt die Art eines Zusammenhanges

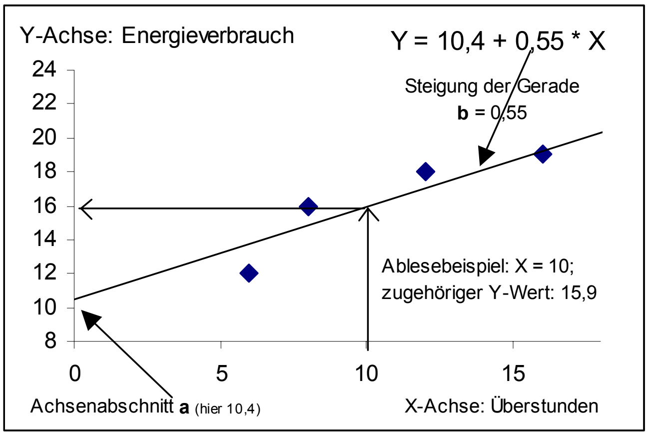
Im Beispiel der Tabelle 7 wurde ein statistischer Zusammenhang zwischen der Anzahl der Überstunden und dem Energieverbrauch ermittelt. Es stellt sich im nächsten Schritt die Frage, *wie* die beiden Merkmale zusammenhängen, welcher Art ihre Beziehung ist. Um dies statistisch zu untersuchen, muss zunächst eine Annahme aufgestellt werden, diese sei:
inhaltlich: Der Energieverbrauch hängt von der Zahl der Überstunden ab
mathematisch: Der Energieverbrauch Y ist eine Funktion der Zahl der Überstunden X
Funktional: $Y = f(X)$ und im linearen Fall: $Y = a + b * X$
In Worten bedeutet dies, dass eine Gerade gesucht wird, die durch das Zentrum der „Punktewolke" geht, wie Abbildung 10 zeigt.

Abbildung 10 Regressionsgrade Überstunden / Energieverbrauch



Die „Regressionsgrade" in Abbildung 10 wird bestimmt durch ihren Schnittpunkt mit der Y-Achse (hier a = 10,4) und ihre Steigung (hier b = 0,55). Mit dieser Grade bzw. der Formel $Y = a + b * X$; hier $Y = 10{,}4 + 0{,}55 * X$ kann für jede denkbare Anzahl von Überstunden ein *erwarteter* Wert für den Energieverbrauch errechnet werden. Daher hat das Modell seinen Namen, denn „re-gressere“ kommt aus dem Lateinischen und bedeutet „zurückführen“; hier wird also der Energieverbrauch auf die Anzahl der Überstunden zurückgeführt. Dies kann zum einen geschehen durch einsetzen von X-Werten in die Formel, so ergibt sich für 10 Überstunden ein erwarteter Energieverbrauch von Y =10,4 + 0,55 * 10 (Stunden) = 15,9 (kWh). X wird auch als das *erklärende* (*unabhängige*) und Y als das *erklärte* (*abhängige*) Merkmal bezeichnet. Dies ist in Abbildung 11 verdeutlicht.

Abbildung 11 Regressionsanalyse



Eine solche Regressionsanalyse kann in Computerprogrammen sehr einfach erzeugt werden. In der Excel-Datei in der beiliegenden Datei finden Sie diese Grafiken, die zugrunde liegenden Zahlen und Hinweise zur Erstellung der Analysen.

Eine Regressionsanalyse bietet somit zwei praktische Möglichkeiten:

- Ein Zusammenhang kann formal beschrieben werden (wie hängen X und Y zusammen?)
- Es kann für gegebene X-Werte ausgerechnet werden, welche Y-Werte zu erwarten sind.

Anwendung finden Regressionsanalysen in verschiedensten Bereichen der betrieblichen Praxis und sind sehr verbreitet.

Gütemaß

Die Güte einer Regressionsanalyse bemisst sich daran, wie gut die Regressionsgrade den tatsächlichen Zusammenhang beschreibt bzw. vorhersagt. Dies wird darin gemessen, wie stark die einzelnen Beobachtungspunkte um die Gerade schwanken. Liegen alle Punkte auf der Gerade, so ist die Regressionsschätzung perfekt. Liegen sie nahe neben der Grade, so ist die Vorhersage, wie im obigen Beispiel, gut – je weiter die Werte von der Grade entfernt liegen, desto „schlechter“ ist die Regression. Diese Darstellung erinnert an die des Korrelationskoeffizienten und tatsächlich ist im bisher besprochen Fall der linearen Einfachregression $\mathbf{R^2} = r^2$ (also das Quadrat des Korrelationskoeffizienten) ein **Gütemaß** für die Regressionsanalyse. (Es wird allgemein als „R-Quadrat“ ausgesprochen, wobei dies als Eigenname zu verstehen ist. Es gibt keine Zahl R, die dann quadriert wird, sondern das Gütemaß heißt R^2, bei machen Autoren aber auch B für Bestimmtheitsmaß). In Worten sagt R^2 aus, wie viel Prozent der Schwankungen der Y-Werte durch die X-Werte vorhergesagt werden. R^2 liegt also zwischen 0 und 1 ($0 \leq R^2 \leq 1$). Im obigen Beispiel ist in Abbildung 10 das Gütemaß mit 93,9 Prozent angegeben, diese Beispielregression beschreibt die Daten also gut, was ja auch grafisch erkennbar ist.

In der Praxis sind allerdings die wenigsten zu untersuchenden Zusammenhänge so einfacher Natur wie das obige Beispiel:

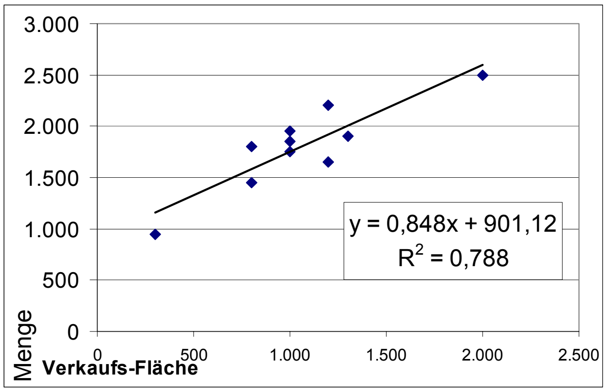
- Der Zusammenhang kann nicht-linear sein, d.h. die Punktwolke kann nicht durch eine Grade, sondern müsste durch eine Kurve beschrieben werden
- Y hängt nicht genau von einer Erklärungsgröße X ab, sondern von mehreren.

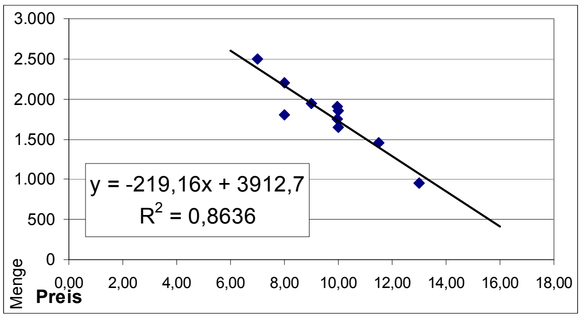
Beide Erweiterungen des Regressionsmodells sind in der Praxis sehr gebräuchlich. Abbildung 12 zeigt ein Beispiel für einen Zusammenhang zwischen einem abhängigen Merkmal, der Absatzmenge eines Produktes und drei Einflussfaktoren, der Verkaufsfläche, der Werbeausgaben und des Preises.

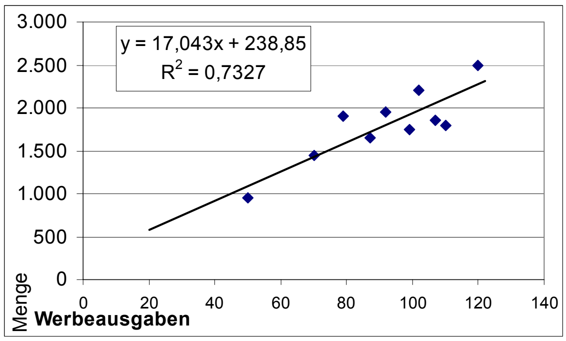
Abbildung 12 Regressionsanalyse mit multiplen Einflussfaktoren

Multivariate Zusammenhänge **Beispiel: Absatzzahlen eines Kosmetikartikels**

	Absatz- Menge	Verkaufs- **Fläche**	**Werbe-** Ausgaben	**Preis** pro Einheit
	Stück	*qm*	*TEuro*	*Euro*
i	y_i	$x1_i$	$x2_i$	$x3_i$
Nr	Absatz	Fläche	Werbung	Preis
1	2.500	2.000	120	7,00
2	1.850	1.000	107	10,00
3	1.750	1.000	99	9,95
4	1.450	800	70	11,50
5	950	300	50	13,00
6	2.200	1.200	102	8,00
7	1.800	800	110	8,00
8	1.950	1.000	92	9,00
9	1.650	1.200	87	10,00
10	1.900	1.300	79	9,95







In Abbildung 12 sind zunächst die Ursprungsdaten und die drei einzelnen Regressionen dargestellt. Anhand der Gütemaße ist zu erkennen, dass der Preis (mit einem R^2 von 86,4 %) den höchsten Erklärungsgrad aufweist, die Verkaufsfläche (R^2 = 78,8 %) den zweithöchsten und auch die Werbeausgaben (R^2 = 73,3 %) einen messbaren Einfluss auf die Absatzmenge haben.

Die inhaltliche Aussage kann der Steigung der Regressionsgraden bzw. dem Vorzeichen von b entnommen werden: Während die Verkaufsfläche, ebenso wie die Werbeausgaben, positiv auf die Absatzmenge wirken, hat der Preis einen negativen Einfluss. Das heißt: je größer die Verkaufsfläche einer Filiale und je höher die dortigen Werbeausgaben, desto höher der Absatz. Je höher jedoch der Preis des Produktes, desto weniger Einheiten werden abgesetzt.

Lineare Mehrfach-regression

Dies kann auch in einer einzigen, *multiplen* Regression errechnet werden. Die Bestimmungsgleichung für den Absatz Y lautet dann:

$$Y = a + b_1 * X_1 + b_2 * X_2 + b_3 * X_3$$

oder hier: Absatz = a + b_1 * Fläche + b_2 * Werbung + b_3 * Preis

Die Durchführung der multiplen Regression in Excel ergibt das folgende Ergebnis:

Absatz = 2398 + 0,4 * Fläche + 1,7 * Werbung – 123,8 * Preis

Damit wird in einer Gleichung das oben dargestellte Ergebnis beschrieben. Eine Erhöhung der Verkaufsfläche erhöht den Absatz um das 0,4-fache, also z.B. 100 qm mehr Verkaufsfläche bringen im Durchschnitt 40 Stück mehr Umsatz. Die Erhöhung der Werbeausgaben um 10 TEuro erhöht den Absatz um 17 Stück und eine Senkung des Preises um 1 Euro würde zu einer Erhöhung des Absatzes um knapp 124 Stück führen. Damit ist auch eine Rangfolge geeigneter Maßnahmen zur Absatzerhöhung erkennbar, die Preissenkung hat in diesem Beispiel die stärkste Wirkung.

Regressionsanalysen finden in verschiedenen Varianten Anwendung. Die hier besprochene lineare Regression wird oft auch als „KQ-Regression" als Abkürzung für Kleinste-Quadrate-Regression (weil mathematisch die Abstände zwischen den Beobachtungspunkten und der Regressionsgrade quadriert werden und die Gerade dann so gewählt, dass die Summe dieser quadrierten Abstände möglichst klein wird) oder OLS-Regression von der englischen Bezeichnung „Ordinary Least Squares"-Regression.

Hinweise zur Durchführung von Regressionsrechnungen finden sich in der Excel-Datei.

1.5 <u>Zeitreihen und Indexzahlen</u>

Ein weiteres Anwendungsgebiet von Zusammenhangmaßen ergibt sich, wenn die Entwicklung eines Merkmals im Zeitablauf betrachtet werden soll, also quasi der Zusammenhang zwischen diesem Merkmal und der Zeit.

Eine **Zeitreihe** wird dabei (künstlich) in mehrere Komponenten zerlegt:

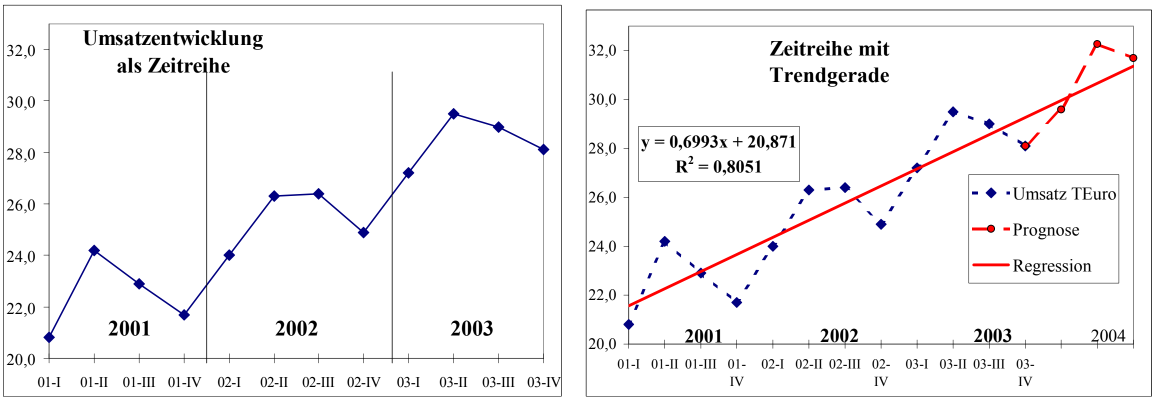
Y = **Trend**-Komponente (+ Konjunktur-Komponente) + **Saison**-Komponente + Rest-Komponente

Beispielsweise seien Umsatzzahlen für die Quartale von 1998 bis 2000 (Tabelle 9) betrachtet, die in Abbildung 13 als Zeitreihe darstellt sind.

Tabelle 9 Zeitreihe einer Umsatzentwicklung

Jahr	2001				2002				2003			
Quartal	01-I	01-II	01-III	01-IV	02-I	02-II	02-III	02-IV	03-I	03-II	03-III	03-IV
Zeitpunkt t	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
y_t	**20,8**	**23,1**	**22,9**	**21,7**	**24,0**	**26,3**	**26,1**	**24,9**	**27,2**	**29,5**	**29,3**	**28,1**

Abbildung 13 Umsatzentwicklung im Zeitablauf



Die linke Grafik in Abbildung 13 zeigt die Entwicklung des Umsatzes in den drei Jahren, wobei erkennbar ist, dass sich in jedem Jahr eine recht regelmäßige Entwicklung wiederholt. Die *saisonale* Komponente zeigt ein Ansteigen des Umsatzes im 1. und 2. Quartal sowie Rückgänge im 3. und 4. Quartal. Um diese Saisoneinflüsse zu bereinigen und den Entwicklungs**trend** betrachten zu können, wird auch hier eine lineare Regression durchgeführt, deren Ergebnis auf der rechten Seite von Abbildung 13 zu erkennen ist. Die Steigung der Regressionsgerade ist positiv, d.h. der Umsatztrend geht über die drei Jahre nach oben, die Rückgänge in der zweiten Jahreshälfte sind nur *saisonbedingt*.

Glättung von Zeitreihenwerten

Es ist erkennbar, dass die Trendkomponente eine **Glättung** der schwankenden Zeitreihe darstellt und damit eine Referenzgröße für die Ermittlung der Saisoneinflüsse darstellen kann. Als weitere Methode zur Glättung von Zeitreihen sind *Gleitende Durchschnitte* (Moving Average) üblich, bei der aus jeweils vier Quartalswerten ein Mittelwert gebildet wird. Die hier verwendete Methode der linearen Trendfunktion hat dabei den Vorteil, dass der Trend für alle Beobachtungszeitpunkte gebildet werden kann und auch Prognosen über diesen Zeitraum hinaus vorgenommen werden können.

Diese Möglichkeit sowie die Saisonbereinigung sind in der Excel-Datei dargestellt.

Indexzahlen

Indexzahlen

Indexzahlen (oder Indizes) sind gewichtete arithmetische Mittelwerte aus Messzahlen. Bekannt ist etwa der Preisindex der Lebenshaltung, der durch das Statistische Bundesamt veröffentlicht wird. Hier werden die Preisentwicklungen aller Güter und Dienstleistungen, die Haushalte im Durchschnitt verbrauchen, zu einer mittleren Preissteigerung zusammengefasst. Dabei werden die (relativen) Mengen und daraus folgend die Ausgabenanteile für diese Produkte berücksichtigt.

Bei Zeitreihenanalysen werden anstelle der absoluten Werte oft **Reihen von Indexzahlen** verwendet. Diese werden dadurch gebildet, dass ein Basiszeitraum = 100 (Prozent) gesetzt wird und alle anderen Werte im Bezug auf dieses Basisjahr umgerechnet werden. Entwicklungen von Preisen, Umsätzen, Marktanteilen können damit für verschiedene Merkmale verglichen werden, die eine unterschiedliche absolute Höhe haben und deshalb (z.B. in einer Grafik) nicht „zusammen passen". Tabelle 10 zeigt hierfür ein Beispiel.

Tabelle 10 Entwicklung von Unternehmer- und Arbeitnehmereinkommen 1991-99

Jahr	Unternehmer	Arbeitnehmer	Unternehmer	Arbeitnehmer	Unternehmer	Arbeitnehmer
	in Mrd. Euro		Steigerungsrate		**Index 1991 = 100**	
1991	323,3	843,9	-	-	100	100
1992	328,6	914,0	1,6%	8,3%	101,6	108,3
1993	320,4	935,4	-2,5%	2,3%	99,1	110,8
1994	344,2	958,5	7,4%	2,5%	106,5	113,6
1995	366,0	992,6	6,3%	3,6%	113,2	117,6
1996	376,3	1.005,0	2,8%	1,3%	116,4	119,1
1997	399,0	1.007,9	6,0%	0,3%	123,4	119,4
1998	420,0	1.023,5	5,3%	1,6%	129,9	121,3
1999	418,5	1.045,4	-0,3%	2,1%	129,5	123,9
	absolute Steigerung 1991 bis 1999:		Durchschnittliche Steigerungsrate*:			
	95,3 Mrd. Euro	**201,4** Mrd. Euro	1,0328	1,0271		
			entspricht			
			3,28%	**2,71%**		

Spaltenbezeichnungen: Unternehmer = Einkommen aus Unternehmertätigkeit und Vermögen
Arbeitnehmer = Arbeitnehmer-Einkommen

* Geometrisches Mittel (da Durchschnitt aus Steigerungsraten; vgl. Punkt 1.3)

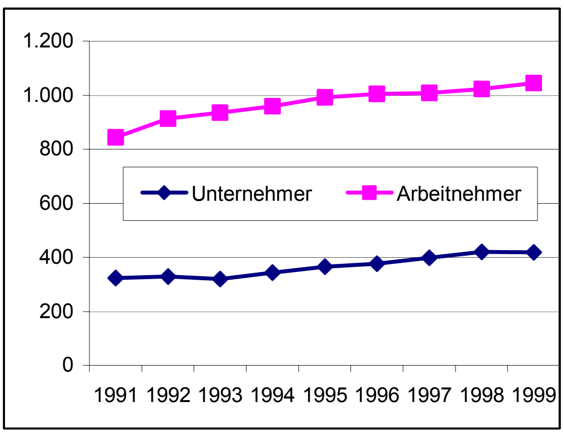
Quelle: eigene Berechnung aus: "Zahlen zur wirtschaftlichen Entwicklung der Bundesrepublik Deutschland" des IW Köln und Statistisches Jahrbuch 2000

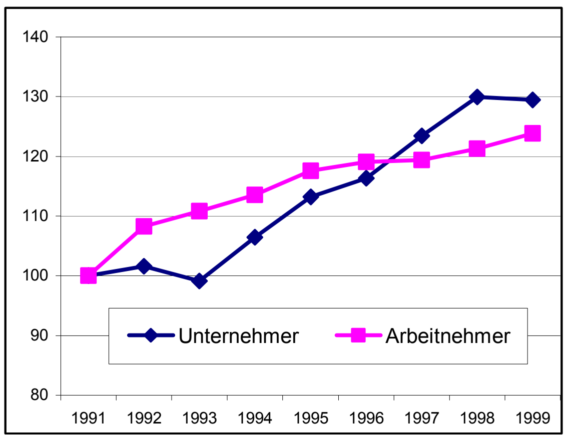
Anmerkung: Inzwischen (Zeitreihe 1995-2003) hat sich dieses Verhältnis umgekehrt, siehe Excel-Datei.

Das Beispiel in Tabelle 10 zeigt den Unterschied zwischen absoluter und relativer Entwicklung. Könnte auf Basis der ersten beiden Spalten formuliert werden, dass das Arbeitnehmereinkommen um mehr als den doppelten *Betrag* gestiegen ist, so zeigt sowohl die Betrachtung der Steigerungsraten als auch der Indexzahlen, die so umgerechnet wurden (Dreisatz), dass das Jahr 1991 den Wert 100,0 annimmt, das gegenteilige Ergebnis. Beide Maßzahlen ergeben, dass die Einkommen aus Unternehmertätigkeit und Vermögen mit 2,9 Prozent stärker gestiegen sind als die Arbeitnehmer-Einkommen mit 2,4 Prozent.

Das Errechnen von Steigerungsraten oder Indexzahlen hat somit den Vorteil der besseren Vergleichbarkeit. Auch lassen sich indizierte Werte besser in einer gemeinsamen Grafik darstellen, wie die folgende Abbildung illustriert.

Abbildung 14 Einkommensentwicklung absolut und als Indexzahlen





2 LITERATURHINWEISE UND WEITERE INFORMATIONEN

Aus der großen Menge guter Statistik-Bücher seien drei herausgegriffen, die jeweils praktische Einführungen in die betriebliche Anwendung darstellen:

Bourier, Günther: „Beschreibende Statistik“ und „Wahrscheinlichkeitsrechung und Schließende Statistik“, Wiesbaden 2003

Hujer, Reinhard: „Betriebliche Statistik“ - Methodenlehre der Betriebsorganisation - REFA Verband für Arbeitsstudien und Betriebsorganisation, München 1993

Puhani, Josef: ”Statistik - Einführung mit praktischen Beispielen”, Würzburg 2001

Spaß an der Statistik und trotzdem - oder eben deshalb - viele interessante Informationen rund um das Thema und seine Anwendungen finden sich in:

Krämer, Walter: „So lügt man mit Statistik” und: „Statistik verstehen”, München 2001

Praktische Arbeit am PC wird mit diesem sehr empfehlenswerten Buch erleichtert, da es neben dem theoretischen Hintergrund auch die praktische Umsetzung in Excel zeigt, Text und Hinweise auf CD-ROM mitliefert und vor allem durch den erfrischenden Schreibstil die statistische Arbeit zur Freude macht:

Monka, Michael und **Voß**, Werner: ”Statistik am PC - Lösungen mit Excel”, München 2002

Weitere Information zur (amtlichen) Statistik sowie interessante Datengrundlagen können im Internet gefunden werden. Wichtige Web-Adressen mit (betriebs-) wirtschaftlich relevanten Informationen finden sich z.B. auf meiner Webseite: http://www.fbw.hs-bremen.de/pschmidt - unter „links“.

3 SCHLAGWORTINDEX