Anhang

1. Einleitung

1.1 Allgemeines

Die **Statistik** ist ein Zweig der Mathematik, der sich mit der Sammlung, Zusammenstellung und Analyse von Zahlenmaterial – den statistischen Daten – für wissenschaftliche, soziale, politische und wirtschaftliche Zwecke beschäftigt. Zu seinem Themengebiet gehört u. a. auch die mathematische Untersuchung von so genannten Zufallsgrößen. Die Statistik arbeitet mit großen Datenmengen ohne Berücksichtigung ihrer inhaltlichen Bedeutung.

Gegenwärtig ist die Statistik ein verläßliches Instrument, um wirtschaftliche, politische, soziale, psychologische, medizinische, biologische und physikalische Daten genau zu beschreiben; zudem dient sie als Werkzeug, um solche Daten miteinander in Beziehung zu setzen und auszuwerten. Die Arbeit des Statistikers beschränkt sich nicht mehr auf das Sammeln und Auflisten von Daten, sondern sie dient der Interpretation der Information. Die Entwicklung der Wahrscheinlichkeitsrechnung erweiterte das Spektrum der statistischen Anwendungen. Viele Werte können durch Wahrscheinlichkeitsverteilungen näherungsweise bestimmt werden, und die Ergebnisse können zur Analyse statistischer Daten verwendet werden.

Zwei grundlegende Bereiche der Statistik sind die **deskriptive** (beschreibende) **Statistik** sowie die **schließende** oder auch **Inferenzstatistik.** Aufgabe der deskriptiven Statistik ist die Beschreibung von Datenmengen durch möglichst „kompakte" Kennwerte. Ein wesentliches Merkmal der deskriptiven Statistik ist, daß sich die Aussagen immer nur auf die vorliegenden Daten beziehen; eine Verallgemeinerung der Aussagen ist nicht zulässig. Um die Ergebnisse der Untersuchungen kleiner Datenmengen (Stichproben) auf große Datenmengen (Grundgesamtheiten) übertragen zu können, benötigt man die Verfahren der Inferenzstatistik.

Dieses Skript soll die für den Praktiker relevanten Verfahren der deskriptiven Statistik behandeln. Auf die Inferenzstatistik hingegen wird nur kurz am Ende des Skripts im Sinne einer allgemeinen Darstellung eingegangen.

1.2 Grundbegriffe der Statistik

Eine **Menge** ist die Gesamtheit gleichartiger Objekte (Individuen oder Ereignisse), an denen ein oder mehrere Merkmale beobachtet werden können. Jedes Objekt heißt Element der Menge. [1]

Beispiel: Menge M: Schulklasse, bestehend aus n Schülern {Schüler 1, Schüler 2, …, Schüler n}. Untersuchtes Merkmal: Körperlänge in cm, charakterisiert durch eine Variable X: $x_1 = 120$, $x_2 = 133$, …, $x_n = 124$.

[1] nach Clauß/Ebner

Die Werte x_i (i = 1, 2, …, n) der Zufallsvariablen X charakterisieren die quantitative Ausprägung des Merkmals an jedem Element der Menge M. Die Werte ändern sich von Element zu Element als Folge „zufälliger" Einflüsse. Die Messung der Körperlänge läßt sich folglich niemals exakt vorhersagen. Es ist aber möglich, von vornherein bestimmte Werte (z.B. 40 cm oder 350 cm) auszuschließen. Andererseits können bestimmte Werte mit mehr oder weniger großer Wahrscheinlichkeit erwartet werden, wenn aufgrund vorheriger Untersuchungen die Durchschnittswerte der Altersgruppe sowie deren Verteilung bereits bekannt sind.

Man unterscheidet zwischen **stetigen** (kontinuierlichen) und **diskreten** Zufallsvariablen. Eine stetige Zufallsvariable kann jeden beliebigen Wert eines bestimmten Intervalls der reellen Zahlengeraden annehmen, z.B. eine Körpergröße im Intervall von 40 - 250 cm, z.B. den Wert 172,456 cm. Eine diskrete Variable kann hingegen nur endlich viele Werte annehmen, z.B. die Werte der Würfe eines Würfels, die immer nur 1, 2, 3, 4, 5 oder 6 annehmen können.

Betrachtet man die Menge aller gleichartigen Objekte, z.B. alle Schüler einer bestimmten Schule, alle Schüler in einem bestimmten Land oder alle Schulkinder überhaupt, so spricht man von der **Grundgesamtheit**. Wird aus dieser Grundgesamtheit eine zufällige Auswahl von n Elementen entnommen, z.B. 100 Schüler aus Hamburg oder 500 Schulkinder im Alter von 10 Jahren, so erhält man eine **Stichprobe**. Die Anzahl n der in der Stichprobe auftretenden Zahlen oder Zufallswerte ist der **Umfang** der Stichprobe. Das primäre Ziel statistischer Untersuchungen besteht darin, von der gerade vorliegenden Stichprobe zu allgemeinen Aussagen über die Grundgesamtheit zu gelangen.

1.3 Skalentypen

Um überhaupt statistische Verfahren anwenden zu können, müssen die Merkmale **quantifizierbar** sein, d.h. sie müssen sich - wie auch immer - „sinnvoll" durch Zahlenwerte ausdrücken lassen.

Die einfachste Form der Quantifizierung ist die Feststellung der Häufigkeit von Ereignissen. Die hierfür erforderliche Operation ist das Zählen. Durch das Zählen wird eine bestimmte Anzahl diskreter Größen ermittelt. Beispiel: In Hamburger Unternehmen wird ermittelt, wie viele Mitarbeiter männlich und wie viele weiblich sind.

Voraussetzung für diese Form der Quantifizierung ist, daß Bedingungen definiert werden, unter denen die Objekte als gleichartig betrachtet werden können. Es würde beispielsweise wenig Sinn ergeben, Schokoladenriegel und Himmelskörper als eine Gruppe zusammenzufassen.

Werden nur zwei Merkmale ermittelt, dann spricht man von einer **dichotomischen** Gruppierung (männlich oder weiblich, verheiratet oder nicht verheiratet, etc.). Selbstverständlich lassen sich auch mehr als zwei Merkmale voneinander unter-

scheiden. So läßt sich z.B. ermitteln, welche Haarfarbe die Mitarbeiter haben (blond, braun, schwarz, etc.).

Die oben dargestellte Form der Quantifizierung beruht auf der Klassifizierung von Objekten, d.h. deren Zuordnung zu Objektklassen nach beobachteten Merkmalen. Für eine Variable werden qualitativ verschiedene Klassen definiert. Anschließend wird gezählt, wie viele Objekte zu jeder Klasse gehören. Diese Form der Quantifizierung wird als **Nominalskalierung** bezeichnet. Für die Anwendung von Nominalskalen ist es lediglich erforderlich, anzugeben, in welcher Hinsicht die Objekte als gleichartig anzusehen sind, um dann einer Klasse zugeordnet werden zu können.

Wenn es darüber hinaus Informationen gibt, wie sich die Klassen zueinander verhalten, so daß man sagen kann, die eine Klasse sei „größer", „kleiner" oder „gleich" einer anderen, so kann man eine weitere Form der Quantifizierung benutzen, die sog. **Ordinalskalierung**. Mit der Ordinalskala können Aussagen über die Unterschiede der Klassen im Sinne einer Rangordnung getroffen werden. Es ist jedoch hier nicht möglich zu sagen, wie stark die Differenzen zwischen den Objekten sind.

Beispiel: Man kann z.B. aussagen, ob ein Mitarbeiter im Vergleich zu einem anderen „mehr" oder „weniger" sympathisch ist. Aufgrund dieser Aussagen kann eine Rangfolge erstellt werden. Nicht möglich jedoch sind Aussagen in der Art von „Mitarbeiter A ist 4,6 mal sympathischer als Mitarbeiter B".

Für viele andere Fragestellungen interessiert jedoch gerade die Größe der Differenz. In diesem Fall brauchen wir eine Skala mit einer **metrischen** Struktur. Die Form der Quantifizierung auf diesem Niveau heißt **Intervallskalierung**. Die Abstände zwischen den Skalenwerten sind konstant.

Beispiel: Die Messung der Temperatur. Der Abstand zwischen 10° C und 20° C ist doppelt so groß wie der Abstand zwischen 55° C und 60° C.

Werden die Meßwerte nicht nur im Hinblick auf ihre qualitative Gleichartigkeit (Nominalskala), relative Größenunterschiede (Ordinalskala) und Intervallgleichheit (Intervallskala), sondern darüber hinaus auch in bezug auf den absoluten Nullpunkt bestimmt, so spricht man von der **Verhältnisskalierung**.

Beispiel: Es ist nicht richtig, zu sagen, daß es bei 20° C doppelt so warm sei wie bei 10° C (der Nullpunkt der Celsius-Skala ist willkürlich mit dem Gefrierpunkt des Wassers festgesetzt worden). Man kann jedoch sagen, daß ein Mensch mit 200 cm Körperlänge doppelt so lang ist, wie ein Mensch mit 100 cm Körperlänge, weil es hier einen festen Nullpunkt gibt.

Von der Art der Skalierung hängen sehr wichtige Entscheidung über die Anwendung statistischer Verfahren ab.

Im folgenden werden die vier Skalentypen, deren Voraussetzungen, zulässige Transformationen, Verteilungscharakteristika sowie Korrelationsverfahren tabellarisch aufgelistet:

Übersicht über Skalierungsformen[2]

Skalentyp	Voraussetzungen	zulässige Transformationen	zulässige Verteilungscharakteristika	zulässige Korrelationsverfahren	Beispiel
Nominalskala	Bestimmbarkeit der Gleichheit oder Ungleichheit von Elementen, ihrer Zugehörigkeit zu einer Klasse	Permutation (Vertauschen), Umbenennung	absolute Häufigkeiten, relative Häufigkeiten (Prozentwerte), Modalwert	Kontingenzkoeffizienten	Zählung von Schülern nach Haarfarbe
Ordinalskala	zusätzlich: Bestimmbarkeit von Größer-Kleiner-Unterschieden und entsprechende Ordnung der Elemente in eine Rangfolge	$y = f(x)$, wobei $f(x)$ eine monotonwachsende Funktion ist	zusätzlich: kumulierte Häufigkeiten, Rangpositionen, Prozentrangwerte, Zentile, Median, Quartile	Rangkorrelationskoeffizienten	Bundesligatabelle
Intervallskala	zusätzlich: Bestimmbarkeit von Einheiten (gleichen Intervallen) und Festlegung eines *relativen* Nullpunkts	lineare Transformation $y = a + bx$, $b > 0$	zusätzlich: arithmetisches Mittel, Standardabweichung, Schiefe	Maßkorrelationskoeffizient, Regressionskoeffizient	Temperaturskala
Verhältnisskala	zusätzlich: Bestimmbarkeit gleicher Proportionen, Existent eines *absoluten* Nullpunkts	Ähnlichkeitstransformation $y = cx$, $c > 0$	zusätzlich: geometrisches Mittel		Kontostand

Die Intervall- und die Verhältnisskala sind metrische Skalen.

Transformationen sind mathematische Operationen, um Werte einer bestimmten Skala in eine andere zu überführen. Transformationen sind zulässig, wenn der Informationsgehalt in beiden Skalen beibehalten wird und sich die Transformation umkehren läßt. Eine solche Transformation wäre z.B. die Überführung der Temperaturen, die in °C gemessen wurden, in Temperaturen in °F. Die Zahlenwerte werden mit einem bestimmten Faktor multipliziert und anschließend wird eine Konstante hinzu addiert. Eine solche Form der Transformation heißt **lineare Transformation**.

[2] nach Gutjahr, W., Zur Skalierung psychischer Eigenschaften, „Probleme und Ergebnisse der Psychologie“, Heft 23, 1968

1.4 Mathematische Symbole

Zufallsvariablen werden mit lateinischen Großbuchstaben bezeichnet; üblicherweise werden Buchstaben aus dem Ende des Alphabets (X, Y, Z) verwendet. Die Werte dieser Variablen werden mit lateinischen Kleinbuchstaben bezeichnet und durch Indizes voneinander unterschieden.

Beispiel: Variable X, Variablenwerte x_1, x_2, …, x_n. Für einen allgemeinen Variablenwert wird häufig der Index i verwendet; also ist x_i der allgemeine Variablenwert der Variablen X.

Sollen alle Variablenwerte einer Variablen addiert werden, so müßte man schreiben: $x_1 + x_2 + x_3$ + usw. Um sich dies zu ersparen, verwendet man das Summenzeichen $\sum$. Das Summenzeichen enthält eine „Fußzeile“, die angibt, von wo an addiert wird, sowie eine „Kopfzeile“, die angibt, wie das letzte Element der Summe heißt.

Beispiel: Für die Summe $x_1 + x_2 + x_3$ schreibt man $\sum_{i=1}^{3} x_i$. Dieser Ausdruck wird als „Summe aller x_i für i = 1 bis 3“ gelesen.

Der allgemeine Fall, das Summieren aller x_i, wird folglich geschrieben als $\sum_{i=1}^{n} x_i$. Wenn dieser allgemeine Fall, in dem alle Werte addiert werden, verwendet wird, dann werden häufig „Kopf-“ und „Fußzeile“ des Summenzeichens weggelassen, also nur $\sum x_i$ geschrieben.

Für das Rechnen mit dem Summenzeichen gelten folgende Sachverhalte:

- $\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i$
- $\sum_{i=1}^{n} a = na$
- $\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$

Hingegen ist zu beachten

- $\sum_{i=1}^{n} (x_i y_i) \neq \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$

Aufgaben:

a) Welches Skalierungsniveau liegt für folgende „Messungen“ vor? Begründen Sie Ihre Aussage!

   Telefonverzeichnis – Schulnoten – Stärke von Kopfschmerzen (schwach/mittel/stark) – Messung von Stromspannung in V – Reihenfolge der Lieblingsspeisen – Aktienkurse – Intelligenzquotient – Tachometeranzeige – Temperatur in Kelvin

b) Zeigen Sie anhand einer linearen Transformation, daß Temperaturskalen in °C und in °F tatsächlich intervallskaliert sind. (Hinweis: Wasser gefriert bei 32 °F und kocht bei 212 °F)

c) Beweisen Sie die o.g. vier Sachverhalte zum Rechnen mit dem Summenzeichen.

2. Univariate Verteilungen

Üblicherweise sind die Meßwerte, die erhoben werden, zunächst ungeordnet und dadurch nicht sehr übersichtlich. Die Liste dieser nicht sortierten Meßwerte wird als Urliste bezeichnet.

Beispiel: Umsätze von 20 Unternehmen in Mio. EUR:

11	13	9	5	19	11	8	12	11	5
5	12	13	16	7	16	12	11	9	13

Um etwas Übersicht in die Daten zu bringen, empfiehlt es sich, diese zunächst der Größe nach zu sortieren. Diese Anordnung wird oft als primäre Tafel bezeichnet:

5	5	5	7	8	9	9	11	11	11
11	12	12	12	13	13	13	16	16	19

Anhand dieser Liste kann man z.B. schon erkennen, daß einige Werte häufiger vorkommen. Weiterhin lassen sich sofort der niedrigste und der höchste Wert erkennen. Die Differenz zwischen diesen beiden Werten ist in diesem Fall 19 – 5; diese Differenz heißt Variationsbreite und wird häufig mit dem Buchstaben v abgekürzt.

Um noch mehr Übersicht in die Daten zu bringen, werden diese in einer sog. Häufigkeitstabelle dargestellt:

x_i (Umsatz)	5	6	7	8	9	10	11	16	13	14	15	16	17	18	19
f_i	3	0	1	1	2	0	4	3	3	0	0	2	0	0	1

2.1 Gruppierung der Daten

Sind viele (verschiedene) Meßwerte vorhanden, so wird die Häufigkeitstabelle leicht unübersichtlich. Es empfiehlt sich in solchen Fällen, mehrere Meßwerte zu **Klassen** zusammenzufassen. Eine Klasse ist die Menge sämtlicher Meßwerte, die innerhalb festgelegter Grenzen liegen. Die **Klassengrenzen** werden durch den kleinsten und den größten Wert einer Klasse gebildet. Die **Klassenmitte** ist das arithmetische Mittel aus beiden Klassengrenzen. Die **Klassenbreite** ist (bei diskreten Variablen) die Anzahl der in der Klasse zusammengefaßten Werte. **Offene Klassen** haben im Gegensatz zu **geschlossenen** Klassen keine Unter- oder Obergrenze.

Die oben dargestellte Häufigkeitstabelle hat eine Klassenbreite von h = 1. Bei h = 3 bzw. h = 5 ergibt sich folgendes Bild:

x_i	**5-7**	**8-10**	**11-13**	**14-16**	**17-19**
f_i	4	3	10	2	1

bzw.

x_i	**5-9**	**10-14**	**15-19**
f_i	7	10	3

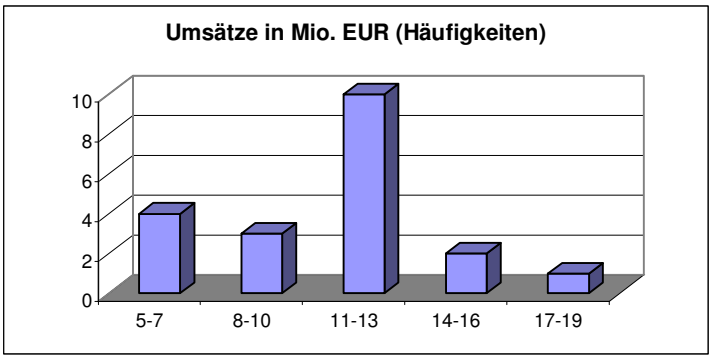
Oft werden nicht nur die tatsächlichen (absoluten) Häufigkeiten, sondern auch die sog. **kumulierten Häufigkeiten** betrachtet. Kumulierte Häufigkeiten sind die Häufigkeiten, die sich durch Summieren aller Häufigkeiten aller vorangegangenen Klassen sowie der betrachteten Klasse ergeben. **Relative Häufigkeiten** bezeichnen die Anteile der absoluten Häufigkeiten an der Gesamtzahl der Meßwerte. Diese Anteile werden üblicherweise als Werte zwischen Null und Eins angegeben; durch Multiplikation mit 100 kann man diese Werte als Prozentwerte angeben. Betrachtet man nun schließlich die kumulierten relativen Häufigkeiten, dann kann man leicht erkennen, wieviel der Meßwerte z.B. in den unteren Klassen oder bis zur Mitte, etc. liegen. Die Erweiterung um kumulierte und relative Häufigkeiten ergibt für die Klassenbreite h = 3:

x_i	**5-7**	**8-10**	**11-13**	**14-16**	**17-19**
f_i	4	3	10	2	1
f_i **kum.**	4	7	17	19	20
f_i **rel.**	0,2	0,15	0,5	0,1	0,05
f_i **rel.kum. %**	20%	35%	85%	95%	100%

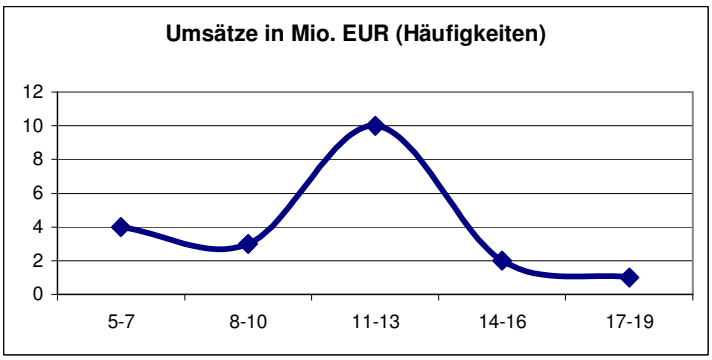
2.2 Graphische Darstellung

Noch übersichtlicher als die Zusammenfassung der Werte in Klassen ist die graphische Darstellung der Häufigkeitsverteilungen. In der graphischen Darstellung werden die Häufigkeiten eines Meßwerts durch eine Fläche dargestellt. Inhaltlich sollen graphische Darstellung und Tabelle dieselbe Information vermitteln.

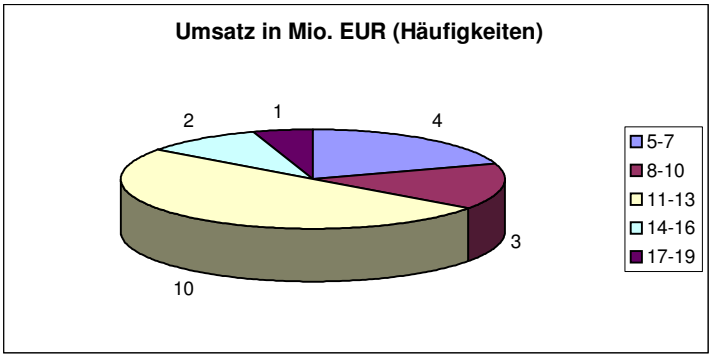
Die einfachste und gebräuchlichste Darstellung ist das **Histogramm**. Im Histogramm wird die Häufigkeit jedes Meßwerts durch eine Fläche abgebildet, die über dem Wert zu bilden ist. Für unser Beispiel ergibt sich folgendes Histogramm:



Ebenfalls gebräuchlich ist die Darstellung der Daten mit Hilfe eines **Polygonzugs**. Diese Darstellung eignet sich besonders, um mehrere Verteilungen in derselben Graphik darzustellen:

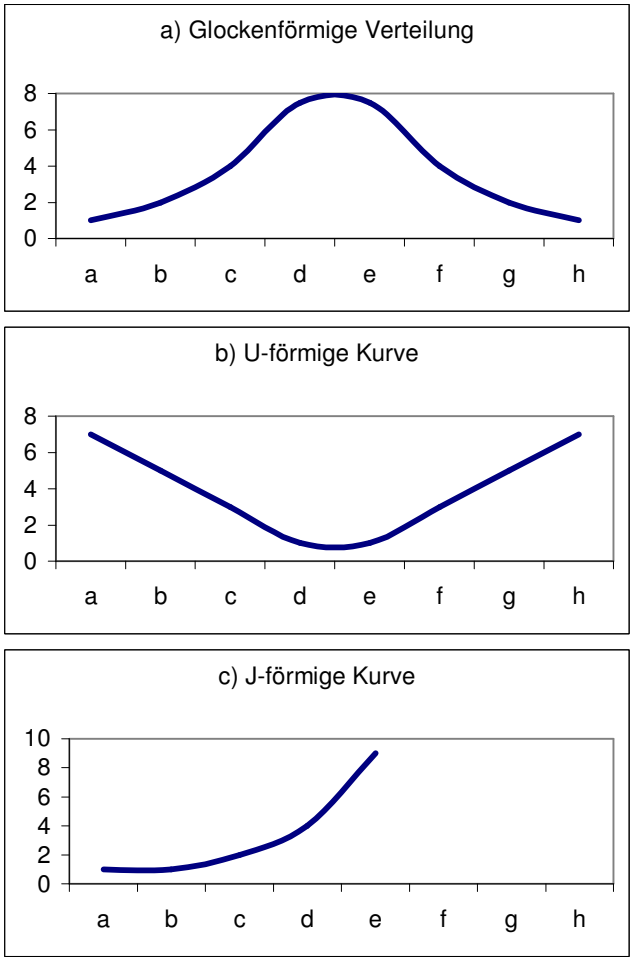


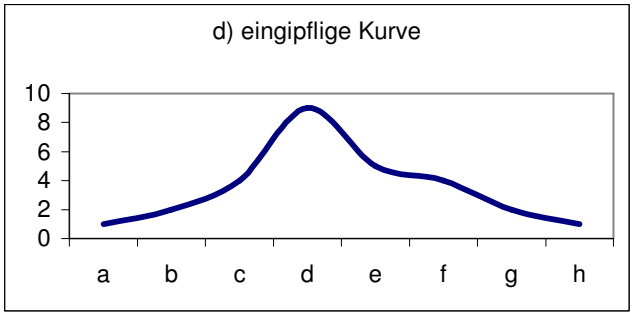
Weiterhin gebräuchlich ist die Darstellung der Daten in einem sog. Kreisdiagramm:

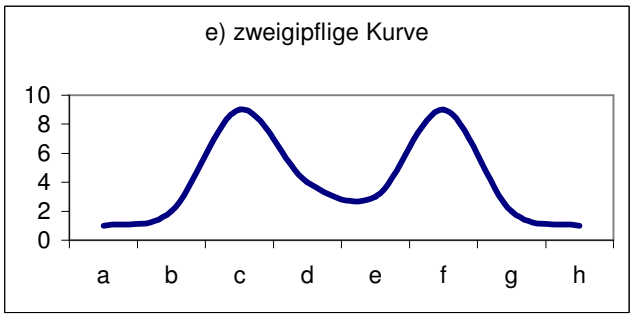


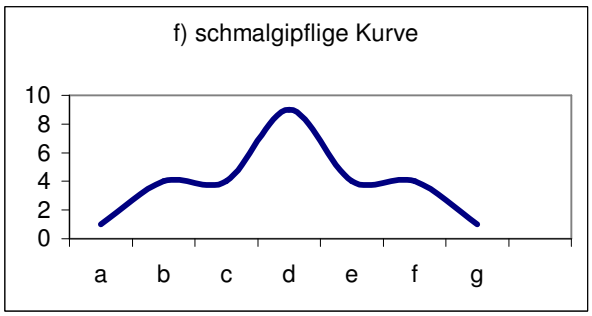
2.3 Typische Formen der Verteilung

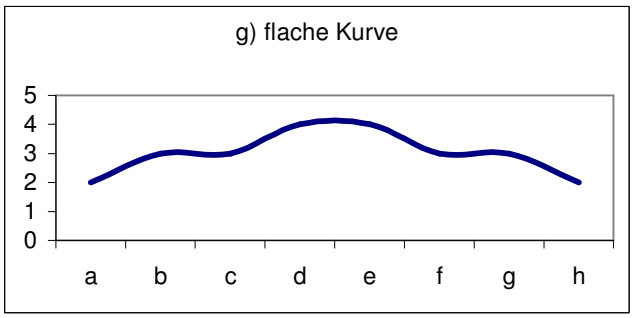
In der Praxis kommen immer wieder bestimmte Formen der Verteilung vor. In der Darstellung als Polygonzug sind die häufigsten sieben Formen:











Die glockenförmige Verteilung (a) kommt sehr häufig vor, wenn stetige Werte gemessen werden. Sie ist Ausdruck dafür, daß mittlere Meßwerte deutlich häufiger auftreten als extreme[3]. Die U-förmige Kurve (b) drückt das Gegenteil aus: extreme Werte kommen hier häufiger vor als mittlere. J-förmige Kurven (c) kommen vor, wenn ein Extrem häufig vorkommt. Im Gegensatz zu den beiden vorangegangenen Verteilungen ist diese Verteilung asymmetrisch. Der Gipfel ist im Gegensatz zur glockenförmigen Verteilung leicht verschoben. Die zweigipflige Kurve (e) hat zwei Häufigkeitsmaxima in der Mitte der Verteilung. In der schmalgipfligen Kurve (f) gibt es sehr wenige

[3] Mehr dazu in Abschnitt 4.

Werte, hier ebenfalls im mittleren Bereich. Die flache Kurve (g) schließlich ist Ausdruck einer geringen Häufung extremer Werte. Das gemessene Merkmal ist stark gestreut.

3. Statistische Kennwerte

Statistische Kennwerte (auch: statistische Maßzahlen, Kennziffern oder Statistiken) dienen dazu, Datenmengen so kompakt wie möglich zu beschreiben. Hierbei gehen Einzelinformationen verloren; dennoch soll die Datenmenge bzw. Verteilung unter Verwendung der Kennwerte möglichst gut beschrieben werden. Man unterscheidet zwischen zwei Gruppen solcher Kennwerte: der Gruppe der **Mittelwerte** und der Gruppe der **Streuungswerte**. Mittelwerte sagen etwas über den „Durchschnitt" der Verteilung aus, während Streuungswerte aussagen, wie „verschieden" die einzelnen Werte der Verteilung sind.

3.1 Mittelwerte

Es gibt unterschiedliche Mittelwerte, die je nach Art der Skalierung der Daten angewendet werden:

- Der **Modalwert,** häufig mit D bezeichnet. Er gibt den häufigsten Wert der Verteilung an und kann bereits für lediglich nominalskalierte Daten verwendet werden.
- Der **Median**, häufig mit Z bezeichnet. Er gibt den mittleren Wert einer Verteilung von Daten an, die der Größe nach sortiert sind. Die Berechnung des Medians setzt ordinalskalierte Daten voraus.
- Das **arithmetische Mittel**, im allgemeinen mit $\bar{x}$ bezeichnet, setzt metrisches Skalenniveau voraus. Das arithmetische Mittel gibt den Durchschnittswert einer Verteilung an. Die Abweichung aller Werte vom arithmetischen Mittel ist minimal.
- Das **geometrische Mittel**, häufig mit G bezeichnet, kann für Daten, die verhältnisskaliert sind, verwendet werden.

3.1.1 Der Modalwert

Der Modalwert ist der in der Verteilung am häufigsten vorkommende Meßwert. Voraussetzung ist, daß sich die Häufigkeiten der Meßwerte voneinander unterscheiden. Beispiel: Gegeben ist die Verteilung 1, 2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 6, 7, 8, 9. Der Wert 4 kommt am häufigsten vor, nämlich fünfmal. 4 ist also der Modalwert dieser Verteilung.

Gibt es in der Verteilung zwei Maxima (man spricht hier von einer zweigipfligen oder bimodalen Verteilung), dann gibt es zwei Modalwerte. Beispiel: Gegeben ist die Verteilung 1, 2, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8. Die Werte 3 und 6 kommen am häufigsten vor, nämlich je dreimal. Die Modalwerte dieser Verteilung sind 3 und 6.

Bei gruppierten Daten ist der Modalwert die Mitte der Modalklasse (der Klasse mit der größten Besetzungszahl).

Beispiel: Gegeben ist die Verteilung

Wert	**3 - 5**	**6 - 8**	**9 - 11**
Häufigkeit	4	7	3

Die Modalklasse ist die Klasse „6 – 8“, weil diese Werte am häufigsten, nämlich siebenmal vorkommen. Die Mitte dieser Klasse ist 7, also ist der Modalwert sieben.

3.1.2 Der Median

Der Median ist der Wert in der nach ihrer Größe geordneten Rangreihe der Meßwerte, der die Reihe halbiert. Beispiel: Gegeben ist die Verteilung 2, 2, 3, 5, 7, 9, 10, 10, 12 mit neun Meßwerten. Der Wert 7 halbiert die Reihe, da genau vier Werte unterhalb und vier Werte oberhalb der 7 liegen.

In der Verteilung 2, 2, 3, 5, 7, 9, 10, 10, 12, 13 liegt mit zehn Werten eine gerade Anzahl von Werten vor. In der Mitte liegen die Werte 7 und 9. Der Median ist hier die Mitte zwischen diesen beiden Werten, nämlich 8.

In einer Verteilung, in der die Daten in Klassen angeordnet sind, läßt sich der Median ebenfalls berechnen:

- Man bilde die kumulierten Häufigkeiten f_c
- Man bestimme n/2. Der Median befindet sich in der Klasse, in der sich dieser Wert befindet.
- Man geht von einer Gleichverteilung der Werte innerhalb dieser Klasse aus.
- Der Median läßt sich berechnen durch $Z = x_u + \frac{\frac{n}{2} - \sum_{i=1}^{m-1} f_i}{f_m} h$, wobei x_u die untere Grenze der entsprechenden Klasse, m die Klasse selbst und h die Klassenbreite ist.

Beispiel: Gegeben ist folgende Verteilung

Wert	**3 - unter 5**	**5 - unter 7**	**7 - unter 9**
Häufigkeit	3	6	2
kum. Häuf.	3	9	11

n ist 11, also ist n/2 = 5,5. In der Klasse „5 – unter 7“ befindet sich der vierte bis neunte Wert, also der gesuchte „5,5-te“. Die Untergrenze dieser Klasse ist $x_u = 5$, es gibt m = 3 Klassen, in der zu untersuchenden Klasse befinden sich $f_m = 6$ Werte, die

Breite dieser Klasse ist h = 2. Also ist der gesuchte Median $Z = x_u + \frac{\frac{n}{2} - \sum_{i=1}^{m-1} f_i}{f_m} h = 5 + \frac{5{,}5 - 3}{6} 2 = 5 + 0{,}4166667 \cdot 2 = 5{,}8333$.

Häufig interessiert nicht nur der mittlere Wert einer Verteilung, sondern z.B. auch der Wert, der die unteren 25% einer Verteilung vom Rest der Verteilung trennt, bzw. die oberen 25%. Würde man so die Verteilung nicht in zwei Teile (wie beim Median), sondern in vier Teile (25%, 50%, 75%) „zerlegen", so spricht man von **Quartilen**. Die Quartile werden mit Q_i bezeichnet, so daß Q_1 den ersten 25% der Verteilung entspricht, usw. Folglich ist Q_2 gleich dem Median Z. Die Berechnung erfolgt nach der o.g. Formel, wobei natürlich n/2 durch n/4 ersetzt wird. Ebenfalls gebräuchlich ist das „Zerlegen" der Verteilung in zehn Teile (Dezile, D_1, …, D_9) bzw. 100 Teile (Zentile, C_1, …, C_{99}). Es gilt dann $Z = Q_2 = D_5 = C_{50}$. Allgemein werden diese Werte (Median, Quartile, Dezile, Zentile) als **Quantile** bezeichnet.

Zwischen dem 1. und dem 3. Quartil liegen die 50% der mittleren Werte. Dieser Bereich ist interessant und wird häufig betrachtet, da die „extremen" Werte hierdurch ausgeblendet werden.

3.1.3 Das arithmetische Mittel

Das arithmetische Mittel von Meßwerten ist deren Summe, geteilt durch ihre Anzahl. Das arithmetische Mittel wird mit $\bar{x}$ („x quer") bezeichnet.

Liegen n Werte x_1, …, x_n vor, dann ist das arithmetische Mittel
$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Beispiel: Gegeben ist die Verteilung 2, 2, 3, 5, 7, 9, 10, 10, 12. Das arithmetische Mittel ist $\frac{1}{9}$ (2+2+3+5+7+9+10+10+12) = $\frac{1}{9} \cdot 60 = 6{,}6667$.
Kommen Meßwerte mehr als einmal vor, dann sind f_1, …, f_k die Häufigkeiten der Meßwerte x_1, …, x_k. In diesem Fall ist das arithmetische Mittel $\bar{x} = \frac{f_1 x_1 + \ldots + f_k x_k}{n} = \frac{1}{n} \sum_{i=1}^{k} f_k x_k$.

3.1.4 Das gewogene arithmetische Mittel

Es kommt vor, daß Erhebungen, die aus unterschiedlichen Stichproben stammen, zusammengefaßt werden müssen. Die einzelnen Stichproben können dabei einen unterschiedlichen Umfang haben. Liegen k Stichproben vor, bezeichnet n_1, …, n_k den Umfang der einzelnen Stichproben und $\bar{x}_1$, …, $\bar{x}_k$ das arithmetische Mittel der einzelnen Stichproben, dann ist

$$\bar{x}_g = \frac{\bar{x}_1 n_1 + \ldots + \bar{x}_k n_k}{n_1 + \ldots + n_k} = \frac{\sum_{i=1}^{k} \bar{x}_i n_i}{\sum_{i=1}^{k} n_i}.$$

Beispiel: Gegeben sind zwei Stichproben vom Umfang $n_1 = 20$ und $n_2 = 50$. Die Mittelwerte sind $\bar{x}_1 = 15$ und $\bar{x}_2 = 10$. Das gewogene arithmetische Mittel ist $\bar{x}_g = \frac{\sum_{i=1}^{k} \bar{x}_i n_i}{\sum_{i=1}^{k} n_i} =$

$$\frac{15 \cdot 20 + 10 \cdot 50}{20 + 50} =$$

$$\frac{800}{70} = 11{,}4286.$$

3.1.5 Die Eigenschaften des arithmetischen Mittels

- Addiert man zu allen Werte einer Verteilung eine Konstante k, so vergrößert sich das arithmetische Mittel dieser Verteilung um den Wert k.
- Die Summe aller Abweichungen der Meßwerte von ihrem arithmetischen Mittel ist Null. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$.
- Die Summe der Quadrate der Abweichungen aller Meßwerte von Ihrem Mittelwert ist kleiner als die Summe der Quadrate der Abweichungen aller Meßwerte von jedem anderen Wert der Verteilung. $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \min$.

3.1.6 Vergleich der Mittelwerte

Alle oben besprochenen Mittelwerte haben die Eigenschaft, empirische Verteilungen durch eine Maßzahl zu charakterisieren. Am einfachsten läßt sich das arithmetische Mittel berechnen. Das arithmetische Mittel hat gegenüber dem Median den Vorteil, daß es

- leicht zu berechnen ist
- zuverlässige Schätzwerte (bei genügend großer Stichprobe) für den Mittelwert der Grundgesamtheit liefert
- eine algebraische Funktion jedes Meßwertes ist

Nicht berechnet werden sollte das arithmetische Mittel hingegen, wenn

- eine zweigipflige Verteilung vorliegt

- die Verteilung offene Klassen hat
- die Daten nicht auf metrischem Niveau vorliegen
- die Verteilung extrem asymmetrisch ist

Der Median sollte hingegen berechnet werden, wenn

- die Verteilung offene Klassen aufweist
- die Verteilungen stark asymmetrisch sind
- die Daten lediglich ordinalskaliert sind

Der Modalwert kann verwendet werden, um mehrgipflige Verteilungen zu kennzeichnen.

Sind die Verteilungen symmetrisch, dann gilt arithmetisches Mittel $\bar{x} \cong$ Median Z $\cong$ Modalwert D. Sind die Verteilungen linksschief, dann gilt D < Z < $\bar{x}$, sind die Verteilungen rechtsschief, dann ist $\bar{x}$ < Z < D.

Aufgaben

a) Gegeben sind folgende Meßwerte (z.B. die Zahlen von Beschäftigten in Unternehmen):

12, 13, 14, 7, 9, 5, 5, 1, 5, 6, 11, 12, 11, 9, 4, 3, 7, 3, 1, 5, 7, 1, 7, 5, 13.

- Bestimmen Sie Modalwert, Median und arithmetisches Mittel.
- Wie verändert sich das arithmetische Mittel, wenn alle Meßwerte mit zwei multipliziert werden und fünf addiert wird?

b) Gegeben ist folgende Verteilung des Einkommens in EUR:

Wert	**bis 1000**	**bis 2000**	**bis 3000**	**bis 4000**	**bis 5000**
Häufigkeit	12	37	25	17	9

- Bestimmen Sie Modalwert, Median und arithmetisches Mittel.
- Was können Sie durch das Verhältnis von Modalwert, Median und arithmetischem Mittel über die Schiefe der Verteilung aussagen?
- Berechnen Sie, in welchem Bereich sich die mittleren 50% der Verteilung befinden ($Q_3 - Q_1$).
- Zu dieser Verteilung kommt ein weiterer Wert, EUR 99.500,- (Klasse 99.000 – 100.000), hinzu. Wie verändern sich Median und arithmetisches Mittel?

3.2 Streuungsmaße

Obwohl Verteilungen in ihren Mittelwerten übereinstimmen, können sie dennoch stark voneinander abweichen. Man betrachte hierzu folgende vier Verteilungen, die alle das arithmetische Mittel 100 haben:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	$\bar{x}$
Verteilung I	97	98	99	100	101	102	103	100
Verteilung II	70	80	90	100	110	120	130	100
Verteilung III	1	30	70	100	130	170	199	100
Verteilung IV	1	98	99	100	101	102	199	100

Trotz gleicher Mittelwerte sind die Werte der Verteilungen offensichtlich unterschiedlich **gestreut.** Es soll im folgenden untersucht werden, wie sich ein Maß für die Streuung angeben läßt.

3.2.1 Die Variationsbreite

Je stärker die extremen Meßwerte voneinander abweichen, desto größer ist die Streuung. Das einfachste Maß, die Streuung zu bestimmen, ist daher die Differenz zwischen dem größten und dem kleinsten Meßwert: $v = x_{max} - x_{min}$.

Folglich wäre die Variationsbreite für die Verteilung I $v_1 = 103 - 97 = 6$. Die Variationsbreite ist also unkompliziert zu bestimmen.

Auf der anderen Seite beschränkt sich die Information lediglich auf die Extremwerte. Die Variationsbreite sagt nichts über die Verteilung der übrigen Werte aus. Dies wird deutlich, wenn man die Verteilungen III und IV vergleicht; während die Werte der Verteilung III in etwa gleichen Abstand zueinander haben, scharen sich die Werte der Verteilung IV – ausgenommen die beiden Extremwerte – um das arithmetische Mittel.

3.2.2 Die durchschnittliche Abweichung

Genauere Aussagen über die Streuung aller Werte einer Verteilung werden geliefert, wenn man die Abstände jedes einzelnen Wertes vom Mittelwert zugrunde legt. Extreme Werte haben einen großen Abstand, Werte, die eng um das arithmetische Mittel gestreut sind, einen kleinen Abstand und das arithmetische Mittel selbst hat den Abstand Null. Die einzelnen Abstände der Meßwerte vom arithmetischen Mittel sind $x_i - \bar{x}$. Es erscheint naheliegend, die Abstände aller Meßwerte vom Mittel zu summieren, um das gewünschte Maß zu erhalten. Allerdings kann dies nicht ohne weiteres getan werden, da die Summe aller Abstände der Meßwerte von ihrem arithmetischen Mittel bekanntlich Null ist. Um dies zu vermeiden, werden nicht die tatsächlichen Abweichungen, sondern die **absoluten Beträge** der Abweichungen betrachtet, also $\left|x_i - \bar{x}\right|$. Die durchschnittliche Abweichung e ist folglich definiert als das arithmetische

Mittel aus den absoluten Beträgen der Abweichungen aller Meßwerte einer Verteilung von ihrem arithmetischen Mittel:

$$e = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n} = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \bar{x} \right| .$$

Wesentlich verbreiteter als die durchschnittliche Abweichung e sind jedoch die Varianz und die Standardabweichung.

3.2.3 Varianz und Standardabweichung

Der Ansatz zur Ermittlung von Varianz und Standardabweichung ist dem Ansatz zur Ermittlung der durchschnittlichen Abweichung sehr ähnlich, allerdings wird nicht der Betrag der Abweichungen, sondern das Quadrat der Abweichungen $(x_i - \bar{x})^2$ zugrunde gelegt. Durch Quadrieren heben sich negative Vorzeichen auf. Ein weiterer Vorteil des Quadrierens besteht darin, daß kleine Abweichungen $|x_i - \bar{x}| < 1$ noch kleiner, Abweichungen > 1 noch größer werden.

Die Varianz s^2 ist die Summe der Abweichungsquadrate aller Meßwerte einer Verteilung von ihrem arithmetischen Mittel, dividiert durch die um Eins verminderte Anzahl der Messungen.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Die Standardabweichung s ist die Quadratwurzel aus der Varianz

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} .$$

Gegenüber der durchschnittlichen Abweichung weisen Standardabweichung und Varianz erhebliche Vorteile auf. Sie werden von den extremen Werten der Verteilung kaum beeinflußt. Weiterhin gehen alle Werte der Verteilung (nicht nur die Extremwerte) in sie ein. Schließlich sind sie Voraussetzung, um mit statistischen Prüfverfahren die Werte der Grundgesamtheit zu berechnen.

Für die oben betrachteten Verteilungen I – IV ergeben sich folgende Streuungsmaße:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	$\bar{x}$	v	e	s^2	s
Verteilung I	97	98	99	100	101	102	103	100	6	1,71	4,67	2,16
Verteilung II	70	80	90	100	110	120	130	100	60	17,14	466,67	21,60
Verteilung III	1	30	70	100	130	170	199	100	198	56,86	5.200,33	72,11
Verteilung III	1	98	99	100	101	102	199	100	198	29,14	3.268,67	57,17

Kann man von einer **Normalverteilung** der Meßwerte ausgehen, dann kann man unter Verwendung vom arithmetischen Mittel und der Streuungsmaße Aussagen über bestimmte Bereiche treffen, welcher Anteil der Meßwerte sich in ihnen befindet. Hierzu im nächsten Abschnitt mehr.

Aufgaben:

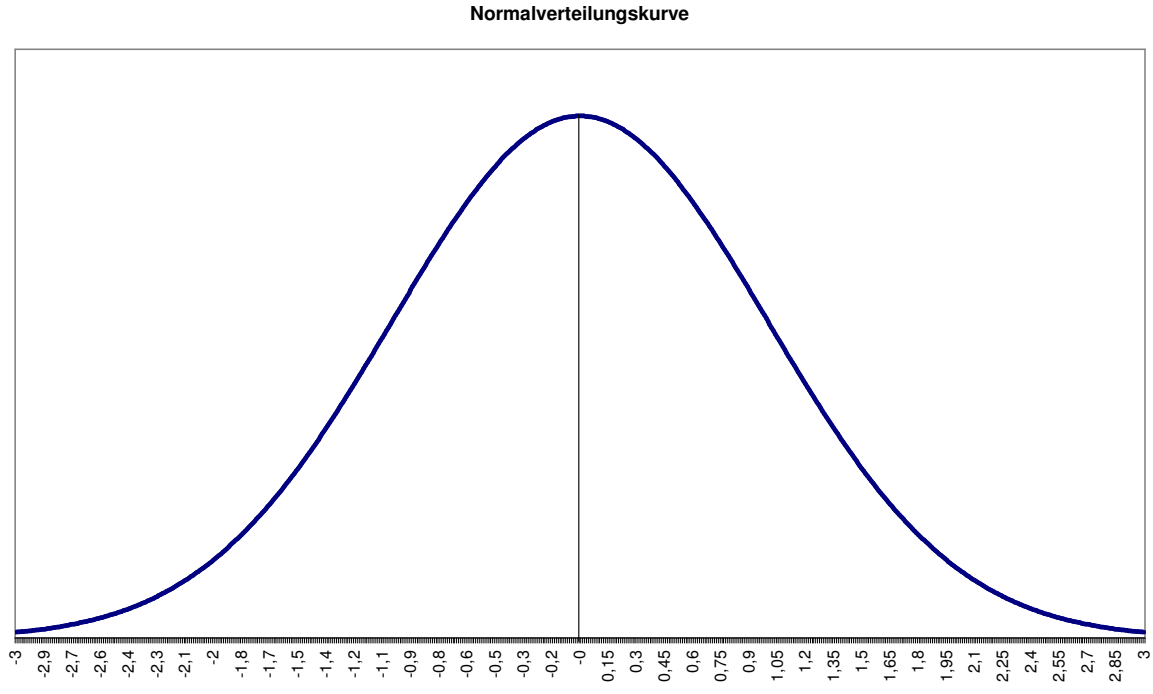
Berechnen Sie alle Streuungsmasse (Variationsbreite, durchschnittliche Abweichung, Varianz und Standardabweichung) für folgende Verteilung:

1, 1, 2, 2, 3, 5, 6, 8, 10, 11, 11, 11, 12, 14, 18, 22

Vergleichen Sie die unterschiedlichen Streuungsmasse.

4. Die Normalverteilung

4.1 Herleitung der Normalverteilung



Im 2. Abschnitt wurden die Möglichkeiten aufgezeigt, Verteilungen darzustellen. Für nicht gleichverteilte Zufallsvariablen hatte sich oft der Effekt eingestellt, daß relativ viele Werte in der Nähe des arithmetischen Mittels lagen, relativ wenige hingegen in den äußeren Bereichen der Verteilung („glockenförmige Verteilung“).

Nehmen wir z.B. die – in diesem Fall symmetrische - Verteilung

x_i	5-7	8-10	11-13	14-16	17-19
f_i	2	3	10	3	2

so ist leicht zu erkennen, daß sich die meisten Meßwerte in unmittelbarer Nähe zum arithmetischen Mittel ($\bar{x} = 12$) befinden. Würde man immer mehr entsprechend verteilte Objekte hinzunehmen und die Intervalle immer weiter verkleinern, bis schließlich eine unendlich große Grundgesamtheit vorliegt, so erhielte man die oben dargestellte Normalverteilungskurve (auch: Gaußsche Kurve). Diese Verteilung zeichnet sich durch einige Merkmale aus:

- Die Kurve ist symmetrisch und eingipflig
- Arithmetisches Mittel, Modus und Median fallen zusammen
- Die Kurvenenden nähern sich asymptotisch der x-Achse (sie werden immer kleiner, erreichen aber niemals den Wert Null)
- Die beiden Punkte mit der größten Steigung liegen bei dem Mittelwert ± einer Standardabweichung
- Es läßt sich genau bestimmen, wie viele Fälle sich in bestimmten Bereichen der Kurve befinden, z.B.

$\bar{x} \pm 1s$: 68,3 % aller Fälle
$\bar{x} \pm 2s$: 95,5 % aller Fälle
$\bar{x} \pm 3s$: 99,7 % aller Fälle

Die Normalverteilung ist in ihrer konkreten Form durch die Größe des arithmetischen Mittels und der Standardabweichung gekennzeichnet.

4.2 Die Standardnormalverteilung

Da in der Praxis unendlich viele unterschiedliche Mittelwerte von Verteilungen sowie unendlich viele Standardabweichungen vorkommen können, können die unterschiedlichsten Ergebnisse für die gleiche Fragestellung vorkommen. Gegeben sind z.B. folgende Verteilungen normalverteilter Werte:

Verteilung 1: Durchschnittliches Nettoeinkommen von deutschen Arbeitnehmern im Jahr 2000; $\bar{x}_1 = 4.000$ EUR, $s_1 = 800$ EUR.

Verteilung 2: Durchschnittlicher Verbrauch von im Jahr 1999 zugelassenen Neuwagen in l/100 km; $\bar{x}_2 = 9$, $s_2 = 0{,}7$.

In beiden Fällen lautet die Frage, in welchem Bereich sich die mittleren 68,3% aller Fälle befinden. Dies läßt sich leicht errechnen:

Verteilung 1: $\bar{x}_1 - 1s = 3.200$ EUR, $\bar{x}_1 + 1s = 4.800$ EUR; 68,3% der Verteilung haben also ein durchschnittliches Nettoeinkommen zwischen EUR 3.200,- und 4.800,-.

Verteilung 2: $\bar{x}_2$ - 1s = 8,3 l/100 km, $\bar{x}_2$ + 1s = 9,7 l/100km; 68,3% der Verteilung haben einen durchschnittlichen Verbrauch zwischen 8,3 und 9,7 l/100 km.

Geht es um eine, zwei oder drei Standardabweichungen, so lassen sich oben genannte Fragestellungen einfach beantworten. Werden Zwischenschritte von Standardabweichungen benötigt, oder die Fragestellung umgekehrt („Person X ein Einkommen von EUR 5.000,-. Welcher Anteil der Bevölkerung hat ein größeres Einkommen?"), so werden weitere Informationen benötigt. Leider ist es unmöglich, für unendlich viele mögliche Verteilungen die erforderlichen Werte in entsprechenden Tabellen anzugeben. Es können jedoch Werte für eine Verteilung mit dem arithmetischen Mittel Null und der Standardabweichung Eins angegeben werden. Die Normalverteilung mit diesen Werten („Parametern") heißt **Standardnormalverteilung.**

Für normalverteilte Variablen wird symbolisch geschrieben $X \sim N(\bar{x};s^2)$, wobei „~" bedeutet, die Variable ist normalverteilt, mit dem Mittelwert $\bar{x}$ und der Varianz s^2. Werte der oben als „Verteilung 1" bezeichneten Verteilung lassen sich also schreiben als $X_1 \sim N(4.000;640.000)$, standardnormalverteilte Werte als $Z \sim N(0;1)$.

Jeder normalverteilte Wert läßt sich in einen standardnormalverteilten Wert z transformieren. Eine solche Transformation heißt **Standardisierung** und wird erreicht durch $z = \frac{x - \bar{x}}{s}$; im Falle der Verteilung 1 würde also der Wert x = 3200 durch den z-Wert $z = \frac{3200 - 4000}{800} = -1$ repräsentiert; wie oben gesehen, ist 3.200 genau –1 Standardabweichung vom arithmetischen Mittel entfernt.

Die gesamte Fläche unter der Standardnormalverteilungskurve hat den Wert Eins. Die Tabelle im **Anhang A)** gibt den Anteil der Fläche unter der Standardnormalverteilungskurve an, der **links** vom entsprechenden z-Wert steht. Unter Verwendung dieser Tabelle lassen sich viele praktische Fragestellungen beantworten.

Es läßt sich sofort erkennen, daß sich links vom arithmetischen Mittel (z = 0) die Fläche 0,5, also 50% aller Fälle befinden. Die angegebene Fläche ist in der Praxis mit der entsprechenden Wahrscheinlichkeit gleichzusetzen.

4.3 Praktische Berechnungen mit der z-Wert-Tabelle

Der Gebrauch der z-Wert-Tabelle soll anhand eines fiktiven Beispiels veranschaulicht werden:

Das Sterbealter von Männern liege in der Bundesrepublik Deutschland bei 73 Jahren, die Standardabweichung bei 6 Jahren, $X \sim N(73;36)$.

a) Wieviele Männer werden älter als 80?

Gesucht ist der Anteil der Männer, die älter werden als 80. Diesem Anteil entspricht die Fläche unter der Standardnormalverteilungskurve, die sich **rechts** vom entsprechenden Standardwert befindet.

Der dem Wert x = 80 entsprechende Standardwert ist $z = \frac{80-73}{6} = 7/6 = 1{,}166666$.

Der z-Wert ist – auf zwei Nachkommastellen gerundet – z = 1,17. Diesem z-Wert entspricht laut Tabelle

z	***0,00***	***0,01***	***0,02***	***0,03***	***0,04***	***0,05***	***0,06***	***0,07***	***0,08***	***0,09***
…										
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	***0,8790***	0,8810	0,8830

die Fläche 0,8790. Diese Fläche befindet sich **links** vom errechneten Wert. Die gesamte Fläche ist Eins, also ist die Fläche **rechts** davon 1 – 0,8790 = 0,1210. Folglich ist der Anteil der Personen, die älter als 80 werden, 0,1210 oder 12,10%.

b) Wieviele Personen sterben im Alter zwischen 60 und 70 Jahren?

Die Standardisierung der Werte $x_1 = 50$ und $x_2 = 60$ liefert $z_1 = \frac{60-73}{6} = -13/6 = -2{,}17$ und $z_2 = \frac{70-73}{6} = -3/6 = 0{,}5$.

Zu berechnen ist die Fläche zwischen z_1 und z_2.

Die **links** von z_1 liegende Fläche ist laut Tabelle 0,0150, die Fläche **links** von z_2 ist 0,5000. Die gesuchte Fläche ist die Fläche, die von z_1 und z_2 **eingegrenzt** wird, also 0,5 – 0,0150 = 0,4850.

Der Anteil der Personen, die im Alter zwischen 60 und 70 Jahren sterben, ist also 0,485 bzw. 48,5%.

c) Gesucht ist das Alter, ab dem die letzten 5% sterben.

Bei dieser Fragestellung ist die Fläche bereits bekannt. Die Fläche, die die ältesten 5% repräsentiert, liegt rechts vom zu bestimmenden z-Wert, also ist die angegebene Fläche **links** davon 0,95. In der Tabelle ist die Fläche 0,9500 nicht angegeben, aber 0,9495. Dieser Fläche entspricht ein z-Wert von 1,64.

Es gilt nun, den z-Wert in einen x-Wert zurückzuverwandeln.

$$z = \frac{x-\bar{x}}{s} \quad \Leftrightarrow \quad zs = x - \bar{x} \quad \Leftrightarrow \quad x = zs + \bar{x}.$$

Der gesuchte Wert ist also $x = zs + \bar{x} = 1{,}64 * 6 + 73 = 82{,}84$.

Die letzten 5% sterben also in einem Alter von 82,84 und darüber.

Aufgaben:

Die durchschnittliche Umsatz X aller Betriebe einer bestimmten Branche in Mio. EUR sei normalverteilt mit $X \sim N(20;12{,}25)$.

a) Geben Sie den Anteil der Unternehmen an, die einen Umsatz zwischen 10 und 30 Mio. EUR haben.
b) Wieviel Prozent der Unternehmen haben einen Umsatz über 25 Mio. EUR?
c) Ein Unternehmen hat einen Umsatz von 22 Mio. EUR. Um wieviel EUR muß sich der Umsatz mindestens vergrößern, damit dieses Unternehmen zu den 10 Prozent mit dem größten Umsatz gehört?

5. Bivariate Verteilungen

In den vorangegangenen Kapiteln wurden ausschließlich Maßzahlen zur Charakterisierung univariater Verteilungen, also Verteilungen einer Variablen, behandelt. In diesem Kapitel werden Verteilungen von zwei Variablen betrachtet.

Bivariaten Häufigkeitsverteilungen liegen Paare von Beobachtungen am gleichen Element einer Stichprobe zugrunde. Liegen vom selben Objekt zwei Beobachtungen vor oder wird an gleichen Objekten mehr als eine Messung vorgenommen, so stellt sich die Frage nach dem Zusammenhang zwischen den Variablen X und Y.

Beispiel:
a) zwei Beobachtungen am selben Objekt: Der Umsatz eines Unternehmens sei die Variable X. Die Variable Y sei die Anzahl der Mitarbeiter. Es soll untersucht werden, ob ein Zusammenhang zwischen Umsatz und Mitarbeiterzahl besteht.

b) Beobachtungen an gleichen Objekten: Der Umsatz von Hamburger Unternehmen sei die Variable X., die von Münchener Unternehmen die Variable Y. Es ist zu untersuchen, welcher Zusammenhang zwischen den beiden Variablen besteht.

Es gibt zwei grundsätzliche Möglichkeiten, den Zusammenhang zwischen zwei Variablen zu ermitteln:
Für lediglich nominalskalierte Variablen kann untersucht werden, wie häufig bestimmte Merkmalskoppelungen vorkommen (z.B.: Tritt hoher Umsatz häufig im Zusammenhang mit bestimmten Regionen auf?). Sind die Variablen ordinalskaliert oder liegt sogar ein metrisches Skalenniveau vor, dann läßt sich untersuchen, ob in beiden Variablen gleichzeitig Veränderungen auftreten (z.B.: Nimmt der Umsatz mit steigender Mitarbeiterzahl zu?).

Das Skalenniveau entscheidet darüber, welche Verfahren zur Ermittlung des Zusammenhanges angewendet werden dürfen:

Bei metrischem Niveau ist die Bestimmung einer **Maßkorrelation** möglich, bei ordinalem Niveau kann die **Rangkorrelation** ermittelt werden. Sind die Variablen lediglich nominalskaliert, so kann man den Zusammenhang durch **Assoziationsmasse** ermitteln.

Für alle Möglichkeiten der Skalierung gibt es grundsätzlich drei Möglichkeiten, welche Zusammenhänge zwischen den Variablen bestehen können:

a) **Übereinstimmung:** Hohen Werten der Variablen X entsprechen auch hohe Werte der Variablen Y; niedrigen Werten der Variablen X entsprechen niedrige Werte der Variablen Y. In diesem Fall liegt eine **positive Korrelation** zwischen den Variablen vor.

b) **Gegensatz**: Hohen Werten der Variablen X entsprechen niedrige Werte der Variablen Y. In diesem Fall liegt ebenfalls ein Zusammenhang zwischen den beiden Variablen vor, allerdings ist dies ein gegensätzlicher Zusammenhang. Man spricht von einer **negativen Korrelation**.

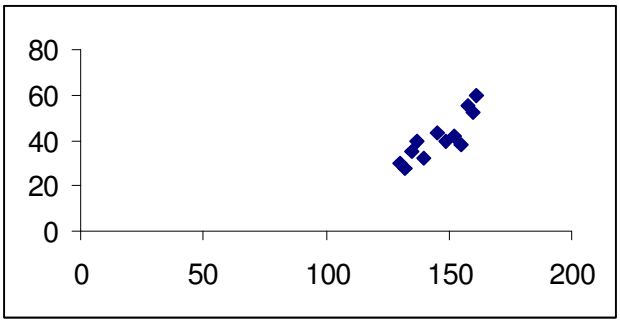
c) **Unabhängigkeit**: Die Werte der Variablen X sind ohne erkennbaren Trend mit Werten der Variablen Y gepaart; hohe Werte treten sowohl mit hohen, mittleren als auch mit niedrigen Werten auf. In diesem Fall besteht kein statistischer Zusammenhang. Die Variablen sind **unabhängig**.

5.1 Darstellung bivariater Verteilungen

Metrische Beobachtungsdaten lassen sich wie auch univariat verteilte Variablen darstellen. Dies soll am Beispiel der Messung von Körperlänge und Gewicht veranschaulicht werden:

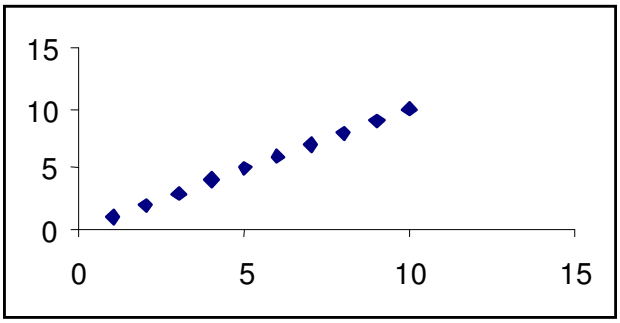
Person	Größe x_i	Gewicht y_i
A	130	30
B	132	28
C	135	35
D	137	40
E	140	32
F	145	43
G	149	40
H	152	42
I	155	38
J	158	55
K	160	52
L	161	60

Zur graphischen Darstellung kann man am besten ein rechtwinkliges Koordinatensystem mit zwei Achsen verwenden, in dem die eine Achse die Werte der Variablen X, die andere die der Variablen Y abbildet. Jedes Wertepaar $x_i y_i$ wird durch einen Punkt gekennzeichnet:

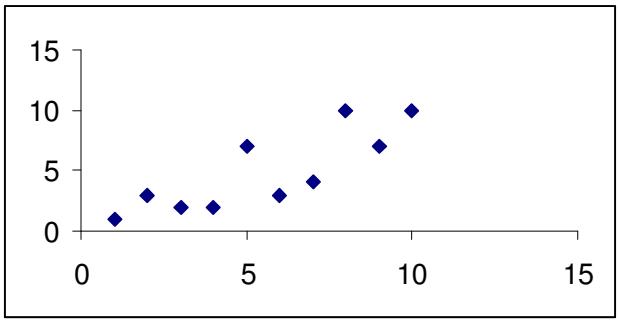


Anhand einer solchen Graphik kann man bereits erkennen, welcher Zusammenhang vorliegt. Im folgenden werden Beispiele für zwei beliebige Variablen aufgezeigt:

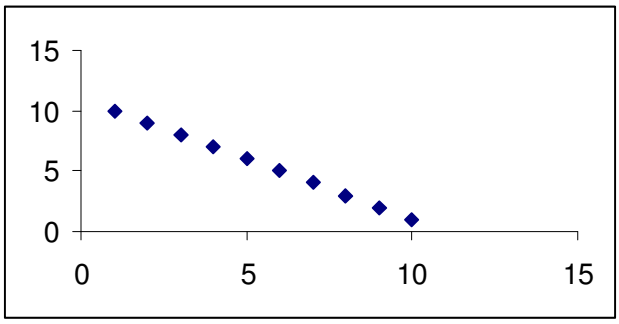
extrem positive Korrelation:



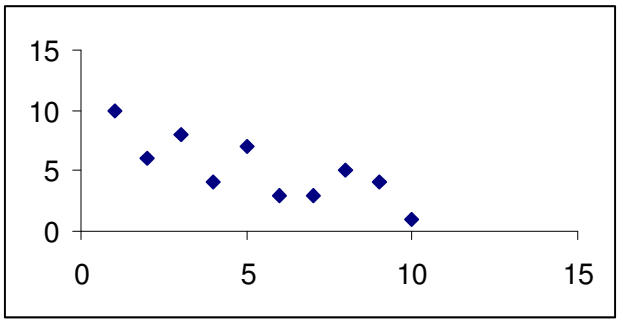
schwach positive Korrelation:



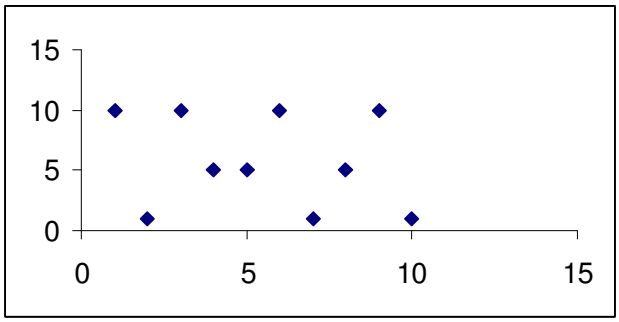
extrem negative Korrelation:



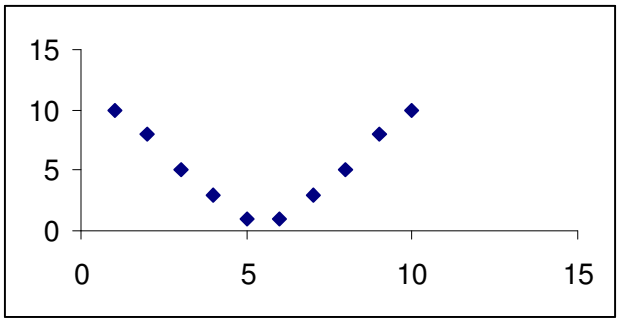
schwach negative Korrelation:



Unabhängigkeit:



nichtlinearer Zusammenhang:



5.2 Überblick über statistische Kennzahlen zur Ermittlung der Korrelation

Wie bereits oben angesprochen hängt die Verwendung statistischer Kennzahlen zur Ermittlung der Korrelation vom vorliegenden Skalenniveau der verwendeten Variablen ab. Die statistischen Kennzahlen, die Aussagen über das Maß des Zusammenhangs treffen, heißen **Koeffizienten**. Man spricht von **Korrelationskoeffizienten,** wenn Variablen mit metrischem Skalenniveau untersucht werden. Liegen Variablen vor, die ordinalskaliert sind, so spricht man von **Rangkorrelationskoeffizienten,** liegt lediglich Nominalskalenniveau vor, so spricht man von **Kontingenzkoeffizienten** oder **Assiziationskoeffizienten.**

Werden zwei Variablen auf unterschiedlichem Skalenniveau untersucht (z.B. Schulnoten und Taschengeld), so muß immer das **niedrigere** Skalenniveau zugrunde gelegt werden, um zu entscheiden, welches Zusammenhangmaß berechnet werden kann.

5.3 Die Maßkorrelation

Sind beide Variablen metrisch skaliert, so läßt sich ihre Korrelation mit dem sog. „Produkt-Moment-Korrelationskoeffizienten“ r beschreiben. Dieser Korrelationskoeffizient wird häufig auch als Pearsonscher Maßkorrelationskoeffizient bezeichnet.

Damit dieser Koeffizient sowohl für große Variablenwerte (z.B. Entfernung zwischen den Planeten) als auch für kleine Variablenwerte (z.B. Entfernung von Atomteilchen) Werte liefert, die einfach zu interpretieren sind, ist sein Wertebereich auf $-1 \leq r \leq 1$ festgelegt.

Die Werte von r lassen sich wie folgt interpretieren:

1: **Strenger positiver Zusammenhang**; hohe Werte der Variablen X fallen mit hohen Werten der Variablen Y zusammen.

0: **Unabhängigkeit**. Es besteht kein Zusammenhang zwischen Werten der Variablen X und Werten der Variablen Y.

-1: **Strenger negativer Zusammenhang**. Hohe Werte der Variablen X fallen mit niedrigen Werten der Variablen Y zusammen (und umgekehrt).

In der Praxis kommen natürlich selten diese extremen Werte vor. Man kann daher die Werte von r wie folgt interpretieren:

$0 < |r| \leq 0{,}3$: schwacher positiver bzw. negativer Zusammenhang

$0{,}3 \leq |r| \leq 0{,}7$: mittlerer Zusammenhang

$0{,}7 \leq |r| < 1$: starker Zusammenhang

Praktisch läßt sich der Produkt-Moment-Korrelationskoeffizient mit folgender Formel berechnen:

$$r = \frac{1}{(n-1)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{s_{xy}}{s_x s_y},$$ [4]

wobei s_x die Standardabweichung der X-Werte ist, s_y die der Y-Werte.

Für das o.g. Beispiel (Gewicht und Körperlänge von 12 Schülern) ergeben sich folgende Werte:

n	= 12
x_i	= 146,17
y_i	= 41,25
s_x	= 11,21
s_y	= 10,00
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	= 1.067,50

Folglich ist

$$r = \frac{1}{(n-1)} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{1}{11} \cdot \frac{1.067{,}50}{11{,}21 \cdot 10{,}00} = 0{,}87.$$

Für das Beispiel liegt also eine starker positiver Zusammenhang vor.

In der Praxis ist es relativ umständlich, erst die Standardabweichungen berechnen zu müssen. Für kleinere Stichprobengrößen kann die Berechnung des Korrelationskoeffizienten r nach folgender Formel erfolgen:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}}$$

Der Korrelationskoeffizient r ist **invariant** gegenüber linearer Transformation der Variablen. Dies bedeutet, daß sich r nicht verändert, wenn beispielsweise Temperaturen in °C in °F transformiert werden. Für die praktische Berechnung hat diese Eigenschaft den Vorteil, daß Sie die Ausgangsdaten nach erlaubten Regeln entsprechend umformen können, um die erforderlichen Bestandteile der Formel einfacher berechnen zu können.

---

[4] $\frac{1}{(n-1)} \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ ist die Kovarianz der Variablen X und Y und wird mit s_{xy} bezeichnet.

Beispiel: Gegeben sind die Datenreihen

A = 5, 10, 20, 40, 45, 60 und
B = 10, 20, 30, 40, 50, 60.

Sie können beide Datenreihen der linearen Transformation f: $x \rightarrow \frac{x}{5}-1$ unterziehen, ohne daß sich das Ergebnis für r verändert:

A´ = 0, 1, 3, 7, 8, 11
B´ = 1, 3, 5, 7, 9, 11

Aufgaben:

a) Berechnen Sie den Korrelationskoeffizienten r für das o.g. Beispiel (Datenreihen A‘ und B‘) nach der „Praxisformel“. Am einfachsten erstellen Sie sich dazu eine Tabelle, in der Sie nebeneinander die x_i, y_i, x_i^2, y_i^2 sowie die x_iy_i bestimmen. Die benötigten Summen können Sie dann sehr einfach bilden.

b) Zeigen Sie, daß der Korrelationskoeffizient r zwischen den o.g. Variablen A und B sowie zwischen A‘ und B‘ gleich ist.

c) In den Ländern X und Y werden im langjährigen Mittel folgende Sonnenscheindauern (in Minuten) gemessen:

	Jan	**Feb**	**Mar**	**Apr**	**Mai**	**Jun**	**Jul**	**Aug**	**Sep**	**Okt**	**Nov**	**Dez**
A	2000	7000	8500	10000	12000	20000	24000	26000	20000	12000	6000	4000
B	22000	20000	13000	10000	8000	4000	8000	10000	10000	13000	20000	23000

Berechnen Sie den Korrelationskoeffizienten r nach einer Formel Ihrer Wahl. Überlegen Sie sich vorher, wie Sie sich die Berechnung erleichtern können. Interpretieren Sie das Ergebnis.

5.4 Die Rangkorrelation

Liegen für die Variablen anstelle metrischer Meßwerte Rangreihen vor, so ist der Zusammenhang zwischen den Variablen mittels eines Rangkorrelationskoeffizienten zu bestimmen. Wie auch im Falle der Maßkorrelation liefert ein Rangkorrelationskoeffizient einen Wert zwischen minus Eins und Eins, der entsprechend zu interpretieren ist.

Es kommt in der Praxis häufig vor, daß sich Eigenschaften von Objekten nicht exakt messen lassen, jedoch Aussagen über den Rang der Objekte getroffen werden können. Dies ist beispielsweise der Fall, wenn es gilt, Produkte aufgrund des subjektiven Eindrucks in eine Rangfolge zu bringen. Der Rangkorrelationskoeffizient mißt in diesem Fall, wie ähnlich die Einschätzungen zweier Personen sind.

Es gibt verschiedene Rangkorrelationskoeffizienten, auf deren Eigenschaften in den folgenden Abschnitten eingegangen wird. Zunächst soll der **Spearmansche Rangkorrelationskoeffizient r_s** betrachtet werden.

5.4.1 Der Spearmansche Rangkorrelationskoeffizient r_s

Voraussetzung für die Berechnung des Spearmanschen Rangkorrelationskoeffizienten r_s ist die Überführung der Ränge in eine Folge natürlicher Zahlen (1, 2, 3, …). Obwohl die Variablen ordinalskaliert sind, wird unterstellt, daß sich der Grad der Ausprägung des Merkmals von Rangplatz zu Rangplatz um denselben Betrag ändert (man rechnet also praktisch mit metrischem Niveau). Ein Problem, das sich häufig ergibt, liegt darin, daß gleiche Ränge vorkommen können, z.B. 1, 2, 2, 4, … In diesem Fall wird einer der doppelt vorkommenden Ränge rechnerisch auf den nächst niedrigeren Rang gesetzt und dann beide Ränge gemittelt, also 1, 2,5, 2,5, 4, …

Der Spearmansche Rangkorrelationskoeffizient ist $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, wobei d_i die Differenz des Rangplatzpaares $(x_i - y_i)$ bezeichnet und n die Anzahl der Rangplätze.

Beispiel: Es soll beurteilt werden, ob bei zehn Unternehmen aus dem Bereich Kosmetik ein Zusammenhang zwischen der Beurteilung des Produkts A (z.B. Deodorant) und B (z.B. Bademittel) besteht. Hierzu werden die Ränge für die beiden Produkte der Unternehmen herangezogen:

Unternehmen	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**Summe**
Produkt A	1	2	3	4	5	6	7	8	9	10	**55**
Produkt B	3	5	2	1	4	8	9	6	7	10	**55**
$\mathbf{d_i}$	-2	-3	1	3	1	-2	-2	2	2	0	**0**
$\mathbf{d_i^2}$	4	9	1	9	1	4	4	4	4	0	**40**

$$r_s = 1 - \frac{6 \cdot 40}{10(100-1)} = 1 - \frac{240}{990} = 0{,}7576.$$

Ob die Rechnung stimmt, kann anhand zweier Eigenschaften kontrolliert werden:

1) Die Summe der Rangplatzdifferenzen ergibt immer Null, also $\sum d_i^2 = 0$

2) Die Summe der Rangzahlen für jede Rangreihe ist $\frac{n(n+1)}{2}$

5.4.2 Kendalls τ_a

Im Gegensatz zu Spearmans Korrelationskoeffizienten basiert Kendalls Koeffizient auf dem Vergleich **konkordanter** und **diskordanter** Paare. Um die Begriffe zu veranschaulichen, kann man von folgendem Beispiel ausgeghen:

Verglichen werden die Ergebnisse zweier Schüler in zwei Tests:

Schüler	**A**	**B**
Deutsch	A	C
Mathematik	B	D

Hierbei bedeutet A ein besseres Ergebnis als B und B ein besseres Ergebnis als C und C ein besseres Ergebnis als D. In diesem Beispiel hat also der Schüler A in beiden Fächern ein besseres Ergebnis erzielt als der Schüler B. Werden die beiden Paare (Ergebnisse in Deutsch/Ergebnisse in Mathematik) miteinander verglichen, so ist die Rangordnung in beiden Fällen gleich (A ist sowohl in Deutsch als auch in Mathematik besser als B). Man spricht in diesem Fall von **konkordanten** Paaren.

In folgendem Beispiel liegen hingegen **diskordante** Paare vor:

Schüler	**A**	**B**
Deutsch	A	B
Mathematik	B	A

Im Fach Deutsch ist A besser, im Fach Mathematik B; die Leistungen in den beiden Fächern sind also in diesem Beispiel gegensinnig.

Der Rangkorrelationskoeffizient τ_a basiert nun auf einer solchen Betrachtung konkordanter und diskordanter Paare; die Differenz der konkordanten und diskordanten Paare wird ins Verhältnis zur Anzahl aller Paare gesetzt. Gibt es mehr konkordante als diskordante Paare, so ist die Differenz $n_c - n_d$ positiv, im umgekehrten Fall negativ und bei gleicher Anzahl Null. Der Rangkorrelationskoeffizient ist

$$\tau_a = \frac{n_c - n_d}{n\frac{n-1}{2}}.$$

Beispiel: Gegeben sind die Ergebnisse von fünf Schülern in den Fächern Deutsch und Mathematik:

Schüler	**A**	**B**	**C**	**D**	**E**
Deutsch	A	C	B	D	E
Mathematik	B	C	D	A	E

Um alle möglichen Paare zu bilden, muß der erste Schüler mit den anderen vier, der zweite mit den anderen drei, der dritte mit den übrigen zwei und schließlich der vierte mit dem fünften verglichen werden. Die Zahl aller möglichen Paare ist also $n(n-1)/2$.

Durch einen Vergleich aller Paare miteinander ergibt sich die Anzahl konkordanter bzw. diskordanter Paare:

Paar	Mathematik	Deutsch	Paartyp
1,2	A – C	B – D	konkordant
1,3	A – B	B – D	konkordant
1,4	A – D	B – A	diskordant
1,5	A – E	B – E	konkordant
2,3	C – B	C –D	diskordant
2,4	C – D	C – A	diskordant
2,5	C – E	C – E	konkordant
3,4	B – D	D – A	diskordant
3,5	B – E	D – E	konkordant
4,5	D – E	A – E	konkordant

Folglich gibt es sechs konkordante und vier diskordante Paare. Kendals τ_a ist also

$$\tau_a = \frac{6-4}{5 \cdot \frac{4}{2}} = 2/10 = 0{,}2.$$

5.4.3 Goodman und Kruskals γ

Der oben vorgestellte Koeffizient τ_a setzt voraus, daß „echte“ Rangreihen vorliegen, d.h., daß kein Rangplatz mehrfach vergeben wird. Ein Koeffizient, der auch solche Rangreihen zuläßt, ist Goodman und Kruskals γ. γ setzt die Differenz konkordanter und diskordanter Paare mit der Summe konkordanter und diskordanter Paare in Beziehung. Die Ermittlung konkordanter und diskordanter Paare erfolgt wie oben beschrieben. Die Berechnung von γ erfolgt durch die Formel

$$\gamma = \frac{n_c - n_d}{n_c + n_d}.$$

Eine spezielle Problematik besteht in der sog. „Bindung“ von Paaren. Bindungen kommen vor, wenn in einer Variablen gleiche Rangplätze vergeben werden, z.B.

Schüler	**A**	**B**
Deutsch	A	C
Mathematik	A	B

Hier tritt der Rangplatz „A“ für den Schüler A zweimal auf. Die Behandlung von Bindungen erfordert spezielle Verfahren, auf denen basierend wiederum weitere Rangkorrelationskoeffizienten aufgebaut sind. Auf die Vorstellung dieser Korrelationskoeffizienten soll jedoch nicht weiter eingegangen werden.

5.4.4 Die Behandlung von gruppierten Daten

Werden nicht einzelne Rangfolgen wie in den oben genannten Beispielen, sondern die Anzahl vorkommender Rangplätze für beide Variablen betrachtet, so kann die Anzahl konkordanter und diskordanter Paare einfach ermittelt werden.

Beispiel: Betrachtet werden die Noten von 30 Schülern in den Fächern Deutsch und Mathematik (die Buchstaben in den Zellen dienen der Bezeichnung der Zellen):

Deutsch/Mathematik	**A**	**B**	**C**
A	5(a)	4(b)	2(c)
B	3(d)	4(e)	2(f)
C	3(g)	3(h)	4(i)

Konkordante Paare: Wegen der Ordnung der Zellen (Zunahme des Rangplatzes nach rechts bzw. unten) sind die konkordanten Paare alle Paare, die **rechts und unterhalb** einer Zelle stehen, also

$$a * (e + f + h + i) + b * (f + i) + d * (h + i) + e * i.$$

Diskordante Paare: Wegen der Ordnung der Zellen sind die diskordanten Paare alle Paare, die **links und unterhalb** einer Zelle stehen, also

$$c * (d + e + g + h) + b * (d + g) + f * (g + h) + e * g.$$

Die Zahl der möglichen Paare ist – wie bisher – $n(n – 1)/2$.

In diesem Beispiel ist also

$$\begin{aligned} n_c &= a*(e + f + h + i) + b*(f + i) + d*(h + i) + e*i = 5*(4 + 2 + 3 + 4) + 4*(2 + 4) \\ &\quad + 3*(3 + 4) + 4*4 \\ &= 65 + 24 + 21 + 16 = 126 \\ n_d &= c * (d + e + g + h) + b * (d + g) + f * (g + h) + e * g = 2 * (3 + 4 + 3 + 3) \\ &\quad + 4 * (3 + 3) + 2 * (3 + 3) + 4 * 3 \\ &= 26 + 24 + 12 + 12 = 74 \\ n &= 30 \end{aligned}$$

Man erhält

$$\tau_a = \frac{126 - 74}{30 \cdot \frac{29}{2}} = 52/435 = 0{,}1195 \text{ bzw.}$$

$$\gamma = \frac{126 - 74}{126 + 74} = 52/200 = 0{,}2600.$$

Die Berechnung der konkordanten und diskordanten Paare für größere Tabellen erfolgt entsprechend.

5.5 Assoziationsmaße für nominalskalierte Daten

Auch im Fall lediglich nominalskalierter Daten ist es möglich, Aussagen über den Zusammenhang zu treffen. Wie in der Einleitung zu Abschnitt fünf dargestellt, handelt es sich allerdings hierbei um die bloße Feststellung, ob es einen Zusammenhang im Sinne von Häufungen gibt. Wegen der simplen Voraussetzung an nominalskalierte Daten kann in diesem Fall ein „Trend" nicht mehr vorhergesagt werden. Die Zusammenhänge von nominalskalierten Daten heißen **Assoziationen**.

Maßzahlen, die Aussagen über die Zusammenhänge mehrfach gestufter nominalskalierter Daten (Häufigkeitstabellen) treffen, heißen **Kontingenzkoeffizienten**. Kontingenzkoeffizienten basieren auf der Berechnung der Größe χ^2 (sprich: „Chi-Quadrat").

5.5.1 Die Berechnung von χ^2

Werden mehrfach gestufte nominalskalierte Daten dargestellt, so hat man eine Variable für die Zeile (Zeilenvariable) und eine Variable für die Spalte (Spaltenvariable). In den Zellen stehen die Besetzungszahlen.

Beispiel: Betriebe nach Bundesland (Zeilenvariable oder Zeilenobjekte) und Umsatz (Spaltenvariable oder Spaltenobjekte):

Betriebe	**< 1 Mio.**	**< 10 Mio.**	**> 10 Mio.**	**Spaltensumme**
Hamburg	10	30	25	65
Hessen	20	20	10	50
Bayern	70	100	50	220
Zeilensumme	100	150	85	335

Man bezeichnet die Zeilen mit der Variablen I = 1, …, i und die Spalten mit der Variablen J = 1, …, J. Das Element in der i-ten Zeile und j-ten Spalte ist n_{ij}. Die Zeilensummen werden mit $n_{i.}$, die Spaltensummen mit $n_{.j}$ bezeichnet. Sie Summe aller Elemente (Summe aller Objekte) ist n.

Besteht kein Zusammenhang zwischen den beiden Variablen, dann verteilen sich die Objekte zufällig auf die Zeilen- und Spaltenvariable, wobei natürlich die unterschiedlichen Größen der Zeilen- bzw. Spaltenobjekte die Verteilung beeinflussen, so daß die Zeilen bzw. Spalten – von zufälligen Abweichungen abgesehen – Vielfache voneinander sind. In diesem Fall sind die Variablen voneinander unabhängig.

Die Zahl χ^2 vergleicht nun die erwartete Häufigkeit im Falle der Unabhängigkeit mit der beobachteten Häufigkeit. Sind die Variablen tatsächlich unabhängig, dann gibt es – bis auf zufällige Differenzen - keine Abweichungen und χ^2 ist nahe Null. Liegt eine maximale Abhängigkeit vor, dann wird $\chi^2 = n\,(q - 1)$, wobei q das Minimum von Zeilen- und Spaltenzahl ist.

Die erwartete Häufigkeit im Fall der Unabhängigkeit in der i-ten Zeile und j-ten Spalte ist

$e_{ij} = n_{i.}n_{.j}/n.$

Die Zahl χ^2 ergibt sich, indem man alle Differenzen zwischen erwarteten und beobachteten Häufigkeiten summiert und quadriert und durch die erwartete Häufigkeit teilt.

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Für die praktische Berechnung empfiehlt sich zunächst die Berechnung der erwarteten Häufigkeiten:

Betriebe	**< 1 Mio.**	**< 10 Mio.**	**> 10 Mio.**	**Spaltensumme**
Hamburg	=65*100/335 = 19,40	29,10	16,49	**65**
Hessen	14,93	22,39	8,33	**50**
Bayern	65,67	98,51	55,82	**220**
Zeilensumme	**100**	**150**	**85**	**335**

Also ist χ^2 = $(10 - 19{,}40)^2/19{,}40 = 4{,}56$
+ $(30 - 29{,}10)^2/29{,}10 = 0{,}03$
+ $(25 - 16{,}49)^2/16{,}49 = 4{,}39$
+ $(20 - 14{,}93)^2/14{,}93 = 1{,}73$
+ $(20 - 22{,}39)^2/22{,}39 = 0{,}25$
+ $(10 - 8{,}33)^2/8{,}33 = 0{,}33$
+ $(70 - 65{,}67)^2/65{,}67 = 0{,}29$
+ $(100 - 98{,}51)^2/98{,}51 = 0{,}02$
+ $(50 - 55{,}82)^2/55{,}82 = 0{,}61$

= 12,20

5.5.2 Der Kontingenzkoeffizient C

Basierend auf der berechneten Zahl χ^2 läßt sich nun der Kontingenzkoeffizient C berechnen:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Für das obige Beispiel ergibt sich $C = \sqrt{\frac{12{,}20}{12{,}20 + 335}} = 0{,}1875$.

Der Wertebereich von C ist $0 \leq C < 1$. Die Ergebnisse sind nach den in den vorangegangenen Abschnitten erläuterten Regeln für Korrelationskoeffizienten zu beurteilen; im Fall des Beispiels liegt also ein schwacher Zusammenhang vor.

5.5.3 Cramers V

Der Kontingenzkoeffizent C kann auch bei vollständigem Zusammenhang der beiden Variablen niemals den Wert Eins erreichen; aus diesem Grunde kann alternativ Cramers V berechnet werden. Diese Größe ist

$V = \sqrt{\frac{\chi^2}{n(q-1)}}$, wobei q das Minimum von Zeilen- und Spaltenzahl ist,

also die Quadratwurzel aus dem errechneten χ^2 geteilt durch das maximal mögliche χ^2.

Für o.g. Beispiel ist $V = \sqrt{\frac{12{,}20}{335 \cdot 2}} = 0{,}1350$.

Es ist zu erkennen, daß die beiden Koeffizienten voneinander Abweichen. Die Stärke der Abweichung hängt von der Besetzungszahl n sowie von der Größe der Zahlentafel (Größe des Minumums von Zeilen- und Spaltenzahl) ab.

5.5.4 Sonderfall Vierfeldertafel

Es kommt häufig vor, daß zwei Variablen mit nur je zwei Ausprägungen betrachtet werden.

Beispiel: Betriebe werden untersucht, ob sie sich in der Stadt oder auf dem Land befinden (Zeilenvariable) und ob sich in den Betrieben mehr oder weniger als zehn Mitarbeiter befinden (Spaltenvariable). Die Variable in den Klammern gibt die Bezeichnung der Felder an.

Betriebe	**< 10 Mitarbeiter**	**10 Mitarbeiter und mehr**
Stadt	100 (a)	500 (b)
Land	400 (c)	300 (d)

Anstelle von χ^2 wird im Fall der Vierfeldertafel der Koeffizient ϕ (sprich: „Phi“) verwendet, wobei

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Für das o.g. Beispiel ist

$$\phi = \frac{100 \cdot 300 - 400 \cdot 500}{\sqrt{(100+500)(400+300)(100+400)(500+300)}} = -0{,}425.$$

Der Koeffizient ϕ hat einen Wertebereich zwischen minus Eins und Eins. Durch Vertauschen der Zeilen bzw. Spalten ändert sich das Vorzeichen von ϕ. Sofern die Zeilen- und Spaltenvariablen ordinales Nieveau haben, gibt ϕ die Richtung des Zusammenhangs an. In diesem Fall dürfen die Zeilen bzw. Spalten natürlich nicht vertauscht werden.

Zwischen χ^2 und ϕ besteht folgender Zusammenhang:

$\phi^2 = \chi^2/n$.

Aufgaben:

a) Berechnen Sie die Rangkorrelationskoeffizienten τ_a und γ für folgende Rangreihen:

X: A B C D E F
Y: B A F E D C

b) Berechnen Sie die Rangkorrelationskoeffizienten τ_a und γ für folgende ordinalskalierte Variablen mit gruppierten Daten:

Betriebe	**< 1 Mio. Umsatz**	**< 10 Mio. Umsatz**	**> 10 Mio. Umsatz**
< 10 Mitarbeiter	10	20	30
< 100 Mitarbeiter	20	40	50
> 100 Mitarbeiter	30	50	60

c) Berechnen Sie den Kontingenzkoeffizienten C und Cramers V für folgende Zahlentafel:

	Werbekosten < 1 Mio. p.a.	Werbekosten < 5 Mio. p.a.	Werbekosten > 5 Mio. p.a.
Hamburg	22	33	44
Berlin	66	55	44
Köln	33	22	11
München	44	88	66

d) Gegeben ist folgende Vierfeldertafel:

Stimmen zur Steuerreform	ja	nein
Angestellte	80	120
Selbständige	70	30

Berechnen Sie ϕ und Cramers V. Gibt es eine Möglichkeit, Cramers V aus ϕ zu berechnen?

6. Lineare Regression

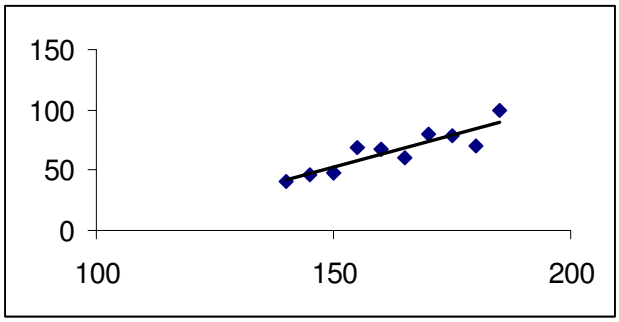
Im vorigen Abschnitt wurde der Zusammenhang zwischen zwei Variablen untersucht. Eng mit dieser Thematik verbunden ist die Frage, wie sich der Wert einer Zufallsvariablen schätzen läßt, wenn der Wert einer anderen Variablen desselben Elements bekannt ist. Eine solche Schätzung läßt sich für metrisch skalierte Daten vornehmen.

Um eine Schätzung vornehmen zu können, müssen die Variablen korreliert sein. Je stärker die Korrelation ist, desto zuverlässiger wird die Vorhersage.

Gegeben sind die Werte zweier Variablen, beispielsweise monatlicher Umsatz (in TEUR) und eingegangene Aufträge eines Unternehmens (es wird davon ausgegangen, daß die einzelnen Aufträge so unterschiedlich sind, daß sich der Umsatz nicht durch die Zahl der Aufträge allein bestimmen läßt):

Monat	1	2	3	4	5	6	7	8	9	10
Umsatz (X) in TEUR	140	145	150	155	160	165	170	175	180	185
Eingegangene Aufträge (Y)	40	46	48	68	67	60	80	78	70	99

Die Aufgabe der linearen Regression besteht darin, eine Gerade zu bestimmen, die sich den gegebenen Punkten optimal annähert:



Die beste Näherung ist gegeben, wenn die Summe der Abweichungen jedes Punkts von der Geraden minimal wird. Mit Hilfe der Regressionsgeraden soll es möglich sein, die eine Variable durch die andere vorherzusagen.

Die allgemeine Funktionsgleichung einer Geraden lautet $y = a + bx$. Im Fall der Regression muß es zwei Geraden geben, nämlich eine, mit deren Hilfe die X-Werte vorausgesagt werden können und eine, mit deren Hilfe die Y-Werte vorausgesagt werden können. Die Gerade zur Vorhersage der Y-Werte auf Basis der X-Werte lautet

$y = a_1 + b_1x$ ("Regression von Y auf X"), die Gerade zur Vorhersage der X-Werte auf Basis der Y-Werte lautet

$x = a_2 + b_2y$ ("Regression von X auf Y").

Es gilt, die Parameter a und b zu bestimmen.

Die sog. Regressionskeoffizienten sind

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad \text{und} \qquad b_2 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (y_i - \bar{y})^2}.$$

Im Zähler dieser beiden Koeffizienten steht die Summe der Produkte der Abstände aller Meßwerte vom Mittelwert, im Zähler steht die Summe der Abweichungsquadrate für $\bar{x}$ bzw. $\bar{y}$. Sind die Regressionskoeffizienten b_1 bzw. b_2 bestimmt, dann können

$$a_1 = \bar{y} - b_1\bar{x} \qquad \text{bzw.} \quad a_2 = \bar{x} - b_2\bar{y}$$

ermittelt werden.

Liegt eine perfekte Korrelation zwischen den Variablen X und Y vor ($r = 1$ bzw. $r = -1$), dann fallen beide Regressionsgeraden zusammen. Je größer die Korrelation ist, desto kleiner wird der Winkel zwischen beiden Regressionsgeraden. Ferner gilt: $r = \sqrt{b_1 b_2}$.

Für o.g. Beispiel lassen sich die Regressionsgeraden wie folgt berechnen:

Monat	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mittelwerte**	**Summen**
Umsatz (X)	140	145	150	155	160	165	170	175	180	185	**163**	
Aufträge (Y)	40	46	48	68	67	60	80	78	70	99	**66**	
$x_i - \bar{x}$	-23	-18	-13	-8	-3	3	8	13	18	23		**0**
$y_i - \bar{y}$	-26	-20	-18	2	1	-6	14	12	4	33		**0**
$(x_i - \bar{x})(y_i - \bar{y})$	576	343	220	-18	-4	-14	108	155	77	752		**2195**
$(x_i - \bar{x})^2$	506	306	156	56	6	6	56	156	306	506		**2063**
$(y_i - \bar{y})^2$	655	384	310	6	2	31	207	154	19	1116		**2884**

Man erhält

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{2195}{2063} = 1{,}06 \quad \text{und} \quad b_2 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (y_i - \bar{y})^2} = \frac{2195}{2884} = 0{,}76.$$

Für r ergibt sich

$$r = \sqrt{b_1 b_2} = \sqrt{1{,}06 \cdot 0{,}76} = 0{,}90$$

Ferner ergibt sich

$$a_1 = \bar{y} - b_1 \bar{x} = 66 - 1{,}06 * 163 = -107 \quad \text{und} \quad a_2 = \bar{x} - b_2 \bar{y} = 163 - 0{,}76 * 66 = 113$$

Hieraus ergeben sich die Gleichungen für die Regressionsgeraden

Regression von Y auf X:

$y = -107 + 1{,}06\ x$

und Regression von X auf Y:

$x = 113 + 0{,}76\ y$

Ist nun für unser Beispiel zu bestimmen, wieviel Aufträge bei einem Umsatz von 200 TEUR eingehen würden, so errechnet sich als Zahl der Aufträge

$y = -107 + 1{,}06 * 200 = 106.$

Lautet die Frage, welcher Umsatz bei einem Eingang von 100 Aufträgen zu erwarten ist, so ergibt sich

x = 113 + 0,76 * 100 = 189 TEUR.

Aufgaben:

Ein Betrieb macht folgende Umsätze pro Jahr (die Variablen Jahr und Umsatz haben metrisches Skalenniveau):

Jahr	1992	1993	1994	1995	1996	1997	1998	1999
Umsatz (in Mio. EUR)	10	12	11	17	14	22	20	24

a) Bestimmen Sie die Gleichungen für die beiden Regressionsgearden.
b) Bestimmen Sie den Korrelationskoeffizienten r.
c) Welche Umsätze sind für die Jahre 2000, 2005 und 2010 zu erwarten?
d) Wann wird sich der Umsatz aus dem Jahr 1992 verdreifacht, wann verfünffacht haben?

7. Ausblick: Schließende Statistik

Die schließende Statistik oder auch Inferenzstatistik ist ein eigenständiges Gebiet der Statistik. Die Teilbereiche der schließenden Statistik sind mindestens so umfangreich wie die der deskriptiven Statistik; aus diesem Grunde würde eine ausführliche Darstellung an dieser Stelle den Rahmen sprengen. Dennoch soll in aller Kürze versucht werden, die grundlegende Idee der schließenden Statistik zu vermitteln.

In der schließenden Statistik werden nicht nur Aussagen über die untersuchten Objekte, sondern darüber hinaus Verallgemeinerungen der Aussagen auf die Grundgesamtheit getroffen.

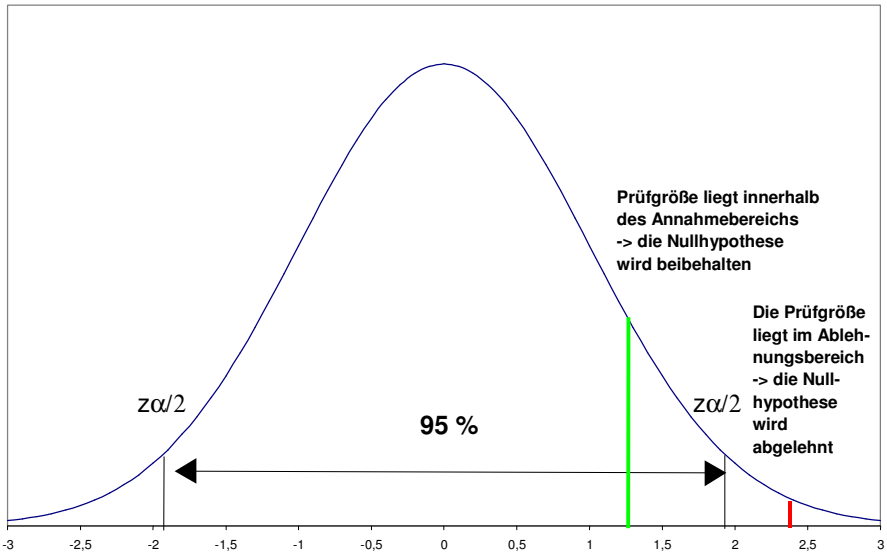
Hierzu läßt sich folgendes Beispiel heranziehen[5]: Zu einem bestimmten Zeitpunkt wurde in einer sehr zeit- und kostenintensiven Erhebung die durchschnittliche Körperlänge von männlichen Schülern der dritten Klasse in einem bestimmten Gebiet ermittelt. 20 Jahre später soll anhand einer Stichprobe überprüft werden, ob sich die Größe verändert hat. Die durchschnittliche Körperlänge betrug in der ersten Untersuchung (n_1 = 2.000 Schüler, Vollerhebung) μ = 142 cm bei einer Standardabweichung von σ = 8 cm, in der 20 Jahre später erhobenen Stichprobe (n_2= 500) $\bar{x}$ = 144 cm bei einer Standardabweichung von s = 8,5 cm. Die schließende Statistik prüft, ob es sich bei der durchschnittlichen Größenzunahme um zwei cm um eine **signifikante** Ver-

[5] Bei diesem Verfahren handelt es sich um den sog. z-Test zum Vergleich eines Stichprobenmittelwerts mit dem Mittelwert einer Grundgesamtheit. Um Parameter der Grundgesamtheit und Statistiken aus den Stichproben voneinander zu unterscheiden, werden für Parameter der Grundgesamtheit griechische Symbole verwendet.

änderung oder nur um eine zufällige Veränderung, z.B. weil die Stichprobe unglücklich zustande kam, handelt.

Die schließende Statistik geht von folgenden Überlegungen aus: Das untersuchte Merkmal, hier die Körperlänge, folgt einer bestimmten **Verteilung**: die Variable Körperlänge ist normalverteilt. Von normalverteilten Werten wissen wir, welcher Anteil sich in bestimmten Bereichen der Verteilung befindet (siehe Abschnitt vier). Kennen wir die Parameter der Grundgesamtheit - hier: durchschnittliche Körperlänge μ = 142 cm und Standardabweichung σ = 8 cm, dann können wir sagen, in welchem Bereich um das arithmetische Mittel sich ein bestimmter Anteil der Objekte befindet. Der Größte Anteil (ca. 95%) befindet sich innerhalb von zwei Standardabweichungen um das arithmetische Mittel.

In der schließenden Statistik werden nun Hypothesen aufgestellt und geprüft. Die erste Hypothese, die sog. „**Nullhypothese**“ besagt, daß kein signifikanter Unterschied zwischen den Mittelwerten vorliegt (H_o: $\mu = \mu_0$); diese Hypothese wird gegen die zweite Hypothese, die sog. „**Alternativhypothese**“ geprüft, die besagt, daß der Unterschied signifikant ist (H_1: $\mu \neq \mu_0$). Für die Prüfung werden rechnerisch beide Mittelwerte standardisiert; der Mittelwert der Grundgesamtheit wird in den Standardwert Null transformiert, der zweite in einen entsprechenden anderen Standardwert. Je nachdem, wie stark sich diese beiden standardisierten Werte voneinander unterscheiden, befindet sich der Mittelwert der Stichprobe innerhalb oder außerhalb des Bereiches von ca. zwei Standardabweichungen (95%) um das arithmetische Mittel der Grundgesamtheit. Wenn sich der Wert **innerhalb** dieser 95% befindet, so wird die Nullhypothese **beibehalten**, befindet er sich **außerhalb**, so wird die Nullhypothese zugunsten der Alternativhypothese **verworfen** und die Unterschiede werden aufgrund ihrer Größe als signifikant eingestuft. Die Nullhypothese wird also abgelehnt, wenn sich die Werte so sehr voneinander unterscheiden, daß der zweite Wert in die äußeren 5% der Verteilung fällt. Man spricht hier von einem **Signifikanzniveau** von α = 5%. Der **kritische Wert,** der über Beibehaltung oder Ablehnung der H_0 entscheidet, ist hier $z_{\alpha/2}$ = +/-1,96.



Im o.g. Beispiel ergibt sich – ohne Herleitung – die Prüfgröße $z = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{144 - 142}{8{,}5}\sqrt{500} = 5{,}26 > 1{,}96$. Damit wird die Nullhypothese zugunsten der Alternativhypothese verworfen, die Größenzunahme ist signifikant bei einer Irrtumswahrscheinlichkeit von 5%.

Es kann jedoch vorkommen, daß die Nullhypothese abgelehnt wird, weil der errechnete Wert im Ablehnungsbereich liegt, obwohl die Nullhypothese tatsächlich zutrifft. Dies kann der Fall sein, wenn die Stichprobe ungünstig gezogen wird, z.B., weil man zufällig ein Gebiet ausgewählt hat, in dem die Unternehmen mehr Mitarbeiter haben als anderswo. Theoretisch kommt dies in 5% aller Fälle vor.

Wird die Nullhypothese in einem solchen Fall verworfen, obwohl sie tatsächlich zutrifft, so spricht man vom sog. **α-Fehler.**

Um den α-Fehler zu minimieren, müßte man den Annahmebereich vergrößern, also das Signifikanzniveau verringern. In diesem Fall würde man die Nullhypothese länger beibehalten.

Andererseits kann es vorkommen, daß die Nullhypothese beibehalten wird, obwohl sie falsch ist. Dies kann der Fall sein, wenn die Stichprobe z.B. in einem Gebiet gezogen wird, in dem die durchschnittliche Körperlänge der Schüler nicht von der in der Vergangenheit ermittelten Anzahl, obwohl sie sich tatsächlich insgesamt stark verändert hat. Wird die Nullhypothese in einem solchen Fall beibehalten, obwohl sie tatsächlich falsch ist, so spricht man vom sog. **β-Fehler**.

		In der Grundgesamtheit gilt	
		H_0	**H_1**
Stichprobe entscheidet:	**H_0**	richtige Entscheidung	β-Fehler
	H_1	α-Fehler	richtige Entscheidung

In der Praxis wird häufig das Signifikanzniveau von $\alpha = 5\%$ verwendet. Dieses Niveau ist ein guter Kompromiß aus nicht zu großer Weite des Annahmebereichs (ca. zwei Standardabweichungen um das arithmetische Mittel) und Sicherheit (95% der Werte fallen liegen im Annahmebereich).

Neben dem Vergleich eines Stichprobenmittelwertes mit dem Mittel der Grundgesamtheit gibt es statistische Testverfahren für die Vergleiche von Mittelwerten aus unterschiedlichen Stichproben. Neben Mittelwerten können aber auch Varianzen oder Korrelationskoeffizienten auf statistische Signifikanz untersucht werden. Diese Größen folgen anderen Verteilungen; auf Einzelheiten soll an dieser Stelle nicht eingegangen werden. Allen statistischen Prüfverfahren ist jedoch - unabhängig von ihrer Verteilung – der hier vorgestellte Ansatz gleich.

8. Analyse von Zeitreihen

In der Praxis kommt es häufig vor, daß Daten über einen bestimmten Zeitraum erhoben werden. Ziel dieser Erhebung ist zum einen, Aufschluß über die bisherige Entwicklung zu erlangen. Zum anderen soll untersucht werden, ob hinter der Entwicklung eine bestimmte Gesetzmäßigkeit steht, mit deren Hilfe Prognosen für die Zukunft erhoben werden können. So kann z.B. die Umsatzentwicklung eines Unternehmens über einen bestimmten Zeitraum betrachtet werden, um die Entwicklung für die nächste vorhersagen zu können.

Im allgemeinen wird davon ausgegangen, daß vier unterschiedliche Faktoren für die Entwicklung ursächlich sind:

1) Die **Grundtendenz (T)** (Beispiel: steigender Bedarf für Multimedia-Produkte)
2) Die **Schwankung der Konjunktur (K)** (Beispiel: im Falle des Aufschwungs wird in der Regel mehr konsumiert als in Zeiten des Abschwungs)
3) Die **Schwankung der Saison (S)** (Beispiel: hoher Absatz von Geschenkartikeln zur Weihnachtszeit, geringer Absatz in den Sommerferien)
4) Die **irregulären Schwankung (I)** (unter diesen Sammelbegriff fallen alle übrigen Schwankungen, die sich nicht durch die drei oben erwähnten Schwankungen erklären lassen, z.B. Lieferengpässe, etc.)

Folglich kann für den Verlauf der Zeitreihen folgende Gleichung aufgestellt werden:
$y = T + K + S + I$.

Die Durchführung einer Analyse von Zeitreihen führt auf das Problem, die Beiträge der einzelnen Komponenten zu ermitteln. Für die Schwankungen läßt sich in der Regel keine lineare Funktion angeben.

Zur Ermittlung des Trends wird in der Praxis häufig die „Methode der gleitenden Durchschnitte“ angewendet. Hierbei wird der bisher betrachtete Zeitraum in kleinere Intervalle zerlegt und für jedes Intervall das arithmetische Mittel der Funktionswerte berechnet. Durch Verbindung der Punkte kann man eine Trendlinie ermitteln, die man in die Zukunft fortschreiben kann. Unterstellt man einen linearen Trend, so empfiehlt sich die Berechnung der Regressionsgeraden wie in Abschnitt sechs dargestellt.

Die Anwendung der linearen Regression stellt in der Regel eine starke Vereinfachung der tatsächlichen Trendverläufe dar. Darüber hinaus kommt es oft vor, daß die tatsächlichen Verläufe der Grundtendenz nicht linear sind.

Häufig vorkommende nicht lineare Verläufe sind Verläufe, die sich durch eine exponentielle Funktion ($y = x^n$) beschreiben lassen sowie Verläufe, deren Funktion sich asymptotisch einer bestimmten Grenze nähert. Ein Beispiel für einen exponentiellen Verlauf ist die Entwicklung der Weltbevölkerung, die immer schneller anwächst. Ein Beispiel für einen asymptotischen Verlauf ist die Entwicklung von Verkäufen langle-

biger Konsumgüter, wie z.B. Videorekorder: Kommen die Güter neu auf den Markt, so ist zunächst ein starker Anwachs zu verzeichnen. Schließlich nähern sich die Verkäufe immer langsamer der sog. Sättigungsgrenze.

Grundsätzlich ist zu bedenken, daß die Erhebung von Prognosen ungenauer wird, je größer der zu betrachtende Zeitraum ist. Die Betrachtung von zukünftigen Zeiträumen wird als Extrapolation bezeichnet, bestimmt man hingegen rückwirkend den Trend für einen Zeitraum, für den keine Daten vorliegen, so spricht man von einer Interpolation.

9. Manipulation von Statistiken

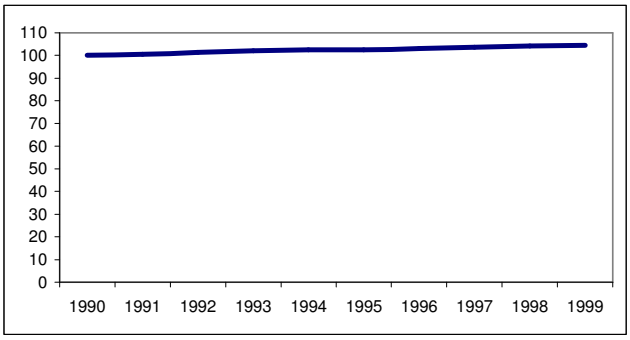
In der Praxis ist es üblich, Statistiken einzusetzen, die den Grundsätzen einer objektiven Darstellung widersprechen. Das Ziel dieser Darstellungen besteht natürlich darin, Werte oder Tendenzen aufzuzeigen, die entweder gar nicht oder zumindest nicht in dieser Form gegeben sind. Hierdurch lassen sich z.B. schwache Umsatzzuwächse ins Dramatische steigern oder die Vorsprünge gegenüber Mitbewerbern deutlich überzeichnen. Da es in der Berufspraxis hilfreich ist, derartige Manipulationen zu erkennen an dieser Stelle kurz die häufigsten Manipulationsmethoden dargestellt werden.

9.1 Beispiel: Umsatzentwicklungen

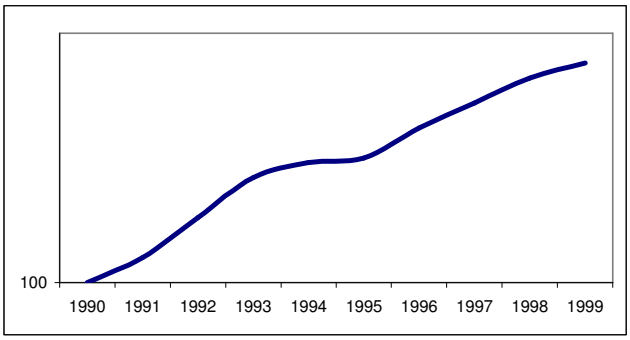
Gegeben sei die Umsatzentwicklung einer Firma für die vergangenen 10 Jahre (Index 1990 = 100):

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
100	100,5	101,3	102,1	102,4	102,5	103,1	103,6	104,1	104,4

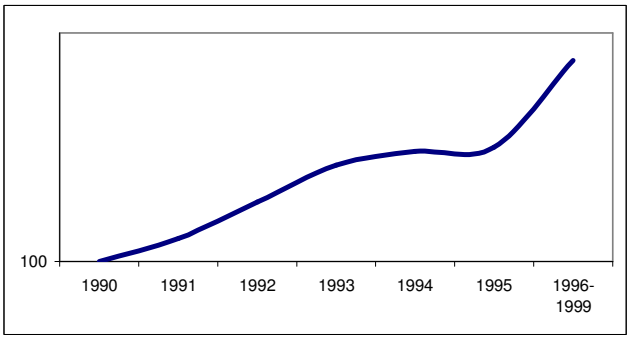
Wie man leicht erkennen kann, ist die Umsatzsteigerung alles andere als besonders groß. Die Darstellung dieser Entwicklung ergibt folgende Graphik, in der kaum eine Veränderung zu erkennen ist:



Wird nun der untere Teil der Graphik abgeschnitten und die y-Achse entsprechend gestreckt, so sieht dies schon ganz anders aus:

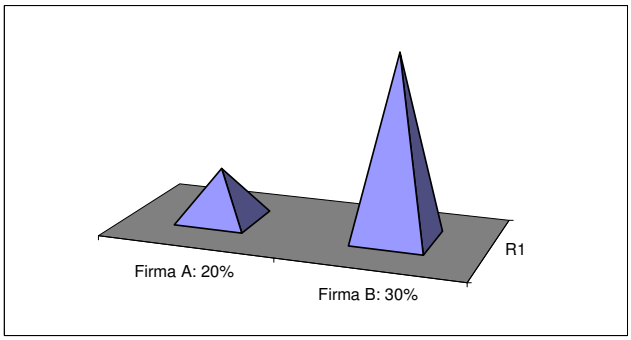


Werden nun noch bestimmte Intervalle zusammengefaßt, kann sich noch eine besondere Steigerung, z.B. in der letzten Zeit, vortäuschen lassen:



9.2 Beispiel: Wettbewerbsvorsprünge

Die Firma A habe einen Marktanteil von 20%, die Firma B einen Marktanteil von 30%. Die Graphik gibt beide Marktanteile im Text richtig wieder, allerdings ist die der Firma B zugeordnete Fläche (bzw. das Volumen in der 3D-Darstellung) vielfach höher. Da meistens der visuelle Eindruck entscheidend ist, dürfte der Vorsprung der Firma B deutlich überbewertet werden:

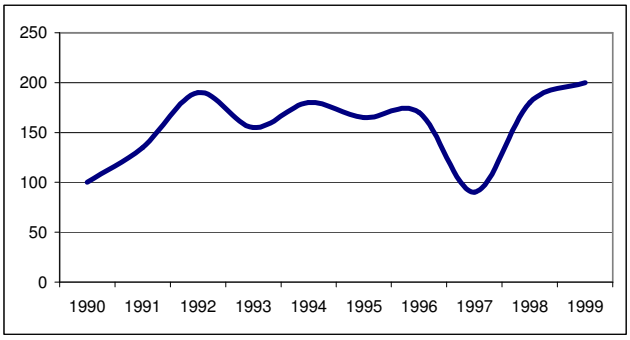


9.3 Die Wahl des richtigen Ausschnitts

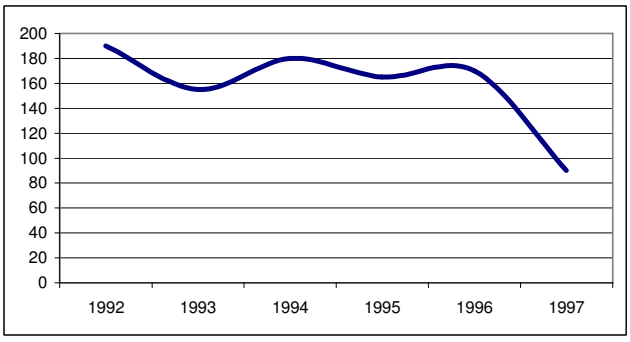
In einem Zeitraum von zehn Jahren haben sich Aktienkurse (oder Umsätze, etc.) wie folgt entwickelt:

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
100	135	190	155	180	165	170	90	180	200

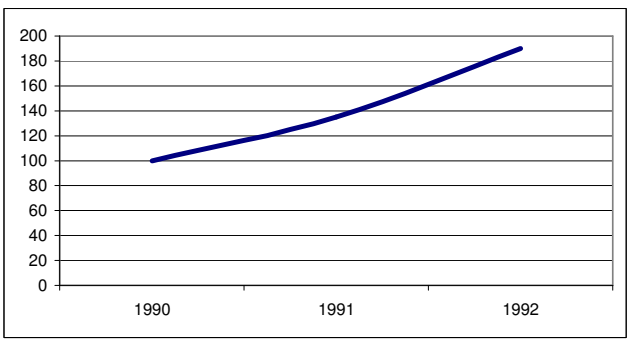
Die entsprechende Graphik sieht wie folgt aus:



Es hängt davon ab, welchen Ausschnitt aus der Graphik man wählt, um entweder einen starken Kursabfall



oder einen starken Kursanstieg aufzuzeigen:

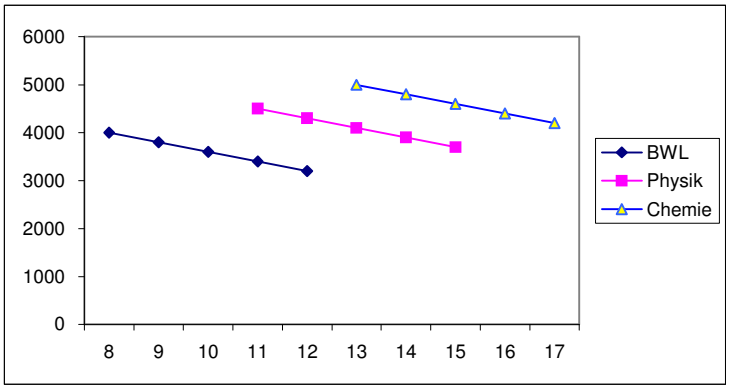


9.4 Scheinkorrelation

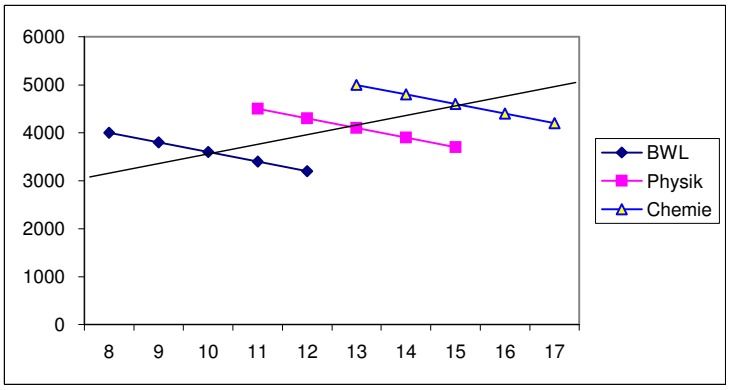
Für die Verwendung von Scheinkorrelationen gibt es unzählige Beispiele. Ein Beispiel sei hier genannt[6]. Betrachtet werden die durchschnittlichen Einstiegsgehälter nach Studiengang und Semesteranzahl:

Fach/Semester	8	9	10	11	12	13	14	15	16	17
BWL	4.000	3.800	3.600	3.400	3.200					
Physik				4.500	4.300	4.100	3.900	3.700		
Chemie						5.000	4.800	4.600	4.400	4.200

Spätestens in der Graphik ist ersichtlich, daß die Einstiegsgehälter für jedes Studienfach mit zunehmender Dauer abnehmen:



Wird hingegen nicht mehr zwischen den einzelnen Fächern unterschieden, so sieht es aus, als ob die Einstiegsgehälter mit steigender Studiendauer anstiegen:



[6] Nach Krämer, S. 165 ff.

Anhang

A) Flächen unter der Normalkurve

z	*0,00*	*0,01*	*0,02*	*0,03*	*0,04*	*0,05*	*0,06*	*0,07*	*0,08*	*0,09*
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

z	*0,00*	*0,01*	*0,02*	*0,03*	*0,04*	*0,05*	*0,06*	*0,07*	*0,08*	*0,09*
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010

B) Lösungen der Übungsaufgaben

Im folgenden sollen die **Ergebnisse** der Übungsaufgaben am Ende der einzelnen Abschnitte zur Kontrolle aufgeführt werden. Sofern nötig, werden zusätzlich Hinweise gegeben. Auf die Darstellung der einzelnen Rechenschritte wird jedoch verzichtet, da sie in den entsprechenden Kapiteln bereits dargestellt wurden.

Abschnitt 1

a) Telefonverzeichnis (nominalskaliert) – Schulnoten (ordinal-) – Stärke von Kopfschmerzen (ordinal-) – Messung von Stromspannung in V (verhältnis-)– Reihenfolge der Lieblingsspeisen (ordinal-) – Aktienkurse (verhältnis-) – Intelligenzquotient (verhältnis-) – Tachometeranzeige (verhältnis-) – Temperatur in Kelvin (verhältnis-)
b) $y = 1{,}8x + 32$ (für die Umwandlung von °C in °F)
c) Falls Ihnen der allgemeine Beweis nicht gelingen sollte: beweisen Sie die Sachverhalte für $i = 1$ bis $n = 3$.

Abschnitt 3.1

a) Der Modalwert ist 5, der Median 7 und das arithmetische Mittel 7,04. Werden alle Werten mit 2 multipliziert und wird 5 addiert, so ist das entsprechende arithmetische Mittel das alte arithmetische Mittel, multipliziert mit 2 plus 5.
b) Der Modalwert ist 1.500, der Median 2.060 und das arithmetische Mittel 2.070. Die Verteilung ist linksschief. Die mittleren 50% liegen im Bereich von 1.703 bis 3.059. Kommt ein weiterer Wert 99.500 hinzu, wird der Median 2.080, das arithmetische Mittel 3.035.

Abschnitt 3.2

Die Variationsbreite v ist 21.
Die durchschnittliche Abweichung e ist 5,0625.
Die Varianz s^2 ist 38,80.
Die Standardabweichung s ist 6,23.

Abschnitt 4

a) Der Anteil beträgt ca. 99,6 %.
b) Der Anteil beträgt ca. 7,7 %.
c) Der Umsatz müßte sich um mindestens 2,52 Mio. EUR vergrößern.

Abschnitt 5.3

a) Sie sollten finden $r = 0{,}9862$.
b) Sie sollten auch für die Korrelation zwischen den Reihen A und B $r = 0{,}9862$ finden.

c) Sie können z.B. alle Werte durch 100 oder durch 60 teilen. Mir r = –0,7987 gibt es einen starken negativen Zusammenhang („je sonniger es im Land A ist, desto weniger Sonne gibt es in Land B“).

Abschnitt 5.5

a) $\tau_a = 1/15$, $\gamma = 1/15$.
b) Sie sollten finden $\tau_a = -0{,}03$ und $\gamma = -0{,}08$.
c) Sie sollten finden $\chi^2 = 20{,}74$. Damit wird C = 0,24 und V = 0,18.
d) $\phi = -0{,}28$. Lesen Sie den Abschnit 5.5.4. Sie sollten erkennen, daß $\phi = \sqrt{V}$.

Abschnitt 6

a) Sie sollten finden: y = -3974,75+2*x und x = 1988,59 + 0,43 * y.
b) Der Korrelationskoeffizient ist r = 0,92.
c) Im Jahr 2000 ist ein Umsatz (in Mio. EUR) von 25,25 zu erwarten, 2005 ein Umsatz von 35,35 und 2010 ein Umsatz von 45,25.
d) Der Umsatz von 1992 wird sich im Laufe des Jahres 2001 verdreifacht, im Laufe des Jahres 2009 verfünffacht haben.

C) Literatur

Zur Erstellung des Skripts wurde folgende Literatur herangezogen:

Benninghaus, H., Deskriptive Statistik. Stuttgart: B. G. Teubner, 1989

Bortz, J., Statistik für Sozialwissenschaftler. Berlin, Heidelberg: Springer-Verlag 1999

Clauß, G./Ebner, H., Statistik für Soziologen, Pädagogen, Psychologen und Mediziner, Band 1. Thun und Frankfurt/M: Verlag Harri Deutsch 1989.

Krämer, Walter, So lügt man mit Statistik, Frankfurt/M: Campus, 1991

Leiner, Bernd, Einführung in die Zeitreihenanalyse, Oldenburg 1991