Mayer Horst (2004): Interview und schriftliche Befragung. Entwicklung, Durchführung und Auswertung, München–Wien; Seite 68–70

Daten und Skalen

Die in einer Stichprobe enthaltenen Untersuchungseinheiten wie Personen, Unternehmen, etc. stellen Merkmalsträger dar, die bestimmte Ausprägungen aufweisen (Geschlecht, Alter, Größe, Einstellungen, Motivation, ...). Messen bedeutet nun, die Merkmalsausprägungen systematisch zu erfassen und ihnen nach ganz bestimmten Regeln numerische Werte zuzuordnen (vgl. dazu z.B. Schnell u.a. 1999, S. 132 ff). Durch diesen Vorgang entstehen Daten, das sind zahlenmäßig erfasste Merkmalsausprägungen, d.h. Messwerte einer bestimmten Variable (vgl. Berekhoven u.a. 1999, S. 69). Diese Daten bilden in Form einer Datenmatrix die Grundlage für die weitere Auswertung.

Beim Messen werden den beobachteten Merkmalsausprägungen nach bestimmten Verfahren Zahlen zugeordnet. In der Physik existieren zur Bestimmung von Länge, Gewicht, Zeit etc. entsprechende Abbildungsvorschriften. Solche Abbildungsvorschriften bzw. Zuordnungsregeln werden als Skalen bezeichnet (Meterstab, Anzeige an einer Waage bzw. einer Uhr etc.). Auch zur Messung von Einstellungen, Motivation, Bildungsstand, Schichtzugehörigkeit, Image, Involvement etc. in der Sozial-, Markt- und Meinungsforschung sind entsprechende Abbildungsvorschriften bzw. Skalen notwendig. Unterschiedliche Abbildungsvorschriften führen zu unterschiedlichen Skalen, wodurch die Messung auf unterschiedlichem Niveau erfolgen kann. Die Bestimmung des verwendeten Messniveaus (oder Skalenniveaus) ist daher wichtig für die weitere Verwendung der Daten.

"Der Informationsgehalt von Daten wird u.a. durch das Messniveau festgelegt, mit dem die Merkmalsausprägungen der Untersuchungsobjekte gemessen werden. Die verschiedenen Messniveaus sind durch eine Reihe von formalen Eigenschaften charakterisiert, die zugleich die bei der Datenanalyse verwendeten Methoden bestimmen." (Mayntz u.a. 1978, S. 38)

So ist es bei einer Befragung zu Ausgaben für das Telefonieren sicherlich von Interesse, wie hoch die durchschnittlichen Telefonkosten pro Monat sind, es macht jedoch keinen Sinn, den Mittelwert der Staatsangehörigkeit der Befragten zu berechnen.

Skalenniveaus
Die Skalentypen lassen sich nach dem Messniveau in vier Kategorien einteilen: Nominalskalen, Ordinalskalen, Intervallskalen und Ratioskala. Für jede Skalenart gelten bestimmte mathematische Eigenschaften.

Messniveau		Eigenschaften	Beschreibung	Beispiele
Nominalniveau		gleich/ungleich	Messwerte sind gleich oder ungleich	Geschlecht, Nationalität
Ordinalniveau		größer/kleiner	Messwerte lassen sich der Größe nach ordnen	Noten, Schicht, Bildung
quantitativ	Intervallniveau	Abstand (+/-)	Abstände zwischen den Messwerten sind angebbar	IQ-Punkte
	Rationalniveau	Verhältnis (./:)	Messwertverhältnisse können berechnet werden	Alter

Der Informationsgehalt der Daten hängt ganz wesentlich von ihrem Messniveau ab. So besteht für Daten auf Nominalskalenniveau lediglich die Möglichkeit sie auf ihre Gleichheit hin zu unterscheiden (z.B. männlich/weiblich). Daten auf Ordinalskalenniveau bieten zusätzlich die Möglichkeit einer Rangordnung (z.B. Unterschicht, Mittelschicht, Oberschicht). Bei Daten auf Intervallskalenniveau sind die Abstände (Intervalle) zwischen den einzelnen Merkmalsklassen gleich groß. Daten auf Rationalskalenniveau haben neben der Unterscheidungs- und Rangordnungsmöglichkeit sowie der gleichen Intervallgröße einen absoluten Nullpunkt. Der Messwert Null entspricht der tatsächlichen Abwesenheit des Merkmals. Daten auf Intervall- sowie Ratioskalenniveau werden vielfach auch als quantitative bzw. metrische Daten bezeichnet.

Mit zunehmendem Messniveau steigen die Aussagekraft und der Informationsgehalt der Daten wobei jedes Messniveau neben seinen charakteristischen Eigenschaften auch alle Eigenschaften der unteren Skalentypen besitzt. Die zentrale Bedeutung der mathematischen Eigenschaften der Skalenniveaus ist, dass sie auch die Rechenmöglichkeiten bestimmen. Alle statistischen Verfahren in der Datenauswertung richten sich nach dem Skalenniveau.

Zöfel Peter (2002): Statistik verstehen.
Ein Begleitbuch zur computergestützten Anwendung. Eine Datenmenge – was tun? Ein Leitfaden zu statistischen Lösungen, Addison-Wesley Verlag; Seite 11-20

Variablenklassifikation

Statistische Analysen können unter Zugrundelegung der verschiedensten Variablen vorgenommen werden. Da gibt es auf der einen Seite die quantitativen Variablen mit stetigen Messwerten wie z. B. Körpergröße oder Körpergewicht, welche im Prinzip beliebig genau gemessen werden können, und auf der anderen Seite qualitative Variablen wie z. B. Schulnoten oder die Codierung eines Merkmals wie den Familienstand in vier Kategorien. Diese qualitativen Variablen können nur diskrete Werte annehmen.

Eine genauere Einteilung der Variablen als die in qualitativ - quantitativ oder diskret - stetig ist diejenige nach vier verschiedenen Skalenniveaus (auch Messniveaus genannt), Bevor auf diese grundlegend wichtige Einteilung ausführlich eingegangen wird, soll zunächst der Begriff des Messens erläutert werden.

Das Messen
Der Begriff des „Messens" soll anhand einer in einer Klinik erhobenen Datenmenge erklärt werden. Von einem bestimmten Patientenkollektiv seien die folgenden Angaben erhoben worden:

- Geschlecht (männlich - weiblich)
- Alter
- Familienstand (ledig - verheiratet - verwitwet - geschieden)
- Körpergröße
- Körpergewicht
- systolischer Blutdruck
- diastolischer Blutdruck
- Cholesterin
- Triglyzeride
- Alkoholkonsum (keiner - mäßig - häufig - sehr häufig)
- Nikotinkonsum (Nichtraucher - mäßig - stark - sehr stark)

Die Werte dieser Variablen bei den einzelnen Fällen (hier: Patienten) bezeichnet man als Variablenwerte. Die Zuordnung der aktuellen Variablenwerte bei den einzelnen Fällen erfolgt mit einem Vorgang, den man „Messen" nennt. Betrachtet man etwa die Variable „Körpergröße", so ist klar, wie diese zu messen ist: Man legt ein Messband an und nimmt die Größe ab, wobei in der Regel eine Messgenauigkeit von 1 cm ausreichend ist. Das Körpergewicht misst man mit einer Waage, den Blutdruck mit einem Blutdruckmessgerät usw.

Etwas anders liegt der Fall bei der Variablen „Alter". Dieses misst man nicht mit Hilfe einer technischen Apparatur; man muss es erfragen oder etwa aus der Geburtsurkunde oder dem Personalausweis erschließen. Trotzdem kann man auch hier von „Messen" reden, wenn man die Definition des Messens wie folgt fasst:

Das Messen einer Variablen ist die Zuordnung von Zahlen zu den einzelnen Fällen.

Mit dieser Definition kann man auch Variablen wie das Geschlecht, den Familienstand oder den Alkohol- und Nikotinkonsum „messen". Beim Geschlecht ordnet man z. B. den Männern die Zahl 1 und den Frauen die Zahl 2 zu; beim Familienstand vergibt man für die gegebenen vier Kategorien die Zahlen 1 bis 4. Ebenso verfährt man beim Alkohol und Nikotinkonsum:

Geschlecht:	**Familienstand**:	**Alkoholkonsum**:	**Nikotinkonsum**:
1 = männlich	1 = ledig	1 = keiner	1 = Nichtraucher
2 = weiblich	2 = verheiratet	2 = mäßig	2 = mäßig
	3 = verwitwet	3 = häufig	3 = stark
	4 = geschieden	4 = sehr häufig	4 = sehr stark

Bei diesen Variablen erfolgt das „Messen" per Augenschein (Geschlecht) oder durch eine entsprechende Befragung. Die Zuordnung („Codierung") von Zahlen zu solchen „kategorialen" Variablen ist spätestens dann notwendig, wenn die statistische Analyse nicht per Hand, sondern unter Einsatz eines entsprechenden Statistik-Programmsystems mit Hilfe eines Computers erfolgen soll.

Skalenniveaus

Von entscheidender Wichtigkeit für die Auswahl eines korrekten statistischen Verfahrens ist die Feststellung des sogenannten Skalenniveaus (auch: Messniveaus) der beteiligten Variablen. Hier unterscheidet man das Nominal-, Ordinal-, Intervall- und Verhältnisniveau. Dabei werden diese Skalenniveaus wie folgt unterschieden.

Skalenniveau	**empirische Relevanz**
Nominal	Gleich - ungleich
Ordinal	Ordnung der Zahlen
Intervall	Differenzen der Zahlen
Rational	Verhältnisse der Zahlen

Nominalniveau

Betrachten wir zunächst das Geschlecht, so stellen wir fest, dass die Zuordnung der beiden Ziffern 1 und 2 willkürlich ist; man hätte sie auch anders herum oder mit anderen Ziffern vornehmen können.

Keinesfalls soll schließlich damit ausgedrückt werden, dass Frauen nach den Männern einzustufen sind; auch soll andererseits nicht die Bedeutung unterlegt werden, dass Frauen mehr wert sind als Männer. Den einzelnen Zahlen kommt also keinerlei empirische Bedeutung zu. Man spricht in diesem Falle von einer nominalskalierten Variablen. In dem hier vorliegenden Spezialfall einer nominalskalierten Variablen mit nur zwei Kategorien spricht man auch von einer dichotomen Variablen.

Eine nominalskalierte Variable ist auch der Familienstand; auch hier hat die Zuordnung der Ziffern zu den Kategorien des Familienstandes keinerlei empirische Relevanz. Im Gegensatz zum Geschlecht ist die Variable aber nicht dichotom; sie beinhaltet vier statt zwei Kategorien.
Nominalskalierte Variablen sind in ihrer Auswertungsmöglichkeit sehr eingeschränkt. Genau genommen können sie nur einer Häufigkeitsauszählung unterzogen werden, Die Berechnung etwa eines Mittelwertes, zumindest bei nicht-dichotomen Variablen, ist sinnlos.

Ordinalniveau
Betrachten wir als Nächstes die Rauchgewohnheit, so kommt den vergebenen Codezahlen insofern eine empirische Bedeutung zu, als sie eine Ordnungsrelation wiedergeben. Die Variable Rauchgewohnheit ist schließlich nach ihrer Wertigkeit aufsteigend geordnet: Ein mäßiger Raucher raucht mehr als ein Nichtraucher, ein starker Raucher mehr als ein mäßiger Raucher und ein sehr starker Raucher mehr als ein starker Raucher. Solche Variablen, bei denen den verwendeten Codezahlen eine empirische Bedeutung hinsichtlich ihrer Ordnung zukommt, nennt man ordinalskaliert.
Die empirische Relevanz dieser Codierung bezieht sich aber nicht auf die Differenz zweier Codezahlen. So ist zwar die Differenz zweier Codezahlen zwischen einem Nichtraucher und einem mäßigen Raucher einerseits und zwischen einem mäßigen Raucher und einem starken Raucher andererseits jeweils 1, man wird aber nicht sagen können, dass der tatsächliche Unterschied zwischen einem Nichtraucher und einem mäßigen Raucher einerseits und einem mäßigen Raucher und einem starken Raucher andererseits gleich ist; dafür sind die Begriffe zu vage. Entsprechendes gilt für den Alkoholkonsum; auch dies ist eine solche ordinalskalierte Variable.

Intervallniveau
Betrachten wir nun etwa die Körpergröße, so geben deren Werte nicht nur eine Rangordnung der beteiligten Personen wieder, auch den Differenzen zweier Werte kommt eine empirische Bedeutung zu. Hat etwa August ein Körpergewicht von 70 kg, Bertram eines von 80 kg und Christian von 90 kg, so kann man sagen, dass Bertram im Vergleich zu August um ebenso viel schwerer ist wie Christian im Vergleich zu Bertram (nämlich um 10 kg). Solche Variablen, bei denen der Differenz (dem Intervall) zwischen zwei Werten eine empirische Bedeutung zukommt, nennt man intervallskaliert. Ihre Bearbeitung unterliegt keinen Einschränkungen; so ist z. B. der Mittelwert ein sinnvoller statistischer Kennwert zur Beschreibung dieser Variablen. Weitere intervallskalierte Variablen im Beispiel der gegebenen Datenmenge sind das Alter, die Körpergröße, systolischer und diastolischer Blutdruck, das Cholesterin und die Triglyzeride.

Verhältnisniveau
Bei allen diesen Variablen kommt nicht nur der Differenz zweier Werte, sondern auch dem Verhältnis zweier Werte empirische Bedeutung zu. Ist etwa Emil 20 Jahre und Fritz 40 Jahre alt, so wird man sagen können, dass Fritz doppelt so alt ist wie Emil. Solche Variablen nennt man verhältnisskaliert. Es sind dies alle intervallskalierten Variablen, die den Wert Null annehmen können und dieser gleichzeitig der niedrigste denkbare Wert ist. Beispiele, bei denen dies nicht der Fall ist, sind etwa die in Grad Celsius gemessene Temperatur (wegen der möglichen Werte kleiner als Null) und der Intelligenzquotient (wegen des nicht möglichen Wertes von Null). Bei den in diesem Buch behandelten statistischen Verfahren kommt der Unterscheidung zwischen intervall- und verhältnisskalierten Variablen keine Bedeutung zu; es gibt nämlich darunter keine Verfahren, die Verhältnisniveau voraussetzen.

Weitere Beispiele für Nominal- und Ordinalniveau
Die Bestimmung des korrekten Skalenniveaus ist eine entscheidende Voraussetzung zur Auswahl des korrekten statistischen Verfahrens. Im folgenden Kapitel wird anhand passender Beispiele noch einmal etwas ausführlicher auf die Unterscheidung von Nominal und Ordinalniveau eingegangen. Häufig ist es nämlich möglich, nominalskaliert erscheinende Variablen durch geschickte Codierung auf Ordinalniveau zu bringen. Eine typische nominalskalierte Variable ist die Angabe des Berufs. Hier könnte etwa folgende Codierung gewählt werden, die beim besten Willen nicht in eine sinnvolle Ordnungsrelation gebracht werden kann:

1 = Angestellter
2 = Beamter
3 = Arbeiter
4 = Selbstständiger
5 = Hausfrau
6 = Auszubildender
7 = Rentner

Auch die Frage nach der Religionsgemeinschaft kann nur mit einer nominalskalierten Variablen realisiert werden, etwa mit folgender Codierung:

1 = evangelisch
2 = katholisch
3 = sonstige christliche Gemeinschaft
4 = andere Religionen
5 = ohne Religionsgemeinschaft

In einer Studie über Einschlafprobleme wurden die Gründe für die Schlafstörungen wie folgt codiert:

1 = Probleme
2 = Geräusche
3 = Tagesereignisse
4 = ungewohnte Umgebung
5 = Sonstiges

Auch hier ist eine andere als eine nominale Skalierung nicht denkbar.
Dichotome nominale Skalierungen sind häufig von der Art

1 =ja	1 = richtig	1 = trifft zu	1 = stimme ich zu
2 = nein	2 = falsch	2 = trifft nicht zu	2 = stimme ich nicht zu

So wie bekanntlich zwei Punkte eine Gerade bestimmen, die ansteigt oder geneigt ist, kann man bei dichotomen nominalskalierten Variablen stets von einer gegebenen Ordnungsrelation sprechen. So bedeutet etwa im Fall des letzten Beispiels eine niedrige Codierung Zustimmung, eine hohe Codierung Ablehnung. Dichotome nominalskalierte Variablen bilden also sozusagen den Übergang zwischen Nominal- und Ordinalniveau. Diesem wollen wir uns nun zuwenden.
Eine häufig gestellte Frage in einem Fragebogen ist die nach der Schulbildung. Eine ordinale Skalierung liegt etwa bei folgender Codierung vor:

1 = Volksschule
2 = Berufsschule
3 = Mittlere Reife
4 = Abitur
5 = Hochschule

Ein typisches Beispiel einer ordinalskalierten Variablen ist die Vorgabe einer Altersklassen-Einteilung in einem Fragebogen:

1 = bis 30 Jahre
2 = 31 - 50 Jahre
3 = über 50 Jahre

Ein solches Vorgehen ist eigentlich nicht empfehlenswert. Da jeder sein eigenes Alter sicherlich ohne Mühe exakt (in Jahren) angeben kann, sollte man dies auch so erfassen. Spätere Klasseneinteilungen können von einem Auswertungsprogramm gegebenenfalls immer noch vorgenommen werden; Sie haben dann aber Variationsmöglichkeiten und können bei Bedarf auch auf den genauen Wert zurückgreifen.

Klasseneinteilungen sollte man nur dann vorgeben, wenn die Ermittlung genauer Angaben zu umständlich oder gar nicht möglich ist. So wurde in einer Erhebung zum allgemeinärztlichen Vorgehen bei psychischen Erkrankungen bei den befragten Ärzten die Anzahl der Patienten pro Quartal abgefragt; dabei wurde folgende Codierung vorgegeben:

1 = unter 500
2 = 500 - 1000
3 = 1 000 -1500
4 = über 1500

Diese grobe Einteilung erscheint vernünftig, da genaue Zahlen wegen der Schwankungen von Quartal zu Quartal nicht angebbar sind. Aus diesem Grund stört es auch nicht, dass die Zahl 1000 einmal als Ober- und einmal als Untergrenze einer Klasse auftritt.

Ordinalskalierte Items treten häufig in psychologischen bzw. psychiatrischen Fragebögen auf. Im Freiburger Fragebogen zur Krankheitsverarbeitung z. B. werden 35 Aussagen der folgenden Art vorgegeben:

- Herunterspielen der Bedeutung und Tragweite
- Wunschdenken und Tagträumen nachhängen
- Aktive Anstrengungen zur Lösung des Problems unternehmen
- Stimmungsverbesserung durch Alkohol oder Beruhigungsmittel suchen
- Trost im religiösen Glauben suchen

Die befragten Personen sollen dann über eine Punktzahl zwischen 1 und 5 angeben, wie weit diese Aussagen für sie zutreffen oder nicht:

1 = gar nicht
2 = wenig
3 = mittelmäßig
4 = ziemlich
5 = sehr stark

In einem anderen Fragebogen über Gefühlslagen, wie man sie bezüglich Arbeit und Beruf haben kann (MBI), werden Aussagen wie die folgenden vorgegeben:

- Nach der Arbeit bin ich völlig fertig.
- Wenn ich zur Arbeit muss, bin ich schon morgens beim Aufstehen müde.
- Ich fühle mich energiegeladen.
- Mein Beruf frustriert mich.
- Ich finde, dass ich in meinem Beruf zu viel arbeite.

Hier wird zur Beantwortung eine Siebenerskala verwandt:

1 = völlig unzutreffend
2 = weitgehend unzutreffend
3 = eher unzutreffend
4 = weder noch bzw. weiß nicht
5 = eher zutreffend
6 = weitgehend zutreffend
7 = völlig zutreffend

Die Codierung bei den beiden letztgenannten Beispielen ist sozusagen um die jeweils mittlere Codierung symmetrisch. Dies ist nicht bei allen solchen Fragebögen der Fall. Betrachten wir etwa einige Aussagen aus dem Trierer Persönlichkeitsfragebogen:

- Ich fühle mich einsam.
- Ich bin unbeschwert und gut aufgelegt.
- Es macht mir Freude, anderen behilflich zu sein.
- Ich bin ein ruhiger, ausgeglichener Mensch.
- Meine Art kommt bei anderen gut an.

Diese Aussagen sind mit Hilfe einer Viererskala zu beantworten:
1 = immer
2 = oft
3 = manchmal
4 = nie

Überzeugungen in verschiedenen Lebenssituationen werden in einem Fragebogen der folgenden Art abgefragt (FKK):

- Ich komme mir manchmal taten- und ideenlos vor.
- Andere Menschen verhindern oft die Verwirklichung meiner Pläne.
- Ich weiß oft nicht, wie ich meine Wünsche verwirklichen soll.
- Ich kann sehr viel von dem, was in meinem Leben passiert, selbst bestimmen.
- Auch in schwierigen Situationen fallen mir immer viele Handlungsalternativen ein.

Hier ist zur Beantwortung eine symmetrische Sechserskala vorgesehen, die aber keine Codierung für eine unentschiedene Beurteilung enthält:

1 = völlig falsch
2 = weitgehend falsch
3 = eher falsch
4 = eher richtig
5 = weitgehend richtig
6 = völlig richtig

Immer wieder auftretende ordinalskalierte Variablen bei zahnmedizinischen Studien sind z. B. der Plaque-Index und der CPITN Wert. Letzterer ist ein pro Sextant ermittelter Behandlungs-Bedürftigkeits-Index mit folgender Codierung:

0 = gesundes Parodont
1 = Blutung
2 = Zahnstein
3 = Taschenbildung von 3,5 bis 5,5 mm
4 = Taschenbildung von 6 mm und mehr

Ähnliches gilt für die Codierung des Plaque-Indexes:

0 = keine Plaque
1 = vereinzelt Plaque-Inseln
2 = deutliche Plaque-Linie entlang des Gingiva-Randes
3 = Plaque-Ausdehnung im zervikalen Drittel des Zahnes
4 = Plaque-Ausdehnung bis ins zweite Zahndrittel
5 = Plaque-Ausdehnung bis über das zweite Drittel hinaus

Bei allen bisher genannten Beispielen liegt die ordinale Skalierung unmittelbar auf der Hand. In vielen anderen Fällen kann man eine solche nach etwas Nachdenken erkennen bzw. durch geschickte Codierung erreichen.

In einer Fragebogen-Untersuchung über die Heimatverbundenheit der Marburger Bevölkerung wurde u.a. nach dem Wohnort gefragt, wobei folgende Antwortmöglichkeiten vorgegeben waren:
1 = Kernstadt
2 = Stadtteil
3 = innerhalb des Landkreises
4 = außerhalb des Landkreises

Diese Variable ist ordinalskaliert, wenn man als Kriterium die Entfernung des Wohnortes vom Stadtzentrum zugrunde legt.

Eine andere Frage lautete „Freuen Sie sich, wenn Sie im Ausland Marburger treffen?"
Die vorgegebenen Antwortmöglichkeiten waren
1 =ja
2 = nein
3 = kommt drauf an

Dies ist eine ungeschickte Codierung; besser wäre die folgende:
1 =ja
2 = kommt drauf an
3 = nein

Dies wäre dann eine ordinale Skalierung: je höher die Codierung, desto geringer die Freude.
In einer biologischen Untersuchung über das Auftreten von Schmetterlingen wurden die meteorologischen Gegebenheiten abgefragt:
1 = Sonne
2 = leicht bewölkt
3 = Wolken

Legt man als Kriterium den Bewölkungsgrad zugrunde, so ist dies eine ordinalskalierte Variable: je höher die Codierung, desto größer der Bewölkungsgrad.
Der Übergang von Ordinal- zu Intervallniveau ist fliegend und eine Einordnung in eines der beiden Niveaus manchmal durchaus strittig. Während man beispielsweise die zwischen den Zahlen 1 und 6 vergebenen Schulnoten als ordinalskaliert ansieht, ist man bei den in der Oberstufe vergebenen Punktwerten von 0 bis 15 wohl eher geneigt, Intervallniveau anzunehmen. Auch bei Variablen, die bestimmte Anzahlen wiedergeben (z. B. Anzahl der Kinder in einer Familie), kann von Intervallniveau ausgegangen werden.
Der Thematik des Skalenniveaus wurde ein breiter Raum eingeräumt, da dessen korrekte Beachtung für die Auswahl des jeweils adäquaten statistischen Verfahrens entscheidend ist.

Elisabeth Aufhauser, unveröffentlichter Text
Unterrichtsmaterial Statistik-UE

Datenniveaus

Die Zeilen unserer Datenmatrix enthalten die Informationen zu den einzelnen Beobachtungen, die Spalten Information zur Variation der Werte auf den einzelnen Variablen über die verschiedenen Beobachtungen hinweg. Die k;i sind die zahlenmäßigen Werte, die für Beobachtung i(i=1,...,I) auf der Variablen k(k=1,...,K) gemessen wurden. I soll die Zahl der Beobachtungen (die Größe der Stichprobe) sein, K die Zahl der erhobenen Variablen.

Computer und Taschenrechner liefern uns im Regelfall nur zahlenmäßige Ergebnisse und benötigen für die statistische Analyse im Allgemeinen auch Zahlen als Ausgangsmaterial. Für unsere statistischen Analysen ist es daher sinnvoll, dass wir unseren Informationen zu den verschiedenen Erhebungsmerkmalen 'zahlenmäßige' Werte zuordnen. Wir müssen uns dabei immer folgende Fragen stellen:

• Mittels welcher Merkmalsvariablen kann das, was wir untersuchen wollen, grundsätzlich erfasst werden? Das entspricht der wichtigen Frage nach der Operationalisierung der Forschungsfrage bzw. einzelner Forschungshypothesen.

• Wie sollen die zu erfassenden Merkmale gemessen werden? Das entspricht der wichtigen Frage nach der Skala, auf der die verschiedenen Variablen gemessen werden sollen.

• Wie und auf welchem Genauigkeitsniveau sollen die gemessenen Werte (zahlenmäßig) verkodet werden? Das entspricht der wichtigen Frage, welche Merkmalsausprägungen wir unterscheiden wollen und wie wir sie in der Datenmatrix (zahlenmäßig) eingeben.

Bedeutung des Datenniveaus für die statistische Analyse
Zahlen haben grundsätzlich die Eigenschaft, dass sie addiert, multipliziert, dividiert .... werden können. Für die den Zahlen zugrundeliegenden (inhaltlichen) Sachverhalte können die Rechenoperationen aber unter Umständen sinnlos sein. Wichtig ist, dass der Computer mit Zahlenwerten alles rechnet, was wir anfordern, und von alleine nie weiß, ob die Rechenoperationen für die Variablen eigentlich auch inhaltlich Sinn machen. Diese Denkarbeit bleibt auch im Zeitalter der Supercomputer noch uns selbst überlassen. Welche statistischen Methoden dürfen wir daher für welche Daten verwenden?
Nach den erlaubten mathematischen Operationen sollten wir zwischen folgenden Typen an Datenniveaus unterscheiden:

(1) Nominaldaten

Beschreiben die Ausprägungen (Werte) einer Variablen nur verschiedene Zustände, Situationen usw. und sind die Ausprägungen nicht im Sinne einer Größer/Kleiner-Relation vergleichbar, handelt es sich um nominalskalierte Variablen. Im Regelfall können wir die einzelnen Ausprägungen dann auch nur verbal aufgrund unterschiedlicher Wertnamen (Worte, Buchstaben oder Zahlen) unterscheiden. Beispiele für Variablen, die typischerweise Nominaldatenniveau besitzen, wären etwa Geschlecht, Beruf, Bundesland u.ä. Ordnen wir den verschiedenen Wertausprägungen auf diesen Variablen Zahlen zu (etwa beim Geschlecht 1 für weiblich und 2 für männlich), so haben diese Zahlen keine inhaltliche Bedeutung und dienen nur dazu, das Rechnen im Computer zu erleichtern. Die Höhe der zahlenmäßigen Werte sagt inhaltlich absolut nichts aus, die Berechnung von Wertabständen ist sinnlos. Insbesonders ist es sinnlos zu sagen, die 2 sei größer oder mehr wert als die 1.

Außer den Vergleichsoperationen = und x dürfen für derartige Variablen aus inhaltlichen Gründen keine mathematischen Operationen durchgeführt werden. Wie wir oben gesehen haben, sind sogar Ordnungsrelationen (kleiner, größer) unsinnig. Auch die Addition ergibt keine sinnvollen und interpretierbaren Werte: !(weiblich) + 2 (männlich) = 3 (Das ergibt maximal ein heterosexuelles Paar, aber noch lange kein Geschlecht). Da die Addition nicht erlaubt ist, ist auch die Berechnung von Mittelwerten u.ä. völlig sinnlos. Wie sähe wohl ein Geschlecht mit dem Wert 1.4 aus? Dennoch gibt es eine Reihe an statistischen Analysemöglichkeiten. So ist es etwa möglich, für/mit nominale/n Variablen

- Häufigkeitsauszählungen durchzuführen,
- den Modus (häufigsten Wert) zu berechnen,
- Kreuztabellen zu erstellen,
- einen Chi-Quadrat-Tests auf Unabhängigkeit durchzuführen,
- sog. Assoziationsmaße zu berechnen,
- sie als unabhängige Faktoren in der Varianzanalyse zu verwenden.

(2) Ordinaldaten

Können die Ausprägungen einer Variable im Sinne von Größer/Kleiner-Relationen miteinander verglichen und in eine Rangordnung gebracht werden, so sprechen wir von ordinal- oder rangskalierten Variablen. Beispiele für Variablen, die typischerweise Ordinaldatenniveau besitzen, sind etwa die Höhe des Schulabschlusses, die Bewertung der Umweltbelastung entlang einer mehrstufigen Skala (von gar nicht bis extrem), Einstellungsmessungen entlang einer Skala der Form stimme zu - stimme eher zu - unentschieden - stimme eher nicht zu - stimme gar nicht zu. Die Zahlen, die den verschiedenen Wertausprägungen auf derartigen Variablen zugeordnet werden, müssen im Sinne einer größenmäßigen Ordnung interpretierbar sein, die größenmäßigen Abstände zwischen den Werten sagen aber nichts über die tatsächlichen Wertunterschiede aus. Das bedeutet dann vor allem, dass bei rangskalierten Variablen auch keine sinnvollen Differenzen gebildet werden können. Beispiel Schulabschluss: 1 Pflichtschule, 2 Lehre, 3 berufsbildende mittlere Schule, 4 allgemeinbildende höhere Schule, 5 berufsbildende höhere Schule, 6 akademische Ausbildung. Sowohl die Differenz zwischen akademischer Ausbildung (6) und allgemeinbildender höherer Schulbildung (4) als auch die Differenz zwischen berufsbildender mittlerer Schule (3) und Pflichtschule (1) beträgt rein rechnerisch 2. Dennoch sind die beiden Differenzen nicht zu vergleichen - und selbstverständlich auch nicht dem Variablenwert 2 (Lehre) gleichzusetzen.

Ordnungsrelationen =, <, > sind die einzigen mathematischen Operationen, die für Variablen auf dem Ordinaldatenniveau erlaubt sind. Addition, Mittelwertbildung u.ä. sind auch für Ordinaldaten nicht erlaubt. So ergibt etwa ein Lehrabschluss (2) plus ein AHS-Abschluss (4) keinen akademischen Titel (6). Und ein mittlerer Ausbildungsgrad von 2.3 in der österreichischen Bevölkerung ist natürlich ebenfalls nicht sehr aussagekräftig. Zusätzlich zu all den statistischen Verfahren, fahren, die mit Nominaldaten möglich sind, können wir mit Variablen auf Ordinaldatenniveau u.a. noch folgende statistische Analysen vornehmen:

- den Median (50% Wertes) und andere Percentil-Werte berechnen;
- Rangkorrelationskoeffizienten berechnen;
- nichtparametrischer Rang- und Verteilungstests durchführen.

Nominal- und Ordinaldaten werden gemeinsam auch als nicht-metrische, diskrete, kategoriale oder qualitative Datenniveaus bezeichnet. Im Unterschied dazu werden Intervall- und Absolutdaten gemeinsam als metrische, kontinuierliche oder quantitative Datenniveaus

bezeichnet. Generell gilt, dass metrischen Skalen konstante Messeinheiten (Grad, Schilling, Altersjahre ...) zugrunde liegen müssen, damit die Berechnung von Wertdifferenzen und Wertsummen inhaltlich Sinn macht. Es kann daher für metrische Daten nicht nur festgestellt werden, ob eine Beobachtung einen größeren Variablenwert aufweist als eine andere, sondern auch, wie groß der Unterschied zwischen den beobachteten Werten ist.

(3) Intervalldaten
Sog. intervallskalierte Variablen besitzen keinen absoluten Nullpunkt, es wird höchstens per Übereinkunft irgendwo ein Nullpunkt festgelegt. Ein typisches Beispiel wären etwa Temperaturskalen. Während Additionen und Subtraktionen für derartige Variablen auch inhaltlich Sinn machen, sind Multiplikationen und Divisionen von Variablenwerten inhaltlich sinnlos! Die Differenz zwischen 15 und 30 Grad Celsius beträgt 15 Grad Celsius, bei 30 Grad Celsius ist es aber physikalisch gesehen nicht doppelt so warm wie bei 15 Grad Celsius. Für die Differenzierung bei den komplexeren statistischen Verfahren ist die Unterscheidung zwischen Intervall- und Absolutdatenniveau an sich unwesentlich.

(4) Daten mit Absolutniveau
Bei Variablen mit Absolutdatenniveau ist der Wert 0 auch der absolute Nullpunkt. Typische Variablen mit Intervalldatenniveau sind etwa Lebensalter, monatliches Erwerbseinkommen, die Einwohnerzahl einer Gemeinde, Distanzen in Kilometer u.ä. Das Vielfache eines Variablenwertes kann auch inhaltlich als Vielfaches interpretiert werden. Es sind damit auch Multiplikationen und Divisionen von Variablenwerten möglich, sinnvoll und interpretierbar. Zumindest vom Datenniveau her dürfen mit derartigen Variablen alle möglichen statistischen Verfahren durchgeführt werden. Aber Achtung: Den meisten statistischen Verfahren liegen neben dem Datenniveau auch noch andere wichtige Annahmen zugrunde, die nicht unbedingt erfüllt sein müssen.
Die Skalenniveaus bilden in der genannten Reihenfolge eine Hierarchie mit zunehmendem Informationsgehalt. Eine Variable mit einem bestimmten Skalenniveau kann immer in eine Variable mit einem niedrigeren Skalenniveau umgewandelt werden. Der umgekehrte Schritt ist nicht möglich.

(5) (0,1) Dummy-Variablen

Dichotome 0-1 oder Dummy-Variablen haben nur zwei Wertausprägungen, etwa ja/nein, trifft zu/trifft nicht zu u.ä. Für die statistische Analyse ist es sinnvoll, die Ausprägungen derartiger Variablen mit den Werten 0(für nein, trifft nicht zu u.ä.) bzw. 1(für ja, trifft zu u.ä.) zu verkoden. Wichtig ist nun, dass derartige 0-1 Variablen metrisches Datenniveau besitzen. Die Addition als grundlegende mathematische Operation und eine darauf basierende Mittelwertbildung machen auch inhaltlich Sinn: Die Summe aller Beobachtungswerte einer (0,1) Variable dividiert durch die Zahl der Beobachtungen entspricht dem Anteil der 1-Werte (der Ja-, Trifft zu-Werte) im Datensatz. Und: Anteils- oder Prozentwerte weisen metrisches Datenniveau auf.

Die Umwandlung von Variablen mit kategorialem Datenniveau (Nominal- oder Ordinaldatenniveau) in eine Reihe von (0,1)-Variablen ist ein durchaus üblicher statistischer Trick, um kategoriale Variablen auch in Methoden verwenden zu können., die eigentlich nur für metrische Variablen geeignet sind (etwa als erklärende, unabhängige Variablen in der Regressionsanalyse).