Deskriptive Statistik

1. Datenstrukturen

im Zusammenhang mit statistischen Untersuchungen in der Biologie

Zur Beantwortung biologischer Fragen werden in der Regel an Pflanzen, Tieren, Menschen oder Mikroorganismen Experimente und Beobachtungen durchgeführt. So könnten wir z.B. Interesse haben an der Frage nach dem Proteingehalt des Blutes von adulten Mäuse des Stammes C57bl. „Alle“ adulten Mäuse dieses Stammes würden wir dann als Grundgesamtheit für die Frage und die untersuchten Tiere als eine Stichprobe der Grundgesamtheit bezeichnen. Um aus der durch den Stichprobenumfang begrenzten Datenmenge (Untersuchungsergebnisse) die optimalen Schlüsse für Grundgesamtheit ziehen zu können, müssen die Daten statistisch aufbereitet werden.

Die gewonnenen Daten können abhängig von der Fragestellung sehr unterschiedlicher Struktur sein. Es ist leicht einzusehen, dass bei der Bearbeitung so verschiedene Daten wie der Masse eines geernteten Mais-Saatguts (Messwert), der Farbe der Rückenflosse eines Fisches (verbale Beschreibung) oder der Anzahl der auf einer Agarplatte gewachsenen Bakterienkolonien (Zählwert), unterschiedliche statistische Verfahren zur Anwendung kommen.

Bei der Auswahl der adäquaten Verfahren sind Kenntnisse der verschiedenen Datenstrukturen notwendig. Solche darzustellen ist die Aufgabe dieses Textes, wobei folgende Begriffe gegeneinander abgegrenzt und erklärt werden sollen.

Grundgesamtheit, statistische Masse, Kollektiv, Stichprobe, Element des Kollektivs, Merkmalsträger, Merkmal, Variable, Merkmalsausprägung, quantitative Daten, qualitative Daten, Nominaldaten, Ordinaldaten, Rangdaten, Kategorialdaten, Scores, diskrete Daten, stetige Daten, quasi-stetige Daten, gekörnte Daten, Proportionaldaten, Intervalldaten, quantifizieren von Daten, qualifizieren von Daten

Beispiel
In einem Zusammenhang, der hier ohne Bedeutung ist, besteht Interesse an morphologischen und physiologischen Eigenschaften adulter Meerschweinchen. So könnte etwa gefragt werden nach Fellstruktur, Fellfarbe, Körperlänge, Organmassen und Blutzellzahlen.

Die Gruppe von Meerschweinchen, aus der ein Tier für die Untersuchung zufällig ausgewählt wird, bezeichnen wir als Grundgesamtheit, statistische Masse oder Kollektiv. Das untersuchte Tier, ist ein Element der Stichprobe bzw. eines Teiles einer Stickprobe und damit auch eines des Kollektivs. Das Tier hat bestimmte Eigenschaften oder Merkmale (X), deren Ausprägungen (x) von Tier zu Tier variieren können. Die Merkmale werden daher auch als Variable bezeichnet und das Tier als Merkmalsträger.

Bei der Untersuchung eines Tieres wurden folgende Daten als Merkmalsausprägungen gewonnen:

Kollektiv: Meerschweinchen	**Element: Tier Nr.4**	
Merkmal (X)	**Merkmals-ausprägung (x)**	**Datenstruktur**
Fellstruktur	glatthaarig	qualitativ, nominal
Geschlecht	weiblich	qualitativ, nominal
Fellfarbe	schwarz	qualitativ, nominal
Länge	21 cm	quantitativ, stetig, quasi-stetig, gekörnt
Masse	1220 g	quantitativ, stetig, quasi-stetig, gekörnt
Motilität	vermindert	qualitativ, ordinal
Pulsfrequenz	270 min^{-1}	quantitativ, diskret
Hb-Konzentration im Blut	15,5 g/100 mL	quantitativ, stetig, quasi-stetig, gekörnt
Anzahl der Erythrozyten	5,7*10^6/µL	quantitativ, diskret
Masse der Leber	31,2 g	quantitativ, stetig, quasi-stetig, gekörnt
Gesundheitszustand	gesund	qualitativ, ordinal
Körpertemperatur	36,9 °C	quantitativ, stetig, quasi-stetig, gekörnt

Tabelle 1

Wenn, wie bei Experimenten üblich, eine Stichprobe von Merkmalsträgern untersucht wird, dann kann es zweckmäßig sein, die Merkmalsausprägungen zu Vergleichszwecken an Skalen graphisch darzustellen. Es ist unmittelbar einleuchtend, dass wir die Fellfarben verschiedener Tiere und deren Körpertemperaturen nicht an einem Skalentyp, darstellen können.

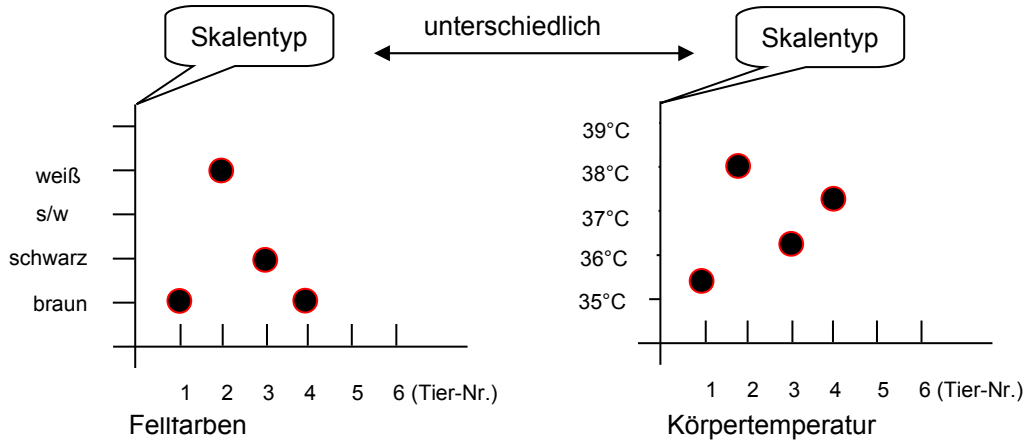


Abb.1 Abb.2

Wir benötigen unterschiedliche Skalentypen und sprechen von unterschiedlich skalierten Daten. Es sollen nun Daten, die primär [siehe scoring] verbal beschrieben werden (z.B. Gesundheitszustand), von solchen, die durch Zahlen dargestellt werden (z.B. Körpertemperatur) gegeneinander abgegrenzt und jeweils differenziert dargestellt werden.

1.1 Qualitative Daten

Die Ausprägung der Merkmale Geschlecht, Fellfarbe, Gesundheitszustand und Motilität des Meerschweinchens seien „weiblich“, „schwarz“, „leicht erkrankt“, „stark vermindert“. Da die Daten nicht durch Zahlen sondern durch Worte dargestellt werden, nennen wir sie qualitative Daten, sie kennzeichnen ein qualitatives Merkmal und werden durch Vergleiche gewonnen. Im Vergleich mit einer Farbskala stellen wir fest, dass das Fell des Tieres schwarz ist. Wir vergleichen den Zustand erkrankter Tiere mit dem bekannten Bild gesunder Tiere.

Je nachdem, ob wir den Daten eine Wertung zusprechen können oder nicht, differenzieren wir in qualitative Daten in Ordinaldaten und Nominaldaten.

1.1.1 Nominaldaten

Die Merkmalsausprägungen wie Fellfarbe und Geschlecht sind wertfrei. Sie können nicht in eine Rangfolge gebracht werden. Ein schwarze Fell hat keinen höheren Wert als ein braunes und die Qualität „weiblich“ ist nicht mehr wert als die Qualität „männlich“. Derartige Merkmalsausprägungen werden an einer Nominalskala dargestellt. Auf dieser sind die Abstände zwischen zwei Marken zwar in der Regel metrisch gleich haben aber keine quantitative Bedeutung. Wir nennen sie Nominaldaten oder nominalskalierte Daten, sie werden durch Vergleich gewonnen. Rechnen können wir mit ihnen zunächst nicht.

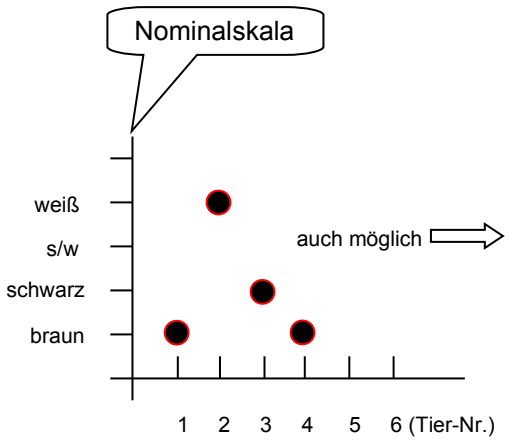


Abb.3

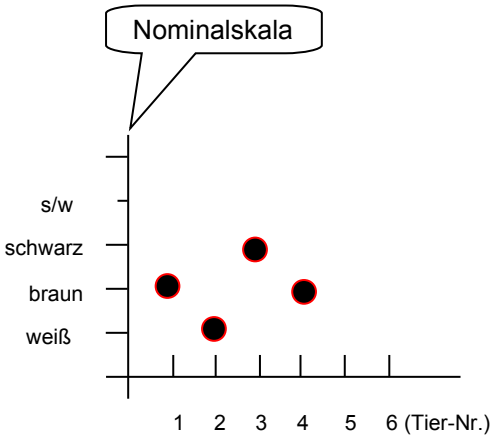


Abb.4

1.1.2 Ordninaldaten, Rangdaten, Kategorialdaten

Anders sieht es bei den Ausprägungen des Gesundheitszustandes oder der Motilität aus. Auch hier werden die Ausprägungen mit Worten beschrieben. Beim Gesundheitszustand können wir bei den Merkmalsausprägungen wählen zwischen z.B. „gesund“,„leicht erkrankt“, „schwer erkrankt“, „agonale Phase“. Wo die Grenzen liegen, steht hier nicht zur Debatte; aber sicher stellen die Ausprägungen eine Rangfolge dar. Das gleiche gilt für die Schätzung der Motilitätsstärke, wir können sie „normal“, „leicht vermindert“, „stark vermindert“, „leicht erhöht“ oder „stark erhöht“ bezeichnen. Daten dieser Art werden an einer Rang- oder Ordinalskala dargestellt, denn wir können sie in eine Rangfolge oder Ordinalfolge bringen oder sie bestimmten Kategorien zuordnen. Wir nennen sie rang- oder ordinalskalierte Daten oder Kategorialdaten. An einer solchen Skala sind die Abstände zwischen den Marken nicht gleichwertig, wenn auch in der Graphik die Abstände metrisch gleich sind, sie stellen jedoch eine Rangabstufung dar. Obwohl auf der Skala „schwer erkrankt“ doppelt so weit von „gesund“ entfernt ist wie „leicht erkrankt“ können wir nicht sagen, „gesund“ ist doppelt so viel wert wie „schwer“ erkrankt“. Auch Ordinaldaten werden durch Vergleich gewonnen und wir können zunächst nicht mit ihnen rechnen.

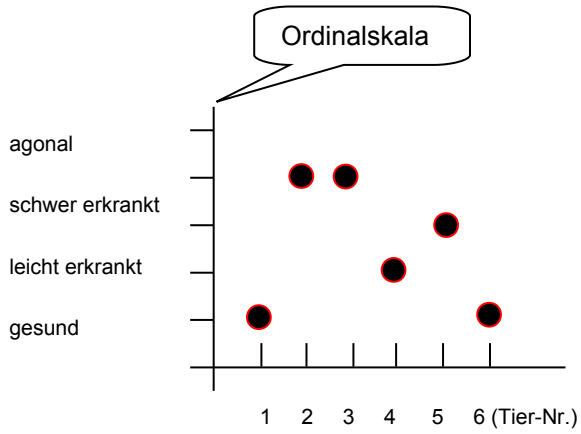


Abb.5

1.1.2.1 scoring (Punkte vergeben)

Um Ordinaldaten in gewisser Weise Rechenoperationen zugänglich zu machen, unterziehen wir sie einem scoring. Hierunter verstehen wir die Zuordnung von Ordinaldaten zu Ziffern, die eine Wertung

darstellen. Um beim Beispiel des Gesundheitszustands zu bleiben: Wir können die Merkmalsausprägungen in der folgenden Weise Zahlen (scores) zuordnen:

Ordinaldaten	**score**
gesund	0
leicht erkrankt	1
schwer erkrankt	2
agonale Phase	3

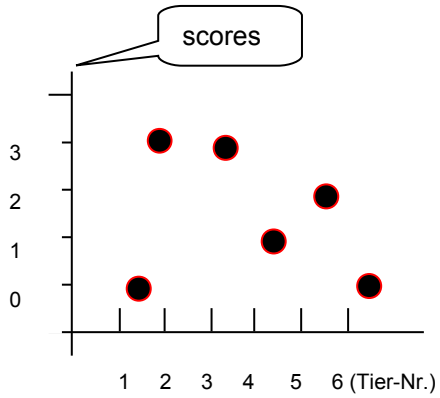


Abb.6

Die Abstände der Punkte (scores) auf der Skala haben keine quantitative Bedeutung. Bedeutsam ist nur die Rangfolge der Punkte. Der Vorteil dieses scoring liegt darin, dass wir mit den scores Rechenoperationen durchführen können.

1.2 Quantitative Daten

Das Gemeinsame der folgenden Daten des Meerschweinchens ist, dass sie Zahlen darstellen. Es sind somit quantitative Daten, die an quantitativen Skalen dargestellt werden. .

Das sind: Masse der Leber 31,2 g; Körperlänge 21 cm; Körpermasse 1220 g; Körpertemperatur 36,9 °C; Hb-Konzentration im Blut 15,5 g/100 mL; Anzahl der Erythrozyten $5{,}7*10^6$/µL; Pulsfrequenz 270 min^{-1}.

Diese Daten unterscheiden sich aber dadurch, dass die einen durch einen Messvorgang (stetige Daten z.B. Körpermasse) gefunden wurden, die anderen durch einen Zählvorgang (diskrete Daten z.B. Erythrozytenzahl).

1.2.1 Diskrete Daten

Die Erys/µL und die Häufigkeit der Herzschläge pro Minute werden durch einen Zählvorgang ermittelt. Zählwerte können prinzipiell nur ganzzahlig sein. Sie werden an einer diskreten, diskontinuierlichen Skala dargestellt, die nur ganze Zahlen trägt. Halbe Erythrozyten und halbe Herzschläge gibt es nicht.

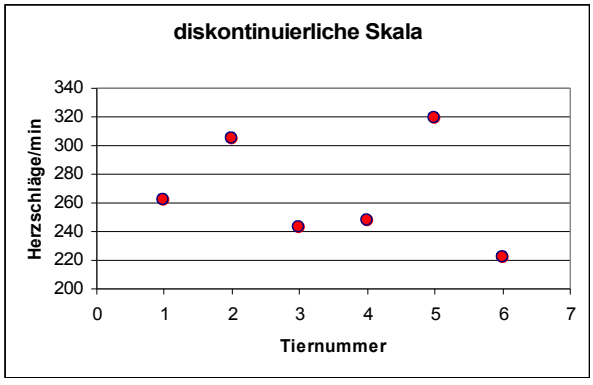


Abb.7

1.2.2 Stetige Daten

Die Masse der Leber 31,2 g; Körperlänge 21 cm; Körpermasse 1220 g; Körpertemperatur 36,9 °C; Hb-Konzentration im Blut 15,5 g/100 mL werden durch Messvorgänge ermittelt. Messwerte sind grundsätzlich kontinuierlich, das bedeutet, sie können jeden beliebigen Wert zwischen zwei ganzen Zahlen inclusive dieser, annehmen. Die Anzahl der Nachkommastellen ist nur durch das Messverfahren begrenzt. Sie werden an einer stetigen oder kontinuierlichen Skala dargestellt.

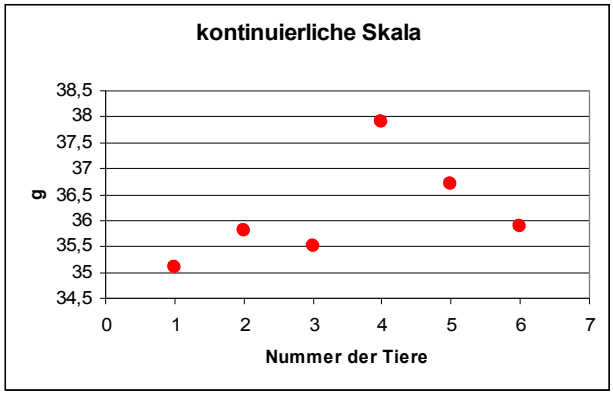


Abb.8

1.2.2.1 Gekörnte Daten, quasi-stetige Daten

Stetige Daten gibt es nicht wirklich, da die Meßgeräte mit denen sie gewonnen werden, immer eine endliche Stellenzahl vorgeben. Es gibt kein Meßgerät, mit dem man bestimmen könnte, dass ein Tier

36,6500753345 g wiegt. Viele Körperwaagen zeigen die Masse nur in kg mit einer Nachkommastelle an. Die Wägung eines Hundes könnte 14,7 kg oder 14,8 kg ergeben. Werte dazwischen werden nicht angezeigt. Das bedeutet, dass (vom Prinzip her) stetige Daten durch das Messgerät diskret geworden sind. Wir nennen sie quasi-stetige Daten oder gekörnte Daten. Mit einer Uhr, deren Zeiger in Minutenabständen springt, können wir nur ganze Minuten messen. Zeitmesswerte sind gekörnte Daten.

1.2.2.2 Proportionaldaten und Intervalldaten

Wenn wir die Körpertemperatur von 36,9 °C mit der Körperlänge von 21 cm vergleichen, dann stellen wir fest, dass ihre Skalen sich bezüglich des Nullpunktes unterscheiden. Stetige Daten können wir wiederum zwei Gruppen zuordnen, je nachdem ob die Skala, an der wir sie darstellen können, einen absoluten Nullpunkt hat oder nicht. Zur Verdeutlichung des Unterschiedes betrachten wir die Körpermasse zweier adulter Mäuse und die Wassertemperaturen zweier Aquarien. Tier Nr.1 wiegt 40 g und Tier Nr.2 wiegt 20 g. In Aquarium Nr.1 messen wir 20 °C und in Nr. 2 messen wir 10 °C.

Tier 1 ist zweifelsfrei doppelt so schwer wie Tier 2. Die Temperatur im Aquarium 1 ist aber nicht doppelt so hoch wie in Aquarium 2. Der Grund: Die Celsiusskala hat keinen absoluten Nullpunkt, er ist willkürlich festgelegt. Solche Skalen, nennen wir Intervallskalen. Die Massenskala der Waage dagegen hat einen absoluten Nullpunkt. Ebenso die Skala, an der wir z.B. die Stoffmengenkonzentration von Hämoglobin im Blut eintragen können. Diese Skalen nennen wir Proportionalskalen oder Verhältnisskalen.

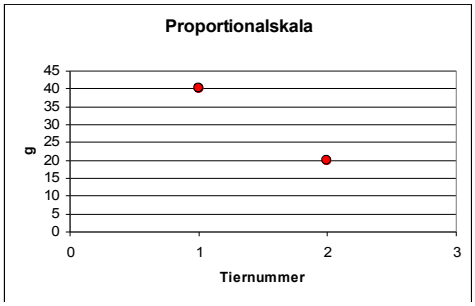


Abb.9

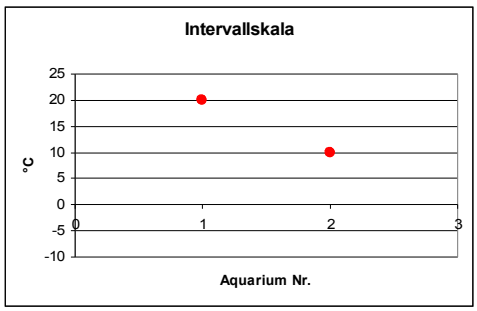


Abb.10

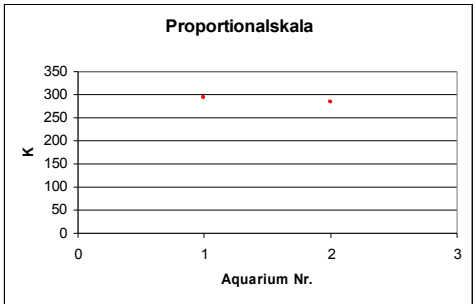


Abb.11

Wenn wir Intervalldaten an einer Proportionalskala darstellen wollen, dann müssen wir sie, wenn es möglich ist, in ein System transformieren, welches einen absoluten Nullpunkt hat. Bei der Temperatur wäre das leicht möglich, wenn wir die Kelvinskala zugrunde legt. 20 °C entsprechen 293 K und 10 °C entsprechen 283 K. Hier sehen wir deutlich, dass von einer Verdopplung der Temperatur keine Rede sein kann.

Ein Beispiel für intervallskalierte Daten sind auch die Kalenderdaten, deren Nullpunkte in verchiedenen Kulturkreisen unterschiedlich festgelegt wurden.

Wozu die Kenntnis diverser Datenstrukturen notwendig ist, zeigt das folgende Beispiel. Bei der Prüfung der Wirkung eines Medikaments zur Steigerung des Hämoglobinkonzentration im Blut erhalten wir Messwerte (stetige Daten). Ein statistischer Test zur Prüfung der Wirkung des Medikaments im Vergleich mit einem anderen Medikament wäre z.B. der t-Test. Dieser darf nur auf stetige Daten, angewendet werden, nicht aber auf Zählwerte. Wir müssen bei der Auswahl eines Tests also die Datenstruktur richtig einschätzen können.

1.3 Quantifizieren qualitativer Daten

Bei statistische Berechnungen können nur Zahlen bearbeitet werden. Qualitative Daten sind aber keine Zahlen. Um solche Daten für Rechenoperationen verwendbar zu machen, kann man sie in geeigneter Weise in Zahlen transformieren. Der Vorgang dieses Quantifizieren von Nominal- und

Ordinaldaten besteht im Auszählen der mit den jeweiligen Merkmalsausprägungen besetzten Merkmalsträger (siehe Scoring).

Beispiel
Bei der Untersuchung einer Gruppe von 17 mit Trypanosomen infizierten Mäusen, die mit einem Medikament behandelt wurden, erhielten wir nach der Therapie das folgende für jedes Tier qualitative Ergebnis:

Maus 1	geheilt	Maus 10	geheilt
Maus 2	geheilt	Maus 11	nicht geheilt
Maus 3	geheilt	Maus 12	geheilt
Maus 4	nicht geheilt	Maus 13	nicht geheilt
Maus 5	nicht geheilt	Maus 14	geheilt
Maus 6	geheilt	Maus 15	nicht geheilt
Maus 7	geheilt	Maus 16	geheilt
Maus 8	geheilt	Maus 17	geheilt
Maus 9	geheilt		

Tabelle 2

Nun können wir durch Zählen der geheilten und nicht geheilten Tiere die für die einzelnen Tiere qualitativen Merkmalsausprägungen für die Gruppe quantifizieren:

geheilt 12 mal (das sind 71%)
nicht geheilt 5 mal (das sind 29%)

Hätten wir eine zweite Gruppe von 19 Mäusen mit einem anderen Medikament behandelt und das Ergebnis wäre 16 geheilt (84%) und 3 nicht geheilt (16%), dann könnten wir jetzt fragen, ob die höhere Heilquote des zweiten Medikaments „wirklich“ für bessere Heilchancen spricht als das erste Medikament, oder ob der Unterschied (71% - 84%) durch den Zufall zu erklären ist. Dies kann anhand der nun vorliegenden quantitativen Daten mit einen Signifikanztest, z.B. den Chi²-Test geprüft werden.

1.4 Qualifizieren quantitativer Daten

Bei manchen Fragen erscheint es sinnvoll, quantitative Daten, z.B. die Massen von Merkmalsträgern, zu qualifizieren.

Beispiel
Es liegen folgende Messwerte der Massen von Meerschweinchen in g vor:

x_1	x_2	x_3	x_4	x_5	x_6
1240	1320	1190	1590	925	1071

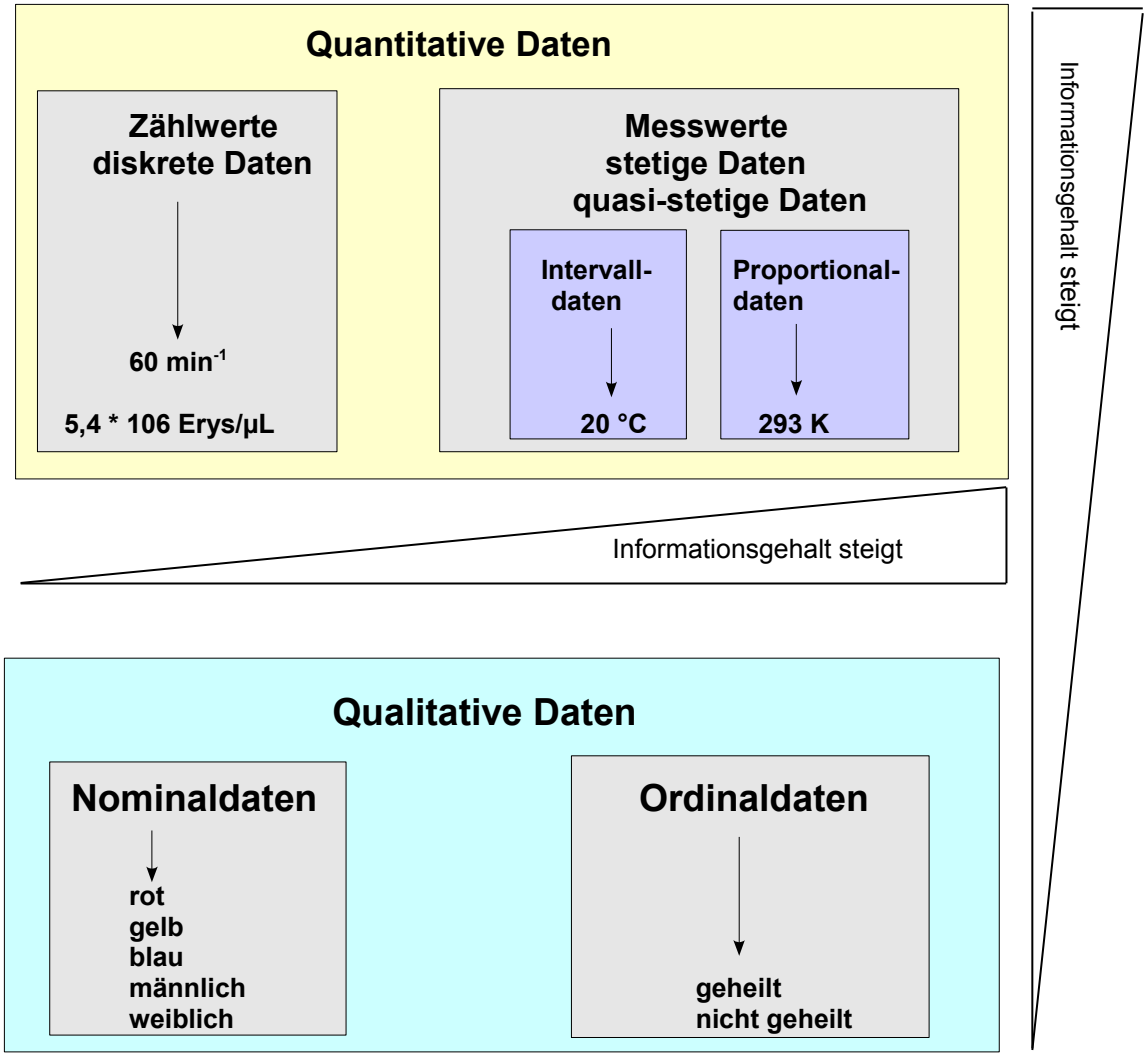
quantitative Daten

Für einen geplanten Versuch können nur Meerschweinchen mit der Masse x, mit $1100\ g \leq x \leq 1500\ g$ verwendet werden. Wir können entsprechend notieren (- = zu leicht, + = zu schwer, 0 verwendbar).

x_1	x_2	x_3	x_4	x_5	x_6	
1240	1320	1190	1590	925	1071	quantitative Daten
↓	↓	↓	↓	↓	↓	↓
0	0	0	+	-	-	qualitative Daten

Es wurde qualifiziert und die nicht verwendbaren Tiere (- und +) können wir nun leicht aussortieren.
Die Qualifikation quantitativer Daten ist immer mit einem Informationsverlust verbunden und sollte daher nicht ohne Not betrieben werden. x_5 und x_6 haben zwar den gleichen qualitativen Wert (-) aber unterschiedliche quantitative Werte (925 und 1017).

1.5 Übersicht über Datenstrukturen



1.6 Übungen

Übung 1
Geben Sie zu den Punkten a bis g an, welche Datenstruktur vorliegt und an welchen Skalen sie dargestellt werden können. Begründen Sie Ihre Entscheidung.

a Fellfarbe bei der Kreuzung von Mäusen
b Extinktion bei einer photometrischen Bestimmung
c Nachweis ob Chloridionen im Wasser sind oder nicht.
d Gesundheitszustand von Meerschweinchen mit pilzinfizierter Haut
e Rectaltemperatur in °C von Ratten in Langzeitnarkose
f Gastemperatur in K in einem Eudiometer.
g Suchen Sie Beispiele für die Qualifizierung quantitativer Daten

Lösungen
a Qualitative Daten — Nominaldaten — Nominalskala
weil: Fellfarbe ist wertfrei

b Quantitative Daten — stetige Daten — stetige Skala
weil: Extinktionen sind Meßwerte, damit im Prinzip stetig, aber durch das Meßgerät gekörnt. Die Skala hat einen natürlichen Nullpunkt, also Proportionalskala

c Qualitative Daten — Ordinaldaten — Ordinalskala
weil: Das Ergebnis der Untersuchung ist „ja“ oder „nein“. Ob das als Nominal- oder Ordinaldatum zu werten ist, hängt davon ab, welche Bedeutung das Ergebnis für die Untersuchung hat. Will man nur wissen ob „ja“ oder „nein“, dann sind es Nominaldaten, Ist der positive Ausfall unerwünscht, dann kann man das Ergebnis als Ordinaldatum werten.

d Qualitative Daten — Ordinaldaten — Ordinatskala
weil: Die möglichen Merkmalsausprägungen sind wertbehaftet.

e Quantitative Daten — stetige Daten — stetige Skala — Intervallskala
weil: Temperaturen sind Meßwerte, daher stetig aber durch das Meßgerät gekörnt. Die Skala hat einen willkürlichen Nullpunkt, daher Intervallskala.

f Quantitative Daten — stetige Daten — stetige Skala — Proportionalskala
weil: Temperaturen sind Meßwerte, daher stetig aber durch das Meßgerät gekörnt. Die Skala hat einen natürlichen Nullpunkt, daher Proportionaiskala.

g Ich habe von gestern auf heute 500 g zugenommen und werde nach meiner Gewichtsänderung gefragt. Ich kann sagen „500 g“ (quantitative Angabe) ich kann aber auch sagen „ein wenig zugenommen“. Das wäre dann qualifiziert.

Übung 2
Geben Sie zu den Punkten a bis g an, welche Datenstruktur vorliegt und an welchen Skalen sie dargestellt werden können. Begründen Sie Ihre Entscheidung.

a Bei einer Gelelektrophorese wurden die Laufstrecken von DNA Fragmenten gemessen. Welche Datenforrn liegt vor?
Lösung: Die Laufstrecken werden in mm gemessen, quasi-stetige Daten, kontinuierliche Skala.

b Bei der Prüfung der Wirksamkeit eines Herbizids wird dieses in verschiedenen Dosen an Kressepflanzen geprüft. Man wertet die Länge der Pflanzen aus. In welcher Datenform liegen die Ergebnisse vor?
Lösung: Die Längen der Pflanzen werden in mm gemessen, quasi-stetige Daten, kontinuierliche Skala.

c An Mäusen wird ein Präparat geprüft, welches einen unsicheren Gang (Ataxie) hervorruft. Welcher Art könnten die Daten sein, die man bei der vergleichenden Untersuchung verschiedener Dosen erhält?
Lösung: Man wird Befunde wie „nicht auffällig“, „schwach gestört“, »stark gestört“, „keine Laufbewegung“ notieren. Das sind Ordinaldaten, die durch Scoring in rechenbare Zahlen transformiert werden.

d In Bakterien wurde eine Plasmid eingeführt welches im positiven Ausfall zu einer roten Koloniefärbung führte. Welche Datenstruktur stellt das Ergebnis der Auszählung der roten und nichtroten Kolonien dar?
Lösung: Es gibt die Katergorien „rot“ und »nicht rot“. Das sind Nominaldaten, deren Häufigkeiten ausgezählt werden kommen.

e Im Venenblut wird im Zusammenhang mit einer Lebererkrankung die Aktivität einer Transaminase bestimmt. In welcher Datenform liegt das zu bewertende Ergebnis vor? Lösung: Es werden U/L gemessen. Das sind quasi-stetige Meßwerte, die an einer kontinierlichen Skala aufgetragen werden.

f Suchen Sie ein Beispiel für die Quantifizierung qualitativer Daten.
Lösung: Bei der Beurteilung des Wachstums von Pflanzen, die unterschiedlich gedüngt wurden, kann man die Bewertung z.B. durch Formulierungen wie „Keine Änderung zu Kontrolle“, »schwache Wachstumsverstärkung“, “starke Wachstumsverstärkung“ vornehmen. Das sind Ordinaldaten, da ihnen eine Wertung zukommt. Sie können über Scoring quantifiziert werden.

g Bei der Untersuchung der Atmungsaktivität von Bakterien werden Gasvolumina an einem Warburg-Manometer ermittelt. Welche Datenform liegt hier vor?
Lösung: Es handelt sich um quasi-stetige Meßwerte, die an einer kontinuierlichen Skala dargestellt werden können.