3. Häufigkeitsverteilungen

Nach der Aufnahme von Stichprobendaten liegt eine Urliste vor, die die Werte in der Reihenfolge ihrer Ermittlung enthält. Sie wurden gesammelt um Fragen, die Merkmalsausprägungen in der Stichprobe betreffen, beantworten zu können. Solche Fragen sind z.B. :

- Welcher ist der kleinste und welcher der größte Wert einer Datenreihe?
- Welcher ist der Wert, der am häufigsten vorkommt?
- Wie verteilen sich die einzelnen Werte in der Datengruppe
- Gibt es viele kleine, viele mittlere oder viele große Werte oder sind alle Werte gleich häufig?
- Ist die Verteilung der Werte in einem Diagramm symmetrisch?

Solche Informationen über die Art der Verteilung sind von Bedeutung, weil viele an den Statistiker gerichtete Fragen mit Methoden der induktiven Statistik (siehe später) bearbeitet werden müssen und weil die Anwendbarkeit solcher Methoden oft von der Art der Verteilung der Daten abhängt.
Wir werden uns hier mit

- Gleichverteilungen
- Normalverteilungen
- Schiefen Verteilungen
- Bimodalen Verteilungen und
- Multimodalen Verteilungen

soweit beschäftigen, wie es zur Einführung in das Thema notwendig ist. In einem späteren Kapitel gehen wir auf bestimmte Verteilungen genauer ein.

Aus einer Urliste können wir die benötigten Informationen oft nicht direkt entnehmen, weil sie zu unübersichtlich ist. Daher führen wir eine sogenannte Verteilungsanalyse durch, deren Ziel es ist, die Daten übersichtlich so darzustellen, dass das Typische ihrer Verteilung erkennbar wird.

Im folgenden Text werden gelegentlich die Begriffe Mittelwert, arithmetisches Mittel, Modalwert und Medianwert verwendet. Auf deren Bedeutung und Berechnung gehen wir im nächsten Kapitel ein.

Beispiel 1
Im Zusammenhang mit einer Seminarvorbereitung haben wir einmal eine Stichprobe von 257 Walnüssen untersucht und dabei auch deren Gewichte ermittelt. Die quasistetigen, also gekörnten Werte wurden in der Urliste in g mit einer Nachkommastelle notiert. Da diese Urliste für die weiteren Überlegungen hier nicht notwendig ist, verzichten wir auf die Darstellung der 257 Werte. Wir können uns auch ohne die Liste zu sehen vorstellen, dass sie die die Messwerte in der Reihenfolge ihrer Notierung, also gänzlich ungeordnet enthält und damit ziemlich unübersichtlich ist.

Wenn wir im Sinne der oben gestellten Fragen nun wissen wollen, wie die einzelnen Messwerte über den gesamten Datenbereich verteilt sind, wie häufig also welche Werte vorkommen und welches die Extremwerte sind, dann ist die ungeordnete Urliste wenig hilfreich. Zur Gewinnung einer besseren Übersicht haben wir die Daten zunächst in steigender Folge geordnet und dann ausgezählt, wie oft jeder Messwerte vorkommt. Das Ergebnis zeigt die folgende Liste, in der jedem Messwert seine Häufigkeit (H) zugeordnet ist.

Aufsteigend geordnete Liste der Walnussgewichte in Gramm (H = absolute Häufigkeit)

Gramm	9	9.2	9.4	9.5	9.7	9.8	9.9	10	10.2	10.3	10.4	10.5
H	2	2	3	5	2	5	2	8	5	6	2	9
Gramm	10.6	410.7	10.8	10.9	11	11.1	11.2	11.3	11.4	11.6	11.7	11.8
H	2	5	12	11	9	12	6	6	8	12	8	6
Gramm	11.9	12	12.1	12.2	12.3	12.4	12.5	12.6	12.7	12.8	12.9	13
H	8	6	6	6	9	2	8	5	6	6	3	8
Gramm	13.1	13.2	13.4	13.5	13.6	13.7	13.8	14	14.1	14.4	14.5	14.8
H	3	5	2	2	2	3	5	3	2	3	2	2
Gramm	15											
H	2											

Tabelle 1

Diese Liste ist schon übersichtlicher. Kleinster Wert (x_{min} = 9,0 g) und größter Wert (x_{max} = 15,0 g) und damit der Datenbereich [Spannweite oder Variationsbreite] (x_{min} = 9,0 g bis x_{max} = 15,0 g) sind sofort erkennbar. Zur Visualisierung stellen wir die Häufigkeiten gegen die Messwerte in einem Säulendiagramm graphisch dar. (Abb.1)

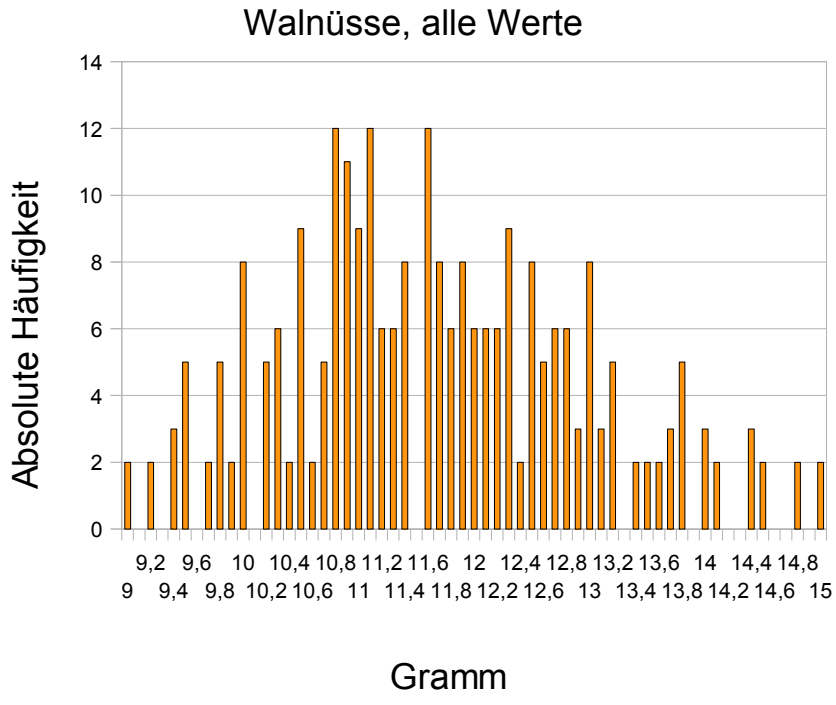

Abb. 1

Die Bewertung einer solchen Graphik ist, wie wir noch sehen werden, relativ subjektiv. Sie zeigt zwar übersichtlicher als die geordnete Liste, dass mittlere Werte häufiger vorkommen als Extremwerte, allerdings erscheint die Graphik mit ihren abwechselnd kleineren und größeren Säulen „unruhig“. Um viele Messwerte graphisch so darzustellen, dass das Charakteristische ihrer Verteilung in einer ansprechenden Form erkennbar wird, fasst man die Messwerte zu Klassen zusammen. Dabei wird dann nicht mehr dargestellt, wie häufig jeder Messwert vorkommt, sondern, wie häufig Messwerte in einer Klasse vorkommen. Dadurch tritt zwar ein Informationsverlust auf (anhand der Klassen wissen wir nicht mehr, wie oft jeder Messwert vorkam) die Darstellung wird aber übersichtlicher.

3.1 Klassenbildung

Klassierung, Kategorisierung
Die Klassenbildung ein wichtiges Instrument bei der Verteilungsanalyse. Wir wollen daher dieses Verfahren in den einzelnen Schritten darstellen. Die 257 Messwerte der Liste bestehen aus 49 verschiedenen, gekörnten Werten, die unterschiedlich häufig vorkommen (2 mal bis 12 mal). Hätten wir die Gewichte der Nüsse in g mit drei oder mehr Nachkommastellen ermittelt, so wäre die Körnung weniger stark ausgeprägt und wir hätten vermutlich statt 49 257 verschiedene Werte erhalten. Im theoretischen Idealfall stetiger Daten (also Zahlen mit unendlich vielen Nachkommastellen) ist es auch bei viel größeren Stichproben sehr wahrscheinlich, dass jeder Wert nur einmal vorkommt. Von einer solchen Datensammlung könnte man keine Analyse der Häufigkeit durchführen, da ja jeder Wert nur einmal auftritt. Da es einerseits aufgrund der Messtechnik praktisch unmöglich ist, wirklich stetige Messwerte zu erhalten, andererseits aber auch unnötig, denn so genaue Werte werden nicht benötigt, ist die Ablesung bzw. Notierung eines Messwertes grundsätzlich schon mit einer Körnung verbunden. Und das bedeutet mit einer primären Klassenbildung. Im Beispiel 1 gab es bei der Versuchsplanung die Vereinbarung, alle Messergebnisse zwischen 8,95 g und < 9,05 g zusammenzufassen zu dem „Messwert“ (Klassenmitte) 9,0 g. Alle Werte zwischen 9,05 g und < 9,15 g zu der Klassenmitte 9,1 g usw. Da das Ergebnis, die Abb.1, die Verteilung der Daten aber noch nicht zufriedenstellend darstellt, sie erscheint uns ja zu „unruhig“, werden wir die 257 Messwerte nun der eigentlichen (einer sekundären) Klassenbildung unterwerfen. Dazu könnten wir etwa alle Werte zwischen 9,0 g und < 10,0 g zu einer Klasse zusammenfassen. Alle Werte dieser Klasse würden dann durch den Mittelwert der Klasse, die Klassenmitte, also 9,5 g repräsentiert.

Im Vorfeld einer Klassenbildung treten dabei folgende Fragen auf:

1. Wie viele Klassen sollen gebildet werden?
2. Wie breit sollen die Klassen sein?
3. Sollen alle Klassen gleich breit sein?
4. Wie sind die Klassengrenzen und Klassenmitten zu bilden?

Anzahl (k) und Breite (b) der Klassen
In der Literatur werden verschiedene Empfehlungen zur Anzahl der Klassen angegeben.

DIN 55302 u.a.	$5 \leq k \leq 20$ oder $k \approx \sqrt[2]{n}$
Andere Stellen	für $n \leq 1000$ gilt $k \approx \sqrt[2]{n}$ für $n > 1000$ gilt $k \approx 10 * \log n$

Daneben gibt eine Reihe spezieller Empfehlungen. Alle diese Angaben sind unverbindlich und wir haben in der Praxis etwas Spielraum bei der Entscheidung. Wir können aber davon ausgehen, dass k mit n wächst. Häufig haben sich 5 – 15 Klassen als zweckmäßig erwiesen. Wenngleich es Empfehlungen gibt, so wird die Anzahl der Klassen oft nach subjektiven Gesichtspunkten gewählt, wobei wir uns etwa an bekannten Beispielen orientieren. Ziel ist es, so viele Klassen zu bilden, dass die Daten übersichtlich so dargestellt werden können, dass das Typische der Verteilung visuell „gut“ herauskommt und wir damit die gewünschten Informationen über die Verteilung erhalten. Dieses Vorgehen ist wegen des Adjektivs „gut“ subjektiv. Wählen wir zu wenig Klassen, so tritt der Verteilungstyp der Daten eventuell nicht klar genug hervor und es gehen viele Informationen verloren. Wählen wir zu viele Klassen, so wird die Darstellung zu unübersichtlich.

Je weniger Klassen gebildet werden,
um so größer ist der Informationsverlust.

Um zu demonstrieren, wie sich die Anzahl der Klassen visuell auswirkt, haben wir die Werte der 257 Nussgewichte viermal unterschiedlich klassiert. Das Ergebnis zeigen im Vergleich mit der schon bekannten Abb.1 die folgenden Säulendiagramme in den Abb. 2 bis 5.

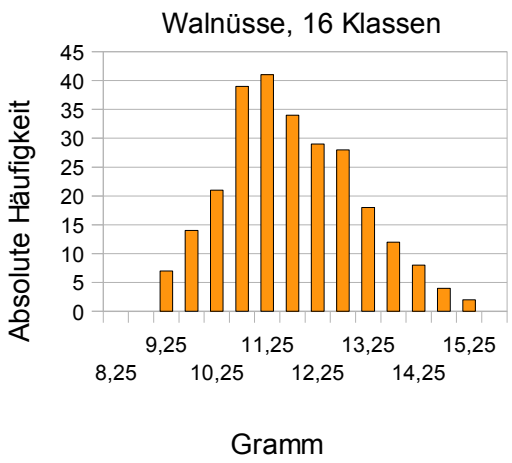

Abb. 2

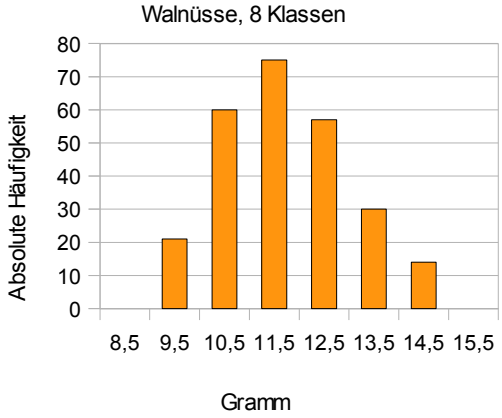

Abb. 3

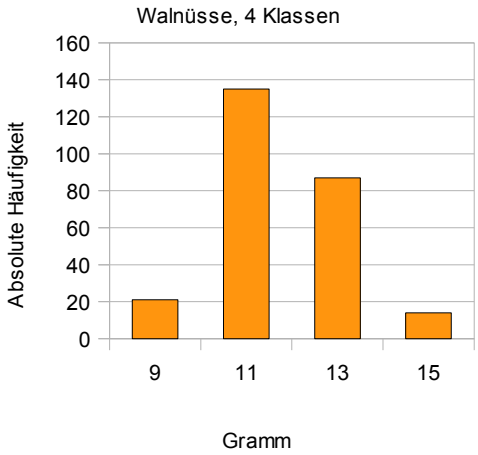

Abb. 4

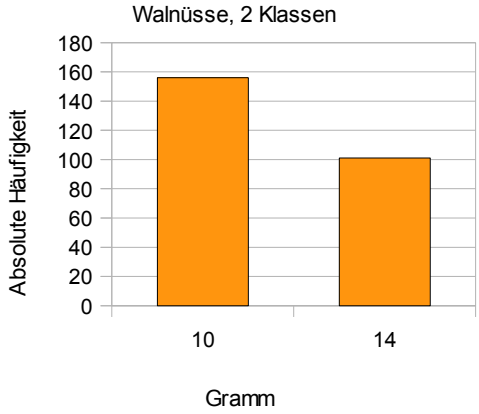

Abb. 5

Alle fünf Diagramme zeigen, dass der Bereich mit der größten Häufigkeit etwas aus der Mitte nach links verschoben ist. Diese Eigenschaft der Verteilung erkennen wir zwar in allen Darstellungen aber die Abb. 1 erscheint uns subjektiv zu differenziert, die Abb. 4 und 5 dagegen zu undifferenziert. Abb. 2 und 3 stellen die Verteilung recht ansprechend dar.

Wie viele Klassen sollen wir im konkreten Beispiel nun aber bilden um eine aussagekräftige Graphik zu erhalten? Aus $k \approx \sqrt[2]{257} \approx 16$ folgen 16 Klassen. Wir könnten uns nun auch für 12 oder 14 Klassen entscheiden, die 16 ist ja nicht verbindlich. Aber was die Klassengrenzen und Klassenmitten angeht, könnten da Probleme auftreten. Sie können es ja mal versuchen. Wir wollen nun sehen, was wir tun müssen, um diese 16 Klassen zu bilden.

Klassengrenzen

Aus der Liste der geordneten Daten lesen wir zunächst die Variationsbreite (Range R) der Messwerte ab.

$$R = x_{min} - x_{max} = 9{,}0\ g - 15{,}0\ g = |6\ g|$$

Die Klassengrenzen müssen so festgelegt werden, dass alle Messwerte eingeordnet werden können. Dabei entspricht der untere Wert der Variationsbreite (hier 9,0) der unteren Grenze der ersten Klasse und der obere Wert der Variationsbreite (hier 15,0) dem oberen Grenzwert der letzten Klasse. Da die Klassenbreite b nach b = R/k berechnet wird, ergibt sich für R = 6 und k = 16 eine Klassenbreite von 0,375 g. Für die Übersichtlichkeit ist es günstiger, als Klassenbreite einen „glatteren“ Wert, also eine ganze Zahl oder 0,5; 1,5 usw. zu haben. Wir müssen nun ein wenig überlegen. Die Differenz der beiden Grenzwerte (Extremwerte) beträgt 6 g. Wäre die Differenz 8 g, so ergäbe das eine Klassenbreite von b = R/k = 8/16 = 0,5 und das wäre besser als 0,375. Um diese Differenz 8 zu

erreichen, müssen wir, was erlaubt ist, die Grenzwerte vertretbar ändern. Wir legen als Grenzen des Variationsbereichs (und damit die Außengrenzen der beiden Randklassen) jetzt 8,0 g statt 9,0 g und 16,0 g statt 15,0 g fest. Die Differenz ist jetzt 8 g. Somit haben wir 16 Klassen mit der Klassenbreite von 0,5 g. Das geht nicht immer auf den ersten Blick so schnell. Gegebenenfalls müssen wir dazu schon ein wenig experimentieren. In den meisten Fällen, wie auch hier, sind alle Klassen nach b = R/k gleich breit (äquidistante Klassenbreite) und alle Klassen sind in dem Sinne geschlossene Klassen, als dass sie eine untere und eine obere Begrenzung haben. Es gibt aber auch Situationen, in denen es zweckmäßig ist, unterschiedlich breite Klassen zu bilden. Und in manchen Fällen ist es notwendig, dass die unterste Klasse keine Untergrenze und die oberste Klasse keine Obergrenze hat (offene Randklassen, siehe später).

Die Klassengrenzen müssen nun so festgelegt werden, dass jeder Messwert eindeutig einer Klasse zugeordnet werden kann. (Siehe Tabelle weiter unten.)

Falsch wäre folgendes Vorgehen:	Klasse 5	10,0 g	bis	10,5 g
	Klasse 6	10,5 g	bis	11,0 g
	Klasse 7	11,0 g	bis	11,5 g usw.

In diesem Falle wären die Messwerte 10,5 und 11,0 nicht eindeutig zuzuordnen. 10,5 könnten wir sowohl der Klasse 5 als auch der Klasse 6, die 11,0 der Klasse 6 als auch der Klasse 7 zuordnen.

Richtig ist es, so vorzugehen:	Klasse 5	10,0 g	bis	< 10,5 g
	Klasse 6	10,5 g	bis	< 11,0 g
	Klasse 7	11,0 g	bis	< 11,5 g usw.

Klassenmitten

Bei der Berechnung der Klassenmitten tritt das Problem auf, dass das arithmetische Mittel aus unterer und oberer Klassengrenze nicht bestimmbar ist, weil der oberen Klassengrenze mit < x keine Zahl zuzuordnen ist mit der wir den Mittelwert bilden könnten. Wir ermitteln die Klassenmitte daher nach:

$$\text{Klassenmitte} = \frac{\text{Summe der benachbarten unteren Klassengrenzen}}{2}$$

Für Klasse 1 liegt die Mitte also nach (8,0 + 8,5)/2 = 8,25 bei 8,25 g. Da es für die spätere Auswertung nach der Klassierung oft günstig ist, als Klassenmitte eine „glatte“ Zahl zu haben, sollten wir das bei der Festlegung der Klassengrenzen beachten. Manchmal ergeben sich dann „krumme“ Klassengrenzen. Es kann aber allerdings auch wünschenswert sein, „glatte“ Zahlen als Klassengrenzen zu haben. Gegebenenfalls lässt es sich auch so einrichten, dass sowohl die Klassengrenzen als auch die Klassenmitten ganzzahlig oder zumindest „glatt“ sind. Dazu müssen wir eventuell ein wenig überlegen und die Grenzen des Variationsbereichs den Zielen sinnvoll und vertretbar anpassen. So ergibt sich für unser Beispiel 1 folgende Tabelle, in die wir schon die absoluten Häufigkeiten für die 16 Klassen eingetragen haben.

Die absoluten Häufigkeiten werden jetzt nicht mehr den einzelnen Messwerten, sondern den Klassenmitten zugeordnet. Die Kenntnisse über die Verteilung der einzelnen Werte in einer Klasse gehen dabei verloren (daher: Urliste aufbewahren!)

Eine Klassenbildung ist bei einem Gewinn an Übersicht immer mit einem Informationsverlust verbunden.

Klasse	untere Klassengrenze	Klassenmitte	obere Klassengrenze	absolute Häufigkeit H
1	8.0	8.25	< 8,5	0
2	8.5	8.75	< 9,0	0
3	9.0	9.25	< 9,5	7
4	9.5	9.75	< 10,0	14
5	10.0	10.25	< 10,5	21
6	10.5	10.75	< 11,0	39
7	11.0	11.25	< 11,5	41
8	11.5	11.75	< 12,0	34
9	12.0	12.25	< 12,5	29
10	12.5	12.75	< 13,0	28
11	13.0	13.25	< 13,5	18
12	13.5	13.75	< 14,0	12
13	14.0	14.25	< 14,5	8
14	14.5	14.75	< 15,0	4
15	15.0	15.25	< 15,5	2
16	15.5	15.75	< 16,0	0

Tabelle 2

Die Abb. 2 zeigt die Verteilung nach der vorliegenden Liste.

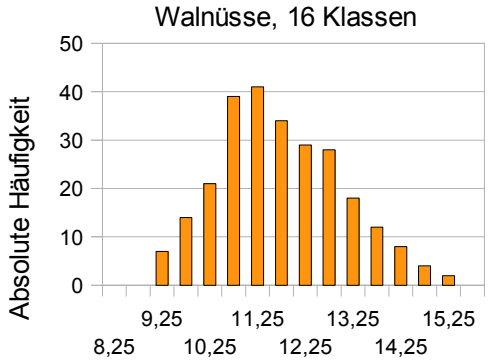

Abb.2

Was können wir jetzt über die Verteilung der 257 Messwerte aus der Graphik mit 16 Klassen ablesen?

1. Extremwerte sind relativ selten. (wie in Abb. 1)
2. Mittlere Werte sind relativ häufig. (wie in Abb. 1)
3. Die Verteilung ist asymmetrisch. Die Klasse mit der größten Häufigkeit (Modus), die modale Klasse mit der Klassenmitte 11,25 g, liegt nicht in der Mitte der Abszisse sondern ist nach links verschoben.
4. Die Häufigkeiten der Klassen steigen bis zum Modus monoton und fallen dann monoton. Demnach haben wir eine monomodale Verteilung, d.h., nur einen Modus. (In Abb.1 hatten wir aber drei Modi.)

Zu der Frage, warum der Modus nach links verschoben ist, kann die Verteilungsanalyse keine Aussagen machen.

Abhängig von den vorliegenden Daten können Verteilungskurven so aussehen wie in Beispiel 1 oder auch ganz anders wie wir an den folgende Verteilungen sehen werden.

3.2 Gleichverteilung

Beispiel 2
Die Gleichverteilung ist eine Verteilung bei der die Häufigkeit des Vorkommens für alle Werte gleich ist. Wenn wir im Gedankenexperiment mit einem fairen 10-seitigen Würfel (0,1,2,3,4,5,6,7,8,9) z.B. 300 Würfe durchführen, dann würden wir (wenn wir uns mit solchen Themen noch nie beschäftigt hätten) erwarten, dass die Ziffern 0 bis 9 mit der gleichen Häufigkeit auftreten. Denn jede Ziffer hat die gleiche Chance, die gleiche Wahrscheinlichkeit, nach einem Wurf oben zu liegen. Wir erwarten also (theoretisch) folgende Häufigkeitsverteilung, wie sie in Abb. 6 dargestellt ist.

Ziffer	0	1	2	3	4	5	6	7	8	9
Absolute Häufigkeit	30	30	30	30	30	30	30	30	30	30

Tabelle 3

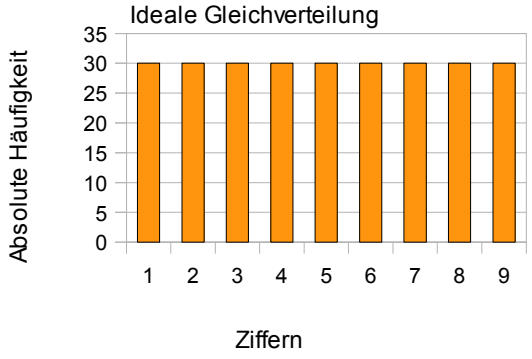

Abb. 6

Soweit die Theorie. Die folgenden Daten zeigen das Ergebnis der empirisch ermittelten Werte nach 300 realen Würfen. Das sieht schon nicht mehr so gleichverteilt aus.

Ziffer	0	1	2	3	4	5	6	7	8	9
Absolute Häufigkeit	26	27	35	26	27	30	27	32	39	31

Tabelle 4

Mit etwas mehr Geduld und einigen Tausend Würfen würden wir dem Erwartungswert für die gleiche Häufigkeit für alle Ziffern sicher etwas näher kommen. 300 Würfe sind offensichtlich zu wenig, um den

Erwartungswert 30 für jede Ziffer zu erreichen. Wie nahe wir einer idealen Gleichverteilung mit den 300 empirischen Werten hier gekommen sind, zeigt die Graphik, in der die grünen Säulen die reale Verteilung zeigen.

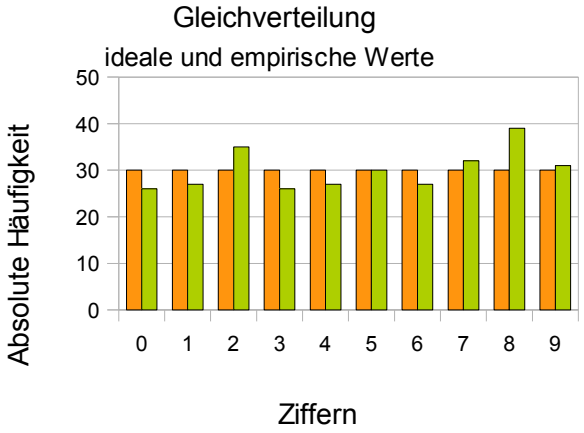

Abb. 7

Gleichverteilungen sind bei hinreichend großen Stichproben dann zu erwarten, wenn für jeden realisierbaren Wert die Wahrscheinlichkeit für sein Auftreten gleich ist. Dies ist in der Regel bei biologischen Daten (biologische Variabilität) nicht gegeben. Ob in einem konkreten Fall die gefundene Abweichung von der idealen Gleichverteilung wie in Beispiel 2 in dem Sinne von Bedeutung ist, als dass etwa der Würfel nicht in Ordnung war, oder (bei einem fairen Würfel) eher zufällig, das können wir durch Betrachten der Graphik nicht entscheiden. Wenn wir es wissen wollen, dann müssen wir die Daten mit Verfahren der induktiven Statistik untersuchen. Damit beschäftigen wir uns später.

3.3 Monomodale Verteilungen

3.3.1 Normalverteilung

Im 19. Jh. stellte der belgische Mathematiker A. Quetelet Untersuchungen zu Körpermaßen beim Menschen an. Er fand bei der Messung des Thoraxumfanges von 7538 schottischen Soldaten, dass Extremwerte selten und mittlere Werte am häufigsten vorkamen. Diese Ergebnisse sowie Untersuchungen zu anderen Themen von A. de Moivre (18.Jh., Glücksspiele, Wahrscheinlichkeitsrechnungen), P.S.Laplace (19.Jh., Messfehler) sowie C.F.Gauß (19.Jh, Normalverteilungsfunktion) führten zu dem, was wir heute unter der

Normalverteilungskurve,
de Moivre-Kurve,
Gauß-Kurve oder
Glockenkurve

verstehen.

Die Normalverteilung, die durch diesen Kurventyp dargestellt wird, ist eine der wichtigsten Verteilungen für Testverfahren in der induktiven Statistik.

Beispiel 3

Eine Reihe biologischer Daten gehören diesem Verteilungstyp an, den wir am Beispiel von 1050 Hühnereiern, die über 12 Monate von 4 Hühnern gesammelt und gewogen wurden, darstellen.

Die für dieses Kapitel wichtigen, typischen Eigenschaften der idealen Normalverteilung sind

1. Die ideale Normalverteilungskurve (blau) ist völlig symmetrisch um den Wert mit der größten Häufigkeit (Modalwert, Modus). Wir haben diese Kurve für diese Darstellung berechnet. (Dazu später mehr.)
2. Das Lot vom höchsten Punkt der Kurve zeigt auf der Abszisse arithmetisches Mittel, Modalwert und Medianwert an. Diese drei Werte sind bei der Normalverteilung identisch.
3. Untere und obere Extremwerte kommen sehr selten und gleich häufig vor.

Auf weitere bedeutende Eigenschaften der Normalverteilung gehen wir in einem späteren Kapitel ein.

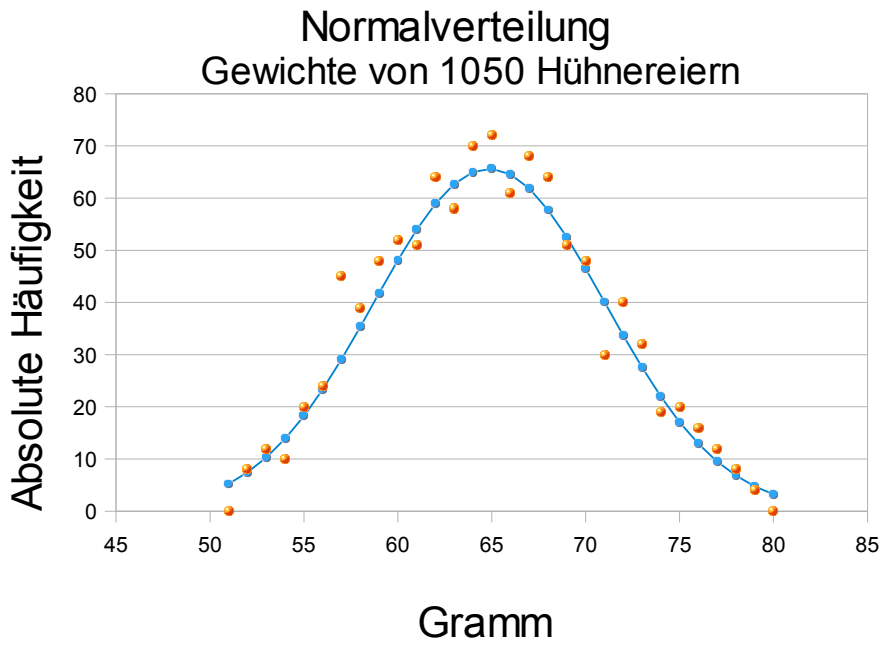

Abb. 8

Daten, die diesem Verteilungstyp entsprechen, nennen wir „normalverteilt“ oder „de Moivreverteilt“. Für die die Gewichte der Hühnereier (orange Punkte) trifft das nach der Graphik genähert zu. Wir können als aus der Graphik den Schluss ziehen, dass die Gewichte der Hühnereier genähert normalverteilt sind. Da es sich bei der Normalverteilung um eine Modellvorstellung handelt, werden wir in der Realität mit empirischen Werten eine solche Idealkurve nicht erreichen, sondern, wie die Eierdaten zeigen, nur mehr oder weniger nahe an sie herankommen. Eine wesentlich größere Stichprobe würde sehr wahrscheinlich eine „glattere“ Glockenkurve ergeben. Wir haben hier das gleiche Problem wie bei der Gleichverteilung. Beides sind theoretische Verteilungen, denen wir uns mit empirischen Daten nur nähern können.

Bei der Analyse vieler quantitativer physiologischer Daten finden wir mehr oder weniger starke Abweichungen von der idealen Glockenkurve. Wir ordnen diese Abweichungen den folgenden beiden Gruppen zu.

1. Exzessive Verteilungen. Sie sind dadurch gekennzeichnet, dass die Höhe des Modalwertes auf der Ordinate angehoben (positiver Exzess) oder abgesunken (negativer Exzess) ist. Auf exzessive Veränderungen werden wir hier nicht eingehen.

2. Schiefe Verteilungen. Hier ist der Modalwert auf der Abszisse nach links (wie in Beispiel 1) oder nach rechts verschoben. Die drei Mittelwerte sind nicht mehr identisch.

Eine Reihe von Testverfahren der induktiven Statistik setzt voraus, dass die zu bearbeitenden Daten normalverteilt sind. Wenn dies nicht gegeben ist, etwa weil der Modalwert nach links oder rechts verschoben ist, dann darf man diese Tests nicht anwenden. In einem solchen Fall kann man durch Transformieren der Daten mit geeigneten Algorithmen erreichen, dass die so bearbeiteten Daten einer Normalverteilung entsprechen. Ist der Modalwert nach links verschoben, dann erreicht man diesen Effekt oft durch Logarithmieren der Daten. Der linke Teil einer logarithmischen Skala ist weiter gespreizt als der rechte Teil. Dadurch werden die nahe beieinanderliegenden Daten im linken Teil der Graphik gespreizt. Die Graphik nähert sich dadurch eine symmetrischen Glockenkurve. Wir sprechen dann von einer logarithmischen Normalverteilung. Auch wenn der Modus nach rechts verschoben ist, können durch geeignete Transformationen die Daten u.U. „normalisiert“ werden. Auf solche Verfahren werden wir in einem späteren Kapitel eingehen.

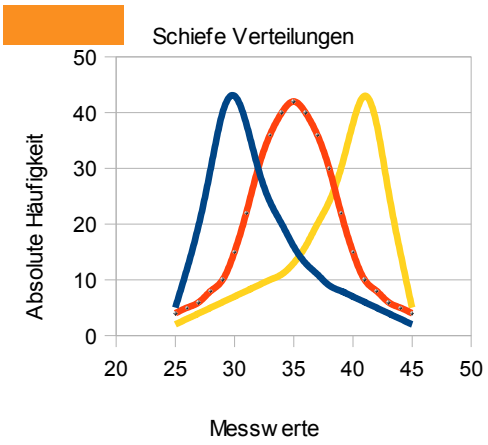

Abb. 9

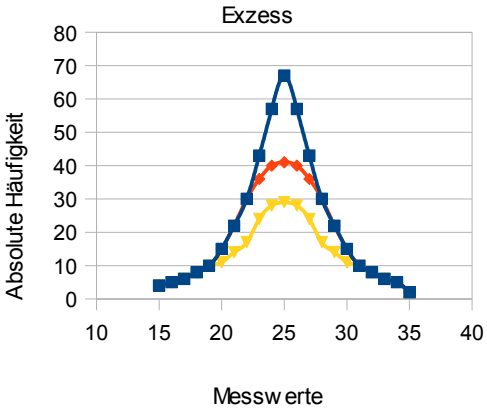

Abb. 10

3.3.2 Schiefe Verteilungen

3.3.2.1 Linksgipfelige Verteilung (linkssteil, positiv schief)

Beispiel 4
Eine Linksverschiebung des Modalwertes tritt bei vielen biologischen Daten auf. Aus einer retrospektiven Erhebung haben wir die klassierten Körpergewichte von 176 Männern im Diagramm dargestellt. Wir sehen, dass diese 176 Werte zwar auch so verteilt sind, dass extreme Werte selten vorkommen, der Modalwert ist aber deutlich nach links verschoben wie in Beispiel 1.

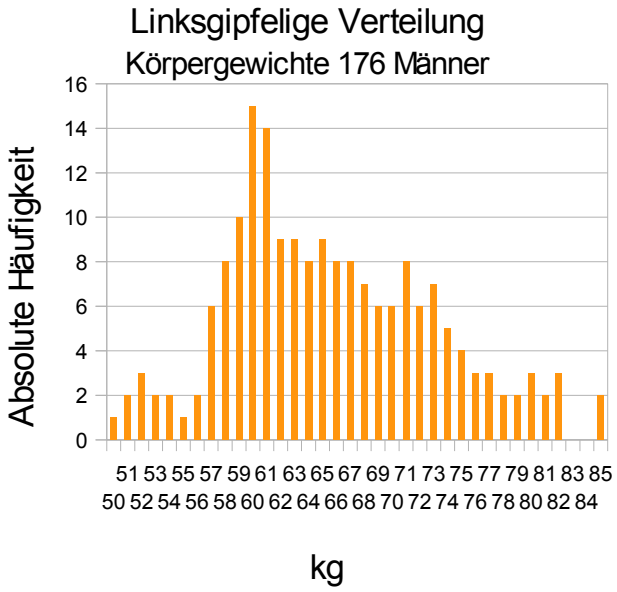

Abb. 11

In dieser Art sind auch andere physiologische Daten, wie Pulsfrequenz, die pharmakologische Wirkung mancher Wirkstoffe und, wie wir in Beispiel 5 sehen, die systolischen Blutdruckwerte beim Menschen verteilt.

Beispiel 5
Eine Analyse der über drei Monate täglich gemessenen systolischen Blutdruckwerte eines männlichen Probanden ergab nach Klassierung folgende Verteilungskurve Abb.12.
Die Kurve ist asymmetrisch linksgipfelig. Als Grund für die Linksgipfeligkeit wird oft angegeben, dass die Abszissenskala links letztlich durch Null begrenzt ist (negative Werte können ja nicht auftreten), während nach rechts die Variationsmöglichkeiten – begrenzt durch physiologische Vorgaben – offen

ist. Einer entsprechenden Senkung der Werte in Richtung Null (die Skala beginnt hier ja erst mit 106 mm Hg) steht entgegen, dass Werte unter ca. 60 mm Hg mit dem Leben nicht vereinbar sind. Diesen Verteilungstyp nennen wir mit Bezug darauf, dass bei weit links liegendem Modus dessen Säule dem Abstrich eines ⌞ entspricht auch ⌞-Kurve.

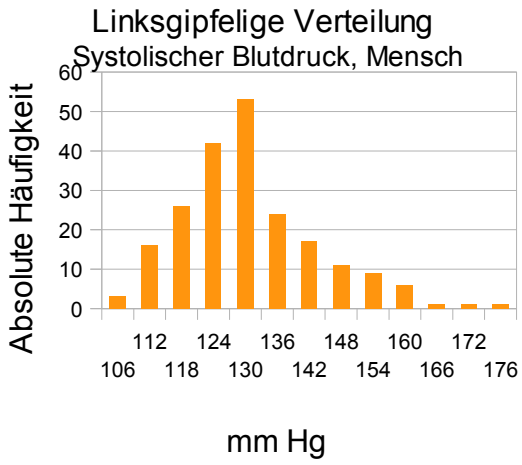

Abb. 12

3.3.2.2 Rechtsgipfelige Verteilung (rechtssteil, negativ schief)

Beispiel 6
Rechtsgipfelige Verteilungen finden wir im biologischen Bereich relativ selten, daher haben wir für dieses Beispiel fiktive Daten gewählt. Wie die Graphik zeigt, ist der Modus deutlich nach rechts verschoben. In Anlehnung an die Bezeichnung L-Kurve sprechen wir hier von einer J-Kurve.

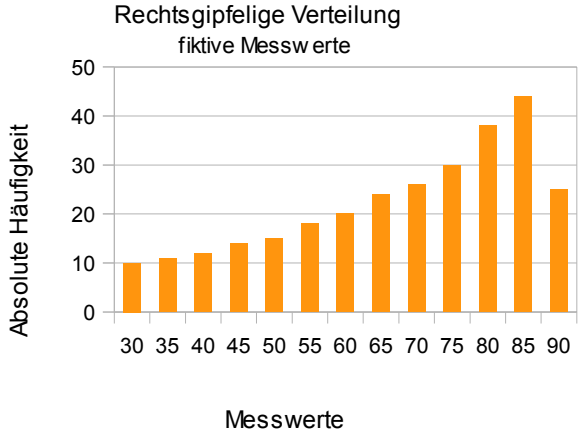

Abb. 13

Einem Literaturhinweis (Immich, Medizinische Statistik, 1974) folgend, haben wir 60 Personen (7 bis 60 Jahre) unabhängig voneinander gebeten, folgender Bitte zu entsprechen: „Bitte nennen Sie spontan eine Zahl zwischen Null und Zwanzig (inclusiv)“.Das Ergebnis lautet

Zahl	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H	1	0	0	4	1	1	1	9	1	2	2	3	4	5	1	4	1	8	10	2	0

Tabelle 5

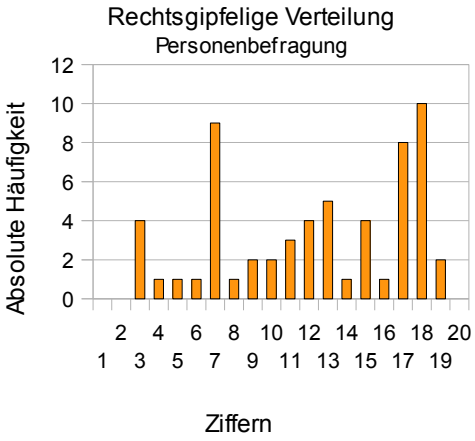

Abb.14

Wir haben hier zwar das Bild einer Verteilung mit mehreren Modalwerten. Die größte Häufigkeit liegt aber deutlich auf der rechten Seite. Aus Gründen, die die Psychologie klären müsste, entscheiden sich offensichtlich deutlich mehr Personen für die Nennung von Zahlen im oberen Bereich. Es gibt hier zwei „Ausrutscher“ wobei ich nur für die sieben eine mögliche Erklärung finde. Der Anteil der Befragten aus Kulturbereichen in denen die sieben eine Glückszahl ist, könnte zu dem „unpassend“ hohen Wert für diese Zahl geführt haben.

3.4 Bimodale Verteilung

Bei den bisherigen Verteilungen haben wir immer – bei sinnvoller Klassierung - nur einen Wert mit der größten Häufigkeit gefunden wobei Beispiel 6 unerwartet aus der Reihe fiel. Das Beispiel 7 zeigt eine Verteilung, bei der die Kurve deutlich zweigipfelig ist. Eine solche bimodale Verteilung, bei der die beiden Modi nicht den gleichen Wert haben müssen, weist auf ein heterogenes Datenmaterial hin.

Beispiel 7
Bei einer retrospektiven Erhebung fanden wir für eine Gruppe von 130 Personen im Alter von 30 – 70 Jahren folgende Verteilung des Körpergewichts.

H	4	6	16	13	12	10	9	15	14	13	8	5	4
kg	68	69	70	71	72	73	74	75	76	77	78	79	80

Tabelle 6

Der Grund für die beiden Modalwerte (70 kg und 75 kg): Die Gruppe bestand aus 70 Männern und 60

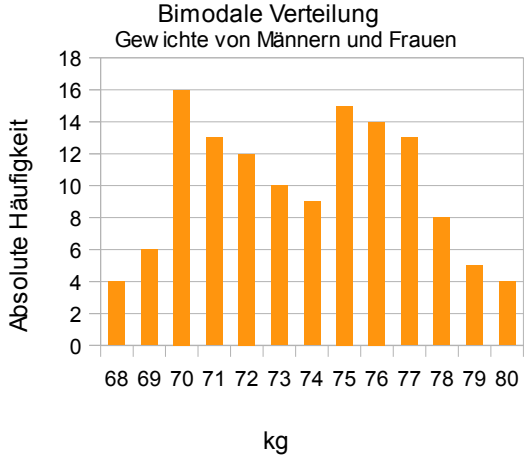

Abb. 15

Frauen, wobei letztere in der Regel ein geringeres Körpergewicht haben.

Die U-Verteilung

Beispiel 8
Wenn bei einer bimodalen Verteilung die beiden Modi ganz nach links bzw. rechts rücken, dann geht die Verteilung in eine U-Verteilung über. Beispiele dafür finden wir bei der Befragung zu spektakulären Vorstellungen. Auf die Frage, ob Menschen je zu fernen Sternen reisen würden, gibt Swoboda, [Knaurs Buch der modernen Statistik, 1971] an, dass die beiden Modi bei „Ja“ und „Nein“ liegen. Undifferenziertere Meinungen in verschiedenen Abstufungen werden seltener genannt.

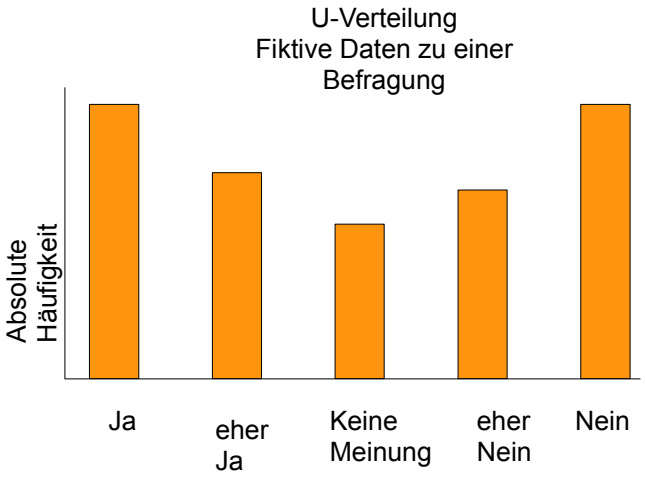

Abb. 16

Immer wenn Fragen polarisierende Meinungen zulassen, kann es zu U-Verteilungen kommen. Anders als bei den bisherigen Beispielen haben wir hier keine Messwerte sondern Ordinaldaten, deren Häufigkeit des Auftretens ausgezählt wurden.

3.5 Multimodale Verteilungen

Zum Schluß wollen wir noch auf die Verteilung von Daten hinweisen, die bei manchen Untersuchungen im Bereich der biologischen und klinischen Chemie auftreten, nämlich solche mit mehreren Modi.

Beispiel 9
Das Blutserum enthält eine Vielzahl von Proteinen, die für diagnostische Zwecke z.B. in Albumine, α1-Globuline, α2-Globuline, ß-Globuline und γ-Globuline eingeteilt werden können. In der Celluloseacetatfolien-Elektrophorese wandern diese Proteine umgekehrt proportional zu ihren molaren Massen unter-schiedlich weit. Nach der densitometrischen Auswer-tung eines Elektropherogramms erhielten wir für ein Rattenserum folgende Kurve.

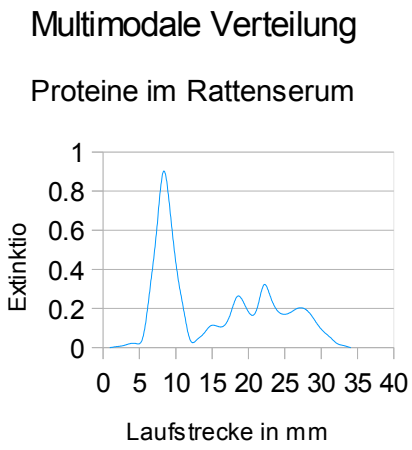

Abb.17

Da Maxima einer Kurve zwischen einem monoton ansteigenden und einem monoton abfallenden Kurventeil liegen, finden wir hier fünf Maxima, also fünf Modalwerte, nämlich, mit steigenden molaren Massen, bei 8 mm (Albumine), 15 mm (α1-Globuline), 18 mm (α2-Globuline), 22 mm (β-Globuline) und 27 mm (γ-Globuline).